Learning Volterra Series via RKHS Methods

Joshua Hanson

May 10, 2019

1 Abstract

This project presents a Fock space framework for identification and modeling of nonlinear systems by Volterra series. By representing approximation models as elements in a Fock space, which is a reproducing kernel Hilbert space (RKHS), least squares regression can be performed to learn the original system dynamics from a collection of input-output sample pairs. The resulting expression can be easily adjusted to incorporate a priori knowledge of the order of nonlinear effects such as distortion, harmonic generation, and asymmetric/aperiodic oscillations, and is optimal in the sense that it is a projection in a Hilbert space of operators (orthogonality principle).

2 Volterra Series for Modeling Nonlinear Operators

Consider a nonlinear operator $H : L^2[0,T] \to C[0,T]$ which we wish to model. Alternative input spaces can be considered, but restricting inputs to L^2 enables approximation schemes with convenient structural properties which will be seen later. One approach to approximating the output of H is to construct a Volterra series, which is defined as follows.

Definition 2.1. *A Volterra series is a functional $\hat{H}_t : L^2[0,T] \to \mathbb{R}$ which models the output of a nonlinear operator at time t of the form*

$$\hat{H}_t u = h_0(t) + \sum_{n=1}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h_n(t, \tau_1, \dots, \tau_n) \prod_{i=1}^{n} u(\tau_i) d\tau_i \tag{1}$$

where $h_n : \mathbb{R}^{n+1} \to \mathbb{R}$ is called the n-th order Volterra kernel. We call the operator $\hat{H} : L^2[0,T] \to C[0,T]$ given by $\hat{H}u(t) = \hat{H}_t u$ for $t \in [0,T]$ the corresponding Volterra operator.

Notice that if the original operator H is causal, we can truncate the upper limits of the integrals in the Volterra series definition from ∞ to t because the desired Volterra kernels must satisfy $h_n(s, \dots) = 0$ for $s > t$ to preserve causality. Volterra series generalize the convolution description of linear operators to nonlinear operators analogously to how Taylor series generalize linear interpolation of real-valued functions to nonlinear functions. In a sense, a Volterra series can be viewed as a Taylor series on L^2 (rather than $\mathbb{R}$) with scalar coefficients replaced by linear integral operators on tensor products of L^2.

The value of Volterra series and Volterra operators depends on their capability to uniformly approximate a large class of nonlinear operators. If there exists an operator that satisfies the desired assumptions but cannot be uniformly approximated arbitrarily accurately by a Volterra operator, than it would be unfruitful to consider this modeling scheme much further. However, the algebra of Volterra series functionals defined on a compact set $K \subset L^2[0,T]$ (which is a Hausdorff space) separates points and vanishes nowhere, so by the Stone-Weierstrass theorem, this algebra is dense in the continuous functions $C(K, \mathbb{R})$. The relatively straightforward proof of this result is shown in [W. J. Rugh, 1981]. Furthermore, if H satisfies a condition called *fading memory* (which will not be defined here but essentially states that the output of H at time t only depends strongly on values of the input in the recent past and depends very weakly on values of the input in the distant past), this approximation can be extended to hold for all bounded Lipschitz inputs and all $t \in \mathbb{R}$, which is shown in [S. Boyd & L. Chua, 1985]. As a remark, the set of bounded Lipschitz inputs is not compact in $L^2(\mathbb{R})$ with respect to either the sup norm or the L^2 norm, but it is compact with respect to a special time-weighted norm defined in [S. Boyd & L. Chua, 1985]. However, bounded Lipschitz inputs are compact in $L^2[0, T]$, which follows from the Arzela-Ascoli theorem.

For the purposes of this project, we will only consider approximation over a finite time interval $[0, T]$, but it is not extremely difficult to extend this to $\mathbb{R}$ under the assumption that H has fading memory. These results show that Volterra series and Volterra operators are in fact good models for nonlinear systems in the sense that the output of any nonlinear operator defined on a compact set $K \subset L^2[0, T]$ can be uniformly approximated to arbitrary tolerance by some Volterra series. The problem now becomes learning the kernels for this Volterra series.

