A Proofs

A.1 Notation and Definitions

Throughout the proofs, we fix $d, k \geq 2$. We denote by $\mathcal{W} = \mathcal{W}^d = \{h_w : w \in \mathbb{R}^{d+1}\}$ the class of linear separators (with bias) over $\mathbb{R}^d$. We assume the following "tie breaking" conventions:

- For $f : [k] \to \mathbb{R}$, $\operatorname{argmax}_{i\in[k]} f(i)$ is the *minimal* number $i_0 \in [k]$ for which $f(i_0) = \max_{i\in[k]} f(i)$;
- $\operatorname{sign}(0) = 1$.

Given a hypotheses class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, denote its restriction to $A \subseteq \mathcal{X}$ by $\mathcal{H}|_A = \{f|_A : f \in \mathcal{H}\}$. Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class and let $\phi : \mathcal{Y} \to \mathcal{Y}'$, $\iota : \mathcal{X} \to \mathcal{X}'$ be functions. Denote $\phi \circ \mathcal{H} = \{\phi \circ h : h \in \mathcal{H}\}$ and $\mathcal{H} \circ \iota = \{h \circ \iota : h \in \mathcal{H}\}$.

Given $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ and a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, denote the approximation error by $\operatorname{Err}^*_{\mathcal{D}}(\mathcal{H}) = \inf_{h\in\mathcal{H}} \operatorname{Err}_{\mathcal{D}}(h)$. Recall that by definition 1.1, $\mathcal{H}$ essentially contains $\mathcal{H}' \subseteq \mathcal{Y}^{\mathcal{X}}$ if and only if $\operatorname{Err}^*_{\mathcal{D}}(\mathcal{H}) \leq \operatorname{Err}^*_{\mathcal{D}}(\mathcal{H}')$ for every distribution $\mathcal{D}$. For a binary hypothesis class $\mathcal{H}$, denote its VC dimension by $\operatorname{VC}(\mathcal{H})$.

Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class and let $S \subseteq \mathcal{X}$. We say that $\mathcal{H}$ *G-shatters* S if there exists an $f : S \to \mathcal{Y}$ such that for every $T \subseteq S$ there is a $g \in \mathcal{H}$ such that

$$\forall x \in T,\ g(x) = f(x), \text{ and } \forall x \in S \setminus T,\ g(x) \neq f(x).$$

We say that $\mathcal{H}$ *N-shatters* S if there exist $f_1, f_2 : S \to \mathcal{Y}$ such that $\forall y \in S,\ f_1(y) \neq f_2(y)$, and for every $T \subseteq S$ there is a $g \in \mathcal{H}$ such that

$$\forall x \in T,\ g(x) = f_1(x), \text{ and } \forall x \in S \setminus T,\ g(x) = f_2(x).$$

The *graph dimension* of $\mathcal{H}$, denoted $d_G(\mathcal{H})$, is the maximal cardinality of a set that is G-shattered by $\mathcal{H}$. The *Natarajan dimension* of $\mathcal{H}$, denoted $d_N(\mathcal{H})$, is the maximal cardinality of a set that is N-shattered by $\mathcal{H}$. Both of these dimensions coincide with the VC-dimension for $|\mathcal{Y}| = 2$. Note also that we always have $d_N(\mathcal{H}) \leq d_G(\mathcal{H})$. As shown in Ben-David et al. [1995], it also holds that $d_G(\mathcal{H}) \leq 4.67 \log_2(|Y|) d_N(\mathcal{H})$.

Proof of Lemma 4.1. Let $A \subseteq \mathcal{X}$ be a G-shattered set with $|A| = d_G(L(\mathcal{H}))$. By Sauer's Lemma, $2^{|A|} \leq |\mathcal{H}|_A|^l \leq |A|^{dl}$, thus $d_G(L(\mathcal{H})) = |A| = O(ld \log(ld))$. □

A.2 Multiclass SVM

Proof of Theorem 3.1. The lower bound follows from Theorems 3.5 and 3.2. To upper bound $d_G := d_G(\mathcal{L})$, let $S = \{x_1, \ldots, x_{d_G}\} \subseteq \mathbb{R}^d$ be a set which is G-shattered by $\mathcal{L}$, and let $f : S \to [k]$ be a function that witnesses the shattering. For $x \in \mathbb{R}^d$ and $j \in [k]$, denote

$$\phi(x, j) = (0, \ldots 0, x[1], \ldots, x[d], 1, 0, \ldots, 0) \in \mathbb{R}^{(d+1)k},$$

where $x[1]$ is in the $(d+1)(j-1)$ coordinate. For every $(i, j) \in [d_G] \times [k]$, define $z_{i,j} = \phi(x_i, f(x_i)) - \phi(x_i, j)$. Denote $Z = \{z_{i,j} \mid (i, j) \in [d_G] \times [k]\}$. Since $\operatorname{VC}(\mathcal{W}^{(d+1)k}) = (d+1)k+1$, by Sauer's lemma,

$$|\mathcal{W}^{(d+1)k}|_Z| \leq |Z|^{(d+1)k+1} = (d_G k)^{(d+1)k+1}.$$

We now show that there is a one-to-one mapping from subsets of S to $\mathcal{W}^{(d+1)k}|_Z$, thus concluding an upper bound on the size of S. For any $T \subseteq S$, choose $W(T) \in \mathbb{R}^{k\times(d+1)}(\mathbb{R})$ such that

$$\{x \in S \mid h_{W(T)}(x) = f(x)\} = T.$$

Such a $W(T)$ exists because of the G-shattering of S by $\mathcal{L}$ using the witness f. Define the vector $w(T) \in \mathbb{R}^{k(d+1)}$ which is the concatenation of the rows of $W(T)$, that is $w(T) = (W(T)_{(1,1)}, \ldots, W(T)_{(1,d+1)}, \ldots, W(T)_{(k,1)}, \ldots, W(T)_{(k,d+1)})$.

