2024 European Control Conference (ECC)
June 25-28, 2024. Stockholm, Sweden

Online Identification of Stochastic Continuous-Time Wiener Models Using Sampled Data

Mohamed Abdalmoaty[1], Efe C. Balta[2], John Lygeros[1], and Roy S. Smith[1]

***Abstract*— It is well known that ignoring the presence of stochastic disturbances in the identification of stochastic Wiener models leads to asymptotically biased estimators. On the other hand, optimal statistical identification, via likelihood-based methods, is sensitive to the assumptions on the data distribution and is usually based on relatively complex sequential Monte Carlo algorithms. We develop a simple recursive online estimation algorithm based on an output-error predictor, for the identification of continuous-time stochastic parametric Wiener models through stochastic approximation. The method is applicable to generic model parameterizations and, as demonstrated in the numerical simulation examples, it is robust with respect to the assumptions on the spectrum of the disturbance process.**

I. INTRODUCTION

Online system identification is a classical problem in the Systems and Control literature. Several methods and algorithms were developed in parallel to the development of adaptive control techniques; see [1], [2]. Apart from its direct interest, online identification has close connections to nonlinear filtering and learning & adaptation; see [3]–[5]. Recursive algorithms are also useful for offline identification when the data sets are large.

The majority of classical prediction error methods (PEM) were designed for linear stochastic models, or nonlinear deterministic models, with disturbances or noise solely at the output, [6, Ch. 11], [3, Ch. 8]. For general nonlinear stochastic discrete-time state-space models, likelihood-based estimators such as the maximum-likelihood estimator or the maximum-a-posteriori estimator are usually used. Both offline and online implementations of such estimators rely on (particle) sequential Monte Carlo approximations, see for example the survey [7] and the references therein. Similar algorithms for continuous-time nonlinear time-series models were proposed as offline methods using sampled measurements [8] and as online algorithms using continuous-time observations [9]. Unfortunately, besides the computational challenges, the maximum-likelihood estimator deviates from its optimal asymptotic properties when there are discrepancies in the data distribution [10].

Most of the existing methods have only asymptotic guarantees, mainly due to the intractability of the finite-sample distributions of the estimators. Some major advantages of asymptotic analysis are the applicability to common model parameterizations in both continuous- and discrete-time, and the ability to work with model misspecification under weak assumptions; a clear drawback is the lack of finite-sample guarantees. System identification of linear systems from a single input-output trajectory using linear least-squares estimators with non-asymptotic guarantees has been studied in [11] and [12]. Most of the existing work employs stochastic input sequences from a predetermined distribution to provide convergence guarantees; see [13] for an overview of recent results. Generalized linear, nonlinear [14], and piecewise-affine systems [15] have also been studied. However, existing algorithms often focus on discrete-time dynamics and are not easily adaptable to continuous-time.

In this work, we focus on the online estimation problem of continuous-time stochastic parametric Wiener models using noisy samples of the output signal. A Wiener model comprises a linear dynamical model, followed by a static nonlinearity at its output. It has found application in different scientific and engineering domains [16] as it can approximate a large class of nonlinear systems. Because the dynamic component is linear, exact time-discretization is possible when the inter-sample behavior of the input signal is known. Moreover, the stability of the model is dictated by the stability of the linear component. Previous work on recursive online identification of Wiener models, as in [17], has only considered the deterministic case in discrete time. In [18], an online PEM algorithm for deterministic nonlinear continuous-time models is given using Euler discretization.

We propose a simple online parameter estimation algorithm for the class of continuous-time stochastic Wiener models. We utilize an output-error predictor and adopt an input-output approach, accommodating a sampled data scenario with additive output measurement noise. The algorithm is developed using an online prediction error framework. Through numerical simulation examples, we showcase the algorithm's performance, especially in cases where the disturbance model is incorrectly specified. It should be noted that even though we consider continuous-time models, the approach is directly applicable to discrete-time models.

II. PROBLEM FORMULATION

We consider the following class of Wiener models

$$\begin{aligned} \mathrm{d}w(t) &= A(\theta)w(t)\,\mathrm{d}t + B(\theta)\,\mathrm{d}\beta(t), \\ x(t) &= G(\mathrm{p};\theta)u(t) + C(\theta)w(t), \\ y(t) &= f(x(t);\theta), \end{aligned} \tag{1}$$

This work has been supported by the Swiss National Science Foundation under NCCR Automation (grant agreement 51NF40_180545)

[1]Automatic Control Laboratory (IfA) and NCCR Automation, Swiss Federal Institute of Technology (ETH Zürich), 8092 Zürich, Switzerland, {mabdalmoaty,jlygeros,rsmith}@control.ee.ethz.ch
[2]Control and Automation Group, Inspire AG, 8005 Zürich, Switzerland efe.balta@inspire.ch

where $G(\mathrm{p};\theta)$ is a single-input single-output continuous-time transfer operator, p is the differential operator, $f(\cdot;\theta)$ is a static parametric function, $\theta \in \Theta \subset \mathbb{R}^d$ is a parameter vector, $u(t) \in \mathbb{R}$ is the input signal, $y(t) \in \mathbb{R}$ is the output signal, $x(t) \in \mathbb{R}$ is a latent signal, and $w(t) \in \mathbb{R}^{n_w}$ is a disturbance driven by a Wiener process $\beta(t) \in \mathbb{R}^{n_w}$. Additionally, $A(\theta), B(\theta), C(\theta)$ are parametric matrices of appropriate sizes. We assume the output is measured at discrete-time instances t_k with additive measurement noise,

$$y_k = y(t_k) + v_k, \quad k = 1, 2, 3, \dots \tag{2}$$

To simplify the exposition, we confine our discussion to a constant sampling period Δ. However, the suggested method is applicable to the more general case with irregular sampling times. Without loss of generality, we assume that v_k has a zero-mean value for all k (non-zero constant mean values can be included in the parameterization of f). We also assume that the input signal is known exactly as a continuous-time signal, and therefore the data set available at time t_N is

