Parsimonious Bayesian Filtering in Markov Jump Systems with Applications to Networked Control

Alexandre Rodrigues Mesquita

***Abstract*—We consider the problem of controlling the precision of the multiple-model multiple-hypothesis filter with Gaussian mixture reduction. The controller adaptively chooses the number of hypotheses kept by the filter to (sub-)optimally seek a tradeoff between filter precision and computational effort. In order to quantify the approximation error due to hypotheses truncation, the controller employs probability divergence measures such as f-divergences and the Wasserstein divergence. The proposed solution is tested on the problem of estimating the states of a networked control system with packet drops on the controller-actuator channel. Theoretical results demonstrate that our strategy leads to a divergence between the true Bayes posterior and the truncated one that remains bounded over time. Numerical results show a good improvement with respect to truncation with a constant number of hypotheses, specially as the number of modes increases and so does the problem dimensionality.**

I. Introduction

A major challenge in the state estimation of hybrid dynamical systems from a Bayesian approach lies in the exponential growth over time of the number of possible continuous state trajectories. This is of particular relevance for Markov Jump Systems (MJSs) since, in the linear case, the Bayes posterior may be computed in closed form. Solving this problem exactly, however, would require a bank of Kalman filters with exponentially growing size over time. To cope with this problem, the multiple model multiple hypothesis filter (M^{3}H) was proposed in [1], [2]. Given that the Bayes posterior in this case is a probability mixture, the M^{3}H truncates and merges the components of this mixture taking into account the history of the discrete states associated to each component.

In order to also incorporate continuous state information in the merging process, the multiple model multiple hypothesis filter with Gaussian mixture reduction (M^{3}HR) was proposed in [3]. This approach merges the mixture components using clustering techniques discussed in [4]. The number of clusters in [3] was pre-selected and kept constant over time. However, by adaptively choosing the number of clusters during the filter operation, one could obtain suitable combinations of estimation error and processing time.

If one considers the possibility of varying the number of clusters with time, we see that, encrusted in the problem of Bayesian filtering of hybrid systems, there is a problem of filter precision control. More precisely, we have an optimal control problem in which one wants to minimize both the time-averaged estimation error and the average computational effort per time-step.

In this work we formulate and solve such a control problem employing different probability measure divergences to allow us to quantify the estimation error. Essential to this formulation is the possibility of aggregating approximation errors made at different times. To this purpose, we use an equivalent of the law of cosines in Euclidean space, to aggregate errors in probability space in a fashion that is less conservative than simply applying the triangle inequality.

This precision control is then applied to the M^{3}HR filter in the fashion of the Runnalls' algorithm [5], which was the most time-efficient clustering algorithm tested in [3]. Numerical results demonstrate reasonable improvement in comparison to the open-loop approach.

As for probability divergences, we study both f-divergences, which take into account only the information content of each distribution regardless of the state space metric, and the Wasserstein distance, which takes into account the state space metric.

This problem is fundamentally different from the standard clustering problem as the latter is static and is not performed in real-time. Because truncation errors may expand over time, a poor choice of divergence or of controller may lead to unbounded aggregated errors in the long run.

It is also fair to notice the distinction between the problem we propose and that of minimizing the estimation error subject to a fixed computational deadline (equal to the sample time for example) at each time-step. The solution to the latter is trivial as the controller should just keep computing until the deadline is reached. Instead we care about the average computational time and do not impose a bound to each time-step. This is motivated by the fact that typical filter computations achieve reasonable precision much earlier than sample times and, therefore, computational time should be constrained due to CPU power conservation and not by the sampling period.

Although the idea of filter precision control is not completely new (see, for example, [6]), it is is new in the context of Bayesian filtering of hybrid systems where computational effort is the control input. Our main contribution, in Section III, is to extend the results in [3] to an optimal control framework that provides formal bounds to the average costs associated with filter precision and computational effort. Other contributions include Theorem 1, which is of general importance to Information Theory and to clustering, as it allows measuring of approximation errors; Theorem 8, which gives new bounds and fast approximations of Wasserstein distances; and Proposition 9, which establishes a promising link between

The author was with the Department of Electronic Engineering, Federal University of Minas Gerais, Belo Horizonte, MG, 31270-901 Brazil e-mail: amesquita12@ufmg.br

This work was partially supported by the project INCT-INSAC under grants CNPq 465755/2014-3, FAPESP 2014/50851-0

Information Theory and Control Theory.

In the next section, we motivate our problem with an example from networked control.

II. A Problem in Networked Control

A common challenge in networked control systems lies in the loss of data packets due to channel noise or channel interference (see [7] for a review of this issue). Packet loss events may be modeled by Markov chains whose transitions are independent on the actual information content of packets. Thus, a control system whose sensors, controllers or actuators are connected by a packet dropping network is a standard example of a Markov Jump System.

In this work we consider the problem of drops in the controller-actuator channel (see [7] for the problem of drops in the sensor-controller channel). Let $x_k \in \mathbb{R}^d$ be the state of a linear system with dynamics given by

$$\begin{aligned} x_{k+1} &= Ax_k + \epsilon_k B u_k + w_k && (1)\\ y_k &= Cx_k + v_k \ , && (2) \end{aligned}$$

where $y_k \in \mathbb{R}^{n_o}$ are the observations corrupted by white Gaussian noise v_k with covariance R_v, $u_k \in \mathbb{R}^{n_i}$ is the controller input and the disturbance w_k is white Gaussian noise, which is independent of v_k and has covariance R_w. The process $\epsilon_k \in \{0,1\}$ accounts for packet drops in the controller-actuator channel and it is modeled by the discrete Hidden Markov Model (ϵ_k, m_k) characterized by

$$\begin{aligned} \Pr\{m_{k+1} = j | m_k = i\} &= \pi_{j|i} && (3)\\ \Pr\{\epsilon_k = j | m_k = i\} &= \varrho_{j|i} && (4) \end{aligned}$$

where $[\pi_{j|i}]$ and $[\varrho_{j|i}]$ define the transition and emission matrices respectively and where the discrete state m_k lies in the set $\{1,\ldots,M\}$.

It is assumed that the controller only has knowledge of the sequence $y_{1:k}$, not observing ϵ_k or m_k directly. Had the controller knowledge of ϵ_k, the optimal state estimator would be a simple Kalman filter.

The Bayes approach to this problem would be to consider all possible sequences $\epsilon_{1:k}$, obtain the posteriors $p(x_k|\epsilon_{1:k}, y_{1:k})$ given by the respective Kalman filters and then weight each posterior according to its likelihood. Unfortunately, the number of possible sequences $\epsilon_{1:k}$ (and of Kalman filters) grows exponentially as 2^k. That is why any Bayesian approach to filtering MJSs needs truncation.

To make it more precise, let $x_{k|k}$ denote the posterior estimates of x_k when (m_0, x_0) is distributed with priors $\pi_{m_0}\phi_{m_0}(x_0)$, where $\phi_{m_0} = \mathcal{N}(\mu_{m_0}, \Sigma_{m_0})$ and $\mathcal{N}$ denotes the multivariate normal distribution with given mean and covariance matrix. For prior $m_0 = i$, define the likelihood function of the output sequence $y_{1:k}$ and the n-th possible mode sequence $\epsilon^{(n)}_{1:k}$, $n = 1,\ldots,2^k$, as

$$\ell_{i,k,n} := \int p\left(y_{1:k}, \epsilon^{(n)}_{1:k} | m_0, x_0\right) \phi_i(x_0) dx_0 \ .$$

Denote by $\mu_{i,k,n}$ the posterior means at time k given by the Kalman filter corresponding to the n-th emission sequence and to prior $m_0 = i$.

