THE HERMITIAN SPLITTING METHOD FOR POSITIVE DEFINITE SYSTEMS
A REVIEW AND SUMMARY

FARID KAVEH

1 HSS and IHSS Iteration: Key Results

In a 2003 paper [1], Bai and Golub introduced a new iterative method for solving large sparse systems of the form

$$Ax = b, \;\; A \in \mathbb{C}^{n\times n}, \; x, b \in \mathbb{C}^n, \tag{1}$$

where A is non-singular, in general non-Hermitian, and positive definite. They called this method the Hermitian/skew-Hermitian splitting iteration (HSS), and provided also an inexact Hermitian/skew-Hermitian splitting iteration (IHSS) method which is more efficient in solving large systems. Note that a non-Hermitian matrix A is said to be positive definite if its Hermitian part $H = \frac{1}{2}(A + A^*)$ is positive definite.

Here we shall analyse the ideas behind the HSS and IHSS techniques, as well as provide some numerical results that demonstrate the theory. We will discuss some methods for optimal parameter estimation and selection that have thus far been presented in the literature, and we will also introduce a new method of optimal parameter estimation. We will state some key results without reproducing the proofs. The exact solution of Equation 1 shall be denoted by $x^* \in \mathbb{C}^n$ throughout. We will need to consider *splittings* of A. These are decompositions of the form $A = M - N$ where M is a non-singular matrix.

1.1 Statements and Explanations

Lemma 1.1 is the main lemma of [1].

Lemma 1.1 *Let $A \in \mathbb{C}^{n\times n}$, $A = M_i - N_i$ $(i = 1, 2)$ be two splittings of the matrix A, and let $x^{(0)} \in \mathbb{C}^n$ be a given initial vector. If $\{x^{(k)}\}$ is a two-step iteration sequence defined by*

$$\begin{cases} M_1 x^{(k+\frac{1}{2})} = N_1 x^{(k)} + b, \\ M_2 x^{(k+1)} = N_2 x^{(k+\frac{1}{2})} + b, \end{cases}$$

$k = 0, 1, 2, \ldots$, then

$$x^{(k+1)} = M_2^{-1} N_2 M_1^{-1} N_1 x^{(k)} + M_2^{-1}(I + N_2 M_1^{-1})b, \;\; k = 0, 1, 2, \ldots \; .$$

Moreover, if the spectral radius $\varrho(M_2^{-1} N_2 M_1^{-1} N_1)$ of the iteration matrix $M_2^{-1} N_2 M_1^{-1} N_1$ is less than 1, then the iterative sequence $\{x^{(k)}\}$ converges to x^* for all initial vectors $x^{(0)} \in \mathbb{C}^n$.

Letting $T : \mathbb{C}^n \to \mathbb{C}^n$ be the mapping given by $T(x) = M_2^{-1} N_2 M_1^{-1} N_1 x + M_2^{-1}(I + N_2 M_1^{-1})b$, T will be a contraction mapping whenever

$$\|T(x) - T(y)\|_2 = \left\|M_2^{-1} N_2 M_1^{-1} N_1 (x - y)\right\|_2 \leq \|x - y\|_2, \;\; \text{for all } x, y \in \mathbb{C}^n \tag{2}$$

Moreover, x^* is a fixed point of T. It is easy to see that Inequality 2 will be satisfied when the spectral norm of the iteration matrix[1] is less than 1. In this case the Banach fixed-point theorem guarantees convergence of the iteration to the unique fixed point $x^* \in \mathbb{C}^n$. The HSS method exploits the fact that if A is a positive definite matrix, then T is necessarily a contraction for a class of pairs of splittings $A = M_i - N_i$, $i = 1, 2$, described below.

[1] The requirement of the lemma is that the spectral *radius* of the iteration matrix be less than 1, and the spectral radius is less than or equal to the spectral norm in general. Although, in the case of the HSS iteration, the iteration matrix is Hermitian, so the spectral radius and the spectral norm coincide.

Let $H = \frac{1}{2}(A + A^*)$, $S = \frac{1}{2}(A - A^*)$ be the Hermitian and skew-Hermitian parts of A, respectively. Then the HSS iteration method is defined by taking the two splittings $A = (H + \alpha I) - (\alpha I - S)$, $A = (\alpha I + S) - (\alpha I - H)$ for $\alpha > 0$ and then proceeding as in Lemma 1.1. More formally, the HSS method can be defined as below.

The HSS iteration method. *Given an initial guess $x^{(0)}$, for $k = 0, 1, 2, \ldots$, until $\{x^{(k)}\}$ converges compute*

$$\begin{cases} (\alpha I + H)x^{(k+\frac{1}{2})} = (\alpha I - S)x^{(k)} + b, \\ (\alpha I + S)x^{(k+1)} = (\alpha I - H)x^{(k+\frac{1}{2})} + b, \end{cases}$$

where α is a given positive constant.