$$D_N := \big\{ (y_k, u(t)) : k = 1, \dots, N,\ t \in [t_1, t_N] \big\}.$$

Moreover, the data is collected open-loop so that u is independent of w and v. To ensure that the data collection process is well-posed, model (1) is assumed to be (asymptoticaly) stable for all $\theta \in \Theta$. The choice of a transfer operator parametrization is

$$G(\mathrm{p};\theta) = \frac{\sum_{j=0}^{m} c_j \mathrm{p}^j}{\mathrm{p}^n + \sum_{j=0}^{n-1} d_j \mathrm{p}^j}, \tag{3}$$

where $m \leq n$, $\theta_G := [c_0 \ \dots \ c_m \ d_0 \ \dots d_{n-1}]^\top \in \mathbb{R}^{d_G}$ and $d_G = n + m + 1$. Other parametrizations, e.g., state space (canonical or not), are also possible.

The parametrization of the Itô stochastic differential equation used to model the disturbance w is done separately from G and f, allowing for the possibility of misspecification only in the disturbance model. The matrices $A(\theta)$, $B(\theta)$, and $C(\theta)$ assume a state-space parametrization that should be identifiable from the marginal second-order moments of y. This typically means that only a few parameters in one of the matrices can be estimated. Nevertheless, the proposed approach can naturally handle cases where G, f, and the model of w are jointly parameterized (see Section IV-A) when the parameterization is identifiable. We order the entries of the parameter vector as $\theta = [\theta_G^\top \ \theta_w^\top \ \theta_f^\top]^\top$, where $\theta_w \in \mathbb{R}^{d_w}$ are parameters appearing in the disturbance model, and $\theta_f \in \mathbb{R}^{d_g}$ are parameters of f. The main objective of the paper is the construction of an online estimation algorithm for θ that, based on the knowledge of the inter-sample behavior of u, maps the current measurement y_k to an estimate $\hat{\theta}_k$. When the system is time-invariant, an appropriate algorithm design ensures the almost sure convergence to a subset of Θ.

III. Proposed Approach

The Output-Error Quadratic PEM (OE-QPEM) [19] estimator based on D_N is defined as the minimizer of

$$V_N(\theta) := \frac{1}{N} \sum_{k=1}^{N} \frac{1}{2} \left(y_k - \mathbb{E}\left[y_k | (u(s))_{s=t_1}^{t_k}; \theta \right] \right)^2, \tag{4}$$

over a suitable compact subset $\Theta \subset \mathbb{R}^d$. The expectation operator is with respect to the process disturbance w and the measurement noise v, and is conditioned on the known input signal u. The OE-QPEM estimator provides a computationally simpler alternative to likelihood-based methods, as it does not require the computation (or approximation) of the predictive densities $p(y_k | y_1, \dots, y_{k-1}, (u(s))_{s=t_1}^{t_k}; \theta)$ of the model's output. The loss of statistical efficiency is often outweighed by the computational simplicity it offers (as shown e.g. in [19] and [20]), and the applicability to complex models; see e.g., [21].

A. Proposed identification method

We propose an online implementation of the OE-QPEM estimator. In a similar vein to (4), we seek to minimize $V(\theta) = \frac{1}{2}\mathbb{E}[\varepsilon_k^2(\theta)]$, where $\varepsilon_k(\theta) := y_k - \mathbb{E}\left[y_k | (u(s))_{s=t_1}^{t_k}; \theta \right]$ is the prediction error, using stochastic approximation. Let us denote the predictor and the gradient vector of the prediction error with respect to θ as

$$\hat{y}_k(\theta) = \mathbb{E}\left[y_k | (u(s))_{s=t_1}^{t_k}; \theta \right], \quad \psi_k(\theta) = -\frac{\mathrm{d}}{\mathrm{d}\theta} \hat{y}_k(\theta),$$

respectively. Then $V'(\theta) = \mathbb{E}\left[\psi_k(\theta) \left(y_k - \hat{y}_k(\theta) \right) \right]$, where we have allowed the interchange of expectation and differentiation. Notice that the outer expectation pertains to the underlying probability space of the data (unknown), while the inner expectations, defining $\hat{y}_k(\theta)$ and $\psi_k(\theta)$, are with respect to the Wiener process $\beta(t)$ in (1). Then, minimizing $V(\theta)$ can be achieved by solving the system of equations

$$\mathbb{E}\left[\psi_k(\theta) (y_k - \mathbb{E}\left[y_k | (u(s))_{s=t_1}^{t_k}; \theta \right] \right] = 0. \tag{5}$$

Applying the Robbins-Monro stochastic approximation scheme [22], we obtain the following recursion

$$\hat{\theta}_k = \hat{\theta}_{k-1} - \gamma_k \psi_k(\hat{\theta}_{k-1}) \left(y_k - \hat{y}_k(\hat{\theta}_{k-1}) \right), \tag{6}$$

in which γ_k are positive scalars tending to zero sufficiently slowly as k grows. There are two primary challenges in computing the OE predictor $\hat{y}_k(\hat{\theta}_{k-1})$ and the gradient vector $\psi_k(\hat{\theta}_{k-1})$. The first is the evaluation of the expected values with respect to β, and the second is doing so online in a recursive manner. The solution to the first challenge is generic in nature, while the second naturally depends on the choice of model class and parameterization.

The *main idea* of our approach is to compute the OE predictor $\hat{y}_k(\hat{\theta}_{k-1})$ and the corresponding prediction error gradient vector $\psi_k(\hat{\theta}_{k-1})$ in (6) by simulating (1) and its output gradient filters using two independent Wiener processes $\beta^{(y)}(t)$ and $\beta^{(\psi)}(t)$, respectively, at $\theta = \hat{\theta}_{k-1}$.

