QUADRATIC CONVERGENCE OF SMOOTHING NEWTON’S METHOD FOR 0/1 LOSS OPTIMIZATION

SHENGLONG ZHOU†, LILI PAN‡, NAIHUA XIU§, AND HOU-DUO QI*

Abstract. It has been widely recognized that the 0/1-loss function is one of the most natural choices for modelling classification errors, and it has a wide range of applications including support vector machines and 1-bit compressed sensing. Due to the combinatorial nature of the 0/1-loss function, methods based on convex relaxations or smoothing approximations have dominated the existing research and are often able to provide approximate solutions of good quality. However, those methods are not optimizing the 0/1-loss function directly and hence no optimality has been established for the original problem. This paper aims to study the optimality conditions of the 0/1 function minimization, and for the first time to develop Newton’s method that directly optimizes the 0/1 function with a local quadratic convergence under reasonable conditions. Extensive numerical experiments demonstrate its superior performance as one would expect from Newton-type methods.

Key words. 0/1-loss function, optimality conditions, Newton’s method, locally quadratic convergence, superior numerical performance

AMS subject classifications. 49M05, 90C26, 90C30, 65K05

1. Introduction. This paper is concerned with the 0/1-loss optimization:

\[
\min_{x \in \mathbb{R}^n} f(x) + \lambda \| (Ax + b)_+ \|_0,
\]

where \( f : \mathbb{R}^n \to \mathbb{R} \) is twice continuously differentiable, \( \lambda > 0 \) is a penalty parameter and \( A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m \). Moreover, \( z_+ := ((z_1)_+, \ldots, (z_m)_+)^\top \) with \( z_+ := \max\{z, 0\} \) and \( \|z\|_0 \) is the \( \ell_0 \) norm of \( z \), counting the number of its non-zero entries. Hence, \( \|z_+\|_0 \) counts the number of positive entries of \( z \), i.e., \( \|z_+\|_0 = \sum_{i=1}^m \ell_{0/1}(z_i) \), where

\[
\ell_{0/1}(z) = \begin{cases} 
1, & z > 0, \\
0, & z \leq 0.
\end{cases}
\]

The function \( \ell_{0/1}(\cdot) \) is known as the Heaviside step function (or the unit step function) in [41, 13] or simply the 0/1-loss function in [17, 19, 6]. It plays an active role in many applications including support vector machines (SVM) [11], the one-bit compressed sensing [5], the maximum rank correlation [18], and the problem of area under curves [31]. However, optimization related to the 0/1-loss function is NP-hard, see [4, 16].

A vast body of work has developed algorithms for optimization involving the 0/1-loss function by making use of its continuous surrogates. A major concern on this part of research is that convergence analysis is often conducted on the surrogate problems rather than on the original ones that involve 0/1-loss functions. On the other hand, there also exists a large body of research that addresses the 0/1-loss optimization directly by taking advantage of the intrinsic appealing feature of the loss function, which captures the discrete nature of the binary classification. We will review two classes of such methods below.

The first class consists of mixed integer programming (MIP), which has become a leading approach to directly optimizing the 0/1-loss function, see [28, 1] for some earlier work. It is straightforward to relate the 0/1 loss to the misclassification minimization for discrimination problems [36, 7].
approach is in general effective with a major issue of scalability for large-sized problems. Much progress has also been made in improving the scalability of MIP by employing various strategies of reducing the problem sizes. Those include, for instance, decomposition strategy [36], local search [32], and convex hull cuts in a branch-and-bound framework [6], to name a few. Some recent work includes [38, 39], where different integer programming reformulations for the 0/1-loss minimizations are built and then tackled via the modern commercial MIP solvers, such as Gurobi and CPLEX. Despite those progresses, the speed of computation is still a bottleneck for the MIP approach.

The second step is to construct a scheme that defines $T_k$ at a given point $w^k$ and approximate the true $T$. The Newton step is to solve the equation $F(w; T_k) = 0$ for the next iterate $w^{k+1}$. Such a computational scheme for $T_k$ is described in (4.1) and (4.5). However, the difficulty is that there
is no guarantee that this scheme will be able to identify the correct \( T \). In other words, we may be encountered with different \( T_k \) each iteration no matter how close our iterate is to the optimal solution. This is where the convergence theory of classical Newton’s method fails to go through. Now we introduce a practically important technique of smoothing motivated by Chen et al. [10]. Instead of solving the equation \( F(w; T_k) = 0 \), we try to solve its perturbed version:

\[
F_{\mu_k}(w; T_k) := F(w; T_k) + \begin{bmatrix} 0 \\ -\mu_k z T_k \end{bmatrix} = 0,
\]

(1.4)

where the smoothing parameter \( \mu_k > 0 \) will be properly chosen as in (4.8). Its Jacobian matrix has the following structure

\[
\nabla F_{\mu_k}(w; T_k) := \begin{bmatrix} \nabla^2 f(x) & A^T_{T_k} \\ A_{T_k} & -\mu_k I \end{bmatrix}
\]

and it is nonsingular if and only if the matrix \( (\nabla^2 f(x) + A^T_{T_k} A_{T_k} / \mu_k) \) is nonsingular (i.e., the Schur complement of \((-\mu_k I)\) in the top 2 \( \times \) 2 block is nonsingular). We note that nonsingularity may still hold even if \( f(\cdot) \) is not convex provided that \( A_{T_k} \) has full row-rank. Due to its connection to [10], we call our method a smoothing Newton method. We also like to point out another interesting connection.

When \( \mu_k \) is fixed, our algorithmic framework is analogous to the primal-dual active-set algorithms extensively studied in [20, 24, 14], whose main targets are quadratic objective functions.

Our last step is to establish the bounds

\[
\norm{F(w^{k+1}; T_{k+1})} = C_1 \norm{F(w^k; T_k)}^2, \quad \norm{w^{k+1} - w^*} = C_2 \norm{w^k - w^*}^2,
\]

where the constants \( C_1 \) and \( C_2 \) only depend on the optimal solution \( w^* \). Those bounds imply the quadratic convergence of our smoothing Newton method provided that the initial point is close to \( w^* \) and complete our theoretical investigation, see Theorem 4.8. The efficiency of the Newton method is confirmed through extensive numerical experiments including 40 SVM problems from real data (16 of them for \( m \leq n \) and 24 for \( m > n \)) and simulated 1-bit compressing data against some existing solvers. As far as we know, this is the first Newton-type method for the 0/1-loss optimization (1.1).

This paper is organized as follows. In the next section, we analyze the 0/1-loss function, calculating its subdifferentials and the proximal operator. In Section 3, we establish the first-order necessary and sufficient optimality conditions of the problem (1.1) through the proximal operator of 0/1-loss function, leading to the well-defined \( P \)-stationary points. In Section 4, we reformulate the \( P \)-stationary condition as a system of nonlinear equations and develop Newton’s method with the promised quadratic convergence. In Section 5, we conduct extensive numerical experiments to demonstrate the outstanding performance of Newton’s method against a few leading solvers for the problems of SVM and 1-bit compressed sensing. We conclude the paper in Section 6.

2. Preliminaries. We first list some notation that is frequently used throughout the paper. Given a subset \( T \subseteq N_m := \{1, 2, \ldots, m\} \), its cardinality and complementary set are \( |T| \) and \( \overline{T} \). The neighbourhood of \( x \in \mathbb{R}^n \) with a radius \( \delta > 0 \) is denoted by \( N(x, \delta) = \{v \in \mathbb{R}^n : \norm{v - x} < \delta\} \). Moreover, \( x_T \) (resp. \( A_T \)) represents the sub-vector (resp. sub-matrix) contains elements (resp. rows) of \( x \) indexed on \( T \). Particularly, \( A_i \) is the \( i \)-th row of \( A \). We combine two vectors as \( (x; y) := (x^T y^T)^T \). The \( r \)-th largest singular value of \( H \in \mathbb{R}^{n \times n} \) is written \( \sigma_r(H) \), namely \( \sigma_1(H) \geq \sigma_2(H) \geq \cdots \geq \sigma_n(H) \). Particularly, we write \( \norm{H} := \sigma_1(H) \) and \( \sigma_{\min}(H) := \sigma_n(H) \). Finally, let \( I \) be the identity matrix and \( 1 \) be the vector with all entries being ones.

Next we describe the formula for computing the subdifferential of \( \norm{z_+}_0 \) and its proximal operator. We note that \( \norm{\cdot}_0 \) is lower semi-continuous (lsc) and \( z_+ \) is obviously continuous, the composition \( \norm{z_+}_0 \) is also lsc. For a proper and lsc function \( g : \mathbb{R}^m \to \mathbb{R} \), its subdifferential \( \partial g(\cdot) \) is well defined as in [35, Definition 8.3]. The following results are easy to prove after simple calculations.

**Lemma 2.1.**

\( (i) \) We have

\[
\partial \norm{y_+} = \left\{ v \in \mathbb{R}^m : \begin{array}{l} v_i \geq 0, \\ y_i = 0, \\ y_i \neq 0, \end{array} \right\}.
\]

(2.1)
(ii) Let \( g(x) := \|(h(x))_+\|_0 \), where \( h : \mathbb{R}^n \to \mathbb{R}^m \) is differentiable, and \( \Gamma := \{ i \in \mathbb{N}_m : h_i(x) = 0 \} \) for a given point \( x \in \mathbb{R}^n \). If

\[
\forall z \in \mathbb{R}^{|\Gamma|}, \ z \geq 0, \quad (\nabla h(x))_+^T z = 0 \quad \implies \quad z = 0,
\]

then the subdifferential of \( g(x) \) at \( x \) is

\[
\partial g(x) = \nabla h(x) \cdot \partial \|(h(x))_+\|_0.
\]

Proof. i) It is straight to check that the regular subdifferential \( \partial \|y_+\|_0 \) (see [35, Definition 8.3]) of \( \|y_+\|_0 \) at \( y \) takes the following form,

\[
\hat{\partial} \|y_+\|_0 = \left\{ v \in \mathbb{R}^m : v_i \left\{ \begin{array}{ll} \geq 0, & y_i = 0, \\ = 0, & y_i \neq 0, \end{array} \right. , i \in \mathbb{N}_m \right\} =: \Omega(y).
\]

We next verify \( \partial \|y_+\|_0 = \Omega(y) \). By letting \( \varphi(\cdot) := \|\cdot\|_0 \), we have

\[
\partial \|y_+\|_0 = \limsup_{z \rightharpoonup y} \hat{\partial} \|z_+\|_0 \quad \text{(by \{35, Equation 8(5)\})}
\]

\[
= \limsup_{z \rightharpoonup y} \Omega(z) \quad \text{(by (2.4))}
\]

\[
= \{ v \in \mathbb{R}^m : \exists z^k \rightharpoonup y, \ v^k \to v \text{ with } v^k \in \Omega(z^k) \} =: \Theta,
\]

where \( z \rightharpoonup y \) represents \( z \to y, \varphi(z) \to \varphi(y) \). Clearly, \( \Omega(y) \subseteq \Theta \). On the other hand, \( \Theta \subseteq \Omega(y) \) follows from that \( \Omega(z^k) \subseteq \Omega(y) \) for any \( z^k \rightharpoonup y \) and \( \Omega(\cdot) \) is closed.

ii) Direct verifications yield the following chain of equations,

\[
\partial^\infty \|y_+\|_0 = \limsup_{\sigma \downarrow 0, z \rightharpoonup y} \sigma \hat{\partial} \|z_+\|_0 \quad \text{(by \{35, Equation 8(5)\})}
\]

\[
= \limsup_{z \rightharpoonup y} \hat{\partial} \|z_+\|_0 \quad \text{(by (2.4))}
\]

\[
= \partial \|y_+\|_0, \quad \text{(by (2.5))}
\]

where \( \partial^\infty \|y_+\|_0 \) is the horizon subdifferential of \( \|y_+\|_0 \). Therefore, we derive that \( \hat{\partial} \varphi = \partial \varphi = \partial^\infty \varphi \). One can easily prove that the horizon cone \( \hat{\partial} \varphi(y) \) (see [35, Definition 3.3]) of \( \hat{\partial} \varphi(y) \) satisfies \( \hat{\partial} \varphi(y) \) = \( \partial^\infty \varphi(y) \). These conditions indicate that the function \( \varphi \) is regular by [35, Corollary 8.11], which together with (2.2), \( g(x) = \varphi(h(x)) \) and [35, Theorem 10.6] derives (2.3) immediately.

