Proximal operator and optimality conditions for ramp loss SVM

Huajun Wang · Yuanhai Shao · Naihua Xiu

Received: 21 September 2020 / Accepted: 25 May 2021 / Published online: 6 June 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract
Support vector machines with ramp loss ($L_r$-SVM) have attracted considerable attention due to the robustness of the ramp loss. However, the corresponding optimization problem is non-convex, and the given Karush–Kuhn–Tucker (KKT) conditions are only first-order necessary conditions. To enrich the optimality theory of $L_r$-SVM, we first introduce and analyze the proximal operator for the ramp loss, and then establish a stronger optimality condition: P-stationarity, which is proved to be the first-order necessary and sufficient conditions for the local minimizer of $L_r$-SVM. Finally, we define the P-support vectors based on the P-stationary point and show that under mild conditions, all of the P-support vectors for $L_r$-SVM are on the two support hyperplanes.

Keywords Ramp loss SVM · Proximal operator · Minimizer · P-stationary point · Support vector

1 Introduction
Support vector machines (SVM) were first introduced by Cortes and Vapnik [9] and have been widely applied in many fields, including text and image classification [5, 11, 15, 20] and disease detection [4, 14, 17]. In this paper, we consider an SVM with linear kernel for the binary classification problem, which is described as follows. The
Fig. 1  a The solid line denotes the ramp loss function.  b The two lines denote the two $\ell_r$ proximal operators with $\gamma C = 0.5, 1.5$ where the dotted line denotes its discontinuity at $s = 1 + \frac{\sqrt{C}}{2}$.  c The two lines denote the two $\ell_r$ proximal operators with $\gamma C = 2, 4$ where the dotted line denotes its discontinuity at $s = \sqrt{2\gamma C}$.

decision hyperplane of SVM classifier, $\langle w, x \rangle + b = 0$ with $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$, is trained from training set $\{(x_i, y_i), i \in \mathbb{N}_m\}$, where $x_i \in \mathbb{R}^n$, $y_i \in \{-1, 1\}$, $m > n$ and $\mathbb{N}_m := \{1, 2, \ldots, m\}$. The well-known SVM is the hinge loss SVM ($L_h$-SVM), which optimizes the following problem:

$$
\min_{w \in \mathbb{R}^n, b \in \mathbb{R}} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{m} \ell_h(1 - y_i(\langle w, x_i \rangle + b)),
$$

(1)

where $C > 0$ is a penalty parameter and $\ell_h(t) = \max\{t, 0\}, t \in \mathbb{R}$ is the hinge loss, which has no cost for $t < 0$, but has a linear cost for $t \geq 0$. However, the $L_h$-SVM tends to be sensitive to outliers because the hinge loss is an unbounded function. To overcome this trouble, Bartlett et al. [1] investigated the robustness of bounded ramp loss to outliers. Here, the ramp loss is defined as follows (see Fig. 1a).

$$
\ell_r(t) := \max\{0, \min\{1, t\}\},
$$

(2)

which is also known as truncated/robust hinge loss in [3,22]. The ramp loss has no cost for $t < 0$, but it has a linear cost for $0 \leq t < 1$ and a fixed cost of 1 for $t \geq 1$, which ensures robustness to outliers because the ramp loss has a fixed cost for outliers. Collobert et al. in [7,8] also extended the ramp loss by replacing all constants with the value of 1 as an adjustable parameter $\mu (\mu > 0)$ in (2). Because the theoretical results have no essential difference between the fixed 1 and parameter $\mu$ for the ramp loss, for simplicity of the proof, we directly use formula (2).

To increase the robustness of $L_h$-SVM, Shen et al. [19] first used the ramp loss to SVM and obtained the ramp loss SVM ($L_r$-SVM). The optimization problem of $L_r$-SVM is

$$
\min_{w \in \mathbb{R}^n, b \in \mathbb{R}} \frac{1}{2} \|w\|^2 + C L_r(1 - A w - b y),
$$

(3)

where $A := [y_1 x_1 y_2 x_2 \cdots y_m x_m]^T \in \mathbb{R}^{m \times n}$, $y := (y_1, y_2, \ldots, y_m)^T \in \mathbb{R}^m$, $1 := (1, 1, \ldots, 1)^T \in \mathbb{R}^m$, $L_r(u) := \sum_{i=1}^{m} \ell_r(u_i)$ with $u := (u_1, u_2, \ldots, u_m)^T \in \mathbb{R}^m$, which computes the sum of positive elements for all $u_i < 1$ and the number of
elements for all $u_i \geq 1$ in $u$. Furthermore, $L_r$-SVM as a stronger robust SVM has been extensively studied in [3,6–8,10,12,21–23].

Due to the non-convexity of the ramp loss, $L_r$-SVM is a non-differentiable and non-convex problem. Thus, to find a satisfying solution, some scholars studied the optimality conditions for $L_r$-SVM. Collobert et al. [7] established the optimality conditions of $L_r$-SVM: Karush–Kuhn–Tucker (KKT) conditions, which are proved to be the first-order necessary conditions for the local minimizers of $L_r$-SVM. To characterize the support vectors for $L_r$-SVM, based on the KKT conditions, the researchers in [7,22] defined the support vectors for $L_r$-SVM and showed that the outliers are not support vectors for $L_r$-SVM. To reduce the scale of the training set of $L_r$-SVM, based on the KKT conditions, Wang et al. [21] designed an efficient working set selection strategy, and Ertekin et al. [10] proposed an outlier filtering mechanism.

A natural question arises. Is there stronger first-order optimality conditions to characterize the minimizer of $L_r$-SVM? In this paper, we provide an affirmative answer to this question. The main results of this paper are summarized as follows: (i) We derive the explicit expression of the proximal operator of ramp loss. (ii) We introduce a novel optimality condition: P-stationarity, which is proved to be the necessary and sufficient conditions for the local minimizer of $L_r$-SVM. (iii) We prove that for $\gamma C \geq 2$, all of the P-support vectors for $L_r$-SVM are on the two support hyperplanes, which means that the outliers do not participate in the construction of the decision function.

This paper is organized as follows. In Sect. 2, we derive the explicit expression of the proximal operator of ramp loss. In Sect. 3, we introduce the definition of the P-stationary point, and reveal the relationship between the P-stationary point and the local/global minimizer of $L_r$-SVM. In Sect. 4, we introduce the P-support vectors based on the P-stationary point and discuss their properties. The conclusions are presented in Sect. 5.

2 Proximal operator for ramp loss

In this section, we derive the explicit expression of the proximal operator of ramp loss, which will be used to study the new first-order optimality conditions of $L_r$-SVM in the next section.

