A Note on Refining the Sample Complexity for Efficient Agnostic Learning of Linear Separators under the Uniform Distribution

Steve Hanneke

The paper HKY15 (Hanneke, Kanade & Yang) on concept drift has a refinement of the sample/query complexity of the agnostic learning algorithm of ABL (Awasthi, Balcan & Long) implicit in it, for homogeneous linear separators under the uniform distribution on (or in) the unit sphere. Specifically, we save a factor of d in the bound.[1] Here I'll just briefly sketch the changes in the part of the ABL proof that gets changed to give the better sample/query complexity. I use the same notation introduced in the original ABL paper, and also use the algorithm given there (rather than the very-slightly different one from HKY15).

Everywhere, C will denote a numerical constant that we can't necessarily set to any value we want, but it could be different values in different places (e.g., $2C \leq C$ would be true under this convention).

In the algorithm, since it's the uniform distribution, we're taking $b_k = c2^{-k}/\sqrt{d}$ for some numerical constant c, $r_k = c2^{-k}$ for another numerical constant c, and $\tau_k = c2^{-k}/\sqrt{d}$ for yet another numerical constant c (the values of these constants aren't very important to the part of the proof we'll discuss here).

I'll just focus on one round of the algorithm. So we're just trying to argue that, there are numerical constants c_0 and c_1 that we can choose to set as large as we want, such that, for sufficiently large (but fixed, numerical) values of c_0 and c_1, setting $\epsilon = c_1\eta$ and

$$m_k = c_0 d \log \frac{1}{\delta},$$

for any fixed $k \leq \log_2(1/\epsilon)$, given $\|w_{k-1} - w^*\| \leq r_k$, with probability at least $1 - c'\delta$,

$$\mathrm{err}_{P_k}(w_k) \leq \kappa,$$

where κ is a given fixed value (e.g., $1/4$ would be sufficiently interesting) and c' is some numerical constant. (To use this in the ABL inductive proof, we'd replace δ everywhere with, say, $\delta/(c'(k+1)^2)$, but this is a minor detail; the final error guarantee would then be $c\eta$ for some $c \gtrsim c_1$).
Here is the argument:

[1] Technically, it also saves a log factor.

By standard VC bounds, with probability at least $1-\delta$,

$$\begin{aligned}\mathrm{err}_{P_k}(w_k) &\le \widehat{\mathrm{err}}(w_k, \mathrm{cleaned}(W)) + C\sqrt{\frac{d+\log(1/\delta)}{m_k}} \\ &\le \widehat{\mathrm{err}}(w_k, \mathrm{cleaned}(W)) + \frac{C}{\sqrt{c_0}}.\end{aligned}$$

(Aside: This step is important to do first, since ABL pick up their extra factor of d by doing their uniform convergence step *after* relaxing to the hinge loss, which has larger range.)
Now

$$\begin{aligned}&\widehat{\mathrm{err}}(w_k, \mathrm{cleaned}(W)) \\ &\le \widehat{\mathrm{err}}(w^*, W) + \widehat{\mathrm{err}}(w_k, W)\end{aligned}$$

which (by a Chernoff bound) with probability at least $1-\delta$ is less than or equal to

$$\begin{aligned}&C\mathrm{err}_{\tilde{P}_k}(w^*) + \frac{C}{m_k}\log\frac{1}{\delta} + \widehat{\mathrm{err}}(w_k, W) \\ &\le C\frac{\eta}{\tilde{P}(x : |w_{k-1}\cdot x| \le b_{k-1})} + \frac{C}{m_k}\log\frac{1}{\delta} + \widehat{\mathrm{err}}(w_k, W) \\ &\le C2^k\eta + \frac{C}{m_k}\log\frac{1}{\delta} + \widehat{\mathrm{err}}(w_k, W) \qquad \text{(different } C \text{ here)} \\ &\le \frac{C}{c_1} + \frac{C}{c_0} + \widehat{\mathrm{err}}(w_k, W).\end{aligned}$$

