Proceedings of the
47th IEEE Conference on Decision and Control
Cancun, Mexico, Dec. 9-11, 2008

On the Optimality of the Proper Orthogonal Decomposition and Balanced Truncation

Seddik M. Djouadi

I. Abstract

The proper orthogonal decomposition (POD), also known as Karhunen-Loève decomposition or principal component analysis, and balanced truncation, are shown to be optimal in the sense of distance minimizations in spaces of Hilbert-Schmidt or trace-class 2 integral operators. Both POD and balanced truncation are shown to be optimal approximations by finite rank operators in the Hilbert-Schmidt norm. Optimality of balanced truncation seems to have been overlooked in the literature, and in fact, it is commonly thought to be non-optimal in any sense. The role of POD and balanced truncation in minimizing different n-widths of specific compact operators is discussed. The n-widths quantify inherent and representation errors due to lack of data or inaccurate measurements and loss of information.

II. Introduction

In this paper, we consider two popular model reduction techniques, the proper orthogonal decomposition (POD), which has been extensively investigated in distributed parameters systems due to its order reduction capability [1]-[13], and balanced truncation, which is a simple yet efficient model reduction technique widely used in reducing model orders of high order linear systems [23], [15]. In particular, we study the optimality of both model reduction techniques and show that, in fact, the two techniques are related, and optimal in the sense of minimizing the Hilbert-Schmidt or trace class 2 norm, although on different spaces. Note that POD is known to solve a certain constrained optimization problem [16], and is optimal in with respect to capturing the energy of the data set [19], but here we show in fact that POD is optimal in a wider sense. Optimality of balanced truncation seems to be missing in the literature. Actually, it has been widely claimed that balanced truncation is not optimal in any sense [23]. We first compute the optimal approximation in the sense of approximating the associated Hankel operator in a specific Hilbert-Schmidt norm. The optimum is a finite rank operator which is not necessarily a Hankel operator. However, by using a particular balanced realization based on the Schmidt pairs of the Hankel operator associated to the original system, we show that the optimal operator can be realized by the corresponding truncated balanced realization. Optimality of both POD and balanced truncation was stated in [14] but without formal proofs.

Geometric interpretation of POD and balanced truncation in terms of optimizing the Kolmogorov, Gel'fand and linear n-widths of the corresponding compact operators is discussed. These n-widths quantify the inherent error generated in the information collecting stage of simulation or identification, due to lack of data and inaccurate measurements, and the representation error due to the loss of information in the information processing stage.

The paper is organized as follows. In section III, POD is showed to be optimal as a shortest distance minimization between an L^2 function of the time and space variables to a particular subspace with explicit computations. In section IV, balanced truncation is shown to be in some sense analogous to POD, in that, it is also optimal in the sense of shortest distance minimization in the Hilbert-Schmidt norm, albeit in different integral operator spaces. Section V discussed optimality of both POD and the balanced truncation in terms of minimization of various n-widths. In section VI we conclude with the summary of our contribution.

III. Optimality of Proper Orthogonal Decomposition

Proper orthogonal Decomposition (POD) has been used extensively to determine efficient bases for dynamical systems, random processes and large data set in general. It was introduced in the context of turbulence by Lumley [16]. It is also known as the Karhunen-Loéve decomposition, principal component analysis, singular systems analysis, and singular value decomposition [17], [18]. The fundamental idea behind POD is as follows: Given a set of simulation data or snapshots $\{S_i\}_{i=1}^N$ of a function $w(t, \mathbf{x})$, in the standard Hilbert space $L^2(T, \Omega)$, where $\mathbf{x} \in \Omega$ for some set Ω of $\mathbb{R}^p$ and T represents a finite or infinite time interval. The nth POD vector $\phi_n(\mathbf{x})$ is chosen recursively so as to minimize the cost function [16], [20]

$$J(\phi_n) := \int_0^T \int_\Omega \Big| S_i(t,\ \mathbf{x}) - \sum_{j=1}^n \alpha_j \phi_j(\mathbf{x}) \Big|^2 \mathbf{dx} dt \qquad (1)$$

subject to the constraints

$$\alpha_j(t) \;=\; \int_\Omega S_i(t,\ \mathbf{x})\phi_j(\mathbf{x})\mathbf{dx} \qquad (2)$$

$$\int_\Omega \phi_i(\mathbf{x})\phi_j(\mathbf{x})d\mathbf{x} \;=\; \delta_{ij}, \text{ for } i,j = 1,2,\cdots,n \qquad (3)$$

The optimal POD basis is given by the eigenfunctions $\{\phi_i\}$ of the averaged autocorrelation function, denoted $R\big(\mathbf{x}, \mathbf{x}'\big)$,

This work was supported in part by NSF Award No. CMMI-082592 and AFoSR Award No. FA9550-08-1-0450. S.M. Djouadi is with the Department of Electrical Engineering and Computer Science at The University of Tennessee, Knoxville, TN 37996 djouadi@eecs.utk.edu.

