A statistical perspective on nonlinear model reduction

Joel Phillips
Cadence Berkeley Laboratories
San Jose, CA 95134, U.S.A.
jrp@cadence.com

ABSTRACT

This paper considers the advantages of statistically motivated reasoning in analyzing the nonlinear model reduction problem. By adopting an information-theoretic analysis, we argue that the general analog macromodeling is tractable on average by nonlinear reduction methods. We provide examples to illustrate the importance of utilizing prior information, and provide a general outline of algorithms based on reproducing kernel Hilbert space machinery.

1. NEED FOR AUTOMATIC MODELING

In the past decade, the need to reduce the time and risk required for implementation of analog/mixed-signal (AMS) designs has become increasingly evident. This is particularly true because of the increasingly complex designs arising in the communications area. As a result, interest has grown, as witnessed by the emergence of the analog modeling languages Verilog-AMS and VHDL-AMS, in more structured design methodologies that utilize behavioral modeling in the specification, design, and verification stages. Good behavioral models are difficult to write and verify by hand. Model creation requires a skill set somewhat distinct from that needed for circuit design. Those designers who do possess the necessary skills are often not the consumers of the final models, thus often not motivated to produce high quality models. The intent of developing automated nonlinear analog modeling tools is to ameliorate these difficulties, while still reaping the advantages of behavioral modeling. Most recent work on automatic macromodeling systems has originated from two major driving forces.

The first driver is the desire to create automatic or semi-automatic analog synthesis systems. Several approaches to analog synthesis[1, 2] depend on the generation of macromodels that describe the *performance* of a circuit or class of circuits, as a function of design parameters. This type of model typically describes the high-level behavior, in terms of figure of merit parameters such as gain, noise figure, and power consumption, of a single circuit block, as a function of high-level design parameters. An interesting reversal on the usual theme is found in [3], where performance macromodels are obtained as a by-product of the data obtained from a synthesis tool operating via circuit level analysis.

The second driver is the need for verification of the operation of analog sub-components when integrated into a larger design such as a wireless communications system. Macromodels are potentially helpful in this context as they can accelerate simulation-based verification, offer isolation from the details of the lower-level implementation, and provide some degree of IP-protection if constructed to hide implementation details. To date, simulation models, unlike the performance type models, typically are constructed for a specific circuit topology and a fixed set of circuit parameters (however see [4] for a contrary example). Their most important feature, which to date has not typically been shared by the synthesis-driven models, is that simulation macromodels, since they can be instantiated in a simulator, can be composed with other macromodels in a circuit testbench familiar to designers. Composability is very important, as it means that complex systems can be modeled by assembling together models of simpler blocks. Composability also means that a macromodel can be utilized or verified in the context of other circuitry. This paper will focus on macromodels intended primarily for simulation applications. In particular, we will be concerned with the *model reduction* paradigm.

2. A MODEL REDUCTION BESTIARY

An automatic macromodeling algorithm must meet three main criteria. First, the models it creates must be accurate, and the accuracy must be controllable. Second, the models must be sufficiently compact to achieve substantial simulation speedups. Third, the models must be generated using reasonable amounts of computational resources. The definition of "reasonable computational resources" is context dependent. Models that are generated "on-the-fly" and used once must be generated very quickly, more quickly than a full simulation would take. Models that are generated off-line and repeatedly re-used can be generated in a langorous fashion. These requirements, particularly the accuracy and compactness requirements, encourage a "white-box" approach to the model generation problem, that is, an approach that uses the maximal amount of information available to the macromodel generation tool, the original circuit itself. This is the model reduction paradigm.

2.1 Linear Model Reduction

Model reduction has met with considerable success for modeling linear, time-invariant passive components algorithms [5, 6] as well as time-varying linear systems[7, 8]. The now-standard means for analyzing these algorithms is the projection formalism. Projection methods reduce linear systems of the form

$$E\frac{dx}{dt} = Ax + Bu, \quad y = Cx + Du \quad (1)$$

where $u(t)$ represents system inputs, $y(t)$ system outputs, $x(t)$ system state, a by drawing an approximate state vector $\hat{x} = Vz$ from a lower-dimensional subspace defined by the column span of V. A matrix W specifies a Petrov-Galerkin condition; the residuals of the system (1), under the substitution $\hat{x} = Vz$, must be orthogonal to the column space of W. Often $W = V$, an orthogonal projection, is used, and this will be assumed henceforth. The projection-based reduced model is of the same form as the original,

$$\hat{E}\frac{dz}{dt} = \hat{A}z + \hat{B}u, \quad y = \hat{C}x + \hat{D}u \quad (2)$$

with the matrices defining the reduced model defined by

$$\hat{A} \equiv V^T A V \quad \hat{B} \equiv V^T B \quad \hat{E} \equiv V^T E V \quad \hat{C} \equiv CV, \hat{D} = D. \quad (3)$$

Common choices for the projectors are Krylov subspaces[6] and, in the truncated balanced realization (TBR) procedure, the principle eigenvectors of the product of controllability and observability Grammians.

