Theorem 2. Let $W^*$ 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,

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

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 $W^{(t+1)} \leftarrow W^{(t)} - \eta^{(t)}\theta^{(t)}$, where $\theta^{(t)} = \nabla^{(t)} + E^{(t)}$, and $E^{(t)} = E_{\text{prune}}^{(t)} + E_{\text{dupl}}^{(t)}$ where we can see that the weight matrix degradation at the $t$th training iteration $E^{(t)}$ is equal to the sum of weight matrix degradation $E_{\text{prune}}^{(t)}$ due to pruning and weight matrix degradation $E_{\text{dupl}}^{(t)}$ due to weight duplication. Clearly, $E_{\text{prune}}^{(t)} = 0$ if no pruning is used, and $E_{\text{dupl}}^{(t)} = 0$ 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 $W^*$ at the $t$th round $D^{(t)}$ can be lower bounded as

$$
D^{(t)} = \frac{\|W^{(t)} - W^*\|^2 - \|W^{(t)} - \eta^{(t)}\nabla^{(t)} - \eta^{(t)}E^{(t)} - W^*\|^2}{\lambda}
$$

$$
= - (\eta^{(t)})^2 \|\theta^{(t)}\|^2 + 2\eta^{(t)}\|E^{(t)}\|W^{(t)} - W^*\| + 2\eta^{(t)}\|\nabla^{(t)}(W^{(t)} - W^*)\|
$$

$$
\geq 1 - (\eta^{(t)})^2 \|\theta^{(t)}\|^2 - 2\eta^{(t)}\|E^{(t)}\| (2 + c)(1 + h) + 2
$$

$$
+ \frac{(2 + c)(1 + h) + 2}{\lambda} \left(\mathcal{L}^{(t)}(W^{(t)}) - \mathcal{L}^{(t)}(W^*) + \frac{\lambda}{2}\|W^{(t)} - W^*\|^2\right),
$$

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

$$
\|W^{(t)}\| \leq \frac{(1 - \eta^{(t-1)})\lambda}{\lambda} \|W^{(t-1)}\| + 2\eta^{(t-1)} + \|\Delta_{\text{prune}} W^{(t-1)}\| + \|\Delta_{\text{dupl}} W^{(t-1)}\|
$$

$$
\leq \frac{t - 2}{t - 1} \|W^{(t-1)}\| + \frac{2}{(t - 1)\lambda} + \frac{c}{(t - 1)\lambda} + \frac{2 + c}{\lambda}
$$

$$
\leq \frac{1}{t - 1} \|W^{(0)}\| + \frac{2(t - 1)}{(t - 1)\lambda} + \frac{(t - 1)c}{(t - 1)\lambda} + \sum_{i=0}^{T-1} p\beta^i \frac{2 + c}{\lambda} \lambda \leq \frac{2 + c}{\lambda} (1 + h),
$$

where, in contrast to (Wang et al., 2011), we added the $\|\Delta_{\text{dupl}} W^{(t-1)}\|$ 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 $\beta$ whenever the weight duplication is performed, introducing the multiplication factor of $\sum_{t=0}^{T-1} p\beta^i$ 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 $\|W^{(t)} - W^*\| \leq (2 + c)(1 + h)/\lambda + 2/\lambda$ by using the fact that $\|W^*\| \leq 2/\lambda$ 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 $\mathcal{L}^{(t)}(W^{(t)})$’s $\lambda$-strong convexity (Shalev-Shwartz & Singer, 2007).

