Received: 7 January 2020 | Revised: 11 June 2020 | Accepted: 18 August 2020

DOI: 10.1002/nla.2335

RESEARCH ARTICLE

WILEY

On the numerical solution of ill-conditioned linear systems by regularization and iteration

Renato Spigler

Department of Mathematics and Physics, Roma Tre University, Roma, Italy

Correspondence
Department of Mathematics and Physics, Roma Tre University, 1, Largo S. Leonardo Murialdo, 00146 Roma, Italy.
Email: spigler@mat.uniroma3.it

Summary

We propose to reduce the (spectral) condition number of a given linear system by adding a suitable diagonal matrix to the system matrix, in particular by shifting its spectrum. Iterative procedures are then adopted to recover the solution of the original system. The case of real symmetric positive definite matrices is considered in particular, and several numerical examples are given. This approach has some close relations with Riley's method and with Tikhonov regularization. Moreover, we identify approximately the aforementioned procedure with a true action of preconditioning.

KEYWORDS

condition number, linear systems, iterative methods, regularization, preconditioning, shift-conditioning

MOS SUBJECT CLASSIFICATION

65F22; 65F10; 65F35; 65F08; 15A12

1 | INTRODUCTION

It is a trivial observation that adding to a given arbitrary $n \times n$ square matrix, A, the matrix αI, where I denotes the $n \times n$ identity matrix, the condition number (in the 2-norm) can be reduced. In fact, asymptotically, $A + \alpha I \sim \alpha I$ as $\alpha \to \infty$, and hence the condition number tends to 1 from above (in any norm such that the identity has norm 1, e.g., any p-norm). Shifting the spectrum of A represents a kind of regularization, indeed approximately a preconditioning strategy for the linear system $Ax = b$,[1] as we will show in Section 3. This procedure, in fact, is strictly connected to the well-known Tikhonov regularization,[2] most often applied to the full-rank overdetermined pseudo-inverse matrix $A^+ := (A^*A)^{-1}A^*$, pertaining to least squares methods, when $A \in \mathbf{R}^{m \times n}$, $m \geq n$.

Shifting may adjust *indefinite* (saddle points)[3] and *quasi-definite*[4] systems (transforming the latter into positive definite systems). Recall that, from the numerical standpoint, almost singular or otherwise ill-conditioned matrices perform equally poorly.

The idea of shifting the spectrum was first proposed, apparently, by J.D. Riley in 1955,[5] to cope with linear systems with real *symmetric positive definite* (SPD) but possibly ill-conditioned matrices. He used a *small* shift parameter α (the regularization parameter), and then followed a perturbative method.

However, aiming at solving $Ax = b$, if we replace A with $A + \alpha I$, with a value of α possibly large (in order to reduce appreciably the condition number of A), we must then recover the solution x to the original problem, for example, by resorting to some iterative procedure. Below, we will show that this is a viable approach, the underlying spectral radius being always less than 1. The subject of iterative methods to solve (large) linear algebraic problems is very broad (see, e.g.,

Dedicated to Gil Strang, on his 85th Birthday

References 6-11 for classical results, and Reference 12 for a nice review up to the year 2000), and here we will try to adapt to our purpose only few well-known schemes.

We will focus in particular on real SPD matrices, and will consider also the more general strategy of replacing A with $A + \mathrm{diag}(\alpha_1, \ldots, \alpha_n)$.

Here is the plan of the article. In Section 2, we describe the general idea of the method and apply it to the SPD case shifting the spectrum by a matrix as αI. Several iterative methods are considered to recover the original solution to $Ax = b$. The simplest fixed point iteration is shown to coincide essentially with Riley's method.[5] A possible way to accelerate the convergence of the latter is discussed. In Section 3, we consider the more general modification of the matrix A, obtained adding to it the diagonal matrix $\mathrm{diag}(\alpha_1, \ldots, \alpha_n)$. Again, different kinds of iterative methods can be adopted. In Section 4, we show that shifting the spectrum may be viewed approximately as a true preconditioning operation, at least when α is sufficiently large. Several numerical examples are given in Section 5, and a short summary concludes the article in Section 6.

2 | SHIFTING THE SPECTRUM: GENERALITIES

Throughout the article we will consider the 2-norm (operator, or spectral norm), and hence the condition number $\kappa(A) := \|A\|_2 \cdot \|A^{-1}\|_2$. When A is SPD,

$$\kappa(A) = \frac{\lambda_M}{\lambda_m}, \tag{1}$$

where λ_M and λ_m are the largest and the smallest eigenvalue of A, $\lambda_M \geq \lambda_m > 0$ (we assume that $\lambda_M > \lambda_m$). Hereafter, we will denote with the subscript M the largest quantity of a given set of numerical values, and with the subscript m the smallest one.

2.1 | Improving conditioning by shift

Setting, for short,

$$A_\alpha := A + \alpha I, \tag{2}$$

we have, for every $\alpha > 0$,

$$\kappa(A_\alpha) = \frac{\lambda_M + \alpha}{\lambda_m + \alpha} < \frac{\lambda_M}{\lambda_m}. \tag{3}$$

Moreover,

$$\frac{\lambda_M + \alpha}{\lambda_m + \alpha} \to 1 \text{ (strictly decreasing)} \quad \text{as } \alpha \to +\infty. \tag{4}$$

But how big should α be chosen in order to conveniently reduce the condition number of A? (and to which value?) Writing for short $\kappa_\alpha := \kappa(A_\alpha) = \frac{\lambda_M+\alpha}{\lambda_m+\alpha}$, hence $\kappa_0 = \frac{\lambda_M}{\lambda_m}$, and claiming that

$$\frac{\kappa_\alpha}{\kappa_0} = \frac{1 + \frac{\alpha}{\lambda_M}}{1 + \frac{\alpha}{\lambda_m}} \leq q \tag{5}$$

for some (suitably small) number q, $q < 1$, to be chosen, we should require that

$$q > \kappa_0^{-1} = \frac{\lambda_m}{\lambda_M}, \quad \text{and} \quad \alpha \geq \frac{1-q}{q\kappa_0 - 1}\, \kappa_0 \lambda_m = \frac{1-q}{q\kappa_0 - 1}\, \lambda_M. \tag{6}$$

Typically, if κ_0 is large, we would like to choose q so small that $q\kappa_0$ might be of order of few units. Thus,

$$\alpha \gtrapprox \frac{\lambda_M}{q\kappa_0 - 1}, \tag{7}$$

that is, an α larger than a fraction of λ_M.

2.2 | Iterative procedures: Effect on the condition number

Shifting the spectrum of A as proposed above leads to a condition number as close to 1 as one wishes, provided that a sufficiently large value of α is chosen. However, if we are interested in solving the linear system $Ax = b$, this treatment leads to a system far from the original one. One way to recover the solution of the original problem is to adopt some iterative procedure.

2.2.1 | Direct iteration

Being $Ax = b$ and $A_\alpha := A + \alpha I$, we have

$$A_\alpha x = Ax + \alpha x = b + \alpha x, \tag{8}$$

and thus we may merely consider the iterative scheme

$$A_\alpha x^{k+1} = b + \alpha x^k, \tag{9}$$

that is

$$x^{k+1} = \alpha\, A_\alpha^{-1} x^k + A_\alpha^{-1} b. \tag{10}$$

The error e^k between x^k and x satisfies

$$e^{k+1} = B_\alpha e^k, \quad B_\alpha := \alpha\, A_\alpha^{-1} \tag{11}$$

for any given $\alpha > 0$, and hence the convergence of the iterative scheme occurs if and only if the spectral radius $\rho(B_\alpha)$ of the matrix B_α is less than 1, that is,

$$\begin{aligned} \rho(B_\alpha) = \rho(\alpha\, A_\alpha^{-1}) &= \alpha \max_i |\lambda_i(A_\alpha^{-1})| \\ &= \frac{\alpha}{\min_i \lambda_i(A) + \alpha} = \frac{\alpha}{\lambda_m + \alpha} < 1, \end{aligned} \tag{12}$$

being $\alpha > 0$ and A_α a real SPD matrix. Moreover,

$$\rho(B_\alpha) \to 1 \ \text{ (strictly increasing) } \ \text{as } \alpha \to +\infty. \tag{13}$$

Therefore, the iterative method in (10) converges always, but its convergence will be slow whenever λ_m is small, hence for matrices that are almost singular. Clearly, the quasi-singularity is a source of ill condition, though not the only one.

The possibly modest rate of convergence of $\rho(B_\alpha)$ to 1 can be assessed, as well as that of $\kappa(A_\alpha)$, looking at the behavior of $f(\alpha) := \kappa(A_\alpha)$ and $g(\alpha) = \rho(B_\alpha)$ as functions of α.

A kind of "*optimal value*" for α can be obtained setting $f(\alpha) = 1/g(\alpha)$, thus choosing

$$\alpha = \overline{\alpha} := \frac{\lambda_m}{\kappa_0 - 2}, \tag{14}$$

provided that $\kappa_0 > 2$ (i.e., $\lambda_M > 2\lambda_m$). This value is optimal in the sense that taking for α either a larger or a smaller value than $\overline{\alpha}$ would worsen either one of the values of κ_α or of $\rho(B_\alpha)$.

Besides assessing the speed of approach of $f(\alpha)$ and $g(\alpha)$ to 1 as $\alpha \to +\infty$, we are interested in the smallest value of f (which occurs for larger α's) and in the smallest value of g (which occurs instead for small α's). These requirements are in competition. Note that $\kappa(A_\alpha)$ depends on λ_M *and* on λ_m (besides α), while $\rho(B_\alpha)$ depends *only* on λ_m (besides α). Therefore, the former could be large even with a not very small λ_m, while $\rho(B_\alpha)$ could still be much less than 1.

