Note on coefficient matrices from stochastic Galerkin methods for random diffusion equations

Tao Zhou\textsuperscript{a}, Tao Tang\textsuperscript{b,}\textsuperscript{*}

\textsuperscript{a}Institute of Computational Mathematics, The Chinese Academy of Sciences, Beijing 100190, PR China
\textsuperscript{b}Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong

\textbf{Article info}

Article history:
Received 18 January 2010
Received in revised form 6 July 2010
Accepted 13 July 2010
Available online 29 July 2010

Keywords:
Galerkin methods
Stochastic diffusion equations
Matrix properties

\textbf{Abstract}

In a recent work by Xiu and Shen [D. Xiu, J. Shen, Efficient stochastic Galerkin methods for random diffusion equations, J. Comput. Phys. 228 (2009) 266–281], the Galerkin methods are used to solve stochastic diffusion equations in random media, where some properties for the coefficient matrix of the resulting system are provided. They also posed an open question on the properties of the coefficient matrix. In this work, we will provide some results related to the open question.

In [5], Xiu and Shen consider simulations of diffusion problems with uncertainties, which yield the following stochastic diffusion equation:

\begin{align}
\frac{\partial u(x, y, t)}{\partial t} &= \nabla \cdot (\kappa(x, y) \nabla u(x, y, t)) + f(x, y, t), \quad \forall x \in \Omega, \ t \in (0, T], \\
u(x, y, 0) &= u_0(x, y), \quad u(x, y, t)|_{\partial \Omega} = 0,
\end{align}

where $x = (x_1, \ldots, x_d)^T \in \Omega \subseteq \mathbb{R}^d$, $d = 1, 2, 3$ are the spatial coordinates, and $y = (y_1, \ldots, y_N) \in \mathbb{R}^N$, $N \geq 1$, is a random vector with independent and identically distributed components.

The steady state counterpart of Eqs. (1) and (2) is

\begin{align}
-\nabla \cdot (\kappa(x, y) \nabla u(x, y, t)) &= f(x, y, t), \quad u(x, y, t)|_{\partial \Omega} = 0, \quad \forall x \in \Omega.
\end{align}

It is assumed that the random diffusion field takes a simple form

\begin{align}
\kappa(x, y) &= \kappa_0(x) + \sum_{i=1}^{N} \kappa_i(x) y_i,
\end{align}

where $\{\kappa_i(x)\}_{i=0}^{N}$ are fixed functions with $\kappa_0(x) > 0$, $\forall x$. For well-posedness, the following assumption is made

\begin{align}
\kappa(x, y) &\geq \kappa_{\min} > 0, \quad \forall x, y.
\end{align}

The Pth-order, gPC approximations of $u(x, y, t)$ and $f(x, y, t)$ are of the form

\begin{align}
&
\text{\textsuperscript{*} Corresponding author.} \\
&\text{E-mail addresses: tzhou@lsec.cc.ac.cn (T. Zhou), ttang@hkbu.edu.hk (T. Tang).}
\end{align}
\[ u(x, y, t) \approx \sum_{m=0}^{M} u_m(x, t) \Phi_m(y), \quad f(x, y, t) \approx \sum_{m=0}^{M} f_m(x, t) \Phi_m(y), \] (6)

where \( M = \binom{N+P}{N} \). \( \Phi_m(y) \) are \( N \)-variate orthonormal polynomials of degree up to \( P \). They are constructed as product of a sequence of univariate polynomials in each directions of \( y_i, i = 1, \ldots, N \), i.e.,

\[ \Phi_m(y) = \phi_{m_1}(y_1) \cdots \phi_{m_N}(y_N), \quad m_1 + \cdots + m_N \leq P, \]

where \( m_i \) is the order of the univariate polynomials of \( \phi(y_i) \) in the \( y_i \) direction for \( 1 \leq i \leq N \). These univariate polynomials are orthonormal (upon proper scaling), i.e.,

\[ \int \phi_i(y_i) \phi_k(y_i) \rho_i(y_i) dy_i = \delta_{ik}, \quad 1 \leq i, k \leq N, \]

where \( \rho_i(y_i) \) is the probability distribution function for random variable \( y_i \). The type of the polynomials \( \phi(y_i) \) is determined by the distribution of \( y_i \). For example, Hermite polynomials are better suited for Gaussian distribution, Jacobi polynomials are better for beta distribution. For detailed discussion, see [6].

