Growing Adaptive Multi-Hyperplane Machines (APPENDIX)

Theorem 2. *Let $\mathbf{W}^*$ be the solution of (6), and T be the total number of training iterations. Further, let the pruning be performed as described above, p be a starting probability of weight duplication, and $0 < \beta < 1$ is a multiplicative factor that reduces p after every weight duplication. Then,*

$$\frac{1}{T}\sum_{t=1}^{T}\big(\mathcal{L}^{(t)}(\mathbf{W}^{(t)}|\mathbf{z}) - \mathcal{L}^{(t)}(\mathbf{W}^*|\mathbf{z})\big) \leq \\ \frac{(2+c)^2(2+p/(1-\beta))}{\lambda} + \frac{(2+c)^2(2+p/(1-\beta))^2}{2T\lambda}\big(\frac{p(2\beta+3)}{(1-\beta)^2} + \ln(T) + 1\big). \tag{1}$$

Proof. The proof closely follows the proof of Theorems 1 and 3 from (Wang et al., 2011). First, we rewrite the update rule of SGD with the pruning step as $\mathbf{W}^{(t+1)} \leftarrow \mathbf{W}^{(t)} - \eta^{(t)}\boldsymbol{\partial}^{(t)}$, where $\boldsymbol{\partial}^{(t)} = \boldsymbol{\nabla}^{(t)} + \mathbf{E}^{(t)}$, and $\mathbf{E}^{(t)} = \mathbf{E}^{(t)}_{prune} + \mathbf{E}^{(t)}_{dupl}$ where we can see that the weight matrix degradation at the t^{th} training iteration $\mathbf{E}^{(t)}$ is equal to the sum of weight matrix degradation $\mathbf{E}^{(t)}_{prune}$ due to pruning and weight matrix degradation $\mathbf{E}^{(t)}_{dupl}$ due to weight duplication. Clearly, $\mathbf{E}^{(t)}_{prune} = \mathbf{0}$ if no pruning is used, and $\mathbf{E}^{(t)}_{dupl} = \mathbf{0}$ if no duplication is used at the t^{th} training iteration. Note that, in contrast to (Wang et al., 2011), we also included the weight duplication degradation. The relative progress towards the optimal solution $\mathbf{W}^*$ at the t^{th} round $D^{(t)}$ can be lower bounded as

$$\begin{aligned} D^{(t)} &= ||\mathbf{W}^{(t)} - \mathbf{W}^*||^2 - ||\mathbf{W}^{(t)} - \eta^{(t)}\boldsymbol{\nabla}^{(t)} - \eta^{(\mathrm{t})}\mathbf{E}^{(t)} - \mathbf{W}^*||^2 \\ &= -(\eta^{(t)})^2||\boldsymbol{\partial}^{(t)}||^2 + 2\eta^{(t)}\|(\mathbf{E}^{(t)})^{\mathrm{T}}(\mathbf{W}^{(t)} - \mathbf{W}^*)\| + 2\eta^{(t)}\|(\boldsymbol{\nabla}^{(t)})^{\mathrm{T}}(\mathbf{W}^{(t)} - \mathbf{W}^*)\| \\ &\geq_1 -(\eta^{(t)})^2||\boldsymbol{\partial}^{(t)}||^2 - 2\eta^{(\mathrm{t})}||\mathbf{E}^{(t)}||\frac{(2+c)(1+h)+2}{\lambda} \\ &\quad + 2\eta^{(t)}\left(\mathcal{L}^{(t)}(\mathbf{W}^{(t)}) - \mathcal{L}^{(t)}(\mathbf{W}^*) + \frac{\lambda}{2}\|\mathbf{W}^{(t)} - \mathbf{W}^*\|^2\right), \end{aligned} \tag{2}$$

where $h = p/(1-\beta)$. For the second term in the r.h.s. of the inequality in (2), we first bounded $\|\mathbf{W}^{(t)}\|$ as