Imposing

$$0 < f(\alpha) - 1 = \frac{(\kappa_0 - 1)\lambda_m}{\lambda_m + \alpha} \leq \varepsilon \tag{15}$$

with $0 < \varepsilon < 1$, we require that

$$\alpha \geq \alpha_1 := \frac{\lambda_m}{\varepsilon}[\kappa_0 - (1 + \varepsilon)], \tag{16}$$

(assuming that $\kappa_0 > 1 + \varepsilon$), while imposing

$$0 \leq g(\alpha) \leq \varepsilon \tag{17}$$

(with the same ε, for simplicity), we should require that

$$\alpha \leq \alpha_2 := \frac{\varepsilon}{1 - \varepsilon}\ \lambda_m. \tag{18}$$

Choosing the same limiting value for the bounds in (16) and (18), $\alpha_1 = \alpha_2$, we obtain the value

$$\varepsilon = \varepsilon^* := 1 - (\kappa_0)^{-1} = 1 - \frac{\lambda_m}{\lambda_M}, \tag{19}$$

and correspondingly we have

$$\alpha = \alpha^* = (\kappa_0 - 1)\lambda_m. = \lambda_M - \lambda_m. \tag{20}$$

Note that if λ_m is small and λ_M is of moderate size, so that A is ill-conditioned because A is almost singular, then $\rho(B_\alpha)$ can be close to 1 (slow convergence of iterations), but α^* may be sufficiently large to make the shift effective. Conversely, if λ_m is not very small and λ_M is large, then $\rho(B_\alpha)$ can be appreciably far from 1 (faster convergence of iterations), while α^* may be relatively large.

Recall that, in order that an iterative method, characterized by the iteration matrix B, attains an error less than ε, at least

$$k = \frac{\log(1/\varepsilon)}{\log(1/\rho(B))} \tag{21}$$

iterations are required. Assuming that α is appreciably larger than λ_m and $\varepsilon = 10^{-p}$, we have for the present case,

$$k = \frac{\log(1/\varepsilon)}{\log\left(1 + \frac{\lambda_m}{\alpha}\right)} = \log_{10} e\ \frac{\log_{10}(\varepsilon^{-1})}{\ln\left(1 + \frac{\lambda_m}{\alpha}\right)} \approx 0.434\ \frac{\alpha}{\lambda_m}\ p \tag{22}$$

iterations. On the contrary, if λ_m is substantially larger than α, we would require only about $0.434\ p/\ln(\lambda_m/\alpha)$ iterations.

The idea to regularize a given ill-conditioned matrix goes back apparently to Riley,[5] who anticipated in 1955 the fundamental 1963 work of Tikhonov.[2,13] We can show that the method that we used above, that is the simplest fixed point iterative method, is essentially equivalent to Riley's method. In fact, Riley's expansion[5] of the solution to $Ax = b$, in our notation, reads

$$x = A^{-1}b = A_\alpha^{-1}b + \alpha A_\alpha^{-2}b + \alpha^2 A_\alpha^{-3}b + \ \ldots\ = y + \alpha A_\alpha^{-1}y + \alpha^2 A_\alpha^{-2}y + \ \ldots\ , \tag{23}$$

where we set $A_\alpha y = b$, while a *repeated* use of the iterative formula in (10) yields

$$x^{k+1} = A_\alpha^{-1} b + \alpha A_\alpha^{-2} b + \alpha^2 A_\alpha^{-3} b + \ \dots \ , \tag{24}$$

which perfectly matches Riley's formula (23). In both cases convergence occurs (only) if inequality (12) holds.

Remark. We can go beyond, trying to *accelerate* the convergence of the algorithm as follows. As the scheme in (10) is suggested by

$$x = A_\alpha^{-1} b + \alpha A_\alpha^{-1} x,$$

we can consider the scheme

$$y^{k+1} = A_\alpha^{-1} b + \alpha A_\alpha^{-2} b + \alpha^2 A_\alpha^{-2} y^k \tag{25}$$

obtained iterating once, as suggested by

$$y = A_\alpha^{-1} b + \alpha A_\alpha^{-2} b + \alpha^2 A_\alpha^{-2} y.$$

The new sequence y^k converges to $y \equiv x$ as $k \to \infty$, and it does it faster since the spectral radius of the corresponding iteration matrix, $C_\alpha := \alpha^2 A_\alpha^{-2}$, is

$$\rho(C_\alpha) = \left(\frac{\alpha}{\lambda_m + \alpha} \right)^2 = (\rho(B_\alpha))^2. \tag{26}$$

Note that the cost we have to pay adopting the scheme in (25) rests on the need to obtain $v := A_\alpha^{-2} b$ in addition to $u := A_\alpha^{-1} b$. The latter task amounts to solve the (well-conditioned) system $A_\alpha u = b$, the former to solve $A_\alpha^2 v = b$. However, this can also be rewritten as $A_\alpha v = A_\alpha^{-1} b = u$, and then only the Cholesky factorization of the SPD matrix A_α is required. The method can be iterated further, at the additional cost of solving successively for z some other, well-conditioned systems like $A_\alpha z = w$, where w has already been computed (but having already made the Cholesky decomposition of A).

2.2.2 | A Jacobi-like iteration

We can consider another iterative method, that is, a Jacobi-like algorithm to solve $A_\alpha x = b + \alpha x$. Writing (as usual) $A = D - (E + F)$, where D is the diagonal matrix "extracted" from A, and $-E$ and $-F$ are its strictly lower and upper triangular parts, we have the scheme

$$(D + \alpha I) x^{k+1} = (E + F + \alpha I) x^k + b,$$

and thus

$$x^{k+1} = B_\alpha x^k + c_\alpha,$$

where

$$B_\alpha := (D + \alpha I)^{-1} (E + F + \alpha I) \tag{27}$$

is the iteration matrix and $c_\alpha := (D + \alpha I)^{-1} b$. Therefore, we will have for the error $e^k := x^k - x$, $e^{k+1} = B_\alpha e^k$, hence

$$e_i^{k+1} = \frac{1}{a_{ii} + \alpha} \left[\sum_{j=1}^{n} (e_{ij} + f_{ij}) \, e_j^k + \alpha \, e_i^k \right] = \frac{\sum_{j=1, j \neq i}^{n} a_{ij} \, e_j^k + \alpha \, e_i^k}{a_{ii} + \alpha}, \tag{28}$$

for every i, $i=1, \dots, n$. Recalling that every SPD matrix must have all its *diagonal entries positive*, it follows that

$$\begin{aligned}\|e^{k+1}\|_\infty &:= \max_i |e_i^{k+1}| = \max_i \frac{|\sum_{j=1, j\neq i}^{k} a_{ij}\, e_j^k + \alpha\, e_i^k|}{a_{ii}+\alpha} \\ &\leq \max_i \frac{\sum_{j=1, j\neq i}^{n} |a_{ij}| + \alpha}{a_{ii}+\alpha}\ \|e^k\|_\infty.\end{aligned}$$

Setting, for short,

$$s_i := \sum_{j\neq i} |a_{ij}|, \quad r_i := \frac{s_i+\alpha}{a_{ii}+\alpha} \tag{29}$$

we can write

$$\|e^{k+1}\|_\infty \leq \max_i \left(\frac{s_i+\alpha}{a_{ii}+\alpha}\right)\ \|e^k\|_\infty \equiv \left(\max_i r_i\right)\ \|e^i\|_\infty. \tag{30}$$

If A is a *strictly diagonally dominant* (sdd) matrix, then $s_i < a_{ii}$, for every i, and hence $r_i < 1$ for every i and thus

$$\|e^{k+1}\|_\infty \leq r\ \|e^k\|_\infty \tag{31}$$

with

$$r := \max_i r_i = \max_i \frac{s_i+\alpha}{a_{ii}+\alpha} < 1. \tag{32}$$

Therefore, the Jacobi-like iterative method converges for every $\alpha > 0$ for every SPD sdd matrix A. Note that strict diagonal dominance is required anyway, but the shift of the quantity α may take into account the *possible smallness* of some of the entries a_{jj}.

Convergence will be fast, however, only if A is *strongly* sdd, that is, if $s_i \ll a_{ii}$ for all i's, but it is *slower* than that of the standard Jacobi method ($\alpha = 0$), since $r_i > s_i/a_{ii}$. To have a good performance, we should have $\alpha/\min_i\ a_{ii}$ small.

It is natural to ask whether the present Jacobi-like method wins over the direct iterative algorithm described in Section 2.2.1. This happens if

$$r_i < \frac{\alpha}{\lambda_m+\alpha}, \quad \text{for all } i,$$

which is true if and only if, for all i,

$$a_{ii} > s_i + \lambda_m, \quad \text{and then} \quad \alpha > \frac{\lambda_m\ s_i}{a_{ii} - (s_i+\lambda_m)}, \tag{33}$$

while the direct iterative method will perform better if either for some i, $a_{ii} \leq s_i + \lambda_m$, or if

$$a_{ii} > s_i + \lambda_m, \quad \text{for all } i \text{ but } \alpha \leq \frac{\lambda_m s_i}{a_{ii} - (s_i+\lambda_m)}.$$

Note that the first condition in (33) amounts to requiring a sufficiently *strong* diagonal dominance.

2.2.3 | A Gauss–Seidel-like iteration

Considering instead a Gauss–Seidel-like method, we have a scheme with the iteration matrix

$$B_\alpha = (D + E + \alpha I)^{-1}(-F + \alpha I).$$

The corresponding spectral radius can be estimated as follows. Being

$$|\det((D+E+\alpha I)^{-1}(-F+\alpha I))| = |\det((D+E+\alpha I)^{-1})|\ |\det(-F+\alpha I)| = \left|\prod_{j=1}^{n}(a_{jj}+\alpha)^{-1}\right| \cdot \alpha^n = \frac{\alpha^n}{\prod_{j=1}^{n}(a_{jj}+\alpha)},$$

which must coincide with $\prod_{j=1}^{n}|\lambda_j(B_\alpha)|$, it follows that

$$(\rho(B_\alpha))^n = \left(\max_j |\lambda_j(B_\alpha)|\right)^n \geq \prod_{j=1}^{n}|\lambda_j(B_\alpha)|,$$

hence

$$\rho(B_\alpha) \geq \frac{\alpha}{\left(\prod_{j=1}^{n}(a_{jj}+\alpha)\right)^{1/n}} =: r. \tag{34}$$

Clearly, $r<1$, and hence the iterative method *may* converge, for any given $\alpha>0$. Note that $r\to 1$ (from below) as $\alpha\to+\infty$, but more precisely,

$$r = \frac{1}{\left(\prod_{j=1}^{n}\left(1+\frac{a_{jj}}{\alpha}\right)\right)^{1/n}} = \exp\left\{-\frac{1}{n}\sum_{j=1}^{n}\log\left(1+\frac{a_{jj}}{\alpha}\right)\right\} > \exp\left\{-\frac{1}{n\alpha}\sum_{j=1}^{n}a_{jj}\right\} > 1-\frac{1}{n\alpha}\sum_{j=1}^{n}a_{jj} > 1-\frac{\lambda_M}{\alpha}, \tag{35}$$

being $\log\left(1+\frac{a_{jj}}{\alpha}\right) < \frac{a_{jj}}{\alpha}$ if $\alpha > \max_j a_{jj}$. We used the fact that $\sum_{j=1}^{n}a_{jj} = \mathrm{tr}(A) = \sum_{j=1}^{n}\lambda_j < n\lambda_M$. Note that $\frac{1}{n}\sum_{j=1}^{n}a_{jj}$ is the *average* value of all diagonal entries, a_{jj}.

