Supplemental Materials of "Reinforced Angle-based Multicategory Support Vector Machines"

Chong Zhang[1] and Yufeng Liu[2,3,4,*] and Junhui Wang[5] and Hongtu Zhu[4]
[1]Department of Statistics and Actuarial Science
University of Waterloo
[2]Department of Statistics and Operations Research
[3]Department of Genetics
[4]Department of Biostatistics
University of North Carolina at Chapel Hill
[5]Department of Mathematics
City University of Hong Kong
[*]Email: yfliu@email.unc.edu

Appendix

Proof of Theorem 1. To prove Theorem 1, we need the following lemma from Zhang and Liu (2014).

Lemma 1 (Zhang and Liu, 2014, Lemma 1)**.** *Suppose we have an arbitrary* $\boldsymbol{f} \in \mathbb{R}^{k-1}$*. For any* $u, v \in \{1, \ldots, k\}$ *such that* $u \neq v$*, define* $\boldsymbol{T}_{u,v} = \boldsymbol{W}_u - \boldsymbol{W}_v$*. For any scalar* $z \in \mathbb{R}$*,* $\langle(\boldsymbol{f} + z\boldsymbol{T}_{u,v}), \boldsymbol{W}_w\rangle = \langle \boldsymbol{f}, \boldsymbol{W}_w\rangle$*, where* $w \in \{1, \ldots, k\}$ *and* $w \neq u, v$*. Furthermore, we have that* $\langle(\boldsymbol{f} + z\boldsymbol{T}_{u,v}), \boldsymbol{W}_u\rangle - \langle \boldsymbol{f}, \boldsymbol{W}_u\rangle = -\langle(\boldsymbol{f} + z\boldsymbol{T}_{v,u}), \boldsymbol{W}_v\rangle + \langle \boldsymbol{f}, \boldsymbol{W}_v\rangle$*.*

The proof consists of two parts. First we show that with $\gamma \leq 1/2$ the RAMSVM is Fisher consistent. Then we show that when $\gamma > 1/2$ the Fisher consistency cannot be guaranteed.

In this proof we assume $P_1 > P_2 \geq \cdots \geq P_k$. We need to show that $\langle \boldsymbol{f}^*(\boldsymbol{x}), \boldsymbol{W}_1\rangle > \langle \boldsymbol{f}^*(\boldsymbol{x}), \boldsymbol{W}_j\rangle$ for $j \neq 1$. First, we show that $\langle \boldsymbol{f}^*(\boldsymbol{x}), \boldsymbol{W}_1\rangle \geq \langle \boldsymbol{f}^*(\boldsymbol{x}), \boldsymbol{W}_j\rangle$ for any j. Note that if this is not true, then by Lemma 1, there exists $\boldsymbol{f}'(\boldsymbol{x}) \in \mathbb{R}^{k-1}$ such that $\langle \boldsymbol{f}^*(\boldsymbol{x}), \boldsymbol{W}_1\rangle = \langle \boldsymbol{f}'(\boldsymbol{x}), \boldsymbol{W}_j\rangle$ and $\langle \boldsymbol{f}^*(\boldsymbol{x}), \boldsymbol{W}_j\rangle = \langle \boldsymbol{f}'(\boldsymbol{x}), \boldsymbol{W}_1\rangle$. One can verify that $E[V(\boldsymbol{f}^*(\boldsymbol{X}), Y)|\boldsymbol{X} = \boldsymbol{x}] > E[V(\boldsymbol{f}'(\boldsymbol{X}), Y)|\boldsymbol{X} = \boldsymbol{x}]$, which contradicts to the definition of $\boldsymbol{f}^*$.

Next, we show that $\langle \boldsymbol{f}^*(\boldsymbol{x}), \boldsymbol{W}_1\rangle \le k-1$. Note that we have $\sum_{j=1}^k \langle \boldsymbol{f}^*(\boldsymbol{x}), \boldsymbol{W}_j\rangle = 0$. If $\langle \boldsymbol{f}^*(\boldsymbol{x}), \boldsymbol{W}_1\rangle > k-1$, there exists q such that $\langle \boldsymbol{f}^*(\boldsymbol{x}), \boldsymbol{W}_q\rangle < -1$. By Lemma 1, there exists $\boldsymbol{f}'(\boldsymbol{x}) \in \mathbb{R}^{k-1}$ such that $\langle \boldsymbol{f}'(\boldsymbol{x}), \boldsymbol{W}_j\rangle = \langle \boldsymbol{f}^*(\boldsymbol{x}), \boldsymbol{W}_j\rangle$ for $j \notin \{1, q\}$, $\langle \boldsymbol{f}'(\boldsymbol{x}), \boldsymbol{W}_1\rangle = \langle \boldsymbol{f}^*(\boldsymbol{x}), \boldsymbol{W}_1\rangle - \epsilon$, and $\langle \boldsymbol{f}'(\boldsymbol{x}), \boldsymbol{W}_q\rangle = \langle \boldsymbol{f}^*(\boldsymbol{x}), \boldsymbol{W}_q\rangle + \epsilon$, where ϵ is a small positive number. Now we have $E[V(\boldsymbol{f}^*(\boldsymbol{X}), Y)|\boldsymbol{X} = \boldsymbol{x}] - E[V(\boldsymbol{f}'(\boldsymbol{X}), Y)|\boldsymbol{X} = \boldsymbol{x}] = \{(1-P_1)(1-\gamma) + P_k\gamma\}\epsilon > 0$, which is a contradiction.