978-1-4244-3124-3/08/$25.00 ©2008 IEEE

of the snapshots, that is, [16], [17]

$$R(\mathbf{x},\mathbf{x}') := \int_0^T S_i(t,\mathbf{x})S_i(t,\mathbf{x}')dt \tag{4}$$

which solves the eigenvalue problem

$$\int_\Omega \int_0^T S_i(t,\mathbf{x})S_i(t,\mathbf{x}')\phi(\mathbf{x}')dtd\mathbf{x}' = \lambda\phi(\mathbf{x}) \tag{5}$$

The Hilbert space $L^2(T,\ \Omega)$ is endowed with the norm

$$\|w(t,\ \mathbf{x})\|_2 := \left(\int_0^T \int_\Omega |w(t,\ \mathbf{x})|^2 d\mathbf{x}dt\right)^{\frac{1}{2}} < \infty \tag{6}$$

For *fixed* n, define the shortest distance minimization in the $\|\cdot\|_2$-norm from the function $w(t,\ \mathbf{x})$ to the subspace $\mathcal{S}$, by

$$\mu := \inf_{s\in\mathcal{S}} \|w(t,\ \mathbf{x}) - s(t,\mathbf{x})\|_2 \tag{7}$$

where the subspace $\mathcal{S}$ is defined as

$$\mathcal{S} := \Big\{\sum_{i=1}^n a_i(t)\varphi_i(\mathbf{x}) :\ a_i(t) \in L^2(T),\ \varphi_i(\mathbf{x}) \in L^2(\Omega)\Big\} \tag{8}$$

Note that this distance problem is posed in an infinite-dimensional space. For finite dimensional spaces, in particular for distances to lower rank matrices see [27], where SVD techniques are used. To compute the distance we view $w(t,\ \mathbf{x})$ as a Hilbert-Schmidt kernel for an integral operator T mapping $L^2(\Omega)$ into $L^2(T)$ both endowed with the standard $\|\cdot\|_2$-norm, and defined by

$$(T\phi)(t) := \int_\Omega w(t,\mathbf{x})\phi(\mathbf{x})d\mathbf{x} \tag{9}$$

It is known that such an operator is compact [21], that is, an operator which maps bounded sets into pre-compact sets. The operator T is said to be a Hilbert-Schmidt or a trace-class 2 operator [25]. Let us denote the class of Hilbert-Schmidt operators acting from $L^2(T)$ into $L^2(\Omega)$, by $\mathcal{C}_2$, and the Hilbert-Schmidt norm $\|\cdot\|_{\mathrm{HS}}$. Define the adjoint of $T^\star$ as the operator acting from $L^2(T)$ into $L^2(\Omega)$ by

$$< Tf, g >_2 := \int_0^T \int_\Omega w(t,\mathbf{x})f(\mathbf{x})d\mathbf{x}g(t)dt$$

$$= \int_\Omega f(\mathbf{x}) \int_0^T w(t,\mathbf{x})g(t)dtd\mathbf{x} =: < f, T^\star g >_1 \tag{10}$$

showing that $(T^\star g)(t) = \int_0^T w(t,\mathbf{x})g(t)dt$.
Using the polar representation of compact operators [25], $T = U(T^\star T)^{\frac{1}{2}}$, where U is a partial isometry and $(T^\star T)^{\frac{1}{2}}$ the square root of T, which is also a Hilbert-Schmidt operator, and admits a spectral factorization of the form [25]

$$(T^\star T)^{\frac{1}{2}} = \sum_i \lambda_i \nu_i \otimes \nu_i \tag{11}$$

where $\lambda_i > 0$, $\lambda_i \searrow 0$ as $i \uparrow \infty$, are the eigenvalues of $(T^\star T)^{\frac{1}{2}}$, and ν_i form the corresponding orthonormal sequence of eigenvectors, i.e., $(T^\star T)^{\frac{1}{2}}\nu_i = \lambda_i\nu_i,\ i = 1, 2, \cdots$. Putting $U\nu_i =: \psi_i$, we can write

$$T = \sum_i \lambda_i\ \nu_i \otimes \psi_i \tag{12}$$

Both $\{\nu_i\}$ and $\{\psi_i\}$ are orthonormal sequences in $L^2(T)$ and $L^2(\Omega)$, respectively. The sum (12) has either a finite or countably infinite number of terms. The above representation is unique. Noting that the polar decomposition of $T^\star = U^\star(TT^\star)^{\frac{1}{2}}$, a similar argument yields

$$(TT^\star)^{\frac{1}{2}} = \sum_i \lambda_i\psi_i \otimes \psi_i T^\star \ = \ \sum_i \lambda_i\psi_i \otimes \nu_i \tag{13}$$

which shows that α_i from an orthonormal sequence of eigenvectors of $(TT^\star)^{\frac{1}{2}}$ corresponding to the eigenvalues λ_i. From (11) and (13) it follows that

$$T\psi_i = U(T^\star T)^{\frac{1}{2}}\psi_i = \lambda_i\nu_i \tag{14}$$

$$T^\star\nu_i = U^\star(TT^\star)^{\frac{1}{2}}\nu_i = \lambda_i\psi_i \tag{15}$$

We say that ψ_i and ν_i constitute a Schmidt pair [21]. In terms of integral operators expressions, identities (14) and (15) can be written, respectively, as

$$\nu_i(t) = \int_\Omega w(t,\mathbf{x})\psi_i(\mathbf{x})d\mathbf{x} \tag{16}$$

$$\psi_i(\mathbf{x}) = \int_0^T w(t,\mathbf{x})\nu_i(t)dt \tag{17}$$

In terms of the eigenvalues λ_i's of T, its Hilbert-Schmidt norm $\|\cdot\|_{HS}$ is given by [25]

$$\|T\|_{HS} = \Big(\sum_i \lambda_i^2\Big)^{\frac{1}{2}} = \left(\int_0^T \int_\Omega |w(t,\ \mathbf{x})|^2 d\mathbf{x}dt\right)^{\frac{1}{2}} \tag{18}$$

Note that since the operator T is Hilbert-Schmidt the sum in (18) is finite. The Hilbert-Schmidt norm is also induced by the operator inner product defined by (21). By interpreting each elements of the subspace $\mathcal{S}$ defined in (8) as a Hilbert-Schmidt operator as we did for $w(t,\mathbf{x})$, we see that $\mathcal{S}$ is the subspace of Hilbert-Schmidt operators of rank n, i.e.,

$$\mathcal{S} = \{s = \sum\nolimits_{j=1}^n \vartheta_j\ f_j(t) \otimes \chi_j(\mathbf{x}) :\ f_j(t) \in L^2(T),$$
$$\chi_j(\mathbf{x}) \in L^2(\Omega),\ \vartheta_j \in \mathbb{R}\} \tag{19}$$

