Scaling SVM and Least Absolute Deviations via Exact Data Reduction Supplement Material

In this supplement, we present all of the details we mentioned in the main text.

A. Proof of Lemma 1

Before we prove Lemma 1, let us cite the following technical lemma.

Lemma 15. *(Hiriart-Urruty & Lemaréchal, 1993) The function f is equal to its biconjugate f^{**} if and only if $f \in \Gamma_0(\Re^n)$.*

We are now ready to derive a simple proof of Lemma 1 based on Lemma 15.

Proof. In order to show $\varphi^{**} = \varphi$, it is enough to show $\varphi \in \Gamma_0(\Re)$ according to Lemma 15. Therefore we only to check the following three conditions:

1). Properness: because $\varphi : \Re \to \Re_+$, i.e., there exists $t \in \Re$ such that $\varphi(t)$ is finite, φ is proper.

2). Lower semi-continuality: φ is lower semicontinuous because it is continuous.

3). Convexity: the convexity of φ is due to the its sublinearity, see Definition 1.

Thus, we have $\varphi \in \Gamma_0(\Re)$, which completes the proof. □

B. Proof of Lemma 2

To prove Lemma 2, we need to following results.

Lemma 16. *(Ruszczyński, 2006) Let $Z \subseteq \Re^n$ be a convex and closed set. Let us define the support function of Z as*

$$\sigma_Z(s) := \sup_{\mathbf{x} \in Z} \mathbf{s}^T \mathbf{x}, \tag{29}$$

and the indicator function ι_Z as

$$\iota_Z(\mathbf{x}) = \begin{cases} 0, & \text{if } \mathbf{x} \in Z, \\ \infty, & \text{otherwise.} \end{cases} \tag{30}$$

Then

$$\sigma_Z^* = \iota_Z, \quad \text{amd } \iota_Z^* = \sigma_Z. \tag{31}$$

Theorem 17. *(Hiriart-Urruty & Lemaréchal, 1993) Let $\sigma \in \Gamma_0(\Re^n)$ be a sublinear function, then σ is the support function of the nonempty closed convex set*

$$S_\sigma := \{\mathbf{s} \in \Re^n : \mathbf{s}^T \mathbf{d} \leq \sigma(\mathbf{d}),\ \forall \mathbf{d} \in \Re^n\}. \tag{32}$$

We are now ready to prove Lemma 2.

Proof. Due to Lemma 16 and Theorem 17, we can see that, there is a nonempty closed convex set $Z \subseteq \Re$ such that

$$\varphi(t) = \sup_{s \in Z} st,\ \forall t \in \Re, \tag{33}$$

where

$$Z := \{s : st \leq \varphi(t),\ \forall t \in \Re\}. \tag{34}$$

Let $t = 1$ and -1 respectively, Eq. (34) implies that

$$\sup_{s\in Z} s \le \varphi(1) \quad \text{and} \quad \inf_{s\in Z} s \ge \varphi(-1). \tag{35}$$

Therefore, Z is a closed and bounded interval, i.e., $Z = [\alpha, \beta]$ with $\alpha, \beta \in \Re$.

Next, let us show that $\alpha \neq \beta$. In fact, in view of the nonnegativity of φ and Eq. (34), it is easy to see that $0 \in Z$. Therefore, if $\alpha = \beta$, we must have $Z = \{0\}$. Thus, Lemma 16 implies that

$$\varphi = \iota_Z^* \equiv 0, \tag{36}$$

which contradicts the fact that φ is a nonconstant function. Hence, we can conclude that $\alpha < \beta$, which completes the proof. □

C. Derivation of the KKT Condition in Eq. (14)

The problem in (12) can be written as follows:

$$\begin{aligned} &\min_{\theta} \frac{C}{2}\|\mathbf{Z}^T\theta\|^2 - \langle \bar{\mathbf{y}}, \theta\rangle, \\ &\text{s.t. } \theta_i \in [\alpha, \beta],\ i = 1, \ldots, l. \end{aligned} \tag{37}$$