The opposite estimate would be more useful. Being $a_{jj}/\alpha>0$, we have

$$r < \frac{\alpha}{\left(\prod_{j=1}^{n}a_{jj}\right)^{1/n}}, \tag{36}$$

and then, for instance, $r<1$ if $\alpha<\min_j a_{jj}$.

3 | ADDING A MORE GENERAL DIAGONAL MATRIX

A more general modification of the given matrix system A, to some extent similar to the previous one, could be realized by changing A into $A+D_\alpha$, with $D_\alpha := \mathrm{diag}(\alpha_1, \alpha_2, \ldots, \alpha_n)$, where the real positive parameters α_i, $i=1, \ldots, n$, do not need to be all equal, and we think of α as the vector of parameters $(\alpha_1, \alpha_2, \ldots, \alpha_n)$. In this case, we cannot say that the spectrum of A is shifted, and there are some differences with respect to the previous case.

3.1 | Effect on the condition number

The condition number of the real SPD matrix $A+D_\alpha$ is now

$$\kappa_\alpha := \kappa(A+D_\alpha) = \frac{\max_j \lambda_j(A+D_\alpha)}{\min_j \lambda(A+D_\alpha)} \leq \frac{\max_j \lambda_j(A)+\max_j(\alpha_j)}{\min_j \lambda_j(A)+\min_j(\alpha_j)} \equiv \frac{\lambda_M+\alpha_M}{\lambda_m+\alpha_m}, \tag{37}$$

where we set $a_M := \max_j a_{jj}$, $a_m := \min_j a_{jj}$, since $\max\lambda(A+B) \leq \max\lambda(A)+\max\lambda(B)$ and $\min\lambda(A+B) \geq \min\lambda(A)+\min\lambda(B)$, for every pair of real symmetric (or Hermitian) matrices, A and B.[14,15] In this case, denoting again

with κ_0 the condition number of A (obtained setting all $\alpha_j = 0$), we see that passing from A to $A + D_\alpha$ the condition number is reduced, that is, $\kappa_\alpha < \kappa_0$, if

$$\frac{\alpha_M}{\alpha_m} < \frac{\lambda_M}{\lambda_m} = \kappa_0. \tag{38}$$

Similarly to the case when all the α_j's are equal, the effect of adding D_σ to A results in strengthening the diagonal dominance of A, if A was already diagonally dominant, or even making it diagonal dominant if it was not. Adding to A a matrix like D_α with possibly different α_j's represents a more careful action in the previous sense, since in so doing we can act on each row of A separately.

3.2 | Effect on the spectral radius of the iteration matrix

Consider again different iterative schemes.

3.2.1 | Direct iteration

Concerning the convergence of the iterative schemes as before, we have, for the "direct" iterative approach,

$$x^{k+1} = (A + D_\alpha)^{-1}\, D_\alpha\, x^k + (A + D_\alpha)^{-1} b,$$

and hence converge occurs if and only if

$$\rho(B_\alpha) < 1, \quad \text{being } B_\alpha := (A + D_\alpha)^{-1}\, D_\alpha = (I + D_\alpha^{-1} A)^{-1}. \tag{39}$$

Note that the last form of B_α reminds the Richardson method with preconditioning matrix D_α. Now, for any pair of real symmetric (or Hermitian) positive definite matrices, say A and B, we have that $\lambda_M(AB) \le \lambda_M(A)\lambda_M(B)$,[15,16] and hence $\lambda_m(AB) \ge \lambda_m(A)\lambda_m(B)$. Therefore,

$$\begin{aligned} \rho(B_\alpha) &= \max_j |\lambda_j((I + D_\alpha^{-1} A)^{-1}| = \frac{1}{\min_j |\lambda_j(I + D_\alpha^{-1} A)|} \\ &= \frac{1}{1 + \min_j |\lambda_j(D_\alpha^{-1} A)|} \le \frac{1}{1 + \frac{\lambda_m}{\alpha_M}} = \frac{\alpha_M}{\alpha_M + \lambda_m}. \end{aligned} \tag{40}$$

This quantity is less than 1, hence the iterative method above converges always. If λ_m is large, choosing, for example, $\alpha_M = \lambda_m$, we can have $\rho(B_\alpha) = 1/2$, and α_M would be large at the same time.

3.2.2 | Jacobi-like iteration

In case we adopt instead a Jacobi-like strategy, we have

$$x^{k+1} = (D + D_\alpha)^{-1}(E + F + D_\alpha)x^k + (D + D_\alpha)^{-1} b, \tag{41}$$

and for the errors,

$$e_i^{k+1} = \frac{1}{a_{ii} + \alpha_i} \sum_{j \ne i} a_{ij} e_j^k + \sum_{j=1}^{n} \alpha_j \delta_{ij} e_j^k, \quad j = 1, 2, \dots, n.$$

Therefore,

$$\|e^{k+1}\|_\infty \leq \left(\max_i \frac{\sum_{j\neq i}|a_{ij}| + \alpha_i}{a_{ii} + \alpha_i} \right) \|e^k\|_\infty,$$

and convergence occurs provided that

$$r := \max_i \frac{\sum_{j\neq i}|a_{ij}| + \alpha_i}{a_{ii} + \alpha_i} < 1. \tag{42}$$

Using the same notation as in (29), (32), we have

$$r = \max_i r_i < 1$$

provided that

$$\frac{s_i + \alpha_i}{a_{ii} + \alpha_i} < 1$$

for all i, that is, if

$$s_i < a_{ii}, \quad \text{for all } i,$$

that is, if the SPD matrix A is sdd. Therefore, for such matrices, the Jacobi-like method described here does converge for every chosen set of numbers $\alpha_i > 0$. The convergence however will be slower than that of the standard Jacobi method (obtained for $\alpha_i = 0$ for every i), but faster (slower) compared with the Jacobi-like iteration with $\alpha_i = \alpha$ for every i, if $\max_i \alpha_i < \alpha$ [$\min_i \alpha_i > \alpha$].

We can accelerate the convergence of the iterations proceeding as in the Remark of Section 2.2.1 for Riley's method, starting from

$$y = B_\alpha^2 y + B_\alpha c_\alpha + c_\alpha, \tag{43}$$

where $B_\alpha := (D + D_\alpha)^{-1}(E + F + D_\alpha)$ and $c_\alpha := (D + D_\alpha)^{-1} b$, which leads to the scheme

$$y^{k+1} = B_\alpha^2 y^k + B_\alpha c_\alpha + c_\alpha, \tag{44}$$

in place of that in (41). Again, the convergence rate will be the square of the previous one, and all comments made in the Remark hold.

Note that the present Jacobi-like algorithms can be implemented *in parallel*, as the standard Jacobi algorithm. We recall that, even when some kind of parallel implementation of the Gauss–Seidel method is made, in general, the parallelized Jacobi method wins.[17]

3.2.3 | Gauss–Seidel-like iteration

Similarly to the case of Section 2.2.3, we consider now, with the same notation, the iteration matrix

$$B_\alpha := (D + D_\alpha + E)^{-1}(-F + D_\alpha),$$

hence

$$|\det(B_\alpha)| = |\det((D + D_\alpha + E)^{-1})| \cdot |\det(-F + D_\alpha)|$$

$$= \left|\prod_{j=1}^{n}(a_{jj}+\alpha_j)^{-1}\right| \cdot \left|\prod_{j=1}^{n}\alpha_j\right| = \prod_{j=1}^{n}|\lambda_j(B_\alpha)| \le \prod_{j=1}^{n}\max_j|\lambda_j(B_\alpha)| = (\rho(B_\alpha))^n.$$

Therefore we obtain (as in Section 2.2.3),

$$\rho(B_\alpha) \ge \left(\prod_{j=1}^{n}\frac{\alpha_j}{a_{jj}+\alpha_j}\right)^{1/n} = \left(\prod_{j=1}^{n}\left(1+\frac{a_{jj}}{\alpha_j}\right)\right)^{-1/n} =: R, \tag{45}$$

and clearly $R<1$. More precisely,

$$R = \exp\left\{-\frac{1}{n}\sum_{j=1}^{n}\log\left(1+\frac{a_{jj}}{\alpha_j}\right)\right\} > 1-\frac{1}{n}\sum_{j=1}^{n}\frac{a_{jj}}{\alpha_k}, \tag{46}$$

cf. (35) and (36).