Recently, there have been several techniques proposed in an attempt to generalize reduction techniques to nonlinear systems which, for simplicity, we will consider in the somewhat restricted form

$$\frac{dx}{dt} = f(x) + Bu, \quad y = Cx + Du \quad (4)$$

where $f(x)$ is an arbitrary nonlinear function. The projection formula may be applied as in the linear case (see [8] for references) to obtain (with the additional constraint $V^T V = I$ imposed for convenience of notation) a reduced model

$$\frac{dz}{dt} = V^T f(Vz) + V^T Bu, \quad y = CVz + Du. \quad (5)$$

So far, this has been a purely formal exercise: we don't know how to choose V, and, assuming we can find a good choice for V (which in turn makes the strong assumption that a linear, or affine, transformation is desirable in the first place) the cost of evaluating $V^T f(Vz)$ is unknown. $V^T f(Vz)$ can of course be evaluated explicitly in the "reduced" model, by reference to the original function $f(x)$, but such a model is not "reduced" by any reasonable definition. Without a more compact representation of $\hat{f}(z) \equiv V^T f(Vz)$, no practical acceleration of computation is achieved over the original system. Several recent attempts have been made to develop practical algorithms. The algorithms differ primarily in how the approximation to $V^T f(Vz)$ is constructed, and to some extent in the construction of the matrix V.

2.2 Volterra Motivated Methods

Several authors (again see [8]) have proposed using multi-dimensional polynomial expansions to simplify the evaluation of nonlinear functions. In this approach, the nonlinear function is approximated with the series,

$$f(x) = A_1 x^{(1)} + A_2 x^{(2)} + A_3 x^{(3)} + \cdots \quad (6)$$

with

$$x^{(1)} \equiv x, \quad x^{(2)} \equiv x \otimes x, \quad x^{(3)} \equiv x \otimes x \otimes x, \text{ etc.} \quad (7)$$

where the $A_k \in R^{n \times n^k}$ are the multi-dimensional (tensor) polynomials coefficients of the expansion. The polynomial terms may be projected by application of the rule

$$\hat{A}_k = V^T A_k (V \otimes V \otimes \cdots V). \quad (8)$$

so that each $\hat{A}_k \in R^{q \times q^k}$ (assuming q is the rank of V). The general complaint with these methods, and the related bilinearization technique[8], is that the number of coefficients in the reduced model grows exponentially with with order k. This makes them impractical for order greater than three or so. Recent improvements[9] in reducing the size of the V matrix postpone but do not eliminate this fundamental underlying problem. What is worse, the reduced models may be larger than the original models. For example, consider a network of simple diodes. Diodes are two-terminal devices, with a single I-V constitutive relation. The tensors A_k for $k \geq 2$ are very sparse, they contain $O(n)$ entries for an n-diode network, independent of k. However, this sparsity is not preserved under projection as written above. Consider representing the original diodes with order 10 polynomials, leading to a total of $10n$ coefficients in

the A_k. But after "reduction" to say ten states, a typical size of the state space for a macromodel, the set of order-10 polynomials in ten dimensions has dimension 2×10^6, and the redundant tensor product representations used in [8] are more than a factor of 2^{10} larger. What is going on? Is there a defect in the projection algorithm, in the use of polynomials, or in our conception of the problem?

2.3 Trajectory Motivated Methods

The perceived limitations of the polynomial-based techniques was one motivation behind the development of the trajectory-piecewise-linear algorithm [10] and its extensions[11]. In this approach, the function f is approximated by a linear combination of affine models,

$$\hat{f}(x) = \sum_k w_k(x) \left[f(x_k) + A_k(x - x_k) \right]. \quad (9)$$

After projection, the reduced $\hat{f}(z)$ then has the similar form

$$\hat{f}(z) = \sum_k w_k(z) \left[V^T \left(f(x_k) - A_k x_x \right) + V^T A_k V z \right]. \quad (10)$$

Piecewise models themselves are subject to the same criticisms as the polynomial-base methods: covering a multi-dimensional space with uniformly sized piecewise regions requires a number of regions that grows exponentially large with dimension. To circumvent this problem, [10] proposed taking the center points x_k only along the trajectory of the ODE system when driven by a specific "test" input. As the number of points is now bounded, the method can represent fairly strong nonlinearities along the trajectory. The suprising aspect of the results in [10] is that the model obtained often exhibits good accuracy when driven by inputs *different* from the "test" input used to construct the model.

