Design of input for data-driven simulation with Hankel and Page matrices

Andrea Iannelli, Mingzhou Yin, Roy S. Smith

***Abstract*— The paper deals with the problem of designing informative input trajectories for data-driven simulation. First, the excitation requirements in the case of noise-free data are discussed and new weaker conditions, which assume the simulated input to be known in advance, are provided. Then, the case of noisy data trajectories is considered and an input design problem based on a recently proposed maximum likelihood estimator is formulated. A Bayesian interpretation is provided, and the implications of using Hankel and Page matrix representations are demonstrated. Numerical examples show the impact of the designed input on the predictive accuracy.**

I. Introduction

Predicting the response of a plant to given initial conditions and input signals is a fundamental task for analysis and control of dynamical systems. While this is a basic problem when a model of the plant is available (either derived from first principles or identified from experiments), increasing interest has been devoted to *direct* data-driven prediction methods, whereby future responses are expressed in terms of past (data) trajectories. The behavioral approach [1] has found undisputed success in recent years within this context. Building on the seminal work [2] providing conditions under which the subspace of trajectories of a linear system can be spanned by noise-free data matrices, extensive research has been done in this direction, especially with the goal of using this non-parametric description for control [3].

As the success of such data-driven controllers highly depends on the accuracy of the predictions, we focus here on the data-driven simulation problem for linear systems and consider two interrelated questions. The first, to which Section II is devoted, is concerned with characterizing the excitation requirements of the data trajectory. This problem was originally addressed in [4] by assuming persistence of excitation of the input and then applying the result from [2]. It is shown in this paper that, for a particular simulation task, the requirements can be greatly relaxed. While these results guarantee that the simulated output can be exactly reconstructed, they all require clean (or noise-free) data trajectories. Building on these weaker excitation requirements, the second contribution, presented in Section III, is a design procedure to choose the input data trajectory in order to maximize the accuracy of the simulated response when the data trajectory is contaminated by noise. The accuracy criterion is formulated from a Bayesian viewpoint by leveraging a maximum likelihood data-driven estimator [5] and the concept of mutual information [6].

We put emphasis, in both results, on the effect of using Hankel and Page matrix representations for the data trajectory. While Hankel matrices have a rich history in system identification [7], Page matrices, proposed in [8] in the context of realization algorithms with noisy Markov parameters, have received less attention. It was shown in [8] that in the Page matrix case, since there are no repeated entries, de-noising by thresholding the lowest singular values is provably optimal. This is not the case for Hankel matrices where low-rank approximations via the SVD implicitly attribute a non-uniform weight on the noisy Markov parameters. The advantages pertaining to the use of Page matrices as distributionally robust predictors were recently discussed in [9]. However, in the behavioural setting Page matrices are reportedly less accurate than Hankel ones, when the comparison is done with equal data length, owing to their less favourable sample efficiency [10]. Numerical results in Section IV show that, by using the proposed input design strategy, this aspect can be ameliorated and Page matrices do offer better predictive accuracy, as conceptually expected as they eliminate noise coupling among entries.

The input design problem in the data-driven setting has recently started receiving attention [11], [12]. In [11], an online procedure to select an input sequence which satisfies the conditions in [2] without requiring persistence of excitation is proposed in the case of clean data. The work in [12] considered the case of impulse response simulation with noisy Hankel matrices, and used the standard excitation requirements [4].

Notation and definitions

The mutual information [6] of two multivariate random variables x and y of size n is defined as

$$I(x;y) = H(x) + H(y) - H(x,y) = H(x) - H(x|y), \quad (1)$$

where $H(x) = -\int p(x)\log(p(x))dx$ is the Shannon's entropy of x. If x has Gaussian distribution with covariance Σ_{x}, it holds $H(x) = \frac{1}{2}\log(2\pi e)^n + \frac{1}{2}\log(\det(\Sigma_{\mathrm{x}}))$.

Given a matrix $X \in \mathbb{R}^{n\times n}$, $X^\dagger$ denotes its pseudo-inverse and $\{\lambda_i(X), i = 1, ..., n\}$ the set of its eigenvalues. The symbol I_n denotes the identity matrix of size n.

This work is supported by the Swiss National Science Foundation under grant no. 200021_178890.
The authors are with the Department of Electrical Engineering, Automatic Control Lab, ETH, Zürich 8092, Switzerland iannelli/myin/rsmith@control.ee.ethz.ch

Given a signal $z \in \mathbb{R}^{n_z}$, we use $z_{[i,j]}$ to denote both the concatenated vector $\left[z_i^\top \ ... \ z_j^\top\right]^\top$ and the sequence $\{z_k\}_{k=i}^{j}$ of length $N = j-i+1$. Given $z_{[i,j]}$, the associated block Hankel and Page matrices with L block rows are defined respectively as

$$\mathcal{H}_L(z_{[i,j]}) := \begin{bmatrix} z_i & z_{i+1} & \cdots & z_{N+i-L} \\ z_{i+1} & z_{i+2} & \cdots & z_{N+i-L+1} \\ \vdots & \vdots & & \vdots \\ z_{i+L-1} & z_{i+L} & \cdots & z_{N+i-1} \end{bmatrix},$$

$$\mathcal{P}_L(z_{[i,j]}) := \begin{bmatrix} z_i & z_{i+L} & \cdots & z_{i+\lfloor \frac{N}{L} \rfloor L-L} \\ z_{i+1} & z_{i+L+1} & \cdots & z_{i+\lfloor \frac{N}{L} \rfloor L-L+1} \\ \vdots & \vdots & & \vdots \\ z_{i+L-1} & z_{i+2L-1} & \cdots & z_{i+\lfloor \frac{N}{L} \rfloor L-1} \end{bmatrix}.$$

where $\mathcal{H}_L(z_{[i,j]}) \in \mathbb{R}^{Ln_z \times (N-L+1)}$ and $\mathcal{P}_L(z_{[i,j]}) \in \mathbb{R}^{Ln_z \times \lfloor \frac{N}{L} \rfloor}$. The signal $z_{[i,j]}$ is persistently exciting (PE) of order L if $\mathcal{H}_L(z_{[i,j]})$ has full row rank [2] and L-Page exciting of order M if the matrix $\begin{bmatrix} \mathcal{P}_L(z_{[i,i+\lfloor \frac{N}{L} \rfloor L-1-(M-1)L]}) \\ \mathcal{P}_L(z_{[L+i,i+\lfloor \frac{N}{L} \rfloor L-1-(M-2)L]}) \\ ... \\ \mathcal{P}_L(z_{[L(M-1)+i,i+\lfloor \frac{N}{L} \rfloor L-1]}) \end{bmatrix}$ has full row rank [9].