Then, by the hidden Markov structure of the process (see [8], [9] for a review of Bayesian filtering), the posterior means are given by the sum of the means for the continuous filters weighted by the posterior probability for each component:

$$x_{k|k} = \sum_{i,n} \frac{\pi_i \ell_{i,k,n}}{\ell_k} \mu_{i,k,n} \ ,$$

where $\ell_k = \sum_{i,n} \pi_i \ell_{i,k,n}$.

In our experiments, we focus on the particular case of memoryless erasure channels, where

$$[\pi_{ij}] = \begin{bmatrix} 1-p_0 & p_0 \\ 1-p_0 & p_0 \end{bmatrix} \quad \text{and} \quad [\varrho_{ij}] = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \ ,$$

such that $m = 1$ always corresponds to a successful transmission and $m = 2$ corresponds to a drop and the drop probability is given by the number p_0. In this case there is no real distinction between the mode variable m_k and the emission variable ϵ_k.

III. A Framework for Precision Control

In this section we compute bounds for the approximation error due to successive truncations of the probability densities in a Bayesian filter. These bounds are then used to propose suboptimal control strategies that trade off computational time and filter precision.

Runnalls' algorithm, which is employed by the M^{3}HR filter, works by recursively merging mixture components pairwise. As a criterion to select the pairs to be merged, it computes a bound on the Kullback-Leibler divergence between the original mixture and the reduced one. The Gaussian mixture composed of this pair of Gaussians is then replaced by the Gaussian that preserves the first two moments. In this section, considering more general divergence measures, we want to quantify the overall error induced by successive pairwise merges.

For a given space $\mathcal{P}$ of probability distributions, consider a mixture probability distribution in $\mathcal{P}$ with components (w_i, ν_i), $i = 1,\ldots,N$. Given a *merging function* $\gamma_t : \mathcal{P}\times\mathcal{P} \mapsto \mathcal{P}$, $t \in [0,1]$, define $\bar{\nu}_n$ as the measure obtained from the consecutive pairwise merging of $\nu_1, \nu_2, \ldots, \nu_n$ as follows

$$\bar{\nu}_n = \gamma_{\left(\frac{w_n}{\bar{w}_n}\right)}(\bar{\nu}_{n-1}, \nu_n), \ n \geq 2, \ \bar{\nu}_1 = \nu_1 \ ,$$

where $\bar{w}_n = \sum_{i=1}^n w_i$.

Now, consider a generic divergence function $\mathcal{D} : \mathcal{P}\times\mathcal{P} \mapsto \mathbb{R}_{\geq 0} \cup \{\infty\}$ and assume that $\mathcal{D}$ is jointly convex.

Assumption 1. *For $t \in [0,1]$, ν_1 and $\nu_2 \in \mathcal{P}$, there exists a function $\bar{D}_t : \mathcal{P}\times\mathcal{P} \mapsto \mathbb{R}_{\geq 0} \cup \{\infty\}$ such that*

$$(1-t)\mathcal{D}(\nu_1,\nu) + t\mathcal{D}(\nu_2,\nu) \leq \mathcal{D}(\gamma_t(\nu_1,\nu_2),\nu) + \bar{\mathcal{D}}_t(\nu_1,\nu_2) \ , \tag{5}$$

for all $\nu \in \mathcal{P}$.

Theorem 1. *Suppose that γ_t and $\bar{\mathcal{D}}_t$ satisfy Assumption 1. Then, the total divergence resulting from consecutive pairwise merges is bounded as*

$$\mathcal{D}\left(\sum_{i=1}^N w_i \nu_i, \nu\right) \leq \mathcal{D}(\bar{\nu}_N, \nu) + \sum_{n=2}^N \bar{w}_n \bar{\mathcal{D}}_{\frac{w_n}{\bar{w}_n}}(\bar{\nu}_{n-1}, \nu_n) \ , \tag{6}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2020.2976274, IEEE Transactions on Automatic Control

for all $\nu \in \mathcal{P}$. Consequently, if we denote by Δ_n the bound associated to the approximation error $\mathcal{D}\left(\sum_{i=1}^{n} w_i\nu_i, \bar{\nu}_n\right)$, we have the recurrence

$$\Delta_n = \Delta_{n-1} + \bar{w}_n \bar{\mathcal{D}}_{\frac{w_n}{\bar{w}_n}}(\bar{\nu}_{n-1}, \nu_n) \ . \qquad (7)$$

Proof. From the convexity of $\mathcal{D}$ we have that

$$\mathcal{D}\left(\sum_{i=1}^{N} w_i\nu_i, \nu\right) \leq w_1\mathcal{D}(\nu_1, \nu) + (1-w_1)\mathcal{D}\left(\sum_{i=2}^{N} \frac{w_i}{1-w_1}\nu_i, \nu\right) \ . \qquad (8)$$

Next, we note that

$$\begin{aligned}
&\bar{w}_{n-1}\mathcal{D}(\bar{\nu}_{n-1}, \nu) + (1-\bar{w}_{n-1})\mathcal{D}\left(\sum_{i=n}^{N} \frac{w_i}{1-\bar{w}_{n-1}}\nu_i, \nu\right) \\
&\quad \leq \bar{w}_{n-1}\mathcal{D}(\bar{\nu}_{n-1}, \nu) + w_n\mathcal{D}(\nu_n, \nu) \\
&\qquad + (1-\bar{w}_n)\mathcal{D}\left(\sum_{i=n+1}^{N} \frac{w_i}{1-\bar{w}_n}\nu_i, \nu\right) \\
&\quad = \bar{w}_n\left(\left(1-\frac{w_n}{\bar{w}_n}\right)\mathcal{D}(\bar{\nu}_{n-1}, \nu) + \frac{w_n}{\bar{w}_n}\mathcal{D}(\nu_n, \nu)\right) \qquad (9) \\
&\qquad + (1-\bar{w}_n)\mathcal{D}\left(\sum_{i=n+1}^{N} \frac{w_i}{1-\bar{w}_n}\nu_i, \nu\right) \\
&\quad \leq \bar{w}_n\bar{\mathcal{D}}_{\frac{w_n}{\bar{w}_n}}(\bar{\nu}_{n-1}, \nu_n) + \bar{w}_n\mathcal{D}(\bar{\nu}_n, \nu) \\
&\qquad + (1-\bar{w}_n)\mathcal{D}\left(\sum_{i=n+1}^{N} \frac{w_i}{1-\bar{w}_n}\nu_i, \nu\right) \ ,
\end{aligned}$$

where the first inequality follows from the convexity of $\mathcal{D}$ and the second is a consequence of (5). Applying inequality (9) successively starting from (8) gives (6). Replacing ν by $\bar{\nu}_n$ in (6) gives the second part of the theorem as stated in (7).