We can also obtain a double estimate **for** R. Using the inequalities $\log x < \log(1+x) < x$, valid for every $x>0$, we have

$$\exp\left\{-\frac{1}{n}\sum_{j=1}^{n}\frac{a_{jj}}{\alpha_j}\right\} < R < \left(\prod_{j=1}^{n}\frac{a_{jj}}{\alpha_j}\right)^{-1/n} = \left(\frac{\prod_{j=1}^{n}\alpha_j}{\prod_{j=1}^{n}a_{jj}}\right)^{1/n}. \tag{47}$$

Also, we can write

$$\begin{aligned}\rho(B_\alpha) &= \max_j \lambda_j((D+D_\alpha+E)^{-1}(-F+D_\alpha))\\ &\le \max_j \lambda_j((D+D_\alpha+E)^{-1})\ \max_j \lambda_k(-F+D_\alpha)\\ &= \frac{\max_j \lambda_j(-F+D_\alpha)}{\min_j \lambda_j(D+D_\alpha+E)} = \frac{\max_j \alpha_j}{\min_j(\alpha_j+a_{jj})},\end{aligned}$$

and thus the upper bound for $\rho(B_\alpha)$,

$$\rho(B_\alpha) \le \frac{\max_j \alpha_j}{\min_j \alpha_j + \min_j a_{jj}}. \tag{48}$$

3.2.4 | SOR-like iteration

Consider now a SOR-like approach to the solution of system $(A+D_\alpha)x = b + D_\alpha x$, that is, the scheme

$$\begin{aligned}(a_{ii}+\alpha_i)\tilde{x}_i^{k+1} &= b_i - \sum_{j=1}^{i-1}a_{ij}x_j^{k+1} - \sum_{j=i+1}^{n}a_{ij}x_j^k + \alpha_i x_i^k,\\ x_i^{k+1} &= \omega\tilde{x}_i^{k+1} + (1-\omega)x_i^k,\end{aligned} \tag{49}$$

where ω is an acceleration parameter. This leads to the iteration matrix

$$\begin{aligned}B_\alpha(\omega) = [I+\omega(D+D_\alpha)^{-1}E]^{-1} \times\\ \times[(1-\omega)I - \omega(D+D_\alpha)^{-1}F + \omega(D+D_\alpha)^{-1}D_\alpha].\end{aligned} \tag{50}$$

Proceeding as in the standard case (recovered by setting $D_\alpha = 0$), we find the following estimate from below for the corresponding spectral radius,

$$\rho(B_\alpha(\omega)) \geq \left(\prod_{j=1}^{n} |1 - \omega\beta_j| \right)^{1/n}, \tag{51}$$

where we set, for short,

$$\beta_j := \frac{a_{jj}}{a_{jj} + \alpha_j}. \tag{52}$$

Here, $0 < \beta_j < 1$ for all j, being A an SPD matrix, hence $a_{jj} > 0$, and $\alpha_j > 0$ for all j.

As in the standard SOR, a *necessary* condition for the convergence of this iterative scheme is that the right-hand side of (51) be less than 1. A sufficient (not necessary) condition for this is

$$|1 - \omega\beta_j| < 1$$

for all j, that is, that

$$0 < \omega < \frac{2}{\beta_j} = 2\left(1 + \frac{\alpha_j}{a_{jj}}\right) \tag{53}$$

for all j, and hence that

$$0 < \omega < 2\left(1 + \frac{\alpha_m}{a_M}\right), \tag{54}$$

where $\alpha_m := \min_j \alpha_j$ and $a_M := \max_j a_{jj}$. This means that the range imposed to the acceleration parameter ω is larger than in the standard SOR method ($\alpha_j = 0$).

In the special case $\alpha_j = \alpha$ for all j, and of a Toeplitz matrix (i.e., "constant on its diagonal"), thus, in particular, with $a_{jj} = a$ for all j's, sometimes considered in the literature, assuming

$$\beta := \frac{a}{a + \alpha},$$

we have

$$\left(\prod_{j=1}^{n} |1 - \omega\beta| \right)^{1/n} = |1 - \omega\beta|,$$

and the necessary condition on ω becomes

$$0 < \omega < 2\left(1 + \frac{\alpha}{a}\right).$$

If we consider, instead of (49), the scheme

$$\begin{aligned} (a_{ii} + \alpha_i)\tilde{x}_i^{k+1} &= b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{k+1} - \sum_{j=i+1}^{n} a_{ij}x_j^k + \alpha_i x_i^{k+1}, \\ x_i^{k+1} &= \omega\tilde{x}_i^{k+1} + (1 - \omega)x_i^k, \end{aligned} \tag{55}$$

we obtain the iteration matrix

$$B_\alpha := [I - \omega(D + D_\alpha)^{-1}(-E + D_\alpha)]^{-1}[(1 - \omega)I - \omega(D + D_\alpha)^{-1}F)], \tag{56}$$

from which we infer

$$\rho(B_\alpha) = \max_j |\lambda_j(B_\alpha)| = |\det(B_\alpha)| = \prod_{j=1}^{n} \left(\frac{1-\omega}{\beta_j} \right). \tag{57}$$

It is certainly true that $\rho(B_\alpha) < 1$ if, for example,

$$\omega > 1 - \min_j b_j.$$

Proceeding as before, we can also establish the estimate

$$\rho(B_\alpha) \geq \left(\prod_{j=1}^{n} (1 - \omega\beta_j) \right)^{1/n}. \tag{58}$$

4 | SHIFTING AS A TRUE PRECONDITIONER?

Can we obtain a shift of the spectrum by either pre-, or pre- and post-multiplication? That is, changing A into PA or into PAP? Setting

$$PAP = A + D_\alpha, \tag{59}$$

being A (given) and P (to be determined), both real *symmetric* and definite positive, this means trying to realize the shift of A on the right-hand side of (59) through the transformation PAP. In general, this problem has *no solution*. In fact, Equation (59) consists of n^2 nonlinear quadratic scalar equations for the $n(n+1)/2$ scalar unknowns P_{ij} (recall that we look for a *symmetric* matrix P).

Confining to the simpler case $D_\alpha = \alpha I$, we look for an approximate solution, P, when α is large. In fact, when $\alpha \to +\infty$, we have $PAP \sim \alpha I$, which is compatible with being $P \sim \alpha^{1/2} A^{-1/2}$. Here, the square root $S^{1/2}$ of the SPD matrix S, is the principal square root of S, that is the only SPD matrix whose square is S. Setting, for convenience, $\beta = \alpha^{1/2}$, hence $P \sim \beta A^{-1/2}$, and expanding formally P in powers of β^{-1} as

$$P = \beta A^{-1/2} + P_0 + \beta^{-1} P_1 + \beta^{-2} P_2 + \mathcal{O}(\beta^3),$$

we obtain from (59) with $D_\alpha = \alpha I$ the conditions

$$A^{1/2}P_0 + P_0 A^{1/2} = 0, \quad A^{1/2}P_1 + P_0 A P_0 + P_1 A^{1/2} = A,$$
$$A^{1/2}P_2 + P_0 A P_1 + P_1 A P_0 + P_2 A^{1/2} = 0, \quad A^{1/2}P_3 + P_0 A P_2 + P_1 A P_1 + P_2 A P_0 = 0, \quad \text{and so forth.}$$

The first equation has the solution $P_0 = 0$, then the second has the solution $P_1 = \frac{1}{2} A^{1/2}$, the third is satisfied by $P_2 = 0$, and the fourth by $P_3 = -\frac{1}{4} A^{3/2}$, and so forth, so that we have a solution of the form

$$P = \alpha^{1/2} A^{-1/2} + \frac{1}{2} \alpha^{-1/2} A^{1/2} - \frac{1}{4} \alpha^{-3/2} A^{3/2} + \mathcal{O}(\alpha^{-2}). \tag{60}$$

Writing $A = D + C$, D being the diagonal matrix "extracted" from it, and if A is *strongly* sdd so that $\|C\|_2 \ll \|D\|_2$, we have

$$\|(D+C)^{1/2} - D^{1/2}\|_2 \leq \frac{1}{\mu_1 + \mu_2} \|C\|_2,$$

and

$$\|(D+C)^{-1/2} - D^{-1/2}\|_2 = \|(D+C)^{-1} C D^{-1}\|_2 \leq \frac{1}{\gamma_1 + \gamma_2} \|A^{-1}\|_2 \, \|D^{-1}\|_2 \, \|C\|_2,$$

TABLE 1 Condition number of shifted matrices, $\kappa(A+\alpha I_2)$

α	0.5	1	2	4	6	8	10	20	50	100
	443.01	222.99	112.24	56.68	38.13	28.85	23.28	12.14	5.45	3.22

TABLE 2 Spectral radius $\rho(B_\alpha)$ in the direct iteration, see (12)

α	0.5	1	2	20
	0.9911	0.9955	0.9978	0.9998

for some positive numbers μ_1, μ_2, γ_1, γ_2, assuming that $A \geq \mu \leq \mu_1^2 I$, $D \geq \mu_2^2$, $A^{-1} \geq \gamma_1^2 I$, $D^{-1} \geq \gamma_2^2$ [Reference 18, lemma 2.2, p. 219]. Thus, we can write

$$(D+C)^{-1/2} = D^{-1/2} + N, \quad (D+C)^{1/2} = D^{1/2} + M,$$

with

$$\|N\|_2 \leq \frac{1}{\gamma_1 + \gamma_2} \, \|A^{-1}\|_2 \, \|D^{-1}\|_2 \, \|C\|_2,$$

and

$$\|M\|_2 \leq \frac{1}{\mu_1 + \mu_2} \, \|C\|_2.$$

Finally we have

$$P = \alpha^{1/2}(D^{-1/2} + N) + \frac{1}{2} \, \alpha^{-1/2}(D^{1/2} + M) + \mathcal{O}(\alpha^{-3/2})), \tag{61}$$

and hence

$$P \approx \alpha^{1/2} D^{-1/2} + \frac{1}{2} \, \alpha^{-1/2} D^{1/2}. \tag{62}$$

Clearly, α should be sufficiently large and $\|C\|_2 = \|A - D\|_2$ sufficiently small.

In conclusion, the answer to the question raised at the beginning of this section is (roughly) affirmative: shifting the spectrum of a given SPD matrix A, adding αI to it, does represent *approximately* a preconditioning, when α is sufficiently large.

5 | NUMERICAL EXAMPLES

In this section we give some numerical examples to illustrate the methods proposed in the previous sections.

Example 1. We start with two extremely simple matrices, A, correspondingly to which we compute the (spectral) condition number and the spectral radius of $A + \alpha I_2$, for several values of the parameter α. Hereafter we will denote with I_n the $n \times n$ identity matrix, n being the dimension of A.

