Jerrad Hampton, October, 2020: Bi-fidelity PCE

This is a casually written document to present a useful estimate with respect to the bi-fidelity PCE method. I am also presenting the work as I see it, to explain that estimate, which may be incorrect, or may be helpful. Traditionally, PCE would be an estimate that did not have a spatial component, only a stochastic component. As such the approximation looks like

$$\hat{u}(\boldsymbol{\xi}) = \sum_{k=1}^{P} c_k \psi_k(\boldsymbol{\xi}), \tag{1}$$

where ξ is of some stochastic dimension d, and $\{\psi_k\}$ are orthogonal polynomials with respect to the distribution of the d-dimensional random vector $\boldsymbol{\Xi}$, and P is identified as a subset of those orthogonal polynomials of lower order, up to e.g. a total-order of p.

The key difference that allows the bi-fidelity PCE to function well, is that the problem now has spatial components, which we will denote by a vector $\boldsymbol{x}$. Now we have a representation of the approximation given by

$$\hat{u}(\boldsymbol{x}, \boldsymbol{\xi}) = \sum_{k=1}^{P} c_k \psi_k(\boldsymbol{x}, \boldsymbol{\xi}). \tag{2}$$

We need basis functions, $\psi_k(\boldsymbol{x}, \boldsymbol{\xi})$, which are orthogonal with respect to a measure related to both the spatial and stochastic domains, but also useful for approximating the function. Fortunately, this problem has already been solved for forward UQ simulation with random spatial functions known as KLE. However, there are different correlation functions, and one or more correlation length parameters. That is, the space of potential KLE functions is large.

In what I received, this identification of the KLE was not clearly explained, and these details may matter. In the examples, it seems that this identification is concerned with reducing a stochastic dimension from P to r, where the spatial variance is absorbed into the coefficients, and that is the approach I take here. Specifically we want to change (2) to

$$\hat{u}^{(r)}(\boldsymbol{x}, \boldsymbol{\xi}) = \sum_{k=1}^{r} c_k(\boldsymbol{x}) \eta_k(\boldsymbol{\xi}), \tag{3}$$

for some new functions η_k. This is a rank-reduction compression we are able to address, with the bi-fidelity approximation error. Let $\boldsymbol{\Psi}_{N,P}$ denote the matrix with $\boldsymbol{\Psi}(i,j) = \psi_j(\boldsymbol{x}, \boldsymbol{\xi}_i)$, let $\boldsymbol{c}_{P,1}$ denote a vector of coefficients, and let $\boldsymbol{u}_{N,1}(i) = u(\boldsymbol{x}, \boldsymbol{\xi}_i)$, the evaluation of the solution for the ith realized random variable. The method employed in what I received is to instead have $\boldsymbol{C}_{P,M}$, and $\boldsymbol{U}_{N,M}$ where M is the number of spatial points sampled, giving (with transposes to this problem in what I was sent, which I'll ignore here)

$$\boldsymbol{\Psi}\boldsymbol{C}_{P,M} = \boldsymbol{U}. \tag{4}$$

In this form, the functions ψ_k above have been separated into a stochastic function, which resides in $\boldsymbol{\Psi}$, and a spatial function, which is absorbed into $\boldsymbol{C}$.

We perform the bi-fidelity method to approximate $\boldsymbol{U}$. The theorem we already have bounds,

$$\|\boldsymbol{U}_H - \bar{\boldsymbol{U}}_H\|_F, \tag{5}$$

for $\bar{\boldsymbol{U}}_H$ constructed through the different bi-fidelity algorithm that does not update the coefficient matrix $\boldsymbol{C}$. We use $\hat{\boldsymbol{U}}$ for this version's construction, which does update the coefficient matrix. The method works best if the spatial information is highly compressible, and if it is not (5) will be large, with or without the coefficient update. However, the recomputation of the coefficient matrix $\boldsymbol{C}$ makes this method more adaptive, but also makes that bound from the non-updating algorithm generally looser, and more difficult to apply.