□

Remark 1. *Note that, if we replaced $\mathcal{D}$ in (5) by the Euclidean distance squared, $\bar{\mathcal{D}}_t$ by $t(1-t)\|x-y\|^2$ and make $\gamma_t(x,y) = (1-t)x+ty$, we would have that (5) is satisfied with equality. This identity is equivalent to the law of cosines and it gives much tighter error bounds than the triangle inequality.*

Let $\nu^{(k)}$ and $\bar{\nu}^{(k)}$ be the posterior distributions for the Bayes filter at time k having different priors $\nu^{(0)}$ and $\bar{\nu}^{(0)}$. Assume the divergence $\mathcal{D}$ admits a contraction rate $\alpha \in (0,1)$, i.e.,

$$\mathcal{D}(\nu^{(k+1)}, \bar{\nu}^{(k+1)}) \leq \alpha\mathcal{D}(\nu^{(k)}, \bar{\nu}^{(k)}), \ \forall k \geq 0, \ \nu^{(0)}, \bar{\nu}^{(0)} \in \mathcal{P} \ .$$

Then, if we denote by $\mathcal{E}_k$ the bound on the truncation error accumulated from all times previous to k, (7) gives that

$$\mathcal{E}_k = \alpha\mathcal{E}_{k-1} + \alpha\Delta^{(k-1)} \ , \qquad (10)$$

where $\Delta^{(k)}$ is the bound on the total truncation error at time k obtained from (7) combining all clusters:

$$\Delta^{(k)} = \sum_{\text{all } j \text{ clusters}} \hat{w}_j\Delta_j^{(k)} \ ,$$

where $\hat{w}_j$ is the total probability mass of the j-th cluster and $\Delta_j^{(k)}$ is the truncation error for that same cluster.

This evolution of the truncation error suggests the formulation of the control problem as a Markov Decision Process (MDP) where the decision variable is the number of components to keep at each time-step and where the cost to be minimized is a function of the truncation error and the computational effort. This would be a MDP with *state* $\mathcal{E}_k$ and with *actions* $N_{k,m}$, defined as the number of components of the reduced measure for mode m at time k. The *instantaneous cost* would be $c(\mathcal{E}_k, N_k) = \mathcal{E}_k + \beta\tau(N_k)$, for some weight $\beta > 0$ and some function $\tau(\cdot)$ that describes the impact of the action vector $N_k = [N_{k,m}]$ on the computational effort. Then, an optimal solution to the MDP is a policy that picks N_k as a function of $\mathcal{E}_k$ in order to minimize the discounted cost

$$\sum_{k=1}^{\infty} \gamma^k \,\mathrm{E}[c(\mathcal{E}_k, N_k)] \ ,$$

where $\gamma \in (0,1)$ is the discount factor.

Next we derive a rollout policy that suboptimally solves this MDP (refer to [10] for an introduction to rollout policies and approximate dynamic programming). We start with a policy θ_0 characterized by the action N_k being constant over time. The value function $V_{\theta_0}(\cdot)$ associated to this policy satisfies the Bellman equation

$$\begin{aligned}
V_{\theta_0}(\mathcal{E}_k) &= \mathrm{E}\left[c(\mathcal{E}_k, N_k) + \gamma V_{\theta_0}(\mathcal{E}_{k+1})\right] \\
&= \mathcal{E}_k + \beta\tau(N_k) + \gamma\,\mathrm{E}[V_{\theta_0}(\alpha\mathcal{E}_k + \alpha\Delta^{(k)})] \ ,
\end{aligned} \qquad (11)$$

where we used (10) in the last equality. Assume now that $\Delta^{(k)}$ reaches an ergodic limit such that $\mathrm{E}[\Delta^{(k)}]$ is constant (this assumption is justified in Section V-C). Then, we can check that $V_{\theta_0}(\mathcal{E}_k) = \mathcal{E}_k/(1-\gamma\alpha) + \eta_0$ solves (11) for some constant η_0. A rollout policy θ_1 is now defined by picking the actions that minimize the total cost predicted by V_{θ_0} at each time-step:

$$\begin{aligned}
N_k &= \arg\min_{N_k} \mathrm{E}\left[c(\mathcal{E}_k, N_k) + \gamma V_{\theta_0}(\mathcal{E}_{k+1})\right] \\
&= \arg\min_{N_k} \frac{\gamma\alpha}{1-\gamma\alpha}\Delta^{(k)} + \beta\tau(N_k) \ .
\end{aligned} \qquad (12)$$

Rollout policies guarantee that the total cost is upper-bounded by $V_{\theta_0}(\mathcal{E}_0)$, but in practice they give much smaller costs. Computing the minimum in (12) would by itself affect the computational time $\tau(N_k)$ if this is to be done online. Instead, with the help of Theorem 1, we can check for a local minimum by looking at the first difference with respect to N_k:

$$\frac{\gamma\alpha}{1-\gamma\alpha}\bar{w}_n\bar{\mathcal{D}}_{\frac{w_n}{\bar{w}_n}}(\bar{\nu}_{n-1}, \nu_n) + \beta(\tau(N_k - \delta_m) - \tau(N_k)) \ ,$$

where n is such that the merge of $\bar{\nu}_{n-1}$ and ν_n would lead to $N_{k,m} - 1$ components; δ_m is the indicator vector at m. This leads to a threshold condition according to which we should truncate one component at a time and stop when the error introduced by the next truncation satisfies

$$\bar{w}_n\bar{\mathcal{D}}_{\frac{w_n}{\bar{w}_n}}(\bar{\nu}_{n-1}, \nu_n) > \frac{1-\gamma\alpha}{\gamma\alpha}\beta(\tau(N_k) - \tau(N_k - \delta_m)) \ . \qquad (13)$$

In words, one should stop truncating when the incremental truncation error becomes larger than a constant times the expected decrease in computational effort. The above condition does not guarantee that the number of components will remain bounded for all time. For this reason it is desirable to add to the stopping criterion the condition that $\sum_m N_{k,m} \leq N_{\max}$, for some constant $N_{\max}$ large enough. Taking into account all these considerations, the proposed strategy is summarized in Algorithm 1.

Algorithm 1 performs a Bayes filtering step, consisting of a bank of Kalman filters, and truncates the number of hypotheses using our suboptimal policy instead of Runnalls' algorithm, which was used in the M^{3}HR filter. The number of hypotheses (filters) increases M-fold at every Bayes step (line 3). Truncation starts by computing the truncation error (left-hand side of (13)) associated to all possible two by two merges and storing them in $d_{i,j,m}$ (line 4). Then we select the merge that provides minimal truncation error (line 7) and check whether this error satisfies (13) (line 8). If so, the algorithm is set to stop as soon as the number of components is less than or equal to $N_{\max}$. Next, the merge is performed (lines 14-17), the d matrix is updated (line 21) and a new merge cycle is started returning to line 6.

Algorithm 1 M^{3}H Filtering with Suboptimal Gaussian Mixture Model Reduction

```
1:  Given the posterior pdf for $(x_0, m_0)$ defined by a mixture
        $$\tilde{\nu} = \sum_{m=1}^{M} \sum_{i=1}^{\mathsf{N}_m} w_{i,m}\nu_{i,m},$$
    and given $\kappa_0$ constant, $N_{\max}$ integer and $k = 1$,
2:  Get observation $y_k$ and input $u_k$,
3:  Get the new posterior pdf:
        $$([w_{i,m}], [\nu_{i,m}]) = \mathsf{BayesFilter}\,(y_k, u_k, [w_{i,m}], [\nu_{i,m}])\,,$$
4:  Compute the truncation error for the $i, j$ merge:
        $$d_{i,j,m} = (w_{i,m} + w_{j,m})\bar{\mathcal{D}}_{w_{j,m}/(w_{i,m}+w_{j,m})}(\nu_{i,m}, \nu_{j,m})$$
    for all $i < j$ and all $m$.
5:  Set StopFlag=FALSE.
6:  while $\sum_m \mathsf{N}_m > M$ do
7:      Find the indices $i^* < j^*$ and $m^*$ that minimize $d_{i,j,m}$.
8:      if $d_{i^*,j^*,m^*} > \kappa_0[\tau([\mathsf{N}_m]) - \tau([\mathsf{N}_m] - \delta_{m^*})]$ then
9:          Set StopFlag=TRUE.
10:         if $\sum_m \mathsf{N}_m \leq N_{\max}$ then
11:             break
12:         end if
13:     end if
14:     Set $w_{i^*,m^*} = w_{i^*,m^*} + w_{j^*,m^*}$.
15:     Set $\nu_{i^*,m^*} = \gamma_{w_{j^*,m^*}/w_{i^*,m^*}}(\nu_{i^*,m^*}, \nu_{j^*,m^*})$.
16:     Remove component $j^*$ from the mixture of index $m^*$.
17:     Set $\mathsf{N}_{m^*} = \mathsf{N}_{m^*} - 1$.
18:     if $\sum_m \mathsf{N}_m \leq N_{\max}$ & StopFlag then
19:         break
20:     end if
21:     Update $d_{i^*,j,m^*}$ for $j > i^*$.
22: end while
23: Increment $k$ and return to step 2.
    { Actual calculations use $\ln(w_{i,m})$ to avoid issues with multiplication precision.}
```

