# A Theoretical Characterization of Linear SVM-Based Feature Selection: Supplement

**Douglas Hardin** DOUG.HARDIN@VANDERBILT.EDU
Department of Mathematics, Vanderbilt University, Stevenson Center, Nashville, TN 37240-0001

**Ioannis Tsamardinos** IOANNIS.TSAMARDINOS@VANDERBILT.EDU
Dept. of Biomedical Informatics, Vanderbilt University, 2209 Garland Ave, Nashville, TN 37232-8340, USA

**Constantin F. Aliferis** CONSTANTIN.ALIFERIS@VANDERBILT.EDU
Dept. of Biomedical Informatics, Vanderbilt University, 2209 Garland Ave, Nashville, TN 37232-8340, USA

## A. Proof of Lemma 1

*Proof of Lemma 1.* As we outline below, the optimization problems (L.1) and (L.2) can be expressed in terms of an objective function $F(\mathbf{w}, b)$ that is minimized over a compact convex set $C \subset \mathbf{R}^N \times \mathbf{R}$ where the objective function $F$ is continuous and can be written in the form

$$F(\mathbf{w}, b) = h(\mathbf{w}) + g(\mathbf{w}, b) \qquad (1)$$

for some convex function $g$ and where $h$ is a strictly convex function. In fact, in all the cases considered here we have $h(\mathbf{w}) := |\mathbf{w}|^2$.

First we show that the respective optimization problems can be restricted to only considering $(\mathbf{w}, b)$ in a compact convex set $C$. Suppose $(\mathbf{X}, Y)$ is separable. For the sample limit hard-margin problem (L.1) we construct $C$ as follows. Let $\mathcal{W}$ denote the collection of "admissible" $(\mathbf{w}, b)$ satisfying the constraint (8). Let $(\mathbf{w}^*, b^*) \in \mathcal{W}$, $r_* = |\mathbf{w}^*|$, and define $C := \{(\mathbf{w}, b) \in \mathcal{W} \mid |\mathbf{w}| \le r_*\}$. By hypothesis there is some $R$ is such that $|\mathbf{X}| \le R$ almost surely. If $(\mathbf{w}, b) \in C$ then $|\mathbf{w}| \le r_*$ and it follows that $|b| \le 1 + r_* R$ showing that $C$ is bounded and hence compact. For the hard-margin case we have $g(\mathbf{w}, b) = 0$. Then $F(\mathbf{w}, b) = |\mathbf{w}|^2 > F(\mathbf{w}^*, b^*) = r_*^2$ for $(\mathbf{w}, b) \in \mathcal{W} \setminus C$ and so $\min_{(\mathbf{w},b)\in\mathcal{W}} F(\mathbf{w}, b) = \min_{(\mathbf{w},b)\in C} F(\mathbf{w}, b)$.

In the sample limit soft-margin (unconstrained form) case (L.3), we have $g(\mathbf{w}, b) = CE([1 - Y(\mathbf{w}\cdot\mathbf{X} + b)]_+^p)$. Fix $(\mathbf{w}^*, b^*)$ and let $r_* := \sqrt{F(\mathbf{w}^*, b^*)}$. If $|\mathbf{w}| > r_*$, then $F(\mathbf{w}, b) \ge h(\mathbf{w}) > r_*^2 = F(\mathbf{w}^*, b^*)$. Since $\min_{|\mathbf{w}|\le r_*} g(\mathbf{w}, b) \to \infty$ as $|b| \to \infty$, there is some $\gamma > 0$ such that $g(\mathbf{w}, b) \ge r_*$ for $|b| > \gamma$ and $|\mathbf{w}| \le r_*$. Hence, letting $C := \{(\mathbf{w}, b) \mid |\mathbf{w}| \le r_* \text{ and } |b| \le \gamma\}$ we have $\min_{(\mathbf{w},b)\in\mathbf{R}^N\times\mathbf{R}} F(\mathbf{w}, b) = \min_{(\mathbf{w},b)\in C} F(\mathbf{w}, b)$.

