Semi-parametric kernel-based identification of Wiener systems

Riccardo Sven Risuleo, Fredrik Lindsten, and Håkan Hjalmarsson

***Abstract*— We present a technique for kernel-based identification of Wiener systems. We model the impulse response of the linear block with a Gaussian process. The static nonlinearity is modeled with a combination of basis functions. The coefficients of the static nonlinearity are estimated, together with the hyperparameters of the covariance function of the Gaussian process model, using an iterative algorithm based on the expectation-maximization method combined with elliptical slice sampling to sample from the posterior distribution of the impulse response given the data. The same sampling method is then used to find the posterior-mean estimate of the impulse response. We test the proposed algorithm on a benchmark of randomly-generated Wiener systems.**

I. Introduction

The Wiener system is a block-oriented nonlinear model where a linear dynamical system is followed by a static nonlinear function in a cascade composition such as the one presented in Figure 1 [1]. Since their introduction, Wiener systems have been successfully used in many applications, for instance, in chemistry [2]–[5], biology [6], [7], and software systems [8], [9]. The Wiener system has also been used as a building block for control approaches such as *Wiener-MPC* [10], [11] and *Neural network Wiener models* [12]–[14].

Many approaches have been proposed for the identification of Wiener systems. For instance, approaches based on maximum likelihood [15]–[18] and Bayesian approaches [19]–[21]. Wiener systems have also been estimated using frequency-domain approaches [22]–[24], subspace approaches [25], [26], and approaches based on orthonormal basis functions [27] or on data from different experiments [28].

In this paper, we propose a kernel-based approach whereby the impulse response of the linear block of the Wiener cascade is modeled as a Gaussian process [29], [30]. The kernel function of the Gaussian process can be used to encode prior information about the impulse response, such as smoothness, exponential decay, or resonant frequencies [31]–[33]. To model the static nonlinearity, we use a basis-function model. The parameters of the kernel and the coefficients of the static nonlinearity are estimated, in an empirical-Bayes fashion, from the marginal likelihood function [34]. Because of the nonlinear dependency of the output on the impulse response, the marginal likelihood is intractable; nonetheless, we can effectively find the estimates of the parameters using the expectation-maximization (EM) method [35]. In particular, we use a stochastic approximation version of the EM method [36] where the intractable expectation step is replaced with a sampling step and a stochastic approximation step. In the sampling step, we use *elliptical slice sampling*, which is an effective sampling algorithm for problems with Gaussian priors and nonlinear likelihood functions [37]. The same sampling algorithm is used to compute the posterior-mean estimate of the impulse response of the linear block.

The rest of the paper is organized as follows. In Section II, we introduce the structure of the Wiener systems we consider. In Section III, we present the kernel-based model and the empirical Bayes approach we propose to identify it. In Section IV, we present an iterative algorithm based on the EM method to estimate the parameters of the model from the marginal likelihood. In Section V, we present some simulation experiments. In Section VI, we draw conclusions and outline future directions of research.

II. Problem formulation

We consider the single-input single-output discrete-time Wiener system in Figure 1.

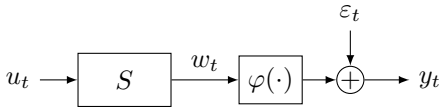

Fig. 1. The Wiener system considered in this paper: S is a linear time-invariant dynamical system and $\varphi(\cdot)$ is a memoryless nonlinear function.

The system is a cascade composition of a linear dynamical system S (which we refer to as the *linear block*) and a memoryless nonlinear function $\varphi(\cdot)$ (which we refer to as the *static nonlinearity*). We suppose that the linear block is time-invariant, asymptotically stable, and strictly causal. Under these assumptions, the intermediate signal w_t is given by the convolution of the input u_t with the impulse response of the linear block g_t according to

$$w_t = \sum_{k=1}^{\infty} g_k u_{t-k}. \tag{1}$$

The output w_t of the linear block passes through the nonlinear function $\varphi(\cdot)$ and is measured with additive noise ε_t:

$$y_t = \varphi(w_t) + \varepsilon_t.$$

We suppose that we have run the system for N time instants, with a known input signal $\{u_t\}_{t=1}^{N}$, and that we have collected N samples of the output $\{y_t\}_{t=1}^{N}$. For notational convenience,

This work was supported by the Swedish Research Council (under contracts 2015-05285, 2016-06079, and 2016-04278) and by the Swedish Foundation for Strategic Research (under contract ICA16-0015)

R. S. Risuleo and H. Hjalmarsson are with the Department of Automatic control, School of Electrical Engineering and Computer Science, KTH - Royal Institute of Technology, Stockholm, Sweden; F. Lindsten is with the Division of Systems and Control, Department of Information Technology, Uppsala University, Uppsala, Sweden. email: risuleo@kth.se, fredrik.lindsten@it.uu.se, hjalmars@kth.se

we use a vector notation where all the measurements of the output are stacked in a vector y, such that the tth element is given by the signal at time t. Furthermore, we suppose that there exists a large enough number n such that the impulse response g_t is zero for all $t \geq n$ (in particular, n may be equal to N). Then, the convolution (1) can be written, in vector form, as the product

$$w = \Phi g$$

where g is an $n \times 1$ vector of the nonzero samples of the impulse response and Φ is the $N \times n$ Toeplitz matrix of the input given by

$$[\Phi]_{i,j} = \begin{cases} u_{i-j+1} & \text{if } i-j+1 < N, \\ 0 & \text{otherwise.} \end{cases}$$

Note that, in the definition of Φ, we have included the initial conditions $\{u_t\}_{t=-n+2}^{0}$ which we assume are known.