The knowledge of the contraction rate α is actually not needed. Given that a rate α exists, we can experimentally try different constants $\kappa_0 > 0$ in Algorithm 1 and pick one that is suitable. This is the equivalent of the user choosing the weight β since, for every $\kappa_0 > 0$, there exists β such that $\kappa_0 = \beta((\gamma\alpha)^{-1} - 1)$ as in (13).

Note in addition that, even when $\alpha \geq 1$ and there is no contraction effectively, the above framework still works for small enough discount factors ($\gamma < \alpha^{-1}$).

The computational time due to the filtering step is linear in N_m since at most $M \sum_m \mathsf{N}_m$ Kalman filters are run after we reduce each mixture to a size of N_m. Thus, the mixture reduction step, which is quadratic in N_m as seen in Algorithm 1, dominates the computational time. From this, we have that the function $\tau(\cdot)$ can be obtained empirically by fitting a second order polynomial in N_m to the computational times.

A more precise, closed-form, structure on $\tau(\cdot)$ can be obtained as follows. Suppose each mixture is reduced to size $N_{k-1,m}$ at time $k-1$. After propagation, each mode will have at most $\sum_m N_{k-1,m} =: \bar{N}_k$ components. From this, line 4 at time k in Algorithm 1 takes time proportional to $M\bar{N}_k(\bar{N}_k - 1)/2$. If we were to reduce each mixture to the minimum size of 1, line 21 in Algorithm 1 would take time proportional to $M(\bar{N}_k - 1)(\bar{N}_k - 2)/2$. However, reducing to $N_{k,m}$ components instead of 1, we save $N_{k,m}(N_{k,m} - 1)/2$ updates in the array $d_{i,j,m}$. Summing the three contributions above from lines 4 and 21 and also that from the Kalman filter, the computational time at time k is proportional to:

$$M(\bar{N}_k - 1)^2 - \sum_{m=1}^{M} \frac{N_{k,m}(N_{k,m} - 1)}{2} + M\tau_0\bar{N}_k \ , \tag{14}$$

where the constant τ_0 corresponds to the computational time of the Kalman filters normalized by the time to compute divergences. The expression above is a function of both $N_{k,m}$ and $N_{k-1,m}$. Taking into account the discount factor, we can rearrange the terms in the total computational cost to isolate the contributions from $[N_{m,k}]$ and obtain

$$\tau([N_{m,k}]) \propto 2\gamma M \left[\left(\sum_{m=1}^{M} N_{k,m} - 1 \right)^2 + \tau_0 \sum_{m=1}^{M} N_{k,m} \right] - \sum_{m=1}^{M} N_{k,m}(N_{k,m} - 1) \ .$$

Remark 2. *Finding an exact value function and such a simple control was possible due to the linear dynamics in (10), which is a consequence of Theorem 1. The same would not be possible if errors were aggregated using the triangle inequality.*

Remark 3. *The given controller is suboptimal in a number of ways. In the first place, we are dealing with upper bounds on error sizes and not the real errors. Secondly, $\mathcal{E}_k$ is not a real state since it does not fully describe the full probability densities. Third, our model does not take into account how the mixture sizes $N_{k.m}$ influence the range of approximation errors at future times. Lastly, we have merely provided a rollout policy and, on top of that, we have no guarantee that (13) gives the global minimum.*

In the next sections we discuss different types of divergences that can be employed with the presented framework.

IV. Precision Control Using f-Divergences

An important class of convex divergences is given by the so-called f-divergences. For a convex function f such that $f(1) = 0$, the f-divergence D_f of the probability measures ν_1 with respect to ν_2 is defined as

$$D_f(\nu_1\|\nu_2) = \int f\left(\frac{d\nu_1}{d\nu_2}\right)\, d\nu_2$$

when ν_1 is absolutely continuous with respect to ν_2 (see [11] for a definition in the general case and for further properties). Due to the convexity of the map $(x, y) \mapsto xf(y/x)$, D_f is jointly convex on (ν_1, ν_2).

Further properties of f-divergences are $D_f(\nu_1\|\nu_2) \geq 0$ and, if f is strictly convex at 1, $D_f(\nu_1\|\nu_2) = 0$ if and only if $\nu_1 = \nu_2$. If P is a Markov transition operator, then $D_f(\nu_1\|\nu_2) \geq D_f(P\nu_1\|P\nu_2)$, which means that f-divergences are non-expansive under the time evolution of dynamical systems. This implies that f-divergences tend to contract (or at least not expand) during the propagation step of a Bayes filter. However, they still may expand during the Bayes step when additional information is added through observation.

Some notorious divergences in probability theory are f-divergences. For $f(t) = |t-1|$, we have the total variation distance $\texttt{TV}(\cdot,\cdot) := D_f(\cdot\|\cdot)$. For $f(t) = t\ln t$, we have the Kullback-Leibler divergence $\texttt{KL}(\cdot,\cdot) := D_f(\cdot\|\cdot)$. For $f(t) = -\ln t$ we have the reverse Kullback-Leibler divergence $\texttt{RKL}(\cdot,\cdot) := D_f(\cdot\|\cdot)$. For $f(t) = \frac{1}{2}(\sqrt{t}-1)^2$, we have the squared Hellinger distance $\mathcal{H}^2(\cdot,\cdot) := D_f(\cdot\|\cdot)$. And, for $f(t) = (t-1)^2$, we have the chi-squared divergence $\chi^2(\cdot,\cdot) := D_f(\cdot\|\cdot)$. From the above divergences, only $\texttt{TV}$ and $\mathcal{H}^2$ are symmetric. In addition, $\texttt{TV}$ and $\mathcal{H}$ are true distances.

The optimal values for (γ_t, ν) in (5) can be defined by means of a $\mathrm{min\text{-}max}$ problem. When a Nash-equilibrium (γ_t^*, ν^*) exists, it is always the case that $\gamma_t^* = \nu^*$. Indeed, given a choice $\nu = \nu^*$, the bound $\bar{\mathcal{D}}$ is minimized by setting $\gamma_t = \nu^*$. For this reason, γ_t^* often coincides with the barycenter

$$\nu^* = \arg\min_{\nu} (1-t)\mathcal{D}(\nu_1, \nu) + t\mathcal{D}(\nu_2, \nu) \ .$$

When $\mathcal{D} = D_f$, [12] showed that the solution ν^* to this problem (the so-called entropic means) are as such: ν^* is the arithmetic mean of the pdfs in the case of $\mathcal{D} = \texttt{KL}$; ν^* is the normalized geometric mean of the pdfs in the case of $\mathcal{D} = \texttt{RKL}$; ν^* is the normalized mean of square-roots of the pdfs in the case of $\mathcal{D} = \mathcal{H}^2$; and ν^* is the normalized harmonic mean of the pdfs in the case of $\mathcal{D}$ being the χ^2-divergence.

