SIGN CONTROLLED SOLVERS FOR THE ABSOLUTE VALUE EQUATION WITH AN APPLICATION TO SUPPORT VECTOR MACHINES

LUTZ LEHMANN, MANUEL RADONS, SIEGFRIED M. RUMP,
AND CHRISTIAN STROHM

ABSTRACT. Let A be a real $n \times n$ matrix and $z, b \in \mathbb{R}^n$. The piecewise linear equation system $z - A|z| = b$ is called an *absolute value equation*. It is equivalent to the general *linear complementarity problem*, and thus NP hard in general. Concerning the latter problem, three solvers are presented: One direct, one semi-iterative and one discrete variant of damped Newton. Their previously proved ranges of correctness and convergence, respectively, are extended. Their performance is compared on instances of the XOR separation problem for *support vector machines* which can be reformulated as an absolute value equation.

1. INTRODUCTION

Denote by $\mathrm{M}_n(\mathbb{R})$ the space of $n \times n$ real matrices and let $A \in \mathrm{M}_n(\mathbb{R})$ and $z, b \in \mathbb{R}^n$. The piecewise linear equation system

$$z - A|z| = b \tag{1.1}$$

is called an *absolute value equation* (AVE) and was first introduced by Rohn in [20]. By decomposing z into nonnegative vectors u and v such that $z = u - v$ and $u^\intercal v = 0$ an equivalent *linear complementarity problem* (LCP)

$$v \;=\; -(I-A)^{-1}b + (I-A)^{-1}(I+A)u \tag{1.2}$$

is obtained. In [17, P. 216-230] Neumaier authored a detailed survey about the AVEs intimate connection to the research field of *linear interval equations*. Especially closely related system types are equilibrium problems of the form

$$Bx + \max(0, x) = c, \tag{1.3}$$

where $B \in \mathrm{M}_n(\mathbb{R})$ and $x, c \in \mathbb{R}^n$. (A prominent example is the first hydrodynamic model presented in [3].) Using the identity $\max(s,t) = (s + t + |s - t|)/2$, equality (1.3) can be reformulated as

$$Bx + \frac{x + |x|}{2} = c \quad \Longleftrightarrow \quad (2B + I)x + |x| = 2c,$$

and for regular $(2B + I)$ the left equality is equivalent to an AVE (1.1).

One connection relevant to the authors of this work concerns nonsmooth optimization: Piecewise affine systems of arbitrary structure may arise as local linearizations of piecewise differentiable objective functions [9] or as intermediary problems

2010 *Mathematics Subject Classification.* Primary .

in the numerical solution of ordinary differential equations with nonsmooth right-hand side [8]. Such system can be, with a *one-to-one solution correspondence*, transformed into an AVE [7, Lem. 6.5].

This position at the crossroads of several interesting problem areas gives relevance to the task of developing efficient solvers for the AVE. The latest publications on the matter include approaches by linear programming [13] and concave minimization [11], as well as a variety of Newton and fixed point methods (see, e.g., [3], [27], [10]). In this article we will present and further analyze three solvers for the AVE – one direct, one semi-iterative, one in the spirit of damped Newton methods – that were developed in [7, 25], [18], and [6], respectively. This especially means that we will recall and further extend convergence results for all three algorithms.

Moreover, we will reformulate the frequently discussed XOR-Problem for support vector machines (SVMs) (see, e.g. [2, 15]) as an AVE and test the algorithms' performance on it.

Content and structure: In Section 2 we will assemble the necessary preliminaries from the literature and (re-)prove some auxiliary results. In Sections 3-5 the three aforementioned solvers are presented; correctness / convergence ranges are proved and key aspects of a performant implementation are addressed. In Section 6 the AVE-formulation of the XOR-problem is derived. Numerical experiments make up Section 7: The algorithms' performances are investigated both with regard to random data as well as the XOR-problem. The article is concluded by some final remarks in Section 8.

2. Preliminaries

We denote by $[n]$ the set $\{1,\dots,n\}$. For vectors and matrices *absolute values and comparisons are used entrywise.* Zero vectors and matrices are denoted by $\mathbf{0}$. Let $c \in \mathbb{R}^n$, then we denote by $\operatorname{diag}_n(c)$ a diagonal matrix in $\mathrm{M}_n(\mathbb{R})$ with entries $c_1,\dots,c_n$. We omit the subscript n and write $\operatorname{diag}(c)$ or $\operatorname{diag}(c_1,\dots,c_n)$ if the dimension is clear from the context.

A *signature matrix* Σ, or, briefly, *a signature*, is a diagonal matrix with entries $+1$ or -1, i.e. $|\Sigma| = I$. The set of n-dimensional signature matrices is denoted by $\mathcal{S}_n$. A single diagonal entry of a signature is a sign σ_i $(i \in [n])$. Let $z \in \mathbb{R}^n$. We write Σ_z for a signature, where $\sigma_i = 1$ if $z_i \geq 0$ and -1 else. Clearly, we then have $\Sigma_z z = |z|$. Using this convention, we can rewrite (1.1) as

$$(I - A\Sigma_z)z \;=\; b\,. \tag{2.1}$$

In this form it becomes apparent that the main difficulty in the computation of a solution for (2.1) is to determine the proper signature Σ for z. That is, to determine in which of the 2^n orthants about the origin z lies. This is NP-hard in general [12].

Denote by $\rho(A)$ the spectral radius of A and let

$$\rho_0(A) \equiv \max\{|\lambda| : \lambda \textit{ real} \text{ eigenvalue of } A\}$$

be the *real spectral radius* of A. Then its *sign-real spectral radius* is defined as follows (see [21, Def. 1.1]):

$$\rho^{\mathbb{R}}(A) \;\equiv\; \max\left\{\rho_0(\Sigma A) : \Sigma \in \mathcal{S}_n\right\}.$$

The exponential number of signatures Σ accounts for the NP-hardness of the computation of $\rho^{\mathbb{R}}(A)$ [21, Cor. 2.9]. It is easy to check that $\mathcal{S}_n$ is a finite subgroup of $\mathrm{Gl}_n(\mathbb{R})$. Thus, for a fixed signature $\bar{\Sigma}$, the sets $\{\Sigma(\bar{\Sigma}A) : \Sigma \in \mathcal{S}_n\}$ and

$\{\Sigma A : \Sigma \in \mathcal{S}_n\}$ are identical modulo a permutation. Furthermore, since all $\Sigma \in \mathcal{S}_n$ are obviously involutive, i.e., $\Sigma^{-1} = \Sigma$, the spectra of A and $\Sigma A \Sigma$ are identical. These observations immediately yield the useful identity

$$\rho^{\mathbb{R}}(A) = \rho^{\mathbb{R}}(\Sigma_1 A) = \rho^{\mathbb{R}}(A\Sigma_2) = \rho^{\mathbb{R}}(\Sigma_1 A \Sigma_2) \qquad \forall\ \Sigma_1, \Sigma_2 \in \mathcal{S}_n.$$

Recall that a real (or complex) square matrix is called a P-matrix if every principal minor is positive [5, p. 147]. An LCP has a unique solution for all right hand sides if and only if its system matrix is a P-matrix [5, p. 148, Thm. 3.3.7]. The solvability properties of (2.1) and the quantity $\rho^{\mathbb{R}}(A)$ are heavily intertwined (cf. [21], [17, p. 220, Thm. 6.1.3-5]):

Theorem 2.1. *Let $A \in \mathrm{M}_n(\mathbb{R})$. Then the following are equivalent:*

(1) $\rho^{\mathbb{R}}(A) \ <\ 1$.
(2) $(I - A)^{-1}(I + A)$ *is a P-matrix.*
(3) *The system $(I - A\Sigma_z)z = b$ has a unique solution for all $b \in \mathbb{R}^n$.*
(4) *The piecewise linear function $\varphi :\ \mathbb{R}^n \to \mathbb{R}^n,\ z \to z + A|z|$ is bijective.*
(5) $\det(I - A\Sigma)\ >\ 0$ *for all $\Sigma \in \mathcal{S}_n$.*
(6) $\det(I - AD)\ >\ 0$ *for all real diagonal matrices $D \in \mathrm{M}_n(\mathbb{R})$ with $\|D\|_\infty \leq 1$.*
(7) *$I - A\Sigma$ is a P-matrix for all $\Sigma \in \mathcal{S}_n$.*

We provide a brief assertion of the statements essential to our investigation. For a complete proof of Theorem 2.1 we refer to the afore cited references.

