Delft University of Technology

Linear Time-Varying Parameter Estimation
Maximum A Posteriori Approach via Semidefinite Programming

Vakili, Sasan; Khosravi, Mohammad; Esfahani, Peyman Mohajerin; Mazo, Manuel

DOI
10.1109/LCSYS.2023.3347198

Publication date
2024

Document Version
Final published version

Published in
IEEE Control Systems Letters

Citation (APA)
Vakili, S., Khosravi, M., Esfahani, P. M., & Mazo, M. (2024). Linear Time-Varying Parameter Estimation: Maximum A Posteriori Approach via Semidefinite Programming. *IEEE Control Systems Letters*, *8*, 73-78. https://doi.org/10.1109/LCSYS.2023.3347198

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

Green Open Access added to TU Delft Institutional Repository

'You share, we take care!' - Taverne project

https://www.openaccess.nl/en/you-share-we-take-care

Otherwise as indicated in the copyright section: the publisher
is the copyright holder of this work and the author uses the
Dutch legislation to make this work public.

Linear Time-Varying Parameter Estimation: Maximum A Posteriori Approach via Semidefinite Programming

Sasan Vakili, Mohammad Khosravi, Peyman Mohajerin Esfahani,
and Manuel Mazo Jr., *Senior Member, IEEE*

***Abstract*—We study the problem of identifying a linear time-varying output map from measurements and linear time-varying system states, which are perturbed with Gaussian observation noise and process uncertainty, respectively. Employing a stochastic model as *prior* knowledge for the parameters of the unknown output map, we reconstruct their estimates from input/output pairs via a Bayesian approach to optimize the *posterior* probability density of the output map parameters. The resulting problem is a non-convex optimization, for which we propose a tractable linear matrix inequalities approximation to warm-start a first-order subsequent method. The efficacy of our algorithm is shown experimentally against classical Expectation Maximization and Dual Kalman Smoother approaches.**

***Index Terms*—Estimation, identification, semidefinite programming, linear matrix inequalities, optimization.**

I. INTRODUCTION

BAYESIAN approaches for estimating characteristics of dynamical systems have been a subject of studies for decades and have recently received extensive attention [1], [2]. In systems theory, the significance of the Bayesian approach is highlighted in state estimation of dynamical systems [3], [4], e.g., through the celebrated recursive Kalman *filter*. The Rauch-Tung-Striebel (RTS) Smoother counterparts [4], on the other hand, are (offline) iterative non-causal algorithms incorporating future measurements into the current state estimation.

An alternative to Bayesian estimation, which requires a prior distribution of the parameters of interest, is the minimax estimation approach, assuming instead the knowledge of ambiguity sets. The *least favourable* uncertainty model from this ambiguity set is then used for estimation [5], [6], [7], [8]. Here, we focus instead on designing a classical smoother for a different problem: system parameters estimation from input/output measurements via Bayesian estimation. This problem arises in, e.g., robot mapping in unknown environments, such as Autonomous Underwater Vehicles operating in the deep sea where global positioning is expensive due to low visibility and lack of radio communications.

Given unknown parameters with random states, applying a Bayesian estimation framework leads to severe non-convexities in the resulting optimization problem. As such, iterative schemes are typically employed to overcome these non-convexities. Assuming the parameters also follow a statistical formulation, two main types of smoother approaches, *Dual Kalman Smoother* (DKS) and *Expectation Maximization* (EM), are available in the literature [9].

Dual Kalman Smoothers (and filters) attempt to maximize the joint probability space of parameters and state (conditioned on input and output observations), iterating between estimating the system states using the last parameters' estimates followed by estimating the parameters from the currently estimated states. Although DKS is computationally efficient according to its recursive structure, its estimation performance can be significantly suboptimal due to the bilinearity between the parameters and states. Unlike DKS, *Expectation Maximization* learns the parameters of statistical models by maximizing the posterior distribution of the parameters from the observed data and their prior density function when incomplete data or hidden variables exist [10], [11], [12].

Considering the states of a dynamical system as hidden variables [13], [14], [15], EM estimates the parameters of a dynamical system in two steps by integrating all possible values of the states in which the model could have generated the observations. The distribution over hidden variables is maximized in the *E-Step* using the parameters estimates from the previous iteration. Subsequently, the *M-Step* maximizes a lower bound of the original cost by fixing the hidden variables distribution to the one optimized in the *E-Step*. A closed-form solution of the *M-Step* is provided in [14] for estimating the parameters of linear time-invariant dynamical systems and in [9, Ch. 6] for estimating the parameters of a Gaussian radial basis function (RBF) approximator. Both solutions consider the maximum likelihood case, where no prior exists for the parameters.

Finding a closed-form expression for the parameters update in the *M-Step* of a Maximum A Posteriori (MAP) smoothing problem when the parameters are time-varying and in the presence of *a priori* knowledge is non-trivial. This challenge leads

Manuscript received 15 September 2023; revised 24 November 2023; accepted 11 December 2023. Date of publication 25 December 2023; date of current version 22 January 2024. This work was supported by the European Union's Horizon 2020 Research and Innovation Programme through the Marie Skłodowska-Curie Grant under Agreement 956200. Recommended by Senior Editor V. Ugrinovskii. *(Corresponding author: Sasan Vakili.)*

The authors are with the Delft Center for Systems and Control, Delft University of Technology, 2628 CN Delft, The Netherlands (e-mail: S.Vakili@tudelft.nl; Mohammad.Khosravi@tudelft.nl; P.MohajerinEsfahani@tudelft.nl; M.Mazo@tudelft.nl).

Digital Object Identifier 10.1109/LCSYS.2023.3347198

2475-1456 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

to a slow convergence of the EM algorithm utilizing computationally demanding approaches to solve the optimization in *M-step*, e.g., first-order methods. The slow convergence of EM is shown experimentally in [16], and further analyses in [17], [18] demonstrate the slow convergence rate of the gradient variant of the EM algorithm for Gaussian Mixture Models.

Alternatives to the iterative schemes can be found in the parameter estimation problem of an elliptically contoured distribution [19, P. 107], employing recent Conic Geometric Optimization methods [20]. These methods, however, require reformulating the MAP problem via techniques, such as those proposed in [21, Sec. 3], which result in losing the output map's original structure. Retaining such a structure to leverage the available *a priori* knowledge is essential to our problem.