In our case, we are interested in the approximation of Gaussian mixtures by a single Gaussian. From the means above, only the normalized geometric mean of Gaussians is again a Gaussian.

In the following proposition we give merging functions and bounds $\bar{\mathcal{D}}_t$ that satisfy condition (5) when $\mathcal{P}$ is the space of multivariate normal distributions on $\mathbb{R}^d$, denoted here by $\mathcal{N}^d$.

Proposition 2. *Suppose $\nu_1 = \mathcal{N}(\mu_1, \Sigma_1)$ and $\nu_2 = \mathcal{N}(\mu_2, \Sigma_2)$ are merged by $\gamma_t(\nu_1, \nu_2) = \mathcal{N}(\bar{\mu}_t, \bar{\Sigma}_t)$. Then, the tuple $(\bar{\mu}_t, \bar{\Sigma}_t, \bar{\mathcal{D}}_t)$ satisfies condition (5) when $\mathcal{P} = \mathcal{N}^d$ and the following f-divergences are used as $\mathcal{D} = D_f$:*

i. For the total variation distance:

$$\begin{aligned}\bar{\mu}_t &= \mu_1\\ \bar{\Sigma}_t &= \Sigma_1\\ \bar{\mathcal{D}}_t &= t\,TV(\nu_1, \nu_2) \ ,\end{aligned}$$

when $t < 0.5$ and vice-versa when $t > 0.5$;

ii. for the Kullback-Leibler divergence:

$$\begin{aligned}\bar{\mu}_t &= (1-t)\mu_1 + t\mu_2\\ \bar{\Sigma}_t &= (1-t)\Sigma_1 + t\Sigma_2 + t(1-t)(\mu_1-\mu_2)(\mu_1-\mu_2)' \quad (15)\\ \bar{\mathcal{D}}_t &= \frac{1}{2}\left(\ln|\bar{\Sigma}_t| - (1-t)\ln|\Sigma_1| - t\ln|\Sigma_2|\right) \ ;\end{aligned}$$

iii. for the reverse Kullback-Leibler divergence:

$$\begin{aligned}\bar{\mu}_t =& \bar{\Sigma}_t\left((1-t)\Sigma_1^{-1}\mu_1 + t\Sigma_2^{-1}\mu_2\right)\\ \bar{\Sigma}_t =& \left((1-t)\Sigma_1^{-1} + t\Sigma_2^{-1}\right)^{-1}\\ \bar{\mathcal{D}}_t =& \frac{1}{2}\Big(t(1-t)(\mu_1-\mu_2)'\tilde{\Sigma}_t^{-1}(\mu_1-\mu_2)\\ & - \ln|\bar{\Sigma}_t| + (1-t)\ln|\Sigma_1| + t\ln|\Sigma_2|\Big) \ ,\end{aligned}$$

where $\tilde{\Sigma}_t = t\Sigma_1 + (1-t)\Sigma_2$;

iv. for the squared Hellinger distance:

$$\begin{aligned}\bar{\mu}_t =& \bar{\Sigma}_t\left((1-t)\Sigma_1^{-1}\mu_1 + t\Sigma_2^{-1}\mu_2\right)\\ \bar{\Sigma}_t =& \left((1-t)\Sigma_1^{-1} + t\Sigma_2^{-1}\right)^{-1} - \epsilon I, \ \epsilon > 0\\ \bar{\mathcal{D}}_{B,t} =& \frac{1}{4}\Big(t(1-t)(\mu_1-\mu_2)'\tilde{\Sigma}_t^{-1}(\mu_1-\mu_2)\\ & - \ln|\bar{\Sigma}_t| + (1-t)\ln|\Sigma_1| + t\ln|\Sigma_2|\Big)\\ \bar{\mathcal{D}}_t =& 1 - e^{-\bar{\mathcal{D}}_{B,t}} \ ,\end{aligned}$$

where $\tilde{\Sigma}_t = t\Sigma_1 + (1-t)\Sigma_2$.

The expressions for merging for the Kullback-Leibler and the reverse Kullback-Leibler divergences are optimal as demonstrated in [13]–[15]. In the case of the squared Hellinger distance, there is no closed form for the optimal merge (see the related problem of computing the Bhattacharyya centroid in [16]).

Proof of Propostion 2 (iv). The bound $\bar{\mathcal{D}}_{B,t}$ is a bound obtained for the Bhattacharyya distance $\mathcal{D}_B := -\ln(1-\mathcal{H}^2)$. From [16], we have that

$$\begin{aligned}\mathcal{D}_B(\nu_1, \nu_2) = & \frac{1}{4}(\mu_1-\mu_2)'(\Sigma_1+\Sigma_2)^{-1}(\mu_1-\mu_2)\\ & + \frac{1}{2}\ln\frac{|(\Sigma_1+\Sigma_2)/2|}{|\Sigma_1|^{1/2}|\Sigma_2|^{1/2}} \ . \qquad (16)\end{aligned}$$

Let us consider the $\mathrm{min\text{-}max}$ problem for the map $\varphi : (\bar{\mu}_t, \bar{\Sigma}_t, \mu, \Sigma) \mapsto (1-t)\mathcal{D}_B(\nu_1, \nu) + t\mathcal{D}_B(\nu_2, \nu) - \mathcal{D}_B(\bar{\nu}_t, \nu)$, where $\nu = \mathcal{N}(\mu, \Sigma)$. Its derivative on μ is

$$\begin{aligned}\varphi_\mu = & \frac{1-t}{2}(\Sigma_1+\Sigma)^{-1}(\mu-\mu_1)\\ & + \frac{t}{2}(\Sigma_2+\Sigma)^{-1}(\mu-\mu_2) - \frac{1}{2}(\bar{\Sigma}_t+\Sigma)^{-1}(\mu-\bar{\mu}_t)\end{aligned}$$

and the derivative on $\bar{\mu}_t$ is $\varphi_{\bar{\mu}_t} = 1/2(\bar{\Sigma}_t + \Sigma)^{-1}(\mu - \bar{\mu}_t)$. Equating both derivatives to zero we obtain

$$\begin{aligned}\mu &= \bar{\mu}_t \\ &= \left[(1-t)(\Sigma_1+\Sigma)^{-1} + t(\Sigma_2+\Sigma)^{-1}\right]^{-1} \\ &\quad \cdot \left((1-t)(\Sigma_1+\Sigma)^{-1}\mu_1 + t(\Sigma_2+\Sigma)^{-1}\mu_2\right) \ .\end{aligned} \tag{17}$$

As long as the Hessian

$$\varphi_{\mu\mu} = \frac{1-t}{2}(\Sigma_1+\Sigma)^{-1} + \frac{t}{2}(\Sigma_2+\Sigma)^{-1} - \frac{1}{2}(\bar{\Sigma}_t+\Sigma)^{-1}$$

is negative definite, the value of μ that maximizes φ is given by (17). This is indeed the case since the concavity of the matrix harmonic mean [17, Thm. 4.1.1] grants that

$$\begin{aligned}&\left[\frac{1-t}{2}(\Sigma_1+\Sigma)^{-1} + \frac{t}{2}(\Sigma_2+\Sigma)^{-1}\right]^{-1} \\ &\quad \geq 2\left[(1-t){\Sigma_1}^{-1} + t{\Sigma_2}^{-1}\right]^{-1} + 2\Sigma \\ &\quad = 2\bar{\Sigma}_t + 2\epsilon I + 2\Sigma > 2\bar{\Sigma}_t + 2\Sigma\end{aligned}$$

and $\varphi_{\mu\mu} < 0$.