The system identification problem we consider is thus as follows.

Problem 1: Given the $N \times 1$ vector y of samples of the output of a Wiener system described by

$$y = \varphi(\Phi g) + \varepsilon, \tag{2}$$

for a known Toeplitz matrix Φ, construct estimates $\hat{g}$ and $\hat{\varphi}$ of the linear block and of the static nonlinearity.

Remark 2: The considered Wiener system is, in general, only identifiable up to a scaling factor: for any $\alpha \neq 0$, the pairs $\big(g, x \mapsto \varphi(x)\big)$ and $\big(\alpha g, x \mapsto \varphi(x/\alpha)\big)$ describe the same input-output relation. To enforce identifiability, we consider linear blocks with unit gain and positive first coefficient of the impulse response (see Section V).

III. Kernel-based identification

We propose a kernel-based semi-parametric model of the Wiener system (2). We model the impulse response of the linear block as a zero-mean Gaussian process,

$$g_{(\cdot)} \sim \mathcal{GP}\left(0, K_\rho(\cdot,\cdot)\right) \tag{3}$$

where the kernel function $K_\rho(\cdot,\cdot)$ encodes the smoothness and exponential decay typical of impulse responses of stable systems. For instance, $K_\rho(\cdot,\cdot)$ may be a kernel in the family of *stable splines* [31] or may be chosen among other kernels for linear-system identification [32], [33], [38]. In general, these kernels depend on some *hyperparameters* ρ which need to be determined from data (see Section IV).

The Gaussian process model (3) implies that the vector of impulse response samples g is a Gaussian random vector,

$$g \sim \mathcal{N}(0, K_\rho), \tag{4}$$

with $\big[K_\rho\big]_{i,j} = K_\rho(i,j)$.

We model the nonlinear block as a combination of P known basis functions,

$$\varphi(\cdot) = \sum_{p=1}^{P} \theta_p f_p(\cdot) \tag{5}$$

where each $f_p(\cdot)$ could be, for instance, an orthonormal basis function, a sigmoid, a wavelet, or a rectifier, among others.

Following our vector notation, we define the matrix-valued vector function $F(\cdot)$ as the function which returns the $N \times P$ matrix whose columns are given by the basis functions:

$$\big[F(w)\big]_{i,j} = f_j(w_i). \tag{6}$$

We model the noise in (2) as Gaussian white-noise process with variance β. Then, we can define the *likelihood function*,

$$\log \mathrm{p}(y|g,\theta,\beta) = -\frac{1}{2\beta}\big\|y - F(\Phi g)\theta\big\|^2 - \frac{N}{2}\log 2\pi\beta,$$

where the $P \times 1$ vector θ contains the stacked coefficients of the combination (5).

To estimate the vector of impulse response samples, we interpret (4) as a prior distribution. Then, we can estimate the system with the posterior distribution of g given the data. In particular, the empirical Bayes minimum-variance estimate of g is given by

$$\hat{g} = \mathbf{E}\left\{g \middle| y, \hat{\theta}, \hat{\rho}, \hat{\beta}\right\}, \tag{7}$$

where the unknown parameters θ, ρ, and β are replaced by their *maximum marginal-likelihood* estimates:

$$\hat{\theta}, \hat{\rho}, \hat{\beta} = \underset{\theta,\,\rho,\,\beta}{\arg\max} \int \mathrm{p}(y|g,\theta,\beta)\mathrm{p}(g|\rho)\,\mathrm{d}g, \tag{8}$$

where $\mathrm{p}(g|\rho)$ is the Gaussian prior distribution (4).

Because y depends nonlinearly on g through the function $F(\cdot)$, it is in general impossible to express the posterior distribution of the impulse response and solve (7) analytically. Therefore, we approximate the expectation using a *Markov-Chain Monte Carlo* (MCMC) approach based on *Elliptical Slice Sampling* (ESS) [37].

In MCMC methods, expectations with respect to intractable distributions are replaced with sample averages. In the case at hand, we approximate the mean (7) with

$$\hat{g} \approx \frac{1}{M}\sum_{m=1}^{M} \bar{g}^{(m)}, \tag{9}$$

where $\{\bar{g}^{(m)}\}_{m=1}^{M}$ are samples drawn from a Markov chain with unique stationary distribution $p(g|y,\theta,\rho,\beta)$. The estimator is consistent and, under ergodicity conditions, converges at a standard Monte Carlo rate of $\sqrt{M}$.