In addition, the distance minimization (7) is then the minimal distance from T to Hilbert-Schmidt operators of rank n. In other terms, we have

$$\mu = \min_{s\in\mathcal{S}} \|T - s\|_{\mathrm{HS}} \tag{20}$$

The space of Hilbert-Schmidt operators is in fact a Hilbert space with the inner product [25], denoted $(\cdot,\ \cdot)$, if A and B are two Hilbert-Schmidt operators defined on $L^2(\Omega)$,

$$(A, B) := tr(B^\star A) \tag{21}$$

where tr denotes the trace, which in this case is given by the sum of the eigenvalues of the operator $B^\star A$ which is necessarily finite [25]. Note that the inner product (21) induces the Hilbert-Schmidt norm $\|A\|_{\mathrm{HS}} = \big(tr(A^\star A)\big)^{\frac{1}{2}}$. In the case where A and B are integral operators with kernels $A(t,\mathbf{x})$ and $B(t,\mathbf{x})$, respectively, the inner product can be realized concretely by

$$(A, B) = \int_0^T \int_\Omega A(t,\mathbf{x})B(t,\mathbf{x})d\mathbf{x}dt \tag{22}$$

The solution to the distance minimization (20) is simply given by the orthogonal projection of T onto $\mathcal{S}$. To compute the latter, note that the eigenvectors of $(TT^\star)^{\frac{1}{2}}$ and $(T^\star T)^{\frac{1}{2}}$ form orthonormal bases (by completing them if necessary) for $L^2(T)$ and $L^2(\Omega)$, respectively. In terms of the eigenvectors ν_j and ψ_j the subspace $\mathcal{S}$ can be written as

$$\mathcal{S} = \text{Span}\{\nu_j \otimes \psi_j,\ j = 1, 2, \cdots, n\} \tag{23}$$

Since the shortest distance minimization (20) is posed in a Hilbert space, by the principle of orthogonality it is solved by the orthogonal projection $P_\mathcal{S}$ acting from $\mathcal{C}_2$ onto $\mathcal{S}$. The latter can be computed by first determining the orthogonal projection P_ν onto $\text{Span}\{\nu_j,\ j = 1, 2, \cdots, n\}$, and the orthogonal projection P_ψ onto $\text{Span}\{\psi_j,\ j = 1, 2, \cdots, n\}$. These projections have finite rank and since the ν_j's and ψ_j's are orthogonal vectors in $L^2(T)$ and $L^2(\Omega)$, respectively, it can be easily verified that P_ν and P_ψ are given by

$$\begin{aligned}(P_\nu f)(t) &= \sum_{j=1}^{n}\Big(\int_0^T f(t)\nu_j(t)dt\Big)\ \nu_j(t) \\ (P_\psi G)(\mathbf{x}) &= \sum_{j=1}^{n}\Big(\int_\Omega G(\mathbf{x})\psi_j(\mathbf{x})d\mathbf{x}\Big)\ \psi_j(\mathbf{x}) \end{aligned} \tag{24}$$

The overall orthogonal projection $P_\mathcal{S}$ can be computed as

$$P_\mathcal{S} = P_\nu \otimes P_\psi \tag{25}$$

That is, if $W \in \mathcal{C}_2$ has spectral decomposition $\sum_{i=1} \eta_i u_i \otimes v_j$, where $u_i \in L^2(T)$, $v_i \in L^2(\Omega)$, then

$$P_\mathcal{S} W = \sum_{i=1} \eta_i P_\mathcal{S}(u_i \otimes v_i) = \sum_{j=1}^{n} \theta_j \nu_j \otimes \psi_j,\ \exists\ \theta_j \tag{26}$$

where the last finite sum is obtained thanks to orthogonality, i.e., only the u_i's and v_i's that live in the span of ν_j's and ψ_j's, respectively, are retained. For the orthogonality property we only need verify that

$$\begin{gathered} x \otimes y - (P_\nu \otimes P_\psi)(x \otimes y) \perp u \otimes v, \\ x \in L^2(T),\ y \in L^2(\Omega),\ u \otimes v \in \mathcal{S} \end{gathered}$$

Computing the inner product, we get

$$< x - P_\nu x,\ u >_1 < y - P_\psi,\ v >_2 = 0$$

because P_ν is the orthogonal projection of $L^2(T)$ onto $\text{Span}\{\nu_j,\ j = 1, 2, \cdots, n\}$, and P_ψ the orthogonal projection of $L^2(\Omega)$ onto $\text{Span}\{\psi_j,\ j = 1, 2, \cdots, n\}$. The minimizing operator $s_o \in \mathcal{S}$ in (20) is then given by

$$s_0 := P_\mathcal{S} T = \sum_{i=1}^{n} \lambda_i \nu_i \otimes \psi_i \tag{27}$$

$$\mu = \|T - P_\mathcal{S} T\|_{\text{HS}} = \Big(\sum_{i=n+1}^{\infty} \lambda_i^2\Big)^{\frac{1}{2}} \tag{28}$$

And as $n \uparrow \infty$, $\|T - P_\mathcal{S} T\|_{\text{HS}} \searrow 0$. Therefore, the minimizing function $s_o(t, \mathbf{x})$ in (7) corresponds to the kernel of s_o, which is given by

$$s_o(t, \mathbf{x}) = \sum_{i=1}^{n} \lambda_i \nu_i(t)\psi_i(\mathbf{x}) \tag{29}$$