3 Learning Volterra Kernels

The learning problem can be formulated as follows. Let $H : L^2[0, T] \to C[0, T]$ be a causal continuous nonlinear operator. It is desired to approximate H over a finite time interval $[0, T]$ by a Volterra operator $\hat{H}$. That is, for any input $u \in L^2[0, T]$ and time $t \in [0, T]$, we wish to construct an approximation for the output of the system at time t, $y(t) = Hu(t)$, with a Volterra series having kernels $(h_n)_{n=0}^{\infty}$. The training data consist of input-output pairs $(u_j, y_j)_{j=1}^{m} \subset L^2[0, T] \times C[0, T]$, where y_j is the output of the system H corresponding to input u_j. As will be seen later, it is convenient to use empirical risk minimization (ERM) with the empirical risk defined as the worst case (over t) average quadratic loss given by

$$L_m(\hat{H}) = \sup_{t \in [0,T]} \frac{1}{m} \sum_{j=1}^{m} (y_j(t) - \hat{H}u_j(t))^2 \tag{2}$$

Before looking more closely at ERM, we will review two previous methods for learning Volterra kernels based on signal processing techniques and identify some disadvantages to these approaches.

3.1 Crosscorrelation and Multiple-Variance Method

Similar to how monomials are not orthogonal with respect to the L^2-inner product, the basis functionals $H_t^{(n)}$ given by $u \mapsto \int_{-\infty}^{t} \cdots \int_{-\infty}^{t} h_n(t, \tau_1, \ldots, \tau_n) \prod_{i=1}^{n} u(\tau_i) d\tau_i$ are not orthogonal in the sense that the outputs of these basis functionals are not uncorrelated when subject to stationary white Gaussian noise input. The crosscorrelaton method as developed in [Y. Lee & M. Schetzen, 1965] orthgonalizes these basis functionals via a Gram-Schmidt-like orthogonalization procedure to produce a set of orthogonal functionals $\{G_t^{(n)}\}_{n=0}^{\infty}$ that when summed, are referred to as a Weiner series. Naturally, we have that the output of the modeled system satisfies $y(t) = \sum_{n=0}^{\infty} H_t^{(n)} u = \sum_{n=0}^{\infty} G_t^{(n)} u$. The orthogonality of Weiner functionals can be stated more precisely as follows. Let $w(t)$ be stationary white Gaussian noise with mean zero and variance σ^2. Then the outputs of the original basis functionals and orthogonalized Weiner functionals satisfy

$$\mathbf{E}[(H_t^{(n)}w)(G_t^{(n')}w)] = 0, \quad n < n' \tag{3}$$

$$\mathbf{E}[(H_t^{(n)}w)(G_t^{(n')}w)] = 0, \quad n \neq n' \tag{4}$$

Let the functionals $H_t^{(n)}$ and $G_t^{(n)}$ be associated with kernels h_n and g_n, respectively. Using this orthogonality condition, one can derive the expressions for the Wiener kernels g_n to be

$$g_n(t, \tau_1, \ldots, \tau_n) = \frac{\mathbf{E}[y(t)x(t-\tau_1)\cdots x(t-\tau_n)]}{\sigma^{2n}n!} \tag{5}$$

for the off-diagonal (i.e., $\tau_i \neq \tau_{i'}$ for all $i \neq i'$) elements, and

$$g_n(t, \tau_1, \ldots, \tau_n) = \frac{\mathbf{E}[(y(t) - \sum_{n'=0}^{n-1} G_t^{(n')}w)x(t-\tau_1)\cdots x(t-\tau_n)]}{\sigma^{2n}n!} \tag{6}$$

for the diagonal elements (i.e., $\tau_i = \tau_{i'}$ for some two (or more) i, i'). These Wiener kernels g_n can then be rearranged via a linear combination to form the desired Volterra kernels h_n, and these expressions are derived in [Y. Lee & M. Schetzen, 1965].

