Supplemental Material: Proofs

Proof to Theorem 1

Proof. Let n^* be the minimal number of training items to ensure a unique solution $\boldsymbol{\theta}^*$. First consider the case $n^* = 0$. It happens if and only if $\boldsymbol{\theta}^* = \mathbf{0}$ and $\text{Rank}(A) = d$, which is a special case of $A\boldsymbol{\theta}^* = \mathbf{0}$. Clearly, this case is consistent with LB1. Next consider the case $n^* \geq 1$. Since $\boldsymbol{\theta}^*$ solves (1), the KKT condition holds:

$$-\lambda A\boldsymbol{\theta}^* \in \sum_{i=1}^{n^*} \partial_1 \ell(\mathbf{x}_i^\top \boldsymbol{\theta}^*, y_i)\mathbf{x}_i. \tag{29}$$

We seek all $\boldsymbol{\delta}$ such that $\boldsymbol{\theta}^* + \boldsymbol{\delta}$ satisfies

$$A(\boldsymbol{\theta}^* + \boldsymbol{\delta}) = A\boldsymbol{\theta}^* \quad \text{and} \quad \mathbf{x}_i^\top(\boldsymbol{\theta}^* + \boldsymbol{\delta}) = \mathbf{x}_i^\top \boldsymbol{\theta}^* \quad \forall i = 1, \cdots, n^*, \tag{30}$$

For any such $\boldsymbol{\delta}$, simple algebra verifies that $\boldsymbol{\theta}^* + t\boldsymbol{\delta}$ satisfies the KKT condition (29) for any $t \in [0, 1]$. Consequently, $\boldsymbol{\theta}^* + \boldsymbol{\delta}$ also solves the problem in (1). To see this, we consider two situations:

- If the loss function $\ell(\cdot, \cdot)$ is convex in the first argument, the KKT condition is a sufficient optimality condition, which means that $\boldsymbol{\theta}^* + \boldsymbol{\delta}$ solves (1).
- If the loss function $\ell(\cdot, \cdot)$ is smooth (not necessary convex) in the first argument, we have $f(\boldsymbol{\theta}^*) = f(\boldsymbol{\theta}^* + \boldsymbol{\delta})$ by using the Taylor expansion (recall f is defined in equation 1):

$$\begin{aligned}
f(\boldsymbol{\theta}^* + \boldsymbol{\delta}) =& f(\boldsymbol{\theta}^*) + \langle \nabla f(\boldsymbol{\theta}^* + t\boldsymbol{\delta}), \boldsymbol{\delta} \rangle \quad (\text{for some } t \in [0, 1]) \\
=& f(\boldsymbol{\theta}^*) + \left\langle \sum_{i=1}^{n^*} \nabla_1 \ell(\mathbf{x}_i^\top(\boldsymbol{\theta}^* + t\boldsymbol{\delta}),\ y_i)\mathbf{x}_i + \lambda A(\boldsymbol{\theta}^* + t\boldsymbol{\delta})),\ \boldsymbol{\delta} \right\rangle \\
=& f(\boldsymbol{\theta}^*) + \left\langle \underbrace{\sum_{i=1}^{n^*} \nabla_1 \ell(\mathbf{x}_i^\top \boldsymbol{\theta}^*,\ y_i)\mathbf{x}_i + \lambda A\boldsymbol{\theta}^*}_{=\mathbf{0} \text{ due to the KKT condition (29)}},\ \boldsymbol{\delta} \right\rangle \\
=& f(\boldsymbol{\theta}^*).
\end{aligned}$$

Therefore, $\boldsymbol{\theta}^* + \boldsymbol{\delta}$ also solves (1). However, the uniqueness of $\boldsymbol{\theta}^*$ requires $\boldsymbol{\delta} = \mathbf{0}$ to be the only value satisfying (30). This is equivalent to say

$$\text{Null}(A) \cap \text{Null}(\text{Span}\{\mathbf{x}_1, \cdots, \mathbf{x}_{n^*}\}) = \{\mathbf{0}\}. \tag{31}$$

It indicates that

$$\text{Rank}(A) + \text{Dim}(\text{Span}\{\mathbf{x}_1, \cdots, \mathbf{x}_{n^*}\}) \geq d.$$

From $n^* \geq \text{Dim}(\text{span}\{\mathbf{x}_1, \cdots, \mathbf{x}_{n^*}\})$, we have $n^* \geq d - \text{Rank}(A)$. We proved the general case for LB1.
If we have $A\boldsymbol{\theta}^* \neq \mathbf{0}$, we can further improve LB1. Let $\mathbf{g}^* = (g_1^*, \ldots, g_{n^*}^*)^\top$ be the vector satisfying

$$-\lambda A\boldsymbol{\theta}^* = \sum_{i=1}^{n^*} g_i^* \mathbf{x}_i \quad \text{and} \quad g_i^* \in \partial_1 \ell(\mathbf{x}_i^\top \boldsymbol{\theta}^*, y_i) \quad \forall i = 1, 2, \cdots, n^*. \tag{32}$$

Since $\boldsymbol{\theta}^*$ satisfies the KKT condition, such vector $\mathbf{g}^*$ must exist. Applying $A\boldsymbol{\theta}^* \neq \mathbf{0}$ to (32), we have $\mathbf{g}^* \neq \mathbf{0}$ and

$$\text{Dim}\left(\text{Span}\{A_{\cdot 1}, A_{\cdot 2}, \cdots, A_{\cdot d}\} \cap \text{Span}\{\mathbf{x}_1, \cdots, \mathbf{x}_{n^*}\}\right) \geq 1. \tag{33}$$