2.4 Analysis

In retrospect, both the polynomial and trajectory methods brought more questions than answers. To make progress in analog macromodeling, we need some way of decomposing the problem into distinct component parts that can be analyzed and compared. In the polynomial methods, from the previous arguments we cannot tell if the ostensibly exponential cost is due to underlying problem complexity, or the algorithm used for choosing the polynomial coefficients. In the trajectory methods, it isn't clear if the observed good performance is due to a superior choice of functional representation, a superior methodology, or even a "lucky" choice of examples. We advocate decomposing the macromodeling problem into three parts:

1. The *specification* of the modeling problem itself, in terms of what fundamental information the macromodel needs to preserve about the original system. The *specification* of the modeling problem determines a lower bound on the complexity of a given macromodel.

2. The *representation* of the model itself, for example, a projection matrix V and polynomials to describe a function $f(x)$.

3. The *algorithms* used to determine a specific choice of macromodel given a candidate *representation* and problem *specification*.

Historically, this type of decomposition played an important part in development of projection methods for linear systems. The canonical representation of models in state-space form with reduction achieved via a projection matrix V cleanly separates the mechanical aspects from the information (specification) carried by the model itself, determined by the choice of column span of V. Once this

realization occurred, it became apparent that there are several possible algorithms to compute V in a numerically stable way.

3. HOW HARD IS NONLINEAR MODEL REDUCTION, REALLY?

The first question to be addressed in developing a nonlinear modeling approach regards the problem *specification*: whether the problem is even tractable at all, in general, or for a specific circuit. Is automated nonlinear analog modeling fundamentally tractable, or are we trying to build a sort of perpetual motion machine? We have some idea from the preceding algorithms of the complications introduced by nonlinear functions, but how severe are these from a fundamental perspective? The thesis of this section is that the difficulty of nonlinear model reduction problem can be described *quantitatively* by adopting a statistical viewpoint and an information-theoretic analysis. The recent adoption by some groups active in analog macromodeling research[3, 12, 13] of techniques from the data mining and machine learning communities is indicative of the attraction of this perspective.

3.1 Analysis of Analog Information

At its root the analog model reduction problem is about removal of information perceived to be redundant, given a level of modeling accuracy. Quantifying likelihood of success requires a precise definition of the "amount of information" in a given base system to be modeled. The typical viewpoint in circuit simulation, and numerical analysis generally, is rather imprecise. Cost of informational representation, or computation, is typically measured by the number of basis functions N used to represent the known or unknown functions in a given computation. These basis functions are classically designed to span a complete function space in some limit, with specific choices (e.g., low-order finite elements vs. high-order spectral discretizations) guided more or less by previous experience. With modern techniques, it is usually possible to develop algorithms that are of low polynomial complexity in N, $O(N) - O(N^2)$, and N is of sufficiently moderate size for most problems of interest. Linear model reduction algorithms adopt very much the same philosophy: the matrix A is represented by explicit enumeration of its entries, the reduced model is defined by basis vectors that are the columns of the projection matrix V. However, the experience with Volterra methods as developed to date indicates that this view of information content is insufficiently sophisticated for effective nonlinear modeling strategies.

Leaving aside for the moment the dynamics of the system, consider only the right-hand-side functions in the n-state linear ($\dot{x} = f_L(x) = Ax$) and nonlinear ($\dot{x} = f_N(x)$) models. In the linear case, the function can be described by a matrix, a member of the set $\mathbb{R}^{n^2}$. The set of general nonlinear functions $\mathbb{R}^n \times \mathbb{R}^n$ is of vastly larger cardinality. Of course, the vast majority of the possible functions are not "reasonable" physical choices, but as the previous example of polynomials shows, explicit enumeration of even a single class of possible "reasonable" functions is computationally impractical. Is there a better measure?