Proof. $(1) \Rightarrow (5)$: $\rho^{\mathbb{R}}(A)\ <\ 1$ implies that the real eigenvalues of all $(I - A\Sigma), \Sigma \in \mathcal{S}_n$, are positive [18].

$(2) \Leftrightarrow (3)$: Follows from (1.2).

$(3) \Leftrightarrow (4)$: Clear.

$(4) \Rightarrow (5)$: The matrices $(I - A\Sigma), \Sigma \in \mathcal{S}_n$, are the Jacobians of the selection functions of φ. It is well known that a bijective piecewise linear functions is coherently oriented in that all its Jacobians have the same nonzero determinant sign. Proof that all are positive: Assume $\det(I - A\Sigma)\ <\ 0$ for all $\Sigma \in \mathcal{S}_n$. Then $A\Sigma$ has at least one real eigenvalue < -1 for all $\Sigma \in \mathcal{S}_n$. This implies that the convex hull of $I - A$ and $I + A$ contains a singular matrix, contradicting the linearity of the determinant for rank-1 updates.

$(5) \Rightarrow (4)$: A coherently oriented piecewise linear function is surjective [23, p.32]. Moreover, the *claim* implies that all generalized Jacobians of φ in the sense of [4] are nonsingular, which implies local injectivity everywhere by [4, Prop. 7.1.16]. But, by [23, p.44, Thm. 2.3.2.1], piecewise linear functions are globally injective if they are locally injective everywhere.

$(5) \Rightarrow (6)$: This follows, again, from the linearity of the determinant for rank-1 updates.

$(6) \Rightarrow (1)$: If it was $\rho^{\mathbb{R}}(A) \geq 1$, we could find a diagonal matrix D wit $\|D\|_\infty \leq 1$ such that $I - AD$ was singular, contradicting the hypothesis. □

There exist various other proofs for the equivalencies listed in Theorem 2.1. See, e.g., [21, 17, 18]. Moreover, note that the sign-real spectral radius is but one facet of the unified Perron-Frobenius theory developed in [22] which extends several key properties of the Perron root of nonnegative real matrices to general real and complex matrices via the concepts of the sign-real and *sign-complex spectral radius*, respectively. A unified expression for these three quantities is derived in [22, Thm.

2.4]:

$$\rho^{\mathbb{K}}(A) \;=\; \max_{0\neq x\in\mathbb{K}^n}\;\min_{x_i\neq 0}\left|\frac{(Ax)_i}{x_i}\right| ,$$

where $\mathbb{K} \in \{\mathbb{R}_+, \mathbb{R}, \mathbb{C}\}$ and $A \in \mathrm{M}_n(\mathbb{K})$.

Remark 2.2. An important fact is that $\rho^{\mathbb{R}}(A)$ *is bounded by all p-norms* [21, Thm. 2.15]. It affirms that all systems considered in the sequel are *uniquely solvable.*

The following simple observation is key to the subsequent discussion:

Proposition 2.1. *Let $A \in \mathrm{M}_n(\mathbb{R})$ and $z, b \in \mathbb{R}^n$ such that they satisfy* (2.1)*. If $\|A\|_\infty < 1$, then for at least one $i \in [n]$ the signs of z_i and b_i have to coincide.*

Proof. Let z_i be an entry of z s.t. $|z_i| \geq |z_j|$ for all $j \in [n]$. If $z_i = 0$, then $z = \mathbf{0}$ and thus $b \equiv z - A|z|$ is the zero vector as well – and the statement holds trivially. If $|z_i| > 0$, then $|e_i^{\mathsf{T}} A|z|| < |z_i|$, due to the norm constraint on A. Thus, $b_i = z_i - e_i^{\mathsf{T}} A|z|$ will adopt the sign of z_i. □

We do not know though, for which indices the signs coincide. The theorem below states restrictions on A which guarantee the coincidence of the signs of z_i and b_i for all $i \in [n]$ where $|b_i| \geq |b_j|$ for all $j \in [n]$ and thus provide the basis for the convergence proofs in Sections 3 and 4. Now let $b \in \mathbb{R}^n$ and define

$$\mathcal{I}^b_{\max} \;\equiv \{1 \leq i \leq n \;:\; |b_i| \geq |b_j| \;\forall\; j \in [n]\} \;.$$

Theorem 2.3. *Let $A \in \mathrm{M}_n(\mathbb{R})$ and $b, z \in \mathbb{R}^n$ such that* (2.1) *is satisfied. Then the set*

$$\mathrm{Eq}(A, b, z) \;\equiv\; \{i \in \mathcal{I}^b_{\max} \;:\; \mathrm{sign}(b_i) \neq \mathrm{sign}(z_i)\} \;,$$

where sign *denotes the signum function, is empty if either of the following conditions is satisfied.*

(1) $\|A\|_\infty < \frac{1}{2}$.
(2) *A is irreducible with $\|A\|_\infty \leq \frac{1}{2}$.*
(3) *A is strictly diagonally dominant and $\|A\|_\infty \leq \frac{2}{3}$.*
(4) *$|A|$ is tridiagonal and symmetric with $\|A\|_\infty < 1$ and $n \geq 2$.*

The first three points are cited from [18, Thm. 3.1]. We will prove the fourth point and reprove the first two by somewhat more elegant means than in the latter reference. This includes a new proof for the following lemma.

Lemma 2.4. *(*[18, Lem. 3.2]*) Let $A \in M_n(\mathbb{R})$ with $\|A\|_\infty < \frac{1}{2}$ or irreducible with norm $\|A\|_\infty \leq \frac{1}{2}$. Then the inverse of $B = I - A$ is strictly diagonally dominant and has a positive diagonal.*

Proof. As $\|A\|_\infty \leq \frac{1}{2} < 1$, the inverse of $(I - A)$ exists and can be expressed via the Neumann series

$$(I - A)^{-1} = I + \sum_{k=1}^{\infty} A^k = I + A(I - A)^{-1} \text{ with } \|A(I - A)^{-1}\| \leq \frac{\|A\|_\infty}{1 - \|A\|_\infty} \leq 1.$$

This already proves diagonal dominance for $\|A\|_\infty < \frac{1}{2}$.

To further explore the diagonal dominance of a matrix sum $I + M + R$, where we will use $M = A^m$ and $R = \sum_{k\neq m} A^k$, we bound the gap in the inequality below as

$$\begin{aligned}|1+M_{ii}+R_{ii}|-\sum_{j\neq i}|M_{ij}+R_{ij}| &\geq |1+M_{ii}|-|R_{ii}|-\sum_{j\neq i}(|M_{ij}|+|R_{ij}|)\\ &= |1+M_{ii}|+|M_{ii}|-\sum_{j=1}^{n}(|M_{ij}|+|R_{ij}|)\ . \end{aligned}\tag{2.2}$$

Thus we get strict diagonal dominance in row i both for $\|M\|_\infty+\|R\|_\infty<1$ and in the case of $\|M\|_\infty+\|R\|_\infty=1$ and $M_{ii}>0$, where the partition $A(I-A)^{-1}=M+R$ can be chosen differently for every $i=1,\ldots,n$.

If $\rho(A)<\frac{1}{2}$, then $(2A)^k$ converges toward zero, so that there is some K with $\|(2A)^K\|\leq\frac{1}{2}$ and thus

$$\sum_{k=1}^{\infty}\|A^k\|\leq\frac{\|A_\infty\|}{1-\|A\|_\infty}\,\frac{1-\|A\|_\infty^K}{1-\|A^K\|_\infty}\leq\frac{1-2^{-K}}{1-2^{-(K+1)}}<1\ ,$$

ensuring strict diagonal dominance of $(I-A)^{-1}$.