To create the Markov chain, we use ESS. It is a variant of slice sampling [39] that is suitable for sampling from the posterior distribution of a random variable with Gaussian prior and nonlinear likelihood. At each state of the Markov chain, an ellipse is generated by sampling a random point ν from the prior. Then, the proposals are generated as points on this ellipse with a step size ω. The step size is then automatically adjusted until a sample is accepted. One step of ESS is presented in Algorithm 1.

Using Algorithm 1, we can easily and effectively draw samples from the posterior distribution of g, for a fixed value of the hyperparameters, to approximate the posterior mean (7).

However, computing the marginal likelihood—that is, the integral in (8)—is also intractable. To address this difficulty,

Algorithm 1 Draw one sample of the posterior of the impulse response of a semiparametric Wiener using ESS.

procedure ESS($g, \Phi, \rho, \theta, \beta$)
  $\nu \sim \mathcal{N}(0, K_\rho), \quad u \sim \mathcal{U}[0,1]$
  $L \leftarrow \log \mathrm{p}(y|g,\theta,\beta) + \log u$ ▷ Threshold
  $\omega \sim \mathcal{U}[0, 2\pi]$ ▷ Step-size parameter
  $\omega_{\min} \leftarrow \omega - 2\pi, \quad \omega_{\max} \leftarrow \omega$
  $g' \leftarrow g\cos\omega + \nu \sin\omega$ ▷ Proposal on ellipse
  if $\log p(y|g',\theta,\beta) > L$ **then** ▷ Accept proposal
    return g'
  else ▷ Adapt step size
    if $\omega < 0$ **then** $\omega_{\min} \leftarrow \omega$ **else** $\omega_{\max} \leftarrow \omega$
    $\omega \sim \mathcal{U}[\omega_{\min}, \omega_{\max}]$
    goto 6 ▷ Update proposal

we interpret it as a *maximum likelihood problem with latent variables*—the latent variable being the impulse response g. We can then use the EM method to find the estimates without having to find an explicit expression for the marginal likelihood. In the next section, we describe the resulting iterative approach.

IV. Iterative identification algorithm

To find the maximum marginal-likelihood estimates (8), we use the EM method. In particular, we use an MCMC-based version of the Stochastic-Approximation EM (SAEM) method [36].

In the general EM framework, a maximum-likelihood problem with latent variables such as (8) is solved, starting from an initial estimate $\hat\theta^{(0)}, \hat\rho^{(0)}, \hat\beta^{(0)}$, iterating two steps:

(E) the marginal likelihood is approximated from the joint distribution of latent variables and data by computing

$$Q^{(k)}(\theta,\rho,\beta) = \mathbf{E}\big\{\log\big[\mathrm{p}(y|g,\theta,\beta)\mathrm{p}(g|\rho)\big]\big\},$$

where the expectation is taken with respect to the posterior distribution $\mathrm{p}(g|y,\hat\theta^{(k)},\hat\rho^{(k)},\hat\beta^{(k)})$.

(M) the estimate is maximized to update the parameters:

$$\hat\theta^{(k+1)}, \hat\rho^{(k+1)}, \hat\beta^{(k+1)} = \arg\max_{\theta,\rho,\beta} Q^{(k)}(\theta,\rho,\beta).$$

Under some mild technical conditions, iterating the E-step and the M-step leads to a sequence of estimates of the parameters that converges to a local solution of the marginal-likelihood problem (8) [40].

In the kernel-based Wiener model that we are considering, the E-step is impossible to compute because we do not know the posterior distribution of the latent variables; however, as shown in the previous section, we can sample this posterior distribution effectively using ESS. Therefore, we can set up a stochastic approximation approach to approximate the function $Q^{(k)}$ using samples from the posterior.

In the SAEM method, the E-step is replaced with a simulation step and a stochastic approximation step [36]. This differentiates SAEM from other Monte-Carlo EM (MCEM) methods [41] and enables a more effective use the available samples [36, Section 3]. This, usually, allows the use of fewer samples per iteration in exchange for a higher number of iterations; this is particularly beneficial if the sampling step is more expensive than the maximization step (which is indeed the case in this application). In addition, SAEM has better convergence properties than MCEM (Among other, the convergence is asymptotic only in the iterations and not in the number of samples needed [36]).

At the kth iteration of the method, we update the estimate of the $Q^{(k)}$ function with

$$\hat Q^{(k)} = (1-\gamma_k)\hat Q^{(k-1)} + \gamma_k \Delta^{(k)}, \tag{10}$$

where the *learning rate* γ_k is a decreasing sequence (with $\gamma_1 = 1$, $\sum_{k=1}^\infty \gamma_k = +\infty$, and $\sum_{k=1}^\infty \gamma_k^2 < \infty$) and where the update term is

$$\Delta^{(k)}(\theta,\rho,\beta) = \frac{1}{M_k}\sum_{m=1}^{M_k} \log \mathrm{p}(y|\bar g^{(m,k)},\theta,\beta)\mathrm{p}(\bar g^{(m,k)}|\rho),$$

with $\{\bar g^{(m,k)}\}_{m=1}^{M_k}$ drawn (using ESS) from the posterior $\mathrm{p}(g|y,\hat\theta^{(k)},\hat\rho^{(k)},\hat\beta^{(k)})$ starting from the state of the chain at the previous iteration (i.e, $\bar g^{(0,k)} = \bar g^{(M_{k-1},k-1)}$). Note that, in general, the number M_k of samples drawn at each iteration may change.