Example 2. The innocent SPD matrix (not diagonally dominant)

$$A := \begin{pmatrix} 149 & 105 \\ 105 & 74 \end{pmatrix},$$

has determinant equal to 1, and the eigenvalues $\lambda_m = 0.0045$ and $\lambda_M = 222.9955$, hence the spectral condition number $\kappa_0 = 4.9727 \times 10^4$. See Tables 1 and 2.

Example 3. The simple SPD matrix (not diagonally dominant)

TABLE 3 Condition number of shifted matrices, $\kappa(A+\alpha I_2)$

α	0	0.1	1	1.5	2	3	4	5	10	20
	9.24×10^4	25.96	3.49	2.66	2.24	1.83	1.62	1.49	1.24	1.12

TABLE 4 Condition number of shifted Hilbert matrices, $\kappa(\texttt{hilb}(n)+\alpha I_n)$

	α							
		0	2	4	6	8	10	100
n	2	19.28	1.58	1.29	1.19	1.14	1.11	1.01
	4	1.55×10^4	1.75	1.37	1.25	1.18	1.15	1.01
	6	$\mathcal{O}(10^7)$	1.80	1.40	1.26	1.20	1.16	1.01
	8	$\mathcal{O}(10^{10})$	1.84	1.42	1.28	1.21	1.16	1.01
	10	$\mathcal{O}(10^{13})$	1.87	1.43	1.29	1.21	1.17	1.01
	20	$\mathcal{O}(10^{18})$	1.95	1.47	1.31	1.23	1.19	1.01
	50	$\mathcal{O}(10^{19})$	2.03	1.51	1.34	1.25	1.20	1.02
	100	$\mathcal{O}(10^{19})$	2.09	1.54	1.36	1.27	1.21	1.02
	1,000	$\mathcal{O}(10^{20})$	2.22	1.61	1.40	1.30	1.24	1.02

$$A := \begin{pmatrix} 2.4298 & 0.4049 \\ 0.4049 & 0.0675 \end{pmatrix},$$

has determinant about equal to 6.7490×10^{-5}; see Table 3.

The corresponding spectral radius is however very close to 1, since λ_m is very small, see (12).

Example 4. We consider Hilbert matrices, $H_n = \{H_{ij}\}_{i,j=1}^n$, of order n, defined by $H_{i,j} := \frac{1}{i+j-1}$, for $i,j = 1, \ldots, n$; see Reference 19. They are real SPD but not diagonally dominant, and very ill-conditioned. There is a theoretical (growth) estimate for their spectral condition number,

$$\kappa(\texttt{hilb}(n)) = \mathcal{O}\left(\frac{(1+\sqrt{2})^{4n}}{\sqrt{n}}\right), \tag{63}$$

which shows that H_n is exponentially ill-conditioned as its size increases. Examples of $\kappa(H_n + \alpha\, I_n)$ for several values of n and α were computed with MATLAB and are given in Table 4.

Example 5. We consider some examples based on the Vandermonde matrix, $V_n := \{(x_i)^{j-1}\}_{i,j=1}^n$. This is also very ill-conditioned. Indeed, a lower bound for its condition number was found to be exponentially growing with n.[20,21] It is very nasty and not SPD. We generate $M_n := V_n V_n^T$, where V_n is an n-dimensional Vandermonde matrix with randomly chosen points (or "nodes"), x_i, using MATLAB. For the vector of nodes

$$\begin{aligned} v &:= \texttt{rand}(10,1) \\ &= [0.2760, 0.6797, 0.6551, 0.1626, 0.1190, 0.4984, 0.9597, 0.3404, 0.5853, 0.2238]^T, \end{aligned}$$

we computed $A := \texttt{vander}(v)$ and $M_{10} := V_{10} V_{10}^T$, for which we have a very small lowest eigenvalue, of order of $\mathcal{O}(10^{-16})$. Thus we consider the matrix $N_{10} := M_{10} + 10^{-4} I_{10}$, for which $\lambda_m = 0.0001$, and few approximate values of $\kappa(M_{10} + \alpha I_{10})$ are in Table 5.

Similarly, choosing a random 20-dimensional vector with $\texttt{rand}(20,1)$, then $V := \texttt{vander}(v)$, and constructing $A := VV^T$, we obtain a symmetric positive definite matrix with the smallest eigenvalue extremely small. Hence we built $M_{20} := A + 10^{-5} I_{20}$, and few values of $\kappa(A + \alpha I_{20})$ are shown in Table 6.

TABLE 5 Condition number of shifted matrices, $\kappa(N_{10}+\alpha I_{10})$

α	0	0.1	1	2	3	4
	2.77×10^{16}	1.50×10^{5}	16.01	8.50	6.00	4.75

TABLE 6 Condition number of shifted matrices, $\kappa(M_{20}+\alpha I_{20})$

α	0	1	2	3	4	5
	7.88×10^{6}	79.87	40.43	27.29	20.71	16.77

TABLE 7 Condition number of shifted matrices, $\kappa(A+\alpha I_{10})$

α	0	0.3	0.4	0.5	1	2	3	4	5	6	10
	48.37	11.07	8.97	7.60	4.55	2.84	2.24	1.944	1.75	1.63	1.38

TABLE 8 Condition number of shifted matrices, $\kappa(A+\alpha I_{20})$

α	0	0.5	0.6	0.7	1	2	3	5
	4,133	8.98	7.65	6.70	4.99	2.99	2.33	1.79

Example 6. For the eigenvalues of the *tridiagonal* Toeplitz matrix with diagonal elements b, superdiagonal elements a, and subdiagonal elements c, the closed-form explicit expression

$$\lambda_k = b + 2\sqrt{ac}\,\cos\left(\frac{k\pi}{n+1}\right), \quad k = 1, 2\ \ldots\ , n,$$

exists;[22] see also Reference 23. Hence, in the SPD case, it is straightforward to obtain its spectral condition number.

Consider the famous tridiagonal Toeplitz matrix A,

$$A = \begin{bmatrix} -2 & 1 & 0 & 0 & \ldots & \ldots & \ldots & 0 \\ 1 & -2 & 1 & 0 & \ldots & \ldots & \ldots & 0 \\ 0 & 1 & -2 & 1 & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & & & & \\ \vdots & & & & \ddots & 1 & -2 & 1 \\ 0 & \ldots & \ldots & \ldots & \ldots & 0 & 1 & -2 \end{bmatrix}, \tag{64}$$

which is SPD (not sdd). For $n = 10$, we have $\lambda_m = 0.0810$, $\lambda_M = 3.9190$. In Table 7 we show some values of $\kappa(A+\alpha I_{10})$.

For $n = 100$, we have $\lambda_m = 0.0010$, $\lambda_M = 3.9990$, and the results of Table 8.

Example 7. The real SPD *pentadiagonal* (not Toeplitz), *not* ssd matrix A, of dimension $n = 10^{19}$

$$A = \begin{bmatrix} 5 & -4 & 1 & 0 & 0 & \ldots & \ldots & 0 \\ -4 & 6 & -4 & 1 & 0 & \ldots & \ldots & \ldots \\ 1 & -4 & 6 & -4 & 1 & 0 & \ldots & \vdots \\ 0 & 1 & -4 & 6 & -4 & 1 & 0 & \ldots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & 1 & -4 & 6 & -4 & 1 \\ \vdots & & & & 1 & -4 & 6 & -4 \\ 0 & \ldots & \ldots & \ldots & 0 & 1 & -4 & 5 \end{bmatrix} \tag{65}$$

TABLE 9 Condition number of shifted matrices, $\kappa(A+\alpha I_{10})$

α	0	0.1	1	2	3	4	5	10
	2,340	145.06	16.25	8.65	6.10	4.83	4.06	2.53

TABLE 10 Condition number of shifted matrices, $\kappa(A+\alpha I_{1,000})$

α	0	0.2	0.4	0.6	0.8	1	3	4	5	50
	12.99	10.99	9.57	8.49	7.66	6.99	3.99	3.39	2.99	1.23

has $\lambda_m = 0.0066$, $\lambda_M = 15.3585$, and the results of Table 9.

Example 8. The *eptadiagonal*, Toeplitz, SPD, of order 1,000, say A,

$$A = \begin{bmatrix} 5 & 2 & 1 & 1 & 0 & 0 & \dots & \dots & 0 \\ 2 & 5 & 2 & 1 & 1 & 0 & & \dots & \dots \\ 1 & 2 & 5 & 2 & 1 & 1 & & \dots & \dots \\ 1 & 1 & 2 & 5 & 2 & 1 & 1 & \dots & \dots \\ 0 & 1 & 1 & 2 & 5 & 2 & 1 & 1 & \dots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & & & \ddots & & 0 \\ \dots & \dots & \dots & 0 & 1 & 1 & 2 & 5 & 2 \\ 0 & \dots & \dots & & 0 & 1 & 1 & 2 & 5 \end{bmatrix} \tag{66}$$

being *not* sdd, the Jacobi method applied to solve the system $Ax = b$ may not converge. In fact this is the case. We have $\lambda_m = 1.0001$, $\lambda_M = 12.9999$. The condition number of A is moderate but numerical experiments show that the Jacobi method does not converge; see the plot in Figure 1, bottom. Some values of $\kappa(A+\alpha I_{1,000})$ are in Table 10.

Numerical experiments, conducted applying the Jacobi method to solve the system with matrix $A+\alpha I_{1,000}$ showed convergence roughly when $\alpha > 2.041$.

So far, we provided examples of reduction of the condition number obtained increasing the value of the parameter α. Below, we show the errors made in a few cases, using the Jacobi-like solver described in Section 2.2.2, as the number of iterations is increased, for some fixed value of α.

We also *compared* the performance of the Jacobi-like and the Gauss-Seidel-like solvers when the number of iterations is increased, for some fixed value of α, and even using a more general, fixed, diagonal matrix, D_α.

