Delft University of Technology

Delft Center for Systems and Control

Technical report 13-002

Exact and approximate approaches to the identification of stochastic max-plus-linear systems*

S.S. Farahani, T. van den Boom, and B. De Schutter

If you want to cite this report, please use the following reference instead:
S.S. Farahani, T. van den Boom, and B. De Schutter, "Exact and approximate approaches to the identification of stochastic max-plus-linear systems," *Discrete Event Dynamic Systems: Theory and Applications*, vol. 24, no. 4, pp. 447–471, Dec. 2014. doi:10.1007/s10626-013-0164-4

Delft Center for Systems and Control
Delft University of Technology
Mekelweg 2, 2628 CD Delft
The Netherlands
phone: +31-15-278.24.73 (secretary)
URL: https://www.dcsc.tudelft.nl

* This report can also be downloaded via https://pub.bartdeschutter.org/abs/13_002.html

Discrete Event Dynamic Systems: Theory and Applications manuscript No.
(will be inserted by the editor)

Exact and Approximate Approaches to the Identification of Stochastic Max-Plus-Linear Systems

Samira S. Farahani · Ton van den Boom · Bart De Schutter

Abstract Stochastic max-plus linear systems, i.e., perturbed systems that are linear in the max-plus algebra, belong to a special class of discrete-event systems that consists of systems with synchronization but no choice. In this paper, we study the identification problem for such systems, considering two different approaches. One approach is based on exact computation of the expected values and consists in recasting the identification problem as an optimization problem that can be solved using gradient-based algorithms. However, due to the structure of stochastic max-plus linear systems, this method results in a complex optimization problem. The alternative approach discussed in this paper, is an approximation method based on the higher-order moments of a random variable. This approach decreases the required computation time significantly while still guaranteeing a performance that is comparable to the one of the exact solution.

Keywords Stochastic discrete event systems · System identification · Stochastic max-plus-linear systems · Analytic integration · Approximation · Moments

1 Introduction

Discrete-event systems form a large class of dynamic systems in which the evolution of the system is specified by the occurrence of certain discrete events, unlike continuous dynamic systems where the state of the system changes as time progresses. Some examples of discrete-event systems are telecommunication networks, railway networks, manufacturing systems, parallel computing, traffic control systems, etc. For such systems there exist different modeling frameworks such as queuing theory, (extended) state machines, formal languages, automata, temporal logic models, generalized semi-Markov processes, Petri nets, and computer simulation models (Cassandras and Lafortune, 1999; Ho, 1992; Peterson, 1981). Models of such systems are in general nonlinear in conventional algebra. However, there exists an important class of discrete event systems, namely the max-plus-linear (MPL) systems, for which the model is *linear* in the max-plus algebra (Baccelli et al, 1992; Cuninghame-Green, 1979; Heidergott et al, 2006). The class of MPL systems consists of discrete-event systems with synchronization but no choice. Synchronization requires the

S. S. Farahani, T. van den Boom, and B. De Schutter
Delft Center for Systems and Control, Delft University of Technology, Delft, the Netherlands
E-mail: {s.safaeifarahani,a.j.j.vandenboom,b.deschutter}@tudelft.nl

availability of several resources at the same time, whereas choice appears, e.g., when some user must choose among several resources (Baccelli et al, 1992). Typical examples of such systems are serial production lines, production systems with a fixed routing schedule, and railway networks. In stochastic systems, processing times and/or transportation times are assumed to be stochastic quantities, since in practice such stochastic fluctuations can, e.g. be caused by machine failure or depreciation (Olsder et al, 1990). Some results on (stochastic) MPL systems including analysis, controller design, etc., can be found in (Akian, 2007; Baccelli et al, 1992; Başar and Bernhard, 1995; Bemporad et al, 2003; Heidergott et al, 2006; Mairesse, 1994; McEneaney, 2004; Olsder et al, 1990; Resing et al, 1990; Somasundaram and Baras, 2011).

The aim of this paper is to identify the model parameters of a stochastic MPL system defined by a state space model. Most identification methods for MPL discrete-event systems use a transfer function approach (Boimond et al, 1995; Gallot et al, 1997) while state space models have certain advantages: they explicitly take the initial state of the system into account, they can reveal "hidden" behavior such as unobservable, unstable modes, the extension from SISO to MIMO is more intuitive and elegant for state space models, and the analysis is often easier. Some examples of state space identification methods for *deterministic* MPL systems are presented in (De Schutter et al, 2002; Schullerus and Krebs, 2001a,b; Schullerus et al, 2003). In the current paper, our focus is on the identification of *stochastic* MPL systems in which modeling errors, noise, and/or disturbances are present. In stochastic MPL systems the influence of noise and disturbances are often max-plus-multiplicative (Baccelli et al, 1992), unlike for conventional linear systems where noise is usually considered to be additive. The noise and disturbances result in a perturbation of system parameters. Consequently, in the identification method, the stochastic properties of the systems have to be taken into account. To this end, we present two different approaches where one is based on exact computation of the expected values using numerical or analytic integration and the other one is an approximation method. In the first approach, we show that the resulting identification problem can be solved using gradient-based search techniques. However, when the order of the stochastic system increases, the computational complexity of the first approach increases drastically. To decrease the computational complexity, we also propose an alternative approximation approach that is based on the idea of the method of Farahani et al (2010) in which only normally distributed noise has been considered. In the current paper, we extend this method such that it is applicable to a much broader range of distributions. This method is based on moments of a random variable and when we have an analytic expression for moments, it simplifies the computations considerably. Hence, we obtain a much faster and more efficient way to solve the identification problem for stochastic MPL systems without increasing the computational complexity and with a comparable performance to the first approach (for the case study considered in this paper).

The structure of this paper is as follows. Section 2 gives a concise description of the max-plus algebra and stochastic MPL systems. In Section 3, an identification problem for stochastic MPL systems is described. Section 4 discusses the first approach, which consists of numerical and analytic integration (in particular in the case of piecewise polynomial probability density functions), to solve the identification problem. Section 5 introduces the second approach, based on the higher-order moments, and describes how it reduces the complexity of the identification problem. The error of this approximation is discussed in this section as well. In Section 6 we provide a brief discussion on complexity analysis of the two approaches. Section 7 presents two worked examples in which both approaches are applied and their performance is examined. Finally, Section 8 concludes the paper.

2 Max-Plus Algebra and Stochastic Max-Plus Linear Systems

In this section, we present a brief overview of the max-plus algebra, followed by a concise description of stochastic max-plus-linear systems. For more information on these topics, the interested reader is referred to (Baccelli et al, 1992; Cuninghame-Green, 1979; Heidergott et al, 2006).

2.1 Max-Plus Algebra

Define $\mathbb{R}_\varepsilon = \mathbb{R} \cup \{\varepsilon\}$ and $\varepsilon = -\infty$. The main operations in max-plus algebra, as suggested by its name, are maximization and addition:

$$x \oplus y = \max(x, y)$$
$$x \otimes y = x + y$$

for $x, y \in \mathbb{R}_\varepsilon$. Based on this definition, the zero element of the max-plus addition is defined as ε, i.e., $x \oplus \varepsilon = x$, and the identity element of the max-plus multiplication as $e = 0$, i.e., $x \otimes e = x$. The corresponding max-plus matrix operations are defined as (Baccelli et al, 1992)

$$(A \oplus B)_{ij} = a_{ij} \oplus b_{ij} = \max(a_{ij}, b_{ij})$$
$$(A \otimes C)_{ij} = \bigoplus_{k=1}^{n} a_{ik} \otimes c_{kj} = \max_{k=1,\ldots,n} (a_{ik} + c_{kj})$$

for $A, B \in \mathbb{R}_\varepsilon^{m \times n}$ and $C \in \mathbb{R}_\varepsilon^{n \times p}$. In this paper, we denote the i-th row of matrix A by $A_{i,\cdot}$ and the j-th column by $A_{\cdot,j}$. To avoid confusion in the sequel, we drop the multiplication sign in conventional algebra expressions while keeping the $\otimes$ sign in max-plus expressions.

Now let $\mathscr{S}_{\text{mpns}}$ denote the set of max-plus-nonnegative-scaling functions, i.e., functions f of the form

$$f(z) = \max_{i=1,\ldots,m} (\tau_{i,1} z_1 + \cdots + \tau_{i,n} z_n + \xi_i)$$

with variable $z \in \mathbb{R}_\varepsilon^n$ and constant coefficients $\tau_{i,j} \in \mathbb{R}^+$ and $\xi_i \in \mathbb{R}$, where $\mathbb{R}^+$ is the set of the nonnegative real numbers. In the sequel, we stress that f is a function of z by writing $f \in \mathscr{S}_{\text{mpns}}(z)$. As shown by van den Boom and De Schutter (2004), the set $\mathscr{S}_{\text{mpns}}$ is closed under the operations $\oplus, \otimes$, and the scalar multiplication by a nonnegative scalar.

2.2 Stochastic MPL Systems

Max-plus-linear systems form a special class of discrete-event systems with synchronization but no choice. The state space representation of such systems is as follows (Baccelli et al, 1992; Cuninghame-Green, 1979):

$$x(k+1) = A(k) \otimes x(k) \oplus B(k) \otimes u(k) \tag{1}$$
$$= \begin{bmatrix} A(k) & B(k) \end{bmatrix} \otimes \begin{bmatrix} x(k) \\ u(k) \end{bmatrix} \tag{2}$$
$$= Q(k) \otimes \phi(k) \tag{3}$$

where

$$Q(k) = \begin{bmatrix} A(k) & B(k) \end{bmatrix} \in \mathbb{R}_{\varepsilon}^{n \times m}, \qquad \phi(k) = \begin{bmatrix} x(k) \\ u(k) \end{bmatrix} \in \mathbb{R}_{\varepsilon}^{m}$$

with $m = n + n_u$ where n is the number of states and n_u is the number of inputs, $x(k)$ is the state of the system at event step k, and $u(k)$ is the input of the system at event step k. In fact $x(k)$ and $u(k)$ contain the time instants at which the internal and the input event occurs for the k-th time, respectively.

In a stochastic system, the system matrices $A(k)$ and $B(k)$ are perturbed by noise and/or modeling errors. Following van den Boom et al (2003), these uncertainties will be presented in a single framework, using one stochastic vector $e(k)$ with a certain probability distribution. Note that the entries of the system matrices belong to $\mathscr{S}_{\text{mpns}}(e(k))$ (van den Boom and De Schutter, 2004), i.e., $A(k) \in \mathscr{S}_{\text{mpns}}^{n \times n}(e(k))$, $B(k) \in \mathscr{S}_{\text{mpns}}^{n \times n_u}(e(k))$. In the sequel, we denote the uncertain system matrices with the matrix $Q(k)$ and the state and input vector with $\phi(k)$ (cf. (3)).

