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

Direct System Identification of Dynamical Networks with Partial Measurements: a Maximum Likelihood Approach

João Victor Galvão da Mata[1], Anders Hansson[2], and Martin S. Andersen[1]

***Abstract*— This paper introduces a novel direct approach to system identification of dynamic networks with missing data based on maximum likelihood estimation. Dynamic networks generally present a singular probability density function, which poses a challenge in the estimation of their parameters. By leveraging knowledge about the network's interconnections, we show that it is possible to transform the problem into a more tractable form by applying linear transformations. This results in a nonsingular probability density function, enabling the application of maximum likelihood estimation techniques. Our preliminary numerical results suggest that when combined with global optimization algorithms or a suitable initialization strategy, we are able to obtain a good estimate of the dynamics of the internal systems.**

***Index Terms*— System identification, maximum likelihood estimation, dynamical networks, singular Gaussian distribution.**

I. Introduction

One of the greatest technical challenges in our society is to efficiently, sustainably and safely control large-scale complex systems. These systems put new demands on control theory. Many of the available methods for modeling, analysis and design do not scale well with increasing complexity. Furthermore, the majority of control theory has been developed in a centralized setting, where all measurements are processed together to compute the control signals. While this paradigm offers conceptual advantages, it is not without inherent limitations. In reality, industrial practice frequently relies on distributed control structures, underscoring the need for more systematic approaches to design and analysis of such structures. The scope of applications spans a wide spectrum, encompassing networks for transportation, communication, and energy supply, as well as industrial production, logistics, and healthcare.

Design methodologies for controller design are often model based, and the time and effort needed for modeling is usually substantially larger than for controller design. Thus, achieving scalable control methods hinges on the ability to obtain these models in an efficient and scalable manner.

Large-scale complex systems are typically described as interconnections of simpler subsystems, i.e., as networked systems. We will here address system identification for such dynamic networks. These networks frequently contain internal variables that are not directly measurable, necessitating methods that are capable of handling what is termed "latent variable" in statistics and "missing data" within the system identification community. In essence, these methods should be able to perform identification tasks while relying on only partially observable variables.

When dealing with interconnected systems, there are essentially two ways to estimate the parameters of the systems: the *direct approach* and the *indirect approach*, which are explained for a single feedback loop in [6]. These approaches can be generalized to networks of systems. The direct approach is typically based on minimizing prediction errors to obtain the system parameters [6], [8]. The indirect approach first estimates parameters that characterize the closed loop transfer function between the inputs and the observable variables, and it then uses these functions and knowledge about the architecture of the system to estimate the parameters of the transfer function for the subsystems [5], [9].

As we will show with an example, the indirect approach suffers from the problem that even if all the true subsystems are stable and the closed loop is stable, the estimates of the subsystems may be unstable. It also suffers from the fact that the number of parameters that are needed to describe the closed loop system can be significantly larger than the total number of parameters describing the individual subsystems. As a result, the variance of the estimated models may be larger than if a direct method is used. However, current direct methods often require more variables to be measured in order for the method to give unbiased estimates using the prediction error method. We will show that using maximum likelihood (ML) estimation, we are able to obtain better estimates by just observing as many variables as used in the indirect method.

Deriving the ML problem is challenging since networks of dynamical systems generally lead to a singular probability density function (pdf) and some variables are not observable. Before addressing this problem, we will first recapitulate the results of [4] for ML estimation when not all variables are observable for a Gaussian distribution in Section II. Then, in Section III, we will show how a singular Gaussian distribution can be transformed into a nonsingular one using linear transformations, and in Section IV, we apply this result to dynamic networks with known interconnections. We present a numerical example in Section V, and we conclude the paper in Section VI.

*This work was supported by the Novo Nordisk Foundation under grant number NNF20OC0061894. It was also supported by ELLIIT.

[1]Department of Applied Mathematics and Computer Science, Technical University of Denmark. Email: {jogal,mskan}@dtu.dk

[2]Department of Electrical Engineering, Linköping University. Email: anders.g.hansson@liu.se

Copyright ©2024 EUCA

II. Nonsingular Gaussian Distribution and Maximum Likelihood Estimation

We start by considering a model of the form

$$\Phi(\theta)x(\theta) + \Gamma(\theta) = e, \tag{1}$$

where $e \in \mathbb{R}^m$ is a realization of a zero-mean Gaussian random variable with covariance λI with $\lambda > 0$, $\theta \in \mathbb{R}^q$ is a vector of unknown model parameters, and $\Gamma(\theta) \in \mathbb{R}^m$.

The vector

$$x(\theta) = \begin{bmatrix} x_o(\theta) \\ x_m(\theta) \end{bmatrix} \in \mathbb{R}^n$$

is composed of observed data $x_o \in \mathbb{R}^{n_o}$ and missing data $x_m(\theta) \in \mathbb{R}^{n_m}$, and we partition $\Phi(\theta) \in \mathbb{R}^{m\times n}$ conformably as $\Phi(\theta) = \begin{bmatrix}\Phi_o(\theta) & \Phi_m(\theta)\end{bmatrix}$. To simplify notation, we will henceforth omit the dependence on θ.

If $m = n$ and Φ is nonsingular, then the pdf for x may be expressed as

$$\begin{aligned} p(x;\lambda,\theta) &= \mathcal{N}(x;\mu,\Psi) \\ &= \frac{1}{\sqrt{(2\pi)^m \det \Psi}} \exp\left\{-\frac{1}{2\lambda}\|\Phi x + \Gamma\|_2^2\right\} \end{aligned}$$

where $\mu = -\Phi^{-1}\Gamma = (\mu_o, \mu_m)$ is the mean and $\Psi = \lambda(\Phi^T\Phi)^{-1}$ is the covariance matrix. Furthermore, the marginal pdf of x_o may be expressed as

$$p(x_o;\lambda,\theta) = \mathcal{N}(x_o;\mu_o,\lambda(\Phi_o^T\Pi\Phi_o)^{-1})$$

where $\Pi = I - \Phi_m(\Phi_m^T\Phi_m)^{-1}\Phi_m^T$ is a projection matrix. As is shown in [4], up to an additive constant, the negative log-likelihood function can be expressed as

$$\begin{aligned} L(\lambda,\theta) =& \frac{1}{2\lambda}(\Phi_o x_o + \Gamma)^T\Pi(\Phi_o x_o + \Gamma) + \frac{m}{2}\ln\lambda \\ &- \frac{1}{2}\ln\det(\Phi^T\Phi) + \frac{1}{2}\ln\det Z \end{aligned} \tag{2}$$

where $Z = \Phi_m^T\Phi_m$. An ML estimate of (λ,θ) can then be obtained as

$$(\hat{\lambda},\hat{\theta}) \in \underset{(\lambda,\theta)}{\arg\min}\ L(\lambda,\theta). \tag{3}$$

We note that the partial derivatives of L with respect to λ and θ can be found in [4].

