On Identification via EM with Latent Disturbances and Lagrangian Relaxation [⋆]

Jack Umenberger [*] **Johan Wågberg** [**] **Ian R. Manchester** [*]
Thomas B. Schön [**]

[*] *School of Aerospace, Mechanical and Mechatronic Engineering,*
University of Sydney, Australia
(e-mail: {j.umenberger, i.manchester}@acfr.usyd.edu.au)
[**] *Department of Information Technology, Uppsala University, Sweden,*
(e-mail: {johan.wagberg, thomas.schon}@it.uu.se}).

Abstract: In the application of the Expectation Maximization (EM) algorithm to identification of dynamical systems, latent variables are typically taken as system states, for simplicity. In this work, we propose a different choice of latent variables, namely, system disturbances. Such a formulation is shown, under certain circumstances, to improve the fidelity of bounds on the likelihood, and circumvent difficulties related to intractable model transition densities. To access these benefits, we propose a Lagrangian relaxation of the challenging optimization problem that arises when formulating over latent disturbances, and fully develop the method for linear models.

Keywords: System identification, expectation maximization, convex relaxation.

1. INTRODUCTION

System identification is the process of building approximate models of dynamical systems from measured data. A common approach is to specify a family of statistical models and choose the parameters of maximum likelihood. Such Maximum Likelihood (ML) methods have been studied extensively and enjoy desirable properties, such as strong consistency; see, e.g., Ljung (1999). Despite this, local maxima in the likelihood function and difficulties in computation of the gradient can mean that application of the method is not always straightforward.

The Expectation Maximization (EM) algorithm is a technique to iteratively solve general ML estimation problems that has proved viable when gradients of the likelihood are difficult to compute. In the context of system identification, it has been used to great effect: see e.g. Schön et al. (2011). The key to the algorithm is the decomposition of the likelihood function based on a user-specified choice of latent variables. Typically, latent variables are chosen so as to simplify the ensuing optimization problem as much as possible; e.g., in Gibson and Ninness (2005) choosing internal system states as latent variables is shown to convexify the search for the ML parameters.

This property has made internal systems states the de facto choice of latent variables in the application of EM to system identification. However, recent research on convexification of simulation error minimization, see e.g. Tobenkin et al. (2010), has expanded the class of value functions that can be efficiently optimized, thereby presenting opportunities for different choices of latent variables.

This paper offers an investigation into the merits of formulating the EM algorithm over latent system disturbances, rather than latent systems states, and proposes a novel strategy for convexifying the challenging optimization problem that ensues.

2. PRELIMINARIES

2.1 Problem formulation

Consider a class of state space models of the form

$$x_{t+1} = a_\theta(x_t, u_t) + w_t, \tag{1a}$$

$$y_t = g_\theta(x_t, u_t) + v_t, \tag{1b}$$

where $x_t \in \mathbb{R}^{n_x}$ denotes the unobserved state variable, and $u_t \in \mathbb{R}^{n_u}$, $y_t \in \mathbb{R}^{n_y}$ denote the observed input and output, respectively. The nonlinear functions a_θ and g_θ are finitely parametrized by θ_a and θ_g, respectively. The disturbance sequence, $W_T = \{w_t\}_{t=1}^T$, is assumed to be distributed according to

$$W_T \sim p_{\theta_w}(W_T), \tag{2}$$

where the PDF is of known form (e.g. Gaussian) but parametrized by unknown θ_w. The measurement noise, v_t, is assumed to be a zero mean Gaussian white noise process, and the initial condition x_1 is also assumed to be normally distributed,

$$v_t \sim \mathcal{N}(0, \Sigma_v), \tag{3a}$$

$$x_1 \sim \mathcal{N}(\mu, \Sigma_1). \tag{3b}$$

All unknown model parameters are grouped into a single vector, $\theta = \{\theta_a, \theta_g, \theta_w, \Sigma_v, \mu, \Sigma_1\}$.

Given sequences of measured inputs $U_T = \{u_t\}_{t=1}^T$ and outputs $Y_T = \{y_t\}_{t=1}^T$ our task is to find the maximum likelihood estimate of the model parameters θ, defined

$$\theta^{ML} = \arg\max_\theta p_\theta(U_T, Y_T), \tag{4}$$

[⋆] This work was supported by the Australian Research Council (DP130100551), and the Swedish Research Council (VR) as part of the project: Probabilistic modeling of dynamical systems (Contract number: 621-2013-5524).

where $p_\theta(U_T, Y_T)$ is the joint density (likelihood) of the observations U_T and Y_T. It is often more convenient to work with the log likelihood, $L_\theta(Y_T) \triangleq \log p_\theta(U_T, Y_T)$. Henceforth, u_t is implicit in the notation, for brevity.

2.2 The Expectation Maximization algorithm

In this section we review the basic principles of ML estimation via the EM algorithm; refer to Dempster et al. (1977) for further details. The algorithm is predicated on the assumption that there exists a set of latent (read: 'hidden' or 'unobserved') variables, Z, such that the 'complete' or joint log likelihood function

$$L_\theta(Y_T, Z) = \log p_\theta(Y_T, Z)$$

is easier to optimize than the incomplete log likelihood $L_\theta(Y_T)$. These latent variables may be thought of as the data that we 'wish' we could observe, in the sense that the problem would be more straightforward if Z was available.