It is readily seen that the iteration matrix for the HSS methods is given by

$$\mathcal{M}(\alpha) = (\alpha I + S)^{-1}(\alpha I - H)(\alpha I + H)^{-1}(\alpha I - S). \tag{3}$$

Crucially, Bai and Golub obtained an upper bound [1] for the spectral radius of $\mathcal{M}(\alpha)$, $\varrho(\mathcal{M}(\alpha)) \leq \sigma(\alpha)$, given by

$$\sigma(\alpha) \equiv \max_{\lambda_i \in \lambda(H)} \left| \frac{\alpha - \lambda_i}{\alpha + \lambda_i} \right|, \tag{4}$$

where $\lambda(H)$ is the spectrum of H. Now if A is positive definite, then so is H, hence $\lambda_i > 0,\ i = 1, 2, \ldots, n$. In such a case we will have

$$\varrho(\mathcal{M}(\alpha)) \leq \sigma(\alpha) < 1, \tag{5}$$

and so the HSS iteration will converge to x^* for positive definite A. Moreover, the upper bound $\varrho(\mathcal{M}(\alpha)) \leq \sigma(\alpha)$ is independent of the spectrum of A or S. Indeed, if $\gamma_R, \gamma_r > 0$ are the maximum and minimum eigenvalues of H respectively, then the optimal value of α is given by [1] $\bar{\alpha} = \sqrt{\gamma_R \gamma_r}$, and consequently

$$\sigma(\bar{\alpha}) = \frac{\sqrt{\gamma_R} - \sqrt{\gamma_r}}{\sqrt{\gamma_R} + \sqrt{\gamma_r}} = \frac{\sqrt{\kappa(H)} - 1}{\sqrt{\kappa(H)} + 1}. \tag{6}$$

Here $\kappa(H)$ is the spectral condition number of H.[2] Notice that $\bar{\alpha}$ minimises $\sigma(\alpha)$ but not necessarily $\varrho(\mathcal{M}(\alpha))$.

Since exact solution to the inner systems of the HSS iteration method can be costly to obtain, the inexact HSS iteration, or IHSS iteration, may be more useful in practice when solving large systems. This method is analogous to the HSS iteration except that the inner systems are now solved via approximate methods. It is stated in [1] as follows.

The IHSS iteration method *Given an initial guess $\bar{x}^{(0)}$, for $k = 0, 1, 2, \ldots$, until $\{x^{(k)}\}$ converges,*

1. *approximate the solution of $(\alpha I + H)\bar{z}^{(k)} = \bar{r}^{(k)}$ ($\bar{r}^{(k)} = b - A\bar{x}^{(k)}$) by iterating the residue satisfies*

$$\frac{\left\| \bar{r}^{(k)} - (\alpha I + H)\bar{z}^{(k)} \right\|}{\left\| \bar{r}^{(k)} \right\|} \leq \epsilon_k,$$

and then compute $\bar{x}^{(k+\frac{1}{2})} = \bar{x}^{(k)} + \bar{z}^{(k)}$;
2. *approximate the solution of $(\alpha I + S)\bar{z}^{(k+\frac{1}{2})} = \bar{r}^{(k+\frac{1}{2})}$ ($\bar{r}^{(k+\frac{1}{2})} = b - A\bar{x}^{(k+\frac{1}{2})}$) by iterating until $\bar{z}^{(k+\frac{1}{2})}$ is such that the residual satisfies*

$$\frac{\left\| \bar{r}^{(k+\frac{1}{2})} - (\alpha I + S)\bar{z}^{(k+\frac{1}{2})} \right\|}{\left\| \bar{r}^{(k+\frac{1}{2})} \right\|} \leq \eta_k,$$

and then compute $\bar{x}^{(k+1)} = \bar{x}^{(k+\frac{1}{2})} + \bar{z}^{(k+\frac{1}{2})}$.

[2] The spectral condition number is defined only for diagonalizable matrices. It is a measure of the sensitivity of the eigenvalues of a diagonalizable matrix to small perturbations. For a definition of the spectral condition number see [8]

The convergence analysis for the IHSS iteration is more complicated than that of the exact HSS iteration, since convergence now depends not only on the spectral condition number of H, but also on the relative errors $\{\epsilon_k\}$ and $\{\eta_k\}$ of the inner system solutions. The following is the main result of Bai and Golub [1] regarding the convergence of the IHSS iteration.

Theorem 1.2 *Let $\{\bar{x}^{(k)}\}$ be the sequence generated by the IHSS iteration given some initial guess $\bar{x}^{(0)}$. Then let*

$$\varrho = \left\|(\alpha I + S)(\alpha I + H)^{-1}\right\|_2, \quad \theta = \left\|A(\alpha I + S)^{-1}\right\|_2,$$

and also $\eta_R = \max\{\eta_k\}$, $\epsilon_R = \max\{\epsilon_k\}$. If $(\sigma(\alpha) + \theta\varrho\eta_R)(1 + \theta\epsilon_R) < 1$, then $\{\bar{x}^{(k)}\}$ converges to x^*.