III. Singular Gaussian Distribution

Dynamical networks generally lead to a singular pdf, making the derivation of the ML problem challenging. These interconnected systems, as we will show in Section IV, can be expressed as an instance of a more general model in the form of

$$A(\theta)x(\theta) + b(\theta) = \begin{bmatrix} e \\ 0 \end{bmatrix}, \tag{4}$$

where $e \in \mathbb{R}^{m_1}$ is a realization of a zero-mean Gaussian random variable with covariance λI with $\lambda > 0$. As in the previous section, we partition $x(\theta)$ as

$$x(\theta) = \begin{bmatrix} x_o(\theta) \\ x_m(\theta) \end{bmatrix} \in \mathbb{R}^{n_o+n_m},$$

corresponding to observed and missing data, respectively, and we then partition $A(\theta) \in \mathbb{R}^{(m_1+m_2)\times(n_o+n_m)}$ and $b(\theta) \in \mathbb{R}^{m_1+m_2}$ conformably as

$$A(\theta) = \begin{bmatrix} A_1(\theta) \\ A_2 \end{bmatrix} = \begin{bmatrix} A_{1o}(\theta) & A_{1m}(\theta) \\ A_{2o} & A_{2m} \end{bmatrix},\ b(\theta) = \begin{bmatrix} b_1(\theta) \\ b_2 \end{bmatrix}.$$

Notice that A_2 and b_2 do not depend on the parameter vector θ. We will see in Section IV that this assumption is satisfied for dynamic networks with known interconnections. Once again, we will omit the dependence on θ to simplify our notation.

In contrast to the situation in the previous section, the pdf for x is now singular, and hence the maximum likelihood estimation is not readily applicable. We will address this issue by transforming the model under the assumption that $A(\theta)$ has full row-rank.

Given a singular value decomposition (SVD) of A_{2m}, i.e.,

$$A_{2m} = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1 & V_2 \end{bmatrix}^T,$$

we can rewrite (4) as

$$\begin{bmatrix} A_{1o} & \bar{A}_{1m1} & \bar{A}_{1m2} \\ \bar{A}_{2o1} & \Sigma_1 & 0 \\ \bar{A}_{2o2} & 0 & 0 \end{bmatrix} \begin{bmatrix} x_o \\ \bar{x}_{m1} \\ \bar{x}_{m2} \end{bmatrix} + \begin{bmatrix} b_1 \\ \bar{b}_{21} \\ \bar{b}_{22} \end{bmatrix} = \begin{bmatrix} e \\ 0 \\ 0 \end{bmatrix}$$

where

$$\bar{A}_{2o} = \begin{bmatrix} \bar{A}_{2o1} \\ \bar{A}_{2o2} \end{bmatrix} = \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix} A_{2o}, \tag{5}$$

$$\bar{A}_{1m} = \begin{bmatrix} \bar{A}_{1m1} & \bar{A}_{1m2} \end{bmatrix} = A_{1m}\begin{bmatrix} V_1 & V_2 \end{bmatrix} \tag{6}$$

and

$$\bar{x}_m = \begin{bmatrix} \bar{x}_{m1} \\ \bar{x}_{m2} \end{bmatrix} = \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} x_m, \quad \bar{b}_2 = \begin{bmatrix} \bar{b}_{21} \\ \bar{b}_{22} \end{bmatrix} = \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix} b_2.$$

Using Σ_1 as a pivot, we can rewrite the system as

$$\begin{bmatrix} \bar{A}_{1o} & 0 & \bar{A}_{1m2} \\ \bar{A}_{2o1} & \Sigma_1 & 0 \\ \bar{A}_{2o2} & 0 & 0 \end{bmatrix} \begin{bmatrix} x_o \\ \bar{x}_{m1} \\ \bar{x}_{m2} \end{bmatrix} + \begin{bmatrix} \bar{b}_1 \\ \bar{b}_{21} \\ \bar{b}_{22} \end{bmatrix} = \begin{bmatrix} e \\ 0 \\ 0 \end{bmatrix}, \tag{7}$$

where $\bar{A}_{1o} = A_{1o} - \bar{A}_{1m1}\Sigma_1^{-1}\bar{A}_{2o1}$ and $\bar{b}_1 = b_1 - \bar{A}_{1m1}\Sigma_1^{-1}\bar{b}_{21}$. The assumption that A has full row-rank implies that $\bar{A}_{2o2}$ has full row-rank, and hence there exists an orthogonal matrix $W = \begin{bmatrix} W_1 & W_2 \end{bmatrix}$ such that $\bar{A}_{2o2}W = \begin{bmatrix} \tilde{A}_{2o21} & 0 \end{bmatrix}$ with $\tilde{A}_{2o21} = \bar{A}_{2o2}W_1$ nonsingular. The matrix W can be obtained by means of an SVD or an LQ decomposition of $\bar{A}_{2o2}$. We then define

$$\begin{bmatrix} \tilde{A}_{1o} \\ \tilde{A}_{2o1} \end{bmatrix} = \begin{bmatrix} \tilde{A}_{1o1} & \tilde{A}_{1o2} \\ \tilde{A}_{2o11} & \tilde{A}_{2o12} \end{bmatrix} = \begin{bmatrix} \bar{A}_{1o} \\ \bar{A}_{2o1} \end{bmatrix} W, \tag{8}$$

$$\bar{x}_o = \begin{bmatrix} \bar{x}_{o1} \\ \bar{x}_{o2} \end{bmatrix} = W^T x_o,$$

which allows us to rewrite equation (7) as

$$\begin{bmatrix} \tilde{A}_{1o1} & \tilde{A}_{1o2} & 0 & \bar{A}_{1m2} \\ \tilde{A}_{2o11} & \tilde{A}_{2o12} & \Sigma_1 & 0 \\ \tilde{A}_{2o21} & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \bar{x}_{o1} \\ \bar{x}_{o2} \\ \bar{x}_{m1} \\ \bar{x}_{m2} \end{bmatrix} + \begin{bmatrix} \bar{b}_1 \\ \bar{b}_{21} \\ \bar{b}_{22} \end{bmatrix} = \begin{bmatrix} e \\ 0 \\ 0 \end{bmatrix}.$$

