A Remark About a Learning Risk Lower Bound

Man Fung Leung [a], Yiqi Lin [a] and Nicolas Wicker [b,*]

[a]The Chinese University of Hong Kong [b]Université de Lille
[*]Corresponding author email: nicolas.wicker@univ-lille.fr

ARTICLE HISTORY
Compiled June 17, 2022

ABSTRACT
In this paper, we correct the learning risk lower bound in the non-realizable case appearing in Anthony and Bartlett (1999) and Mohri *et al.* (2012). Our contribution is mainly technical as we follow closely the proof of Anthony and Bartlett by correcting first the lemma they use and adapting then the lower bound proof itself.

In supervised learning, given a sample S with $|S| = m$ over a set $\mathcal{X}$ with labels $y_1, \ldots, y_m$, the objective is to learn a concept, that is a function over $\mathcal{X}$ with let say values in $\mathcal{R}$. In the realizable case, this function exists in a set H which achieves zero expected error. Here we focus on the non-realizable case, where this function does not necessarily exist and the learning risk is then compared to the one obtained by the best possible function in $\mathcal{H}$. The risk $R_D(h)$ for a function h and a distribution D over $\mathcal{X} \times \{0,1\}$ is defined as $R_D(h) = \mathbb{E}_{(x,y)\sim D}\big(h(x) \neq y\big)$. Then, the main result stands as follows:

Theorem 1. *For a set of $\{0,1\}$-valued functions $\mathcal{H}$ over a set $\mathcal{X}$ with Vapnik-Chervonenkis dimension d, there exists a distribution D over $\mathcal{X} \times \{0,1\}$ verifying:*

$$P_{S\sim D^m}\left(R_D(h_S) - \inf_{h\in\mathcal{H}} R_D(h) \geq \frac{0.014}{\sqrt{\frac{m}{d} + \log\left(\frac{4}{3}\right)}} \wedge 0.01\right) \geq 0.01,$$

where h_S is the hypothesis inferred from $\mathcal{H}$ using a sample S of size m.

This is essentially theorem 5.2 in Anthony and Bartlett [1999] and theorem 3.7 in Mohri et al. [2012]. It has been outperformed recently by Kontorovich and Pinelis [2019]. However, the proof in the cited books is pretty straight forward while the counterpart in Kontorovich and Pinelis [2019] utilizes more advanced tools. We would like to set it right for didactical reasons using similar arguments in cited book directly.

To begin with, we first rectify a technical lemma used to prove the lower bound. The setting is as follows. In a game there are two coins with biases such that the probabilities of getting a head are p_- and p_+. The target is to guess which coin was thrown from the outcomes. Mathematically, let p be taken randomly in $\{p_-, p_+\}$ where $p_- = (1-\alpha)/2$ and $p_+ = (1+\alpha)/2$ with $\alpha \in (0,1)$. Then the m outcomes of the game $X_1, \ldots, X_m$ are i.i.d. observations of a Bernoulli random variable $B(1,p)$. As the goal is to guess which coin was thrown, a function $h(Y_m) \in \{p_-, p_+\}$ where $Y_m = \sum_{i=1}^m X_i$

can be built to guess p according to the outcomes. It goes without saying, that the more observations we have, the easier it is to guess the true p.

Before that, we recall Slud's inequality (Theorem 2.1 in Slud [1977]) briefly, which will be important to approximate the prediction error of $h(Y_m)$ in the game. To the best of our knowledge, it is also the sharpest binomial tail bound comparing with several alternatives in Telgarsky [2009].