In the case $\rho(A)=\frac{1}{2}$ the assumptions of the Wielandt theorem [14, 26] are satisfied, $\rho(A)=\rho(|A|)=\|A\|_\infty$, such that there is a sign s and a signature matrix $\Sigma=\mathrm{diag}(\sigma_1,\ldots,\sigma_n)$ with $|s|=|\sigma_i|=1$, $i=1,\ldots,n$, so that $A=s\,T^{-1}\,|A|\,T$ and thus for the powers of A

$$\left|A^k\right|=\left|\,s^kT^{-1}\,|A|^k\,T\,\right|=|A|^k.$$

The diagonal elements of $|A|^k$ are sums of products over k-cycles of positive elements. Since $|A|$ is irreducible there is at least one k_i-cycle, $k_i\in[n]$, for each diagonal element at position i. Thus we find $(|A|^{k_i})_{ii}=|(A^k)_{ii}|>0$. For the square of that power we note that the diagonal element satisfies the identity

$$\left|\sum_{j=1}^{n}(A^{k_i})_{ij}(A^{k_i})_{ji}\right|=\left(\left|A^{2k_i}\right|\right)_{ii}=\left(\left|A^{k_i}\right|^2\right)_{ii}=\sum_{j=1}^{n}(|A^{k_i}|)_{ij}(|A^{k_i}|)_{ji}\ .\tag{2.3}$$

By the triangle inequality, the identity of the leftmost and rightmost terms is only possible if all the terms in the sum on the left have the same sign. As $(A^{k_i})_{ii}^2>0$, all those terms are positive and consequently $(A^{2k_i})_{ii}>0$, which proves diagonal dominance in row i by setting $M=A^{2k_i}$ and $R=\sum_{m\neq 2k_i}A^m$ in the separation inequality (2.2). This can be done for any index thus proving overall diagonal dominance of $(I-A)^{-1}$. □

Proof. (**Theorem 2.3**) (1) and (2): Let $A\in M_n(\mathbb{R})$ with $\|A\|_\infty<\frac{1}{2}$ or irreducible with norm $\|A\|_\infty\leq\frac{1}{2}$. Moreover, assume set $\Sigma_z\equiv\Sigma$. Then $(I-A\Sigma)^{-1}$ is strictly diagonally dominant by Lemma 2.4 since both irreducibility and the norm constraint are invariant under scalings of A with a signature matrix. Hence, $z_i=e_i^\intercal(I-A\Sigma)^{-1}b$ will adopt the sign of b_i for all $i\in\mathcal{I}^b_{\max}$.

(3): See [18].

(4): The proof is performed inductively. The (2×2)-case can be verified by brute force computation, which we omit for the sake of readability. (We note that it is the only part of the proof that makes use of the symmetry of A.) Now assume the statement of the theorem holds for an $N\geq 2$, but the tuple (A,z,b) contradicts it in dimension $N+1$. We restate two observations from the proof of [18, Thm. 3.1.3]:

- Since $\|A\|_\infty< 1$, it is $\operatorname{sign}(z_i) = \operatorname{sign}(b_i)$ for all $i \in [n]$, if $|z_1| = \cdots = |z_n|$. Thus we may assume that not all entries of z have the same absolute value.
- Define $\mathcal{I}^z_{\max}$ analogously to $\mathcal{I}^b_{\max}$. If $\|A\|_\infty< 1$ and $i \in \mathcal{I}^z_{\max}$, we have $\sum_j |a_{ij}z_j| < |z_i|$ and hence $\operatorname{sign}(b_i) = \operatorname{sign}(z_i)$. Consequently, if there existed a tuple (A, z, b) which violated the claim of the theorem, for any $i \in \operatorname{Eq}(A, b, z)$ we would have $i \notin \mathcal{I}^z_{\max}$.

As $N + 1 \geq 3$, we can organize the system without loss of generality such that $N \in \mathcal{I}^z_{\max}$, while the last row does not hold the (only) contradiction. Then there exists a scalar $\zeta \in [0, 1]$ such that

$$\zeta \cdot |z_N| = |z_{N+1}| \quad \Longrightarrow \quad \zeta \cdot a_{j,N+1} \cdot |z_N| = a_{j,N+1} \cdot |z_{N+1}|.$$

Let $A \in \mathrm{M}_n(\mathbb{R})$ and denote by $A_{a,b}$ a matrix in $M_{n-1}(\mathbb{R})$ that is derived from A by eliminating its a-th row and b-th column. Then

$$\bar{A} \;\equiv\; A_{N+1,N+1} + \operatorname{diag}_N(0, \ldots, 0, \zeta a_{N,N+1})$$

is still tridiagonal with $\|A\|_\infty< 1$ and $|A|$ symmetric. Accordingly, for $\bar{z} \equiv (z_1, \ldots, z_N)^\intercal$ and $\bar{b} \equiv (b_1, \ldots, b_N)^\intercal$ we have

$$\bar{z} + \bar{A}|\bar{z}| \;=\; \bar{b}.$$

Hence, the tuple $(\bar{A}, \bar{z}, \bar{b})$ contradicts the induction hypothesis for dimension N. □

3. Signed Gaussian Elimination

If one is sure of the sign σ_k of z_k one can remove this variable from the right hand side of the AVE. Let A_{*k} denote the k-th column $Ae_k = (A_{jk})_j$ and A_{j*} the j-th row $e_j^\intercal A$. Then the removal of the variable is reflected in the formula

$$(I - A_{*k}e_k^\intercal\sigma_k)z = b + (A - A_{*k}e_k^\intercal)|z|.$$

The inverses of rank-1-modifications are well-known to be (see, e.g. [1])

$$(I - uv^\intercal)^{-1} = I + \frac{1}{1 - v^\intercal u}uv^\intercal \tag{3.1}$$

so that it is easy to remove the matrix factor on the left side. We then have

$$z = \bar{b} + \bar{A}|z| \;, \tag{3.2}$$

where

$$\bar{b} = b + \frac{1}{1 - A_{kk}\sigma_k}\sigma_k A_{*k}\, b_k \tag{3.3}$$

and

$$\bar{A} = A_{red} + \frac{1}{1 - A_{kk}\sigma_k}\sigma_k A_{*k}\,(A_{red})_{k*} \;, \tag{3.4}$$

with

$$A_{red} = A - A_{*k}e_k^\intercal = A(I - e_ke_k^\intercal) \;. \tag{3.5}$$

In Python this can be achieved as:

Function 1. Sign Controlled Elimination Step

```
def elim(A,b,k,sigk):
sk = A[:,k]*sigk;
A[:,k] = 0;
sk = sk/(1-sk[k])
b = b + sk*b[k];
A = A + sk*A[k,:];
```

This procedure corresponds to one step of Gaussian elimination. Now let $J \subseteq [n]$ be an index set and define

$$\text{bmaxJ} \ \equiv \{i \in J \ : \ |b_i| \geq |b_j| \ \forall j \in J\} \ .$$

Using this convention we can give the pseudocode of a slight modification of the algorithm that was introduced as *signed Gaussian elimination* (SG) in [18]:

Algorithm 2. Signed Gaussian Elimination

```
sge(A,b):
set J = [n];
while (#J > 1) do:
determine bmaxJ;
forall k in bmaxJ set sigk = sig(bk);
forall k in bmaxJ do elim(A,b,k,sigk);
J = J\ bmaxJ;
endwhile
perform reverse substitution for (I-A)z = b;
return z
```

Theorem 3.1. *Let $A \in \mathrm{M}_n(\mathbb{R})$ and $z, b \in \mathbb{R}^n$ such that* (1.1) *is satisfied. If A conforms to any of the conditions listed in Theorem 2.3, then the signed Gaussian elimination computes the unique solution of the AVE* (1.1) *correctly.*

Proof. It was already noted that criteria (1)-(4) imply the unique solvability of the AVE (Remark 2.2). We may thus focus on proving the correctness of the algorithm:

Theorem 2.3 ensures the correctness of the sign-picks. The conditions listed in Theorem 2.3 are clearly invariant under the (sign controlled) elimination step. Hence, the argument applies recursively down to the scalar level. □

For dense A the SG has a cubic computational cost. For A with band structure it was shown in [18] that the computation has the asymptotic cost of sorting n floating point numbers. Moreover, note that the SG is numerically stable, since $I - A\Sigma$ is strictly diagonally dominant if $\|A\|_\infty < 1$.

For counterexamples which demonstrate the sharpness of the conditions (1)-(3) in Theorem 2.3 with respect to the SG's correctness, see [18]. For the remaining point, consider the identity as A.

