Past-future information bottleneck in dynamical systems

Felix Creutzig[*]
Berkeley Institute of the Environment, University of California, Berkeley, California 94720, USA

Amir Globerson[†]
School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem 91904, Israel

Naftali Tishby[‡]
School of Computer Science and Engineering and The Interdisciplinary Center for Neural Computation,
The Hebrew University of Jerusalem, Jerusalem 91904, Israel

(Received 15 September 2008; revised manuscript received 12 March 2009; published 27 April 2009)

Biological systems need to process information in real time and must trade off accuracy of presentation and coding costs. Here we operationalize this trade-off and develop an information-theoretic framework that selectively extracts information of the input past that is predictive about the output future, obtaining a generalized eigenvalue problem. Thereby, we unravel the input history in terms of structural phase transitions corresponding to additional dimensions of a state space. We elucidate the relation to canonical correlation analysis and give a numerical example. Altogether, this work relates information-theoretic optimization to the joint problem of system identification and model reduction.

DOI: 10.1103/PhysRevE.79.041925 PACS number(s): 87.10.Vg, 89.70.−a, 45.30.+s

I. INTRODUCTION

Biological sensory systems need to encode the most relevant incoming information and transmit this information successfully under noisy conditions in real time. Extracting the most predictive information from incoming temporal signals is crucial for two reasons. First, predictive adaptive coding is a natural implementation of redundancy reduction of data, thus making efficient use of scarce resources and therefore called efficient coding. Second, it is only information about the future that can be behaviorally relevant. Hence, organisms should get sufficiently accurate predictions that allow them to behave with resulting positive benefit while coding costs, e.g., energy resources or requirements on the architecture of the neural system are kept low. These biological tasks can be investigated from an information-theoretic perspective. The task of encoding a signal with minimal coding space is called source coding. The task of successful information transmission under noisy condition is done by adding redundancy counteracting the noisy elements and is called channel coding. The task of encoding and transmitting information simultaneously in an efficient manner is called joint source channel coding [1]. We are here interested in joint lossy source-channel coding in time. Here the term *lossy* indicates that not all information is encoded and transmitted but all the remaining, *relevant* information is encoded and transmitted in an efficient manner.

This framework can also be interpreted from a learning-theory perspective, where a complexity measure is desirable to quantify a preference for simpler models [2]. Hence, the extraction of sufficiently accurate predictive information, transmitting some but not all information, can also be regarded as the construction of an internal model mirroring external signal statistics but with limited complexity.

This work aims not only to characterize predictive information in a signal, but also to find *which properties* of the past are those that are relevant and sufficient for predicting the future. In particular, we describe the data stream as a dynamical system and seek to isolate the most predictive components of the past, relating them to parameters of the underlying system.

We formalize our problem in terms of a dynamical system, characterized by input U, state space X, and output Y. The information bottleneck (IB) method is ideally suited to extract approximate minimal sufficient statistics [3] and is, hence, a natural approach to find the relevant features between past and future. It has already been shown that predicting the next step of a multivariate time series with the information bottleneck method is formally equivalent to slow feature analysis—a successful algorithm modeling neural processes [4]. Here we are interested in the more general case: given past signal values U_p, how is the information of the past compressed into a model $\hat{Y}_f$ such that information about the complete future Y_f is preserved. We provide an analytic solution of the trade-off problem on the basis of previously obtained results for the IB when the variables are jointly Gaussian [5]. Our results show that as the trade-off parameter increases, the compressed state space goes through a series of structural phase transitions, gradually increasing its dimension. Thus, for example, to obtain *little* information about the future, it turns out one can use a one-dimensional (scalar) state space. As more information is required about the future, the dimension of the required state space increases up to its maximum n. The structure and location of the phase transitions are related to the eigenvalues of the so-called Hankel matrix. We will use a modified Ho-Kalman algorithm to obtain dynamical systems with reduced state space ap-

[*] creutzig@nature.berkeley.edu
[†] gamir@cs.huji.ac.il
[‡] tishby@cs.huji.ac.il

©2009 The American Physical Society

proximating the information-theoretic optimal mapping. We also clarify the relation to canonical correlation analysis and characterize the optimal trade-off function: the information curve. Finally, we specify the information curve of the well-known spring-mass system, thus giving an example to demonstrate the numerical feasibility of the past-future information bottleneck.