Theorem 2 (Slud's Inequality)**.** *Let $Y_m \sim B(m, p)$. If k is an integer such that*

$$0 \leq p \leq \frac{1}{4} \text{ and } mp \leq k \leq m, \text{ or } 0 \leq p \leq \frac{1}{2} \text{ and } mp \leq k \leq m(1-p),$$

then

$$P(Y_m \geq k) \geq P\left(U \geq \frac{k - mp}{\sqrt{mp(1-p)}}\right), \text{ where } U \sim N(0,1).$$

Building on Slud's inequality, our Lemma 3 rectifies lemma 5.1 in Anthony and Bartlett [1999] and lemma 3.2 in Mohri et al. [2012] in the sense that we consider both odd and even number of outcomes m carefully. To be specific, when m is odd, $m/2$ is not an integer and the conditions in Slud's inequality may fail even with rounding. In both books, it is proposed in the odd case to bound $P(Y_m \geq m/2)$ by $P(Y_m \geq (m+1)/2)$, which is right initially. However, with k and p replaced by $(m+1)/2$ and $(1-\alpha)/2$ respectively, the condition $mp \leq k \leq m(1-p)$ leads to $1/m \leq \alpha$, which is not always verified particularly when this property is used in the proof of theorem of 5.2 in Anthony and Bartlett [1999] (or theorem 3.7 in Mohri et al. [2012]) and now in Theorem 1. We remark that when m is even, $m/2$ is an integer and $k = m/2$ leads to $m(1-\alpha)/2 \leq m/2 \leq m(1+\alpha)/2$, which holds automatically.

Therefore, to derive the correct learning risk lower bound, the following modified lemma is necessary with no restriction on m:

Lemma 3. *The prediction error in the aforementioned game satisfies:*

$$P(h(Y_m) \neq p) \geq \Phi(m, \alpha),$$

where

$$\Phi(m, \alpha) = \frac{1-\alpha}{8}\left[1 - \sqrt{1 - \exp\left(-\frac{m\alpha^2}{1-\alpha^2}\right)}\right].$$

Proof. Let $P(p = p_+) = P(p = p_-) = 1/2$ and $Y_m \sim B(m, p)$. Let the decision rule be $h(Y_m) = p_-$ if $Y_m < m/2$ and $h(Y_m) = p_+$ otherwise. First, we consider the case where m is even and obtain:

$$
\begin{aligned}
P(h(Y_m) \neq p) &= P(h(Y_m) \neq p_- \mid p = p_-)P(p = p_-) + P(h(Y_m) \neq p_+ \mid p = p_+)P(p = p_+) \\
&\geq \frac{1}{2}P(h(Y_m) \neq p_- \mid p = p_-) \\
&= \frac{1}{2}P\left(Y_m \geq \frac{m}{2} \mid p = p_-\right) \\
&\geq \frac{1}{2}P\left(U \geq \frac{\frac{m}{2} - mp_-}{\sqrt{mp_-(1-p_-)}}\right), \text{ where } U \sim N(0,1) \text{ by Theorem 2} \\
&\quad \text{as } \frac{1-\alpha}{2} \leq \frac{1}{2} \text{ and } m\frac{1-\alpha}{2} \leq \frac{m}{2} \leq m\frac{1+\alpha}{2} \\
&= \frac{1}{2}P\left(U \geq \frac{\frac{m}{2} - m\frac{1-\alpha}{2}}{\sqrt{m\frac{1-\alpha^2}{4}}}\right) \\
&= \frac{1}{2}P\left(U \geq \frac{\sqrt{m}\alpha}{\sqrt{1-\alpha^2}}\right).
\end{aligned}
$$

Using the tail inequality $P(U \geq u) \geq [1 - \sqrt{1 - \exp(-u^2)}]/2$ (see, e.g., Mohri et al. [2012]), we get:

$$
P(h(Y_m) \neq p) \geq \frac{1}{4}\left[1 - \sqrt{1 - \exp\left(-\frac{m\alpha^2}{1-\alpha^2}\right)}\right]. \tag{1}
$$