Note that convergence does not require the limit of ϵ_k for large k to approach zero. Likewise for the limit of η_k.

In terms of computational complexity, the cost of the k-th IHSS iteration is $\mathcal{O}(4n + 2a + \chi_k(H) + \chi_k(S))$, where a is the cost of computing Ay for some $y \in \mathbb{C}^n$,[3] and $\chi_k(H)$ and $\chi_k(S)$ are the costs of solving the first and second inner systems, respectively, on the k-th iteration.

1.2 Discussion

The HSS and IHSS iteration methods have the great benefit that the convergence of the generated sequence can be controlled through the conditioning of H, instead of the condition number of A. This is evident from Equation 6, where the convergence rate of the HSS method is bounded above by considering only the condition number of H and not that of S. However, even the optimal $\sigma(\bar{\alpha})$ is in many cases a very generous upper bound for $\varrho(\mathcal{M}(\alpha))$, meaning that our control over the convergence through the condition number of H is weak.[4]

Moreover, the HSS and IHSS methods are limited by the efficiency of the solvers used to resolve the inner systems. Hence, more analysis is required to understand the structure of the inner systems for a given application, and to select appropriate solvers accordingly.

2 Optimal Parameter Selection

Of great interest to the application of HSS iteration and related methods is the optimal selection of the parameter α. As discussed in the foregoing section, while $\bar{\alpha}$ minimises the upper bound for the convergence rate of the iteration, it does not in fact minimise the convergence rate itself. Great effort has been made in finding a prescriptive method of determining $\alpha^* = \arg\min_\alpha\{\varrho(\mathcal{M}(\alpha))\}$. In [2] Bai and collaborators consider the problem for two-by-two block matrices with some success. In [7], Huang considers the problem more generally and obtains a cubic polynomial in terms of the traces of some matrices. A positive real solution to this cubic then gives an estimate of α^*.

2.1 Optimal Parameter Selection for Two-by-Two Matrices

Suppose the matrix A in Equation 1 is of the form

$$A = \begin{pmatrix} \lambda_1 I_r & E \\ -E^* & \lambda_2 I_s \end{pmatrix}, \quad \lambda_1 \geq \lambda_2 > 0 \tag{7}$$

where $I_r \in \mathbb{C}^{r\times r}$, $I_s \in \mathbb{C}^{s\times s}$ are the identities in their respective spaces, and $E \in \mathbb{C}^{r\times s}$. In [2] the authors seek α^* for the class of matrices in (7). In doing so they first consider the general case for real two-by-two matrices. Let $A \in \mathbb{R}^{2\times 2}$ and $H = \frac{1}{2}(A + A^*)$ with eigenvalues $\lambda_1 \geq \lambda_2 > 0$, and $S = \frac{1}{2}(A - A^*)$ has determinant $\det(S) = q^2$, $q \in \mathbb{R}$. Then it is found that the eigenvalues of α are given by

[3] this may be $o(n^2)$ if A has sparse structure

[4] see for example Bai and Golub's numerical experiments in applying the HSS iteration to the problem of the convection-diffusion equation in the unit cube [1]

$$\lambda_{\pm} = \frac{(\alpha^2 - \lambda_1\lambda_2)(\alpha^2 - q^2) \pm \sqrt{(\alpha^2 - \lambda_1\lambda_2)^2(\alpha^2 - q^2)^2 - (\alpha^2 - \lambda_1^2)(\alpha^2 - \lambda_2^2)(\alpha^2 + q^2)^2}}{(\alpha + \lambda_1)(\alpha + \lambda_2)(\alpha^2 + q^2)} \tag{8}$$

and so if

$$(\alpha^2 - \lambda_1\lambda_2)^2(\alpha^2 - q^2) \geq (\alpha^2 - \lambda_1^2)(\alpha^2 - \lambda_2^2)(\alpha^2 + q^2)^2 \tag{9}$$

then $\varrho(\mathcal{M}(\alpha))$ is given by

$$\frac{\left|\alpha^2 - \lambda_1\lambda_2\right|\left|\alpha^2 - q^2\right| + \sqrt{(\alpha^2 - \lambda_1\lambda_2)^2(\alpha^2 - q^2)^2 - (\alpha^2 - \lambda_1^2)(\alpha^2 - \lambda_2^2)(\alpha^2 + q^2)^2}}{(\alpha + \lambda_1)(\alpha + \lambda_2)(\alpha^2 + q^2)}. \tag{10}$$

Otherwise if

$$(\alpha^2 - \lambda_1\lambda_2)^2(\alpha^2 - q^2)^2 < (\alpha^2 - \lambda_1^2)(\alpha^2 - \lambda_2^2)(\alpha^2 + q^2)^2 \tag{11}$$

then $\varrho(\mathcal{M}(\alpha))$ is equal to

$$\sqrt{\frac{(\alpha - \lambda_1)(\alpha - \lambda_2)}{(\alpha + \lambda_1)(\alpha + \lambda_2)}}. \tag{12}$$

Minimising $\varrho(\mathcal{M}(\alpha))$ by simply differentiating the above expressions and solving for local minima is not straightforward due to the square root term in (10). However, it is possible through some analysis to derive the result stated in Lemma 2.1.

