Backward error and condition number of a generalized Sylvester equation, with application to the stochastic Galerkin method

Pranesh, Srikara

2019

ISSN 1749-9097

Backward error and condition number of a generalized Sylvester equation, with application to the stochastic Galerkin method[1]

Srikara Pranesh[a,*]

[a]*School of Mathematics, University of Manchester, Manchester M139PL*

Abstract

The governing equations of the stochastic Galerkin method can be formulated as a generalized Sylvester equation. Therefore developing solvers for it is attracting a lot of attention from the uncertainty quantification community. In this regard Krylov subspace based iterative solvers, which are used for standard linear systems are being used for the generalized Sylvester equations as well. This is achieved by converting the generalized Sylvester equation to a standard linear system using the Kronecker product. Accordingly the residual is used as a stopping criterion for the iterations, and the condition number of linear systems is used for the generalized Sylvester equations as well. For a linear system a small residual implies a small backward error, and hence using residual as a stopping criterion is justified. In this work we prove that this need not be the case for the generalized Sylvester equation. We introduce two definitions for the backward error, and then derive an upperbound on each of them. We also verify the predictions of the analysis using numerical experiments. For the special case of the stochastic Galerkin method we show that the upper bound on the backward error can be computed with minimal computational overhead, and hence it can be used as a stopping criterion in the iterative solvers. For the matrices from the stochastic Galerkin method we numerically demonstrate that the actual backward error can be upto 2 orders of magnitude higher than the relative residual. Finally by taking into account the structure of the equation we derive an expression for the condition number, and discuss an algorithm for their computation in the special case of the stochastic Galerkin method.

Keywords: generalized Sylvester equation, residual, backward error, condition number, stochastic Galerkin method.
2010 MSC: 15A24, 35R60, 65F10, 65F35

[1] Version of June 19, 2019. **Funding:** This work was supported by Engineering and Physical Sciences Research Council grant EP/P020720/1. The opinions and views expressed in this publication are those of the authors, and not necessarily those of the funding bodies.

[*] Corresponding author

Email address: srikara.pranesh@manchester.ac.uk (Srikara Pranesh)

1. Introduction

The equation

$$\sum_{i=1}^{p} A_i X B_i = C, \tag{1.1}$$

where $\{A_i\}_{i=1}^{p} \in \mathbb{R}^{n\times n}$, $\{B_i\}_{i=1}^{p} \in \mathbb{R}^{m\times m}$, and $X, C \in \mathbb{R}^{n\times m}$ is called as the generalized Sylvester equation. Using the Kronecker product notation, this can also be written as

$$\left[\sum_{i=1}^{p} (B_i^T \otimes A_i)\right] \text{ Vec}(X) = \text{ Vec}(C), \tag{1.2}$$

where $A \otimes B = (a_{ij}B)$ is the Kronecker product, and the $\text{Vec}(\cdot)$ stacks the columns of a matrix one above the other forming a vector. Until recently this equation was thought to be of only theoretical interest [1, Sec 16.5], and now it is known that the governing equation in the stochastic Galerkin method can be reformulated as the generalized Sylvester equation [2]. This development has initiated a great interest in the uncertainty quantification community to develop efficient algorithms, which exploits its matrix structure. Several attempts in this directions have already been made, for example [2], [3], [4], [5], and they are based on the Krylov subspace based iterative methods. All these Krylov subspace based algorithms share the following common features.

1. The norm of the residual is used to measure the stability of the computed solution.
2. If $A \in \mathbb{R}^{n\times n}$ and $M \in \mathbb{R}^{n\times n}$ is a preconditioner for A, then $\kappa_2(M^{-1}A) = \| (M^{-1}A)^{-1} \|_2 \|M^{-1}A\|_2$ — 2-norm condition number — is used to measure the effectiveness of the preconditioner.

The above two criterions are standard procedures adopted in the solution of linear systems, but (1.1) is a matrix equation, that is unknown is a matrix rather than a vector. Therefore in this work we examine the suitability of the above two criterions for the solution of (1.1) in the context of the stochastic Galerkin method. The main contribution of this work is to demonstrate that the tools for the analysis of a standard linear system do not carry over in a straightforward manner to the generalized Sylvester equation, and it should be analysed separately considering the matrix structure of the equation.

The residual is usually used as a stopping criterion for the solution of linear systems by iterative methods, because a small residual implies a small backward error [6, Sec 4.2]. However this need not be true for (1.1) since the unknown is a matrix. If it is the case then the computed solution would be an exact solution to a problem which is a much larger perturbation of the original problem than stipulated. This is indeed the case for the Sylvester and the Lyapunov equations, as shown by Higham in [7]. Similar result for a two term generalized Sylvester equation was proved by Kågström in [8]. In this work we will prove a similar result for a generalized Sylvester equation with p terms. One unique feature of the backward error analysis of the generalized Sylvester equation is that, unlike Sylvester or

Lyapunov equations, the perturbations appear non-linearly. To address this issue we consider two alternative definitions of the backward error, and derive an upper bound on each of them. All the predictions made by the theory are verified using numerical experiments. We simplify the results for the special case of the stochastic Galerkin method, and demonstrate that the bound on the backward error can be derived with minimal computational overhead. Using numerical experiments we will compare the actual backward error and the relative residual and demonstrate that the former can be upto 2 order of magnitude higher. For a complete discussion on the perturbation theory of the generalized Sylvester equation we refer to [9, Ch 8], however they do not discuss the backward error analysis.

The condition number of a function is its sensitivity to the change in its output with respect to the input. Therefore the condition number depends on the structure of the function under consideration. Since (1.1) can be represented as a standard linear system, condition number of a linear system is used for it as well. A major drawback of this approach is that it completely ignores the structure of the equation. In this work we derive the Frobenius norm condition number of the generalized Sylvester equation by considering its structure. Again for the special case of the matrices from stochastic Galerkin method we discuss an efficient way to estimate the condition number.

Only recently the p term generalized Sylvester equation is being considered by the numerical linear algebra community, and work in this regard is very sparse. Fundamental results such as the existence and uniqueness of the solution of a generalized Sylvester equation for a general A_i and B_i are not available. However for the specific case of the stochastic Galerkin application this result is available [2, Sec 2], and therefore in the perturbation analysis we will assume that (1.2) is nonsingular. A discussion regarding the general case is beyond the scope of this work.

The rest of the paper is organised as follows. In the next section we will briefly describe the stochastic Galerkin method. In section 3 we introduce two definitions for the backward error of a generalized Sylvester equation and derive an upper bound on both of them, and in section 4 we derive an expression for the condition number of the generalized Sylvester equation. Next in section 5 we specialise the results of sections 3 and 4 for the stochastic Galerkin application, and discuss algorithms for their computation. In Section 6 we perform numerical experiments to verify the predictions made by the analysis of sections 3 and 4. Further in this section we consider matrices from the stochastic Galerkin discretisation of an elliptic stochastic partial differential equation. We summarise and enumerate a few open problems in Section 7.

2. Stochastic Galerkin formulation

To clarify the context of the problem, and for the sake of completeness, in this section we provide a very brief introduction to the stochastic Galerkin formulation. Specifically we consider an elliptic stochastic partial differential equation (spde). If $\boldsymbol{x} \in \mathcal{D} \subset \mathbb{R}^d$, where usually $d = 1, 2, 3$, then an elliptic spde is given as

$$\begin{aligned} -\nabla \cdot [\kappa(\boldsymbol{x}, \theta) \nabla u(\boldsymbol{x}, \theta)] &= f(\boldsymbol{x}, \theta) \quad a.s. \\ \text{with } u(\boldsymbol{x}) &= 0 \quad \text{on } \partial\mathcal{D}, \end{aligned} \tag{2.1}$$

where $a.s$ means *almost surely*, $\partial\mathcal{D}$ is the boundary of $\mathcal{D}$. Let $(\Omega, \mathcal{F}, \mathcal{P})$ be the probability space, then $\theta \in \Omega$, where Ω is the sample space. Further the input coefficients $\kappa(\boldsymbol{x}, \theta)$ and $f(\boldsymbol{x}, \theta)$ are modelled as real valued random fields and are assumed to be positive and finite. This assumption guarantees the positive definiteness of the final matrix [2, Sec 2]. Now using the finite element method to discretise (2.1) we obtain

$$A(\theta)x(\theta) = b(\theta), \tag{2.2}$$

where $A(\theta) \in \mathbb{R}^{n\times n}$, $c(\theta), x(\theta) \in \mathbb{R}^n$, and n is the dimension of the finite element space. At this point there are various possible ways to solve (2.2), and the stochastic Galerkin method is one of them, which consists of the following two steps.

