Numer. Math.
DOI 10.1007/s00211-016-0829-7

Numerische Mathematik

Random reordering in SOR-type methods

Peter Oswald[1] · **Weiqi Zhou**[2]

Received: 23 October 2015 / Revised: 30 May 2016

Abstract When iteratively solving linear systems $By = b$ with Hermitian positive semi-definite B, and in particular when solving least-squares problems for $Ax = b$ by reformulating them as $AA^*y = b$, it is often observed that SOR type methods (Gauß-Seidel, Kaczmarz) perform suboptimally for the given equation ordering, and that random reordering improves the situation on average. This paper is an attempt to provide some additional theoretical support for this phenomenon. We show error bounds for two randomized versions, called shuffled and preshuffled SOR, that improve asymptotically upon the best known bounds for SOR with cyclic ordering. Our results are based on studying the behavior of the triangular truncation of Hermitian matrices with respect to their permutations.

Mathematics Subject Classification 65F10 · 15A60

1 Introduction

In this paper, we discuss the influence of the equation ordering in a linear system $By = b$ on deriving upper bounds for the convergence speed of the classical successive over-relaxation (SOR) method. We assume that B is a complex $n \times n$ Hermitian positive semi-definite matrix with positive diagonal part D. If we write $B = L + D + L^*$, where L denotes the strictly lower triangular part of B and $*$ stands for Hermitian conjugation, then one step of the classical SOR iteration reads

✉ Peter Oswald
agp.oswald@gmail.com

Weiqi Zhou
weiqizhou@mathematik.uni-marburg.de

[1] Institute for Numerical Simulation, University of Bonn, Bonn, Germany

[2] FB Mathematics and Informatics, Philipps-University Marburg, Marburg, Germany

Published online: 27 July 2016

$$y^{(k+1)} = y^{(k)} + \omega(D + \omega L)^{-1}(b - By^{(k)}), \qquad k = 0, 1, \ldots. \tag{1}$$

The classical Gauß-Seidel method for solving $By = b$ emerges if one takes $\omega = 1$. If one attempts to solve a general linear system $Ax = b$ in the least-squares sense, then one has the choice to apply the SOR method to either the normal equation $A^*Ax = A^*b$ or to $AA^*y = b$. In the latter case, the algorithm resulting from applying (1) to $B = AA^*$ is equivalent to the Kaczmarz method (here approximations to the solution of $Ax = b$ are recovered by setting $x^{(k)} = A^*y^{(k)}$), see [6].

To make the paper more readable and avoid technical detail, we make two additional assumptions. First, we consider only consistent systems ($b \in \mathrm{Ran}(B)$). This guarantees convergence of (1) for any $0 < \omega < 2$ and any $y^{(0)}$ to a solution of $By = b$, while for inconsistent systems the method diverges (this does not contradict the known convergence of the Kaczmarz method for inconsistent systems $Ax = b$ since the divergence manifests itself only in the $\mathrm{Ker}(B) = \mathrm{Ker}(AA^*)$ component of $y^{(k)}$ which is annihilated when recovering $x^{(k)} = A^*y^{(k)}$). Secondly, we assume that B has unit diagonal ($D = I$) which can always be achieved by transforming to the equivalent rescaled system $D^{-1/2}BD^{-1/2}\tilde{y} = D^{-1/2}b$ (for the Kaczmarz algorithm, one would simply use row normalization in A). Alternatively, the analysis of the SOR method can be carried out with arbitrary $D > 0$, with minor changes in some places, see [11] for some details. With both approaches, D enters the final results via the spectral properties of the transformed B or its norm, respectively. Note that with $D = I$, one step of (1) consists of n consecutive projection steps onto the i-th coordinate direction, $i = 1, 2, \ldots, n$, and the method thus becomes an instance of the alternating direction method (ADM). Unless stated otherwise, these two assumptions are silently assumed from now on.

Since any positive semi-definite B can be factored, in a non-unique way, as

$$B = AA^*,$$

we can always assume that B is produced by some $n \times m$ matrix A with unit norm rows. Denote by $r = \mathrm{rank}(B) \le \min(n, m)$ its rank, the spectral properties of A and B are obviously related: the non-zero eigenvalues $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_r$ of B are given by the squares of the non-zero singular values of A. Thus, if we define the essential condition number $\bar{\kappa}(A)$ of a matrix as the quotient of its largest and smallest non-zero singular values then

$$\bar{\kappa} := \bar{\kappa}(B) = \bar{\kappa}(A)^2 = \frac{\lambda_1}{\lambda_r}.$$

The unit diagonal assumption $D = I$ for B implies $0 < \lambda_r \le 1 \le \lambda_1 \le n$. In the convergence analysis below, we will use the energy semi-norm $|y|_B = \langle By, y\rangle^{1/2} = \|A^*y\|^{1/2}$ associated with B, it is a norm if and only if B is non-singular, i.e., positive definite. Here, $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ denote the usual Euclidian scalar product and norm in $\mathbb{C}^n$, respectively. Later, we will use the notation $\|\cdot\|$ also for matrices (then it stands for their spectral norm) which should be clear from the context and not lead to any confusion.