Lemma 2.1 *Let $A \in \mathbb{R}^{2\times 2}$. Let also $H = \frac{1}{2}(A + A^T)$ have eigenvalues $\lambda_1 \geq \lambda_2 > 0$ and $S = \frac{1}{2}(A - A^T)$ have $\det(S) = q^2$, $q \in \mathbb{R}$. Define also the polynomials*

$$p_1(x) = (x^2 + q^2)^2(x^2 - \lambda_1^2)(x^2 - \lambda_2^2) - (x^2 - q^2)^2(x^2 - \lambda_1\lambda_2),$$
$$p_2(x) = (x^2 + q^2)^2(\lambda_1^2 - x^2)(x^2 - \lambda_2^2) - (x^2 - q^2)^2(x^2 - \lambda_1\lambda_2).$$

Then $\alpha^* = \arg\min_\alpha\{\varrho(\mathcal{M}(\alpha))\}$ is contained in the finite set

$$S = \{\alpha > 0 : p_1(\alpha) = 0 \ \text{ or } \ p_2(\alpha) = 0\} \tag{13}$$

The key insight in the proof of Lemma 2.1 is that $\varrho(\mathcal{M}(\alpha))$ is minimised when $\lambda_\pm$ have the same modulus. Then one derives a sufficient condition for this, in this case that one of $p_1(\alpha) = 0$ or $p_2(\alpha) = 0$ is satisfied. If some additional conditions are met, this result can be extended to matrices of the form introduced in (7). The idea of the proof in this case is to decompose A into a direct sum of real two-by-two matrices and then proceed as in the proof of Lemma 2.1 (although the details are much more involved).

Theorem 2.2 *With A as in (7) and with the non-zero singular values of $E \in \mathbb{C}^{r\times s}$ satisfying $q_1 \geq q_2 \geq \ldots \geq q_k$. Then if $\lambda_1 = \lambda_2 = \lambda^*$, we will have $\alpha^* = \lambda^*$. Otherwise if $\lambda_1 > \lambda_2 > 0$, then α^* is either a solution to*

$$(x - \sqrt{\lambda_1\lambda_2})(x - \sqrt{q_1 q_2}) = 0,$$

or it is a root of one of the following polynomials.

$$p_1^j(x) = (x^2 + q_j^2)^2(x^2 - \lambda_1^2)(x^2 - \lambda_2^2) - (x^2 - q_j^2)^2(x^2 - \lambda_1\lambda_2), \ \ j \in \{1, k\},$$
$$p_2^j(x) = (x^2 + q_j^2)^2(\lambda_1^2 - x^2)(x^2 - \lambda_2^2) - (x^2 - q_j^2)^2(x^2 - \lambda_1\lambda_2), \ \ j \in \{1, k\}.$$

2.2 A Very General Estimate

Another approach (taken in [7]) is to recast the HSS iteration method as a single-step iteration and consider the behaviour of this mathematically equivalent formulation. Returning now to the general case of (1), If we define

$$M(\alpha) \coloneqq \frac{1}{2\alpha}(\alpha I + H)(\alpha I + S), \quad N(\alpha) \coloneqq \frac{1}{2\alpha}(\alpha I - H)(\alpha I - S) \tag{14}$$

This gives a splitting $A = M(\alpha) - N(\alpha)$. This splitting then generates a single-step iteration process defined by the recursive relationship

$$M(\alpha)x^{(k+1)} = N(\alpha)x^{(k)} + b, \quad k = 0, 1, 2, \dots . \tag{15}$$

Indeed, this iterative process converges to x^* if the spectral radius of the iteration matrix $M(\alpha)^{-1}N(\alpha)$ is less than 1. Now, since H is positive definite, $(\alpha I + H)$ is never singular for $\alpha > 0$, so we can formally write

$$(\alpha I + H)^{-1} = \sum_{k=0}^{\infty} (-1)^k \alpha^{-k} H^k. \tag{16}$$

This implies that the commutator $[H, (\alpha I + H)^{-1}] = 0$ and so

$$\begin{aligned} M(\alpha)^{-1}N(\alpha) &= (\alpha I + S)^{-1}(\alpha I + H)^{-1}(\alpha I - H)(\alpha I - S) \\ &= (\alpha I + S)^{-1}(\alpha I - H)(\alpha I + H)^{-1}(\alpha I - S) = \mathcal{M}(\alpha). \end{aligned}$$