Now, the derivative of φ on Σ is

$$\begin{aligned}&-\frac{1-t}{4}(\Sigma_1+\Sigma)^{-1}(\mu-\mu_1)(\mu-\mu_1)'(\Sigma_1+\Sigma)^{-1} \\ &\quad -\frac{t}{4}(\Sigma_2+\Sigma)^{-1}(\mu-\mu_2)(\mu-\mu_2)'(\Sigma_2+\Sigma)^{-1} \\ &\quad +\frac{1}{4}(\bar{\Sigma}_t+\Sigma)^{-1}(\mu-\bar{\mu}_t)(\mu-\bar{\mu}_t)'(\bar{\Sigma}_t+\Sigma)^{-1} \\ &\quad +\frac{1-t}{2}(\Sigma_1+\Sigma)^{-1} + \frac{t}{2}(\Sigma_2+\Sigma)^{-1} - \frac{1}{2}(\bar{\Sigma}_t+\Sigma)^{-1} \ .\end{aligned}$$

Replacing μ as in (17), this derivative equals

$$-\frac{t(1-t)}{4}(\tilde{\Sigma}_t+\Sigma)^{-1}(\mu_2-\mu_1)(\mu_2-\mu_1)'(\tilde{\Sigma}_t+\Sigma)^{-1} + \varphi_{\mu\mu} \ .$$

Then, this derivative is negative definite and the maximum of φ is achieved by $\Sigma = 0$. Replacing $\mu = \bar{\mu}_t$ and $\Sigma = 0$ in the expressions for φ and $\mathcal{D}_B$, we find that $\varphi \leq \bar{\mathcal{D}}_{B,t}$.

Given this bound, from the convexity of the exponential and the definition of $\mathcal{D}_B$, we have that

$$\begin{aligned}e^{-\bar{\mathcal{D}}_{B,t}} &\leq e^{-\varphi} \\ &\leq \frac{(1-t)(1-\mathcal{H}^2(\nu_1,\nu)) + t(1-\mathcal{H}^2(\nu_2,\nu))}{1-\mathcal{H}^2(\nu,\bar{\nu}_t)} \ .\end{aligned}$$

Rearranging this inequality, we find that

$$\begin{aligned}&(1-t)\mathcal{H}^2(\nu_1,\nu) + t\mathcal{H}^2(\nu_2,\nu) \\ &\quad \leq \mathcal{H}^2(\nu,\bar{\nu}_t)e^{-\bar{\mathcal{D}}_{B,t}} + 1 - e^{-\bar{\mathcal{D}}_{B,t}} \leq \mathcal{H}^2(\nu,\bar{\nu}_t) + 1 - e^{-\bar{\mathcal{D}}_{B,t}} \ .\end{aligned}$$

□

All of the divergences in the proposition have a similar behavior when approaching zero. In particular, if we are at equilibrium with $\bar{\Sigma} = \Sigma_1 = \Sigma_2$, then, in the limit of small mean deviations, we have

$$\bar{\mathcal{D}}_t \propto \frac{1}{2}t(1-t)(\mu_1-\mu_2)'\bar{\Sigma}^{-1}(\mu_1-\mu_2) \tag{18}$$

in the case of the last three divergences.

Notably, Runnalls' algorithm [5] employs the same merging function and the same error bound as those of the Kullback-Leibler divergence in the proposition without, however, controlling the reduced mixture size.

For future reference, we give the χ^2-divergence [18] between two multivariate normals with $2\Sigma_1 > \Sigma_2$:

$$\begin{aligned}\ln(\chi^2(\nu_1,\nu_2))) &= \frac{1}{2}(2\mu_1-\mu_2)'(2\Sigma_1-\Sigma_2)^{-1}(2\mu_1-\mu_2) \\ &\quad + \frac{1}{2}\ln|2\Sigma_1-\Sigma_2| - \mu_1'\Sigma_1^{-1}\mu_1 \\ &\quad + \frac{1}{2}\mu_2'\Sigma_2^{-1}\mu_2 - \ln|\Sigma_1| + \frac{1}{2}\ln|\Sigma_2| \ .\end{aligned} \tag{19}$$

V. Precision Control using the Wasserstein Distance

The limit behavior in (18) shows that information divergences always weight mean deviations according to the posterior covariance matrix. However, there might be situations in which we want to weight mean components differently, according to some metric of interest in $\mathbb{R}^d$. This case is captured nicely by the so-called Wasserstein distance. In the next sections we give the main facts about this distance and derive suitable merging functions and bounds for it. Related work that also applies the Wasserstein distance in Gaussian mixture reduction is found in [19], [20]. Because our focus is on real-time applications, we derive alternative faster approximate solutions.

In Section V-C, we show how this distance is connected with the mean absolute error for matrix weighted norms $\|\cdot\|_Q$ in $\mathbb{R}^d$. In particular, we find that, in order to control the Q-norm of the error, one must replace the inverse of the equilibrium posterior covariance in (18) by a combination of the form $\bar{\Sigma}^{-1} + f(Q)$.

A. The Wasserstein Distance

We denote by $\mathcal{P}_2(\mathbb{R}^d)$ the space of probability measures on $\mathbb{R}^d$ with finite second moment.

Definition 1. *For $\nu_1, \nu_2 \in \mathcal{P}_2(\mathbb{R}^d)$, we define the Wasserstein distance $\mathcal{W}_2(\nu_1,\nu_2)$ between them as*

$$\begin{aligned}\mathcal{W}_2^2(\nu_1,\nu_2) :=& \inf\left\{\int \|x-y\|^2\nu(dx,dy) : \right. \\ &\qquad \left. \int \nu(x,dy) = \nu_1, \int \nu(dx,y) = \nu_2\right\} \\ =& \inf\left\{\mathrm{E}\left[\|X-Y\|^2\right] : X \sim \nu_1, Y \sim \nu_2\right\} \ .\end{aligned}$$

Intuitively, $\mathcal{W}_2^2$ measures the expected squared distance in $\mathbb{R}^d$ between random variables $X \sim \nu_1$ and $Y \sim \nu_2$ considering the best possible coupling between them. Differently than f-divergences, the Wasserstein distance takes into account the distance on the space where X and Y lie.

Endowed with the distance $\mathcal{W}_2$, $\mathcal{P}_2(\mathbb{R}^d)$ is a metric space. Specifically, $\mathcal{W}_2$ is a metrization of the weak topology in $\mathcal{P}_2(\mathbb{R}^d)$ [21, Thm. 6.9]. The space $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$ is geodesic given that any two probability measures are connected by a minimizing geodesic and, moreover, if one of the measures is absolutely continuous with respect to the Lebesgue measure, this geodesic is unique [21, Cor. 7.22, Cor.7.23].

