Identifying Regions of Trusted Predictions Supplementary Material

Nivasini Ananthakrishnan[1] **Shai Ben-David**[1, 2] **Tosca Lechner**[1] **Ruth Urner**[3]

[1]David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, ON, N2L 3G1, Canada
[2]Vector Institute, Toronto, ON, M5G 1M1, Canada
[3]Lassonde School of Engineering, EECS Department, York University, Toronto, ON, M3J 1P3, Canada

A PROOFS

A.1 USEFUL LEMMAS

Lemma 1 (Hoeffding's inequality for general bounded random variables). *Let $X_1, \dots, X_N$ be independent random variables. Assume that $X_i \in [m_i, M_i]$ for every i. Then, for any $t > 0$, we have*

$$\Pr\left[\sum_{i=1}^{N}(X_i - \mathbb{E}[X_i]) \geq t\right] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{N}(M_i - m_i)^2}\right).$$

Lemma 2 (Bretagnolle-Huber inequality). *Let P and Q be probability measures on the same measurable space $(\Omega, \mathcal{F})$, and let $A \in \mathcal{F}$ be an arbitrary event. Then, $P(A) + Q(A^c) \geq \frac{1}{2}\exp(-KL(P,Q))$. Here, $KL(P,Q)$ is the KL-divergence between P and Q.*

Lemma 3. *Let P be a distribution over domain X. Let X' be a subset of X. Let S be an i.i.d. sample of size m drawn from the distribution P. Let $\hat{p}(X', S)$ be the fraction of the m samples that are in X'. For any $\delta > 0$, with probability $1 - \delta$ over the generation of the samples S,*

$$|P(X') - \hat{p}(X', S)| \leq w_p(m, \delta)$$

where

$$w_p(m, \delta) = \sqrt{\frac{1}{2m}\ln\frac{2}{\delta}}.$$

Proof. Let X_i be a random variable indicating if the i^{th} sample belongs to set X'. $X_i = 1$ if the i^{th} sample belongs to X' and zero otherwise. For each i, $\mathbb{E}[X_i] = P(X')$. $\hat{p}(X', S) = \frac{\sum_{i=1}^{N} X_i}{m}$. Applying Hoeffding's inequality, we get the inequality of the theorem. □

Lemma 4. *Let D be distribution over $X \times \{0, 1\}$. Let X' be a subset of X. Let S be an i.i.d. sample of size m drawn from D. Let $\hat{\ell}(X', S)$ be the fraction of the m labelled samples with label 1 in $S \cap X'$. For any $\delta > 0$, with probability $1 - \delta$ over the generation of the samples S, if $\hat{p}(X', S) - w_p(m, \delta/2) > 0$, then*

$$|\bar{\ell}_P(X') - \hat{\ell}(X', S)| < w_\ell(m, \delta, \hat{p}(X', S))$$
$$w_\ell(m, \delta, \hat{p}(X', S)) = \frac{1}{\hat{p}(X', S) - w_p(m, \delta/2)} \cdot \left(w_p(m, \delta/2) + \sqrt{\frac{1}{2m}\ln\frac{4}{\delta}}\right),$$

where $\hat{p}(X', S)$ is the fraction of the samples from S in X' that have label 1, $w_p(m, \delta/2))$ is as defined in Lemma 3.

Proof of Lemma 4. Let X_i be a random variable such that

$$X_i = \begin{cases} 1 & \text{If } i^{\text{th}} \text{ sample belongs to the set } X' \text{ and has label one.} \\ 0 & \text{otherwise.} \end{cases}$$

$\mathbb{E}[X_i] = P(X')\bar{\ell}_P(X')$, for each i. $\sum_{i=1}^{m} X_i = m\hat{p}\hat{\ell}_P(X', S)$. Note that by triangle inequality,

$$\begin{aligned} &|P(X')\hat{\ell}_P(X', S) - P(X')\bar{\ell}_P(X')| \\ &\leq |\hat{p}\hat{\ell}_P(X', S) - P(X')\bar{\ell}_P(X')| + |\hat{p} - P(X')|\hat{\ell}_P(X', S) \\ &\leq |\hat{p}\hat{\ell}_P(X', S) - P(X')\bar{\ell}_P(X')| + w_p. \end{aligned}$$

For any $\epsilon > 0$,

$$\begin{aligned} &\Pr[|\bar{\ell}_P(X') - \hat{\ell}(X', S)| > \epsilon] \\ &= \Pr[P(X') \cdot |\bar{\ell}_P(X') - \hat{\ell}(X', S)| > P(X')\epsilon] \\ &\leq \Pr[|\hat{p}\hat{\ell}_P(X', S) - P(X')\bar{\ell}_P(X')| + w_p > (\hat{p} - w_p)\epsilon] \\ &= \Pr[|m\hat{p}\hat{\ell}(X', S) - mP(X')\bar{\ell}_P(X')| > m(\hat{p} - w_P)\epsilon - w_p] \\ &= \Pr[\sum_{i=1}^{m} |X_i - \mathbb{E}[X_i]| > m((\hat{p} - w_P)\epsilon - w_p)]] \\ &\leq 2\exp\left(-2m((\hat{p} - w_P)\epsilon - w_p)^2\right) && \text{(By Hoeffding's inequailty).} \end{aligned}$$

When $\hat{p} - w_p > 0$, choosing

$$w_l(m, \delta, \hat{p}) > \frac{w_p}{\hat{p} - w_p} + \frac{1}{\hat{p} - w_p}\sqrt{\frac{1}{2m}\ln\frac{4}{\delta}},$$

we get that with probability $1 - \delta$, $|\bar{\ell}_P(X') - \hat{\ell}(X', S)| < w_\ell(m, \delta, \hat{p})$. □

A.2 CONFIDENCE SCORES USING LIPSCHITZ ASSUMPTION

Proof of Theorem 1. The algorithm partitions the space into r^d cells. Let p_c be the probability weight of a cell c and let $\hat{p}_c$ be the estimate of p_c that is calculated based on a sample to be the fraction of sample points in the cell c. From Lemma 3 and a union bound, we know that with probability $1 - \frac{\delta}{2}$, for every cell c,

$$p_c \in [\hat{p}_c - w_p(c), \hat{p}_c + w_p(c)].$$

Here $w_p(c) = w_p(m, \delta/2r^d)$ (as defined in Lemma 3).

The algorithm also estimates the average label of a cell c - ℓ_c as $\hat{\ell}_c$. This is the fraction of the sample point in the cell that have the label one. This is the same as the labelling probability estimate defined in Lemma 4. When the true probability weights of cells lie within the calculated confidence interval, by Lemma 4, we know that with probability $1 - \frac{\delta}{2}$, for every cell c,

$$\hat{\ell}_c \in [\hat{\ell}_c - w_\ell(c), \hat{\ell}_c - w_\ell(c))].$$

Here $w_\ell(c) = w_\ell(m, \delta/2r^d, \hat{p}_c)$ (as defined in Lemma 4).