Hence, to minimise $\varrho(\mathcal{M}(\alpha))$ we may consider minimising $N(\alpha)$ so that $M(\alpha)^{-1}N(\alpha) \approx 0$. This may be done by considering the minimum of the function

$$\Phi(\alpha) \coloneqq \|N(\alpha)\|_F^2 = \operatorname{Tr}(N(\alpha)N(\alpha)^*), \tag{17}$$

where $\|\cdot\|_F$ is the Frobenius norm. Through straightforward calculation, minimising $\Phi(\alpha)$ turns out to be equivalent to minimising the polynomial [7][5]

$$\phi(\alpha) \coloneqq 4n\alpha^3 + 3a\alpha^2 + 2b\alpha + c. \tag{18}$$

with

$$a = -2\operatorname{Tr}(H), \quad b = \operatorname{Tr}(H^2) - \operatorname{Tr}(S^2), \quad c = 2\operatorname{Tr}(HS^2), \quad d = -\operatorname{Tr}(H^2S^2). \tag{19}$$

Since in the HSS iteration method we have $\alpha > 0$, of interest are positive real roots of $\phi(\alpha)$. It is also a fact that for $V \in \mathbb{C}^{n\times n}$ Hermitian positive definite and $U \in \mathbb{C}^{n\times n}$ skew-Hermitian, $\operatorname{Tr}(VU^2) \leq 0$ [7], in particular we have $c \leq 0$. Clearly if $c < 0$ we will have a positive real root of $\phi(\alpha)$. Otherwise if $c = 0$ then $\alpha = 0$ is one root and the other two roots are

$$\alpha_{\pm} = \frac{-3a \pm \sqrt{9a^2 - 32nb}}{8n},$$

so we will have a positive real solution whenever $9a^2 - 32nb \geq 0$. If in any case there is more than one positive root of $\phi(\alpha)$, then we will have to check which is the global minimiser of $\Phi(\alpha)$.

Numerical experiments show that this method outperforms that purposed in [2] in required CPU times and iteration count, even when the system is of the type seen in (7).[7]

[5] It does seem that in [7] the minimised quantity is $4\alpha^2\Phi(\alpha)$. Based on numerical experiments, this appears to be more than good enough for applications, however it is an estimate of the optimal parameter and not an exact figure.

2.3 A New Method of Estimation

The estimation method for the selection of α discussed in the previous subsection has the advantage that it is quite general and makes very few assumption about the structure of A. However, in some cases, such as when using finite difference methods to solve the convection-diffusion equation, we do in fact know a lot about the sparsity structure of A. In particular, using the scheme discussed in Section 3, we will have for the convection-diffusion problem that

$$H = H_x \otimes I \otimes I + I \otimes H_y \otimes I + I \otimes I \otimes H_z, \tag{20}$$

where we have defined

$$H_x = \text{tridiag}(1, 6, -1), \quad H_y = H_z = \text{tridiag}(1, 0, -1). \tag{21}$$

We will also have the skew-Hermitian part

$$S = S_x \otimes I \otimes I + I \otimes S_y \otimes I + I \otimes I \otimes S_z, \tag{22}$$

where

$$S_\xi = \text{tridiag}(-r_\xi, 0, r_\xi), \quad \xi \in \{x, y, z\}, \tag{23}$$

with the r_ξ known parameters of the problem. For this class of problem we introduce a new method of estimating the optimal value of α. The idea is again to minimise $N(\alpha)$ introduced in (14) for some choice of metric, but instead of minimising the Frobenius norm $\|N(\alpha)\|_F$ we will instead minimise an upper bound on the spectral norm

$$\Psi(\alpha) \coloneqq \|N(\alpha)\|_2. \tag{24}$$

In the interest of brevity we will consider the case where $r_\xi = r$ for all $\xi \in \{x, y, z\}$, but the method can be immediately extended to the more general case. We will also have need of the following bound, sometimes called 'Schur's test'.[9]

For any $A \in \mathbb{C}^{m\times n}$,

$$\|A\|_2 \leq \sqrt{\|A\|_\infty \|A\|_1}. \tag{25}$$

Moreover, since $\|\cdot\|_2$ is induced by the 2-norm on $\mathbb{C}^n$ it is submultiplicative. Note first of all that given the structure of H and S, we have

$$\|\alpha I - H\|_\infty = \|\alpha I - H\|_1 = (|6 - \alpha| + 4) \tag{26}$$
$$\|\alpha I - S\|_\infty = \|\alpha I - S\|_1 = (\alpha + 4r). \tag{27}$$

These together give

$$\begin{aligned}
\Psi(\alpha) &= \frac{1}{2\alpha}\|(\alpha I - H)(\alpha I - S)\| \\
&\leq \frac{1}{2\alpha}\|\alpha I - H\|_2 \|\alpha I - S\|_2 \\
&\leq \frac{1}{2\alpha}\|\alpha I - H\|_1 \|\alpha I - S\|_1 \\
&= \frac{1}{2\alpha}(\alpha + 6r)(|6 - \alpha| + 6)
\end{aligned} \tag{28}$$

From Figure 1 it is clear that $\alpha = 6$ is the value that always minimises this upper bound, regardless of the value of r. In the numerical experiments detailed in Section 3 we compare this estimate of α^* to that obtained by using the method outlined in Subsection 3.2. Note that our method can be easily generalised to the case of the convection diffusion equation in d-dimensions. We need only make minor modifications to the expressions for $\|\alpha I - H\|_1$ and $\|\alpha I - S\|_1$.