In this letter, we consider systems with known linear time-varying dynamics affected by process and measurement Gaussian noise but with unknown time-varying output maps. We propose a method to estimate the unknown parameters of the output map having as *a priori* information a linear stochastic system encoding the evolution of the parameters. We derive an optimization problem applying a fully Bayesian approach, maximizing the exact posterior distribution of the parameters when unfolded over the whole time horizon. A tractable conservative approximation to the resulting optimization problem is derived via semidefinite programming (SDP) using linear matrix inequalities (LMIs) techniques. The solution from this approximation then provides a warm-start for a first-order quasi-Newton algorithm that enjoys a locally superlinear convergence rate. This combination allows us to enjoy both the computational advantage of DKS and outperform the statistical performance of EM. We illustrate the efficacy and performance of our proposed method in comparison with DKS and EM through a Monte Carlo experiment with different signal-to-noise ratios (SNRs) in Section V.

Notation: Throughout this letter, $\mathbb{Z}_+$, $\mathbb{R}$ and $\mathbb{R}^{n\times m}$ denote the set of positive integers, real numbers, and n by m real matrices, respectively. We indicate $\mathrm{diag}(\mathrm{A}_1, \dots, \mathrm{A}_k)$ as a block diagonal matrix with diagonal entries of given matrices $\mathrm{A}_1, \dots, \mathrm{A}_k$. The symbol $\mathbb{I}$ denotes the identity matrix, and tr is the trace operator. Given $\mathrm{A} \in \mathbb{R}^{m\times n}$, a matrix with columns $\mathrm{a}_1, \dots, \mathrm{a}_n \in \mathbb{R}^m$, we define $\mathrm{vec}(\mathrm{A})$ as the vector $[\mathrm{a}_1^\top, \dots, \mathrm{a}_n^\top]^\top \in \mathbb{R}^{mn}$. For a positive symmetric matrix $\mathrm{A} \in \mathbb{R}^{n\times n}$, $\Lambda(\mathrm{A}) := (\lambda_i(\mathrm{A}))_{i=1}^n$ indicates the vector of eigenvalues of A in descending order, i.e., $\lambda_i(\mathrm{A})$ is the i^{th} largest eigenvalue of A. A multivariate normal (Gaussian) distribution with mean μ and covariance matrix Σ is denoted by $\mathcal{N}(\mu, \Sigma)$, and the symbol $\sim$ stands for *"distributed according to"*.

II. PROBLEM DEFINITION

Consider a discrete-time linear time-varying dynamical system described by the process model:

$$\mathrm{x}_{k+1} = \mathrm{A}_k\mathrm{x}_k + \mathrm{B}_k\mathrm{u}_k + \mathrm{w}_k, \qquad k \in \mathbb{Z}_+, \tag{1}$$

where k denotes the time index, $\mathrm{x}_k \in \mathbb{R}^{n_\mathrm{x}}$ is the vector of state variables, $\mathrm{A}_k \in \mathbb{R}^{n_\mathrm{x}\times n_\mathrm{x}}$ is the state transition matrix, $\mathrm{u}_k \in \mathbb{R}^{n_\mathrm{u}}$ is the vector of inputs, $\mathrm{B}_k \in \mathbb{R}^{n_\mathrm{x}\times n_\mathrm{u}}$ is the input matrix, and $\mathrm{w}_k \in \mathbb{R}^{n_\mathrm{x}}$ is an independent realization at time k of the process noise with Gaussian distribution $\mathcal{N}(0, \Sigma_{\mathrm{w}_k})$. The initial state of system (1), denoted by x_0, is also assumed to be drawn from a Gaussian distribution $\mathcal{N}(\mu_{\mathrm{x}_0}, \Sigma_{\mathrm{x}_0})$. For $k \in \mathbb{Z}_+$, the state of the system is observed at time instant k through a perturbed linear time-varying map:

$$\mathrm{y}_k = \mathrm{C}_k\mathrm{x}_k + \mathrm{v}_k, \qquad k \in \mathbb{Z}_+, \tag{2}$$

where $\mathrm{y}_k \in \mathbb{R}^{n_\mathrm{y}}$ denotes the output measurements, $\mathrm{C}_k \in \mathbb{R}^{n_\mathrm{y}\times n_\mathrm{x}}$ is an *unknown time-varying* observation matrix, and $\mathrm{v}_k \in \mathbb{R}^{n_\mathrm{y}}$ is the vector of measurement noise signals with Gaussian distribution $\mathcal{N}(0, \Sigma_{\mathrm{v}_k})$. Let θ_k be the vector of all parameters at each time index k:

$$\theta_k := \mathrm{vec}\big(\mathrm{C}_k^\top\big), \tag{3}$$

which implies that C_k and θ_k uniquely characterize each other. We introduce the following assumption, providing a form of *a priori* information. This plays a role akin to that of a regularizer in non-Bayesian techniques, such as in *Supervised Learning*, where algorithms without such regularizers are prone to overfitting.

Assumption 1 (Output Map Dynamics): The dynamics of the output map are governed by the difference equation

$$\theta_{k+1} = \theta_k + \eta_k, \qquad k \in \mathbb{Z}_+, \tag{4}$$

where k denotes the time index, $\theta_k \in \mathbb{R}^{n_\mathrm{y}n_\mathrm{x}}$ is the vector of parameters driven by the vector of process noise $\eta_k \in \mathbb{R}^{n_\mathrm{y}n_\mathrm{x}}$ with Gaussian distribution $\mathcal{N}(\mu_{\eta_k}, \Sigma_{\eta_k})$. Further, assume that the initial parameter of system (4), denoted by θ_0, is drawn from the normal distribution $\mathcal{N}(\mu_{\theta_0}, \Sigma_{\theta_0})$.

Assumption 1 imposes a Gaussian random walk dynamics on the evolution of the parameters, which is the minimal structure and assumption on the variations of the parameters because of the maximum entropy feature of the Gaussian distributions. This assumption allows us to employ a stochastic belief of a deterministic reality in the *Bayesian* viewpoint.