The maximum likelihood problem can be related to the joint likelihood by marginalizing over the latent variables

$$\theta^{ML} \triangleq \arg\max_\theta L_\theta(Y_T) = \arg\max_\theta \log \int p_\theta(Y_T, Z) dZ.$$

This is a formidable optimization problem, as marginalization has separated the logarithm from the likelihood. The idea behind the EM algorithm is to take some estimate θ_k of the parameters, use this to build a lower bound for $L_\theta(Y_T)$, then maximize the lower bound in place of the likelihood function to improve our estimate of θ.

For an arbitrary distribution $\rho(Z)$, Jensen's inequality gives

$$\int \rho(Z) \log \frac{p_\theta(Y_T, Z)}{\rho(Z)} dZ \leq \log \int \rho(Z) \frac{p_\theta(Y_T, Z)}{\rho(Z)} dZ,$$

where the right hand side is simply $L_\theta(Y_T)$. Therefore, we may define a lower bound for the likelihood function by

$$B_\rho(\theta, \theta_k) \triangleq \int \rho(Z) \log \frac{p_\theta(Y_T, Z)}{\rho(Z)} dZ. \tag{5}$$

Notice that Jensen's inequality has 'reunited' the logarithm with the likelihood, thereby making the bound more amenable to optimization. Choosing $\rho(Z) = p_{\theta_k}(Z|Y_T)$ yields an 'optimal bound' in the sense that $B_\rho(\theta_k, \theta_k) = L_{\theta_k}(Y_T)$ and so intuitively, maximizing $B_\rho(\theta, \theta_k)$ w.r.t θ will result in $L_\theta(Y_T) > L_{\theta_k}(Y_T)$.

It is convenient to express the optimal bound in the form

$$B_\rho(\theta, \theta_k) = Q(\theta, \theta_k) + H(\theta_k),$$

where $Q(\theta, \theta_k)$ represents

$$\int p_{\theta_k}(Z|Y_T) \log p_\theta(Y_T, Z)\ dZ = \mathrm{E}_{\theta_k}\big[\log p_\theta(Y_T, Z)|Y_T\big]$$

and $H(\theta_k)$ denotes the differential entropy of $p_{\theta_k}(Z|Y_T)$. As $H(\theta_k)$ is independent of θ, maximizing the bound reduces to maximizing $Q(\theta, \theta_k)$.

To summarize, each iteration of the EM algorithm consists of an expectation (E) step to compute $Q(\theta, \theta_k)$, and a maximization (M) step in which $Q(\theta, \theta_k)$ is maximized to deliver an improved θ_{k+1}, such that $L_{\theta_{k+1}}(Y_T) \geq L_{\theta_k}(Y_T)$.

2.3 Latent variables for dynamical systems

In the application of EM to the identification of dynamical systems, there are two natural candidates for the choice of latent variables: systems states, x_t, and system disturbances, w_t. Choosing latent states yields a joint likelihood function of the form

$$p_\theta(Y_T, X_T) = \left[\prod_{t=1}^{T} p_\theta(y_t|x_t)\right]\left[\prod_{t=1}^{T-1} p_\theta(x_{t+1}|x_t)\right] p_\theta(x_1). \tag{8}$$

whereas latent disturbances leads to

$$p_\theta(Y_T, x_1, W_T) = \left[\prod_{t=1}^{T} p_\theta(y_t|x_t)\right] p_\theta(W_T) p_\theta(x_1), \tag{9}$$

where $x_{t+1} = a_\theta(x_t, u_t) + w_t$ for $t \in [1, T]$. We denote this state sequence by $\mathcal{X}_T(\theta, x_1, W_T)$, which, for given θ, is a deterministic mapping from initial conditions and disturbances to system states.

Algorithm 1 Expectation Maximization algorithm

```
(1) Set $k = 0$ and initialize $\theta_k$ such that $L_{\theta_k}(Y_T)$ is finite.
(2) Expectation (E) Step:
        $Q(\theta, \theta_k) = \mathrm{E}_{\theta_k}\big[\log p_\theta(Y_T, Z)|Y_T\big]$    (6)
(3) Maximization (M) Step:
        $\theta_{k+1} = \arg\max_\theta Q(\theta, \theta_k)$    (7)
(4) If not converged, $k \leftarrow k + 1$ and return to step 2.
```

It is widely recognized that formulating the EM algorithm over latent states can simplify the optimization problem in (7). In contrast, $\mathcal{X}_T(\theta, x_1, W_T)$ renders (9) a highly nonconvex function of θ, thereby complicating the M step. Despite this, there are some compelling reasons to believe that latent disturbances may offer benefits over latent states, particularly related to the fidelity of the bound $B_\rho(\theta, \theta_k)$; see Section 5.

3. CONVEXIFICATION OF EM WITH LATENT DISTURBANCES

The standard choice of latent states can be interpreted as a way of simplifying, and in some cases convexifying, the M step. In this section, we propose an alternative convexification strategy that enables the EM algorithm to be formulated over latent disturbances.