$$\begin{aligned} ||\mathbf{W}^{(t)}|| &\leq ||(1-\eta^{(t-1)}\lambda)\mathbf{W}^{(t-1)}|| + 2\eta^{(t-1)} + \|\Delta_{prune}\mathbf{W}^{(t-1)}\| + \|\Delta_{dupl}\mathbf{W}^{(t-1)}\| \\ &\leq \frac{t-2}{t-1}||\mathbf{W}^{(t-1)}|| + \frac{2}{(t-1)\lambda} + \frac{c}{(t-1)\lambda} + \frac{2+c}{\lambda} \\ &\leq \frac{1}{t-1}||\mathbf{W}^{(0)}|| + \frac{2(t-1)}{(t-1)\lambda} + \frac{(t-1)c}{(t-1)\lambda} + \sum_{t=0}^{T-1} p\beta^t\frac{2+c}{\lambda} \leq \frac{2+c}{\lambda}(1+h), \end{aligned} \tag{3}$$

where, in contrast to (Wang et al., 2011), we added the $\|\Delta_{dupl}\mathbf{W}^{(t-1)}\|$ term equal to the norm of the duplicated weight. This term is upper bounded by $(2+c)/\lambda$, as the norm of any weight is upper bounded by the weight matrix norm when weight duplication is not used during training (Wang et al., 2011). The duplication probability p drops by a factor of β whenever the weight duplication is performed, introducing the multiplication factor of $\sum_{t=0}^{T-1} p\beta^t$ to the total weight matrix norm degradation due to duplication, where the sum of geometric sequence of duplication probabilities is upper bounded by $h = p/(1-\beta)$. We then use triangle inequality to bound $||\mathbf{W}^{(t)} - \mathbf{W}^*|| \leq (2+c)(1+h)/\lambda + 2/\lambda$ by using the fact that $\|\mathbf{W}^*\| \leq 2/\lambda$ according to the result in (Kivinen et al., 2002). Lastly, the third term in the r.h.s. of the inequality in (2) was obtained using function $\mathcal{L}^{(t)}(\mathbf{W}^{(t)})$'s λ-strong convexity (Shalev-Shwartz & Singer, 2007).

Dividing both sides of inequality (2) by $2\eta^{(t)}$ and rearranging, we obtain

$$\mathcal{L}^{(t)}(\mathbf{W}^{(t)}) - \mathcal{L}^{(t)}(\mathbf{W}^*) \leq \frac{D^{(t)}}{2\eta^{(t)}} - \frac{\lambda}{2}||\mathbf{W}^{(t)} - \mathbf{W}^*||^2 + \frac{\eta^{(t)}||\boldsymbol{\partial}^{(t)}||^2}{2} + \frac{(2+c)(2+h)}{\lambda}\|\mathbf{E}^{(t)}\|, \tag{4}$$

Summing over all t and dividing by T, we obtain

$$\frac{1}{T}\left(\sum_{t=1}^{T}\mathcal{L}^{(t)}(\mathbf{W}^{(t)})-\sum_{t=1}^{T}\mathcal{L}^{(t)}(\mathbf{W}^{*})\right)\leq\frac{1}{T}\sum_{t=1}^{T}\frac{D^{(t)}}{2\eta^{(t)}}-\frac{1}{T}\sum_{t=1}^{T}\frac{\lambda}{2}\|\mathbf{W}^{(t)}-\mathbf{W}^{*}\|^{2}$$
$$+\frac{1}{2T}\sum_{t=1}^{T}\eta^{(t)}\|\boldsymbol{\partial}^{(t)}\|^{2}+\frac{(2+c)(2+h)}{T\lambda}\sum_{t=1}^{T}\|\mathbf{E}^{(t)}\|. \tag{5}$$

We bound the first and second terms in the r.h.s. of inequality (5) as

$$\frac{1}{2T}\sum_{t=1}^{T}\left(\frac{D^{(t)}}{\eta^{(t)}}-\lambda||\mathbf{W}^{(t)}-\mathbf{W}^{*}||^{2}\right)=\frac{1}{2T}\left((\frac{1}{\eta^{(1)}}-\lambda)||\mathbf{W}^{(1)}-\mathbf{W}^{*}||^{2}+\right.$$
$$\left.\sum_{t=2}^{T}\left(\frac{1}{\eta^{(t)}}-\frac{1}{\eta^{(t-1)}}-\lambda\right)||\mathbf{W}^{(t)}-\mathbf{W}^{*}||^{2}-\frac{1}{\eta^{(T)}}||\mathbf{W}^{(T+1)}-\mathbf{W}^{*}||^{2}\right) \tag{6}$$
$$=_1-\frac{1}{2T\eta^{(T)}}||\mathbf{W}^{(T+1)}-\mathbf{W}^{*}||^{2}\leq 0.$$

In $=_1$, the first and second terms vanish after plugging in $\eta_t \equiv 1/(\lambda t)$.