Condition numbers and other spectral properties often enter the asymptotic error estimates of iterative schemes for solving linear systems, the best known examples are the standard bounds for the Jacobi-Richardson and conjugate gradient methods for systems with positive definite B, see e.g., [6]. For the SOR method, such upper estimates have been established in [10] for non-singular B, and recently improved in [11] within the framework of the Kaczmarz iteration to include the semi-definite case:

Theorem 1 *Let B be a given $n \times n$ Hermitian positive semi-definite matrix with unit diagonal, and assume that $By = b$ is consistent, i.e., possesses at least one solution $\bar{y}$. Then the SOR iteration (1) converges for $0 < \omega < 2$ in the energy semi-norm associated with B according to*

$$|\bar{y} - y^{(k)}|_B^2 \leq (1 - \frac{(2-\omega)\omega\lambda_1}{(1 + \frac{1}{2}\lfloor \log_2(2n) \rfloor \omega\lambda_1)^2 \bar{\kappa}})^k |\bar{y} - y^{(0)}|_B^2, \quad k \geq 1. \tag{2}$$

If B is singular, then for sufficiently small rank r the term $\frac{1}{2}\log_2(2n)$ can be replaced by the smaller term $C_0 \ln r$, where C_0 is an absolute constant.

The proof of (2) rests on rewriting the squared energy semi-norm of Qy, where

$$Q = I - \omega(I + \omega L)^{-1} B$$

is the error iteration matrix associated with (1), as

$$|Qy|_B^2 = |y|_B^2 - \omega(2-\omega)\|(I + \omega L)^{-1} By\|^2 \leq |y|_B^2 - \frac{\omega(2-\omega)\|By\|^2}{\|I + \omega L\|^2}.$$

and using a spectral norm inequality for L from [10],

$$\|L\| \leq \frac{1}{2}\lfloor \log_2(2n) \rfloor \|B\|, \tag{3}$$

to estimate the term $\|I + \omega L\| \leq 1 + \omega\|L\|$. For singular B with small rank $r < n$, the estimate (3) has been improved in [11] to

$$\|L\| \leq C_0 \ln r \|B\|, \quad r \geq 2, \tag{4}$$

where C_0 is a fixed positive constant. It is well known that the estimate (3) is sharp in its logarithmic dependency on n, more precisely

$$b_n := \sup_{B \neq 0} \frac{\|L\|}{\|B\|} \asymp \frac{1}{\pi} \ln n, \quad n \to \infty,$$

where the supremum is taken with respect to all $n \times n$ matrices B, and $\asymp$ stands for asymptotic equality (see [3,9] for sharp estimates and examples). Similar lower estimates hold also for Hermitian positive semi-definite matrices B with unit diagonal $D = I$, and examples exist that show the necessity of the logarithmic terms in (2), see [10].

For non-singular B, i.e., when $|\cdot|_B$ becomes a norm and the system has full rank $r = n$, the outlined idea of proof for Theorem 1 has been carried out in detail in [10]. The changes for singular B are minimal, the proof of (4) for this case can be found in [11, Theorem 4], see also the proof of Part (b) of Theorem 4 in Sect. 3.

The crucial inequalities (3) and (4), and consequently the error bounds in Theorem 1, suffer from one serious drawback: they are invariant under simultaneously reordering rows and columns in $B = AA^*$ resp. reordering the rows in A. Indeed, $B_\sigma = P_\sigma B P_\sigma^*$ has the same spectrum and spectral norm as B for any permutation σ of the index set $\{1, \ldots, n\}$ (P_σ denotes the associated $n \times n$ row permutation matrix), while the spectral properties of the lower triangular part L_σ of B_σ depend on σ. As a matter of fact, in practice it is often observed (for example, see [5,14,16]) that reordering improves the convergence behavior of SOR methods as well as other, more general, alternating directions, subspace correction, and projection onto convex sets (POCS) methods. The interest in explaining this observation theoretically has been further stimulated by convergence results for a randomized Kaczmarz iteration in [13]. In the language of SOR for solving a consistent system $By = b$ with $D = I$, instead of performing the n consecutive projection steps on the ith coordinate that compose the SOR iteration step (1) in the fixed order $i = 1, 2, \ldots, n$, the method in [13] performs the projection steps on coordinate directions by randomly selecting i uniformly and independently from $\{1, \ldots, n\}$ in each single step. For a fair comparison with the original SOR iteration (1), and the randomized SOR methods discussed below, it is appropriate to combine n single projection steps on randomly and independently chosen coordinate directions into one iteration step. The iterates $y^{(k)}$ of this method which we call for short *single step randomized SOR iteration* are now random variables. Under the same assumptions as in Theorem 1, the following estimate for the expectation of the squared energy semi-norm error can be deduced from [13]:

$$\mathbb{E}(|y^{(k)} - y^*|_B^2) \le \left(1 - \frac{(2-\omega)\omega\lambda_1}{n\bar{\kappa}}\right)^{kn} |y^{(0)} - y^*|_B^2, \quad k \ge 1. \tag{5}$$

The two upper estimates (2) and (5) are obtained by different techniques, and although a rough comparison of the upper bounds suggests that the single step randomized SOR beats the original SOR, in practice this is generally not true, and depends on the given system and the ordering of the equations in it.