The assumption in (2.2) can be regarded as a constraint qualification for the chain rule in (2.3) to hold. We finish this section with a formula to compute the proximal operator of \( \|\cdot\|_0 \). Let \( \alpha > 0 \), the proximal operator of \( \alpha\|\cdot\|_0 \) at \( \nu \) is defined by

\[
\text{Prox}_{\alpha\|\cdot\|_0}(\nu) = \arg\min_{y \in \mathbb{R}^m} \frac{1}{2} \|y - \nu\|^2 + \alpha \|y_+\|_0.
\]

As shown in [40, Lemma 2.2], the proximal operator admits a closed form as

\[
\text{Prox}_{\alpha\|\cdot\|_0}(\nu)_i = \begin{cases} 0, & \nu_i \in (0, \sqrt{2\alpha}), \\ 0 \text{ or } \nu_i, & \nu_i \in [0, \sqrt{2\alpha}], \\ \nu_i, & \nu_i \in (-\infty, 0) \cup (\sqrt{2\alpha}, \infty). \end{cases}
\]

3. Optimality Conditions. In this section, we study the optimality conditions of (1.1) and characterize the conditions in terms of Proximal-stationarity (i.e., P-stationarity) using the proximal operator. Those results will lay down the foundation for Newton’s method in the next section. For a given point \( x^* \in \mathbb{R}^n \), we denote

\[
\Gamma_* := \{ i \in \mathbb{N}_m : A_i x^* + b_i = 0 \}.
\]

Our first result is to characterize a local minimizer of (1.1).
LEMMA 3.1. The following relationships hold for the problem (1.1).

i) A local minimizer $x^*$ satisfies the following condition if $A_{\Gamma_*}$ is full row rank,

$$ -\nabla f(x^*) \in A^T \partial \|(Ax^* + b)_+\|_0. $$

ii) A point $x^*$ satisfying (3.2) is a local minimizer if the function $f$ is locally convex around $x^*$.

Proof. i) It follows from [35, Theorem 10.1] that a local minimizer of (1.1) must satisfy $-\nabla f(x^*) \in \lambda \partial g(x^*)$, where $g(x) := \|(Ax + b)_+\|_0$. This together with Lemma 2.1 and $\lambda \partial \|\cdot\|_+ = \partial \|\cdot\|_0$ by (2.1) derives the result immediately.

ii) Since the problem (1.1) is equivalent to the following problem,

$$ \begin{align*}
\min_{x \in \mathbb{R}^n, y \in \mathbb{R}^m} & f(x) + \lambda \|y_+\|_0,
\text{s.t.} & \ Ax + b - y = 0,
\end{align*} $$

it suffices to show that $(x^*:y^*)$ is a local minimizer of the problem (3.3), where $x^*$ satisfies (3.2) and $y^* = Ax^* + b$, namely, there is a $z^*$ such that

$$ \nabla f(x^*) + A^T z^* = 0, \ Ax^* + b - y^* = 0, \ \partial \|y^*_+\|_0 \ni z^*. $$

It follows from $y^* = Ax^* + b$ and (3.1) that $\Gamma_* = \{ i \in \mathbb{N}_m : y^*_i = 0 \}$. This together with $\partial \|y^*_+\|_0 \ni z^*$ and the expression of the $\partial \|y^*_+\|_0$ in (2.1) indicates

$$ \begin{align*}
y^*_{\Gamma_*} = 0, \ z^*_{\Gamma_*} \geq 0, \ y^*_0 \neq 0, \ z^*_0 = 0.
\end{align*} $$

Define a radius $\delta := \min\{\delta_1, \delta_2\}$, where

$$ \delta_1 := \begin{cases} +\infty, & A_{\Gamma_*}^T z^*_{\Gamma_*} = 0, \\ \min_{\lambda \in \mathcal{A}_{\Gamma_*}} & \lambda \leq \lambda_{\Gamma_*}, \\ \text{otherwise}, \end{cases} \quad \delta_2 := \begin{cases} +\infty, & y^*_0 \leq 0, \\ \min_{i \in \mathcal{I}}\{y^*_i : y^*_i > 0\}, & \text{otherwise}, \end{cases} $$

and consider a local region of $w^* := (x^*:y^*)$ by

$$ N(w^*, \delta) = \{ (x:y) \in \mathbb{R}^n \times \mathbb{R}^m : Ax + b - y = 0, \ |w - w^*| < \delta \}. $$

Indeed, $N(w^*, \delta)$ is a neighbourhood of $w^*$ since $Ax^* + b - y^* = 0$ from (3.4). Next we show that, for any $w \in N(w^*, \delta),

$$ \|y^*_+\|_0 \leq \|y_+\|_0. $$

Obviously, it is true if $y^*_0 \leq 0$ as $\|y^*_+\|_0 = 0$. For $y^*_0 \not\leq 0$, to guarantee (3.8), it suffices to show that for any $i$, $y^*_i > 0 \implies y_i > 0$. Suppose there is a $j \in \mathbb{N}_m$ such that $y^*_j > 0$ but $y_j \leq 0$. This incurs the following contradiction

$$ \delta_2 \geq \delta > \|w - w^*\| \geq |y_j - y^*_j| = y^*_j - y_j \geq y^*_j \geq \delta_2. $$

Again, for any $w \in N(w^*, \delta)$, we have $Ax + b - y = 0$, which and (3.4) generate

$$ y - y^* = A(x - x^*). $$

Next, the convexity of $f$ gives rise to

$$ f(x) - f(x^*) \geq \langle \nabla f(x^*), x - x^* \rangle = -\langle A^T z^*, x - x^* \rangle = -\langle A_{\Gamma_*}^T z^*_{\Gamma_*}, x - x^* \rangle = -\langle A_{\Gamma_*}(x - x^*), z^*_{\Gamma_*} \rangle =: \phi $$

Now we make the conclusion by two cases. If $\|y^*_+\|_0 = \|y_+\|_0$, then $y^*_{\Gamma_*} \leq 0$ due to $y^*_{\Gamma_*} = 0$. This and (3.9) yield that

$$ 0 \geq y_{\Gamma_*} - y^*_{\Gamma_*} = A_{\Gamma_*}(x - x^*), $$
which together with $z^*_T \geq 0$ from (3.5) indicates $\phi \geq 0$, namely $f(x) \geq f(x^*)$. So

$$f(x) + \lambda \|y^+\|_0 \geq f(x^*) + \lambda \|y^+_*\|_0.$$ 

If $\|y^*_+\|_0 \neq \|y^*_+\|_0$, we must have $\|y^+_+\|_0 \geq 1 + \|y^*_+\|_0$ by (3.8). If $A^T_iz^*_T = 0$, then $\phi = 0 \geq -\lambda$. Otherwise, it follows

$$\phi \geq -\|A^T_i z^*_T\|_1 \|x - x^*\| \geq -\|A^T_i z^*_T\|_1 \|w - w^*\|$$

$$\geq -\|A^T_i z^*_T\| \delta \geq -\|A^T_i z^*_T\| \delta_1 = -\lambda. \quad \text{(by (3.6))}$$

Both lead to $\phi \geq -\lambda$, which results in

$$f(x) + \lambda \|y^+\|_0 \geq f(x^*) + \phi + \lambda \|y^+\|_0 \quad \text{(by (3.10))}$$

$$\geq f(x^*) - \lambda + \|y^+\|_0 \quad \text{(by $\phi \geq -\lambda$)}$$

$$\geq f(x^*) + \lambda \|y^*_+\|_0. \quad \text{(by $\|y^+_+\|_0 \geq 1 + \|y^*_+\|_0$)}$$

Overall, the two cases show that $(x^*; y^*)$ is a local minimizer to (3.3). Namely, $x^*$ is a local minimizer to (1.1).

The characterization (3.2) is nice and it is in the classic form of differential inclusion. However, the challenge is that it is difficult to extract second-order information which is essential to Newton’s method. To this purpose, we continue to characterize it in terms of P-stationarity.

**Definition 3.2.** A point $x^*$ is called a P-stationary point of the problem (1.1) if there exist a constant $\tau > 0$ and a point $z^* \in \mathbb{R}^n$ such that

$$\begin{cases}
\nabla f(x^*) + A^T z^* = 0 \\
\text{Prox}_{\tau \lambda \|c\|+\|b\|_0}(Ax^* + b + \tau z^*) \ni Ax^* + b.
\end{cases}$$

(3.12)

We also say a point $(x^*; z^*)$ is a P-stationary point of the problem (1.1) if it satisfies the conditions in (3.12). For a point $x^*$, we denote two constants by

$$\tau_1 := \begin{cases}
+\infty, & y^* \leq 0, \\
\min \left\{ \frac{(y^*_+)^2}{2}, y^*_+ > 0 \right\}, & \text{otherwise},
\end{cases} \quad \tau_2 := \begin{cases}
+\infty, & \Gamma_* = \emptyset, \\
\max_{i \in [n]} |p_i|^2, & \text{otherwise},
\end{cases}$$

(3.13)

where $y^* := Ax^* + b$ and $p^* := -(A^T_i A^T_*)^{-1} A^T_i \nabla f(x^*)$. Clearly, both $\tau_1 > 0$ and $\tau_2 > 0$. Based on these notation, we have the following main result of this section.

**Theorem 3.3.** The following relationships hold for the problem (1.1).

i) A local minimizer $x^*$ is a P-stationary point for any $0 < \tau < \tau_* := \min\{\tau_1, \tau_2\}$ if $A_{\tau_*}$ is full row rank.

ii) A P-stationary point with $\tau > 0$ is a local minimizer if the function $f$ is locally convex around $x^*$.

iii) A P-stationary point with $\tau \geq \|A\|^2/c_f$ is a global minimizer if the function $f$ is strongly convex with a constant $c_f > 0$.