The outputs of these two simulations, denoted $y_{1,k}(\hat{\theta}_{k-1})$ and $\psi_{1,k}(\hat{\theta}_{k-1})$, are unbiased estimators of $\hat{y}_k(\hat{\theta}_{k-1})$ and $\psi_k(\hat{\theta}_{k-1})$, and are independent by construction. This important property means that the vector

$$\psi_{1,k}(\hat{\theta}_{k-1}) \left(y_k - y_{1,k}(\hat{\theta}_{k-1}) \right) \tag{7}$$

is an unbiased estimator of the estimating function in (5). While only two Wiener processes are needed, the performance of the algorithm may be improved by considering the average of $M \geq 1$ independent simulations:

$$\bar{y}_k(\hat{\theta}_{k-1}) = \frac{1}{M}\sum_{m=1}^{M} y_{m,k}(\hat{\theta}_{k-1}),$$

and similarly for $\bar{\psi}_k(\hat{\theta}_{k-1})$. These unbiased estimators can be thought of as "measurements" of $\hat{y}_k(\hat{\theta}_{k-1})$ and $\psi_k(\hat{\theta}_{k-1})$, respectively. With this in mind, an approximation of (6) is

$$\hat{\theta}_k = \hat{\theta}_{k-1} - \gamma_k \bar{\psi}_k(\hat{\theta}_{k-1})\left(y_k - \bar{y}_k(\hat{\theta}_{k-1})\right).$$

Note that M can be either fixed or changed with k, and could be thought of as a tuning parameter of the algorithm.

To further improve the convergence properties, a stochastic Newton direction can be used. For a fixed θ, the Hessian $V''(\theta) = \mathbb{E}[\psi_k(\theta)\psi_k^\top(\theta)]$ can be determined as the solution R of $\mathbb{E}[\psi_k(\theta)\psi_k^\top(\theta) - R] = 0$ where the expectation is with respect to the data distribution. Using the ideas outlined above, we arrive at the following algorithm

$$R_k = R_{k-1} + \gamma_k\left[\bar{\psi}_k(\theta_{k-1})\left[\bar{\psi}_k(\theta_{k-1})\right]^\top - R_{k-1}\right],$$
$$\hat{\theta}_k = \left[\hat{\theta}_{k-1} - \gamma_k R_k^{-1}\bar{\psi}_k(\hat{\theta}_{k-1})\left(y_k - \bar{y}_k(\hat{\theta}_{k-1})\right)\right]_\Theta,$$

where $[\,\cdot\,]_\Theta$ is a projection operator. A particularly simple implementation defines $[\,\theta\,]_\Theta := \theta$ if $\theta \in \Theta$, otherwise $[\,\theta\,]_\Theta = \hat{\theta}_{k-1}$ (see [6, (11.50)]). The only issue with this algorithm is that it is not recursive. The estimates of the OE predictor and its gradient vector at time t_k are the outputs of filters with infinite impulse responses, and hence they rely on all past data in general. This issue is fixed below using approximation processes similar to those used in classical online PEM algorithms.

B. Recursive computation of the OE predictor

We construct a natural recursive approximation $\bar{y}_k$ of $\bar{y}_k(\hat{\theta}_{k-1})$. Define $z(t;\theta) = G(\mathrm{p};\theta)u(t)$, $z_k(\theta) = z(t_k;\theta)$. For the sake of clarity, we let the input be constant over the sampling interval. The sampled-data transfer function[1] is then

$$G_\Delta(\mathrm{z}^{-1};\theta) := (1-\mathrm{z}^{-1})\mathcal{Z}\left\{\mathcal{L}^{-1}\left\{\frac{G(\mathrm{p};\theta)}{\mathrm{p}}\right\}_{t=k\Delta}\right\}$$
$$=: \frac{\sum_{r=1}^{n} b_r(\theta)\mathrm{z}^{-r}}{1+\sum_{r=1}^{n} a_r(\theta)\mathrm{z}^{-r}}$$

When the input is not constant over the sampling interval, and/or the sampling times are irregular, the model is discretized exactly by using the knowledge of the inter-sample behaviour (see Section IV-A). For any fixed θ, the recursion

$$z_k(\theta) = -a_1(\theta)z_{k-1}(\theta) - \cdots - a_n(\theta)z_{k-n}(\theta) + b_1(\theta)u_{k-1} + \cdots + b_n(\theta)u_{k-n}.$$

holds exactly. A recursive approximation z_k of $z_k(\hat{\theta}_{k-1})$ is then obtained with the current estimate $\hat{\theta}_{k-1}$, using previous values of z_k as initial values. We denote the approximation of $z_k(\hat{\theta}_{k-1})$ compactly by

$$z_k = \varphi_{k-1}^\top \eta(\hat{\theta}_{k-1}), \tag{8}$$

where $\varphi_{k-1} = \left[-z_{k-1}\ \ldots\ -z_{k-n}\ u_k\ \ldots\ u_{k-n}\right]^\top$, and

$$\eta(\hat{\theta}_{k-1}) = \left[a_1(\hat{\theta}_{k-1})\ldots a_n(\hat{\theta}_{k-1})\ b_1(\hat{\theta}_{k-1})\ldots b_n(\hat{\theta}_{k-1})\right]^\top.$$

[1]This is simply zero-order hold sampling. $\mathcal{L}^{-1}\{\cdot\}$ is the inverse Laplace transform, $\mathcal{Z}\{\cdot\}$ is the Z-transform, and z is the Z-transform variable.

