Rademacher complexity of k-fold maxima of hyperplanes

Aryeh Kontorovich

September 17, 2018

Let $B \subset \mathbb{R}^d$ be the Euclidean d-dimensional unit ball. It is well-known (see, e.g., [12, Theorem 4.3]) that for any $x_1, \ldots, x_n \in B$, we have the following upper bound on the Rademacher complexity

$$R_n := \frac{1}{n}\mathbb{E} \sup_{w \in B} \sum_{i=1}^{n} \sigma_i (w \cdot x_i) \le \frac{1}{\sqrt{n}},$$

where the expectation is over the Rademacher sequence σ, distributed uniformly in $\{-1, 1\}^n$. Suppose we are interested in the following quantity:

$$R_{n,k} := \frac{1}{n}\mathbb{E} \sup_{w_1, w_2, \ldots, w_k \in B} \sum_{i=1}^{n} \sigma_i \max\{w_1 \cdot x_i, w_2 \cdot x_i, \ldots, w_k \cdot x_i\}.$$

It was observed[1] in [9] that trivially, $R_{n,k} = O(k/\sqrt{n})$, and the question of the optimal dependence on k was posed. It was pointed out by F. Nazarov therein that $R_{n,k} = \Omega(\sqrt{k/n})$. In this note, we prove the following estimate:

$$R_{n,k} = O\left(\sqrt{\frac{k \log k \cdot (\log n)^3}{n}}\right). \tag{1}$$

An analogous bound with additional logarithmic factors has been obtained by S. Mendelson and in [6, Lemma 6]; both communicated to us by Sasha Rakhlin [13].

The $\log k$ factor in (1), in general, cannot be removed — as recently shown by [4]. It remains an open question whether the $\log n$ is necessary.

[1] Actually, the inner max in [9] was replaced by min but the symmetry of B makes it immaterial.

Bounding fat-shattering dimension via the VC-dimension. As a warm-up, let's start with the family F of homogeneous hyperplanes, i.e., functions of the form $f_w : x \mapsto w \cdot x$, where $x, w \in B$. It is a classic fact (see [2, Theorem 4.6]) that $\mathrm{fat}_\gamma(F) \le 1/\gamma^2$. We now give an alternate and perhaps simpler proof, which will also be useful in the sequel. Suppose that F is able to γ-shatter some set $S = \{x_1, \ldots, x_\ell\}$. Now in general, this means that there is some $r \in \mathbb{R}^\ell$ such that for each $y \in \{-1, 1\}^\ell$ there is an $f_w \in F$ such that

$$\min_{i \in [\ell]} y_i (w \cdot x_i - r_i) \ge \gamma. \tag{2}$$

However, for the specific case of hyperplanes, we can take $r \equiv 0$ in (2) without loss of generality — this is the *shattering at zero* notion defined and employed in [7, Lemma 2]. Renormalizing the F to enforce $|w \cdot x| \ge 1$ for all $x \in S$ we obtain $\tilde{F} = \{x \mapsto w \cdot x : x \in S, \|w\| \le 1/\gamma\}$. Now S is γ-shattered by F iff S is 1-shattered by $\tilde{F}$. Define further $\bar{F} = \mathrm{sgn}(\tilde{F})$; now if $\tilde{F}$ 1-shatters S then certainly $\bar{F}$ shatters S in the ordinary VC sense. Since $\|x\| \le 1$ for all $x \in S$ and $\|w\| \le 1/\gamma$ for all $f_w \in \tilde{F}$, a standard argument (see, e.g., [12, Theorem 4.2] or [8, Lemma 6]) yields $|S| \le 1/\gamma^2$. This proves that $\mathrm{fat}_\gamma(F) \le 1/\gamma^2$ via reduction to a VC argument.

Extending to unions of fat hyperplanes. Now define the function class F_k on $B \subset \mathbb{R}^d$ as follows. Each $f \in F_k$ is parametrized by $(w_1, \ldots, w_k) \in B^k$ and acts on $x \in B$ by:

$$f(x) = \max_{j \in [k]} w_j \cdot x.$$

Suppose that F_k γ-shatters some $S = \{x_i, i \in [\ell]\}$. Shattering (at zero, with $r \equiv 0$) implies that for each $y \in \{-1, 1\}^\ell$ there is a $(w_1, \ldots, w_k) \in B^k$ such that

$$\min_{i \in [\ell]} y_i \max_{j \in [k]} w_j \cdot x_i \ge \gamma.$$

Renormalizing by γ as before, we obtain the function class $\tilde{F}_k$, which 1-shatters S. Taking signs as before (note that the sgn and max operators commute), we convert $\tilde{F}_k$ to the binary function class $\bar{F}_k$. But now observe that $\bar{F}_k$ is precisely the concept class obtained by k-fold unions of half-spaces. Formally, let $\bar{F}^{(j)}$ be the concept class defined by $\{x \mapsto \mathrm{sgn}(w \cdot x) : x \in S, \|w\| \le 1/\gamma\}$ (thus all of the $\bar{F}^{(j)}$'s are the same). Then a concept C belongs to $\bar{F}_k$ iff