Therefore, we can see that the Lagrangian is

$$\begin{aligned} L(\theta, \mu, \nu) = &\frac{C}{2}\|\mathbf{Z}^T\theta\|^2 - \langle \bar{\mathbf{y}}, \theta\rangle \\ &+ \sum_{i=1}^{l} \mu_i(\alpha - \theta_i) + \sum_{i=1}^{l} \nu_i(\theta_i - \beta), \end{aligned} \tag{38}$$

where $\mu = (\mu_1, \ldots, \mu_l)^T$, $\nu = (\nu_1, \ldots, \nu_l)^T$, and $\mu_i \ge 0$, $\nu_i \ge 0$ for all $i = 1, \ldots, l$. μ and ν are in fact the vector of Lagrangian multipliers.

For simplicity, let us denote $\theta^*(C)$ by θ^*. Then the KKT conditions (Boyd & Vandenberghe, 2004) are

$$\frac{\partial L(\theta, \mu, \nu)}{\partial \theta}|_{\theta^*} = 0 \Rightarrow C\mathbf{Z}\mathbf{Z}^T\theta^* - \bar{\mathbf{y}} - \mu + \nu = 0, \tag{39}$$

$$\begin{aligned} \mu_i(\alpha - \theta_i^*) &= 0, \\ \nu_i(\theta_i^* - \beta) &= 0, \end{aligned} \quad i = 1, \ldots, l. \tag{40}$$

Eq. (40) is known as the complementary slackness condition. The equation in (39) actually involves l equations. We can write down the i^{th} equation as follows:

$$C\langle \mathbf{Z}^T\theta^*, a_i\mathbf{x}_i\rangle - \mu_i + \nu_i = b_i y_i. \tag{41}$$

Recall that the i^{th} column of $\mathbf{Z}$ is $a_i\mathbf{x}_i$. In view of Eq. (40) and Eq. (41), we can see that:

1. if $\theta_i^* = \alpha$, then $\nu_i = 0$ and Eq. (41) results in

$$C\langle \mathbf{Z}^T\theta^*, a_i\mathbf{x}_i\rangle \ge b_i y_i; \tag{42}$$

2. if $\theta_i^* \in (\alpha, \beta)$, then $\mu_i = \nu_i = 0$ and Eq. (41) results in

$$C\langle \mathbf{Z}^T\theta^*, a_i\mathbf{x}_i\rangle = b_i y_i; \tag{43}$$

3. if $\theta_i^* = \beta$, then $\mu_i = 0$ and Eq. (41) results in

$$C\langle \mathbf{Z}^T\theta^*, a_i\mathbf{x}_i\rangle \le b_i y_i. \tag{44}$$

Then, in view of the inequalities in (42), (43) and (44), and Eq. (13), it is straightforward to derive the KKT condition in (14).

D. Proof of Lemma 3

Proof. The first part of the statement is trivial by the definition of $\hat{\mathcal{R}}$ and $\hat{\mathcal{L}}$. Therefore, we only consider the second part of the statement.

Let $\mathbf{G} = \mathbf{Z}\mathbf{Z}^T$. By permuting the columns and rows of $\mathbf{G}$, we have

$$\hat{\mathbf{G}} = \begin{pmatrix} \hat{\mathbf{G}}_{11} & \hat{\mathbf{G}}_{12} \\ \hat{\mathbf{G}}_{21} & \hat{\mathbf{G}}_{22} \end{pmatrix} = \begin{pmatrix} [\mathbf{X}^T]_{\hat{\mathcal{S}}^c}^T [\mathbf{X}^T]_{\hat{\mathcal{S}}^c} & [\mathbf{X}^T]_{\hat{\mathcal{S}}^c}^T [\mathbf{X}^T]_{\hat{\mathcal{S}}} \\ [\mathbf{X}^T]_{\hat{\mathcal{S}}}^T [\mathbf{X}^T]_{\hat{\mathcal{S}}^c} & [\mathbf{X}^T]_{\hat{\mathcal{S}}}^T [\mathbf{X}^T]_{\hat{\mathcal{S}}} \end{pmatrix}.$$