Given a state-space model (A, B, C, D), the block Toeplitz matrix of impulse response coefficients, the extended observability matrix, and the reversed extended controllability matrix are defined respectively as

$$\mathcal{T}_i = \begin{bmatrix} D & 0 & 0 & 0 \\ CB & D & 0 & 0 \\ \vdots & \ddots & D & 0 \\ CA^{i-2}B & CA^{i-3}B & \cdots & D \end{bmatrix},$$

$$\mathcal{O}_i = \left[C^\top \ \ (CA)^\top \ ... \ (CA^{i-1})^\top\right]^\top,$$

$$\mathcal{C}_i = \left[A^{i-1}B \ ... \ AB \ \ B\right].$$

II. Data-driven simulation

A. Problem setting and available results

Consider a linear time-invariant (LTI) system

$$x_{t+1} = Ax_t + Bu_t, \tag{2a}$$

$$y_t = Cx_t + Du_t, \tag{2b}$$

where $x \in \mathbb{R}^{n_x}$ is the state, $u \in \mathbb{R}^{n_u}$ is the input, and $y \in \mathbb{R}^{n_y}$ is the output. It is assumed that (2) is minimal, i.e. the system is controllable, observable, and n_x is its McMillan degree. The lag l is defined as the smallest integer i such that $\mathcal{O}_i$ has rank n_x. The data-driven (or model-free) simulation problem for (2) is stated next.

Problem 1: Given an input-output *data* trajectory $(u_{\mathrm{d}\,[0,N-1]}, y_{\mathrm{d}\,[0,N-1]})$, an input *simulation* trajectory $u_{\mathrm{s}\,[0,L_s-1]}$, and an initial condition x_{ini}, find the (unique) output *simulation* trajectory $y_{\mathrm{s}\,[0,L_s-1]}$.

It was shown in [4] that Problem 1 can be solved in a fully input-output (or representation-free) setting by leveraging results from behavioral system theory [1]. The first observation is that an input-output *initial* trajectory $(u_{\mathrm{ini}\,[0,L_0-1]}, y_{\mathrm{ini}\,[0,L_0-1]})$ of (2) can be used, if $L_0 \geq l$, to uniquely define the initial condition x_{ini} of Problem 1. Define $L = L_0 + L_s$.

Lemma 1: [4] Given $(u_{\mathrm{ini}\,[0,L_0-1]}, y_{\mathrm{ini}\,[0,L_0-1]})$, with $L_0 \geq l$, $(u_{\mathrm{d}\,[0,N-1]}, y_{\mathrm{d}\,[0,N-1]})$, and $u_{\mathrm{s}\,[0,L_s-1]}$. Assume that the data generating system is controllable and that $u_{\mathrm{d}\,[0,N-1]}$ is persistently exciting of order $L + n_x$. Partition Hankel matrices built using the data trajectory according to the indices L_0 (p) and L_s (f) as follows

$$\left[\begin{array}{c} U \\ \hline Y \end{array}\right] = \left[\begin{array}{c} U_{\mathrm{p}} \\ U_{\mathrm{f}} \\ \hline Y_{\mathrm{p}} \\ Y_{\mathrm{f}} \end{array}\right] := \left[\begin{array}{c} \mathcal{H}_L(u_{\mathrm{d}\,[0,N-1]}) \\ \hline \mathcal{H}_L(y_{\mathrm{d}\,[0,N-1]}) \end{array}\right]. \tag{3}$$

Then, $y_{\mathrm{s}\,[0,L_s-1]} = Y_{\mathrm{f}} g$, where g satisfies

$$\begin{bmatrix} u_{\mathrm{ini}\,[0,L_0-1]} \\ y_{\mathrm{ini}\,[0,L_0-1]} \\ u_{\mathrm{s}\,[0,L_s-1]} \end{bmatrix} = \begin{bmatrix} U_{\mathrm{p}} \\ Y_{\mathrm{p}} \\ U_{\mathrm{f}} \end{bmatrix} g. \tag{4}$$

The proof is an application of the so-called *Fundamental Lemma* (FL) [2], which guarantees, under the assumptions of Lemma 1, that any trajectory generated by (2) is spanned by $\left[\begin{array}{c} \mathcal{H}_L(u_{\mathrm{d}\,[0,N-1]}) \\ \hline \mathcal{H}_L(y_{\mathrm{d}\,[0,N-1]}) \end{array}\right]$. Uniqueness of the output *simulation* trajectory is guaranteed by the fact that x_{ini} is uniquely defined by an input-output trajectory of appropriate length. Lemma 1 thus provides an answer to Problem 1 by virtue of Hankel data matrices that characterize the system's behavior. It has been recently shown in [9] that Page data matrix representations can be used as well to span the system's trajectories. A straightforward application of the results from [9] to the same arguments used in [4] yields the following result on data-driven simulation with Page matrices.

Lemma 2: Given $(u_{\mathrm{ini}\,[0,L_0-1]}, y_{\mathrm{ini}\,[0,L_0-1]})$, with $L_0 \geq l$, $(u_{\mathrm{d}\,[0,N-1]}, y_{\mathrm{d}\,[0,N-1]})$, and $u_{\mathrm{s}\,[0,L_s-1]}$. Assume that the data generating system is controllable and that $u_{\mathrm{d}\,[0,N-1]}$ is L-Page exciting of order n_x+1. Partition Page matrices built using the data trajectory according to the indices L_0 (p) and L_s (f) as follows

$$\left[\begin{array}{c} U \\ \hline Y \end{array}\right] = \left[\begin{array}{c} U_{\mathrm{p}} \\ U_{\mathrm{f}} \\ \hline Y_{\mathrm{p}} \\ Y_{\mathrm{f}} \end{array}\right] := \left[\begin{array}{c} \mathcal{P}_L(u_{\mathrm{d}\,[0,N-1]}) \\ \hline \mathcal{P}_L(y_{\mathrm{d}\,[0,N-1]}) \end{array}\right]. \tag{5}$$