Using $\tilde{A}_{2o21}$ as a pivot, we obtain the system

$$\begin{bmatrix} 0 & \tilde{A}_{1o2} & 0 & \bar{A}_{1m2} \\ \tilde{A}_{2o11} & \tilde{A}_{2o12} & \Sigma_1 & 0 \\ \tilde{A}_{2o21} & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \bar{x}_{o1} \\ \bar{x}_{o2} \\ \bar{x}_{m1} \\ \bar{x}_{m2} \end{bmatrix} + \begin{bmatrix} \tilde{b}_1 \\ \bar{b}_{21} \\ \bar{b}_{22} \end{bmatrix} = \begin{bmatrix} e \\ 0 \\ 0 \end{bmatrix},$$

where $\tilde{b}_1 = \bar{b}_1 - \tilde{A}_{1o1}\tilde{A}_{2o21}^{-1}\bar{b}_{22}$. Finally, we rewrite this system as two systems,

$$\begin{bmatrix} \tilde{A}_{1o2} & \bar{A}_{1m2} \end{bmatrix} \begin{bmatrix} \bar{x}_{o2} \\ \bar{x}_{m2} \end{bmatrix} + \tilde{b}_1 = e, \tag{9}$$

$$\begin{bmatrix} \tilde{A}_{2o11} & \Sigma_1 \\ \tilde{A}_{2o21} & 0 \end{bmatrix} \begin{bmatrix} \bar{x}_{o1} \\ \bar{x}_{m1} \end{bmatrix} + \begin{bmatrix} \bar{b}_{21} + \tilde{A}_{2o12}\bar{x}_{o2} \\ \bar{b}_{22} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{10}$$

From this we realize that only $\bar{x}_{02}$ and $\bar{x}_{m2}$ are directly related to e. The matrix $\begin{bmatrix} \tilde{A}_{1o2} & \bar{A}_{1m2} \end{bmatrix}$ has full row-rank, so if it is a square matrix, then it is also invertible, and otherwise we can make use of column compression to further reduce the number of variables.

The equation (9) is of the form (1) with

$$\Phi = \begin{bmatrix} \tilde{A}_{1o2} & \bar{A}_{1m2} \end{bmatrix}, \quad \Gamma = \tilde{b}_1,$$

and hence we can processed as for the nonsingular Gaussian distribution in Section II. We note that Φ is a linear transformation of A_1. To see this, first note that $\tilde{A}_{1o2} = \bar{A}_{1o}W_2$, which follows from (8), and recall that

$$\bar{A}_{1o} = A_{1o} - \bar{A}_{1m1}\Sigma_1^{-1}\bar{A}_{2o1} = A_{1o} - A_{1m}V_1\Sigma_1^{-1}U_1^T A_{2o}.$$

Using (6), we arrive at

$$\begin{aligned} \Phi &= \begin{bmatrix} \Phi_o & \Phi_m \end{bmatrix} \\ &= \begin{bmatrix} A_{1o} & A_{1m} \end{bmatrix} \begin{bmatrix} W_2 & 0 \\ -V_1\Sigma_1^{-1}U_1^T A_{2o}W_2 & V_2 \end{bmatrix}. \end{aligned}$$

Similarly, tracing the transformations applied to b, we find that

$$\Gamma = b_1 - \begin{bmatrix} A_{1o} & A_{1m} \end{bmatrix} \begin{bmatrix} Hb_2 \\ V_1\Sigma_1^{-1}U_1^T(I - A_{2o}H)b_2 \end{bmatrix},$$

where $H = W_1(U_2^T A_{2o}W_1)^{-1}U_2^T$. Thus, the reduced nonsingular problem is obtained as a linear transformation of (A_1, b_1) where the transformation matrices are functions of (A_2, b_2), which are known and do not depend on θ.

IV. Dynamic Networks

We will now consider a dynamic network model that can be cast in the form of (4) where A_2 and b_2 do not dependent on the parameters vector θ. Specifically, will consider a network of M systems, where the ith system is described by an ARMAX model of the form

$$y_k^i = -\sum_{j=1}^{n_a^i} a_j^i y_{k-j}^i + \sum_{j=0}^{n_b^i} b_j^i u_{k-j}^i + e_k^i + \sum_{j=1}^{n_c^i} c_j^i e_{k-j}^i \tag{11}$$

for $k = 1, 2, \ldots, N$. We will assume that y_k^i, u_k^i, and e_k^i are zero for all $k \leq 0$. To simplify notation, we define $u^i = (u_1^i, \ldots, u_N^i)$, $y^i = (y_1^i, \ldots, y_N^i)$, and $e^i = (e_1^i, \ldots, e_N^i)$, corresponding to the ith system's input, output, and disturbance signals. We also define $a^i = (a_1^i, \ldots, a_{n_a^i}^i)$, $b^i = (b_0^i, \ldots, b_{n_b^i}^i)$, and $c^i = (c_1^i, \ldots, c_{n_c^i}^i)$ as well as lower-triangular Toeplitz matrices $T_{a^i} \in \mathbb{R}^{N\times N}$, $T_{b^i} \in \mathbb{R}^{N\times N}$ and $T_{c^i} \in \mathbb{R}^{N\times N}$ whose first columns are

$$\begin{bmatrix} 1 \\ a^i \\ 0 \end{bmatrix}, \quad \begin{bmatrix} b^i \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ c^i \\ 0 \end{bmatrix},$$

respectively, c.f., [10]. This allows us to express the ARMAX model (11) as

$$T_{a^i}y^i = T_{b^i}u^i + T_{c^i}e^i.$$

The interconnections are defined in terms of sparse matrices $\Lambda \in \mathbb{R}^{M\times M}$ and $\Omega \in \mathbb{R}^{M\times Q}$ with ± 1 as nonzero entries such that

$$\begin{bmatrix} u_k^1 \\ u_k^2 \\ \vdots \\ u_k^M \end{bmatrix} = \Lambda \begin{bmatrix} y_k^1 \\ y_k^2 \\ \vdots \\ y_k^M \end{bmatrix} + \Omega \begin{bmatrix} r_k^1 \\ r_k^2 \\ \vdots \\ r_k^Q \end{bmatrix}, \quad 1 \leq k \leq N,$$

where $r^i = (r_1^i, r_2^i, \ldots, r_N^i)$, $i = 1, \ldots, Q$, are exogenous signals. We will assume that $r_k^i = 0$ for all $k \leq 0$.