Now, suppose that $T_1 \neq T_2$ for $T_1, T_2 \subseteq S$. We now show that $w(T_1)|_Z \neq w(T_2)|_Z$. Suppose w.l.o.g. that there is some $x_i \in T_1 \setminus T_2$. Thus, $f(x_i) = h_{W(T_1)}(x_i) \neq h_{W(T_2)}(x_i) =: j$. It

follows that the inner product of x_i with row $f(x_i)$ of $W(T_1)$ is greater than the inner product of x_i with row j of $W(T_1)$, while for $W(T_2)$, the situation is reversed. Therefore, $\text{sign}(\langle w(T_1), z_{i,j}\rangle) \neq \text{sign}(\langle w(T_2), z_{i,j}\rangle)$, so $w(T_1)$ and $w(T_2)$ induce different labelings of Z. It follows that the number of subsets of S is bounded by the size of $\mathcal{W}^{(d+1)k}|_Z$, thus $2^{d_G} \leq (kd_G)^{(d+1)k+1}$. We conclude that $d_G \leq O(dk\log(dk))$. □

A.3 Simple classes that can be represented by the class of linear separators

In this section we define two fairly simple hypothesis classes, and show that the class of linear separators is richer than them. We will later use this observation to prove lower bounds on the Natarajan dimension of various multiclass hypothesis classes.

Let $l \geq 2$. For $f \in \{-1,1\}^{[d]},\ i \in [l],\ j \in \{-1,1\}$ define $f^{i,j} : [d] \times [l] \to \{-1,1\}$ by

$$f^{i,j}(u,v) = \begin{cases} f(u) & v = i \\ j & v \neq i, \end{cases}$$

And define the hypothesis class $\mathcal{F}^l$ as

$$\mathcal{F}^l = \{f^{i,j} : f \in \{\pm 1\}^{[d]},\ i \in [l],\ j \in \{-1,1\}\}.$$

For $g \in \{-1,1\}^{[d]},\ i \in [l],\ j \in \{\pm 1\}$ define $g^{i,j} : [d] \times [l] \to \{-1,1\}$ by

$$g^{i,j}(u,v) = \begin{cases} h(u) & v = i \\ j & v > i \\ -j & v < i, \end{cases}$$

And define the hypothesis class $\mathcal{G}^l$ as

$$\mathcal{G}^l = \{g^{i,j} : g \in \{-1,1\}^{[d]},\ i \in [l],\ j \in \{\pm 1\}\}.$$

Let $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}, \mathcal{H}' \subset \mathcal{Y}^{\mathcal{X}'}$ be two hypotheses classes. We say that $\mathcal{H}$ is *richer* than $\mathcal{H}'$ if there is a mapping $\iota : \mathcal{X}' \to \mathcal{X}$ such that $\mathcal{H}' = \mathcal{H} \circ \iota$. It is clear that if $\mathcal{H}$ is richer than $\mathcal{H}'$ then $d_N(\mathcal{H}') \leq d_N(\mathcal{H})$ and $d_G(\mathcal{H}') \leq d_G(\mathcal{H})$. Thus, the notion of richness can be used to establish lower and upper bounds on the Natarajan and Graph dimension, respectively. The following lemma shows that $\mathcal{W}$ is richer than $\mathcal{F}^l$ and $\mathcal{G}^l$ for every l. This will allow us to use the classes $\mathcal{F}^l,\ \mathcal{G}^l$ instead of $\mathcal{W}$ when bounding from below the dimension of an ECOC or TC hypothesis class in which the binary classifiers are from $\mathcal{W}$.

Lemma A.1. *For any integer $l \geq 2$, $\mathcal{W}$ is richer than $\mathcal{F}^l$ and $\mathcal{G}^l$.*

Proof. We shall first prove that $\mathcal{W}$ is richer than $\mathcal{F}^l$. Choose l unit vectors $e_1, \ldots, e_l \in \mathbb{R}^d$. For every $i \in [l]$, choose d affinely independent vectors such that

$$x_{1,i}, \ldots, x_{d,i} \in \{x \in \mathbb{R}^d : \langle x, e_i\rangle = 1,\ \forall i' \neq i, \langle x, e_{i'}\rangle < 1\}.$$

This can be done by choosing d affinely independent vectors in $\{x \in \mathbb{R}^d : \langle x, e_i\rangle = 1\}$ that are very close to e_i. Define $\iota(m,i) = x_{m,i}$. Now fix $i \in [l]$ and $j \in \{-1,+1\}$, and let $f^{i,j} \in \mathcal{F}^l$. We must show that $f^{i,j} = h \circ \iota$ for some $h \in \mathcal{W}$. We will show that there exists an affine map $\Lambda : \mathbb{R}^d \to \mathbb{R}$ for which $f^{i,j} = \text{sign} \circ \Lambda \circ \iota$. This suffices, since $\mathcal{W}$ is exactly the set of all functions of the form $\text{sign} \circ \Lambda$ where Λ is an affine map. Define $M = \{x \in \mathbb{R}^d : \langle x, e_i\rangle = 1\}$, and let $A : M \to \mathbb{R}$ be the affine map defined by

$$\forall m \in [d],\ A(x_{m,i}) = f(m,i).$$

Let $P : \mathbb{R}^d \to M$ be the orthogonal projection of $\mathbb{R}^d$ on M. For $\alpha \in \mathbb{R}$, define an affine map $\Lambda_\alpha : \mathbb{R}^d \to \mathbb{R}$ by

$$\Lambda_\alpha(x) = A(P(x)) + \alpha \cdot \langle x - e_i, e_i\rangle.$$

Note that, $\forall m \in [d],\ \Lambda_\alpha(x_{m,i}) = f(m,i)$. Moreover, for every $i' \neq i$ and $m \in [d]$ we have $\langle x_{m,i'} - e_i, e_i\rangle < 0$. Thus, by choosing $|\alpha|$ sufficiently large and choosing $\text{sign}(\alpha)$ depending on j, we can make sure that $f^{i,j} = \text{sign} \circ \Lambda_\alpha \circ \iota$.