In order to identify the unknown system parameters, we need to distinguish between the parameters that are known a priori, i.e., the parameters that are either constant or determined in advance such as the nominal transportation times in a production system, and the parameters that have to be identified. Therefore, the i-th row of the matrix $Q(k)$ can be written as:

$$Q_{i,\cdot}(k) = \Xi_{i,\cdot} + \theta^T \Delta^{(i)} + e^T(k) \Lambda S^{(i)} \tag{4}$$

where Ξ represents the parameters that are known a priori, θ is a vector of unknown parameters, $e(k) = [e_1(k), \ldots, e_{n_e}(k)]^T$ is the stochastic vector and all its elements are assumed to be independent random variables, the diagonal matrix $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_{n_e})$ contains the amplitude of the noise, and $\Delta^{(i)}$ and $S^{(i)}$ are selection matrices for the i-th row with zeros and ones as entries. The role of the selection matrices is indeed to determine which elements of the vectors θ and $e(k)$ will appear in the i-th row of $Q(k)$. For example for the first element of the first row, i.e., $i = 1$, let $\Delta^{(1)} = [1 \;\; 0 \;\; 1]^T$ and $S^{(1)} = [0 \;\; 1 \;\; 1]^T$; then $\theta^T \Delta^{(1)} = [\theta_1 \;\; \theta_2 \;\; \theta_3] \cdot [1 \;\; 0 \;\; 1]^T = \theta_1 + \theta_3$ and $e^T(k) \Lambda S^{(1)} = [e_1(k) \;\; e_2(k) \;\; e_3(k)] \cdot \text{diag}(\lambda_1, \lambda_2, \lambda_3) \cdot [0 \;\; 1 \;\; 1]^T = e_2(k)\lambda_2 + e_3(k)\lambda_3$. As mentioned before, the influence of noise here is max-plus-multiplicative and as a result, it appears in the system matrices.

3 Identification of Stochastic Max-Plus Linear Systems

The identification procedure in this paper is based on input-state data. Note that in MPL systems the state contains the time instants at which the state events occur. Since the state is observable by assumption, these instants can be measured easily and so we usually have full state information. We assume that the probability density function of $e(k)$, denoted by $f(e)$, and the matrices Ξ, Δ, and S are known a priori and that the only parameters that have to be identified are the components of θ and the diagonal elements of Λ, denoted by $\lambda = [\lambda_1, \ldots, \lambda_{n_e}]^T$. Therefore, in the identification procedure we will derive estimates $\hat{\theta}$ and $\hat{\lambda}$ for θ and λ, respectively. Consider the measured input-state sequence $\{(u_{\text{meas}}(k), x_{\text{meas}}(k))\}_{k=1}^{N}$ of a system of the form (3) and assume that the system parameters $\hat{\theta}$ and $\hat{\lambda}$ have to be identified using this sequence. Further, we assume that the input-state sequence is sufficiently rich[1] to capture all the relevant information about the system (see

[1] Intuitively, this can be characterized as follows. Note that (3) and (4) imply that each component of $x(k+1)$ can be written as a max expression of terms in which the unknown parameters θ and λ appear.

also Schullerus and Krebs (2001b); Schullerus et al (2003)). Now consider the following identification problem:

$$\min_{(\hat{\theta},\hat{\lambda})} J(\hat{\theta},\hat{\lambda}) \quad \text{subject to} \quad \hat{\lambda} > 0 \tag{5}$$

with

$$J(\hat{\theta},\hat{\lambda}) = \sum_{k=1}^{N-1}\sum_{i=1}^{n} \left(\mathbb{E}[x_i(k+1|k)] - x_{\text{meas},i}(k+1)\right)^2 \tag{6}$$

where $\mathbb{E}[\cdot]$ denotes the expected value operator and $\mathbb{E}[x_i(k+1|k)]$ is the one-step-ahead prediction of x_i for event step $k+1$, using the knowledge from event step k. Considering (3) and (4), we can rewrite the one-step-ahead prediction as

$$\mathbb{E}[x_i(k+1|k)] = \mathbb{E}\left[\left(\Xi_{i,\cdot} + \hat{\theta}^T \Delta^{(i)} + e^T(k)\hat{\Lambda} S^{(i)}\right) \otimes \phi(k)\right]$$

and hence, the one-step-ahead prediction error will be given by

$$\begin{aligned}\hat{\eta}_i(k+1,\hat{\theta},\hat{\lambda}) &= \mathbb{E}[x_i(k+1|k)] - x_{\text{meas},i}(k+1) \\ &= \mathbb{E}\Big[\max_{j=1,\ldots,m}\left(\xi_{ij} + \hat{\theta}^T \Delta^{(i)}_{\cdot,j} + e^T(k)\hat{\Lambda} S^{(i)}_{\cdot,j} + \phi_j(k) - x_{\text{meas},i}(k+1)\right)\Big]\end{aligned} \tag{7}$$

Now for a specific realization of the noise vector $e(k)$, let:

$$\eta_i(k+1,\hat{\theta},\hat{\lambda},e(k)) = \max_{j=1,\ldots,m}\left(\xi_{ij} + \hat{\theta}^T \Delta^{(i)}_{\cdot,j} + e^T(k)\hat{\Lambda} S^{(i)}_{\cdot,j} + \phi_j(k) - x_{\text{meas},i}(k+1)\right)$$

which is indeed a max-plus-nonnegative-scaling function. Hence,

$$\hat{\eta}_i(k+1,\hat{\theta},\hat{\lambda}) = \mathbb{E}[\eta_i(k+1,\hat{\theta},\hat{\lambda},e(k))]$$

To have a more compact notation, let $\alpha_{ij}(k) = \xi_{ij} + \phi_j(k) - x_{\text{meas},i}(k+1)$, $\Pi_{ij} = \Delta^{(i)}_{\cdot,j}$, and $\Gamma_{ij} = \text{diag}((S^{(i)})_{1,j},\ldots,(S^{(i)})_{n_e,j})$. Since $e^T(k)\hat{\Lambda} S^{(i)}_{\cdot,j}$ is a scalar and $\hat{\Lambda}$ is a diagonal matrix, we have:

$$e^T(k)\hat{\Lambda} S^{(i)}_{\cdot,j} = (S^{(i)}_{\cdot,j})^T \hat{\Lambda} e(k) = \hat{\lambda}^T \Gamma_{ij} e(k)$$

Therefore, we can rewrite $\eta_i(k+1,\hat{\theta},\hat{\lambda},e(k))$ as

$$\eta_i(k+1,\hat{\theta},\hat{\lambda},e(k)) = \max_{j=1,\ldots,m}\left(\alpha_{ij}(k) + \Pi_{ij}^T\hat{\theta} + \hat{\lambda}^T\Gamma_{ij}e(k)\right) \tag{8}$$

Hence, it is only left to compute the expected value of $\eta_i(k+1,\hat{\theta},\hat{\lambda},e(k))$, i.e., $\hat{\eta}_i(k+1,\hat{\theta},\hat{\lambda})$, for which different methods are proposed in Sections 4 and 5.

4 First Approach: Numerical or Analytic Integration

In this section, first we show how the computation of $\hat{\eta}_i(k+1,\hat{\theta},\hat{\lambda})$ leads to the computation of an integral. Then we propose two different methods to deal with the integration.

An input signal is then said to be sufficiently rich if it is such that each of these terms is the maximal one sufficiently often (this is also related to the idea of persistent excitation in conventional system identification (Ljung, 1999)).

4.1 Computation of the Expected Value

By considering the definition of the expected value, we can compute $\hat{\eta}_i(k+1,\hat{\theta},\hat{\lambda})$ as follows:

$$\begin{aligned}\hat{\eta}_i(k+1,\hat{\theta},\hat{\lambda}) &= \mathbb{E}[\eta_i(k+1,\hat{\theta},\hat{\lambda},e(k))] \\ &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\eta_i(k+1,\hat{\theta},\hat{\lambda},e)f(e)\mathrm{d}e \qquad (9)\\ &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\max_{j=1,\ldots,m}(\alpha_{ij}(k)+\Pi_{ij}^T\hat{\theta}+\hat{\lambda}^T\Gamma_{ij}e)f(e)\mathrm{d}e \qquad (10)\\ &= \sum_{j=1}^{m}\int\cdots\int_{e\in\Omega_{ij}(\hat{\theta},\hat{\lambda},k)}(\alpha_{ij}(k)+\Pi_{ij}^T\hat{\theta}+\hat{\lambda}^T\Gamma_{ij}e)f(e)\mathrm{d}e \qquad (11)\end{aligned}$$

where $\mathrm{d}e=\mathrm{d}e_1,\ldots,\mathrm{d}e_{n_e}$ and the polyhedral sets $\Omega_{ij}(\hat{\theta},\hat{\lambda},k), i=1,\ldots,n,\ j=1,\ldots,m$ are defined such that

$$\mathrm{int}(\Omega_{i\ell})\cap\mathrm{int}(\Omega_{i\nu})=\emptyset \ \text{ for } \ \ell\neq\nu$$

where $\mathrm{int}(\Omega_{ij})$ denotes the interior of Ω_{ij}, and such that for all $e\in\Omega_{ij}(\hat{\theta},\hat{\lambda},k)$,

$$\eta_i(k+1,\hat{\theta},\hat{\lambda},e)=\alpha_{ij}(k)+\Pi_{ij}^T\hat{\theta}+\hat{\lambda}^T\Gamma_{ij}e$$

and that for any i it holds that $\bigcup_{j=1}^{m}\Omega_{ij}(\hat{\theta},\hat{\lambda},k)=\mathbb{R}^{n_e}$, i.e., for all realizations of e, the j-th term in (8) gives the maximum, and the sets $\Omega_{ij}(\hat{\theta},\hat{\lambda},k)$ cover the whole space of $\mathbb{R}^{n_e}$ and only overlap at the boundaries of the regions[2].

Remark 1 Note that the sets $\Omega_{ij}, i=1,\ldots,n, j=1,\ldots,m$ are polyhedra. This follows from the fact that Ω_{ij} is described by a system of linear inequalities. In fact, for $e\in\Omega_{ij}(\hat{\theta},\hat{\lambda},k)$ we have:

$$\max_{j=1,\ldots,m}(\alpha_{ij}(k)+\Pi_{ij}^T\hat{\theta}+\hat{\lambda}^T\Gamma_{ij}e)=\alpha_{ij}(k)+\Pi_{ij}^T\hat{\theta}+\hat{\lambda}^T\Gamma_{ij}e.$$

Hence, $\alpha_{ij}(k)+\Pi_{ij}^T\hat{\theta}+\hat{\lambda}^T\Gamma_{ij}e\geq\alpha_{i\ell}(k)+\Pi_{i\ell}^T\hat{\theta}+\hat{\lambda}^T\Gamma_{i\ell}e$ for $\ell=1,\ldots,m$.