A basic theory of information is the Shannon theory[14]. Information is associated with the probability of occurance p of a random variable X as $-\log p$. The entropy $H(X)$ of a random variable is defined to be

$$H(X) \equiv -E_p\{\log p(x)\} \tag{11}$$

where $E_p\{\}$ denotes expected value over the distribution p. For a discrete variable, $H(X) = -\sum p(x)\log p(x)$. A fundamental result in information theory is that the entropy represents a lower bound on the average length of a code needed to describe a random variable X drawn with probability density p. In a very abstract view of information of any sort, the best achievable, i.e. minimal representation, is tied to entropy.

With a probabilistic interpretation of appropriate variables in the analog modeling problem, systems for which good compact macromodels exists are those of low entropy. We can, in principle, quantify the smallest achievable model, and compare the size of this optimal model to the original detailed description.

EXAMPLE 1. *Consider the statement: "A number between one and ten." This is a reasonable model for a box that provides as output a an integer $I \in [1, 10]$. The entropy is unspecified without knowledge of the distribution. A reasonable assumption might be a uniform distribution. This is in fact the maximum entropy distribution – all other probability distributions of I have lower entropy. The uniform distribution is* non-informative. *As $\log 10$ is a bound on the average length of an optimal code (thus a model) to describe I with uniform distribution, in principle we can always find a "model" of size $\log 10$ bits to describe a box that emits "a number between one and ten."*

3.2 Impact of imperfect prior knowledge

Entropy based reasoning does not seem very useful so far. Consider the Volterra models. Cubic Volterra models are models "with third order multi-dimensional polynomials." This statement is as equally non-informative, in its context, as Example 1 above. In general, if a variable X is drawn from a set $\mathcal{H}$, the entropy is bounded for all distributions by[14]

$$H(X) \leq \log |\mathcal{H}| \tag{12}$$

where $|\cdot|$ denotes cardinality of the set. The uniform distribution achieves this bound. Example 1 above illustrates a model that is sized more or less regardless of the actual details of the behavior of the random variable. The model not surprisingly is of maximal size. In practice, we always have some additional information about the probability distribution that can be formalized as an additional variable Y. The entropy of interest is the conditional entropy

$$H(X|Y) = -E_{p(x,y)}\{\log p(X|Y)\} \tag{13}$$

where $p(x, y)$ is the joint probability distribution of X, Y and $p(X|Y)$ the conditional probability distribution of X (given Y). Prior information reduces the entropy, $H(X|Y) \leq H(X)$ with equality holding only if X, Y are independent (Y gives no information about X).

EXAMPLE 2. *Suppose in the previous example that we know in addition the mean of the distribution to be drawn. Again, for any given mean μ, there is a unique maximal entropy distribution[15] $p_\mu(x)$. Choosing $\mu = 5.5$ gives the uniform distribution, with entropy $\log 10$. Any other mean results in a skewed distribution with lower entropy; we have added information about the problem. Means close to 1 or 10 almost completely specify the distribution and have very low entropies. Figure 1 illustrates. These distributions have low information content, and samples from them are easy, on average, to represent. The stronger the prior knowledge, the farther the distribution from the uninformative uniform distribution, and the lower the entropy.*

This example suggests that the difficulty with exponential growth of cost in the Volterra methods is not fundamental, but stems from not exploiting prior knowledge in the use of the polynomial basis. Of course, in actual computation, the information picture is more complicated than the idealized situation above portrays. First, we
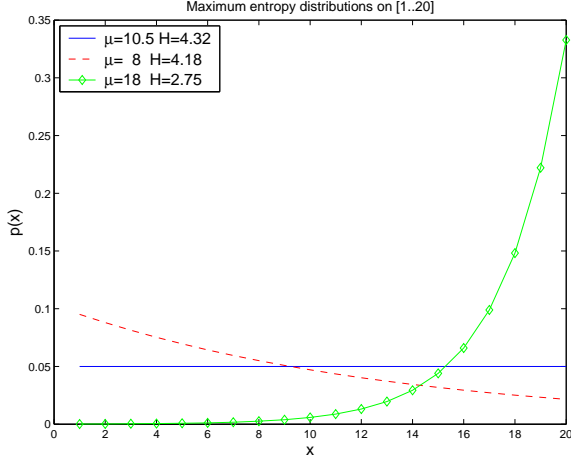
Figure 1: Maximum entropy probability distributions on [1..20] with constrained mean μ.