Next, if m is odd, we rewrite Y_m as

$$
Y_m = Y_{m-1} + X_m,
$$

where $Y_{m-1} \sim B(m-1, p)$ and $X_m \sim B(1, p)$. Then, since $m-1$ is even, we can apply Slud's inequality (Theorem 2) to Y_{m-1}. Therefore, for odd m we have

$$
\begin{aligned}
P(h(Y_m) \neq p) &\geq \frac{1}{2} P\left(Y_m \geq \frac{m}{2} \mid p = p_-\right) \\
&= \frac{1}{2} P\left(Y_{m-1} + X_m \geq \frac{m+1}{2} \mid p = p_-\right) \\
&= \frac{1}{2} P\left(Y_{m-1} \geq \frac{m+1}{2}, X_m = 0 \mid p = p_-\right) + \frac{1}{2} P\left(Y_{m-1} \geq \frac{m-1}{2}, X_m = 1 \mid p = p_-\right) \\
&= \frac{1+\alpha}{4} P\left(Y_{m-1} \geq \frac{m+1}{2} \mid p = p_-\right) + \frac{1-\alpha}{4} P\left(Y_{m-1} \geq \frac{m-1}{2} \mid p = p_-\right) \\
&\geq \frac{1-\alpha}{4} P\left(Y_{m-1} \geq \frac{m-1}{2} \mid p = p_-\right) \\
&\geq \frac{1-\alpha}{4} P\left(U \geq \frac{\frac{m-1}{2} - (m-1)\frac{1-\alpha}{2}}{\sqrt{(m-1)\frac{1-\alpha^2}{4}}}\right), \text{ using again Theorem 2} \\
&\quad \text{as } \frac{1-\alpha}{2} \leq \frac{1}{2} \text{ and } (m-1)\frac{1-\alpha}{2} \leq \frac{m-1}{2} \leq (m-1)\frac{1+\alpha}{2} \\
&= \frac{1-\alpha}{4} P\left(U \geq \frac{\sqrt{(m-1)}\alpha}{\sqrt{1-\alpha^2}}\right).
\end{aligned}
$$

We apply the tail inequality $P(U \geq u) \geq [1 - \sqrt{1 - \exp(-u^2)}]/2$ again to obtain

$$
P(h(Y_m) \neq p) \geq \frac{1-\alpha}{8}\left[1 - \sqrt{1 - \exp\left(-\frac{(m-1)\alpha^2}{1-\alpha^2}\right)}\right]. \tag{2}
$$

Gathering (1) and (2) leads to:

$$
P(h(Y_m) \neq p) \geq \frac{1-\alpha}{8}\left[1 - \sqrt{1 - \exp\left(-\frac{m\alpha^2}{1-\alpha^2}\right)}\right].
$$

□

Before proving the main proposition, we need one more lemma to lower bound the tail probability. This lemma is part of lemma 3.3 in Mohri et al. [2012], which we restate here:

Lemma 4. *For a random variable Z taking values in $[0,1]$ and a constant $\gamma \in [0,1)$,*

$$
P(Z > \gamma) \geq \mathbb{E}(Z) - \gamma.
$$

Proof.

$$\begin{aligned}
\mathbb{E}(Z) &= \int_{z\leq\gamma} z\mathrm{d}P(z) + \int_{z>\gamma} z\mathrm{d}P(z) \\
&\leq \int_{z\leq\gamma} \gamma\mathrm{d}P(Z) + \int_{z>\gamma} \mathrm{d}P(Z) \text{ as } z \leq 1 \\
&\leq \gamma P(Z \leq \gamma) + P(Z > \gamma) \\
&\leq \gamma + (1-\gamma)P(Z > \gamma).
\end{aligned}$$

Hence, as $\gamma \in [0, 1)$,

$$P(Z > \gamma) \geq \frac{\mathbb{E}(Z) - \gamma}{1-\gamma} \geq \mathbb{E}(Z) - \gamma.$$

□

Now we can prove Theorem 1. We show that for a class of function $\mathcal{H}$ over $\mathcal{X}$ of Vapnik-Chervonenkis dimension d, the difference between the optimal function risk in $\mathcal{F}$ and the Bayes classifier risk can be lower bounded by a constant. More specifically, we show that there exists a distribution over $\mathcal{X} \times \{0, 1\}$ such that this difference is bounded. The proof follows the one in Anthony and Bartlett [1999] using the probabilistic method pioneered by Erdös, with the only difference in its use of the modified lemma.