Next, we bound the third term in the r.h.s. of inequality (5) as follows,

$$\begin{aligned}\frac{1}{2T}\sum_{t=1}^{T}\eta^{(t)}\|\boldsymbol{\partial}^{(t)}\|^{2}&=\frac{1}{2T}\sum_{t=1}^{T}\eta^{(t)}(\|\boldsymbol{\nabla}^{(t)}\|+\|\mathbf{E}_{prune}^{(t)}\|+\|\mathbf{E}_{dupl}^{(t)}\|)^{2}\\
&\leq\frac{1}{2T}\sum_{t=1}^{T}\frac{1}{\lambda t}\left((2+c)(1+h)+2+c+pt\beta^{t}(2+c)(1+h)\right)^{2}\\
&\leq\frac{1}{2T\lambda}\sum_{t=1}^{T}\frac{1}{t}\left((2+c)(2+h)+pt\beta^{t}(2+c)(2+h)\right)^{2}\\
&=\frac{(2+c)^{2}(2+h)^{2}}{2T\lambda}\sum_{t=1}^{T}\frac{1}{t}(1+pt\beta^{t})^{2}\\
&=\frac{(2+c)^{2}(2+h)^{2}}{2T\lambda}(\sum_{t=1}^{T}\frac{1}{t}+2p\sum_{t=1}^{T}\beta^{t}+p^{2}\sum_{t=1}^{T}t\beta^{2t})\\
&\leq_1\frac{(2+c)^{2}(2+h)^{2}}{2T\lambda}(\ln(T)+1+\frac{2p}{1-\beta}+\frac{p^{2}\beta^{2}}{(1-\beta^{2})^{2}})\\
&\leq\frac{(2+c)^{2}(2+h)^{2}}{2T\lambda}(\frac{3p}{(1-\beta)^{2}}+\ln(T)+1).\end{aligned} \tag{7}$$

In $\leq_1$ we bound the terms in the parentheses according to the divergence rate of the harmonic series, as well as according to upper bounds on the sum of low-order power series.

Next, we bound the fourth term in the r.h.s. of inequality (5) as follows,

$$\begin{aligned}\frac{(2+c)(2+h)}{T\lambda}\sum_{t=1}^{T}\|\mathbf{E}^{(t)}\|&\leq\frac{(2+c)(2+h)}{T\lambda}\sum_{t=1}^{T}(\|\mathbf{E}_{prune}^{(t)}\|+\|\mathbf{E}_{dupl}^{(t)}\|)\\
&\leq\frac{(2+c)(2+h)}{T\lambda}\sum_{t=1}^{T}\left(c+pt\beta^{t}(2+c)(1+h)\right)\\
&\leq\frac{(2+c)(2+h)c}{\lambda}+\frac{(2+c)^{2}(2+h)^{2}}{T\lambda}p\sum_{t=1}^{T}t\beta^{t}\\
&\leq\frac{(2+c)^{2}(2+h)}{\lambda}+\frac{(2+c)^{2}(2+h)^{2}}{T\lambda}\frac{p\beta}{(1-\beta)^{2}}.\end{aligned} \tag{8}$$

We bounded $\|\mathbf{E}^{(t)}_{prune}\|$ using the bound on $\|\Delta\mathbf{W}^{(t)}_{prune}\|$, and bounded $\|\mathbf{E}^{(t)}_{dupl}\|$ using the bound on $\|\mathbf{W}^{(t)}\|$. We obtain (1) by combining inequality (5) with inequalities (6), (7), and (8). □

Theorem 3. *Let $\mathcal{F}$ be a class of functions that MM can implement, and w.l.o.g. $\|\mathbf{x}\| \leq 1$. Then, with probability of at least $1-\delta$, the risk of any function $f \in \mathcal{F}$ is bounded from above as*

$$R(f) \leq \widetilde{R}_N(f) + \frac{4 + 4K\|\mathbf{W}\|}{\sqrt{N}} + (\|\mathbf{W}\| + 1)\sqrt{\frac{\ln\frac{1}{\delta}}{2N}}, \tag{9}$$

where $K = \sum_{i=1}^{M} b_i \sum_{j\neq i}^{M} b_j$, and b_i is the number of weights for the i^{th} class.