Let the inputs and outputs of system (1)-(2) be measured for $k = 0, \dots, n_\mathcal{T}$, where $(n_\mathcal{T} + 1) \in \mathbb{Z}_+$ denotes the length of the measurement data. More precisely, the input-output trajectory data is given by $\mathcal{D} = \{(\mathrm{u}_k, \mathrm{y}_k) \,|\, k = 0, \dots, n_\mathcal{T}\}$. Additionally, we assume:

Assumption 2 (Noise): The process, measurement, and output map noise realizations, i.e., w_k, v_k and η_k, respectively, for all $k \in \{0, \dots, n_\mathcal{T}\}$, are independent. Furthermore, the means μ_{x_0}, μ_{θ_0}, μ_{η_k}, and covariance matrices Σ_{x_0}, Σ_{w_k}, ΣV_k, Σ_{θ_0} and Σ_{η_k}, for $k \in \{0, \dots, n_\mathcal{T}\}$, are known.

Remark 1 (A Priori Knowledge): While we assume μ_{θ_0}, μ_{η_k}, Σ_{θ_0}, and Σ_{η_k} to be readily known, in practical applications, these parameters can be obtained through various means depending on the context, e.g., employing prior knowledge of the nominal model, empirically from previous experiments' data, or employing a suitable hyperparameter estimation method when $\mu_{\eta_k} = \mu_{\theta_0}$ and $\Sigma_{\eta_k} = \Sigma_{\theta_0}$, for $k \in \mathbb{Z}_+$.

Ultimately, the question is whether the observation model (2) could be estimated. More precisely, we would like to address the following problem:

Problem 1: Given the process and observation models (1) and (2), input-output measurement data $\mathcal{D}$, and under Assumptions 1 and 2, estimate the *unknown time-varying* observation matrices C_k in an efficient and tractable way.

To address the problem 1, we develop a MAP approach in the next section, followed by a tractable reformulation using LMI techniques in Section IV.

III. MAXIMUM A POSTERIORI ESTIMATION

In this section, we propose a Bayesian method for estimating the unknown observation matrices $\mathrm{C}_0, \dots, \mathrm{C}_{n_\mathcal{T}}$. The three main elements in Bayesian estimation methods are a prior density function, an observation model, and a loss function, which we briefly explain for solving our problem with the MAP approach.

A. Lifted Process and Observation Model

Let us represent the process model (1) in the following *lifted matrix form*:

$$\mathrm{x} = \mathrm{A}(\mathrm{u} + \mathrm{w_x}), \tag{5}$$

where $\mathrm{x} = \left[\mathrm{x}_0^\mathsf{T}, \dots, \mathrm{x}_{n_\mathcal{T}}^\mathsf{T}\right]^\mathsf{T}$ includes the system states over the entire horizon up to time $n_\mathcal{T}$, while the input vector is modified to include the initial state $\mathrm{u} = \left[\mu_{\mathrm{x}_0}^\mathsf{T}, (\mathrm{B}_0\mathrm{u}_0)^\mathsf{T}, \dots, (\mathrm{B}_{n_\mathcal{T}-1}\mathrm{u}_{n_\mathcal{T}-1})^\mathsf{T}\right]^\mathsf{T}$, and the noise vector consists of the uncertainty of the initial state and process noises $\mathrm{w_x} = \left[\mathrm{w}_{\mathrm{x}_0}^\mathsf{T}, \mathrm{w}_0^\mathsf{T}, \dots, \mathrm{w}_{n_\mathcal{T}-1}^\mathsf{T}\right]^\mathsf{T}$ with $\mathrm{w}_{\mathrm{x}_0} \sim \mathcal{N}(0, \Sigma_{\mathrm{x}_0})$. Given that the process noises and initial state uncertainty are uncorrelated from Assumption 2, we can specify $\mathrm{w_x}$ in terms of a multivariate normal distribution $\mathcal{N}(0, \Sigma_{\mathrm{w_x}})$ in which $\Sigma_{\mathrm{w_x}} = \mathrm{diag}(\Sigma_{\mathrm{x}_0}, \Sigma_{\mathrm{w}_0}, \dots, \Sigma_{\mathrm{w}_{n_\mathcal{T}-1}})$. The *lifted* transition matrix A has the lower triangular form:

$$\mathrm{A} = \begin{bmatrix} \mathbb{I} & & & & \\ A_0 & \mathbb{I} & & & \\ A_1 A_0 & A_1 & \mathbb{I} & & \\ \vdots & \vdots & \ddots & \ddots & \\ A_{n_\mathcal{T}-1} \dots A_0 & A_{n_\mathcal{T}-1} \dots A_1 & \dots & A_{n_\mathcal{T}-1} & \mathbb{I} \end{bmatrix}.$$

Similarly, the observation model (2) for the entire trajectory can be expressed as:

$$\mathrm{y} = \mathrm{C}\mathrm{x} + \mathrm{v}, \tag{6}$$

where $\mathrm{y} = \left[\mathrm{y}_0^\mathsf{T}, \dots, \mathrm{y}_{n_\mathcal{T}}^\mathsf{T}\right]^\mathsf{T}$ is the vector of all measurements, $\mathrm{v} \sim \mathcal{N}(0, \Sigma_\mathrm{v})$ is the vector of all measurement noise realizations with $\Sigma_\mathrm{v} = \mathrm{diag}(\Sigma_{\mathrm{v}_0}, \Sigma_{\mathrm{v}_1}, \dots, \Sigma_{\mathrm{v}_{n_\mathcal{T}}})$, and C is the lifted observation matrix:

$$\mathrm{C} = \mathrm{diag}\big(\mathrm{C}_0, \mathrm{C}_1, \dots, \mathrm{C}_{n_\mathcal{T}}\big).$$