4. Full Step Newton Method

In this section we analyze the full step Newton method (FN) which is defined by the recursion

$$z^{k+1} \ = \ (I - A\Sigma_k)^{-1} b \ , \tag{4.1}$$

where $\Sigma_k \equiv \Sigma_{z^k}$. The iteration has the terminating criterion

$$z^k = z^{k+1} \ .$$

It was developed in [7] and is equivalent to the semi-iterative solver for the equilibrium problem (1.3) developed in [3]. A first, albeit rather restrictive, convergence result is [7, Prop. 7.2]:

Proposition 4.1. *If $\|A\|_p < 1/3$ for any p-norm, then the iteration* (4.1) *converges for all b in finitely many iterations from any z_0 to the unique solution of* (2.1). *Moreover, the p-norms of both $z_i - z$ as well as $(I - A\Sigma_i)z_{i+1} - b$ are monotonically reduced.*

Moreover, in [7, Prop. 7] convergence was proved for the first two restrictions on A in Theorem 2.3. The following extends this result to the criteria in Theorem 2.3.3-4.

Theorem 4.1. *Let $A \in \mathrm{M}_n(\mathbb{R})$ and $z, b \in \mathbb{R}^n$ such that* (1.1) *is satisfied. If A conforms to any of the conditions listed in Theorem 2.3, then for any initial vector $z^0 \in \mathbb{R}^n$ the full step Newton method* (4.1) *computes the unique solution of the AVE* (1.1) *correctly in at most $n + 1$ iterations.*

Proof. Note that all conditions listed in Theorem 2.3 are invariant under scalings of A by a signature matrix. Now assume that z satisfies the equality

$$z - Az \ = \ b$$

and set $\Sigma \equiv \Sigma_z$. Then, since $\Sigma\Sigma = I$, we have

$$b \ = \ z - A\Sigma\Sigma z \ \equiv z - A'|z| \ ,$$

and A' is still strictly diagonally dominant with $\|A\|_\infty \leq 2/3$. This implies $\mathrm{Eq}(A', b, z)$ is empty. Hence, for all edges (Σ, Σ') of $\mathcal{G}(A, b)$ we have

$$\sigma'_{ii} \ = \ 1 \qquad \text{if} \quad b_i \geq 0 \qquad \text{and} \qquad \sigma'_{ii} \ = \ -1 \qquad \text{else} \qquad \forall \ i \in b_{max} \ .$$

That is, the signs with index in $\mathcal{I}^b_{\max}$ are fixed throughout all iterations. Now assume $i \in \mathcal{I}^b_{\max}$. Then for all $k \geq 1$ we will have $\operatorname{sign}(z_i^k) = \operatorname{sign}(b_i)$. This allows us to rewrite the i-th equation in (1.1) and express the z_i^k as a linear combination of the other z_j^k by the transformations (3.2)-(3.5) of A and b to $\bar{A}$ and $\bar{c}$, respectively, which corresponds to one step of Gaussian elimination. As mentioned in the proof of Theorem 3.1, all restrictions listed in Theorem 2.3.1-4 are invariant under the latter operation, which implies that the argument applies recursively and all signs of z are fixed correctly in at most $n + 1$ iterations. Again, we remark that the conditions in Theorem 2.3.1-4 imply the uniqueness of the solution at which we arrive via the afore described procedure. □

Example 4.2. Let

$$A \equiv \begin{bmatrix} \frac{\varepsilon}{2} & \frac{1+\varepsilon}{2} \\ 0 & \frac{1}{2} \end{bmatrix} \qquad \text{and} \qquad z \equiv \begin{bmatrix} \frac{\varepsilon}{2} \\ 1 \end{bmatrix} \ .$$

Then, for $b \equiv z - A|z|$ we have $b = (-\frac{2+\varepsilon}{4}, \frac{1}{2})^\intercal$. And clearly $|b_1| > |b_2|$, but $\mathbf{sign}(b_1) \neq \mathbf{sign}(z_1)$. This example with $\|A\|_\infty = 1/2 + \varepsilon$ leads the SG astray [18, Prop.5.2]. But an elementary calculation shows that for $n \leq 2$ we have convergence of the FN method if $\|A\|_\infty < 1$ [19]. Moreover, it is easy to see that for the cyclic counterexample presented below, the SG will immediately fix *all* signs correctly and thus compute the solution of the system, while the FN may cycle for certain initial signatures.

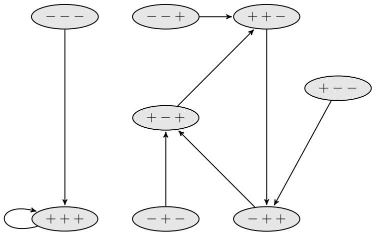

FIGURE 1. Iteration graph of Example 4.3 for $n = 3$.

This demonstrates that, while the correctness and convergence proofs of SG and FN, respectively, are based on mostly analogous constructions, the algorithms are neither equivalent nor does either correctness or convergence range encompass the other.

4.1. **Limitations.** If for all signature matrices the inverse of $(I - A\Sigma)$ is defined, then the iteration (4.1) induces a directed graph with vertex set $\mathcal{S}_n$, where $\Sigma(z^k)$ is connected by an outgoing edge to $\Sigma(z^{k+1})$. We call this graph the *iteration graph* of the tuple (A, b) and denote it by $\mathcal{G}(A, b)$.

Example 4.3. For $n > 2$ set $b = (1, \dots, 1)^\intercal \in \mathbb{R}^n$ and define $A \in \mathrm{M}_n(\mathbb{R})$ as the cyclic Töplitz matrix

$$A = \begin{bmatrix} 0 & 0 & \dots & 0 & a \\ a & 0 & \dots & 0 & 0 \\ 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & 0 \\ 0 & \dots & 0 & a & 0 \end{bmatrix}.$$

In [7] it was shown that, if $a \in \mathbb{R}$ satisfies

$$\frac{1}{2} + \frac{1}{2^n} \;\le\; a \;\le\; \frac{1}{\sqrt{2}},$$

iteration (4.1) cycles between n distinct and definite points when the initial signature Σ contains exactly one negative component and no zeros.

Birthday attack. Consider the following approach at the solution of (2.1): Treat $\mathcal{S}_n$ as an urn; remove a signature Σ_0 uniformly at random, along with all signatures in $\mathcal{G}(A, b)$ that can be reached from Σ. Then repeat the procedure for another random signature Σ_1, and so forth.

If $\mathcal{G}(A, b)$ has a sufficiently small number of connected components, this method yields a high probability of finding one that contains a solution. In cryptology a similar approach at finding collisions of hash functions is known as *birthday attack*, cf. [24]. We denote the probability of "drawing" a solution for (2.1) in this manner with k sign picks by $p^k_{A,b}$. Even for uniquely solvable systems $p^k_{A,b}$ may be insignificantly small for reasonably sized k (in the sense of "polynomial in n").

Proposition 4.2. *There exist irreducible, uniquely solvable systems* (2.1) *such that*

$$p^k_{A,b} \;\le\; \frac{2^{\lfloor \frac{n}{3} \rfloor} + 4\sum_{i=0}^{k-1} i}{2^n},$$

where n is the dimension of the AVE.

Proof. Let $C \in \mathrm{M}_3(\mathbb{R})$ be a cyclic Töplitz matrix as in the counterexample presented above and define $c \equiv (1,1,1)^\intercal$. Then $\mathcal{G}(C,c)$ has 8 vertices and a simple calculation shows:

(1) $\mathcal{G}(C,c)$ has two connected components. 6 of its vertices belong to the subgraph that contains the cycle and 2 vertices belong to the subgraph which contains the solution.
(2) No vector $(I - C\Sigma)^{-1}c$ has a zero component.

Now define the following block diagonal matrix

$$\begin{bmatrix} C & 0 & . & 0 & 0 \\ 0 & C & . & 0 & 0 \\ . & . & . & . & . \\ 0 & 0 & . & C & 0 \\ 0 & 0 & . & 0 & D \end{bmatrix} \equiv \bar{A} \in \mathrm{M}_n(\mathbb{R}),$$

where $D \equiv C$ if $n \bmod 3 = 0$ or else $D \equiv \mathbf{0}$, where $\mathbf{0}$ denotes the zero-matrix in dimension $n \bmod 3$. Moreover, set $b^\intercal \equiv (1,\ldots,1)^\intercal \in \mathbb{R}^n$.