In this paper, we consider two different randomization strategies for SOR closer to the original method. In the first, given the k-th iterate $y^{(k)}$, we choose (independently and randomly) a permutation σ of $\{1, \ldots, n\}$, and do one full SOR iteration step (1) with $By = b$ and $y^{(k)}$ replaced by $B_\sigma y_\sigma = b_\sigma$ and $y_\sigma^{(k)} = P_\sigma y^{(k)}$, where $y_\sigma = P_\sigma y$, $b_\sigma = P_\sigma b$. Then the original order is restored by setting $y^{(k+1)} = P_\sigma^* y_\sigma^{(k+1)}$. This approach which we call for short *shuffled SOR iteration* is equivalent to a random ordering without repetition in each sweep of n steps of the single step randomized SOR iteration. In practice, random ordering without repetition is considered superior to random ordering with repetition although theoretical proof for this observation is yet missing, see the conjectures in [4,12]. In [15], where the counterpart of the shuffled SOR iteration for coordinate descent methods in convex optimization appears as algorithm EPOCHS, similar statements can be found.

It is also tempting to investigate the effect of an one-time reordering, followed by the application of the SOR iteration in the classical, cyclic fashion (we call this *preshuffled SOR iteration*). In other words, the preshuffled SOR iteration coincides with a shuffled SOR iteration if we reuse the randomly generated σ from the iteration step at $k = 0$ for all further iteration steps at $k > 0$. Observe that in terms of the Kaczmarz iteration these two schemes merely correspond to shuffling the rows in the row-normalized matrix A, i.e., $Ax = b$ is replaced by $P_\sigma Ax = P_\sigma b$. The numerical experiments presented in [11] suggest that shuffled and preshuffled SOR iterations often perform in expectation equally good, and better than the single step randomized iteration.

The present paper is an attempt to gain some insight into what can be expected from these randomization strategies. Speaking in mathematical terms, if

$$Q_\sigma = (I + \omega L_\sigma)^{-1}((1-\omega)I - \omega L_\sigma^*),$$

denotes the error iteration matrix of the SOR method applied to $B_\sigma y_\sigma = b_\sigma$, then we aim at investigating the quantity

$$\mathbb{E}[|Qy|_B^2] := \frac{1}{n!}\sum_\sigma |Q_\sigma y_\sigma|_{B_\sigma}^2, \quad |y|_B = 1, \tag{6}$$

to obtain upper bounds for the expected square energy semi-norm error in the shuffled SOR iteration.

As was outlined above, obtaining estimates for the norm behavior of Q_σ, and of relevant averages such as (6), must be closely related to studying the behavior of L_σ which will be at the heart of our considerations in Sect. 2. In particular, we apply a corollary of the recently proved paving conjecture to show that for any positive semi-definite B with $D = I$ there is a permutation σ (depending on B) with the property

$$\|L_\sigma\| \le C_1 \|B\|, \tag{7}$$

where C_1 is an absolute constant. We further establish that

$$\|\mathbb{E}[LL^*]\| < \|B\|^2, \quad \mathbb{E}[LL^*] := \frac{1}{n!}\sum_\sigma P_\sigma^* L_\sigma L_\sigma^* P_\sigma, \tag{8}$$

which will lead to bounds for (6).

In Sect. 3, we apply the results of Sect. 2 to establish two new error decay bounds for the above mentioned shuffled SOR iterations. First of all, we show that the quantity in (6) satisfies

$$\mathbb{E}(|Qy|_B^2) \le (1 - \frac{(2-\omega)\omega\lambda_1}{(1+\omega\lambda_1)^2\bar{\kappa}(B)})|y|_B^2,$$

which implies a bound for the expected square energy semi-norm error decay of the shuffled SOR iteration that compares favorably with the bounds in Theorem 1, as the

logarithmic dependence on n and r is removed. Next, we prove using (7) that there exists a σ such that the preshuffled SOR iteration can achieve the same effect, i.e., replacing the $\frac{1}{2}\lfloor \log_2(2n)\rfloor$ resp. $C_0 \ln r$ factor by the constant C_1 from (7). Although asymptotic in nature, and in case of the preshuffled SOR iteration due to the currently available estimates for C_1 not yet practical, the bounds established in Theorem 4 should be viewed as theoretical support for the numerically observed convergence behavior of shuffled and preshuffled SOR iterations.

2 Triangular truncation and reordering

If not stated otherwise, in this section $B = L + D + L^*$ belongs to $\mathscr{H}_n$, the set of all $n \times n$ Hermitian matrices, with no assumptions on positive semi-definiteness or normalization of its diagonal elements (i.e., not assuming $D = I$). The notation of Sect. 1 is reused for this slightly more general situation.

Theorem 2 *If $B \in \mathscr{H}_n$ then the average operator $\mathbb{E}[LL^*]$ defined in (8) satisfies*

$$\|\mathbb{E}[LL^*]\| \le 4\|B\|^2.$$

Moreoever, if $D = I$ and B is positive semi-definite, then (8) holds.

Proof For given $B \in \mathscr{H}_n$, set $H = L + L^*$. Since $\|D\| \le \|B\|$, we have

$$\|H\| = \|B - D\| \le \|B\| + \|D\| \le 2\|B\|, \tag{9}$$

while for positive semi-definite B

$$\|H\| = \|B - I\| \le \max(\|B\| - 1, 1) \le \|B\|. \tag{10}$$

Thus, establishing (8) with B replaced by H is enough.