Then, $y_{\mathrm{s}\,[0,L_s-1]} = Y_{\mathrm{f}} g$, where g satisfies

$$\begin{bmatrix} u_{\mathrm{ini}\,[0,L_0-1]} \\ y_{\mathrm{ini}\,[0,L_0-1]} \\ u_{\mathrm{s}\,[0,L_s-1]} \end{bmatrix} = \begin{bmatrix} U_{\mathrm{p}} \\ Y_{\mathrm{p}} \\ U_{\mathrm{f}} \end{bmatrix} g. \tag{6}$$

While, by uniqueness, they provide the same output simulation trajectory, the Hankel and Page matrices formulations are markedly different as far as data length requirements are concerned, because of the different excitation requirements. To guarantee the conditions of Lemma 1, it must hold $N \geq (L + n_x)(n_u + 1) - 1$. Instead, one needs $N \geq L((n_u L + 1)(n_x + 1) - 1)$

to apply Lemma 2. This of course implies that the Page formulation typically requires a much longer data set, with potential negative effects on experimental and computational costs. Another more subtle consequence is related to the fact that, for a given N, the number of columns c_H of a Hankel matrix will always be larger than the number of columns c_P of a Page matrix. Specifically, $L(c_H(N) - c_P(N)) = N(L-1) - L^2 + L$, thus this gap increases linearly with N. The implication is that, in the range of N for which the two formulations can be compared (i.e. $N \geq L((n_u L + 1)(n_x + 1) - 1)$), $c_H \gg c_P$. This represents a strong disadvantage of the Page matrix when working with noisy data trajectories, since the effect of noise can be averaged by increasing the number of columns of the data matrices. This fact plays a decisive role when comparing Hankel and Page predictive performance with noisy data, as recently observed in [10].

B. Weaker excitation conditions

The excitation conditions prescribed by Lemmas 1 and 2 are required because the respective data matrices are used to solve the simulation problem for *any* input simulation trajectory and initial condition. However, when these are known in advance, the requirements on the data trajectory can be significantly relaxed, as shown next.

As a preamble, it is recalled that a necessary and sufficient condition for a generic input-output trajectory $(u_{[0,T-1]}, y_{[0,T-1]})$ to be generated by (2) is that there exists $x_0 \in \mathbb{R}^{n_x}$ such that

$$y_{[0,T-1]} = \mathcal{O}_T x_0 + \mathcal{T}_T u_{[0,T-1]}, \tag{7}$$

as can be shown by writing (2b) for $t \in [0, T-1]$. Since by assumption $(u_{\text{ini }[0,L_0-1]}, y_{\text{ini }[0,L_0-1]})$ is a trajectory of (2), it satisfies (7), and we denote by $\hat{x}$ the initial condition of the state for which (7) is verified. We also denote by $x_{\text{d }[0,N-1]}$ the state trajectory associated with the input-output data trajectory.

Lemma 3: Given $(u_{\text{ini }[0,L_0-1]}, y_{\text{ini }[0,L_0-1]})$, with $L_0 \geq l$, $(u_{\text{d }[0,N-1]}, y_{\text{d }[0,N-1]})$, and $u_{\text{s }[0,L_s-1]}$. Partition Hankel matrices as in (3). Assume that

$$\left[\begin{array}{c} u_{\text{ini }[0,L_0-1]} \\ u_{\text{s }[0,L_s-1]} \\ \hline \hat{x} \end{array}\right] \in \text{Im}\left(\left[\begin{array}{c} U_\text{p} \\ U_\text{f} \\ \hline X_\text{p} \end{array}\right]\right), \tag{8}$$

where $X_\text{p} = [x_{\text{d},0}\ x_{\text{d},1}\ \dots\ x_{\text{d},N-L}]$.

Then, $y_{\text{s }[0,L_s-1]} = Y_\text{f} g$, where g satisfies

$$\begin{bmatrix} u_{\text{ini }[0,L_0-1]} \\ y_{\text{ini }[0,L_0-1]} \\ u_{\text{s }[0,L_s-1]} \end{bmatrix} = \begin{bmatrix} U_\text{p} \\ Y_\text{p} \\ U_\text{f} \end{bmatrix} g. \tag{9}$$

Proof: Specializing (7) to the initial trajectory gives

$$y_{\text{ini }[0,L_0-1]} = \mathcal{O}_{L_0}\hat{x} + \mathcal{T}_{L_0} u_{\text{ini }[0,L_0-1]}, \tag{10a}$$

$$\hat{x} = \mathcal{O}_{L_0}^\dagger y_{\text{ini }[0,L_0-1]} - \mathcal{O}_{L_0}^\dagger \mathcal{T}_{L_0} u_{\text{ini }[0,L_0-1]}, \tag{10b}$$

where the vector $\hat{x}$ is unique, since by assumption $L_0 \geq l$ and thus $\mathcal{O}_{L_0}$ has full rank. Using (8), we conclude from (10) that $\begin{bmatrix} u_{\text{ini }[0,L_0-1]} \\ y_{\text{ini }[0,L_0-1]} \\ u_{\text{s }[0,L_s-1]} \end{bmatrix} \in \text{Im}\left(\begin{bmatrix} U_\text{p} \\ \mathcal{O}_{L_0} X_\text{p} + \mathcal{T}_{L_0} U_\text{p} \\ U_\text{f} \end{bmatrix}\right)$. Moreover, it can be shown starting from (2), see e.g. the subspace identification literature [7], that the data trajectory $(u_{\text{d }[0,N-1]}, y_{\text{d }[0,N-1]})$ satisfies the matrix equation $Y_\text{p} = \mathcal{O}_{L_0} X_\text{p} + \mathcal{T}_{L_0} U_\text{p}$. Therefore

$$\begin{bmatrix} u_{\text{ini }[0,L_0-1]} \\ y_{\text{ini }[0,L_0-1]} \\ u_{\text{s }[0,L_s-1]} \end{bmatrix} \in \text{Im}\left(\begin{bmatrix} U_\text{p} \\ Y_\text{p} \\ U_\text{f} \end{bmatrix}\right). \tag{11}$$