Proposition 3. *The function $\mathcal{W}_2^2(\cdot,\cdot)$ is jointly convex:*

$$\begin{aligned}&\mathcal{W}_2^2\left(w_1p_1 + w_2p_2, w_1q_1 + w_2q_2\right) \leq \\ &\qquad w_1\mathcal{W}_2^2(p_1,q_1) + w_2\mathcal{W}_2^2(p_2,q_2) \ .\end{aligned} \tag{20}$$

Proof. From Definition 1, there exists a random variable $i \in \{1, 2\}$ and a sequence X_n and Y_n such that $\Pr\{X_n \mid i\} = p_i$, $\Pr\{Y_n \mid i\} = q_i$, $\Pr\{i\} = w_i$ and $\mathrm{E}[\|X_n - Y_n\|^2 \mid i] \xrightarrow{n} \mathcal{W}_2^2(p_i, q_i)$. Then, $X_n \sim w_1 p_1 + w_2 p_2$, $Y_n \sim w_1 q_1 + w_2 q_2$ and $\mathrm{E}[\|X_n - Y_n\|^2] \xrightarrow{n} \sum_i w_i \mathcal{W}_2^2(p_i, q_i)$ and (20) follows from Definition 1 applied to the left hand side. □

Proposition 4 (Sec. 6.2 in [22]; Sec. 2 in [23])**.** *Let $\gamma_{\nu_1,\nu_2}(t)$, $t \in [0, 1]$, be a constant-speed geodesic curve from $\nu_1 \in \mathcal{P}_2(\mathbb{R}^d)$ to $\nu_2 \in \mathcal{P}_2(\mathbb{R}^d)$. Then, $\gamma_{\nu_1,\nu_2}(t)$ is also a barycenter of ν_1 and ν_2:*

$$\gamma_{\nu_1,\nu_2}(t) = \arg \min_{\nu \in \mathcal{P}_2(\mathbb{R}^d)} \{(1-t)\mathcal{W}_2^2(\nu_1, \nu) + t\mathcal{W}_2^2(\nu, \nu_2)\} \ .$$

Moreover, the barycenter is unique when one of the measures is absolutely continuous with respect to the Lebesgue measure.

The next result states the fact that the space $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$ is a positively curved space and allows us to find an upper bound for the merging error that is considerably tighter than the bound that would be obtained by a mere application of the triangle inequality. Indeed, for the case of Dirac measures, the bound below recovers the corresponding error in Euclidean space, which is a direct consequence of the law of cosines.

Lemma 5 (Thm 7.3.2 in [24])**.** *For $w_1 \in [0, 1]$, $w_2 = 1 - w_1$, and probability measures $\nu_1, \nu_2, \nu \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$\begin{aligned} &w_1 \mathcal{W}_2^2(\nu_1, \nu) + w_2 \mathcal{W}_2^2(\nu_2, \nu) \\ &\qquad \le w_1 w_2 \mathcal{W}_2^2(\nu_1, \nu_2) + \mathcal{W}_2^2(\gamma_{\nu_1,\nu_2}(w_2), \nu) \end{aligned}$$

where γ is a geodesic curve as in Proposition 4. Moreover, the inequality reduces to equality when the local curvature of $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$ is zero, which is the case when ν_1, ν_2 and ν are Dirac measures.

From Lemma 5, we have that the Wasserstein distance satisfies condition (5) with the geodesic γ as a merging function and with $\bar{\mathcal{D}}_t(\nu_1, \nu_2) = t(1-t)\mathcal{W}_2^2(\nu_1, \nu_2)$.

Proposition 6 (Thm 2.2 [25], [26])**.** *The Wasserstein distance between two Gaussian distributions is given in closed-form by*

$$\begin{aligned} \mathcal{W}_2^2\left(\mathcal{N}(\mu_1, \Sigma_1), \mathcal{N}(\mu_2, \Sigma_2)\right) = \|\mu_1 - \mu_2\|^2 + \operatorname{tr}\Sigma_1 + \operatorname{tr}\Sigma_2 \\ - 2\operatorname{tr}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2} \ . \end{aligned}$$

Consider now the subspace $\mathcal{N}_0^d \subset \mathcal{P}_2(\mathbb{R}^d)$ composed of d-dimensional zero-mean Gaussian probability measures and denote by $\mathbb{P}(d)$ the set of positive semidefinite matrices in $\mathbb{R}^{d\times d}$. The next proposition states that $\mathcal{N}_0^d$ is a totally geodesic submanifold of $\mathcal{P}_2(\mathbb{R}^d)$, i.e., any two points in $\mathcal{N}_0^d$ are connected by a geodesic that lies in $\mathcal{N}_0^d$.

Proposition 7 ([27], Example 1.7; [25])**.** *For $\mathcal{N}(0, V) \in \mathcal{N}_0^d$ and $\mathcal{N}(0, U) \in \mathcal{N}_0^d$, with U, V positive definite, define*

$$T := U^{1/2}(U^{1/2}VU^{1/2})^{-1/2}U^{1/2}$$

and

$$\Gamma(t) := [tI + (1-t)T]V[tI + (1-t)T] \ .$$

Then $\mathcal{N}(0, \Gamma(t))$ is a geodesic from $\mathcal{N}(0, V)$ to $\mathcal{N}(0, U)$ in $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$. In addition, $\Gamma(t)$ is itself a geodesic on the space $\mathbb{P}(d)$ endowed with the metric $\mathcal{W}_2(U, V) = \mathcal{W}_2(\mathcal{N}(0, U), \mathcal{N}(0, V))$.

From Proposition 6, we see that the submanifold of Gaussian measures can be parametrized by the direct sum of $\mathbb{R}^d$, equipped with the Euclidean distance, and $\mathbb{P}^d$ equipped with the Wasserstein metric. Therefore, the full geodesic from $\mathcal{N}(\mu_1, U)$ to $\mathcal{N}(\mu_2, V)$ is given by $\mathcal{N}((1-t)\mu_1 + t\mu_2, \Gamma(t))$.

B. Approximations of the Wasserstein Geodesics

The geodesics given by Proposition 7 require the computation of matrix square roots, which is disadvantageous from the perspective of computational time. For example, this operation may be many times slower than a matrix inversion or the Cholesky decomposition. We investigate faster alternatives from approximations of the Wasserstein geodesic.

Two alternative merging functions that come easily to mind are the arithmetic and harmonic matrix means. One can show that the harmonic mean leads to a bounded $\bar{\mathcal{D}}_t$ whereas the arithmetic mean leads to unboundedness. However, it turns out (as verified empirically) that the arithmetic mean gives a tighter approximation of the Wasserstein geodesic for small distances. This assertion is related to the following theorem.

Theorem 8. *For positive definite matrices Σ_1 and Σ_2, the Wasserstein distance between them is upper bounded as*

$$\mathcal{W}_2^2\left(\Sigma_1, \Sigma_2\right) \le \frac{1}{4}\operatorname{tr}(\Sigma_1 - \Sigma_2)\Sigma_1^{-1}(\Sigma_1 - \Sigma_2) \ .$$

Proof. Let $\gamma(t) = \Sigma_1 + t(\Sigma_2 - \Sigma_1))$, $t \in [0, 1]$, be a non-geodesic curve connecting Σ_1 and Σ_2 on $\mathbb{P}(d)$. Since $\mathcal{W}_2$ is a geodesic distance, it is upper bounded by the length of $\gamma(t)$. In order to compute the length of $\gamma(t)$, we first consider the expression for the metric tensor g that induces $\mathcal{W}_2$ and that is given in [26, Equation (32)]:

$$g_\Sigma(U, U) = \sum_{i=1}^{d}\sum_{j=1}^{d} \sigma_i \frac{u_{ij}^2}{(\sigma_i + \sigma_j)^2} \ ,$$

where $\Sigma = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_d) \in \mathbb{P}(d)$ and the tangent vector $U = [u_{ij}]$ is a symmetric matrix in $\mathbb{R}^{d\times d}$. Making use of the fact that $4\sigma_i\sigma_j \le (\sigma_i + \sigma_j)^2$, we have that

$$\begin{aligned} g_\Sigma(U, U) &= \sum_{i=1}^{d}\sum_{j=1}^{d} \frac{\sigma_i\sigma_j}{(\sigma_i + \sigma_j)^2}\sigma_j^{-1}u_{ij}^2 && (21) \\ &\le \sum_{i=1}^{d}\sum_{j=1}^{d} \frac{1}{4}\sigma_j^{-1}u_{ij}^2 = \frac{1}{4}\operatorname{tr} U\Sigma^{-1}U \ . && (22) \end{aligned}$$

Since $\operatorname{tr} U\Sigma^{-1}U$ is invariant under similarity transformations (and so are arc lengths), it also defines an upper bound when Σ is non-diagonal. Incidentally, one can verify that this bound is tight in the sense that, under the metric $\bar{g}_\Sigma(U, U) = 1/4 \operatorname{tr} U\Sigma^{-1}U$, the geodesic $\Gamma(t)$ in Proposition 7 has constant speed and has length equal to the Wasserstein distance. Moreover, we see from (22) that the two metrics coincide in the scalar case as the inequality becomes equality.