*Proof.* i) As $x^*$ is a local minimizer of (1.1), condition (3.2) is valid by Lemma 3.1 if $A_{\tau_*}$ is full row rank. In other words, there is a $z^*$ such that

$$\nabla f(x^*) + A^T z^* = 0, \quad Ax^* + b - y^* = 0, \quad z^* \in \partial \|y^*_+\|_0.$$ 

Therefore, to show (3.12), we only need to verify that, for any $0 < \tau < \tau^*$,

$$z^* \in \partial \|y^*_+\|_0 = y^* \in \mathbb{P} := \text{Prox}_{\tau \lambda \|c\|+\|b\|_0}(y^* + \tau z^*).$$

Recall the definition of $\Gamma_*$ in (3.1) and the second condition in (3.14), we have $y_{\Gamma_*} = (Ax^* + b)_{\Gamma_*} = 0$. Same reasoning also allows for obtaining (3.5) due to $z^* \in \partial \|y^*_+\|_0$. As $A_{\tau_*}$ is full row rank, the first condition in (3.14) and $A^T z^* = A^T_i z^*_T$ derive that

$$z^*_T = p^*.$$
Now, $0 < \tau < \tau_* = \min\{\tau_1, \tau_2\}$ in (3.13) results in

\begin{equation}
\begin{aligned}
y^*_i \geq \min_{i: y^*_i > 0} y^*_i = \sqrt{2\tau_1} \lambda \geq \sqrt{2\tau_2} \lambda > \sqrt{2\tau} \lambda \quad \text{if } y^*_i > 0, \\
z^*_i \leq \max_i |p^*_i| = \sqrt{2\lambda/\tau_2} \leq \sqrt{2\lambda/\tau} < \sqrt{2\lambda/\tau} \quad \text{if } z^*_i > 0.
\end{aligned}
\end{equation}

These and (3.5) yield the following condition,

\[ y^*_i \begin{cases} = 0, & z^*_i = 0 \text{ or } 0 < z^*_i < \sqrt{2\lambda/\tau}, \\ < 0 \text{ or } > \sqrt{2\lambda/\tau}, & z^*_i = 0. \end{cases} \]

It is easy to see that the above condition satisfies that

\begin{equation}
\begin{aligned}
y^*_i \in \mathbb{P}_i = \begin{cases} 0, & y^*_i + \tau z^*_i \in (0, \sqrt{2\tau} \lambda), \\
0 \text{ or } y^*_i + \tau z^*_i \in \{0, \sqrt{2\tau} \lambda\}, \\
y^*_i + \tau z^*_i \in (-\infty, 0) \cup (\sqrt{2\tau} \lambda, \infty). \end{cases}
\end{aligned}
\end{equation}

ii) Note that the second condition in (3.12) means $y^* \in \mathbb{P}$, which by (3.16) implies $z^*_i = 0$ if $y^*_i \neq 0$ and $z^*_i \geq 0$ if $y^*_i = 0$, resulting in $z^* \in \partial y^*_+ \|y^*_+\|$ by (2.1). Consequently, we obtain (3.14) and (3.2). The claim follows from Lemma 3.1 ii).

iii) Let $(x^*; z^*)$ be a $P$-stationary point with $\tau \geq \|A\|^2/c_f$. Then the second condition in (3.12) indicates that

\[ (1/2\tau)\|Ax^* + b - (Ax^* + b + \tau z^*)\|^2 + \lambda\|(Ax^* + b)_+\|_0 \leq (1/2\tau)\|Ax + b - (Ax^* + b + \tau z^*)\|^2 + \lambda\|(Ax + b)_+\|_0 \]

for any $x \in \mathbb{R}^n$, which after simplifying leads to

\begin{equation}
\begin{aligned}
\lambda\|(Ax^* + b)_+\|_0 - \|A(x - x^*)\|^2/(2\tau) \\
\leq \lambda\|(Ax + b)_+\|_0 - (z^*, A(x - x^*)) \\
= \lambda\|(Ax + b)_+\|_0 - (A^T z^*, x - x^*) \\
= \lambda\|(Ax + b)_+\|_0 + \langle \nabla f(x^*), x - x^* \rangle. \quad \text{(by (3.12))}
\end{aligned}
\end{equation}

The strong convexity of $f$ implies

\begin{align*}
f(x) + \lambda\|(Ax + b)_+\|_0 - f(x^*) - \lambda\|(Ax^* + b)_+\|_0 \\
\geq (c_f/2)\|x - x^*\|^2 + \langle \nabla f(x^*), x - x^* \rangle + \lambda\|(Ax + b)_+\|_0 - \lambda\|(Ax^* + b)_+\|_0 \\
\geq (c_f/2)\|x - x^*\|^2 - 1/(2\tau)\|A(x - x^*)\|^2 \quad \text{(by (3.17))} \\
\geq (c_f/2 - \|A\|^2/(2\tau))\|x - x^*\|^2 \geq 0. \quad \text{(by } \tau \geq \|A\|^2/c_f). \end{align*}

This shows the global optimality of $x^*$ to the problem (1.1).

4. Smoothing Newton’s Method. The main purpose of this section is to formulate Newton’s method and establish its quadratic convergence. We first state two assumptions for this purpose.

Assumption 4.1. Suppose $f$ is twice continuously differentiable, $\nabla^2 f(x^*)$ is positive definite and $A_{\Gamma_*}$ is full row rank, where $\Gamma_*$ is given by (3.1).

Assumption 4.2. Suppose $\nabla^2 f$ is locally Lipschitz continuous around $x^*$ with a constant $L_*> 0$, namely

\[ \|\nabla^2 f(x) - \nabla^2 f(x')\| \leq L_*\|x - x'\| \]

for any $x$ and $x'$ in the neighbourhood of $x^*$.
4.1. Stationary equations. For a point \( w := (x; z) \), we define the sets

\[
\begin{align*}
S &= \left\{ i \in \mathbb{N}_m : A_i x + b_i + \tau z_i \in (0, \sqrt{2T\lambda}) \right\}, \\
E &= \left\{ i \in \mathbb{N}_m : A_i x + b_i + \tau z_i \in \{0, \sqrt{2T\lambda}\} \right\}, \\
O &= \left\{ i \in \mathbb{N}_m : A_i x + b_i + \tau z_i \in (-\infty, 0) \cup (\sqrt{2T\lambda}, \infty) \right\}, \\
E^o &= \left\{ i \in \mathbb{N}_m : A_i x + b_i = 0, \tau z_i \in [0, \sqrt{2T\lambda}] \right\},
\end{align*}
\]

(4.1)

for a given \( \tau > 0 \). Obviously, \( E^o \subseteq E \). It is worth mentioning that all sets depend on \( w \). For simplicity, we drop their dependence whenever there is no confusion to be caused. Same rules are also applied into \( S, E, O \) and \( E^o \) for \( w^* := (x^*; z^*) \). A key step towards the Newton method is the construction of the following system of equations. For a given subset \( \Gamma \subseteq \mathbb{N}_m \) and a scalar \( \mu \geq 0 \), it follows the definitions in (1.3) and (1.4) that

\[
(2.6) \quad \nabla F(\mu; \Gamma) = \begin{bmatrix} \nabla^2 f(x) + A_i^T z_i \end{bmatrix} = 0 \quad \nabla F_{\mu}(w; \Gamma) = \begin{bmatrix} \nabla^2 f(x) & A_i^T \\ A_i & -\mu I \\ 0 & 0 & I \end{bmatrix}.
\]

We note that the matrix \( \nabla F_{\mu} \) is a slight perturbation of the Jacobian matrix of \( F \) and \( \nabla F(w; \Gamma) = \nabla F_0(w; \Gamma) \). The following result relates a P-stationary point to a system of equations.

**Theorem 4.3.** A point \( w^* = (x^*; z^*) \) is a P-stationary point with \( \tau > 0 \) of the problem (1.1) if and only if \( F(w^*; \Gamma^*) = 0 \) and \( \Gamma^* = (S \cup E^o) \), where \( \Gamma^* \) is defined by (3.1). The Jacobian \( \nabla F(w^*; \Gamma^*) \) is non-singular if Assumption 4.1 holds.

**Proof.** The second claim is obvious due to Assumption 4.1. We only need to prove the first claim. We start with the sufficiency. The definitions in (4.1) show the following relationships,

\[
\Gamma^* = S \cup E^o, \quad \Gamma_o = O \cup (E \setminus E^o).
\]

(4.3)

We recall \( y^* = Ax^* + b \) and \( P := \text{Prox}_{\tau\|\cdot\|_0}(y^* + \tau z^*) \). It follows that

\[
P_i = \begin{cases} 
0, & i \in S, \\
0 \text{ or } y_i^* + \tau z_i^*, & i \in E, \\
y_i^* + \tau z_i^*, & i \in O, \\
\end{cases}
\]

where the first equation is by (2.6) and second one is by (4.2) and (4.3), which indicates \( y_i^* \in P_i \) due to \( y_i^* \neq 0 \) from (4.2). Moreover, the first and third equations in (4.2) suffice to \( \nabla f(x^*) + A^T z^* = 0 \), showing (3.12). Namely, \( w^* \) is a P-stationary point.

Necessity. Let \( w^* \) be a P-stationary point satisfying (3.12) and \( T^* := S \cup E^o \). Then \( T^* = O \cup (E \setminus E^o) \). It follows from (4.1), (4.2) that

\[
y_i^* \in \begin{cases} 
0, & i \in S, \\
0 \text{ or } y_i^* + \tau z_i^*, & i \in E, \\
y_i^* + \tau z_i^*, & i \in O, \\
\end{cases}
\]

where the equality is by the definition of \( E^o \), which together with \( y_i^* = 0, i \in E^o \) and \( y_i^* \neq 0, i \in E \setminus E^o \) suffices to

\[
\begin{cases} 
y_i^* = 0, & i \in S \cup E^o = T^*, \\
y_i^* = y_i^* + \tau z_i^*, & i \in E \setminus E^o, \\
y_i^* = y_i^* + \tau z_i^*, & i \in O, \\
\end{cases}
\]

This gives rise to the last two conditions \( y_i^* = 0, z_i^* = 0 \) in (4.2). Furthermore, the first condition in (3.12) and \( z_i^* = 0 \) derive the first condition of (4.2). Overall, we have \( F(w^*; T^*) = 0 \). Now we
show \( T_* = \Gamma_* \). By (3.1) that \( \Gamma_* = \{ i \in N_m : y_i^* = 0 \} \), it follows \( T_* \subseteq \Gamma_* \). Suppose, there is a \( j \in \Gamma_* \) but \( j \notin T_* \), then we have \( y_j^* = z_j^* = 0 \) and thus \( j \in E_o^* \subseteq T_* \) by (4.1), a contradiction. Therefore, \( T_* = \Gamma_* \), finishing the proof.

Remark 4.4. It is interesting to note that Theorem 4.3 suggests a threshold value for \( \lambda \) to exclude the zero solutions when \( b_i \neq 0, \forall i \in N_m \). Suppose \( x^* = 0 \) is a P-stationary point. The second equation \( \Gamma_*, x + b_{\Gamma_*} = 0 \) in (1.2). This implies that the Hessian matrix \( \nabla c \) has diagonal blocks such as SVM and 1-bit CS, their functions are separable and have block structures such as

This manuscript is for review purposes only.

4.2. Algorithmic design. Theorem 4.3 lays the foundation for developing Newton’s method, which is to solve the stationarity equation in (4.2). Let \( w^k := (x^k; z^k) \) be the current iterate. We define \( S_k \) and \( E_k \) by (4.1) with \( w \) being replaced by \( w^k \) and let

\[
T_k := S_k \cup E^o_k.
\]

Let \( d^k = (u^k; v^k) \) with \( u^k \in \mathbb{R}^n \) and \( v^k \in \mathbb{R}^m \). For such a defined \( T_k \), a Newton direction \( d^k \) for the equation (4.2) solves the following linear equations:

\[
\nabla F(w^k; T_k) d = -F(w^k; T_k).
\]

To improve the nonsingularity of the Jacobian matrix \( \nabla F(w^k; T_k) \), we replace it with \( \nabla F_{\mu_k}(w^k; T_k) \). That is, at \( w^k \), we solve the equation:

\[
\nabla F_{\mu_k}(w^k; T_k) d = -F(w^k; T_k),
\]

where \( \nabla F_{\mu_k}(w^k; T_k) \) is defined in (4.2). The Newton direction \( d^k \) satisfies

\[
\nabla^2 f(x^k) A_{T_k}^T 0 \\
A_{T_k} -\mu_k I 0 \\
0 0 I
\]

\[
= -
\begin{bmatrix}
\nabla f(x^k) + A_{T_k}^T z^k_{T_k} \\
A_{T_k} x^k + b_{T_k} \\
z^k_{T_k}
\end{bmatrix}
\]

Here, the rule to update \( \mu_k \) is as follows:

\[
\mu_k = \min \{ \alpha \mu_{k-1}, \rho \| F(w^k; T_k) \| \},
\]

where \( \alpha \in (0, 1) \) and \( \rho > 0 \). Now we summarize the proposed method in Algorithm 4.1.