As a result, the objective function of problem (12) can be rewritten as

$$\frac{C}{2}[\theta]_{\hat{\mathcal{S}}^c}^T \hat{\mathbf{G}}_{11}[\theta]_{\hat{\mathcal{S}}^c} - \hat{\mathbf{y}}^T[\theta]_{\hat{\mathcal{S}}^c} + R([\theta]_{\hat{\mathcal{S}}}) \tag{45}$$

where

$$\hat{\mathbf{y}} = \mathbf{y}_{\hat{\mathcal{S}}^c} - C\hat{\mathbf{G}}_{12}[\theta]_{\hat{\mathcal{S}}}, \ , \tag{46}$$

$$R([\theta]_{\hat{\mathcal{S}}}) = \frac{C}{2}[\theta]_{\hat{\mathcal{S}}}^T \hat{\mathbf{G}}_{22}[\theta]_{\hat{\mathcal{S}}} - \mathbf{y}_{\hat{\mathcal{S}}}^T[\theta]_{\hat{\mathcal{S}}} \tag{47}$$

Due to the assumption that $[\theta^*(C)]_{\hat{\mathcal{S}}}$ is known, $\hat{\mathbf{y}}$ and $R([\theta]_{\hat{\mathcal{S}}})$ can be treated as constants, and thus problem (12) reduces to problem (15). □

E. Review of the Dual Coordinate Descent Method

Problem (15) can be efficiently solved by the dual coordinate descent method (Hsieh et al., 2008).

More precisely, the optimization procedure starts from an initial point $\hat{\theta}^0 \in \Re^{|\hat{\mathcal{S}}^c|}$ and generates a sequence of points $\{\hat{\theta}^k\}_{k=0}^{\infty}$. The process from $\hat{\theta}^k$ to $\hat{\theta}^{k+1}$ is referred to as an outer iteration. In each outer iteration, we update the components of $\hat{\theta}^k$ one at a time and thus get a sequence of points $\hat{\theta}^{k,i} \in \Re^{|\hat{\mathcal{S}}^c|}$, $i = 1, \ldots, |\hat{\mathcal{S}}^c|$. Suppose we are at the k^{th} outer iteration. To get $\hat{\theta}^{k,i}$ from $\hat{\theta}^{k,i-1}$, we need to solve the following optimization problem:

$$\begin{aligned} \min_t \ & \frac{C}{2}(\hat{\theta}^{k,i-1} + t\mathbf{e}_i)^T \hat{\mathbf{G}}_{11}(\hat{\theta}^{k,i-1} + t\mathbf{e}_i) \\ & \qquad - \hat{\mathbf{y}}^T(\hat{\theta}^{k,i-1} + t\mathbf{e}_i) \\ \text{s.t.} \ & [\hat{\theta}^{k,i-1}]_i + t \in [\alpha, \beta], \ i = 1, \ldots, l, \end{aligned} \tag{48}$$

where $\mathbf{e}_i = (0, \ldots, 1, \ldots, 0)^T$. Clearly, problem (48) is equivalent to the following 1D optimization problem:

$$\begin{aligned} \min_t \ & \frac{C}{2}[\hat{\mathbf{G}}_{11}]_{i,i}t^2 + (C\mathbf{e}_i^T\hat{\mathbf{G}}_{11}\hat{\theta}^{k,i-1} - [\hat{\mathbf{y}}]_i)t \\ \text{s.t.} \ & [\hat{\theta}^{k,i-1}]_i + t \in [\alpha, \beta], \end{aligned} \tag{49}$$

which admits a closed form solution t^*. Once t^* is available, we can set $\hat{\theta}^{k,i} = \hat{\theta}^{k,i-1} + t^*\mathbf{e}_i$. For more details, please refer to Hsieh et al. (2008).

