A Simulated Maximum Likelihood Method for Estimation of Stochastic Wiener Systems

Mohamed Rasheed Abdalmoaty and Håkan Hjalmarsson

Abstract— **This paper introduces a simulation-based method for maximum likelihood estimation of stochastic Wiener systems. It is well known that the likelihood function of the observed outputs for the general class of stochastic Wiener systems is analytically intractable. However, when the distributions of the process disturbance and the measurement noise are available, the likelihood can be approximated by running a Monte-Carlo simulation on the model. We suggest the use of Laplace importance sampling techniques for the likelihood approximation. The algorithm is tested on a simple first order linear example which is excited only by the process disturbance. Furthermore, we demonstrate the algorithm on an FIR system with a cubic nonlinearity. The performance of the algorithm is compared to the maximum likelihood method and other recent techniques.**

I. INTRODUCTION

Stochastic Wiener models is a subclass of the general class of nonlinear state-space dynamical systems. A Wiener system is formed by two building blocks as shown in Figure 1. The first part is a linear dynamical system and the second part is a general static nonlinear function. Although this subclass might seem limited, it is flexible enough to describe many interesting physical systems where the nonlinearity is at the output. This might be as simple as a linear system with a nonlinear measurement sensor or more complicated processes such as those considered in [27], [11], and [9]. It has also been recognized in [2] that if the two building blocks of the Wiener model are multivariable, then the class can be used to approximate fairly general nonlinear models with arbitrary accuracy.

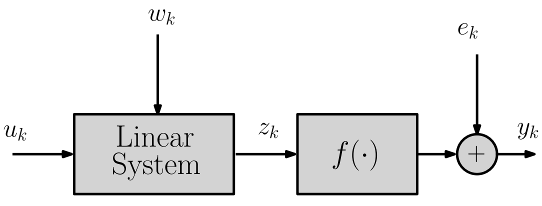

Fig. 1: Stochastic Wiener model

In this paper, we focus on single-input single-output parametric models.

The interest in the class of Wiener models within the system identification community is apparent by the number of available identification methods. An approach that was suggested in several contributions, see [1], [23], and [24] for example, disregards the process noise (sets $w_k = 0$ for all k) and tries to adapt classical methods such as the prediction-error method and subspace identification techniques [12]. Most of such approaches come with assumptions that can not handle common nonlinearities such as dead-zones and saturation. Nonparametric techniques that disregard the process noise or the process disturbance have also been considered as in [7] and [16].

It is known, as shown in [8], that ignoring the process disturbance leads to biased estimates. Instead, [8] tried to construct the maximum-likelihood estimator. Under the assumption that the disturbance and noise processes are white, the problem is approached by approximating a number of independent integrals over the reals using Simpson's rule. If the whiteness assumption is relaxed, the integrals are not independent anymore. A solution has been suggested in [26] using the celebrated Expectation-Maximization algorithm in combination with particle smoothers. The particle smoother is required to approximate the Q-step of the EM algorithm. Such an approach is an example of the recent development that employs nonlinear filtering techniques for identification of nonlinear systems as in [20] and [14]. These are filtering methods based on sequential Monte Carlo approximations. Markov Chains Monte Carlo (MCMC) methods has also been used. The recent survey paper [19] describes the available approaches in a common framework.

More recently, in [22] a simulation-based method known as indirect inference was used for the identification of Wiener models. This method is an instance of a large family of simulation-based techniques that are developed and used in econometrics, see for example the survey paper [5] or the book [6]. Simulation-based methods are techniques that only require the possibility of simulating data from the model once a parameter is fixed. They can be used for parameter estimation in fairly general statistical dynamical models. The role of these methods is to approximate cost functions in which some intractable integral appears, by arguments involving a version of the law of large numbers. When applied to approximate the likelihood function, the method is known as Simulated Maximum-Likelihood (SML). Perhaps the first appearance of such an idea was in [15] in which an analytically intractable log-likelihood function was approximated by simulations. Several contributions in econometrics have suggested the simulated-maximum likelihood method for models with latent variables such as [4], [10], [13] and [3].

This work was supported by the Swedish Research Council under contract 2015-05285 and by the European Research Council under the advanced grant LEARN, contract 267381.

Mohamed Rasheed Abdalmoaty and Håkan Hjalmarsson are with the Automatic Control Department and ACCESS Linnaeus Center, KTH Royal institute of Technology, Stockholm, Sweden (e-mail: {mohamed.abdalmoaty, hakan.hjalmarsson}@ee.kth.se).