II. BACKGROUND: THE INFORMATION BOTTLENECK METHOD AND THE HO-KALMAN ALGORITHM

Let us first introduce the IB method. The information bottleneck method was presented as a general information theoretic approach of extracting informative components or approximate minimal sufficient statistics [3]. Given two variables X and Y, the IB method seeks an optimally compressed version of X denoted by $\hat{X}$ which keeps as much information about the *relevant variable* Y as possible. The above two quantities, namely, compression level and relevant information, are complementary and in general we need to trade one for the other. They are both quantified by the Shannon's mutual information. One, $I(X;\hat{X})$, quantifies the compression and should be minimized, while the other, $I(\hat{X};Y)$, quantifies the relevant information and should be maximized. It can be shown that both conditions are satisfied simultaneously by solving the following variational problem [3]:

$$\min \mathcal{L}:\mathcal{L}_{IB} \equiv I(X;\hat{X}) - \beta I(\hat{X};Y). \tag{1}$$

The positive parameter β determines the trade-off between compression and preserved relevant information and, by this, makes apparent a natural order: by increasing β one unravels features ("statistics") in X that are informative about Y, where more informative features are revealed first. For Gaussian statistics, the exact solution of the information bottleneck problem can be described as an eigenvalue problem [5]. We will explicitly rely on this work.

We will apply the IB method to the input and output of a linear dynamical system and from this infer the optimal state-space realization corresponding to a specific β value. From system identification theory, it is well known how to identify a state-space realization [6]. In the classical realization framework, the minimal state-space realization is described on the basis of knowledge of the pulse response parameters of a multivariable system. A solution for the case in which one has knowledge of all the coefficients is known as the Ho-Kalman algorithm and was first described in [7]. It is based on full-rank decompositions of the block Hankel matrix built from the pulse response coefficients of the system.

III. STATE SPACE AS THE INFORMATION BOTTLENECK BETWEEN PAST AND FUTURE

We formalize the problem as the past-future information bottleneck optimization of data streams:

$$\min \mathcal{L}:\mathcal{L}_{PFIB} \equiv I(\text{past, internal representation}) - \beta I(\text{internal representation, future}). \tag{2}$$

Given past signal values we are interested in a compressed version of the past such that information about the future is preserved. When varying β, we obtain the optimal trade-off curve—also known as the *information curve*—between compression and prediction, which is a more complete characterization of the complexity of the process. Our aim is to make the underlying predictive structure of the process explicit, and capture it by the states of a dynamical system.

In the following, we will investigate a specific state-space description of the general objective function [Eq. (2)]. We focus on the discrete-time case where the lumped linear dynamic system with process noise can be written as follows:

$$x_{t+1} = Ax_t + Bu_t, \tag{3}$$

$$y_t = Cx_t + Du_t. \tag{4}$$

Here u, x, and y are $p\times 1$, $m\times 1$, and $q\times 1$ vectors and A, B, C, D are $m\times m$, $m\times p$, $q\times m$, and $q\times p$ matrices, respectively. We denote the system's parameters given by the above equations by DS. Our focus is on the bottleneck function of the state space, and hence, we set $D=0$, as it directly links the input to the output. The dimension m of the state space corresponds to the number of poles of the transfer function [8]. We also assume for simplicity that u_t is a zero mean Gaussian process with unit covariance. However, results are also valid for more general Gaussian processes. Since we are only interested on the effect of past input on future output at the present moment $t=0$, we clamp the input to zero for times $t\geq 0$. Extensions, including also input future and output past into the analysis, are possible using the same techniques as, e.g., in [9]. However, as we gain only numerical accuracy at the cost of making the system non-causal, we stick to the direct causal input-output relation.

Assuming stationarity of the input signal, we can focus on the case where the past is measured up to $t=0$ and the future for $t>0$. Our aim is to find an optimal model output $\hat{y}_t$ that compresses the information of the input past but keeps information on the output future. The model output is specified as

$$\hat{x}_{t+1} = A_r(\beta)\hat{x}_t + B_r(\beta)u_t, \tag{5}$$

$$\hat{y}_t = C_r(\beta)\hat{x}_t + D_r(\beta)u_t + \xi. \tag{6}$$

Hence, the IB Lagrangian can be written as

$$\min_{DS^r|DS} \mathcal{L}:\mathcal{L} \equiv I(U_p,\hat{Y}_f) - \beta I(Y_f,\hat{Y}_f), \tag{7}$$