Connected to the bi-fidelity method, we have to identify the r new basis functions η_i, and the matrix,

$$\boldsymbol{\eta}_{N,r} := \boldsymbol{\Psi}_{N,P}\boldsymbol{P}_{P,r}(\boldsymbol{U}^{(r)}), \tag{6}$$

have the name motivated by η in what I was sent. Specifically, η_i represents the function

$$\eta_i(\boldsymbol{\xi}) = \sum_{k=1}^{P} \boldsymbol{\beta}(i,k)\psi_k(\boldsymbol{\xi}), \tag{7}$$

where the affect of the associated projection is given in the coefficients of the orthogonal matrix, $\boldsymbol{\beta}$. These η_i are thus orthogonal, and this orthogonality is critical to the ℓ_2-approximation theory.

This gives the matrix equation which will be used for the final approximation for the PCE itself,

$$\boldsymbol{\eta}_{N_H,r}\boldsymbol{C}_{r,M_H} = \boldsymbol{U}^H_{N_H,M_H}(\boldsymbol{\eta}). \tag{8}$$

That is, $\{u_i\}_{i=1}^{M_H}$ are approximated by $\{\eta_i\}_{i=1}^{r}$, by a set of r coefficients for each of the M_H points. Here, as in what I received, I perform this approximation by M_H independent least squares regressions, which is equivalent to doing the Frobenius regression on the entire matrix. This is the key approximation with high-fidelity samples.

The key low-fidelity computation is the basis identification, specifically how well can the high-fidelity problem be reconstructed by r spatially varying functions. In what I was sent, this has a claimed connection to a KLE, and there are similarities, but this approach does not progress through identifying a covariance kernel, and eigenvalues/functions of it. However, being a KLE or not

does not matter for the approximation here.

To this end, I state our previous bi-fidelity theorem with the matrices that we need. Here we use $\boldsymbol{U}_H^T := \boldsymbol{U}_{N,M_H}$ for $\boldsymbol{H}$, and $\boldsymbol{U}_L^T := \boldsymbol{U}_{N,M_L}$ for $\boldsymbol{L}$, and using a hat to denote the rank r bi-fidelity or low-fidelity reconstruction, respectively, of the matrix. Both of these use only the realizations of the model, and neither of these involve computing PCE coefficients. The transposes are to avoid making changes to the theorem, and do not effect the outcome as the $\|A\|_2 = \|A^T\|_2$. I've also changed $\rho_k(\tau)$ as as to avoid needing to put minimums in the later estimates.

Theorem 0.0.1. *For any $\tau \geq 0$, let*

$$\epsilon(\tau) = \|\boldsymbol{U}_H^T\boldsymbol{U}_H - \tau\boldsymbol{U}_L^T\boldsymbol{U}_L\|_2. \tag{9}$$

Let $\bar{\boldsymbol{U}}_H$ and $\bar{\boldsymbol{U}}_L$ be corresponding static coefficient bi-fidelity estimates of rank r with coefficients $\boldsymbol{C}_L$, and let σ_k denote the kth largest singular value of $\boldsymbol{U}_L$. Then,

$$\|\boldsymbol{U}_H - \bar{\boldsymbol{U}}_H\|_2 \leq \rho_k(\tau); \tag{10}$$

where $\rho_k(\tau)$ is defined by

$$\rho_k(\tau) := \min_{\tau, k \leq rank(\boldsymbol{U}_L)} (1 + \|\boldsymbol{C}_L\|)\sqrt{\tau\sigma_{k+1}^2 + \epsilon(\tau)} + \|\boldsymbol{U}_L - \bar{\boldsymbol{U}}_L\|\sqrt{\tau + \epsilon(\tau)\sigma_k^{-2}}. \tag{11}$$

When $k = rank(\boldsymbol{A}_L)$, we set $\sigma_{k+1} = 0$.