In this paper, we investigate the simulated maximum-likelihood method for estimation of Wiener models with process disturbances. The method is to be seen as an alternative to the MCMC methods when the main goal is parameter estimation. The latent variables in this case are considered nuisance parameters. The outline of the paper goes as follows. In Section II, the estimation problem is defined. In Section III, the simulated maximum-likelihood method is introduced. First, in Section III-A we consider models with white disturbance process. Then, in Section III-B we describe the method for the more general case of colored disturbance process. In Section IV, some computational issues are discussed. Section V evaluates the performance of the method on several numerical examples. Finally, the paper is concluded in Section VI.

II. The problem

Consider the following stochastic Wiener model

$$\mathcal{M}(\theta)\begin{cases} x_{k+1}(\theta) = A(\theta)x_k(\theta) + B(\theta)u_k \\ z_k(\theta) = C(\theta)x_k(\theta) + H(q,\theta)w_k \\ y_k(\theta) = f(z_k(\theta),\theta) + e_k \\ x_0 = 0, \quad u_0 = 0 \end{cases} \tag{1}$$

For some fixed finite integer N, assume that for each $k = 1,\dots,N$ the input sequence $\{u_k\}$ is known, and both $\{e_k\}$ and $\{w_k\}$ are independent stochastic processes with known probability density functions

$$e_k \sim p_e(\cdot), \qquad \text{and } w_k \sim p_w(\cdot), \qquad k = 1,\dots,N$$

defined on the appropriate spaces according to the dimensions of the signals. Both processes are assumed to be white and stationary. The parameter θ is a finite dimentional vector parameterizing the linear dynamical system state-space matrices, A, B, and C. The disturbance process is modelled by the transfer operator $H(q,\theta)$ which is also parameterized by θ. If the disturbance process model $H(q,\theta) = 1$, the process disturbance is white and coincides with $\{w_k\}$. The symbol q denotes the forward shift operator that acts on time sequences. The function $f(\cdot,\theta)$ represents the static nonlinearity at the output and can be parameterized by θ. Furthermore, we assume that the measurement is collected in open-loop so that the input $\{u_k\}$ is independent of both the disturbance and noise processes. Let

$$Y = \begin{bmatrix} y_1 & y_2 & \dots & y_N \end{bmatrix}^T$$

be a vector of observations. For simplicity, we assume here y_k to be scalar. The maximum-likelihood estimation method requires the evaluation of the likelihood function, denoted $p(Y;\theta)$, of the observations. The maximum-likelihood estimate (MLE) is defined by the maximization problem

$$\hat{\theta}_N := \arg\max_{\theta\in\Theta} p(Y;\theta).$$

Unfortunately, for the general stochastic Wiener model, the joint likelihood function of the observations is not analytically tractable. In the following section, we introduce a simulation-based technique for the likelihood approximation.

III. Simulated Maximum-Likelihood

We first consider the cases in which the process disturbance is known to be white, i.e. $H(q,\theta) = 1$. In this case, one can directly write the joint likelihood of the observations as a product of the likelihood functions for a single observation.

A. Direct sampling for white disturbance

Under the assumptions that the noise and disturbance processes are white, $\{y_k\}$ is a sequence of independent random variables. The joint likelihood of Y is then given by

$$p(Y;\theta) = \prod_{k=1}^{N} p(y_k;\theta).$$

Therefore, it is enough to find the likelihood function for a single observation y_k. It is more convenient from the numerical point of view to work with the negative log-likelihood function defined by

$$-\log(p(Y;\theta)) = -\sum_{k=1}^{N} \log(p(y_k;\theta)) \tag{2}$$

in which $\log(\cdot)$ is the logarithmic function. The MLE is then defined as the global minimizer of the negative log-likelihood.