Example 9. Figure 1, top, refers to the simple case of the SPD sdd matrix

$$A = \begin{pmatrix} 9 & 8.99 \\ 8.99 & 9 \end{pmatrix}, \tag{67}$$

whose condition number is $\kappa(A) = 1.799 \times 10^3$. We constructed an example choosing the expected solution $x = (12)^T$, and $x_0 := 0.5\,x$ as the initial point of the iterations, computed $b := Ax$, and tried to recover x by the Jacobi-like method of Section 2.2.2, with $\alpha = 0$, $\alpha = 2$, and $\alpha = 5$. Hereafter, in the MATLAB programs, we set to 0.001 the tolerance we required. The errors, relative to the initial residual, are plotted as a function of the iteration numbers.

In Figure 1, bottom, we did the same for the eptadiagonal Toeplitz SPD matrix of Example 8, whose dimension is $n = 1,000$. Recall that it is *not* sdd.

Example 10. Here we constructed an SPD pentadiagonal matrix, A, of dimension $n = 1,000$, with random entries. The diagonal was obtained using uniformly distributed numbers in the interval $(0, 1)$, provided by the MATLAB code `rand(1,000,1)`, and similarly for the other diagonals. The resulting matrix was sdd, with $\kappa(A) \approx 1.939 \times 10^3$. We also

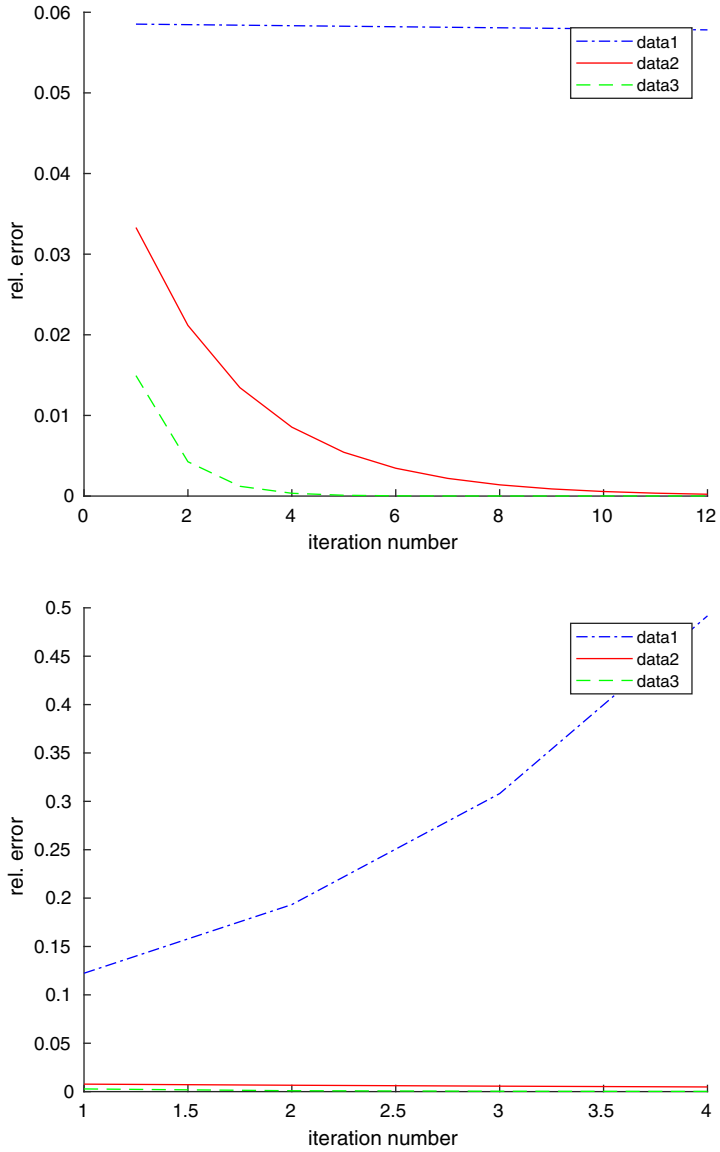

FIGURE 1 The relative errors of the Jacobi-like method for $\alpha = 0$ (data 1), 2 (data 2), and 5 (data 3), are plotted versus the number of iterations, for the simple ill-conditioned 2×2 matrix in (67), top, and the eptadiagonal matrix of dimension $n = 1,000$ of Example 8, bottom (color online)

chose a random vector x (the expected solution), the initial point $x_0 = 0.6\,x$, and the right-hand side $b := A\,x$. In Figure 2, we show the relative errors made using the Jacobi-like (continuous blue line) and the Gauss–Seidel-like (dashed red line) solvers of Section 2.2, versus the iteration number, for the fixed value of α, $\alpha = 2$.

Example 11. We considered the SPD pentadiagonal, not Toeplitz, *not* sdd matrix of Example 7. We chose $x = \texttt{rand}(10, 1)$ (the expected solution), the initial point $x_0 = 0.5\,x$, and computed $b := A\,x$. In Figure 3, we plotted the relative errors made by the Jacobi-like and the Gauss–Seidel-like solvers versus the number of iterations. The four plots (from upper left to lower right) refer to the values $\alpha = 0$, $\alpha = 2$, $\alpha = 5$, and $\alpha = 10$.

Example 12. Here we considered an SPD pentadiagonal matrix, say A, similar to that of Example 7, but with *all* diagonal entries equal to 6, and of much higher dimension, $n = 3,000$. The condition number turns out to be $\kappa(A) \approx 2.5959 \times 10^{12}$, and A is *not* sdd. In Figure 4, we compare the relative errors obtained adopting the Jacobi-like and the Gauss–Seidel-like methods, versus the number of iterations, when (from upper left to lowest): $\alpha = 0$, $\alpha = 0.1$, and $\alpha = 4.2$. Note that the Jacobi method does *not* converge when $\alpha = 0$, and it still does not for $\alpha = 0.1$.

Example 13. We considered the SPD pentadiagonal matrix $B := A + 0.1\,I_{300}$, where A is that of the previous Example 12. The condition number now is $\kappa(B) \approx 160.99$. In Figure 5, the relative errors made by the Jacobi-like and the Gauss–Seidel-like methods are compared, for (from upper left to lower right): $\alpha = 0$, $\alpha = 1$, $\alpha = 2$, and $\alpha = 10$.

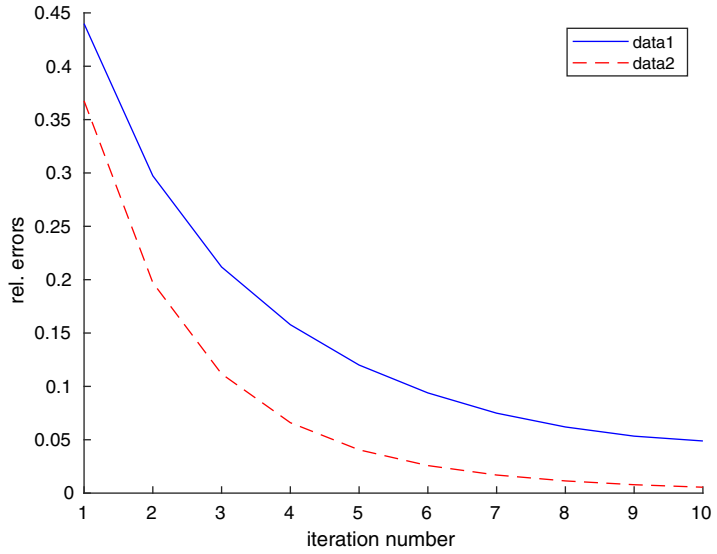

FIGURE 2 Comparing Jacobi-like (data 1, continuous blue line) and Gauss–Seidel-like (data 2, dashed red line) methods for an SPD pentadiagonal matrix of dimension $n = 1,000$, with $\alpha = 2$ (color online)

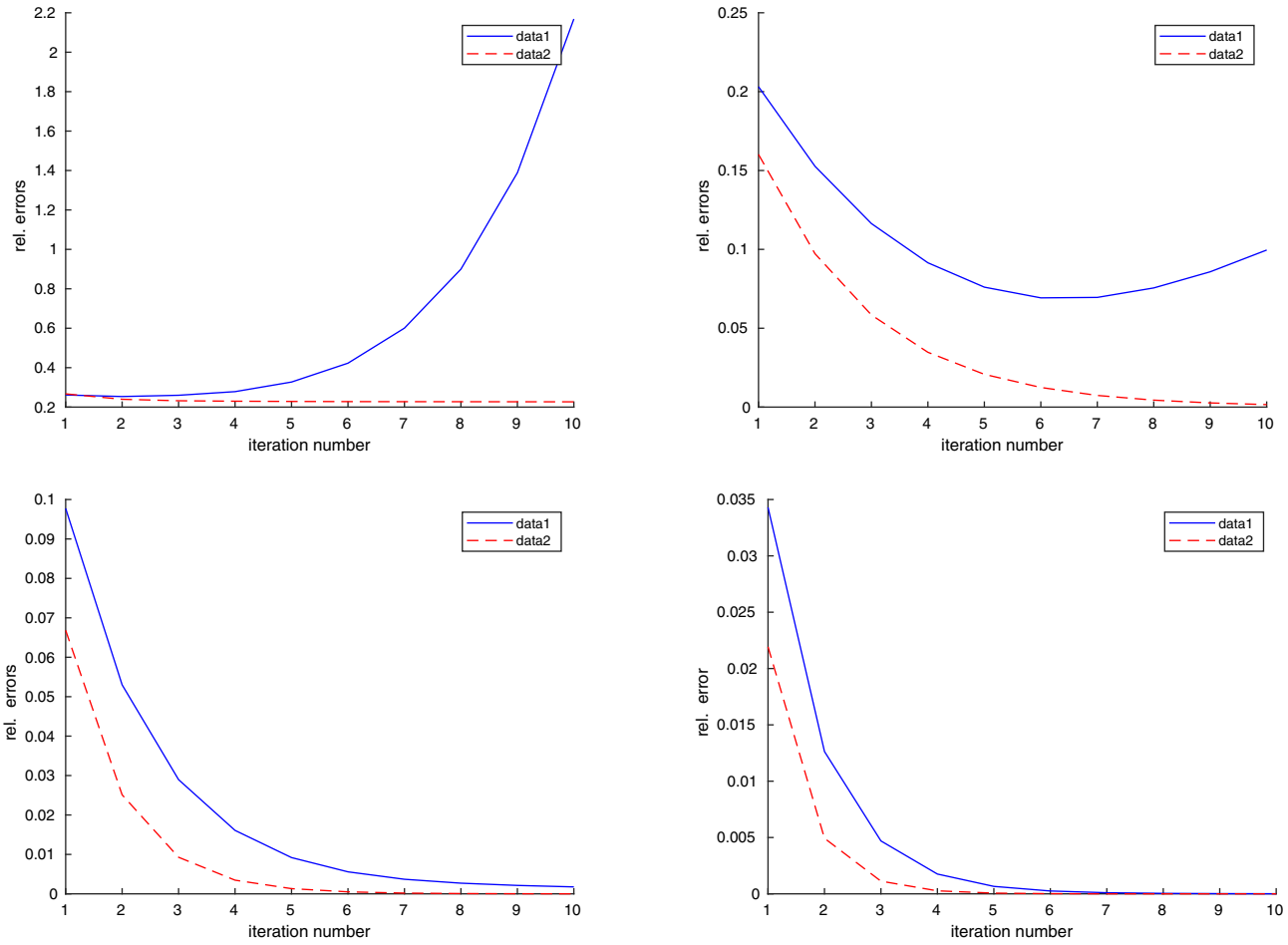

FIGURE 3 Comparing Jacobi-like and Gauss–Seidel-like solvers for the SPD pentadiagonal matrix of Example 7, from upper left to lower right: for $\alpha = 0$, $\alpha = 2$, $\alpha = 5$, and $\alpha = 10$ (color online)