For the \( N \)-variate basis polynomials \( \Phi_m(y) \), each index \( 1 \leq m \leq M \) corresponds to a unique sequence \( m_1, \ldots, m_N \), and they are also orthonormal

\[ \mathbb{E}[\Phi_m(y) \Phi_n(y)] = \int \Phi_m(y) \Phi_n(y) \rho(y) dy = \delta_{mn}, \]

where \( \rho(y) = \Pi_{i=1}^{N} \rho_i(y_i) \).

Upon substituting Eq. (6) into the governing Eq. (1) and projecting the resulting equation onto the subspace spanned by the first \( M \) gPC basis polynomials, we obtain for all \( k = 1, \ldots, M \),

\[ \frac{\partial v_k}{\partial t} = \sum_{j=1}^{M} \nabla \cdot (a_{jk}(x) \nabla v_k) + f_k(x, t), \] (7)

where

\[ a_{jk} = \sum_{i=0}^{N} \kappa_i(x) e_{jk}, \quad 1 \leq k, j \leq M, \]

\[ e_{jk} = \int y_i \phi_i(y) \phi_k(y) \rho(y) dy, \quad 1 \leq k, j \leq M, \quad 0 \leq i \leq N. \]

Denote \( \mathbf{v} = (v_1, \ldots, v_M)^T \) and \( \mathbf{A}(x) = (a_{jk})_{1 \leq j, k \leq M} \). By definition, \( \mathbf{A} = \mathbf{A}^T \) is symmetric. The gPC Galerkin equations (7) can be written as:

\[ \frac{\partial \mathbf{v}}{\partial t}(x, t) = \nabla \cdot (\mathbf{A} \nabla \mathbf{v}) + \mathbf{f}, \]

\[ \mathbf{v}(x, 0) = \mathbf{v}_0(x), \quad \mathbf{v}(\cdot, t)|_{\partial \Omega} = 0. \]

This is a coupled system of diffusion equations, where \( \mathbf{v}_0(x) \) is the gPC expansion coefficient vector obtained by expressing the initial condition of (2) in the form of Eq. (6).

Similarly, the following steady state decomposition can be obtained:

\[ -\nabla \cdot (\mathbf{A} \nabla \mathbf{v}) = \mathbf{f}, \quad \mathbf{v}(x, 0) = \mathbf{v}_0(x). \]

(10)

After some simple algebra, it can be verified that the components of \( \mathbf{A}(x) \) satisfy

\[ a_{ij} = \kappa_0(x) + \sum_{i=1}^{N} \kappa_i(x) b_i, \]

(11)

\[ \sum_{i=0}^{N} |a_{ij}| = \sum_{i=1}^{N} |\kappa_i(x)||a_j + c_j|, \]

(12)

where \( b_j, a_j, c_j \) are the coefficients of the recurrence relation corresponding normalized polynomials in \( y_i \) direction, namely,

\[ x^N \int_j P^{\beta \gamma}(x) = a_j P^{\beta \gamma}_{j+1}(x) + b_j P^{\beta \gamma}_j(x) + c_j P^{\beta \gamma}_{j+1}(x). \]

(13)

Since for every \( i, y_i \) are assumed to be have the same distribution, we then have the same recurrence relation formula for every direction, thus for donation simplicity, we can drop the subscribe \( i \) and just use \( a_j, b_j, c_j \) in the recurrence relation. We have the following lemmas [5]:
Lemma 1. The matrix $A(x)$ is positive definite for any $x \in D$, moreover, each row of $A(x)$ has at most $(2N + 1)$ non-zero entries.