Finally, we describe the dynamics of the output map parameters θ for the entire trajectory as:

$$\theta = \mu_\theta + \mathrm{w}_\theta, \tag{7}$$

where $\theta = \left[\theta_0^\mathsf{T}, \dots, \theta_{n_\mathcal{T}}^\mathsf{T}\right]^\mathsf{T}$, and μ_θ results from the summation of the biases of the initial parameter and the noise: $\mu_\theta = \left[\mu_{\theta_0}^\mathsf{T}, \mu_{\theta_0}^\mathsf{T} + \mu_{\eta_0}^\mathsf{T}, \dots, \mu_{\theta_0}^\mathsf{T} + \sum_{i=0}^{n_\mathcal{T}-1} \mu_{\eta_i}^\mathsf{T}\right]^\mathsf{T}$. Similarly, the noise signal w_θ results from the integration over the entire horizon including the uncertainty of the initial parameter as: $\mathrm{w}_\theta = D\eta$, where $\eta = \left[\eta_{\theta_0}^\mathsf{T}, \eta_0^\mathsf{T}, \dots, \eta_{n_\mathcal{T}-1}^\mathsf{T}\right]^\mathsf{T}$ and $\eta_{\theta_0} \sim \mathcal{N}(0, \Sigma_{\theta_0})$, with

$$D = \begin{bmatrix} \mathbb{I} & & & \\ \mathbb{I} & \mathbb{I} & & \\ \vdots & \vdots & \ddots & \\ \mathbb{I} & \mathbb{I} & \dots & \mathbb{I} \end{bmatrix}.$$

Since the parameters are assumed to be independent (Assumption 2), we have $\mathrm{w}_\theta \sim \mathcal{N}(0, \Sigma_{\mathrm{w}_\theta})$, $\Sigma_{\mathrm{w}_\theta} = \mathrm{D}\Sigma_\eta\mathrm{D}^\mathsf{T}$, with $\Sigma_\eta = \mathrm{diag}(\Sigma_{\theta_0}, \Sigma_{\eta_0}, \dots, \Sigma_{\eta_{n_\mathcal{T}-1}})$. Ultimately, the model (7) is used to specify the prior density function.

In what follows, we represent C as $\mathrm{C}(\theta)$ to emphasize the dependence of C on θ according to (3). Consequently, substituting x in (6) with the expression from (5) results in the observation model, with unknown $\mathrm{C}(\theta)$, describing the measurements y as a function of the applied inputs u:

$$\mathrm{y} = \mathrm{C}(\theta)\mathrm{A}\mathrm{u} + \mathrm{w_y}(\theta), \tag{8}$$

where $\mathrm{w_y}(\theta) = \mathrm{C}(\theta)\mathrm{A}\mathrm{w_x} + \mathrm{v}$. Also, from Assumption 2:

$$\mathrm{w_y}(\theta)|\theta \sim \mathcal{N}\big(0, \Sigma_{\mathrm{w_y}}(\theta)\big),$$

where $\Sigma_{\mathrm{w_y}}(\theta) = \mathrm{C}(\theta)\mathrm{A}\Sigma_{\mathrm{w_x}}\mathrm{A}^\mathsf{T}\mathrm{C}(\theta)^\mathsf{T} + \Sigma_\mathrm{v}$. This model is used later to specify the conditional probability density function of the measurements. Note that $w_y(\theta)|\theta$ is conditioned on knowing the parameters and remains Gaussian with the derived covariance since both noise sources, $\mathrm{w_x}$ and v, are Gaussian and independent.

B. MAP Loss Function

In MAP estimation, one aims to find an estimate $\hat{\theta}$ for the parameters by minimizing the cost function [22]: $\mathbb{E}[1 - \mathbb{1}_{\theta:\left\|\theta-\hat{\theta}\right\|_\infty \le \epsilon}(\theta)]$, where θ is the vector of random variables, $\mathbb{1}(.)$ is an indicator function, and ϵ is a small scalar. It is further shown in [23, Chapter 4] that minimizing the expectation of such a loss function implies maximizing the conditional probability density of θ given the vector of observations and inputs, i.e.,

$$\hat{\theta} = \mathrm{argmax}_\theta \; p(\theta|y, u), \tag{9}$$

where $\hat{\theta}$ is the estimate of the true parameter θ. The following lemma formalizes this first step to compute the MAP estimation.

Lemma 1 (MAP Optimization Problem): Let us define the function $\mathcal{J} : \mathbb{R}^{n_\mathrm{y} n_\mathrm{x}(n_\mathcal{T}+1)} \to \mathbb{R}$ as

$$\mathcal{J}(\theta) := \mathrm{logdet}\big(\Sigma_{\mathrm{w_y}}(\theta)\big) + \left\|\mathrm{y} - \mathrm{C}(\theta)\mathrm{A}\mathrm{u}\right\|^2_{\Sigma^{-1}_{\mathrm{w_y}}(\theta)} + \left\|\theta - \mu_\theta\right\|^2_{\Sigma^{-1}_{\mathrm{w}_\theta}}. \tag{10}$$

The MAP estimation (9) is equivalent to:

$$\hat{\theta} = \underset{\theta}{\mathrm{argmin}} \; \mathcal{J}(\theta). \tag{11}$$

Proof: Using Bayes' rule, the MAP estimation (9) can be reformulated as

$$\max_\theta \; p(\theta|\mathrm{y}, \mathrm{u}) = \max_\theta \; \frac{p(\mathrm{y}|\theta, \mathrm{u})p(\theta|\mathrm{u})}{p(\mathrm{y}|\mathrm{u})}. \tag{12}$$

We first note that the denominator of (12) does not depend on θ and, hence, can be neglected without changing the optimizer. Moreover, we note that the dynamics of θ in (4) (or equivalently in the lifted form in (7)) do not depend on the input sequence of u (i.e., $p(\theta|\mathrm{u}) = p(\theta)$). Next, using a straightforward computation, one can derive the probability density functions $p(\theta)$ and $p(\mathrm{y}|\theta, \mathrm{u})$. Specifically, from (7), we know that the variable θ is Gaussian with the probability density function

$$p(\theta) = \frac{\exp\Big(-\frac{1}{2}(\theta - \mu_\theta)^\mathsf{T}\,\Sigma^{-1}_{\mathrm{w}_\theta}(\theta - \mu_\theta)\Big)}{\sqrt{(2\pi)^{(n_\mathcal{T}+1)n_\mathrm{y}n_\mathrm{x}}\det\big(\Sigma_{\mathrm{w}_\theta}\big)}}.$$

Similarly, we know from (8) that given θ and the input sequence u, the output sequence y is also Gaussian with the conditional probability density function

$$p(\mathrm{y}|\theta, \mathrm{u}) = \frac{\exp\Big(-\frac{1}{2}(\mathrm{y} - \mathrm{C}(\theta)\mathrm{A}\mathrm{u})^\mathsf{T}\,\Sigma^{-1}_{\mathrm{w_y}}(\theta)(\mathrm{y} - \mathrm{C}(\theta)\mathrm{A}\mathrm{u})\Big)}{\sqrt{(2\pi)^{(n_\mathcal{T}+1)n_\mathrm{y}}\det\big(\Sigma_{\mathrm{w_y}}(\theta)\big)}}.$$

Finally, applying the monotonically increasing function log and observing that all terms in the denominators except $\det(\Sigma_{\mathrm{w_y}}(\theta))$ are constant, we arrive at the minimization problem of the function $\mathcal{J}$ defined in (11). ■

In the next section, we propose a tractable conservative approximation using LMI techniques to tackle the non-convex objective function $\mathcal{J}(\theta)$ defined in (10).