3.1 Expectation step

We begin by outlining the consequences of formulating the EM algorithm over latent disturbances. To perform the E step, i.e. compute $Q(\theta, \theta_k)$ as in (6), it is convenient to use the following decomposition of (9),

$$\begin{aligned}&\underbrace{\mathrm{E}_{\theta_k}\big[\log p_\theta(Y_T, x_1, W_T)|Y_T\big]}_{Q(\theta,\theta_k)} = \underbrace{\mathrm{E}_{\theta_k}\big[\log p_\theta(x_1)|Y_T\big]}_{Q_1(\theta,\theta_k)} + \\ &\underbrace{\mathrm{E}_{\theta_k}\big[\log p_\theta(W_T)|Y_T\big]}_{Q_2(\theta,\theta_k)} + \underbrace{\mathrm{E}_{\theta_k}\big[\log p_\theta(Y_T|\mathcal{X}_T(\theta, x_1, W_T))|Y_T\big]}_{Q_3(\theta,\theta_k)}\end{aligned} \tag{10}$$

and compute each term separately.

Ignoring constant terms, $Q_1(\theta, \theta_k)$ can be expressed as

$$-\Big[\log\det\Sigma_1 + \mathrm{tr}\big(\Sigma_1^{-1}\big((\hat{x}_{1|T} - \mu)(\hat{x}_{1|T} - \mu)' + \hat{\Sigma}_{1|T}\big)\big)\Big] \tag{11}$$

where

$$\hat{x}_{1|T} = \mathrm{E}_{\theta_k}\big[x_1|Y_T\big], \quad \hat{\Sigma}_{1|T} = \mathrm{Var}_{\theta_k}\big[x_1|Y_T\big]. \tag{12}$$

Calculating the quantities in (12) amounts to solving a state smoothing problem. For general nonlinear models, recent developments in particle smoothing methods, see e.g. Lindsten and Schön (2013), may be leveraged.

Computation of $Q_2(\theta,\theta_k)$ can be more challenging, depending on the complexity of the distribution $p_\theta(W_T)$. We propose a Monte Carlo approximation: if we can generate realizations W_T^i from the distribution $p_{\theta_k}(W_T|Y_T)$ then

$$\tilde{Q}_2(\theta,\theta_k) \triangleq \frac{1}{M}\sum_{i=1}^{M}\log p_\theta(W_T^i) \approx \mathrm{E}_{\theta_k}\big[\log p_\theta(W_T)|Y_T\big],$$

with equality in the limit $M \to \infty$. To sample from $p_{\theta_k}(W_T|Y_T)$, first generate realizations from $p_{\theta_k}(X_T|Y_T)$, again using particle smoothing methods. Then W_T^i can be recovered by substituting X_T^i into (1a) and solving for w_t.

Finally, we turn our attention to computation of $Q_3(\theta,\theta_k)$, and once more employ a Monte Carlo (MC) approximation

$$\begin{aligned}&\mathrm{E}_{\theta_k}\big[\log p_\theta(Y_T|\mathcal{X}_T(\theta,x_1,W_T))|Y_T\big]\\ &\approx \frac{1}{M}\sum_{i=1}^{M}\log p_\theta(Y_T|\mathcal{X}_T(\theta,x_1^i,W_T^i)),\end{aligned} \tag{13}$$

where $\{x_1^i, W_T^i\}$ are sampled from $p_{\theta_k}(x_1,W_T|Y_T)$. By the Markov property of (1b), (13) can be expressed as

$$\tilde{Q}_3(\theta,\theta_k) \triangleq -\left[\frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}|y_t - g_\theta(x_t^i,u_t)|^2_{\Sigma_v^{-1}} + T\log\det\Sigma_v\right] \tag{14}$$

where $x_{1:T}^i = \mathcal{X}_T(\theta,x_1^i,W_T^i)$, and constants are ignored.

3.2 Maximization step

To perform the M step, i.e. maximize $Q(\theta,\theta_k)$ as in (7), we will utilize the same decomposition as in (10), and optimize each of the conditional expectations separately. This is valid because each term in (10) is a function of different parameters: μ and Σ_1 appear only in $Q_1(\theta,\theta_k)$; θ_w in $\tilde{Q}_2(\theta,\theta_k)$; and θ_a,θ_g and Σ_v in $\tilde{Q}_3(\theta,\theta_k)$.

Maximization of $Q_1(\theta,\theta_k)$ is straightforward once the smoothed quantities of (12) have been computed; closed form expressions for the global optimizer are given in Section 4.2. Maximization of the Monte Carlo approximation $\tilde{Q}_2(\theta,\theta_k)$ can be handled by numerical optimization methods, e.g. gradient descent, initialized at θ_k.

Finally, we must consider maximization of $\tilde{Q}_3(\theta,\theta_k)$. This is a challenging problem, as the choice of latent variables $Z = \{x_1, W_T\}$ means that the states $\mathcal{X}_T(\theta,x_1,W_T)$ now change as a function of θ, rendering $\tilde{Q}_3(\theta,\theta_k)$ a nonconvex function of the model parameters.