We will now show that the dynamic network model can be written as in (4). To this end, we define vectors $y = (y^1, \ldots, y^M)$, $u = (u^1, u^2, \ldots, u^M)$, $e = (e^1, e^2, \ldots, e^M)$, and $r = (r^1, r^2, \ldots, r^Q)$, and matrices $T_{y^i} = T_{c^i}^{-1}T_{a^i}$ and $T_{u^i} = T_{c^i}^{-1}T_{b^i}$ for $i = 1, \ldots, M$. We also define two block-diagonal matrices,

$$\begin{aligned} T_y &= \mathrm{blkdiag}(T_{y^1}, \ldots, T_{y^M}) \\ T_u &= \mathrm{blkdiag}(T_{u^1}, \ldots, T_{u^M}). \end{aligned}$$

This allows us to express the dynamic network model as an instance of equation (4) with

$$\begin{aligned} A &= \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} T_y & -T_u \\ -\Lambda \otimes I & I \end{bmatrix} (P \otimes I) \\ x &= (P \otimes I)^T \begin{bmatrix} y \\ u \end{bmatrix} \\ b &= \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = -\begin{bmatrix} 0 \\ \Omega \otimes I \end{bmatrix} r, \end{aligned}$$

where $P \in \mathbb{R}^{2M\times 2M}$ is a permutation matrix that is defined such that the observed parts of y and u correspond to the leading entries of x. The parameter vector θ represents the unknown model parameters $(a^1, \ldots, a^M, b^1, \ldots, b^M, c^1, \ldots, c^M)$.

The matrix A satisfies $m_1 = m_2 = MN$ and $n = 2MN$, and hence it is square. Using the fact that the matrices T_{a^i} and T_{c^i} are nonsingular for all i, we see that T_y is nonsingular and then, using the Schur complement, we have that A is nonsingular if $T_y - T_u(\Lambda \otimes I)$ is full rank. As a result, the transformation described in Section III yields a square and nonsingular matrix $\Phi = \begin{bmatrix} \bar{A}_{1o2} & \bar{A}_{1m2} \end{bmatrix}$ as in (9). Furthermore, A_2 only depends on the network topology and is independent of any model parameters. Thus, the linear transformations that are needed to obtain Φ and Γ are independent of the model parameters.

V. Numerical Experiment

We will now illustrate some properties of the proposed method using a numerical example based on the network of $M = 3$ systems shown in Fig. 1. The interconnections are described by the matrices

$$\Lambda = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \quad \Omega = I.$$

The example matches the example in Fig. 2 in [5] except for the way the disturbances enter the system.

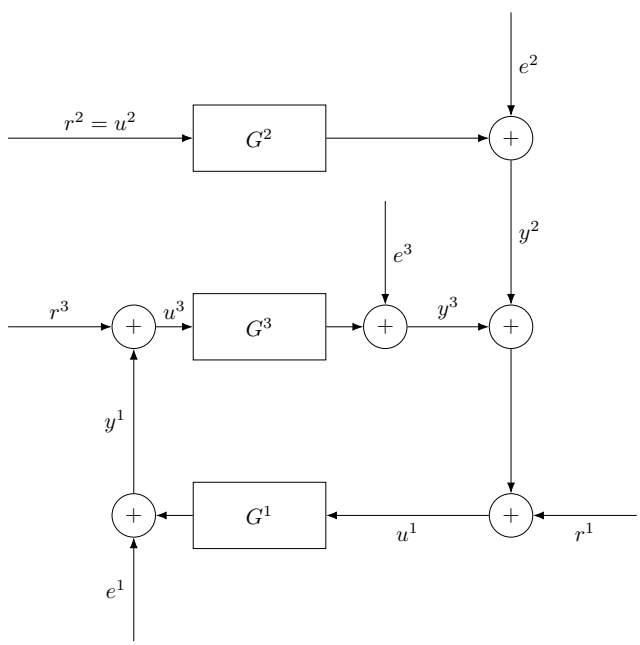

Fig. 1. Block diagram for a dynamic network.

Each subsystem is a second-order system, and we only have one observable variable, namely $x_o = (u^3)$. As we will show soon, the indirect approach does not handle ARMAX models, so we will limit our attention to the case where all subsystems are ARX models in order to facilitate a fair comparison. We generated $N = 500$ measurements where the error (e^1, e^2, e^3) is a realization of a zero-mean Gaussian random variable with covariance $0.1 \cdot I$, and the input (r^1, r^2, r^3) is a vector of independent samples from the Rademacher distribution.

The true systems that we use to generate both the observable and the missing states are zero-order hold discretizations of continuous systems whose Laplace domain transfer functions are given by

$$G^1 = \frac{0.5}{(s+2)^2}, \quad G^2 = \frac{1}{s^2 + 2s + 0.75}, \quad G^3 = \frac{0.5}{(s+0.5)^2}.$$

The three systems are stable, and hence the poles of the discrete models will be inside the unit circle.

A. Direct Approach

We will use gradient descent combined with a backtracking line search to find local minima of the negative log-likelihood function, which is not a convex function. Given the observed data x_o, we generate 100 initial parameter vectors θ such that the closed-loop system is stable. The parameters for each system are drawn from a uniform distribution on $[-1, 1]$.