don't know the exact densities. Second, in computation we will choose some parametric form for the possible densities, and the "exact" densities may not be on the candidate list thus formed. A minimal complexity density estimation of p will asymptotically choose a representation with distribution q that minimizes $H(p) + D(p||q)$ [16]. The relative entropy $D(p||q) \equiv E_p\{\log(p/q)\}$ is the fundamental penalty for description with $q \neq p$. Third, we must represent continuous quantities. These considerations complicate the analysis, but are ameliorated by the fact that, since we are designing approximate models, we can accept some degree of error in our representation. Precise discussion of minimal representations of continuous data in the presence of an error criterion takes us into the area of lossy data compression and rate distortion theory and is beyond the scope of this paper. Minimal representations will arise from minimizing over distributions q that meet an admissible expected error ϵ. The rate distortion function $R(\epsilon)$ for a random variable X with approximation $\hat{X}$ and error function d

$$R(\epsilon) = \min_{p(x|\hat{x}):E_p\{d\}<\epsilon} E\left\{\log \frac{p(x,\hat{x})}{p(x)p(\hat{x})}\right\} \qquad (14)$$

determines the achievable lossy data description length[17]. Note that the quantity in brackets is the mutual information $I(X,\hat{X}) = H(X) - H(X|\hat{X})$, and so is bounded by the entropy $H(X)$. In the simplest contexts our length estimates with error will pick up a $\log \epsilon$ factor.

EXAMPLE 3 (TBR). *Consider the controllability operator $\mathcal{L}$: $[-\infty,0]^m \to \mathbb{R}^n$ which maps the inputs of an m-input n-state linear state-space model to the state $x_0 = x(0)$ at time zero, $x_0 = \mathcal{L}u$. Suppose for simplicity that the system is in balanced coordinates. The controllability Gramian X_c is*

$$X_c = \mathcal{L}\mathcal{L}^\dagger = \int_{-\infty}^0 e^{At}BB^T e^{A^Tt}dt \qquad (15)$$

(with $\mathcal{L}^\dagger$ denoting the adjoint operator). Consider interpreting $u(t)$ as a zero mean random variable with Gaussian distribution, inputs uncorrelated, and each input having autocorrelation function $R_u(t_1,t_2) = \delta(t_1 - t_2)$. The time-domain inputs are unit power. x_0 is then also a Gaussian random variable with correlation matrix

$$E\{x_0 x_0^T\} = E\{\mathcal{L}uu^\dagger\mathcal{L}^\dagger\} = \mathcal{L}E\{uu^\dagger\}\mathcal{L}^\dagger = X_c. \qquad (16)$$

The entropy of an n variable Gaussian distribution is[14]

$$H(X_0) = \frac{n}{2}\log(2\pi e) + \frac{1}{2}\sum_i \log \sigma_i \qquad (17)$$

where σ_i are the eigenvalues of the covariance matrix. By this metric a low entropy state-space model is one whose Grammian has many small singular values. If the admissible error after truncating states is more than the usual twice-the-sum-of-the-tail bound, the small singular values are irrelevant to the description length as described by the rate distortion function. In the modeling context, if most of the entropy is concentrated in the first few normal models, model reduction is very effective. The probabilistic interpretation and the classical criterion agree.

Furthermore, note that the stronger the degree of prior information, the lower the entropy. Consider the impact of input choice on the degree of reducibility of the model. A large circuit may have many possible ways to construct the input/output ports, i.e., the B/C matrices. Let $\aleph_j$ and $\aleph_k$ denote two possible subsets of possible input vectors $B_{\aleph_j}, B_{\aleph_k}$. The conditioning property of entropy implies

$$\aleph_j \subset \aleph_k \Rightarrow H(X_j) < H(X_k). \qquad (18)$$

EXAMPLE 4 (TBR, CONTINUED: PRIOR RESTRICTION). *In the context of the previous linear modeling example, it is easy to verify explicitly that addition of input vectors monotonically increases the entropy. The Gramian X_c is the solution to the Lyapunov equation $AX_c + X_cA^T = BB^T$. Let*

$$B_1 = B_{\aleph_j \cap \aleph_k}, \; B_2 = B_{\aleph_k - \aleph_j \cap \aleph_k}. \qquad (19)$$

If

$$AP_1 + P_1A^T = B_1B_1^T \qquad (20)$$

$$AP_2 + P_2A^T = B_2B_2^T \qquad (21)$$

and

$$AP + PA^T = [B_1B_2][B_1B_2]^T \qquad (22)$$

then it follows that $P = P_1 + P_2$. The claim follows from the strict concavity of the function $\log\det P$ and the symmetric positive definiteness of P_1, P_2.