Remark 0.0.1. *I note again that this theorem does not easily apply to this algorithm, as the coefficient matrix $\boldsymbol{C}_L$ is not reused for this bound. However, this bound can still provide an estimate of truncation error. This estimate relies on the Gramian estimate, which relies on high-fidelity samples, and it may be more efficient to simply perform the a posteriori error analysis with the new $\boldsymbol{C}$.*

We also have the theorem from our ℓ_2 paper, that I find more easily applied, though other versions can be used. First, let's define the coherence. As we aren't using importance sampling, this can be simplified by removing the weight function. For presentation, I will ignore the more complicated coherence, and in the paper that complicated coherence has a the probability for $\mathcal{E}$ having an extra $1/r$ term which could be quite large. Let the coherence be defined as

$$\mu_2 := \sup_{\boldsymbol{\xi} \in \Omega} \sum_{k=1}^{r} |\eta_k(\boldsymbol{\xi})|^2. \tag{12}$$

We note that under coherence optimal conditions, $\mu_2 = r$, and such conditions can be guaranteed by importance sampling. Additionally, define the truncation error $\delta_i(\boldsymbol{\xi})$ to be the function achieving the minimum L_2 distance between $u_i(\boldsymbol{\xi})$

and the space of approximations from linear combinations of the basis functions $\{\eta_k(\boldsymbol{\xi})\}_{k=1}^r$. That is δ is the ϵ of the ℓ_2 paper, which has been changed as it may be confused with the $\epsilon(\tau)$ used in the bi-fidelity estimate. We now present Theorem 2.1 of the ℓ_2 paper. We note that the error ν_i in this theorem is the useful bound that we seek.

Theorem 0.0.2. *Let*

$$\hat{u}_i(\boldsymbol{\Xi}) = \sum_{k=1}^{r} \hat{c}(i,k)\eta_k(\boldsymbol{\xi}), \tag{13}$$

where $\hat{c}_i$ is the least-squares solution. It follows that for $\mathcal{E}$, which is independent of i, and is a sampling event that occurs with probability

$$\mathbb{P}(\mathcal{E}) \geq 1 - 2r\,exp(-0.1 N_H \mu_2^{-1}), \tag{14}$$

that

$$\nu_i := \mathbb{E}\left(\|u_i(\boldsymbol{\Xi}) - \hat{u}_i(\boldsymbol{\Xi})\|^2_{L_2(\Omega,f)}; \mathcal{E}\right) \leq \left(1 + \frac{4\mu_2}{N_H}\right)\mathbb{E}(\delta_i^2(\boldsymbol{\Xi})), \tag{15}$$

where μ_2 is as in (12), and

$$\mathbb{E}\left(X; \mathcal{E}\right) = \int_{\mathcal{E}} X(\boldsymbol{\xi}) f(\boldsymbol{\xi}) d\boldsymbol{\xi} = \mathbb{E}(X|\mathcal{E})\mathbb{P}(\mathcal{E}) \tag{16}$$

denotes the expectation restricted to the event (also known as restricted expectation), and is closely related to conditional expectation.

Remark 0.0.2. *I note here that this error bound is for each point in space. The error from the bi-fidelity approximation can be concentrated in certain spatial regions, and this is likely to apply to the PCE approximations for these M_H points as well. The error from each point can be accumulated, as we do in the final corollary here.*

Now we have the new, original, theorem that ties these approximations together.

Theorem 0.0.3. *For $\mathcal{E}$ as in Theorem 0.0.2, and from that theorem, it follows that,*

$$\nu_i := \mathbb{E}\left(\|u_i(\boldsymbol{\Xi}) - \hat{u}_i(\boldsymbol{\Xi})\|^2_{L_2(\Omega,f)}; \mathcal{E}\right) \tag{17}$$

$$\leq \left(1 + \frac{4\mu_2}{N_H}\right)\left(\frac{\Theta_{N_H}}{N_H}\|\boldsymbol{U}_H(:,i) - \hat{\boldsymbol{U}}_H(:,i)\|_2^2\right), \tag{18}$$

Here $\boldsymbol{U}_H$ is the $N_H \times M_H$ matrix of high-fidelity data; $\boldsymbol{C}_L$ is the coefficient matrix associated with $\hat{\boldsymbol{U}}_H$, which is the bi-fidelity approximation to $\boldsymbol{U}_H$; and Θ_{N_H} is a random variable that converges almost surely to 1 as $N_H \to \infty$.