Assume that θ is fixed and w_k is given. Then an expression for the likelihood of y_k at θ can easily be obtained by conditioning on the unobserved random variable w_k. If w_k is given, the observation y_k has a simple computable density function $p(y_k|w_k;\theta) = p_e\left(y_k - f(z_k^w(\theta),\theta)\right)$. Furthermore, the likelihood can be written

$$\begin{aligned} p(y_k;\theta) &= \int_{\mathbb{R}} p(y_k|w_k;\theta)p_w(w_k)dw_k \\ &= \int_{\mathbb{R}} p_e\left(y_k - f(z_k^w(\theta),\theta)\right)p_w(w_k)dw_k \\ &= \int_{\mathbb{R}} p_e\left(y_k - f(C(\theta)x_k(\theta) + w_k,\theta)\right)p_w(w_k)dw_k \\ &= \mathbf{E}_{w_k}[p_e\left(y_k - f(z_k^w(\theta),\theta)\right)] \end{aligned} \tag{3}$$

Here, $z_k^w(\theta)$ denotes a simulated version of the unobserved signal z_k using the first row in (1), the known input sequence $\{u_j\}$ with $j = 1,\dots,k$ and the given w_k. The expression in (3) suggests that it is possible to approximate the negative log-likelihood (2) by a Monte-Carlo sum

$$-\widehat{\log(p(Y;\theta))} = -\sum_{k=1}^{N} \log(\hat{p}(y_k;\theta)) \tag{4}$$

in which y_k are the observations and

$$\begin{aligned} \hat{p}(y_k;\theta) &= \frac{1}{M}\sum_{m=1}^{M} p_e\left(y_k - f(C(\theta)x_k(\theta) + w_k^m,\theta)\right), \\ x_{k+1}(\theta) &= A(\theta)x_k(\theta) + B(\theta)u_k, \end{aligned}$$

and w_k^m are samples drawn according to the distribution of the disturbance process

$$\{w_k^m\} \overset{\text{iid over } m}{\sim} p_w$$

At this point, there exist two possibilities. First, it is possible to use the same values $\{w^m\}$ for all k in the Monte-Carlo sum (4). We will refer to this approximation by SML(fixed) which stands for Simulated Maximum-Likelihood with fixed samples. The second possibility is to generate an independent sequence $\{w_k^m\}$ for each different k. We will refer to this approximation by SML(indpt) which stands for Simulated Maximum-Likelihood with independent samples. A comparison between the two possibilities is made in Section V.

Under the assumption that $p(y_k|w_k;\theta)$ has a finite variance under $p_w\, dw$, the variance of $\hat{p}$ is

$$\mathbf{var}(\hat{p}(y_k;\theta)) = \frac{1}{M}\mathbf{E}_{w_k}\left[(p(y_k|w_k;\theta) - \mathbf{E}_{w_k}\left[p(y_k|w_k;\theta)\right])^2\right].$$

The observation to be made here is that, for each fixed θ, the variance depends only on M. A direct application of the strong law of large numbers implies that

$$\hat{p}(y_k;\theta) \overset{\text{a.s.}}{\to} \mathbf{E}_{w_k}\left[p(y_k|w_k;\theta)\right] \quad \text{as } M \to \infty.$$

The central limit theorem gives the asymptotic distribution. For large M and given y_k we have $\hat{p}(y_k;\theta)$ approximately

$$\mathcal{N}\left(\mathbf{E}_{w_k}\left[p(y_k|w_k;\theta)\right], \mathbf{var}(\hat{p}(y_k;\theta))\right).$$

This can be used to develop an expression for the accuracy of the likelihood approximation.

The suggested Simulated Maximum Likelihood (SML) method amounts to the minimization of the expression in (4) with respect to θ to get an approximation of the maximum likelihood estimate.

B. Laplace importance sampling for colored disturbance

If the process disturbance is not white, the observations $\{y_k\}$ cannot be assumed independent, and we are forced to approximate the joint likelihood function directly in $\mathbb{R}^N$. Again, assuming that the model can be simulated for each fixed θ, we can condition on the vector

$$W = \begin{bmatrix} w_1 & w_2 & \dots & w_N \end{bmatrix}^T.$$

The joint likelihood of the output is given by

$$\begin{aligned} p(Y;\theta) &= \int_{\mathbb{R}^N} p(Y|W;\theta)p_W(W)dW \\ &= \int_{\mathbb{R}^N} \left(\prod_{k=1}^{N} p_e(y_k - f(z_k^w(\theta),\theta))\right) p_W(W)dW \\ &= \mathbf{E}_W\left[p(Y|W;\theta)\right] \end{aligned} \tag{5}$$

This is a multidimensional integral over $\mathbb{R}^N$. It is possible to use the model (1) to simulate the output M times, and approximate the above integrals by a Monte-Carlo sum. First, generate M sequences $\{W_m\}_{m=1}^M$ each of length N using the known density $p_W(W)$, then generate the corresponding output. The joint likelihood (5) is approximated directly by

$$\hat{p}(Y;\theta) = \frac{1}{M}\sum_{m=1}^{M}\left(\prod_{k=1}^{N} p_e\left(y_k - y_{k,m}^s(\theta)\right)\right)$$

Here, $y_{k,m}^s(\theta)$ denotes the simulated output at time k.