The maximum distance between any two points in any cell is $r\sqrt{2}$. By the λ-Lipschitz, any point in the cell has labelling probability within $\lambda r\sqrt{2}$ of the average labelling probability of the cell. Therefore, with probability $1 - \delta$, for each cell c, for every point x in the cell c, the labelling probability of x satisfies:

$$\ell_P(x) \in [\hat{\ell}_c - w_\ell(c) - \lambda r\sqrt{2}, \hat{\ell}_c + w_\ell(c) + \lambda r\sqrt{2}].$$

This is the interval returned by the algorithm. Now we lower bound true confidence based on the confidence interval of the labelling probability. For a point x, let $c(x)$ denote the cell containing the point.

$$\begin{aligned}C_P(x,0) &= \ell_P(x)\\ &\geq \hat{\ell}_{c(x)} - w_\ell(c(x)) - \lambda r\sqrt{2}\\ C_P(x,1) &= 1-\ell_P(x)\\ &\geq 1-\hat{\ell}_{c(x)} - w_\ell(c(x)) - \lambda r\sqrt{2}.\end{aligned}$$

□

Proof of Theorem 2. We choose the input to the algorithm to be $r = \frac{1}{m^{1/8d}}$. With probability $1-\frac{\delta}{2}$, for all cells with probability weight greater than $\gamma = \frac{1}{m^{1/4}}$, the length of the confidence interval of the labelling probability is less than

$$\begin{aligned}&\frac{\frac{1}{m^{1/2}}}{\frac{1}{m^{1/4}}+\frac{1}{m^{1/2}}} - \frac{1}{\frac{1}{m^{1/4}}-\frac{1}{m^{1/2}}}\sqrt{\frac{1}{2m}\ln\frac{4m^{1/8}}{\delta}} + \frac{\lambda\sqrt{2}}{m^{1/8}}\\ &\leq \frac{1}{m^{1/4}-1} + \frac{1}{m^{1/4}-1}\sqrt{\frac{1}{16}\ln\frac{4m}{\delta}} + \frac{\lambda\sqrt{2}}{m^{1/8}}.\end{aligned}$$

This quantity decreases with increase in m and converges to zero. Therefore, for every $\epsilon_c > 0$, there is $M_1(\epsilon_c, \delta)$ such that this interval is less than ϵ_c. When sample size is larger than $M_1(\epsilon_c, \delta)$, with probability $1-\frac{\delta}{2}$, the size of confidence intervals for labelling probabilities of cells with weights greater than $\gamma = \frac{1}{m^{1/4}}$, is smaller than ϵ_c.

The points for which we can't say anything about the interval lengths are points in cells with weight at most γ. The total weight of such points is at most $\gamma\frac{1}{r^d} = \frac{1}{m^{1/8}}$. For any $\epsilon_x > 0$, let $M_2(\epsilon_x)$ be such that $\frac{1}{M_2(\epsilon_x)^{1/8}} < \epsilon_x$.

Choosing a sample size M greater than $M_1(\epsilon_c, \delta)$ and $M_2(\epsilon_x)$, we get that

$$\Pr_{S\sim P^M}[w_\ell > \epsilon_c] < \epsilon_x.$$

□

Proof of Theorem 1. For any $m > 0$, let $\mathcal{D}_m = \frac{1}{2}\mathcal{N}(m,1)\times\{1\} + \frac{1}{2}\mathcal{N}(-m,1)\times\{0\}$. The labelling probability for a point x according to this distribution is

$$\begin{aligned}f_m(x) &:= \frac{\exp(-(x-m)^2)}{\exp(-(x-m)^2)+\exp(-(x+m)^2)}\\ &= \frac{1}{1+\exp(-2xm)}\end{aligned}$$

Consider $\mu(\epsilon, M) > \frac{1}{\sqrt{M}}\ln\frac{1-\epsilon}{\epsilon} + \frac{1}{\sqrt{M}}$. We will show that the ϵ labelling probability learning problem for the class $\mathcal{F}_{\mu(\epsilon,M)}$ can be reduced to the problem of finding, for each $\mathcal{D}_m \in \mathcal{F}_{\mu(\epsilon,M)}$, finding an m' s.t. $|m-m'| < \frac{1}{\sqrt{M}}$ with probability $\frac{2}{3}$, based on samples from D_m. This problem as sample complexity at most $\sqrt{M}$

For any $m \geq \mu$, for any m' s.t. $|m-m'| < \Delta(M) = \frac{1}{\sqrt{M}}$, we will show that for every $x \in \mathbb{R}$, $|f_m(x) - f_{m'}| < \epsilon$.

$$\begin{aligned}|f_m(x) - f_{m'}(x)| &= \frac{|\exp(-2xm) - \exp(-2xm')|}{(1+\exp(-2xm))(1+\exp(-2xm'))}\\ &\leq |\exp(-2xm) - \exp(-2xm')|\end{aligned}$$

For $x < \bar{x} = \frac{1}{2\Delta(M)}\ln\frac{1}{1-\epsilon}$,

$$\begin{aligned}|\exp(-2xm) - \exp(-2xm')| &\leq 1 - \exp(-2x|m-m'|)\\ &< 1-\exp(-2\bar{x}|m-m'|)\\ &\leq \epsilon\end{aligned}$$

For $x \geq \bar{x}$,

$$\begin{aligned}
|\exp(-2xm) - \exp(-2xm')| &< \exp\left(-2x \min\{m, m'\}\right) \\
&\leq \exp\left(-2\bar{x}(\mu(\epsilon, M) - \Delta)\right) \\
&\leq \exp\left(-2 \cdot \frac{\sqrt{M}}{2} \ln \frac{1}{1-\epsilon} \cdot \frac{1}{\sqrt{M}} \ln \frac{1-\epsilon}{\epsilon}\right) \\
&= \epsilon
\end{aligned}$$

□

A.3 FUNCTION CLASS WITH LOW APPROXIMATION ERROR

Proof of Theorem 3. Let, for any classifier h, $\mathcal{L}^{0/1}_{P,B}(h) = \mathbb{P}_{(X,Y)\sim P}[h(X) \neq Y, X \in B]$. Then the quantity we want to bound - $\mathcal{L}^{0/1}_{P|B}(h_H(S_l))$ can be expressed as $\mathcal{L}^{0/1}_{P|B}(h_{H(S_l)}) = \frac{\mathcal{L}^{0/1}_{P,B}(h_H(S_l))}{P(B)}$. We will show how to estimate an upper bound for the numerator - $\mathcal{L}^{0/1}_{P,B}(h_H(S_l))$ using S_l and a lower bound for the denominator - $P(B)$ using S_u.

Upper bound for $\mathcal{L}^{0/1}_{P,B}(h_H(S_l))$: By uniform convergence, with probability $1 - \frac{\delta}{2}$, $\mathcal{L}^{0/1}_{P,B}(h_H(S_l)) < \mathcal{L}^{0/1}_{S_l,B}(h_H(S_l)) + \epsilon_{UC}(|S_l|, \delta/2)$.

Lower bound for $P(B)$: By Lemma 3, the estimate of $P(B)$ based on unlabelled samples - $\frac{|S_u \cap B|}{|S_u|}$ satisfies, with probability $1 - \frac{\delta}{2}$, $P(B) > \frac{|S_u \cap B|}{|S_u|} - \sqrt{\frac{1}{2|S_u|} \ln \frac{4}{\delta}}$.

Therefore, with probability $1 - \delta$, the upper bound for $\mathcal{L}^{0/1}_{P|B}(h_{\mathcal{H}}(S_l))$ holds. □

Proof of Theorem 4. Let $h* = \text{argmin}_{h \in H} \mathcal{L}^{0/1}_P(h)$. With probability at least $1 - \frac{\delta}{4}$, by uniform convergence, $\mathcal{L}^{0/1}_{S_l}(h^*) \leq \mathcal{L}^{0/1}_{S_l}(h_H(S_l)) + 2\epsilon_{UC}(|S_l|, \delta/4)$. So, $h^* \in H_{2\epsilon_{UC}(|S_l|,\delta/4)}$.