We now present the main contribution of this paper: a convexification of the maximization of $\tilde{Q}_3(\theta,\theta_k)$. To make the connection to existing work more transparent, we introduce the 'weighted simulation error', defined

$$\mathcal{E}(\theta,x_1,W_T) \triangleq \sum_{t=1}^{T}|y_t - g_\theta(x_t,u_t)|^2_{\Sigma_v^{-1}}, \tag{15}$$

where $x_{1:T} = \mathcal{X}_T(\theta,x_1,W_T)$. Applying this definition to (14), $-\tilde{Q}_3(\theta,\theta_k)$ can now be expressed as

$$\frac{1}{M}\sum_{i=1}^{M}\mathcal{E}(\theta,x_1^i,W_T^i) + T\log\det\Sigma_v \tag{16}$$

and thus maximization of $\tilde{Q}_3(\theta,\theta_k)$ is equivalent to minimization of (16). To convexify (16), let us first replace the concave term $\log\det\Sigma_v$ with an affine upper bound

$$\log\det\Sigma_{v_k} + \mathrm{tr}\big(\Sigma_{v_k}^{-1}\Sigma_v\big),$$

which is tight at Σ_{v_k}, our best guess of the covariance Σ_v.

Next, suppose there exists a function $\hat{J}_\lambda(\theta,x_1,W_T)$, convex in θ, that upper bounds the simulation error, i.e.

$$\hat{J}_\lambda(\theta,x_1,W_T) \geq \mathcal{E}(\theta,x_1,W_T) \quad \forall\theta\in\Theta.$$

where Θ is a convex set. It is then easy to see that

$$\frac{1}{M}\sum_{i=1}^{M}\hat{J}_{\lambda_i}(\theta,x_1^i,W_T^i) \geq \frac{1}{M}\sum_{i=1}^{M}\mathcal{E}(\theta,x_1^i,W_T^i) \quad \forall\theta\in\Theta.$$

Therefore, provided such a function $\hat{J}_\lambda(\theta,x_1,W_T)$ exists,

$$\frac{1}{M}\sum_{i=1}^{M}\hat{J}_{\lambda_i}(\theta,x_1^i,W_T^i) + T\log\det\Sigma_{v_k} + T\mathrm{tr}\big(\Sigma_{v_k}^{-1}\Sigma_v\big), \tag{17}$$

represents a convex upper bound for (16).

3.3 Convex relaxation of simulation error minimization

To realize the upper bound in (17) we draw inspiration from recent research presented in Megretski (2008); Tobenkin et al. (2010); Manchester et al. (2012), which provides a candidate for $\hat{J}_\lambda(\theta,x_1,W_T)$. To derive this function, first consider the widely studied problem of minimizing simulation error, which may be formalized as

$$J^* \triangleq \min_{\theta,X_T} J(\theta,X_T) \triangleq \sum_{t=1}^{T}|y_t - g_\theta(x_t,u_t)|^2_{\Sigma_v^{-1}} \tag{18a}$$

$$\text{s.t. } \mathcal{F}(\theta,X_T,W_T) = 0. \tag{18b}$$

Here $\mathcal{F}(\theta,X_T,W_T)$ encodes the dynamic constraints on x_t, such that $\mathcal{F}(\theta,\mathcal{X}_T(\theta,x_1,W_T),W_T) = 0$.

The key idea is the application of Lagrangian relaxation, or the S-Procedure; a technique used extensively in robust control to approximate difficult constrained optimization problems with 'easier' unconstrained problems. In our context, Lagrangian relaxation takes the form

$$\hat{J}_\lambda(\theta,x_1,W_T) \triangleq \sup_{X_{2:T}}\{J(\theta,X_T) - \lambda'\mathcal{F}(\theta,X_T,W_T)\} \tag{19}$$

where λ can be interpreted as a Lagrange multiplier. For arbitrary λ, the function $\hat{J}_\lambda(\theta)$ has the two essential properties that we require:

1) It is convex in θ. To see this, observe that when a and g are linearly parametrized, J and $\mathcal{F}$ are convex and affine functions of θ, respectively. Therefore, $\hat{J}_\lambda(\theta)$ is convex in θ, as it is the supremum of an infinite family of convex functions; see Section 3.2.3 of Boyd and Vandenberghe (2004).
2) It is an upper bound for the simulation error. To see this, observe that if $X_T = \mathcal{X}_T(\theta,x_1,W_T)$, then
$$J(\theta,X_T) - \lambda'\mathcal{F}(\theta,X_T,W_T) = \mathcal{E}(\theta,x_1,W_T),$$
implying that the supremum over X_T can be no smaller.

The original simulation error minimization (18) may then be approximated by the convex optimization problem

$$\hat{J}_\lambda^* \triangleq \min_\theta \; \hat{J}_\lambda(\theta, x_1, W_T). \tag{20}$$

The remaining challenge is to choose the Lagrange multiplier λ such that $\hat{J}_\lambda(\theta)$ is a useful upper bound, i.e., such that $\hat{J}_\lambda^* \approx J^*$. Unfortunately, the simultaneous search for λ and θ is not jointly convex, due to the coupling between λ and $\mathcal{F}$, and so λ must be specified in advance.