where the input past, the output future, and the model future are given by

$$U_p = \begin{pmatrix} u_t \\ u_{t-1} \\ \cdots \\ u_{k-1} \end{pmatrix}, \quad Y_f = \begin{pmatrix} y_{t+1} \\ y_{t+2} \\ \cdots \\ y_{t+k} \end{pmatrix}, \quad \hat{Y}_f = \begin{pmatrix} \hat{y}_{t+1} \\ \hat{y}_{t+2} \\ \cdots \\ \hat{y}_{t+k} \end{pmatrix},$$

with $u_t=[u_1(t),\ldots,u_p(t)]^T$, $y_t=[y_1(t),\ldots,y_q(t)]^T$, $\hat{y}_t=[\hat{y}_1(t),\ldots,\hat{y}_q(t)]^T$, and $k\to\infty$. The Lagrangian is optimized with respect to the matrices of the reduced system that are, in fact, a function of the trade-off parameter β: DS^r

$=[A_r(\beta), B_r(\beta), C_r(\beta)]$. These will be derived in what follows.

We minimize Eq. (7). First, we can rewrite the mutual information quantities in terms of differential entropies,

$$\mathcal{L} = h(\hat{Y}_f) - h(\hat{Y}_f|U_p) - \beta h(\hat{Y}_f) + \beta h(\hat{Y}_f|Y_f). \tag{8}$$

For differential entropies, $h(X) = -\int_X f(x)\log f(x)dx$. In particular, for Gaussian variables

$$h(X) = \frac{1}{2}\log(2\pi e)^d|\Sigma_X|,$$

where $|\Sigma_X|$ denotes the determinant of Σ_X and $\Sigma_X := \langle XX^T\rangle$ is the covariance matrix of X [1]. Hence, we have to find the covariance matrices of the quantities in Eq. (8). Recall that, for finite time steps k,

$$x_0 = [J_C]_k U_p,$$

$$Y_f = [J_O]_k x_0,$$

where the $m \times (p*k)$ controllability matrix and the $(q*k) \times m$ observability matrices are given, respectively, by

$$[J_C]_k = (BAB \ldots A^{k-1}B), \quad [J_O]_k = \begin{pmatrix} C \\ CA \\ \cdots \\ CA^{k-1} \end{pmatrix}.$$

Define the mapping of the input past into the output future as $Y_f = [J_O]_k[J_C]_k U_p \equiv HU_p$, where H is the so-called Hankel operator given as

$$H = \begin{pmatrix} CB & CAB & CA^2B & \ldots & CA^kB \\ CAB & CA^2B & CA^3B & \ldots & CA^{k+1}B \\ CA^2B & CA^3B & CA^4B & \ldots & CA^{k+2}B \\ \ldots & & & & \end{pmatrix}.$$

The rank of the Hankel matrix, corresponding to a system transfer function, is equal to m, the order of its minimal state-space realization [8]. The analysis here is asymptotically true for $k \to \infty$, if all eigenvalues of A are inside the unit disk. Equivalently, relying on Eqs. (5) and (6) and adding model noise for regularization [5], we obtain $\hat{Y}_f = H(\beta)U_p + \xi$ for the model future. We seek to identify $H(\beta)$, or equivalently, $A_r(\beta)$, $B_r(\beta)$, and $C_r(\beta)$. Based on the Hankel operator, we compute $\Sigma_{\hat{Y}_f} = H(\beta)\Sigma_{U_p}H(\beta)^T + \Sigma_\xi$ and $\Sigma_{\hat{Y}_f|U_p} = \Sigma_\xi$. The last covariance matrix is given by

$$\begin{aligned} \Sigma_{\hat{Y}_f|Y_f} &= \Sigma_{\hat{Y}_f} - \Sigma_{\hat{Y}_f, Y_f}\Sigma_{Y_f}^{-1}\Sigma_{Y_f,\hat{Y}_f} \\ &= H(\beta)\Sigma_{U_p}H(\beta)^T + \Sigma_\xi - H(\beta)\Sigma_{U_p,Y_f}\Sigma_{Y_f}\Sigma_{Y_f,U_p}H(\beta)^T \\ &= H(\beta)\Sigma_{U_p|Y_f}H(\beta)^T + \Sigma_\xi, \end{aligned}$$

where we used Schur's formula in the first and last step [10]. Neglecting irrelevant constants, Eq. (8) then becomes

$$\begin{aligned} \mathcal{L} = {} & (1-\beta)\log|H(\beta)\Sigma_{U_p}H(\beta)^T + \Sigma_\xi| - \log|\Sigma_\xi| \\ & + \beta\log|H(\beta)\Sigma_{U_p|Y_f}H(\beta)^T + \Sigma_\xi|. \end{aligned}$$