Likewise, a recursive approximation, denoted as $w_{m,k}$, is derived for $w_m(t_k;\hat{\theta}_{k-1})$. Because the model of w is linear, it can be sampled exactly (in the sense that the statistical properties of the sampled model are identical to that of the continuous-time one at the sampling times; see e.g., [23, Ch.3, Sec.10, pages 82-83]) to get

$$w_{m,k+1}(\theta) = A_\Delta(\theta)w_{m,k}(\theta) + B_\Delta(\theta)\beta_{m,k}^{(y)},$$
$$\beta_{m,k}^{(y)} \overset{\text{iid}}{\sim} \mathcal{N}(0, I_{n_w}) \tag{9}$$

where $A_\Delta(\theta) = e^{A(\theta)\Delta}$ and $B_\Delta(\theta)$ is a square root of $\int_0^\Delta e^{A(\theta)s}B(\theta)B^\top(\theta)e^{A^\top(\theta)s}\,\mathrm{d}s$. The approximation process is then computed recursively as

$$w_{m,k} = A_\Delta(\hat{\theta}_{k-1})w_{m,k-1} + B_\Delta(\hat{\theta}_{k-1})\beta_{m,k}^{(y)} \tag{10}$$

where at time t_k we only need to store $\{w_{m,k}\}_{m=1}^M$. The random variables $\beta_{m,k}^{(y)}$ are sampled in run-time and not stored. Finally, we define

$$\bar{y}_k = \frac{1}{M}\sum_{m=1}^{M} f\left(z_k + C(\hat{\theta}_{k-1})w_{m,k}\,;\hat{\theta}_{k-1}\right)$$

which only requires storing φ_k, $\{w_{m,k}\}_{m=1}^M$, and $\hat{\theta}_k$ at t_k.

C. Recursive computation of the gradient vector

Define $x_m(t;\theta) = z(t;\theta) + C(\theta)w_m(t;\theta)$, where $w_m(t;\theta)$ is driven by $\beta_m^{(\psi)}$. Applying the chain rule,

$$\partial_{\theta_j}[x_m(t;\theta)] = \partial_{\theta_j}[z(t;\theta)] + C_j(\theta)\,w_m(t;\theta) + C(\theta)\partial_{\theta_j}[w_m(t;\theta)], \tag{11}$$

where $C_j(\theta)$ is the entry-wise derivative of $C(\theta)$ with respect to θ_j. Similarly, we have

$$\partial_{\theta_j}[y_m(t;\theta)] = \partial_{\theta_j}[f(a;\theta)]\big|_{a=x_m(t;\theta)} + \partial_x[f(x;\theta)]\big|_{x=x_m(t;\theta)}\partial_{\theta_j}[x_m(t;\theta)], \tag{12}$$

and $\partial_{\theta_j}[z(t;\theta)] = G_j'(p;\theta)u(t)$, $1 \le j \le d_G$ with

$$G_j'(\mathrm{p};\theta) = \begin{cases} \frac{\mathrm{p}^j}{\mathrm{p}^n + \sum_{j=1}^n d_j\mathrm{p}^j}, & 1 \le j \le m+1 \\ \frac{-\mathrm{p}^{j-m-1}}{\mathrm{p}^n + \sum_{j=1}^n d_j\mathrm{p}^j}G(\mathrm{p};\theta)u(t), & m+2 \le j \le d_G \end{cases}$$

The gradient filters $G_j'(\mathrm{p};\theta)$ can be discretized similarly to $G(\mathrm{p};\theta)$, based on the inter-sample behaviour of $u(t)$, to get

$$G_j'(\mathrm{z}^{-1};\theta) := \frac{\sum_{r=1}^{n_j} b_r^{(j)}(\theta)\mathrm{z}^{-r}}{1+\sum_{r=1}^{n_j} a_r^{(j)}(\theta)\mathrm{z}^{-r}},$$

in which $n_j = n$ for $1 \le j \le m+1$, while $n_j = 2n$ for $m+2 \le j \le d_G$. Analogous to (8), a recursive approximation $z_k^{(j)}$ of $\partial_{\theta_j}[z(t_k;\theta)]$ is defined as

$$z_k^{(j)} = [\varphi_{k-1}^{(j)}]^\top \eta_j(\hat{\theta}_{k-1}),$$

in which $\varphi_{k-1}^{(j)} = \left[-z_{k-1}^{(j)}\ \ldots\ -z_{k-n_j}^{(j)}\ u_{k-1}\ \ldots\ u_{k-n_j}\right]^\top$,

$$\eta_j(\hat{\theta}_{k-1}) = \left[a_1^{(j)}(\hat{\theta}_{k-1})\ldots a_{n_j}^{(j)}(\hat{\theta}_{k-1})\ b_1^{(j)}(\hat{\theta}_{k-1})\ldots b_{n_j}^{(j)}(\hat{\theta}_{k-1})\right]^\top.$$

On the other hand, the gradients $\partial_{\theta_j}[w_m(t;\theta)]$, denoted as $w_m^{(j)}(t;\theta)$ in the sequel, are obtained by differentiating the

stochastic integral equations defining $w_m(t;\theta)$ with respect to θ. It can be shown that they satisfy

$$\mathrm{d}\zeta^{(j)}(t;\theta) = F^{(j)}(\theta)\zeta^{(j)}(t;\theta)\,\mathrm{d}t + L^{(j)}(\theta)\,\mathrm{d}\beta_m^{(\psi)}(t) \quad (13)$$

in which $\zeta^{(j)}(t;\theta) = \begin{bmatrix}[w_m(t;\theta)]^\top & [w_m^{(j)}(t;\theta)]^\top\end{bmatrix}^\top$, with the following drift and dispersion matrices

$$F^{(j)}(\theta) = \begin{bmatrix} A(\theta) & 0 \\ A_j(\theta) & A(\theta)\end{bmatrix}, \qquad L^{(j)}(\theta) = \begin{bmatrix} B(\theta) \\ B_j(\theta)\end{bmatrix}.$$

Here, $A_j(\theta)$ and $B_j(\theta)$ are defined similarly to $C_j(\theta)$. Notice that $\beta_m^{(\psi)}$ in (13) and $\beta_m^{(y)}$ in (10) are independent processes by definition. The sampled versions of $\zeta^{(j)}(t;\theta)$ are

$$\begin{aligned}\zeta_{m,k+1}^{(j)}(\theta) &= F_\Delta^{(j)}(\theta)\zeta_{m,k}^{(j)}(\theta) + L_\Delta^{(j)}(\theta)\beta_{m,k}^{(\psi)} \\ \beta_{m,k}^{(\psi)} &\overset{\text{iid}}{\sim} \mathcal{N}(0, I_{2n_w}),\end{aligned}$$

where $F_\Delta^{(j)}(\theta) = e^{F^{(j)}(\theta)\Delta}$, and $L_\Delta^{(j)}(\theta)$ is a square root of $\int_0^\Delta e^{F^{(j)}(\theta)s} L^{(j)}(\theta)[L^{(j)}(\theta)]^\top e^{[F^{(j)}(\theta)]^\top s}\,\mathrm{d}s$. For irregular sampling times, Δ is simply replaces by $\Delta_k = t_{k+1} - t_k$.