As the theory suggests, using the above approximation for the likelihood function will give a consistent estimator of θ as both M and N approach infinity. However, in practice this approximation would be (computationally) inefficient and it would require a prohibitively large M. To see this, observe that in high dimensional spaces small local perturbations lead to large global errors. More importantly, any distribution that does not depend on the observation will not be concentrated. This means that most of the draws will have 0 contribution to the likelihood function. This behavior is demonstrated by a numerical example in the simulation study in Section V.

The solution to this issue is to make a change of measure and use instead an "efficient" sampling density. The word efficient here is used in the sense of minimizing the variability in the estimates and the required samples M. This method of changing the measure is known as importance sampling and it is used here to increase the computational efficiency.

To use importance sampling we first write the likelihood in the following form

$$p(Y;\theta) = \int_{\mathbb{R}^N} p(Y|W;\theta)\frac{p_W(W)}{\tilde{p}_W(W|Y;\theta)}\tilde{p}_W(W|Y;\theta)dW$$

with an arbitrary density function $\tilde{p}_W(W|Y,\theta)$ for W which may depend on the observation Y and the parameter θ. The likelihood approximation is then given by generating M sequences $\{W_m\}_{m=1}^M$ each of length N using the importance sampling density $\tilde{p}_W(W|Y;\theta)$, and calculate

$$\hat{p}(Y;\theta) = \frac{1}{M}\sum_{m=1}^{M}\left(\prod_{k=1}^{N} p_e\left(y_k - y_{k,m}^s(\theta)\right)\right)\frac{p_W(W_m)}{\tilde{p}_W(W_m|Y;\theta)}.$$

This approximation is then minimized with respect to θ by a numerical iterative algorithm. In iteration j of the algorithm, $y_{k,m}^s(\theta_j)$ and $\tilde{p}_W(W_m|Y;\theta_j)$ are to be evaluated for the current available value θ_j.

It is easy to see that if we choose $\tilde{p}_W(W|Y;\theta) = p_W(W|Y;\theta)$, the conditional density of W given Y, then only one sample is needed to recover $p(Y;\theta)$. This follows since

$$\frac{p(Y|W;\theta)p_W(W)}{p_W(W|Y;\theta)} = p(Y;\theta), \tag{6}$$

does not depend on W and $\int p_W(W|Y;\theta)dW = 1$

This suggests that a "good" choice for the importance sampling density should be close (in some sense) to the unknown conditional density $p_W(W|Y;\theta)$. The expression in (6) also shows that the computation of $p_W(W|Y;\theta)$ is essentially equivalent to computing $p(Y;\theta)$ (recall that computing $p(Y|W;\theta)$ is simple). It is therefore clear that the real challenge lies in the choice of the importance sampling density.

One method for choosing the importance sampling density is based on the Laplace approximation. The idea behind the Laplace approximation is simple. The method aims at finding an importance sampling density with mean and variance matching the mode and curvature of the unknown conditional density of W. Laplace approximation of the importance sampling density has been used in the context of simulation-based methods in different ways, see for example [4], [3].

To arrive at the approximation, we start by the expression of the likelihood function

$$p(Y;\theta)=\int_{\mathbb{R}^N} p(Y,W;\theta)dW,$$

and assume that the density $p(Y,W;\theta)$ is twice differentiable in W and that for a given observation vector Y and a parameter θ, it has a peak at W_s

$$\begin{aligned} W_s(\theta) &:= \arg\max_{W} \quad p(Y,W;\theta) \\ &= \arg\max_{W} \quad p(W|Y;\theta)p(Y;\theta). \end{aligned}$$

Observe that this is the Maximum a Posteriori (MAP) estimate of W given Y, considering θ as a fixed known value. Then, a Taylor expansion of the log of the joint density around W_s reads

$$\begin{aligned} &\log p(Y,W;\theta) \approx \log p(Y,W_s(\theta);\theta) \\ &+\frac{1}{2}(W-W_s(\theta))^T\frac{\partial^2 \log p(Y,W_s(\theta))}{\partial W \partial W^T}(W-W_s(\theta)). \end{aligned} \tag{7}$$