What makes the SAEM method very effective in our case is that (10) reduces to an update of certain sample moments of the latent variables. In particular, if we define the following updates at the kth iteration of the SAEM method,

$$\begin{aligned}
\hat S^{(k)} &= (1-\gamma_k)\hat S^{(k-1)} + \frac{\gamma_k}{M_k}\sum_{m=1}^{M_k} \bar g^{(m,k)}\bar g^{(m,k)\,T},\\
\hat F^{(k)} &= (1-\gamma_k)\hat F^{(k-1)} + \frac{\gamma_k}{M_k}\sum_{m=1}^{M_k} F(\Phi\bar g^{(m,k)}),\\
\hat W^{(k)} &= (1-\gamma_k)\hat W^{(k-1)} + \frac{\gamma_k}{M_k}\sum_{m=1}^{M_k} F(\Phi\bar g^{(m,k)})^T F(\Phi\bar g^{(m,k)}).
\end{aligned} \tag{11}$$

then, we have the following result.

Theorem 3: At the kth iteration of the SAEM method, the solution to $\arg\max_{\theta,\rho,\beta}\hat Q^{(k)}$, with $\hat Q^{(k)}$ given in (10), is

$$\begin{aligned}
\hat\rho^{(k+1)} &= \arg\min_\rho \operatorname{Tr}\left\{K_\rho^{-1}\hat S^{(k)}\right\} + \log\det K_\rho,\\
\hat\theta^{(k+1)} &= \left[\hat W^{(k)}\right]^{-1}\hat F^{(k)\,T} y,\\
\hat\beta^{(k+1)} &= \frac{1}{N}\left(\|y\|^2 - y^T\hat F^{(k)}\hat\theta^{(k+1)}\right).
\end{aligned}$$

where $\hat S^{(k)}$, $\hat F^{(k)}$, and $\hat W^{(k)}$ are expressed in (11).

Proof: See Appendix A.

Note that, by letting $\gamma_1 = 1$, we have an automatic initialization of the iterations with $\hat Q^{(1)} = \Delta^{(1)}$. The updates in Theorem 3 should converge to a local maximum of the marginal likelihood [36], [42], [43]. Then, we can run a final ESS sampler to compute the estimate of the impulse response with (9).

The whole proposed approach for solving Problem 1 is presented in Algorithm 2.

Algorithm 2 Estimate a semi-parametric Wiener model using SAEM with ESS sampling. At each SAEM iteration M_k samples are used after a burn-in B_k. The final estimate is computed with M samples after a burn-in B.

procedure W-ESS($y, \Phi, F(\cdot)$)
  initialize $g, \hat{g}, \hat{\rho}, \hat{\theta}, \hat{\beta}, \hat{S}, \hat{F}, \hat{W}, k \leftarrow 1$
  while not converged **do** ▷ SAEM
    $\hat{S} \leftarrow (1-\gamma_k)\hat{S}$, $\hat{F} \leftarrow (1-\gamma_k)\hat{F}$, $\hat{W} \leftarrow (1-\gamma_k)\hat{W}$
    for $i = -B_k : M_k$ **do** ▷ SAE-step
      $g \leftarrow \text{ESS}(g, \Phi, \hat{\rho}, \hat{\theta}, \hat{\beta})$
      if $i > 0$ **then** ▷ Burn-in done
        $\hat{S} \leftarrow \hat{S} + \gamma_k g g^T / M_k$
        $\hat{F} \leftarrow \hat{F} + \gamma_k F(\Phi g)/M_k$
        $\hat{W} \leftarrow \hat{W} + \gamma_k F(\Phi g)^T F(\Phi g)/M_k$
    $\hat{\theta} \leftarrow \hat{W}^{-1}\hat{F}^T y$ ▷ M-step
    $\hat{\rho} \leftarrow \arg\min_\rho \operatorname{Tr}\left\{K_\rho^{-1}\hat{S}\right\} + \log\det K_\rho$
    $\hat{\beta} \leftarrow \left(\|y\|^2 - y^T\hat{F}\hat{\theta}\right)/N$
    $k \leftarrow k+1$
  for $i = -B : M$ **do**
    $g \leftarrow \text{ESS}(g, \Phi, \hat{\rho}, \hat{\theta}, \hat{\beta})$
    if $i > 0$ **then**
      $\hat{g} \leftarrow \hat{g} + g/M$ ▷ Compute posterior mean
  return $\hat{\theta}, \hat{g}$ ▷ Return estimate