Notice that the output kernels depend on the variance of the input noise. Input noise with high variance excites higher order nonlinearities more so than small variance input, and vice versa for lower order nonlinearities. To obtain more accurate kernels, it is advantageous to adapt the variance of the input noise to suit the order of the kernel being estimated, using small variance noise for lower order kernels and high variance noise for higher order kernels. This is referred to as the multiple-variance method. Implementing this approach to the previous crosscorrelation method modified formulas for the Wiener kernels, which are derived in [S. Orchioni, 2014] and will not be listed here, but have similar form to the equations above with some additional off-diagonal terms inside the expectation.

One primary disadvantage to both of these methods is the requirement for restrictively large sample sizes for acceptable convergence (even more so for high dimensional input spaces), often exceeding tens of thousands of input-output pairs [M. Franz & B. Schölkopf, 2006]. Other disadvantages include the requirement that the input be Gaussian noise (which cannot be created perfectly in experiment, leading to unavoidable estimation errors), and the assumption that the output is noise-free (any disturbance in the measured output will be modeled).

3.2 Fock Space Framework and RKHS Kernel Trick

Using kernel methods can overcome all three of the difficulties mentioned in the previous section. The following Fock space framework for continuous-time system identification and regression in a reproducing kernel Hilbert space (RKHS) is introduced in [L. V. Zyla, &R. J. P. deFigueiredo, 1983] and further developed in [R. J. P. deFigueiredo & T. A. W. Dwyer, 1980]. The same approach for discrete-time systems is also considered in [M. Franz & B. Schölkopf, 2006] with additional comments on time and memory complexity.

Definition 3.1. *A Fock space is a direct sum of tensor products of identical copies of a Hilbert space, given by*

$$\begin{aligned} F(L^2[0,T]) &= \bigoplus_{n=0}^{\infty} L^2[0,T]^{\otimes n} \\ &= \mathbb{R} \oplus L^2[0,T] \oplus (L^2[0,T] \otimes L^2[0,T]) \oplus \ldots \end{aligned} \tag{7}$$

Since the direct sum of Hilbert spaces remains a Hilbert space, the tensor product of Hilbert spaces remains a Hilbert space, and $L^2[0,T]$ is a Hilbert space, it follows that $F(L^2[0,T])$ is also a Hilbert space. Note that elements in $F(L^2[0,T])$ are given by sequences $(h_n)_{n=0}^{\infty}$ such that for each fixed t and n, $h_n(t,\dots) \in L^2[0,T]^{\otimes n}$. (An additional summability condition also needs to be satisfied, and this will be discussed in a few sentences.)

Suppose G_t and H_t are Volterra series functionals. By the Riesz representation theorem applied to each $L^2[0,T]^{\otimes n}$, G_t and H_t can be represented by elements $(g_n)_{n=0}^{\infty}$ and $(h_n)_{n=0}^{\infty}$, respectively, in $F(L^2[0,T])$ satisfying the relation given in the definition of a Volterra series. Therefore, Volterra series functionals are also elements of the Fock space $F(L^2[0,T])$ (because Hilbert spaces are self-dual). We can define the Fock space inner product of two Volterra series functionals by

$$\begin{aligned}\langle G_t, H_t\rangle_{F(L^2[0,T])} &= \sum_{n=0}^{\infty} \frac{1}{n!}\langle g_n, h_n\rangle^n_{L^2[0,T]^{\otimes n}} \\ &= \sum_{n=0}^{\infty} \frac{1}{n!}\int_0^T \cdots \int_0^T g_n(t,\tau_1,\dots,\tau_n)h_n(t,\tau_1,\dots,\tau_n)d\tau_1\cdots d\tau_n\end{aligned} \tag{8}$$