This shows that $p(Y,W;\theta)=\exp(\log(p(Y,W;\theta))$ is approximately given by the exponential of the second term of the right hand side of (7). This is indeed a normal distribution centered around $W_s(\theta)$ and with the covariance matrix $\Sigma(\theta):=-\left[\frac{\partial^2 \log p(Y,W_s(\theta);\theta)}{\partial W \partial W^T}\right]^{-1}$. This suggests using the importance sampling density

$$\tilde{p}(W|Y;\theta)=\mathcal{N}\left(W_s(\theta),\Sigma(\theta)\right)$$

for a given Y and θ. The reader is refered to Theorem 7.108 in [18] for a rigorous justification of the Laplace approximation.

IV. Computational Issues

In this section, we discuss some computational issues related to the suggested methods.

A. Fixed vs. Independent samples

Consider again the situation of Section III-A where the disturbance process is white. It is clear that using fixed samples $\{w^m\}$ for all k to calculate the approximation

$$\hat{p}(Y;\theta)=\prod_{k=1}^{N}\hat{p}(y_k;\theta)$$

requires shorter computational time than when using independent samples over k. However, doing this will lead to a correlation between the estimates $\hat{p}(y_k;\theta)$. On the other hand, independent samples over k leads to independent estimates $\hat{p}(y_k;\theta)$. In this case,

$$\begin{aligned} \mathbf{E}_W[\hat{p}(Y;\theta)] &= \mathbf{E}_W\prod_{k=1}^{N}\hat{p}(y_k;\theta)=\prod_{k=1}^{N}\mathbf{E}_{w_k}[\hat{p}(y_k;\theta)] \\ &= \prod_{k=1}^{N}\hat{p}(y_k;\theta)=p(Y;\theta) \end{aligned}$$

B. Computation of the likelihood approximation in high dimension

We now describe a solution for a numerical problem that arises when the method in Section III-B is implemented. Assume that both e_k and w_k are Gaussians with variances λ_e and λ_w respectively. For a given θ, and a corresponding importance sampling density, the likelihood approximation is evaluated by calculating $p(Y|W;\theta)p_W(W)/\tilde{p}_W(W|Y;\theta)=$

$$\frac{c_1\exp(-\frac{1}{2\lambda_e}\|Y-Y^W(\theta)\|^2)c_2\exp(-\frac{1}{2\lambda_w}\|W\|^2)}{c_3(\theta)\exp(\frac{1}{2}\|W-W_s(\theta)\|^2}), \tag{8}$$

$$\text{with} \qquad c_1=\frac{1}{(2\pi\lambda_e)^{\frac{N}{2}}}, \qquad c_1=\frac{1}{(2\pi\lambda_w)^{\frac{N}{2}}}$$

$$c_3(\theta)=\frac{1}{(2\pi)^{\frac{N}{2}}\det\Sigma(\theta)^{-1}}$$

For large values of N, a direct calulation of this expression is not possible. First, the value $\det\Sigma(\theta)^{-1}$ will be a very small number. Second, the constants c_1 and c_2 will be very large. Furthermore, it is likely that the arguments of the exponential function will be too large making the exponential function equal to zero for any computer with finite precision. This issue can be solved by first taking the logarithm of the fraction in (8) and then applying the exponential function to the result. The logarithm transforms products into sums and exponents into scaling factors, which makes the numbers more tractable. In addition, notice that the whole expression can be normalized by any constant that is independent of θ.

V. Simulation Study

All the numerical examples were implemented in MATLAB 2015b on an Intel-based laptop with a 2.7 GHz processor and 8 Gbyte RAM. The function `fminsearch` was used to find the minimizer of the negative log of the likelihood approximation. For the optimization step of the Laplace importance sampling, the Matlab package IPOPT [21] was used. This is a software package for large-scale nonlinear optimization based on interior-point algorithms. It is designed to find (local) solutions of constrained nonlinear optimization problems. IPOPT requires the gradient and the Hessain of the objective functions to be calculated analytically and provided to the solver as function handles.