Proof. The proof closely follows the proof of Theorem 6 from (Guermeur, 2010). For the clarity of notation, we introduce $f_i(\mathbf{x}) = g(i, \mathbf{x}) = \max_j \mathbf{w}^T_{i,j}\mathbf{x}$, and $f_{i,j}(\mathbf{x}) = \mathbf{w}^T_{i,j}\mathbf{x}$, $i \in \{1, \ldots, M\}, j \in \{1, \ldots, b_i\}$. Then, let $\overline{\mathcal{F}}$ stand for the product space $\mathcal{F}^M$, so that $(f_1(\cdot), \ldots, f_M(\cdot)) \in \overline{\mathcal{F}}$. Additionally, in order to retain the generality of the Theorem and its proof, in the following we use κ to denote a kernel function as in (Guermeur, 2010), and $\Phi(\mathbf{x})$ to denote a kernel mapping from the original input space to the feature space induced by the kernel function κ. However, note that the MM model, although being non-linear classifier, uses a linear kernel to compare each weight $\mathbf{w}_{i,j}$ to a new data point, and in the following we can also set $\Phi(\mathbf{x}) = \mathbf{x}$. Further, let $\|\mathbf{w}\|_\infty \leq \Lambda_w$ and let $\forall \mathbf{x} \in \mathbb{R}^D$, $\|\mathbf{x}\| \leq \Lambda_{\Phi(\mathbb{R}^D)}$.

It follows,

$$\forall \overline{f} \in \overline{\mathcal{F}}, R(\overline{f}) \leq \widetilde{R}(\overline{f}). \tag{10}$$

Consequently,

$$\forall \overline{f} \in \overline{\mathcal{F}}, R(\overline{f}) \leq \widetilde{R}_N(\overline{f}) + \sup_{\overline{f}\in\overline{\mathcal{F}}} \left(\widetilde{R}(\overline{f}) - \widetilde{R}_N(\overline{f})\right). \tag{11}$$

The rest of the proof consists in the computation of an upper bound on the supremum of the empirical process appearing in (11). Let Z denote a random pair (X, Y) and Z_i its copies which constitute the N-sample $D_N : D_N = (Z_i)_{1\leq i\leq N}$. After simplifying notation this way, the bounded differences inequality can be applied to the supremum of interest by setting $n = N$, $(T_i)_{1\leq i\leq n} = D_N$ (i.e., $T_i = Z_i$), and $f(T_1, \ldots, T_n) = \sup_{\overline{f}\in\overline{\mathcal{F}}}\left(\widetilde{R}(\overline{f}) - \widetilde{R}_N(\overline{f})\right)$. The functions $\overline{f} \in \overline{\mathcal{F}}$ take their values in the interval $[-B_{\overline{\mathcal{F}}}, B_{\overline{\mathcal{F}}}]^M$, with $B_{\overline{\mathcal{F}}} = \Lambda_w \Lambda_{\Phi(X)}$. Consequently, the loss function associated with the risk $\widetilde{R}$ takes its values in the interval $[0, K_{\overline{\mathcal{F}}}]$. We can then get the following result (Guermeur, 2010): With probability of at least $1-\delta$,

$$\sup_{\overline{f}\in\overline{\mathcal{F}}}(\widetilde{R}(\overline{f}) - \widetilde{R}_N(\overline{f})) \leq \mathbb{E}_{D_N} \sup_{\overline{f}\in\overline{\mathcal{F}}}(\widetilde{R}(\overline{f}) - \widetilde{R}_N(\overline{f})) + K_{\overline{\mathcal{F}}}\sqrt{\frac{\ln(\frac{1}{\delta})}{2N}}. \tag{12}$$