V. Simulations

A. Random Wiener systems

In this section, we present a simulation experiment in which we estimate a dataset of 150 Wiener systems. We generate the Wiener systems with the following procedure:

1) We draw an order $m \in \{3, \dots, 10\}$ for the linear block and an order $P \in \{3, \dots, 10\}$ for the static nonlinearity, uniformly at random.
2) We generate a transfer function for the linear block by randomly drawing m poles and $m-1$ zeros in complex conjugate pairs. The magnitudes of the zeros are drawn uniformly in $[0, 0.9]$ and the magnitudes of the poles are draw uniformly in $[0.4, 0.8]$. All phases are drawn uniformly in $[0, \pi/3]$. To make the system identifiable, we normalized the linear block such that it has unit gain (see Remark 2).
3) We generate the static nonlinearity by randomly drawing P coefficients, uniformly in $[-1, 1]$ and letting $f_p(\cdot)$ be the pth Legendre polynomial.
4) We generate a uniform white-noise input signal in $[-1, 1]$ and we collect $N = 500$ samples of the output, with Gaussian white measurement noise with variance 10% of the corresponding noiseless output variance. All systems are simulated from zero initial conditions.

To ensure that the data were informative enough, we discarded all systems where the least-squares estimate of θ from y and w had negative fit.

Using the described procedure, we generated 150 systems and we estimated the impulse response of the linear block $\hat{g}$ and the coefficients of the static nonlinearity $\hat{\theta}$ and we normalized them such that $\|\hat{g}\|^2 = 1$ and $[\hat{g}]_1 > 1$. To evaluate the performance of the different methods, we used the following normalized measures of fit:

$$\text{FIT}_g = 1 - \frac{\|\hat{g} - g_0\|_2}{\|g_0 - \text{mean}(g_0)\|_2},$$
$$\text{FIT}_f = 1 - \frac{\|F(x)\hat{\theta} - F(x)\theta_0\|_2}{\|F(x)\theta_0 - \text{mean}(F(x)\theta_0)\|_2},$$

where g_0 and θ_0 are the true values of the impulse response and of the coefficients, respectively, and where x is a uniform grid of 100 points in $[-1, 1]$. The function "$\text{mean}(\cdot)$" denotes the sample mean of a vector.

In the simulation, we compared the following methods.

W-SA The proposed kernel-based method. It uses SAEM with ESS sampling. The impulse response was modeled as a stable spline, $\left[K_\rho\right]_{i,j} = \rho^{\max(i,j)}$, and the basis functions were chosen as Legendre polynomials with the correct order. There is no scaling in front of the kernel because of the nonidentifiability in Remark 2. At each iteration, $M_k = 50$ samples were used to compute the updates (11). The stochastic approximation rate was set to $\gamma_k = 1/k^{0.6}$ and the iterations were initialized with $\rho = 0.5$, $\beta = 1$, $[\theta]_1 = 1$, and $[\theta]_k = 0$ for $k > 1$. The update of $\hat{\rho}$ was done with search on a uniform grid of 20 values in $[0.4, 0.9]$. For the first iteration we used a burn-in of $B_1 = 10$ (then, $B_k = 0$ for $k > 1$). The iterations were stopped once the relative change in the parameter values dropped below 10^{-3}. The final estimates were computed with $M = 3000$ and $B = 100$.

W-MC The proposed kernel-based method with MCEM instead of SAEM. We used the same kernel and basis functions as W-SA. At each iteration the sample moments in (11) were computed afresh (i.e. $\gamma_k = 1$) using $M_k = 400$ samples. The rest of the parameters were the same as W-SA.

W-MAP A maximum-a-posteriori estimation of the proposed semi-parametric model. The system was estimated by solving, using gradient descent,

$$g_\rho, \theta_\rho = \arg\max_{g,\theta} N \log\left\|\tilde{y} - \tilde{F}(\tilde{\Phi}g)\theta\right\|^2 + g^T K_\rho^{-1} g,$$

for 20 uniformly-spaced values of $\rho \in [0.4, 0.9]$, where $\tilde{y}$ and $\tilde{\Phi}$ are the first $2/3$ of the available data and $\tilde{F}$ is a truncated version of the function (6) (note that, for fixed value of ρ, this cost function is proportional to the posterior density of g). Then, the value of ρ that gives smallest prediction error on the remaining $1/3$ of the data was chosen and the system was re-estimated using the whole dataset.

NLHW The `nlhw` method [15] implemented in Matlab with polynomial models for the linear block and for the static nonlinearity, with the correct orders.