As $\bar{A}$ is clearly reducible, there are no interactions between the subsystems, which means that the number of connected components of $\mathcal{G}(A,b)$ is simply

$$[\#\text{ conn. comp. } \mathcal{G}(C,c)]^{\#\ C-\text{blocks}} \cdot [\#\text{ conn. comp. } \mathcal{G}(D,d)] \ = \ 2^{\left\lfloor \frac{n}{3} \right\rfloor} .$$

Denote by $\mathrm{Sl}(C,c)$ the connected component of $\mathcal{G}(C,c)$ that contains the solution to $c = z + C|z|$. Then, if we pick a signature uniformly at random, the probability of it belonging to the subgraph that contains the solution is

$$\left(\frac{\#\text{ vertices in } \mathrm{Sl}(C,c)}{\#\text{ vertices in } \mathcal{G}(C,c)} \right)^{\left\lfloor \frac{n}{3} \right\rfloor} \ = \ \left(\frac{2}{8} \right)^{\left\lfloor \frac{n}{3} \right\rfloor} .$$

Hence, the probability of a single signature, picked uniformly at random, lying in one of the subgraphs that do not contain a solution is

$$1 - \left(\frac{2}{8} \right)^{\left\lfloor \frac{n}{3} \right\rfloor} \ \geq \ 1 - \frac{2^{\left\lfloor \frac{n}{3} \right\rfloor}}{2^n} .$$

A simple calculation shows that every sign pick removes at most 4 signatures from $\mathcal{S}_n$. For simplicity we assume that the number of removed signatures is always 4. Then we get, by the Bernoulli inequality for variable factors (see, e.g., [16]), that

$$p \ \geq \ \prod_{i=0}^{k-1} \left(1 - \frac{2^{\left\lfloor \frac{n}{3} \right\rfloor} + 4i}{2^n} \right) \geq 1 - \frac{2^{\left\lfloor \frac{n}{3} \right\rfloor} + 4\sum_{i=0}^{k-1} i}{2^n} ,$$

where p is the probability of k picks landing in a subgraph that contains no solution.

Due to observation (2) and the continuity of the matrix inversion $\mathcal{G}(\bar{A},b)$ is stable under perturbations of $\bar{A}$. Hence there exists an irreducible matrix A in a neighborhood of $\bar{A}$ such that $\mathcal{G}(\bar{A},b)$ and $\mathcal{G}(A,b)$ are isomorphic. This concludes the proof. □

4.2. **Sherman-Morrison-Woodbury-Updates.** The iterations of the FN algorithm have the significant practical drawback that in each step the whole updated linear system is inverted, which is rather excessive if only a few signs are updated – and we will see that this is the case for the majority of the steps. Using the the general form of the Sherman-Morrison-Woodbury-formula (SMW)

$$(B + uv^\intercal)^{-1} = B^{-1} - \frac{B^{-1}uv^\intercal B^{-1}}{1 + v^\intercal B^{-1}u} \,, \tag{4.2}$$

which was already cited in a specialized form in (3.1), the iteration costs can be reduced significantly. In the above form the SMW formula has a cost of roughly $3n^2$ multiply-adds for the update of a dense system. The structure of the updates in iteration (4.1) allows to reduce the cost further to roughly n^2 multiply-adds per updated sign: Define

$$B \equiv I - A\Sigma \,.$$

If we update B to $B' \equiv I - A\Sigma'$, where Σ' differs from Σ only in the i-th component, i. e. $\sigma'_i = -\sigma_i$, we can choose u and v in (4.2) as follows

$$u \equiv 2A\Sigma e_i \qquad \text{and} \qquad v \equiv e_i,$$

so that

$$B' = I - A\Sigma + 2A\Sigma e_i e_i^\intercal = I - A\Sigma(I - 2e_i e_i^\intercal)$$

implements the sign change in the ith column of A. Plugging this into the right hand side of (4.2) and using $B^{-1}A\Sigma = B^{-1} - I$ gives:

$$\begin{aligned}(B + uv^\intercal)^{-1} &= B^{-1} - \frac{2[B^{-1}A\Sigma e_i][e_i^\intercal B^{-1}]}{1 + 2e_i^\intercal B^{-1}A\Sigma e_i} \\ &= B^{-1} - \frac{2}{2e_i B^{-1} e_i^\intercal - 1}\left[B^{-1}e_i - e_i\right]\left[e_i^\intercal B^{-1}\right] . \qquad (4.3)\end{aligned}$$

A direct Python implementation of this update for one changed sign matrix is, using $H = B^{-1}$ as the stored inverse and zn$= B^{-1}b$ as the solution according to the current signature:

Function 3. Sherman Morrison Woodbury Update

```
def switch_sign(H,zn, k):
u = H[:,k];
u[k] -= 1;
u /= H[k,k]-0.5;
zn -= u*zn[k]
H -= u*H[k,:]
```

It can easily be seen that the cost of an iteration step is roughly $m \cdot n^2$ multiply-adds plus $m \cdot n$ multiplications, where m denotes the number of updated signs.

5. Piecewise (Linear) Newton Method

If $\|A\|_\infty < 1$ then the map $F(z) \equiv z - A|z| - b$ is bijective (Theorem 2.1 and Remark 2.2). Given a fixed initial point z_0 the pre-image of the ray through $F(z_0)$,

$$R(z_0) \equiv \{z \in \mathbb{R}^n | \exists t \geq 0 : F(z) = tF(z_0)\}$$

is a piecewise linear affine curve where the linear segments are the intersections of $R(z_0)$ with the orthants of $\mathbb{R}^n$. Indeed, if the sign of $z(s) = F^{-1}(sF(z_0))$ is constant $\sigma \in \{\pm 1\}^n$ on a segment $s \in (t - \Delta t, t)$, then with $\Sigma = \operatorname{diag}(\sigma)$ and

$z_\sigma = (I - A\Sigma)^{-1}b$ one has $F(z(s)) = (I - A\Sigma)z(s) - b = sF(z_0)$ for $s \in [t - \Delta t, t]$ and thus

$$tz(s) - sz(t) = (t - s)z_\sigma \iff z(s) = \tfrac{s}{t}z(t) + (1 - \tfrac{s}{t})z_\sigma$$

so that $z(s)$ is a point on the line from $z(t)$ to z_σ. To find a solution to $F(z) = 0$ it is sufficient to follow this curve from $z_0 = z(1)$ to the solution $z(0)$.

5.1. **Exact path-following.** Starting from z_0 one can pass from one end-point z_j of $R(z_0)$ to the next by repeatedly executing

(1) Determine the sign vector σ so that with $\Sigma = \operatorname{diag}(\sigma)$, $z_\sigma = (I - A\Sigma)^{-1}b$ the sign vector in direction $z_\sigma - z_j$ is equal to σ, that is, $\operatorname{sign}(z_j + \varepsilon(z_\sigma - z_j)) = \sigma$ for all sufficiently small $\varepsilon > 0$. If z_j is an inner point of its orthant, then this task is trivial. The end-points z_j of the linear segments will lie in at least one coordinate plane for $j \geq 1$ which still gives a unique choice for the next sign vector. If z_j lies in the intersection of $m > 1$ coordinate planes, then one has to try up to $2^m - 1$ sign combinations to find the correct signature.

(2) Determine the largest $\tau \in (0, 1]$ so that $\operatorname{sign}(z_j + t(z_\sigma - z_j)) = \sigma$ for all $t \in (0, \tau)$.

(3) Set $z_{j+1} = z_j + \tau(z_\sigma - z_j)$. If $\tau = 1$ then obviously $z_{j+1} = z_\sigma$ is the solution of the equation $F(z) = 0$.

To resolve the sign-determination in the first step one could employ a lower-dimensional variant of the signed Gaussian elimination. To be successful the more restrictive assumptions of Theorem 2.3 apply.

5.2. **Systematic Perturbation of the Path.** The fact that the path $R(z_0)$ meets the intersection of multiple coordinate planes is not stable under perturbation of z_0. Instead of changing the initial point one may apply some small perturbations to the computation of the sequence z_j. This can be achieved by different methods. One can restart the iteration from a random perturbation of $z_j + \frac{\tau}{2}(z_\sigma - z_j)$ which has the advantage to lie inside the orthant of signature σ, requiring no immediate updates of the inverse $(I - A\Sigma)^{-1}$. Another way to introduce some quasi-randomness systematically is to over-shoot the orthant boundary and select z_{j+1} from the extension of the current segment into the next orthant. This has the additional advantage that the signature determination in the first step is always trivial.

(1) Compute the signature $\sigma = \operatorname{sign}(z_j)$, set $\Sigma = \operatorname{diag}(\sigma)$ and $z_\sigma = (I - A\Sigma)^{-1}b$.