1. The matrix $A(\theta)$ and the vector $b(\theta)$ are discretised as

$$\begin{aligned} A(\theta) &= A_1 + \sum_{i=2}^{p} A_i \xi_i, \text{and} && (2.3) \\ b(\theta) &= b_1 + \sum_{i=p+1}^{p_1} b_i \xi_i. && (2.4) \end{aligned}$$

   where ξ_i are bounded, zero mean, unit variance, and uncorrelated random variables. Furthermore A_i are sparse, symmetric and positive definite. (2.3) and (2.4) are called as the finite-noise assumption and can be computed using the Karhunen-Löeve expansion [10].

2. The solution $x(\theta)$ is discretised using orthogonal polynomials, which are function of the random variables ξ_i in (2.3) and (2.4). Orthogonality is defined with respect to the probability measure $\mathcal{P}$, and they are called as the generalised polynomial chaos (gPC) functions[2]. Accordingly we obtain

$$x(\theta) = \sum_{i=1}^{m} x_i \psi_i(\theta), \tag{2.5}$$

   where ψ_i-s are the gPC bases functions.

$$m = (p_1 + q - 3)!/(p_1 - 3)!q! \tag{2.6}$$

   and q is the maximum degree of the polynomials $\psi_i(\theta)$.

Now substituting (2.3), (2.4), (2.5) in (2.2), and applying Galerkin projection we obtain (1.1), where $\{A_i\}_{i=1}^{p}$ is given by (2.3), and $X = [x_1, x_2, \cdots, x_m]$, where $x_i \in \mathbb{R}^n$ are from (2.5). Further $C = [c_1, c_2, \cdots c_m]$, where

$$\begin{aligned} c_i &= b_1\mathbb{E}(\psi_i) + \sum_{j=p+1}^{p_1} b_j \mathbb{E}(\xi_j\psi_i), \qquad \text{for } i = 1, 2, \cdots, m, \\ (B_i)_{jk} &= \mathbb{E}(\xi_i\psi_j\psi_k), \qquad \text{for } i = 1, 2, \cdots, p, \quad j, k = 1, 2, \cdots, m, && (2.7) \end{aligned}$$

[2] These gPC function are the bases of a subspace of $L^2(\Omega)$ [11]

and $\mathbb{E}(\cdot)$ denotes the expectation operator. For a detailed description of the formulation we refer interested reader to [12, Ch 1] and references therein. In the next section we will consider a generalized Sylvester equation— no structure is assumed on A_i, B_i — and derive an upper bound on a backward error. However we will return to the special case of the generalized Sylvester equation arising from the stochastic Galerkin method in section 5. In this work we consider only matrices with real entries, as the stochastic Galerkin method results in real matrices. However results for the complex case follows directly from the analysis in the next section.

3. Backward error

For the Sylvester equation, which is $AX - XB = C$ with appropriate matrix dimensions, Higham in [1, Ch 16] showed that the backward error can be written in terms of the perturbation in the component matrices. That is, if we define the normwise backward error as

$$\eta^{'}(Y) = \min\Big\{\epsilon : (A + \Delta A)Y - Y(B + \Delta B) = C + \Delta C, \quad \|\Delta A\|_F \leq \epsilon\alpha, \\ \|\Delta B\|_F \leq \epsilon\beta, \quad \|\Delta C\|_F \leq \epsilon\gamma\Big\},$$

where $\|\cdot\|_F$ is the Frobenius norm, $\alpha, \beta, \gamma > 0$, and Y is an approximate solution. Then it was shown that

$$\eta^{'}(Y) \leq \mu\frac{\|R'\|_F}{(\alpha + \beta)\|Y\|_F + \gamma},$$

where $R' = C - AY + YB$ is the residual,

$$\mu := \frac{(\alpha + \beta)\|Y\|_F + \gamma}{(\alpha^2\sigma_m^2 + \beta^2\sigma_n^2 + \gamma^2)^{1/2}}, \tag{3.1}$$

and σ_n, σ_m are the n-th and m-th singular values of Y. Similar idea was used in [8] for a two term generalized Sylvester equation.

Adopting a similar strategy, again if Y is an approximate solution of a generalized Sylvester equation, then we define

$$\eta(Y) = \min\Big\{\epsilon : \sum_{i=1}^{p}(A_i + \Delta A_i)Y(B_i + \Delta B_i) = C + \Delta C, \quad \|\Delta A_i\|_F \leq \epsilon\alpha_i, \\ \|\Delta B_i\|_F \leq \epsilon\beta_i, \text{ for } i \in \{1, 2, 3, \cdots, p\}, \text{ and } \|\Delta C\|_F \leq \epsilon\gamma\Big\}, \tag{3.2}$$

where α_i, β_i and γ are positive and indicators of the extent of perturbation. The choice $\alpha_i = \|A_i\|_F$, $\beta_i = \|B_i\|_F$, and $\gamma = \|C\|_F$ are of practical interest, and in such cases we refer to $\eta(Y)$ as the *normwise relative backward error.* Now the perturbed generalized Sylvester equation

$$\sum_{i=1}^{p}(A_i + \Delta A_i)Y(B_i + \Delta B_i) = C + \Delta C$$

can be written as

$$\sum_{i=1}^{p} (\Delta A_i Y B_i + A_i Y \Delta B_i + \Delta A_i Y \Delta B_i) - \Delta C = R, \tag{3.3}$$

where $R = C - \sum_{i=1}^{p} A_i Y B_i$. Note that, to determine ΔA_i, ΔB_i and ΔC from (3.3) involves solution of a non-linear least square problem. This aspect poses a major difficulty in the analysis unlike Sylvester or Lyapunov equations, where the perturbations appear linearly. A similar situation also arises in the backward error analysis of a matrix-matrix multiplication, and this is the reason for the absence of a unique expression for its backward error. For various alternative definitions for the backward error of a matrix-matrix multiplication we refer to [1, p 77]. In this work we consider the following two definitions for the backward error.

$$\eta_A(Y) = \min \Big\{ \epsilon : \sum_{i=1}^{p} (A_i + \Delta A_i) Y B_i = C + \Delta C, \quad \|\Delta A_i\|_F \leq \epsilon \alpha_i, \\ \text{for } i \in \{1, 2, 3, \cdots, p\}, \text{ and } \|\Delta C\|_F \leq \epsilon \gamma \Big\}, \tag{3.4}$$

$$\eta_B(Y) = \min \Big\{ \epsilon : \sum_{i=1}^{p} A_i Y (B_i + \Delta B_i) = C + \Delta C, \quad \|\Delta B_i\|_F \leq \epsilon \beta_i, \\ \text{for } i \in \{1, 2, 3, \cdots, p\}, \text{ and } \|\Delta C\|_F \leq \epsilon \gamma \Big\}, \tag{3.5}$$

Perturbations appears linearly in (3.4) and (3.5), and therefore the analysis will simplify greatly. Now instead of deriving a bound on $\eta_A(Y)$ and $\eta_B(Y)$ separately, we analyse the linear part of (3.3), and at the end of the section demonstrate that bounds for (3.4) and (3.5) can be obtained as a special case. Before we begin with the analysis, we would like to emphasize that we are interested in estimating the error incurred by the floating-point computation, and they are modest compared to truncation error[3] in (2.3), (2.4), and discretization error. Therefore it is very unlikely that the second order terms in (3.3) will be important. However for the sake of rigour we use the definitions (3.4) and (3.5).