To satisfy (31), we must have

$$d = \text{Dim}\left(\text{Span}\{A_{\cdot 1}, A_{\cdot 2}, \cdots, A_{\cdot d}, \mathbf{x}_1, \cdots, \mathbf{x}_{n^*}\}\right).$$

Using the fact in linear algebra

$$\begin{aligned}
&\text{Dim}\left(\text{Span}\{A_{.1}, A_{.2}, \cdots, A_{.d}, \mathbf{x}_1, \cdots, \mathbf{x}_{n^*}\}\right) \\
=&\underbrace{\text{Dim}\left(\text{Span}\{A_{.1}, A_{.2}, \cdots, A_{.d}\}\right)}_{=\text{Rank}(A)} + \\
&\underbrace{\text{Dim}\left(\text{Span}\{\mathbf{x}_1, \cdots, \mathbf{x}_{n^*}\}\right)}_{\leq n^*} - \\
&\underbrace{\text{Dim}\left(\text{Span}\{A_{.1}, A_{.2}, \cdots, A_{.d}\} \cap \text{Span}\{\mathbf{x}_1, \cdots, \mathbf{x}_{n^*}\}\right)}_{\geq 1 \text{ (from (33))}}
\end{aligned}$$

We conclude that $n^* \geq d - \text{Rank}(A) + 1$. We completed the proof for LB1. □

Proof to Theorem 2

Proof. When A has full rank we have an equivalent expression for the KKT condition (29):

$$-\lambda A^{\frac{1}{2}}\boldsymbol{\theta}^* \in \sum_{i=1}^{n^*} A^{-\frac{1}{2}}\mathbf{x}_i \partial_1 \ell(\mathbf{x}_i^\top \boldsymbol{\theta}^*, y_i) \quad \forall i = 1, \cdots, n^*. \tag{34}$$

Let us decompose $A^{-\frac{1}{2}}\mathbf{x}_i$ for all $i = 1, \cdots, n^*$ into $A^{-\frac{1}{2}}\mathbf{x}_i = \alpha_i A^{\frac{1}{2}}\boldsymbol{\theta}^* + \mathbf{u}_i$, where $\mathbf{u}_i$ is orthogonal to $A^{\frac{1}{2}}\boldsymbol{\theta}^*$: $\mathbf{u}_i^\top A^{\frac{1}{2}}\boldsymbol{\theta}^* = 0$. Equivalently $\mathbf{x}_i = \alpha_i A\boldsymbol{\theta}^* + A^{\frac{1}{2}}\mathbf{u}_i$. Applying this decomposition, we have

$$\mathbf{x}_i^\top \boldsymbol{\theta}^* = \alpha_i \|\boldsymbol{\theta}^*\|_A^2 + \mathbf{u}_i^\top A^{\frac{1}{2}}\boldsymbol{\theta}^* = \alpha_i \|\boldsymbol{\theta}^*\|_A^2.$$

Putting it back in (34) we obtain

$$-\lambda A^{\frac{1}{2}}\boldsymbol{\theta}^* \in \sum_{i=1}^{n^*} \left(\alpha_i A^{\frac{1}{2}}\boldsymbol{\theta}^* + \mathbf{u}_i\right) \partial_1 \ell(\alpha_i \|\boldsymbol{\theta}^*\|_A^2, y_i) \quad \forall i = 1, \cdots, n^*. \tag{35}$$

Since $\mathbf{u}_i$ is orthogonal to $A^{\frac{1}{2}}\boldsymbol{\theta}^*$, (35) can be rewritten as

$$\exists \alpha_i \in \mathbb{R},\ \exists y_i \in \mathcal{Y},\ \exists g_i \in \partial_1 \ell(\alpha_i \|\boldsymbol{\theta}^*\|_A^2, y_i) \quad \forall i = 1, \cdots, n^* \tag{36}$$

$$\text{satisfying} \quad \sum_{i=1}^{n^*} g_i \mathbf{u}_i = 0$$

$$-\lambda A^{\frac{1}{2}}\boldsymbol{\theta}^* = A^{\frac{1}{2}}\boldsymbol{\theta}^* \sum_{i=1}^{n^*} \alpha_i g_i \tag{37}$$

Since $A\boldsymbol{\theta}^* \neq 0$, we have $A^{\frac{1}{2}}\boldsymbol{\theta}^* \neq 0$ and (37) is equivalent to $-\lambda = \sum_{i=1}^{n^*} \alpha_i g_i$. It follows that

$$\lambda = -\sum_{i=1}^{n^*} \alpha_i g_i \leq n^* \sup_{\alpha \in \mathbb{R}, y \in \mathcal{Y}, g \in \partial_1 \ell(\alpha \|\boldsymbol{\theta}^*\|_A^2, y)} -\alpha g = n^* \sup_{\alpha \in \mathbb{R}, y \in \mathcal{Y}, g \in -\partial_1 \ell(\alpha \|\boldsymbol{\theta}^*\|_A^2, y)} \alpha g$$

It indicates the lower bound for n^*

$$n^* \geq \left\lceil \frac{\lambda}{\sup_{\alpha \in \mathbb{R}, y \in \mathcal{Y}, g \in -\partial_1 \ell(\alpha \|\boldsymbol{\theta}^*\|_A^2, y)} \alpha g} \right\rceil.$$

□

Proof to Theorem 3

Proof. Let $D = \{\mathbf{x}_i, y_i\}_{i=1,\cdots,n}$ be a teaching set for $[\mathbf{w}^*; b^*]$. The following KKT condition needs to be satisfied:

$$\mathbf{0} \in \sum_{i=1}^{n} \partial \ell(y_i(\mathbf{x}_i^\top \mathbf{w}^* + b^*)) y_i \begin{bmatrix} \mathbf{x}_i \\ 1 \end{bmatrix} + \begin{bmatrix} \lambda A \mathbf{w}^* \\ 0 \end{bmatrix}. \tag{38}$$