Eq. (18) seems to match with our intuition based on linear model reduction algorithms. However, Eq. 18 is quite general, and suggests the remarkable speculation that many or most nonlinear circuits are "reducible" with sufficiently strong prior conditions on the input vectors. This lends some theoretical support to the observations[10, 13] that even seemingly vague priors seem to lead to practically useful results.

CONJECTURE 1. *Substantially all practical analog modeling problems are low-entropy relative to non-informative distributions over the circuit class.*

It also suggests why the trajectory methods seem to have an advantage over polynomials in treating more nonlinear circuits with a given size model. More prior information is available, specifically about the expected form of the input waveforms, when trajectories are computed. However this observation raises the interesting question of whether polynomial methods would exhibit increased efficiency given a means of exploiting similar prior information. The answer turns out to be yes, see [13]. Given that, a broader question arises: how should we compare the two methods? A deeper understanding of the methods requires factoring out differences due to

the choice of functional representation (piecewise linear vs. polynomial vs. piecewise polynomial etc.) from advantages due to exploiting different underlying prior information. In information theoretic terms, we wish to quantify separately the impact of reducing entropy by prior information $H(X) \to H(X|Y)$ from the relative entropy penalty $D(p||q)$ due to imperfect parametrization, as well as, eventually, the errors due to imperfect estimation. To make such a comparison it will be useful to have a common mechanism to describe the various choices of representation, particularly later when we desire to combine disparate choices in a single algorithm.

4. REPRESENTATION

To say a function is *non*-linear says only that: it is *not* linear. Nonlinearity is not a property, it is a *lack* of a property. Without further specific information about the function class under consideration, we are forced to consider the possible inclusion of a very large class of functions. This is the fundamental issue in nonlinear model reduction: before we approach any given problem, we need to include the *possibility* of an exponentially large number of basic functions. But by the arguments of the previous section, we hypothesize that any *given* problem leads to a relatively small class of necessary basis functions. The particular function class must be algorithmically deduced from the problem at hand. We advocate the machinery of *reproducing kernel Hilbert spaces* as a mechanism for implicit specification of high dimensional function spaces; [18] is the standard reference from which we crib

DEFINITION 1 (RKHS). *A reproducing kernel Hilbert space S is a Hilbert space of functions f on an index set $\mathfrak{I}$ such that for each $t \in \mathfrak{I}$ the evaluation functional $L_t f \mapsto f(t)$ is a bounded linear functional.*

Definition 1 is a fairly abstract condition on a function space, probably sufficently general for most residual functions $f(x)$ encountered in circuit modeling, and it leads to a general form for nonlinear representations that is useful because it is not overly verbose a-priori. By the Riesz representation theorem for each $t \in \mathfrak{I}$ there is an $R_t \in S$ called the representer of evaluation such that $f(t) = <R_t, f>$, and this representer is associated with a function of present interest:

THEOREM 1. *For every RKHS there is a unique positive definite kernel function $R(t,s) : \mathfrak{I} \times \mathfrak{I} \to \mathbb{R}$ and vice versa. S is spanned by the functions $R(t, \cdot) \quad \forall t \in \mathfrak{I}$.*

The convenience of the RKHS machinery for our purposes stems from two sources: the representer theorem and the duality with stochastic processes.

THEOREM 2 (REPRESENTER THEOREM). *Given a set $\mathfrak{X}$, a function $L : (\mathfrak{X} \times \mathbb{R} \times \mathbb{R})^n \to \mathbb{R}$, and a strictly monotonic function $\Omega : \mathbb{R}_+ \to \mathbb{R}$, each minimizer f of $L(x_1, y_1, f(x_1), \ldots, x_1, y_1, f(x_n))) + \Omega(||f||^2)$ admits a representation of the form*

$$f(x) = \sum_{k=1}^{n} \alpha_k K(x_k, x). \tag{23}$$

The implication of the representer theorem is that we may specify a space of functions in which to work, by selecting a kernel function $K(s,t)$, independently of the precise needs of the problem at hand. The particular basis functions, the $K(x_k, x)$, are only chosen once we choose x_k, which is done based on the particular problem data. Theorem 2 guarantees that these are indeed the right

---

Algorithm 1. **Nonlinear Model Reduction**

1. *Select a starting system $g(x, \dot{x}) = 0$ where $x \in \mathbb{R}^n$, $g : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$.*

2. *Use prior assumptions on the model properties to obtain a set $\{x_k\}$ of samples $x_k \in \mathbb{R}^n$.*