The approximation process is then computed as

$$\begin{aligned}\zeta_{m,k+1}^{(j)} &= F_\Delta^{(j)}(\hat\theta_{k-1})\zeta_{m,k}^{(j)} + L_\Delta^{(j)}(\hat\theta_{k-1})\beta_{m,k}^{(\psi)}, \\ w_{m,k} &= [\zeta_{m,k}^{(j)}]_{1:n_w}, \qquad w_{m,k}^{(j)} = [\zeta_{m,k}^{(j)}]_{n_w:2n_w}.\end{aligned}$$

Notice that, (i) $w_{m,k}$ here is driven by $\beta_{m,k}^{(\psi)}$ and is independent of that in (10), and (ii) $w_{m,k}$ is the same for all j. The recursive approximations of (11) and (12) are

$$\begin{aligned}x_{m,k}^{(j)} &= z_k^{(j)} + C_j(\hat\theta_{k-1})w_{m,k} + C(\hat\theta_{k-1})w_{m,k}^{(j)}, \\ y_{m,k}^{(j)} &= \partial_{\theta_j}\left[f(a;\theta)\right]\Big|_{\substack{a=x_{m,k}\\ \theta=\hat\theta_{k-1}}} + \partial_x[f(x;\hat\theta_{k-1})]\big|_{x=x_{m,k}} x_{m,k}^{(j)}.\end{aligned}$$

Finally, with $\psi_{m,k} = -\begin{bmatrix} y_{m,k}^{(1)} & y_{m,k}^{(2)} & \dots & y_{m,k}^{(d)}\end{bmatrix}^\top$,

$$\bar\psi_k = \frac{1}{M}\sum_{m=1}^M \psi_{m,k}$$

which provides a recursive approximation of $\bar\psi_k(\hat\theta_{k-1})$. It only requires storing $\{\varphi_k^{(j)}\}_j$ and $\{\zeta_{m,k}^{(j)}\}_{m,j}$ at t_k.

D. Algorithm Summary

A summary of the estimation algorithm is collected in Algorithm 1. It starts from a given parameterization as in (1), and considers the general case of estimating parameters in G, f, and the disturbance model. The algorithm is started with an initial value $\hat\theta_0 \in \Theta$ that can be obtained by an a priori offline/patch identification, or using prior knowledge. The initial regressors φ_0 and $\{\varphi_0^j\}$ can be obtained from previous input-output data or simply set to zero.

Notice that, at each iteration, Lines 6 and 8 in Algorithm 1 require the discretization of the transfer functions $G(\mathrm{p};\hat\theta_{k-1})$ and $G_j'(\mathrm{p};\hat\theta_{k-1})$. Likewise, Lines 5 and 10 require the discretization of the Itô stochastic differential equation of w and its gradient with respect to θ. For irregular sampling times and general inputs, the discretization is to be done exactly by using the knowledge of the inter-sample behaviour of the input.

Algorithm 1: OE-QPEM online estimator

output: Sequence of estimates $\{\hat\theta_k\}_{k\geq 1}$
input : Gain sequence $\{\gamma_k\}_{k\geq 1}$, $M \geq 1$, initial parameter $\hat\theta_0$, sampling period Δ, initial Hessian $R_0 = cI$ (for relatively large $c > 0$), initial regressors φ_0, $\{\varphi_0^j\}_{j=1}^{n_G}$, parameter vectors $\eta(\hat\theta_0)$, $\{\eta_j(\hat\theta_0)\}_{j=1}^{n_G}$.

1 Set $w_{m,0} = 0$ and $\zeta_{m,0}^{(j)} = 0$ for all m and j
2 Set index $k \leftarrow 1$ and collect data point (y_1, u_1)
3 **while** *true* **do**
4 $\quad\beta_{m,k}^{(y)} \overset{\text{iid}}{\sim} \mathcal{N}(0, I_{n_w})$, $m = 1,\dots,M$
5 $\quad w_{m,k} = A_\Delta(\hat\theta_{k-1})w_{m,k-1} + B_\Delta(\hat\theta_{k-1})\beta_{m,k}^{(y)}$
6 $\quad z_k = \varphi_{k-1}^\top \eta(\hat\theta_{k-1})$
7 $\quad \bar y_k = \frac{1}{M}\sum_{m=1}^M f\left(z_k + C(\hat\theta_{k-1})w_{m,k}\,;\hat\theta_{k-1}\right)$
8 $\quad z_k^{(j)} = [\varphi_{k-1}^{(j)}]^\top \eta_j(\hat\theta_{k-1})$
9 $\quad \beta_{m,k}^{(\psi)} \overset{\text{iid}}{\sim} \mathcal{N}(0, I_{2n_w})$, $m = 1,\dots,M$
10 $\quad \zeta_{m,k}^{(j)} = F_\Delta^{(j)}(\hat\theta_{k-1})\zeta_{m,k-1}^{(j)} + L_\Delta^{(j)}(\hat\theta_{k-1})\beta_{m,k}^{(\psi)}$
11 $\quad w_{m,k} = [\zeta_{m,k}^{(j)}]_{1:n_w}$
12 $\quad w_{m,k}^{(j)} = [\zeta_{m,k}^{(j)}]_{n_w:2n_w}$
13 $\quad x_{m,k}^{(j)} = z_k^{(j)} + C_j(\hat\theta_{k-1})w_{m,k} + C(\hat\theta_{k-1})w_{m,k}^{(j)}$
14 $\quad y_{m,k}^{(j)} = \partial_{\theta_j}[f(a;\theta)]\big|_{\substack{a=x_{m,k}\\ \theta=\hat\theta_{k-1}}} + \partial_x[f(x;\hat\theta_{k-1})]\big|_{x=x_{m,k}} x_{m,k}^{(j)}$
15 $\quad \psi_{m,k} = -\begin{bmatrix} y_{m,k}^{(1)} & y_{m,k}^{(2)} & \dots & y_{m,k}^{(d)}\end{bmatrix}^\top$
16 $\quad \bar\psi_k = \frac{1}{M}\sum_{m=1}^M \psi_{m,k}$
17 $\quad \varepsilon_k = y_k - \bar y_k$
18 $\quad R_k = R_{k-1} + \gamma_k\left[\bar\psi_k\bar\psi_k^\top - R_{k-1}\right]$
19 $\quad \hat\theta_k = \left[\hat\theta_{k-1} - \gamma_k R_k^{-1}\bar\psi_k\varepsilon_k\right]_\Theta$
20 $\quad$ store φ_0, $\{\varphi_0^j\}_{j=1}^{n_G}$ and $\{w_{m,k}\}$, $\{\zeta_{m,0}^{(j)}\}$ $\forall m, j$
21 $\quad$ set index $k \leftarrow k+1$, and collect data (y_k, u_k)
22 $\quad \varphi_k = [-z_k\ [\varphi_{k-1}]_{1:n-1}^\top\ u_k\ [\varphi_{k-1}]_{n+2:2n-1}^\top]^\top$
23 $\quad \varphi_k^{(j)} = [-z_k^{(j)}\ [\varphi_{k-1}^{(j)}]_{1:n-1}^\top\ u_k\ [\varphi_{k-1}^{(j)}]_{n+2:2n-1}^\top]^\top$
24 **end**