Since any (x, y) with $\mathrm{sign}(w_k \cdot x) \neq y$ has $y(w_k \cdot x) \le 0$ and hence $y(v_k \cdot x) \le 0$, we know any such (x, y) has $\ell(v_k, x, y) \ge 1$; also, every (x, y) has $\ell(v_k, x, y) \ge 0$. Therefore,

$$\widehat{\mathrm{err}}(w_k, W) \le \ell(v_k, W).$$

The next bit follows part of the original argument of ABL. By assumption $\|w_{k-1} - w^*\| \le r_k$, and by definition $\ell(v_k, W) \le \ell(w, W) + \kappa/8$ among vectors $w \in B(w_{k-1}, r_k)$, so that

$$\ell(v_k, W) \le \ell(w^*, W) + \frac{\kappa}{8}.$$

Lemma 3.5 of ABL (2nd arXiv version 1307.8371v2) implies every $(x, y) \in W$ has $0 \le \ell(w^*, x, y) \le C\sqrt{d}$. Using this as the a-priori range of these random variables, Hoeffding's inequality implies that with probability at least $1-\delta$,

$$\begin{aligned}&\ell(w^*, W) \\ &\le \mathbf{E}_{(x,y)\sim\tilde{P}_k}[\ell(w^*, x, y)] + C\sqrt{\frac{d\log(1/\delta)}{m_k}} \\ &\le \mathbf{E}_{(x,y)\sim\tilde{P}_k}[\ell(w^*, x, y)] + \frac{C}{\sqrt{c_0}}.\end{aligned}$$

(Aside: This was another step that was important to change from ABL. They again used their uniform convergence bound, because *why not* given that they'd already used it once; but of course that picks up a factor of d from the pseudo-dimension, which multiplies with the d from the squared range of $\ell(w^*, x, y)$ to give a factor d^2; since we avoided needing uniform convergence of $\ell(w, W)$ above, it makes sense to just use a bound on $\ell(w^*, W)$ for the single vector w^* here, avoiding the pseudo-dimension factor and thus saving a factor of d in the required size of m_k.)
From here, Lemma 5.1 of ABL (2013 arXiv v2 version[2]) implies the last line above is less than or equal to

$$\begin{aligned}
&\mathbf{E}_{(x,y)\sim P_k}[\ell(w^*, x, y)] + C\sqrt{\frac{\eta}{\epsilon}}\frac{\sqrt{r_k^2/(d-1) + b_{k-1}^2}}{\tau_k} + \frac{C}{\sqrt{c_0}} \\
&\leq \mathbf{E}_{(x,y)\sim P_k}[\ell(w^*, x, y)] + \frac{C}{\sqrt{c_1}} + \frac{C}{\sqrt{c_0}} \qquad \text{(different } C\text{)}
\end{aligned}$$

and Lemma 3.7 of ABL (JACM version) implies this is at most

$$C\frac{\tau_k}{b_{k-1}} + \frac{C}{\sqrt{c_1}} + \frac{C}{\sqrt{c_0}}.$$

Thus, if we pick the constants in the definition of τ_k and b_{k-1} so that $\tau_k \leq \kappa b_{k-1}/(6C)$, say, then this is at most

$$\frac{\kappa}{16} + \frac{C}{\sqrt{c_1}} + \frac{C}{\sqrt{c_0}}.$$

Altogether,

$$\mathrm{err}_{P_k}(w_k) \leq \frac{C}{\sqrt{c_0}} + \frac{C}{c_1} + \frac{C}{c_0} + \frac{\kappa}{8} + \frac{\kappa}{16} + \frac{C}{\sqrt{c_1}} + \frac{C}{\sqrt{c_0}}.$$

We can pick c_0 and c_1 as large as we wish, so for instance, any $c_0 \geq 144C^2/\kappa^2$ and $c_1 \geq 16C^2/\kappa^2$ would suffice to conclude $\mathrm{err}_{P_k}(w_k) \leq \kappa$.

[2]It's the JACM version's Lemma 3.8, but for uniform $z_k = \sqrt{r_k^2/(d-1) + b_{k-1}^2}$ instead.