Since $F$ is continuous and $C$ is compact then there must be some global minimizer $(\boldsymbol{\omega}, \beta) \in C$. Suppose $F(\boldsymbol{\omega}', \beta') = F(\boldsymbol{\omega}, \beta)$ then $(\tilde{\boldsymbol{\omega}}, \tilde{\beta}) := ((\boldsymbol{\omega} + \boldsymbol{\omega}')/2, \beta + \beta')/2) \in C$ since $C$ is convex. If $\boldsymbol{\omega}' \ne \boldsymbol{\omega}$ then we have (since $g$ is convex and $h$ is strictly convex) $g(\tilde{\boldsymbol{\omega}}, \tilde{\beta}) \le (g(\boldsymbol{\omega}, \beta) + g(\boldsymbol{\omega}', \beta'))/2$ and $h(\tilde{\boldsymbol{\omega}}) < (h(\boldsymbol{\omega}) + g(\boldsymbol{\omega}'))/2$ and thus $F(\tilde{\boldsymbol{\omega}}, \tilde{\beta}) < F(\boldsymbol{\omega}, \beta)$ contradicting the fact that $F(\boldsymbol{\omega}, \beta) = \min_{(\mathbf{w},b))\in C} F(\mathbf{w}, b)$. Hence, $\boldsymbol{\omega}' = \boldsymbol{\omega}$.

The $m$-sample optimization problems (F.1) and (F.3) can be expressed in terms of minimizing an objective function $F_m(\mathbf{w}, b) = h(\mathbf{w}) + g_m(\mathbf{w}, b)$ of the form (1) over a compact convex set $C_m$. In the hard-margin case $g_m = 0$ and $C = \bigcap_m C_m$ almost surely. In the soft-margin case we can choose $C_m = C$ and

$$g_m(\mathbf{w}, b) = \frac{C}{m} \sum_{k=1}^{m} [1 - y_k(\mathbf{w}\cdot\mathbf{x}_k - b)]_+^p.$$

In both cases we have with probability 1 that $g_m \to g$ uniformly on compact sets in $\mathbf{R}^N \times \mathbf{R}$ as $m \to \infty$ and that $C_m \supset C$ and $C = \bigcap_m C_m$. Let $(\mathbf{w}_m, b_m) \in C_m$ be a global minimizer for $F_m$. Suppose $(\mathbf{w}^*, b^*)$ is a limit point of the sequence $(\mathbf{w}_m, b_m)$ (that is, some subsequence $(\mathbf{w}_m, b_m)$ converges to $(\mathbf{w}^*, b^*)$) and that $\mathbf{w}^* \ne \boldsymbol{\omega}$. Then

$$\limsup_{m\to\infty} F_m(\mathbf{w}_m, b_m) \ge F(\mathbf{w}^*, b^*) > F(\boldsymbol{\omega}, \beta). \qquad (2)$$

On the other hand, we have

$$\limsup_{m\to\infty} F_m(\mathbf{w}_m, b_m) \le \limsup_{m\to\infty} F_m(\boldsymbol{\omega}, \beta) = F(\boldsymbol{\omega}, \beta)$$

which contradicts (2). Hence any limit point $(\mathbf{w}^*, b^*)$ of the sequence $(\mathbf{w}_m, b_m)$ must have $\mathbf{w}^* = \boldsymbol{\omega}$. Since the $(b_m)$ is a bounded sequence any subsequence of $(b_m)$ must contain a convergent subsequence showing that

$\boldsymbol{\omega}$ is the only limit possible limit point of the sequence $(\mathbf{w}_m)$. Since $(\mathbf{w}_m)$ is a bounded sequence in $\mathbf{R}^N$ this implies that $\mathbf{w}_m \to \boldsymbol{\omega}$ almost surely.

Finally, we address the uniqueness of $\beta$. In the hard margin case, there must be some $(\mathbf{x}, y)$ in the support of $(\mathbf{X}, Y)$ such that $y(\boldsymbol{\omega} \cdot \mathbf{x} + b) = 1$ otherwise $(\boldsymbol{\omega}, \beta)$ is an interior point of the constraint set and cannot be a minimizer. For the soft-margin case with $p > 1$ it is more direct to consider the equivalent constrained problem (L.2). In this case the objective function $\mathcal{F}_s(\mathbf{w}, b, \xi) = h(\mathbf{w}) + q(\xi)$ where $q$ is a strictly convex function of $\xi = \xi(\mathbf{X}, Y)$ and it follows that $(\mathbf{w}, \xi)$ are uniquely defined for any minimizer $(\mathbf{w}, b, \xi)$ almost surely. Then, as in the hard-margin case, there must be some $(\mathbf{x}, y)$ in the support of $(\mathbf{X}, Y)$ for which we have equality in the constraints and thus $b$ is uniquely determined. □