The proof that $\mathcal{W}$ is richer than $\mathcal{G}^l$ is similar and simpler. Let $e_1, \ldots, e_d \in \mathbb{R}^{d-1}$ be affinely independent. Define

$$\iota(m,i) = (e_m, i) \in \mathbb{R}^{d-1} \times \mathbb{R} \cong \mathbb{R}^d,$$

Given $g^{i,j} \in \mathcal{G}^{d,l}$, let $A : \mathbb{R}^{d-1} \times \{i\} \to \mathbb{R}$ be the affine map defined by $A(e_m, i) = g^{i,j}(m,i)$ and let $P : \mathbb{R}^d \to \mathbb{R}^{d-1} \times \{i\}$ be the orthogonal projection. Define $\Lambda : \mathbb{R}^d \to \mathbb{R}$ by

$$\Lambda(x,y) = A(P(x,y)) + j \cdot 10 \cdot (y-i).$$

It is easy to check that $\text{sign} \circ \Lambda \circ \iota = g^{i,j}$. □

Note A.2. *From Lemma A.1 it follows that* $\text{VC}(\mathcal{F}^l), \text{VC}(\mathcal{G}^l) \leq d+1$. *On the other hand, both* $\mathcal{F}^l$ *and* $\mathcal{G}^l$ *shatter* $([d] \times \{1\}) \cup \{(1,2)\}$. *Thus,* $\text{VC}(\mathcal{F}^l) = \text{VC}(\mathcal{G}^l) = d+1$

A.4 Trees

Proof of Theorem 3.2. We first prove the upper bound. Let $A \subseteq \mathcal{X}$ be a G-shattered set with $|A| = d_G(\mathcal{H}_{\text{trees}})$. By Sauer's Lemma, and since the number of trees is bounded by k^k, we have $2^{|A|} \leq k^k \cdot |\mathcal{H}|_A|^k \leq k^k \cdot |A|^{dk}$, thus $d_G(\mathcal{H}_{\text{trees}}) = |A| = O(dk \log(dk))$.

To prove the lower bound, by Lemma A.1, it is enough to show that $d_N(\mathcal{G}^l_T) \geq d \cdot (k-1)$ for some l. We will take $l = |N(T)| = k-1$. Linearly order $N(T)$ such that for every node v, the nodes in the left sub-tree emanating from v are smaller than the nodes in the corresponding right sub-tree. We will identify $[l]$ with $N(T)$ by an order-preserving map, thus $\mathcal{G}^l \subset \{-1,1\}^{[d] \times N(T)}$. We also identify the labels with the leaves.

Define $g_1 : [d] \times N(T) \to \text{leaf}(T)$ by setting $g_1(i,v)$ to be the leaf obtained by starting from the node v, going right once and then going left until reaching a leaf. Similarly, define $g_2 : [d] \times N(T) \to leaf(T)$ by setting $g_2(i,v)$ to be the leaf obtained by starting from the node v, going left once and then going right until reaching a leaf.

We shall show that g_1, g_2 witness the N-shattering of $[d] \times N(T)$ by $\mathcal{G}^l_T$. Given $S \subset [d] \times N(T)$ define $C : N(T) \to \mathcal{G}^l$ by

$$C(v)(i,u) = \begin{cases} -1 & u < v \\ 1 & u > v \\ 1 & u = v,\ (i,u) \in S \\ -1 & u = v,\ (i,u) \notin S. \end{cases}$$

It is not hard to check that $\forall (i,u) \in S,\ h_C(i,u) = g_1(i,u)$, and $\forall (i,u) \notin S,\ h_C(i,u) = g_2(i,u)$. □

Note A.3. *Define* $\tilde{\mathcal{G}}^l = \{g^{i,1} : g \in \{-1,1\}^{[d]},\ i \in [l]\}$. *The proof shows that* $d_N(\tilde{\mathcal{G}}^l_T) \geq d \cdot (k-1)$. *Since* $\text{VC}(\tilde{\mathcal{G}}^l) = d$, *we obtain a simpler proof of Theorem 23 from Daniely et al. [2011], which states that for every tree* T *there exists a class* $\mathcal{H}$ *of VC dimension* d *for which* $d_N(\mathcal{H}_T) \geq d(k-1)$.

A.5 ECOC, One vs. All and All Pairs

To prove the results for ECOC and its special cases, we first prove a more general theorem, based on the notion of a sensitive vector for a given code. Fix a code $M \in \mathbb{R}^{k \times l}(\mathbb{R})$. We say that a binary vector $u \in \{\pm 1\}^l$ is q-*sensitive* for M if there are q indices $j \in [l]$ for which $\tilde{M}(u) \neq \tilde{M}(u \oplus e_j)$. Here, $u \oplus e_j := (u[1], \ldots, -u[j], \ldots, u[l])$.