3. *Using $\{x_k\}$, compute the projector $V : \mathbb{R}^q \to \mathbb{R}^n$.*

4. *Obtain a set $\{z_k\}$ of samples $z_k \in \mathbb{R}^q$ by $z_k = V^\dagger(x_k)$.*

5. *Using $\{z_k\}$, compute a representation for the reduced DAE $\hat{g}(z, \dot{z}) = V^\dagger[g(V(z), (\partial V/\partial z)\dot{z})]$.*

6. *Assess the accuracy and increase the order q or sample space $\{x_k\}$ size as needed.*

---

functions to fit data to minimize a cost function (e.g. square error) L in the space of interest. Note that we did not unduly restrict the choice of norm here. In particular norms involving derivatives are of interest as they lead to splines. Thus the RKHS is minimal in a certain important sense: in any particular problem we include just the basis functions we observe in the sample space defined by the x_k, which are in turn determined by the problem data. We have no reason to believe any particular basis function is relevant for our problem unless and until we observe it in data, and until then it remains latent in the RKHS. At this point we have opened a route to bypass intractability, but created an experiment design problem. We must somehow choose the x_k, and a poor choice will lead us back to the intractably large full basis. Assuming a sufficiently rich RKHS, low-entropy arguments imply that such x_k exist; now we must look for them.

5. ALGORITHMS

A high level outline of modeling algorithms is shown as Algorithm 1. The critical steps are 2, 3, and 5. A key point is that Step 2 is the proxy for experiment design. The hope is that we have sufficient prior knowledge to choose the set $\{x_k\}$ such that (1) the projector will be moderate in size in step 3, (2) the choice leads to $\{z_k\}$ that are good choices for the approximations in Step 5. In linear projection based reduction, we choose $\{x_k\}$ to span a Krylov space because we know (prior information!) from the theory of projection based models that such subspaces are highly correlated with the desired frequency response of the model, which in turn is derived from our knowledge of frequency characteristics of the inputs. In trajectory methods, prior assumptions are made on the inputs, with $\{x_k\}$ taken from the simulation response to those inputs. In the linear case with uniform priors on the inputs, under the trajectory method the vectors x_k will be drawn from the controllable space of the model with probabilities according to the weights of the principle controllable modes as given by eigenvalues of the controllability Grammian. More selective choice of $\{x_k\}$ can reduce computational complexity, and also lead to superior models.

EXAMPLE 5 (TBR: BAYESIAN INTERPRETATION). *The assumption of inputs uncorrelated in time is equivalent to assumption of a flat spectrum. The TBR model is the best (in an entropy, i.e. maximal entropy, sense) q-state model under the prior assumption of flat spectrum, i.e. a non-informative prior on frequency content of input.*

In standard polynomial (and by extension Volterra) models, the

$\{x_k\}$ can be interpreted as representative of a noninformative prior distribution. This is not a good choice as it leads to the need for an exponentially large number of x_k as previously discussed. Superior choices were illustrated in [13], where structural information or trajectory information was used to form $\{x_k\}$. These choices were demonstrated to lead to strong regularizers (the standard priors implicit in kernels are quite weak in contrast).

Assuming reasonable choices in Step 2, Steps 3 and 5, dealing with nonlinear function representation, are the core of the modeling computation. If we adopt the RHKS representations, then in each case the nonlinear functions ($V(z)$ and $g(z, \dot{z})$) are represented by kernel expansions of the form

$$f(z) = \sum_{k,p} \alpha_k K_\lambda^{(p)}(z, \hat{z}_k). \tag{24}$$

The algorithm must choose the coefficients α, the kernel parameters λ, the kernel type itself $K^{(p)}$ as indexed by p, and the support vectors $\hat{z}_k$ that determine the usable portion of the RKHS. Each of the parameters may in turn be parametrized in a model with hierarchical form. For example, it is common to regularize coefficients in an expansion by assigning a statistical distribution (such as the normal distribution) to them; the distribution in turn is specified by *hyperparameters*. Ignoring the distinctions introduced by hierarchical modeling, for each of the parameters in the model we have several choices for its treatment: **Choice 1**: Guess!. **Choice 2**: optimize for minimum error or maximum probability. Classical numerical optimization, statistically motivated procedures such as cross-validation, as well as "boosting" fall into this category. **Choice 3**: A committee or averaging strategy, e.g. bootstrap, based on Choice 2. **Choice 4**: A full Bayesian analysis that integrates the model parameters over a posterior probability distribution. With good choice of the $\{x_k\}$, the $\hat{z}_k$ can be taken from the set $\{z_k\}$[13].