In order to reduce the complexity of the optimization problem (3) we rewrote the cost function to eliminate the dependence on λ. Minimizing (2) with respect to λ leads to

$$\lambda^\star = \frac{1}{m}(\Phi_o x_o + \Gamma)^T \Pi (\Phi_o x_o + \Gamma),$$

and then the cost function can be rewritten as

$$\begin{aligned} L(\theta) = &\frac{m}{2} \ln\left(\frac{1}{m}(\Phi_o x_o + \Gamma)^T \Pi (\Phi_o x_o + \Gamma)\right) \\ &- \frac{1}{2} \ln\det(\Phi^T \Phi) + \frac{1}{2} \ln\det Z + \frac{m}{2}. \end{aligned}$$

Our experiments showed that performing this variable reduction leads to better convergence properties than solving the original problem.

Our implementation is based on the Python library JAX [1], which uses automatic differentiation to compute the partial derivatives of the cost function.

B. Indirect Approach

The indirect approach first estimates the overall transfer function between the inputs and the observable variables, and then it uses this information to estimate a transfer function for each subsystem separately.

To analyze the example using transfer functions, we first write the ARMAX model in (11) as

$$y_k^i = G^i(q) u_k + H^i(q) e_k^i$$

where q denotes the forward shift operator such that $q u_k = u_{k+1}$, and where

$$G^i(q) = \frac{B^i(q)}{A^i(q)}, \qquad H^i(q) = \frac{C^i(q)}{A^i(q)}$$

and

$$\begin{aligned} A^i(q) &= q^{n_a^i} + a_1^i q^{n_a^i - 1} + \cdots + a_{n_a^i}^i \\ B^i(q) &= b_0^i q^{n_b^i} + b_1^i q^{n_b^i - 1} + \cdots + b_{n_b^i}^i \\ C^i(q) &= c_0^i q^{n_c^i} + c_1^i q^{n_c^i - 1} + \cdots + c_{n_c^i}^i. \end{aligned}$$

It is straight forward to show using algebraic manipulations that

$$\begin{aligned} \Delta(q) u_k^3 = &G^1(q) r_k^1 + G^1(q) G^2(q) r_k^2 + r_k^3 + H^1(q) e_k^1 \\ &+ G^1(q) H^2(q) e_k^2 + G^1(q) H^3(q) e_k^3 \end{aligned}$$

where $\Delta(q) = 1 - G^1(q) G^3(q)$. Equivalently, we can write

$$\begin{aligned} A^2(A^1 A^3 - B^1 B^3) u_k^3 = &B^1 A^2 A^3 r_k^1 + B^1 B^2 A^3 r_k^2 \\ &+ A^1 A^2 A^3 r_k^3 + C^1 A^2 A^3 e_k^1 \\ &+ B^1 C^2 A^3 e_k^3 + B^1 A^2 C^3 e_k^3 \end{aligned}$$

where we have omitted the dependence on q to simplify notation. From this we realize that it is possible to use an ARMAX model of the form

$$\bar{A}u_k^3 = \bar{B}^1 r_k^1 + \bar{B}^2 r_k^2 + \bar{B}^3 r_k^3 + \bar{C}\bar{e}_k \tag{12}$$

where we only have measurements of u_k^3 and (r_k^1, r_k^2, r_k^3) in order to estimate the model using any traditional identification technique. In our numerical experiments, we used the System Identification Toolbox [7] from MATLAB. It then follows that

$$\begin{aligned}\frac{B^1}{A^1} &= \frac{\bar{B}^1}{\bar{B}^3}\\ \frac{B^2}{A^2} &= \frac{\bar{B}^2}{\bar{B}^1}\\ \frac{B^3}{A^3} &= \frac{\bar{B}^3 - \bar{A}}{\bar{B}^1}\end{aligned}$$

from which A^i and B^i can be found. Notice that it is not possible to recover C^i. Thus, we will from now on assume that we have ARX models, i.e., $C^i = 1$.

If we assume that all the internal systems are order two, the polynomials $A^i(q)$ and $B^i(q)$ will have degree two, then the number of parameters are 5 for each ARX model, i.e., a total of 15 parameters. However, we have that $\bar{A}$ and $\bar{B}^i$ are of degree 6 whereas $\bar{C}$ is of degree of degree 4. We will then need 26 parameters when estimating the reformulated model described in (12). To prevent an increase in the number of parameters, one can either add constraints to the problem or perform a reparametrization of the cost function. Opting for this approach comes with the trade-off of increasing the computational complexity.

Here we have normalized such that $\bar{A}$ and $\bar{C}$ are monic polynomials.

For this example the number of parameters have increased using the indirect approach. This may impair the quality of the estimate, since it is known that the more parameters that are estimated, the higher is the variance of the estimate. However, the indirect approach benefits from that a standard prediction error method may be used to obtain unbiased estimates, and these estimates may be utilized to initialize the algorithm. Moreover, the analysis shows what signals need to be measured and which do not need to be measured. Here we see that it is enough to measure u^3 in addition to r. Similar analysis can be carried out to show that either of u^1, u^3, y^1 or y^3 are sufficient to measure, c.f. the example related to Fig. 2 in [5]. This reference also discusses what signals need to be measured for a general dynamic network in order to estimate all parameters when using indirect approaches.

C. Results

When observing only $x_o = u^3$, we found that all 100 different initial points generated for the direct approach converged to local minima corresponding to stable systems. For most of these estimates, the missing states x_m were poorly estimated, while the observed states x_o were well estimated. Only the best local solutions (in terms of the cost function value) led to good estimates of both the observable and missing states.

Using the indirect approach, $\bar{B}^1$ presented unstable zeros, which would result in unstable models for the subsystems 2 and 3. To mitigate this, the unstable zeros where removed from $\bar{B}^1$ and the static gain was adjusted in order to match its original value before the exclusion of the unstable zeros.

Fig. 2 presents the comparison between the true observed variable (u_3) and the one generated when we simulated the system using the estimated parameter vector $\hat{\theta}$ from the proposed direct approach with the lowest cost function value. Fig. 3 features a similar comparison for the missing variables (y_1, y_2, y_3, u_1).

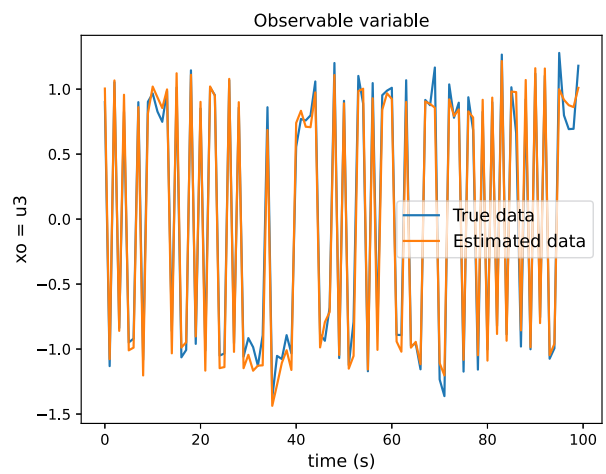

Fig. 2. Estimated and true observable variables.