E. Theoretical Motivation

The validity of the stochastic approximation step is achieved by the simulation of (1) using independent Wiener processes. Indeed, under this setting (and open-loop operation), (7) is an unbiased estimator of (5).

The development in Section III-A implicitly assumes that $\{\varepsilon_k(\theta)\}$ is independent and weakly stationary. This does not need to be the case: the validity of the stochastic approximation can be established for a more general class of (ergodic) statistically dependent prediction error processes; see [24]. This in particular means that the convergence of the algorithm can be established even for cases with undermodelling/misspecification, such that $\hat\theta_k \to \vartheta$ almost surely as $k \to \infty$ where ϑ is the set of roots of (5). Moreover, under mild regularity assumptions and an identifiability condition implying $\vartheta = \{\theta_\circ\}$, the asymptotic distribution $\sqrt{k}(\hat\theta_k - \theta_\circ)$ can be characterized; see [6, App.11A].

Additionally, interchanging the order of ordinary differentiation and stochastic integration can be justified (see [25]), and therefore the gradient filters in (13) and the gradients in (11) are well-defined. A detailed analysis of the proposed method is deferred to a dedicated future contribution.

IV. Numerical Examples

A. Example 1

Consider the model

$$\begin{aligned} \mathrm{d}x(t) &= ax(t)\,\mathrm{d}t + bu(t)\,\mathrm{d}t + \sigma\,\mathrm{d}\beta(t), \\ y(t) &= x^2(t), \end{aligned} \tag{14}$$

where the measurement $y_k = y(t_k) + v_k$ is recorded with *irregular* sampling times: $\Delta_k = t_{k+1} - t_k$ are random with uniform distribution over the interval $[0.5, 1]$, and $v_k \sim \mathcal{N}(0, 0.01^2)$. Let $\theta = \begin{bmatrix} a & b & \sigma \end{bmatrix}^\top$, and notice that the plant and disturbance models are jointly parameterized. Consider a case where the data is generated by (14) when the known input is a sum of 10 sinusoids: $u(t) = \sum_{\ell=1}^{10} A_\ell \cos(\omega_\ell t + \phi_\ell)$, with $A_\ell = 6$ for all ℓ, frequencies ω_ℓ in $\{\frac{\pi}{5}, \frac{2\pi}{5}, \ldots, 10\pi\}$ selected uniformly at random, and Schroeder phases $\phi_\ell = \frac{\ell(\ell-1)}{10}\pi$. The true parameter $\theta_\circ = \begin{bmatrix} -1 & 1 & 1 \end{bmatrix}^\top$, and the constraint set $\Theta := \{\theta = [a\ b\ \sigma]^\top \in \mathbb{R}^3 \ : \ a < 0\}$; namely, only a is constrained to guarantee stability. We applied Algorithm 1 to ten independent data sets, using zero initial parameters, zero initial regressors φ_0, with $R_0 = 5I$, a gain sequence $\gamma_k = 1/k^{0.9}$ and constant $M = 100$.

The results are shown in Figure 1 indicating the successful convergence of the algorithm; at the end of one of the runs $\hat{\theta} = \begin{bmatrix} -1.01 & -1.00 & -1.05 \end{bmatrix}^\top$. Note that b and σ are identifiable only in magnitude due to the quadratic nonlinearity.

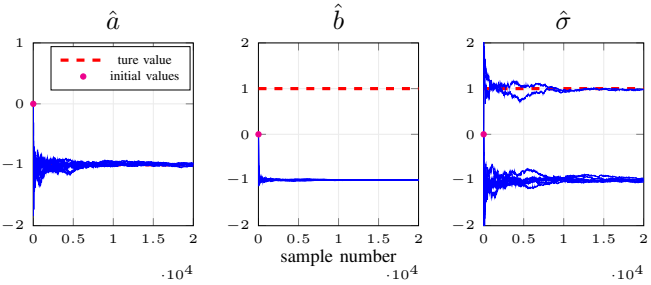

Fig. 1: Ten MC simulations of Algorithm 1 applied to (14)

B. Example 2

Consider now the model given by

$$\begin{aligned} \mathrm{d}w(t) &= \sigma\ \mathrm{d}\beta(t) \\ x(t) &= \frac{c}{p^2 + ap + b} u(t) + w(t) \\ y(t) &= \frac{1}{1 + |x(t)|^\alpha} \end{aligned} \tag{15}$$

and let $\theta = \begin{bmatrix} a & b & c & \sigma & \alpha \end{bmatrix}$. The static nonlinear function in (15) is known as the Hill function and is commonly used in biochemistry, particularly in pharmacology, to describe the dose-response relationship [26].