(2) Determine $0 < \tau_1 < \tau_2 < 2$ so that $\operatorname{sign}(z_j + t(z_\sigma - z_j))$ is constant for $t \in (0, \tau_1)$ and for $t \in (\tau_1, \tau_2)$.

(3) Set $\tau = \lambda \min(1, \tau_1) + (1 - \lambda)\min(1, \tau_2)$ and $z_{j+1} = z_j + \tau(z_\sigma - z_j)$ with some constant $\lambda \in (0, 1)$. If $\tau = 1$ then $z_{j+1} = z_\sigma$ is the solution of the AVE.

In pseudocode this is captured by

Function 4. Piecewise Newton Method

```
pnm(A,b,z):
H = linalg.inverse( I - A*diag(sig(z)) )
zn = H*b
while norm(z-zp) > eps*norm(z):
dz = z - zp;
determine tau from z, dz;
set z = z + tau*dz
determine in idx the sign change indices
```

```
forall k in idx do switch_sign(H,zp,k);
endwhile
return z
```

The motivation of a small perturbation of path $R(z_0)$ would demand that $\lambda > 0$ is a small parameter. Experiments show that the number of steps to the solution does not significantly differ if λ is varied in the interval $(0.1, 0.9)$.

5.3. **Remarks on non-contractive A.** In case that the sign-real spectral radius of A is ≤ 1, one loses the bijectivity of F and possibly also its surjectivity. One consequence is that the set $R(z_0) = \{z | \exists t \geq 0 : F(z) = tF(z_0)\}$ may have multiple components, and the component of z_0 may contain no root of F. For instance, descending branches may meet at some orthant boundary going in with no outgoing branch from that point. This situation is locally stable, a small value of λ will lead to a sequence that oscillates along the orthant boundary for some iterations. Using a large value of λ gives better chances to jump out of this trap.

5.4. **Remarks on SMW-Updates.** Let J and J' be the Jacobians that are active on the orthants P and P' which are identified by the signatures σ and σ', respectively. Moreover, denote by k the number of signs wherein σ and σ' are not equal. Then the number of orthants that intersect in $P \cap P'$, the common facet of P and P' equals 2^k and we have

$$\dim(P \cap P') \ = \ n - k \ .$$

The exact path following algorithm, in a traversal from P to P', requires precisely k SMW-updates as discussed in Section 4.2 to transform J into J'. The algorithm based on systematic perturbation (ideally) requires only one.

6. Support Vector Machines and XOR-Problem as AVE

The adaptation of Support Vector Machines to a given data set is a quadratic programming problem with linear inequality constraints. The associated KKT system can be formulated as a Linear Complementarity Problem with a P-matrix or as a Kojima system, which can, in turn, be reformulated as an Absolute Value Equation. By Theorem 2.1.(2) this problem has a sign-real spectral radius < 1 and thus a unique solution. However, the stronger conditions for correctness of the signed Gaussian elimination are not necessarily satisfied. Hence, this class of problems allows to construct interesting test cases for the AVE solvers. We will use 2-dimensional data sets for which the solutions can be visualized.

6.1. **Support Vector Machines.** Support vector machines are functions which separate two point sets of data in some space $\mathbb{R}^n$ and thus are building blocks for classification algorithms. The separating function for a SVM has the form

$$S(w, x) = \langle w, \phi(x) \rangle$$

where $\phi : \mathbb{R}^n \to H$ is a function into some Hilbert space H, $w \in H$ is the vector representing the parameters of that function class and $\langle \cdot, \cdot \rangle$ is the scalar product on H.

To separate two disjoint point sets A and B the value -1 is assigned to A and 1 to B so that samples can be combined as pairs (x, y) with $y = -1$ for $x \in A$

and $y = 1$ for $x \in B$. Instead of demanding $f(w, x) = y$ resp. $yf(w, x) = 1$ one demands the weakened condition

$$y\, S(w, x) \geq 1$$

so that the sets are separated by the pre-image of $[-1, 1]$.

Common examples for Hilbert spaces and embeddings include:

- Linear functions: $H = \mathbb{R}^{n+1}$, $\phi(x) = (1, x)$, $w = (b, \bar{w})$ and $S(w, x) = b + \bar{w}^\intercal x$, so that both sets are on the opposite sides of the plane $\langle w, x\rangle = -b$ with a minimal distance of $\frac{1}{\|w\|}$ from the plane. This implies the restriction that not only the sets A, B but also their convex hulls have to be disjoint.
- Gauss kernels: $H = L^2(\mathbb{R}^n)$, $\phi(x) = \left[z \mapsto \exp\left(\frac{\|z-x\|}{2\sigma^2}\right)\right]$ with some locality constant σ. Here the resulting approximations of the sets A and B are blobs around balls of radius $\sim \sigma$ about the sampling points.
- Polynomial functions: $H = \mathbb{R}^{|E|}$, $\phi(x) = (x^\alpha : \alpha \in E)$ where $E \subset \mathbb{N}^n$ is a set of multi-indices α enumerating the monomials $x^\alpha = x_1^{\alpha_1} \cdot \cdot x_2^{\alpha_2} \cdots x_n^{\alpha_n}$. Using a scaled euclidean scalar product $\langle u, v\rangle = \sum_{\alpha \in E} c_\alpha u_\alpha v_\alpha$ results in the polynomial function $S(w, x) = \sum_{\alpha \in E} c_\alpha w_\alpha x^\alpha$. The separating set $S(w)^{-1}([-1, 1]) = \{x : S(w, x) \in [-1, 1]\}$ has a flexible shape, increasingly so with increasing degree.

6.2. **The Learning Task.** Given are samples (x_k, y_k), $k = 1, \ldots, N$ with $y_k = \pm 1$ which fall into one of two disjoint sets A and B. To select a parameter vector w among the admissible ones for a given sample

$$\{w \in H : y_k f(w, x_k) \geq 1 \quad \forall k = 1, \ldots, N\}\ ,$$

where each inequality defines a half-space in H, one can choose the norm minimal vector. This leads to the optimization problem

$$\min f(w) = \frac{1}{2}\|w\|^2 \quad \text{such that} \quad g_k(w) = 1 - y_k f(w, x_k) \leq 0\ .$$

At the minimizer, if it exists, some of the constraints will be exactly satisfied. These are the active constraints. Let I_0 be the set of active indices, then the minimizer takes the form

$$w^* = \sum_{k \in I_0}^{N} \lambda_k\, y_k \phi(x_k)\ ,$$

where the Lagrange multipliers λ_k satisfy the linear system

$$1 = \sum_{j \in I_0} y_k y_j \left\langle \phi(x_k), \phi(x_j)\right\rangle \lambda_j, \qquad \forall k \in I_0.$$

As the solution of the adaptation task only depends on this restricted set of sampling points, these are called the *support vectors* of the separating function resp. SVM.

6.3. **Kernel Function.** Using the solution of the adaptation task, as derived in the previous section, both the computation of the Lagrange multipliers and the evaluation of the separating function can be reduced to the knowledge of the kernel function $K(x, y) = \langle \phi(x), \phi(y) \rangle$,

$$\begin{aligned} S(w^*, x) &= \sum_{k \in I_0} \lambda_k y_k \, K(x_k, x) \\ 1 &= \sum_{j \in I_0} y_k y_j \, K(x_k, x_j) \, \lambda_j, \qquad \forall k \in I_0. \end{aligned}$$

For applications it is thus desirable to have a simple evaluation for this kernel function on $\mathbb{R}^n \times \mathbb{R}^n$ that no longer involves the Hilbert space it was defined on.

The given examples have this property. No reduction is necessary in the case of *linear functions*, while for the *Gauss kernels* one has

$$\begin{aligned} K(x, y) &= \int_{\mathbb{R}^n} \exp\left(-\frac{\|z - x\|^2 + \|z - y\|^2}{2\sigma^2} \right) \\ &= \exp\left(\frac{\|x - y\|^2}{4\sigma^2} \right) \sigma^n \left(\int_{\mathbb{R}} e^{-s^2} ds \right)^n \\ &= (\sqrt{2\pi}\sigma)^n \exp\left(-\frac{\|x - y\|^2}{4\sigma^2} \right) . \end{aligned}$$

The separation function is a linear combination of radial basis functions where the parameter σ determines the width of the basis function. This radius should be considered relative to the geometric dimensions of the data set. Thus for very small σ the classification sets $S(w^*, x) \geq 1$ and $S(w^*, x) \leq -1$ will consist of small circles around the support vectors, which requires that almost all data points are support vectors. With increasing σ the number of required support vectors reduces. Very large values of σ will lead to numerical instabilities. The results seem robust for values of σ in the range of the diameter of the data set.

