Statistical learning theory and linguistic typology: a learnability perspective on OT's strict domination

Online supplement

Émile Enguehard emile.enguehard@ens.fr
Edward Flemming flemming@mit.edu
Giorgio Magri magrigrg@gmail.com

This note contains a proof of Corollary 1 on page 9 of the paper, repeated below. The corollary is stated as an example in Koltchinskii and Panchenko (2005, p. 1465). The proof of the corollary is omitted in that paper but has been provided in unpublished class notes by Panchenko (2004, class 21), that we follow here.

Corollary 1. *Consider a classifier $f = \sum_{i=1}^{K} w_k h_k$ in $\mathcal{F}$ which classifies correctly a training set $T = ((x_1, y_1), \ldots, (x_n, y_n))$ with margin $\delta = \delta_T(f)$, namely $y_1 f(x_1), \ldots, y_n f(x_n) > \delta$.*

- *If the weights w_k decay polynomially, i.e. $w_k \leq k^{-B}$ for some $B > 1$, KP's bound (8) becomes:*

$$Err_{\mathbb{P}}(f) \leq K \left(\frac{C_B}{\delta^{2/(2B-1)}} \frac{V}{n} \log^2 \frac{n}{\delta} + \frac{t}{n} \right) \quad (1)$$

where $C_B \to 1$ as $B \to \infty$.

- *If the weights w_k decay exponentially, namely $w_k \leq e^{-k}$, KP's bound (8) becomes:*

$$Err_{\mathbb{P}}(f) \leq K \left(\frac{V}{n} \log^2 \frac{n}{\delta} + \frac{t}{n} \right) \quad (2)$$

Proof. Recall that the effective dimension $d_T(f)$ of a classifier $f = \sum_{i=1}^{K} w_k h_k \in \mathcal{F} = conv(\mathcal{H})$ which succeeds on a training set T of cardinality n with a margin of confidence $\delta_T(f) = \delta$ is defined as follows:

$$d_T(f) = \min_{0 \leq d \leq K} \left[d + \left(\sum_{j=d+1}^{K} w_j \right)^2 \frac{2 \log n}{\delta^2} \right]$$

Suppose that the weights w_k in the representation of f decay polynomially, i.e. $w_k \leq k^{-B}$ for some $B > 1$. For any integer $d \in \{1, \ldots, K-1\}$, the following chain of inequalities thus holds. Step (a) uses the assumption $w_k \leq k^{-B}$. Step (b) uses a well-know integral upper bound for finite series with a decreasing summand (see for instance Cormen *et al.* 1990, appendix A.2). Step (c) uses the identity $\int x^\alpha \mathrm{d}x = x^{\alpha+1}/(\alpha+1)$.

$$\begin{aligned} \sum_{k=d+1}^{K} w_k &\overset{(a)}{\leq} \sum_{k=d+1}^{K} k^{-B} \\ &\overset{(b)}{\leq} \int_{d}^{\infty} x^{-B} \mathrm{d}x \\ &\overset{(c)}{=} \frac{1}{B-1} \left(\frac{1}{d} \right)^{B-1} \end{aligned}$$

The effective dimension of the classifier f can therefore be bounded as follows:

$$d_T(f) \leq \min_{0 \leq d \leq T} \underbrace{\left[d + \frac{1}{(B-1)^2} \left(\frac{1}{d} \right)^{2(B-1)} \frac{2 \log n}{\delta^2} \right]}_{F(d)}$$

By setting the derivative of the function $F(d)$ equal to zero, we obtain:

$$\begin{aligned} F'(d) = 0 &\iff 1 - \frac{2}{B-1} \frac{1}{d^{2B-1}} \frac{2 \log n}{\delta^2} = 0 \\ &\iff \frac{1}{d^{2B-1}} = \frac{B-1}{2} \frac{\delta^2}{2 \log n} \\ &\iff d = D_B \left(\frac{2 \log n}{\delta^2} \right)^{\frac{1}{2B-1}} \end{aligned}$$

where we have used the position:

$$D_B = \left(\frac{2}{B-1} \right)^{\frac{1}{2B-1}}$$

The effective dimension of the classifier f can therefore be bounded further as follows:

$$\begin{aligned} d_T(f) \leq \\ &\leq D_B \left(\frac{2\log n}{\delta^2}\right)^{\frac{1}{2B-1}} + \\ &\quad + \frac{C^{-2(B-1)}}{(B-1)^2}\left(\frac{2\log n}{\delta^2}\right)^{\frac{-2(B-1)}{2B-1}} \frac{2\log n}{\delta^2} \\ &= D_B \left(\frac{2\log n}{\delta^2}\right)^{\frac{1}{2B-1}} + \\ &\quad + \frac{C^{-2(B-1)}}{(B-1)^2}\left(\frac{2\log n}{\delta^2}\right)^{\frac{1}{2B-1}} \\ &= C_B \left(\frac{\log n}{\delta^2}\right)^{\frac{1}{2B-1}} \\ &\leq C_B \log n \left(\frac{1}{\delta^2}\right)^{\frac{1}{2B-1}} \end{aligned}$$

where we have used the position:

$$C_B = 2^{\frac{1}{2B-1}} \left\{ D_B + \frac{D_B^{-2(B-1)}}{(B-1)^2} \right\}$$