A. First order stochastic Wiener system with direct sampling

We first consider the following FIR model with cubic nonlinearity at the output

$$\begin{aligned} z_k &= \theta u_k + u_{k-1} + w_k, \\ y_k &= \ \ z_k^3 + e_k, \quad u_0 = 0 \\ u_k &\sim \mathcal{N}(0,\lambda_u) \quad \forall k, \quad \mathbf{E}[u_k u_j]=0 \ \ \forall k\neq j, \\ e_k &\sim \mathcal{N}(0,\lambda_e) \quad \forall k, \quad \mathbf{E}[e_k e_j]=0 \ \ \forall k\neq j, \\ w_k &\sim \mathcal{N}(0,\lambda_w) \quad \forall k, \quad \mathbf{E}[w_k w_j]=0 \ \ \forall k\neq j, \\ e_k &\perp w_j \perp u_l \qquad \forall k,j,l \in \{1,2,3,\dots\} \end{aligned}$$

this example is taken from [22] where the authors compare the MLE and an indirect inference method based on a best linear approximation model. The process disturbance here is white. Therefore, we can apply the direct sampling method from section III-A. The suggested simulated-maximum likelihood was implemented for a model with

$$\theta = 0.5, \quad \lambda_u = \frac{1}{3}, \quad \lambda_e = 0.1 \quad \lambda_w = 0.2$$

and compared to the ML estimated calculated with Gauss-Hermite approximation as suggested in [22]. The number of observations $N = 1000$, and the number of Monte-Carlo samples $M = 5000$ and the number of points for the Gauss-Hermite approximation is taken to be the same as M. The Output-Error estimate (OE) (considering $W = 0$) is also calculated. The average results over 1000 disturbance, noise and input realizations are summarized in the following table

	Mean	std	MSE
ML:	0.4991	0.0311	0.0010
SML(fixed):	0.5071	0.0507	0.0026
OE:	0.6246	0.0709	0.0205

It is clear that the simulated maximum-likelihood estimate is unbiased but worse than the maximum-likelihood estimate given by the Gauss-Hermite quadrature. On the other hand, the OE estimate that ignores the process disturbance is clearly biased.

If we increase M to 50000, and use independent samples for each k as explained in section III-A, we get the following average result over 100 Monte-Carlo iterations

	Mean	std	MSE
ML:	0.4988	0.0310	9.53×10^{-4}
SML(indpt):	0.4985	0.0319	0.0010
OE:	0.6129	0.0722	0.0179

We see a clear improvement in the resulting estimate as suggested by the theory. Comparing this result to the results obtained in [22] for the same example, we see that by increasing the number of Monte-Carlo samples used for the approximation, we can achieve better accuracy compared to the indirect inference method with optimal weighting.

We repeat the last experiment and compare the case of fixed sample for all k against independent samples over k. First, for $M = 5000$, we get the following average result over 500 Monte-Carlo iterations

	Mean	std	MSE
ML:	0.5008	0.0320	0.0010
SML(fixed):	0.5061	0.0517	0.0027
SML(indpt):	0.5060	0.0535	0.0029
OE:	0.6295	0.0698	0.0192

Both cases with fixed and independent samples have comparable performance. On the other hand, for $M = 50000$, we get the following average result over 100 Monte-Carlo iterations we get

	Mean	std	MSE
ML:	0.5011	0.0315	9.82×10^{-4}
SML (fixed):	0.5034	0.0374	0.0014
SML (indpt):	0.5003	0.0321	0.0010
OE:	0.6215	0.0695	0.0195

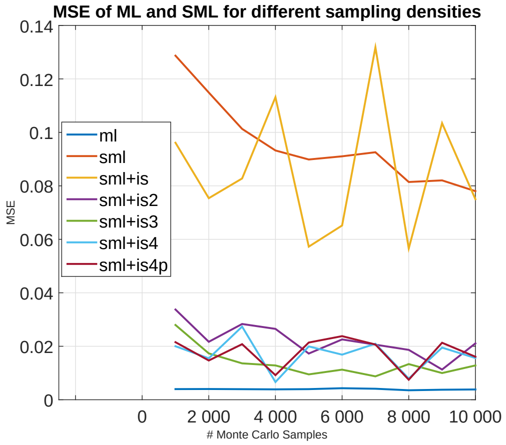

Fig. 2: Illustration of Importance Sampling

From this we see that using a fixed sample gives a slightly worse result, as was discussed in Section IV-A.