As explained in Section V-B, the indirect approach yields degree six polynomials for the subsystems, requiring a model reduction technique to allow a direct comparison between the parameters. Such a reduction may introduce additional errors in the estimate, so we instead compare the methods using a frequency domain analysis. Fig. 4 presents such a comparison based on Bode plots for each subsystem, considering the ten best local solutions found using our approach, the solution found with the indirect approach, and the true system.

From the Bode plots in Fig. 4, we can see that both approaches were able to obtain an estimate that somehow captures the dynamics of subsystems 1 and 3. However, when looking to the subsystem 2 the plots show a high variance when using the direct approach with different initial guesses. This highlights the necessity of implementing a good initialization strategy.

Next, we compare the model fit for the parameter estimates obtained using the direct and the indirect approach. Given a parameter estimate $\hat{\theta}$ and new realizations of the signals (e^1, e^2, e^3) and the inputs (r^1, r^2, r^3), the fit is defined as

$$\text{fit}(\hat{x}) = 1 - \frac{\|\hat{x} - x^{\text{ref}}\|_2}{\|\hat{x} - \mathbf{1}\bar{x}^{\text{ref}}\|_2},$$

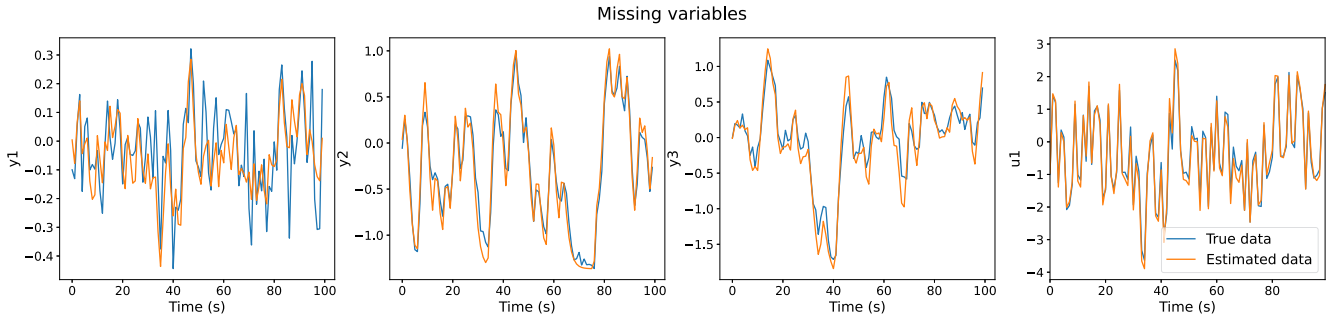

Fig. 3. Comparison between the true missing variables and the estimate obtained using the proposed direct approach.

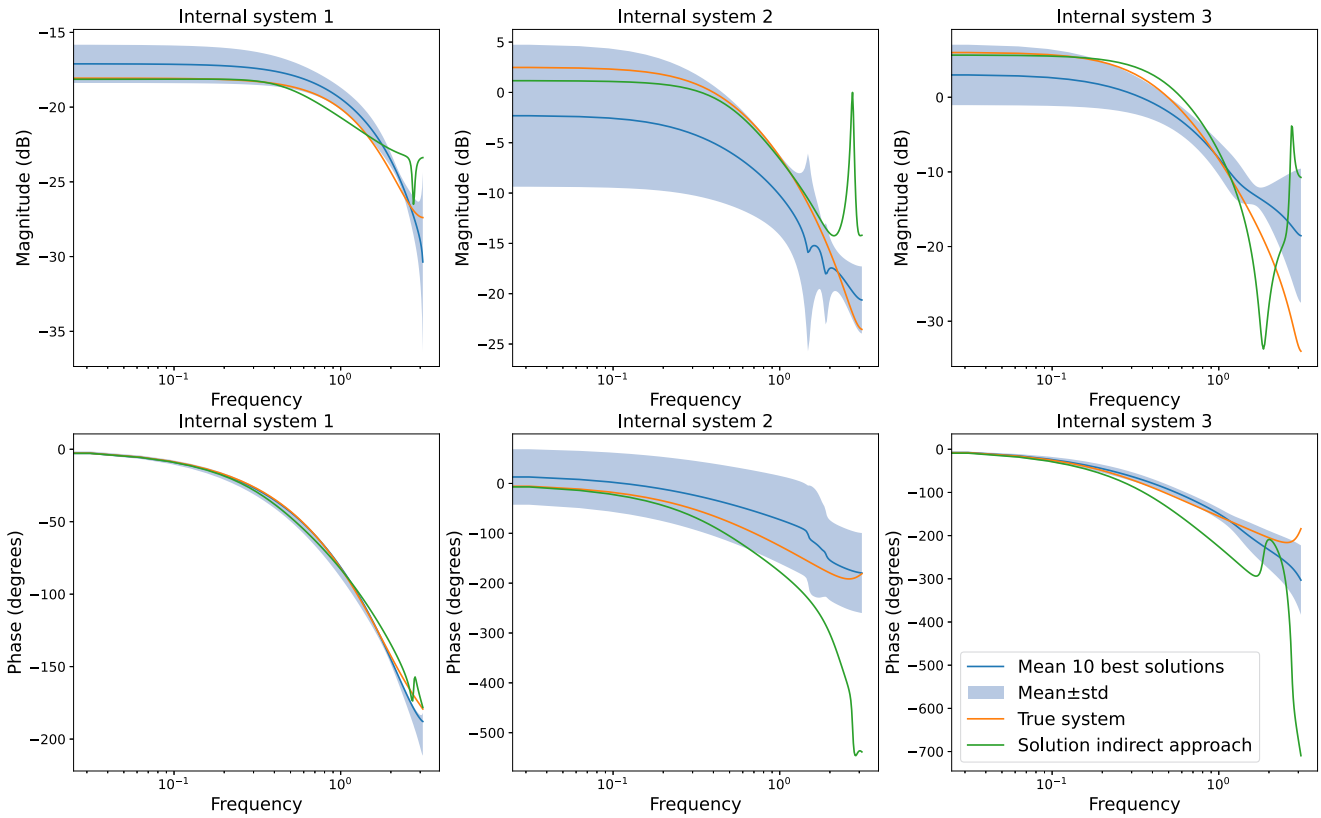

Fig. 4. Comparison of indirect and direct approach: magnitude and phase Bode plots for each internal system.