If we construct a new training set

$$\hat{D} = \left\{ \hat{\mathbf{x}}_i = \mathbf{x}_i + \frac{b^*}{\|\mathbf{w}^*\|_A^2} A\mathbf{w}^*,\ \hat{y}_i = y_i \right\}_{i=1,\cdots,n}$$

then $[\mathbf{w}^*; 0]$ satisfies the KKT condition defined on $\hat{D}$. This can be verified as follows:

$$\begin{aligned}
&\sum_{i=1}^{n} \partial \ell(\hat{y}_i(\hat{\mathbf{x}}_i^\top \mathbf{w}^*)) \hat{y}_i \begin{bmatrix} \hat{\mathbf{x}}_i \\ 1 \end{bmatrix} + \begin{bmatrix} \lambda A \mathbf{w}^* \\ 0 \end{bmatrix} \\
=&\sum_{i=1}^{n} \partial \ell(y_i(\mathbf{x}_i^\top \mathbf{w}^* + b^*)) y_i \begin{bmatrix} \mathbf{x}_i + \frac{b^*}{\|\mathbf{w}^*\|_A^2} A\mathbf{w}^* \\ 1 \end{bmatrix} + \begin{bmatrix} \lambda A \mathbf{w}^* \\ 0 \end{bmatrix} \\
=&\underbrace{\sum_{i=1}^{n} \partial \ell(y_i(\mathbf{x}_i^\top \mathbf{w}^* + b^*)) y_i \begin{bmatrix} \mathbf{x}_i \\ 1 \end{bmatrix} + \begin{bmatrix} \lambda A \mathbf{w}^* \\ 0 \end{bmatrix}}_{\ni \mathbf{0} \text{ from (38)}} + \begin{bmatrix} \frac{b^*}{\|\mathbf{w}^*\|_A^2} A\mathbf{w}^* \\ 0 \end{bmatrix} \underbrace{\sum_{i=1}^{n} \partial \ell(y_i(\mathbf{x}_i^\top \mathbf{w}^* + b^*)) y_i}_{\ni 0 \text{ from (38)}} \\
\ni& \mathbf{0}
\end{aligned} \tag{39}$$

where $0 \in \sum_{i=1}^{n} \partial \ell(y_i(\mathbf{x}_i^\top \mathbf{w}^* + b^*)) y_i$ is from the bias dimension in (38). It follows that

$$\mathbf{0} \in \sum_{i=1}^{n} \partial \ell(\hat{y}_i \hat{\mathbf{x}}_i^\top \mathbf{w}^*) \hat{y}_i \hat{\mathbf{x}}_i + \lambda A \mathbf{w}^*$$

which is equivalent to

$$\begin{aligned}
\mathbf{0} &\in \sum_{i=1}^{n} \partial \ell(\hat{y}_i \hat{\mathbf{x}}_i^\top \mathbf{w}^*) A^{-\frac{1}{2}} \underbrace{\hat{y}_i \hat{\mathbf{x}}_i}_{=:\mathbf{z}_i} + \lambda A^{\frac{1}{2}} \mathbf{w}^* \\
&= \sum_{i=1}^{n} \partial \ell(\mathbf{z}_i^\top \mathbf{w}^*) A^{-\frac{1}{2}} \mathbf{z}_i + \lambda A^{\frac{1}{2}} \mathbf{w}^*.
\end{aligned} \tag{40}$$

We decompose $A^{-\frac{1}{2}} \mathbf{z}_i = \alpha_i A^{\frac{1}{2}} \mathbf{w}^* + \mathbf{u}_i$ where $\mathbf{u}_i$ satisfies $\mathbf{u}_i^\top A^{\frac{1}{2}} \mathbf{w}^* = 0$. Applying this decomposition to (40), we have

$$\lambda A^{\frac{1}{2}} \mathbf{w}^* \in \sum_{i=1}^{n} -\partial \ell(\alpha_i \|\mathbf{w}^*\|_A^2)(\alpha_i A^{\frac{1}{2}} \mathbf{w}^* + \mathbf{u}_i). \tag{41}$$

Since $\mathbf{u}_i$ is orthogonal to $A^{\frac{1}{2}} \mathbf{w}^*$, (41) implies that

$$\lambda A^{\frac{1}{2}} \mathbf{w}^* \in \sum_{i=1}^{n} -\partial \ell(\alpha_i \|\mathbf{w}^*\|_A^2) \alpha_i A^{\frac{1}{2}} \mathbf{w}^*.$$

Since $\mathbf{w}^* \neq \mathbf{0}$ we have

$$\lambda \in \sum_{i=1}^{n} -\partial \ell(\alpha_i \|\mathbf{w}^*\|_A^2) \alpha_i.$$

Together with

$$\sum_{i=1}^{n} -\partial \ell(\alpha_i \|\mathbf{w}^*\|_A^2) \alpha_i \leq n \sup_{\alpha \in \mathbb{R}, g \in -\partial \ell(\alpha \|\mathbf{w}^*\|_A^2)} \alpha g,$$

we obtain LB3. □

Proof to Proposition 1

Proof. We simply verify the KKT condition to see that $\boldsymbol{\theta}^*$ is a solution to (10) by applying the construction in (11). The uniqueness of $\boldsymbol{\theta}^*$ is guaranteed by the strong convexity of (10). □

Proof to Proposition 2

Proof. We only need to verify that the KKT condition holds for $\boldsymbol{\theta}^*$. Due to the strong convexity of (12) uniqueness is guaranteed automatically. We denote the subgradient $\partial_a \max(1-a,0) = -\partial_1 \max(1-a,0) = -\mathbf{I}(a)$, where

$$\mathbf{I}(a) = \begin{cases} 1, & \text{if } a < 1 \\ [0,1], & \text{if } a = 1 \\ 0, & \text{otherwise} \end{cases}. \tag{42}$$