Accordingly (3.3) is linearized by neglecting the second order perturbation terms, and can be written as,

$$\begin{bmatrix} H_1 & H_2 & -\gamma I_{mn} \end{bmatrix} \begin{bmatrix} \text{Vec } (\Delta A') \\ \text{Vec } (\Delta B') \\ \text{Vec } (\Delta C)/\gamma \end{bmatrix} = \text{Vec}(R), \tag{3.6}$$

[3] Usually $\mathcal{O}(10^{-3})$

where

$$
\begin{aligned}
H_1 &= \left[\alpha_1(B_1^T Y^T \otimes I_n), \alpha_2(B_2^T Y^T \otimes I_n), \cdots, \alpha_p(B_p^T Y^T \otimes I_n)\right], \qquad (3.7)\\
H_2 &= \left[\beta_1(I_m \otimes A_1 Y), \beta_2(I_m \otimes A_2 Y), \cdots, \beta_p(I_m \otimes A_p Y)\right],\\
\text{Vec }(\Delta A') &= \left[\text{ Vec }(\Delta A_1)^T/\alpha_1, \text{ Vec }(\Delta A_2)^T/\alpha_2, \cdots, \text{ Vec }(\Delta A_p)^T/\alpha_p\right]^T,\\
\text{Vec }(\Delta B') &= \left[\text{ Vec }(\Delta B_1)^T/\beta_1, \text{ Vec }(\Delta B_2)^T/\beta_2, \cdots, \text{ Vec }(\Delta B_p)^T/\beta_p\right]^T,
\end{aligned}
$$

$I_m \in \mathbb{R}^{m\times m}$, $I_n \in \mathbb{R}^{n\times n}$ and $I_{mn} \in \mathbb{R}^{mn\times mn}$ are identity matrices. Compactly written, (3.6) is $Hz = r$, which is an underdetermined system and is of full rank if $\gamma \neq 0$. Therefore there exists a unique minimum two norm solution $z = H^\dagger r$, where $H^\dagger$ is the pseudo inverse. Using the relation between the 2-norm and the infinity norm it follows that

$$\frac{1}{\sqrt{2p+1}}\|H^\dagger r\|_2 \leq \|z\|_\infty \leq \|H^\dagger r\|_2.$$

Note that $\|z\|_\infty = \eta_A(Y)$ if $\Delta B_i = 0$, and $\|z\|_\infty = \eta_B(Y)$ if $\Delta A_i = 0$. Further if the second order terms are negligible in (3.3), then $\|z\|_\infty \approx \eta(Y)$. Now using the relation $\|H^\dagger r\|_2 \leq \|H^\dagger\|_2\|r\|_2$ we can deduce that

$$\|z\|_\infty \leq \|H^\dagger\|_2\|r\|_2 = \|r\|_2/\sigma_{\min}(H), \qquad (3.8)$$

where $\sigma_{\min}(H)$ is the minimum singular value of H. From (3.8) we can see that even if the norm of the residual is low, $\|z\|_\infty$ can be high, as it also depends on $\sigma_{\min}(H)$, which in turn depends on A_i, B_i, C, and Y. To quantify this dependency we now derive a lower bound on $\sigma_{\min}(H)$.

To achieve this consider the singular value decomposition (SVD) $Y = U\Sigma V^T$, where $U \in \mathbb{R}^{n\times n}$, $V \in \mathbb{R}^{m\times m}$ have orthonormal columns, and $\Sigma \in \mathbb{R}^{m\times n}$ is a diagonal matrix with entries the singular values σ_i. The singular values are assumed to have the ordering $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min(m,n)}$, and $\sigma_{\min(m,n)+1} = \cdots = \sigma_{\max(m,n)} = 0$. Now substitute the SVD in the expression of H in (3.6). Since the singular values are invariant under orthonormal transformations $\sigma_{\min}(H) = \sigma_{\min}(\tilde{H})$, where

$$
\begin{aligned}
&\tilde{H} = Q_1 H Q_2, \text{ where}\\
&Q_1 = V^T \otimes U^T,\\
&Q_2 = \text{diag}\left((U\otimes U)e_p^T, (V\otimes V)e_p^T, V\otimes U\right), \text{ and}
\end{aligned}
$$

$e_p^T = [1, 1, \cdots, 1] \in \mathbb{R}^p$. Further

$$
\begin{aligned}
\tilde{H} &= \left[\tilde{H}_1 \quad \tilde{H}_2 \quad -\gamma I_{mn}\right], \text{ where}\\
\tilde{H}_1 &= \left[\alpha_1(\tilde{B}_1^T\Sigma^T \otimes I_n), \alpha_2(\tilde{B}_2^T\Sigma^T \otimes I_n), \cdots, \alpha_p(\tilde{B}_p^T\Sigma^T \otimes I_n)\right],\\
\tilde{H}_2 &= \left[\beta_1(I_m \otimes \tilde{A}_1\Sigma), \beta_2(I_m \otimes \tilde{A}_2\Sigma), \cdots, \beta_p(I_m \otimes \tilde{A}_p\Sigma)\right],\\
\tilde{A}_i &= U^T A_i U, \quad i = 1, 2, 3, \cdots, p, \text{ and}\\
\tilde{B}_i &= V^T B_i V, \quad i = 1, 2, 3, \cdots, p.
\end{aligned}
$$

Now recall that the singular values of $\tilde{H}$ are the eigenvalues of

$$\tilde{H}\tilde{H}^T = \sum_{i=1}^{p} \left\{ \alpha_i^2 (\tilde{B}_i^T \Sigma^T \Sigma \tilde{B}_i \otimes I_n) + \beta_i^2 (I_m \otimes \tilde{A}_i \Sigma \Sigma^T \tilde{A}_i^T) \right\} + \gamma^2 I_{mn}.$$

Using the Courant-Fischer min-max theorem [13, Coro. 7.7.4], and a property of the eigenvalues of the Kronecker product, [4] we can show that

$$\lambda_{\min}(\tilde{H}\tilde{H}^T) \geq \sum_{j \in \mathcal{J}} \alpha_j^2 \lambda_{\min}(\tilde{B}_j^T \Sigma^T \Sigma \tilde{B}_j) + \sum_{k \in \mathcal{J}'} \beta_k^2 \lambda_{\min}(\tilde{A}_k \Sigma \Sigma^T \tilde{A}_k^T) + \gamma^2, \tag{3.9}$$

where $\lambda_{\min}(\cdot)$ denotes the minimum eigenvalue, $\mathcal{J}, \mathcal{J}' \subset (1, 2, 3, \cdots, p)$ are the indices for which the matrices B_i and A_i respectively are nonsingular. We can further simplify the summands of (3.9) as

$$\lambda_{\min}(\tilde{B}_i^T \Sigma^T \Sigma \tilde{B}_i) = \frac{1}{\|\tilde{B}_i^{-1} (\Sigma^T \Sigma)^\dagger \tilde{B}_i^{-T}\|_2} \geq \frac{\sigma_m(Y)^2}{\|B_i^{-1}\|_F^2}, \tag{3.10}$$

and

$$\lambda_{\min}(\tilde{A}_i \Sigma \Sigma^T \tilde{A}_i^T) = \frac{1}{\|\tilde{A}_i^{-T} (\Sigma \Sigma^T)^\dagger \tilde{A}_i^{-1}\|_2} \geq \frac{\sigma_n(Y)^2}{\|A_i^{-1}\|_F^2}. \tag{3.11}$$

To obtain the last inequality we have used $\|AB\|_2 \leq \|A\|_2 \|B\|_2$, the unitary invariance of the two norm, and $\|A\|_2 \leq \|A\|_F$. Substituting (3.10), (3.11) in (3.9) we can show that

$$\lambda_{\min}(\tilde{H}\tilde{H}^T) \geq \sum_{j \in \mathcal{J}} \alpha_j^2 \frac{\sigma_m(Y)^2}{\|B_j^{-1}\|_F^2} + \sum_{k \in \mathcal{J}'} \beta_k^2 \frac{\sigma_n(Y)^2}{\|A_k^{-1}\|_F^2} + \gamma^2,$$

and therefore

$$\sigma_{\min}(H) \geq \left(\sum_{j \in \mathcal{J}} \alpha_j^2 \frac{\sigma_m(Y)^2}{\|B_j^{-1}\|_F^2} + \sum_{k \in \mathcal{J}'} \beta_k^2 \frac{\sigma_n(Y)^2}{\|A_k^{-1}\|_F^2} + \gamma^2 \right)^{1/2}. \tag{3.12}$$

Finally using (3.12) in (3.8) we obtain

$$\|z\|_\infty \leq \mu \frac{\|R\|_F}{\tau}, \tag{3.13}$$

$$\mu = \frac{\tau}{\left(\sum_{j \in \mathcal{J}} \alpha_j^2 \frac{\sigma_m(Y)^2}{\|B_j^{-1}\|_F^2} + \sum_{k \in \mathcal{J}'} \beta_k^2 \frac{\sigma_n(Y)^2}{\|A_k^{-1}\|_F^2} + \gamma^2 \right)^{1/2}}, \tag{3.14}$$

$$\tau = \sum_{i=1}^{p} \left(\alpha_i \|B_i\|_F + \|A_i\|_F \beta_i \right) \|Y\|_F + \gamma. \tag{3.15}$$

[4] If $A \in \mathbb{R}^{n \times n}$,$B \in \mathbb{R}^{m \times m}$, then $\lambda_k(A \otimes B) \in \{\lambda_i(A)\lambda_j(B) : i \in (1, 2, \cdots, n),\ j \in (1, 2, \cdots, m)\}$ for $k \in (1, 2, \cdots, mn)$ [14, Theorem 4.2.15].