The induced Fock space norm is then given by

$$\begin{aligned}\|H_t\|_{F(L^2[0,T])} &= \sum_{n=0}^{\infty} \frac{1}{n!}\|h_n\|^n_{L^2[0,T]^{\otimes n}} \\ &= \sum_{n=0}^{\infty} \frac{1}{n!}\int_0^T \cdots \int_0^T |h_n(t,\tau_1,\dots,\tau_n)|^2 d\tau_1\cdots d\tau_n < \infty\end{aligned} \tag{9}$$

where the finiteness of this sum is the required summability condition mentioned earlier.

From here, we introduce the weighted Fock space $F_\rho(L^2[0,T])$ for any bounded positive sequence $\rho = \{\rho_0, \rho_1, \dots\}$. The sequence ρ is to incorporate any a priori knowledge of the original system to be modeled by weighting the orders of the nonlinearities, somewhat similarly to the weighting of the input noise variance in the multiple-variance method to suit the order of the nonlinearity. This new space is defined exactly as before, except the new inner product and norm are now given by

$$\begin{aligned}\langle G_t, H_t\rangle_{F_\rho(L^2[0,T])} &= \sum_{n=0}^{\infty} \frac{\rho_n}{n!}\langle g_n, h_n\rangle^n_{L^2[0,T]^{\otimes n}} \\ &= \sum_{n=0}^{\infty} \frac{\rho_n}{n!}\int_0^T \cdots \int_0^T g_n(t,\tau_1,\dots,\tau_n)h_n(t,\tau_1,\dots,\tau_n)d\tau_1\cdots d\tau_n\end{aligned} \tag{10}$$

We can define the induced Fock space norm by

$$\begin{aligned}\|H_t\|_{F(L^2[0,T])} &= \sum_{n=0}^{\infty} \frac{\rho_n}{n!}\|h_n\|^n_{L^2[0,T]^{\otimes n}} \\ &= \sum_{n=0}^{\infty} \frac{\rho_n}{n!}\int_0^T \cdots \int_0^T |h_n(t,\tau_1,\dots,\tau_n)|^2 d\tau_1\cdots d\tau_n < \infty\end{aligned} \tag{11}$$

It is quite easy to see that $F_\rho(L^2[0,T])$ is an RKHS. The associated Mercer kernel $K_\rho : L^2[0,T] \times L^2[0,T] \to \mathbb{R}$ is given by

$$K_\rho(u,v) = \sum_{n=0}^{\infty} \frac{1}{\rho_n n!}\langle u, v\rangle^n_{L^2[0,T]} \tag{12}$$

By equation (10), for any input $u \in L^2[0,T]$, the element $H_t \in F_\rho(L^2[0,T])$ can be expressed as the inner product

$$H_t u = \langle K_\rho(u,\cdot), H_t \rangle_{F_\rho(L^2[0,T])} \tag{13}$$

which shows that $F_\rho(L^2[0,T])$ is an RKHS.

Now that it has been estabished that the space of Volterra series functionals is an RKHS, we can express the data as projections of the output of the original operator H at time t along the representation of the inputs in $F_\rho(L^2[0,T])$. From now on, we will represent the Volterra series functional H_t by the output of the corresponding Volterra operator evaluated at time t, given by $\hat{H}(\cdot)(t)$. Trivially, the data $(u_j, y_j)_{j=1}^m$ necessarily satisfy

$$H(u_j\big|_{[0,t]})(t) = y_j(t), \quad 1 \le j \le m,\ t \in [0,T] \tag{14}$$

where the restriction is to ensure the resulting Volterra operator $\hat{H}$ will be causal. Since each input u_j is represented by a Mercer kernel $K_\rho(u_j,\cdot) \in F_\rho(L^2[0,T])$, considering the projection of $H(\cdot)(t)$ along the input representors in $F_\rho(L^2[0,T])$ gives the following equivalent form for equation (14).