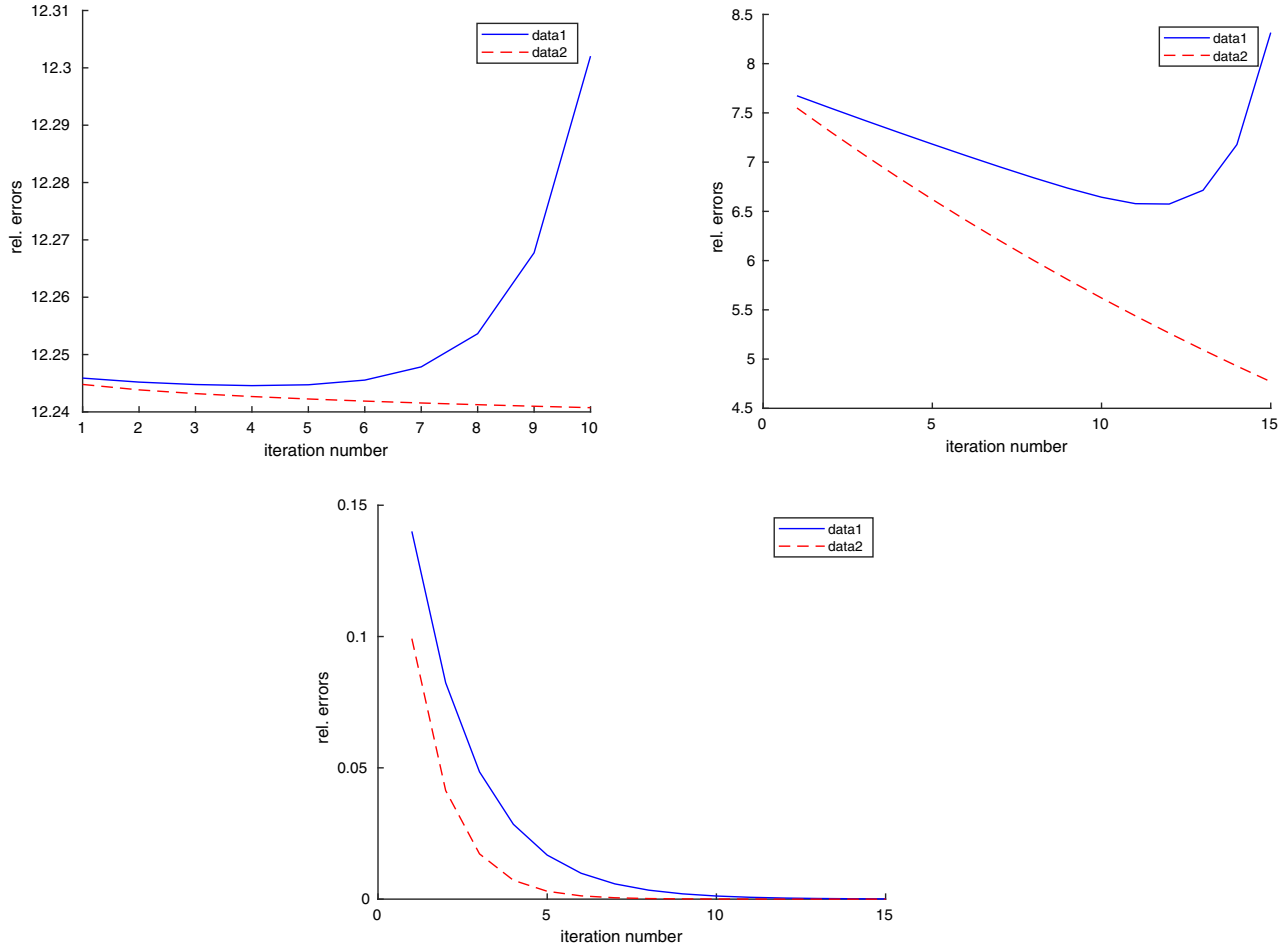

FIGURE 4 Comparing Jacobi-like and Gauss–Seidel-like solvers for an SPD pentadiagonal matrix of dimension $n = 3,000$: from upper left to lowest, for $\alpha = 0$, $\alpha = 0.1$, $\alpha = 4.2$ (color online)

Finally, we give some simple examples of regularization achieved by adding a more general diagonal matrix, as described in Section 3.

Example 14. We considered the two apparently harmless matrices

$$A_3 := \begin{pmatrix} 14 & 32 & 50 \\ 32 & 77 & 122 \\ 50 & 122 & 194 \end{pmatrix}, \quad A_4 := \begin{pmatrix} 30 & 70 & 110 & 150 \\ 70 & 174 & 278 & 382 \\ 110 & 278 & 446 & 614 \\ 150 & 382 & 614 & 846 \end{pmatrix}, \tag{68}$$

obtained multiplying for its own transpose the 3 matrix A_3 whose rows are $(1, 2, 3)$, $(4, 5, 6)$, $(7, 8, 9)$, and the 4×4 matrix A_4 whose rows are $(1, 2, 3, 4)$, $(5, 6, 7, 8)$, $(9, 10, 11, 12)$, $(13, 14, 15, 16)$, so that we had two SPD matrices. They have the extremely high condition numbers, $\kappa(A_3) \approx 1.4167 \times 10^{17}$ and $\kappa(A_4) \approx 2.8759 \times 10^{17}$, and are *not* sdd. We chose the (expected) solution $x := (2, 2, 2)^T$ and $x := (2, 2, 2, 2)^T$, respectively, the initial point $x_0 = 0.5\,x$, and computed $b = A\,x$, in both cases. In Figure 6, top and bottom, the relative errors made by the Jacobi-like and the Gauss–Seidel-like solvers are plotted as functions of the number of iterations, but here we adopted for regularization the more general diagonal matrices $D_\alpha^{(3)} := \mathrm{diag}(68.1, 77.2, 0)$ and $D_\alpha^{(4)} := \mathrm{diag}(300.1, 556.1, 556.1, 300.1)$, respectively. The resulting matrices, $A_3 + D_\alpha^{(3)}$ and $A_4 + D_\alpha^{(4)}$, are sdd. In the second case, for instance, D_α is such that, in each row, the diagonal entry exceeds by just 0.1 the sum of all the other entries.

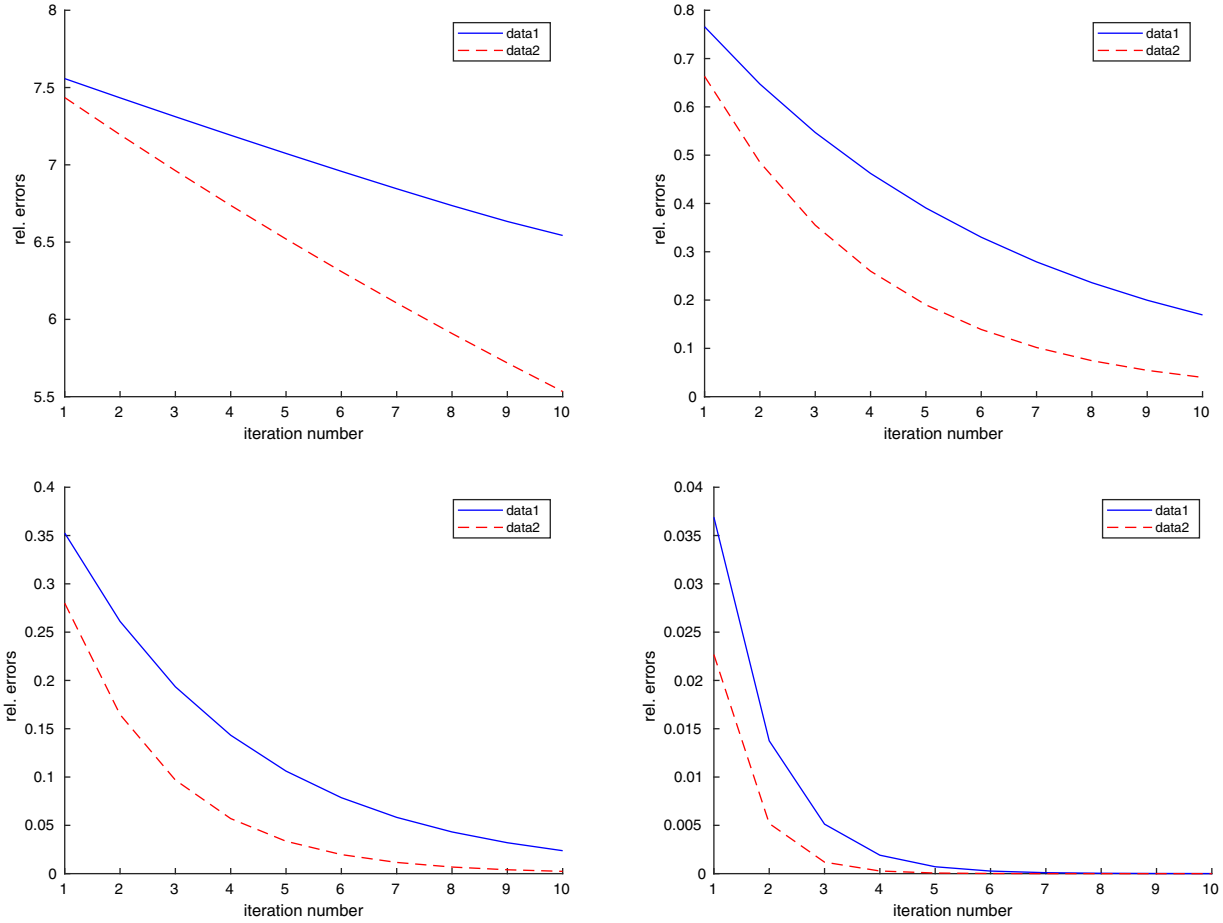

FIGURE 5 Comparing Jacobi-like and Gauss–Seidel-like solvers for the SPD pentadiagonal matrix of dimension $n = 3,000$ of Example 13: from upper left to lower right, for $\alpha = 0$, $\alpha = 1$, $\alpha = 2$, and $\alpha = 10$ (color online)