Consider now (7) for the simulation trajectory

$$y_{\text{s }[0,L_s-1]} = \mathcal{O}_{L_s} x_\text{ini} + \mathcal{T}_{L_s} u_{\text{s }[0,L_s-1]}, \tag{12}$$

where, to obtain x_ini, write (2a) in the interval $[0, L_0]$

$$x_\text{ini} = x_{L_0} = A^{L_0}\hat{x} + \mathcal{C}_{L_0} u_{\text{ini }[0,L_0-1]}. \tag{13}$$

Because $\hat{x}$ is unique, x_ini is uniquely defined. Substituting the explicit expression of $\hat{x}$ (10b) in (13) yields

$$x_\text{ini} = \underbrace{\left(\mathcal{C}_{L_0} - A^{L_0}\mathcal{O}_{L_0}^\dagger \mathcal{T}_{L_0}\right)}_{P} u_\text{ini} + \underbrace{A^{L_0}\mathcal{O}_{L_0}^\dagger}_{Q} y_\text{ini}.$$

Rewrite (12) equivalently as

$$\begin{aligned} y_{\text{s }[0,L_s-1]} &= \begin{bmatrix} \mathcal{O}_{L_s}P & \mathcal{O}_{L_s}Q & \mathcal{T}_{L_s} \end{bmatrix} \begin{bmatrix} u_{\text{ini }[0,L_0-1]} \\ y_{\text{ini }[0,L_0-1]} \\ u_{\text{s }[0,L_s-1]} \end{bmatrix}, \\ &= \underbrace{\begin{bmatrix} \mathcal{O}_{L_s}P & \mathcal{O}_{L_s}Q & \mathcal{T}_{L_s} \end{bmatrix} \begin{bmatrix} U_\text{p} \\ Y_\text{p} \\ U_\text{f} \end{bmatrix}}_{K} g, \end{aligned} \tag{14}$$

where the existence of the vector g in the last equality is guaranteed by (11). Because (12) uniquely defines $y_{\text{s }[0,L_s-1]}$, showing $K = Y_f$ proves the statement. To see this, note that writing (2) in matrix form for the data trajectory $(u_{\text{d }[0,N-1]}, y_{\text{d }[0,N-1]})$ also yields $Y_\text{f} = \mathcal{O}_{L_s} X_\text{f} + \mathcal{T}_{L_s} U_\text{f}$, where $X_\text{f} = [x_{\text{d},L_0}\ x_{\text{d},L_0+1} \dots\ x_{\text{d},L_0+N-L}]$. Note also that, by their definition, P and Q predict the current state given previous input and output sequences. This of course also holds for the data trajectory, and thus we have that $X_\text{f} = \begin{bmatrix} P & Q \end{bmatrix} \begin{bmatrix} U_p \\ Y_p \end{bmatrix}$. Therefore

$$K = \mathcal{O}_{L_s} \begin{bmatrix} P & Q \end{bmatrix} \begin{bmatrix} U_\text{p} \\ Y_\text{p} \end{bmatrix} + \mathcal{T}_{L_s} U_\text{f} = Y_\text{f}, \tag{15}$$

which concludes the proof. ■

The result is proved for Hankel data matrices, but it can be seamlessly extended to Page matrices by simply changing the notation. That is, define $X_\text{p} = [x_{\text{d},0}\ \ x_{\text{d},L} \dots\ \ x_{\text{d},(\lfloor \frac{N}{L} \rfloor - 1)L}]$ and $X_\text{f} = [x_{\text{d},L_0}\ x_{\text{d},L_0+L} \dots x_{\text{d},L_0+(\lfloor \frac{N}{L} \rfloor - 1)L}]$, and replace partition (3) by (5). Notably, the excitation requirements do not change between the two data representations, in contrast to the previous results discussed in Section II-A.

Remark 1: Equation 8 can in principle be satisfied with rank 1 data matrices, i.e. the minimum number of

columns is 1 and the minimum length N of the data trajectory is L. In this case, the range condition on $\hat{x}$ is satisfied if and only if the initial conditions of the initial and data trajectories coincide up to a scaling factor, while for $N > L$ this is only a sufficient condition. Note also that the assumption on $\hat{x}$ in (8) is only required to guarantee (11). The latter is a condition that can be checked directly from the data, and can thus be used in lieu of (8) to define the set of $y_{\text{ini}\,[0,L_0-1]}$ which can be simulated given an input trajectory satisfying the excitation conditions in (8). For controllable systems this is in principle w.l.o.g. because, as also shown in (13), any x_{ini} can be generated by $u_{\text{ini}\,[0,L_0-1]}$ alone.

III. Input design for simulation with noisy data

The main question to be addressed is: What is the best input data trajectory $u_{\text{d}\,[0,N-1]}$ to solve Problem 1? Lemmas 1, 2, and 3 are already experiment design results, as also recognized in [11] with respect to the Fundamental Lemma, because they provide both the minimum experiment length and the required signal properties by characterizing the span of the associated data matrices. In fact, when the initial condition and the input simulation trajectory are set in advance, Lemma 3 gives the less restrictive excitation requirements. However, all of these results rely on the assumption of working with clean data. If the data trajectory is contaminated with noise, computing exactly the output simulation trajectory is not possible as these results no longer hold. A possible strategy to make use of these methodologies while reducing the effect of the noise was presented in [5], under the name of signal matrix model (SMM). This is reviewed next, with an emphasis on some novel aspects, since it will be used later to frame the input design problem.

For the sake of readability, the time indexes are dropped from the sequences' subscripts. The same time indexes employed in the previous section apply for the respective trajectories (e.g. y_{s} will denote $y_{\text{s}\,[0,L_s-1]}$).

A. The SMM estimator

The signal matrix model is a maximum likelihood estimator of the vector g introduced in the previous lemmas. This was proposed in [5] in conjunction with Lemma 1, and, under its assumptions and Hankel data matrices partitions, provided a predicted output simulation trajectory $y_{\text{s}} = Y_{\text{f}} g_{\text{SMM}}$ with favourable statistical properties.

In the same spirit, a maximum likelihood estimator g_{SMM} can be defined building on Lemma 3. We consider the case where the output data trajectory ($n_y = 1$ is assumed here for simplicity of representation) is subject to i.i.d. Gaussian noise

$$\tilde{y}_{\text{d},i} = y_{\text{d},i} + w_i, \;\; w_i \sim \mathcal{N}(0, \sigma^2), \;\; i = 1, ..., N-1. \quad (16)$$

The rest of the problem's data (u_{d}, u_{ini}, and y_{ini}) are assumed to be noise-free because either they will be optimized over later (i.e. the input data trajectory) or are fixed by the analyst (i.e. the initial trajectory).