Theorem 1. *For a set of $\{0, 1\}$-valued functions $\mathcal{H}$ over a set $\mathcal{X}$ with Vapnik-Chervonenkis dimension d, there exists a distribution D over $\mathcal{X} \times \{0, 1\}$ verifying:*

$$P_{S\sim D^m}\left(R_D(h_S) - \inf_{h\in\mathcal{H}} R_D(h) \geq \frac{0.014}{\sqrt{\frac{m}{d} + \log\left(\frac{4}{3}\right)}} \wedge 0.01\right) \geq 0.01,$$

where h_S is the hypothesis inferred from $\mathcal{H}$ using a sample S of size m.

Proof. Let us consider a set $\mathcal{X} = \{x_1, \ldots, x_d\}$ shattered by $\mathcal{H}$, which existence is possible given that the Vapnik-Chervonenkis dimension of $\mathcal{H}$ is d. We define a distribution $D_{\boldsymbol{\sigma}}$ over $\mathcal{X} \times \{0, 1\}$ satisfying:

$$\begin{aligned}
P_{D_{\boldsymbol{\sigma}}}(x_i, 1) &= \frac{1}{d}(\frac{1}{2} + \frac{\sigma_i\alpha}{2}), \\
P_{D_{\boldsymbol{\sigma}}}(x_i, 0) &= \frac{1}{d}(\frac{1}{2} - \frac{\sigma_i\alpha}{2}),
\end{aligned}$$

with $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_d) \in \{-1, 1\}^d$. For a given $\boldsymbol{\sigma}$, we will sample d labels in a biased way either towards 1 or towards 0 and compare the Bayes classifier $h^*_{D_{\boldsymbol{\sigma}}} = \mathrm{argmax}_{y\in\{0,1\}} P_{D_{\boldsymbol{\sigma}}}(x_i, y) = \mathbb{1}_{\sigma_i>0}$ with any classifier $h \in \mathcal{H}$,

$$R_{D_{\boldsymbol{\sigma}}}(h) - R_{D_{\boldsymbol{\sigma}}}(h^*_{D_{\boldsymbol{\sigma}}}) = \frac{1}{d}\sum_{i=1}^{d}\Big(\frac{1}{2} + \frac{\sigma_i\alpha}{2}\Big)\Big(\mathbb{1}_{h(x_i)\neq 1} - \mathbb{1}_{\sigma_i<0}\Big) + \frac{1}{d}\sum_{i=1}^{d}\Big(\frac{1}{2} - \frac{\sigma_i\alpha}{2}\Big)\Big(\mathbb{1}_{h(x_i)\neq 0} - \mathbb{1}_{\sigma_i>0}\Big).$$

Indeed, that $\mathcal{X}$ can be shattered by $\mathcal{H}$ gives rise to $h^*_{D_{\boldsymbol{\sigma}}} \in \mathcal{H}$. The right-hand side can be rewritten as:

$$\frac{1}{d}\sum_{i=1}^{d}\left[\frac{\sigma_i\alpha}{2}\left(\mathbb{1}_{h(x_i)\neq 1} - \mathbb{1}_{\sigma_i<0}\right) - \frac{\sigma_i\alpha}{2}\left(\mathbb{1}_{h(x_i)\neq 0} - \mathbb{1}_{\sigma_i>0}\right)\right].$$

Looking separately at each term, we get:

- If $\sigma_i = 1$:

$$\frac{\alpha}{2}\mathbb{1}_{h(x_i)\neq h^*_{D_{\boldsymbol{\sigma}}}(x_i)} + \frac{\alpha}{2}\mathbb{1}_{h(x_i)\neq h^*_{D_{\boldsymbol{\sigma}}}(x_i)}.$$

- If $\sigma_i = -1$:

$$\frac{-\alpha}{2}(-1)\mathbb{1}_{h(x_i)\neq h^*_{D_{\boldsymbol{\sigma}}}(x_i)} + \frac{\alpha}{2}\mathbb{1}_{h(x_i)\neq h^*_{D_{\boldsymbol{\sigma}}}(x_i)}.$$