2.1 $\ell_r$ proximal operator

The proximal operator, first introduced by Moreau [16], is an effective tool for analyzing smooth/non-smooth problems and plays an important role in optimality theory and algorithmic design [2,18]. Next, we give the definition of $\ell_r$ proximal operator for the one-dimensional case.

**Definition 1 ($\ell_r$ proximal operator [18])** For any given $\gamma, C > 0$ and $s \in \mathbb{R}$, the proximal operator of $\ell_r(v)$ (dubbed as $\ell_r$ proximal operator) is defined as

$$\text{prox}_{\gamma C\ell_r}(s) = \arg\min_{v \in \mathbb{R}} C\ell_r(v) + \frac{1}{2\gamma}(v - s)^2.$$  (4)
The following two propositions state that the $\ell_r$ proximal operator admits a closed-form solution for $0 < \gamma C < 2$ or $\gamma C \geq 2$.

**Proposition 1** (Solution to $\ell_r$ proximal operator for $0 < \gamma C < 2$) For any given $\gamma, C > 0$ and $0 < \gamma C < 2$, the solution to $\ell_r$ proximal operator at $s \in \mathbb{R}$ is given as

$$
\text{prox}_{\gamma C \ell_r}(s) := \begin{cases} 
  s, & s > 1 + \frac{\gamma C}{2}, \\
  s - \gamma C, & s = 1 + \frac{\gamma C}{2}, \\
  s - \gamma C, & \gamma C \leq s < 1 + \frac{\gamma C}{2}, \\
  0, & 0 < s < \gamma C, \\
  s, & s \leq 0.
\end{cases}
$$

**Proof** For $0 < \gamma C < 2$, it follows from (4) that $\text{prox}_{\gamma C \ell_r}(s)$ is the minimizer of the following function

$$
\phi(v) := \begin{cases} 
  \phi_1(v) := \gamma C + \frac{(v-s)^2}{2}, & v > 1, \\
  \phi_2(v) := \gamma C + \frac{(1-s)^2}{2}, & v = 1, \\
  \phi_3(v) := \gamma C v + \frac{(v-s)^2}{2}, & 0 < v < 1, \\
  \phi_4(v) := \frac{s^2}{2}, & v = 0, \\
  \phi_5(v) := \frac{(v-s)^2}{2}, & v < 0.
\end{cases}
$$

Because $\phi_1(v)$ for $v > 1$, $\phi_3(v)$ for $0 < v < 1$ and $\phi_5(v)$ for $v < 0$ are strongly convex and twice continuously differentiable, the unique minimal values of $\phi_1(v)$, $\phi_3(v)$ and $\phi_5(v)$ are attained at $v = s$, $v = s - \gamma C$ and $v = s$. Moreover, $\phi_2(v)$ and $\phi_4(v)$ attain their minimal values at non-differentiable points at $v = 1$ and $v = 0$. The rest part is to compare the five values $\phi_1(v)$ for $v > 1$, $\phi_2(v)$ for $v = 1$, $\phi_3(v)$ for $0 < v < 1$, $\phi_4(v)$ for $v = 0$ and $\phi_5(v)$ for $v < 0$:

(i) As $s > 1 + \frac{\gamma C}{2} \Leftrightarrow \min\{\phi_2(v), \phi_3(v), \phi_4(v), \phi_5(v)\} > \phi_1(v)$, we obtain $v = s$.

(ii) As $s = 1 + \frac{\gamma C}{2} \Leftrightarrow \min\{\phi_2(v), \phi_4(v), \phi_5(v)\} > \phi_1(v) = \phi_3(v)$, we have $v = s$ or $v = s - \gamma C$.

(iii) As $\gamma C \leq s < 1 + \frac{\gamma C}{2} \Leftrightarrow \min\{\phi_1(v), \phi_2(v), \phi_4(v), \phi_5(v)\} > \phi_3(v)$, we obtain $v = s - \gamma C$.

(iv) As $0 < s < \gamma C \Leftrightarrow \min\{\phi_1(v), \phi_2(v), \phi_3(v), \phi_5(v)\} > \phi_4(v)$, we get $v = 0$.

(v) As $s \leq 0 \Leftrightarrow \min\{\phi_1(v), \phi_2(v), \phi_3(v), \phi_4(v)\} > \phi_5(v)$, we get $v = s$.

Summarizing the above analysis, we have (5), which completes the proof. \hfill $\square$

**Proposition 2** (Solution to $\ell_r$ proximal operator for $\gamma C \geq 2$) For any given $\gamma, C > 0$ and $\gamma C \geq 2$, the solution to $\ell_r$ proximal operator at $s \in \mathbb{R}$ is given as

$$
\text{prox}_{\gamma C \ell_r}(s) := \begin{cases} 
  s, & s > \sqrt{2\gamma C}, \\
  s or 0, & s = \sqrt{2\gamma C}, \\
  0, & 0 < s < \sqrt{2\gamma C}, \\
  s, & s \leq 0.
\end{cases}
$$
Proof For $\gamma C \geq 2$, it follows from (4) that $\text{prox}_{\gamma C \ell_r}(s)$ is the minimizer of the following function

$$
\psi(\tau) := \begin{cases} 
  \psi_1(\tau) := \gamma C + \frac{(\tau - s)^2}{2}, & \tau > 1, \\
  \psi_2(\tau) := \gamma C + \frac{(1-s)^2}{2}, & \tau = 1, \\
  \psi_3(\tau) := \gamma C \tau + \frac{(\tau - s)^2}{2}, & 0 < \tau < 1, \\
  \psi_4(\tau) := s^2, & \tau = 0, \\
  \psi_5(\tau) := \frac{(\tau - s)^2}{2}, & \tau < 0.
\end{cases}
$$

As in Proposition 1, the unique minimal values of $\psi_1(\tau)$, $\psi_2(\tau)$, $\psi_3(\tau)$, $\psi_4(\tau)$ and $\psi_5(\tau)$ are attained at $\tau = s$, $\tau = 1$, $\tau = s - \gamma C$, $\tau = 0$ and $\tau = s$ respectively. The rest part is to compare the five values $\psi_1(\tau)$ for $\tau > 1$, $\psi_2(\tau)$ for $\tau = 1$, $\psi_3(\tau)$ for $0 < \tau < 1$, $\psi_4(\tau)$ for $\tau = 0$ and $\psi_5(\tau)$ for $\tau < 0$:

(i) As $s > \sqrt{2\gamma C} \iff \min\{\psi_2(\tau), \psi_3(\tau), \psi_4(\tau), \psi_5(\tau)\} > \psi_1(\tau)$, we have $\tau = s$.
(ii) As $s = \sqrt{2\gamma C} \iff \min\{\psi_2(\tau), \psi_3(\tau), \psi_5(\tau)\} > \psi_1(\tau) = \psi_4(\tau)$, we obtain $\tau = s$ or $\tau = 0$.
(iii) As $0 < s < \sqrt{2\gamma C} \iff \min\{\psi_1(\tau), \psi_2(\tau), \psi_3(\tau), \psi_5(\tau)\} > \psi_4(\tau)$, we get $\tau = s$.
(iv) As $s \leq 0 \iff \min\{\psi_1(\tau), \psi_2(\tau), \psi_3(\tau), \psi_4(\tau)\} > \psi_5(\tau)$, we have $\tau = s$.