The scalar $\mu \geq 1$ is a magnification factor, therefore small residual does not necessarily imply small $\|z\|_\infty$.

Now we demonstrate that the analysis of this section generalizes the results of [7] and [8, p 1051]. Consider just the Sylvester equation $AX - XB = C$, then it can be recast into a generalized Sylvester equation in the following way,

$$(AXI_m) - (I_nXB) = C.$$

For the above equation (3.13) and (3.14) reduce to

$$\begin{aligned} \mu &= \frac{(\alpha_1 + \beta_2)\|Y\|_F + \gamma}{(\alpha_1^2 \sigma_m(Y)^2 + \beta_2^2 \sigma_n(Y)^2 + \gamma^2)^{1/2}}, \\ \eta(Y) &\leq \mu \frac{\|R\|_F}{(\alpha_1 + \beta_2)\|Y\|_F + \gamma}. \end{aligned} \tag{3.16}$$

(3.16) is same as (3.1), and the result of [8, p 1051]. Furthermore if $B = 0$ and $X \in \mathbb{R}^{n\times 1}$, Sylvester equation simplifies to a system of linear equations, and (3.16) further reduces to the backward error of a linear system of equation within a factor of $\sqrt{2}$.

To explore the conditions under which μ is very large, let us consider the case where $m = n$, A_i-s and B_i-s are non-singular. Further we consider the normwise relative perturbation, that is $\alpha_i = \|A_i\|_F$, $\beta_i = \|B_i\|_F$, and $\gamma = \|C\|_F$. Since all A_i and B_i are non-singular $\mathcal{J} = \mathcal{J}' = (1, 2, 3, \cdots, p)$, and (3.14) simplifies to

$$\begin{aligned} \mu &= \frac{2\left(\sum_{i=1}^p \|A_i\|_F \|B_i\|_F\right)\|Y\|_F + \|C\|_F}{\left(\sum_{i=1}^p \|A_i\|_F^2 \frac{\sigma_m(Y)^2}{\|B_i^{-1}\|_F^2} + \sum_{i=1}^p \|B_i\|_F^2 \frac{\sigma_n(Y)^2}{\|A_i^{-1}\|_F^2} + \|C\|_F^2\right)^{1/2}}, \\ &\geq \frac{2\left(\sum_{i=1}^p \|A_i\|_F \|B_i\|_F\right)\|Y\|_F + \|C\|_F}{\sigma_{\min}(Y)\left(\sum_{i=1}^p \frac{\|A_i\|_F\|B_i\|_F}{\kappa_F(B_i)} + \sum_{i=1}^p \frac{\|B_i\|_F\|A_i\|_F}{\kappa_F(A_i)}\right) + \|C\|_F}, \end{aligned} \tag{3.17}$$

where $\kappa_F(B_i) = \|B_i^{-1}\|_F\|B_i\|_F$ and $\kappa_F(A_i) = \|A_i^{-1}\|_F\|A_i\|_F$. From (3.17) we can conclude that $\mu \gg 1$ when either $\|Y\|_F \gg \sigma_{\min}(\mathrm{Y})$, and/or matrices A_i, B_i are ill conditioned. To further clarify the dependence on the condition numbers of Y, A_i, and B_i, we consider the case when $\gamma = 0$, that is the right hand side matrix is assumed to be known exactly. Under this assumption (3.17) simplifies to

$$\begin{aligned} \mu &\geq \frac{2\left(\sum_{i=1}^p \|A_i\|_F \|B_i\|_F\right)\|Y\|_F}{\sigma_{\min}(Y)\left(\sum_{i=1}^p \frac{\|A_i\|_F\|B_i\|_F}{\kappa_F(B_i)} + \sum_{i=1}^p \frac{\|B_i\|_F\|A_i\|_F}{\kappa_F(A_i)}\right)}, \\ &\geq \min_{i\in(1,2,\cdots,p)} [\kappa_F(A_i), \kappa_F(B_i)]\|Y\|_F\|Y^\dagger\|_2, \\ &\geq \kappa_2(Y) \min_{i\in(1,2,\cdots,p)} [\kappa_F(A_i), \kappa_F(B_i)], \end{aligned} \tag{3.18}$$

where $\kappa_2(Y) = \|Y^\dagger\|_2\|Y\|_2$. From the above inequality it becomes clear that, even if some A_i and B_i are ill conditioned the backward error can still be small, in Section 6 we will verify this numerically. Therefore the only remaining question is, under what conditions the computed solution is ill conditioned? This seems to be an open problem even for Sylvester equations, [1, Chapter 16].

Recall that $\|z\|_\infty = \eta_A(Y)$ if $\Delta B_i = 0$, therefore by considering $\beta_k = 0$ in (3.14) we obtain

$$
\begin{aligned}
\eta_A(Y) &\leq \mu_A \frac{\|R\|_F}{\tau}, \qquad (3.19)\\
\mu_A &= \frac{\tau}{\left(\sum_{j\in\mathcal{J}} \alpha_j^2 \frac{\sigma_m(Y)^2}{\|B_j^{-1}\|_F^2} + \gamma^2\right)^{1/2}},\\
\tau &= \sum_{i=1}^{p} \left(\alpha_i \|B_i\|_F + \|A_i\|_F \beta_i\right) \|Y\|_F + \gamma.
\end{aligned}
$$

In the above equation μ_A is the magnification factor, which is a ratio of the backward error $\eta_A(Y)$ and the relative residual $\|R\|_F/\tau$. Similarly a bound on $\eta_B(Y)$ can be derived by setting $\alpha_j = 0$ in (3.14). Therefore,

$$
\begin{aligned}
\eta_B(Y) &\leq \mu_B \frac{\|R\|_F}{\tau}, \qquad (3.20)\\
\mu_B &= \frac{\tau}{\left(\sum_{k\in\mathcal{J}'} \beta_k^2 \frac{\sigma_n(Y)^2}{\|A_k^{-1}\|_F^2} + \gamma^2\right)^{1/2}},\\
\tau &= \sum_{i=1}^{p} \left(\alpha_i \|B_i\|_F + \|A_i\|_F \beta_i\right) \|Y\|_F + \gamma.
\end{aligned}
$$

An interpretation similar to μ_A can also be given for μ_B as well.

4. Condition number

In this section we derive the condition number of a generalized Sylvester equation using the perturbation theory. To achieve this we first consider

$$\sum_{i=1}^{p} \Big((A_i + \Delta A_i)(X + \Delta X)(B_i + \Delta B_i) \Big) = C + \Delta C.$$

Unlike backward error analysis, only small perturbations are of interest here, and accordingly second order terms can be dropped. Hence we obtain

$$\sum_{i=1}^{p} A_i \Delta X B_i = \Delta C - \sum_{i=1}^{p} \left(\Delta A_i X B_i + A_i X \Delta B_i \right). \qquad (4.1)$$

Using the Kronecker product notation (4.1) can be restated as

$$P\ \mathrm{Vec}(\Delta X) = \gamma\,(\ \mathrm{Vec}(\Delta C)/\gamma) - \sum_{i=1}^{p} \left\{ \begin{bmatrix} \alpha_i(B_i^T X^T \otimes I_n) & \beta_i(I_m \otimes A_i X) \end{bmatrix} \begin{bmatrix} \mathrm{Vec}(\Delta A_i)/\alpha_i \\ \mathrm{Vec}(\Delta B_i)/\beta_i \end{bmatrix} \right\}, \tag{4.2}$$

where $P = \sum_{i=1}^{p} B_i^T \otimes A_i$, and α, β, γ are positive and a measure of the extent of perturbation. If we consider a normwise perturbation, and

$$\epsilon = \max\{\|\Delta A_i\|_F/\alpha_i, \|\Delta B_i\|_F/\beta_i, \|\Delta C\|_F/\gamma\},\ \text{ for } i \in (1, 2, 3, \cdots, p),$$

then

$$\frac{\|\Delta X\|_F}{\|X\|_F} \leq (2p+1)^{1/2}\Psi\epsilon, \tag{4.3}$$

where

$$\begin{aligned} \Psi &= \|P^{-1} \begin{bmatrix} H_1 & H_2 & -\gamma I_{mn} \end{bmatrix} \|_2/\|X\|_F, \qquad (4.4)\\ H_1 &= \left[\alpha_1(B_1^T X^T \otimes I_n), \alpha_2(B_2^T X^T \otimes I_n), \cdots, \alpha_p(B_p^T X^T \otimes I_n)\right], \\ H_2 &= \left[\beta_1(I_m \otimes A_1 X), \beta_2(I_m \otimes A_2 X), \cdots, \beta_p(I_m \otimes B_p X)\right]. \end{aligned}$$