Lemma 4: Given $(u_{\text{ini}}, y_{\text{ini}})$, with $L_0 \geq l$, $(u_{\text{d}}, \tilde{y}_{\text{d}})$, and u_{s} satisfying the assumptions of Lemma 3. Partition the Hankel matrices as in (3). Define the random variable $\bar{y} := Yg - \left[\begin{smallmatrix} y_{\text{ini}} \\ 0 \end{smallmatrix}\right]$ where the usual partition applies. The value of g that maximizes $\mathbb{E}(\bar{y}|g)$, i.e. the conditional probability of observing the realization $\bar{y}$ corresponding to the available data given g, is given by

$$\min_{g \in \mathcal{G}} \operatorname{logdet}(\Sigma_{\text{y}}(g)) + \begin{bmatrix} Y_{\text{p}} g - y_{\text{ini}} \\ 0 \end{bmatrix}^{\top} \Sigma_{\text{y}}^{-1}(g) \begin{bmatrix} Y_{\text{p}} g - y_{\text{ini}} \\ 0 \end{bmatrix}, \quad (17)$$

where $\mathcal{G}$ is the set

$$\mathcal{G} = \left\{ g \in \mathbb{R}^{N-L+1} \,\middle|\, \begin{bmatrix} U_{\text{p}} \\ U_{\text{f}} \end{bmatrix} g = \begin{bmatrix} u_{\text{ini}} \\ u_{\text{s}} \end{bmatrix} \right\}, \quad (18)$$

and

$$(\Sigma_{\text{y}})_{i,j} = (\operatorname{cov}(\bar{y}|g))_{i,j} = \sigma^2 \sum_{k=1}^{N-L+1-|i-j|} g_k g_{k+|i-j|}. \quad (19)$$

The maximum likelihood simulation trajectory is then

$$\hat{y}_{\text{s,SMM}} = Y_{\text{f}} g_{\text{SMM}}, \quad (20)$$

where $g_{\text{SMM}} \in \arg\min(17)$.

Lemma 4 combines the maximum likelihood formulation from [5], which leads to problem (17), with the result in Lemma 3. The SMM estimator can also be postulated for Page matrix representations. Besides the notational difference of replacing partition (3) by (5), an important distinction is that for Page matrices

$$\Sigma_{\text{y}} = \operatorname{cov}(\bar{y}|g) = \sigma^2 \|g\|_2^2 I_L.$$

That is, the covariance matrix appearing in the optimization problem (17) is diagonal. This comes from the fact that $\operatorname{cov}(\bar{y}|g) = \left(g^{\top} \otimes I_L\right) \operatorname{cov}\left(\operatorname{vec}(Y)\right) (g \otimes I_L)$. Due to the absence of repeated entries in the Page matrix (5), $\operatorname{vec}(Y)$ is a vector of uncorrelated random variables with covariance $\sigma^2 I_N$. This is of course not the case when the structured Hankel matrix is used, due to the repetitions in each column, which leads to the banded structure in (19). Nonetheless, setting to zero the off-diagonal terms simplifies the solution of problem (17), thus the approximation to a diagonal Σ_{y} was proposed in [5] even when working with Hankel matrices.

B. Input design for SMM

In [12] the input design problem for identification of the truncated infinite impulse response using the SMM estimator with Hankel matrices was investigated. The mean-square error (MSE) matrix [13] of the estimated response was chosen to measure the accuracy of the estimates. The input design problem was then formulated as the minimization of A-, D-, and E- optimality criteria. The main finding was that, if the off-diagonal entries of Σ_{y} are neglected, minimizing these criteria is equivalent to minimizing the Euclidean norm of g_{SMM}.

As observed earlier, this is true only for Page matrices and not in general for Hankel matrices. In the latter case, minimization of the Euclidean norm of g_{SMM} is justified from an A-optimality viewpoint, since this consists of minimizing the trace of the MSE matrix. Inspired by the recent work in [14] concerning experiment design using tools from information theory, we provide here a Bayesian formulation of the SMM input design problem and show the implications of Page and Hankel matrix representations.

1) Bayesian perspective: The data-based SMM output simulation trajectory can be modelled as a Gaussian random variable (the subscript SMM will be dropped)

$$y_{\mathrm{s}}^{\mathrm{D}} \sim \mathcal{N}(\hat{y}_{\mathrm{s}}, \Sigma_{\mathrm{y,f}}), \tag{21}$$

where $\hat{y}_{\mathrm{s}}$ is given in (20) and $\Sigma_{\mathrm{y,f}}$ is the matrix made of the last L_s rows and columns of Σ_{y} (evaluated at g). Assume that prior knowledge on y_{s} is encoded in a positive definite kernel matrix Σ_{K}, resulting in a prior distribution $y_{\mathrm{s}}^0 \sim \mathcal{N}(0, \Sigma_{\mathrm{K}})$. When the simulation problem consists of estimating the first L_s coefficients of the truncated infinite impulse response [15], one can refer to an extensive literature on choices of kernel matrices encoding priors related to system's theoretic properties [16], [17]. However, the idea of using priors to improve on and regularize the estimate from data (21) can in principle be used for other simulation problems as well. Examples of priors might include smoothness and decay rate of the response. We can then, in the spirit of Kalman filtering, combine prior and data-based estimates to provide the MSE estimate. This can be interpreted as the posterior distribution of y_{s} given the data trajectory

$$(y_{\mathrm{s}}|(u_{\mathrm{d}}, y_{\mathrm{d}})) \sim \mathcal{N}(K\hat{y}_{\mathrm{s}}, \Sigma_{\mathrm{post}}), \tag{22}$$

where $K = \Sigma_{\mathrm{K}} \left(\Sigma_{\mathrm{K}} + \Sigma_{\mathrm{y,f}}\right)^{-1}$ is the Kalman gain and $\Sigma_{\mathrm{post}} = \Sigma_{\mathrm{K}} - \Sigma_{\mathrm{K}} \left(\Sigma_{\mathrm{K}} + \Sigma_{\mathrm{y,f}}\right)^{-1} \Sigma_{\mathrm{K}}$ is the posterior covariance. These expressions were already presented in [5] and are standard filtering relationships [18]. The novelty here is their interpretation in the data-driven simulation setting. This is important as it enables the formal definition of the input design problem as the maximization of the distance between the prior and posterior distributions of y_{s}. By using an information theoretic result [6], the expected value of the KL divergence between these distributions coincides with the mutual information of y_{s} and the data $(u_{\mathrm{d}}, y_{\mathrm{d}})$. Therefore, maximizing $I(y_{\mathrm{s}}; (u_{\mathrm{d}}, y_{\mathrm{d}}))$ yields, from a Bayesian viewpoint, an informative experiment.