Also by uniform convergence, with probability $1 - \frac{\delta}{2}$, for any $h \in H_{2\epsilon_{UC}(|S_l|,\delta/4)}$, $\Delta_P(h, h^*, B) \leq \Delta_{S_u}(h, h^*, B) + \epsilon_{UC}(|S_u|, \delta/4) \leq DC_{B,H}\left(S_l, S_u, 2\epsilon_{UC}(|S_l|, \delta/4)\right) + \epsilon_{UC}(|S_u|, \delta/4)$. In particular, the above inequality holds for $h_H(S_l)$.

$$\begin{aligned}
\mathcal{L}^{0/1}_{P,B}(h_H(S_l)) &\leq \mathcal{L}^{0/1}_{P,B}(h^*) + \Delta_P(h_H(S_l), h^*, B) \\
&\leq \epsilon_{\text{approx}} + DC_{B,H}\left(S_l, S_u, 2\epsilon_{UC}(|S_l|, \delta/4)\right) \\
&+ \epsilon_{UC}(|S_u|, \delta/4)
\end{aligned}$$

Using S_u to obtain a lower bound on $P(B)$ with Lemma 3, we obtain the upper bound on $\mathcal{L}^{0/1}_{P|B}(h_H(S_l)) = \frac{\mathcal{L}^{0/1}_{P,B}(h_H(S_l))}{P(B)}$ provided by the theorem. □

Proof of Lemma 1. Let c be a label coverage hypothesis having α-region-conditional validity relative to $\{B_1, B_2\}$.

$$\begin{aligned}
&\mathbb{P}_{(X,Y)\sim P}[Y \in c(X) | X \in B_1 \cup B_2] \\
&= \frac{\mathbb{P}_{(X,Y)\sim P}[Y \in c(X), X \in B_1] + \mathbb{P}_{(X,Y)\sim P}[Y \in c(X), X \in B_2]}{P_{\mathcal{X}}(B_1) + P_{\mathcal{X}}(B_2)} \\
&= \frac{\mathbb{P}_{(X,Y)\sim P}[Y \in c(X) | X \in B_1] \cdot P_{\mathcal{X}}(B_1)}{P_{\mathcal{X}}(B_1) + P_{\mathcal{X}}(B_2)} \\
&+ \frac{\mathbb{P}_{(X,Y)\sim P}[Y \in c(X) | X \in B_2] \cdot P_{\mathcal{X}}(B_2)}{P_{\mathcal{X}}(B_1) + P_{\mathcal{X}}(B_2)}
\end{aligned}$$

By α-region-conditional validity,

$$\leq \frac{\alpha \mathbb{P}_{(X,Y)\sim P}[X \in B_1] + \alpha \mathbb{P}_{(X,Y)\sim P}[X \in B_2]}{P_{\mathcal{X}}(B_1) + P_{\mathcal{X}}(B_2)} = \alpha$$

□

A.4 GENERATIVE MODELS

Proof of Theorem 5. For some distribution P with PDF $p(x,y) = \begin{cases} ap_1(x) & \text{, if } y = 1 \\ (1-a)p_0(x) & \text{, if } y = 0 \end{cases}$, we can write the corresponding CLF as $l_P(x) = \frac{ap_1(x)}{ap_1(x)+(1-a)p_0(x)}$. Assume we get an iid sample S from P of size m. Let S_1 denote the subset of S with label 1 and S_0 be the subset of S labelled 0. Furthermore let $m_1 = |S_1|$ and $m_0 = |S_0|$. Then let $\hat{a} = \frac{m_1}{m_1+m_0}$. According to our assumptions there exists a learner $\mathcal{A}$ of $\mathcal{F}$ with sample complexity $m_{\mathcal{F}}$. Let $\hat{p}_1 = \mathcal{A}(S_1)$ and $\hat{p}_0 = \mathcal{A}(S_0)$. We then define our CLF-learner by $\mathcal{A}'(S)(x) = \frac{\hat{a}\hat{p}_1(x)}{\hat{a}\hat{p}_1(x)+(1-\hat{a})\hat{p}_0(x)}$.

$$\begin{aligned}
&\mathbb{E}_{X\sim P_X}[|l_P(X) - \mathcal{A}'(S)(X)|] \\
&= \int |l_P(x) - \mathcal{A}'(S)(x)|p(x)dx \\
&= \int |\frac{ap_1(x)}{ap_1(x) + (1-a)p_0(x)} - \frac{\hat{a}\hat{p}_1(x)}{\hat{a}\hat{p}_1(x) + (1-\hat{a})\hat{p}_2(x)}|(ap_1(x) + (1-a)p_0(x))dx \\
&= \int |\frac{ap_1(x)(\hat{a}\hat{p}_1(x) + (1-\hat{a})\hat{p}_0(x)) - \hat{a}\hat{p}_1(x)(ap_1(x) + (1-a)p_0(x))}{\hat{a}\hat{p}_1(x) + (1-\hat{a})\hat{p}_0(x)}|dx \\
&= \int |\frac{ap_1(x)(1-\hat{a})\hat{p}_2(x) - \hat{a}\hat{p}_1(x)(1-a)p_0(x)}{\hat{a}\hat{p}_1(x) + (1-\hat{a})\hat{p}_2(x)}|dx \\
&= \int |\frac{ap_1(x)(1-\hat{a})\hat{p}_2(x) - \hat{a}\hat{p}_1(x)(1-\hat{a})\hat{p}_2(x) + \hat{a}\hat{p}_1(x)(1-\hat{a})\hat{p}_2(x)\hat{a}\hat{p}_1(x)(1-a)p_0(x)}{\hat{a}\hat{p}_1(x) + (1-\hat{a})\hat{p}_0(x)}|dx \\
&= \int |\frac{(1-\hat{a})\hat{p}_0(x)(ap_1(x) - \hat{a}\hat{p}_1(x)) + \hat{a}\hat{p}_1(x)((1-\hat{a})\hat{p}_0(x) - (1-a)p_0(x))}{\hat{a}\hat{p}_1(x) + (1-\hat{a})\hat{p}_0(x)}|dx \\
&\leq \int |\frac{(1-\hat{a})\hat{p}_0(x)\max\{|\hat{a}\hat{p}_1(x) - ap_1(x)|, |(1-\hat{a})\hat{p}_0(x) - (1-a)p_0(x)|\} + \hat{a}\hat{p}_1(x)\max\{|\hat{a}\hat{p}_1(x) - ap_1(x)|, |(1-\hat{a})\hat{p}_0(x) - (1-a)p_0(x)|\}}{\hat{a}\hat{p}_1(x) + (1-\hat{a})\hat{p}_0(x)}|dx \\
&= \int |\max\{|\hat{a}\hat{p}_1(x) - ap_1(x)|, |(1-\hat{a})\hat{p}_0(x) - (1-a)p_0(x)|\}|dx \\
&= \int |\max\{|\hat{a}\hat{p}_1(x) - \hat{a}p_1(x) + \hat{a}p_1(x) - ap_1(x)|, |(1-\hat{a})\hat{p}_0(x) - (1-\hat{a})p_0(x) + (1-\hat{a})p_0(x) - (1-a)p_0(x)|\}|dx \\
&\leq \int |\max\{\hat{a}|\hat{p}_1(x) - p_1(x)| + p_1(x)|\hat{a} - a|, (1-\hat{a})|\hat{p}_0(x) - p_0(x)| + p_0|(1-\hat{a}) - (1-a)|\}|dx = \\
&\leq \int |\max\{\hat{a}|\hat{p}_1(x) - p_1(x)|, (1-\hat{a})|\hat{p}_0(x) - p_0(x)|\} + \max\{p_1(x)|\hat{a} - a|, p_0|(1-\hat{a}) - (1-a)|\}|dx = \\
&\leq \int |\max\{\hat{a}|\hat{p}_1(x) - p_1(x)|, (1-\hat{a})|\hat{p}_0(x) - p_0(x)|\}dx + |\hat{a} - a| \\
&\leq \int \hat{a}|\hat{p}_1(x) - p_1(x)| + (1-\hat{a})|\hat{p}_0(x) - p_0(x)|dx + |\hat{a} - a| \\
&= \hat{a}d_{TV}(p_1, \hat{p}_1) + (1-\hat{a})d_{TV}(p_0, \hat{p}_0) + |\hat{a} - a|
\end{aligned}$$