$(2p+1)^{1/2}\Psi$ is the condition number of a generalized Sylvester equation, and the above bound is attainable. For the Sylvester equation, $p = 2$ and we obtain

$$5^{1/2}\|P^{-1} \begin{bmatrix} \alpha_1(X^T \otimes I_n) & \beta_1(I_m \otimes X) & -\gamma I_{mn} \end{bmatrix} \|_2/\|X\|_F,$$

which is the condition number derived in [7] to within a factor of $(3/5)^{1/2}$. Similar conclusion holds for the results of [8], which are derived for a two term generalized Sylvester equation. The bound in (4.3) can be weakened to

$$\frac{\|\Delta X\|_F}{\|X\|_F} \leq (2p+1)^{1/2}\Phi\epsilon, \quad \text{where}$$

$$\Phi = \|P^{-1}\|_2 \left(\sum_{i=1}^{p} \left[\alpha_i \|B_i\|_F + \beta_i\|A_i\|_F\right] \|X\|_F + \gamma \right) /\|X\|_F. \tag{4.5}$$

Note that Φ can be much greater than Ψ as demonstrated in [1] for the Sylvester equation.

For a linear system $Ax = b$, the condition number is given by

$$\kappa_{A,b}(A, x) = \frac{\|A^{-1}\|\|b\|}{\|x\|} + \|A^{-1}\|\|A\|, \tag{4.6}$$

where $\|\cdot\|$ can be any norm of choice, and we will consider the Frobenius norm for consistency with (4.3). The system of equations in stochastic Galerkin were initially identified as a standard linear system [15, Section 3.3.6], and therefore their condition number is estimated using (4.6). One major drawback of using (4.6) for estimating the condition number is that it does not take into account the actual structure of the equation. For the case of the generalized Sylvester equation, (4.6) can be written as

$$\kappa_{A,b}(A,x) = \frac{\| \left(\sum_{i=1}^{p}(B_i^T \otimes A_i)\right)^{-1} \|_F \|C\|_F}{\|X\|_F} + \left\| \left(\sum_{i=1}^{p}(B_i^T \otimes A_i)\right)^{-1} \right\|_F \left\| \sum_{i=1}^{p}(B_i^T \otimes A_i) \right\|_F . \tag{4.7}$$

Note that in deriving the above expression, a perturbation of $\sum_{i=1}^{p}(B_i^T \otimes A_i)$ is considered rather than individual matrices, we refer to [1, Sec 7.1] for further details. We will compare (4.7) and (4.5) numerically in Section 6.

5. Application to the stochastic Galerkin method

In the stochastic Galerkin method the matrices $\{A_i\}_{i=1}^{p}$ are symmetric and positive definite, B_1 is a diagonal matrix with positive entries and $\{B_i\}_{i=2}^{p}$ are sparse, symmetric, singular matrices, and for problems of practical interest $m \ll n$ [16, Section 4] . Therefore the expression for μ_A in (3.19) simplifies to

$$\mu_A^{(sg)} = \frac{2\left(\sum_{i=1}^{p} \|A_i\|_F \|B_i\|_F\right) \|Y\|_F + \|C\|_F}{\left(\frac{\|A_1\|_F^2}{\|B_1^{-1}\|_F^2}\sigma_m(Y)^2 + \|C\|_F^2\right)^{1/2}}. \tag{5.1}$$

From a computational point of view evaluating (5.1) is not expensive because B_1 is a diagonal matrix, and $m \ll n$, that is Y is a tall, skinny matrix, and therefore $\sigma_m(Y)$ can be computed very efficiently [17, Sec 5.4]. Similarly the expression for μ_B in (3.20) can also be simplified, which gives

$$\mu_B^{(sg)} = \frac{2\left(\sum_{i=1}^{p} \|A_i\|_F \|B_i\|_F\right) \|Y\|_F + \|C\|_F}{\|C\|_F}. \tag{5.2}$$

From (5.1) and (5.2) we can infer that the solution of the generalized Sylvester equation in the stochastic Galerkin method is backward stable in B_i, and conditionally backward stable in A_i.

Next we consider the estimation of the condition number Φ given by (4.5). As mentioned in section 1, iterative solvers are used for the solution of (1.1) in the context of the stochastic Galerkin method. Further, since (1.2) is a symmetric and positive definite matrix, preconditioned conjugate gradient (PCG) is the most popular choice. Also extremely good preconditioners are available as well [2], [18]. Now note that in (4.5), $\|P^{-1}\|_2 = \lambda_{\min}(P)$, that is the minimum eigenvalue of P, and this can be estimated using the Lancsoz method. Now exploiting the connection between the PCG and Lancsoz algorithm [6, Sec. 5.1], the condition number Φ can be computed as a bi-product of solving the generalized Sylvester equation. Further since the Lancsoz iteration converges to the extreme eigenvalues very quickly [6, Sec. 4.2.3], this method is computationally efficient.

Estimating the condition number Ψ would involve either inversion of P or solution of a linear system with multiple right hand sides, that is each column of H_1, H_2, and I_{mn}. Therefore even though Ψ is desirable, for problem of practical interest, estimating it would be extremely expensive.

6. Numerical Experiments

In this section we perform numerical experiments to achieve three objective

1. Verify the predictions made by the analysis in Sections 3 regarding the conditions in which μ of (3.13) has a large value.
2. Compare the actual backward error $\eta_A(Y)$ and the relative residual — to be defined later — for matrices from the stochastic Galerkin method.
3. Compare the condition number given by (4.4), (4.5) and (4.7) for matrices from the stochastic Galerkin method.

All the experiments are performed on a Mac laptop with Intel Core i5, and 8 Gb RAM, using MATLAB 2018b. We have made our codes available at `https://github.com/SrikaraPranesh/GeneralizedSylvester`.

To numerically verify the predictions made by the analysis of the Sections 3 regarding μ in (3.13), we consider a two term generalized Sylvester equation, that is

$$A_1XB_1 + A_2XB_2 = C. \tag{6.1}$$

We consider 1000 square matrices of size 4 × 4, and the matrices are generated using the `randsvd` command of MATLAB. To make the results reproducible we seed the random number generator using `rng(s)`, where $s = [1 : 1 : 1000]$. Right hand side C is generated using the `randn` command. (6.1) is solved by converting it to a linear system using the Kronecker product notation of (1.2), and using Gaussian elimination with partial pivoting, namely the '\' command of MATLAB. We will refer to '\' as the backslash command. Backward error for the linear system, which we refer to as the relative residual is computed using

$$\frac{\|\left[\sum_{i=1}^{2}(B_i^T \otimes A_i)\right]\widehat{x} - b\|_\infty}{\|\sum_{i=1}^{2} B_i^T \otimes A_i\|_\infty \|\widehat{x}\|_\infty - \|b\|_\infty}, \tag{6.2}$$

where $\widehat{x}$ is the computed solution, and the actual backward error $\eta(Y)$ is computed using (3.6) again by using the backslash command. Recall that μ in (3.13) was derived by considering only the first order perturbation terms, and therefore to be consistent with the analysis the test matrices were chosen so that the second order terms are negligible. We verify this by solving the non-linear least square problem (3.3) using `lsqnonlin` function of MATALB with default options. Throughout the numerical experiments we will consider the combinations of 2-norm condition numbers of A_i and B_i, which are listed in Table 6.1.