Using its definition (1), the mutual information for the case of interest can be defined as

$$\begin{aligned} I(y_{\mathrm{s}}; (u_{\mathrm{d}}, y_{\mathrm{d}})) &= H(y_{\mathrm{s}}) - H(y_{\mathrm{s}}|(u_{\mathrm{d}}, y_{\mathrm{d}})), \\ &= \frac{1}{2}\left(\log(\det(\Sigma_{\mathrm{K}})) - \log(\det(\Sigma_{\mathrm{post}}))\right), && \text{(23a)} \\ &= \frac{1}{2}\log(\det(I_{L_s} + \Sigma_{\mathrm{K}}\Sigma_{\mathrm{y,f}}^{-1})), && \text{(23b)} \end{aligned}$$

where (23b) comes from the fact that

$$\begin{aligned} \Sigma_{\mathrm{post}} &= \Sigma_{\mathrm{K}} - \left(\Sigma_{\mathrm{K}}\Sigma_{\mathrm{y,f}}^{-1} + I_{L_s}\right)^{-1} \Sigma_{\mathrm{K}}\Sigma_{\mathrm{y,f}}^{-1}\Sigma_{\mathrm{K}}, \\ &= \Sigma_{\mathrm{K}} - \left(\Sigma_{\mathrm{y,f}}^{-1} + \Sigma_{\mathrm{K}}^{-1}\right)^{-1} \Sigma_{\mathrm{y,f}}^{-1}\Sigma_{\mathrm{K}}, \\ &= \left(\Sigma_{\mathrm{y,f}}^{-1} + \Sigma_{\mathrm{K}}^{-1}\right)^{-1} \left[\left(\Sigma_{\mathrm{y,f}}^{-1} + \Sigma_{\mathrm{K}}^{-1}\right)\Sigma_{\mathrm{K}} - \Sigma_{\mathrm{y,f}}^{-1}\Sigma_{\mathrm{K}}\right], \\ &= \left(\Sigma_{\mathrm{y,f}}^{-1} + \Sigma_{\mathrm{K}}^{-1}\right)^{-1}. \end{aligned}$$

The following result shows an important relationship between the mutual information and g.

Lemma 5: If $\Sigma_{\mathrm{y,f}} = \sigma^2 \|g\|_2^2 I_{L_s}$, then there exist functions f and h such that

$$I(y_{\mathrm{s}}; (u_{\mathrm{d}}, y_{\mathrm{d}})) = f(\|g\|_2^2, \Sigma_{\mathrm{K}}) + h(\Sigma_{\mathrm{K}}),$$

where $f(\cdot,\cdot)$ is monotonically decreasing with respect to the first argument irrespective of the second, and h does not depend on g.

Proof: Substituting $\Sigma_{\mathrm{y,f}} = \sigma^2 \|g\|_2^2 I_{L_s}$ in (23a), the mutual information can be written as

$$\begin{aligned} I &= \frac{1}{2}\log\left(\det\left(\frac{1}{\sigma^2\|g\|_2^2} I_{L_s} + \Sigma_{\mathrm{K}}^{-1}\right)\right) + \underbrace{\frac{1}{2}\log(\det(\Sigma_{\mathrm{K}}))}_{h(\Sigma_{\mathrm{K}})}, \\ &= \underbrace{-\frac{1}{2}L_s\log(z) + \frac{1}{2}\sum_{i=1}^{L_s}\log(1 + z\lambda_i(\Sigma_{\mathrm{K}}))}_{f(z,\Sigma_{\mathrm{K}})} + h(\Sigma_{\mathrm{K}}). \end{aligned}$$

where $z = \sigma^2\|g\|_2^2 \geq 0$ and in the second equality the definition of the characteristic polynomial is used. It can then be shown that $f(z, \Sigma_{\mathrm{K}})$ is monotonically decreasing with respect to z, and thus $\|g\|_2^2$, irrespective of Σ_{K}. Rewrite f and its derivative as

$$f = \frac{1}{2}\log\left(\underbrace{\prod_{i=1}^{L_s}\left(\frac{1 + z\lambda_i}{z}\right)}_{p(z,\lambda_i)}\right), \quad \frac{\partial f}{\partial z} = \frac{1}{2p(z,\lambda_i)}\frac{\partial p(z,\lambda_i)}{\partial z},$$

and observe first that $\lambda_i(\Sigma_{\mathrm{K}}^{-1}) > 0$ for all i, because they are eigenvalues of the inverse of a positive definite matrix. Define $p_i(z,\lambda_i) = \left(\frac{1+z\lambda_i}{z}\right)$. Since $p_i(z,\lambda_i) > 0$, it follows that $p(z,\lambda_i) > 0$ and thus the derivatives of f and p with respect to z have the same sign. Monotonic decrease of f with respect to z then follows from the fact that

$$\begin{aligned} \frac{\partial p_i(z,\lambda_i)}{\partial z} &= -\frac{1}{z^2}, \\ \frac{\partial p(z,\lambda_i)}{\partial z} &= \underbrace{\left(\prod_{i=1}^{L_s} p_i(z,\lambda_i)\right)}_{\geq 0} \underbrace{\left(\sum_{j=1}^{L_s}\frac{\partial p_j}{\partial z}\frac{1}{p_j}\right)}_{\leq 0}. \end{aligned}$$

■

Recall that, when Page matrices are used, $\Sigma_{\mathrm{y,f}}$ has the diagonal structure assumed in the lemma. This has two important implications for the input design problem.