Let $\epsilon_{\mathcal{F},TV}(m,\delta) = \arg\min_{\epsilon' : m \geq m_{\mathcal{F},TV}(\epsilon',\delta)} \epsilon'$. Now if we have $m \geq \max\{m_{TV,\mathcal{F}}(\frac{\epsilon}{3}, \frac{\delta}{3}), \frac{-9\ln(\frac{\delta}{6})}{2\epsilon^2}\}$, we have:

- From Hoeffding inequality we get $Pr_{S\sim P^m}[|a - \hat{a}|| \geq \frac{\epsilon}{3}] \leq 2\exp(-\frac{2}{9}\epsilon^2 m) = \frac{\delta}{3}$
- From the sample complexity $\hat{a}d_{TV}(p_1, \hat{p}_1) \leq \hat{a}\epsilon_{\mathcal{F},TV}(\hat{a}m, \delta) \leq \frac{\epsilon}{3}$ with probability $1 - \frac{\delta}{3}$.
- similarly $(1-\hat{a})d_{TV}(p_1, \hat{p}_1) \leq (1-\hat{a})\epsilon_{\mathcal{F},TV}((1-\hat{a})m, \delta) \leq \frac{\epsilon}{3}$ with probability $1 - \frac{\delta}{3}$.

Thus get $\mathbb{E}_{X\sim P_X}[|f(X) - \mathcal{A}'(S)(X)|] \leq \epsilon$ with probability $1 - \delta$. □

A.5 SAMPLE COMPLEXITY OF CLF-COVERAGE LEARNING COMPARED TO OTHER LEARNING PROBLEMS

Proof of Lemma 5. From Thm 1.3 of Devroye et al. [2018], we get that TV distance approximation implies approximating the mean to precision $C\epsilon$. The KL-divergence between two one-dimensional, unit variance Gaussians with means having difference $C\epsilon$ is $\frac{1}{2}C^2\epsilon^2$. Then we apply the Bretagnolle-Huber inequality (stated in the appendix as Lemma 2) to get the lower bound on estimating the mean of the Gaussian to be $\frac{1}{2}\exp\left(-C^2\epsilon^2/2\right)$. □

Proof of Theorem 1. For any $m > 0$, let $\mathcal{D}_m = \frac{1}{2}\mathcal{N}(m,1) \times \{1\} + \frac{1}{2}\mathcal{N}(-m,1) \times \{0\}$. The labelling probability for a point x according to this distribution is

$$\begin{aligned} f_m(x) &:= \frac{\exp(-(x-m)^2)}{\exp(-(x-m)^2) + \exp(-(x+m)^2)} \\ &= \frac{1}{1+\exp(-2xm)} \end{aligned}$$

Consider $\mu(\epsilon, M) > \frac{1}{\sqrt{M}} \ln \frac{1-\epsilon}{\epsilon} + \frac{1}{\sqrt{M}}$. We will show that the ϵ labelling probability learning problem for the class $\mathcal{F}_{\mu(\epsilon,M)}$ can be reduced to the problem of finding, for each $\mathcal{D}_m \in \mathcal{F}_{\mu(\epsilon,M)}$, finding an m' s.t. $|m - m'| < \frac{1}{\sqrt{M}}$ with probability $\frac{2}{3}$, based on samples from D_m. This problem as sample complexity at most $\sqrt{M}$

For any $m \geq \mu$, for any m' s.t. $|m - m'| < \Delta(M) = \frac{1}{\sqrt{M}}$, we will show that for every $x \in \mathbb{R}$, $|f_m(x) - f_{m'}| < \epsilon$.

$$\begin{aligned} &|f_m(x) - f_{m'}(x)| \\ &= \frac{|\exp(-2xm) - \exp(-2xm')|}{(1+\exp(-2xm))(1+\exp(-2xm'))} \\ &\leq |\exp(-2xm) - \exp(-2xm')| \end{aligned}$$

For $x < \bar{x} = \frac{1}{2\Delta(M)} \ln \frac{1}{1-\epsilon}$,

$$\begin{aligned} |\exp(-2xm) - \exp(-2xm')| &\leq 1 - \exp(-2x|m-m'|) \\ &< 1 - \exp(-2\bar{x}|m-m'|) \\ &\leq \epsilon \end{aligned}$$

For $x \geq \bar{x}$,

$$\begin{aligned} |\exp(-2xm) - \exp(-2xm')| &< \exp\left(-2x\min\{m, m'\}\right) \\ &\leq \exp\left(-2\bar{x}(\mu(\epsilon, M) - \Delta)\right) \\ &\leq \exp\Bigg(- 2 \cdot \frac{\sqrt{M}}{2} \ln \frac{1}{1-\epsilon} \\ &\cdot \frac{1}{\sqrt{M}} \ln \frac{1-\epsilon}{\epsilon}\Bigg) \\ &= \epsilon \end{aligned}$$

□

Proof of Observation 2. We know that there is a CLF-learner $\mathcal{A}_CLF$ for $\mathcal{P}$. As learner for classification to excess risk, we define our learner $\mathcal{A}$ by

$$\mathcal{A}(S) = \begin{cases} 1 & \text{if } \mathcal{A}_{CLF} \geq \frac{1}{2} \\ 0 \text{ otherwise} & \end{cases}.$$

Let $S \sim P^m$ with $m \geq m_{CLF,\mathcal{P}}(\epsilon, \delta)$. With probability $1 - \delta$ over the generation of S, $\mathcal{A}_{CLF}(S)$ returns a hypothesis $\hat{l}_P$ with

$$\int |l_P(x) - \hat{l}_P(x)| dP \leq \epsilon.$$

Then with probability at least $1-\delta$ the learner $\mathcal{A}$ returns a hypothesis h with

$$
\begin{aligned}
&\mathbb{P}_{(X,Y)}[h(X) \neq Y] - \mathbb{P}_{(X,Y)\sim P}[h^*_P \neq Y]] = \\
&\int |h(x) - l_P(x)| \mathbb{1}\left[h(x) = 0\right] dP + \int |h(x) - 1 + l_P(x)| \mathbb{1}\left[h(x) = 1\right] \\
&- \int |h^*_P(x) - l_P(x)| \mathbb{1}\left[h(x) = 0\right] dP + \int |h^*_P(x) - 1 + l_P(x)| \mathbb{1}\left[h^*_P(x) = 1\right] = \\
&\int (|h(x) - l_P(x)| - |h^*_P(x) - 1 + l_P(x))| \mathbb{1}\left[h(x) = 0 \wedge h^*_P(x) = 1\right] dP \\
&+ \int (|h(x) - 1 + l_P(x)| - |h^*_P(x) - l_P(x))| \mathbb{1}\left[h(x) = 1 \wedge h^*_P(x) = 0\right] dP = \\
&\int |2l_P(x) - 1| \mathbb{1}\left[h(x) = 0 \wedge h^*_P(x) = 1\right] dP + \\
&\int |2l_P(x) - 1| \mathbb{1}\left[h(x) = 1 \wedge h^*_P(x) = 0\right] dP \\
&= \int |2l_P(x) - 1| \mathbb{1}\left[\hat{l}_P(x) < \frac{1}{2} \wedge l_P(x) \geq \frac{1}{2}\right] dP + \\
&\int |2l_P(x) - 1| \mathbb{1}\left[\hat{l}_P(x) \geq \frac{1}{2} \wedge l_P(x) < \frac{1}{2}\right] dP \\
&\leq \int |2l_P(x) - 1| \mathbb{1}\left[|\hat{l}_P(x) - l_P| \leq |\frac{1}{2} - l_P(x)|\right] dP \\
&\leq 2 \int |l_P(x) - \hat{l}_P(x)| dP \leq 2\epsilon.
\end{aligned}
$$

□

Proof of Lemma 6. Due to α-domain validity and (β,γ)-domain non-triviality, the probability weight of points satisfying at least one of the following two (bad) conditions is at most $1-\alpha+1-\beta$.