Straightforward computation shows that

$$(P_\sigma^* L_\sigma L_\sigma^* P_\sigma)_{st} = \sum_{k=1}^{\min(s,t)-1} H_{s\sigma(k)} H_{\sigma(k)t}, \quad s,t = 1,\dots,n.$$

By counting the number of permutations for which $\sigma(k) = l$ for some $k = 1,\dots,\min(s,t)-1$ we get

$$\frac{1}{n!}\sum_\sigma (P_\sigma^* L_\sigma L_\sigma^* P_\sigma)_{st} = \frac{(n-1)!}{n!}(\min(s,t)-1)\sum_{l=1}^n H_{sl}H_{lt} = \frac{\min(s,t)-1}{n}(H^2)_{st}.$$

Hence

$$\frac{1}{n!}\sum_\sigma (P_\sigma^* L_\sigma L_\sigma^* P_\sigma) = \frac{1}{n} K \circ H^2,$$

where $\circ$ denotes Hadamard multiplication, and

$$K = \begin{pmatrix} 0\,0\,0\,\ldots & 0 & 0 \\ 0\,1\,1\,\ldots & 1 & 1 \\ 0\,1\,2\,\ldots & 2 & 2 \\ \vdots\;\vdots\;\vdots\;\ddots & \vdots & \vdots \\ 0\,1\,2\,\ldots\,n-2 & n-2 & n-2 \\ 0\,1\,2\,\ldots\,n-2 & n-2 & n-1 \end{pmatrix}. \tag{11}$$

In other words, the above Hadamard product can be written as the linear combination of $n-1$ diagonally projected submatrices of H^2, each of norm $\leq \|H^2\|$. This gives

$$\|\frac{1}{n!}\sum_{\sigma} P_\sigma^* L_\sigma L_\sigma^* P_\sigma\| \leq \frac{n-1}{n}\|H^2\| < \|H\|^2,$$

which completes the proof. □

The following result was suggested to the first author by B. Kashin (Steklov Institute, Moscow), and is included here with his permission.

Theorem 3 *There is an absolute constant C_2 such that for any $B \in \mathscr{H}_n$ there exists a permutation σ for which*

$$\|L_\sigma\| \leq C_2\|B\|. \tag{12}$$

Moreover, if $D = I$ and B is positive semi-definite then (7) holds with an absolute constant $C_1 \leq C_2$.

Proof Weaker versions of (12), where the spectral norm $\|L\| = \|L\|_{\ell_2^n \to \ell_2^n}$ is replaced by $\|L\|_{\ell_2^n \to \ell_q^n}$ with $1 \leq q < 2$, have been proved in [7] and [1].

For the proof of (12) we explore the following particular result on matrix paving, which for a long time was known as Anderson's Paving Conjecture for Hermitian matrices with small diagonal. This conjecture is equivalent to the Kadison-Singer Problem, a positive solution of which was recently given in [8]. We formulate it for $B \in \mathscr{H}_n$ with zero diagonal, and refer to the recent expository paper [2] for details.

Theorem (Anderson's Paving Conjecture) *For any $0 < \epsilon < 1$, there is an integer $\gamma(\epsilon) \geq 2$ such that for any $n \in \mathbb{N}$ and any $B \in \mathscr{H}_n$ with zero diagonal, there exists a partition*

$$w_1 \cup w_2 \ldots \cup w_\gamma = \{1, 2, \ldots, n\}, \qquad w_i \cap w_j = \emptyset, \quad i \neq j,$$

into $\gamma \leq \gamma(\epsilon)$ non-empty index subsets such that

$$\|B_{w_k}\| \leq \epsilon\|B\|, \quad k = 1, \ldots, \gamma.$$

Here B_{w_k} is the $|w_k| \times |w_k|$ submatrix corresponding to the index set $w_k \times w_k$.

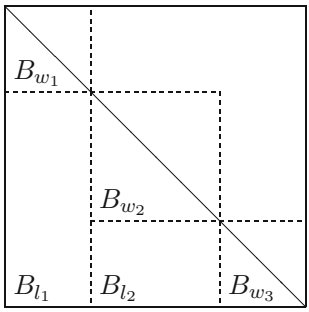

Fig. 1 Block structure of B_σ

Returning to the proof of Theorem 3, by (9) it is enough to consider matrices $B \in \mathscr{H}_n$ with zero diagonal. For given $0 < \epsilon < 1$ we proceed by induction in n to establish (12) with a constant $C_\epsilon := (\gamma(\epsilon) - 1)/(1 - \epsilon)$, where $\gamma(\epsilon)$ is defined in the above paving theorem. To find an estimate for the best constant C_2 in (12), we then optimize with respect to ϵ resp. γ. Any σ will do for $n = 2$ since $C_\epsilon > 1$ and in this case $\|L\| = \|B\|$. Suppose the statement holds for all matrix dimensions less than n. For $B \in \mathscr{H}_n$ with zero diagonal, consider the partition $w_1, w_2, \ldots, w_\gamma$ in the above paving theorem, and denote by σ_0 the permutation that makes B_{σ_0} contain the submatrices B_{w_k} as consecutive diagonal blocks, as depicted in Fig. 1 for $\gamma = 3$. Let B_{l_k} be the rectangular submatrices below B_{w_k} in this B_{σ_0}, $k = 1, \ldots, \gamma - 1$.

For each $k = 1, \ldots, \gamma$ we have $|w_k| < n$, and by the induction assumption there exist permutations σ_k such that

$$\|(L_{w_k})_{\sigma_k}\| \le C_\epsilon \|B_{w_k}\|, \quad k = 1, \ldots, \gamma,$$

where $(L_{w_k})_{\sigma_k}$ is the strictly lower triangular part of $(B_{w_k})_{\sigma_k}$.

By superposing the permutations σ_k within each block with σ_0, we get the desired σ: In each diagonal block of B_σ we have now $(B_{w_k})_{\sigma_k}$ instead of B_{w_k}, and the rectangular submatrices B'_{l_k} below the diagonal blocks are row and column permuted copies of the previous B_{l_k}.