Lemma 2. Assume that the random variables $y_i, 1 \leq i \leq N$, have either an identical Beta distribution in $(-1, 1)$ with $\rho(y_i) = (1 - y_i)(1 + y_i)^\alpha$, or an identical Gamma distribution in $(0, +\infty)$ with $\rho(y_i) = y_i^\alpha e^{-\gamma}$, where $\alpha > -1$ and the scaling constants are omitted. Then, the matrices $A(x)$ derived via the corresponding gPC basis are strictly diagonal dominant $\forall x \in \Omega$. More precisely, we have

$$
a_{ij} \geq K_{\min} + \sum_{k=1}^{M} |a_{ik}|, \quad 1 \leq j \leq M, \quad \forall x \in \Omega.
$$

Open questions in [5]: it is stated in [5] that it is an open question whether Lemma 2 holds for general Beta distributions in $(-1, 1)$ with $\rho(y_i) \sim (1 - y_i)(1 + y_i)^\alpha$, $\alpha \neq \beta, \gamma > -1$.

The purpose of this paper is to provide an answer to the above open question. More precisely, we will prove the following result:

Theorem 1. (I) Assume that the random variables $y_i, 1 \leq i \leq N$, have an identical Beta distribution in $(-1, 1)$ with $\rho(y_i) = (1 - y_i)(1 + y_i)^\alpha$, satisfy that

$$|\alpha| \geq \frac{1}{2}, \quad ||\beta| | \geq \frac{1}{2}, \quad \text{(of course, } \alpha > -1)$$

or (II) the random variables $y_i, 1 \leq i \leq N$, have an identical Gamma distribution in $(0, +\infty)$ with $\rho(y_i) = y_i^\alpha e^{-\gamma}$, where $\alpha > -1$ and the scaling constants are omitted.

Then, the matrices $A(x)$ derived via the corresponding gPC basis are strictly diagonal dominant $\forall x \in \Omega, i.e.,$

$$a_{ij} \geq K_{\min} + \sum_{k=1}^{M} |a_{ik}|, \quad 1 \leq j \leq M, \quad \forall x \in \Omega.$$

Proof. We recall the classical Jacobi-polynomials (see, e.g., [1,3,4]) $\{P_n^{\alpha, \beta}\}_{n=0}^\infty$, which the terminal value follows

$$P_n^{\alpha, \beta}(1) = \binom{n + \alpha}{n},$$

where for integer $n$,

$$\binom{z}{n} = \frac{\Gamma(z + 1)}{\Gamma(n + 1)\Gamma(n - z + 1)},$$

and $\Gamma(z)$ is the usual Gamma function. For notation simplicity, we will denote $\omega = 2n + \alpha + \beta$ in the following. The Jacobi-polynomials satisfy the orthogonality condition

$$\int_{-1}^{1} (1 - x)^\alpha (1 + x)^\beta P_m^{\alpha, \beta}(x) P_n^{\alpha, \beta}(x) dx = \frac{2^{\alpha + \beta + 1}}{2n + \alpha + \beta + 1} \frac{\Gamma(n + \alpha + 1)\Gamma(n + \beta + 1)}{\Gamma(n + \alpha + \beta + 1)n!} \delta_{mn}$$

and recurrence relation

$$x P_n^{\alpha, \beta}(x) = a_n P_{n+1}^{\alpha, \beta}(x) + b_n P_n^{\alpha, \beta}(x) + c_n P_{n-1}^{\alpha, \beta}(x)$$

with

$$a_n = \frac{2(n + 1)(n + \alpha + \beta + 1)}{(\omega)(\omega + 2)}, \quad b_n = \frac{\beta^2 - \alpha^2}{\omega(\omega + 2)}, \quad c_n = \frac{2(n + \alpha)(n + \beta)}{\omega(\omega + 1)}.$$

To work with the $L^2(\omega)$-normalized polynomials, we define

$$\tilde{P}_n^{\alpha, \beta} = P_n^{\alpha, \beta}/\sqrt{h_n^{\alpha, \beta}},$$

where

$$h_n^{\alpha, \beta} = \frac{2^{\alpha + \beta + 1}}{2n + \alpha + \beta + 1} \frac{\Gamma(n + \alpha + 1)\Gamma(n + \beta + 1)}{\Gamma(n + \alpha + \beta + 1)n!},$$

which are orthonormal.