Further, it can be shown that

$$\mathbb{E}_{D_N} \sup_{\overline{f}\in\overline{\mathcal{F}}}(\widetilde{R}(\overline{f}) - \widetilde{R}_N(\overline{f})) \leq 4\left(\frac{1}{\sqrt{N}} + \mathbb{E}_{\sigma,D_N}\left[\sup_{\overline{f}\in\overline{\mathcal{F}}} \frac{1}{N}\left|\sum_{i=1}^{N} \sigma_i \frac{1}{2}\left(\overline{f}_{Y_i}(X_i) - \max_{k\neq Y_i} \overline{f}_k(X_i)\right)\right|\right]\right). \tag{13}$$

In order to address the specific case of the considered MM model, we will introduce a different definition of cat than in the proof of Theorem 6 in (Guermeur, 2010). For $n \in \mathbb{N}^*$, let $z^n = ((x_i, y_i))_{1\leq i\leq n} \in (\mathbb{R}^D \times \mathcal{Y})^n$ and let cat be a mapping from $\overline{\mathcal{F}} \times \mathbb{R}^D \times \mathcal{Y}$ into $\{1, \ldots, M\}^2 \times \mathbb{N}^2$ such that

$$\begin{aligned}\forall(\overline{f}, x, y) \in \overline{\mathcal{F}} \times \mathbb{R}^D \times \mathcal{Y}, cat(\overline{f}, \mathbf{x}, y) =&(k, l, p, q) \Rightarrow (k = y) \wedge (l \neq y) \wedge \left(\overline{f}_l(\mathbf{x}) = \max_{i\neq y} \overline{f}_i(\mathbf{x})\right) \\ &\wedge (p = \arg\max_j \mathbf{w}^{\mathrm{T}}_{k,j}\mathbf{x}) \wedge (q = \arg\max_j \mathbf{w}^{\mathrm{T}}_{l,j}\mathbf{x}).\end{aligned} \tag{14}$$

The rest of the proof is straightforward modification of the proof of Theorem 6 in (Guermeur, 2010). By construction of the mapping cat,

$$\forall z^N \in (\mathbb{R}^D \times \mathcal{Y})^N, \frac{1}{2}\mathbb{E}_\sigma \left[\sup_{\overline{f} \in \overline{\mathcal{F}}} \left| \sum_{i=1}^{N} \sigma_i \left(\overline{f}_{y_i}(\mathbf{x}_i) - \max_{k \neq y_i} \overline{f}_k(\mathbf{x}_i) \right) \right| \right]$$

$$\leq \frac{1}{2}\mathbb{E}_\sigma \left[\sup_{\overline{f} \in \overline{\mathcal{F}}} \left| \sum_{k \neq l, p, q} \sum_{i: cat(\overline{f}, \mathbf{x}, y) = (k,l,p,q)} \sigma_i \left(\overline{f}_{k,p}(\mathbf{x}_i) - \overline{f}_{l,q}(\mathbf{x}_i) \right) \right| \right] \tag{15}$$

$$\leq \Lambda_w \mathbb{E}_\sigma \left[\sup_{\overline{f} \in \overline{\mathcal{F}}} \sum_{k \neq l, p, q} \left\| \sum_{i: cat(\overline{f}, \mathbf{x}, y) = (k,l,p,q)} \sigma_i \kappa(x_i, \cdot) \right\| \right].$$

Then, let Π_N be the set of all mappings π_N from $\{1, \ldots, N\}$ into $(k, l, p, q) \in \{1, \ldots, M\}^2 \times \mathbb{N}^2$, such that for all values of i, the pair (k, l) is always made up of two different values, while $p \in \{1, \ldots, b_k\}$ and $q \in \{1, \ldots, b_l\}$. It follows

$$\Lambda_w \mathbb{E}_\sigma \left[\sup_{\overline{f} \in \overline{\mathcal{F}}} \sum_{k \neq l, p, q} \left\| \sum_{i: cat(\overline{f}, \mathbf{x}, y) = (k,l,p,q)} \sigma_i \kappa(\mathbf{x}_i, \cdot) \right\| \right] \leq \Lambda_w \sum_{k \neq l, p, q} \mathbb{E}_\sigma \left[\sup_{\pi_N \in \Pi_N} \left\| \sum_{i: \pi_N = (k,l,p,q)} \sigma_i \kappa(\mathbf{x}_i, \cdot) \right\| \right]. \tag{16}$$