We split L_σ into the sum of a block-diagonal matrix L_1 containing all $(L_{w_k})_{\sigma_k}$, and another lower triangular matrix L_2 containing all rectangular submatrices B'_{l_k}. Since

$$\|L_1\| \le \max_{k=1,\ldots,\gamma} \|(L_{w_k})_{\sigma_k}\| \le C_\epsilon \max_{k=1,\ldots,\gamma} \|B_{w_k}\| \le C_\epsilon \epsilon \|B\|,$$

and

$$\|L_2\| \le \sum_{k=1}^{\gamma-1} \|B'_{l_k}\| \le (\gamma - 1)\|B\| \le C_\epsilon (1 - \epsilon)\|B\|$$

(note that each B'_{l_k} is a row and column permuted version of a rectangular submatrix of the original B, thus $\|B'_{l_k}\| \le \|B\|$). Therefore,

$$\|L_\sigma\| \le \|L_1\| + \|L_2\| \le C_\epsilon \|B\|, \tag{13}$$

which concludes the induction step.

To find numerical estimates for C_2, we need bounds for $\gamma(\epsilon)$. The bounds given in [2, Sect. 4]) are very rough, therefore we rely on Corollary 26 from Tao's blog on the Kadison-Singer problem accessible at https://terrytao.wordpress.com/2013/11/04/ which implies the following: For given $\gamma \ge 2$, there exists a partition into γ^2 index subsets such that the statement of Theorem 2 holds with $\epsilon = \epsilon(\gamma) = 2/\gamma + 2\sqrt{2/\gamma}$. For $\gamma \ge 12$, one has $\epsilon < 1$, and we conclude that

$$C_2 \le 2 \inf_{0<\epsilon<1} C(\epsilon) \le 2 \inf_{\gamma \ge 12} C(\epsilon(\gamma)) = 2 \min_{\gamma \ge 12} \frac{\gamma^2 - 1}{1 - 2/\gamma - 2\sqrt{2/\gamma}} = 2907,$$

with the minimum achieved for $\gamma = 18$. The factor 2 comes from taking into account (9). This bound is overly pessimistic (note that results closer to the known lower bound $\gamma(\epsilon) \ge 1/\epsilon^2$ would result in much smaller values of C_2).

It is therefore worth looking for improvements if B is positive semi-definite and has unit diagonal $D = I$. Then $B - I$ is a Hermitian matrix with zero diagonal and spectrum in $[-1, \|B\| - 1]$ satisfying (10), and Corollary 25 of Tao's blog yields, for any $\gamma \ge 2$, the existence of a partition into γ index subsets such that in Theorem 2 we can take $\epsilon = \epsilon'(\gamma) = 1/\gamma + 2/\sqrt{\gamma}$. Repeating the above proof steps for this case, we see that

$$C_1 \le \min_{\gamma \ge 6} \frac{\gamma - 1}{1 - 1/\gamma - 2/\sqrt{\gamma}} \le 32.42$$

(here $\epsilon'(\gamma) < 1$ for $\gamma \ge 6$, and the minimum is achieved for $\gamma = 12$). □

It remains an open question if an inequality similar to (12) also holds for the average of the norms $\|L_\sigma\|$, namely if

$$\mathbb{E}[\|L\|] := \frac{1}{n!} \sum_\sigma \|L_\sigma\| \le C_n \|B\|, \qquad B \in \mathscr{H}_n, \tag{14}$$

holds for some (bounded or slowly increasing) sequence of positive constants $C_n = o(\ln(n))$ (for a related result, see [1, Theorem 8.4]). A proof of (14) would imply improved asymptotic estimates for the expected convergence rate of the preshuffled SOR iteration, and not only for the best possible convergence rate, as established in Part (b) of Theorem 4 below.

3 Application to SOR iterations

In this section we show *a priori* convergence estimates for the shuffled and preshuffled SOR iterations that improve upon the one for the standard SOR iteration (1) stated in Theorem 1, at least asymptotically. These estimates are formulated in terms of the energy semi-norm associated with B, and are equivalent to estimates in the usual

Euclidian norm for the Kaczmarz iteration applied to a consistent linear system $Ax = b$, where $B = AA^*$. The result is summarized in the following

Theorem 4 *Let $By = b$ be a consistent linear system with positive semi-definite $B = L + I + L^* \in \mathscr{H}_n$, and denote by $\bar{y}$ an arbitrary solution of it. Fix any $\omega \in (0, 2)$.*

a) The expected squared energy semi-norm error of the shuffled SOR iteration converges exponentially with the bound

$$\mathbb{E}(|\bar{y} - y^{(k)}|_B^2) \le \left(1 - \frac{\omega(2-\omega)\lambda_1}{(1+\omega\lambda_1)^2\bar{\kappa}}\right)^k |\bar{y} - y^{(0)}|_B^2, \quad k \ge 1,$$

for any $\omega \in (0, 2)$.

b) There exists some ordering σ such that the classical SOR iteration on the system $B_\sigma y_\sigma = b_\sigma$ converges for any $\omega \in (0, 2)$ with square energy semi-norm error decay

$$|\bar{y} - y^{(k)}|_B^2 \le \left(1 - \frac{\omega(2-\omega)\lambda_1}{(1+C_1\omega\lambda_1)^2\bar{\kappa}}\right)^k |\bar{y} - y^{(0)}|_B^2, \quad k \ge 1,$$

where the constant C_1 satisfies (7).