Algorithm 4.1 MMO1: Newton’s method for 0/1-loss optimization

1. Initialize \( w^0 = (x^0; z^0) \) and \( \mu_{-1} > 0 \). Set the parameter \( \tau, \lambda, \rho > 0, \alpha \in (0, 1) \).

   Compute \( T_0 \) by (4.5) and set \( k := 0 \).

2. if \( \| F(w^k; T_k) \| > 0 \) then

3. Update \( \mu_k \) by (4.8).

4. Update \( d^k \) by solving (4.7).

5. Update \( w^{k+1} = w^k + d^k \).

6. Update \( T_{k+1} \) by (4.5) and set \( k := k + 1 \).

7. end if

8. return \( w^k \).

Remark 4.5. In general, the computational complexity for solving the equation (4.7) is approximately \( O(n^2 \max \{n, |T_k| \}) \). This is fine for small-sized problems. When \( n \) is large, the computational cost is too high and existing first-order algorithms would be faster. Fortunately, for many real applications, such as SVM and 1-bit CS, their functions are separable and have block structures such as in (1.2). This implies that the Hessian matrix \( \nabla^2 f(x^k) \) is of diagonal blocks and is invertible. The worst-case computational complexity can be reduced to \( O(|T_k|^2 \max \{n, |T_k| \}) \). For SVM or 1-bit CS problems, \( T_k \) coincides with the indices of incorrectly classified samples that take a relatively small portion of the total samples. Hence, \( |T_k| \) can be on a small scale and computation of the Newton direction can be very cheap.
4.3. Quadratic convergence. Let us first explain why it is a challenging task to establish the quadratic convergence of the proposed Newton method. Suppose \( \mathbf{w}^* \) satisfies the stationarity equation \( F(\mathbf{w}^*; \Gamma_*) = 0 \) (see Theorem 4.3). If we know \( \Gamma_* \) beforehand, then by fixing \( T_k = \Gamma_* \), our proposed method reduces to the standard Newton’s method that solves equations with smooth functions. The quadratic convergence follows under Assumption 4.1 and Assumption 4.2. However, the difficulty we are facing is that the set \( T_k \) may change from iteration to iteration. A different \( T_k \) leads to a different system of equations instead of a fixed system. This is where the standard proof for quadratic convergence breaks down. As we will see below, it takes a great deal of effort in establishing quadratic convergence.

The first technical result is about extending the stationarity equation to some indices that are given in a neighborhood of \( \mathbf{w}^* \). In the proof, we recall \( \mathbf{y} = A\mathbf{x} + \mathbf{b} \) and \( \mathbf{y}^* = A\mathbf{x}^* + \mathbf{b} \).

**Lemma 4.6.** Let \( \mathbf{w}^* \) be a P-stationary point with \( 0 < \tau < \tau_* := \min\{\tau_1, \tau_2\} \) of the problem (1.1), \( \tau_1, \tau_2 \) and \( \Gamma_* \) be given by (3.13) and (3.1). Then there is a \( \delta^*_1 > 0 \) such that, for any \( \mathbf{w} \in N(\mathbf{w}^*, \delta^*_1) \) with its associated indices \( \mathcal{S} \) and \( \mathcal{E} \), it holds

\[
F(\mathbf{w}^*; T) = 0 \quad \text{and} \quad T := (\mathcal{S} \cup \mathcal{E}) \subseteq \Gamma_*.
\]

**Proof.** i) Theorem 4.3 states that the P-stationary point \( \mathbf{w}^* \) of (1.1) satisfies

\[
\nabla f(x^*) + A_{\Gamma_*} \mathbf{z}_{\Gamma_*}^* = 0, \quad \mathbf{y}_{\Gamma_*}^* = 0, \quad \mathbf{z}_{\Gamma_*}^* = 0
\]  

for \( 0 < \tau < \tau_* \), where \( \Gamma_* = \mathcal{S} \cup \mathcal{E} \). Note that \( \mathcal{E} \setminus \mathcal{E}^* \subseteq \Gamma_* \) which by (4.10) leads to

\[
\mathbf{z}_{\mathcal{E} \setminus \mathcal{E}^*}^* = 0.
\]

Using the same reasoning for proving (3.15), we can prove for \( 0 < \tau < \tau_* = \min\{\tau_1, \tau_2\} \) in (3.13) that

\[
y_i^* > \sqrt{2\tau_1} \quad \text{if} \quad y_i^* > 0, \quad \tau z_i^* < \sqrt{2\tau_1} \quad \text{if} \quad z_i^* > 0.
\]

Therefore, we have the following facts

\[
\mathcal{E}^*_0 = \{i \in \mathbb{N}_m : y_i^* = 0, \tau z_i^* \notin \{0, \sqrt{2\tau_1}\}\} \quad \text{by (4.1)}
\]

\[
\mathcal{E} = \mathcal{S} \cup \mathcal{E}^*_0 \cup \mathcal{E} \supseteq \mathcal{S} \cup \mathcal{E}_*, \quad \Gamma_* = \mathcal{O}.
\]

For a sufficiently small \( \delta^*_1 \), any \( \mathbf{w} \in N(\mathbf{w}^*, \delta^*_1) \) satisfies

\[
|y_i + \tau z_i - y_i^* - \tau z_i^*| \leq c \delta^*_1, \quad \forall \ i \in \mathbb{N}_m,
\]

where \( c > 0 \) is a constant relied on \( A \) and \( \tau \). The definitions of \( \mathcal{S} \) and \( \mathcal{E}_* \) in (4.1) mean that, for any \( i \in \mathcal{S} \) or \( i \in \mathcal{S}_* \),

\[
y_i + \tau z_i \in (0, \sqrt{2\tau_1}) \iff |y_i + \tau z_i - \sqrt{2\tau_1}| < \sqrt{2\tau_1},
\]

\[
y_i^* + \tau z_i^* \in (0, \sqrt{2\tau_1}) \iff |y_i^* + \tau z_i^* - \sqrt{2\tau_1}| < \sqrt{2\tau_1}.
\]
Using this fact, if $S_* \notin S$, then there is an $i \in S_*$ but $i \notin S$ such that

\[
|y_i + \tau z_i - y_i^* - \tau z_i^*| = |y_i + \tau z_i - \sqrt{\tau \lambda / 2} - y_i^* - \sqrt{\tau \lambda / 2}| \\
\geq \sqrt{\tau \lambda / 2} - |y_i^* + \tau z_i^* - \sqrt{\tau \lambda / 2}| \quad \text{(by $i \notin S$ and (4.16))} \\
\geq \sqrt{\tau \lambda / 2} - \max_i |y_i^* + \tau z_i^* - \sqrt{\tau \lambda / 2}| =: \delta_s \\
> 0, \quad \text{(by (4.16))}
\]

Since $c\delta_i$ can be smaller than $\delta_s$, the above fact contradicts with (4.15). Hence, it holds $S_* \subseteq S$. Similar reasoning also derives $O_* \subseteq O$. These allow us to obtain

\[
\mathcal{E} = \mathbb{N}_m \setminus (S \cup O) \subseteq \mathbb{N}_m \setminus (S_* \cup O_*) = \mathcal{E}_*.
\]

Overall, for any $w \in \mathcal{N}(w^*, \delta_1^{\tau})$, it holds

\[
(4.17) \quad S_* \subseteq S, \quad O_* \subseteq O, \quad \mathcal{E} \subseteq \mathcal{E}_*.
\]

The above relations enable us to claim that

\[
(4.18) \quad (S \setminus S_*) \subseteq \mathcal{E}_*, \quad (O \setminus O_*) \subseteq \mathcal{E}_*.
\]

Now, we can show

\[
\begin{aligned}
y_S^* &= 0, \quad \text{(by $S_* \subseteq \Gamma_*$ and (4.10))} \\
y_{S \setminus S_*}^* &= 0, \quad \text{(by $S \setminus S_* \subseteq \mathcal{E}_*$ from (4.18) and (4.13))} \\
y_{\mathcal{E}_*}^* &= 0, \quad \text{(by $\mathcal{E}_* \subseteq \mathcal{E}_*$ from (4.17) and (4.13))} \\
z_{\mathcal{O}_*}^* &= 0, \quad \text{(by $\mathcal{O}_* \subseteq \mathcal{T}_*$ and (4.10))} \\
z_{\mathcal{O} \setminus \mathcal{O}_*}^* &= 0, \quad \text{(by $\mathcal{O} \setminus \mathcal{O}_* \subseteq \mathcal{E}_*$ from (4.18) and (4.13))} \\
z_{\mathcal{E} \setminus \mathcal{E}_*}^* &= 0, \quad \text{(by $\mathcal{E} \setminus \mathcal{E}_* \subseteq \mathcal{E}_*$ from (4.17) and (4.13))}
\end{aligned}
\]

These conditions combining with

\[
T = S \cup \mathcal{E}_* = S_* \cup (S \setminus S_*) \cup \mathcal{E}_* \\
\mathcal{T} = O \cup (\mathcal{E} \setminus \mathcal{E}_*) = O_* \cup (O \setminus O_*) \cup (\mathcal{E} \setminus \mathcal{E}_*)
\]

imply $y_T^* = 0$ and $z_T^* = 0$. As a consequence of this and $z_T^* = 0$ from (4.10),

\[
0 = f(x^*) + A_T^* z_T^* = f(x^*) + A_T^* y_T^* = f(x^*) + A_T^* y_T^*.
\]

Overall, we verify $F(w^*; T) = 0$, as desired. Finally, we observe that

\[
T = S \cup \mathcal{E}_* = S_* \cup (S \setminus S_*) \cup \mathcal{E}_* \\
\subseteq S_* \cup \mathcal{E}_* \cup \mathcal{E}_* \quad \text{(by $S \setminus S_* \subseteq \mathcal{E}_*$ from (4.18))} \\
\subseteq S_* \cup \mathcal{E}_* \cup \mathcal{E}_* \quad \text{(by $\mathcal{E}_* \subseteq \mathcal{E}_*$ from (4.17))} \\
= \Gamma_* \quad \text{(by (4.14))}
\]

The whole proof is completed. \hfill \Box

The second technical result is about the uniform nonsingularity of the perturbed Jacobian matrix $\nabla F_\mu(w; T)$ over a neighborhood of $w^*$.