Theorem A.4. *If there exists a* q*-sensitive vector for a code* $M \in \mathbb{R}^{k \times l}(\mathbb{R})$ *then* $d_N(\mathcal{W}_M) \geq d \cdot q$.

Proof. By Lemma A.1, it suffices to show that $d_N(\mathcal{F}^l_M) \geq d \cdot q$. Let $u \in \{\pm 1\}^l$ be a q-sensitive vector. Assume w.l.o.g. that the sensitive coordinates are $1, \ldots, q$. We shall show that $[d] \times [q]$ is N-shattered by $\mathcal{F}^l_M$. Define $g_1, g_2 : [d] \times [q] \to [k]$ by

$$g_1(x,y) = \tilde{M}(u),\ g_2(x,y) = \tilde{M}(u \oplus e_y)$$

Let $T \subset [d] \times [q]$. Define $h_1, \ldots, h_l \in \mathcal{F}^l$ as follows. For every $j > q$, define $h_j \equiv u[j]$. For $j \leq q$ define

$$h_j(x,y) = \begin{cases} u[j] & y \neq j \\ u[j] & y = j,\ (x,y) \in T \\ -u[j] & y = j,\ (x,y) \in [d] \times [q] \setminus T. \end{cases}$$

For $h = (h_1, \ldots, h_l)$, it is not hard to check that

$$\forall (x,y) \in T, \quad \tilde{M}(h_1(x,y), \ldots, h_l(x,y)) = g_1(x,y), \text{ and}$$
$$\forall (x,y) \in [d] \times [q] \setminus T, \quad \tilde{M}(h_1(x,y), \ldots, h_l(x,y)) = g_2(x,y).$$

□

The following lemma shows that a code with a large distance is also highly sensitive. In fact, we prove a stronger claim: the sensitivity is actually at least as large as the distance between any row and the row closest to it in Hamming distance. Formally, we consider $\Delta(M) = \max_i \min_{j \neq i} \Delta_{\mathrm{h}}(M[i], M[j]) \geq \delta(M)$.

Lemma A.5. *For any binary code $M \in \mathbb{R}^{k \times l}(\pm 1)$, there is a q-sensitive vector for M, where $q \geq \frac{1}{2}\Delta(M) \geq \frac{1}{2}\delta(M)$.*

Proof. Let i_1 the row in M such that its hamming distance to the row closest to it is $\Delta(M)$. Denote by i_2 the index of the closest row (if there is more than one such row, choose one of them arbitrarily). We have $\Delta_{\mathrm{h}}(M[i_1], M[i_2]) = \Delta(M)$. In addition, $\forall i \neq i_1, i_2, \Delta_{\mathrm{h}}(M[i_1], M[i]) \geq \Delta(M)$. Assume w.l.o.g. that the indices in which rows i_1 and i_2 differ are $1, \ldots, \Delta(M)$. Consider first the case that $i_1 < i_2$. Define $u \in \{\pm 1\}^{[l]}$ by

$$u[j] = \begin{cases} M_{(i_1,j)} & j \leq \lceil \frac{\Delta}{2} \rceil \\ M_{(i_2,j)} & \text{otherwise.} \end{cases}$$

Is is not hard to check that for every $1 \leq j \leq \lceil \frac{\Delta}{2} \rceil$, $i_1 = \tilde{M}(u)$ and $\tilde{M}(u \oplus e_j) = i_2$, thus u is $\lceil \frac{\Delta}{2} \rceil$-sensitive. If $i_1 > i_2$, the proof is similar except that u is defined as

$$u[j] = \begin{cases} M_{(i_2,j)} & j \leq \lceil \frac{\Delta}{2} \rceil \\ M_{(i_1,j)} & \text{otherwise.} \end{cases}$$

□

Proof of Theorem 3.3. The upper bound follows from Lemma 4.1. The lower bound follows form Theorem A.4 and Lemma A.5. □

Proof of Theorem 3.4. The upper bounds follow from Theorem 3.3. To show that $d_N(\mathcal{W}_{\mathrm{OvA}}) \geq (k-1)d$, we note that the all-negative vector $u = (-1, \ldots, -1)$ of length k is $(k-1)$-sensitive for the code M^{OvA}, and apply Theorem A.4.

To show that $d_N(\mathcal{W}_{\mathrm{AP}}) \geq d\binom{k-1}{2}$, assume for simplicity that k is odd (a similar analysis can be given when k is even). Define $u \in \{\pm 1\}^{\binom{k}{2}}$ by

$$\forall i < j,\ u[i,j] = \begin{cases} 1 & j - i \leq \frac{k-1}{2} \\ -1 & \text{otherwise.} \end{cases}$$

For every $n \in [k]$, we have $\sum_{1 \leq i < j \leq k} u[i,j] \cdot M^{\mathrm{AP}}_{n,(i,j)} = 0$, as the summation counts the number of pairs (i,j) such that $n \in \{i,j\}$ and $M^{\mathrm{AP}}_{n,(i,j)}$ agrees with $u[i,j]$. Thus, $\tilde{M}^{\mathrm{AP}}(u) = 1$, by our tie-breaking assumptions. Moreover, it follows that for every $1 < i < j \leq k$, we have $\tilde{M}^{\mathrm{AP}}(u \oplus e_{(i,j)}) \in \{i,j\}$, since flipping entry $[i,j]$ of u increases $(M^{AP}u)_j$ or $(M^{AP}u)_i$ by 1 and does not increase the rest of the coordinates of the vector $M^{AP}u$. This shows that u is $\binom{k-1}{2}$-sensitive. □

A.6 Approximation

Proof of Theorem 3.5. We first show that for any tree for k classes T, $\mathcal{L}$ essentially contains $\mathcal{W}_T$. It follows that $\mathcal{L}$ essentially contains $\mathcal{W}_{\text{trees}}$ as well. Let $\mathcal{D}$ a distribution over $\mathbb{R}^d$, let $C : N(T) \to \mathcal{W}$ be a mapping associating nodes in T to binary classifiers in $\mathcal{W}$, and let $\epsilon > 0$. We will show that there exists a matrix $W \in \mathbb{R}^{k \times (d+1)}$ such that $\Pr_{x \sim \mathcal{D}}[h_W(x) \neq h_C(x)] < \epsilon$.