Proposition 1 *The function $\hat{\eta}_i(k+1,\hat{\theta},\hat{\lambda})$, defined in* (7)*, is convex in $\hat{\theta}$ and $\hat{\lambda}$, and its subgradients with respect to $\hat{\theta}$ and $\hat{\lambda}$ are*

$$g_{i,\hat{\theta}}(\hat{\theta},\hat{\lambda},k)=\sum_{j=1}^{m}\Big(\int\cdots\int_{e\in\Omega_{ij}(\hat{\theta},\hat{\lambda},k)}f(e)de\Big)\Pi_{ij} \qquad (12)$$

$$g_{i,\hat{\lambda}}(\hat{\theta},\hat{\lambda},k)=\sum_{j=1}^{m}\Big(\int\cdots\int_{e\in\Omega_{ij}(\hat{\theta},\hat{\lambda},k)}e^Tf(e)de\Big)\Gamma_{ij} \qquad (13)$$

respectively.

[2] If there are two identical affine arguments in the max expression in (10), then the corresponding sets Ω_{ij} coincide. So in general the Ω_{ij} sets either coincide or they only overlap at the boundaries (see also Remark 1). For the sake of simplicity of the exposition, we assume here that any identical arguments in the max expression in (10) have already been eliminated.

Proof The proof of this proposition is the straightforward application of the definition of a convex function. Consider vectors $\hat{\theta}_0$ and $\hat{\lambda}_0$ with the same size as $\hat{\theta}$ and $\hat{\lambda}$, respectively. Recall that (cf. (11))

$$\hat{\eta}_i(k+1,\hat{\theta}_0,\hat{\lambda}_0) = \sum_{j=1}^{m} \int\cdots\int_{e\in\Omega_{ij}(\hat{\theta}_0,\hat{\lambda}_0,k)} (\alpha_{ij}(k) + \Pi_{ij}^T\hat{\theta}_0 + \hat{\lambda}_0^T\Gamma_{ij}e)f(e)\mathrm{d}e$$

Then, using the fact that $\bigcup_{j=1}^{m}\Omega_{ij}(\hat{\theta}_0,\hat{\lambda}_0,k) = \mathbb{R}^{n_e}$, there holds for any $\hat{\theta}$ and $\hat{\lambda}$:

$$\begin{aligned}
\hat{\eta}_i(k+1,\hat{\theta},\hat{\lambda}) &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \max_{j=1,\ldots,m}(\alpha_{ij}(k) + \Pi_{ij}^T\hat{\theta} + \hat{\lambda}^T\Gamma_{ij}e)f(e)\mathrm{d}e \\
&= \sum_{j=1}^{m} \int\cdots\int_{e\in\Omega_{ij}(\hat{\theta}_0,\hat{\lambda}_0,k)} \max_{j=1,\ldots,m}(\alpha_{ij}(k) + \Pi_{ij}^T\hat{\theta} + \hat{\lambda}^T\Gamma_{ij}e)f(e)\mathrm{d}e \\
&\geq \sum_{j=1}^{m} \int\cdots\int_{e\in\Omega_{ij}(\hat{\theta}_0,\hat{\lambda}_0,k)} (\alpha_{ij}(k) + \Pi_{ij}^T\hat{\theta} + \hat{\lambda}^T\Gamma_{ij}e)f(e)\mathrm{d}e \qquad (14)
\end{aligned}$$

Note that the sets $\Omega_{ij}(\cdot,\cdot,k)$ in (14) are computed for $\hat{\theta}_0$ and $\hat{\lambda}_0$, whereas for $\hat{\eta}_i(k+1,\hat{\theta},\hat{\lambda})$, they are computed for $\hat{\theta}$ and $\hat{\lambda}$ (cf. (11)). Now consider:

$$\begin{aligned}
&\sum_{j=1}^{m} \int\cdots\int_{e\in\Omega_{ij}(\hat{\theta}_0,\hat{\lambda}_0,k)} (\alpha_{ij}(k) + \Pi_{ij}^T\hat{\theta} + \hat{\lambda}^T\Gamma_{ij}e)f(e)\mathrm{d}e \\
&= \sum_{j=1}^{m} \int\cdots\int_{e\in\Omega_{ij}(\hat{\theta}_0,\hat{\lambda}_0,k)} (\alpha_{ij}(k) + \Pi_{ij}^T\hat{\theta}_0 + \hat{\lambda}_0^T\Gamma_{ij}e)f(e)\mathrm{d}e \\
&\quad + \sum_{j=1}^{m} \int\cdots\int_{e\in\Omega_{ij}(\hat{\theta}_0,\hat{\lambda}_0,k)} (\Pi_{ij}^T(\hat{\theta}-\hat{\theta}_0))f(e)\mathrm{d}e + \sum_{j=1}^{m} \int\cdots\int_{e\in\Omega_{ij}(\hat{\theta}_0,\hat{\lambda}_0,k)} ((\hat{\lambda}-\hat{\lambda}_0)^T\Gamma_{ij}e)f(e)\mathrm{d}e \\
&= \sum_{j=1}^{m} \int\cdots\int_{e\in\Omega_{ij}(\hat{\theta}_0,\hat{\lambda}_0,k)} (\alpha_{ij}(k) + \Pi_{ij}^T\hat{\theta}_0 + \hat{\lambda}_0^T\Gamma_{ij}e)f(e)\mathrm{d}e \\
&\quad + \sum_{j=1}^{m} \Big(\int\cdots\int_{e\in\Omega_{ij}(\hat{\theta}_0,\hat{\lambda}_0,k)} f(e)\mathrm{d}e\Big)\Pi_{ij}^T(\hat{\theta}-\hat{\theta}_0) + \sum_{j=1}^{m} \Big(\int\cdots\int_{e\in\Omega_{ij}(\hat{\theta}_0,\hat{\lambda}_0,k)} e^T f(e)\mathrm{d}e\Big)\Gamma_{ij}^T(\hat{\lambda}-\hat{\lambda}_0) \\
&= \hat{\eta}_i(k+1,\hat{\theta}_0,\hat{\lambda}_0) + g_{i,\hat{\theta}}^T(\hat{\theta}_0,\hat{\lambda}_0,k)(\hat{\theta}-\hat{\theta}_0) + g_{i,\hat{\lambda}}^T(\hat{\theta}_0,\hat{\lambda}_0,k)(\hat{\lambda}-\hat{\lambda}_0)
\end{aligned}$$

with $g_{i,\hat{\theta}}$ and $g_{i,\hat{\lambda}}$ defined in (12) and (13), respectively. Hence, we conclude that

$$\hat{\eta}_i(k+1,\hat{\theta},\hat{\lambda}) \geq \hat{\eta}_i(k+1,\hat{\theta}_0,\hat{\lambda}_0) + g_{i,\hat{\theta}}(\hat{\theta}_0,\hat{\lambda}_0,k)(\hat{\theta}-\hat{\theta}_0) + g_{i,\hat{\lambda}}(\hat{\theta}_0,\hat{\lambda}_0,k)(\hat{\lambda}-\hat{\lambda}_0)$$

which proves that $\hat{\eta}_i$ is convex in $\hat{\theta}$ and $\hat{\lambda}$, and that $g_{i,\hat{\theta}}$ and $g_{i,\hat{\lambda}}$ are subgradients [3] of $\hat{\eta}_i$ with respect to $\hat{\theta}$ and $\hat{\lambda}$. □

[3] For the case that $f(e)$ is a continuous function, $\hat{\eta}_i(k+1,\hat{\theta},\hat{\lambda})$ is continuously differentiable and since it is a convex function, the subgradients are unique and they are equal to the gradients.

Therefore, $J(\hat{\theta},\hat{\lambda})$ in the identification problem (5) can be written as

$$J(\hat{\theta},\hat{\lambda}) = \sum_{k=1}^{N}\sum_{i=1}^{n}\left(\hat{\eta}_i(k+1,\hat{\theta},\hat{\lambda})\right)^2 \tag{15}$$

with the gradients

$$\nabla_{\hat{\theta}} J(\hat{\theta},\hat{\lambda}) = \sum_{k=1}^{N}\sum_{i=1}^{n} 2\hat{\eta}_i(k+1,\hat{\theta},\hat{\lambda})\, g_{i,\hat{\theta}}(\hat{\theta},\hat{\lambda},k)$$
$$\nabla_{\hat{\lambda}} J(\hat{\theta},\hat{\lambda}) = \sum_{k=1}^{N}\sum_{i=1}^{n} 2\hat{\eta}_i(k+1,\hat{\theta},\hat{\lambda})\, g_{i,\hat{\lambda}}(\hat{\theta},\hat{\lambda},k).$$

Consequently, despite the fact that $J(\hat{\theta},\hat{\lambda})$ is not convex anymore since $\hat{\eta}_i(k+1,\hat{\theta},\hat{\lambda})$ is not always positive, the optimal $\hat{\theta}$ and $\hat{\lambda}$ can be found using gradient-based search methods (see e.g. (Pardalos and Resende, 2002, Chapter 5)).

Note that to compute the cost function (15), we first need to find the value of $\hat{\eta}_i(k+1,\hat{\theta},\hat{\lambda})$, which leads to the solution of integrals (9) or (11). The rest of this section is dedicated to two different methods that solve the above-mentioned integrals.

4.2 Numerical Integration

To obtain the value of $\hat{\eta}_i(k+1,\hat{\theta},\hat{\lambda})$, one can compute the integral (9) using numerical integration. The common methods for numerical integration are (non)adaptive integration, (non)iterative integration, exponential quadrature, Monte Carlo integration, the Nyström method, the Quasi-Monte Carlo method, and the Multi-step method (Davis and Rabinowitz, 1984). However, numerical integration is in general both cumbersome and time-consuming, and it becomes even more complicated as the probability density function f becomes more and more complex. Hence, in the following, an alternative method is proposed based on analytic integration.

4.3 Analytic Integration for Piecewise Polynomial Probability Density Functions

As mentioned, numerical integration is not an optimal way of computing the integral (9), since it is quite complex and time-inefficient. One way to avoid this complexity is to consider a piecewise polynomial probability density function defined on polyhedral sets. In this case, either the stochastic vector has a piecewise polynomial probability density function or we approximate the real probability density function with a piecewise polynomial probability density function.