4. IDENTIFICATION OF LGSS MODELS

We now apply the strategy proposed in Section 3 to the special case of linear Gaussian state space (LGSS) models of the form

$$x_{t+1} = Ax_t + Bu_t + Gw_t, \tag{21a}$$

$$y_t = Cx_t + Du_t + v_t, \tag{21b}$$

where x_1 and v_t are distributed according to (3), and $w_t \sim \mathcal{N}(0, \Sigma_w)$.

4.1 Expectation step

To compute $Q_1(\theta, \theta_k)$ we require the quantities in (12), which in the LGSS case may be calculated, e.g., by the methods of (Durbin and Koopman, 2012, Section 4.4). To compute $Q_2(\theta, \theta_k)$, the assumption of independent disturbances implies that it can be expressed as

$$-\Big[T \log\det \Sigma_w + \sum_{t=1}^{T} \mathrm{tr}\big(\Sigma_w^{-1} \mathrm{E}_{\theta_k}\left[w_t w_t' | Y_T\right]\big)\Big] \tag{22}$$

where constants are ignored. Again, in the LGSS case, $\mathrm{E}_{\theta_k}\left[w_t w_t' | Y_T\right]$ can be computed exactly by standard disturbance smoothers; see, e.g., (Durbin and Koopman, 2012, Section 4.5). Finally, to compute $\tilde{Q}_3(\theta, \theta_k)$, realizations of $p_{\theta_k}(x_1, W_T | Y_T)$ can be efficiently generated by mean correction methods, such as those in (Durbin and Koopman, 2012, Section 4.9).

4.2 Maximization step

To maximize $Q_1(\theta, \theta_k)$, notice that (11) is concave w.r.t μ and Σ_1^{-1}. Therefore, setting the gradient to zero gives the global maximizers $\mu = \hat{x}_{1:T}$ and $\Sigma_1 = \hat{\Sigma}_{1:T}$. Maximization of $Q_2(\theta, \theta_k)$ can be handled in a similar way. Defining

$$\hat{\Sigma}_w = \frac{1}{T} \sum_{t=1}^{T} \mathrm{E}_{\theta_k}\left[w_t w_t' | Y_T\right] \tag{23}$$

and substituting into (22), $Q_2(\theta, \theta_k)$ is proportional to

$$-\Big[T \log\det \Sigma_w + T \mathrm{tr}\big(\Sigma_w^{-1} \hat{\Sigma}_w\big)\Big]$$

and thus, by the same arguments used above, the global maximizer is given by $\Sigma_w = \hat{\Sigma}_w$.

Finally, maximization of $\tilde{Q}_3(\theta, \theta_k)$ is handled by the Lagrangian relaxation proposed in Section 3.2. To apply this method to LGSS models, first consider an implicit representation of the dynamics in (21a)

$$Ex_{t+1} = Fx_t + Ku_t + Lw_t, \tag{24}$$

where the original form can be recovered as $A = E^{-1}F$, $B = E^{-1}K$ and $G = E^{-1}L$. It is known that equivalent constraints can give non-equivalent bounds in Lagrangian relaxation, and so these implicit dynamic constraints improve performance, in general. For these linear dynamics, the dynamic constraint (18b) becomes

$$\mathcal{F}(\theta, X_T, W_T) = \bar{F}(\theta) X_{2:T} + \epsilon(\theta, x_1, W_T) = 0, \tag{25}$$

where $\bar{F} \in \mathbb{R}^{(T-1)n_x \times (T-1)n_x}$ and $\epsilon \in \mathbb{R}^{(T-1)n_x}$ are given by

$$\begin{bmatrix} E & 0 & \dots & & \\ -F & E & 0 & & \\ 0 & -F & E & 0 & \\ \vdots & & & \ddots & \ddots \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} -Fx_1 - K\tilde{u}_1 - Gw_1 \\ -K\tilde{u}_2 - Gw_2 \\ \vdots \\ -K\tilde{u}_{T-1} - Gw_{T-1} \end{bmatrix}$$

respectively.

4.3 Lagrange multipliers

To utilize the convex bound (17) proposed in 3.2, we must supply suitable Lagrange multipliers, $\{\lambda_i\}_{i=1}^M$. In passing, we note that the number of particles, M, used in (14) is a trade-off between accuracy of the MC approximation, and size of the optimization problem. In practice, we have had success with M on the order of tens of particles.

Recall that the key to the EM algorithm is that optimization of $Q(\theta, \theta_k)$ guarantees improvement of $L_\theta(Y_T)$, and hence, we must ensure that optimization of (17) results in improvement of $\tilde{Q}_3(\theta, \theta_k)$. The simplest way to achieve this is to find λ_i such that $\hat{J}_{\lambda_i}(\theta_k, x_1^i, W_T^i) = \mathcal{E}(\theta_k, x_1^i, W_T^i)$ for all $\{x_1^i, W_T^i\}_{i=1}^M$ appearing in (17).