Now note that $\alpha_i(t) = \lambda_i \nu_i(t)$, $\phi(\mathbf{x}) = \psi(\mathbf{x})$, we see that $s_o(t, \mathbf{x})$ solves the optimization problem (1) since it minimizes the cost function $J(\phi_n)$ and $\alpha_i(t)$, $\phi_i(\mathbf{x})$ satisfy constraints (2) and (3), respectively. Moreover, (16) and (17) imply that $\phi_i(\mathbf{x})$ is related to $\alpha_i(t)$ by

$$\phi_i(\mathbf{x}) = \frac{1}{\lambda_i}\int_0^T w(t, \mathbf{x})\alpha_i(t)dt \tag{30}$$

In the next section, we show that balanced truncation is in some sense similar to POD, in that, it is also optimal in the sense of distance minimization in the Hilbert-Schmidt norm, albeit in different operator spaces. The techniques developed for POD will help us in the context of showing the optimality of balanced truncation as well.

IV. Optimality of Balanced Truncation

Balanced truncation is a simple and popular model reduction technique, which can be described as follows [23], [15]: Suppose we have a stable linear time invariant (LTI) system described by the following n-dimensional state space equation

$$\dot{x}(t) = Ax(t) + Bu(t),\ y(t) = Cx(t) \tag{31}$$

where $x(t)$ is the $n \times 1$-state vector of the system, $u(t)$ is an $m \times 1$-input vector, and $y(t)$ is an $p \times 1$-output or measurement vector. A, B, and C are constant matrices of appropriate dimensions.

The underlying idea of balanced truncation is to take into account both the input and output signals of the system when deciding which states to truncate with appropriate scaling. The latter is performed by transforming the controllability and observability gramians, denoted W_c and W_o respectively, so that they are equal and diagonal. Computing a state balancing transformation M is achieved by first calculating the matrix [15], $W_{co} = W_c W_o$, and determining its eigenmodes $W_{co} = M\Lambda M^{-1}$.

$$\begin{aligned} \dot{z}(t) &= \tilde{A}z(t) + \tilde{B}u(t),\ y(t) = \tilde{C}z(t) && (32) \\ \tilde{A} &:= M^{-1}AM,\ \tilde{B} := M^{-1}B,\ \tilde{C} := CM && (33) \end{aligned}$$

The transformation M is chosen such that the controllability and observability gramians for the transformed system satisfy [15]

$$\tilde{W}_c = \tilde{W}_o = M^{-1}W_c M^{-1T} = M^T W_o M =: \Sigma \tag{34}$$

where Σ is a diagonal matrix that satisfies $\Sigma^2 = \Lambda$, and the diagonal elements of Σ, σ_i's, are known as the Hankel singular vales of the system, i.e.,

$$\Sigma = \text{diag}\{\sigma_1,\ \sigma_2,\ \cdots,\ \sigma_n\} \tag{35}$$

where σ_i's are arranged in non-increasing order $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$. In balanced truncation only states corresponding to large Hankel singular values are retained. Small Hankel singular values correspond to states which are deemed weakly controllable and weakly observable, and therefore deleted from the state-space model. For instance, if the first n_r states are retained then the resulting transformation is given by $M_r = P_r M$, where P_r is the orthogonal projection of rank r. The reduced order model is obtained by letting $x_r = P_r M x$ as follows

$$\dot{x}_r(t) = A_r x_r(t) + B_r u(t),\ y_r(t) = C_r x_r(t) \quad (36)$$
$$A_r := P_r M^{-1} A M P_r;\ B_r := P_r M^{-1} B,\ C_r := C M P_r$$

Balanced truncation is optimal in a precise sense when starting from a balanced realization. To see this define a causal bounded input-output operator G acting on the standard space $L^2(-\infty,\ \infty)$ of absolutely square integrable functions defined on $(-\infty,\ \infty)$, into $L^2(-\infty,\ \infty)$ described by the convolution [15]

$$(Gu)(t) := \int_{-\infty}^{t} Ce^{A(t-\tau)}Bu(\tau)d\tau \qquad (37)$$

Now, define the Hankel operator of G by

$$\Gamma_G : L^2(-\infty,\ 0] \longmapsto L^2[0,\ \infty),\ \Gamma_G := P_+ G|_{L^2(-\infty,\ 0]}$$

where $G\big|_{L^2(-\infty,\ 0]}$ denotes the restriction of G to $L^2(-\infty,\ 0]$, and P_+ is the orthogonal projection acting from $L^2(-\infty,\ \infty)$ into $L^2[0,\ \infty)$, i.e., P_+ is the truncation operator

$$P_+ f(t) = \begin{cases} f(t) & \text{if } t \geq 0 \\ 0 & \text{if } t < 0 \end{cases},\ f(t) \in L^2(-\infty,\ \infty) \quad (38)$$

Then, the Hankel operator Γ_G can be written as

$$\Gamma_G u(t) = \int_{-\infty}^{0} Ce^{A(t-\tau)}Bu(\tau)d\tau, \quad \text{for } \ t \geq 0 \qquad (39)$$