$$\langle K_\rho(u_j\big|_{[0,t]},\cdot), H(\cdot)(t)\rangle_{F_\rho(L^2[0,T])} = y_j(t), \quad 1 \le j \le m,\ t \in [0,T] \tag{15}$$

From the Hilbert space projection theorem (a.k.a., orthogonality principle), for any point $H(\cdot)(t)$ in $F_\rho(L^2[0,T])$, there is a unique point $\hat{H}(\cdot)(t)$ in the closed subspace C spanned by the input representors $(K_\rho(u_j\big|_{[0,t]},\cdot))_{j=1}^m$ which minimizes the distance $\left\|H(\cdot)(t) - \hat{H}(\cdot)(t)\right\|_{F_\rho(L^2[0,T])}$ over C. Furthermore, the error vector $H(\cdot)(t) - \hat{H}(\cdot)(t)$ is orthogonal to every element in C. Note that if $H(\cdot)(t)$ cannot be expressed as an element in $F_\rho(L^2[0,T])$ itself, there exists an element in $F_\rho(L^2[0,T])$ arbitrarily close to $H(\cdot)(t)$, as mentioned in section 2. The optimal approximation $\hat{H}(\cdot)(t)$ is the projection of $H(\cdot)(t)$ onto the closed span of $(K_\rho(u_j\big|_{[0,t]},\cdot))_{j=1}^m$, hence the output of $\hat{H}$ at time t for input v can be written as the linear combination

$$\hat{H}(v)(t) = \sum_{j=1}^m c_j(t) K_\rho(u_j\big|_{[0,t]}, v) \tag{16}$$

It remains to solve for the coefficients $(c_j(t))_{j=1}^m$ for each time t, which are obtained via the equation

$$c(t) = G^{-1}(t) y(t) \tag{17}$$

where the Gramian matrix $G(t) \in \mathbb{R}^{m\times m}$ has elements $G_{ij} = \langle K_\rho(u_i\big|_{[0,t]},\cdot), K_\rho(u_i\big|_{[0,t]},\cdot)\rangle_{F_\rho(L^2[0,T])}$, and $c(t) = [c_1(t),\ldots,c_m(t)]^T$ and $y(t) = [y_1(t),\ldots,y_m(t)]^T$ are vector representation of the coefficients and output samples. This equation is simply a rearragement of equation (16) when the training measurements $(y_j)_{j=1}^m$ are substituted for the desired outputs $\hat{H}(\cdot)(t)$ from the Volterra series model. $G^{-1}(t)$ can be calculated as the solution to the differential equation given by differentiating $G^{-1}(t)G(t) = I$

$$\begin{aligned} \dot{G}^{-1}(t) &= -G^{-1}(t)G(t)G^{-1}(t), \quad t \ge 0 \\ G^{-1}(0) &= \lim_{t\to 0} G(t) \end{aligned} \tag{18}$$

where the conditions of the existence of this limit are discussed in [L. Zyla & R. deFigueiredo, 1983]. Essentially, if the limits of all the training inputs u_j and outputs y_j as time goes to zero from above exist and the determinant of the kth order time derivative of $G(t)$ for any $k \geq m$ is non-zero, then $\lim_{t\to 0} G(t)$ exists. This is easily satisfied for non-pathological choices of training inputs u_j and weighting sequence ρ. If the system is time-invariant, this is a linear vector equation which can be solved using simple techniques from linear algebra.