EXAMPLE 6 (REF. [13]). *In [13], a full instance of Algorithm 1 was given. In Step 3, a linear projector was chosen via SVD of the samples $\{x_k\}$, and kernel principal components used to pick the $\hat{z}_k$ for the final construction of the reduced DAE; again interpretable as using a maximum entropy argument in conjunction with utilizing a specific prior condition on the accessible model space. Cross-validation was used to choose the (unique) kernel types and parameters for Step 5.*

6. SUMMARY

In this paper we have tried to motivate analyzing some of the recent attempts at automated nonlinear modeling in a statistical light. We see four advantages from this perspective. First, it is a suitable way to reason in the presence of uncertainty. In almost all cases we, as algorithm or model designers, are lacking some information about the way models are used. Second, in the very high dimensional spaces that are the home of nonlinear models, we cannot account for *all* possible occurences. A possible approach is to account for all situations that we *reasonably believe* may occur, and then provide tests for robustness (that may be statistical). The third motivator is the connection to information theory, classically formulated in probabilistic terms, which may not directly suggest algorithms, but may suggest new ways of thinking about modeling. Finally, we can exploit the large and growing literature on statistical learning, inferences, and data mining.

REFERENCES

[1] M. Hershenson, S. P. Boyd, and T. H. Lee. Optimal deisgn of a CMOS op-amp via geometric programming. *IEEE Trans. Computer-Aided Design*, 20:1–21, 2001.

[2] W. Daems, G. Gielen, and W. Sansen. Simulation-based generation of posynomial performance models for the sizing of analog integrated circuits. *IEEE Trans. Computer-Aided Design*, 22:517–534, 2003.

[3] H. Liu, A. Singhee, R. Rutenbar, and L. R. Carley. Remembrance of circuits past: macromodeling by data mining in large analog design spaces. In 39th *ACM/IEEE Design Automation Conference*, New Orleans, LA, June 2002.

[4] L. Daniel, C. S. Ong, S. C. Low, K. H. Lee, and J. K. White. Geometrically parametrized interconnect performance models for interconnect synthesis. In *Proc. Int. Symp. Physical Design*, pages 202–207, April 2002.

[5] Lawrence T. Pillage and Ronald A. Rohrer. Asymptotic Waveform Evaluation for Timing Analysis. *IEEE Transactions on Computer-Aided Design*, 9(4):352–366, April 1990.

[6] P. Feldmann and R. W. Freund. Efficient linear circuit analysis by Padé approximation via the Lanczos process. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 14:639–649, 1995.

[7] J. Roychowdhury. Reduced-order modeling of time-varying systems. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 46:1273–1288, 1999.

[8] Joel R. Phillips. Projection-based approaches for model reduction of weakly nonlinear, time-varying systems. *IEEE Trans. Computer-Aided Design*, 22:171–187, 2003.

[9] P. Li and L. T. Pileggi. NORM: Compact model order reduction of weakly nonlinear systems. In 40th *ACM/IEEE Design Automation Conference*, pages 472–477, Anaheim, CA, June 2003.

[10] M. Rewieński and J. White. A trajectory piecewise-linear approach to model order reduction and fast simulation of nonlinear circuits and micromachined devices. *IEEE Trans. Computer-Aided Design*, 22:155–170, 2003.

[11] N. Dong and J. Roychowdhury. Piecewise polynomial nonlinear model reduction. In 40th *ACM/IEEE Design Automation Conference*, pages 484–489, Anaheim, CA, June 2003.

[12] F. De Bernardinis, M. I. Jordan, and A. Sangiovanni-Vincetelli. Support vector machines for analog circuit performance representation. In 40th *ACM/IEEE Design Automation Conference*, Anaheim, CA, June 2003.

[13] J. R. Phillips, J. Afonso, A. Oliveira, and L. M. Silveira. Analog macromodeling using kernel methods. In *International Conference on Computer Aided-Design*, Santa Clara, CA, November 2003.

[14] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley, 1991.

[15] J. O. Berger. *Statistical decision theory and Bayesian analysis*. Springer-Verlag, New York, 1985.

[16] A. Barron and T. Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37(4), 1991.

[17] J. C. Kieffer. Strong converses in source coding relative to a fidelity criterion. *IEEE Trans. Information Theory*, 37:257–262, 1991.

[18] G. Wahba. *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM, Philadelphia, 1990.