The Hankel operator Γ_G maps past inputs to future outputs. Expression (38) shows that the Hankel operator Γ_G is an integral operator mapping $L^2(-\infty,\ 0]$ into $L^2[0,\ \infty)$, with kernel the impulse response $k(t,\ \tau)$ defined by

$$k(t,\ \tau) := Ce^{A(t-\tau)}B, \quad \tau < 0, \quad t \geq 0 \qquad (40)$$

Balanced truncation is commonly thought to be a model reduction technique that is not optimal in any sense [23]. We show that this is not the case, and in fact balanced truncation is indeed optimal in the sense of the Hilbert-Schmidt norm. The techniques we use are reminiscent of the previous section and guarantee for the optimum to be a Hankel operator. This contrasts, for example, with the minimization in various norms addressed in [24], [26]. To see this note that the Hankel operator Γ_G has finite rank $k \leq n$ [15], and therefore belongs to the Hilbert-Schmidt class of operators acting from $L^2(-\infty,\ 0]$ into $L^2[0,\ \infty)$. Let its spectral factorization be given by

$$\Gamma_G = \sum_{i=1}^{n} \sigma_i \chi_i \otimes \zeta_i,\ \chi_i \in L^2(-\infty,\ 0],\ \zeta_i \in L^2[0,\ \infty) \ (41)$$

where σ_i are the Hankel singular values of the system G ordered in decreasing order, i.e., $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{n-1} \geq \sigma_n$, and $\{\chi_i\}_1^n$ and $\{\zeta_i\}_1^n$ are orthonormal sets in $L^2(-\infty,\ 0]$ and $L^2[0,\ \infty)$, respectively. Next, consider the optimal distance minimization

$$\mu_{n_r} := \min_{n_r < k} \|\Gamma_G - \Gamma_{G_r}\|_{\text{HS}} \qquad (42)$$

where Γ_{G_r} is an operator acting from $L^2(-\infty,\ 0]$ into $L^2[0,\ \infty)$ of rank $n_r < n$. An application of identities (27) and (28) to the minimization (42) yields the *unique* optimum (since the distance minimization is posed in a Hilbert space)

$$\Gamma_{G_r} = \sum_{i=1}^{n_r} \sigma_i \chi_i \otimes \zeta_i \qquad (43)$$

and the shortest distance

$$\mu_{n_r} = \left\| \sum_{i=n_r+1}^{n} \sigma_i \chi_i \otimes \zeta_i \right\|_{HS} = \Big(\sum_{n_r+1}^{n} \sigma_i^2 \Big)^{\frac{1}{2}} \qquad (44)$$

The operator Γ_{G_r} is not necessarily a Hankel operator, however, we will show that starting from a specific balanced realization for the original system, the minimizing operator can be chosen to be a Hankel operator corresponding to the reduced order model. To do so let $\Gamma_G = U_G\big(\Gamma_G^\star \Gamma_G\big)^{\frac{1}{2}}$ be a polar decomposition of Γ_G, applying (14) and (15) to Γ_G the vectors χ_i and ζ_i satisfy

$$\begin{aligned} \Gamma_G\, \chi_i &= U_G(\Gamma_G^\star \Gamma_G)^{\frac{1}{2}} \chi_i = \sigma_i \zeta_i,\ i = 1, \cdots, n \quad (45) \\ \Gamma_G^\star\, \zeta_i &= U_G^\star(\Gamma_G \Gamma_G^\star)^{\frac{1}{2}} \zeta_i = \sigma_i \chi_i,\ i = 1, \cdots, n \quad (46) \end{aligned}$$

That is, χ_i and ζ_i form a Schmidt pair for Γ_G. In terms of the Schmidt pair (41) implies that the Hankel operator Γ_G can be expressed as

$$(\Gamma_G u)(t) = \sum_{i=1}^{n} \zeta_i(t) \sigma_i \int_{-\infty}^{0} \chi_i(\tau) u(\tau) d\tau \qquad (47)$$

We propose the following realization for the impulse response $k(t,\ \tau)$ given in (40), for $i,\ j = 1,\ 2,\ \cdots,\ n$,

$$\begin{aligned} \tilde{A} &= (a_{ij}) := \Big(\frac{\sigma_j}{\sigma_i}\Big)^{\frac{1}{2}} \int_{-\infty}^{0} \zeta_i^\star(\tau) \dot{\zeta}_j(\tau) d\tau \\ \tilde{B} &:= (\sqrt{\sigma_1}\chi_1(0),\ \sqrt{\sigma_2}\chi_2(0),\ \cdots,\ \sqrt{\sigma_n}\chi_n(0))^T \\ \tilde{C} &:= (\sqrt{\sigma_1}\zeta_1(0),\ \sqrt{\sigma_2}\zeta_2(0),\ \cdots,\ \sqrt{\sigma_n}\zeta_n(0)) \quad (48) \end{aligned}$$

The corresponding semi-group can be computed as

$$e^{\tilde{A}t} = \Big(\frac{\sigma_j}{\sigma_i}\Big)^{\frac{1}{2}} \int_{-\infty}^{0} \zeta_i^\star(\tau) \zeta_j(t-\tau) d\tau \qquad (49)$$

since

$$\lim_{t \longrightarrow 0} \frac{1}{t} \int_{-\infty}^{0} \zeta_i^\star(\tau) \zeta_j(\tau - t) d\tau = \int_{-\infty}^{0} \zeta_i^\star(\tau) \dot{\zeta}_j(\tau) d\tau \quad (50)$$

Define the controllability and observability operators denoted Ψ_c and Ψ_o, respectively by [15]

$$\Psi_c : L^2(-\infty,\ 0] \longmapsto \mathbb{R}^n,\ \Psi_c u := \int_{0}^{\infty} e^{\tilde{A}\tau} \tilde{B} u(\tau) d\tau$$
$$\Psi_o : \mathbb{R}^n \longmapsto L^2[0,\ \infty),\ \Psi_o x_0 := \tilde{C} e^{\tilde{A}t} x_0, \ \ t \geq 0$$