1. The true CLF lies outside the CLF-coverage set.
2. The CLF-coverage set has length more than γ.

For points in such a set, we can trivially bound the difference between the true CLF and the CLF estimate obtained from the CLF-coverage set by one. The weight of the points not in this set can be trivially bounded by one. For all points in the complement set, the difference between the true CLF and the CLF estimate can be bounded by γ. Therefore, we can bound the ℓ_1 norm of the difference between the true CLF and the CLF estimate from the coverage hypothesis by $2-\alpha-\beta+\gamma$. □

B EXTENDED DISCUSSION RELATED WORK

B.1 CONFORMAL LEARNING

One earlier approach to providing confidence estimates to prediction is through the notion of *conformal mappings* Shafer and Vovk [2008]. This notion usually applies to regression or multiclass learning problem and outputs regions in the label space that are guaranteed, with high probability, to contain the true label value.

The conformal mappings literature differs from this work in several respects: In most setups the probability here is the joint probability over the training data and the probability over a newly arriving test-point. Most guarantees discussed in the literature on conformal mappings are distribution-free, requiring only that the data is exchangable (the common i.i.d. assumption is a special case of exchangeablility). Furthermore, most of the conformal prediction literature considers online-settings.

There is some work on conformal prediction that explores giving guarantees conditioned on subsets or elements of the domain [Lei and Wasserman, 2014, Vovk, 2013, Foygel Barber et al., 2020]. The probabilities here are still aggregated over the generation of the training set and on the randomness of the instance to be classified. Lei and Wasserman [2014] show that it is impossible to give point-wise-guarantees in the distribution-free regression setting. Vovk [2013] extend this result to a general prediction setting that includes classification. Furthermore, Foygel Barber et al. [2020] show that it is also impossible to give distribution-free validity guarantees for all regions with mass greater than some tolerance parameter δ while providing non-trivial coverage sets in a regression setup. However, they also show that if the collection of possible regions is pre-definded and has finite VC-dimension it is possible to provide validity guarantees for these regions and non-trivial coverage sets in a distribution-free setting. We note that this definition, while similar to ours, gives slightly worse validity guarantees, as they only require validity for regions with mass higher than some δ, while our definition of region-validity gives a guarantee on all regions of a predefined collection of subset. It is left as an open question in their work whether the impossibility of non-trivial point-wise guarantees or region-specific guarantees for greater collections of subsets can be overcome by additional distributional assumptions. We address this question in our work. Our setup also distinguishes the randomness that comes from the sampling of the training set and the randomness that comes from sampling a new instance. Furthermore, we consider a binary classification setting. In addition to analysing point-wise and region-specific guarantees for label coverage sets we also propose coverage set is the conditional labelling function(CLF) instead of the label itself.

Selective Classification/ Classification with Abstention: Another line of work that is related to our paper is learning with abstention. Similar to our setting, the classification problem does not only consist of the goal of classifying correctly, but to also allows the classifier to abstain from making a prediction, if the confidence of a prediction is too low. Many works in this line provide accuracy guarantees that hold with high probability over the domain [Bartlett and Wegkamp, 2008, Yuan and Wegkamp, 2010, Freund et al., 2004, Herbei and Wegkamp, 2006, Kalai et al., 2012, Michael, 2010]. In contrast to their work we also provide point-wise-guarantees and guarantees that hold for specific subregions of the domain. Furthermore we also consider the problem of learning a coverage set for the conditional labeling function to distinguish if the uncertainty of a point comes from undersampling or from a lack of informative features/inherent stochasticity of the process. Such a distinction is not made in the classification with abstention setting.

Point-wise guarantees are provided in earlier work [El-Yaniv and Wiener, 2010, Wiener and El-Yaniv, 2015].The former study gave a theoretical analysis of the selective classification setup in which a classification function and a selective function are learned simultaneously [El-Yaniv and Wiener, 2010]. The risk of a classification is then only accessed on the set of instances that was selected for classification. As a non-triviality requirement they propose high coverage over the set. They analyse the trade-off between risk and coverage (which is similar to our trade-off between validty and non-triviality), and introduce a notion of "perfect classification" which requires risk 0 with certainty. In our setting this corresponds to point-wise validity guarantees for our coverage sets. Their results for such point-wise guarantees are developed under an assumption of realizability by a hypothesis class. Under this assumption they provide an optimal learning strategy for their notion of perfect classification and then also show that for some hypothesis classes they can give non-trivial point-wise prediction guarantees. In contrast to this work, our setup also considers probabilistic labeling functions (which are excluded by a realizability assumption). In addition, our analysis extends to different types of assumptions on the family of probability distributions and extends their framework to also include region-wise guarantees.

Learning Gaussians An example of our mixture model CLF-learning setup is to learn conditional labeling functions of mixtures of Gaussians. Our aim here differs from the literature of learning Gaussians we are aware of. There are many results for parameter estimation for Gaussian mixtures models if the components of the mixture are well separated [e.g. Kwon and Caramanis [2020]]. In the case of Gaussians that are not well separated the problem of parameter estimation of a mixture of Gaussians is known to be hard [Moitra and Valiant, 2010]. In contrast, we give a finite sample complexity on the labelled data needed to learn the conditional labelling function that is independent of separation criteria. Thus we show that learning the conditional labelling function of Gaussians is easier than parameter estimation for a mixture of Gaussians. Another line of work looks at learning the marginal of Gaussian mixtures in total variation distance [Ashtiani et al., 2020], giving finite sample complexities for this learning problems without separability assumptions. We show that a learner for generative mixture models in total variation distance can be used to learn the labelling function of a mixture of homogeneously labelled generating distributions with the same sample complexity. However we also show that the problem of learning the marginal of a Gaussian mixture model can have higher sample complexity than learning the associated conditional labelling function. There is also work that looks at learning classification of homogeneously labelled Gaussians with respect to excess risk [Li et al., 2017]. We show that CLF learning implies learning classification with respect to excess risk. However, we also show that the problem of learning the CLF is in general harder and give examples that separate the sample complexities of the two problems.

C EXTENDED DISCUSSION OF THE RELATION BETWEEN LABEL COVERAGE AND CLF COVERAGE

Given a label coverage hypothesis c_{label} with α-point-validity and β-domain-non-triviality, we can construct a CLF coverage hypothesis c_{CLF} by

$$c_{CLF}(x) = \begin{cases} [0,\alpha] & \text{, if } c_{label}(x) = \{0\} \\ [1-\alpha, 1] & \text{, if } c_{label}(x) = \{1\} \\ [0,1] & \text{, if } c_{label}(x) = \{0,1\} \end{cases}$$

This hypothesis has point-wise-validity and β, α-domain-non-triviality. Given access to a CLF coverage hypothesis h'_{CLF} with point-wise validity, we can construct a label coverage hypothesis $h'_{label,\alpha}$ with α-point-wise validity by

$$c'_{label}(x) = \begin{cases} \{0\} & \text{, if } c'_{CLF}(x) = [a,b] \text{ with } b \leq \alpha \\ \{1\} & \text{, if } c'_{CLF}(x) = [a,b] \text{ with } a \geq 1-\alpha \\ \{0,1\} & \text{, otherwise.} \end{cases}$$

While this hypothesis satisfies point-wise validity guarantees we cannot bound the non-triviality of c'_{label} in terms of the non-triviality of c'_{CLF}, as it is possible that the true CLF is close to $\frac{1}{2}$ for every point, yielding trivial label coverage sets, even if we have a tight CLF coverage. However we can compare the non-triviality of c'_{label} to the best possible label coverage possible for a given distribution. Let the α-level Bayes-coverage hypothesis be defined by

$$c^*_{P,\alpha}(x) = \begin{cases} \{0\} & \text{, if } l_P(x) \leq \alpha \\ \{1\} & \text{, if } l_P(x) \geq 1-\alpha \\ \{0,1\} & \text{, otherwise.} \end{cases}$$

It is easy to see that this is the optimal label coverage hypothesis in terms of non-triviality that fulfills point-wise α-level coverage. If $c^*_{P,\alpha}$ has beta-non-triviality β and c'_{CLF} has β, γ-domain-non-triviality, then c'_{label} has $(\gamma + (1-\gamma)(\beta' + \beta))$-domain-non-triviality.