Since they are the same, we have:

$$R_{D_{\boldsymbol{\sigma}}}(h) - R_{D_{\boldsymbol{\sigma}}}(h^*_{D_{\boldsymbol{\sigma}}}) = \frac{\alpha}{d}\sum_{i=1}^{d}\mathbb{1}_{h(x_i)\neq h^*_{D_{\boldsymbol{\sigma}}}(x_i)} = \frac{\alpha}{d}\sum_{x\in\mathcal{X}}\mathbb{1}_{h(x)\neq h^*_{D_{\boldsymbol{\sigma}}}(x)}.$$

As this works for any $\boldsymbol{\sigma}$, with U denoting the uniform distribution over $\{-1,1\}^d$ and h_S denoting the hypothesis inferred from S:

$$\begin{aligned}
\mathbb{E}_{\substack{\boldsymbol{\sigma}\sim U\\ S\sim D^m_{\boldsymbol{\sigma}}}}\left[R_{D_{\boldsymbol{\sigma}}}(h_S) - R_{D_\sigma}(h^*_{D_{\boldsymbol{\sigma}}})\right] =& \frac{\alpha}{d}\sum_{x\in\mathcal{X}}\mathbb{E}_{\substack{\boldsymbol{\sigma}\sim U\\ S\sim D^m_{\boldsymbol{\sigma}}}}\left(\mathbb{1}_{h_S(x)\neq h^*_{D_{\boldsymbol{\sigma}}}(x)}\right)\\
=& \frac{\alpha}{d}\sum_{x\in\mathcal{X}}\mathbb{E}_{\boldsymbol{\sigma}\sim U}\left[P_{S\sim D^m_{\boldsymbol{\sigma}}}(h_S(x)\neq h^*_{D_{\boldsymbol{\sigma}}}(x))\right]\\
=& \frac{\alpha}{d}\sum_{x\in\mathcal{X}}\sum_{n=0}^{m}\mathbb{E}_{\boldsymbol{\sigma}\sim U}\left[P_{S\sim D^m_{\boldsymbol{\sigma}}}(h_S(x)\neq h^*_{D_{\boldsymbol{\sigma}}}(x)) \;\middle|\; |S|_x = n\right]P(|S|_x = n),
\end{aligned}$$

where $|S|_x$ stands for the number of occurences of x in S. The shattering of $\mathcal{X}$ implies here that the labels of S can be chosen among the 2^d possibilities with no restriction on h_S so we can use Lemma 3 inverting if necessary the roles of 0 and 1 and Jensen's inequality:

$$\begin{aligned}
\mathbb{E}_{\substack{\boldsymbol{\sigma}\sim U\\ S\sim D^m_{\boldsymbol{\sigma}}}}\left[R_{D_{\boldsymbol{\sigma}}}(h_S) - R_{D_{\boldsymbol{\sigma}}}(h^*_{D_{\boldsymbol{\sigma}}})\right] \geq& \frac{\alpha}{d}\sum_{x\in\mathcal{X}}\sum_{n=0}^{m}\Phi(n,\alpha)P(|S|_x = n)\\
\geq& \frac{\alpha}{d}\sum_{x\in\mathcal{X}}\Phi(m/d,\alpha) \text{ by convexity of } \Phi(.,\alpha) \text{ and}\\
& x \text{ is picked uniformly in } \mathcal{X}\\
=& \alpha\Phi(m/d,\alpha).
\end{aligned}$$

As this is true for expectation with respect to $\boldsymbol{\sigma} \sim U$, there exists some $\boldsymbol{\sigma'}$ verifying:

$$\begin{aligned}\mathbb{E}_{S\sim D^m_{\boldsymbol{\sigma}'}}\Big[R_{D_{\boldsymbol{\sigma}'}}(h_S) - R_{D_{\boldsymbol{\sigma}'}}(h^*_{D_{\boldsymbol{\sigma}'}})\Big] \geq & \alpha\Phi(m/d,\alpha) \\ \geq & \alpha\frac{1-\alpha}{8}\left[1-\sqrt{1-\exp\left(-\frac{(m/d)\alpha^2}{1-\alpha^2}\right)}\right].\end{aligned}$$