Thus, we have (6), which completes the proof. □

Next, we use a simple example to illustrate the $\ell_r$ proximal operator with $0 < \gamma C < 2$ in (5) and $\gamma C \geq 2$ in (6), which are shown in Fig. 1b, c.

2.2 $L_r$ proximal operator

Based on the separate property of $L_r(\cdot)$, we extend (5) and (6) to a multi-dimensional case.

Definition 2 ($L_r$ proximal operator) For any given $\gamma, C > 0$, the proximal operator of $L_r(v)$ (dubbed as $L_r$ proximal operator) at $s = (s_1, s_2, \ldots, s_m)^\top \in \mathbb{R}^m$ is defined as

$$
\text{prox}_{\gamma CL_r}(s) := \arg \min_{v \in \mathbb{R}^m} CL_r(v) + \frac{1}{2\gamma} \|v - s\|^2.
$$

The following proposition states that the $L_r$ proximal operator admits a closed-form solution for two cases: $0 < \gamma C < 2$ or $\gamma C \geq 2$.

Proposition 3 (Solution to $L_r$ proximal operator) For a given $\gamma, C > 0$, the solution to $L_r$ proximal operator at $s = (s_1, s_2, \ldots, s_m)^\top \in \mathbb{R}^m$ is given as

$$
\text{prox}_{\gamma CL_r}(s) := \begin{bmatrix} \text{prox}_{\gamma C\ell_r}(s_1) \\ \vdots \\ \text{prox}_{\gamma C\ell_r}(s_m) \end{bmatrix}.
$$
where $\text{prox}_{\gamma C\ell_r}(s_i)$ takes the formula (5) as $0 < \gamma C < 2$ or (6) as $\gamma C \geq 2$.

**Proof** It follows from (7) that $[\text{prox}_{\gamma C\ell_r}(s)]_i = \text{prox}_{\gamma C\ell_r}(s_i)$, where

$$
\text{prox}_{\gamma C\ell_r}(s_i) = \arg\min_{v \in \mathbb{R}} C\ell_r(v) + \frac{1}{2\gamma} (v - s_i)^2, \quad i \in \mathbb{N}_m.
$$

Using (5) or (6) completes the proof. □

### 3 First-order optimality conditions

In this section, we develop a first-order necessary and sufficient optimality conditions for (3). To proceed with this, we introduce a variable $u \in \mathbb{R}^m$ to equivalently reformulate (3) as

$$
\begin{aligned}
\min_{w \in \mathbb{R}^n, b \in \mathbb{R}, u \in \mathbb{R}^m} & \quad \frac{1}{2} \|w\|^2 + CL_r(u) \\
\text{s.t.} & \quad u + Aw + by = 1.
\end{aligned}
$$

Now let us define some notation

$$
B := [A \ y] \in \mathbb{R}^{m \times (n+1)}, \quad E := \begin{bmatrix} I_{n \times n} & 0 \\ 0 & 0 \end{bmatrix}, \quad H := EB^\dagger,
$$

where $B^\dagger \in \mathbb{R}^{(n+1) \times m}$ is the generalized inverse of $B$, and $I_{n \times n}$ is the identity matrix of order $n$. Denote $\lambda_H := \lambda_{\max}(H^\top H)$, where $\lambda_{\max}(H^\top H)$ is the maximum eigenvalue of $H^\top H$.

**Definition 3** (P-stationary point of (9)) For a given $C > 0$, we call $(w^*; b^*; u^*)$ is a proximal stationary (P-stationary) point of (9) if there exists a Lagrangian multiplier vector $\lambda^* \in \mathbb{R}^m$ and a constant $\gamma > 0$ such that

$$
\begin{aligned}
\left\{ 
\begin{array}{l}
w^* + A^\top \lambda^* = 0, \\
(y, \lambda^*) = 0, \\
u^* + Aw^* + b^*y = 1, \\
\text{prox}_{\gamma CL_r}(u^* - \gamma \lambda^*) \ni u^*.
\end{array}
\right.
\end{aligned}
$$

Based on the above definition, we obtain the desired result in this section.

**Theorem 1** (First-order necessary optimality conditions) Let $B$ be the full column rank. For a given $C > 0$, if $(w^*; b^*; u^*)$ is a global minimizer of (9), then it is a P-stationary point with $\lambda^* \in \mathbb{R}^m$ and $0 < \gamma < 1/\lambda_H$. 

\(\mathbb{C}\) Springer
Proof Because $B$ is the full column rank, then the $B^\top B$ is the inverse, which means $(B^\top B)^{-1}B^\top B = I_{m \times m}$. Because $B^\dagger B = I_{m \times m}$, then we have $(B^\dagger)^\top = B(B^\top B)^{-1}$ by [13]. Furthermore, the equality constraint in (9) is equivalent to $(w; b) = B^\dagger (1 - u)$ due to $1 - u = A w + b y = B(w; b)$. By (10) and $H(1 - u) = E B^\dagger B \left[ \begin{array}{c} w \\ b \\ 0 \end{array} \right]$, the problem (9) becomes

$$\min_{u \in \mathbb{R}^m} \frac{1}{2} \| H(u - 1) \|^2 + C L_r(u). \quad (12)$$

Since $(u^*; b^*; u^*)$ is a global minimizer of (9), $u^*$ is also a global minimizer of (12). To prove (11), we prove that (a) the global minimizer $u^*$ of (12) satisfies

$$u^* = \operatorname{prox}_{\gamma C L_r}(u^* - \gamma H^\top H(u^* - 1)), \quad (13)$$

and (b) (13) means (11), where $0 < \gamma < 1/\lambda H$. For simplicity, denote

$$g(u) := \| H(u - 1) \|^2/2, \quad \lambda^* := H^\top H(u^* - 1). \quad (14)$$

To prove (a), we denote $P := \operatorname{prox}_{\gamma C L_r}(u^* - \gamma \lambda^*)$ and consider any point $z \in P$. From the definition of the $L_r$ proximal operator at $u^* - \gamma \lambda^*$, we have

$$C L_r(z) + \frac{1}{2\gamma} \| z - (u^* - \gamma \lambda^*) \|^2 \leq C L_r(u^*) + \frac{1}{2\gamma} \| u^* - (u^* - \gamma \lambda^*) \|^2$$

$$= C L_r(u^*) + \frac{\gamma}{2} \| \lambda^* \|^2. \quad (15)$$

Since $g$ is a quadratic function on $u$, we have

$$g(z) - g(u^*) \leq \langle \lambda^*, z - u^* \rangle + \frac{\lambda_H}{2} \| z - u^* \|^2. \quad (16)$$