Note that [15] $\Gamma_G = \Psi_o \Psi_c$, and using the realization (48), we have

$$(\Psi_o \Psi_c u)(t) \ = \ \sum_{i=1}^{n} \sigma_i \zeta_i \int_{-\infty}^{0} \chi_i^\star(\tau) u(\tau) d\tau = (\Gamma_G u)(t)$$

and the observability gramian is given by

$$\begin{aligned}\Psi_o^\star\Psi_o &= \int_0^\infty e^{\tilde{A}^\star t}\tilde{C}^\star\tilde{C}e^{\tilde{A}t}dt \qquad (51)\\ &= \Big(\int_0^\infty \sum_{i=1}^n\sum_{j=1}^n \sqrt{\sigma_i}\sqrt{\sigma_j}\zeta_i^\star(t)\zeta_j(t)dt\Big)\\ &= (\sigma_i\delta_{ij}) = \Sigma = diag(\sigma_1,\ \sigma_2,\ \cdots,\ \sigma_n) \ (52)\end{aligned}$$

where δ_{ij} the usual Kronecker delta.

Similarly the controllability gramian $\Psi_c\Psi_c^\star = \Sigma$, and the realization $(\tilde{A}, \tilde{B}, \tilde{C})$ is therefore balanced. By the same token as POD using a similar expression as (25), define the Hankel operator corresponding to the n_r-th order model Γ_{Gn_r} as

$$\begin{aligned}(\Gamma_{G_{n_r}}u)(t) &= \sum_{i=1}^{n_r}\zeta_i(t)\sigma_i\int_{-\infty}^0 \chi_i(\tau)u(\tau)d\tau \qquad (53)\\ &= \int_{-\infty}^0 \tilde{C}P_r(P_re^{\tilde{A}(t-\tau)}P_r)P_r\tilde{B}d\tau \ (54)\end{aligned}$$

The last equality follows by (48), (49) and the fact that $P_r^2 = P_r$ (since P_r is a projection). Putting $\tilde{C}_r := \tilde{C}P_r$, $\tilde{B}_r := P_r\tilde{B}$ correspond to truncating $\tilde{C}$ and and $\tilde{B}$, respectively, and (50) implies that $P_re^{\tilde{A}(t-\tau)}P_r = e^{P_r\tilde{A}P_r(t-\tau)}$, and $\tilde{A}_r := P_r\tilde{A}P_r$ correspond to truncating the state space model $(\tilde{A},\ \tilde{B},\ \tilde{C})$ to n_r states, and the Hankel operator has rank n_r. Moreover,

$$\mu_{n_r} = \|\Gamma_G - \Gamma_{G_{n_r}}\|_{\text{HS}} = (\sum_{i=n_r+1}^n \sigma_i^2)^{\frac{1}{2}} \qquad (55)$$

By uniqueness of the minimizer in (42), expressions (53) and (55) imply that we must have $\Gamma_r \equiv \Gamma_{G_{n_r}}$.

In terms of kernel approximation, balanced truncation is a particular case of POD in the sense that the kernel we want to approximate is the impulse response of the system $k(t,\ \tau)$ defined in (42). The optimization index μ_{n_r} can then be written as in POD

$$\begin{aligned}\mu_{n_r}^2 &= \min\Big\{\int_0^\infty\int_{-\infty}^0 \Big|k(t,\ \tau) - \sum_{i=1}^{n_r} f_i(t)g_i(\tau)\Big|^2 d\tau dt :\\ &\qquad f_i \in L^2[0,\ \infty);\ g_i \in L^2(-\infty,\ 0]\Big\} \qquad (56)\\ &= \int_0^\infty\int_{-\infty}^0 \Big|k(t,\ \tau) - \tilde{C}_re^{\tilde{A}_r(t-\tau)}\tilde{B}_r\Big|^2 d\tau dt \qquad (57)\end{aligned}$$

Expressions (45) and (57) show that balanced truncation is optimal in the sense of optimal approximation in the Hilbert-Schmidt norm of the Hankel operator Γ_G, and optimal in the sense of the $\|\cdot\|_2$-norm of kernels corresponding to impulse responses of linear time-invariant systems defined over $[0,\ \infty) \times (-\infty,\ 0]$. The linear time-invariant system framework allows the exact computations of the optimal lower order model approximation. This contrasts with POD which uses simulation data and particular open-loop inputs to generate snapshots.

V. Geometric Interpretation

The eigenvalues λ_i's of $(T^\star T)^{\frac{1}{2}}$ (or singular values of T) defined in (12), and the Hankel singular values σ_i's of Γ_G have a geometric interpretation in terms of the computation of the n-widths of compact operators T and Γ_G that are defined on Hilbert spaces $L^2(T)$ and $L^2(-\infty,\ 0]$, respectively. In this section, we discuss the role of POD and balanced truncation in optimizing different n-widths defined in [22] (and references therein.)

We start by defining the Kolmogorov n-width of $T\big(L^2(\Omega)\big)$ into $L^2(h)$ as the optimization [22]

$$\begin{aligned}d_n\Big(T\big(L^2(\Omega)\big);\ L^2(h)\Big) = \inf_{X_n}\ &\sup_{\|f\|_2\le 1,\ f\in L^2(\Omega)}\\ &\inf_{g\in X_n}\|Tf - g\|_2 \qquad (58)\end{aligned}$$

where X_n is an n-dimensional subspace of $L^2(h)$.