By Eqs. (18)–(22), after some algebra, we can drive the recurrence relation for the normalized Jacobi-polynomials

$$x \tilde{P}_n^{\alpha, \beta}(x) = \tilde{a}_n \tilde{P}_{n+1}^{\alpha, \beta}(x) + \tilde{b}_n \tilde{P}_n^{\alpha, \beta}(x) + \tilde{c}_n \tilde{P}_{n-1}^{\alpha, \beta}(x)$$

(23)
with
\[
\tilde{a}_n = 2 \sqrt{\frac{(n+1)(n+\alpha+1)(n+\beta+1)(n+\alpha+\beta+1)}{(\omega+1)(\omega+2)^2(\omega+3)}},
\]
\[
\tilde{b}_n = \frac{\beta^2 - \alpha^2}{\omega(\omega+2)}, \quad \tilde{c}_n = 2 \sqrt{\frac{n(n+2)(n+\beta)(n+\alpha+\beta)}{(\omega-1)(\omega)^2(\omega+1)}}.
\]

It is easy to see that
\[
\tilde{a}_n = \tilde{c}_{n+1}, \quad \tilde{a}_n, \tilde{c}_n \geq 0, \quad 0 \leq \tilde{b}_n < 1, \quad \forall \alpha, \beta > -1, \ n \geq 1.
\]

(1i): Case \( \alpha = \beta \). For this special case, we have
\[
\tilde{a}_n = \sqrt{\frac{(n+1)(n+2\alpha+1)}{(\omega+1)(\omega+3)}}, \quad \tilde{b}_n = 0, \quad \tilde{c}_n = \tilde{a}_{n-1}.
\]

and
\[
\tilde{a}_n = \sqrt{\frac{(n+1)(\omega+1)}{(n+2\alpha+1)(\omega+3)}} = \sqrt{\frac{1}{4} + \frac{1-4\alpha^2}{4(\omega+1)(\omega+3)}}.
\]

We can see that \( \tilde{a}_n \leq 1/2 \), when \( 1 - 4\alpha^2 \leq 0 \), for all \( n \), which yields
\[
0 < \tilde{a}_n + \tilde{c}_n = \tilde{a}_n + \tilde{a}_{n-1} < 1, \quad |\alpha| > \frac{1}{2}, \quad \forall n > 1.
\]

Then, substituting \( y_i = \pm (\tilde{a}_j + \tilde{c}_j) \) into Eq. (5) separately and using together Eqs. (11) and (12) yields the inequality (15).

(1ii): Case \( \alpha \neq \beta \). We first assume that \( \beta^2 > \alpha^2 \). Note
\[
\tilde{a}_n = 2 \sqrt{\frac{(n+1)(n+\alpha+1)(n+\beta+1)(n+\alpha+\beta+1)}{(\omega+1)(\omega+2)^2(\omega+3)}} = 2 \sqrt{\frac{1}{4} + \frac{1-(\alpha+\beta)^2}{4(\omega+1)(\omega+3)}} \sqrt{\frac{1}{4} + \frac{1-(\alpha+\beta)^2}{4(\omega+2)^2}}.
\]

We know form the above equations that \( \tilde{a}_n < 1/2 \) if \( 1 - (\alpha+\beta)^2 \leq 0 \). Therefore,
\[
-1 < \tilde{a}_n + \tilde{a}_{n-1} + \tilde{b}_n < 1, \quad \forall n > 1.
\]

Now we check for \( \tilde{a}_n + \tilde{a}_{n-1} + \tilde{b}_n \). We have
\[
\tilde{a}_n + \tilde{a}_{n-1} = 2 \sqrt{\frac{1}{4} + \frac{1-(\alpha+\beta)^2}{4(\omega+1)(\omega+3)}} \sqrt{\frac{1}{4} + \frac{1-(\alpha-\beta)^2}{4(\omega+2)^2}} + 2 \sqrt{\frac{1}{4} + \frac{1-(\alpha+\beta)^2}{4(\omega+1)(\omega+3)}} \sqrt{\frac{1}{4} + \frac{1-(\alpha-\beta)^2}{4(\omega+2)^2}}.
\]