The KKT condition is

$$\begin{aligned} &\sum_{i=1}^{n} -y_i \mathbf{x}_i \partial_1 \max(1 - y_i \mathbf{x}_i^\top \boldsymbol{\theta}^*, 0) + \lambda \boldsymbol{\theta}^* \\ =&\sum_{i=1}^{n} -y_i \mathbf{x}_i \mathbf{I}(y_i \mathbf{x}_i^\top \boldsymbol{\theta}^*) + \lambda \boldsymbol{\theta}^* \\ =&-n \frac{\lambda \boldsymbol{\theta}^*}{\lceil \lambda \|\boldsymbol{\theta}^*\|^2 \rceil} \mathbf{I}\left(\frac{\lambda \|\boldsymbol{\theta}^*\|^2}{\lceil \lambda \|\boldsymbol{\theta}^*\|^2 \rceil} \right) + \lambda \boldsymbol{\theta}^* \\ =&-\lambda \boldsymbol{\theta}^* \mathbf{I}\left(\frac{\lambda \|\boldsymbol{\theta}^*\|^2}{\lceil \lambda \|\boldsymbol{\theta}^*\|^2 \rceil} \right) + \lambda \boldsymbol{\theta}^* \\ \ni&\mathbf{0} \end{aligned}$$

where the last line is due to $\mathbf{I}\left(\frac{\lambda \|\boldsymbol{\theta}^*\|^2}{\lceil \lambda \|\boldsymbol{\theta}^*\|^2 \rceil} \right)$ giving either the set $[0,1]$ or the value 1. □

Proof to Corollary 2

Proof. We show this number matches LB2. Let $A = I$, $\ell(a,b) = \max(1-ab, 0)$, and consider the denominator of (7):

$$\begin{aligned} \sup_{\alpha \in \mathbb{R}, y \in \mathcal{Y}, g \in -\partial_1 \ell(\alpha \|\boldsymbol{\theta}^*\|^2, y)} \alpha g &= \sup_{\alpha, y \in \{-1,1\}, g \in y\mathbf{I}(y\alpha \|\boldsymbol{\theta}^*\|^2)} \alpha g \\ &= \sup_{\alpha, g \in \mathbf{I}(\alpha \|\boldsymbol{\theta}^*\|^2)} \alpha g \\ &= \frac{1}{\|\boldsymbol{\theta}^*\|^2} \end{aligned}$$

where the first equality is due to $\partial_1 \ell(a,b) = -b\mathbf{I}(ab)$. Therefore, $LB2 = \lceil \lambda \|\boldsymbol{\theta}^*\|^2 \rceil$ which matches the construction in (13). □

Proof to Proposition 3

Proof. We first verify that $\boldsymbol{\theta}^*$ is a solution to (14) based on the teaching set construction in (16). We only need to verify

the gradient of (14) is zero. Computing the gradient of (14), we have

$$\begin{aligned}
&\sum_{i=1}^{n} \frac{-y_i \mathbf{x}_i}{1+\exp\{y_i \mathbf{x}_i^\top \boldsymbol{\theta}^*\}} + \lambda \boldsymbol{\theta}^* \\
=& -n \frac{\mathbf{x}_i}{1+\exp\left\{\tau^{-1}\left(\lambda\|\boldsymbol{\theta}^*\|^2 \left\lceil \frac{\lambda\|\boldsymbol{\theta}^*\|^2}{\tau_{\max}} \right\rceil^{-1}\right)\right\}} + \lambda \boldsymbol{\theta}^* \\
=& -n \frac{\tau^{-1}\left(\lambda\|\boldsymbol{\theta}^*\|^2 \left\lceil \frac{\lambda\|\boldsymbol{\theta}^*\|^2}{\tau_{\max}} \right\rceil^{-1}\right)}{1+\exp\left\{\tau^{-1}\left(\lambda\|\boldsymbol{\theta}^*\|^2 \left\lceil \frac{\lambda\|\boldsymbol{\theta}^*\|^2}{\tau_{\max}} \right\rceil^{-1}\right)\right\}} \frac{\boldsymbol{\theta}^*}{\|\boldsymbol{\theta}^*\|^2} + \lambda \boldsymbol{\theta}^* \\
=& -n\lambda\|\boldsymbol{\theta}^*\|^2 \left\lceil \frac{\lambda\|\boldsymbol{\theta}^*\|^2}{\tau_{\max}} \right\rceil^{-1} \frac{\boldsymbol{\theta}^*}{\|\boldsymbol{\theta}^*\|^2} + \lambda \boldsymbol{\theta}^* \\
=& \mathbf{0},
\end{aligned}$$

where the third equality uses the fact $\lambda\|\boldsymbol{\theta}^*\|^2 \left\lceil \frac{\lambda\|\boldsymbol{\theta}^*\|^2}{\tau_{\max}} \right\rceil^{-1} \leq \tau_{\max}$ and the property $a = \frac{\tau^{-1}(a)}{1+e^{\tau^{-1}(a)}}$. The strong convexity of (14) automatically implies uniqueness. □