Lemma A.1 in [5] states, that without loss of generality, we can set $\Sigma_\xi = I$. Then minimizing the Lagrangian gives

$$\begin{aligned} \frac{d\mathcal{L}}{dH(\beta)} = {} & (1-\beta)[H(\beta)\Sigma_{U_p}H(\beta)^T]^{-1}2H(\beta)\Sigma_{U_p} \\ & + \beta[H(\beta)\Sigma_{U_p|Y_f}H(\beta)^T + I]^{-1}2H(\beta)\Sigma_{U_p|Y_f}. \end{aligned}$$

Equating this to zero and rearranging, we obtain conditions for the weight matrix A,

$$\begin{aligned} & \frac{\beta-1}{\beta}\{[H(\beta)\Sigma_{U_p|Y_f}H(\beta)^T + I][H(\beta)\Sigma_{U_p}H(\beta)^T + I]^{-1}\}H(\beta) \\ & \quad = H(\beta)[\Sigma_{U_p|Y_f}\Sigma_{U_p}^{-1}]. \end{aligned} \tag{9}$$

Let us denote the singular value decomposition of the Hankel matrix as

$$H = W^T\Sigma_H V. \tag{10}$$

We now state the past-future information bottleneck of dynamical systems (PFIB). The solution to Eq. (9) is

$$H(\beta) = W^T\Sigma_H(\beta)V, \tag{11}$$

where $\Sigma_H(\beta) = \mathrm{diag}[\sigma_1(\beta), \sigma_2(\beta), \ldots, \sigma_m(\beta)]$ and $\sigma_i(\beta) \equiv \sqrt{\frac{\sigma_i^2(\beta-1)-1}{r_i}}$ and $r_i = v_i\Sigma_{U_p}v_i^T$ is the norm induced by Σ_{U_p} and v_i are row vectors of V.

Proof. Let us first calculate $\Sigma_{U_p|Y_f}\Sigma_{U_p}^{-1}$, using Schur's formula and Eq. (10):

$$\Sigma_{U_p|Y_f}\Sigma_{U_p}^{-1} = I - H^T(HH^T + I)^{-1}H = V^T(I + \Sigma_H^2)^{-1}V. \tag{12}$$

Hence, $\Sigma_{U_p|Y_f} = V^T(I + \Sigma_H^2)^{-1}V\Sigma_u$. Consider the positive definite bilinear form induced by Σ_{U_p}:

$$v_i\Sigma_{U_p}v_j^T = r_i\delta_{i,j},$$

where the v_i are the row vectors of V. Denote by R the matrix with r_i on its diagonal. We substitute Eqs. (10)–(12) into Eq. (9) and obtain

$$\begin{aligned} & \frac{\beta-1}{\beta}\{[W^T\Sigma_H(\beta)(I+\Sigma_H^2)^{-1}V\Sigma_{U_p}V^T\Sigma_H(\beta)W + I] \\ & \quad \times [W^T\Sigma_H(\beta)V\Sigma_{U_p}V^T\Sigma_H(\beta)W + I]^{-1}\}W^T\Sigma_H(\beta)V \\ & = W^T\Sigma_H(\beta)VV^T(I+\Sigma_H^2)^{-1}V\Sigma_{U_p}\Sigma_{U_p}^{-1} \\ & \Leftrightarrow \frac{\beta-1}{\beta}\{[W^T\Sigma_H(\beta)(I+\Sigma_H^2)^{-1}R\Sigma_H(\beta)W + I] \\ & \quad \times [W^T\Sigma_H^2(\beta)RW + I]^{-1}\}W^T\Sigma_H(\beta)V \\ & = W^T\Sigma_H(\beta)(I+\Sigma_H^2)^{-1}V. \end{aligned}$$