Let $f(e)$ be a piecewise polynomial function defined on polyhedral sets P_ℓ, $\ell = 1,\ldots,n_p$, such that

$$\bigcup_{\ell=1}^{n_p} P_\ell = \mathbb{R}^{n_e}$$
$$\text{int}(P_i)\cap\text{int}(P_j) = \emptyset \ \text{ for } \ i \neq j$$

where $\text{int}(P_i)$ denotes the interior of P_i, and for $e \in P_\ell$ the probability density function is given by $f_\ell(e)$, where

$$f_\ell(e) = \sum_{i_1=0}^{M_1} \sum_{i_2=0}^{M_2} \cdots \sum_{i_{n_e}=0}^{M_{n_e}} \zeta_{i_1,i_2,\ldots,i_{n_e}} e_1^{i_1} e_2^{i_2} \cdots e_{n_e}^{i_{n_e}}$$

for some integers $M_1,\ldots,M_{n_e}$ and coefficients $\zeta_{i_1,i_2,\ldots,i_{n_e}} \in \mathbb{R}$. Consider the signal $\eta(k+1,\hat{\theta},\hat{\lambda})$. Let $\Psi_{ij\ell}(\hat{\theta},\hat{\lambda},k) = \Omega_{ij}(\hat{\theta},\hat{\lambda},k) \cap P_\ell$ for $j = 1,\ldots,m$, $\ell = 1,\ldots,n_p$. Then by Remark 1, $\Psi_{ij\ell}(\hat{\theta},\hat{\lambda},k)$ is a polyhedron, and $\hat{\eta}_i(k+1,\hat{\theta},\hat{\lambda})$ can be written as

$$\hat{\eta}_i(k+1,\hat{\theta},\hat{\lambda}) = \sum_{\ell=1}^{n_p} \sum_{j=1}^{m} \int\cdots\int_{e\in\Psi_{ij\ell}(\hat{\theta},\hat{\lambda},k)} (\alpha_{ij}(k) + \Pi_{ij}^T \hat{\theta} + \hat{\lambda}^T \Gamma_{ij} e) f_\ell(e) de. \tag{16}$$

This is a sum of integrals of polynomial functions in e and then can be solved analytically for each polyhedron $\Psi_{ij\ell}$ (Büeler et al, 2000; Lasserre, 1998). Note that if a piecewise polynomial probability density function is used as an approximation of "true" non-polynomial probability function, the quality of the approximation can be improved by increasing the number of sets n_p.

5 Second Approach: Approximation Method

As will be discussed in Section 6, the complexity of the method of Section 4.3, increases exponentially as n_e increases and polynomially as n and n_u increase. It also increases in the case of having non-piecewise polynomial probability density functions, such as normal probability density function, that cannot be directly applied to the first approach and hence, have to be approximated by piecewise polynomial probability density functions. Therefore, in this section, we propose to adopt and extend the approach presented in (Farahani et al, 2010), which is based on the p-th moment of a stochastic random variable, in order to approximate $\mathbb{E}[\eta_i(k+1,\hat{\theta},\hat{\lambda},e(k))]$ and to significantly decrease the computational burden. Note that this extended approach is now applicable to a wide range of distributions, while in (Farahani et al, 2010) the method was only proposed for the case of normally distributed noise. As will be shown in Section 6, the complexity of this method in general increases quadratically as n increases, linearly as n_u increases, and polynomially as n_e increases, which offers a great advantage compared to the first approach.

5.1 Description of the Approximation Method

This approximation approach is inspired by the relation between the ∞-norm and the p-norm. Assume that $x = [x_1,\ldots,x_m]^T$ is a vector in $\mathbb{R}^m$; accordingly, for $p \geq 1$, $\|x\|_p = \left(|x_1|^p + \cdots + |x_m|^p\right)^{1/p}$ defines the p-norm and $\|x\|_\infty = \max(|x_1|,\ldots,|x_m|)$ defines the ∞-norm of x. The relation between these norms is as follows (Golub and Van Loan, 1996):

$$\|x\|_\infty \leq \|x\|_p \leq m^{1/p} \|x\|_\infty$$

Theorem 1 (Jensen's Inequality (Rudin, 1987)) *Let x be an integrable real-valued random variable and let φ a concave function such that $\varphi(x)$ is integrable. Then $\mathbb{E}[\varphi(x)] \leq \varphi(\mathbb{E}[x])$.*

Now by assuming that x is an integrable stochastic vector, and by considering the monotonicity and linearity of the expected value operator, we can derive the following inequalities:

$$\begin{aligned}\mathbb{E}\big[\max(x_1,\ldots,x_m)\big] &\overset{(i)}{\leq} \mathbb{E}\big[\max(|x_1|,\ldots,|x_m|)\big]\\ &\overset{(ii)}{\leq} \mathbb{E}\big[(|x_1|^p+\cdots+|x_m|^p)^{1/p}\big]\\ &\overset{(iii)}{\leq} \left(\sum_{j=1}^{m}\mathbb{E}\big[|x_j|^p\big]\right)^{1/p} \end{aligned} \tag{17}$$

where the last inequality is the result of applying Jensen's inequality for concave functions, since $\varphi(x)=x^{1/p}$ is a concave function for $p\geq 1$ and $x>0$, and in our case the argument x is $\sum_{j=1}^{m}|x_j|^p$ which is positive and integrable by our assumption.

Inequality (i) reduces to an equality if all variables x_j are nonnegative. Hence, in order to reduce the error in inequality (i), we define $x_j=y_j-L$ for some offset L such that x_j is almost always positive. Note that if y_j is from a distribution with a finite domain (such as the uniform distribution), L can be defined such that $L\leq y_j$ for $j=1,\ldots,m$ and hence, inequality (i) turns into an equality. However, if y_j has no finite bounds (such as in case of the normal distribution), inequality (i) never reduces to an equality and we can only decrease the error by defining L such that it is less than or equal almost all $y_j,\ j=1,\ldots,m$. For example if $y_j,\ j=1,\ldots,m$ are normally distributed with mean μ_j and variance σ_j, then L can be defined as $L=\min_{j=1,\ldots,m}(\mu_j-3\sigma_j)$. This choice of L has been made based on the 3σ-rule, which states that 99.7% of observations of a normally distributed random variable with mean μ and standard deviation σ falls within the interval $[\mu-3\sigma,\mu+3\sigma]$.

Remark 2 For a positive even integer $p=2q,\ q\in\mathbb{N}\setminus\{0\}$, we have $\mathbb{E}[x^p]=\mathbb{E}[|x|^p]$. Hence, without loss of generality, we can use $\mathbb{E}[x^p]$ in our problem by only considering even values for p in the sequel. So from now on, p is an even integer larger than or equal to 2.

Now let[4] $y_{ij}=\alpha_{ij}+\Pi_{ij}^T\hat{\theta}+\hat{\lambda}^T\Gamma_{ij}e$ (cf. Section 3, eq. (8)). Hence, the random variable $x_{ij}=y_{ij}-L$ can be written in a compact form as $x_{ij}=\beta_{ij}+\gamma_{ij}^Te$ where $\beta_{ij}=\alpha_{ij}+\Pi_{ij}^T\hat{\theta}-L$ and $\gamma_{ij}=(\hat{\lambda}^T\Gamma_{ij})^T$. Now by adopting (17) and considering (8), we have:

$$\begin{aligned}\mathbb{E}[\eta_i(k+1,\hat{\theta},\hat{\lambda},e(k))]-L &= \mathbb{E}[\max(x_{i1},\ldots,x_{im})]\\ &\leq \left(\sum_{j=1}^{m}\mathbb{E}\big[x_{ij}^p\big]\right)^{1/p}\\ &= \left(\sum_{j=1}^{m}\mathbb{E}\big[(\beta_{ij}+\gamma_{ij}^Te)^p\big]\right)^{1/p}\\ &= \left(\sum_{j=1}^{m}\mathbb{E}\big[(\underbrace{\beta_{ij}}_{z_{ij,0}}+\underbrace{\gamma_{ij,1}e_1}_{z_{ij,1}}+\cdots+\underbrace{\gamma_{ij,n_e}e_{n_e}}_{z_{ij,n_e}})^p\big]\right)^{1/p}\\ &\overset{(*)}{=} \left(\sum_{j=1}^{m}\mathbb{E}\Big[\sum_{k_0+k_1+\cdots+k_{n_e}=p}\binom{p}{k_0,k_1,\ldots,k_{n_e}}\prod_{t=0}^{n_e}z_{ij,t}^{k_t}\Big]\right)^{1/p}\end{aligned}$$

[4] In the rest of this section, the index k will be dropped (except for η_i and $\hat{\eta}_i$) for the sake of simplicity of notation.

$$= \left(\sum_{j=1}^{m} \sum_{k_0+k_1+\cdots+k_{n_e}=p} \frac{p!}{k_0!k_1!\cdots k_{n_e}!} \mathbb{E}\big[\prod_{t=0}^{n_e} z_{ij,t}^{k_t}\big] \right)^{1/p}$$

$$\overset{(**)}{=} \left(\sum_{j=1}^{m} \sum_{k_0+k_1+\cdots+k_{n_e}=p} \frac{p!}{k_0!k_1!\cdots k_{n_e}!} \prod_{t=0}^{n_e} \mathbb{E}\big[z_{ij,t}^{k_t}\big] \right)^{1/p} \tag{18}$$

where $(*)$ is based on the multinomial theorem for $(z_{ij,0}+\cdots+z_{ij,n_e})^p$, which is the generalization of the binomial theorem to polynomials (Graham et al, 1994), and $(**)$ is due the fact that, by assumption, the elements of the stochastic vector e, i.e., $e_1,\ldots,e_{n_e}$ are independent and for independent random variables $Z_1,\ldots,Z_{n_e}$, $\mathbb{E}[\prod_{t=1}^{n_e} Z_t] = \prod_{t=1}^{n_e} \mathbb{E}[Z_t]$.

Consequently, we can approximate the function $\hat{\eta}_i(k+1,\hat{\theta},\hat{\lambda}) = \mathbb{E}[\eta_i(k+1,\hat{\theta},\hat{\lambda},e(k))]$ by $\hat{\eta}_{\text{app},i}(k+1,\hat{\theta},\hat{\lambda})$ for an appropriate choice of p where

$$\hat{\eta}_{\text{app},i}(k+1,\hat{\theta},\hat{\lambda}) = \left(\sum_{j=1}^{m} \sum_{k_0+k_1+\cdots+k_{n_e}=p} \frac{p!}{k_0!k_1!\cdots k_{n_e}!} \prod_{t=0}^{n_e} \mathbb{E}\big[z_{ij,t}^{k_t}\big] \right)^{1/p} + L \tag{19}$$

where $z_{ij,0} = \alpha_{ij} + \Pi_{ij}^T \hat{\theta} - L$ and $z_{ij,t} = (\hat{\lambda}^T \Gamma_{ij})_t e_t$ for $t = 1,\ldots,n_e$.

