Supplementary Material for Paper: Learning Reliable Rules under Class Imbalance

A Additional Terminology

Literature on imbalanced data commonly uses the following terminology.

DEFINITION A.1. *Given a hypothesis $h \in \mathcal{H}$, a target concept $c \in \mathcal{C}$ and a sample $S = (x_1, \ldots, x_m)$, we define the following.*

$$\begin{cases} \textit{The true positives } \mathrm{TP} = \sum_{i=1}^{m} \mathbf{1}_{h(x_i)=1 \wedge c(x_i)=1} \\ \textit{The false positives } \mathrm{FP} = \sum_{i=1}^{m} \mathbf{1}_{h(x_i)=1 \wedge c(x_i)=0} \\ \textit{The true negatives } \mathrm{TN} = \sum_{i=1}^{m} \mathbf{1}_{h(x_i)=0 \wedge c(x_i)=0} \\ \textit{The false negatives } \mathrm{FN} = \sum_{i=1}^{m} \mathbf{1}_{h(x_i)=0 \wedge c(x_i)=1} \end{cases}$$

Figure 1 presents the above terminology in a graphical way.

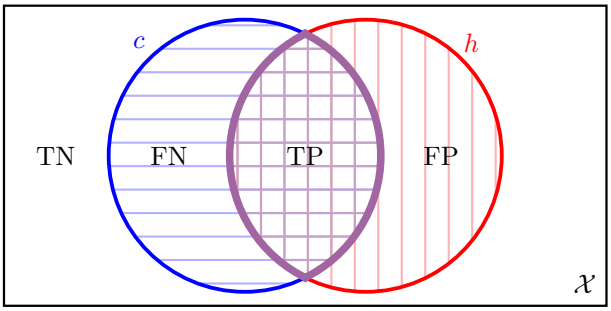

Figure 1: Given a target concept c (shown with horizontal lines) and a hypothesis h (shown with vertical lines) defined over the instance space $\mathcal{X}$, we can see the different subsets of $\mathcal{X}$ that contribute towards the quantities TN, FN, TP, FP.

DEFINITION A.2. (EMPIRICAL RECALL) *Given a hypothesis $h \in \mathcal{H}$, a target concept $c \in \mathcal{C}$, and a sample $S = (x_1, \ldots, x_m)$ where for at least one $i \in \{1, \ldots, m\}$ it holds $c(x_i) = 1$, the* empirical recall *of h is defined by*

$$\widehat{\mathrm{REC}}_S(h, c) = \frac{\sum_{i=1}^{m} \mathbf{1}_{h(x_i)=1 \wedge c(x_i)=1}}{\sum_{i=1}^{m} \mathbf{1}_{c(x_i)=1}} = \frac{TP}{TP + FN}.$$

DEFINITION A.3. (EMPIRICAL PRECISION) *Given a hypothesis $h \in \mathcal{H}$, a target concept $c \in \mathcal{C}$, and a sample $S = (x_1, \ldots, x_m)$ where for at least one $i \in \{1, \ldots, m\}$ it holds $h(x_i) = 1$, the* empirical precision *of h is,*

$$\widehat{\mathrm{PREC}}_S(h, c) = \frac{\sum_{i=1}^{m} \mathbf{1}_{h(x_i)=1 \wedge c(x_i)=1}}{\sum_{i=1}^{m} \mathbf{1}_{h(x_i)=1}} = \frac{TP}{TP + FP}.$$

B Omitted Proofs

Below we provide proofs for the claims that we made in the paper.