Consequently, to complete the derivation of the bound, it suffices to find a uniform upper bound on the expressions of the form

$$\mathbb{E}_\sigma \left\| \sum_{i \in \mathcal{I}_N} \sigma_i \kappa(\mathbf{x}_i, \cdot) \right\|, \tag{17}$$

where $\mathcal{I}_N$ is a subset of $\{1, \ldots, N\}$. By applying Jensen's inequality and using the fact that $\kappa(\mathbf{x}_i, \mathbf{x}_i) \geq 0$, a uniform upper bound of the above expression can be shown to be equal to

$$\mathbb{E}_\sigma \left\| \sum_{i \in \mathcal{I}_N} \sigma_i \kappa(\mathbf{x}_i, \cdot) \right\| \leq \Lambda_{\Phi(\mathbb{R}^D)} \sqrt{N}. \tag{18}$$

By substitution in the right-hand side of (16), and then in the right-hand side of (15), we get

$$\forall z^N \in (\mathbb{R}^D \times \mathcal{Y})^N, \frac{1}{2}\mathbb{E}_\sigma \left[\sup_{\overline{f} \in \overline{\mathcal{F}}} \left| \sum_{i=1}^{N} \sigma_i \left(\overline{f}_{y_i}(\mathbf{x}_i) - \max_{k \neq y_i} \overline{f}_k(\mathbf{x}_i) \right) \right| \right] \leq K \Lambda_w \Lambda_{\Phi(\mathbb{R}^D)} \sqrt{N}, \tag{19}$$

where $K = \sum_{i=1}^{M} b_i \sum_{j \neq i}^{M} b_j$, which implies that

$$\frac{1}{2}\mathbb{E}_{\sigma, D_N} \left[\sup_{\overline{f} \in \overline{\mathcal{F}}} \frac{1}{N} \left| \sum_{i=1}^{N} \sigma_i \left(\overline{f}_{y_i}(\mathbf{x}_i) - \max_{k \neq y_i} \overline{f}_k(\mathbf{x}_i) \right) \right| \right] \leq \frac{K \Lambda_w \Lambda_{\Phi(\mathbb{R}^D)}}{\sqrt{N}}. \tag{20}$$

In the case of MM, it is easy to see that $K_{\overline{\mathcal{F}}} = 1 + \Lambda_w \Lambda_{\Phi(\mathbb{R}^D)}$. Also, due to the assumptions of the Theorem, we can set $\Lambda_{\Phi(\mathbb{R}^D)} = 1$ and $\Lambda_w = \|\mathbf{W}\|$. Finally, combining inequalities (11), (12), (13), and (20) produces the bound (9), which concludes the proof.

As a concluding remark, we note that the main difference between proofs of Theorem 6 from (Guermeur, 2010) and the proof of Theorem 4 is in the definition of *cat* mapping. Unlike in (Guermeur, 2010), where the image of *cat* mapping is of cardinality $M \cdot (M - 1)$, the image of *cat* mapping for MM is of cardinality K, due to a larger number of weights per class. □

References

Guermeur, Y. Sample complexity of classifiers taking values in $\mathbb{R}^Q$, application to multi-class SVMs. *Communications in Statistics - Theory and Methods*, 39(3):543–557, 2010.

Kivinen, J., Smola, A. J., and Williamson, R. C. Online learning with kernels. *IEEE Transactions on Signal Processing*, 52(8):2165–2176, 2002.

Shalev-Shwartz, S. and Singer, Y. Logarithmic regret algorithms for strongly convex repeated games (Technical Report). The Hebrew University, 2007.

Wang, Z., Djuric, N., Crammer, K., and Vucetic, S. Trading representability for scalability: Adaptive multi-hyperplane machine for nonlinear classification. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2011.