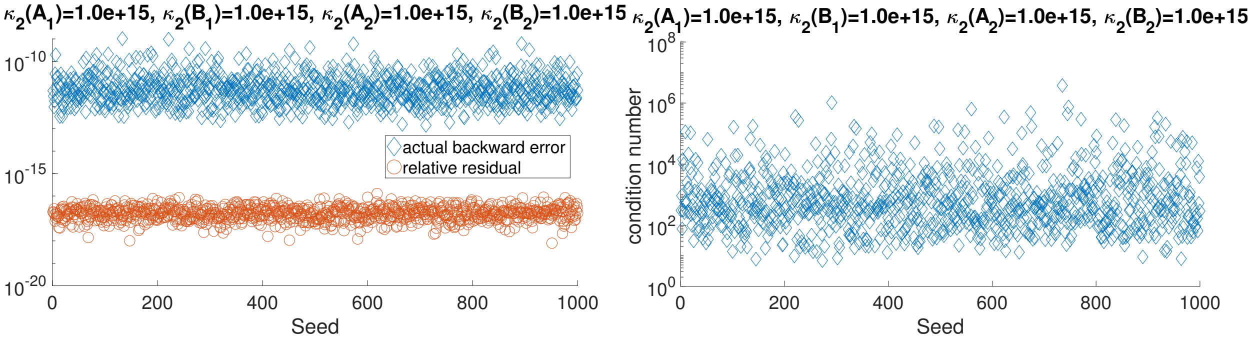
We choose the condition numbers of all the matrices to be $1e15$, that is case 1 in Table 6.1. We deliberately choose a high condition number to demonstrate the dependency of

Table 6.1: Combinations of the 2-norm condition numbers of the matrices A_1, A_2, B_1, and B_2 in (6.1), which are considered in the numerical experiments.

	$\kappa_2(A_1)$	$\kappa_2(B_1)$	$\kappa_2(A_2)$	$\kappa_2(B_2)$
case 1	1e15	1e15	1e15	1e15
case 2	1e15	1e15	1e2	1e15
case 3	1e15	1e15	1e2	1e2

the backward error on the condition number of the component matrices. The results are displayed in Figure 6.1a and we can observe that the actual backward error is up to 6 orders of magnitude higher than the relative residual which is $\mathcal{O}(u)$, where $u = 2.22e - 16$ is the unit round off of the double precision, therefore $\mu \approx 10^6$. From (3.18) we can see that the magnification factor depends on the condition number of the solution as well. In Figure 6.1b we display the condition number of the computed solution and it can be observed that the solution matrix is well conditioned compared to the input matrices. Therefore the high value of μ is because of the high condition number of the input matrices.

Next we set the condition numbers of A_1, B_1, B_2 to 1×10^{15}, and the condition number of A_2 to 1×10^2, that is case 2 in Table 6.1. In Figure 6.2a we display the actual backward error and the relative residual. From the figure we can observe that the actual backward error is up to 12 orders of magnitude higher than the relative residual which is $\mathcal{O}(u)$, that is $\mu \approx 10^{12}$. Next in Figure 6.2b we have displayed the condition number of the computed solution for all the test matrices, and we can observe that the condition number of the solution is also extremely high, thus leading to a high value of μ.

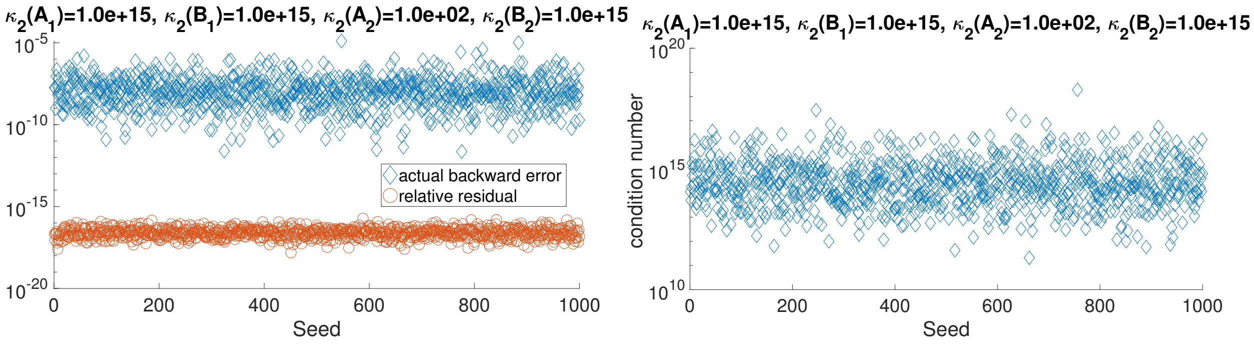


(a)

(b)

Figure 6.1: (a) Compares the actual backward error (3.2) and the relative residual (6.2). (b) Condition number of the solution matrix. Results are computed for the case 1 in Table 6.1.

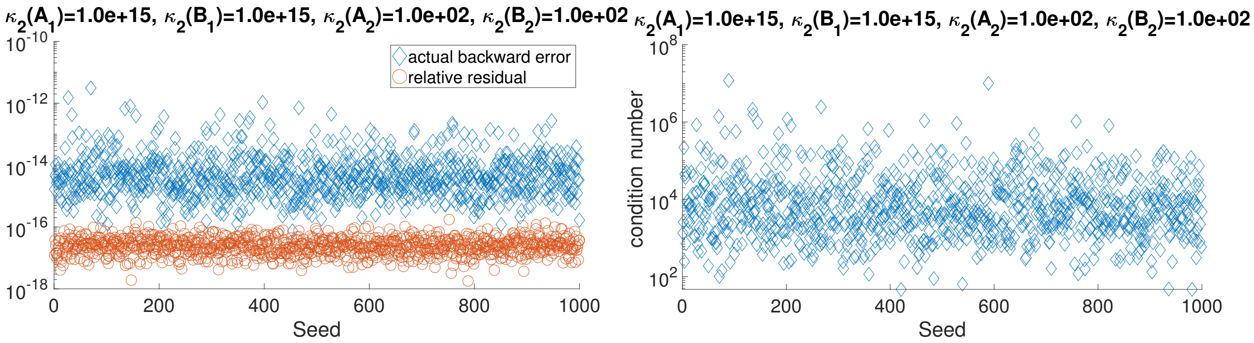
From these two experiments we can conclude that the magnification factor μ is extremely high when the input matrices are either ill-conditioned or the solution matrix is



(a) (b)

Figure 6.2: (a) Compares the actual backward error (3.2) and the relative residual (6.2). (b) Condition number of the solution matrix. Results are computed for the case 2 in Table 6.1.

ill-conditioned. However (3.18) predicts that the magnification is a function of the minimum over condition number of the input matrices. To verify this we set the condition number of A_1, B_1 to 1×10^{15}, and A_2, B_2 to 1×10^{2}, that is case 3 in Table 6.1. In Figure 6.3a we display the actual backward error and the relative residual, and we can observe that the magnification factor is relatively low, that is $\mu = \mathcal{O}(10^3)$. Further from Figure 6.3b we can see that the solution is relatively well conditioned as well.