Fortunately, in the case of the LGSS model (21), such a λ is readily found, as shown in the following lemma:

Lemma 1. Consider a Lagrange multiplier of the form

$$\lambda = X_{2:T} + h = [(x_2 + h_2)', \dots, (x_T + h_T)']'.$$

Suppose, for a given θ_k, we set

$$h = (\bar{F}')^{-1}\big[\Psi \mathcal{X}_{2:T}(\theta_k, x_1, W_T) - \bar{C}'\bar{\Sigma}_v^{-1}(Y_{2:T} - \bar{D}U_{2:T}) - \epsilon\big] \tag{26}$$

where

$$\Psi = \bar{C}'\bar{\Sigma}_v^{-1}\bar{C} - \bar{F} - \bar{F}', \tag{27}$$

$$\bar{C} = I_{T-1} \otimes C, \;\; \bar{D} = I_{T-1} \otimes D \;\; \text{and} \;\; \bar{\Sigma}_v = I_{T-1} \otimes \Sigma_v.$$

Here, I_n denotes the $n \times n$ identity matrix and $\otimes$ the Kronecker product. Then, with this choice of λ, we have

$$\hat{J}_\lambda(\theta_k, x_1, W_T) = \mathcal{E}(\theta_k, x_1, W_T).$$

Proof. Due to space restrictions, we merely sketch the proof. In this case, $J(\theta, X_T) - \lambda'\mathcal{F}(\theta, X_T, W_T)$ becomes

$$X_{2:T}' \Psi X_{2:T} - 2\psi' X_{2:T} + \gamma \tag{28}$$

and so the supremum in (19) can be computed analytically, assuming $\Psi < 0$. Setting the supremizing $X_{2:T}$ to be $\mathcal{X}_{2:T}(\theta_k, x_1, W_T)$ and solving for h yields (26). □

The proof of Lemma 1 assumed negative definite Ψ; the following lemma clarifies when this assumption holds:

Lemma 2. Consider the set of models of the form (24) defined by the constraint

$$\Theta = \{\theta : F'PF - E - E' + P^{-1} + C'\Sigma_v^{-1}C < 0\} \tag{29}$$

for some $P = P' > 0$. This set has the following properties:

1. Ψ is negative definite when constructed from $\theta \in \Theta$.
2. A model of the form (21) is stable if and only if $\theta \in \Theta$.

3. $\{\Theta, P\}$ is a convex set.

Proof. Properties 2 and 3 are straightforward extensions of Lemma 4 and Section 2.1 of Manchester et al. (2012), respectively. Property 1 can be inferred from (Megretski, 2008, Theorem 2).

Therefore, by restricting our model class to Θ defined in (29), we ensure stability of the identified model and $\Psi < 0$. Algorithm 2 provides a summary of the steps required to implement the proposed method.

Algorithm 2 LGSS identification via EM

```
(1) Set $k = 0$ and initialize $\theta_k$ such that $L_{\theta_k}(Y_T)$ is finite.
(2) Expectation (E) Step:
    (2.1) Compute $\hat{x}_{1|T}$ and $\hat{\Sigma}_{1|T}$ in (12).
    (2.2) Compute $\hat{\Sigma}_w$ in (23).
    (2.3) Generate $M$ realizations from $p_{\theta_k}(x_1, W_T|Y_T)$.
(3) Maximization (M) Step:
    (3.1) Set $\mu = \hat{x}_{1|T}$, $\Sigma_1 = \hat{\Sigma}_{1|T}$ and $\Sigma_w = \hat{\Sigma}_w$.
    (3.2) Compute $\{\lambda^i\}_{i=1}^M$ from $\{x_1^i, W_T^i\}_{i=1}^M$ using (26).
    (3.3) Update $\theta_a, \theta_g$ and $\Sigma_v$ by solving (20), with
          $\{\theta_a, \theta_g, \Sigma_v\} \in \Theta$ defined by (29).
(4) If not converged, $k \leftarrow k + 1$ and return to step 2.
```

5. ON THE MERITS OF LATENT DISTURBANCES

5.1 Fidelity of bound on the likelihood function

The premise of the EM algorithm is the optimization of a tractable lower bound $B_\rho(\theta, \theta_k)$ in lieu of $L_\theta(Y_T)$. In practice, there is a trade-off between the fidelity of $B_\rho(\theta, \theta_k)$ and ease of optimization; e.g. a high fidelity bound is of little value if it is riddled with local maxima. Nonetheless, suppose one was given a method of optimizing any $B_\rho(\theta, \theta_k)$. Then, one may wonder: which choice of latent variables produces the bound that represents $L_\theta(Y_T)$ most faithfully?

Recall from Section 2.2, that the quality of the bound depends on the arbitrary distribution $\rho(Z)$, with $\rho = p_{\theta_k}(Z, |Y_T)$ optimal in the sense that then $B_\rho(\theta_k, \theta_k) = L_{\theta_k}(Y_T)$. It thus follows that when the sensitivity of $p_{\theta_k}(Z|Y_T)$ to θ_k is low, $B_\rho(\theta, \theta_k)$ bounds $L_\theta(Y_T)$ more tightly. In fact, if the Kullback-Leibler (KL) divergence of $p_{\theta_k}(Z|Y_T)$ from $p_\theta(Z|Y_T)$, which we denote $D_{\mathrm{KL}}(\theta_k, \theta, Z)$, is used to approximate this sensitivity, it gives the exact error between $B_\rho(\theta, \theta_k)$ and $L_\theta(Y_T)$ at θ.