Next, we show that if $\gamma \le 1/2$, then $\langle \boldsymbol{f}^*(\boldsymbol{x}), \boldsymbol{W}_j\rangle \ge -1$ for any j. Suppose this is not true and $\langle \boldsymbol{f}^*(\boldsymbol{x}), \boldsymbol{W}_j\rangle < -1$ for a fixed $j \ne 1$. Because the dot product $\langle \boldsymbol{f}^*(\boldsymbol{x}), \boldsymbol{W}_1\rangle \le k-1$ is the maximum among all such dot products, we have that $-1 < \langle \boldsymbol{f}^*(\boldsymbol{x}), \boldsymbol{W}_q\rangle \le k-1$ for some q. Define $\boldsymbol{f}'(\boldsymbol{x})$ such that $\langle \boldsymbol{f}'(\boldsymbol{x}), \boldsymbol{W}_i\rangle = \langle \boldsymbol{f}^*(\boldsymbol{x}), \boldsymbol{W}_i\rangle$ for $i \notin \{j, q\}$, $\langle \boldsymbol{f}'(\boldsymbol{x}), \boldsymbol{W}_q\rangle = \langle \boldsymbol{f}^*(\boldsymbol{x}), \boldsymbol{W}_q\rangle - \epsilon$, and $\langle \boldsymbol{f}'(\boldsymbol{x}), \boldsymbol{W}_j\rangle = \langle \boldsymbol{f}^*(\boldsymbol{x}), \boldsymbol{W}_j\rangle + \epsilon$. One can verify that $E[V(\boldsymbol{f}'(\boldsymbol{X}), Y)|\boldsymbol{X} = \boldsymbol{x}] - E[V(\boldsymbol{f}^*(\boldsymbol{X}), Y)|\boldsymbol{X} = \boldsymbol{x}] = \{P_q - 1 + (1-P_j)\gamma\}\epsilon$. As $\gamma \le 1/2$, $\{P_q - 1 + (1-P_j)\gamma\} < 0$, hence this is a contradiction. Therefore, we have $\langle \boldsymbol{f}^*(\boldsymbol{x}), \boldsymbol{W}_j\rangle \ge -1$.

Lastly, using the above results and an argument similar to Lemma A.2 in Liu and Yuan (2011), we have that $\langle \boldsymbol{f}^*(\boldsymbol{x}), \boldsymbol{W}_1\rangle = k-1$ and $\langle \boldsymbol{f}^*(\boldsymbol{x}), \boldsymbol{W}_j\rangle = -1$ for $j \ne 1$. This completes the first part of the proof.

For the second part, we show that if $\gamma > 1/2$, then the RAMSVM can be inconsistent. We do so by giving a counter example. Let $k = 3$ and $P_3 = 0$. Then $E[V(\boldsymbol{f}'(\boldsymbol{X}), Y)|\boldsymbol{X} = \boldsymbol{x}] = \gamma P_1\{2 - \langle \boldsymbol{f}^*(\boldsymbol{x}), \boldsymbol{W}_1\rangle\}_+ + P_2(1-\gamma)\{1 + \langle \boldsymbol{f}^*(\boldsymbol{x}), \boldsymbol{W}_1\rangle\}_+ + \gamma P_2\{2 - \langle \boldsymbol{f}^*(\boldsymbol{x}), \boldsymbol{W}_2\rangle\}_+ + P_1(1-\gamma)\{1 + \langle \boldsymbol{f}^*(\boldsymbol{x}), \boldsymbol{W}_2\rangle\}_+ + (1-\gamma)\{1 + \langle \boldsymbol{f}^*(\boldsymbol{x}), \boldsymbol{W}_3\rangle\}_+$. If $1/2 < P_1 < \gamma$, then one can verify that the minimizer $\boldsymbol{f}^*$ is such that $\langle \boldsymbol{f}^*(\boldsymbol{x}), \boldsymbol{W}_1\rangle = \langle \boldsymbol{f}^*(\boldsymbol{x}), \boldsymbol{W}_2\rangle = 2$ and $\langle \boldsymbol{f}^*(\boldsymbol{x}), \boldsymbol{W}_3\rangle = -4$. Therefore, it is not Fisher consistent. $\square$

Next, we provide the dual problems of Guermeur (2012) and Liu and Yuan (2011). The MSVM framework proposed in Guermeur (2012) used K_1, K_2, K_3 and p as hyperparameters to denote different MSVM methods. The values of the hyperparameters that correspond to MSVMs 2-4 and 6 are reported in Table A1.