PROPOSITION 3.1. (LOWER BOUND FOR RECALL) *Let p_b be given such that $\mathbf{Pr}_{x \sim D}(c(x) = 1) \geq p_b > 0$. Let $h \in \mathcal{H}$ be a hypothesis with risk $R_D(h, c)$. Then, for this hypothesis h it holds*

$$\mathrm{REC}_D(h, c) \geq 1 - \frac{R_D(h, c)}{p_b}.$$

Proof. We have that

$$\begin{aligned} \mathbf{Pr}_{x \sim D}(c(x) = 1) = \mathbf{Pr}_{x \sim D}(c(x) = 1 \wedge h(x) = 1) \\ + \mathbf{Pr}_{x \sim D}(c(x) = 1 \wedge h(x) = 0). \end{aligned}$$

Dividing with $\mathbf{Pr}_{x \sim D}(c(x) = 1)$, using Definition 3.1 and rearranging, we obtain:

$$\mathrm{REC}_D(h, c) = 1 - \frac{\mathbf{Pr}_{x \sim D}(h(x) = 0 \wedge c(x) = 1)}{\mathbf{Pr}_{x \sim D}(c(x) = 1)}.$$

We now use the fact that $\mathbf{Pr}_{x \sim D}(h(x) = 0 \wedge c(x) = 1) \leq R_D(h, c)$ and the fact that $\mathbf{Pr}_{x \sim D}(c(x) = 1) \geq p_b$ and we obtain the statement. ∎

PROPOSITION 3.2. (LOWER BOUND FOR PRECISION) *Let p_b be given such that $\mathbf{Pr}_{x \sim D}(c(x) = 1) \geq p_b > 0$. Let $h \in \mathcal{H}$ be a hypothesis with risk $R_D(h, c)$ and for which it holds $\mathrm{REC}_D(h, c) \geq 1 - \gamma$ for some $0 \leq \gamma < 1$. Then, for this hypothesis h it holds*

$$\mathrm{PREC}_D(h, c) \geq 1 - \frac{R_D(h, c)}{(1 - \gamma) p_b}.$$

Proof. It holds that $\mathbf{Pr}_{x \sim D}(h(x) = 1) \geq \mathbf{Pr}_{x \sim D}(h(x) = 1 \wedge c(x) = 1)$. Furthermore, by Definition 3.1 and by what is given, we have $\mathbf{Pr}_D(h(x) = 1 \wedge c(x) = 1) = \mathrm{REC}_D(h, c) \cdot \mathbf{Pr}_D(c(x) = 1) \geq (1 - \gamma) p_b$. Hence, $\mathbf{Pr}_{x \sim D}(h(x) = 1) \geq (1 - \gamma) p_b > 0$.

On the other hand, working in a manner similar to Proposition 3.1 we have

$$\begin{aligned} \mathbf{Pr}_{x \sim D}(h(x) = 1) = \mathbf{Pr}_{x \sim D}(h(x) = 1 \wedge c(x) = 1) \\ + \mathbf{Pr}_{x \sim D}(h(x) = 1 \wedge c(x) = 0), \end{aligned}$$

where by dividing with $\mathbf{Pr}_{x \sim D}(h(x) = 1)$, which is nonzero, using Definition 3.2 and rearranging, we obtain:

$$\mathrm{PREC}_D(h, c) = 1 - \frac{\mathbf{Pr}_{x \sim D}(c(x) = 0 \wedge h(x) = 1)}{\mathbf{Pr}_{x \sim D}(h(x) = 1)}.$$

We now use the fact that $\mathbf{Pr}_{x \sim D}(c(x) = 0 \wedge h(x) = 1) \leq R_D(h, c)$ and the fact that $\mathbf{Pr}_{x \sim D}(h(x) = 1) \geq (1 - \gamma) p_b$ and we obtain the statement. ∎

THEOREM 3.1. *Let L be a learner such that, for every $0 < \varepsilon, \delta < 1$, L can produce an $h \in \mathcal{H}$ that achieves the PAC criterion (Definition 2.3) when learning $c \in \mathcal{C}$ using hypotheses from $\mathcal{H}$ under a set of distributions $\mathcal{D}$ over $\mathcal{X}$. Let p_b be an input parameter that is known to the learner such that $\mathbf{Pr}_{x\sim D}(c(x)=1) \geq p_b > 0$. Then, for any $0 < \xi < 1$ and any $0 < \gamma \leq 1/2$, using L to generate an $h \in \mathcal{H}$ for which it holds $R_D(h,c) \leq \min\{\varepsilon, \gamma p_b, \xi p_b/2\}$ implies for the same h that $\text{REC}_D(h,c) \geq 1-\gamma$ as well as $\text{PREC}_D(h,c) \geq 1-\xi$. That is, L PAC learns $\mathcal{C}$ with high recall and high precision using $\mathcal{H}$.*