F. Proof of Theorem 6

Proof. We will prove the first half of the statement. The second half can be proved analogously. To show $[\theta^*(C)]_i = \alpha$, i.e., $i \in \mathcal{R}$, (R1) implies that we only need to show $C\langle \mathbf{Z}^T\theta^*(C), a_i\mathbf{x}_i\rangle > b_i y_i$. Thus, we can see that

$$\begin{aligned} C\langle \mathbf{Z}^T\theta^*(C), a_i\mathbf{x}_i\rangle &= C\left\langle \mathbf{Z}^T\theta^*(C) - \frac{C_0+C}{2C}\mathbf{Z}^T\theta^*(C_0), a_i\mathbf{x}_i\right\rangle + C\left\langle \frac{C_0+C}{2C}\mathbf{Z}^T\theta^*(C_0), a_i\mathbf{x}_i\right\rangle \\ &\geq \frac{C_0+C}{2}\langle \mathbf{Z}^T\theta^*(C_0), a_i\mathbf{x}_i\rangle - C\left\|\mathbf{Z}^T\theta^*(C) - \frac{C_0+C}{2C}\mathbf{X}^T\theta^*(C_0)\right\|\|a_i\mathbf{x}_i\| \\ &\geq \frac{C_0+C}{2}\langle \mathbf{Z}^T\theta^*(C_0), a_i\mathbf{x}_i\rangle - \frac{C-C_0}{2}\|\mathbf{Z}^T\theta^*(C_0)\|\|a_i\mathbf{x}\| \\ &> b_i y_i. \end{aligned}$$

Note that, the second inequality is due to Theorem 5, and the last line is due to the statement. This completes the proof. □

G. Improving SSNSV via VI

In this section, we describe how to strictly improve SSNSV by using the same technique used in DVI rules in a detailed manner.

Estimation of $\mathbf{w}^*$ via VI

We show that $\Omega_{[s_b,s_a]}$ in Eq. (25) can be strictly improved by the variational inequalities. Consider $\mathcal{F}_{s_a}$. Because $s_a > s_b$, we can see that $\mathbf{w}^*(s_b) \in \mathcal{F}_{s_a}$. Therefore, by Theorem 4, we have

$$\langle \mathbf{w}^*(s_a), \mathbf{w}^*(s) - \mathbf{w}^*(s_a)\rangle \geq 0, \tag{50}$$

which is the first constraint in (25). Similarly, consider $\mathcal{F}_{s_b}$. Since $\hat{\mathbf{w}}(s_b) \in \mathcal{F}_{s_b}$, Theorem 4 implies that

$$\langle \mathbf{w}^*(s), \hat{\mathbf{w}}(s_b) - \mathbf{w}^*(s)\rangle \geq 0,$$

which is equivalent to

$$\|\mathbf{w}^*(s) - \tfrac{1}{2}\hat{\mathbf{w}}(s_b)\| \leq \tfrac{1}{2}\|\hat{\mathbf{w}}(s_b)\|. \tag{51}$$

Clearly, the radius determined by the inequality (51) is only a half of the radius determined by the second constraint in (25). In view of the inequalities in (50) and (51), we can see that $\mathbf{w}^*(s)$ can be bounded inside the following region:

$$\Omega'_{[s_b,s_a]} := \left\{ \mathbf{w} : \begin{array}{l} \langle \mathbf{w}^*(s_a), \mathbf{w} - \mathbf{w}^*(s_a)\rangle \geq 0, \\ \|\mathbf{w} - \frac{1}{2}\hat{\mathbf{w}}(s_b)\| \leq \frac{1}{2}\|\hat{\mathbf{w}}(s_b)\| \end{array} \right\}$$

It is easy to see that $\Omega'_{[s_b,s_a]} \subset \Omega_{[s_b,s_a]}$. As a result, the bounds in (R1′) and (R2′) with $\Omega'_{[s_b,s_a]}$ are tighter than that of $\Omega_{[s_b,s_a]}$. Thus, SSNSV (Ogawa et al., 2013) can be strictly improved by the estimation in (26). In fact, we have the following theorem:

Theorem 18. *Suppose we are given two parameters $s_a > s_b > 0$, and let $\mathbf{w}^*(s_a)$ and $\hat{\mathbf{w}}(s_b)$ be the optimal solution at $s = s_a$ and a feasible solution at $s = s_b$, respectively. Moreover, let us define*

$$\begin{aligned} \rho &= -\|\mathbf{w}^*(s_a)\|^2 + \tfrac{1}{2}\langle \mathbf{w}^*(s_a), \hat{\mathbf{w}}(s_b)\rangle \\ \mathbf{v}^\perp &= \mathbf{v} - \tfrac{\mathbf{v}^T\mathbf{w}^*(s_a)}{\|\mathbf{w}^*(s_a)\|^2}\mathbf{w}^*(s_a), \forall \mathbf{v} \in \Re^n. \end{aligned}$$