Dual Problems of Soft Margin MSVM in Guermeur (2012). The dual problem of soft margin MSVM in Guermeur (2012) is

$$\max -\frac{1}{4}\left\{\boldsymbol{\alpha}^T M_1 \boldsymbol{\alpha}\right\} + M_2\boldsymbol{\alpha},$$

$$s.t. \begin{cases} 0 \le (1-K_3)(2-p)\alpha_{i,j} \le (2-p)m(i,j), & i=1,\ldots,n,\ j\neq y_i, \\ 0 \le K_3(2-p)\sum_{j\neq y_i}\alpha_{i,j} \le (2-p)\sum_{j\neq y_i} m(i,j), & i=1,\ldots,n, \\ (p-1)\alpha_{i,j} \ge 0, & i=1,\ldots,n,\ j\neq y_i, \\ \sum_{i=1}^n\sum_{l=1}^k \{K_1\delta_{y_i,j} + (1-K_1)/k - \delta j, l\}\alpha_{i,l} = 0, & j=1,\ldots,k-1. \end{cases} \tag{A.1}$$

Here M_1 and M_2 are fixed matrices, $m(i,j)$ is a real number, and δ is the Kronecker symbol. Both M_1 and M_2 depend only on the MSVM method, and $m(i,j)$ depends on i, j and the MSVM method. For more details about M_1, M_2 and $m(i,j)$, see Guermeur (2012). One can verify that for any set of hyperparameters, the equality constraints in (A.1) do not vanish. Notice that (A.1) includes the dual problems of (2) as a special case.

Dual Problems of Hard Margin MSVM in Guermeur (2012). The dual problems of hard margin MSVM in Guermeur (2012) can be written as

$$\max -\frac{1}{4}\left\{\boldsymbol{\alpha}^T M_1' \boldsymbol{\alpha}\right\} + M_2'\boldsymbol{\alpha},$$

$$s.t. \begin{cases} \alpha_{i,j} \ge 0, & i=1,\ldots,n,\ j\neq y_i, \\ \sum_{i=1}^n\sum_{l=1}^k \{K_1\delta_{y_i,j} + (1-K_1)/k - \delta j, l\}\alpha_{i,l} = 0, & j=1,\ldots,k-1. \end{cases} \tag{A.2}$$

Here M_1' and M_2' are fixed matrices, similar to M_1 and M_2 in the soft margin case.

MSVM	p	K_1	K_2	K_3
MSVM2	1	1	1	0
MSVM3	1	1	1	1
MSVM4	1	0	1/(k-1)	0
MSVM6	2	0	1/(k-1)	0

Table A1: Hyperparameters for different MSVM methods in Guermeur (2012).

Dual Problems in Liu and Yuan (2011). For the optimization in Liu and Yuan (2011), its dual problem can be written as

$$\min \boldsymbol{\beta}^T H\boldsymbol{\beta} + g^T\boldsymbol{\beta},$$

$$s.t. \begin{cases} 0 \le \alpha_{i,j} \le A_{i,j}, & i=1,\ldots,\ j=1,\ldots,k, \\ E\boldsymbol{\beta} = 0, & j=1,\ldots,k, \end{cases} \tag{A.3}$$

where $\boldsymbol{\beta} = (\boldsymbol{\alpha}_{\cdot 1}^T, \ldots, \boldsymbol{\alpha}_{\cdot k}^T)^T$, and H, E are fixed matrices that depend only on the observed predictors and the kernel function $K(\cdot,\cdot)$. For more information about H and E, see Section 3 in Liu and Yuan (2011).

References

Guermeur, Y. (2012). A Generic Model of Multi-Class Support Vector Machine. *International Journal of Intelligent Information and Database Systems*, **6**(6), 555–577.

Liu, Y. and Yuan, M. (2011). Reinforced Multicategory Support Vector Machines. *Journal of Computational and Graphical Statistics*, **20**(4), 901–919.

Zhang, C. and Liu, Y. (2014). Multicategory Angle-based Large-margin Classification. *Biometrika*, **101**(3), 625–640.