First, searching for the input sequence which minimizes the Euclidean norm of g also maximizes the mutual information, and thus the distance (in expectation) between prior and posterior distributions of the simulated trajectory. Second, the prior on the simulation output, introduced via the kernel Σ_{K}, has no effect on the input design problem. That is, maximizing the mutual information coincides with minimizing uncertainty in the data estimate (21). Precisely, it coincides with the D-optimality criterion applied to $\Sigma_{\mathrm{y,f}}$. This is in contrast with recent results on input design for kernel-based impulse response identification [14], and is a property of the SMM estimator with the Page matrix representation.

2) Optimization problem: Motivated by the Bayesian interpretation, the experiment design optimization problem can be defined, at an abstract level, as the solution to the bi-level optimization problem

$$\min_{g,u_{\mathrm{d}}} \|g\|_2^2, \tag{24a}$$

$$\text{s.t.} \;\; u_{\mathrm{d}} \in \mathcal{U}, \tag{24b}$$

$$g \in \arg\min_{g\in\mathcal{G},u_{\mathrm{d}}}(17), \tag{24c}$$

where $\mathcal{U}$ defines input constraints. Using the relaxations of the SMM objective function (17) suggested in [5], and the idea to replace the inner optimization problem by its KKT conditions used in [12], problem (24) can be formulated as the following nonlinear program

$$\min_{g,u_{\mathrm{d}},\nu} \|g\|_2^2, \tag{25a}$$

$$\text{s.t.} \;\; u_{\mathrm{d}} \in \mathcal{U}, \tag{25b}$$

$$\begin{bmatrix} F(u_{\mathrm{d}}) & U^\top \\ U & 0 \end{bmatrix} \begin{bmatrix} g \\ \nu \end{bmatrix} = \begin{bmatrix} \hat{Y}_p(u_{\mathrm{d}})^\top y_{\mathrm{ini}} \\ \bar{u} \end{bmatrix}, \tag{25c}$$

where: $\nu \in \mathbb{R}^L$ are Lagrangian multipliers; $\bar{u} = \begin{bmatrix} u_{\mathrm{ini}}^\top & u_{\mathrm{s}}^\top \end{bmatrix}^\top$; and $F(u^d) = L\sigma^2 I_M + \hat{Y}_p(u_{\mathrm{d}})^\top \hat{Y}_p(u_{\mathrm{d}})$, where $M = c_H$ for Hankel matrices and $M = c_P$ for Page. The matrix $\hat{Y}_p$ is built with an output *predicted* data trajectory $\hat{y}_{\mathrm{d}}$ which linearly depends on the optimized input data trajectory u_{d}. This can be done using a baseline model, e.g. a previously identified impulse response model of the system. Analyses in [12] showed that the accuracy of the baseline model has generally small impact on the input design problem. It is important to recognize that the trajectory optimized via (25) automatically satisfies the input condition of Lemma 3, since the constraint $Ug = \bar{u}$ effectively enforces the range constraint. As for the state condition of Lemma 3, its fulfillment will depend on the choice of initial condition (see Remark 1). Finally, it is noted that, even though the Bayesian interpretation of problem (25) only holds exactly for Page matrices, the same program will later be tested on Hankel matrices. This is justified by the fact that, in the case of Hankel matrix, the objective function (25a) can be interpreted in terms of A-optimality criterion applied to the simulated response's mean-square error matrix [12].

IV. Numerical examples

We consider the design of an input data trajectory u_{d} for a data-driven simulation of the SISO system

$$G(z) = \frac{0.1159(z^3 + 0.5z)}{z^4 - 2.2z^3 + 2.42z^2 - 1.87z + 0.7225},$$

which was originally investigated in [16] and then also studied in [5], [12]. The accuracy of the estimated output simulation trajectory $\hat{y}_{\mathrm{s}}$ is quantified via the following fit

$$W = 100\left(1 - \left[\frac{\sum_{i=0}^{L_s-1}(y_{\mathrm{s},i} - \hat{y}_{\mathrm{s},i})^2}{\sum_{i=0}^{L_s-1}(y_{\mathrm{s},i} - \bar{y}_{\mathrm{s}})^2)^2}\right]^{1/2}\right),$$

where $\bar{y}_{\mathrm{s}}$ is the mean of the true output sequence y_{s}.

For the input constraint set $\mathcal{U}$, a bound on the total energy is imposed here by defining $\mathcal{U} = \left\{u_{\mathrm{d}} | \sum_{i=0}^{N-1}(u_{\mathrm{d},i})^2 \leq E_0 N\right\}$. Other constraint sets (e.g. magnitude constraints) could be studied as well. The following parameters are kept fixed throughout the analyses: $L_0 = 4, L_s = 10, L = 14, E_0 = 0.1$. Note that L_0 is equal to the system order, thus the condition $L_0 \geq l$ is satisfied. The baseline model consists of a truncated impulse response model of length $4L_s$ estimated with the *impulseest* MATLAB function using a prior experiment with i.i.d. Gaussian inputs of length 100 with SNR=10. The solver IPOPT [19] is employed to solve the nonlinear program (25). The optimized input data trajectory is then used to do data-driven simulation with SMM, i.e.: compute g by solving the relaxed version of (17) proposed in [5]; and then predict $\hat{y}_{\mathrm{s}}$ with (20).

In the first experiment, we compare the accuracy of Page and Hankel data matrices when data trajectories of *same* length N are used to predict the response to an impulse. Figure 1 shows the mean fit W over 200 realizations of the noise as a function of N. Noise contaminates the output data trajectory as in (16), and two cases are considered, $\sigma^2 = 0.001$ and $\sigma^2 = 0.01$, which correspond respectively to a signal-to-noise ratio (SNR) of 100 and 10.

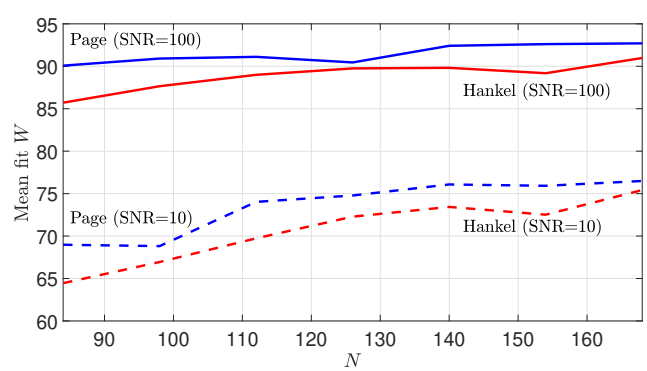

Fig. 1: Mean fit of Hankel (red) and Page (blue) matrices with optimally designed input as a function of N and σ^2.