Because $u^*$ is the global minimizer of (12), we derive that

$$C L_r(u^*) + g(u^*) \leq C L_r(z) + g(z). \quad (17)$$
From these three facts, we conclude that

\[
0 \leq CL_r(z) + g(z) - (CL_r(u^*) + g(u^*)) \\
\leq CL_r(z) + \langle \lambda^*, z - u^* \rangle + \frac{\lambda H}{2} \|z - u^*\|^2 - CL_r(u^*) \\
= CL_r(z) + \langle \lambda^*, z - u^* \rangle + \frac{1}{2\gamma} \|z - u^*\|^2 - CL_r(u^*) - \frac{\lambda H - \frac{1}{2\gamma}}{2} \|z - u^*\|^2 \\
\leq \frac{\lambda H - \frac{1}{2\gamma}}{2} \|z - u^*\|^2,
\]

which indicates that \( \|z - u^*\|^2 \leq 0 \) due to \( 0 < \gamma < 1/\lambda H \). Hence, we have \( z = u^* \). Since \( z \) is arbitrary in \( P \) and \( z = u^* \), thus \( P \) is a singleton containing only \( u^* \), which completes the proof of result a).

Next we prove (b). From (10) and (14), we have

\[-\lambda^* = H^T H(1 - u^*) = H^T E B^\dagger (1 - u^*) = H^T E \begin{bmatrix} w^* \\ b^* \end{bmatrix},\]

which leads to

\[-B^T \lambda^* = B^T H^T E \begin{bmatrix} w^* \\ b^* \end{bmatrix} = B^T (B^\dagger)^T E^T E \begin{bmatrix} w^* \\ b^* \end{bmatrix} = \begin{bmatrix} w^* \\ 0 \end{bmatrix},\]

where we used two facts: \( B^T (B^\dagger)^T = B^T B (B^T B)^{-1} = I_{m \times m} \) and \( E^T E = E \). From \( B = [A \ y] \), the above equation yields

\[
\begin{cases}
    w^* + A^\top \lambda^* = 0, \\
    \langle y, \lambda^* \rangle = 0.
\end{cases}
\]

Summarizing the above equations, the feasibility of \((w^*; b^*; u^*)\) and (13), we obtain (11), which completes the proof.

\[\square\]

**Theorem 2** (First-order sufficient optimality conditions) For a given \( C > 0 \), if \((w^*; b^*; u^*)\) with \( \lambda^* \in \mathbb{R}^m \) and \( \gamma > 0 \) is a \( P \)-stationary point, then it is a local minimizer of (9) and the \( \lambda^* = (\lambda_1^*, \lambda_2^*, \ldots, \lambda_m^*)^\top \in \mathbb{R}^m \) satisfies

\[
\lambda_i^* \begin{cases}
    \in [-C, 0], & 0 < \gamma C < 2, \\
    \in [-\sqrt{2C/\gamma}, 0], & \gamma C \geq 2,
\end{cases} \quad i \in \mathbb{N}_m.
\]
**Proof** Denote \( \phi^* := (w^*; b^*; u^*) \) and let \( \Theta \) be the feasible region of (9), i.e.,

\[
\Theta := \{ \phi := (w; b; u) : u + Aw + by = 1 \}.
\]

Then, for any \( \phi \in \Theta \), we have \( u + Aw + by = 1 \), which and (11) suffice to

\[
-A(w - w^*) = u - u^* + (b - b^*)y.
\]  \( \tag{19} \)

The convexity of \( \|w\|^2 \) allows us to derive that

\[
\|w\|^2 - \|w^*\|^2 \geq 2\langle w - w^*, w^* \rangle \overset{11}{=} -2\langle A(w - w^*), \lambda^* \rangle
\]

\[
\overset{19}{=} 2\langle u - u^*, \lambda^* \rangle + 2(b - b^*)(y, \lambda^*) \overset{11}{=} 2\langle u - u^*, \lambda^* \rangle.
\]  \( \tag{20} \)

Now, we show that \( \phi^* \) is a local minimizer of (9). Namely, there exists a neighborhood \( U(\phi^*, \delta) \) of \( \phi^* \in \Theta \) with \( \delta > 0 \) such that

\[
\frac{1}{2} \|w^*\|^2 + CL_r(u^*) \leq \frac{1}{2} \|w\|^2 + CL_r(u), \quad \forall \phi \in \Theta \cap U(\phi^*, \delta).
\]  \( \tag{21} \)

For this purpose, we let

\[
\delta := \begin{cases} \frac{\gamma C}{2}, & 0 < \gamma C < 2, \\ 1, & \gamma C \geq 2, \end{cases} \quad \delta_m := \delta/\sqrt{2\gamma},
\]

and denote a local region of \( \phi^* = (w^*; b^*; u^*) \) by

\[
U(\phi^*, \delta) := \{ \phi : \|(w; b) - (w^*; b^*)\| \leq \delta/\sqrt{2}, |u_i - u_i^*| \leq \delta_m, i \in \mathbb{N}_m \}.
\]  \( \tag{23} \)

To show that (21) and \( \lambda^* \in \mathbb{R}^m \) satisfies (18), we consider the following two cases:

**Case i)** \( 0 < \gamma C < 2 \) and \( \phi \in \Theta \cap U(\phi^*, \delta) \). Denote \( r^* := u^* - \gamma \lambda^* \in \mathbb{N}_m \), and

\[
\begin{align*}
\Gamma_1^* &:= \{ i \in \mathbb{N}_m : r_i^* \leq 0 \}, & \Gamma_2^* &:= \{ i \in \mathbb{N}_m : r_i^* \in (0, \gamma C) \}, \\
\Gamma_3^* &:= \{ i \in \mathbb{N}_m : r_i^* \in [\gamma C, 1 + \frac{\gamma C}{2}) \cup \{ i \in \mathbb{N}_m : r_i^* = 1 + \frac{\gamma C}{2}, \lambda_i^* \neq 0 \}, \\
\Gamma_4^* &:= \{ i \in \mathbb{N}_m : r_i^* > 1 + \frac{\gamma C}{2} \cup \{ i \in \mathbb{N}_m : r_i^* = 1 + \frac{\gamma C}{2}, \lambda_i^* = 0 \}.
\end{align*}
\]  \( \tag{24} \)