In the approximation function $\hat{\eta}_{\text{app},i}(k+1,\hat{\theta},\hat{\lambda})$, we have to compute the k_t-th moment of each random variable $z_{ij,t}$, $t = 0,\ldots,n_e$ and $j = 1,\ldots,m$. By definition, the κ-th moment of a random variable z with can be computed as follows:

$$\mathbb{E}[z^\kappa] = \int_{-\infty}^{\infty} z^\kappa f(z) dz \tag{20}$$

where $f(z)$ is the probability density function of z and without loss of generality, we assume that κ is an even integer larger than or equal to 2 (cf. Remark 2). Note that since $z_{ij,0} = \alpha_{ij} + \Pi_{ij}^T \hat{\theta} - L$ does not include any elements of the stochastic vector e and hence, it is not a random variable, we have $\mathbb{E}[z_{ij,0}^\kappa] = z_{ij,0}^\kappa$ for any integer κ.

In general, moments of a random variable can be finite or infinite. Hence, to be able to apply $\hat{\eta}_{\text{app},i}(k+1,\hat{\theta},\hat{\lambda})$ as an approximation of $\hat{\eta}_i(k+1,\hat{\theta},\hat{\lambda})$, we need to consider random variables with finite moments for which a closed-form expression exists, such as variables with a uniform distribution, normal distribution, Beta distribution, etc. Note that if moments do not have a closed-form, one has to solve the integral (20) numerically. In that case, there is no advantage of using this approximation method since it is not time-efficient any more, and using numerical or analytic integration presented in Sections 4.2 and 4.3 would even be better options. In the following, we present some examples of finite moments of few distributions that have a closed-form as well: the uniform distribution, the Beta distribution, and the normal distribution. To find the moments of other distributions, the interested reader is referred to (Papoulis, 1991). For the case of a uniformly distributed random variable z on an interval $[a,b]$, i.e., $z \sim \mathscr{U}(a,b)$, the κ-th moment can be computed as

$$\mathbb{E}[z^\kappa] = \frac{1}{\kappa+1} \sum_{l=0}^{\kappa} a^l b^{\kappa-l} \tag{21}$$

and for a random variable z that has a Beta distribution with parameters α and β, i.e., $z \sim \mathscr{B}(\alpha,\beta)$, the κ-th moment can be written in a recursive form as

$$\mathbb{E}[z^\kappa] = \frac{\alpha+\kappa-1}{\alpha+\beta+\kappa-1} \mathbb{E}(z^{\kappa-1}). \tag{22}$$

In case of a normally distributed random variable z with mean μ and variance σ^2, i.e., $z \sim \mathcal{N}(\mu, \sigma^2)$, the κ-th moment has a closed-form that can be expressed as (Willink, 2005):

$$\mathbb{E}\big[z^{\kappa}\big] = \sigma^{\kappa} i^{-\kappa} H_{\kappa}(i\mu/\sigma) \tag{23}$$

where

$$H_{\kappa}(z) \equiv (-1)^{\kappa} \exp(z^2/2) \frac{d^{\kappa}}{dz^{\kappa}} \exp(-z^2/2)$$

is the κ-th Hermite polynomial. Note that the right-hand side of (23) is in fact real because $H_{\kappa}(z)$ contains only even powers of z if κ is even (note that here we assume that $\kappa = 2q$, $q \in \mathbb{N} \setminus \{0\}$). Considering equations (26.2.51) and (22.3.11) in Abramowitz and Stegun (1964) leads to

$$H_{\kappa}(z) = \kappa! \sum_{l=0}^{\kappa/2} \frac{(-1)^l z^{\kappa-2l}}{2^l l! (\kappa-2l)!} \tag{24}$$

where $\kappa/2 \in \mathbb{N} \setminus \{0\}$ since κ is an even integer in our case.

Remark 3 For the case of a normally distributed stochastic vector e, the random variable $x_{ij} = \beta_{ij} + \gamma_{ij}^T e$ is also normally distributed with a certain mean and standard deviation, using the property of the normal distribution that sum of the independent normally distributed random variables has also a normal distribution (Dekking et al, 2005). Hence, we can immediately compute the p-th moment in (17) and we do not need to use (18). In this way, our computation will be faster since we have less terms (compare (17) with (18)). In general, this remark is valid for all distributions that are preserved under the summation and for which a closed form of their higher-order moments exists, such as Poisson and Gamma distributions (Papoulis, 1991).

Furthermore, we can obtain gradients of $\hat{\eta}_{\text{app},i}(\hat{\theta}, \hat{\lambda})$ with respect to $\hat{\theta}$ and $\hat{\lambda}$. Recall that $z_{ij,0} = \alpha_{ij} + \Pi_{ij}^T \hat{\theta} - L$ and $z_{ij,t} = (\hat{\lambda}^T \Gamma_{ij})_t e_t$ for $j = 1, \ldots, m$ and $t = 1, \ldots, n_e$ with the stochastic vector $e = [e_1, \ldots, e_{n_e}]^T$ and Γ_{ij} being a diagonal matrix. Hence, only $z_{ij,0}$ depends on $\hat{\theta}$, and the rest of $z_{ij,t}$, $t = 1, \ldots, n_e$ depend only on $\hat{\lambda}$. Accordingly, by applying the chain rule, we obtain the following subgradients:

$$\nabla_{\hat{\theta}} \hat{\eta}_{\text{app},i}(k+1, \hat{\theta}, \hat{\lambda}) = \Bigg(\sum_{j=1}^{m} \sum_{k_0+k_1+\cdots+k_{n_e}=p} \frac{p!}{k_0! k_2! \cdots k_{n_e}!} \prod_{t=0}^{n_e} \mathbb{E}\big[z_{ij,t}^{k_t}\big] \Bigg)^{1/p-1} \times \Bigg(\sum_{j=1}^{m} \sum_{k_0+k_1+\cdots+k_{n_e}=p-1} \frac{(p-1)!}{k_0! k_2! \cdots k_{n_e}!} k_0 z_{ij,0}^{k_0-1} \Pi_{ij} \prod_{t=1}^{n_e} \mathbb{E}\big[z_{ij,t}^{k_t}\big] \Bigg)$$

and

$$\nabla_{\hat{\lambda}} \hat{\eta}_{\text{app},i}(k+1, \hat{\theta}, \hat{\lambda}) = \Bigg(\sum_{j=1}^{m} \sum_{k_0+k_1+\cdots+k_{n_e}=p} \frac{p!}{k_0! k_2! \cdots k_{n_e}!} \prod_{t=0}^{n_e} \mathbb{E}\big[z_{ij,t}^{k_t}\big] \Bigg)^{1/p-1} \times \Bigg(\sum_{j=1}^{m} \sum_{k_0+k_1+\cdots+k_{n_e}=p} \frac{(p-1)!}{k_0! k_2! \cdots k_{n_e}!} z_{ij,0}^{k_0} \sum_{\ell=1}^{n_e} k_\ell (\Gamma_{ij})_{\ell\ell} \mathbb{E}[e_\ell z_{ij,\ell}^{k_\ell-1}] \prod_{\substack{t=1 \\ t\neq\ell}}^{n_e} \mathbb{E}\big[z_{ij,t}^{k_t}\big] \Bigg)$$

As a result, we can approximate $J(\hat{\theta},\hat{\lambda})$ in (6) by replacing $\hat{\eta}_i(k+1,\hat{\theta},\hat{\lambda})$ with $\hat{\eta}_{\text{app},i}(k+1,\hat{\theta},\hat{\lambda})$ as follows:

$$J_{\text{app}}(\hat{\theta},\hat{\lambda}) = \sum_{k=1}^{N-1}\sum_{i=1}^{n}\Big(\hat{\eta}_{\text{app},i}(k+1,\hat{\theta},\hat{\lambda})\Big)^2 \tag{25}$$

with the gradients

$$\nabla_{\hat{\theta}} J_{\text{app}}(\hat{\theta},\hat{\lambda}) = \sum_{k=1}^{N-1}\sum_{i=1}^{n} 2\hat{\eta}_{\text{app},i}(k+1,\hat{\theta},\hat{\lambda})\,\nabla_{\hat{\theta}}\hat{\eta}_{\text{app},i}(k+1,\hat{\theta},\hat{\lambda})$$

$$\nabla_{\hat{\lambda}} J_{\text{app}}(\hat{\theta},\hat{\lambda}) = \sum_{k=1}^{N-1}\sum_{i=1}^{n} 2\hat{\eta}_{\text{app},i}(k+1,\hat{\theta},\hat{\lambda})\,\nabla_{\hat{\lambda}}\hat{\eta}_{\text{app},i}(k+1,\hat{\theta},\hat{\lambda})$$

and solve the optimization problem by means of a gradient-based optimization method, such as a steepest descent method or a Quasi-Newton (DFP, BFGS) method (Pardalos and Resende, 2002).

5.2 On the Error of the Approximation Method

Note that $\mathbb{E}[\max(x_1,\ldots,x_m)]$ in (17) is bounded from above and from below. Its upper bound has been presented in (17) and its lower bound can be obtained by using Jensen's inequality for convex functions (the max function in this case) (Boyd and Vandenberghe, 2004). Hence,

$$\max(\mathbb{E}[x_1],\ldots,\mathbb{E}[x_m]) \le \mathbb{E}[\max(x_1,\ldots,x_m)] \le \left(\sum_{j=1}^{m}\mathbb{E}\big[|x_j|^p\big]\right)^{1/p} \tag{26}$$

Consequently, the error of approximating $\mathbb{E}[\max(x_1,\ldots,x_m)]$ by its upper bound in (17) is always bounded from above by

$$\left(\sum_{j=1}^{m}\mathbb{E}\big[|x_j|^p\big]\right)^{1/p} - \mathbb{E}[\max(x_1,\ldots,x_m)] \le \left(\sum_{j=1}^{m}\mathbb{E}\big[|x_j|^p\big]\right)^{1/p} - \max(\mathbb{E}[x_1],\ldots,\mathbb{E}[x_m]) \tag{27}$$

and since in our case $x_j,\ j=1,\ldots,m$ are assumed to have finite moments, this upper bound is finite and the error of the approximation cannot be larger than this value.