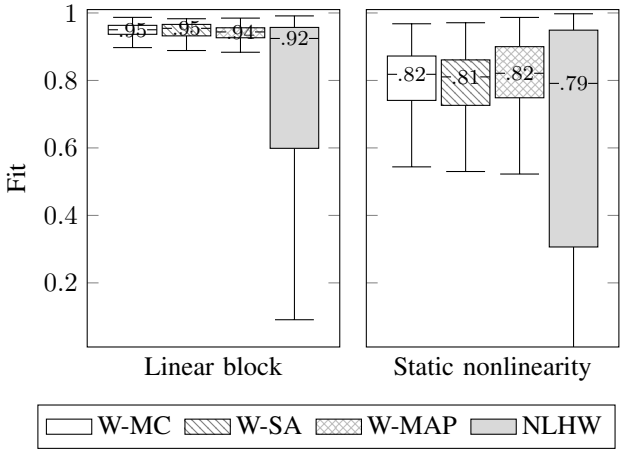

Fig. 2. Boxplots of the estimated impulse responses and static nonlinearities over the 150 systems in the dataset.

We used default initialization and convergence criteria.

The results of this simulation are presented in Figure 2. In the figure, we see the boxplots of the fit scores FIT_g (left plot) and FIT_f (right plot) over the 150 systems in the dataset. From the figure, it appears that the proposed kernel-based methods, on average, outperform NLHW.

In addition, it appears that W-SA has performance that is comparable to W-MC and W-MAP. On average, estimating one system took 143.30 seconds with W-MAP, 37.33 seconds with W-MC, and 0.45 seconds with W-SA (All code was implemented in Julia-0.6.2 and run on an i7-3770 @3.4GHz CPU; the same implementation was used for W-MC and W-SA). Therefore, it appears that the proposed SAEM approach is indeed effective at leveraging the information from previous iterations of the EM method to improve the estimates and speed up the convergence (see Section IV).

B. Choice of learning rate

To evaluate the impact of the choice of the learning rate γ_k on the performance of the method, we ran another set of simulations. In each simulation, we estimated the 150 systems in the dataset with W-SA using $\gamma_k = 1/k^\alpha$ for different values of the exponent α. In Figure 3 we present the results. The solid lines show the median fits of the estimates over the 150 systems in the dataset (Full circles: FIT_g; empty circles FIT_f) as a function of α. The dashed line (triangles) shows the median running time of the estimation algorithm. From this simulation, it seems that the method is quite robust to the choice of α and the performance seems good also for values that do not fulfill the stochastic approximation convergence conditions (i.e., for $\alpha < 0.5$).

VI. Conclusions

In this paper, we have proposed a kernel-based semi-parametric model for Wiener system identification. The impulse response of the linear block was modeled as a Gaussian process and the static nonlinearity using basis functions. Learning of these models was done via stochastic approximation EM and elliptical slice sampling.

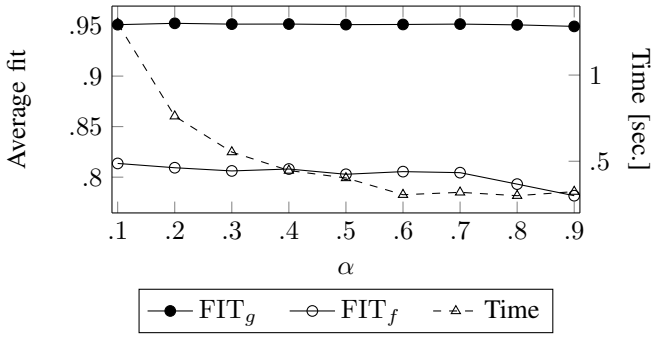

Fig. 3. Median fit (circles) and wallclock time (triangles) of W-SA run with different learning rates $\gamma_k = 1/k^\alpha$.

We have tested the proposed approach on a benchmark of randomly generated Wiener systems. From the simulation, it appears that the proposed approach compares favorably to the maximum-likelihood approach in the Matlab system identification toolbox.

In future publications, we plan to extend the proposed kernel-based model to include a nonparametric model for the static nonlinearity. This will allow for a very flexible and general model. This model may also be extended to more general Hammerstein-Wiener and Wiener-Hammerstein type of structures. To estimate the components of these more general models, we are considering approximation approaches based on sampling and on variational methods. These approaches will be explored in future publications.

Appendix

A. Proof of Theorem 3

Consider the complete-data likelihood[1]

$$\log\big[\mathrm{p}(y|g,\theta,\beta)\mathrm{p}(g|\rho)\big] \cong -\frac{1}{2\beta}\big\|y - F(\Phi g)\big\|^2 - \frac{N}{2}\log\beta - \frac{1}{2}g^T K_\rho^{-1} g - \frac{1}{2}\log\det K_\rho.$$

At the kth iteration of the SAEM method, we have (up to an additive constant) the stochastic update term

$$\Delta^{(k)} \cong \frac{1}{M_k}\sum_{m=1}^{M_k}\Big[-\frac{1}{2\beta}\Big\|y - F(\Phi\bar{g}^{(m,k)})\theta\Big\|^2 - \frac{N}{2}\log\beta - \frac{1}{2}\mathrm{Tr}\Big\{K_\rho^{-1}\bar{g}^{(m,k)}\bar{g}^{(m,k)\,T}\Big\} - \frac{1}{2}\log\det K_\rho\Big]$$