where $\hat{x}$ is the signal of interest from the estimated model defined by $\hat{\theta}$, x^{ref} is the same signal from the true model, $\bar{x}^{\text{ref}}$ denotes the mean of the true signal x^{ref}, and $\mathbf{1}$ is a vector of all ones. With the estimate given by the indirect approach, the fit values for the observable and missing variables were 0.8934 and 0.6312, respectively, and with the best local solution from the direct approach, the corresponding fit values were 0.9008 and 0.8344, respectively. Analyzing these results we can check a clear advantage for the proposed direct approach, but when considering the other local solutions, we found that the fit associated with the missing variables x_m rapidly decrease when going from the second best solution to the worst one, which also highlights the impact of and sensitivity to the choice of initial point.

As a simple initialization strategy, we tried using the reduced model obtained via the indirect approach as the initiation for the proposed direct approach. When comparing with the original fit value obtained with the indirect approach we had an improvement in the fit value for both the observed and the missing variables, which were 0.8973 and 0.7253, respectively. We repeated this experiment ten times with different realizations of both the error signals (e^1, e^2, e^3) and the inputs (r^1, r^2, r^3). Table I shows the fit of x_o and x_m for the high order models obtained by the indirect method and for the model obtained with the direct approach initialized based on a reduced model. The table also show the fit improvement, which is the difference between the direct approach and the indirect approach. We see that the two

TABLE I

FIT VALUES FOR THE SOLUTIONS FROM THE INDIRECT AND THE DIRECT APPROACH, AND THEIR RESPECTIVE INCREMENTS.

Exp	$100 \times$ Fit($\hat{x}_o$)			$100 \times$ Fit($\hat{x}_m$)		
	IND	DIR	IMPROV	IND	DIR	IMPROV
1	88.76	89.48	0.71	61.07	72.05	10.97
2	88.86	89.78	0.92	38.14	68.38	30.24
3	88.82	89.65	0.83	65.09	71.54	6.45
4	88.87	89.41	0.54	66.31	76.68	10.37
5	88.77	89.49	0.72	60.41	73.54	13.12
6	89.24	89.75	0.50	63.08	76.60	13.52
7	88.51	89.36	0.84	38.46	69.35	30.88
8	88.94	89.43	0.49	61.78	67.24	5.46
9	89.08	89.11	0.02	56.61	60.97	4.35
10	88.87	86.42	-2.44	57.71	62.23	4.51

methods obtain a similar fit for $\hat{x}_o$, but we see a significant improvement in the fit for x_m. The mean improvement for x_o and x_m are 0.32% and 12.99%, respectively.

As indicated by the results, we observed that when only observing a few signals, our method often converges to a nearby local minimum at which the observable data x_o is approximated well. However, if the starting point is not in a neighborhood of a global optimum, then we cannot ensure that estimate of x_m will be accurate, but the estimate of x_o is typically still good. We also observed that initialization based on another approach such as the indirect method can substantially improve the solution for the proposed direct approach.

In contrast to the indirect approach where the knowledge of additional signals does not necessarily translate to better estimates, the proposed method can benefit from observing more signals, thereby improving the estimation accuracy and reducing the variance and sensitivity of the solutions for the different initial points. Fig. 5 presents the Bode plots for each of the internal subsystems. The figure shows the true system and the estimate corresponding to the best local minimum along the mean and standard deviation for the 50 best local minima when observing both u^1 and u^3.

From the Bode plots in Fig. 5, we can see that observing u^1 together with u^3 leads to a considerable increase in the estimation accuracy while reducing the sensitivity to initialization.

Comparing the results in Figures 4 and 5, we see that with fewer observed signals, we are more likely to converge to a poor local minimum, and when we observe more variables, local optimization tends to end up in good local minima. This suggests that global optimization should be considered when only limited data are available.

VI. CONCLUSION

We propose a new direct approach based on the ML estimation to perform system identification of dynamic networks with missing data. Dynamic networks generally lead to a singular pdf, and we show that when we know how the systems are interconnected, we are able to derive a nonsingular pdf by rewriting the problem using linear transformations, allowing us to perform the ML estimation.

The obtained results suggest that our approach is suitable for estimating the parameters of dynamic networks when combined with global optimization or a suitable initialization strategy. Using random initialization, we found that choosing the best local solution among many led to a precise prediction of the observable state and a good estimation of the missing ones.

As discussed in Section V-B, we are not able to recover C^i using the indirect approach. In contrast, our direct approach allows us to include this term, making it more suitable for ARMAX models. Another drawback of indirect approaches is that they can lead to unstable models for the subsystems even when the internal transfer functions are known to be stable. Moreover, the number of parameters that are required to describe the closed-loop system can be significantly larger. This either increases the variance of the estimator or requires the addition of constraints to the problem or a reparametrization of the cost function to avoid increasing the number of parameters.

The proposed direct approach can benefit from additional observed variables, while for the indirect approach, observing some of the signals does not add useful information. In the numerical example presented in Section V, knowing the sets (u^3, u^1), (u^3, y^3) or (u^3, y^1) does not add more information when using the indirect approach than knowing just u^3. In contrast, using our method, each new observable variable would contribute to a better estimation. With our approach, the measurement of (u^3, u^1) led to a considerable improvement in the fit values for both observed and missing variables, while decreasing the variance and sensitivity of the solutions for the different initial points.

When compared to prediction error based direct approaches, the proposed approach requires fewer observable variables to be able to estimate the system parameters, allowing the estimation of the network presented in Section V while observing only u^3, which would not be possible with prediction error based approaches [3].