Proof We start with b). Take the σ for which $\|L_\sigma\| \le C_1\|B\|$ according to (7). To simplify notation, let us drop the subscript σ so that now $B_\sigma = B = L + I + L^*$, $b_\sigma = b$, $P_\sigma = I$, and $\|L\| \le C_1\|B\|$. Recall the notation $Q = I - \omega(I + \omega L)^{-1}B$ for the error iteration matrix, and check that $\mathbb{C}^n = U \oplus V$, where $U = \mathrm{Ker}(B)$ and $V = (I + \omega L)^{-1}\mathrm{Ran}(B)$ are Q-invariant subspaces (obviously, Q is the identity when restricted to U). Write the SOR iterates as $y^{(k)} = u^{(k)} + v^{(k)}$, $u^{(k)} \in U$, $v^{(k)} \in V$. Since $y^{(k+1)} = Qy^{(k)} + \omega(I + \omega L)^{-1}b$, by induction it follows that

$$u^{(k)} = u^{(0)}, \quad v^{(k)} = Q^k v^{(0)} + \omega(I + Q + \cdots + Q^{k-1})(I + \omega L)^{-1}b, \quad k \ge 1. \tag{15}$$

Now, any solution $\bar{y}$ of $By = b$ can be written as $\bar{y} = u + \bar{v}$, where $u \in U$ is arbitrary, and $\bar{v} \in V$ is unique. Because

$$|\bar{y} - y^{(k)}|_B^2 = \langle B(\bar{y} - y^{(k)}), \bar{y} - y^{(k)}\rangle = \langle B(\bar{v} - v^{(k)}), \bar{v} - v^{(k)}\rangle = |\bar{v} - v^{(k)}|_B^2,$$

and $\bar{v} - v^{(k+1)} = \bar{v} - Qv^{(k)} - \omega(I + \omega L)^{-1}B\bar{v} = Q(\bar{v} - v^{(k)})$, all we need is an estimate of the form

$$|Qv|_B^2 \le \rho|v|_B^2, \quad v \in V.$$

By substituting $\omega B = (I + \omega L) + (I + \omega L^*) - (2 - \omega)I$ below, we get

$$\begin{aligned} |Qv|_B^2 &= \langle Bv, v\rangle - \omega\langle B((I + \omega L)^{-1} + (I + \omega L^*)^{-1})Bv, v\rangle \\ &\quad + \omega\langle B(I + \omega L^*)^{-1}(\omega B)(I + \omega L)^{-1}Bv, v\rangle \\ &= \langle Bv, v\rangle - \omega(2 - \omega)\|(I + \omega L)^{-1}Bv\|^2. \end{aligned}$$

Using (7), the last term can be bounded from below as

$$\|(I+\omega L)^{-1}Bv\|^2 \geq \frac{\|Bv\|^2}{\|I+\omega L\|^2} \geq \frac{\lambda_r |v|_B^2}{(1+\omega C_1\|B\|)^2} = \frac{\lambda_1 |v|_B^2}{(1+\omega C_1\lambda_1)^2\bar{\kappa}}.$$

Therefore, we obtain

$$\rho = 1 - \frac{\omega(2-\omega)\lambda_1}{(1+\omega C_1\lambda_1)^2\bar{\kappa}},$$

which gives the bound stated in Part b). Moreover, since $\|v\|$ and $|v|_B$ are equivalent norms on V, we see that $v^{(k)} \to \bar{v}$. According to (15)

$$y^{(k)} \to u^{(0)} + \bar{v},$$

so the SOR iteration converges in the usual sense as well, with the $U = \mathrm{Ker}(B)$ component in the limit depending on the starting vector $y^{(0)}$ if B is singular. Returning to the original formulation as preshuffled SOR iteration, the $\mathrm{Ker}(B)$ component of the limit would also depend on σ.

The result of Part a) requires a similar, yet slightly more subtle analysis. Recall that in each step of the shuffled iteration, given the current iterate $y^{(k)}$, we choose a permutation σ at random, apply the SOR step (with matrix $B_\sigma = P_\sigma B P_\sigma^*$ and right-hand side $b_\sigma = P_\sigma b$) to $P_\sigma y^{(k)}$, and return afterwards to the original ordering by multiplying with P_σ^*. In other words, the iteration step is now

$$\begin{aligned} y^{(k+1)} &= P_\sigma^*[(I-\omega(I+\omega L_\sigma)^{-1}B_\sigma)P_\sigma y^{(k)} + \omega(I+\omega L_\sigma)^{-1}b_\sigma] \\ &= \underbrace{(I-\omega P_\sigma^*(I+\omega L_\sigma)^{-1}P_\sigma B)}_{=Q_\sigma} y^{(k)} + \omega P_\sigma^*(I+\omega L_\sigma)^{-1}P_\sigma b. \end{aligned}$$