**Lemma 4.7.** Let $w^*$ be a P-stationary point with $0 < \tau < \tau_*$ of the problem (1.1) and $\tau_*$ be given by (3.13). Assume Assumption 4.1 and Assumption 4.2. It holds

\[
(4.19) \quad C_* \geq \|\nabla F_\mu(w; T)\| \geq \sigma_{\text{min}}(\nabla F_\mu(w; T)) \geq c_* > 0,
\]
for any \( w \in N(w^*, \delta_2^*) \) and any \( 0 \leq \mu \leq c_s/2 \), where \( T = S \cup E^o \) and

\[
(4.20) \quad C_s := 2 \max\{1, \|H(\Gamma_*)\|\}, \quad c_s := 0.5 \min\{1, \min_{\Gamma \subseteq \Gamma_*} \sigma_{\min}(H(\Gamma))\},
\]

\[
(4.21) \quad \delta_2^* := \min\left\{ \delta_1^*, \frac{c_s}{2}L_s \right\}, \quad H(\Gamma) := \begin{bmatrix} \nabla^2 f(x^*) & A_\Gamma^T \\ A_\Gamma & 0 \end{bmatrix}.
\]

**Proof.** Since \( \nabla^2 f(x^*) \) is positive definite and \( A_\Gamma \) is full row rank by Assumption 4.1, \( A_\Gamma \) is full row rank for any \( \Gamma \subseteq \Gamma_* \) and thus \( H(\Gamma) \) is non-singular. We have \( \sigma_{\min}(H(\Gamma)) > 0 \) for any \( \Gamma \subseteq \Gamma_* \) and \( c_s > 0 \). Now we build the bounds of \( \nabla F_\mu(w; T) \). For any given two matrices \( D' \) and \( D \), we have the first fact

\[
\|D' - D\| \geq \max \{ |\sigma_i(D') - \sigma_i(D)| \} \geq |\sigma_{i_0}(D') - \sigma_{i_0}(D)|
\]

\[
(4.22) \quad \geq \sigma_{i_0}(D') - \sigma_{\min}(D) \geq \sigma_{\min}(D') - \sigma_{\min}(D),
\]

where the first inequality is from [30, Reminder (2), on Page 76] and \( i_0 \) satisfies that \( \sigma_{i_0}(D) = \sigma_{\min}(D) \). Recall that

\[
H(\Gamma) = \begin{bmatrix} \nabla^2 f(x^*) & A_\Gamma^T \\ A_\Gamma & 0 \end{bmatrix}, \quad \nabla F(w^*; T) = \begin{bmatrix} H(T) & 0 \\ 0 & I \end{bmatrix}.
\]

For any \( w \in N(w^*, \delta_2^*) \) and \( \delta_2^* \leq \delta_1^* \), Lemma 4.6 contributes to \( T \subseteq \Gamma_* \). So \( H(T) \) is a submatrix of \( H(\Gamma_*) \). Hence,

\[
\|H(T)\| \leq \|H(\Gamma_*)\|, \quad \sigma_{\min}(H(T)) \geq \min_{\Gamma \subseteq \Gamma_*} \sigma_{\min}(H(\Gamma)),
\]

where the latter is by \( T \subseteq \Gamma_* \), which gives us the second fact

\[
(4.23) \quad \sigma_{\min}(\nabla F(w^*; T)) = \min\{1, \sigma_{\min}(H(T))\}
\]

\[
(4.24) \quad \geq \min\{1, \min_{\Gamma \subseteq \Gamma_*} \sigma_{\min}(H(\Gamma))\} = 2c_s,
\]

\[
(4.25) \quad \|\nabla F(w^*; T)\| = \max\{1, \|H(T)\|\} \leq \max\{1, \|H(\Gamma_*)\|\} = C_s/2.
\]

The locally Lipschitz continuity of \( \nabla^2 f \) around \( x^* \) with \( L_s \) yields the third fact,

\[
(4.26) \quad \|\nabla F^2(w^*; T) - \nabla F^2(w; T)\| \leq L_s \|x - x^*\| \leq L_s \|w - w^*\| \leq L_s \delta_2^* \leq c_s/2. \quad (\text{by } (4.21))
\]

Now these three facts allow us to derive

\[
\sigma_{\min}(\nabla F_\mu(w; T)) \geq \sigma_{\min}(\nabla F(w; T)) - \|\nabla F(w; T) - \nabla F_\mu(w; T)\| \quad (\text{by } (4.22))
\]

\[
\geq \sigma_{\min}(\nabla F_\mu(w; T)) - \mu \quad (\text{by } (4.2))
\]

\[
\geq \sigma_{\min}(\nabla F(w^*; T)) - c_s/2 - \mu \quad (\text{by } (4.25))
\]

\[
\geq \sigma_{\min}(\nabla F(w^*; T)) - c_s \quad (\text{by } (4.22))
\]

\[
\geq c_s. \quad (\text{by } (4.23))
\]
Similarly, we also have
\[
\|\nabla F_{\mu}(w;T)\| \\
\leq \|\nabla F(w;T)\| + \|\nabla F(w;T) - \nabla F_{\mu}(w;T)\| \\
= \|\nabla F(w;T)\| + \mu \\
\leq \|\nabla F(w^*;T)\| + \|\nabla F(w^*;T) - \nabla F(w;T)\| + \mu \\
\leq \|\nabla F(w^*;T)\| + c_s/2 + \mu \\
\leq C_s + c_s \\
\leq C_s/2 + c_s \\
\leq C_s. 
\] (4.27)

The whole proof is completed. \(\square\)

Now we are ready to claim the following local quadratic convergence.

**Theorem 4.8.** Let \(w^*\) be any \(P\)-stationary point with \(0 < \tau < \tau_0\) of (1.1), \(\tau_s\) and \(\delta_s, c_s, C_s\) be given by (3.13) and (4.20). Assume Assumption 4.1 and Assumption 4.2. Let \(\{w^k\}\) be the sequence generated by Algorithm 4.1 and \(0 \leq \mu_{-1} \leq c_s/2\). If the initial point satisfies \(w^0 \in N(w^*, \delta_s)\), where
\[
\delta_s := \min\{\delta^2_s, c_s/(2(L_s + 2\rho C_s))\},
\] (4.28)
then the following results hold.

a) The sequence \(\{d^k\}_{k \geq 0}\) is well defined and \(\lim_{k \to \infty} d^k = 0\).

b) The whole sequence \(\{w^k\}\) converges to \(w^*\) quadratically, namely,
\[
\|w^{k+1} - w^*\| \leq (\|L_s + 2\rho C_s\|/c_s)\|w^k - w^*\|^2.
\]

c) The halting condition satisfies
\[
\|F(w^{k+1}; T_{k+1})\| \leq (\|L_s + 2\rho C_s\|C_s/c_s^3)\|F(w^k; T_k)\|^2
\]
and Algorithm 4.1 reaches \(\|F(w^k; T_k)\| < \epsilon\) for a given tolerance \(\epsilon > 0\) when
\[
k \geq \left\lceil \log_2 \left(2 \sqrt{(\|L_s + 2\rho C_s\|C_s/c_s^3)\|w^0 - w^*\|} - \log_2(\sqrt{\epsilon}) \right) \right\rceil.
\] (4.29)

**Proof.** a) It is easily observed
\[
0 < \mu_k \leq \mu_{k-1}, \quad k = 1, 2, 3, \ldots, \quad \lim_{k \to \infty} \mu_k = 0.
\] (4.30)

It follows from Lemma 4.6 and the facts \(\delta_s \leq \delta^2_s \leq \delta^2_s\) and \(w^0 \in N(w^*, \delta_s)\) that
\[
F(w^*; T_0) = 0
\] (4.31)
with \(T_0 = S_0 \cup E_0\) and from Lemma 4.7 that
\[
C_s \geq \|\nabla F_{\mu_0}(w^0; T_0)\| \geq \sigma_{\min}(\nabla F_{\mu_0}(w^0; T_0)) \geq c_s.
\] (4.32)

Here, we used the fact that \(\mu_0 \leq \mu_{-1} \leq c_s/2\) by (4.30). From (4.6), we have
\[
\nabla F_{\mu_0}(w^0; T_0) \cdot d^0 = -F(w^0; T_0).
\] (4.33)

Lemma 4.7 states that \(\nabla F_{\mu_0}(w^0; T_0)\) is non-singular and thus \(d^0\) is well defined. Let
\[
w^0_\beta = w^* + \beta(w^0 - w^*) = (x^0_\beta; z^0_\beta)
\] (4.34)
where \(\beta \in [0, 1]\). One can easily check that \(w^0_\beta \in N(w^*, \delta_s)\) as
\[
\|w^0_\beta - w^*\| = \beta\|w^0 - w^*\| \leq \delta_s.
\]
The definition in (4.2) enables us to obtain
\[
\|\nabla F_{\mu_0}(w^0; T_0) - \nabla F(w^0_{\beta}; T_0)\| \\
\leq \|\nabla^2 f(x^0) - \nabla^2 f(x^0_{\beta})\| + \mu_0 \quad \text{(by (4.2))}
\]
\[
\leq L_* \|x^0 - x^0_{\beta}\| + \rho \|F(w^0; T_0)\| \quad \text{(by (4.8))}
\]
\[
\leq L_* \|w^0 - w^0_{\beta}\| + \rho \|\nabla F_{\mu_0}(w^0; T_0)\||d^0|| \\
\leq L_* \|w^0 - w^0_{\beta}\| + \rho \left( \|F(w^0; T_0)\| + \|d^0\| \right) \quad \text{(by (4.6))}
\]
\[
\leq L_* (1 - \beta) \|w^0 - w^*\| + \rho (C_\gamma \|w^1 - w^*\| + \|w^0 - w^*\|) \quad \text{(by (4.32))}
\]
\[
= (L_* (1 - \beta) + \rho C_\gamma) \|w^0 - w^*\| + \rho C_\gamma \|w^1 - w^*\|.
\]

Denote \( \Theta_{\mu_0} := \nabla F_{\mu_0}(w^0; T_0) \) and \( \Theta(\beta) := \nabla F(w^0_{\beta}; T_0) \). The above condition yields
\[
\int_0^1 \|\Theta_{\mu_0} - \Theta(\beta)\|d\beta \leq (L_* / 2 + \rho C_\gamma) \|w^0 - w^*\| + \rho C_\gamma \|w^1 - w^*\|. \tag{4.35}
\]

For the fixed \( T_0 \), the function \( F(\cdot, T_0) \) is differentiable, which by (4.31) derives
\[
F(w^0; T_0) = F(w^*; T_0) + \int_0^1 \Theta(\beta)(w^0 - w^*)d\beta = \int_0^1 \Theta(\beta)(w^0 - w^*)d\beta. \tag{4.40}
\]

Now the following chain of inequalities holds.
\[
c_* \|w^1 - w^*\| = c_* \|w^0 + d^0 - w^*\|
\]
\[
= c_* \|w^0 - w^* - \Theta_{\mu_0}^{-1}F(w^0; T_0)\| \quad \text{(by (4.33))}
\]
\[
\leq \|\Theta_{\mu_0}(w^0 - w^*) - F(w^0; T_0)\| \quad \text{(by (4.32))}
\]
\[
= \|\Theta_{\mu_0}(w^0 - w^*) - \int_0^1 \Theta(\beta)(w^0 - w^*)d\beta\| \quad \text{(by (4.36))}
\]
\[
\leq \int_0^1 \|\Theta_{\mu_0} - \Theta(\beta)\| \|w^0 - w^*\|d\beta
\]
\[
= (\theta_*/2) \|w^0 - w^*\|^2 + \rho C_\gamma \|w^1 - w^*\| \|w^0 - w^*\| \quad \text{(by (4.35))}
\]
\[
\leq (\theta_*/2) \|w^0 - w^*\|^2 + \rho C_\gamma \|w^1 - w^*\| \tag{4.37}
\]
\[
\leq (\theta_*/2) \|w^0 - w^*\|^2 + (c_*/2) \|w^1 - w^*\|,
\]

where \( \theta_* := L_* + 2 \rho C_\gamma \) and the last inequality is from \( \delta_* \leq c_*/(2 \theta_*) \) by (4.28) and
\[
\|w^0 - w^*\| < \delta_* \leq c_*/(2 \theta_*) < c_*/(2 \rho C_\gamma). \tag{4.37}
\]

The above chain of inequalities suffices to the following fact
\[
\|w^1 - w^*\| \leq (\theta_* / c_*) \|w^0 - w^*\|^2. \tag{4.38}
\]

This together with \( \|w^0 - w^*\| < \delta_* \) and (4.37) derives
\[
\|w^1 - w^*\| \leq (\theta_* / c_*) \delta_* \|w^0 - w^*\| \leq (1/2) \|w^0 - w^*\| < \delta_*,
\]

which means \( w^1 \in N(w^*, \delta_*) \). In addition, \( \mu_1 \leq \mu_0 \leq c_*/2 \) by (4.30). Hence, replacing \( T_0 \) by \( T_1 \), the same reasoning allows us to show that \( d^1 \) is well defined and
\[
\|w^2 - w^*\| \leq (\theta_* / c_*) \|w^1 - w^*\|^2.
\]