It follows from (8) and (5) that \( u^* \in \text{prox}_{\gamma CL_r}(u^* - \gamma \lambda^*) \) is equivalent to

\[
u^* = \begin{bmatrix} (\text{prox}_{\gamma CL_r}(u^* - \gamma \lambda^*))_{\Gamma_1^*} \\ (\text{prox}_{\gamma CL_r}(u^* - \gamma \lambda^*))_{\Gamma_2^*} \\ (\text{prox}_{\gamma CL_r}(u^* - \gamma \lambda^*))_{\Gamma_3^*} \\ (\text{prox}_{\gamma CL_r}(u^* - \gamma \lambda^*))_{\Gamma_4^*} \end{bmatrix} = \begin{bmatrix} (u^* - \gamma \lambda^*)_{\Gamma_1^*} \\ 0_{\Gamma_2^*} \\ (u^* - \gamma \lambda^* - \gamma C 1)_{\Gamma_3^*} \\ (u^* - \gamma \lambda^*)_{\Gamma_4^*} \end{bmatrix},
\]

which is identical to

\[
\lambda^*_{\Gamma_1^*} = 0_{\Gamma_1^*}, \quad u^*_{\Gamma_2^*} = 0_{\Gamma_2^*}, \quad \lambda^*_{\Gamma_3^*} = -C 1_{\Gamma_3^*}, \quad \lambda^*_{\Gamma_4^*} = 0_{\Gamma_4^*}.
\]
This and (24) lead to
\[
\lambda_i^* = 0, u_i^* \leq 0, i \in \Gamma_1^*, \quad \lambda_i^* \in (-C, 0), u_i^* = 0, i \in \Gamma_2^*,
\]
\[
\lambda_i^* = -C, u_i^* \in [0, 1 - \frac{\gamma C}{2}], i \in \Gamma_3^*, \quad \lambda_i^* = 0, u_i^* \geq 1 + \frac{\gamma C}{2}, i \in \Gamma_4^*,
\]
which shows that \(\lambda_i^* \in [-C, 0]\) for \(0 < \gamma C < 2\). Denote \(\Gamma^* := \Gamma_2^* \cup \Gamma_3^*, \quad \Gamma^* := \Gamma_1^* \cup \Gamma_4^*\). To prove (21), from (20), (25) and (26), we only need to verify the following two results:
\[
C L_r(u_{\Gamma^*}) - C L_r(u_{\Gamma^*}) + (u_{\Gamma^*} - u_{\Gamma^*}, \lambda_{\Gamma^*}) \geq 0, \quad (27)
\]
\[
C L_r(u_{\Gamma^*}) - C L_r(u_{\Gamma^*}) \geq 0. \quad (28)
\]
From (2) and (22)–(26), we obtain the desired conclusion:
\begin{enumerate}
\item For \(i \in \Gamma_1^*\), by \(u_i^* \in [0, 1 - \frac{\gamma C}{2}]\) and \(|u_i - u_i^*| \leq \delta_m\), we obtain \(u_i \leq u_i^* + \delta_m < 1\), which implies \(\ell_r(u_i) \geq u_i^*\). This together with \(C > -\lambda_i^* > 0\) from (25) and \(C = -\lambda_i^*\) from (26) allows us to obtain (27).
\item For \(i \in \Gamma_2^*\), from \(u_i^* \leq 0\) and \(|u_i - u_i^*| \leq \delta_m\), we get \(u_i \leq u_i^* + \delta_m < 1\), which means \(\ell_r(u_i) \geq \lambda_i^*\). Hence (28) holds.
\item For \(i \in \Gamma_3^*\), it follows from \(u_i^* \geq 1 + \frac{\gamma C}{2}\) and \(|u_i - u_i^*| \leq \delta_m\) that we have \(u_i \geq u_i^* - \delta_m > 1\), which obtains \(\ell_r(u_i) = \ell_r(u_i^*)\). Hence, (28) holds.
\end{enumerate}
Summarizing the above (a1), (a2) and (a3), we have (27) and (28), which completes the proof of Case i).

Case ii) \(\gamma C \geq 2\) and \(\phi \in \Theta \cap U(\phi^*, \delta)\). Denote \(r^* = u^* - \gamma \lambda^* \in N_m\), and
\[
\begin{align*}
F_1^* := \{i \in N_m : r_i^* \leq 0\}, \\
F_2^* := \{i \in N_m : r_i^* \in (0, \sqrt{2\gamma C}) \cup \{i \in N_m : r_i^* = \sqrt{2\gamma C}, u_i^* = 0\}\}, \\
F_3^* := \{i \in N_m : r_i^* > \sqrt{2\gamma C} \cup \{i \in N_m : r_i^* = \sqrt{2\gamma C}, u_i^* \neq 0\}\}.
\end{align*}
\]
From (8) and (6), we have that \(u^* \in \text{prox}_{\gamma C L_r}(u^* - \gamma \lambda^*)\) is equivalent to
\[
\begin{bmatrix}
(u^* - \gamma \lambda^*)_{F_1^*} \\
(u^* - \gamma \lambda^*)_{F_2^*} \\
(u^* - \gamma \lambda^*)_{F_3^*}
\end{bmatrix} = \begin{bmatrix}
0_{F_1^*} \\
0_{F_2^*} \\
0_{F_3^*}
\end{bmatrix},
\]
which is identical to
\[
\begin{align*}
\lambda_{F_1^*} &= 0, \\
u_{F_2^*} &= 0, \\
\lambda_{F_3^*} &= 0.
\end{align*}
\]
This and (29) lead to
\[
\begin{align*}
\lambda_i^* &= 0, u_i^* \leq 0, i \in F_1^*, \\
\lambda_i^* &= \sqrt{2\gamma C}, u_i^* = 0, i \in F_2^*, \\
\lambda_i^* &= 0, u_i^* \geq 2\gamma C, i \in F_3^*.
\end{align*}
\]
which shows that \( \lambda^*_i \in [-\sqrt{2C/\gamma}, 0] \) for \( \gamma C \geq 2 \). Denote \( \mathcal{F}^* := \mathcal{F}_2^* \), \( \bar{\mathcal{F}}^* := \mathcal{F}_1^* \cup \mathcal{F}_3^* \). To show (21), we only need to prove the following two inequalities:

\[
C L_r(u_{\mathcal{F}^*}) - C L_r(u_{\bar{\mathcal{F}}^*}) + (u_{\mathcal{F}^*} - u_{\bar{\mathcal{F}}^*}, \lambda_{\mathcal{F}^*}) \geq 0, \tag{32}
\]

\[
C L_r(u_{\mathcal{F}^*}) - C L_r(u_{\bar{\mathcal{F}}^*}) \geq 0. \tag{33}
\]

From (2), (22), (23) and (29)–(31), we obtain the desired conclusion.