Proof to Corollary 3

Proof. We show that the number matches LB2. In (7) let $A = I$ and $\ell(a, b) = \log(1 + \exp\{-ab\})$. The denominator of LB2 is:

$$\begin{aligned}
\sup_{\alpha\in\mathbb{R}, y\in\mathcal{Y}, g\in-\partial_1\ell(\alpha\|\boldsymbol{\theta}^*\|^2, y)} \alpha g =& \sup_{\alpha, y\in\{-1,1\}, g=y(1+\exp\{y\alpha\|\boldsymbol{\theta}^*\|^2\})^{-1}} \alpha g \\
=& \sup_{\alpha, g=(1+\exp\{\alpha\|\boldsymbol{\theta}^*\|^2\})^{-1}} \alpha g \\
=& \sup_{\alpha} \frac{\alpha}{1+\exp\{\alpha\|\boldsymbol{\theta}^*\|^2\}} \\
=& \|\boldsymbol{\theta}^*\|^{-2} \sup_{t} \frac{t}{1+\exp\{t\}} \\
=& \frac{\tau_{\max}}{\|\boldsymbol{\theta}^*\|^2},
\end{aligned}$$

which implies $LB2 = \left\lceil \frac{\lambda\|\boldsymbol{\theta}^*\|^2}{\tau_{\max}} \right\rceil$. □

Proof to Proposition 4

Proof. We first prove the case for $\mathbf{w}^* = \mathbf{0}$. We can verify that the KKT condition is satisfied by designing $\mathbf{x}_1$ and y_1 as in (18):

$$\begin{aligned}
(\mathbf{x}_1^\top \mathbf{w}^* + b^* - y_1)\mathbf{x}_1 + \lambda \mathbf{w}^* =& \mathbf{0} \\
\mathbf{x}_1^\top \mathbf{w}^* + b^* - y_1 =& 0.
\end{aligned}$$

The uniqueness of $[\mathbf{w}^*; b^*]$ is indicated by the strong convexity of (17) when $n = 1$.

We then prove the case for $\mathbf{w}^* \neq \mathbf{0}$. With simple algebra, we can verify the KKT condition holds via the construction in (19):

$$\begin{aligned}
(\mathbf{x}_1^\top \mathbf{w}^* + b^* - y_1)\mathbf{x}_1 + (\mathbf{x}_2^\top \mathbf{w}^* + b^* - y_2)\mathbf{x}_2 + \lambda \mathbf{w}^* =& \mathbf{0} \\
(\mathbf{x}_1^\top \mathbf{w}^* + b^* - y_1) + (\mathbf{x}_2^\top \mathbf{w}^* + b^* - y_2) =& 0.
\end{aligned}$$

Similarly, the uniqueness is implied by the strong convexity of (17) when $n = 2$. □

Proof to Corollary 4

Proof. We match the lower bound LB1 in (6). Note $\boldsymbol{\theta}^* = [\mathbf{w}^*; b^*] \in \mathbb{R}^{d+1}$, and A in this case is a $(d+1) \times (d+1)$ matrix with the $d \times d$ identity matrix I_d padded with one additional row and column of zeros for the offset. Therefore $Rank(A) = Rank(I_d) = d$. When $\mathbf{w}^* = \mathbf{0}$, $A\boldsymbol{\theta}^* = \mathbf{0}$ and $LB1 = (d+1) - Rank(A) = 1$. When $\mathbf{w}^* \neq \mathbf{0}$, $A\boldsymbol{\theta}^* \neq \mathbf{0}$ and $LB1 = (d+1) - Rank(A) + 1 = 2$. These lower bounds match the teaching set sizes in (18) and (19), respectively. □

Proof to Proposition 5

Proof. Unlike in previous learners (including homogeneous SVM), we no longer have strong convexity w.r.t. b. In order to prove that (21) is a teaching set, we need to verify the KKT condition and verify solution uniqueness.

We first verify the KKT condition to show that the solution under (21) includes the target model $[\mathbf{w}^*; b^*]$. From (21), we have

$$\mathbf{x}_+^\top \mathbf{w}^* + b^* = 1, \quad \mathbf{x}_-^\top \mathbf{w}^* + b^* = -1. \tag{43}$$

Applying them to the KKT condition and using the notation in (42) we obtain

$$\begin{aligned}
&-\frac{n}{2}\mathbf{I}(\mathbf{x}_+^\top \mathbf{w}^* + b^*)\begin{bmatrix}\mathbf{x}_+\\ 1\end{bmatrix} + \frac{n}{2}\mathbf{I}(-\mathbf{x}_-^\top \mathbf{w}^* - b^*)\begin{bmatrix}\mathbf{x}_-\\ 1\end{bmatrix} + \begin{bmatrix}\lambda \mathbf{w}^*\\ 0\end{bmatrix}\\
=&-\frac{n}{2}\mathbf{I}(1)\begin{bmatrix}\mathbf{x}_+\\ 1\end{bmatrix} + \frac{n}{2}\mathbf{I}(1)\begin{bmatrix}\mathbf{x}_-\\ 1\end{bmatrix} + \begin{bmatrix}\lambda \mathbf{w}^*\\ 0\end{bmatrix}\\
\supset&\frac{n}{2}\mathbf{I}(1)\begin{bmatrix}\mathbf{x}_- - \mathbf{x}_+\\ 0\end{bmatrix} + \begin{bmatrix}\lambda \mathbf{w}^*\\ 0\end{bmatrix} \quad \text{setting the last dimension to 0}\\
=&\mathbf{I}(1)\begin{bmatrix}-\frac{n}{\|\mathbf{w}^*\|^2}\mathbf{w}^*\\ 0\end{bmatrix} + \begin{bmatrix}\lambda \mathbf{w}^*\\ 0\end{bmatrix} \quad \text{applying (21)}\\
\supseteq&\mathbf{I}(1)\begin{bmatrix}-\lambda \mathbf{w}^*\\ 0\end{bmatrix} + \begin{bmatrix}\lambda \mathbf{w}^*\\ 0\end{bmatrix} \quad \text{observing } n \geq \lambda\|\mathbf{w}^*\|^2\\
\ni&\mathbf{0}.
\end{aligned}$$

It proves that $[\mathbf{w}^*; b^*]$ solves (20) by our teaching set construction.