Thus, as before

$$|e_\sigma^{(k+1)}|_B^2 = |Q_\sigma(\bar{y}-y^{(k)})|_B^2 = |e^{(k)}|_B^2 - \omega(2-\omega)\|(I+\omega L_\sigma)^{-1}P_\sigma B(e^{(k)})\|^2,$$

where for short we have set $e_\sigma^{(k+1)} := \bar{y} - y^{(k+1)}$ and $e^{(k)} := \bar{y} - y^{(k)}$ (indicating that $y^{(k+1)}$ depends on σ, while $y^{(k)}$ is considered fixed at the moment). The expected square semi-norm error after $k+1$ iterations (conditioned on the error $e^{(k)}$) is thus

$$\frac{1}{n!}\sum_\sigma |e_\sigma^{(k+1)}|_B^2 = |e^{(k)}|_B^2 - \frac{\omega(2-\omega)}{n!}\sum_\sigma \|(I+\omega L_\sigma)^{-1}P_\sigma B e^{(k)}\|^2. \tag{16}$$

We give a lower estimate for the last term in (16) with $Be^{(k)}$ temporarily replaced by any unit vector z. Since for positive definite $S \in \mathscr{H}_n$ we have

$$\langle Sy, y\rangle\langle S^{-1}y, y\rangle \geq 1, \quad \|y\| = 1,$$

(indeed, $1 = \|y\|^4 = \langle S^{1/2}y, S^{-1/2}y\rangle^2 \le \|S^{1/2}y\|^2\|S^{-1/2}y\|^2$), applying this inequality with $S = (I + \omega L_\sigma)(I + \omega L_\sigma^*)$ and $y = P_\sigma z$, we get

$$\|(I + \omega L_\sigma)^{-1} P_\sigma z\|^{-2} \le \|(I + \omega L_\sigma^*) P_\sigma z\|^2 = \langle \|z\|^2 + \omega(Hz, z) + \omega^2 (\frac{1}{n!} \sum_\sigma P_\sigma^* L_\sigma L_\sigma^* P_\sigma) z, z\rangle,$$

where as before $H = B - I = L + L^*$. Thus, by the arithmetic-harmonic-mean inequality,

$$\begin{aligned}
&\frac{1}{n!} \sum_\sigma \|(I + \omega L_\sigma)^{-1} P_\sigma z\|^2 \\
&\ge n! \left(\sum_\sigma \|(I + \omega L_\sigma)^{-1} P_\sigma z\|^{-2} \right)^{-1} \\
&\ge n! \left(\sum_\sigma \|(I + \omega L_\sigma^*) P_\sigma z\|^2 \right)^{-1} \\
&= \left((\|z\|^2 + \omega(Hz, z) + \omega^2 (\frac{1}{n!} \sum_\sigma P_\sigma^* L_\sigma L_\sigma^* P_\sigma z, z) \right)^{-1}.
\end{aligned}$$

By Theorem 2, the sum in the last expression can be estimated by

$$\|z\|^2 + \omega(Hz, z) + \omega^2 (\frac{1}{n!} \sum_\sigma P_\sigma^* L_\sigma L_\sigma^* P_\sigma z, z) \le (1 + w\|H\|)^2 \le (1 + \omega\lambda_1)^2.$$

This gives the needed auxiliary result

$$\frac{1}{n!} \sum_\sigma \|(I + \omega L_\sigma)^{-1} P_\sigma z\|^2 \ge (1 + \omega\lambda_1)^{-2}, \quad \|z\| = 1.$$

Going back to the notation of (16), we therefore have

$$\frac{1}{n!} \sum_\sigma |e_\sigma^{(k+1)}|_B^2 = |e^{(k)}|_B^2 - \frac{\omega(2-\omega)}{(1+\omega\lambda_1)^2} \|Be^{(k)}\|^2 \le \left(1 - \frac{\omega(2-\omega)\lambda_r}{(1+\omega\lambda_1)^2}\right) |e^{(k)}|_B^2, \tag{17}$$

which implies the desired estimate for the expected square energy semi-norm error after one iteration step, conditioned on the previous iterate. Since the random choice of σ is considered independent from iteration step to iteration step, we can take the expectation of $|e^{(k)}|_B^2$ in (17) and arrive at the statement of Part a). Finally, we note that for singular B, the result of Part a) only implies that the unique solution component in $\mathrm{Ran}(B)$ is recovered at an exponential rate from the iterates (in expectation). □

We conclude with a few further comments on the estimates for shuffled SOR iterations obtained in Theorem 4. First of all, they are worst-case upper bounds for the

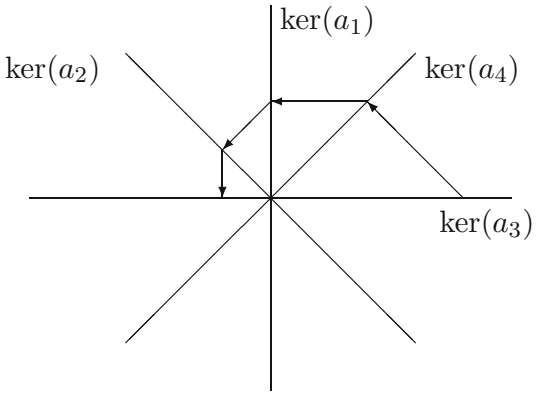

Fig. 2 Hyperplanes for ADM example with $m = 4$

class of all consistent systems $By = b$ with Hermitian positive semi-definite matrix B and normalization condition $D = I$. As such, they improve upon the worst-case upper bounds for fixed cyclic ordering from Theorem 1, at least in the asymptotic regime $n \to \infty$. The current estimate $C_1 \leq 32.42$ entering the bound for the preshuffled SOR iteration is certainly too pessimistic compared to our numerical experience reported in [11], it is due to our reliance on Theorem 2 for which currently only suboptimal quantitative versions, i.e., crude estimates for $\gamma(\epsilon)$, are available. Finding better estimates for $\gamma(\epsilon)$ and the constant C_1 in (7), or replacing the use of simple norm estimates for L by more subtle techniques, would be desirable. We leave this for future work.