Then, for all $s \in [s_b, s_a]$,

$$\langle \mathbf{w}^*(s_a), \bar{x}_i\rangle > \tfrac{2\|\bar{\mathbf{x}}_i\|}{\|\hat{\mathbf{w}}(s_b)\|}\rho \text{ and } \ell_i > 1 \Rightarrow i \in \mathcal{R} \Leftrightarrow \alpha_i = 0, \tag{52}$$

where

$$\begin{aligned} \ell_i = &-\tfrac{\langle \mathbf{w}^*(s_a), \bar{\mathbf{x}}_i\rangle}{\|\mathbf{w}^*(s_a)\|^2}\rho + \tfrac{1}{2}\langle \hat{\mathbf{w}}(s_b), \bar{\mathbf{x}}_i\rangle \\ &- \|\bar{\mathbf{x}}_i^\perp\|\sqrt{\frac{1}{4}\|\hat{\mathbf{w}}(s_b)\|^2 - \tfrac{\rho^2}{\|\mathbf{w}^*(s_a)\|^2}}. \end{aligned} \tag{53}$$

Similarly,

$$u_i < 1 \Rightarrow i \in \mathcal{L} \Leftrightarrow \alpha_i = c, \tag{54}$$

where

$$u_i = \begin{cases} \frac{1}{2}\left(\langle \hat{\mathbf{w}}(s_b), \bar{\mathbf{x}}_i\rangle + \|\hat{\mathbf{w}}(s_b)\|\|\bar{\mathbf{x}}_i\|\right), \\ \qquad\qquad \text{if } \langle \mathbf{w}^*(s_a), \bar{x}_i\rangle \geq -\frac{2\|\bar{\mathbf{x}}_i\|}{\|\hat{\mathbf{w}}(s_b)\|}\rho \\ -\frac{\langle \mathbf{w}^*(s_a), \bar{\mathbf{x}}_i\rangle}{\|\mathbf{w}^*(s_a)\|^2}\rho + \frac{1}{2}\langle \hat{\mathbf{w}}(s_b), \bar{\mathbf{x}}_i\rangle \\ \qquad +\|\bar{\mathbf{x}}_i^{\perp}\|\sqrt{\frac{1}{4}\|\hat{\mathbf{w}}(s_b)\|^2 - \frac{\rho^2}{\|\mathbf{w}^*(s_a)\|^2}}, \\ \qquad\qquad \text{otherwise.} \end{cases} \tag{55}$$

For convenience, we call the screening rule presented in Theorem 18 as the "enhanced" SSNSV (ESSNSV).

To prove Theorem 18, we first establish the following technical lemma.

Lemma 19. *Consider the problem as follows:*

$$\min_{\mathbf{w}} f(\mathbf{w}) = \mathbf{v}^T\mathbf{w}, \text{ s.t. } \mathbf{u}^T\mathbf{w} \leq d, \|\mathbf{w} - \mathbf{o}\| \leq r, \tag{56}$$

where $r > 0$. Let $d' = d - \mathbf{u}^T\mathbf{o}$ and the optimal solution of problem (56) be f^*. Then we have

1. *If $\mathbf{v}^T\mathbf{u} + \frac{\|\mathbf{v}\|d'}{r} \geq 0$, then*

$$f^* = \mathbf{v}^T\mathbf{o} - r\|\mathbf{v}\|.$$

2. *Otherwise,*

$$f^* = \mathbf{v}^T\mathbf{o} - \|\mathbf{v}^{\perp}\|\sqrt{r^2 - \frac{(d')^2}{\|\mathbf{u}\|^2}} + \frac{\mathbf{v}^T\mathbf{u}d'}{\|\mathbf{u}\|^2},$$

where $\mathbf{v}^{\perp} = \mathbf{v} - \frac{\mathbf{v}^T\mathbf{u}}{\|\mathbf{u}\|^2}\mathbf{u}$.