By Lemma 4, we have:

$$P_{S\sim D^m_{\boldsymbol{\sigma}'}}\Big[\frac{1}{\alpha}\big(R_{D_{\boldsymbol{\sigma}'}}(h_S) - R_{D_{\boldsymbol{\sigma}'}}(h^*_{D_{\boldsymbol{\sigma}'}})\big) > \gamma u\Big] > (1-\gamma)u,$$

where

$$u = \frac{1-\alpha}{8}\left[1-\sqrt{1-\exp\Big(-\frac{(m/d)\alpha^2}{1-\alpha^2}\Big)}\right],$$

provided that $\gamma u < 1$ to comply with Lemma 4, which will be checked later.

Now, we want our final result to be of the following form:

$$P_{S\sim D^m_{\boldsymbol{\sigma}'}}\Big[\big(R_{D_{\boldsymbol{\sigma}'}}(h_S) - R_{D_{\boldsymbol{\sigma}'}}(h^*_{D_{\boldsymbol{\sigma}'}})\big) > \epsilon\Big] > \delta, \tag{3}$$

so that it is coherent with the form in Anthony and Bartlett [1999] and Mohri et al. [2012]. Therefore, we look for α, γ and δ such that: $\alpha\gamma u \geq \epsilon$ and $(1-\gamma)u \geq \delta$. This leads to the sufficient condition:

$$\frac{1-\alpha}{8}\left[1-\sqrt{1-\exp\left(-\frac{(m/d)\alpha^2}{1-\alpha^2}\right)}\right] \geq \frac{\delta}{1-\gamma}. \tag{4}$$

Choosing $\gamma = 1 - 16\delta/(1-\alpha)$ leads to:

$$\begin{aligned}&\frac{1-\alpha}{8}\left[1-\sqrt{1-\exp\left(-\frac{(m/d)\alpha^2}{1-\alpha^2}\right)}\right] \geq \frac{1-\alpha}{16} \\ &\Leftrightarrow -\sqrt{1-\exp\Big\{-\frac{(m/d)\alpha^2}{1-\alpha^2}\Big\}} \geq -\frac{1}{2} \\ &\Leftrightarrow 1-\exp\Big\{-\frac{(m/d)\alpha^2}{1-\alpha^2}\Big\} \leq \frac{1}{4} \\ &\Leftrightarrow \exp\Big\{-\frac{(m/d)\alpha^2}{1-\alpha^2}\Big\} \geq \frac{3}{4} \\ &\Leftrightarrow \frac{(m/d)\alpha^2}{1-\alpha^2} \leq \log\Big(\frac{4}{3}\Big) \\ &\Leftrightarrow \frac{m}{d} \leq \Big(\frac{1}{\alpha^2}-1\Big)\log\Big(\frac{4}{3}\Big).\end{aligned} \tag{5}$$

Besides, as $(1-\gamma)u \geq \delta$,

$$\alpha\gamma u \geq \frac{\delta}{1-\gamma}\alpha\gamma = \frac{\alpha}{\frac{16}{1-\alpha}}\Big(1 - \frac{16\delta}{1-\alpha}\Big) = \frac{(1-\alpha)\alpha}{16} - \alpha\delta.$$

This should be larger than ϵ, so we can study the inequality:

$$-\frac{\alpha^2}{16} + \alpha\Big(\frac{1}{16} - \delta\Big) - \epsilon \geq 0.$$

Then, α should stay in the interval $[\alpha', \alpha'']$ where

$$\alpha' = \frac{\delta - \frac{1}{16} + \big(\frac{1}{16} - \delta\big)\sqrt{1 - \frac{\epsilon}{4\left(\frac{1}{16}-\delta\right)^2}}}{-1/8} \text{ provided that } 4\Big(\frac{1}{16} - \delta\Big)^2 - \epsilon \geq 0 \text{ and } \frac{1}{16} > \delta \tag{6}$$

and

$$\alpha'' = \frac{\delta - \frac{1}{16} - \big(\frac{1}{16} - \delta\big)\sqrt{1 - \frac{\epsilon}{4\left(\frac{1}{16}-\delta\right)^2}}}{-1/8} \text{ with the same conditions.}$$