In the case of the *polynomial functions* one can impose the additional condition of rotational invariance which gives the kernel as a function of $\langle x, y \rangle$. One simple expression of that form is

$$K(x, y) = (R^2 + \langle x, y \rangle)^d \ ,$$

which corresponds to the set of all multi-indices of total degree up to d. A higher degree in principle increases the separability of the images of the data sets. The parameter R again allows to scale the kernel function according to the geometric extends of the data set. Both very small and very large values lead to ill-conditioned matrices and numerical instabilities in the evaluation of the separation function. Further, the workable range of medium sized values depends on the degree.

6.4. **KKT System and Kojima Function.** The resulting KKT system for the full Lagrangian

$$L(w, \lambda) = \frac{1}{2}\|w\|^2 + \sum_{k=1}^{N} \lambda_k g_k(w), \quad g_k(w) \leq 0 \ ,$$

with minimizer at $w(\lambda) = \sum_{k=1}^{N} \lambda_k y_k \phi(x_k)$ leads to an LCP

$$\begin{aligned} -g(w(\lambda)) &= K\lambda - \mathbb{1} \geq 0 \\ \lambda &\geq 0 \\ \lambda^\mathsf{T}(-g(w(\lambda))) &= \lambda^\mathsf{T}(K\lambda - \mathbb{1}) = 0 \ , \end{aligned}$$

or, equivalently, a Kojima system

$$\begin{aligned} 0 &= \nabla f(w) + \sum_k \lambda_k^+ \nabla g_k(w) = w - \sum_k \lambda_k^+ \, y_k \phi(x_k) \\ \lambda_k^- &= g_k(w) \qquad\qquad\qquad\quad = 1 - \sum_j y_k y_j \langle \phi(x_k), \phi(x_j) \rangle \, \lambda_j^+ \ , \end{aligned}$$

where $\lambda_k^+ = \max(0, \lambda_k) = \frac{\lambda_k + |\lambda_k|}{2}$ and $\lambda_k^- = \min(0, \lambda_k) = \frac{\lambda_k - |\lambda_k|}{2}$. After elimination of w this has the compact form

$$\begin{aligned} & \lambda^- = \mathbb{1} - K\lambda^+ \\ \iff & \lambda - |\lambda| = 2 \cdot \mathbb{1} - K(\lambda + |\lambda|) \\ \iff & 2 \cdot \mathbb{1} = (I+K)\lambda - (I-K)|\lambda| \ , \end{aligned}$$

which is one inversion away from the form of an AVE, i.e.

$z = b + A|z|$ with $b = 2 \cdot (I+K)^{-1}\mathbb{1}$ and $A = (I+K)^{-1}(I-K) = -I + 2(I+K)^{-1}$.

As K is at least positive semi-definite by construction, the spectrum of the symmetric matrix A is contained in $(-1, 1)$. Additionally, $K = (I-A)^{-1}(I+A)$ is a P-matrix. By the general theory of LCP this means that the solution of the system exists and is unique (Theorem 2.1).

7. Experiments

The majority of this section will deal with the presentation of the performance data of SG, FN and PN (without randomization) when applied to several instances of the AVE-formulation of the XOR-problem. Before that we give a brief account of how they cope with randomly generated data.

7.1. **Randomly Generated Data.** 500 tuples (A, b) of dimension 2.000 were generated uniformly at random and A was scaled by $1/(\|A\|_\infty + 1/n)$ to achieve unique solvability of the system. The results were rather encouraging.

(1) Signed Gaussian elimination: *All systems were solved.* We remark that this is no surprise since, for random systems, the SG is a most likelihood estimator for the signature of z (cf. [18]).
(2) Full step Newton: *All systems were solved.* The average number of iterations, when started with $\Sigma(b)$ as the initial signature, was approx. 3. Moreover, the number of updated signs never exceeded 20 and was approx. 10 on average. Hence, the computational cost was virtually identical to that of the SG. We remark though that this implies a practical advantage of FN over SG, since a performant implementation of FN can be assembled from refined building blocks, e.g. BLAS equation solvers or the matlab-backslash, while SG requires an implementation from scratch.
(3) Piecewise Newton: *All systems were solved.* In terms of the number of steps, the performance was virtually identical to that of FN.

With random data all three algorithms perform at eye level. However, the structured problems considered below will allow us to differentiate between them indeed.

7.2. **SVM-training-Problem.** In this section we will test SG, FN and PN on two instances of the XOR problem: The classical 2×2 chessboard or XOR (eXclusive OR) pattern as well as a pattern that is loosely shaped like the letter “T”. Both have the property that the two sets can not be separated by a simple line. The following experiments were performed:

- Gaussian and polynomial kernels were applied, the dimension of the polynomial kernels was varied (denoted in the format DX, where ”X” is the degree). The radius of both kernel types (as described in Section 6.4) was varied as well (denoted in the format RX). The influence of the radius is illustrated in the Figures 2 and 3 that have ”interesting” level sets resulting from too small radii on the left and instances of practically useful discriminator functions using few support vectors on the right.
- The iterative solvers, FN and PN, were started with two different best-guess values: The signature of b (denoted below by FN$_\Sigma$ and PN$_\Sigma$, resp.) and the signature of the (possibly incorrect) solution of the signed Gaussian elimination (based on the hope that the SG might pick a significant number of correct signs before diverging from the proper signature; FN$_{SG}$ and PN$_{SG}$, resp.). Convergent iterations were recorded in the format $X(Y, Z)$, where X denotes the number of iterations, Y the number of iterations where the number of updated signs is larger $n/3$, i.e., where a full inversion is more efficient than the SMW updates, and Z is the overall number of sign updates during iteration steps with less than $n/3$ updated signs, i.e., steps where SMW updates are more efficient than a full inversion.
- For each of the aforementioned system variations, both iterative solvers were started with 100 random initial values. The probability of their convergence was recorded.

We remark that for a proper implementation of the SMW-updates some sort of singularity detection should be implemented to avoid numerical instabilities in case of small denominators in (4.3).

7.2.1. *Chessboard Pattern.* The pattern XOR10 (four 10×10 squares arranged in a chessboard-pattern) results in a system with dimension $n = 400$. Analogously, the system corresponding to the XOR4 pattern has dimension $n = 64$.

XOR4

kernel	Gauss		polynomial	
setting	R2	R6	R6 D4	R6 D8
FN	1.0	1.0	0.0	1.0
PN	1.0	1.0	0.99	1.0

XOR10

kernel	Gauss		polynomial	
setting	R5	R15	R10 D5	R10 D8
FN	0.32	0.0	0.0	0.0
PN	1.0	1.0	0.96	0.97

Table 1. XOR4 and XOR10 pattern: probability of success with randomly generated initial values.

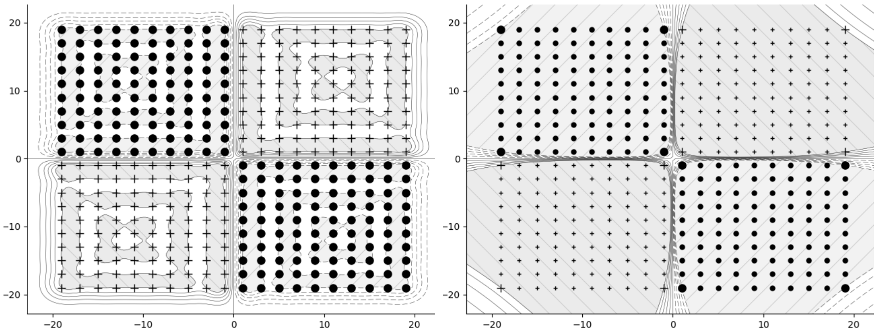

Figure 2. Plot for pattern XOR10 solved with Gaussian kernel using radius 2 and 10 as in Table 1.