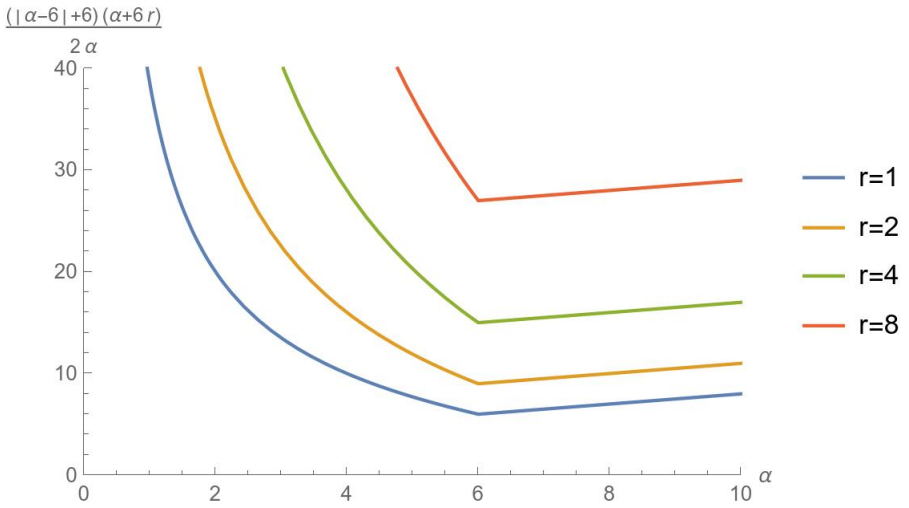

Figure 1: Plot of the bound obtained on $\Psi(\alpha)$ for different values of r. Visually it is clear that the minimum of this bound always occurs at $\alpha = 6$.

3 Examples and Experiments

3.1 A Prototypical Use Case

The prototypical application case for the HSS and IHSS iteration methods is the equilibrium state convection-diffusion equation. [1, 4, 3, 7] The convection-diffusion equation is in general given by (cite Thomas Stocker)

$$\frac{\partial v}{\partial t} = \nabla \cdot (D\nabla v) - \nabla(v\vec{u}) + f, \tag{29}$$

where v is the unknown, D is the diffusion coefficient, $\vec{u}$ is the velocity field of the medium, and f is a source term. In the case of equilibrium flow, and assuming also that D and $\vec{u}$ are constant, the convection-diffusion equation reduces to

$$-\nabla^2 v + \vec{u} \cdot \nabla v = f. \tag{30}$$

Equation 30 considered on the unit cube (or square) with Dirichlet boundary conditions serves as the main test for the HSS method. This is for two reasons. Firstly, it is time-independent: if there was a time-dependence, (and hence initial or final conditions) a brute force method of solving the PDE would simply involve a single term recurrence relation without a need for solving linear systems. [cite something] Moreover, unlike the Poisson equation (i.e. the case where $\vec{u} = 0$) which is Hermitian positive definite, the differential operator of the equilibrium state convection diffusion equation is non-Hermitian positive definite, hence it demonstrates more fully the advantages of the HSS method. Suppose we were to convert (30) into a linear set of equations by using a uniform step size $h = \frac{1}{N+1}$ in every dimension. We will denote the function v evaluated at $(x_i = ih, y_j = jh, z_k = kh)$ by $v_{i,j,k}$, with $i, j, k = 0, 1, 2, \ldots, N+1$. The Dirichlet boundary conditoins impose the constraints $v_{0,j,k} = v_{i,0,k} = v_{i,j,0} = 0$ and also $v_{N+1,j,k} = v_{i,N+1,k} = v_{i,j,N+1} = 0$. In our linearalisation, both the first and second derivatives are approximated by the central difference scheme, i.e.

$$\frac{\partial v}{\partial x}|_{(x_i,y_j,z_k)} \approx \frac{v_{i+1,j,k} - v_{i-1,j,k}}{2h} \tag{31}$$

$$\frac{\partial^2 v}{\partial x^2}|_{(x_i,y_j,z_k)} \approx -\frac{2v_{i,j,k} - v_{i-1,j,k} - v_{i+1,j,k}}{h^2} \tag{32}$$

and similar in the y and z directions. Here we will introduce the following definitions.