(b1) For \( i \in \mathcal{F}^* \), from \( u_i^* = 0 \) and \( |u_i - u_i^*| \leq \delta_m \), we have \( u_i \leq u_i^* + \delta_m < 1 \). This means \( \ell_r(u_i) \geq u_i \) and \( \ell_r(u_i^*) = u_i^* = 0 \), which together with \( C \geq \frac{2}{\gamma} \geq -\lambda_i^* > 0 \) by (30) leads to (32).

(b2) For \( i \in \mathcal{F}_1^* \), from \( u_i^* \leq 0 \) and \( |u_i - u_i^*| \leq \delta_m \), we obtain \( u_i \leq u_i^* + \delta_m < 1 \). This implies that \( \ell_r(u_i) \geq \ell_r(u_i^*) \), and hence (33) holds.

(b3) For \( i \in \mathcal{F}_3^* \), from \( u_i^* \geq \sqrt{2\gamma C} \) and \( |u_i - u_i^*| \leq \delta_m \), we have \( u_i \geq u_i^* - \delta_m > 1 \). This obtains \( \ell_r(u_i) = \ell_r(u_i^*) \), and hence (33) holds.

Summarizing the above (b1), (b2) and (b3), we have (32) and (33), which completes the proof of Case ii.

At the end of this section, we analyze the relationship between the P-stationary point and the KKT point given by [7]. Based on the problem (9) and the necessary conditions results in [7], if \((\bar{w}; \bar{b}; \bar{u})\) is a local minimizer of (9), then we have the following KKT conditions:

\[
\begin{cases}
\bar{w} + A^\top \bar{\lambda} = 0, \\
(y, \bar{\lambda}) = 0, \\
\bar{u} + A\bar{w} + \bar{b}y = 1, \\
C \partial L_r(\bar{u}) + \bar{\lambda} \geq 0,
\end{cases}
\]

where \( \bar{\lambda} \in \mathbb{R}^m \) is a multiplier vector, \((\bar{w}; \bar{b}; \bar{u})\) is called the KKT point of (9), and \( \partial L_r(\bar{u}) = (\partial \ell_r(\bar{u}_1), \partial \ell_r(\bar{u}_2), \ldots, \partial \ell_r(\bar{u}_m))^\top \in \mathbb{R}^m \) is the subdifferential of \( L_r(\bar{u}) \) with

\[
\partial \ell_r(\bar{u}_i) = \begin{cases}
\in [0, 1], & \bar{u}_i \in [0, 1], \\
1, & \bar{u}_i \in (0, 1), \\
0, & \bar{u}_i < 0 \text{ or } \bar{u}_i > 1,
\end{cases}
\]

Furthermore, from the above formula and the last formula of (34), we have

\[
0 \in C \partial L_r(\bar{u}) + \bar{\lambda} \iff \bar{\lambda} \in \mathbb{R}^m : \bar{\lambda}_i \begin{cases}
\in [-C, 0], & \bar{u}_i \in [0, 1], \\
= -C, & \bar{u}_i \in (0, 1), \\
= 0, & \bar{u}_i < 0 \text{ or } \bar{u}_i > 1.
\end{cases}
\]

**Theorem 3** For a given \( C > 0 \), if \((w^*; b^*; u^*)\) with \( \lambda^* \in \mathbb{R}^m \) and \( \gamma > 0 \) is a P-stationary point of (9), then it is also a KKT point of (9), but the converse does not hold.
\textbf{Proof} The former conclusion is obvious since the P-stationary point is the local minimizer of (9). The latter conclusion is based on the following counterexample. Consider the training set with the positive vectors \( x_1 = (3, 3)^\top, x_2 = (6, -2)^\top \) and negative vector \( x_3 = (1, 1)^\top \). Namely,

\[
A := [y_1 x_1 \ y_2 x_2 \ y_3 x_3]^\top = \begin{bmatrix} 3 & 6 & -1 \\ 3 & -2 & -1 \end{bmatrix}, \quad y = [1, \ 1, \ -1]^\top.
\]

For a given \( C = 0.25 \), we can verify that

\[
\bar{w} = (0.5, 0.5)^\top, \quad \bar{b} = -2, \quad \bar{u} = (0, 1, 0)^\top, \quad \lambda = (-0.25, 0, -0.25)^\top, \quad \partial L_r(\bar{u}) = (1, 0, 1)^\top
\]

satisfy (34). This means \((\bar{w}; \bar{b}; \bar{u})\) with \( \lambda \) is a KKT point of (9). In particular, for \( i = 2 \), we have \( \bar{u}_2 = 1 \) and \( \lambda_2 = 0 \). However, for \( 0 < \gamma < 8 \) and \( C = 0.25 \), that is, \( 0 < \gamma C < 2 \), from (5) and \( \bar{u}_2 - \gamma \lambda_2 \in (0, 1 + \frac{\gamma C}{2}) \), we obtain \( \text{prox}_{\gamma C \ell_r}(\bar{u}_2 - \gamma \lambda_2) = 0 \) or \( \bar{u}_2 - \gamma C \) but the both does not equal to \( \bar{u}_2 \). For \( \gamma \geq 8 \) and \( C = 0.25 \), i.e., \( \gamma C \geq 2 \), from (6) and \( \bar{u}_2 - \gamma \lambda_2 \in (0, \sqrt{2} \gamma C) \), we get \( \text{prox}_{\gamma C \ell_r}(\bar{u}_2 - \gamma \lambda_2) = 0 \neq \bar{u}_2 \). In summary, we have \( \bar{u}_2 \notin \text{prox}_{\gamma C \ell_r}(\bar{u}_2 - \gamma \lambda_2) \), which means that \((\bar{w}; \bar{b}; \bar{u})\) is not a P-stationary point of (9).

\[\square\]

\[\textbf{4 P-support vectors}\]

In this section, we define the support vectors for \( L_r \)-SVM by the P-stationary point of (9), which are named as P-support vectors since they are selected by the proximal operator. Before proceeding, we first review the definitions of support vectors for hard margin SVM, hinge loss SVM and ramp loss SVM, respectively.

(i) \textit{Support vectors for hard margin SVM} [9] If \((\tilde{w}; \tilde{b})\) is a global minimizer of hard margin SVM and \( \tilde{\alpha} = (\tilde{\alpha}_1, \tilde{\alpha}_2, \ldots, \tilde{\alpha}_m)^\top \) with \( \tilde{\alpha}_i \geq 0 \) is a solution to its dual problem, then \( \tilde{w} \) satisfies

\[
\tilde{w} = \sum_{i \in \tilde{J}^*} \tilde{\alpha}_i y_i x_i,
\]

where \( \tilde{J}^* := \{ i \in \mathbb{N}_m : \tilde{\alpha}_i > 0 \} \). The training vectors \( \{ x_i, i \in \tilde{J}^* \} \) are called support vectors. For any \( i \in \tilde{J}^* \), the support vector \( x_i \) satisfies

\[
y_i((\tilde{w}, x_i) + \tilde{b}) = 1. \tag{36}\]

Namely, all of the support vectors \( \{ x_i, i \in \tilde{J}^* \} \) are on the two support hyperplanes.