Remark 2 (Robust Estimation): Alternatively, a robust minimax estimation formulation similar to [8] could be employed. This approach, however, requires finding an ambiguity set to approximate non-Gaussian observation uncertainties due to the multiplication of Gaussian variables in $\mathrm{w_y}(\theta)$.

IV. PROPOSED SOLUTION

The optimization problem (11) is non-convex not only because of the weight $\Sigma_{\mathrm{w_y}}^{-1}(\theta)$, being quadratic in the parameters θ, in the second term but also because of the log-determinant operator in the first term. A typical approach is to use first-order algorithms to find a solution due to the mentioned non-convexities. These algorithms, however, only guarantee convergence to a local optimum. Therefore, selecting an appropriate initial starting point is crucial to the obtained quality of the solution. We propose to solve the problem in two steps: first, we perform a convex relaxation of (11) into a set of LMIs, which we use to compute an initial approximate minimizer; next, we employ this approximate minimizer to initialize (warm-start) a first-order optimization method, e.g., steepest descent [24] or quasi-Newton algorithms [25], to solve (11) thus refining our initial minimizer estimate.

Theorem 1 (LMI Conservative Approximation): Consider the following LMIs:

$$\begin{aligned}
\min_{\mathrm{S},\theta,\gamma,\beta} \quad & \mathrm{tr}(\mathrm{S}-\mathbb{I})+\gamma+\beta \\
\text{s.t.} \quad & \begin{bmatrix} -\Sigma_{\mathrm{w_x}}^{-1} & \mathrm{A}^\top \mathrm{C}(\theta)^\top \\ \mathrm{C}(\theta)\mathrm{A} & \Sigma_{\mathrm{v}}-\mathrm{S} \end{bmatrix} \preceq 0, \\
& \begin{bmatrix} -\mathrm{S} & (\mathrm{y}-\mathrm{C}(\theta)\mathrm{Au}) \\ (\mathrm{y}-\mathrm{C}(\theta)\mathrm{Au})^\top & -\gamma \end{bmatrix} \preceq 0, \\
& \begin{bmatrix} -\Sigma_{\mathrm{w}_\theta} & (\theta-\mu_\theta) \\ (\theta-\mu_\theta)^\top & -\beta \end{bmatrix} \preceq 0.
\end{aligned} \tag{13}$$

Then, the optimal value of the nonlinear program (11) is upper bounded by $J^* + \|\mathrm{y}-\mathrm{C}(\theta^*)\mathrm{Au}\|^2_{\Sigma_{\mathrm{w_y}}^{-1}(\theta^*)-\mathrm{S}^{*-1}}$, where J^* and (S^*, θ^*) are the optimal value and the optimizer of (13), respectively.

Proof: Consider a matrix $\mathrm{S} \succ 0$ upper bounding the covariance matrix $\Sigma_{\mathrm{w_y}}(\theta) \succ 0$ as

$$\Sigma_{\mathrm{w_y}}(\theta) = \mathrm{C}(\theta)\mathrm{A}\Sigma_{\mathrm{w_x}}\mathrm{A}^\top \mathrm{C}(\theta)^\top + \Sigma_{\mathrm{v}} \preceq \mathrm{S}. \tag{14}$$

Thus, $\lambda_i(\Sigma_{\mathrm{w_y}}(\theta)) \le \lambda_i(\mathrm{S})$, for $i = 1, \dots, n_{\mathrm{y}}(n_{\mathcal{T}}+1)$, which implies that

$$\mathrm{logdet}\big(\Sigma_{\mathrm{w_y}}(\theta)\big) \le \mathrm{logdet}(\mathrm{S}).$$

Since $\mathrm{logdet}(\mathrm{S}) = \sum_{i=1}^{n_{\mathrm{y}}(n_{\mathcal{T}}+1)} \log \lambda_i(\mathrm{S})$ and $\mathrm{tr}(\mathrm{S}) = \sum_{i=1}^{n_{\mathrm{y}}(n_{\mathcal{T}}+1)} \lambda_i(\mathrm{S})$, we also have

$$\mathrm{logdet}\big(\Sigma_{\mathrm{w_y}}(\theta)\big) \le \mathrm{logdet}(\mathrm{S}) \le \mathrm{tr}(\mathrm{S}-\mathbb{I}).$$

Using the *Schur complement*, one can see that (14) is equivalent to the following linear matrix inequality

$$\begin{bmatrix} -\Sigma_{\mathrm{w_x}}^{-1} & \mathrm{A}^\top \mathrm{C}(\theta)^\top \\ \mathrm{C}(\theta)\mathrm{A} & \Sigma_{\mathrm{v}}-\mathrm{S} \end{bmatrix} \preceq 0.$$