For every $v \in N(T)$, denote by $w(v) \in \mathbb{R}^{d+1}$ the linear separator such that $C(v) = h_{w(v)}$. For every $w \in \mathbb{R}^{d+1}$ define $\tilde{w} = w + (0, \ldots, 0, \gamma)$. Recall that for $x \in \mathbb{R}^d$, $\bar{x} \in \mathbb{R}^{d+1}$ is simply the concatenation $(x, 1)$. Choose $r > 0$ large enough so that $\Pr_{x \sim \mathcal{D}}[||\bar{x}|| > r] < \epsilon/2$ and $\forall v \in N(T),\ ||\tilde{w}(v)|| < r$. Choose $\gamma > 0$ small enough so that

$$\Pr_{x \sim \mathcal{D}}[\exists v \in N(T), \langle \tilde{w}(v), \bar{x} \rangle \in (-\gamma, \gamma)] = \Pr_{x \sim \mathcal{D}}[\exists v \in N(T), \langle w(v), \bar{x} \rangle \in (-2\gamma, 0)] < \epsilon/2.$$

Let $a = 2r^2/\gamma + 1$. For $i \in [k]$, let $v_{i,1}, \ldots, v_{i,m_i}$ be the path from the root to the leaf associated with label i. For each $1 \leq j < m_i$ define $b_{i,j} = 1$ if $v_{i,j+1}$ is the right son of $v_{i,j}$, and $b_{i,j} = -1$ otherwise. Now, define $W \in \mathbb{R}^{k \times (d+1)}$ to be the matrix whose i'th row is $w_i = \sum_{j=1}^{m_i - 1} a^{-j} \cdot b_{i,j} \tilde{w}(v_{i,j})$.

To prove that $\Pr_{x \sim \mathcal{D}}[h_W(x) \neq h_C(x)] < \epsilon$, it suffices to show that $h_W(x) = h_C(x)$ for every $x \in \mathbb{R}^d$ satisfying $||\bar{x}|| < r$ and $\forall v \in N(T), \langle \tilde{w}(v), \bar{x} \rangle \notin (-\gamma, \gamma)$, since the probability mass of the rest of the vectors is less than ϵ. Let $x \in \mathbb{R}^d$ be a vector that satisfies these assumptions. Denote $i_1 = h_C(x)$. It suffices to show that for all $i_2 \in [k] \setminus \{i_1\}$, $\langle w_{i_1}, \bar{x} \rangle > \langle w_{i_2}, \bar{x} \rangle$, since this would imply that $h_W(x) = i_1$ as well.

Indeed, fix $i_2 \neq i_1$, and let j_0 be the length of the joint prefix of the two root-to-leaf paths that match the labels i_1 and i_2. In other words, $\forall j \leq j_0,\ v_{i_1,j} = v_{i_2,j}$ and $v_{i_1,j_0+1} \neq v_{i_2,j_0+1}$. Note that

$$\langle \bar{x}, (b_{i_1,j_0} - b_{i_2,j_0}) \tilde{w}(v_{i_1,j_0}) \rangle = \langle \bar{x}, 2b_{i_1,j_0} \tilde{w}(v_{i_1,j_0}) \rangle = 2|\langle \bar{x}, \tilde{w}(v_{i_1,j_0}) \rangle| \geq 2\gamma.$$

The last equality holds because b_{i_1,j_0} and $\langle \bar{x}, w(v_{i_1,j_0}) \rangle$ have the same sign by definition of $b_{i,j}$. We have

$$\begin{aligned}
\langle w_{i_1}, \bar{x} \rangle - \langle w_{i_2}, \bar{x} \rangle &= \langle \bar{x}, \sum_{j=1}^{m_{i_1}-1} a^{-j} b_{i_1,j} \tilde{w}(v_{i_1,j}) - \sum_{j=1}^{m_{i_2}-1} a^{-j} b_{i_2,j} \tilde{w}(v_{i_2,j}) \rangle \\
&= \langle \bar{x}, a^{-j_0}(b_{i_1,j_0} - b_{i_2,j_0}) \tilde{w}(v_{i_1,j_0}) \rangle + \langle \bar{x}, \sum_{j=j_0+1}^{m_{i_1}-1} a^{-j} b_{i_1,j} \tilde{w}(v_{i_1,j}) - \sum_{j=j_0+1}^{m_{i_2}-1} a^{-j} b_{i_2,j} \tilde{w}(v_{i_2,j}) \rangle \\
&\geq \langle \bar{x}, a^{-j_0}(b_{i_1,j_0} - b_{i_2,j_0}) \tilde{w}(v_{i_1,j_0}) \rangle - \sum_{j=j_0+1}^{\infty} a^{-j} 2r^2 \\
&\geq 2a^{-j_0} \left(\gamma - \frac{r^2}{a-1} \right) > 0.
\end{aligned}$$

Since this holds for all $i_2 \neq i_1$, it follows that $h_W(x) = i_1$. Thus, we have proved that $\mathcal{L}$ essentially contains W_{trees}.