(ii) \textit{Support vectors for hinge loss SVM} [9] If \((\hat{w}; \hat{b})\) is a global minimizer of hinge loss SVM and \( \hat{\alpha} = (\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_m)^\top \) with \( \hat{\alpha}_i \in [0, C] \) is a solution to its dual

\[\square\]
problem, then \( \hat{\mathbf{w}} \) satisfies
\[
\hat{\mathbf{w}} = \sum_{i \in \hat{J}^*} \hat{\alpha}_i y_i \mathbf{x}_i,
\]
where \( \hat{J}^* := \{ i \in \mathbb{N}_m : \hat{\alpha}_i \in (0, C] \} \). The training vectors \( \{ \mathbf{x}_i, i \in \hat{J}^* \} \) are called support vectors. For any \( i \in \hat{J}^* \), the support vector \( \mathbf{x}_i \) satisfies
\[
\begin{align*}
y_i((\hat{\mathbf{w}}, \mathbf{x}_i) + \hat{b}) &= 1, \quad i \in \{ i \in \hat{J}^* : \hat{\alpha}_i \in (0, C) \}, \\
y_i((\hat{\mathbf{w}}, \mathbf{x}_i) + \hat{b}) &\leq 1, \quad i \in \{ i \in \hat{J}^* : \hat{\alpha}_i = C \}.
\end{align*}
\]
That is, a part of the support vectors \( \{ \mathbf{x}_i, i \in \hat{J}^* \} \) are on the two support hyperplanes.

(iii) KKT support vectors for ramp loss SVM [7,22] If \( (\overline{\mathbf{w}}, \overline{b}; \overline{u}) \) with \( \overline{\lambda} \in \mathbb{R}^m \) is a KKT point of (9), from (35) and the first equation of (34), then \( \overline{\mathbf{w}} \) satisfies
\[
\overline{\mathbf{w}} = -\sum_{i \in \overline{J}^*} \overline{\lambda}_i y_i \mathbf{x}_i,
\]
where \( \overline{J}^* := \{ i \in \mathbb{N}_m : \overline{\lambda}_i \in [-C, 0) \} \). The training vectors \( \{ \mathbf{x}_i, i \in \overline{J}^* \} \) are called KKT support vectors. From (35) and the third equation of (34), for any \( i \in \overline{J}^* \), the KKT support vector \( \mathbf{x}_i \) satisfies
\[
\begin{align*}
y_i((\overline{\mathbf{w}}, \mathbf{x}_i) + \overline{b}) &\in [0, 1], \quad i \in \{ i \in \overline{J}^* : \overline{\lambda}_i \in (-C, 0) \}, \\
y_i((\overline{\mathbf{w}}, \mathbf{x}_i) + \overline{b}) &\in [0, 1], \quad i \in \{ i \in \overline{J}^* : \overline{\lambda}_i = -C \}.
\end{align*}
\]
That is, a few of the KKT support vectors \( \{ \mathbf{x}_i, i \in \overline{J}^* \} \) are on the two support hyperplanes.

The above results show that the hard margin SVM and hinge loss SVM define the support vectors at their global minimizer. However, authors in [7,22] defined the KKT support vectors at the KKT point of \( L_r \)-SVM. In the following, we define the P-support vectors for \( L_r \)-SVM at the P-stationary point, which is also the local minimizer of \( L_r \)-SVM.

**Theorem 4** (P-support vectors for \( 0 < \gamma C < 2 \)) For \( 0 < \gamma C < 2 \), if \( (\mathbf{w}^*; \mathbf{b}^*; \mathbf{u}^*) \) with \( \lambda^* \in \mathbb{R}^m \) and \( \gamma > 0 \) is a P-stationary point of (9), then \( (\mathbf{w}^*; \mathbf{b}^*) \) is the local minimizer of (3) and \( \mathbf{w}^* \) satisfies
\[
\mathbf{w}^* = -\sum_{i \in T^*} \lambda_i^* y_i \mathbf{x}_i,
\]
where \( T^* := \{ i \in \mathbb{N}_m : \lambda_i^* \in [-C, 0) \} \). The training vectors \( \{ \mathbf{x}_i : i \in T^* \} \) are called the P-support vectors of (3). For any \( i \in T^* \), the P-support vector \( \mathbf{x}_i \) satisfies
\[
\begin{align*}
y_i((\mathbf{w}^*, \mathbf{x}_i) + \mathbf{b}^*) &= 1, \quad i \in T^* := \{ i \in T^* : \lambda_i^* \in (-C, 0) \}, \\
y_i((\mathbf{w}^*, \mathbf{x}_i) + \mathbf{b}^*) &\in [\frac{-\gamma C}{2}, 1], \quad i \in T_2^* := \{ i \in T^* : \lambda_i^* = -C \}.
\end{align*}
\]
That is, a part of the $P$-support vectors $\{x_i, i \in T^*\}$ are on the two support hyperplanes.

**Proof** For $0 < \gamma C < 2$, since $(w^*; b^*; u^*)$ with $\lambda^* \in \mathbb{R}^m$ and $\gamma > 0$ is a P-stationary point, from Theorem 2, then $(w^*; b^*)$ is a local minimizer of (3) and we have $\lambda^*_i \in [-C, 0], i \in \mathbb{N}_m$. Furthermore, it follows from (25) and (26) that we have

$$T^* = T_1^* \cup T_2^*, \quad \overline{T}^* := \mathbb{N}_m \setminus T^*,$$

where $T_1^* = \Gamma_2^*$ and $T_2^* = \Gamma_3^*$. From the first equation of (11) and using the fact $A = [y_1 x_1 \cdots y_m x_m]^T \in \mathbb{R}^{m \times n}$, we derive

$$w^* = -A^T \lambda^* - A^T \lambda^* = -A^T \lambda^* = -\sum_{i \in T^*} \lambda^*_i y_i x_i.$$  

It follows from (25) and (26) that $u^*_i = 0, i \in T_1^*$ and $u^*_i \in [0, 1 - \gamma C/2], i \in T_2^*$, which together with the third equation of (11) means $(Aw^* + b^*y)_i = 1, i \in T_1^*$ and $(Aw^* + b^*y)_i \in \left[\frac{\gamma C}{2}, 1\right], i \in T_2^*$. This and the definition of $A$ lead to (38), which completes the proof. \hfill \Box

**Theorem 5** (P-support vectors for $\gamma C \geq 2$) For $\gamma C \geq 2$, if $(w^*; b^*; u^*)$ with $\lambda^* \in \mathbb{R}^m$ and $\gamma > 0$ is a P-stationary point of (9), then $(w^*; b^*)$ is the local minimizer of (3) and $w^*$ satisfies

$$w^* = -\sum_{i \in I^*} \lambda^*_i y_i x_i,$$

where $I^* := \{i \in \mathbb{N}_m : \lambda^*_i \in [-\sqrt{2C/\gamma}, 0)\}$. The training vectors $\{x_i : i \in I^*\}$ are called the P-support vectors of (3). For any $i \in I^*$, the P-support vector $x_i$ satisfies

$$y_i ((w^*, x_i) + b^*) = 1.$$  

(39)

Namely, all of the P-support vectors $\{x_i, i \in I^*\}$ are on the two support hyperplanes.