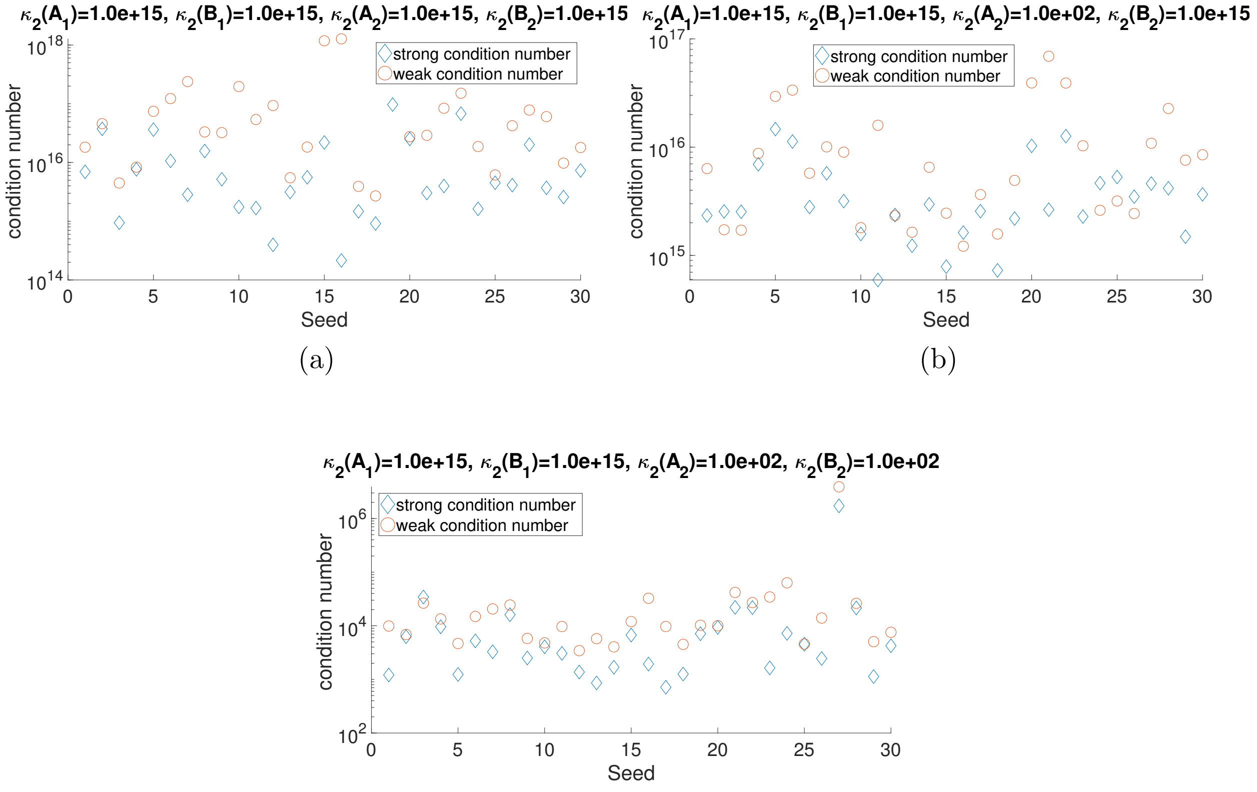
(a) (b)

Figure 6.3: (a) Compares the actual backward error (3.2) and the relative residual (6.2). (b) Condition number of the solution matrix. Results are computed for the case 3 in Table 6.1.

From these numerical experiments we can conclude that our analysis successfully captures the behaviour of the first order perturbation terms in the backward error. Further the solution of a generalized Sylvester equation is only conditionally backward stable, as the backward error depends on the condition number of the input matrices and the solution matrix.

Next we perform numerical experiments to compare the condition number obtained using

(4.4) and (4.5). The action of P^{-1} is obtained by using the backslash command of MATLAB. We consider the same three combination of condition numbers listed in Table 6.1, and the corresponding actual condition numbers of the generalized Sylvester equation are displayed in Figure 6.4. For the sake of clarity we display the result for only first 30 matrices, the trend is broadly similar for the remaining matrices as well. In the plot *strong condition number* refers to (4.4), and *weak condition number* refers to (4.5). From the plots we can observe that the condition number of the generalized Sylvester equation computed using (4.4) is always lesser than (4.5), and in some cases it is up to three orders of magnitude lower.



(a)

(b)

(c)

Figure 6.4: Condition number of the generalized Sylvester equation, *strong condition number* is computed using (4.4), and *weak condition number* is computed using (4.5), condition number of the component matrices are listed in Table 6.1 (a) case 1, (b) case 2, and (c) case 3. Condition numbers of only the first thirty matrices are displayed.

Now for the matrices obtained from the stochastic Galerkin discretisation of an elliptic sPDE we perform numerical experiments to compare the actual backward error and the relative residual. The actual backward error is given by

$$\eta_A(Y) = \|H_A^{\dagger} r\|_{\infty}, \tag{6.3}$$

where $H_A = \begin{bmatrix} H_1 & -\|C\|_F I_{mn} \end{bmatrix}$, H_1 is given by (3.7), and r is the residual; bound for the backward error given by

$$\eta(Y) \le \mu_A^{(sg)} \frac{\|R\|_F}{2\left(\sum_{i=1}^p \|A_i\|_F \|B_i\|_F\right) \|Y\|_F + \|C\|_F}, \tag{6.4}$$

where $\mu_A^{(sg)}$ is given by (5.1), and the relative residual is computed as

$$\frac{\| \operatorname{Vec}(C) - \left(\sum_{i=1}^p B_i^T \otimes A_i\right) \operatorname{Vec}(X)\|_\infty}{\| \sum_{i=1}^p B_i^T \otimes A_i\|_\infty \| \operatorname{Vec}(X)\|_\infty + \| \operatorname{Vec}(C)\|_\infty}. \tag{6.5}$$

Recall from (5.2) that for the stochastic Galerkin method, small residual implies small $\eta_B(Y)$, therefore we do not consider it here. Further the condition numbers estimated using (4.4), (4.5), and (4.7) are also compared.

We solve (2.1) on a square domain between $[0, 0.5] \times [0, 0.5]$. $f(\boldsymbol{x}, \theta)$ is assumed to be deterministic and a constant function of unit magnitude. A Homogeneous boundary condition is applied at $x = 0$ and $x = 0.5$. For the spatial discretisation a linear triangles are used, and a finite element space of dimensions 31 and 127 are considered. The random field $\kappa(\boldsymbol{x}, \theta)$ is assumed to have a mean of 200, variance of 1000, and a Gaussian covariance model with a correlation length of 2.5 is adopted. The algorithm proposed in [10] is used to discretise the covariance function. Two cases for the random variables ξ_i in (2.3) are considered (i) Standard normal, then ψ_i in (2.5) are chosen to be Hermite polynomials, (ii) uniform between $[-\sqrt{3}, \sqrt{3}]$, then ψ_i in (2.5) are chosen to be Legendre polynomials. The matrices B_i in (2.7) are computed using the stochastic Galerkin toolbox in `https://github.com/ezander/sglib`. Since the coefficient of variation of the input random field is small, the positive definiteness can be guaranteed even when standard normal random variables are used in (2.3). $M = B_1^T \otimes A_1$ is used as the preconditioner, this is commonly known as the mean based preconditioner in the uncertainty quantification community. The preconditioned conjugate Gradient is terminated when the 2-norm of the residual — numerator of (6.5) — is less than 1×10^{-6}. For estimating the actual backward error $\eta_A(Y)$, $H_A^\dagger$ is computed using the `pinv` command in (6.3). Further the action of P^{-1} in (4.4) is achieved using the backslash command, and $\|P^{-1}\|_2$ in (4.5) is computed using the `eig` command.

In Table 6.2 for two sizes of matrices A_i we display the actual backward error $\eta_A(Y)$, its bound given by (6.4), and the relative residual computed using (6.5) for varying size of matrices B_i. In (2.6) we can observe that the size of matrices B_i depends on two quantities, namely p_1 and q, these values are displayed in parenthesis in the second column of Table 6.2. From Table 6.2 we can observe that the actual backward error is up to 2 orders of magnitude higher than the relative residual. Further more the upper bound on the backward error is tight. Therefore rather than using a norm of the residual as the stopping criterion $\|R\|_F \mu_A^{(sg)}$ should be used as the stopping criterion.

Next in Table 6.3 the estimates of Ψ, Φ, and $\kappa_{A,b}(A, x)$ of (4.4), (4.5), and (4.7) respectively are displayed. The size of matrices A_i, and B_i are same as that of Table 6.2. From Table 6.3 we can observe that Ψ is up to two order of magnitude lower than Φ and $\kappa_{A,b}(A, x)$. Furthermore Φ and $\kappa_{A,b}(A, x)$ are broadly similar, however the perturbations considered in deriving Φ is in accordance with the structure of the equation.