Similarly, considering $\gamma \ge 0$ and $\beta \ge 0$ such that

$$\begin{aligned}
&(\mathrm{y}-\mathrm{C}(\theta)\mathrm{Au})^\top \mathrm{S}^{-1}(\mathrm{y}-\mathrm{C}(\theta)\mathrm{Au}) \\
&\quad \le (\mathrm{y}-\mathrm{C}(\theta)\mathrm{Au})^\top \Sigma_{\mathrm{w_y}}(\theta)^{-1}(\mathrm{y}-\mathrm{C}(\theta)\mathrm{Au}) \le \gamma,
\end{aligned} \tag{15}$$

and

$$(\theta-\mu_\theta)^\top \Sigma_{\mathrm{w}_\theta}^{-1}(\theta-\mu_\theta) \le \beta, \tag{16}$$

we can apply again the *Schur complement* to the inequalities in (15) and (16) to obtain the last two LMIs in (13). Finally, replacing the terms in the cost function $\mathcal{J}(\theta)$ in (10) with their bounds and including the corresponding LMIs as constraints arrives at the LMIs (13). Note further that by definition, we have

$$\begin{aligned}
&J^* + \|\mathrm{y}-\mathrm{C}(\theta^*)\mathrm{Au}\|^2_{\Sigma_{\mathrm{w_y}}^{-1}(\theta^*)-\mathrm{S}^{*-1}} \\
&\quad = \mathcal{J}\big(\theta^*\big) + \mathrm{tr}\big(\mathrm{S}^*-\mathbb{I}\big) - \mathrm{logdet}\big(\Sigma_{\mathrm{w_y}}(\theta^*)\big) \ge \mathcal{J}\big(\theta^*\big),
\end{aligned} \tag{17}$$

where the function $\mathcal{J}(\theta^*)$ is defined in (10), and the last inequality follows from (14). ■

The tightness of the inequality in (17) mainly depends on the gap between $\log\det(\mathrm{S})$ and $\mathrm{tr}(\mathrm{S}-\mathbb{I})$ since $\log\det(\mathrm{S})$ is bounded from above by $\mathrm{tr}(\mathrm{S}-\mathbb{I})$, which is negligible when $\lambda_i(\mathrm{S}) \approx 1$, for $i = 1, \dots, n_{\mathrm{y}}(n_{\mathcal{T}}+1)$. One may employ a suitable matrix W to scale the eigenvalues of S, replace $\log\det(\mathrm{S})$ with $\log\det(\mathrm{WSW}) - 2\log\det(\mathrm{W})$ and approximate $\log\det(\mathrm{WSW})$ with $\mathrm{tr}(\mathrm{WSW}-\mathbb{I})$. Furthermore, the closeness of J^* and $\mathcal{J}(\theta^*)$ in (17) is proportional to the fitness quality of the measurements and whether S^* is close to the covariance matrix accordingly.

In addition, Theorem 1 provides an approximation of (11), producing an initial near-optimal solution. As indicated earlier, we propose to employ this solution to warm-start a local (non-convex) optimizer. Due to its fast convergence, we propose to employ, as a refining optimizer, the BFGS algorithm [25, Ch. 6], a variant of quasi-Newton methods. The BFGS algorithm approximates the Hessian matrix for its search directions relying on an analytical expression of the gradient $\nabla_\theta \mathcal{J}(\theta)$. The gradient of the cost function (10) with respect to the parameters θ

$$\nabla_\theta \mathcal{J}(\theta) = \left[\frac{\partial \mathcal{J}}{\partial \theta_1}, \dots, \frac{\partial \mathcal{J}}{\partial \theta_{n_{\mathrm{y}} n_{\mathrm{x}}(n_{\mathcal{T}}+1)}}\right]^\top \tag{18}$$

can be easily derived applying the chain rule:

$$\begin{aligned}
\frac{\partial \mathcal{J}}{\partial \theta_{ijk}} = 2\mathrm{tr}\Big[&\Big(\mathrm{A}\Sigma_{\mathrm{w_x}}\mathrm{A}^\top \mathrm{C}(\theta)^\top \Sigma_{\mathrm{w_y}}(\theta)^{-1}\Big)\mathrm{C}_k^{ij}(\theta) \\
&-\Big(\mathrm{A}\Sigma_{\mathrm{w_x}}\mathrm{A}^\top \mathrm{C}(\theta)^\top \Sigma_{\mathrm{w_y}}(\theta)^{-1} \\
&\times (\mathrm{y}-\mathrm{C}(\theta)\mathrm{Au})(\mathrm{y}-\mathrm{C}(\theta)\mathrm{Au})^\top \Sigma_{\mathrm{w_y}}(\theta)^{-1}\Big)\mathrm{C}_k^{ij}(\theta) \\
&-\Big(\mathrm{Au}(\mathrm{y}-\mathrm{C}(\theta)\mathrm{Au})^\top \Sigma_{\mathrm{w_y}}(\theta)^{-1}\Big)\mathrm{C}_k^{ij}(\theta) \\
&+\Big((\theta-\mu_\theta)^\top \Sigma_{\mathrm{w}_\theta}^{-1}\Big)\theta^{ijk}\Big],
\end{aligned}$$

where $\mathrm{C}_k^{ij}(\theta)$ is the single-entry matrix of $\mathrm{C}(\theta)$ with the block matrix of $\mathrm{C}_k(\theta)$ having 1 at index (i, j) and zero elsewhere, and θ^{ijk} is the single-entry vector of θ with 1 at index ijk and zero elsewhere.

The LMIs (13) initializes the original non-convex problem with a locally optimal solution. Thus, the computational complexity of the proposed method consists of the well-known computational complexity of solving the SDP problems [26], i.e., a one-time solution of (13), and the computation of the gradient (18) per iteration of the first-order method:

$$\mathcal{O}\Bigg(\frac{n_{\mathrm{x}}^3(n_{\mathcal{T}}+1)^3+n_{\mathrm{x}}^2(n_{\mathcal{T}}+1)^2}{2}+n_{\mathrm{y}}n_{\mathrm{x}}^2(n_{\mathcal{T}}+1)^2 +n_{\mathrm{x}}n_{\mathrm{y}}^2(n_{\mathcal{T}}+1)^3+\big(n_{\mathrm{y}}\big)^3(n_{\mathcal{T}}+1)^3\Bigg),$$

which is $\mathcal{O}(n_{\mathcal{T}}^3)$ when $n_{\mathrm{x}}, n_{\mathrm{y}} \ll n_{\mathcal{T}}$.

V. NUMERICAL EXAMPLE

In this section, we provide a numerical example to verify the efficacy and performance of the proposed method: employing the LMIs (13) to warm-start the solution of (11) via the BFGS optimizer. Additionally, we compare the resulting solution with the estimates obtained from EM and DKS algorithms. To have a fair comparison, we also employ the same BFGS optimizer for the *M-Step* of the EM estimation.