Since the error bound (27) is a rough upper bound, we introduce another upper bound for the approximation error that is tighter than (27). This new upper bound is exact for the case of having a stochastic vector that has a probability distribution with a bounded domain (such as in case of a uniform distribution), and is an approximate upper bound for the case that the stochastic vector has a probability distribution with an unbounded domain (such as for the normal distribution). To obtain the new upper bound, we consider the three inequalities in (17) and their corresponding error. The first error, due to (i), approaches zero if L becomes more and more negative for the case of a probability distribution with an unbounded domain. In case of a probability distribution with a bounded domain this error is zero for an appropriate choice of L. Regardless of the type of the probability distribution, the second error due to (ii) approaches zero if $p \to +\infty$, since by definition $\|x\|_\infty = \lim_{p\to+\infty}\|x\|_p$. However, the third error, which is in fact the error of Jensen's inequality, needs more discussion. In (Simić, 2009a, Theorem[5] 2.1) and (Simić, 2009b, Theorem 2.1) two upper bounds

[5] This theorem is in fact a special case of the results appeared in (Pečarić and Beesack, 1987), as explained in (Ivelić and Pečarić, 2011)

for Jensen's inequality are presented for the fractional and the absolute error, respectively. For a (strictly) positive, twice continuously differentiable, concave function f defined on an interval $[a,b]$, Jensen's inequality can be stated in the form

$$1 \leq \frac{f(\mathbb{E}[x])}{\mathbb{E}[f(x)]}$$

for which an upper bound can be formulated as follows (Simić, 2009a):

$$1 \leq \frac{f(\mathbb{E}[x])}{\mathbb{E}[f(x)]} \leq \max_{q\in[0,1]}[\frac{f(qa+(1-q)b)}{qf(a)+(1-q)f(b)}] := S_f(a,b)$$

and it has been proven in Simić (2009a) that there exists a unique q_0 for which $S_f(a,b)$ is maximal. In a similar way, the absolute error can be also defined as follows (Simić, 2009b): For a differentiable, concave function f defined on an interval $[a,b]$ we have

$$0 \leq f(\mathbb{E}[x]) - \mathbb{E}[f(x)] \leq \max_{\omega\in[0,1]}[f(\omega a+(1-\omega)b) - \omega f(a) - (1-\omega)f(b)] := T_f(a,b)$$

and again it has been shown that there exists a unique ω_0 for which $T_f(a,b)$ is maximal (Simić, 2009b).

In our case the concave function is $f(x) = x^{1/p}$ and since we assume that p is a positive even integer greater than or equal to 2, the argument x has to be larger or equal to zero, which is the case since $x = \sum_{j=1}^{m} x_j^p$. Note that with this choice of p, $f(x)$ is in fact a strictly concave function. Now, by substituting f in the above formulas and determining the optimal value of q and ω for each case, the following expressions are obtained for $S_f(a,b)$ and $T_f(a,b)$:

$$S_f(a,b) = \Big(\frac{\frac{1}{p}(ab^{\frac{1}{p}+1} - a^2b^{\frac{1}{p}} - a^{\frac{1}{p}}b^2 + a^{\frac{1}{p}+1}b)}{-(\frac{p-1}{p})(a^{\frac{1}{p}+1} + b^{\frac{1}{p}+1} - a^{\frac{1}{p}}b - ab^{\frac{1}{p}})}\Big)^{\frac{1}{p}} \cdot \frac{-(\frac{p-1}{p})(a^{\frac{1}{p}+1} + b^{\frac{1}{p}+1} - a^{\frac{1}{p}}b - ab^{\frac{1}{p}})}{-a^{\frac{1}{p}}b^{\frac{1}{p}+1} - a^{\frac{1}{p}+1}b^{\frac{1}{p}} + a^{\frac{2}{p}}b + ab^{\frac{2}{p}}}$$

$$T_f(a,b) = (\frac{a-b}{p(a^{\frac{1}{p}} - b^{\frac{1}{p}})})^{\frac{1}{p-1}} - \Big(\frac{1}{a-b}[(a^{\frac{1}{p}} - b^{\frac{1}{p}})(\frac{a-b}{p(a^{\frac{1}{p}} - b^{\frac{1}{p}})})^{\frac{p}{p-1}}) - a^{\frac{1}{p}}b + ab^{\frac{1}{p}}]\Big) \quad (28)$$

Note that from the above formulas for different values of a,b and p, we can conclude the following:

$$\text{if } (a \to \infty \text{ or } b \to \infty) \text{ and } (p < \infty): \begin{cases} S_f(a,b) \to \infty \\ T_f(a,b) \to \infty \end{cases}$$

$$\text{if } (a < \infty \text{ and } b < \infty) \text{ and } (p \to \infty): \begin{cases} S_f(a,b) \to 1 \\ T_f(a,b) \to 0 \end{cases}$$

$$\text{if } (a = 0 \text{ and } b < \infty) \text{ and } (p < \infty): \begin{cases} S_f(a,b) \to \infty \\ T_f(a,b) < \infty \end{cases}$$

Therefore, to have finite upper bounds for Jensen's inequality, a and b have to be finite, and $a \neq 0$ for computing $S_f(a,b)$. Note that for the case of a probability distribution with a bounded domain, a and b can be easily obtained. Assume that each $x_{ij} = y_{ij} - L$ belongs to the interval $[c_{ij,1}, c_{ij,2}]$, and since $L \leq y_{ij}$ for all $j = 1,\ldots,m$, we can conclude that $0 \leq c_{ij,1} \leq c_{ij,2}$ and hence, $c_{ij,1}^p \leq c_{ij,2}^p$. Therefore,

$$c_{ij,1}^p \leq x_{ij}^p \leq c_{ij,2}^p \Rightarrow \underbrace{\sum_{j=1}^{m} c_{ij,1}^p}_{a} \leq \underbrace{\sum_{j=1}^{m} x_{ij}^p}_{x} \leq \underbrace{\sum_{j=1}^{m} c_{ij,2}^p}_{b} \quad (29)$$

However, if we have a probability distribution with an unbounded domain, we need to define an approximate a and b. For example, assume that $x_{ij} = y_{ij} - L$ is normally distributed, i.e., $x_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma_{ij}^2)$ for $j = 1, \ldots, m$, which by nature is not bounded. However, since by 3σ-rule, each x_{ij} is in the interval $[\mu_{ij} - 3\sigma_{ij} - L, \mu_{ij} + 3\sigma_{ij} - L]$ with probability 99.7%, we make an assumption that

$$\mu_{ij} - 3\sigma_{ij} - L \leq y_{ij} - L \leq \mu_{ij} + 3\sigma_{ij} - L \quad \Rightarrow \quad c_{ij,1} \leq x_{ij} \leq c_{ij,2}$$

where $c_{ij,1} := \mu_{ij} - 3\sigma_{ij} - L$ and $c_{ij,2} := \mu_{ij} + 3\sigma_{ij} - L$. Since we have $L = \min_j(\mu_{ij} - 3\sigma_{ij})$, it follows that $0 \leq c_{ij,1} \leq c_{ij,2}$ and consequently $c_{ij,1}^p \leq c_{ij,2}^p$. Hence,

$$c_{ij,1}^p \leq x_{ij}^p \leq c_{ij,2}^p \Rightarrow \quad \underbrace{\sum_{j=1}^{m} c_{ij,1}^p}_{a} \leq \underbrace{\sum_{j=1}^{m} x_{ij}^p}_{x} \leq \underbrace{\sum_{j=1}^{m} c_{ij,2}^p}_{b} \tag{30}$$

Note that these error bounds are only an approximation, since we leave out the cases where $x_{ij} > \mu_{ij} + 3\sigma_{ij} - L$ and $x_{ij} < \mu_{ij} - 3\sigma_{ij} - L$ for $j = 1, \ldots, m$.

As mentioned above, for finite a and b, if $p \to \infty$ then $S_f(a,b)$ converges to 1 and $T_f(a,b)$ converges to 0. This suggests that in order to get a good approximation, p should not be selected too small. However, since in our case both a and b depend on p (as shown in (29) and (30)), they will approach infinity if $p \to \infty$ and consequently both $S_f(a,b)$ and $T_f(a,b)$ become infinite. This suggests that p should not be selected too large either. Hence, there is a trade-off between the choice of p and the magnitude of the (approximation) error.

Moreover, in our case $a = 0$ is very improbable to occur. Recall that the random variable x_{ij} has the following form:

$$x_{ij} = \alpha_{ij} + \Pi_{ij}^T \hat{\theta} + \hat{\lambda}^T \Gamma_{ij} e - L$$

where all the elements of e are independent. Hence, $a = 0$ only if all the elements of the vector α_i and matrices Π_i and Γ_{ij} are equal, and this is very unlikely to be the case. Consequently, by considering the cases in which a is not zero, we can compute both upper bounds $S_f(a,b)$ and $T_f(a,b)$, and if $a = 0$, we can only use $T_f(a,b)$ as an upper bound for Jensen's inequality.

6 Complexity analysis of the two proposed approaches

Even if the integral in (16) can be computed analytically, the computational load is still quite heavy. This is because the method in Section 4.3 contains two time-consuming steps: In the first step all polyhedra $\Psi_{ij\ell}$ have to be specified. Note that $\Psi_{ij\ell}(\hat{\theta}, \hat{\lambda}, k) = \Omega_{ij}(\hat{\theta}, \hat{\lambda}, k) \cap P_\ell$ where the number of polyhedra Ω_{ij} is equal to $nm = n^2 + n_u \cdot n$ and the number of polyhedra P_ℓ is n_p. Hence, in the worst case the number of polyhedra $\Psi_{ij\ell}$ that has to be considered is $O(n(n+n_u)n_p)$, which becomes more and more time-consuming as n_p, n_u, and n become larger. In the second step, the integral over each of these regions has to be calculated, for which in case of a uniform probability density function, we need to compute all the vertices of each polyhedron $\Psi_{ij\ell}$. As explained in (Mattheiss and Rubin, 1980), we have the following upper bound for the number of the vertices of a polytope defined by m (non-redundant) inequality constraints in an n_e-dimensional space:

$$\binom{m - \lfloor \frac{n_e+1}{2} \rfloor}{m - n_e} + \binom{m - \lfloor \frac{n_e+2}{2} \rfloor}{m - n_e}$$

This means that in our case with $m = n + n_u$ inequality constraints in an n_e-dimensional space, the number of vertices for the worst case can be $O((n+n_u)^{\lfloor \frac{n_e}{2} \rfloor})$ if $m \gg n_e \gg 1$, which is again time-consuming as n, n_u and n_e increase. In the case of having other piecewise polynomial probability density functions, the order of complexity of the second step becomes even bigger since then, the integral computation is more complex than in the case of the uniform distribution. Accordingly, the complexity of the whole procedure in the worst case is of the order $O(nn_p(n+n_u)^{\lfloor \frac{n_e}{2} \rfloor})$ for the first approach in the case of a uniformly distributed noise.

However, due to the structure of the second approach, none of the two above-mentioned steps are present. Hence by considering (19), the total number of terms in the first sum is $m = n + n_u$ and in the second sum, i.e., the multinomial sum, is $\binom{p+n_e-1}{p}$ and assuming that $n_e \gg p > 1$, the order of the error for this sum is $O(\frac{n_e^p}{p!})$. Also, the total number of the expected values that have to computed is pn_e. Also, since i changes from 1 to n, there are n terms. Hence, the complexity of this approximation method is of the order $O(n(n+n_u)n_e p\binom{p+n_e-1}{p}) = O(n(n+n_u)n_e p\frac{n_e^p}{p!})$, which increases polynomially as n, n_u, or n_e increase and exponentially as p increases.