B. First order linear system

Consider the linear-time invariant state-space model structure

$$\begin{aligned} x_{k+1} &= \theta x_k + w_k, \\ y_k &= \quad x_k + e_k, \quad x_0 = 0 \\ e_k &\sim \mathcal{N}(0, \lambda_e) \quad \forall k, \quad \mathbf{E}[e_k e_j] = 0 \quad \forall k \neq j, \\ w_k &\sim \mathcal{N}(0, \lambda_w) \quad \forall k, \quad \mathbf{E}[w_k w_j] = 0 \;\; \forall k \neq j, \\ e_k &\perp w_j \;\; \forall k, j, l \in \{1, 2, 3, \dots\} \end{aligned}$$

Here we did not apply a nonlinearity at the ouput. However, doing this is straightforward. The motivation for considering a linear system is to be able to calculate the MLE analytically for comparison. Observe that adding a nonlinearity at the output results in a blind identification problem similar to the one considered in [25]. Such a problem has attracted some attention; however usually the proposed solutions require stringent assumptions.

For a first experiment we fix $N = 200$ and generate data using the above linear system with the following parameters

$$\theta = 0.7, \quad \lambda_w = 1.5, \quad \text{and } \lambda_e = 1$$

Since we have a linear model with Gaussian disturbance and measurements noise, the likelihood function has a known analytical form. The MLE can be computed directly using the analytical likelihood function and is to be used as a reference for performance evaluation.

We start by showing the effect of importance sampling. We simulated the estimator over 1000 realizations for each $M = 1000 : 1000 : 10000$ using different importance sampling densities.

Figure 2 shows the result for the MLE and different sampling densities. The MLE is denoted by `ml`, and is of course independent of M.
`sml` denotes direct sampling using $p_W(W)$.
`sml+is` denotes sampling using $\mathcal{N}(W_s, \lambda_w I_N)$,
`sml+is2` denotes sampling using $\mathcal{N}(W_s, \frac{\lambda_w}{2} I_N)$,

sml+is3 denotes sampling using $\mathcal{N}(W_s, \frac{\lambda_w}{3} I_N)$,
sml+is4 denotes sampling using $\mathcal{N}(W_s, \frac{\lambda_w}{4} I_N)$,
Finally, sml+is4p denotes sampling from $\mathcal{N}(W_s + \varepsilon, \frac{\lambda_w}{4} I_N)$ with $\varepsilon = ((-1 + 2 * \texttt{rand}(N, 1)) * 0.1)$

It is clear from this result that naïve sampling using the distribution of the process disturbance is extremely inefficient. Using any sampling density centered at the maximizer W_s and with variance less than λ_w decreases the MSE considerably. It also seems that the performance is not very sensitive to the used mean and variance as long as the resulting $\tilde{p}$ is a reasonable representation of $p(W|Y;\theta)$

By applying the Laplace importance sampling method, we can achieve good results. To demonstrate this, we fix $N = 1000$ and generate data using the above linear system with the same choice for the parameters. We choose $M = 1000$ and simulate the method over 100 disturbance and noise realizations. We get the following

	Mean	std	MSE
ML:	0.6963	0.0280	7.88×10^{-4}
SML+IS:	0.6963	0.0280	7.88×10^{-4}

The result shows that the simulated likelihood method with the importance sampling is exact for linear systems (only one sample was actually needed).

C. First order stochastic Wiener system with Importance Sampling

In this last experiment, we evaluate the behavior of the simulated maximum-likelihood method on the same Wiener model as in Section V-A. Again, the number of observations $N = 1000$, and the number of Monte-Carlo samples $M = 5000$ and the number of points for the Gauss-Hermite approximation is taken to be the same as M. The average result over 100 Monte-Carlo iterations are as follows

	Mean	std	MSE
ML:	0.4944	0.0319	0.0010
SML:	0.5151	0.0469	0.0024
OE:	0.6234	0.0701	0.0201

The result shows that using the importance sampling in high dimension (N=1000) gives almost the same result as the direct sampling with the whiteness assumption.

VI. CONCLUSIONS

We have introduced a simulated maximum likelihood method that can be used for stochastic Wiener systems identification. The initial simulation study shows that the method has a good performance compared to known analytical and other approximate methods.

The computation time for all the presented examples is in the order of minutes. However, the required time increases with the number of estimated parameters. For example, estimating a system with 8 unknown parameters using 500 input-output samples and 10^4 Monte-Carlo samples and the true parameters as initial value takes about 3.5 hours. Nevertheless, in this preliminary study, computational speed has not been addressed, and the implementation is not optimized in any way.

REFERENCES

[1] E.-W. Bai. Frequency domain identification of wiener models. *Automatica*, 39(9):1521 – 1530, 2003.