By the induction, we can conclude that \( w^k \in N(w^*, \delta_*) \), \( d^k \) is well defined and
\[
\|w^{k+1} - w^*\| \leq (\theta_* / c_*) \|w^k - w^*\|^2, \tag{4.39}
\]
\[
\leq (\theta_* / c_*) \|w^k - w^*\| \delta_* \leq (1/2) \|w^k - w^*\|. \tag{4.37}
\]
Therefore, (4.39) claims b). The conclusion of a) can be made by (4.40) that
\[ w^k \to w^*, \quad \| d^k = w^{k+1} - w^k = w^{k+1} - w^* + w^* - w^k \to 0. \]

\[ (4.41) \]

\[ F(w^*; T_k) = 0, \]

where \( T_k = S_k \cup \mathcal{E}^c_k \). By letting \( w^k_{\beta} = w^* + \beta (w^k - w^*) \), where \( \beta \in [0, 1] \), we have \( w^k_{\beta} \in N(w^*, \delta_a) \). To show (4.19) in Lemma 4.7, we verified the lower and upper bounds by (4.26) and (4.27). Similarly, we can prove these bounds hold for \( \nabla F(w^k_{\beta}; T_k) \). (In fact, since \( w^k_{\beta} \in N(w^*, \delta_a) \subseteq N(w^*, \delta^2_a) \) by \( \delta_a \leq \delta^2_a \), one just needs to set \( \mu = 0 \) and \( w = w^k_{\beta} \) in (4.26) and (4.27).

\[ (4.42) \]

\[ C_* \geq \| \nabla F(w^k_{\beta}; T_k) \| \geq \sigma_{\text{min}}(\nabla F(w^k_{\beta}; T_k)) \geq c_* \]

Again, the function \( F(\cdot, T_k) \) is differentiable for the fixed \( T_k \), so the Mean-value theorem states that there is a \( \beta_0 \in (0, 1) \) satisfying
\[ \| F(w^k; T_k) \| = \| F(w^*; T_k) + \nabla F(w^k_{\beta_0}; T_k)(w^k - w^*) \| \]
\[ = \| \nabla F(w^k_{\beta_0}; T_k)(w^k - w^*) \| \quad \text{(by (4.41))} \]
\[ \leq [c_* \| w^k - w^* \|, C_* \| w^k - w^* \| ] \quad \text{(by (4.42))} \]

This contributes to
\[ \| F(w^k; T_k) \| \leq C_* \| w^k - w^* \| \leq (\theta_* C_* / c_*) \| w^{k-1} - w^* \|^2 \quad \text{(by (4.39))} \]
\[ \leq (\theta_* C_* / c_*) \| F(w^{k-1}; T_{k-1}) \|^2 \quad \text{(by (4.43))} \]
\[ \leq (\theta_* C_* / c_*) \| w^{k-1} - w^* \|^2 \quad \text{(by (4.43))} \]
\[ \leq (\theta_* C_* / c_*) 2^{-2k} \| w^{k-2} - w^* \|^2 \quad \text{(by (4.40))} \]
\[ \vdots \]
\[ \leq (\theta_* C_* / c_*) 2^{-2k} \| w^0 - w^* \|^2, \quad \text{(by (4.40))} \]

where the third inequality yields the first conclusion in c). This also enable to verify that \( \| F(w^k; T_k) \| < \epsilon \) if \( k \) satisfies (4.29). The whole proof is completed.

\[ \square \]

Remark 4.9. Relationship to primal-dual active-set algorithms. It is interesting to note that when \( \mu_k \equiv \mu \) (a constant) for all indices \( k \), Algorithm 4.1 shares a similar framework to the primal-dual active-set algorithm in [14, Alg. 1], whose main target is the convex quadratic programming in compressed sensing with \( \ell_1 \) regularization. In terms of convergence theory, both Theorem 4.8 and [14, Thm. 2] require the initial point to be close to the interested solution point. There are two key differences. (i) Theorem 4.8 is able to identify the quadratic convergence region \( N(w^*, \delta_a) \) with \( \delta_a \) being given by (4.28), while [14, Thm. 2] does not have such a characterization and is only about it local convergence (not its convergence rate). (ii) However, [14, Thm. 2] can be globalized via a continuation technique, while it is challenging to globalize Algorithm 4.1 because we are dealing with 0/1-loss function and there are no merit functions available for globalization.

Remark 4.10. On the choice of the smoothing parameter \( \mu_k \). An interesting question raised by one referee is whether the particular choice of \( \mu_k \) in (4.8) may play a role in globalization of Algorithm 4.1. From the smoothing perspective, there exists a number of good strategies to update \( \mu \) as long as it drives \( \mu_k \to 0 \). For example, we may update \( \mu_k \) by solving the equation \( e^\mu - 1 = 0 \) via Newton’s method as done in [34], see also [23] for other options. To incorporate such a strategy in a globalization scheme, we must find a merit function to work with. As commented in Remark 4.9, it is not easy to construct a merit function because the composition of the operator \( \| \cdot \|_0 \) with the inequality constraint \( Ax \leq b \) leads to the scenario where the sparsity is not over a symmetric set any more. We refer to [3, 29] for detailed discussion on algorithmic advantages of sparsity being over symmetric sets.
5. Numerical Experiments. In this part, we will conduct extensive numerical experiments of \texttt{NM01} in Algorithm 4.1 by using MATLAB (R2019a) on a laptop with 32GB memory and Inter(R) Core(TM) i9-9880H 2.3Ghz CPU, against a few leading solvers for solving SVM and 1-bit CS problems.

5.1. Experiments for SVM. There exists a large body of SVM literature. We only focus on the binary classification, which has a training dataset \{\((a_i^k, c_i) : i \in N_m\)} with \(a_i^k \in \mathbb{R}^{n-1}\) being samples and \(c_i \in \{-1, 1\}\) being the two classes. It is widely recognized that the data are often linearly inseparable and (1.1) is an ideal model to deal with this case with the following setup

\[
f(x) = \|Dx\|^2, \quad A = -[c_1a_1, \ldots, c_ma_m]^\top, \quad b = 1,
\]

where \(D\) is a diagonal matrix with \(D_{ii} = 1, i \in N_{n-1}\) and \(D_{nn} \geq 0\) (e.g., \(D_{nn} = 10^{-4}\)), and \(a_i = (a_i^1; 1) \in \mathbb{R}^n, i \in N_m\). We will consider two types of datasets: synthetic data and real data described below.

**Example 5.1 (Synthetic data in \(\mathbb{R}^2\)).** Give four samples \((0, 0), (0, 1), (1, 0), (1, a)\) with labels \(+1, +1, -1, -1\), where the last point can be treated as an outlier when \(a > 1\).

**Example 5.2 (Real data in higher dimensions).** We select 40 datasets from three libraries: \texttt{libsvm}, \texttt{uci} and \texttt{kaggle}. All datasets are feature-wisely scaled to \([-1, 1]\) and all the classes not being 1 are treated as -1. Their details are presented in Table 1. There are 16 datasets with \(m \leq n\) and 24 datasets with \(m > n\).

There are large numbers of methods that have been proposed for SVMs, each with its advantages/disadvantages. It is more reasonable to compare \texttt{NM01} with those methods that aim at optimizing the 0/1-loss function directly, such as MIP-based methods. However, it is known that MIP-based methods prefer the datasets on small scales (see the numerical experiments reported in [32, 38, 39]) and behave very slowly for the datasets on mediate/large scales, such as most datasets in Table 1. Therefore, we will not include them in the following numerical comparisons.

On the other hand, there is very limited work on developing methods that directly optimize 0/1-loss from the perspective of continuous optimization. Because of this, we are unable to find an available Matlab implementation for such kinds of methods. Hence, we only select five leading solvers, with available Matlab implementations from the machine learning community. These methods solve the surrogate/relaxations of 0/1-loss involved SVMs. They are \texttt{HSV} from the library \texttt{libsvm}\footnote{https://www.csie.ntu.edu.tw/~cjilin/libsvm/}, \texttt{SSVM} \footnote{https://www.esat.kuleuven.be/sista/lssvmlab/} [37] implemented by \texttt{liblssvm}\footnote{https://www.csie.ntu.edu.tw/~cjilin/liblinear/}, \texttt{RVM} \footnote{https://www.csie.ntu.edu.tw/~cjilin/liblinear/}, \texttt{LSVM} from the library \texttt{liblinear}\footnote{https://www.csie.ntu.edu.tw/~cjilin/liblinear/}, and \texttt{FSSVM} (a MATLAB built-in function \texttt{fitclinear})\footnote{https://mathworks.com/help/stats/fitclinear.html}. All involved parameters are set as their default values. To demonstrate the performance of one method, let \(x\) be its obtained solution and \(A_0 = [a_1, \ldots, a_m]^\top\). We will report the CPU Time and the classification accuracy \(Acc\) defined by \(Acc := 1 - \|\text{sgn}(A_0x) - c\|_2/m\).

5.1.1. Implementation of Algorithm 4.1. We terminate our algorithm if one of the conditions is satisfied: \(k \geq 1000\) or \(\|F(w_k; T_k)\| < 10^{-4}\). We initialize \(x^0 = 0\) and \(z^0 = 1\), and set \(\mu_1 = 0.05\) if \(m < n\) and \(\mu_1 = 5\) otherwise. Moreover, we update \(\mu_k\) by (4.8) with \(\rho = 1\) and \(\alpha = 0.5\) if \(k\) is a multiple of 5. The rest of this part is about setting the parameters \(\tau\) and \(\lambda\). We try to suggest general principles, but bearing in mind that the best strategy of setting \(\tau\) and \(\lambda\) is problem dependent. For the validation purpose, we conducted the performance comparison of Algorithm 4.1 on four test problems \texttt{arco}, \texttt{cocl}, \texttt{dbw1} and \texttt{facb} where we vary one parameter while the other is being fixed. Fig. 1 is for fixed \(\lambda\) and Fig. 2 is for fixed \(\tau\).