The question now becomes: is $p_\theta(X|Y_T)$ or $p_\theta(x_1, W_T|Y_T)$ least sensitive to θ? To gain some insight into this question we have plotted $D_{\mathrm{KL}}(\theta_k, \theta, X_T)$ and $D_{\mathrm{KL}}(\theta_k, \theta, \{x_1, W_T\})$ for a generic second order LGSS model; see Fig. 1. Such a model is parametrized by $\theta = \{\zeta, \omega_n\}$ where ζ denotes the damping ratio and ω_n the natural frequency. Refer to the figure caption for experimental details.

Fig. 1 clearly depicts $D_{\mathrm{KL}}(\theta_k, \theta, \{x_1, W_T\})$ uniformly under bounding $D_{\mathrm{KL}}(\theta_k, \theta, X_T)$. While this result is by no means conclusive, it lends some support to the intuitive hypothesis that $p_\theta(x_1, W_T|Y_T)$ may be less sensitive than $p_\theta(X|Y_T)$ to θ. A deeper understanding of this relationship is deserving of further study.

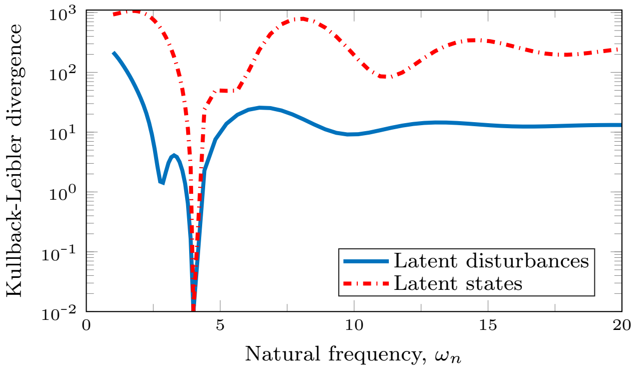

Fig. 1. $D_{\mathrm{KL}}(\theta_k, \theta, X_T)$ and $D_{\mathrm{KL}}(\theta_k, \theta, \{x_1, W_T\})$ for a LGSS model, with true $\hat{\theta} = \{0.3, 10\}$, $\theta_k = \{0.3, 4\}$ and $\theta = \{0.3, \omega_n\}$. The covariance of disturbances and measurement noise, both Gaussian, is equal to 0.1.

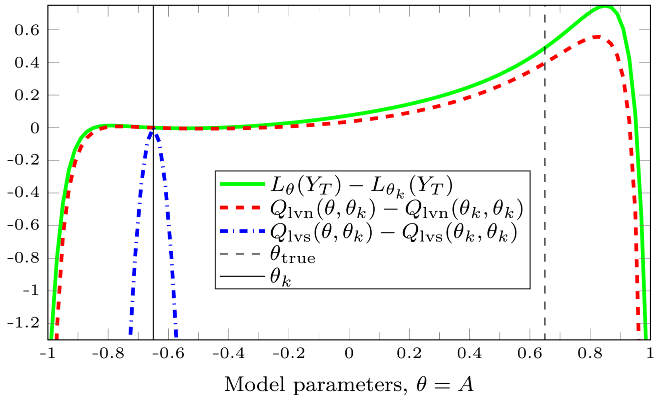

Fig. 2. Lower bounds on the likelihood function: $Q_{\mathrm{lvn}}(\theta, \theta_k)$ and $Q_{\mathrm{lvs}}(\theta, \theta_k)$ are formulated with latent disturbances and latent states, respectively. Additive disturbances and measurement noise are zero-mean Gaussian, with covariance 0.01 and 0.1, respectively.

5.2 Effect of disturbances on bound degeneration

It has been observed, see e.g. Schön et al. (2011), that the bound $B_\rho(\theta, \theta_k)$ built with latent states, begins to degenerate when the magnitude of the disturbances are small. In fact, when $w_t \equiv 0$, $p_\theta(x_{t+1}|x_t)$ becomes deterministic, and so $p_{\theta_k}(X_T|Y_T)$ reduces to a δ function; thus the bound collapses to a single point at $\theta = \theta_k$.

Conversely, as was shown in Section 2.3, when we formulate the EM algorithm over latent disturbances, the problematic $p_\theta(x_{t+1}|x_t)$ is eliminated, and so the bound remains well behaved. In fact, for given x_1 and $w_t \equiv 0$, $B_\rho(\theta, \theta_k) = L_\theta(Y_T)$, and thus the bound exactly reproduces the likelihood function.

This cursory analysis suggests that in regions of the parameter space where the magnitude of disturbances is small, at least relative to the measurement noise, higher fidelity bounds may be obtained by choosing latent disturbances instead of latent states. We illustrate this principle for a simple first order LGSS model in Fig. 2. The disturbances are an order of magnitude smaller than the measurement noise, and so as expected, the bound formed with latent disturbances represents $L_\theta(Y_T)$ most faithfully.

5.3 Circumventing problematic transition densities

To formulate the EM algorithm over latent states, one must construct the joint likelihood function (8) from the *transition density* $p_\theta(x_{t+1}|x_t)$. Unfortunately, for many models of interest $p_\theta(x_{t+1}|x_t)$ may not have a closed form. A simple example of this is a LGSS model with rank deficient disturbances, however more interesting instances arise, e.g., when modeling diffusion processes. In such cases, EM based on latent states breaks down, although some solutions to this problem have been proposed for linear models; see e.g. Solo (2003).