[2] S. Boyd and L. Chua. Fading memory and the problem of approximating nonlinear operators with volterra series. *IEEE Transactions on Circuits and Systems*, 32(11):1150–1161, Nov 1985.

[3] C. N. Brinch. Efficient simulated maximum likelihood estimation through explicitly parameter dependent importance sampling. *Computational Statistics*, 27(1):13–28, 2012.

[4] J. Durbin and S. Koopman. Monte Carlo maximum likelihood estimation for non-Gaussian state space models. 84(3):669–684, 1997.

[5] C. Gourieroux and A. Monfort. Simulation-based inference: A survey with special reference to panel data models. *Journal of Econometrics*, 59(12):5 – 33, 1993.

[6] C. Gouriroux and A. Monfort. *Simulation-based Econometric Methods*. Oxford University Press, 1996.

[7] W. Greblicki. Nonparametric identification of wiener systems. *IEEE Transactions on Information Theory*, 38(5):1487–1493, Sep 1992.

[8] A. Hagenblad, L. Ljung, and A. Wills. Maximum likelihood identification of wiener models. *Automatica*, 44(11):2697 – 2705, 2008.

[9] W. I. Hunter and M. J. Korenberg. The identification of nonlinear biological systems: Wiener and hammerstein cascade models. *Biol. Cybern.*, 55(2-3):135–144, Nov. 1986.

[10] W. Jank. Efficient simulated maximum likelihood with an application to online retailing. *Statistics and Computing*, 16(2):111–124, 2006.

[11] A. Kalafatis, N. Arifin, L. Wang, and W. Cluett. A new approach to the identification of ph processes based on the wiener model. *Chemical Engineering Science*, 50(23):3693 – 3701, 1995.

[12] L. Ljung. *System Identification (2Nd Ed.): Theory for the User*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1999.

[13] C. E. Mcculloch. Maximum Likelihood Algorithms for Generalized Linear Mixed Models. 92(437):162–170, 2011.

[14] B. Ninness, A. Wills, and T. Schon. Estimation of general nonlinear state-space systems. In *49th IEEE Conference on Decision and Control, Atlanta, Georgia, USA*, pages 1–6, dec 2010.

[15] R. J. G. Peter J. Diggle. Monte carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(2):193–227, 1984.

[16] G. Pillonetto. Consistent identification of wiener systems: A machine learning viewpoint. *Automatica*, 49(9):2704 – 2712, 2013.

[17] J. F. Richard and W. Zhang. Efficient high-dimensional importance sampling. *Journal of Econometrics*, 141(2):1385–1411, 2007.

[18] M. Schervish. *Theory of Statistics*. Springer Series in Statistics. Springer New York, 1996.

[19] T. B. Schön, F. Lindsten, J. Dahlin, J. Wågberg, C. A. Naesseth, A. Svensson, and L. Dai. Sequential monte carlo methods for system identification. *IFAC-PapersOnLine*, 48(28):775 – 786, 2015. 17th IFAC Symposium on System Identification SYSID 2015 Beijing, China, 1921 October 2015.

[20] T. B. Schön, A. Wills, and B. Ninness. System identification of nonlinear state-space models. *Automatica*, 47(1):39 – 49, 2011.

[21] A. Wächter and T. L. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106(1):25–57, 2005.

[22] B. Wahlberg, J. Welsh, and Lennart. Identification of stochastic wiener systems using indirect inference. *IFAC-PapersOnLine*, 48(28):620 – 625, 2015. 17th IFAC Symposium on System Identification SYSID 2015 Beijing, China, 1921 October 2015.

[23] D. Westwick and M. Verhaegen. Identifying mimo wiener systems using subspace model identification methods. *Signal Processing*, 52(2):235 – 258, 1996. Subspace Methods, Part II: System Identification.

[24] T. Wigren. Recursive prediction error identification using the nonlinear wiener model. *Automatica*, 29(4):1011 – 1025, 1993.

[25] A. Wills, T. Schön, L. Ljung, and B. Ninness. Blind identification of wiener models. volume 18, pages 5597–5602, 2011.

[26] A. Wills, T. B. Schön, L. Ljung, and B. Ninness. Identification of hammerstein-wiener models. *Automatica*, 49(1):70–81, Jan. 2013.

[27] Y. Zhu. Distillation column identification for control using wiener model. In *American Control Conference, 1999. Proceedings of the 1999*, volume 5, pages 3462–3466 vol.5, 1999.