Next we prove uniqueness by contradiction. We use $f(\mathbf{w}, b)$ to denote the objective function in (20) under the teaching set. It is easy to verify that $f(\mathbf{w}^*, b^*) = \frac{\lambda}{2}\|\mathbf{w}^*\|^2$. Assume that there exists another solution $[\bar{\mathbf{w}}; \bar{b}]$ different from $[\mathbf{w}^*; b^*]$. We can obtain $\|\bar{\mathbf{w}}\|^2 \leq \|\mathbf{w}^*\|^2$ due to

$$\frac{\lambda}{2}\|\mathbf{w}^*\|^2 = f(\mathbf{w}^*, b^*) = f(\bar{\mathbf{w}}, \bar{b}) \geq \frac{\lambda}{2}\|\bar{\mathbf{w}}\|^2.$$

The second equality is due to $[\bar{\mathbf{w}}; \bar{b}]$ being a solution; the inequality is due to whole-part relationship. Therefore, there are only two possibilities for the norm of $\bar{\mathbf{w}}$: $\|\bar{\mathbf{w}}\| = \|\mathbf{w}^*\|$ or $\|\bar{\mathbf{w}}\| = t\|\mathbf{w}^*\|$ for some $0 \leq t < 1$. Next we will show that both cases are impossible.

(Case 1) For the case $\|\bar{\mathbf{w}}\| = \|\mathbf{w}^*\|$, we have

$$\begin{aligned}
f(\bar{\mathbf{w}}, \bar{b}) =&\frac{n}{2}\max\left(1 - (\mathbf{x}_+^\top \bar{\mathbf{w}} + \bar{b}), 0\right) + \frac{n}{2}\max\left(1 + (\mathbf{x}_-^\top \bar{\mathbf{w}} + \bar{b}), 0\right) + \frac{\lambda}{2}\|\bar{\mathbf{w}}\|^2\\
=&\frac{n}{2}\max\left(\underbrace{\mathbf{x}_+^\top(\mathbf{w}^* - \bar{\mathbf{w}}) + (b^* - \bar{b})}_{=:\Delta_+}, 0\right) + \frac{n}{2}\max\left(\underbrace{-\mathbf{x}_-^\top(\mathbf{w}^* - \bar{\mathbf{w}}) - (b^* - \bar{b})}_{=:\Delta_-}, 0\right)\\
&+ \frac{\lambda}{2}\|\mathbf{w}^*\|^2\\
=&\frac{n}{2}\max(\Delta_+, 0) + \frac{n}{2}\max(\Delta_-, 0) + f(\mathbf{w}^*, b^*).
\end{aligned}$$

From $f(\bar{\mathbf{w}}, \bar{b}) = f(\mathbf{w}^*, b^*)$, it follows $\Delta_+ \leq 0$ and $\Delta_- \leq 0$. Since

$$0 \geq \Delta_+ + \Delta_- = (\mathbf{x}_+ - \mathbf{x}_-)^\top (\mathbf{w}^* - \bar{\mathbf{w}}) = \frac{2(\mathbf{w}^*)^\top (\mathbf{w}^* - \bar{\mathbf{w}})}{\|\mathbf{w}^*\|^2} = 2 - 2\frac{\bar{\mathbf{w}}^\top \mathbf{w}^*}{\|\mathbf{w}^*\|^2},$$

we have $\bar{\mathbf{w}}^\top \mathbf{w}^* \geq \|\mathbf{w}^*\|^2$. But because $\|\bar{\mathbf{w}}\| = \|\mathbf{w}^*\|$, we must have $\bar{\mathbf{w}} = \mathbf{w}^*$. Applying this new observation to $\Delta_+ \leq 0$ and $\Delta_- \leq 0$, we obtain $b^* = \bar{b}$. It means that $[\mathbf{w}^*; b^*] = [\bar{\mathbf{w}}; \bar{b}]$, contradicting our assumption $[\mathbf{w}^*; b^*] \neq [\bar{\mathbf{w}}; \bar{b}]$.

(Case 2) Next we turn to the case $\|\bar{\mathbf{w}}\| = t\|\mathbf{w}^*\|$ for some $t \in [0, 1)$. Recall our assumption that $[\bar{\mathbf{w}}; \bar{b}]$ solves (20). Then it follows that the following specific construction $[\hat{\mathbf{w}}, \hat{b}]$ solves (20) as well:

$$\hat{\mathbf{w}} = t\mathbf{w}^*, \quad \hat{b} = tb^*. \tag{44}$$

To see this, we consider the following optimization problem:

$$\begin{aligned} \min_{\mathbf{w}, b} \quad & L(\mathbf{w}, b) := \frac{n}{2} \max(1 - (\mathbf{x}_+^\top \mathbf{w} + b), 0) + \frac{n}{2} \max(1 + (\mathbf{x}_-^\top \mathbf{w} + b), 0) \\ \text{s.t.} \quad & \|\mathbf{w}\| \leq t\|\mathbf{w}^*\|. \end{aligned} \tag{45}$$

Since $[\bar{\mathbf{w}}; \bar{b}]$ solves (20), it is easy to see that $[\bar{\mathbf{w}}; \bar{b}]$ solves (45) too, otherwise there exists a solution for (45) which gives a lower function value on (20). Then we can verify that $[\hat{\mathbf{w}}; \hat{b}]$ solves (45) as well by showing the following optimality condition holds:

$$-\left[\begin{array}{c} \frac{\partial L(\mathbf{w},b)}{\partial \mathbf{w}} \\ \frac{\partial L(\mathbf{w},b)}{\partial b} \end{array}\right]\Bigg|_{[\hat{\mathbf{w}};\hat{b}]} \cap \underbrace{\mathcal{N}_{\|\mathbf{w}\| \leq t\|\mathbf{w}^*\|}(\hat{\mathbf{w}}, \hat{b})}_{\text{Normal cone to the set } \{[\mathbf{w}; b] : \|\mathbf{w}\| \leq t\|\mathbf{w}^*\|\} \text{ at } [\hat{\mathbf{w}}; \hat{b}]} \neq \emptyset \tag{46}$$