**Proof** For $\gamma C \geq 2$, again, because $(w^*; b^*; u^*)$ with $\lambda^* \in \mathbb{R}^m$ and $\gamma > 0$ is a P-stationary point, from Theorem 2, then $(w^*; b^*)$ is a local minimizer of (3) and we get $\lambda^*_i \in [-\sqrt{2C/\gamma}, 0], i \in \mathbb{N}_m$. Based on (30) and (31), we have

$$I^* = I^*_2, \quad \overline{I}^* := \mathbb{N}_m \setminus I^*,$$

which together with the first equation of (11) and $A = [y_1 x_1 \cdots y_m x_m]^T \in \mathbb{R}^{m \times n}$ leads to

$$w^* = -A^T \lambda^*_I - A^T \lambda^*_I = -A^T \lambda^*_I = -\sum_{i \in I^*} \lambda^*_i y_i x_i.$$  

Furthermore, from (30), we have $u^*_i = 0, i \in I^*$, which together with the third equation of (11) means $1_{I^*} = (Aw^* + b^*y)_{I^*}$. This and the definition of $A$ allow us to obtain (39), which completes the proof. \hfill \Box

Springer
The above theorem shows that the P-support vectors can sufficiently reduce the scale of the training set, and the $L_r$-SVM is resistant to the influence of outliers.

5 Conclusions

In this paper, with the help of explicit expression of the proximal operator for ramp loss, we introduced and characterized a novel first-order necessary and sufficient optimality conditions for the local minimizers of $L_r$-SVM, which is called the P-stationarity. We defined the P-support vectors based on the definition of the P-stationary point and showed that for $\gamma C \geq 2$, all of the P-support vectors for $L_r$-SVM are on two support hyperplanes, that is to say, the decision hyperplane only required very few important training samples from the training set. This result allowed us to define a working set by the P-support vectors and design an iterative algorithm with working set strategy for optimizing $L_r$-SVM in the future. In addition, an extension of the presented theoretical results in this paper to nonlinear kernel $L_r$-SVM is also left for future work.

Acknowledgements The authors would like to thank the associate editor and two anonymous referees for their constructive comments, which have significantly improved the quality of the paper. This work was supported by the National Natural Science Foundation of China (11971052, 11926348-9, 61866010, 11871183), and the Natural Science Foundation of Hainan Province (120RC449).

References

1. Bartlett, P.L., Mendelson, S.: Rademacher and gaussian complexities: risk bounds and structural results. J. Mach. Learn. Res. 3, 463–482 (2002)
2. Beck, A.: First-Order Methods in Optimization. SIAM Publisher, Philadelphia (2017)
3. Brooks, J.P.: Support vector machines with the ramp loss and the hard margin loss. Oper. Res. 59(2), 467–479 (2011)
4. Carrizosa, E., Martin-Barragan, B., Morales, D.R.: Binarized support vector machines. Informs J. Comput. 22(1), 154–167 (2010)
5. Carrizosa, E., Morales, D.R.: Supervised classification and mathematical optimization. Comput. Oper. Res. 40(1), 150–165 (2013)
6. Carrizosa, E., Nogales-Gmez, A., Romero, M.D.: Heuristic approaches for support vector machines with the ramp loss. Optim. Lett. 8(3), 1125–1135 (2014)
7. Collobert, R., Sinz, F., Weston, J., Bottou, L.: Trading convexity for scalability. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 201–208 (2006)
8. Collobert, R., Sinz, F., Weston, J., Bottou, L.: Large scale transductive SVMs. J. Mach. Learn. Res. 7, 1687–1712 (2006)
9. Cortes, C., Vapnik, V.: Support vector networks. Mach. Learn. 20(3), 273–297 (1995)
10. Ertekin, S., Bottou, L., Giles, C.L.: Nonconvex online support vector machines. IEEE Trans. Pattern Anal. Mach. Intell. 33(2), 368–381 (2011)
11. Gaudioso, M., Gorgone, E., Hiriart-Urruty, J.B.: Feature selection in SVM via polyhedral k-norm. Optim. Lett. 14(1), 19–36 (2020)
12. Hess, E.J., Brooks, J.P.: The support vector machine and mixed integer linear programming: ramp loss SVM with $L_1$-norm regularization. In: Proceedings of the 14th Informs Computing Society Conference Richmond, pp. 226–235 (2015)
13. Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press, Cambridge (2013)
14. Khemchandani, R., Suresh, C.: Optimal kernel selection in twin support vector machines. Optim. Lett. 3(1), 77–88 (2009)
15. Montaes, D.C., Quiroz, A.J., Dulcerubio, M.D., Riascosvillegas, A.J.: Efficient nearest neighbors methods for support vector machines in high dimensional feature spaces. Optim. Lett. (2020). https://doi.org/10.1007/s11590-020-01616-w
16. Moreau, J.J.: Fonctions convexes duales et points proximaux dans un espace hilbertien. C.R.Acad.Sci.Paris. 255, 2897–2899 (1962)
17. Polyak, R., Ho, S.S., Griva, I.: Support vector machine via nonlinear rescaling method. Optim. Lett. 1(4), 367–378 (2007)
18. Rockafellar, R.T., Wets, R.J.: Variational analysis. Springer Science and Business Media (2009)
19. Shen, X.T., Tseng, G.C., Zhang, X.G., Wong, W.H.: On ψ-Learning. J. Am. Stat. Assoc. 98(1), 724–734 (2003)
20. Wang, X.M., Fan, N., Pardalos, P.M.: Stochastic subgradient descent method for large-scale robust chance-constrained support vector machines. Optim. Lett. 11(5), 1013–1024 (2017)
21. Wang, Z., Vucetic, S.: Fast online training of ramp loss support vector machines. In: Ninth IEEE International Conference on Data Mining, pp. 569–577 (2009)
22. Wu, Y., Liu, Y.: Robust truncated hinge loss support vector machines. J. Am. Stat. Assoc. 102, 974–983 (2007)
23. Xu, L., Crammer, K., Schuurmans, D.: Robust support vector machine training via convex outlier ablation. In: Proceedings of the 21st National Conference on Artificial Intelligence, pp. 536–542 (2006)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.