By left-hand multiplication with W, inserting W^TW between the brackets and right-hand multiplication with V^T, we obtain

$$\frac{\beta-1}{\beta}\{[\Sigma_H(\beta)(I+\Sigma_H^2)^{-1}R\Sigma_H(\beta)+I][\Sigma_H(\beta)^2R+I]^{-1}\}\Sigma_H(\beta)$$
$$=\Sigma_H(\beta)(I+\Sigma_H^2)^{-1}.$$

In this form, all matrices are diagonal and we can proceed in solving the individual Hankel singular values,

$$\frac{\beta-1}{\beta}\left(\frac{\sigma(\beta)_i^2 r_i}{\sigma_i^2+1}+1\right)\left(\frac{1}{\sigma(\beta)_i^2 r_i+1}\right)-\frac{1}{1+\sigma_i^2}=0.$$

After some reshaping, we obtain for $\sigma(\beta)_i^2$:

$$\sigma(\beta)_i^2=\frac{\sigma_i^2(\beta-1)-1}{r_i}\qquad \text{Q.E.D.}$$

Note that $H(\beta)$ is not a Hankel matrix, in general. However, $H(\beta)$ can be translated into reduced matrices $A(\beta)$, $B(\beta)$, and $C(\beta)$ by the algorithm of Ho and Kalman [7] obtaining a dynamical system that approximates the information-theoretic optimal mapping. Define $\gamma(\beta)\equiv[\Sigma_H(\beta)\Sigma_H^{-1}]^{1/2}$ and $[J_C(\beta)]_k=\gamma(\beta)[J_C]_k$, $[J_O(\beta)]_k=[J_O]_k\gamma(\beta)$. We can then factorize $H(\beta)$ into $H(\beta)=[J_O(\beta)]_k[J_C(\beta)]_k$. Then

$$B(\beta)=[J_C(\beta)]_1,\quad C(\beta)=[J_O(\beta)]_1. \tag{13}$$

Define the submatrices $[J_C(\beta)]_{1:k-1}$ and $[J_C(\beta)]_{2:k}$ obtained from J_C by deleting the last and first row respectively. Then $A(\beta)$ can be computed as

$$A(\beta)=[J_C(\beta)]_{1:k-1}^{+}[J_C(\beta)]_{2:k}, \tag{14}$$

where $(\,)^+$ denotes the Moore-Penrose pseudoinverse. Similarly to balanced model truncation [6,11], the PFIB procedure also relies on Hankel singular values. The difference is continuous weighting in PFIB versus discrete weighting in balanced model truncation.

Relation to CCA: in canonical correlation analysis, the eigenvectors and eigenvalues of $\Sigma_{Y_fU_p}\Sigma_{Y_f}^{-1}\Sigma_{U_pY_f}\Sigma_{U_p}^{-1}$ are computed. In the past-future information bottleneck the target matrix is $\Sigma_{U_p|Y_f}\Sigma_{U_p}^{-1}=I-\Sigma_{Y_fU_p}\Sigma_{Y_f}^{-1}\Sigma_{U_pY_f}\Sigma_{U_p}^{-1}$. Hence, eigenvectors of both procedures are identical [5]. The eigenvalues of CCA, called (squared) canonical correlation coefficients, are denoted as $\lambda_i^{\text{CCA}}=\rho_i^2$. The eigenvalues of $\Sigma_{U_p|Y_f}\Sigma_{U_p}^{-1}$ can be calculated from Eq. (12) as $\lambda_i^{\text{PFIB}}=(1+\sigma_i)^{-1}$. Hence, the relationship between the Hankel singular values and the canonical correlation coefficients can be calculated by $(1+\sigma_i)^{-1}=\lambda_i^{\text{PFIB}}=1-\lambda_i^{\text{CCA}}=1-\rho_i^2$ to give

$$\rho_i^2=\frac{\sigma_i^2}{\sigma_i^2+1}\quad\text{or}\quad\sigma_i^2=\frac{\rho_i^2}{1-\rho_i^2}.$$