	XOR4 R2	XOR4 R6	XOR10 R2	XOR10 R10
SG	√	-	√	√
FN_Σ	1 (-,-)	12 (6,48)	1 (-,-)	-
FN_{SG}	1 (-,-)	2 (-,4)	1 (-,-)	1 (-,-)
PN_Σ	1 (-,-)	57 (-,64)	1 (-,-)	282 (1,280)
PN_{SG}	1 (-,-)	2 (-,4)	1 (-,-)	1 (-,-)

Table 2. XOR pattern with Gaussian kernel.

	XOR4 R6 D4	XOR4 R6 D8	XOR10 R10 D5	XOR10 R10 D8
SG	√	√	√	√
FN_Σ	-	14 (6,51)	-	-
FN_{SG}	1 (-,-)	1 (-,-)	1 (-,-)	1 (-,-)
PN_Σ	78 (2,32)	45 (1,44)	-	290 (1,294)
PN_{SG}	1 (-,-)	1 (-,-)	1 (-,-)	1 (-,-)

Table 3. XOR patterns with polynomial kernel.

7.2.2. *"T"-Pattern.* The "T"-pattern has 65 data points and thus results in a system of dimension $n = 65$.

kernel	Gauss		polynomial	
setting	R2	R6	R6 D5	R6 D8
FN	1.0	1.0	0.0	1.0
PN	1.0	1.0	0.97	1.0

Table 4. "T"-pattern with 100 randomly generated initial values.

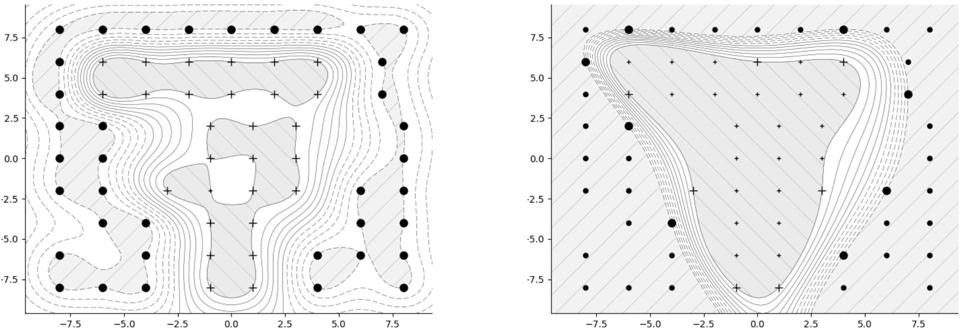

Figure 3. Plot for "T"-pattern solved with Gaussian kernel using radius 2 and polynomial kernel of degree 5 using radius 6 as in Table 4.

kernel	Gauss		polynomial	
setting	R2	R6	R6 D5	R6 D8
SG	√	-	-	-
FN_Σ	1 (-,1)	13 (4,78)	-	19 (6,67)
FN_{SG}	1 (-,-)	12 (4,63)	-	18 (5,72)
PN_Σ	1 (-,1)	58 (-,58)	52 (2,79)	40 (2,45)
PN_{SG}	1 (-,-)	5 (-,5)	40 (-,53)	17 (-,17)

Table 5. "T"-pattern with Gaussian and polynomial kernels.

8. Observations and Final Remarks

SG is an interesting solver from a theoretical point of view due to its fixed cubic cost which roughly equals that of solving a linear system with identical structure. However, as the experiments on the "T"-problem show, both FN and PN seem more robust when confronted with problems that lack strong symmetries. Moreover, SG has the significant practical drawback that an efficient implementation would have to be written from scratch, which is a nontrivial problem in its own right, while the other solvers can, in large parts, be assembled from mature implementations of standard linear solvers.

If they converge and if the SMW-updates are used, FN and PN have, on average, cubic costs as well due to the usually small number of sign updates per iteration step. Of the two iterative solvers, PN is certainly more performant on the investigated problems which were chosen to test out the solvers' limitations. It deserves mentioning though, that FN is highly performant for several well-known system types such as the equilibrium- and (elitist) lasso-problems presented in [3] and [27], respectively, while having the practical appeal of a very simple, straightforward-implementable structure.

The randomized experiments, which either failed or succeeded completely, demonstrate that the initial vector, at least with regard to the present examples, generally does not influence the convergence, but only its speed. So, while the matrix constructed in Proposition 4.2 is generic, the resulting immunity of the system against randomization may also appear in relevant practical problems.

Finally, we remark that one rather favorable property of the investigated triad of algorithms is that among the considered example problems there was not one where neither of them provided a proper solution.

References

1. M.S. Bartlett, *An Inverse Matrix Adjustment Arising in Discriminant Analysis*, Annals of Mathematical Statistics **22** (1951), no. 1, 107–111.
2. C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.
3. L. Brugnano and V. Casulli, *Iterative solution of piecewise linear systems*, SIAM Journal on Scientific Computing **30** (2008), no. 1, 463–472.
4. C.F. Carke, *Optimization and nonsmooth analysis*, Les publications CRM, Universit de Montral, 1989.
5. R.W. Cottle, J.-S. Pang, and R.E. Stone, *The Linear Complementarity Problem*, Academic Press, 1992.
6. A. Griewank, *On stable piecewise linearization and generalized algorithmic differentiation*, Optimization Methods and Software **28** (2013), no. 6, 1139–1178.
7. A. Griewank, J.U. Bernt, M. Radons, and T. Streubel, *Solving piecewise linear equations in abs-normal form*, Linear Algebra and Its Applications **471** (2015), 500–530.
8. A. Griewank, R. Hasenfelder, M. Radons, L. Lehmann, and T. Streubel, *Integrating Lipschitzian Dynamical Systems using Piecewise Algorithmic Differentiation*, Submitted 2016 (2016).
9. A. Griewank, T. Streubel, L. Lehmann, M. Radons, and R. Hasenfelder, *Piecewise linear secant approximation via Algorithmic Piecewise Differentiation*, Submitted 2016.
10. S.-L. Hu, Z.-H. Huang, and Q. Zhang, *A generalized Newton method for absolute value equations associated with second order cones*, Journal of Computational and Applied Mathematics **235** (2011), no. 5, 1490–1501.
11. O.L. Mangasarian, *Absolute value equation solution via concave minimization*, Optimization Letters **1** (2007), no. 1, 3–8.
12. ———, *Absolute value programming*, Computational Optimization and Applications **36** (2007), no. 1, 43–53.
13. ———, *Absolute Value Equation Solution Via Linear Programming*, Journal of Optimization Theory and Applications **161** (2014), no. 3, 870–876.
14. C. Meyer, *Matrix analysis and applied linear algebra*, SIAM, 2000.
15. M.L. Minsky and S. Papert, *Perceptrons*, MIT Press, 1969.
16. D.S. Mitrinovic, *Analytic Inequalities*, Springer, 1970.
17. A. Neumaier, *Interval methods for systems of equations.*, Cambridge University Press, 1990.
18. M. Radons, *Direct solution of piecewise linear systems*, Theoretical Computer Science **626** (2016), 97–109.
19. ———, *Efficient solution of piecewise linear systems*, Master-Thesis, 2016.
20. J. Rohn, *Systems of linear interval equations*, Linear Algebra and Its Applications **126** (1989), 39–78.
21. S.M. Rump, *Theorems of Perron-Frobenius type for matrices without sign restrictions*, Linear Algebra and Its Applications **266** (1997), 1–42.
22. ———, *Perron-Frobenius theory for complex matrices*, Linear Algebra and Its Applications **363** (2002), 251–273.
23. S. Scholtes, *Introduction to piecewise differentiable equations*, Springer, 2012.
24. D. Stinson, *Cryptography: Theory and Practice*, CRC Press, 1995.
25. T. Streubel, A. Griewank, M. Radons, and J.U. Bernt, *Representation and Analysis of Piecewise Linear Functions in Abs-normal form*, Proc. of the IFIP TC 7 (2014), 323–332.
26. H. Wielandt, *Unzerlegbare, nicht negative Matrizen*, Mathematische Zeitschrift **52** (1950), no. 1, 642–648.
27. X.-T. Yuan and S. Yan, *Nondegenerate piecewise linear systems: A finite newton algorithm and applications in machine learning*, Neural Computation **24** (2012), no. 4, 1047–1084.

Institute of Mathematics, Humboldt University of Berlin, Germany

Workgroup Discrete and Algorithmic Mathematics, Technical University of Berlin, Germany
E-mail address: radons@math.tu-berlin.de

Institute for Reliable Computing, Hamburg University of Technology, Germany

Institute of Mathematics, Humboldt University of Berlin, Germany