Given a convex closed set Ω and a point $\boldsymbol{\theta} \in \Omega$, the normal cone at point $\boldsymbol{\theta}$ is defined to be a set

$$\mathcal{N}_\Omega(\boldsymbol{\theta}) = \{\boldsymbol{\phi} : \ \langle \boldsymbol{\phi}, \boldsymbol{\psi} - \boldsymbol{\theta} \rangle \leq 0 \ \forall \boldsymbol{\psi} \in \Omega\}.$$

The optimality condition basically suggests that at the optimal point, the negative (sub)gradient direction overlaps with the normal cone. In other words, there does not exist any valid direction to decrease the objective at the optimal point. Readers can refer to Nocedal & Wright (2006) or Bertsekas & Nedic (2003) for more explanations about the geometric optimality condition.

Because of (43) and (44), we have $\mathbf{x}_+^\top \hat{\mathbf{w}} + \hat{b} = t < 1$. Thus at $[\hat{\mathbf{w}}; \hat{b}]$ the subgradient is

$$-\left[\begin{array}{c} \frac{\partial L(\mathbf{w},b)}{\partial \mathbf{w}} \\ \frac{\partial L(\mathbf{w},b)}{\partial b} \end{array}\right]\Bigg|_{[\hat{\mathbf{w}};\hat{b}]} = \frac{n}{2}\begin{bmatrix} \mathbf{x}_+ - \mathbf{x}_- \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{n\mathbf{w}^*}{\|\mathbf{w}^*\|^2} \\ 0 \end{bmatrix} \tag{47}$$

And the normal cone is

$$\mathcal{N}_{\|\mathbf{w}\| \leq t\|\mathbf{w}^*\|}(\hat{\mathbf{w}}, \hat{b}) = \left\{ s \begin{bmatrix} \mathbf{w}^* \\ 0 \end{bmatrix} \ \middle| \ s \geq 0 \right\}. \tag{48}$$

The intersection is non-empty by choosing $s = \frac{n}{\|\mathbf{w}^*\|^2}$. Since both $[\hat{\mathbf{w}}; \hat{b}]$ and $[\bar{\mathbf{w}}; \bar{b}]$ solve (45), we have $L(\hat{\mathbf{w}}, \hat{b}) = L(\bar{\mathbf{w}}, \bar{b})$. Together with $\|\hat{\mathbf{w}}\| = \|\bar{\mathbf{w}}\|$, we have

$$f(\hat{\mathbf{w}}, \hat{b}) = L(\hat{\mathbf{w}}, \hat{b}) + \frac{\lambda}{2}\|\hat{\mathbf{w}}\|^2 = f(\bar{\mathbf{w}}, \bar{b}) = f(\mathbf{w}^*, b^*).$$

Therefore, we proved that $[\hat{\mathbf{w}}; \hat{b}]$ solves (20) as well. To see the contradiction, let us check the function value of $f(\hat{\mathbf{w}}, \hat{b})$

via a different route:

$$\begin{aligned}
f(\hat{\mathbf{w}}, \hat{b}) =& f(t\mathbf{w}^*, tb^*) \\
=& \sum_{i=1}^{\frac{n}{2}} \max\left(1 - t(\mathbf{x}_+^\top \mathbf{w}^* + b^*), 0\right) + \sum_{i=1}^{\frac{n}{2}} \max\left(1 + t(\mathbf{x}_-^\top \mathbf{w}^* + b^*), 0\right) + \frac{\lambda}{2}\|\mathbf{w}^*\|^2 t^2 \\
=& \sum_{i=1}^{\frac{n}{2}} \max\left(1 - t, 0\right) + \sum_{i=1}^{\frac{n}{2}} \max\left(1 - t, 0\right) + \frac{\lambda}{2}\|\mathbf{w}^*\|^2 t^2 \\
=& n(1-t) - \frac{\lambda}{2}\|\mathbf{w}^*\|^2(1-t^2) + \frac{\lambda}{2}\|\mathbf{w}^*\|^2 \\
\geq& n(1-t) - \frac{n}{2}(1-t^2) + \frac{\lambda}{2}\|\mathbf{w}^*\|^2 \\
=& \frac{n}{2}(1-t)^2 + f(\mathbf{w}^*, b^*) \\
>& f(\mathbf{w}^*, b^*),
\end{aligned}$$

where the first inequality uses the fact that $n \geq \lambda\|\mathbf{w}^*\|^2$. It contradicts our early assertion $f(\hat{\mathbf{w}}, \hat{b}) = f(\mathbf{w}^*, b^*)$. Putting cases 1 and 2 together we prove uniqueness. □

Proof to Corollary 5

Proof. The upper bound directly follows Proposition 5. We only need to show the lower bound $LB3 = \lceil \lambda\|\mathbf{w}^*\|^2 \rceil$ in Theorem 3. Let $A = I$, $\ell(a) = \max(1-a, 0)$, and consider the denominator of (9):

$$\sup_{\alpha \in \mathbb{R}, g \in -\partial\ell(\alpha\|\mathbf{w}^*\|^2)} \alpha g = \sup_{\alpha, g \in \mathbf{I}(\alpha\|\mathbf{w}^*\|^2)} \alpha g = \frac{1}{\|\mathbf{w}^*\|^2}$$

where the first equality is due to $\partial\ell(a) = -\mathbf{I}(a)$. Therefore, $LB3 = \lceil \lambda\|\mathbf{w}^*\|^2 \rceil$ which proves the lower bound. □

Proof to Proposition 6

Proof. We first point out that for t to be well-defined the argument to $\tau^{-1}()$ has to be bounded $\frac{\lambda\|\mathbf{w}^*\|^2}{n} \leq \tau_{\max}$. This implies $n \geq \frac{\lambda\|\mathbf{w}^*\|^2}{\tau_{\max}}$. The size of our proposed teaching set is the smallest among all such symmetric construction that satisfy this constraint.