The information curve of predictive information illustrates the trade-off between model accuracy, here: predictive information $I(\hat{Y}_f,Y_f)$, and model complexity, here: required or compressed information from the past $I(U_p,\hat{Y}_f)$. This curve is similar to the rate-distortion curve of lossy source coding. As can be deduced from [5], the theoretical information curve for PFIB is given as

$$I(U_p,\hat{Y}_f)_\beta=\frac{1}{2}\sum_{i=1}^{n(\beta)}\log(\beta-1)\sigma_i^2, \tag{15}$$

$$I(\hat{Y}_f,Y_f)_\beta=I(U_p,\hat{Y}_{fut})-\frac{1}{2}\sum_{i=1}^{n(\beta)}\log\beta\frac{\sigma_i^2}{\sigma_i^2+1}$$
$$=\frac{1}{2}\sum_{i=1}^{n(\beta)}\log\frac{\beta-1}{\beta}(\sigma_i^2+1)$$
$$=\frac{1}{2}\sum_{i=1}^{n(\beta)}\log\frac{\beta-1}{\beta}\frac{1}{1-\rho_i^2},$$

where $n(\beta)$ indicates the maximal index i such that $\beta\geq 1+\frac{1}{\sigma_i^2}$.

As $\beta\rightarrow\infty$ the predictive information converges to

$$I(\hat{Y}_f,Y_f)_\infty=\frac{1}{2}\sum_{i=1}^{n}\log\frac{1}{1-\rho_i^2}.$$

Reassuringly, this is identical to Akaike result on the mutual information between past and future of a stochastic system [12].

IV. SPRING-MASS SYSTEM AS AN EXAMPLE

As an example system we apply the past-future information bottleneck on a spring-mass system with two different masses, both fixed with a spring k_1 at the wall, a spring k_2 connects both masses. Two forces u_1 and u_2 perturb the masses such that they are displaced by y_1 and y_2 from their idle position. This can be modeled as a dynamic system

$$A=\begin{pmatrix}0&1&0&0\\ \frac{-(k_1+k_2)}{m_1}&\frac{-c}{m_1}&\frac{k_2}{m_1}&0\\ 0&0&0&1\\ \frac{k_2}{m_2}&0&\frac{-(k_1+k_2)}{m_1}&\frac{-c}{m_2}\end{pmatrix},$$

$$B=\begin{pmatrix}0&0\\ \frac{1}{m_1}&0\\ 0&0\\ 0&\frac{1}{m_2}\end{pmatrix},\quad C=\begin{pmatrix}1&0&0&0\\0&0&1&0\end{pmatrix}. \tag{16}$$

For the frictionless system, $c=0$. The input is given by the forces $u=[u_1,u_2]^T$, the output by the resulting displacement $y=[y_1,y_2]^T$. The two-dimensional output $[y_1,y_2]^T$ represents the displacements of the two masses.

We calculated reduced realizations for a set of different β values. The different reduced realizations correspond to points in the information plain spanned by $I(U_p,\hat{Y}_f)$ and $I(\hat{Y}_f,Y_f)$. The coordinates can be calculated as follows:

$$I(U_p,\hat{Y}_f)=\log|H(\beta)H^T(\beta)+I|,$$

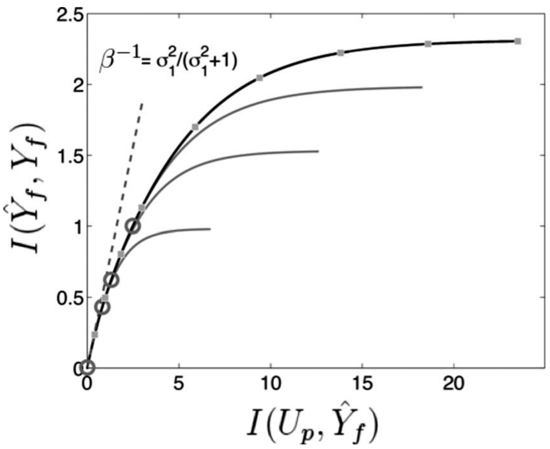

FIG. 1. Information curve for the spring-mass system. The dynamics is given by the matrices in Eq. (16) and the following parameters: $m_1=5$; $m_2=\sqrt{15}$; $k_1=1$; $k_2=0.5$; $c=0$. This particular system has Hankel singular values $\sigma_1=0.3358$, $\sigma_2=0.3109$, $\sigma_3=0.2374$, and $\sigma_4=0.2103$. The reduced systems $H(\beta)$, shown as gray squares, all lie on the information curve as given by the Hankel singular values. Circles indicate critical beta values.

$$I(\hat{Y}_f,Y_f) = I(U_p,\hat{Y}_f) - \log\big|H(\beta)H^T(\beta) - H(\beta)H^T(HH^T+I)^{-1}HH^T(\beta)+I\big|.$$