At the first iteration of the SAEM method ($k = 1$) we have

$$\hat{Q}^{(1)} = \Delta^{(1)} \cong -\frac{1}{2\beta}\Big(\|y\|^2 - 2y^T\hat{F}^{(1)}\theta + \theta^T\hat{W}^{(1)}\theta\Big) - \frac{N}{2}\log\beta - \frac{1}{2}\mathrm{Tr}\Big\{K_\rho^{-1}\hat{S}^{(1)}\Big\} - \frac{1}{2}\log\det K_\rho$$

[1]In the following "$\cong$" denotes equality up to an additive constant.

where we have defined the first sample moments from (11) (with $\gamma_1 = 1$). Consider now $k > 1$; from (10) we have that

$$\hat{Q}^{(k)} \cong -\frac{1}{2\beta}\left(\|y\|^2 - 2y^T\hat{F}^{(k)}\theta + \theta^T\hat{W}^{(k)}\theta\right) - \frac{N}{2}\log\beta - \frac{1}{2}\mathrm{Tr}\left\{K_\rho^{-1}\hat{S}^{(k)}\right\} - \frac{1}{2}\log\det K_\rho,$$

where we have used the definitions in (11). Straightforward maximization gives the update for ρ. From the first-order optimality condition,

$$\frac{\partial\hat{Q}^{(k)}}{\partial\theta} = \theta^T\hat{W}^{(k)} - y^T\hat{F}^{(k)} = 0,$$

we find the update for θ. Replacing this into the equation and maximizing with respect to β, we have the result. ■

REFERENCES

[1] F. Giri and E. W. Bai, *Block-oriented nonlinear system identification*. Springer, 2010.

[2] G. A. Pajunen, "Identification of a ph process represented by a nonlinear Wiener model," in *Adaptive Systems in Control and Signal Processing 1983*, pp. 91–95, Elsevier, 1984.

[3] A. Kalafatis, N. Arifin, L. Wang, and W. R. Cluett, "A new approach to the identification of pH processes based on the Wiener model," *Chem. Eng. Sci.*, vol. 50, no. 23, pp. 3693–3701, 1995.

[4] J. C. Gomez, A. Jutan, and E. Baeyens, "Wiener model identification and predictive control of a pH neutralisation process," *IEE Proc.-Control Theory Appl.*, vol. 151, no. 3, pp. 329–338, 2004.

[5] S. Mahmoodi, J. Poshtan, M. R. Jahed-Motlagh, and A. Montazeri, "Nonlinear model predictive control of a pH neutralization process based on Wiener–Laguerre model," *Chem. Eng. J.*, vol. 146, no. 3, pp. 328–337, 2009.

[6] P. Z. Marmarelis and K.-I. Naka, "White-noise analysis of a neuron chain: an application of the Wiener theory," *Science*, vol. 175, no. 4027, pp. 1276–1278, 1972.

[7] M. J. Korenberg and I. W. Hunter, "The identification of nonlinear biological systems: LNL cascade models," *Biol. Cybern.*, vol. 55, no. 2-3, pp. 125–134, 1986.

[8] T. Patikirikorala, L. Wang, A. Colman, and J. Han, "Hammerstein–Wiener nonlinear model based predictive control for relative QoS performance and resource management of software systems," *Control Eng. Pract.*, vol. 20, no. 1, pp. 49–61, 2012.

[9] D. Aryani, L. Wang, and T. Patikirikorala, "Control oriented system identification for performance management in virtualized software system," *IFAC Proc. Vol.*, vol. 47, no. 3, pp. 4122–4127, 2014.

[10] S. J. Norquay, A. Palazoglu, and J. Romagnoli, "Model predictive control based on Wiener models," *Chem. Eng. Sci.*, vol. 53, no. 1, pp. 75–84, 1998.

[11] H. H. J. Bloemen and T. J. J. Van Den Boom, "MPC for Wiener systems," in *Proc. IEEE Conf. Decis. Control (CDC)*, vol. 5, pp. 4595–4600, IEEE, 1999.

[12] H. Al-Duwaish, M. N. Karim, and V. Chandrasekar, "Use of multilayer feedforward neural networks in identification and control of Wiener model," *IEE Proc.-Control Theory Appl.*, vol. 143, no. 3, pp. 255–258, 1996.

[13] A. Janczak, "2 Neural network Wiener models," in *Identification of Nonlinear Systems Using Neural Networks and Polynomial Models*, pp. 31–75, Springer Berlin Heidelberg, 2004.

[14] O. Nelles, *Nonlinear system identification: from classical approaches to neural networks and fuzzy models*. Springer Science & Business Media, 2013.

[15] A. Hagenblad, L. Ljung, and A. Wills, "Maximum likelihood identification of Wiener models," *Automatica*, vol. 44, no. 11, pp. 2697–2705, 2008.

[16] A. Wills and L. Ljung, "Wiener system identification using the maximum likelihood method," in *Block-oriented nonlinear system identification*, pp. 89–110, Springer, 2010.