$C = C_1 \cup C_2 \cup \ldots \cup C_k$, where $C_j \in F^{(j)}$. From above, we know that the VC-dimension of $F^{(j)}$ is at most $1/\gamma^2$. Now we invoke [3, Lemma 3.2.3]: the class of k-fold unions of concepts taken from a concept class C has VC-dimension less than $2k\,\mathrm{VC}(C)\log(3k)$. We conclude that

$$\mathrm{fat}_\gamma(F_k) \le \frac{2k\log(3k)}{\gamma^2}. \tag{3}$$

From fat-shattering to Rademacher. As pointed out to us by Sasha Rakhlin [13], the fat-shattering estimate above can be used to upper-bound $R_{n,k}$ by converting the former to a covering number bound and plugging it into Dudley's chaining integral [5]:

$$R_{n,k}(F) \le \inf_{\alpha\ge 0}\left(4\alpha + 12\int_\alpha^\infty \sqrt{\frac{\log \mathcal{N}(t, F, \|\cdot\|_2)}{n}}dt\right), \tag{4}$$

where $\mathcal{N}(\cdot)$ are the ℓ_2 covering numbers. A proof of this result as stated in (4) may be found in [10].

It remains to bound the covering numbers. A simple way of doing so is to invoke Lemmas 2.6, 3.2, and 3.3 in [1] — but this incurs superfluous logarithmic factors in n. Instead, we opt for a more advanced tool: [11, Theorem 1] states that

$$\mathcal{N}(t, F, \|\cdot\|_2) \le \left(\frac{2}{t}\right)^{K\,\mathrm{fat}_{ct}(F)}$$

for all F bounded by 1, where K, c are universal constants. Plugging this into (4), we get

$$\begin{aligned} R_{n,k}(F_{n,k}) &\le \inf_{\alpha\ge 0}\left(4\alpha + 12\int_\alpha^1 \sqrt{\frac{\log \mathcal{N}(t, F, \|\cdot\|_2)}{n}}dt\right) \\ &\le \inf_{\alpha\ge 0}\left(4\alpha + 12c'\sqrt{\frac{2k\log(3k)}{n}}\int_\alpha^1 \frac{1}{t}\sqrt{\log\frac{2}{t}}dt\right) \end{aligned}$$

where $c' = \sqrt{K}/c$. Now

$$\int_\alpha^1 \frac{1}{t}\sqrt{\log\frac{2}{t}}dt = \frac{2}{3}\left(\log(2/\alpha)^{3/2} - (\log 2)^{3/2}\right)$$

and choosing $\alpha = 1/\sqrt{n}$ yields

$$\begin{aligned} R_{n,k}(F_{n,k}) &\leq \frac{4}{\sqrt{n}} + 8c'\sqrt{\frac{2k\log(3k)}{n}}\left(\log(2\sqrt{n})^{3/2} - (\log 2)^{3/2}\right) \\ &= O\left(\sqrt{\frac{k\log k \cdot (\log n)^3}{n}}\right). \end{aligned}$$

Acknowledgement. Thanks to Idan Attias for pointing out a mistake in a previous calculation.

References

[1] Noga Alon, Shai Ben-David, Nicolò Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44(4):615–631, 1997.

[2] Peter Bartlett and John Shawe-Taylor. *Generalization performance of support vector machines and other pattern classifiers*, pages 43–54. MIT Press, Cambridge, MA, USA, 1999.

[3] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. Assoc. Comput. Mach.*, 36(4):929–965, 1989.

[4] Monika Csikos, Andrey Kupavskii, and Nabil H. Mustafa. Optimal bounds on the vc-dimension. *CoRR*, abs/1807.07924, 2018.

[5] R.M Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290 – 330, 1967.

[6] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *J. Amer. Math. Soc.*, 29(4):983–1049, 2016.

[7] Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Efficient regression in metric spaces via approximate lipschitz extension. *IEEE Transactions on Information Theory*, 63(8):4838–4849, Aug 2017.

[8] Steve Hanneke and Aryeh Kontorovich. Optimality of SVM: Novel proofs and tighter bounds. 2017.

[9] Aryeh Kontorovich (https://mathoverflow.net/users/12518/aryeh kontorovich). Rademacher average involving minima. MathOverflow. URL:https://mathoverflow.net/q/266457 (version: 2017-04-05).

[10] Sham Kakade and Ambuj Tewari. Dudley's theorem, fat shattering dimension, packing numbers. Lecture 15, Toyota Technological Institute at Chicago. 2008.

[11] S. Mendelson and R. Vershynin. Entropy and the combinatorial dimension. *Invent. Math.*, 152(1):37–55, 2003.

[12] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations Of Machine Learning.* The MIT Press, 2012.

[13] Alexander Rakhlin, 2017. Private communication.