Consequently,
\[
\xi = 1 - (\tilde{a}_n + \tilde{a}_{n-1} + \tilde{b}_n) \geq \frac{(\alpha-\beta)^2}{4(\omega+2)^2} + \frac{(\alpha-\beta)^2}{4(\omega+1)(\omega+3)} + \frac{(\alpha+\beta)^2}{4(\omega-1)(\omega+3)} - \frac{\beta^2 - \alpha^2}{\omega(\omega+2)} - \frac{(\alpha+\beta)^2}{4(\omega+2)^2} - \frac{(\alpha+\beta)^2 - 1}{2(\omega-1)(\omega+3)} - \frac{\beta^2 - \alpha^2}{\omega(\omega+2)}
\]
\[
= \frac{8\alpha^2 - 2[\alpha^4 + 4\alpha^3 + \alpha^2] + \mu\alpha^2 + \nu\omega + \lambda}{4\omega(\omega-1)(\omega+2)^2(\omega+3)},
\]

where
\[
\lambda = 12[(\beta^2 - \alpha^2)\omega - (\alpha - \beta)^2],
\]
\[
\mu = 6[(\alpha + \beta)^2 - 1] + 4(\alpha - \beta)^2,
\]
\[
\nu = 8(\beta - \alpha) + 8(\beta^2 - \alpha^2).
\]

Note that \( f(\omega), \lambda, \mu, \nu \) are all positive due to the assumption \( \beta^2 > \alpha^2 \) and the facts \( (\alpha + \beta)^2 \geq 1 \) and \( \omega > 1, \forall n > 1 \). Furthermore, it is clear that \( \xi > 0 \) if we require that (for the leading term \( \omega^4 \)) \( 8\alpha^2 - 2 \geq 0 \), i.e.,
\[
|\tilde{a}_n + \tilde{a}_{n-1} + \tilde{b}_n| < 1, \quad \forall |\beta| > |\alpha| \geq \frac{1}{2}, \quad \forall n > 1.
\]

For cases of \( n = 0, 1 \) the proof is trivial. So again we can substitute \( y_i = \tilde{b}_j \pm (\tilde{a}_j + \tilde{c}_j) \) into Eq. (5) separately and this will lead to the inequality (15).
As to the case $\chi^2 > \beta^2$, we can use the same methods as above and we will arrive at $|\chi| > |\beta| > 1/2$. Using together the two cases yields the desired results of part I in Theorem 1.

(II) Case of the Gamma distribution.

Now we recall the generalized Laguerre polynomials

$$L^{(\alpha)}_n(x) = e^x \cdot \sum_{i=0} \binom{x+n+i}{n} \frac{x^i}{i!},$$

which satisfy the orthogonality condition

$$\int_0^{\infty} x^e e^{-x} L^{(\alpha)}_n(x)L^{(\beta)}_m(x)dx = \frac{\Gamma(n+\alpha+1)}{n!} \delta_{n,m}$$

and the recurrence relations

$$x L^{(\alpha)}_n(x) = -(n + 1)L^{(\alpha)}_{n+1}(x) + (2n + \alpha + 1)L^{(\alpha)}_n(x) - (n + \alpha)L^{(\alpha)}_{n-1}(x).$$

We shall prefer to work with the normalized polynomials, that is

$$\tilde{L}^{(\alpha)}_n(x) = \frac{L^{(\alpha)}_n(x)}{\sqrt{h_n}} \quad h_n^2 = \frac{\Gamma(n+\alpha+1)}{n!}.$$  

For the normalized gLPs, we have the following recurrence relations

$$x \tilde{L}^{(\alpha)}_n(x) = a_n \tilde{L}^{(\alpha)}_{n+1}(x) + b_n \tilde{L}^{(\alpha)}_n(x) + c_n \tilde{L}^{(\alpha)}_{n-1}(x)$$

with

$$a_n = -(n + \alpha + 1), \quad b_n = 2n + \alpha + 1, \quad c_n = -n.$$  

For the Gamma distribution, since $y \in [0, \infty)$, letting $y = 0$ and $y \to \infty$ in Eq. (5) leads to

$$K_0(x) > K_{\min}, \quad K_i(x) > 0, \quad 1 \leq i \leq N, \quad \forall x \in D.$$  

This, together with the facts that $a_n + c_n = -b_n$ and Eqs. (11) and (12), lead to the desired result (15).