The plot shows that Page matrices always outperform Hankel matrices. It is remarked that, since the comparison is made for equal length N of the experiments,

Hankel data matrices have a larger number of columns (precisely, $c_H(N) - c_P(N) = N(L-1)/L - L + 1$ as discussed in Section II-A). In this type of comparison, typically done using i.i.d. Gaussian inputs, Page matrices are reportedly less accurate than Hankel matrices [10]. The favourable trend reported in Figure 1 is achieved thanks to the relaxed excitation conditions from Lemma 3 and the optimally designed input using problem (25). It is also noted that a comparison with standard persistently exciting inputs is not possible, for the Page matrix, because in the considered range of N the classic excitation requirements (Lemma 2) do not hold.

In Figure 2, a comparison is made between the prediction accuracy of Page and Hankel matrices for two different simulation problems where u_{d} is respectively a heavily and lightly damped sine wave.

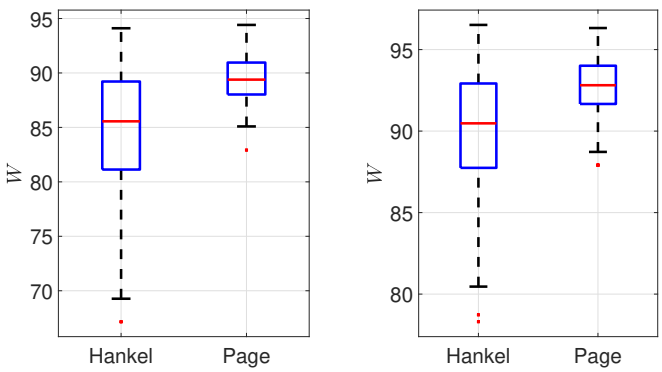

Fig. 2: Fit box plots for Page and Hankel matrices with $N = 84$ and SNR=100 (200 realizations of the noise). Left: heavily, right: lightly, damped sine wave.

The advantage of employing Page matrices can be appreciated here also in terms of dispersion. There are a few aspects that can provide an explanation for these results. It is intuitively expected that input design is more effective with the Page matrix representation, as the matrix U can be designed without structural constraints. In addition, Page matrices are known to have favourable properties when dealing with noisy data matrices [8]. Moreover, the input design criterion and the expression of the covariance Σ_{y} used in the SMM problem are only exact for Page representations. All these reasons provide valuable justifications for the use of Page matrices in data-driven simulation problems. Lemma 3 is a key enabler that makes this possible in the regime of data length N where a comparison with Hankel matrices is still meaningful.

V. Conclusions

Less restrictive excitation requirements have been presented for the data-driven simulation problem with clean data and Page and Hankel matrix representations. Building on these, a Bayesian input design problem for the case of noisy data has been formulated. For the Page matrix case, this can be interpreted as the choice of input trajectory maximizing the distance between prior and posterior distributions of the output response, while for the Hankel matrix this only holds approximately. Numerical results show that, by leveraging the results presented in the paper, the Page matrix representation can markedly outperform the Hankel one with the same data length. It is an interesting research question whether a similar approach to data informativity can be used when the data matrices are used for control.

References

[1] J. W. Polderman and J. C. Willems, *Introduction to Mathematical Systems Theory: A Behavioral Approach*. Springer-Verlag, 1997.

[2] J. C. Willems, P. Rapisarda, I. Markovsky, and B. L. M. De Moor, "A note on persistency of excitation," *Systems & Control Letters*, vol. 54, no. 4, pp. 325–329, 2005.

[3] I. Markovsky and F. Dörfler, "Behavioral systems theory in data-driven analysis, signal processing, and control," Tech. Rep., 2021.

[4] I. Markovsky and P. Rapisarda, "Data-driven simulation and control," *International Journal of Control*, vol. 81, no. 12, pp. 1946–1959, 2008.

[5] M. Yin, A. Iannelli, and R. S. Smith, "Maximum likelihood estimation in data-driven modeling and control," *arXiv:2011.00925*, 2020.

[6] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley-Interscience, 2006.

[7] P. van Overschee and L. de Moor, *Subspace identification for linear systems: theory, implementation, applications*. Kluwer Academic Publishers, 1996.

[8] A. Damen, P. Van den Hof, and A. Hajdasinski, "Approximate realization based upon an alternative to the Hankel matrix: the Page matrix," *Systems & Control Letters*, vol. 2, no. 4, pp. 202–208, 1982.

[9] J. Coulson, J. Lygeros, and F. Dorfler, "Distributionally robust chance constrained data-enabled predictive control," *IEEE Transactions on Automatic Control*, 2021.

[10] I. Markovsky and F. Dörfler, "Identifiability in the Behavioral Setting," Tech. Rep., 2020.

[11] H. J. van Waarde, "Beyond persistent excitation: Online experiment design for data-driven modeling and control," *IEEE Control Systems Letters*, vol. 6, pp. 319–324, 2022.

[12] A. Iannelli, M. Yin, and R. S. Smith, "Experiment design for impulse response identification with signal matrix models," in *19th IFAC Symposium on System Identification*, 2021.

[13] L. Ljung, *System identification: theory for the user*. Prentice Hall, 1999.

[14] Y. Fujimoto and T. Sugie, "Informative input design for kernel-based system identification," *Automatica*, vol. 89, pp. 37–43, 2018.

[15] I. Markovsky, J. C. Willems, P. Rapisarda, and B. L. De Moor, "Algorithms for deterministic balanced subspace identification," *Automatica*, vol. 41, no. 5, pp. 755 – 766, 2005.

[16] G. Pillonetto and G. De Nicolao, "A new kernel-based approach for linear system identification," *Automatica*, vol. 46, no. 1, pp. 81–93, 2010.

[17] T. Chen, H. Ohlsson, and L. Ljung, "On the estimation of transfer functions, regularizations and gaussian processes - revisited," *Automatica*, vol. 48, no. 8, pp. 1525–1535, 2012.

[18] D. Simon, *Optimal State Estimation: Kalman, H Infinity, and Nonlinear Approaches*. Wiley, 2006.

[19] A. Wächter and L. T. Biegler, "On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming," *Mathematical Programming*, vol. 106, no. 1, pp. 25–57, 2006.