Next, we show that $\mathcal{L}$ strictly contains $\mathcal{W}_{\text{trees}}$, by showing a distribution over labeled examples such that the approximation error using $\mathcal{L}$ is strictly smaller than the approximation error using $\mathcal{W}_{\text{trees}}$. Assume w.l.o.g. that $d = 2$ and $k = 3$: even if they are larger we can always restrict the support of the distribution to a subspace of dimension 2 and to only three of the labels. Consider the distribution $\mathcal{D}$ over $\mathbb{R}^2 \times [3]$ such that its marginal over $\mathbb{R}^2$ is uniform in the unit circle, and $\Pr_{(X,Y) \sim D}[Y = i \mid X = x] = \mathbb{I}[x \in D_i]$, where D_1, D_2, D_3 be subsets sectors of equal angle of the unit circle (see Figure 1):

Clearly, by taking the rows of W to point to the middle of each sector (dashed arrows in the illustration), we get $\text{Err}^*_{\mathcal{D}}(\mathcal{L}) = 0$. In contrast, no linear separator can split the three labels into two gropus without error, thus $\text{Err}^*_{\mathcal{D}}(\mathcal{W}_{\text{trees}}) > 0$.

Finally, to see that $\mathcal{L}$ essentially contains $\mathcal{W}_{\text{OvA}}$, we note that $\mathcal{W}_{\text{OvA}} = \mathcal{W}_T$ where T is a tree such that each of its internal nodes has a leaf corresponding to one of the labels as its left son. Thus $\mathcal{W}_{\text{OvA}}$ is essentially contained in $\mathcal{W}_{\text{trees}}$. □

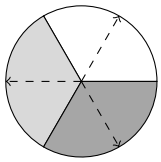

Figure 1: Illustration for the proof of Theorem 3.5

Proof of Theorem 3.7. It is easily seen that $\mathcal{W}_{AP}$ contains $\mathcal{L}$: Let $W \in \mathbb{R}^{d+1\times k}$, and denote its i'th row by $W[i]$. For each column (i,j) of M^{AP}, define the binary classifier $h_{i,j} \in \mathcal{W}$ such that $\forall x \in \mathbb{R}^d, h_{i,j}(\bar{x}) = \text{sign}(\langle W[j] - W[i], \bar{x}\rangle)$. Then for all x, $h_W(x) = \tilde{M^{\text{AP}}}(h_{1,1}(x), \ldots, h_{k-1,k}(x))$.

To show that the inclusion is strict, as in the proof of Theorem 3.5, we can and will assume that $d = 2$. Choose k^* to be the minimal number such that for every $k \geq k^*$, $d_N(\mathcal{W}_{AP}) > d_N(\mathcal{L})$: This number exists by Theorems 3.4 and 3.1 (note that though we chose k^* w.r.t. $d = 2$, the same k^* is valid for every d). For any $k \geq k^*$, it follows that there is a set $S \subseteq \mathbb{R}^2$ that is N-shattered by $\mathcal{W}_{AP}$ but not by $\mathcal{L}$. Thus, there is a hypothesis $h \in \mathcal{W}_{AP}$ such that for every $g \in \mathcal{L}$, $g|_S \neq h|_S$. Define the distribution $\mathcal{D}$ to be uniform over $\{(x, h(x)) : x \in S\}$. Then clearly $\text{Err}^*_{\mathcal{D}}(\mathcal{L}) > \text{Err}^*_{\mathcal{D}}(\mathcal{W}_{AP}) = 0$. □

Next, we prove Theorem 3.6, which we restate more formally as follows. Note that the result on OvA is implied since there exists a tree that implements OvA.

Theorem A.6. *(Restatement of Theorem 3.6) If there exists an embedding $\iota : \mathbb{R}^d \to \mathbb{R}^{d'}$ and a tree T such that $\mathcal{W}_T^{d'} \circ \iota$ essentially contains $\mathcal{L}$, then necessarily $d' \geq \tilde{\Omega}(dk)$.*

Proof. Assume that $i \in [k]$ is the class corresponding to the leaf with the least depth, l. Note that $l \leq \log_2(k)$. Let $\phi : [k] \to \{\pm 1\}$ be the function that is 1 on $\{i\}$ and -1 otherwise. It is not hard to see that $\phi \circ \mathcal{L}$ is the hypothesis class of convex polyhedra in $\mathbb{R}^d$ having $k-1$ faces. Thus,

$$\text{VC}(\phi \circ \mathcal{L}) \geq (k-1)d, \tag{3}$$

[see e.g. Takacs, 2009]. On the other hand, $\phi \circ \mathcal{W}_T^{d'}$, is the class of convex polyhedra in $\mathbb{R}^{d'}$ having $l \leq \log_2(k)$ faces. Thus, by Lemma 4.1

$$\text{VC}(\phi \circ \mathcal{W}_T^{d'} \circ \iota) \leq \text{VC}(\phi \circ \mathcal{W}_T^{d'}) \leq O(ld' \log(ld')) \leq O(\log(k)d' \log(\log(k)d')) \tag{4}$$

By the assumption that $\mathcal{W}_T^{d'} \circ \iota$ essentially contains $\mathcal{L}$, $\text{VC}(\phi \circ \mathcal{L}) \leq \text{VC}(\phi \circ \mathcal{W}_T^{d'} \circ \iota)$. Combining with equations (3) and (4) it follows that $d(k-1) = O(\log(k)d' \log(\log(k)d'))$. Thus, $d' = \tilde{\Omega}(dk)$. □

To prove Lemma 3.8, we first state the classic VC-dimension theorem, which will be useful to us.