Proof. We want L to generate a hypothesis h that satisfies $\text{REC}_D(h,c) \geq 1-\gamma$ and $\text{PREC}_D(h,c) \geq 1-\xi$. In order to guarantee $\text{REC}_D(h,c) \geq 1-\gamma$ we will use Proposition 3.1 where it is enough if $1 - R_D(h,c)/p_b \geq 1-\gamma \Leftrightarrow R_D(h,c) \leq \gamma p_b$. Furthermore, in order to guarantee $\text{PREC}_D(h,c) \geq 1-\xi$, by Proposition 3.2 it is enough if $1 - R_D(h,c)/((1-\gamma)p_b) \geq 1-\xi \Leftrightarrow R_D(h,c) \leq \xi(1-\gamma)p_b$. Requiring $\gamma \leq 1/2$ it follows that $\xi p_b/2 \leq \xi(1-\gamma)p_b$ and hence we can use the constraint $R_D(h,c) \leq \xi p_b/2$ when using Proposition 3.2 for every $0 < \gamma \leq 1/2$.

Hence, in order to satisfy the constraints $R_D(h,c) \leq \varepsilon$, $\text{REC}_D(h,c) \geq 1-\gamma$, and $\text{PREC}_D(h,c) \geq 1-\xi$ in Definition 3.3, it is enough if we require L to perform PAC learning according to Definition 2.3 and produce a hypothesis h that satisfies $R_D(h,c) \leq \min\{\varepsilon, \gamma p_b, \xi p_b/2\}$.

Finally, since L satisfies Definition 2.3, there exists a polynomial $p(1/\varepsilon, 1/\delta, n, size(c))$ that provides enough samples for PAC learning $\mathcal{C}$ so that it holds $R_D(h,c) \leq \varepsilon$ at the end of the learning process. Therefore, the polynomial $p(1/(\min\{\varepsilon, \gamma p_b, \xi p_b/2\}), 1/\delta, n, size(c))$ (or the even looser $p(1/(\varepsilon\gamma\xi p_b/2), 1/\delta, n, size(c))$) is a good enough polynomial to justify Definition 3.3 for L. ∎

For Lemma 3.1 we will use the following fact.

PROPOSITION B.1. (HOEFFDING'S BOUND) *Let $X_1, \ldots, X_R$ be R independent random variables, each taking values in the range $\mathfrak{I} = [\alpha, \beta]$. Let μ denote the mean of their expectations. Then* $\mathbf{Pr}\left(\left|\frac{1}{R}\sum_{i=1}^{R} X_i - \mu\right| \geq \epsilon\right) \leq 2e^{-2R\epsilon^2/(\beta-\alpha)^2}$.

LEMMA 3.1. *Let $\mathbf{Pr}_{x\sim D}(c(x)=1) = p > 0$. Let $m_i \geq \lceil 2^{3+2i} \ln(2^{1+i}/\delta) \rceil$ for $i \in \{1,2,\ldots\}$. Then, with probability more than $1-\delta$, Algorithm 1 halts within $\lceil \lg(3/2p) \rceil$ iterations and provides a lower bound p_b such that $0 < p/8 \leq p_b < p$.*

Proof. In round $i \in \{1,2,\ldots\}$ we draw m_i examples in order to form an estimate $\widehat{p}_i$ of the true rate p of the minority class within $\epsilon_i = 2^{-(2+i)}$ of its true value. We require this to happen except with probability $\delta/2^i$ and using Proposition B.1 with $\alpha = 0$ and $\beta = 1$ we get that $m_i = \lceil 2^{3+2i} \ln(2^{1+i}/\delta) \rceil$ training examples are enough for this purpose.