We demonstrate our solution on the following system:

$$\mathrm{x}_{k+1} = \begin{bmatrix} 0.7 & 0.25 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0.25 & 0.7 \end{bmatrix}\mathrm{x}_k + \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}(3.5+\cos(2k)) + \mathrm{w}_k,$$

with $\mu_{\mathrm{x}_0} = \begin{bmatrix}1, 0.5, 2\end{bmatrix}^{\mathsf{T}}$. The observation is a two-dimensional model, i.e., the number of measurements per time instant is $n_{\mathrm{y}} = 2$. The system has sampled input and measurement pairs in $\mathcal{D}$ every 100 milliseconds for 10 seconds, i.e., $n_{\mathcal{T}} = 100$. Thus, the number of parameters to be estimated is $n_{\mathrm{y}}n_{\mathrm{x}}(n_{\mathcal{T}}+1) = 606$. The noise covariance of process, observation and output map dynamics, Σ_{w_k}, Σ_{v_k} and Σ_{η_k}, are assumed to remain constant across the entire horizon. Moreover, the initial state and parameter noise covariance are similar to the noise covariance of process and output map dynamics, respectively (i.e., $\Sigma_{\mathrm{x}_0} = \Sigma_{\mathrm{w}_k}$ and $\Sigma_{\theta_0} = \Sigma_{\eta_k}$). The output map noise biases μ_{η_k} are generated such that $\mu_{1,\eta_k} = 5+\mathrm{e}^{-0.6k}\cos(0.4k)$, $\mu_{2,\eta_k} = 1.5+\mathrm{e}^{-0.6k}\sin(0.025k)$, $\mu_{3,\eta_k} = 2$, $\mu_{4,\eta_k} = 5+\mathrm{e}^{-0.6k}\cos(0.4k)$, $\mu_{5,\eta_k} = 1.5+\mathrm{e}^{-0.6k}\sin(0.025k)$, and $\mu_{6,\eta_k} = 2$. The initial parameter bias, μ_{θ_0}, is derived from μ_{η_0} by setting $k = 0$, and the DKS and EM algorithms' initialization is according to these noise bias values. We examine the performance of our algorithm, SDP-GD, compared with EM and DKS on four scenarios generated by employing High/Low SNRs for the process and observation noise, particularly 30 and 10 dB, and *High/Low* parameter variation of:

$$\begin{aligned} \text{High}: \Sigma_{\eta_k} &= \mathrm{diag}(2.17,\ 0.076,\ 1.19,\ 1.38,\ 0.87,\ 1.27) \\ \text{Low}: \Sigma_{\eta_k} &= \mathrm{diag}(6.9,\ 0.2,\ 3.8,\ 4.4,\ 2.8,\ 4)\ \cdot 10^{-2}. \end{aligned}$$

Combined results from 100 experiments for each of the four scenarios, keeping the same ground-truth realization in each scenario, are illustrated in Figures 1 and 2. The figures demonstrate the median (vertical dotted lines) and distribution across experiments based on the mean squared error (MSE), i.e., $\mathrm{MSE} = \frac{1}{n_{\mathcal{T}}+1}\sum_{k=0}^{n_{\mathcal{T}}}\|\theta_k - \hat{\theta}_k\|_2^2$. One can observe how the DKS underperforms compared to the EM and our SDP-GD solutions in more challenging scenarios where the process-observation model noise is high or in the presence of *High* parameter variation.

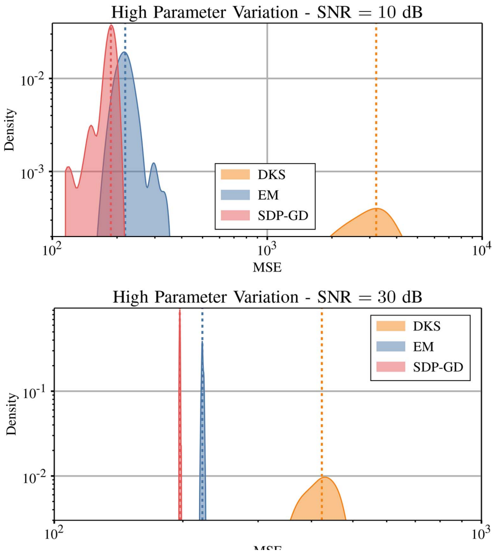

Fig. 1. The Mean Squared Error of the three methods in high noise of Σ_η and two different SNRs for 100 experiments.

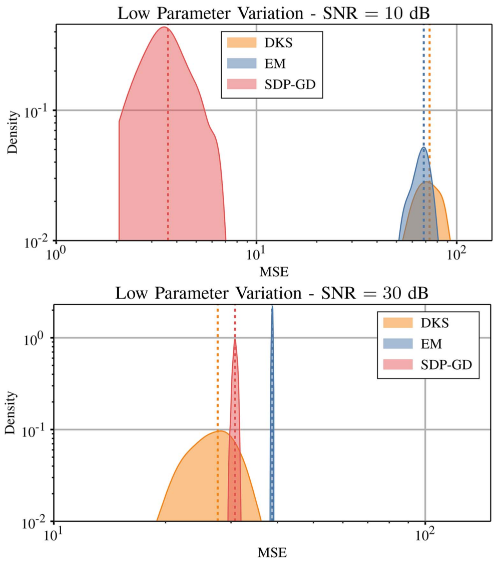

Fig. 2. The Mean Squared Error of the three methods in low noise of Σ_η and two different SNRs for 100 experiments.

TABLE I
THE AVERAGE COMPUTATION PERFORMANCE ON ALL SCENARIOS FOR 100 EXPERIMENTS

	Experiment Scenario	**DKS**	**EM**	**SDP-GD**
10 dB	*Low* Parameter Variation	18 ± 4	14265 ± 3112	1265 ± 109
	High Parameter Variation	27 ± 13	22547 ± 6768	1542 ± 178
30 dB	*Low* Parameter Variation	9 ± 3	7999 ± 802	1543 ± 132
	High Parameter Variation	21 ± 13	3796 ± 412	1605 ± 128

The average and standard deviation of the computation time of each method across 100 experiments are reported in Table I. We performed all the experiments on a cluster node with 384G memory and 40 CPU cores (2 Intel Xeon Gold 6148 @ 2.40GHz). The elapsed execution times confirm our hypothesis that EM is computationally more expensive than the other alternatives.