Theorem A.7 (Vapnik [1998])**.** *There exists a constant $C > 0$ such that for every hypothesis class $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$ of VC dimension d, a distribution $\mathcal{D}$ over $\mathcal{X}$, $\epsilon, \delta > 0$ and $m \geq C\frac{d+\ln(\frac{1}{\delta})}{\epsilon^2}$ we have*

$$\Pr_{S\sim\mathcal{D}^m}\left[\text{Err}^*_{\mathcal{D}}(\mathcal{H}) \geq \inf_{h\in\mathcal{H}} \text{Err}_S(h) - \epsilon\right] \geq 1 - \delta.$$

We also use the following lemma, which proves a variant of Hoeffding's inequality.

Lemma A.8. *Let $\beta_1, \ldots, \beta_k \geq 0$ and let $\gamma_1, \ldots, \gamma_k \in \mathbb{R}$, such that $\forall i,\ |\gamma_i| \leq \beta_i$. Fix an integer $j \in \{1, \ldots, \lfloor \frac{k}{2} \rfloor\}$ and let $\mu = j/k$. Let $(X_1, \ldots, X_k) \in \{\pm 1\}^k$ be a random vector sampled uniformly from the set $\{(x_1, \ldots, x_k) : \sum_{i=1}^k \frac{x_1+1}{2} = \mu k\}$. Define $Y_i = \beta_i + X_i\gamma_i$ and denote $\alpha_i = \beta_i + |\gamma_i|$. Assume that $\sum_{i=1}^k \alpha_i = 1$. Then*

$$\Pr\left[\sum_{i=1}^k Y_i \leq \mu - \epsilon\right] \leq 2\exp\left(-\frac{\epsilon^2}{2\sum_{i=1}^k \alpha_i^2}\right).$$

Proof. First, since $\mu < \frac{1}{2}$, it suffices to prove the claim for the case $\forall i, \gamma_i \geq 0$ since this is the "harder" case. Let $Z_1, \ldots, Z_k \in \{\pm 1\}$ be independent random variables such that $\Pr[Z_i = 1] = \mu - \frac{\epsilon}{2}$. Denote $W_i = \beta_i + Z_i\gamma_i$. Further denote $\bar{W} = \sum_{i=1}^k W_i$ and $\bar{Z} = \sum_{i=1}^k \frac{Z_i+1}{2}$.

Note that for every $j_0 \leq j = \mu k$, given that $\bar{Z} = j_0$, $\bar{W}$ can be described as follows: We start with the value $\sum_{i=1}^k \beta_i - \gamma_i$ and then choose j_0 indices uniformly from $[k]$. For each chosen index i, the value of $\bar{W}$ is increased by $2\gamma_i$. $\sum_{i=1}^k Y_i$ can be described in the same way, except that that $j \geq j_0$ indices are chosen. Thus, $\Pr\left[\sum_{i=1}^k Y_i \leq \mu - \epsilon\right] \leq \Pr\left[\bar{W} \leq \mu - \epsilon \mid \bar{Z} = j_0\right]$. Thus, we have

$$\begin{aligned}
\Pr\left[\sum_{i=1}^k Y_i \leq \mu - \epsilon\right] &\leq \Pr\left[\bar{W} \leq \mu - \epsilon \mid \bar{Z} \leq \mu k\right] \\
&\leq \Pr\left[\bar{W} \leq \mu - \epsilon\right] / \Pr\left[\bar{Z} \leq \mu k\right] \\
&\leq 2\Pr\left[\bar{W} \leq \mu - \epsilon\right] \\
&\leq 2\exp\left(-\frac{\epsilon^2}{2\sum_{i=1}^k \alpha_i^2}\right).
\end{aligned}$$

The last inequality follows from Hoeffding's inequality and noting that

$$E[W_i] = \beta_i + (2(\mu - \frac{\epsilon}{2}) - 1)\gamma_i = (\mu - \frac{\epsilon}{2})(\beta_i + \gamma_i) + (1 - \mu + \frac{\epsilon}{2})(\beta_i - \gamma_i) \geq (\mu - \frac{\epsilon}{2})\alpha_i.$$

So that $\sum_{i=1}^k E[W_i] \geq (\mu - \frac{\epsilon}{2})\sum_{i=1}^k \alpha_i = \mu - \frac{\epsilon}{2}$. □

Proof of Lemma 3.8. The idea of this proof is as follows: Using a uniform convergence argument based on the VC dimension of the binary hypothesis class, we show that there exists a labeled sample S such that $|S| \approx \frac{d+k}{\nu^2}$, and for all possible mappings ϕ, the approximation error of the hypothesis class on the sample is close to the approximation error on the distribution $\mathcal{D}_\phi$. This allows us to restrict our attention to a finite set of hypotheses, based on their restriction to the sample. For these hypotheses, we show that with high probability over the choice of ϕ, the approximation error on the sample is high. Using a union bound on the possible hypotheses, we conclude that the approximation error on the distribution will be high, with high probability over the choice of ϕ.

For $i \in [k]$, denote $p_i = \Pr_{x\sim\mathcal{D}}[f(x) = i]$. Let $S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \subseteq \mathcal{X} \times [k]$ be an i.i.d. sample drawn according to $\mathcal{D}$ where $m = \lceil C\frac{d+(k+2)\ln(2)}{(\nu/2)^2}\rceil$, for the constant from C from Theorem A.7. Given S, denote $S_\phi\{(x_1, \phi(y_1)), \ldots, (x_m, \phi(y_m))\} \subseteq \mathcal{X} \times \{\pm 1\}$. For $i \in [k]$, let $\hat{p}_i = \frac{|\{j:y_j=i\}|}{m}$.