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

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

(4)
Summing over all $t$ and dividing by $T$, we obtain
\[
\frac{1}{T} \left( \sum_{t=1}^{T} \mathcal{L}^{(t)}(W^{(t)}) - \sum_{t=1}^{T} \mathcal{L}^{(t)}(W^*) \right) \leq \frac{1}{T} \sum_{t=1}^{T} \frac{D^{(t)}}{2\eta^{(t)}} - \frac{1}{T} \sum_{t=1}^{T} \frac{\lambda}{2} \|W^{(t)} - W^*\|^2
\]
\[
+ \frac{1}{2T} \sum_{t=1}^{T} \eta^{(t)} \|\theta^{(t)}\|^2 + \frac{(2 + c)(2 + h)}{T\lambda} \sum_{t=1}^{T} \|E^{(t)}\|.
\]
\[
\text{(5)}
\]
We bound the first and second terms in the r.h.s. of inequality (5) as follows,
\[
\frac{1}{2T} \sum_{t=1}^{T} \left( \frac{D^{(t)}}{\eta^{(t)}} - \lambda \|W^{(t)} - W^*\|^2 \right) = \frac{1}{2T} \left( \frac{1}{\eta^{(1)}} - \lambda \right) \|W^{(1)} - W^*\|^2 + \frac{1}{\eta^{(T)}} \|W^{(T+1)} - W^*\|^2
\]
\[
\sum_{t=2}^{T} \left( \frac{1}{\eta^{(t)}} - \frac{1}{\eta^{(t-1)}} - \lambda \right) \|W^{(t)} - W^*\|^2 - \frac{1}{\eta^{(T)}} \|W^{(T+1)} - W^*\|^2
\]
\[
= 1 - \frac{1}{2T\eta^{(T)}} \|W^{(T+1)} - W^*\|^2 \leq 0.
\]
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,
\[
\frac{1}{2T} \sum_{t=1}^{T} \eta^{(t)} \|\theta^{(t)}\|^2 = \frac{1}{2T} \sum_{t=1}^{T} \eta^{(t)} (\|\nabla^{(t)}\| + \|E_{\text{prune}}^{(t)}\| + \|E_{\text{dupl}}^{(t)}\|)^2
\]
\[
\leq \frac{1}{2T} \sum_{t=1}^{T} \frac{1}{\lambda t} \left( (2 + c)(1 + h) + 2 + c + pt\beta^t(2 + c)(1 + h) \right)^2
\]
\[
\leq \frac{1}{2T\lambda} \sum_{t=1}^{T} \frac{1}{t} \left( (2 + c)(2 + h) + pt\beta^t(2 + c)(2 + h) \right)^2
\]
\[
= \frac{(2 + c)^2(2 + h)^2}{2T\lambda} \sum_{t=1}^{T} \frac{1}{t}(1 + pt\beta^t)^2
\]
\[
= \frac{(2 + c)^2(2 + h)^2}{2T\lambda} \left( \sum_{t=1}^{T} \frac{1}{t} + 2p \sum_{t=1}^{T} \beta^t + p^2 \sum_{t=1}^{T} t\beta^{2t} \right)
\]
\[
\leq \frac{1}{2T\lambda} \frac{(2 + c)^2(2 + h)^2}{(1 - \beta)^2} \left( \ln(T) + 1 + \frac{2p}{1 - \beta} + \frac{p^2 \beta^2}{1 - \beta^2} \right)
\]
\[
\leq \frac{(2 + c)^2(2 + h)^2}{2T\lambda} \left( \frac{3p}{(1 - \beta)^2} + \ln(T) + 1. \right)
\]
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,
\[
\frac{(2 + c)(2 + h)}{T\lambda} \sum_{t=1}^{T} \|E^{(t)}\| \leq \frac{(2 + c)(2 + h)}{T\lambda} \sum_{t=1}^{T} (\|E_{\text{prune}}^{(t)}\| + \|E_{\text{dupl}}^{(t)}\|)
\]
\[
\leq \frac{(2 + c)(2 + h)}{T\lambda} \sum_{t=1}^{T} (c + pt\beta^t(2 + c)(1 + h))
\]
\[
\leq \frac{(2 + c)(2 + h)c}{\lambda} + \frac{(2 + c)^2(2 + h)^2}{T\lambda} p \sum_{t=1}^{T} t\beta^t
\]
\[
\leq \frac{(2 + c)^2(2 + h)}{\lambda} + \frac{(2 + c)^2(2 + h)^2}{T\lambda} \frac{p\beta}{(1 - \beta)^2}.
\]
We bounded \( \|E_{prune}\| \) using the bound on \( \|\Delta W_{prune}\| \), and bounded \( \|E_{dupl}\| \) using the bound on \( \|W\| \). We obtain (1) by combining inequality (5) with inequalities (6), (7), and (8).