Conditioning on the samples used in each iteration to be representative of the underlying true rate p, the process stops in the worst case when $p > (1/4) \cdot (1/2)^{i-1} + (1/2)^{2+i} = 2^{-(1+i)} + 2^{-(2+i)} = 3 \cdot 2^{-(2+i)}$. The reason is that smaller values of p, when approximated within $2^{-(2+i)}$ of their true value may still return an empirical estimate less than or equal to the threshold of $3 \cdot 2^{-(2+i)} - 2^{-(2+i)} = 2^{-(1+i)}$ even if the sample is indicative of the underlying distribution D. In other words, recursive halving of our guesses stops when $3 \cdot 2^{-(i+2)} < p \Leftrightarrow i > \lg(3/(4p))$. Hence, it is enough if $i \geq 1 + \lg(3/4p)$ or equivalently, if $i \geq \lceil \lg(3/2p) \rceil$.

As mentioned above, the estimate in each iteration $i \in \{1,2,\ldots\}$ is computed within $2^{-(2+i)}$ of its true value except with probability $\delta/2^i$. Hence, the probability that some empirical estimate is not within $2^{-(2+i)}$ of its true value, by the union bound is $\sum_{i=1}^{\lceil \lg(3/(2p)) \rceil} \delta/2^i < \delta \cdot \sum_{i=1}^{\infty} 2^{-i} = \delta$.

This last observation allows us to argue that p_b satisfies $p/8 \leq p_b < p$ *when the entire process succeeds* and all the empirical estimates are computed within $\epsilon_i = 2^{-(2+i)}$ of their true values, for $i \in \{1,2,\ldots\}$. First, Algorithm 1 stops when the empirical estimate $\widehat{p}_i$ of p satisfies $\widehat{p}_i > 2^{-(1+i)}$. Since the estimate is accurate within $\epsilon_i = 2^{-(2+i)}$ of its true value, it follows that the true rate p cannot be less than $\widehat{p}_i - \epsilon_i$ for which it holds $\widehat{p}_i - \epsilon_i > 2^{-(1+i)} - 2^{-(2+i)} = 2^{-(2+i)}$, which is precisely the value that we assign to p_b. This explains that $p_b < p$. On the other hand, for the statement $p/8 \leq p_b$ we distinguish cases. If Algorithm 1 terminates at iteration 1, then $p_b = 1/8 \geq p/8$ for any probability p and hence the claim is trivial. If however Algorithm 1 terminates at iteration $i > 1$, this has happened because the estimates $\widehat{p}_{i-1}$ and $\widehat{p}_i$ satisfied $\widehat{p}_{i-1} \leq 2^{-i}$ and $\widehat{p}_i > 2^{-(1+i)}$. However, correctly approximating p within $\epsilon_{i-1} = 2^{-(1+i)}$ and obtaining $\widehat{p}_{i-1} \leq 2^{-i}$ implies that $p \leq 2^{-i} + 2^{-(1+i)} = 3 \cdot 2^{-(1+i)}$. Therefore we have $p/p_b \leq 3 \cdot 2^{-(1+i)}/2^{-(2+i)} \Leftrightarrow p_b \geq p/6$. Hence, in either case we have $p_b \geq p/8$. ∎

COROLLARY 3.2. *Lemma 3.1 requires total sample size $\mathcal{O}\left(\frac{1}{p^2} \cdot \ln\left(\frac{1}{p\delta}\right)\right)$.*

Proof. We will use the fact that Lemma 3.1 lasts no more than $i = \lceil \lg(3/(2p)) \rceil < 1 + \lg(3/(2p)) = \lg(3/p)$

iterations. For the total sample size we have,

$$\begin{aligned} m &= \sum_{i=1}^{\lceil \lg(3/(2p)) \rceil} m_i = \sum_{i=1}^{\lceil \lg(3/(2p)) \rceil} \left\lceil 2^{3+2i} \cdot \ln\left(2^{1+i}/\delta\right) \right\rceil \\ &< \sum_{i=1}^{\lceil \lg(3/(2p)) \rceil} \left(1 + 2^{3+2i} \cdot \ln\left(2^{1+i}/\delta\right)\right) \\ &< 8 \cdot \ln\left(2 \cdot 2^{\lg(3/p)}/\delta\right) \cdot \sum_{i=1}^{\lceil \lg(3/(2p)) \rceil} 4^i + \sum_{i=1}^{\lceil \lg(3/(2p)) \rceil} 1 \\ &= 8 \cdot \ln\left(\frac{6}{p\delta}\right) \cdot \frac{4}{3} \cdot \left(4^{\lceil \lg(3/(2p)) \rceil} - 1\right) + \lceil \lg(3/(2p)) \rceil \end{aligned}$$