Suppose the measurements $y_k = y(t_k) + v_k$ are recorded with a constant sampling period $\Delta = 0.5$, $v_k \overset{\text{iid}}{\sim} \mathcal{N}(0, 0.05^2)$, and that the underlying data-generating system is given with the true parameters $a_\circ = 1.2$, $b_\circ = 0.27$, $c_\circ = 1$, and the Hill coefficient $\alpha_\circ = 1.7$. The set Θ only constrains a and b such that the model is stable.

Consider first the following Gaussian disturbance in continuous-time

Case 1: $\mathrm{d}w(t) = -0.75 w(t)\,\mathrm{d}t + 1.5\,\mathrm{d}\beta(t)$

Notice that in this case, (15) misspecifies the spectrum of the disturbance, while it matches the true parameterization of the plant model. We applied the proposed algorithm to fit (15) to ten independent data sets, with random parameter initialization (uniformly within a 50% interval of the true values), $R_0 = 10I$. The initial regressors were constructed using the first two data samples, and the algorithm started at the third sample. The gain sequence is $\gamma_k = 1/k^{0.85}$ and $M = 100$. In all cases, the input is pseudo-random binary input of amplitude ± 5 applied through a zero-order hold.

The results in Figure 2 show that regardless of the specification of the disturbance model, the algorithm successfully converges to the true parameters. In particular, $\hat{\sigma}$ converges to the stationary marginal variance of w.

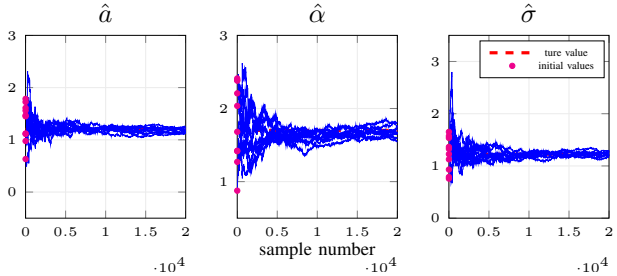

Fig. 2: Ten MC simulations of Algorithm 1 applied to (15) when the true disturbance model is given by Case 1. The estimates of b and c (not shown) exhibit the same behaviour.

For comparison we also applied, to the same data set and settings, an online OE-QPEM algorithm ignoring $w(t)$ completely by assuming that $w(t) = 0$ for all t. The results are shown in Figure 3, where the resulting bias is clear.

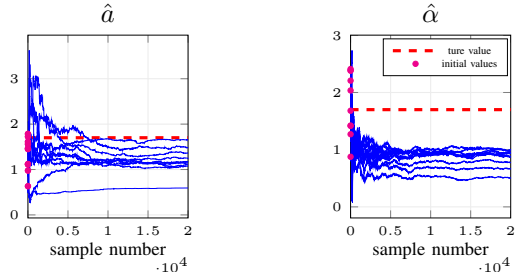

Fig. 3: Ten MC simulations of an online OE-QPEM algorithm that ignores $w(t)$. The estimates of b and c (not shown) exhibit the same behaviour, and σ is not estimated here.

Finally, to check the performance of the algorithm for cases under distributional misspecification we considered the following two additional cases for the disturbance model

Case 2: $\mathrm{d}\xi(t) = -0.75\xi(t)\,\mathrm{d}t + 1.5\,\mathrm{d}\beta(t)$
$w(t_k) = \xi(t_k)\rho_k$, and $\rho_k \sim \mathcal{U}(0, 1)$

Case 3: $\mathrm{d}\xi(t) = -0.75\xi(t)\,\mathrm{d}t + 1.5\,\mathrm{d}\beta(t)$

$$w(t_k) = \begin{cases} \xi(t_k) & \text{with prob. } 0.8 \\ w \sim \mathcal{N}(0, 0.5) & \text{with prob. } 0.2 \end{cases}$$

These cases correspond to mixed continuous-discrete non-Gaussian disturbances under which (15) misspecifies both the marginal distribution and the dependence structure of $w(t)$.

Distributional misspecification is not accounted for by the OE-QPEM estimator; hence, in these cases an asymptotic bias in its estimates is inevitable. The bias depends on the true nonlinearity, the input, and the moments of the true disturbance process. Still, as the results given in Figures 4 and 5 show, the estimates converge to points close to the true values for the considered cases.

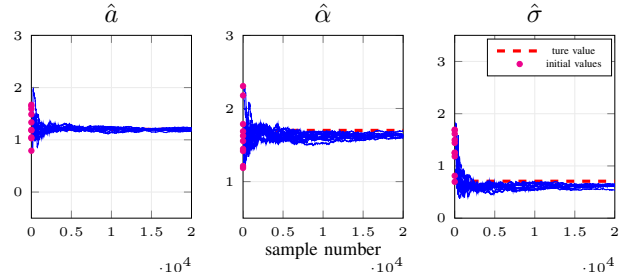

Fig. 4: Ten MC simulations of Algorithm 1 applied to (15) when the true disturbance model is given by Case 2. The estimates of b and c (not shown) exhibit the same behaviour.

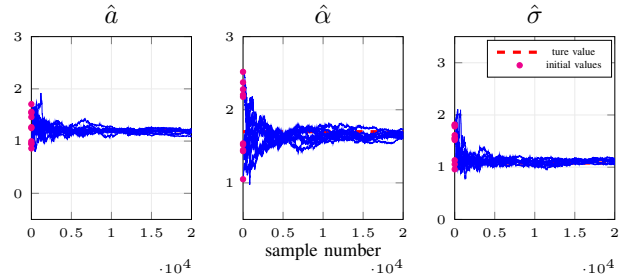

Fig. 5: Ten MC simulations of Algorithm 1 applied to (15) when the true disturbance model is given by Case 3. The estimates of b and c (not shown) exhibit the same behaviour.