The points for some realizations are represented as gray squares in Fig. 1. The theoretical information curve as given by Eq. (15) is displayed as black line in Fig. 1. At each value of $I(U_p,\hat{Y}_f)$ the curve is bounded by a tangent with a slope $\beta^{-1}[I(U_p,\hat{Y}_f)]$ depicted in Fig. 1 as a dashed line. In the general information bottleneck method, the data processing inequality yields an upper bound on the slope at the origin, $\beta^{-1}(0)<1$. Here we obtain, similar to GIB [5], a tighter bound: $\beta^{-1}(0)<1-\lambda_1$. The asymptotic slope of the curve is always zero, as $\beta\to\infty$, reflecting the law of diminishing return: adding more and more bits to the description of $\hat{Y}_f$ provides less and less accuracy gain about U_p. We see that all sample realizations lie on the optimal information curve. This demonstrates the numerical feasibility of the past-future information bottleneck. The gray lines display the information curves when restricting the dimensionality of the state space to one, two, or three dimensions, respectively.

V. DISCUSSION

In a similar spirit to our results, the use of information between past and future for model selection has already been employed by calculating the information either with canonical correlation coefficients or by spectral densities [12–14] and can be traced back to [15]. In other work discrete-time stochastic processes are considered where the input is not entirely known [16]. A maximum likelihood ansatz is used to derive system matrices for a finite data set. Still other work shows that autoregressive moving average systems can be asymptotically efficient estimated by CCA and emphasizes that this approach provides accessible information on the appropriateness of the chosen model complexity [17]. Furthermore, the canonical correlation coefficients estimated by CCA of past-future data were shown to be equal to the cosines of the principal angles between the linear subspaces spanned by input past and output future [18].

The difference between our work and these results is that the information bottleneck allows a continuous rather than a discrete trade-off between two objectives and provides the computation of the optimal information curve. Fundamentally, our work shows that the problems of system identification and model reduction in dynamical system theory can be approached simultaneously by information-theoretic optimization. Such an approach may prove to be suitable for describing biological signal processing systems.

ACKNOWLEDGMENTS

F.C. would like to thank Boehringer Ingelheim Founds and the German National Merit Foundation for support. N.T. and A.G thank Israeli Science Foundation center of excellence grant 8006/00.

[1] T. Cover and J. Thomas, *The Elements of Information Theory* (Plenum Press, New York, 1991).
[2] W. Bialek, I. Nemenman, and N. Tishby, Neural Comput. **13**, 2409 (2001).
[3] N. Tishby, F. C. Pereira, and W. Bialek, Proceedings of 37th Allerton Conference on Communication and Computation (unpublished).
[4] F. Creutzig and H. Sprekeler, Neural Comput. **20**, 1026 (2008).
[5] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss, J. Mach. Learn. Res. **6**, 165 (2005).
[6] T. Katayama, *Subspace Methods for System Identification* (Springer, New York, 2005).
[7] B. Ho and R. Kalman, Regelungstechnik **14**, 545 (1966).
[8] C. T. Chen, *Linear System Theory and Design* (Oxford University Press, New York, 1999).
[9] R. Lim, M. Phan, and R. Longman, Princeton University Tech. Report No. 3045, 1998 (unpublished).
[10] J. R. Magnus and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics* (Wiley, New York, 1988).
[11] S. Gugercin and A. C. Antoulas, Int. J. Control **77**, 748 (2004).
[12] H. Akaike, in *Canonical Correlation Analysis of Time Series and the Use of an Information Criterion*, edited by R. Mehra and D. Lainiotis (Academic, New York, 1976), pp. 27–96.
[13] N. Jewell and P. Bloomfield, Ann. Stat. **11**, 837 (1983).
[14] L. Li, J. Time Ser. Anal. **27**, 309 (2005).
[15] I. Gelfand and A. Yaglom, Am. Math. Soc. Transl. **12**, 199 (1959).
[16] S. Soatto and A. Chiuso, Department of Computer Science, UCLA Tech. Report No. UCLA-CSD- 010001, 2000 (unpublished).
[17] D. Bauer, Econometric Theory **21**, 181 (2005).
[18] K. D. Cock and B. D. Moor, citeseer.ist.psu.edu/564422.html