D COMPARING SAMPLE COMPLEXITIES OF TASKS

In this section, we compare the CLF-coverage learning problem with the CLF-learning problem and the problem of learning the Bayes classifier, in terms of sample complexity. We construct classes of distributions for which the following hold:

- The hardest CLF-coverage learning problem requiring point-wise validity has lower sample complexity than the problem of learning the CLF in TV distance.
- The easiest CLF-coverage learning problem requiring domain validity has higher sample complexity than the problem of learning the Bayes optimal classifier.

For $\mu > 0$, let $\mathcal{F}_\mu$ be the class of distributions:

$$\mathcal{F}_\mu := \left\{ \frac{1}{2}\mathcal{N}(x,1) \times \{1\} + \frac{1}{2}\mathcal{N}(-x,1) \times \{0\} : x \geq \mu \right\}.$$

We start by providing a lower bound on the sample complexity of the CLF-learning problem for every class $\mathcal{F}_\mu$. This lower bound is stated as the following lemma:

Lemma 5. *[Lower bound for distribution learning sample complexity] For $\epsilon < 0.004$, for any $\mu > 0$, the sample complexity for TV-learning the class $\mathcal{F}_\mu$ is at least $m_{TV,\mathcal{F}_\mu}(\epsilon, \frac{1}{3}) \geq C\frac{1}{\epsilon^2}$ for a universal constant C.*

In the following theorem, we show that there is a class $\mathcal{F}_\mu$ with arbitrarily small sample complexity for learning CLF-coverage sets with point-wise validity and point-wise non-triviality. Combined with the previous lemma, this theorem shows that there is a class for which CLF learning is more difficult than CLF-coverage sets learning.

Theorem 1. *For every $M \in \mathbb{N}$ and every $\epsilon > 0$, there exists a $\mu > 0$ such that the sample complexity for CLF-learning $\mathcal{F}_\mu$ - $m_{CLF,\mathcal{F}_\mu}(\epsilon, \frac{1}{3})$ is at most M.*

D.1 CONNECTION BETWEEN CLF-LEARNING AND CLASSIFICATION, EXTENDED VERSION

We will now look at the connection between CLF learning and learning a good classification rule. As it is not always possible to define a function with classification loss our optimality criterion will be defined with respect to the excess risk, which is defined as the loss of the best possible classifier, i.e. the Bayes classifier for the distribution.

Definition 1. *Classification Learning with respect to excess risk We say the family of probabilities $\mathcal{P}$ can be learned with respect to excess risk with sample complexity $m_{ex,\mathcal{P}}$, if there is a learner $\mathcal{A}$ such that for every $\epsilon, \delta \in (0,1)$ for all $m \geq m(\epsilon, \delta)$ and all $P \in \mathcal{P}$, we have*

$$\mathbb{P}_{S \sim P^m}[\mathcal{L}_P^{0/1}(\mathcal{A}(S)) \leq \mathcal{L}_P^{0/1}(h_P^*) + \epsilon] \geq 1 - \delta$$

We first show that CLF learning implies learning the classification problem up to excess risk.

Observation 2. *If a family of distributions $\mathcal{P}$ is CLF-learnable with sample complexity $m_{CLF,\mathcal{P}}(\epsilon, \delta)$, then $\mathcal{P}$ learnable with respect to excess risk with sample complexity at most $m(\epsilon, \delta) \leq m_{CLF,\mathcal{P}}(2\epsilon, \delta)$.*

Now we show that learning CLF-coverage sets can be harder than learning the Bayes classifier. The Bayes classifier for any distribution in any class $\mathcal{F}_\mu$ is the classifier that thresholds at zero. We don't need any samples to learn the Bayes classifier for the classes $\mathcal{F}_\mu$. However, to provide CLF-coverage sets, we will need samples. This is even for the easiest CLF-covering problem requiring domain validity and domain non-triviality.

To show that samples are required for the CLF-coverage learning problem requiring domain validity and domain non-triviality, we first show that solving this problem implies approximating the CLF in ℓ_1 distance. Then we note a positive lower bound on the difference in ℓ_1 distance between distributions in $\mathcal{F}_\mu$. Together, these statements imply that for every distribution class $\mathcal{F}_\mu$, there exist $\alpha(\mu), \beta(\mu), \gamma(\mu) > 0$ such that the problem of providing CLF-covers satisfying $\alpha(\mu)$-domain wide validity and $(\beta(\mu), \gamma(\mu))$-domain non-triviality has higher sample complexity than learning the Bayes classifier.

The following lemma (Lemma 6) shows how the CLF-coverage learning problem with the easiest validity requirement implies approximating the CLF in ℓ_1 distance:

Lemma 6. *A CLF-coverage set that has α-marginal coverage and (β, γ)-domain non-triviality yields an approximation to the true CLF that has at most $(2 - \alpha - \beta + \gamma)$ - ℓ_1 distance from the true CLF. The approximation is one obtained by simply choosing any element of the CLF-coverage set for each point.*

Now, we note a positive lower-bound in - $\epsilon(\mu)$ on the ℓ_1 norm of the pair-wise differences of distributions in the class $\mathcal{F}_\mu$.

Observation 1. *For every $\mu > 0$, there is an $\epsilon(\mu) > 0$ such that the distribution class $\mathcal{F}_\mu$ contains distributions with ℓ_1 norm of difference at least $\epsilon(\mu)$.*

Therefore, the CLF learning problem to precision $\epsilon(\mu)$ for the class $\mathcal{F}_\mu$ requires samples and is hence harder than the problem of learning the Bayes classifier. Combined with Lemma 6, Observation 1 shows that CLF-coverage learning with domain-validity is harder than learning the Bayes classifier for $\mathcal{F}_\mu$.

E SPLIT CONFORMAL ALGORITHM

The split conformal was algorithm introduced by Vovk et al. [Vovk et al., 2005]. This algorithm is shown to provide distribution-free, marginal coverage by Lei and Wasserman [Lei and Wasserman, 2014]. The split conformal method partitions the sample into two parts - the training set and the validation set. The training set is used to train a model. The validation set is used to evaluate that model. The size of the coverage sets is determined by how well the trained model fits the validation set. When the model fits the validation set well, the algorithm outputs small coverage sets. A modification of this algorithm for regression conformal prediction satisfying a more refined guarantee than the marginal coverage is provided by Barber et al [Foygel Barber et al., 2020]. Rather than simply guaranteeing coverage with high probability over test points drawn from the domain, the refined guarantee is for coverage with high probability over test-points drawn conditioned on membership in predefined subsets of the domain.

We now state this form of the split conformal algorithm by Barber et al. [Foygel Barber et al., 2020] as Algorithm 1.