Applying an online OE-QPEM algorithm that ignores w completely to these two cases yields a significant bias similar to that observed in Case 1 (see Figure 3), and therefore the corresponding results are omitted.

V. CONCLUSIONS

We propose a simple online identification algorithm suitable for the identification of stochastic continuous-time parametric Wiener models from discrete sampled measurements. The proposed method is based on a stochastic approximation that approximates online, in a recursive fashion, an output-error quadratic PEM estimator. The simulation examples illustrate the convergence of the algorithm as expected, even when the disturbance model is incorrectly specified. We also show that misspecification in the dependence structure of the disturbance does not affect the convergence points of the plant and nonlinearity parameter estimates.

For the sake of clarity, we used several assumptions that are stronger than what is actually required. The restriction to single-input single-output systems, black-box canonical parameterization of G, uniform sampling, and the assumption that the input is constant over the sampling interval are not needed. What is required is the knowledge of the inter-sampling behavior of the input. In addition, the main requirement of the parameterization is its identifiability via the second-order moments of y [19]. A detailed asymptotic and finite-sample analysis of the proposed method is deferred to an extended future contribution.

REFERENCES

[1] P. Hammond, *Theory of self-adaptive control systems*. Springer, 1965.

[2] P. Eykhoff, *System Identification. Parameter and State Estimation*. John Wiley, 1974.

[3] G. Goodwin and K. Sin, *Adaptive Filtering Prediction and Control*. Dover Books on Electrical Engineering, Dover Publications, 2014.

[4] H.-F. Chen and W. Zhao, *Recursive identification and parameter estimation*. CRC Press, 2014.

[5] L. Ljung and T. Söderström, *Theory and practice of recursive identification*. MIT press, 1983.

[6] L. Ljung, *System Identification: Theory for the User*. Prentice Hall, 2nd ed., 1999.

[7] N. Kantas, A. Doucet, S. S. Singh, J. Maciejowski, and N. Chopin, "On particle methods for parameter estimation in state-space models," *Statist. Sci.*, vol. 30, no. 3, pp. 328–351, 2015.

[8] H. Singer, *Langevin and Kalman Importance Sampling for Nonlinear Continuous-Discrete State-Space Models*. Springer, 2018.

[9] S. C. Surace and J.-P. Pfister, "Online maximum-likelihood estimation of the parameters of partially observed diffusion processes," *IEEE Transactions on Automatic Control*, vol. 64, no. 7, pp. 2814–2829, 2018.

[10] M. Abdalmoaty, H. Hjalmarsson, and B. Wahlberg, "The Gaussian Maximum-Likelihood Estimator Versus the Optimally Weighted Least-Squares Estimator [Lecture Notes]," *IEEE Signal Processing Magazine*, vol. 37, no. 6, pp. 195–199, 2020.

[11] S. Oymak and N. Ozay, "Non-asymptotic identification of LTI systems from a single trajectory," in *2019 American Control Conference (ACC)*, pp. 5655–5661, IEEE, 2019.

[12] M. Simchowitz, K. Singh, and E. Hazan, "Improper learning for non-stochastic control," in *Conference on Learning Theory*, pp. 3320–3436, PMLR, 2020.

[13] N. Matni and S. Tu, "A tutorial on concentration bounds for system identification," in *2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 3741–3749, IEEE, 2019.

[14] D. Foster, T. Sarkar, and A. Rakhlin, "Learning nonlinear dynamical systems from a single trajectory," in *Learning for Dynamics and Control*, pp. 851–861, PMLR, 2020.

[15] H. Mania, M. I. Jordan, and B. Recht, "Active learning for nonlinear system identification with guarantees," *arXiv preprint arXiv:2006.10277*, 2020.

[16] E.-W. Bai and F. Giri, *Introduction to Block-oriented Nonlinear Systems*, pp. 3–11. London: Springer London, 2010.

[17] T. Wigren, "Recursive prediction error identification using the nonlinear Wiener model," *Automatica*, vol. 29, no. 4, pp. 1011 – 1025, 1993.

[18] T. Wigren, "Recursive identification of a nonlinear state space model," *International Journal of Adaptive Control and Signal Processing*, vol. 37, no. 2, pp. 447–473, 2023.

[19] M. Abdalmoaty and H. Hjalmarsson, "Linear prediction error methods for stochastic nonlinear models," *Automatica*, vol. 105, pp. 49–63, 2019.

[20] M. Abdalmoaty and H. Hjalmarsson, "Application of a linear PEM estimator to a stochastic Wiener-Hammerstein benchmark problem," *IFAC-PapersOnLine*, vol. 51, no. 15, pp. 784 – 789, 2018.

[21] R. Bereza, O. Eriksson, M. R.-H. Abdalmoaty, D. Broman, and H. Hjalmarsson, "Stochastic approximation for identification of nonlinear differential-algebraic equations with process disturbances," in *2022 IEEE 61st Conference on Decision and Control (CDC)*, pp. 6712–6717, IEEE, 2022.

[22] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, pp. 400–407, 09 1951.

[23] K. J. Åström, *Introduction to Stochastic Control Theory*. Mathematics in Science and Engineering, Academic Press, 1970.

[24] L. Ljung, "Strong convergence of a stochastic approximation algorithm," *The Annals of Statistics*, vol. 6, no. 3, pp. 680–696, 1978.

[25] J. E. Hutton and P. I. Nelson, "Interchanging the order of differentiation and stochastic integration," *Stochastic processes and their applications*, vol. 18, no. 2, pp. 371–377, 1984.

[26] C. Minto, T. Schnider, T. Short, K. Gregg, A. Gentilini, and S. Shafer, "Response Surface Model for Anesthetic Drug Interactions," *Anesthesiology*, vol. 92, pp. 1603–1616, 06 2000.