The Kolmogorov n-width measures the extent to which the space $L^2(h)$ can be approximated by n-dimensional subspaces of $T\big(L^2(\Omega)\big)$, it is a measure of the "massivity" of $T\big(L^2(\Omega)\big)$. It represents the minimum representation error of $T\big(L^2(\Omega)\big)$ by the n-dimensional subspace X_n of $L^2(h)$. In other words, the Kolmogorov n-width quantifies the representation error due to innacurate representattion of the set $T\big(L^2(\Omega)\big)$: It represents the loss of information in the information processing stage. The n-width in the sense of Gel'fand, is defined as

$$d^n\Big(T\big(L^2(\Omega)\big);\ L^2(h)\Big) := \inf_{L^n}\ \sup_{\|f\|_2\le 1,\ f\in L^n}\|Tf\|_2 \qquad (59)$$

where the infimum is taken over all subspaces L^n of $T\big(L^2(\Omega)\big)$ of codimension at most n. If

$$d^n\Big(T\big(L^2(\Omega)\big);\ L^2(h)\Big) = \sup\{\|f\|_2\ :\ f \in T\big(L^2(\Omega)\big)\cap L^n\}$$

where L^n is a subspace of codimension at most n, then L^n is an optimal subspace for $d^n\Big(T\big(L^2(\Omega)\big);\ L^2(h)\Big)$. A subspace L^n is of codimension n if there exist n continuous linear functionals $\{f_i\}_{i=1}^n$ on $L^2(h)$ for which

$$L^n = \{g :\ g \in L^2(h),\ f_i(g) = 0,\ i = 1, 2, \ldots, n\} \qquad (60)$$

The Gel'fand n-width characterizes the experimental complexity of the information collecting stage using simulation or identification. It is related to the inherent error due to lack of data and inaccurate measurements. The inverse of the Gel'fand n-width gives the least number of measurements needed to reduce the modelling uncertainty to a predetermined value. The linear n-width is defined is defined by

$$\delta_n\big(T\big(L^2(\Omega)\big);\ L^2(h)\big) := \inf_{P_n}\ \sup_{\|\phi\|_2\le 1,\ \phi\in L^2(\Omega)}\|T\phi - P_n\phi\|_2$$

where P_n is any continuous linear operator from $L^2(\Omega)$ into $L^2(h)$ of rank at most n. Similar definitions for the Hankel operator range $\Gamma_G\big(L^2[0,\ -\infty)\big)$ hold. The basic results of this section are the following theorems which tell us that the different n-widths can be computed, and provide us with explicit optimal subspaces and operators [22].

Theorem 1: Let the operator T be defined as above, and let $\{\lambda_i\}$, $\{\alpha_i\}$, $\{\psi_i\}$ be defined as above. Then $d^n\Big(T\big(L^2(\Omega)\big);\quad L^2(h)\Big)= \quad d_n\Big(T\big(L^2(\Omega)\big);\quad L^2(h)\Big)= \delta_n\Big(T\big(L^2(\Omega)\big);\ L^2(h)\Big)= \lambda_{n+1},\ n = 0,\ 1,\ 2,\ \cdots$. Furthermore, the temporal coefficients $\{\alpha_i\}$ and POD basis $\{\psi_i\}$ are optimal for the n-widths in the following sense

i) the subspace spanned by the coefficients $\{\alpha_i\}$, $X_n = \text{Span}\{\alpha_1, \cdots, \alpha_n\}$, is optimal for $d_n\Big(T\big(L^2(\Omega)\big);\ L^2(h)\Big)$.

ii) the subspace $L^n = \Big\{\phi \in L^2(\Omega),\ <\phi,\ \psi_i>_1 = 0,\ i = 1, 2, \cdots, n\Big\}$ is optimal for $d^n\Big(T\big(L^2(\Omega)\big);\ L^2(h)\Big)$.

iii) the linear operator $P_n\phi = \sum_{i=1}^n <\phi,\ \psi_i>_1\ \psi_i$ is optimal for $\delta_n\big(T\big(L^2(\Omega)\big);\ L^2(h)\big)$.

A similar Theorem holds for the Hankel operator Γ_G and is stated next.

Theorem 2: Let the operator Γ_G be defined as above, and let $\{\sigma_i\}$, $\{\chi_i\}$, $\{\zeta_i\}$ be defined as above. Then $d^n\big(T\big(L^2(-\infty,\ 0]\big);\ L^2[0,\ \infty)\big) = d_n\big(T\big(L^2(-\infty,\ 0]\big);\ L^2[0,\ \infty)\big) = \delta_n\big(T\big(L^2(-\infty,\ 0]\big);\ L^2[0,\ \infty)\big)= \lambda_{n+1},\ n = 0,\ 1,\ 2,\ \cdots$. Furthermore, the temporal coefficients $\{\chi_i\}$ and POD basis $\{\zeta_i\}$ are optimal for the n-widths in the following sense

i) the subspace spanned by the vectors $\{\zeta_i\}$, $X_n = \text{Span}\{\zeta_1, \cdots, \zeta_n\}$, is optimal for $d_n\Big(\Gamma_G\big(L^2(-\infty, 0]\big);\ L^2[0,\ \infty)\big)$.

ii) the subspace $L^n = \Big\{\chi \in L^2(-\infty,\ 0],\ \int_{-\infty}^0 \chi(\tau)\chi_i(\tau)\, d\tau = 0,\ i = 1, 2, \cdots, n\Big\}$ is optimal for $d^n\big(\Gamma_G\big(L^2(-\infty,\ 0]\ \big);\ L^2[0,\ \infty)\big)$.

iii) the linear operator $Q_n\phi = \sum_{i=1}^n \int_{-\infty}^0 \phi(\tau)\chi_i(\tau)d\tau\chi_i$ is optimal for $\delta_n\big(\Gamma_G\big(L^2(-\infty,\ 0]\big);\ L^2[0,\ \infty)\big)$.