Hence, $m < \frac{32}{3} \cdot \ln\left(\frac{6}{p\delta}\right) \cdot 4^{\lg(3/p)} + \lceil \lg(3/(2p)) \rceil$. In other words, we have $m < \frac{96}{p^2} \cdot \ln\left(\frac{6}{p\delta}\right) + \lceil \lg(3/(2p)) \rceil$. ∎

PROPOSITION 4.1. *Let D be a product distribution over $\{0,1\}^n$ where each variable is satisfied with the same probability λ. Consider a target c and a hypothesis h as in (4.1). Then,*

$$\begin{cases} R_D(h,c) &= \lambda^m\left(\lambda^u + \lambda^w - 2\lambda^{u+w}\right) \\ \text{REC}_D(h,c) &= \lambda^w \\ \text{PREC}_D(h,c) &= \lambda^u \end{cases}$$

Proof. We have $\mathbf{Pr}_{x\sim D}(c(x) = 1 \wedge h(x) = 0) = \lambda^{m+u} \cdot (1-\lambda^w)$. Moreover, we have, $\mathbf{Pr}_{x\sim D}(h(x) = 1 \wedge c(x) = 0) = \lambda^{m+w} \cdot (1-\lambda^u)$. Therefore, for the risk $R_D(h,c)$ we have, $R_D(h,c) = \mathbf{Pr}_{x\sim D}(c(x) = 1 \wedge h(x) = 0) + \mathbf{Pr}_{x\sim D}(h(x) = 1 \wedge c(x) = 0) = \lambda^m\left(\lambda^u + \lambda^w - 2\lambda^{u+w}\right)$. Using Definition 3.1 we have, $\text{REC}_D(h,c) = \lambda^w$. Using Definition 3.2 we have, $\text{PREC}_D(h,c) = \lambda^u$. ∎

C Omitted Discussion

First we note that one can prove a simpler version of Theorem 3.1 by requiring only high recall, as shown below.

THEOREM C.1. *Let L be a learner such that, for every $0 < \varepsilon, \delta < 1$, L can produce an $h \in \mathcal{H}$ that achieves the PAC criterion (Definition 2.3) when learning $c \in \mathcal{C}$ using hypotheses from $\mathcal{H}$ under a set of distributions $\mathcal{D}$ over $\mathcal{X}$. Let p_b be an input parameter that is known to the learner such that $\mathbf{Pr}_{x\sim D}(c(x) = 1) \geq p_b > 0$. Then, for any $0 < \gamma \leq 1$, using L to generate an $h \in \mathcal{H}$ for which it holds $R_D(h,c) \leq \min\{\varepsilon, \gamma p_b\}$ implies for the same h that $\text{REC}_D(h,c) \geq 1-\gamma$. That is, L PAC learns $\mathcal{C}$ with high recall using $\mathcal{H}$.*

Proof. We want to produce a hypothesis $h \in \mathcal{H}$ that satisfies Definition 3.3 when ξ is omitted (since we do not care about precision in the statement). In other words, we want the hypothesis that L generates, to have recall $\text{REC}_D(h,c) \geq 1-\gamma$. By Proposition 3.1 it is enough if $1 - R_D(h,c)/p_b \geq 1-\gamma \Leftrightarrow R_D(h,c) \leq \gamma p_b$.

Hence, in order to satisfy the constraints $R_D(h,c) \leq \varepsilon$ and $\text{REC}_D(h,c) \geq 1-\gamma$ in Definition 3.3, it is enough if we require L to perform PAC learning according to Definition 2.3 and produce a hypothesis h that satisfies $R_D(h,c) \leq \min\{\varepsilon, \gamma p_b\}$.