6 | CONCLUSIONS

In this article, we considered iterative methods to solve possibly ill-conditioned linear systems, based on first shifting conveniently the spectrum of the given matrix' system, A. Replacing A with $A + D_\alpha$, where D_α is a more general diagonal matrix is also considered. We recovered in particular the Riley's method, who in 1955 anticipated the Tikhonov regularization method, and propose a way to accelerate it. A number of examples have been considered, and the corresponding condition numbers for several values of α were shown in some tables. Plots, illustrating the behavior of the errors as functions of the number of iterations, for fixed α, and comparing this behavior using some different iterative solvers have also been made.

Adding a diagonal matrix with entries sufficiently large may make it more strongly sdd or even make it ssd if it was not. It may also be useful to transform a symmetric quasi-definite system,[4] whose matrix has the form

$$\begin{pmatrix} \mathbf{M} & \mathbf{A} \\ \mathbf{A}^T & -\mathbf{N} \end{pmatrix},$$

where M and N are symmetric positive definite matrices (of some interest nowadays, due to a number of applications[24]).

Future directions should include considering adapting other iterative methods, in particular "block methods" (in place of "point methods"), as well as determining whether optimal values for the acceleration parameter in the SOR-like methods exist, and, if so, what they are. Varying the value of this parameter, as done in standard cases, could also be an issue to address. Extensive simulations for challenging large-size problems should be performed, examining the peculiarities of practical implementations.

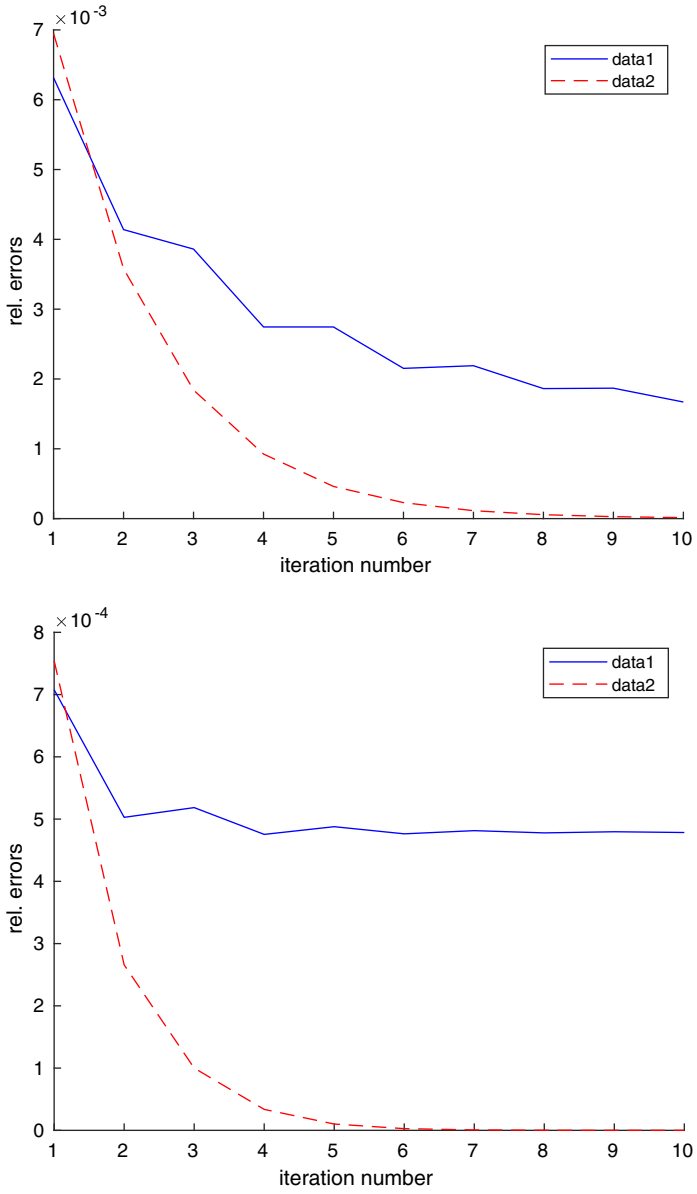

FIGURE 6 Comparing Jacobi-like (data 1) and Gauss–Seidel (data 2) methods, using diagonal matrices with different entries, for two simple ill-conditioned matrices: (top) of dimension $n = 3$, with $D_\alpha = \text{diag}(68.1, 77.2, 0.0)$, and (bottom) of dimension $n = 4$, with $D_\alpha = \text{diag}(300.1, 556.1, 556.1, 300.1)$ (color online)

ACKNOWLEDGEMENTS

The author wishes to thank Gil Strang for his interest and encouragement, and Stefano De Marchi for some useful suggestions. The author indebted with his students Marta Moretti, Lorenzo Cappelli, and Claudio Guidarelli, who participated in part of this work running some examples.

ORCID

Renato Spigler https://orcid.org/0000-0002-4561-4845

REFERENCES

1. Benzi M. Preconditioning techniques for large linear systems: A survey. J Comp Phys. 2002;182:418–477. https://doi.org/10.1006/jcph.2002.7176.
2. Neumaier A. Solving ill-conditioned and singular linear systems: A tutorial on regularization. SIAM Rev. 1998;40:636–666.
3. Benzi M, Golub GH, Liesen J. Numerical solution of saddle point problems. Acta Numer. 2005;14:1–137. https://doi.org/10.1017/S0962492904000212.
4. Benzi M. Iterative solution of symmetric quasi-definite linear systems [book review]. SIAM Rev. 2018;60(3):757–759.
5. Riley JD. Solving systems of linear equations with a positive definite, symmetric, but possibly ill-conditioned matrix. Math Tables Aids Comput. 1955;9:96–101.

6. Berman A, Plemmons RJ. Nonnegative matrices in the mathematical sciences. Philadelphia, PA: SIAM, 1994.
7. Hackbusch W. Iterative solution of large sparse systems of equations. 2nd ed. New York, NY: Springer, 2016.
8. Varga RS. Matrix iterative analysis. Englewood Cliffs, NJ: Prentice-Hall, 1962 xiii+322 pp.
9. Young DM. Iterative methods for solving partial difference Q3 equations of elliptic type [PhD thesis]. Cambridge, MA: Harvard University, 1950.
10. Young DM. Iterative solution of large linear systems. Unabridged republication of the 1971 edition. New York-Mineola, NY and London, UK: Dover, Academic Press, 2003 xxvi+570 pp.
11. Young DM. Historical overview of iterative methods. Comput Phys Comm. 1989;53:1–17.
12. Hadjidimos A. Successive overrelaxation (SOR) and related methods. J Comput Appl Math. 2000;123:177–199.
13. Tikhonov AN. Solution of incorrectly formulated problems and the regularization method. Soviet Math Dokl. 1963;4:1035–1038.
14. Fulton W. Eigenvalues of sums of Hermitian matrices (after A. Klyachko). Séminaire Bourbaki. 1998;1997/98(5):255–269.
15. Wielandt H. On the eigenvalues of $A + B$ and AB. J Res Nat Bur Stand Sect B. 1973;77B:61–63.
16. Zhang F, Zhang Q. Eigenvalue inequalities for matrix product. IEEE Trans Automat Control. 2006;51:1506–1509.
17. Tsitsiklis JN. A comparison of Jacobi and Gauss-Seidel parallel iterations. Appl Math Lett. 1989;2(2):167–170. https://doi.org/10.1016/0893-9659(89)90014-1.
18. Schmitt B. Perturbation bounds for matrix square roots and Pythagorean sums. Lin Algebra Appl. 1992;174:215–227.
19. Todd J. The condition of certain matrices II. Arch Math. 1954;5:249–257.
20. Gautschi W, Inglese G. Lower bounds for the condition number of Vandermonde matrices. Numer Mathematik. 1987;52:241–250.
21. Camargo A. An exponential lower bound for the condition number of real Vandermonde matrices. Appl Numer Math. 2018;128:81–83. https://doi.org/10.1016/j.apnum.2018.01.020.
22. Noschese S, Pasquini L, Reicher L. Tridiagonal Toeplitz matrices: Properties and novel applications. Numer Linear Algebra Appl. 2013;20(2):302–326. https://doi.org/10.1002/nla.1811.
23. Todd J. The condition of certain matrices, III. J Res Nat Bur Stand. 1958;60(1):1–7. Research Paper 2815.
24. Orban D, Arioli M. Iterative solution of symmetric quasi-definite linear systems. Philadelphia, PA: SIAM, 2017.

How to cite this article: Spigler R. On the numerical solution of ill-conditioned linear systems by regularization and iteration. *Numer Linear Algebra Appl*. 2020;e2335. https://doi.org/10.1002/nla.2335