VI. Conclusion

In this paper, tools borrowed from the theory of operators were used to show that POD and balanced truncation are optimal in a precise sense. Optimality is quantified in terms of shortest distance minimizations, or optimal approximations by finite or lower rank Hilbert-Schmidt (integral) operators in Hilbert-Schmidt norms. The difference in the two model reduction techniques lies in the fact, that the optimizations occur in different integral operators defined on different L^2 spaces. However, both optimal approximations are posed in Hilbert operator spaces, i.e., the spaces of Hilbert-Schmidt operators, where the geometry is "nice" and the principle of orthogonality holds for both, allowing for the optimal approximations to be computed explicitly. Geometric interpretation of POD and balanced truncation in terms of optimizing the Kolmogorov, Gel'fand and linear n-widths is discussed. These n-widths quantify the inherent and representation errors generated in the information collecting and processing stages in simulation or identification.

References

[1] Atwell, J. and King, B., "Computational Aspects of Reduced Order Feedback Controllers for Spatially Distributed Systems," *Proceedings of the 38th IEEE Conference on Control and Desicion*, December 1999, pp. 4301-4306.

[2] P.D. Christophides, Nonlinear and Robust Control of PDE Systems, Birkhauser, 2000.

[3] Armaou, O.M. and Krstić, M., *Flow Control by Feedback: Stabilization and Feedback*, Springer, London, 2003

[4] Banks, H., del Rosario, R., and Smith, R., "Reduced Order Model Feedback Control Design: Numerical Implentation in a Thin Shell Model," Technical Report CRSC-TR98-27, Center for Research in Scientific Computation, North Carolina State University, June, 1998.

[5] K. Willcox and J. Peraire, " Balanced Model Reduction via the Proper Orthogonal Decomposition," *AIAA JOURNAL* Vol. 40, No. 11, November 2002, pp. 2323-2330

[6] Caraballo, E., Samimy, M., and DeBonis, J., "Low Dimensional Modeling of Flow for Closed-Loop Flow Control," AIAA Paper 2003-0059, January 2003.

[7] Carlson, H., Glauser, M., Higuchi, H., and Young, M., "POD Based Experimental Flow Control on a NACA-4412 Airfoil," AIAA Paper 2004-0575, January 2004.

[8] Carlson, H., Glauser, M., and Roveda, R., "Models for Controlling Airfoil Lift and Drag," AIAA Paper 2004-0579, January 2004.

[9] Cohen, K., Siegel, S., McLaughlin, T., and Myatt, J., "Proper Orthogonal Decomposition Modeling of a Controlled Ginzburg-Landau Cylinder Wake Model," AIAA Paper 2003-2405, January 2003.

[10] Ilak, M. and Rowley, C. W., "Reduced-Order Modeling of Channel Flow Using Traveling POD and Balanced POD,". *3rd AIAA Fow Control Conference*, San Francisico, June 2006

[11] Efe, M. and Ozbay, H., "Proper Orthogonal Decomposition for Reduced Order Modeling: 2D Heat Flow," *Proc. of 2003 IEEE Conference on Control Applications*, June 23-25, 2003, pp. 1273-1277.

[12] Camphouse, R.C., "Boundary Feedback Control Using Proper Orthogonal Decomposition Models," *Journal of Guidance, Control, and Dynamics*, Vol. 28, No. 5, September-October 2005, pp 931-938.

[13] Camphouse, R. C. and Myatt, J. H., "Reduced Order Modelling and Boundary Feedback Control of Nonlinear Convection," AIAA Paper 2005-5844, August 2005.

[14] R.C. Camphouse, S.M. Djouadi and J.H. Myatt, "Feedback Control for Aerodynamics", Proc. of the IEEE Conference on Decision and Control, 2006.

[15] K. Zhou, J.C. Doyle and K. Glover, Robust and Optimal Control, Prentice-Hall, 1996.

[16] J. Lumley, The Structures of Inhomogeneous Turbulent Flow, Atmospheric Turbulence and Radio Wave Propagation, edited by A. M. Yaglom and V.I. Tatarski, Nauka, Moscow, pp. 166-178, 1967.

[17] Holmes, P., Lumley, J., and Berkooz, G., *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*, Cambridge University Press, New York, 1996, pp. 86-127.

[18] Berkooz, G., Holmes, P., and Lumley, J., The Proper Orthogonal Decomposition in the Analysis of Turbulent Flows, Annual Review of Fluid Mechanics, Vol. 25, 1993, pp. 539-575.

[19] W.R. Graham, J. Peraire, and K.T. Yang, Optimal control of vortex shedding using low order models. Part I: Open-loop model development. Int. J. for Num. Methods in Eng. 44(7):973-990, 1999

[20] J. Borggaard, Optimal Reduced-Order Modeling for Nonlinear Distributed Parameter Systems, Proceedings of the American Control Conference, Minneapolis, pp. 1150-1154, 2006.

[21] F. Riesz and B.Sz.-Nagy, Functional Analysis, Dover, 1990.

[22] A. Pinkus, n-Widths in Approximation Theory, Springer-Verlag, 1985.

[23] Moore, B., Principal Component Analysis in Linear Systems: Controllability, Observability, and Model Reduction, IEEE Transactions on Automatic Control, Vol. AC-26, No. 1, 1981, pp. 17-31.

[24] L.M. Silverman and M. Bettayeb, Optimal Approximation of Linear Systems, Proceedings of the American Control Conference, 1981.

[25] Schatten R. *Norm Ideals of Completely Continuous Operators*, Springer-Verlag, Berlin, Gottingen, Heidelberg, 1960.

[26] K. Glover, R.F. Curtain and R. Partington, Realisation and Approximation of Linear Infinite Dimensional Systems with Erroe Bounds, SIAM J. Control and Optiz., vol. 26, No. 4, (1988), 863-898.

[27] L. Mirsky, Symmetric Gauge Functions and Unitarily Invariant Norms, Quart. J. Math (2), 11 (1960), 50-59.