(i) **For SVM problems.** It follows from (4.4) in Remark 4.4 that if \(2\lambda \tau \leq \min_i\{b_i : b_i > 0\}\), then \(x^* = 0\) and \(z^* = 0\) is a P-stationary point. For SVM, this condition turns into \(2\lambda \tau \leq 1\) since \(b = 1\). This phenomenon can be observed in our numerical experiments. For example, zero solutions were obtained by Algorithm 4.1 when \(\tau = 1/(2\lambda)\) for fixed \(\lambda = 15\) in Fig. 1 and when \(\lambda \leq 1/(2\tau)\) for fixed \(\tau = 5\) in Fig. 2. Hence, it is recommended to set \(\tau\) and \(\lambda\) to satisfy \(2\lambda \tau > 1\) for SVM problems.
SMOOTHING NEWTON’S METHOD FOR 0/1-LOSS OPTIMIZATION

Table 1: Descriptions of real datasets.

| Data | Datasets | Source     | n     | m     | Sparse |
|------|----------|------------|-------|-------|--------|
| arce | Arcene   | uci        | 10000 | 100   | No     |
| colc | Colon-cancer | libsvm   | 2000  | 62    | No     |
| dbv1 | uci      | 4702       | 64    | Yes   |
| dbv2 | uci      | 3721       | 64    | Yes   |
| dbv3 | uci      | 242        | 64    | Yes   |
| dbv4 | uci      | 229        | 64    | Yes   |
| dext | Dexter   | uci        | 19999 | 300   | Yes    |
| dmea | Detect malicious executable | uci    | 531   | 373   | Yes    |
| doro | Dorothea | uci        | 100000| 800   | Yes    |
| dubb | Duke breast-cancer | libsvm | 7129  | 38    | No     |
| fabc | Farm ads binary classification | libsvm | 54877 | 4143  | Yes    |
| leuq | Leukemia | libsvm     | 7129  | 38    | No     |
| aips | Airline passenger satisfaction | kaggle | 22    | 103904| No     |
| ccfd | Credit card fraud detection | kaggle | 28    | 284807| No     |
| covt | Covtype.binary | libsvm | 54    | 581012| Yes    |
| dccc | Default of credit card clients | kaggle | 23    | 30000 | No     |
| escd | Email spam classification dataset | kaggle | 3000  | 5172  | Yes    |
| gise | Gisette | libsvm     | 5000  | 6000  | Yes    |
| hepm | Hepmass  | uci        | 28    | 700000| No     |
| hxff | Hedge fund x: financial mod. chal. | kaggle | 88    | 10000 | No     |
| higg | Higgs    | uci        | 28    | 1100000| No    |
| hmeq | Hmeq_data | kaggle | 10    | 5960  | No     |
| htr2 | Htru2    | uci        | 8     | 17898 | No     |
| idac | Ida2016challenge | uci    | 170   | 60000 | Yes    |
| jcin | Jcnn1    | libsvm     | 22    | 49990 | Yes    |
| mrpe | Malware analysis datasets: raw pe | kaggle | 1024  | 51959 | No     |
| mpte | Malware analysis datasets: top-1000 | kaggle | 1000  | 47580 | Yes    |
| ospi | Online shoppers purchasing intention | uci   | 17    | 12330 | No     |
| pssr | Parkinson speech database | uci    | 26    | 1039  | No     |
| qgot | Qsar oral toxicity | uci    | 1024  | 8992  | Yes    |
| reas | Real-sim | libsvm     | 20958 | 72309 | Yes    |
| retb | Real time bidding | kaggle | 88    | 1000000| No    |
| scct | Santander customer transaction | kaggle | 200   | 200000| No     |
| skin | Skin nonskin | libsvm | 3     | 245056| No     |
| spli | Splice    | libsvm     | 60    | 1000  | No     |
| susy | Susy     | uci        | 18    | 500000| No     |

(ii) On the choice of $\tau$. Despite that a sufficient condition $0 < \tau < \tau_*$ is provided in Theorem 4.8, it is still difficult to set a proper $\tau$ as $\tau_*$ is not known. However, as the condition is sufficient, it is unnecessary to choose it from $(0, \tau_*)$ strictly. To see its effect, we tested it with varying $\tau \in [10^{-3}, 10]$, fixed $\lambda = 15$ and report the results in Fig. 1. It can be clearly seen that bigger values of $\tau$ (e.g., $\tau \geq 1$) lead to better accuracy $\text{ACC}$. An underlying heuristic explanation is as follows: Algorithm 4.1 solves the system (4.6) with index set $T_k = S_k \cup E_k^2$ being decided by the parameter $\tau$, see (4.1). We observed that setting $\tau$ too small often led to infrequent change of $T_k$ and this often forced the algorithm fell into (possibly undesirable) local regions too quickly. By contrast, setting $\tau$ slightly bigger enabled altering $T_k$ frequently enough to make Algorithm 4.1 escape from undesirable local regions so as to achieve better solutions. Since the theoretical convergence is in favour of small values of $\tau < \tau_*$, it is not suggested to set the values of $\tau$ too large.

(iii) On the choice of $\lambda$. For the parameter $\lambda$, we varied values $\lambda \in [10^{-2}, 10^2]$ and report its effect in Fig. 2. As expected, zero solutions were achieved when $\lambda \leq 1/(2\tau)$. From the left sub-figure, ACCs are in favour of bigger values of $\lambda > 1/(2\tau)$ which is reasonable since it penalizes the 0/1 loss in (1.1). This choice of $\lambda$ is consistent with what we have observed in
(iv) **Estimating** $\tau_\ast$. Although $\tau_\ast$ is unknown, we may be able to numerically estimate it by using the obtained solution and (3.13) provided that some information was available a priori. Note that if $x_\ast^\prime = 0$ and $z_\ast^\prime = 0$ then $\tau_\ast = \min\{\tau_1, \tau_2\} = 1/(2\lambda)$ by (3.13), which can be seen in Fig. 1 and Fig. 2. On the other hand, if a solution satisfies that $Ax_\ast^\prime + b < 0$, then $\tau_1 = \tau_2 = \infty$ (3.13) and hence $\tau_\ast = \infty$. This phenomena can be observed for datasets *arce* and *colc* since they are linearly separable, see the results for $\tau > 1/(2\lambda)$ in Fig. 1 and for $\lambda > 1/(2\tau)$ in Fig. 2.

Therefore, in the following experiments, we set $\tau = 5$ and $\lambda = 15$ for simplicity.

**5.1.2. Numerical comparisons.** We first employ five methods to solve Example 5.1 under different $a = 1, 10, 100$ to test their robustness to the outliers. For such data, the classifier with a maximum margin is $x_\ast^1 = 1/2$. The classifiers by each method are plotted in Fig. 3, where HSVM is omitted since it solves the dual problem and does not provide the solution $x$. Obviously, NM01 finds the true classifiers for all scenarios, while the other methods are influenced significantly by $a$.

For Example 5.2, we have 40 datasets with sample size from a few to ten million (e.g., *higg* having 11,000,000 samples). Results of six methods are reported in Table 2, where "---" denotes the results are not obtained if a solver takes too much time or requires a large memory that is out of the capacity of our desktop. For example, HSVM consumes more than 10,000 seconds on the data *covt* and SSVM requires at lest 32GB memory to solve *mtpe*. In general, NM01 renders the highest Acc for most datasets. For the computational time, F SVM and L SVM are very fast for datasets of moderate sizes. However, our method is more competitive especially when the data size is in million scale, such as *higg*, *retb*, *susy*, *hepm* with more than $10^6$ samples, NM01 runs the fastest. For instance, F SVM...
and L SVM respectively took 80.69 seconds and 65.65 seconds for higg, which is solved by our method within 5.63 seconds.

Table 2: Results of six solvers for Example 5.2

| data  | Acc | Time (seconds) |
|-------|-----|----------------|
|       | FSVM | HSVM | L SVM | RSVM | SSVM | NM01 |
|       | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| arcc  | 0.952 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| colc  | 0.984 | 0.984 | 0.984 | 0.984 | 0.984 | 0.984 |
| dbv1  | 0.984 | 0.984 | 0.984 | 0.984 | 0.984 | 0.984 |
| dbv2  | 0.984 | 0.984 | 0.984 | 0.984 | 0.984 | 0.984 |
| dbv3  | 0.984 | 0.984 | 0.984 | 0.984 | 0.984 | 0.984 |
| dext  | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| dmea  | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| doro  | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| dubc  | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| fabc  | 0.996 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 |
| leuc  | 0.992 | 0.984 | 0.984 | 0.973 | 1.000 | 1.000 |
| lsst  | 0.905 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 |
| newb  | 0.995 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 |
| rcvb  | 0.990 | 0.990 | 0.997 | 0.997 | 0.997 | 0.997 |
| scad  | 0.986 | 1.000 | 1.000 | 0.971 | 1.000 | 1.000 |
| alps  | 0.976 | 0.974 | 0.974 | 0.974 | 0.974 | 0.974 |
| cfdf  | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 |
| covt  | 0.763 | 0.757 | 0.757 | 0.757 | 0.757 | 0.757 |
| dccc  | 0.810 | 0.809 | 0.809 | 0.799 | 0.820 | 0.820 |
| escd  | 0.993 | 0.992 | 0.996 | 0.986 | 0.971 | 0.996 |
| gise  | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| hepm  | 0.837 | 0.836 | 0.836 | 0.836 | 0.836 | 0.836 |
| hxfx  | 0.589 | 0.589 | 0.589 | 0.589 | 0.589 | 0.589 |
| higg  | 0.641 | 0.641 | 0.641 | 0.641 | 0.641 | 0.641 |
| hmmq  | 0.860 | 0.859 | 0.860 | 0.803 | 0.862 | 0.865 |
| htrue | 0.977 | 0.977 | 0.977 | 0.971 | 0.979 | 0.979 |
| idac  | 0.991 | 0.992 | 0.992 | 0.992 | 0.992 | 0.992 |
| icen  | 0.924 | 0.924 | 0.923 | 0.923 | 0.923 | 0.923 |
| mper  | 0.948 | 0.951 | 0.951 | 0.951 | 0.951 | 0.951 |
| mpe  | 0.968 | 0.968 | 0.968 | 0.968 | 0.968 | 0.968 |
| ospi  | 0.884 | 0.884 | 0.884 | 0.884 | 0.884 | 0.884 |
| pssr  | 0.641 | 0.643 | 0.654 | 0.626 | 0.647 | 0.663 |
| qスト  | 0.946 | 0.969 | 0.976 | 0.849 | 0.945 | 0.971 |
| reas  | 0.989 | 0.989 | 0.994 | 0.994 | 0.994 | 0.994 |
| retb  | 0.998 | 0.998 | 0.998 | 0.998 | 0.998 | 0.998 |
| sctp  | 0.910 | 0.909 | 0.909 | 0.914 | 0.914 | 0.914 |
| skin  | 0.929 | 0.929 | 0.924 | 0.924 | 0.943 | 0.943 |
| spli  | 0.839 | 0.839 | 0.840 | 0.805 | 0.840 | 0.840 |
| susy  | 0.788 | 0.787 | 0.787 | 0.790 | 0.790 | 0.790 |
5.2. Simulations for 1-bit CS. The aim of 1-bit CS is to recover a sparse signal \( x \) from \( c = \text{sgn}(A_0x) \), where \( A_0 := [a_1, \ldots, a_m]^\top \in \mathbb{R}^{m \times n} \) and \( c_i \in \{1, -1\}, i \in \mathbb{N}_m \). The original optimization model for 1-bit CS [5] takes the following form:

\[
\min \|x\|_0, \quad \text{s.t.} \quad c_i\langle a_i, x \rangle \geq 0, \quad i \in \mathbb{N}_m.
\]

Various relaxation methods have been proposed. Here we adopt the smoothing technique using the popular \( \ell_q \) norm \( (0 < q < 1) \) to approximate the \( \ell_0 \) norm [26] and use the 0/1-loss function to deal with the constraints. This leads to the model (1.1) with

\[
f(x) = \sum_{i=1}^n (\epsilon_i^2 + \epsilon^2)^{q/2}, \quad A = [-c_1a_1, \ldots, c_ma_m], \quad b = \epsilon 1,
\]

where \( \epsilon > 0, \epsilon > 0 \). Here, \( b = \epsilon 1 \) is adopted from [12]. In our test, we set \( q = 0.5, \epsilon = 0.05 \) but update \( \epsilon \) by \( \epsilon_k = 0.5 \) and \( \epsilon_{k+1} = \epsilon_k/2 \). The test problems are taken from [21] and are described as follows.

**Example 5.3.** Rows of \( A_0 \) are the independent and identically distributed (iid) samples of \( N(0, \Sigma) \) with \( \Sigma_{ij} = v^{-|i-j|}, i, j \in \mathbb{N}_n \) and \( v \in (0, 1) \). The nonzero entries of the ground truth \( s \)-sparse vector \( x^* \in \mathbb{R}^n \), namely, \( \|x^*\|_0 \leq s \), are generated from the i.i.d. samples of the standard Gaussian distribution \( N(0, 1) \), followed by a normalization of \( x^* \) to be a unit vector. Let \( c^* = \text{sgn}(A_0x^*) \) and \( \tilde{c} = \text{sgn}(A_0x^* + \xi) \), where entries of the noise \( \xi \) are the i.i.d. samples of \( N(0, 0.1^2) \). Finally, we randomly select \( \lfloor rm \rfloor \) entries in \( \tilde{c} \) and flip their signs, and the flipped vector is denoted by \( c \), where \( r \) is the flipping ratio.