Algorithm 1 Split conformal algorithm for restricted conditional coverage

```
Input: Validity parameter: $\alpha$, Collection of subsets of the domain: $\mathcal{B}$,
Labelled samples: $S = (x_i, y_i)_{i=1}^{m}$, Binary classification learning algorithm: $\mathcal{A}$,
Test point: $x$.
Output: Label coverage set for $x$
$S_t = \{(x_1, y_1), \ldots, (x_n, y_n)\}$
$S_v = \{(x_{n+1}, y_{n+1}), \ldots, (x_m, y_m)\}$
Default coverage set for $x$ is $\hat{C}(x) = \{\mathcal{A}(S_t)(x)\}$.
for $B \in \mathcal{B}$ such that $x \in B$ do
    $N_{S_v}(B) = |S_v \cap B|$
    $\mathcal{L}^{0/1}_{S_v,B}(\mathcal{A}(S_t)) = |\{(x', y') : (x', y') \in S_v \cap B$ and $A(S_t)(x') \neq y'\}|$
    $\mathrm{threshold} = N_{S_v}(B) - \left\lceil \left(1 - \alpha + \frac{1}{m-n}\right)(N_{S_v}(B) + 1)\right\rceil$
    if $\mathcal{L}^{0/1}_{S_v,B}(\mathcal{A}(S_t)) \geq \mathrm{threshold}$ then
        Set coverage set of $x$ to be trivial i.e. $\hat{C}(x) = \{0, 1\}$
    end if
end for
Return $\hat{C}(x)$
```

We refer the reader to the paper of Barber et al.[Foygel Barber et al., 2020] for a proof that this algorithm yields coverage sets satisfying region-conditional validity. While this algorithm has the desired quality of providing distribution-free validity guarantees, under certain distributional assumptions, this algorithm could provide coverage sets with sub-optimal non-triviality. We show that this is the case under the distributional assumptions we consider in this work.

F COMPARISON OF METHODS

We have seen a few approaches for constructing label-coverage sets with region-conditional validity so far. In Section 4, we proposed two methods that stem from the two conditional error bounds provided by Theorem 3 and Theorem 4. We will refer to these methods as 'conditional error bound methods' (abbreviated as CEB methods). We will refer to the CEB method based on Theorem 3 as the 'baseline CEB method' and the CEB method based on Theorem 4 as the 'decisiveness-based CEB method'. Another method for coverage sets with region-conditional validity is the modified split conformal algorithm proposed by Barber et al. [Foygel Barber et al., 2020] (see Algorithm 1 in Section E of the appendix for a description of this algorithm). In this section, we will discuss some differences among these approaches. We will focus on differences in the case of the data generating distribution satisfying the assumption we have been studying in this chapter - low approximation error by a function class H.

The modified split conformal algorithm and the baseline CEB both have the advantage of providing distribution-free validity. The baseline CEB method uses the whole labelled training set to both train an ERM classifier and evaluate that classifier. The split conformal algorithm on the other hand partitions the labelled training set into two parts and uses one part for training a model and the other part for evaluating that model. Due to this, the classifier used for coverage sets construction in the baseline CEB method is likely to have lower error (by a constant factor) than the classifier in the split conformal method. This could result in the baseline CEB method's coverage sets having higher non-triviality compared to the coverage sets from the split conformal algorithm. However, the split conformal method allows for more general training algorithms and is therefore likely to adapt better even when the probability distribution is not approximated well by the function class.

Compared to the split conformal algorithm and the baseline CEB method, the decisiveness-based CEB method has the disadvantage of requiring knowledge of an upper bound on the approximation error of the function class in order to construct coverage sets. However, the decisiveness-based CEB method makes better use of the distributional assumption to provide coverage sets with higher non-triviality in some cases. The distributional assumption allows the decisiveness-based CEB method to better utilize unlabelled data. Recall that both conditional error bounds are obtained by estimating the error on the region $\left(\mathcal{L}^{0/1}_{P,B}\right)$ and the probability weight of the region $(P(B))$. Unlabelled data is used to estimate the probability weight by both the baseline and the decisiveness-based CEBs. The decisiveness-based CEB also uses the unlabelled data to estimate region's error whereas the baseline CEB uses only labelled data for this. The rest of this section describes an example where decisiveness-based CEB method provides better non-triviality compared to the baseline CEB method and the split conformal method.

We use the following notation for the example: The domain $\mathcal{X}$ is the unit real interval - $[0, 1]$. The class of threshold classifiers over this domain is denoted by $H_{\text{thresholds}} = \{h_a : a \in [0, 1]\}$. The threshold classifier denoted by h_a for $a \in [0, 1]$ is such that $h(x) = 0$ for every $x \leq a$ and $h(x) = 1$ for every $x > a$. For $\epsilon > 0$, $\mathcal{P}_{\text{thresholds},\epsilon}$ denotes the class of probability distributions that are approximated by the class H with approximation error - $\text{opt}_P(H)$ at most ϵ. That is, a probability distribution P belongs to the class $\mathcal{P}_{\text{thresholds},\epsilon}$ if and only if $\min_{h \in H} \mathcal{L}_h^{0/1} \leq \epsilon$.

Example 1. *Let the domain $\mathcal{X}$ be the unit interval $[0, 1] \subseteq \mathbb{R}$. Let the marginal distribution $P_{\mathcal{X}}$ be the uniform distribution. Let the conditional distribution be:*

$$P(y = 1|x) = \begin{cases} 1 - 0.001, & \textit{if } x \geq \frac{1}{2} \\ 0.001, & \textit{if } x < \frac{1}{2}. \end{cases}$$

We want to construct label-coverage sets based on samples drawn from P using prior knowledge that P has approximation error by the threshold class $opt_P(\mathcal{H}_{thresholds}) = 0.001$. We have access to 100 labelled samples drawn i.i.d from P and 10^7 unlabelled samples drawn i.i.d from $P_{\mathcal{X}}$. The goal is to provide $(0.85, 0.85, \{B\})$-region-conditional coverage sets for $B = [0, 0.01]$. The following hold:

- *With probability more than $\frac{1}{2}$ over the samples drawn, the split conformal method assigns trivial label-coverage sets for all points in B.*
- *With probability more than $\frac{1}{2}$ over the samples drawn, the baseline CEB method assigns trivial label-coverage sets for all points in B.*
- *With probability more than $\frac{1}{2}$, the decisiveness-based CEB method assigns non-trivial label-coverage sets for all points in B.*

Note that in this example, the claim we make about the split conformal algorithm is for the algorithm with parameter $(1 - \alpha)$ equalling 0.85. Barber et al. Foygel Barber et al. [2020] show that with this parameter, the coverage sets satisfy a notion called 0.85-restricted conditional coverage. This is a weaker validity guarantee that implies $(0.85, 0.85)$-region conditional validity. Now we prove the correctness of the above example. We outline a sketch of this proof here and defer the full details of the proof to the appendix.

We will now give an outline to the proof:

1. Showing that the baseline CEB method returns trivial coverage sets: We show that the bound provided by Theorem 3 for B with sample failure parameter $\delta = 0.15$ is vacuous (greater than 1.0). This implies trivial coverage sets. Note that $\frac{\epsilon_{UC}(|S_l|, 0.15/2)}{\frac{|S_u \cap B|}{|S_u|}}$ is a lower bound on the baseline CEB given by Theorem 3. The numerator of this lower bound is a constant value that we can calculate. The denominator is close to $P(B) = 0.01$ with high probability. The denominator is less than the numerator with high probability and hence the baseline CEB is vacuous.
2. Showing that the split conformal algorithm returns trivial coverage sets: First we show that with probability, there are only few validation samples in B. Then we show that this implies that the split conformal algorithm returns trivial coverage sets.
3. Showing that the decisiveness-based CEB method returns non-trivial coverage sets: We first show that with high probability over the samples, the decisiveness of the region B is the highest value - one. This will imply that the error of the region B is low. The unlabelled samples provide an estimate of the probability weight of B that is larger than the bound on the region's error. This results in a small conditional error bound due to the decisiveness-based CEB given by Theorem 4.