Definition 3.1 *Let $X \in \mathbb{C}^{m\times n}$, $Y \in \mathbb{C}^{p\times q}$. Then the Kronecker product of X and Y, denoted $X \otimes Y$, is the $(m \cdot p) \times (n \cdot q)$ matrix*

$$\begin{pmatrix} X_{11}Y & \dots & X_{1n}Y \\ \vdots & & \vdots \\ X_{m1}Y & \dots & X_{mn}Y \end{pmatrix}$$

Definition 3.2 *Let T be a $n \times m \times p$ two-tensor. Then* $\mathrm{vec}(T)$ *is the nmk dimensional vector given by arranging the elements T through the first dimension, then through the second, and lastly through the third. In other words*

$$[\mathrm{vec}(T)]_{i+mj+kp} = T_{i,j,k}$$

For details about the Kronecker product and the $\mathrm{vec}(\cdot)$ function see [5, pp 274-276]. Let $\vec{u} = (u_x, u_y, u_z)$ and define the mesh Reynold's numbers

$$r_x = \frac{u_x h}{2}, \;\; r_y = \frac{u_y h}{2}, \;\; r_z = \frac{u_z h}{2}. \tag{33}$$

Then, using the approximation scheme described, (30) can be written [6, 5]

$$A \cdot \mathrm{vec}(v) = h^2 \,\mathrm{vec}(f), \tag{34}$$

where

$$A = T_x \otimes I \otimes I + I \otimes T_y \otimes I + I \otimes I \otimes T_z, \tag{35}$$

with

$$T_x = \mathrm{tridiag}(-1-r_x, 6, -1+r_x), \;\; T_y = \mathrm{tridiag}(-1-r_y, 0, -1+r_y), \;\; T_z = \mathrm{tridiag}(-1-r_z, 0, -1+r_z). \tag{36}$$

3.2 Numerical Results

We implemented a numerical solver for the convection-diffusion equation in (30) using the IHSS iteration method. We used the source term

$$f(x,y,z) = S_0 \exp\left\{-\left((x-1/4)^2 + (y-1/4)^2 + (z-1/4)^2\right)/r^2\right\} \tag{37}$$

with $S_0 = 10$ and $r = 0.1$ and the initial guess $x^{(0)} = \mathrm{vec}(f)$. Additionally, in each trial iteration was terminated once the condition

$$\left\| Ax^{(k)} - b \right\|_2 \leq 1.0 \times 10^{-5} \tag{38}$$

was met. The GMRES method was used to resolve the inner systems in each iteration. We also took $u_\xi = u, \;\; \xi \in \{x, y, z\}$. To select a suitable value for α, we compare the method proposed in [7] (as discussed in Subsection 3.2) with the method we introduced in Subsection 3.3 and we share the results in Tables 2-4. where the system times, number of outer iterations and the average time per outer iteration are presented for some parameters. In each case the total number of grid points is N^3.

We denote the estimate of optimal alpha obtained via our method by $\alpha_\Psi = 6$, while that obtained via the method in [7] is denoted by α_Φ. Table (something) shows the calculated values of α_Φ for some different values of u. From this data it seems that $\alpha_\Phi \to 6 = \alpha_\Psi$ in the limit of large u. Moreoever, α_Ψ gives superior results for small u as documented in Table 1. These facts taken together seems to indicate that α_Ψ is indeed a better estimate than α_Φ for α^* in the use case considered here.

The Python code can be found in the `HSS_exper.py` file. The `test_HSS_exper.py` file tests the HSS and IHSS iteration implementation in `HSS_exper.py`. The machine used was a personal computer with an

Intel Core i7-8565 CPU and 16 GB of memory running Ubuntu 20.04.3. Options to record precise CPU times for a process are limited on the Python platform, so instead we elected to measure the average system time for each iteration. The product of this average time with the number of iterations then gives an estimate of the CPU time.

$n = 8$	
u	α_Φ
1	0.00235
10	0.252
100	5.842
1000	5.998
10000	5.99998

Table 1

$N = 8$, $u = 1$		
	$\alpha_\Phi = 0.00235$	α_Ψ
Outer iterations	$500+^6$	118
Average time per iteration (ms)	302	279
Estimated CPU time(s)	-	32.9

Table 2

$N = 8$, $u = 10$		
	$\alpha_\Phi = 0.252$	α_Ψ
Outer iterations	191	40
Average time per iteration (ms)	2501	465
Estimated CPU time(s)	477.6	18.6

Table 3

$N = 8$, $u = 1000$		
	$\alpha_\Phi = 5.998$	α_Ψ
Outer iterations	80	80
Average time per iteration (ms)	2186	2536
Estimated CPU time(s)	174.8	202.8

Table 4

3.3 Discussion

Given that A fulfils the assumptions outlined in (20) and (22), the method proposed by us for the selection of α outperforms that proposed in [7]. Moreover, $\alpha_\Phi \to \alpha_\Psi$ in the limit of large u, and the performace of the two methods converges. This new method is also more suitable for matrix-free implementations since no explicit information about A, H, or S is needed to compute $\alpha_\Psi = 6$ (other than their sparsity structure which is assumed). However, this new method is far more limited in its use cases as it makes very strong assumption about the structure of H and S.

References

[1] Zhong-Zhi Bai, Gene H Golub, and Michael K Ng. "Hermitian and skew-Hermitian splitting methods for non-Hermitian positive definite linear systems". In: *SIAM Journal on Matrix Analysis and Applications* 24.3 (2003), pp. 603–626.