We report the CPU time, the signal-to-noise ratio \( \text{SNR} := -10\log_{10}(\|x - x^*\|_2^2) \), the hamming error \( \text{HE} := \|\text{sgn}(A_0x) - c^*\|_0/m \), and the hamming distance \( \text{HD} := \|\text{sgn}(A_0x) - c\|_0/m \), where \( x \) is the solution obtained by a method. We note that larger \( \text{SNR} \) (smaller \( \text{HE} \), or smaller \( \text{HD} \)) corresponds to better recovery.

5.2.1. Implementation and benchmark methods. The stopping criteria and the rule for updating \( \mu_k \) for Algorithm 4.1 are the same as for the SVM case. We initialize \( x^0 = 0 \) and \( z^0 = 1 \). We tested the algorithm under different choices of \( \tau \) and \( \lambda \), and only report the numerical results with \( \tau = 1 \) and \( \lambda = 1 \), which satisfy \( 2\tau \lambda > \min\{b_i^2 : b_i > 0\} = \epsilon^2 \) and could yield good overall performance. Moreover, it is observed that the generated solutions had many tiny values, see Fig. 4. Therefore, we apply a refinement step that keeps the \( s \) largest elements in the magnitude of the solution and sets the rest to zeros.

![Fig. 4: Refinement of the solution.](image-url)
Fig. 5: Effects of $m$, $s$, $v$ and $r$ for Example 5.3.
Four leading solvers are selected for comparison. They are PDASC\textsuperscript{5} [21], BIHT\textsuperscript{6} [25], AOP\textsuperscript{7} [44] and PBAOP\textsuperscript{7} [22], where the last three require to specify the true sparsity level \(s\), and the last two also need a flipping ratio \(L\). As in [44], we choose \(L = \text{HD}\), where \(\text{HD}\) is the hamming distance generated by BIHT. We also apply the refinement step to PDASC so that all five methods produce \(s\)-sparse solutions. Finally, all methods start with \(x^0 = 0\) and their solutions are normalized to have a unit length.

5.2.2. Comparison. We now apply the five methods to solve Example 5.3 under different scenarios. For each scenario, we report average results over 200 instances if \(n \leq 1000\) and 20 instances otherwise. For small scale instances, we set five parameters as \((m, n, s, v, r) = (500, 250, 5, 0.5, 0.05)\). To see the effect of each of these parameters, we tested one parameter while the others being fixed.

- Effect of \(m \in \{0.1, 0.2, \ldots, 1\}n\). We note that the bigger \(m\) enables the better performance, since more samples are available to recover the signal. It can be clearly seen from Fig. 5 that NM01 gets the largest SNR and the smallest \(\text{HD}\) and \(\text{HE}\), leading to a better performance than the others.

- Effect of \(s \in \{2, 3, \ldots, 10\}\). The three sub-figures in the second row of Fig. 5 indicate that it is getting more difficult to recover the ground truth signal when \(s\) increases. In comparison with other methods, NM01 delivers the best recoveries as it achieves the highest SNR and the smallest \(\text{HD}\) and \(\text{HE}\).

- Effect of \(v \in \{0.1, 0.2, \ldots, 0.9\}\). The third row sub-figures in Fig. 5 demonstrate that the bigger values of \(v\) degrade the performance of each method, because each pair of two rows of \(A_n\) is more correlated with increasing \(v\). It is observed that NM01 delivers the best results when \(v < 0.9\).

- Effect of \(r \in \{0.02, 0.04, \ldots, 0.2\}\). As expected, the bigger \(r\) is (i.e., the more signs are flipped), the harder the recovery is. This can be seen in the sub-figure in the last row of Fig. 5. NM01 outperforms the others.

- Effect of \(n \in \{2000, 4000, \ldots, 10000\}\). For the higher dimensional instances, we fix \(m = n/2, s = 5n/1000, v = 0.5\) and \(r = 0.05\). We record the average results in Table 3 where NM01 achieves the most desirable recovery accuracy. For the computational time, the other methods are naturally expected to run super-fast since they belong to the family of greedy methods that exploit the sparse structure of the solutions. Nevertheless, NM01 is relatively competitive in terms of the computational speed.

\begin{table}
\centering
\begin{tabular}{|c|c|c|c|c|c|c|c|c|c|}
\hline
\textbf{n} & \textbf{BIHT} & \textbf{AOP} & \textbf{PBAOP} & \textbf{PDASC} & \textbf{NM01} & \textbf{BIHT} & \textbf{AOP} & \textbf{PBAOP} & \textbf{PDASC} & \textbf{NM01} \\
\hline
\textbf{SNR} & & & & & & 0.034 & 0.414 & 0.176 & 0.061 & 0.170 \\
\hline
\textbf{TIME} & & & & & & 0.276 & 1.832 & 0.773 & 0.255 & 0.880 \\
\hline
\textbf{HE} & & & & & & 0.803 & 4.149 & 2.062 & 0.636 & 1.884 \\
\hline
\textbf{HD} & & & & & & 1.466 & 7.279 & 3.778 & 1.186 & 3.715 \\
\hline
\textbf{2000} & & & & & & 2.414 & 11.76 & 6.777 & 2.081 & 5.675 \\
\hline
\textbf{4000} & & & & & & 0.201 & 0.204 & 0.209 & 0.129 & 0.170 & 0.176 & 0.175 & 0.145 & 0.091 \\
\hline
\textbf{6000} & & & & & & 0.207 & 0.203 & 0.201 & 0.125 & 0.177 & 0.174 & 0.171 & 0.198 & 0.087 \\
\hline
\textbf{8000} & & & & & & 0.203 & 0.204 & 0.206 & 0.134 & 0.173 & 0.174 & 0.176 & 0.247 & 0.097 \\
\hline
\textbf{10000} & & & & & & 0.205 & 0.202 & 0.202 & 0.133 & 0.174 & 0.171 & 0.172 & 0.262 & 0.094 \\
\hline
\end{tabular}
\caption{Effect of the higher \(n\) for Example 5.3.}
\end{table}

Remark 5.1. \textbf{(On nonsingularity of the Jacobian matrix.)} We finish this section by discussing the important issue of nonsingularity of the (smoothing) Jacobian matrix \(\nabla^2 f(x^k) + A_k^T A_k / \mu_k\) used in Algorithm 4.1. As pointed out in Introduction, its nonsingularity is equivalent to the the nonsingularity of \(M_k := \nabla^2 f(x^k) + A_k^T A_k / \mu_k\). If \(\nabla^2 f(x^k)\) is positive definite, then \(M_k\) is always nonsingular.

\begin{itemize}
\item [5]http://jszy.whu.edu.cn/jiaoyuling/en/lwcg/1349484/content/54893.htm#lwcg
\item [6]https://laurentjacques.gitlab.io/publication/
\item [7]http://www.esat.kuleuven.be/stadius/ADB/huang/downloads/1bitCSLab.zip
\end{itemize}
This is the case for the SVM problems tested. For the problem of 1-bit compressed sensing, we let
\[ C_x := \nabla^2 f(x^*) = \text{diag} \left\{ \frac{q [\varepsilon^2 - (1-q)(x_i^*)^2]}{\varepsilon^2 + (x_i^*)^2} - \frac{q/2}{(x_i^*)^2} - q/2, \quad i = 1, \ldots, n \right\}, \]
where \( x^* \) is the limit of the sequence \( \{x^k\} \). Suppose \( C_x \) is nonsingular, then \( C_k := \nabla^2 f(x^k) \) is also nonsingular when \( x^k \) is close to \( x^* \). The Woodbury matrix identity implies that the nonsingularity of \( M_k \) is equivalent to that of the matrix
\[ \tilde{M}_k := I + \frac{1}{\mu_k} A_k C_k^{-1} A_k^T. \]
Since \( \mu_k \) converges to 0, \( \Delta_k \) cannot have \((-1)\) as its eigenvalue when \( x^k \) is close to \( x^* \) and hence \( M_k \) is always nonsingular. To slightly generalize the above argument, as long as \( \Delta_k \) does not have \((-1)\) among its eigenvalues, \( \tilde{M}_k \) (hence \( M_k \)) is always nonsingular. And the chance for \( \Delta_k \) to have \((-1)\) as its eigenvalue is extremely small in general. This is what we experienced in our test. The argument above does raise the question how to ensure the nonsingularity. It comes back to the globalization issue of Algorithm 4.1. Our proposal is to use a gradient method whenever singularity becomes an issue. We leave this to future research.

6. Conclusion. Optimizing the 0/1-loss function has been a challenging task for several decades, and few optimality conditions or theoretical convergence guarantees have been established for most of the 0/1-loss function minimizations. This paper is the first to develop Newton's method with guaranteed quadratic convergence. This has come a long way by first proposing a \( P \)-stationarity condition that leads to stationarity equations, and then establishing the desired convergence results with help of very technical control over the growth of residue equations. The excellent numerical performance of the proposed method for solving the SVM and 1-bit CS problems indicate that it might work well for other related applications. We strongly feel that the techniques developed in this paper can be extended to a more general case, where \( Ax + b \) in (1.1) is replaced by some non-linear functions.

Acknowledgements. We would like to thank both the referees for their detailed comments that have helped to improve the quality of the paper. In particular, we thank one referee for suggesting the current title and the other for pointing out the link of the proposed algorithm to the primal-dual active-set algorithms extensively studied among semi-smooth Newton methods.

REFERENCES

[1] S. M. Bajgier and A. V. Hill, An experimental comparison of statistical and linear programming approaches to the discriminant problem, Decision Sciences, 13 (1982), pp. 604–618.
[2] A. Beck and Y. C. Eldar, Sparsity-constrained nonlinear optimization: Optimality conditions and algorithms, SIAM Journal on Optimization, 23 (2013), pp. 1480–1509.
[3] A. Beck and N. Hallak, On the minimization over sparse symmetric sets: projections, optimality conditions, and algorithms, Mathematics of Operations Research, 41 (2016), pp. 196–223.
[4] S. Ben-David, N. Eiron, and P. M. Long, On the difficulty of approximately maximizing agreements, Journal of Computer and System Sciences, 66 (2003), pp. 496–514.
[5] P. T. Boufounos and R. G. Baraniuk, 1-bit compressive sensing, in 2008 42nd Annual Conference on Information Sciences and Systems, IEEE, 2008, pp. 16–21.
[6] J. P. Brooks, Support vector machines with the ramp loss and the hard margin loss, Operations Research, 59 (2011), pp. 467–479.
[7] J. P. Brooks and E. K. Lee, Analysis of the consistency of a mixed integer programming-based multi-category constrained discriminant model, Annals of Operations Research, 174 (2010), pp. 147–167.
[8] E. Carrizosa, B. Martín-Barragan, and D. R. Morales, Binarized support vector machines, INFORMS Journal on Computing, 22 (2010), pp. 154–167.
[9] C. C. Chang and C. J. Lin, LIBSVM: A library for support vector machines, ACM Transactions on Intelligent Systems and Technology (TIST), 2 (2011), pp. 1–27.
[10] X. Chen, L. Qi, and D. Sun, Global and superlinear convergence of the smoothing Newton method and its application to general box constrained variational inequalities, Mathematics of computation, 67 (1998), pp. 519–540.
[11] C. Cortes and V. Vapnik, Support-vector networks, Machine Learning, 20 (1995), pp. 273–297.
[12] D. Dai, L. Shen, Y. Xu, and N. Zhang, Noisy 1-bit compressive sensing: models and algorithms, Applied and Computational Harmonic Analysis, 40 (2016), pp. 1–32.