   To show that decisiveness is one with high probability, we show that any classifier in $H_{\text{thresholds}}$ that labels any point in $S_u \cap B$ zero, has high sample error with high probability. This shows that all classifiers with low empirical error have the same behaviour on $S_u \cap B$ i.e., label zero for all points in $S_u \cap B$. Therefore the decisiveness is one.

We will now give a detailed proof with full calculations.

Proof of validity of Example 1. Using the enumeration from the proof outline we will now show the three parts.

1. Showing that the baseline CEB method returns trivial coverage sets: We show that the bound provided by Theorem 3 for B with sample failure parameter $\delta = 0.15$ is greater 1.0. This implies trivial coverage sets.

   Note that a lower bound for this bound is

   $$\begin{aligned}\frac{\epsilon_{UC}(|S_l|, 0.15/2)}{\frac{|S_u \cap B|}{|S_u|}} &= \frac{\sqrt{\frac{9(1+\log(2/0.15))}{150}}}{\frac{|S_u \cap B|}{|S_u|}} \\ &> \frac{0.35 \cdot 10^7}{|S_u \cap B|}.\end{aligned}$$

   The expected value of $|S_u \cap B|$ is $10^7 \cdot 0.01 = 10^5$. With high probability, $|S_u \cap B|$ is not much larger than 10^5. By the Hoeffding inequality,

   $$\mathbb{P}\left[|S_u \cap B| > 10^5 + \sqrt{\frac{10^7}{2} \ln 4}\right] \leq \frac{1}{10}.$$

   With probability at least 0.9,

   $$\begin{aligned}\frac{|S_u \cap B|}{|S_u|} &< \frac{10^5}{10^7} + 10^{-3.5}\sqrt{\frac{\ln 10}{2}} \\ &< 0.013.\end{aligned}$$

   Since $0.35 > 0.013$, the bound given by Theorem 3 is bigger than 1.0.

2. Showing that the split conformal algorithm returns trivial coverage sets: First we show that with probability at least 0.52, the number of validation samples that lie in the region B is at most six. Then we show that this low number of validation samples in B implies that the split conformal method returns trivial coverage sets.

   (a) Showing that there are few validation samples in B. The expected size of $|S_v \cap B|$ is $P(B)|S_v| = 0.01 \cdot 75 = 0.75$. By applying the Hoeffding inequality, we get that

   $$\begin{aligned}\mathbb{P}\left[|S_v \cap B| \geq 6\right] &\leq \exp\left(-\frac{2(6 - P(B)|S_v|)^2}{|S_v|}\right) \\ &\leq \exp\left(-\frac{2(6-0.75)^2}{75}\right) \\ &< 0.48.\end{aligned}$$

   (b) Showing that split conformal algorithm returns trivial coverage sets when $|S_v \cap B| \leq 6$. Recall that the split conformal algorithm (Algorithm 1) calculates a threshold value and the number of errors in $S_v \cap B$ made by the empirical risk minimizer from $\mathcal{H}_{\text{thresholds}}$. The split conformal algorithm returns trivial coverage sets if the errors in $S_v \cap B$ is greater than the threshold value. If the threshold value is negative, then the split conformal algorithm returns trivial coverage sets regardless of the number of errors in $S_v \cap B$. We will now show that when $|S_v \cap B| \leq 6$, the threshold value is negative. The threshold is at most

   $$|S_v \cap B| - \left\lceil \left(1 - \alpha + \frac{1}{|S_v|}\right)(|S_v \cap B| + 1)\right\rceil.$$

   When $|S_v \cap B| \leq 6$, this threshold value is negative.

3. Showing that the decisiveness-based CEB method returns non-trivial coverage sets: We first show that with probability at least 0.62 over samples, the decisiveness of the region B is the highest value - one. Like in the first part, we also show that the fraction of sample points that lie in B is at least 0.013 with probability at least 0.9. The high decisiveness combined with the lower bound on the fraction of samples that lie in B results in the conditional error bound in Theorem 4 with sample failure parameter $\delta = 0.15$ being less than 0.85. This results in non-trivial coverage sets for $(0.85, 0.85)$-region-conditional validity.

   To show that decisiveness is one with high probability, we show that any classifier in $H_{\text{thresholds}}$ that labels any point in $S_u \cap B$ zero, has high sample error with high probability. This shows that all classifiers with low empirical error have the same behaviour on $S_u \cap B$ - labelling all points in $S_u \cap B$ zero. And therefore the decisiveness is one.

(a) Showing that the decisiveness of set B is one with high probability over the samples. We show this by showing that all classifiers in $\mathcal{H}_{\text{thresholds}}$ having sample error within $2\epsilon_{UC}\left(|S_l|, \frac{1}{4}\cdot 0.15\right)$ of the optimal sample error all label all points in $S_u \cap B$ zero. First we show that the sample error of the classifier $h_{\frac{1}{2}}$ is at most 0.0686 with probability at least 0.84. This implies that the sample of the empirical risk minimizing classifier is also at most 0.0686. We show this by applying the Hoeffding inequality. The expected sample error is 0.001.

$$\begin{aligned}\mathbb{P}\left[\mathcal{L}^{0/1}_{S_l,B} \geq 0.1\right] &\leq \exp\left(-\frac{2(12-0.15)^2}{150}\right)\\ &< 0.16.\end{aligned}$$

All classifiers with sample error within $2\epsilon_{UC}(100, 0.15/4)$ have sample error at most $0.0686 + 2\cdot 0.34 = 0.77$.

$$\begin{aligned}0.686 + 2\epsilon_{UC}\left(|S_l|, \frac{1}{4}\cdot 0.15\right) &= 0.686 + 2\sqrt{\frac{9(1+\log(4/0.15))}{150}}\\ &< 0.832\end{aligned}$$

Now we show that with high probability any classifier that labels some point in B one (a classifier $h_a \in \mathcal{H}_{\text{thresholds}}$ with $a < 0.01$) has sample error larger than 0.832. We first show that there are few labelled samples in B similar to how we showed that there are few validation samples in B. With probability at least 0.77, $|S_l \cap B| \leq 12$.

$$\begin{aligned}\mathbb{P}\left[|S_l \cap B| \geq 12\right] &\leq \exp\left(-\frac{2(12-P(B)|S_l|)^2}{|S_l|}\right)\\ &< 0.23.\end{aligned}$$

Next we show that of the labelled sample points not in B, most have labels different from the labels assigned by a classifier $h_a \in \mathcal{H}_{\text{thresholds}}$ with $a < 0.01$. A labelled sample not in B with label agreeing with a classifier $h_a \in \mathcal{H}_{\text{thresholds}}$ differs from the label assigned to it by the classifier $h_{\frac{1}{2}}$. We have already shown that most labelled samples have labels agreeing with the classifier $h_{\frac{1}{2}}$. Therefore, the number of labelled samples in $S_l \setminus B$ that are correctly labelled by a classifier h_a with $a < 0.01$ is upper bounded by the number of labelled samples on which $h_{\frac{1}{2}}$ makes an error. We have shown that this is at most 12 with probability at least 0.832. Therefore, any classifier h_a with $a < 0.01$ makes error on at least $(150 - 12 - 12)$ labelled sample points with probability at least 0.62. This concludes our proof that the decisiveness of the set B is one with probability at least 0.62.

(b) Showing a lower-bound on the number of unlabelled samples in B. By Lemma 3, with probability at least $\frac{1}{10}$, $\frac{|S_u \cap B|}{|S_u|} > 0.01 - \sqrt{\frac{1}{2|S_u|}\ln 10}$.

Therefore, with probability at least 0.52, the conditional generalization error of the empirical risk minimizer can be bounded above by 0.15, using Theorem 4. The decisiveness-based method for $(1-\alpha) = 0.85$-region conditional validity returns non-trivial coverage sets for all points in the set B.

□