In contrast, by formulating the EM algorithm over latent disturbances, we obtain the joint likelihood function of (9). Comparing (9) to (8), notice that the problematic transition density has been replaced with the disturbance distribution, $p_\theta(W_T)$. Therefore, by reformulating over latent disturbances, one may elegantly extend the class of models that can be identified via EM, to include those that lack closed form expressions for the transition density.

5.4 Preliminary numerical experiments

Finally, we demonstrate the performance of Algorithm 2 by a comparison with the latent states based algorithm in Gibson and Ninness (2005). Specifically, we apply each algorithm to the identification of 500 different first order LGSS models, randomly generated by Matlab's `drss` function. Further details are found in the caption of Fig. 3.

To quantify the predictive power of the identified models we use the normalized simulation error, defined

$$\tilde{\mathcal{E}}(\theta) \triangleq \frac{\sum_{t=1}^{T} |y_t - g(x_t, u_t)|^2}{\sum_{t=1}^{T} y_t' y_t}, \quad x_{1:T} = \mathcal{X}_T(\theta, \mu, 0)$$

where $\{u_t, y_t\}_{t=1}^{T}$ represents validation data.

Of course, no general conclusions can be drawn from such a study, especially given the simplicity of the model structure. Nonetheless, on average Algorithm 2 performed marginally better, achieving lower $\tilde{\mathcal{E}}(\theta)$ in 57% of the trials. More interestingly, the dense column of scores concentrated along the vertical axis of Fig. 3, comprising approx. 18% of the data points, indicates that there were a number of trials in which Algorithm 2 performed significantly better than the latent states based alternative.

6. CONCLUSION

In this paper, we have proposed a system identification strategy based on a formulation of the EM algorithm over latent system disturbances, rather than latent system states. The main contribution is the application of a Lagrangian relaxation that enables the challenging 'maximization step' to be formulated as a convex optimization problem. Such a formulation was shown to alleviate difficulties related to the identification of models with intractable transition densities, and, in some circumstances, to improve the fidelity of the bound on the likelihood. The proposed strategy was fully developed for identification of LGSS models, and preliminary results from numerical experiments suggest that it could represent a competitive alternative to existing EM based identification methods.

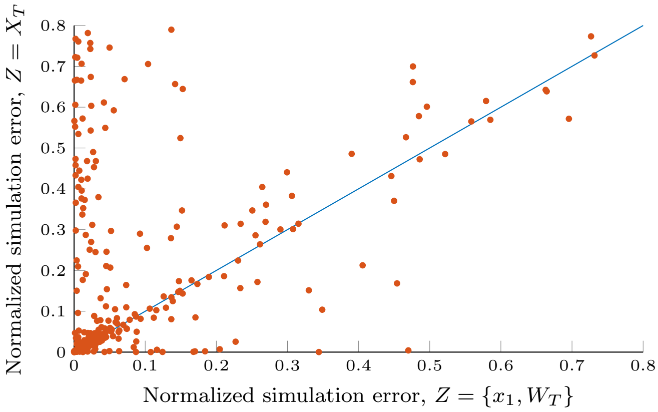

Fig. 3. Comparison of normalized simulation error. The true disturbance and measurement noise covariance was set to $\Sigma_w = 10^{-4}$ and $\Sigma_v = 0.2$. Initial guesses, θ_0, were generated by Matlab's `drss`. $T = 50$ observations were used for identification.

Specialized algorithms for minimization of the Lagrangian relaxation are the subject of current research; preliminary results indicate a dramatic reduction in computation time compared to general purpose SDP solvers, thereby enabling efficient identification of higher order systems.

REFERENCES

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, Cambridge.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1–38.

Durbin, J. and Koopman, S.J. (2012). *Time series analysis by state space methods*. 38. Oxford University Press.

Gibson, S. and Ninness, B. (2005). Robust maximum-likelihood estimation of multivariable dynamic systems. *Automatica*, 41(10), 1667–1682.

Lindsten, F. and Schön, T.B. (2013). Backward simulation methods for Monte Carlo statistical inference. *Foundations and Trends in Machine Learning*, 6(1), 1–143.

Ljung, L. (1999). *System Identification: Theory for the User (2nd Edition)*. Prentice Hall, 2 edition.

Manchester, I., Tobenkin, M.M., and Megretski, A. (2012). Stable nonlinear system identification: Convexity, model class, and consistency. In *Proceedings of the 16th IFAC Symposium on System Identification (SYSID)*. Brussels, Belgium.

Megretski, A. (2008). Convex optimization in robust identification of nonlinear feedback. *2008 47th IEEE Conference on Decision and Control*, 1370–1374.

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

Solo, V. (2003). An EM algorithm for singular state space models. In *Decision and Control, 2003. Proceedings. 42nd IEEE Conference on*, volume 4, 3457–3460. IEEE.

Tobenkin, M.M., Manchester, I.R., Wang, J., Megretski, A., and Tedrake, R. (2010). Convex optimization in identification of stable non-linear state space models. *49th IEEE Conference on Decision and Control (CDC)*, (4), 7232–7237.