The major limitation of our approach results from the nonconvexity of the cost function, requiring an initial set of parameters in the neighborhood of the global optimum to ensure the convergence towards it. Thus, a suitable initialization strategy is needed or, alternatively, the use of global optimization is called for. To investigate the issue, we explored two different approaches: the first approach was to consider 100 randomly generated initializations and use the best local solution with respect to the cost function value, and the second approach was to compute an initial guess from the estimate obtained using the indirect method. The results show that both approaches were able to converge to solutions for which the observable and the missing variables could be well estimated.

In future work, we plan to apply the proposed approach to bigger dynamical network architectures, explore different initialization strategies and other optimization methods, e.g., trust-region methods or cheaper stochastic methods such as

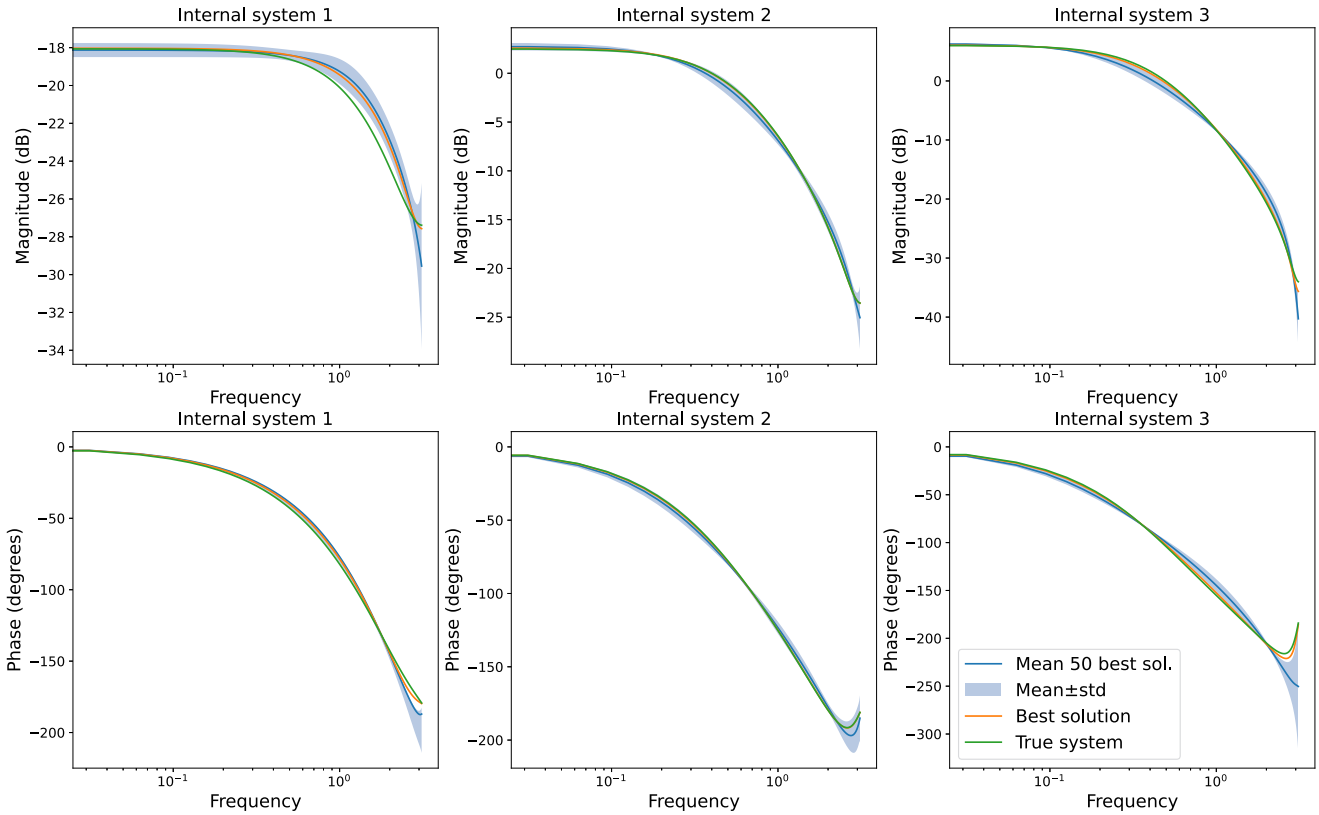

Fig. 5. Magnitude and phase Bode plots for each internal system, considering the 50 best solutions, when observing both u^3 and u_1 using the proposed direct approach.

AdaHessian [11] or AdaSub [2] that incorporate second-order information. We also plan to derive sufficiency conditions that characterize which signals need to be observed to obtain consistent estimates of each internal system in the dynamical networks when using the proposed maximum likelihood approach.

References

[1] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018.

[2] João Victor Galvão da Mata and Martin S. Andersen. Adasub: Stochastic optimization using second-order information in low-dimensional subspaces, 2023.

[3] Paul M.J. Van den Hof. Personal communication.

[4] A. Hansson and R. Wallin. Maximum likelihood estimation of Gaussian models with missing data—Eight equivalent formulations. *Automatica*, 48:1955–1962, 2012.

[5] J. M. Hendrickx, M. Gevers, and A. S. Bazanella. Identifiability of dynamical networks with partial node measurements. *IEEE Transactions on Automatic Control*, 64(6):2240–2253, 2019.

[6] L. Ljung. *System Identification: Theory for the User*. Prentice Hall information and system sciences series. Prentice Hall PTR, 1999.

[7] Lennart Ljung. *System identification toolbox*. Math Works Natick, MA, USA, 1995.

[8] Paul M.J. Van den Hof, Arne Dankers, Peter S.C. Heuberger, and Xavier Bombois. Identification of dynamic models in complex networks with prediction error methods—basic methods for consistent module estimates. *Automatica*, 49(10):2994–3006, 2013.

[9] Paul M.J. Van Den Hof and Ruud J.P. Schrama. An indirect method for transfer function estimation from closed loop data. *Automatica*, 29(6):1523–1527, 1993.

[10] R. Wallin and A. Hansson. Maximum likelihood estimation of linear siso models subject to missing output data and missing input data. *International Journal of Control*, 87(11):2354–2364, 2014.

[11] Zhewei Yao, Amir Gholami, Sheng Shen, Kurt Keutzer, and Michael W. Mahoney. ADAHESSIAN: an adaptive second order optimizer for machine learning. *CoRR*, abs/2006.00719, 2020.