Table 6.2: Comparison of the actual backward error $\eta_A(Y)$ (6.3), its bound (6.4), and the relative residual (6.5). Matrices are obtained from the stochastic Galerkin method. The size of A_i is n, m is the size of B_i, and the numbers in the parenthesis are p_1 and q in (2.6). 'Hermite' and 'Legendre' indicate the type of gPC polynomials used in (2.5).

n	m		Hermite			Legendre		
			$\eta_A(Y)$	$\eta_A(Y)$ bound	relative residual	$\eta_A(Y)$	$\eta_A(Y)$ bound	relative residual
31	6	(2,2)	1.79e-07	5.96e-07	1.20e-08	2.68e-08	9.79e-08	3.96e-09
	10	(3,2)	1.73e-07	5.88e-07	1.17e-08	2.52e-08	1.20e-07	3.71e-09
	15	(4,2)	1.73e-07	5.88e-07	1.17e-08	2.53e-08	1.21e-07	3.73e-09
	10	(2,3)	7.04e-08	2.22e-07	1.60e-09	1.80e-07	6.15e-07	2.65e-08
	20	(3,3)	6.81e-08	2.19e-07	1.54e-09	1.74e-07	6.24e-07	2.57e-08
127	6	(2,2)	2.94e-07	9.84e-07	1.05e-08	1.36e-08	1.02e-07	1.08e-09
	10	(3,2)	2.88e-07	9.74e-07	1.03e-08	1.27e-08	1.26e-07	1.14e-09
	15	(4,2)	2.88e-07	9.74e-07	1.03e-08	1.27e-08	1.27e-07	1.15e-09
	10	(2,3)	1.07e-07	3.52e-07	1.27e-09	5.52e-08	7.22e-07	9.78e-09
	20	(3,3)	1.05e-07	3.48e-07	1.25e-09	5.47e-08	7.28e-07	9.58e-09

Table 6.3: Comparison of the actual condition number Ψ (4.4), its upper bound Φ (4.5), and the condition number of the corresponding linear system computed by (4.7). Matrices are obtained from the stochastic Galerkin method. The size of A_i is n, m is the size of B_i, and the numbers in the parenthesis are p_1 and q in (2.6). 'Hermite' and 'Legendre' indicate the type of gPC polynomials used in (2.5).

n	m		Hermite			Legendre		
			Ψ	Φ	$\kappa_{A,b}(A,x)$	Ψ	Φ	$\kappa_{A,b}(A,x)$
31	6	(2,2)	8.10e+01	7.30e+02	8.76e+02	7.77e+01	1.68e+03	1.61e+03
	10	(3,2)	9.18e+01	1.02e+03	1.43e+03	8.81e+01	2.02e+03	2.59e+03
	15	(4,2)	1.01e+02	1.33e+03	2.09e+03	9.74e+01	2.37e+03	3.68e+03
	10	(2,3)	8.29e+01	2.04e+03	2.62e+03	7.77e+01	2.86e+03	3.54e+03
	20	(3,3)	9.41e+01	2.93e+03	4.74e+03	1.21e+02	6.26e+03	7.29e+03
127	6	(2,2)	3.15e+02	2.84e+03	6.58e+03	3.02e+02	6.47e+03	1.20e+04
	10	(3,2)	3.57e+02	3.99e+03	1.07e+04	3.43e+02	7.82e+03	1.94e+04
	15	(4,2)	3.95e+02	5.20e+03	1.58e+04	3.79e+02	9.16e+03	2.75e+04
	10	(2,3)	3.16e+02	7.97e+03	1.97e+04	3.02e+02	1.11e+04	2.64e+04
	20	(3,3)	3.58e+02	1.14e+04	3.57e+04	3.43e+02	2.42e+04	5.45e+04

7. Conclusion and future direction

In this work we derived an upper bound on the backward error of the generalized Sylvester equation, and demonstrated that a small residual need not imply a small backward error. We proceed by introducing two definitions for the backward error, where we consider the perturbations in A_i and B_i in (1.1) separately, and derived an upperbound on each of them. Predictions of the analysis are verified using numerical experiments. Further for the stochastic Galerkin method a computationally efficient method to estimate a bound on the actual backward error is discussed. Using numerical experiments we demonstrate that the actual backward error can be up to two orders of magnitude higher than the relative residual. Therefore accounting for the magnification factor it would be advisable to use $\|r\|_2\mu_A^{(sg)}$ as the stopping criterion, rather than just $\|r\|_2$ [5]. We also derived an expression for the condition number of the generalized Sylvester equation by considering the structure of the equation. Since efficient preconditioners are available for the stochastic Galerkin method, using Krylov subspace methods to estimate the actual condition number can be computationally efficient.

This work highlights the fact that the generalized Sylvester equation should be analysed in its own right as matrix equation, rather than as a usual linear system. As indicated in the introduction, only recently the generalized Sylvester equation is being considered by the numerical linear algebra community, therefore many open problems still remain.

1. Conditions for the existence and uniqueness of the solution of a generalized Sylvester equation, for a general A_i, B_i, and C.
2. Backward error for the generalized Sylvester equation, by using the definition (3.2).

We will consider these questions in our future research.

Acknowledgement

Author would like to thank Prof. Nick Higham for insightful comments on multiple drafts of this manuscript. Also the help of Gopika Ajith with stochastic Galerkin experiments is acknowledged.

References

[1] N. J. Higham, Accuracy and Stability of Numerical Algorithms, 2nd Edition, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2002. doi:10.1137/1.9780898718027.

[2] C. E. Powell, H. C. Elman, Block-diagonal preconditioning for spectral stochastic finite-element systems, IMA J. Numer. Anal. 29 (2) (2009) 350–375. doi:10.1093/imanum/drn014.

[3] C. E. Powell, D. Silvester, V. Simoncini, An efficient reduced basis solver for stochastic Galerkin matrix equations, SIAM J. Sci. Comput. 39 (1) (2017) A141–A163. doi:10.1137/15M1032399.

[4] S. Pranesh, D. Ghosh, Cost reduction of stochastic Galerkin method by adaptive identification of significant polynomial chaos bases for elliptic equations, Computer Methods in Applied Mechanics and Engineering 340 (2018) 54–69. doi:10.1016/j.cma.2018.04.043.

[5] r is the residual

[5] A. Bouhamidi, K. Jbilou, A note on the numerical approximate solutions for generalized Sylvester matrix equations with applications, Applied Mathematics and Computation 206 (2) (2008) 687–694. doi:10.1016/j.amc.2008.09.022.

[6] R. Barrett, M. W. Berry, T. F. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, H. Van der Vorst, Templates for the solution of linear systems: building blocks for iterative methods, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1994. doi:10.1137/1.9781611971538.

[7] N. J. Higham, Perturbation theory and backward error for $AX - XB = C$, BIT 33 (1993) 124–136. doi:10.1007/978-94-015-8196-7_39.

[8] B. Kågström, A Perturbation Analysis of the Generalized Sylvester Equation $(AR - LB, DR - LE) = (C, F)$, SIAM J. Matrix Anal. Appl. 15 (4) (1994) 1045–1060. doi:10.1137/S0895479893246212.

[9] M. Konstantinov, D. W. Gu, V. Mehrmann, P. Petkov, Perturbation theory for matrix equations, Vol. 9, Gulf Professional Publishing, 2003.
URL https://www.elsevier.com/books/perturbation-theory-for-matrix-equations/konstantinov/978-0-444-51315-1

[10] S. Pranesh, D. Ghosh, Faster computation of the Karhunen–Loève expansion using its domain independence property, Computer Methods in Applied Mechanics and Engineering 285 (2015) 125–145. doi:10.1016/j.cma.2014.10.053.

[11] D. Xiu, G. E. Karniadakis, The Wiener–Askey polynomial chaos for stochastic differential equations, SIAM J. Sci. Comput. 24 (2) (2002) 619–644. doi:10.1137/S1064827501387826.

[12] S. Pranesh, Development of an efficient domain decomposition algorithm for solving large stochastic mechanics problems, Ph.D. thesis, Indian Institute of Science, Bangalore, India 560012 (2018).

[13] R. A. Horn, C. R. Johnson, Matrix analysis, Cambridge University, 2012.
URL https://www.cambridge.org/core/books/matrix-analysis/FDA3627DC2B9F5C3DF2FD8C3CC136B48

[14] R. A. Horn, C. R. Johnson, Topics in matrix analysis, Cambridge University Press, 2011.
URL https://www.cambridge.org/core/books/topics-in-matrix-analysis/B988495A235F1C3406EA484A2C477B03

[15] R. G. Ghanem, P. D. Spanos, Stochastic finite elements: A spectral approach, Dover publications, 2003.
URL http://store.doverpublications.com/0486428184.html

[16] S. Pranesh, D. Ghosh, A FETI-DP based parallel hybrid stochastic finite element method for large stochastic systems, Computers & Structures 195 (2018) 64–73. doi:10.1016/j.compstruc.2017.09.011.

[17] J. Dongarra, M. Gates, A. Haidar, J. Kurzak, P. Luszczek, S. Tomov, I. Yamazaki, The Singular Value Decomposition: Anatomy of Optimizing an Algorithm for Extreme Scale, SIAM Rev. 60 (4) (2018) 808–865. doi:10.1137/17M1117732.

[18] E. Ullmann, A Kronecker product preconditioner for stochastic Galerkin finite element discretizations, SIAM J. Sci. Comput. 32 (2) (2010) 923–946. doi:10.1137/080742853.