The performance of EM and DKS algorithms highly depends on the initialization, while in contrast, our proposed solution takes advantage of a warm-start initializer obtained from solving a convexified approximation of the original optimization problem. This initialization helps converge to a better local optimum faster than the EM algorithm. In addition, the *M-step* of the EM algorithm, in this problem, does not hold a closed-form solution, which results in utilizing a first-order method. This gradient *M-Step* also plays a part in the general slowness of the EM algorithm. Our algorithm, however, requires a one-time execution of the set of LMIs followed by an iterative quasi-Newton method with a superlinear convergence rate. Hence, it provides the best of both worlds, i.e., better estimations than EM and DKS with less computation time than EM.

VI. CONCLUSION

We have introduced a method for estimating an unknown output map of a linear time-varying system by employing a stochastic characterization of the evolution of the output map parameters, which serves as *a priori* information for MAP estimation. The derived MAP optimization problem is solved by relaxing the optimization as a set of LMIs, whose solution provides a warm-start for a gradient descent algorithm. Compared with standard approaches to solve this problem, namely EM and DKS, we showed experimentally the superiority of our method in estimation performance and lower computational demands compared to EM. Future work will explore the incorporation of other types of *a priori* knowledge on the output map, the development of efficient causal filters following similar approaches, the minimax formulation for robust estimation, considering noise models with more general structures, and introducing methods for efficient design of the control input.

REFERENCES

[1] T. D. Barfoot, *State Estimation for Robotics*. Cambridge, U.K.: Cambridge Univ. 2017.

[2] R. van de Schoot et al., "Bayesian statistics and modelling," *Nat. Rev. Methods Primers*, vol. 1, no. 1, p. 1, 2021.

[3] R. E. Kalman and R. S. Bucy, "New results in linear filtering and prediction theory," *J. Basic Eng.*, vol. 83, no. 1, pp. 95–108, 1961.

[4] H. Rauch, "Solutions to the linear smoothing problem," *IEEE Trans. Autom. Control*, vol. 8, no. 4, pp. 371–372, Oct. 1963.

[5] S. Yi and M. Zorzi, "Robust Kalman filtering under model uncertainty: The case of degenerate densities," *IEEE Trans. Autom. Control*, vol. 67, no. 7, pp. 3458–3471, Jul. 2022.

[6] S. Shafieezadeh-Abadeh, V. A. Nguyen, D. Kuhn, and P. M. M. Esfahani, "Wasserstein distributionally robust Kalman filtering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 8483–8492.

[7] V. A. Nguyen, S. Shafieezadeh-Abadeh, D. Kuhn, and P. M. Esfahani, "Bridging Bayesian and minimax mean square error estimation via Wasserstein distributionally robust optimization," *Math. Oper. Res.*, vol. 48, no. 1, pp. 1–37, 2023.

[8] S. Yi and M. Zorzi, "Robust fixed-lag smoothing under model perturbations," *J. Frankl. Inst.*, vol. 360, no. 1, pp. 458–483, 2023.

[9] S. Haykin, *Kalman Filtering and Neural Networks*. Hoboken, NJ, USA: Wiley, 2004.

[10] L. E. Baum and J. A. Eagon, "An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology," *Bull. Amer. Math. Soc.*, vol. 73, no. 3, pp. 360–363, 1967.

[11] S. Liu, X. Zhang, L. Xu, and F. Ding, "Expectation–maximization algorithm for bilinear systems by using the Rauch–Tung–Striebel smoother," *Automatica*, vol. 142, Aug. 2022, Art. no. 110365.

[12] M. Zheng and Y. Ohta, "Bayesian positive system identification: Truncated gaussian prior and hyperparameter estimation," *Syst. Control Lett.*, vol. 148, Feb. 2021, Art. no. 104857.

[13] T. B. Schön, A. Wills, and B. Ninness, "System identification of nonlinear state-space models," *Automatica*, vol. 47, no. 1, pp. 39–49, 2011.

[14] Z. Ghahramani and G. E. Hinton, "Parameter estimation for linear dynamical systems," Dept. Comput. Sci., Toronto, ON, Canada, Rep. TR-96-2, 1996.

[15] N. Sammaknejad, Y. Zhao, and B. Huang, "A review of the expectation maximization algorithm in data-driven process identification," *J. Process Control*, vol. 73, pp. 123–136, Jan. 2019.

[16] I. Naim and D. Gildea, "Convergence of the EM algorithm for Gaussian mixtures with unbalanced mixing coefficients," 2012, *arXiv:1206.6427*.

[17] S. Balakrishnan, M. J. Wainwright, and B. Yu, "Statistical guarantees for the EM algorithm: From population to sample-based analysis," *Ann. Statist.*, vol. 45, no. 1, pp. 77–120, 2017.

[18] B. Yan, M. Yin, and P. Sarkar, "Convergence of gradient EM on multi-component mixture of Gaussians," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6959–6969.

[19] M. E. Johnson, *Multivariate Statistical Simulation: A Guide to Selecting and Generating Continuous Multivariate Distributions*. vol. 192. Hoboken, NJ, USA: Wiley, 1987,

[20] S. Sra and R. Hosseini, "Conic geometric optimization on the manifold of positive definite matrices," *SIAM J. Optim.*, vol. 25, no. 1, pp. 713–739, 2015.

[21] R. Hosseini and S. Sra, "An alternative to EM for Gaussian mixture models: Batch and stochastic Riemannian optimization," *Math. Program.*, vol. 181, no. 1, pp. 187–223, 2020.

[22] Z. Chen, "Bayesian filtering: From Kalman filters to particle filters, and beyond," *Statistics*, vol. 182, no. 1, pp. 1–69, 2003.

[23] B. C. Levy, *Principles of Signal Detection and Parameter Estimation*. Berlin, Germany: Springer, 2008.

[24] D. G. Luenberger, *Optimization by Vector Space Methods*. Hoboken, NJ, USA: Wiley, 1997.

[25] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed., New York, NY, USA: Springer, 2006.

[26] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*. Philadelphia, PA, USA: Soc. Ind. Appl. Math., 1994.