Plugging this bound on the effective dimension into the general expression (8) of KP's bound, we obtain:

$$\begin{aligned} Err_{\mathbb{P}}(f) &\leq K\left(d_T(f)\log\frac{n}{\delta}\frac{V}{n} + \frac{t}{n}\right) \\ &\leq K\left(\frac{C_B\log n}{\delta^{2/(2B-1)}}\log\frac{n}{\delta}\frac{V}{n} + \frac{t}{n}\right) \\ &\leq K\left(\frac{C_B}{\delta^{2/(2B-1)}}\log^2\frac{n}{\delta}\frac{V}{n} + \frac{t}{n}\right) \end{aligned}$$

Finally, we note that $D_B \to 1$ when $B \to \infty$, in fact:

$$\log D_B = \frac{\log 2}{2B-1} - \frac{\log(B-1)}{2B-1} \to 0$$

Furthermore, $\frac{D_B^{-2(B-1)}}{(B-1)^2} \to 0$ when $B \to \infty$, in fact:

$$\begin{aligned} \log \frac{D_B^{-2(B-1)}}{(B-1)^2} = \\ = \quad -\frac{2B-2}{2B-1}\log 2 - \frac{2B}{2B-1}\log(B-1) \to -\infty \end{aligned}$$

Thus, $C_B \to 1$ as $B \to \infty$.

Suppose next that the weights w_k in the representation of the classifier $f = \sum_{i=1}^K w_k h_k$ decay exponentially, namely $w_k \leq e^{-k}$. For any integer $d \in \{0, \ldots, K-1\}$, the following chain of inequalities thus holds. Step (c) uses the identity $\int e^{\alpha x} dx = e^{\alpha x}/\alpha$.

$$\begin{aligned} \sum_{k=d+1}^K w_k &\leq \sum_{k=d+1}^K e^{-k} \\ &\leq \int_d^\infty e^{-x}\mathrm{d}x \\ &= e^{-d} \end{aligned}$$

The effective dimension of the classier f can therefore be bounded as follows:

$$d_T(f) \leq \min_{0\leq d\leq T} \underbrace{\left[d + e^{-2d}\frac{2\log n}{\delta^2}\right]}_{F(d)}$$

By setting the derivative of the function $F(d)$ equal to zero, we obtain:

$$\begin{aligned} F'(d) = 0 \quad &\iff \quad 1 - 2\frac{2\log n}{\delta^2}e^{-2d} = 0 \\ &\iff \quad e^{-2d} = \frac{\delta^2}{4\log n} \\ &\iff \quad d = -\frac{1}{2}\log\frac{\delta^2}{4\log n} \\ &\iff \quad d = \log\frac{2\sqrt{\log n}}{\delta} \end{aligned}$$

The effective dimension of the classifier f can therefore be bounded further as follows:

$$\begin{aligned} d_T(f) &\leq \log\frac{2\sqrt{\log n}}{\delta} + \frac{\delta^2}{4\log n}\frac{2\log n}{\delta^2} \\ &= \log\frac{2\sqrt{\log n}}{\delta} + \frac{1}{2} \simeq \log\frac{2\sqrt{\log n}}{\delta} \end{aligned}$$

Plugging this bound on the effective dimension into the general expression (8) of KP's bound, we obtain:

$$\begin{aligned} Err_{\mathbb{P}}(f) &\leq K\left(d_T(f)\log\frac{n}{\delta}\frac{V}{n} + \frac{t}{n}\right) \\ &\leq K\left[\log\frac{2\sqrt{\log n}}{\delta}\log\frac{n}{\delta}\frac{V}{n} + \frac{t}{n}\right] \\ &\leq K\left[\log^2\frac{n}{\delta}\frac{V}{n} + \frac{t}{n}\right] \end{aligned}$$

concluding the proof of the corollary. □

References

Thomas Cormen, Charles Leiserson, Ronald Rivest, and Clifford Stein. 1990. *Introduction to Algorithms*. MIT Press, Cambridge, MA. Third edition.

Vladimir Koltchinskii and Dmitry Panchenko. 2005. Complexities of convex combinations and bounding the generalization error in classification. *Ann. Statist.*, 33.4:1455–1496.

Dmitry Panchenko. 2004. Statistical learning theory. Lecture notes for the class 18.465 (Topics in Statistics), Department of Mathematics, MIT.