Hence, the complexity order of these two methods shows that the second approach is computationally more time-efficient than the first one.

7 Example

In this section we present two examples to study the performance of the first and the second approach. In the first example, we consider a uniformly distributed noise vector, which has a bounded domain, and we compare the performance of the first and the second approach with each other and with the approach that uses Monte Carlo simulation for the computation of the expected values. In the second example, a normally distributed noise vector, which has no bounded domain, is considered. Note that if we apply the analytic integration approach of Section 4.3 to the case with normally distributed random variables, we would need an approximation using piecewise-polynomial functions. This would introduce approximation errors as well as an increase in computational complexity. Hence, to avoid the additional complexity, we do not compare the performance of these two methods with one another in this example. We only compare the performance of the second approach with the first approach using the numerical integration (cf. Section 4.2) and with the one Monte Carlo simulation for the computation of the expected values.

7.1 Example 1: uniform distribution

In this example we apply the first method to estimate the parameters θ and λ, for the case with uniformly distributed noise. We consider the following state space model:

$$x(k) = A(k) \otimes x(k-1) \oplus B(k) \otimes u(k) \tag{31}$$

$$y(k) = C(k) \otimes x(k) \tag{32}$$

with the system matrices

$$A(k) = \begin{bmatrix} \theta_1(k) & 0 \\ \varepsilon & \theta_2(k) \end{bmatrix} \qquad B(k) = \begin{bmatrix} \theta_3(k) \\ \theta_4(k) \end{bmatrix} \qquad C(k) = \begin{bmatrix} 0 & 0 \end{bmatrix}$$

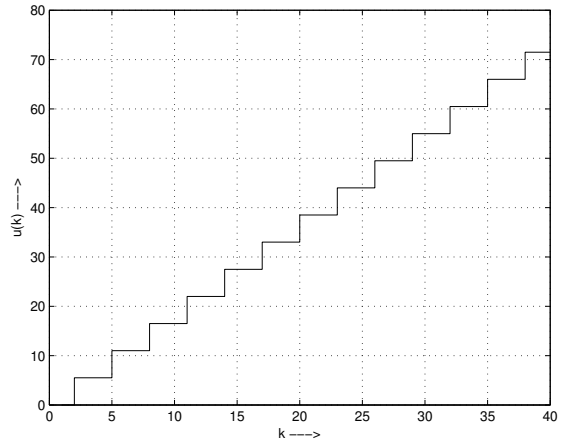

Fig. 1 The first 40 samples of the input signal $u(k)$ for the first and the second example.

to obtain a system of the form (1), where the true parameter vector θ is given by

$$\theta = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 & \theta_4 \end{bmatrix}^T = \begin{bmatrix} 0.3 & 0.3 & 0.7 & 0.6 \end{bmatrix}^T$$

These parameters are perturbed by uniformly distributed noise components $e_t(k)$ with $e_t(k) \sim \mathscr{U}(-1,1)$ for $t = 1,\ldots,4$, and with scaling factor

$$\lambda = \begin{bmatrix} \lambda_1 & \lambda_2 & \lambda_3 & \lambda_4 \end{bmatrix}^T = \begin{bmatrix} 0.3 & 0.3 & 0.3 & 0.3 \end{bmatrix}^T.$$

In this simulation study we simulate the system for 400 event steps, i.e., for $k = 1,\ldots,400$. The parameter estimation is done with input-state data where the input signal is a staircase signal with an average slope of about 1.83, given by

$$u(k) = 5.5 \cdot \Big(1 + \lfloor k/3 \rfloor\Big)$$

where $\lfloor x \rfloor$ denotes the largest integer less than or equal to x. The input signal $u(k)$ in shown in Figure 1 for $k = 1,\ldots,40$.

As a first step, we estimate the parameter θ for a deterministic model, i.e., a noiseless model with $\hat{\lambda} = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}^T$, using the residuation-based estimation techniques described by (Baccelli et al, 1992; Cuninghame-Green, 1979; Menguy et al, 2000). Note that in this case, we do not expect to have a good estimation since we are ignoring the effect of noise. The optimization result is as follows:

$$\hat{\theta} = \begin{bmatrix} 0.0167 & 0.0009 & 0.4056 & 0.3011 \end{bmatrix}^T.$$

As we expected, due to the absence of a noise model the estimation fails and the estimated parameters are quite far from the true values.

The second step is to estimate the parameters θ and λ for the stochastic system (31)-(32). We minimize the cost function (15) based on the one-step ahead prediction, i.e., we predict the behavior of the system at the event step $k+1$ based on the information that we have at the event step k. We use a multi-start, sequential quadratic programming (SQP)

method, considering 30 different initial values that are chosen randomly with both larger and smaller values than the real ones, to start the optimization with, and then we report the estimated parameters for which the cost function has the lowest value.

Methods to compute the expected value	Monte Carlo simulation	Numerical integration	Analytic integration	Approximation method
$\hat{\theta}$	$\begin{bmatrix}0.2824\\0.2944\\0.6910\\0.5885\end{bmatrix}$	$\begin{bmatrix}0.2976\\0.3018\\0.6967\\0.5898\end{bmatrix}$	$\begin{bmatrix}0.2841\\0.2926\\0.6954\\0.5808\end{bmatrix}$	$\begin{bmatrix}0.3012\\0.2823\\0.6746\\0.5991\end{bmatrix}$
$\hat{\lambda}$	$\begin{bmatrix}0.1575\\0.3882\\0.6224\\0.0027\end{bmatrix}$	$\begin{bmatrix}0.2838\\0.4694\\0.4529\\0.0896\end{bmatrix}$	$\begin{bmatrix}0.4591\\0.3239\\0.0858\\0.2700\end{bmatrix}$	$\begin{bmatrix}0.0596\\0.0670\\0.2613\\0.0479\end{bmatrix}$
CPU time	73549 s	44992 s	1523 s	1024 s

Table 1 Estimation results for θ and λ, using four different methods to calculate the expected value in (7) with uniformly distributed noise, and the average computation time (CPU time) of each method.

We use four different methods to compute the expected value in (7): Monte Carlo simulation (Kalos and Whitlock, 2008), numerical integration (cf. Section 4.2), the analytic integration method explained in Section 4.3, and the approximation method of Section 5 (cf. (19)). By means of experiments, we have found out that $p = 14$ gives good approximation in this specific example. The results of the optimization are presented in Table 1. As shown, the estimated parameter $\hat{\theta}$ is quite close to the exact value of θ for the above-mentioned methods. However, for λ we do not have a good estimation. Note that, in general, in prediction error identification, one can obtain the correct system model, i.e., θ, but it is much more difficult to estimate the noise model, i.e., λ (Goodwin and Payne, 1977; Ljung, 1999).

The reason that the analytic integration method of Section 4.3 and the numerical integration give different results (cf. Table 1) is – apart from the numerical integration accuracy – mainly due to the fact that here we have independent experiments with different random initial values. As reported in Table 1, for 400 event steps, the computation time[6] of the optimization procedure for one initial value using the approximation method and using the analytic integration approach is quite close (it is about a factor 1.5 lower for the approximation method) since both methods are analytic. However, the computation time of the optimization problem using the analytic integration approach is about a factor 30 lower than using numerical integration with 10^5 samples. If we increase the number of samples to 10^7 the computation time using numerical integration becomes about a factor 3000 larger than the one using the analytic integration, and for 10^{10} it is not even tractable anymore. For the numerical integration the relative error[7] between the analytic integration and the numerical integration using 10^5 samples is 0.03% and using 10^7 samples is 0.008%. Based on a trade-off between the CPU time and the relative error, it has been decided to do the experiments with 10^5 samples. The computation time of the optimization procedure using Monte Carlo simulation, reported in Table 1, is also for 10^5 samples and the relative error between

[6] These times were obtained running Matlab 7.5.0 (R2007b) on a 2.33 GHz Intel Core Duo E655 processor.

[7] The relative error is defined as $\frac{|x_0-x|}{|x|}$ where x is the true value and x_0 is the estimated value.

the analytic integration and the Monte Carlo simulation using this number of samples is 0.06% and using 10^7 samples, it is 0.003%. Hence, due to the same trade-off as before, we chose 10^5 samples. As a result, by comparing the CPU times of these four methods we can conclude that the analytic integration method 4.3 and the approximation method of Section 5 are considerably faster (at least 30 times and 45 times, respectively) than the numerical integration and the Monte Carlo simulation.

7.2 Example 2: normal distribution

Here we consider the same input-output stochastic max-plus-linear system as the one in Example 7.1. The true parameter vector θ is the same as before, i.e.,

$$\theta = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 & \theta_4 \end{bmatrix}^T = \begin{bmatrix} 0.3 & 0.3 & 0.7 & 0.6 \end{bmatrix}^T$$

except that now, each of its elements is perturbed by one of the noise components $e_t(k)$, $t = 1,\ldots,4$ that are independent and have a standard normal distribution, i.e., $e_t(k) \sim \mathcal{N}(0,1)$, with the scaling factor[8]

$$\lambda = \begin{bmatrix} \lambda_1 & \lambda_2 & \lambda_3 & \lambda_4 \end{bmatrix}^T = \begin{bmatrix} 0.1 & 0.1 & 0.1 & 0.1 \end{bmatrix}^T.$$

We also consider the same input signal as the one specified in Example 7.1.

Similar to the previous example, first we estimate the parameter θ for a deterministic model, using the mentioned residuation-based estimation techniques. The optimization result is as follows:

$$\hat{\theta} = \begin{bmatrix} 0.0725 & -0.0218 & 0.8416 & 0.7035 \end{bmatrix}^T.$$

As we expected and as we have also seen in Example 7.1, by neglecting the effect of noise, we do not obtain a good estimation.

As the next step, we estimate the parameters θ and λ for the above-mentioned stochastic system (31)-(32). To this end, we minimize the cost function (15) using three different methods: Monte Carlo simulation, the direct numerical integration of (9), and the approximation method of Section 5 (cf. (19)). As we did in Example 7.1, we minimize these cost functions based on the one-step ahead prediction, using a multi-start, SQP method with 30 different initial values, and reporting the estimated parameter with the lowest cost function value. We have chosen $p = 30$ for the approximation method. As before, this choice was done by means of experiments, for which we obtain good approximation in this specific example. The estimation results are presented in Table 2.

Comparing the results, we can conclude that the approximation method gives a good estimation for θ that is very close to the results obtained from the exact solution using numerical integration and Monte Carlo simulation. Similar to the first example, again we obtain an unsatisfactory estimation for λ. Note that the estimated values for λ using numerical integration is very close to the exact values. However, this result is random and is not repeated using different initial values.