Remark 0.0.3. *We note that ν_i does not tend to zero as $N_H \to \infty$. The first term tends to 1, as μ_2 is finite and does not vary with N_H. However, the second term's $\|\boldsymbol{U}_H(:,i) - \hat{\boldsymbol{U}}_H(:,i)\|_2^2$, grows with N_H, such that the second term converges to a constant, a mean squared error in approximation from using the r basis functions to approximate $\boldsymbol{U}_H$, derived from $\mathbb{E}(\delta_i^2(\boldsymbol{\Xi}))$.*

We also note that Θ_{N_H} depends on the realizations of the random variables $u(\boldsymbol{\Xi})$, and the bi-fidelity reconstruction which fill the vector $\boldsymbol{U}_H(:,i) - \hat{\boldsymbol{U}}_H(:,i)$. Θ_{N_H} is a multiplicative correction to correct the sample mean to the true mean.

Finally, though this bound is ultimately simple, and has a moment estimate in it that is stochastic, it is rather tight. The corollary that uses Theorem 0.0.1, has looser bounds as the bi-fidelity bound is loose, but they are as tight as that loose bound can reasonably permit.

Proof:
We begin with (15), seeking to estimate the truncation error $\mathbb{E}(\delta_i^2(\boldsymbol{\Xi}))$. Recall that, $\boldsymbol{U}_H(k,i)$ contains $u_i(\boldsymbol{\xi}_k)$, and that $\hat{\boldsymbol{U}}_H(k,i)$ contains the corresponding $\hat{u}_i(\boldsymbol{\xi}_k)$ computed via the bi-fidelity PCE. It follows that $\|\boldsymbol{U}_H(:,i) - \hat{\boldsymbol{U}}_H(:,i)\|_2^2$ is an estimate for $\mathbb{E}(\delta_i^2(\boldsymbol{\Xi}))$, where Θ_{N_H} thus converges to 1 almost surely as $N_H \to \infty$ by the strong law of large numbers. ■

We then have the following corollary, where we replace the high-fidelity bound with the associated bi-fidelity bound.

Corollary 1. *Under the conditions of Theorem 0.0.3, it follows that,*

$$\nu_i \leq \left(1 + \frac{4\mu_2}{N_H}\right)\left(\frac{X_{N_H}}{N_H}\rho_k^2(\tau)\right); \tag{19}$$

$$\sum_{i=1}^{M_H} \nu_i \leq \left(1 + \frac{4\mu_2}{N_H}\right)\left(\frac{r\bar{X}_{N_H}}{N_H}\rho_k^2(\tau)\right). \tag{20}$$

Here $\rho_k(\tau)$ is as in Theorem 0.0.1; X_{N_H}, and $\bar{X}_{N_H}$ are random variables which converge a.s. to values in $(0,1]$; r is the rank of $\boldsymbol{U}_H - \bar{\boldsymbol{U}}_H$, and the rest is as in Theorem 0.0.3.