[2] Zhong-Zhi Bai, Gene H. Golub., and Chi-Kwong Li. "Optimal Parameter in Hermitian and Skew-Hermitian Splitting Method for Certain Two-by-Two Block Matrices". In: *SIAM Journal on Scientific Computing* 28.2 (2006), pp. 583–603. DOI: 10.1137/050623644. eprint: https://doi.org/10.1137/050623644. URL: https://doi.org/10.1137/050623644.

[3] Zhong-Zhi Bai and Xue-Ping Guo. "On Newton-HSS methods for systems of nonlinear equations with positive-definite Jacobian matrices". In: *Journal of computational mathematics* (2010), pp. 235–260.

[4] Zhong-Zhi Bai et al. "Block triangular and skew-Hermitian splitting methods for positive-definite linear systems". In: *SIAM Journal on Scientific Computing* 26.3 (2005), pp. 844–863.

[5] James W. Demmel. *Applied Numerical Linear Algebra*. Society for Industrial and Applied Mathematics, 1997. DOI: 10.1137/1.9781611971446. eprint: https://epubs.siam.org/doi/pdf/10.1137/1.9781611971446. URL: https://epubs.siam.org/doi/abs/10.1137/1.9781611971446.

[6] Chen Greif and James Varah. "Block stationary methods for nonsymmetric cyclically reduced systems arising from three-dimensional elliptic equations". In: *SIAM journal on matrix analysis and applications* 20.4 (1999), pp. 1038–1059.

[7] Yu-Mei Huang. "A practical formula for computing optimal parameters in the HSS iteration methods". In: *Journal of Computational and Applied Mathematics* 255 (2014), pp. 142–149. ISSN: 0377-0427. DOI: https://doi.org/10.1016/j.cam.2013.01.023. URL: https://www.sciencedirect.com/science/article/pii/S0377042713002483.

[8] Erxiong Jiang and Peter C.B. Lam. "An upper bound for the spectral condition number of a diagonalizable matrix". In: *Linear Algebra and its Applications* 262 (1997), pp. 165–178. ISSN: 0024-3795. DOI: https://doi.org/10.1016/S0024-3795(97)80029-8. URL: https://www.sciencedirect.com/science/article/pii/S0024379597800298.

[9] Roy Mathias. "The spectral norm of a nonnegative matrix". In: *Linear Algebra and its Applications* 139 (1990), pp. 269–284. ISSN: 0024-3795. DOI: https://doi.org/10.1016/0024-3795(90)90403-Y. URL: https://www.sciencedirect.com/science/article/pii/002437959090403Y.

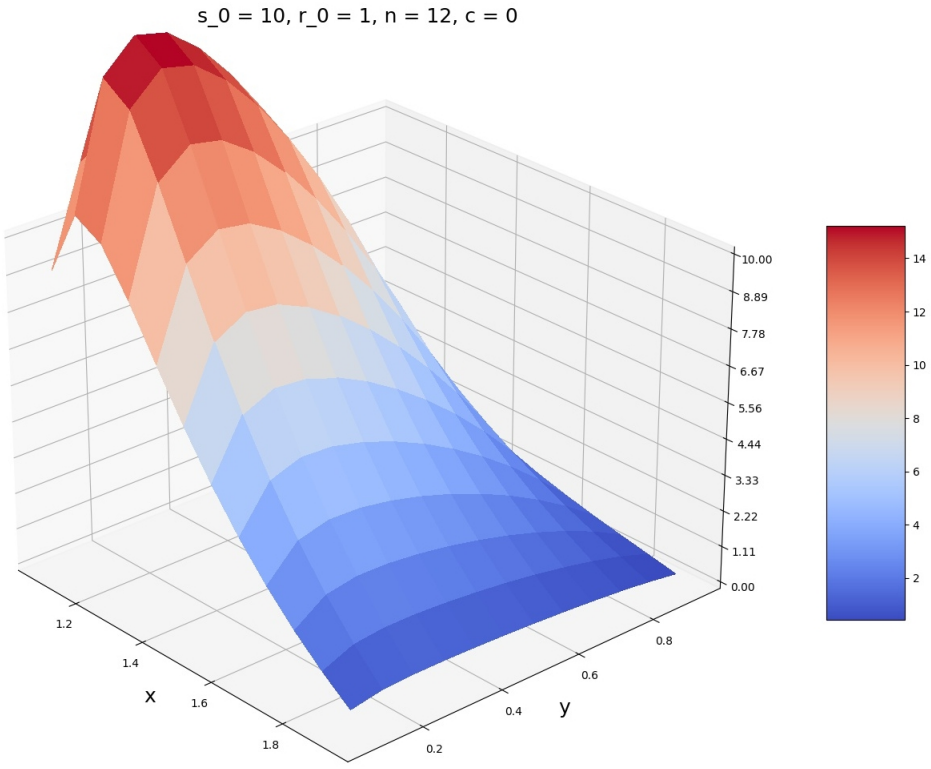

Figure 2: A typical solution obtained using our solver. Included for no particular reason.