Recall that one of the goals of using the proposed approximation method (19) is to decrease the computation time. For 400 event steps, the computation times of the optimization procedure using the three above-mentioned methods are presented in Table 2. As explained

[8] Note that here due to the 3σ-rule, we choose λ one third of the one in Example 7.1.

Methods to compute the expected value	Monte Carlo simulation	Numerical integration	Approximation method
$\hat{\theta}$	$\begin{bmatrix}0.3030\\0.2984\\0.6881\\0.5940\end{bmatrix}$	$\begin{bmatrix}0.3092\\0.2969\\0.6974\\0.5937\end{bmatrix}$	$\begin{bmatrix}0.2750\\0.2824\\0.6799\\0.5781\end{bmatrix}$
$\hat{\lambda}$	$\begin{bmatrix}0.0449\\0.0417\\0.0479\\0.0398\end{bmatrix}$	$\begin{bmatrix}0.0976\\0.1016\\0.1236\\0.0945\end{bmatrix}$	$\begin{bmatrix}0.0409\\0.0400\\0.0403\\0.0419\end{bmatrix}$
CPU time	110796 s	83890 s	899 s

Table 2 Estimation results for θ and λ, using three different methods to calculate the expected value in (7) with a normally distributed noise, and the average computation time (CPU time) of each method.

in Example 7.1, the reported CPU time for Monte Carlo simulation and numerical integration in this example is also for 10^5 samples. Therefore, the approximation method increases the time efficiency significantly (it is about 80 times faster than the two other methods) while still guaranteeing a comparable performance to the exact solution.

8 Conclusions

This paper has discussed the identification problem of stochastic max-plus-linear systems. Since we deal with stochastic systems, the solution of this problem leads to the computation of an expected value. We have proposed two approaches to for the computation of this expectation. The first approach uses either numerical integration or analytic integration. The analytic integration method can be applied to distributions that have a piecewise affine polynomial probability density function, or when their probability density functions can be approximated by such functions. The second approach is an approximation method based on higher-order moments of a random variable and we applied it with the assumption of having an error vector with independent components. This method is applicable to any distribution with finite moments, and it involves no analytic or numerical integration provided that a closed-form expression of the higher-order moments of that random variable exists. Since both the analytic integration method and the approximation method, using closed-form moments, result in an analytic solution, they are computationally much faster than the numerical integration or the Monte Carlo simulation. Moreover, since in the first and the second approach, an explicit expression for the gradient can be calculated, the parameter estimation can be done using gradient-based optimization methods.

One topic for future research is the development of algorithms for stochastic max-plus linear systems based on input-output data (instead of input-state data) or with only partial state information, based on the approaches proposed in this paper. Another interesting topic is to explore the possibilities of improving the estimation of the noise amplitude. Yet another topic would be to find a method to specify the most appropriate order of moments p, in order to obtain a better estimation in the second approach. It is also interesting to apply the proposed approaches to solve the identification problem of other classes of discrete-event systems such as max-min-plus systems and max-min-plus-scaling systems.

Acknowledgements The authors would like to thank Dr. Ioan Landau for his useful comments and suggestions and Dr. Hans van der Weide for his help in the derivation of the approximation method presented in Farahani et al (2010). This research is partially funded by the Dutch Technology Foundation STW project "Model-predictive railway traffic management" (11025), and by the European Union Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 257462 HYCON2 Network of Excellence.

References

Abramowitz MA, Stegun I (1964) Handbook of Mathematical Functions. National Bureau of Standards, US Government Printing Office, Washington DC

Akian M (2007) Representation of stationary solutions of Hamilton-Jacobi-Bellman equations: a max-plus point of view. In: Proceedings of the International Workshop "Idempotent and Tropical Mathematics and Problems of Mathematical Physics", Moscow, Russia

Baccelli F, Cohen G, Olsder G, Quadrat J (1992) Synchronization and Linearity. John Wiley & Sons, New York

Başar T, Bernhard P (1995) H_∞-Optimal Control and Related Minimax Design Problems, 2nd edn. Birkhauser, Boston, Massachusetts

Bemporad A, Borrelli F, Morari M (2003) Min-max control of constrained uncertain discrete-time linear systems. IEEE Transactions on Automatic Control 48(9):1600–1606

Boimond J, Hardouin L, Chiron P (1995) A modeling method of SISO discrete-event systems in max-algebra. In: Proceedings of the 3rd European Control Conference, Rome, Italy, pp 2023–2026

van den Boom T, De Schutter B (2004) Model predictive control for perturbed max-plus-linear systems: A stochastic approach. International Journal of Control 77(3):302–309

van den Boom T, De Schutter B, Verdult V (2003) Identification of stochastic max-plus-linear systems. In: Proceedings of the 2003 European Control Conference (ECC'03), Cambridge, UK, paper 104

Boyd S, Vandenberghe L (2004) Convex Optimization. Cambridge University Press, London, UK

Büeler B, Enge A, Fukuda K (2000) Exact volume computation for convex polytopes: A practical study. In: Kalai G, Ziegler G (eds) Polytopes – Combinatorics and Computation, Birkäuser Verlag, Basel, Switzerland, pp 131–154

Cassandras C, Lafortune S (1999) Introduction to Discrete Event Systems. Kluwer Academic Publishers, Boston, Massachusetts

Cuninghame-Green R (1979) Minimax Algebra, Lecture Notes in Economics and Mathematical Systems, vol 166. Springer-Verlag, Berlin, Germany

Davis PJ, Rabinowitz P (1984) Methods of Numerical Integration, 2nd edn. Academic Press, New York

Dekking FM, Kraaikamp C, Lopuhaä HP, Meester LE (2005) A Modern Introduction to Probability and Statistics. Springer, London, UK

Farahani SS, van den Boom T, van der Weide H, De Schutter B (2010) An approximation approach for model predictive control of stochastic max-plus linear systems. In: Proceedings of the 10th International Workshop on Discrete Event Systems (WODES'10), Berlin, Germany, pp 386–391

Gallot F, Boimond J, Hardouin L (1997) Identification of simple elements in max-algebra: Application to SISO discrete event systems modelisation. In: Proceedings of the European Control Conference (ECC'97), Brussels, Belgium, paper 488

Golub G, Van Loan C (1996) Matrix Computations, 3rd edn. The John Hopkins University Press, Baltimore, Maryland

Goodwin GC, Payne RL (1977) Dynamic System Identification: Experiment Design and Data Analysis. Academic Press, New York

Graham R, Knuth D, Patashnik O (1994) Concrete Mathematics: A Foundation for Computer Science, 2nd edn. Addison-Wesley, Boston, Massachusetts

Heidergott B, Olsder G, van der Woude J (2006) Max Plus at Work. Princeton University Press, Princeton, New Jersey

Ho Y (ed) (1992) Discrete Event Dynamic Systems: Analyzing Complexity and Performance in the Modern World. IEEE Press, Piscataway, New Jersey

Ivelić S, Pečarić JE (2011) Remarks on "on a converse of jensen's discrete inequality" of s. simić. Hindawi Publishing Corporation, Journal of Inequalities and Applications Article ID 309565

Kalos MH, Whitlock PA (2008) Monte Carlo Methods. Wiley-VCH, Weinheim, Germany

Lasserre J (1998) Integration on a convex polytope. Proceedings of the American Mathematical Society 126(8):2433–2441

Ljung L (1999) System Identification: Theory for the User, 2nd edn. Prentice-Hall, Upper Saddle River, New Jersey

Mairesse J (1994) Stochastic linear systems in the (max,+) algebra. In: 11th International Conference on Analysis and Optimization of Systems Discrete Event Systems, Lecture Notes in Control and Information Sciences, vol 199, Springer-Verlag, London, UK, pp 257–265

Mattheiss T, Rubin D (1980) A survey and comparison of methods for finding all vertices of convex polyhedral sets. Mathematics of Operations Research 5(2):167–185

McEneaney WM (2004) Max-plus eigenvector methods for nonlinear H-infinity problems: Error analysis. SIAM Journal of Control and Optimization 43:379–412

Menguy E, Boimond JL, Hardouin L, Ferrier JL (2000) A first step towards adaptive control for linear systems in max algebra. Discrete Event Dynamic Systems 10(4):347–368

Olsder G, Resing J, de Vries R, Keane M, Hooghiemstra G (1990) Discrete event systems with stochastic processing times. IEEE Transactions on Automatic Control 35(3):299–302

Papoulis A (1991) Probability, Random Variables, and Stochastic Processes. McGraw-Hill, Singapore

Pardalos P, Resende M (eds) (2002) Handbook of Applied Optimization. Oxford University Press, Oxford, UK

Peterson J (1981) Petri Net Theory and the Modeling of Systems. Prentice-Hall, Englewood Cliffs, New Jersey

Pečarić JE, Beesack PR (1987) On knopp's inequality for convex functions. Canadian Mathematical Bulletin 30(3):267–272

Resing J, de Vries R, Hooghiemstra G, Keane M, Olsder G (1990) Asymptotic behavior of random discrete event systems. Stochastic Processes and their Applications 36:195–216

Rudin W (1987) Real and Complex Analysis, 3rd edn. McGraw-Hill Series in Higher Mathematics, McGraw-Hill, New York

Schullerus G, Krebs V (2001a) Diagnosis of batch processes based on parameter estimation of discrete event models. In: Proceedings of the European Control Conference 2001 (ECC'01), Porto, Portugal, pp 1612–1617

Schullerus G, Krebs V (2001b) Input signal design for discrete event model based batch process diagnosis. In: Proceedings of the 4th IFAC Workshop on On-Line Fault Detection and Supervision in the Chemical Process Industries, Chejudo Island

Schullerus G, Krebs V, De Schutter B, van den Boom T (2003) Optimal input signal design for identification of max-plus-linear systems. In: Proceedings of the 2003 European

Control Conference (ECC'03), Cambridge, UK, paper 026

De Schutter B, van den Boom T, Verdult V (2002) State space identification of max-plus-linear discrete event systems from input-output data. In: Proceedings of the 41st IEEE Conference on Decision and Control, Las Vegas, Nevada, pp 4024–4029

Simić S (2009a) On a converse of Jensen's discrete inequality. Journal of Inequalities and Applications Article ID 153080

Simić S (2009b) On an upper bound for Jensen's inequality. Journal of Inequalities in Pure and Applied Mathematics 10(2), article 60

Somasundaram KK, Baras JS (2011) Solving multi-metric network problems an interplay between idempotent semiring rules. Linear Algebra and its Applications 435(7):1494–1512

Willink R (2005) Normal moments and Hermite polynomials. Statistics & Probability Letters 73(3):271–275