\[ R(f) \leq \tilde{R}_N(f) + \frac{4 + 4K\|W\|}{\sqrt{N}} + (\|W\| + 1)\sqrt{\frac{\ln \frac{1}{\delta}}{2N}}, \]  

(9)

where \( K = \sum_{i=1}^{M} b_i \sum_{j \neq i} b_j \), 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 \( f_i(x) = g_i(x) \) and \( f_{i,j}(x) = w_i^j x \), \( i \in \{1, \ldots, M\}, j \in \{1, \ldots, b_i\} \). Then, let \( \mathcal{F} \) stand for the product space \( \mathcal{F}^M \), so that \( (f_1(\cdot), \ldots, f_M(\cdot)) \in \mathcal{F} \). Additionally, in order to retain the generality of the Theorem and its proof, in the following we use \( \kappa \) to denote a kernel function as in (Guermeur, 2010), and \( \Phi(x) \) to denote a kernel mapping from the original input space to the feature space induced by the kernel function \( \kappa \). However, note that the MM model, although being non-linear classifier, uses a linear kernel to compare each weight \( w_{i,j} \) to a new data point, and in the following we can also set \( \Phi(x) = x \). Further, let \( \|w\|_{\infty} \leq \Lambda_w \) and let \( \forall x \in \mathbb{R}^D, \|x\| \leq \Lambda_{\Phi(\mathbb{R}^D)} \).

It follows,

\[ \forall \tilde{f} \in \mathcal{F}, R(\tilde{f}) \leq \tilde{R}(\tilde{f}). \]  

(10)

Consequently,

\[ \forall \tilde{f} \in \mathcal{F}, R(\tilde{f}) \leq \tilde{R}_N(\tilde{f}) + \sup_{\tilde{f} \in \mathcal{F}} \left( \tilde{R}(\tilde{f}) - \tilde{R}_N(\tilde{f}) \right). \]  

(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 \( f(T_1, \ldots, T_n) = \sup_{\tilde{f} \in \mathcal{F}} \left( R(\tilde{f}) - \tilde{R}_N(\tilde{f}) \right) \). The functions \( \tilde{f} \in \mathcal{F} \) take their values in the interval \([-B_{\mathcal{F}}, B_{\mathcal{F}}]^M\), with \( B_{\mathcal{F}} = \Lambda_w \Lambda_{\Phi(X)} \). Consequently, the loss function associated with the risk \( \tilde{R} \) takes its values in the interval \([0, K_{\mathcal{F}}]\). We can then get the following result (Guermeur, 2010): With probability of at least \( 1 - \delta \),

\[ \sup_{\tilde{f} \in \mathcal{F}} \left( \tilde{R}(\tilde{f}) - \tilde{R}_N(\tilde{f}) \right) \leq \mathbb{E}_{D_N} \sup_{\tilde{f} \in \mathcal{F}} (R(\tilde{f}) - \tilde{R}_N(\tilde{f})) + K_{\mathcal{F}} \sqrt{\frac{\ln \frac{1}{\delta}}{2N}}. \]  

Further, it can be shown that

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

(13)

In order to address the specific case of the considered MM model, we will introduce a different definition of \( \text{cat} \) 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 \( \text{cat} \) be a mapping from \( \mathcal{F} \times \mathbb{R}^D \times \mathcal{Y} \) into \( \{1, \ldots, M\}^2 \times \mathbb{N}^2 \) such that

\[ \forall (\tilde{f}, x, y) \in \mathcal{F} \times \mathbb{R}^D \times \mathcal{Y}, \text{cat}(\tilde{f}, x, y) = (k, l, p, q) \Rightarrow (k = y) \land (l \neq y) \land \left( \tilde{f}_k(x) = \max_{i \neq y} \tilde{f}_i(x) \right) \land (p = \arg \max_j w^T_{k,j} x) \land (q = \arg \max_j w^T_{l,j} x). \]  

(14)

The rest of the proof is straightforward modification of the proof of Theorem 6 in (Guermeur, 2010). By construction of
By substitution in the right-hand side of (16), and then in the right-hand side of (15), we get

\[ \Lambda \text{cat} \]

As a concluding remark, we note that the main difference between proofs of Theorem 6 from (Guermeur, 2010) and the concludes the proof.


References