With the optimal coefficients $c_j(t)$ determined for all $t \in [0, T]$, the desired approximation is given by

$$\hat{H}(v)(t) = \sum_{j=1}^{m} c_j(t) \sum_{n=0}^{\infty} \frac{1}{\rho_n n!} \langle u_j\Big|_{[0,t]}, v\Big|_{[0,t]}\rangle^n_{L^2[0,T]} \tag{19}$$

and the Volterra kernels can be immediately identified as

$$h_n(t, \tau_1, \ldots, \tau_n) = \frac{1}{\rho_n n!} \sum_{j=1}^{m} c_j(t) \prod_{i=1}^{n} u_j(\tau_i) \tag{20}$$

If $\rho_n = \rho^n$, $\hat{H}$ reduces to

$$\hat{H}(v)(t) = \sum_{j=1}^{m} c_j(t) \text{exp} \frac{1}{\rho} \langle u_j\Big|_{[0,t]}, v\Big|_{[0,t]}\rangle_{L^2[0,T]} \tag{21}$$

which is easily computed. It is easy to see that minimizing the worst-case distance between $H(\cdot)(t)$ and $\hat{H}(\cdot)(t)$ in Fock space over t also minimizes the empirical risk defined earlier (they are, in fact, the same quantity).

4 Remarks and Conclusion

It was mentioned that one of the disadvantages of the crosscorrelation method and multiple-variance method is that any noise in the measured output would be modeled. Noise was not explicitly considered in the above RKHS approach to learning the Volterra kernels, but can be compensated for easily. Suppose that the outputs y_j are affected by noise and are replaced by $y_j + V_j$, with V_j being the projection of the noise into the output space. Assume V_i is independent of V_j for $i \neq j$. Then the new coefficients $(c_j(t))_{j=1}^m$ are now given by the solution to the differential equation $c(t) = (I + \Sigma_V^{-1} G^{-1}(t))^{-1} \Sigma_V^{-1} y(t)$ where Σ_V is the diagonal matrix of the L^2 norms of each sample of the noise process (i.e., $\Sigma_V^{-1} = \text{diag}(\|V_1\|_{L^2[0,T]}, \ldots, \|V_m\|_{L^2[0,T]})$). This result is briefly derived in [L. V. Zyla, & R. J. P. deFigueiredo, 1983].

In the discrete-time case with bounded inputs, it is easy to define a probability measure on the input space since it would be isomorphic to a compact subset of $\mathbb{R}^T$, for inputs defined for T samples. However, it is more difficult (although, not impossible) to assign a probability measure to spaces of continuous functions (something like a Haar measure on a topological group representation of $L^2[0, T]$). In this case, it is easier to deterministically choose a collection of inputs wisely to excite the desired nonlinearities of the original system to be modeled rather than randomly select inputs. This invalidates the $O(m^{-\frac{1}{2}})$ error bound for the ERM algorithm presented in Theorem 9.1 in the notes, but nevertheless the performance of the RKHS method should still significantly exceed the older crosscorrelation method.

To summarize, a weighted Fock space framework for system identification and modeling has been presented which enables least squares regression in an RKHS. The resulting input-output map can be expressed as a Volterra series operator and can be implemented by a bank of linear filters followed by a polynomial output map (or other universally approximating memoryless nonlineary, such as a feedforward neural network or rational function).

5 References

(1) W. J. Rugh, *Nonlinear System Theory: The Volterra/Wiener Approach*, 1981

(2) S. Boyd, L. O. Chua, *Fading Memory and the Problem of Approximating Nonlinear Operators with Volterra Series*, 1985

(3) Y. W. Lee, M. Schetzen, *Measurement of the Wiener Kernels of a Non-linear System by Cross-correlation*, 1965

(4) S. Orchioni, *Improving the approximation ability of Volterra series identified with a cross-correlation method*, 2014

(5) M. Franz, B. Schölkopf, *A Unifying View of Wiener and Volterra Theory and Polynomial Kernel Regression*, 2006

(6) L. V. Zyla, R. J. P. deFigueiredo, *Nonlinear system identification based on a Fock space framework*, 1983

(7) R. J. P. deFigueiredo, T. A. W. Dwyer, *A Best Approximation Framework and Implementation for Simulation of Large-Scale Nonlinear Systems*, 1980