Notice that, we assume problem (56) is feasible, i.e., $\frac{|\mathbf{u}^T\mathbf{o} - d|}{\|\mathbf{u}\|} \leq r$.

Proof. Let $\mathbf{z} = \mathbf{w} - \mathbf{o}$, problem (56) can be rewritten as:

$$\min_{\mathbf{z}} \mathbf{v}^T\mathbf{z} + \mathbf{v}^T\mathbf{o}, \text{ s.t. } \mathbf{u}^T\mathbf{z} \leq d - \mathbf{u}^T\mathbf{o}, \|\mathbf{z}\| \leq r. \tag{57}$$

Problem (57) reduces to

$$\min_{\mathbf{z}} \mathbf{v}^T\mathbf{z}, \text{ s.t. } \mathbf{u}^T\mathbf{z} \leq d', \|\mathbf{z}\| \leq r. \tag{58}$$

To solve problem (58), we make use of the Lagrangian multiplier method. For notational convenience, let $\mathcal{F} := \{\mathbf{z} : \mathbf{u}^T\mathbf{z} \leq d', \|\mathbf{z}\| \leq r\}$.

$$\begin{aligned} \min_{\mathbf{z}\in\mathcal{F}} \mathbf{v}^T\mathbf{z} &= \min_{\mathbf{z}} \max_{\substack{\mu\geq 0,\\ \nu\geq 0}} \mathbf{v}^T\mathbf{z} + \nu(\mathbf{u}^T\mathbf{z} - d') + \frac{\mu}{2}(\|\mathbf{z}\|^2 - r^2) \\ &= \max_{\substack{\mu\geq 0,\\ \nu\geq 0}} \min_{\mathbf{z}} \mathbf{v}^T\mathbf{z} + \nu(\mathbf{u}^T\mathbf{z} - d') + \frac{\mu}{2}(\|\mathbf{z}\|^2 - r^2) \\ &= \max_{\substack{\mu\geq 0,\\ \nu\geq 0}} -\frac{1}{2\mu}\|\mathbf{v} + \nu\mathbf{u}\|^2 - \nu d' - \frac{\mu r^2}{2}. \end{aligned} \tag{59}$$

Notice that, in Eq. (59), we make the assumption that $\mu > 0$. However, we can not simply exclude this possibility. In fact, if $\mu = 0$, we must have

$$\mathbf{v} + \nu\mathbf{u} = 0, \tag{60}$$

since otherwise the function value of

$$\mathbf{v}^T\mathbf{z} + \nu(\mathbf{u}^T\mathbf{z} - d')$$

in the second line of Eq. (59) can be made arbitrarily small. As a result, we will have

$$g(\mu, \nu) = -\infty,$$

which contradicts the strong duality of problem (56) (Boyd & Vandenberghe, 2004). [Problem (56) is clearly lower bounded since the feasible set is compact.] Therefore, in view of Eq. (60), we can conclude that $\mu = 0$ only if $\mathbf{v}$ point in the opposite direction of $\mathbf{u}$.

Let us first consider the general case, i.e., $\mathbf{v}$ does not point in the opposite direction of $\mathbf{u}$. In view of Eq. (59), let $g(\mu, \nu) = -\frac{1}{2\mu}\|\mathbf{v} + \nu\mathbf{u}\|^2 - \nu d' - \frac{\mu r^2}{2}$. It is easy to see that

$$\frac{\partial g(\mu, \nu)}{\partial \nu} = 0 \Leftrightarrow \nu = -\frac{\mathbf{v}^T\mathbf{u} + \mu d'}{\|\mathbf{u}\|^2}. \tag{61}$$

Since ν has to be nonnegative, we have

$$\nu = \max\left\{0, -\frac{\mathbf{v}^T\mathbf{u} + \mu d'}{\|\mathbf{u}\|^2}\right\}. \tag{62}$$

Case 1. If $-\frac{\mathbf{v}^T\mathbf{u}+\mu d'}{\|\mathbf{u}\|^2} \leq 0$, then $\nu = 0$ and thus

$$\frac{\partial g(\mu, \nu)}{\partial \mu} = 0 \Leftrightarrow \mu = \frac{\|\mathbf{v}\|}{r}. \tag{63}$$