We verify that the KKT condition to show the construction in (23) includes the solution $[\mathbf{w}^*; b^*]$. From (23), we have

$$\mathbf{x}_+^\top \mathbf{w}^* + b^* = t \quad \mathbf{x}_-^\top \mathbf{w}^* + b^* = -t.$$

We apply them and the teaching set construction to compute the gradient of (22):

$$\begin{aligned}
& -\frac{n}{2}\frac{1}{1+\exp\{\mathbf{x}_+^\top \mathbf{w}^* + b^*\}}\begin{bmatrix}\mathbf{x}_+ \\ 1\end{bmatrix} + \frac{n}{2}\frac{1}{1+\exp\{-\mathbf{x}_-^\top \mathbf{w}^* - b^*\}}\begin{bmatrix}\mathbf{x}_- \\ 1\end{bmatrix} + \begin{bmatrix}\lambda\mathbf{w}^* \\ 0\end{bmatrix} \\
=& -\frac{n}{2}\frac{1}{1+\exp\{t\}}\begin{bmatrix}\mathbf{x}_+ \\ 1\end{bmatrix} + \frac{n}{2}\frac{1}{1+\exp\{t\}}\begin{bmatrix}\mathbf{x}_- \\ 1\end{bmatrix} + \begin{bmatrix}\lambda\mathbf{w}^* \\ 0\end{bmatrix} \\
=& -\frac{n}{\|\mathbf{w}^*\|^2}\frac{t}{1+\exp\{t\}}\begin{bmatrix}\mathbf{w}^* \\ 0\end{bmatrix} + \begin{bmatrix}\lambda\mathbf{w}^* \\ 0\end{bmatrix} \\
=& -\frac{n}{\|\mathbf{w}^*\|^2}\frac{\lambda\|\mathbf{w}^*\|^2}{n}\begin{bmatrix}\mathbf{w}^* \\ 0\end{bmatrix} + \begin{bmatrix}\lambda\mathbf{w}^* \\ 0\end{bmatrix} \\
=& \mathbf{0}.
\end{aligned}$$

This verifies the KKT condition.

Finally we show uniqueness. The Hessian matrix of the objective function (22) under our training set (23) is:

$$\underbrace{\frac{n}{2}\frac{\exp\{t\}}{(1+\exp\{t\})^2}}_{:=a}\underbrace{\begin{bmatrix}\mathbf{x}_+\mathbf{x}_+^\top+\mathbf{x}_-\mathbf{x}_-^\top & \mathbf{x}_++\mathbf{x}_- \\ \mathbf{x}_+^\top+\mathbf{x}_-^\top & 2\end{bmatrix}}_{:=A}+\lambda\underbrace{\begin{bmatrix}I & \mathbf{0}\\ \mathbf{0}^\top & 0\end{bmatrix}}_{:=B}.$$

Note $a>0$ and $A=\begin{bmatrix}\mathbf{x}_+\\1\end{bmatrix}\begin{bmatrix}\mathbf{x}_+ & 1\end{bmatrix}+\begin{bmatrix}\mathbf{x}_-\\1\end{bmatrix}\begin{bmatrix}\mathbf{x}_- & 1\end{bmatrix}$ is positive semi-definite. We show that $aA+\lambda B$ is positive definite. Suppose not. Then there exists $[\mathbf{u};v]\neq\mathbf{0}$ such that $[\mathbf{u};v]^\top(aA+\lambda B)[\mathbf{u};v]=0$. This implies $[\mathbf{u};v]^\top(aA)[\mathbf{u};v]+\lambda\mathbf{u}^\top\mathbf{u}=0$. Since the first term is non-negative due to A being positive semi-definite, $\mathbf{u}=\mathbf{0}$. But then we have $2av^2=0$ which implies $[\mathbf{u};v]=\mathbf{0}$, a contradiction. Therefore uniqueness is guaranteed.

□

Proof to Corollary 6

Proof. The upper bound directly follows Proposition 6. We only need to show the lower bound $\left\lceil\frac{\lambda\|\mathbf{w}^*\|^2}{\tau_{\max}}\right\rceil$ by applying $LB3$ in Theorem 3. Let $A=I$ and $\ell(a)=\log(1+\exp\{-a\})$ and consider the denominator of (9):

$$\begin{aligned}\sup_{\alpha\in\mathbb{R},g\in\partial\ell(-\alpha\|\mathbf{w}^*\|^2)}\alpha g &= \sup_{\alpha,g=(1+\exp\{\alpha\|\mathbf{w}^*\|^2\})^{-1}}\alpha g\\ &=\sup_\alpha\frac{\alpha}{1+\exp\{\alpha\|\mathbf{w}^*\|^2\}}\\ &=\|\mathbf{w}^*\|^{-2}\sup_t\frac{t}{1+\exp\{t\}}\\ &=\frac{\tau_{\max}}{\|\mathbf{w}^*\|^2},\end{aligned}$$

which implies $LB3=\left\lceil\frac{\lambda\|\mathbf{w}^*\|^2}{\tau_{\max}}\right\rceil$. □