Proof:
Note that for Y_{N_H} a random variable that depends on N_H,

$$\|\boldsymbol{U}_H(:,i) - \hat{\boldsymbol{U}}_H(:,i)\|_2^2 \leq Y_{N_H}\|\boldsymbol{U}_H(:,i) - \bar{\boldsymbol{U}}_H(:,i)\|_2^2, \tag{21}$$

where we recall that the difference between $\hat{\boldsymbol{U}}_H$ and $\bar{\boldsymbol{U}}_H$ is that the former recomputes the coefficient matrix using the high-fidelity data while the latter does not, using the coefficients computed using low-fidelity data. As a result, the LHS converges to $\mathbb{E}(\delta_i^2(\boldsymbol{\Xi}))$, while the RHS converges to an unknown finite quantity. As the LHS converges to the minimum possible value, the random

variable Y_{N_H} converges to some unknown value in $(0, 1]$. Using a a simple matrix norm inequality, and Theorem 0.0.1,

$$\|\boldsymbol{U}_H(:, i) - \bar{\boldsymbol{U}}_H(:, i)\|_2 \leq \|\boldsymbol{U}_H - \bar{\boldsymbol{U}}_H\|_2, \tag{22}$$

$$\leq \rho_k(\tau). \tag{23}$$

In (19), we have $X_{N_H} = \Theta_{N_H} Y_{N_H}$, where Θ_{N_H} is as in Theorem 0.0.3. As Θ_{N_H} converges a.s to 1 and Y_{N_H} converges a.s. to some unknown value in $(0, 1]$, X_{N_H} converges a.s. to the same unknown value as Y_{N_H}.

To show (20), we note that

$$\sum_{i=1}^{M_H} \nu_i \leq \sum_{i=1}^{M_H} \left(1 + \frac{4\mu_2}{N_H}\right) \left(\frac{\Theta_{N_H}(i)}{N_H} \|\boldsymbol{U}_H(:, i) - \hat{\boldsymbol{U}}_H(:, i)\|_2^2\right), \tag{24}$$

where the $\Theta_{N_H}(i)$ are a set of M_H random variables, each convergent to 1 a.s. As a result the maximum of these M_H random variables, which we refer to as $\bar{\Theta}_{N_H}$ also converges to 1 a.s. Having $\bar{X}_{N_H}$ be similarly defined as the maximum of the X_{N_H} above, it follows that $\bar{X}_{N_H}$ converges a.s. to some value in $(0, 1]$. We thus have

$$\sum_{i=1}^{M_H} \nu_i \leq \sum_{i=1}^{M_H} \left(1 + \frac{4\mu_2}{N_H}\right) \left(\frac{\bar{\Theta}_{N_H}}{N_H} \|\boldsymbol{U}_H(:, i) - \hat{\boldsymbol{U}}_H(:, i)\|_2^2\right), \tag{25}$$

$$= \left(1 + \frac{4\mu_2}{N_H}\right) \left(\frac{\bar{\Theta}_{N_H}}{N_H} \sum_{i=1}^{M_H} \|\boldsymbol{U}_H(:, i) - \hat{\boldsymbol{U}}_H(:, i)\|_2^2\right), \tag{26}$$

$$\leq \left(1 + \frac{4\mu_2}{N_H}\right) \left(\frac{\bar{X}_{N_H}}{N_H} \sum_{i=1}^{M_H} \|\boldsymbol{U}_H(:, i) - \bar{\boldsymbol{U}}_H(:, i)\|_2^2\right), \tag{27}$$

$$= \left(1 + \frac{4\mu_2}{N_H}\right) \left(\frac{\bar{X}_{N_H}}{N_H} \|\boldsymbol{U}_H - \bar{\boldsymbol{U}}_H\|_F^2\right), \tag{28}$$

$$\leq \left(1 + \frac{4\mu_2}{N_H}\right) \left(\frac{r\bar{X}_{N_H}}{N_H} \rho_k^2(\tau)\right), \tag{29}$$

where the last line shows (20) and follows from Theorem 0.0.1 and the matrix inequality

$$\|\boldsymbol{U}_H - \bar{\boldsymbol{U}}_H\|_F \leq \sqrt{r} \|\boldsymbol{U}_H - \bar{\boldsymbol{U}}_H\|_2, \tag{30}$$

for r the rank of $\boldsymbol{U}_H - \bar{\boldsymbol{U}}_H$. ■.