Then $g(\mu, \nu) = -r\|\mathbf{v}\|$ and the optimal value of problem (56) is given by

$$\mathbf{v}^T\mathbf{o} - r\|\mathbf{v}\|. \tag{64}$$

Case 2. If $-\frac{\mathbf{v}^T\mathbf{u}+\mu d'}{\|\mathbf{u}\|^2} > 0$, then $\nu = -\frac{\mathbf{v}^T\mathbf{u}+\mu d'}{\|\mathbf{u}\|^2}$ and

$$g(\mu, \nu) = -\frac{1}{2\mu}\|\mathbf{v}^\perp\|^2 - \frac{\mu}{2}\left(r^2 - \frac{(d')^2}{\|\mathbf{u}\|^2}\right) + \frac{\mathbf{v}^T\mathbf{u}d'}{\|\mathbf{u}\|^2}, \tag{65}$$

Thus,

$$\frac{\partial g(\mu, \nu)}{\partial \mu} = 0 \Leftrightarrow \mu = \frac{\|\mathbf{v}^\perp\|}{\sqrt{r^2 - \frac{(d')^2}{\|\mathbf{u}\|^2}}} \tag{66}$$

Then $g(\mu, \nu) = -\|\mathbf{v}^\perp\|\sqrt{r^2 - \frac{(d')^2}{\|\mathbf{u}\|^2}} + \frac{\mathbf{v}^T\mathbf{u}d'}{\|\mathbf{u}\|^2}$ and the optimal value of problem (56) is given by

$$\mathbf{v}^T\mathbf{o} - \|\mathbf{v}^\perp\|\sqrt{r^2 - \frac{(d')^2}{\|\mathbf{u}\|^2}} + \frac{\mathbf{v}^T\mathbf{u}d'}{\|\mathbf{u}\|^2}. \tag{67}$$

Now let us consider the case with $\mathbf{v}$ pointing in the opposite direction of $\mathbf{u}$. We can see that there exists $\gamma = -\frac{\mathbf{u}^T\mathbf{v}}{\|\mathbf{u}\|^2} > 0$ such that $\mathbf{v} = -\gamma\mathbf{u}$. By plugging $\mathbf{v} = -\gamma\mathbf{u}$ in problem (56) and following an analogous argument as before, we can see that the statement in Lemma 19 is also applicable to this case.

Therefore, the proof of the statement is completed. □

We are now ready to prove Theorem 18.

Proof. To prove the statements in (52) and (53), we only need to set

$$\begin{aligned} &\mathbf{v} := \bar{\mathbf{x}}_i, \quad \mathbf{u} : -\mathbf{w}^*(s_a), \quad d := -\|\mathbf{w}^*(s_a)\|^2, \\ &\mathbf{o} := \tfrac{1}{2}\hat{\mathbf{w}}, \quad r := \tfrac{1}{2}\|\hat{\mathbf{w}}\|, \\ &d' := \rho = -\|\mathbf{w}^*(s_a)\|^2 + \frac{1}{2}\langle \mathbf{w}^*(s_a), \hat{\mathbf{w}}\rangle, \end{aligned}$$

and then apply Lemma 19. Notice that, for case 1, the optimal value

$$f^* = \frac{1}{2}\left(\langle \hat{\mathbf{w}}(s_b), \bar{\mathbf{x}}_i\rangle - \|\hat{\mathbf{w}}(s_b)\|\|\bar{\mathbf{x}}_i\|\right) \leq 0,$$

and thus none of the non-support vectors can be identified [recall that, according to (R1′), f^* has to be larger than 1 such that $\bar{\mathbf{x}}_i$ can be detected as a non-support vector]. As a result, we only need to consider case 2.

The statement in (54) and (55) follows with an analogous argument by noting that

$$\max_{\mathbf{w}\in\Theta'_{[s_b,s_a]}} \langle \mathbf{w}, \bar{\mathbf{x}}_i\rangle = -\min_{\mathbf{w}\in\Theta'_{[s_b,s_a]}} -\langle \mathbf{w}, \bar{\mathbf{x}}_i\rangle.$$

□