Using the inequality $\sqrt{1-x} \geq 1 - x$ for $x \in [0,1]$,

$$\alpha' \leq \frac{\delta - \frac{1}{16} + \frac{1}{16} - \delta - \frac{\epsilon}{4\left(\frac{1}{16}-\delta\right)}}{-1/8} = \frac{2\epsilon}{\frac{1}{16} - \delta}$$

and

$$\alpha'' \geq \frac{\delta - \frac{1}{16} - \big(\frac{1}{16} - \delta\big) + \frac{\epsilon}{4\left(\frac{1}{16}-\delta\right)}}{-1/8} = 16\Big(\frac{1}{16} - \delta\Big) - \frac{2\epsilon}{\frac{1}{16} - \delta}.$$

We take then $\alpha = 2\epsilon/(1/16 - \delta)$ and check

$$\alpha = \frac{2\epsilon}{\frac{1}{16} - \delta} \leq 16\Big(\frac{1}{16} - \delta\Big) - \frac{2\epsilon}{\frac{1}{16} - \delta}, \tag{7}$$

so as to ensure that $\alpha \leq \alpha''$. Inequality 7 yields:

$$\epsilon \leq 4\Big(\frac{1}{16} - \delta\Big)^2 \text{ which is the same as inequality 6.}$$

Consequently inequality 5 becomes:

$$\frac{m}{d} \leq \Big(\frac{\big(\frac{1}{16} - \delta\big)^2}{4\epsilon^2} - 1\Big) \log(\frac{4}{3})$$
$$\Leftrightarrow \epsilon \leq \frac{\big(\frac{1}{16} - \delta\big)\sqrt{\log\big(\frac{4}{3}\big)}}{2\sqrt{\frac{m}{d} + \log\big(\frac{4}{3}\big)}}.$$

Now, we take $\delta = 0.01$. By condition 6, $\epsilon \leq 4\big(1/16 - \delta\big)^2 = 0.011025$. Consequently, $\alpha = 2\epsilon/(1/16 - \delta) \in [0, 0.42]$ and $\gamma = 1 - 16\delta/(1 - \alpha) \in [0.73, 0.84]$. which confirm that α and γ stay in the interval $[0, 1]$ as well as $\gamma u < 1$. The conclusion follows as:

$$\frac{\big(\frac{1}{16} - \delta\big)\sqrt{\log\big(\frac{4}{3}\big)}}{2\sqrt{\frac{m}{d} + \log\big(\frac{4}{3}\big)}} = \frac{0.014}{\sqrt{\frac{m}{d} + \log\big(\frac{4}{3}\big)}} \text{ for } \delta = 0.01.$$

□

In essence, the result is the same as the one of Anthony and Bartlett, which shows that the Vapnik dimension is the crucial parameter of a class of functions and that the number of instances in the learner should be at least of the same order of magnitude.

Acknowledgements

N. Wicker wishes to thank CUHK for welcoming him as visiting professor as part of this work has been done there. The authors also wish to thank the following persons who have contributed to this research as a working group: Sihan Chen, Hui Gong, Adrian Patrick Kennedy, Sze Him Isaac Leung, Guohao Shen, Ling Wang and Boning Yang.

References

M. Anthony and P.L. Bartlett. *Neural Network Learning: Theoretical Foundations.* Cambridge University Press, 1999.

A. Kontorovich and I. Pinelis. Exact lower bounds for the agnostic probably-approximately-correct (pac) machine learning model. *The Annals of Statistics*, 47(5):2822–2854, 2019.

M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning.* MIT press, 2012.

E. V. Slud. Distribution inequalities for the binomial law. *The Annals of Probability*, 5(3): 404–412, 1977.

M. Telgarsky. Central binomial tail bounds. *arXiv e-prints*, 2009.