Another issue is the formal superiority of the bound (5) for the single-step randomized SOR iteration compared to our results which is not reflected in the actual performance of the methods in many tests, where shuffled and preshuffled SOR iterations compete well. The appearance of an additional factor λ_1 in the denominator of our convergence rate estimates in Theorems 1 and 4 compared to (5) is inherent to our approach of analyzing the error reduction per sweep rather than estimating the single-step error reduction. Due to the assumed normalization $D = I$, we have $1 \leq \lambda_1 \leq n$, however in many practical cases (and for typical ensembles of random matrices) the actual value of λ_1 remains close to 1 which partly mitigates the issue. We conclude with an academic example showing that the extra λ_1 factor in the denominator of the bound in Theorem 1 is necessary (whether this is also true for the bounds in Theorem 4 remains open).

For each $m \in \mathbb{N}$, consider the homogenous linear system $By = 0$, where $B = AA^*$ is induced by the $2m \times 2$ matrix A with unit norm row vectors a_j given by

$$a_j = \big(\cos((j-1)\theta_m),\, \sin((j-1)\theta_m)\big)\,, \quad j = 1, \ldots, 2m, \quad \theta_m := \frac{\pi}{2m}. \tag{18}$$

It is easy to check that $A^*A = mI$. Thus, B has rank $r = 2$, essential condition number 1, and spectral norm $\|B\| = \lambda_1 = m$. As mentioned in the introduction, applying the Gauß-Seidel method ($\omega = 1$) to $By = 0$ is the same as applying the Kaczmarz aka ADM method to $Ax = 0$. From a geometric point of view (see Fig. 2), since the $2m$ hyperplanes the ADM method for $Ax = 0$ projects on split the plane with equal

angles θ_m, the error reduction rate per single step of the ADM iteration with cyclic ordering is simply $\cos\theta_m$, and becomes increasingly slow as m grows (see Fig. 2). The convergence rate of the squared error per sweep is thus

$$(\cos\theta_m)^{2m} \approx (1 - \frac{\pi^2}{8m^2})^{2m} \approx 1 - \frac{\pi^2}{4m}, \quad m \to \infty.$$

This shows that without the $\lambda_1 = m$ factor in the denominator of the bound (2) from Theorem 1 we would arrive at a contradiction.

References

1. Bourgain, J., Tzafriri, L.: Invertibility of large submatrices with applications to the geometry of Banach spaces and harmonic analysis. Israel J. Math. **57**(2), 137–224 (1987)
2. Casazza, P.G., Tremain, J.C.: Consequences of the marcus/spielman/srivastava solution of the Kadison-Singer problem. In: Aldroubi, A., Cabrelli, C., Jaffard, S., Molter, U. (eds.) Trends in applied harmonic analysis, pp. 191–214. Springer, Cham (2016)
3. Davidson K.R.: Nest algebras. Pitman Research Notes in Mathematical Sciences. Longman Sci. Tech. Pub. vol. 191, Harlow (1988)
4. Duchi, J.C.: Commentary on "Towards a noncommutative arithmetic-geometric mean inequality" by B. Recht and C. Ré. J. Mach. Learn. Res. **23**, 11.25–11.27 (2012). (W&CP COLT 2012)
5. Feichtinger H.G., Gröchenig K.: Theory and practice of irregular sampling. In: Benedetto, J.J., Frazier, M.W. (eds.) Wavelets: mathematics and applications, Stud. Adv. Math., pp. 305–363. CRC, Boca Raton (1993)
6. Hackbusch, W.: Iterative solution of large sparse systems of equations. Springer, New York (1994)
7. Kashin, B.S.: On some properties of matrices of bounded operators from the space l_2^n into l_2^m (Russian). Izv. Akad. Nauk Arm SSR Mat. **15**, 379–394 (1980)
8. Marcus, A., Spielman, D.A., Srivastava, N.: Interlacing families II: mixed characteristic polynomials and the Kadison-Singer problem. Ann. Math. **182**(1), 327–350 (2015)
9. Mathias, R.: The Hadamard operator norm of a circulant and applications. SIAM J. Matrix Anal. Appl. **14**(4), 1152–1167 (1993)
10. Oswald, P.: On the convergence rate of SOR: a worst case estimate. Computing **52**(3), 245–255 (1994)
11. Oswald, P., Zhou, W.: Convergence analysis for Kaczmarz-type methods in a Hilbert space framework. Linear Algebra Appl. **478**, 131–161 (2015)
12. Recht, B., Ré, C.: Toward a noncommutative arithmetic-geometric mean inequality: conjectures, case-studies, and consequences. J. Mach. Learn. Res. **23**, 11.1–11.24 (2012). (W&CP COLT 2012)
13. Strohmer, T., Vershynin, R.: A randomized Kaczmarz algorithm with exponential convergence. J. Fourier Anal. Appl. **15**(2), 262–278 (2009)
14. Varga, R.: Orderings of the successive overrelaxation scheme. Pac. J. Math. **9**(3), 925–939 (1959)
15. Wright, S.J.: Coordinate descent algorithms. Math. Program. **151**(1), 3–34 (2015)
16. Young, D.: Iterative methods for solving partial difference equations of elliptic type. Trans. Am. Math. Soc. **76**(1), 92–111 (1954)