[17] B. Wahlberg, J. Welsh, and L. Ljung, "Identification of Wiener systems with process noise is a nonlinear errors-in-variables problem," in *Proc. IEEE Conf. Decis. Control (CDC)*, pp. 3328–3333, 2014.

[18] M. R. Abdalmoaty and H. Hjalmarsson, "A simulated maximum likelihood method for estimation of stochastic Wiener systems," in *Proc. IEEE Conf. Decis. Control (CDC)*, pp. 3060–3065, 2016.

[19] G. Pillonetto and A. Chiuso, "Gaussian processes for Wiener–Hammerstein system identification," in *Proc. IFAC Symp. System Identification (SYSID)*, vol. 15, pp. 838–843, 2009.

[20] F. Lindsten, T. B. Schön, and M. I. Jordan, "Bayesian semiparametric Wiener system identification," *Automatica*, vol. 49, no. 7, pp. 2053–2063, 2013.

[21] A. Wills, T. B. Schön, L. Ljung, and B. Ninness, "Identification of Hammerstein–Wiener models," *Automatica*, vol. 49, no. 1, pp. 70–81, 2013.

[22] G. Vandersteen, Y. Rolain, and J. Schoukens, "Non-parametric estimation of the frequency-response functions of the linear blocks of a Wiener-hammerstein model," *Automatica*, vol. 33, no. 7, pp. 1351–1355, 1997.

[23] M. Schoukens and Y. Rolain, "Parametric identification of parallel Wiener systems," *IEEE Trans. Instrum. Meas.*, vol. 61, no. 10, pp. 2825–2832, 2012.

[24] F. Giri, Y. Rochdi, A. Radouane, A. Brouri, and F. Z. Chaoui, "Frequency identification of nonparametric Wiener systems containing backlash nonlinearities," *Automatica*, vol. 49, no. 1, pp. 124–137, 2013.

[25] D. Westwick and M. Verhaegen, "Identifying MIMO Wiener systems using subspace model identification methods," *Signal Process.*, vol. 52, no. 2, pp. 235–258, 1996.

[26] M. Lovera, T. Gustafsson, and M. Verhaegen, "Recursive subspace identification of linear and non-linear Wiener state-space models," *Automatica*, vol. 36, no. 11, pp. 1639–1650, 2000.

[27] K. Tiels and J. Schoukens, "Wiener system identification with generalized orthonormal basis functions," *Automatica*, vol. 50, no. 12, pp. 3147–3154, 2014.

[28] G. Bottegal, R. Castro-Garcia, and J. A. K. Suykens, "On the identification of Wiener systems with polynomial nonlinearity," in *Proc. IEEE Conf. Decis. Control (CDC)*, 2017.

[29] C. E. Rasmussen and C. K. Williams, *Gaussian processes for machine learning*. the MIT Press, 2006.

[30] G. Pillonetto, F. Dinuzzo, T. Chen, G. De Nicolao, and L. Ljung, "Kernel methods in system identification, machine learning and function estimation: A survey," *Automatica*, vol. 50, no. 3, pp. 657–682, 2014.

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

[32] T. Chen and L. Ljung, "Constructive state space model induced kernels for regularized system identification," in *Proc. IFAC World Cong.*, vol. 19, pp. 1047–1052, 2014.

[33] F. Dinuzzo, "Kernels for linear time invariant system identification," *SIAM J. Control Optim.*, vol. 53, no. 5, pp. 3299–3317, 2015.

[34] J. S. Maritz and T. Lwin, *Empirical Bayes methods*. Chapman and Hall London, 1989.

[35] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc. Ser. B (Methodol.)*, pp. 1–38, 1977.

[36] E. Moulines, M. Lavielle, and B. Delyon, "Convergence of a stochastic approximation version of the EM algorithm," *Ann. Statist.*, vol. 27, no. 1, pp. 94–128, 1999.

[37] I. Murray, R. Adams, and D. MacKay, "Elliptical slice sampling," in *Proc. Int. Conf. Artif. Intell. Stat. (AISTAT)*, pp. 541–548, 2010.

[38] T. Chen, A. Chiuso, G. Pillonetto, and L. Ljung, "Rank-1 kernels for regularized system identification," in *Proc. IEEE Conf. Decis. Control (CDC)*, 2013.

[39] R. M. Neal, "Slice sampling," *Ann. Statist.*, vol. 31, no. 3, pp. 705–767, 2003.

[40] C. F. J. Wu, "On the convergence properties of the EM algorithm," *Ann. Statist.*, vol. 11, no. 1, pp. 95–103, 1983.

[41] G. C. G. Wei and M. A. Tanner, "A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms," *J. Am. Stat. Assoc.*, vol. 85, no. 411, pp. 699–704, 1990.

[42] C. Andrieu, É. Moulines, and P. Priouret, "Stability of stochastic approximation under verifiable conditions," *SIAM Journal on Control and Optimization*, vol. 44, no. 1, pp. 283–312, 2005.

[43] C. Andrieu and M. Vihola, "Markovian stochastic approximation with expanding projections," *Bernoulli*, vol. 20, no. 2, pp. 545–585, 2014.