For any fixed $\phi : [k] \to \{\pm 1\}$, with probability $> 1 - 2^{-(k+2)}$ over the choice of S we have, by Theorem A.7, that $\mathrm{Err}^*_{\mathcal{D}_\phi}(\mathcal{H}) > \inf_{h\in\mathcal{H}} \mathrm{Err}_{S_\phi}(h) - \nu$. Since $|\{\pm 1\}^{[k]}| = 2^k$, w.p. $> 1 - \frac{1}{4}$,

$$\forall \phi \in \{\pm 1\}^{[k]}, \quad \mathrm{Err}^*_{\mathcal{D}_\phi}(\mathcal{H}) > \inf_{h\in\mathcal{H}} \mathrm{Err}_{S_\phi}(h) - \frac{\nu}{2}. \tag{5}$$

Moreover, we have

$$E[\sum_{i=1}^k \hat{p}_i^2] = \frac{1}{m^2}\sum_{i=1}^k \left(\binom{m}{2}p_i^2 + mp_i\right) \leq k \cdot \left(\frac{m(m-1)}{2m^2}\frac{100}{k^2} + \frac{10}{mk}\right) \leq \frac{60}{k}.$$

Thus, by Markov's inequality, w.p. $\geq \frac{1}{2}$ we have

$$\sum_{i=1}^k \hat{p}_i^2 < \frac{120}{k}. \tag{6}$$

Thus, with probability at least $1 - \frac{1}{4} - \frac{1}{2} > 0$, both (6) and (5) holds. In particular, there exists a sample S for which both (6) and (5) hold. Let us fix such an $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$.

Assume now that $\phi \in \{\pm 1\}^{[k]}$ is sampled according to the first condition. Denote

$$Y_i = |\{j : h(x_j) \neq \phi(y_j) \text{ and } y_j = i\}|/m.$$

For a fixed $h \in \mathcal{H}$ we have

$$\Pr_{\phi}\left[\mathrm{Err}_{S_\phi}(h) < \mu - \frac{\nu}{2}\right] = \Pr_{\phi}\left[\sum_{i=1}^{k} Y_i < \mu - \frac{\nu}{2}\right]$$

We note that Y_i are independent random variables with $E[Y_i] \geq \mu\hat{p}_i$ and $0 \leq Y_i \leq \hat{p}_i$. Thus, by Hoeffding's inequality,

$$\Pr_{\phi}\left[\mathrm{Err}_{S_\phi}(h) < \mu - \frac{\nu}{2}\right] \leq \exp\left(-\frac{\nu^2}{2\sum_{i=1}^{k}\hat{p}_i^2}\right) \leq \exp\left(-\frac{\nu^2 k}{240}\right).$$

By Sauer's lemma, $|\mathcal{H}|_{\{x_1,\ldots,x_m\}}| \leq \left(\frac{em}{d}\right)^d$. Thus, with probability $\geq 1 - \left(\frac{em}{d}\right)^d \exp\left(-\frac{\nu^2 k}{240}\right)$ over the choice of ϕ, $\inf_{h\in\mathcal{H}} \mathrm{Err}_{S_\phi}(h) \geq \mu - \frac{\nu}{2}$ and by (5) also

$$\mathrm{Err}^*_{\mathcal{D}_\phi}(\mathcal{H}) \geq \frac{1}{2} - \nu. \tag{7}$$

Finally, since $m = O\left(\frac{k+d}{\nu^2}\right)$, if $k = \Omega\left(\frac{d\ln(1/\nu)+\ln(1/\delta)}{\nu^2}\right)$ then Eq. (7) holds w.p $> 1-\delta$, concluding the proof for the case when the first condition holds. If the second condition holds, the proof is very similar, with the sole difference that Lemma A.8 is used instead of Hoeffding's inequality. □

Proof of Corollary 3.9. The Corollary follows from Lemma 3.8, by noting that $\mathrm{Err}^*_{\mathcal{D}}(\mathcal{H}_T) \geq \mathrm{Err}^*_{\mathcal{D}_\phi}(\mathcal{H})$, where $\phi : [k] \to \{\pm 1\}$ is defined as $\phi(i) = 1$ if and only if $\lambda^{-1}(i)$ is in the right subtree emanating from the root of T. □

Proof of Corollary 3.10. Let $\phi : [k] \to \{\pm 1\}$ be the function that is -1 on $\left[\lfloor \frac{k}{2} \rfloor\right]$ and 1 otherwise. By Lemma 4.1, applied to $L(\mathcal{H}) = \phi \circ \mathcal{H}_{(M,\mathrm{Id})}$, $\mathrm{VC}(\phi \circ \mathcal{H}_{(M,\mathrm{Id})}) = O(dl\log(dl))$, so that, by Lemma 3.8 (applied to a random choice of λ instead of ϕ), $\mathrm{Err}^*_{\mathcal{D}_{\phi\circ\lambda}}(\phi \circ \mathcal{H}_{(M,\mathrm{Id})}) \geq \frac{1}{2} - \nu$ with probability $> 1 - \delta$ over the choice of λ. The proof follows as we note that for every λ, $\mathrm{Err}^*_{\mathcal{D}}(\mathcal{H}_{(M,\lambda^{-1})}) = \mathrm{Err}^*_{\mathcal{D}_\lambda}(\mathcal{H}_{(M,Id)}) \geq \mathrm{Err}^*_{\mathcal{D}_{\phi\circ\lambda}}(\phi \circ \mathcal{H}_{(M,\mathrm{Id})})$. □