Finally, since L satisfies Definition 2.3, there exists a polynomial $p(1/\varepsilon, 1/\delta, n, size(c))$ that provides enough samples for PAC learning $\mathcal{C}$ so that it holds $R_D(h,c) \leq \varepsilon$ at the end of the learning process. Therefore, the polynomial $p(1/(\min\{\varepsilon, \gamma p_b\}), 1/\delta, n, size(c))$ (or the even looser $p(1/(\varepsilon\gamma p_b), 1/\delta, n, size(c))$) is a good enough polynomial to justify Definition 3.3 for L (with ξ omitted). ∎

Similarly to Theorem 3.1, Theorem C.1 applies to situations where we may, or may not be working with realizable learning problems. However, combining Theorem C.1 and Theorem 2.2 we can obtain the following corollary for realizable learning problems.

COROLLARY C.1. *Let $\mathcal{H}$ be a hypothesis class with VC-dim$(\mathcal{H}) = d < \infty$. Let p_b be an input parameter that is known to the learner such that $\mathbf{Pr}_{x\sim D}(c(x) = 1) \geq p_b > 0$. Under the realizability assumption, a concept class $\mathcal{C}$ is PAC-learnable with high recall by $\mathcal{H}$ with sample complexity $m \leq \left\lceil \frac{4}{r}\left(d \lg\left(\frac{12}{r}\right) + \lg\left(\frac{2}{\delta}\right)\right)\right\rceil$. where, $r = \min\{\varepsilon, \gamma p_b\}$.*

Comparing Theorem 3.1 and Theorem C.1 First of all, in the proof of Theorem 3.1 we used $\gamma \leq 1/2$. Indeed, our interest in Definitions 2.3 and 3.3 is to understand the behavior of the sample size as the parameters that are associated with learning, approach 0. Hence, with our requirement $\gamma \leq 1/2$, Theorem 3.1 always provides the guarantee of a hypothesis $h \in \mathcal{H}$ that has recall at least 1/2. The reason we do this simplification is because otherwise, the sample size would depend inversely on the quantity $\xi(1-\gamma)p_b$ and such a polynomial dependence on $1/(1-\gamma)$ is not compatible with Definition 3.3. We believe that Definition 3.3 should remain as is and it is a question if a better bound can be derived for precision having the desired dependence with γ for every $0 < \gamma < 1$. (The issue arises for large values of γ, which imply that we have very weak requirements for the recall on the generated hypothesis.)

Second, requiring $\gamma \leq 1/2$ in Theorem 3.1 is perhaps natural and insignificant as one can trivially produce a solution h that achieves $\text{REC}_D(h,c) \geq 1/2$. For example, h can be a randomized predictor that tosses

a fair coin and decides about the label 0 or 1. Going one step further, one could also return the deterministic predictor h_1 that always returns 1 for every $x \in \mathcal{X}$. Clearly in this latter situation $\textsc{Rec}_D(h_1, c) = 1$ regardless of c. Of course the issue these two solutions have is that even if they provide 'high' recall, nevertheless they have potentially prohibitive risk. After all, we are dealing with imbalanced data where the probability of the positive (minority) class is typically some small constant close to 0. On the other hand though, along the same lines one can argue even for the vanilla version of PAC learning (Definition 2.3), where one can always generate a hypothesis that has risk at most $1/2$ by deterministically predicting the label of the majority class. Therefore, we argue that large values of risk, or low values of recall, as determined by a threshold of $1/2$, are of little importance, and thus our requirement of $\gamma \leq 1/2$ is largely inconsequential. In fact, this is the interesting and important case.

Finally, we remark that both theorems assume a lower bound for the quantity $\mathbf{Pr}_{x\sim D}(c(x) = 1)$, which is typically unknown to the learner. However, we have waived this requirement with the introduction of the pre-processing phase that is discussed in Section 3.3, which can also be applied to Theorem C.1.