Remark 1. We remark that in the proof for the Beta distribution, it is essential to make sure that $\tilde{a}_n + \tilde{c}_n < 1$. Otherwise, we may fail to get a strictly diagonal dominant matrix. To see this, let us recall normalized Legendre polynomials ($x = \beta = 0$) which satisfy

$$x \tilde{L}_n(x) = \sqrt{\frac{(n+1)^2}{(2n+1)(2n+3)}} \tilde{L}_{n+1}(x) + \sqrt{\frac{n^2}{(2n-1)(2n+1)}} \tilde{L}_{n-1}(x).$$

We can see that for a fix $n$ (for example, $n = 3$), $\tilde{a}_n + \tilde{c}_n > 1$. We assume one dimension for the parameter space and set in Eq. (4)

$$\frac{1}{\tilde{a}_n + \tilde{c}_n} < |K_1(x)| < 1$$

such that

$$\kappa(x, y) = 1 + K_1(x)y.$$  

It is clear that $\kappa(x, y)$ satisfy assumption (5). Moreover, it is also true that

$$\sum_{k=n} |\Delta k| = |K_1(x)| (\tilde{a}_n + \tilde{c}_n) > 1 = K_0(x) = a_n,$$

which means that the matrix $A(x)$ is no longer strictly diagonal dominant.

Remark 2. Note that the random media is assumed to have a structure of truncated K-L expansion [2], namely,

$$\kappa(x, y) = K_0(x) + \sum_{i=1}^N \sqrt{\lambda_i} K_i(x) y_i,$$

where $K_0(x)$ is the mean diffusivity. $\sum_{i=1}^N \sqrt{\lambda_i}$ is the fluctuations, and $\{y_i\}$ are independent random variables, and $\{\lambda_i, K_i(x)\}$ are the eigenvalues and eigenfunctions for

$$\int C(x, z) K_i(z) dz = \lambda_i K_i(x), \quad i = 1, 2, \ldots.$$
where $C(x,z)$ is the covariance function for $\kappa(x,y)$, and usually $\lambda_i$ decay to zero as $i$ goes to infinity. The point is that if the fluctuation is small, we have that

$$K_0(x) \approx K_{\min} > 0, \quad \forall x, y.$$  

Assume, especially,

$$K_0(x) \geq K_{\min} > \frac{K_0(x)}{2}, \quad \forall x, y.$$  

Then we can prove in this case that the resulting matrix $A(x)$ is strictly diagonal dominant for all $\alpha, \beta > -1$ for the Beta distribution. More precisely, we have

**Theorem 2.** If the random field satisfies a further condition like Eq. (42), then the matrix $A(x)$ is strictly diagonal dominant for all $\alpha, \beta > -1$.

**Proof.** In fact, we only need to fill the gap of $|x|$, $|\beta| < \frac{1}{4}$. Indeed, for these cases, we have for all $n$

$$\hat{a}_n = 2 \sqrt{\frac{(n + 1)(n + \alpha + 1)(n + \beta + 1)(n + \alpha + \beta + 1)}{(\alpha + 1)(\alpha + 2)(\omega + 3)}} = 2 \sqrt{\frac{1}{4} + \frac{1 - (\alpha + \beta)^2}{4(\alpha + 1)(\omega + 3)}},$$

$$\hat{b}_n = \frac{\beta^2 - \alpha^2}{2(\alpha + 2)} < \frac{1}{3},$$

Consequently

$$\hat{b}_n \pm (\hat{a}_n + \hat{c}_n) < 1.$$  

Then substituting $y_i = (\hat{b}_j \pm (\hat{a}_j + \hat{c}_j))\frac{1}{2}$ into Eq. (5) gives

$$\frac{K_0(x)}{2} + \sum_{i=1}^{N} K_i(x) \frac{b_j}{2} \geq K_{\min} \frac{K_0(x)}{2} + \sum_{i=1}^{N} |K_i(x)| \frac{(a_j + c_j)}{2},$$  

which implies

$$K_0(x) + \sum_{i=1}^{N} K_i(x)b_j \geq \sum_{i=1}^{N} |K_i(x)| (a_j + c_j).$$  

This completes the proof. $\square$

**Remark 3.** It is noted that the numerical example used in [5] for steady state case satisfies condition (42), which explains why the authors find their method is stable for all cases of $\alpha$ and $\beta$.

**Acknowledgments**

This research is supported in part by Hong Kong Research Grants Council and the FRG grants of Hong Kong Baptist University. Tang’s research was also supported by Collaborative Research Fund of National Science Foundation of China (NSFC) under Grant No. G10729101.

**References**