Now we can find an upper bound on the arc length of $\gamma(t)$ using the upper bound on the metric above. From the definition of arc length:

$$\mathcal{W}_2^2\left(\Sigma_1, \Sigma_2\right) = \left(\int_0^1 \sqrt{g_{\Gamma(t)}(\dot{\Gamma}(t), \dot{\Gamma}(t))}\; dt\right)^2$$

$$\leq \left(\int_0^1 \sqrt{g_{\gamma(t)}(\dot{\gamma}(t), \dot{\gamma}(t))}\; dt\right)^2 \leq \int_0^1 g_{\gamma(t)}(\dot{\gamma}(t), \dot{\gamma}(t))\; dt \;\;,$$

where the first inequality follows from the minimizing property of geodesics and the second one follows from convexity. Using the metric $\bar{g}$ above, and rearranging terms such that we have an analytic function of the matrix $Z = \Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2} - I$ in the integrand, we have

$$\begin{aligned}
&\mathcal{W}_2^2\left(\Sigma_1, \Sigma_2\right)\\
&\quad\leq \int_0^1 \frac{1}{4}\operatorname{tr}(\Sigma_2 - \Sigma_1)\left(\Sigma_1 + t(\Sigma_2 - \Sigma_1)\right)^{-1}\left(\Sigma_2 - \Sigma_1\right)\, dt\\
&\quad= \frac{1}{4}\operatorname{tr}(\Sigma_2 - \Sigma_1)\Sigma_1^{-1/2}\\
&\qquad\times \int_0^1 \left(I + t(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2} - I)\right)^{-1}\; dt \times\; \Sigma_1^{-1/2}(\Sigma_2 - \Sigma_1)\\
&\quad= \frac{1}{4}\operatorname{tr}(\Sigma_2 - \Sigma_1)\Sigma_1^{-1/2}Z^{-1}\ln(I + Z)\Sigma_1^{-1/2}(\Sigma_2 - \Sigma_1)\;\;.
\end{aligned}$$

Using the fact that $\ln(I + X) \leq X$ for a semidefinite matrix X, we have

$$\mathcal{W}_2^2\left(\Sigma_1, \Sigma_2\right) \leq \frac{1}{4}\operatorname{tr}(\Sigma_2 - \Sigma_1)\Sigma_1^{-1}(\Sigma_2 - \Sigma_1)\;\;.$$

□

From Theorem 8 and its proof we see that the Wasserstein geometry approximates the Euclidean geometry when Σ_1^{-1} and Σ_2^{-1} are close enough so that $\Gamma(t)^{-1}$ is approximately constant. This suggests the approximation of γ_t by the arithmetic mean $M_{\Sigma_1,\Sigma_2}(t)$ and of $\bar{\mathcal{D}}_t$ by $(1-t)\mathcal{W}_2^2(\Sigma_1, M_{\Sigma_1,\Sigma_2}(t)) + t\mathcal{W}_2^2(\Sigma_2, M_{\Sigma_1,\Sigma_2}(t))$ so that

$$\begin{aligned}
\bar{\mathcal{D}}_t(\Sigma_1, \Sigma_2) \approx{}& \frac{1}{4}t^2(1-t)\min\left\{\operatorname{tr}(\Sigma_1 - \Sigma_2)\Sigma_1^{-1}(\Sigma_1 - \Sigma_2),\right.\\
&\left.\operatorname{tr}(\Sigma_1 - \Sigma_2)M_{\Sigma_1,\Sigma_2}(t)^{-1}(\Sigma_1 - \Sigma_2)\right\}\\
&\quad+ \frac{1}{4}t(1-t)^2\min\left\{\operatorname{tr}(\Sigma_1 - \Sigma_2)\Sigma_2^{-1}(\Sigma_1 - \Sigma_2),\right.\\
&\left.\operatorname{tr}(\Sigma_1 - \Sigma_2)M_{\Sigma_1,\Sigma_2}(t)^{-1}(\Sigma_1 - \Sigma_2)\right\}\;\;,
\end{aligned}$$

where the symmetry of the Wasserstein distance was used to choose the smallest of the two possible bounds. In our experiments, to avoid the computation of inverses, we adopt the loose approximation

$$\begin{aligned}
&\min\{\Sigma_1^{-1}, M_{\Sigma_1,\Sigma_2}(t)^{-1}\}\\
&\quad\approx \operatorname{diagm}(\max\{\operatorname{diag}(\Sigma_1)^{-1}, \operatorname{diag}(M_{\Sigma_1,\Sigma_2}(t))^{-1}\})\;\;,
\end{aligned}\tag{23}$$

where the maximum is taken elementwise and where $\operatorname{diag}(\cdot)$ denotes the vector of diagonal elements of a matrix and $\operatorname{diagm}(\cdot)$ indicates the diagonal matrix with given entries.

C. Controlling the Mean Absolute Estimation Bias

In this section we show how a proper choice of a Wasserstein distance may be used to control the mean absolute estimation bias. To this purpose, we extend the definition of $\mathcal{W}_2$ above to $\mathcal{W}_{2,H}$ by replacing the Euclidean norm by the matrix weighted norm $\|\cdot\|_H$, for some positive definite matrix H.

We restrict our analysis to the system presented in Section II to take advantage of its mode-independent dynamics in order to obtain formal bounds between the time-averaged mean absolute estimation bias and the Wasserstein distance.

Proposition 9. *For the hybrid system presented in Section II, consider the prior probability ϕ given by the N-component Gaussian mixture*

$$\phi = \sum_{i=1}^{N} w_i\phi_i := \sum_{i=1}^{N} w_i\mathcal{N}(\mu_i, \bar{\Sigma})$$

and its $\mathbf{N}_c$-cluster approximation given by the probability density

$$\tilde{\phi} = \sum_{j=1}^{\mathbf{N}_c} \tilde{w}_j\tilde{\phi}_j := \sum_{j=1}^{\mathbf{N}_c} \tilde{w}_j\mathcal{N}(\tilde{\mu}_j, \bar{\Sigma})$$

where, for each cluster C_j, $\tilde{w}_j = \sum_{i\in\mathsf{C}_j} w_i$, $\tilde{\mu}_j = \sum_{i\in\mathsf{C}_j} w_i/\tilde{w}_j\;\mu_i$ and $\bar{\Sigma}$ is the posterior covariance at equilibrium. Let $x_{k|k}$ and $\tilde{x}_{k|k}$ denote the posterior estimates of x_k when x_0 is distributed with priors ϕ and $\tilde{\phi}$ respectively. Assume that the covariance of posterior means contracts over time (due to observations) with a rate $\bar{\alpha}$:

$$\mathrm{E}\left[\mathsf{Cov}_{\tilde{w}_{j,k,n}}(\tilde{\mu}_{j,k,n}) \mid x_0 \sim \tilde{\phi}\right] \leq \bar{\alpha}^k\mathsf{Cov}_{\tilde{w}_j}(\tilde{\mu}_j)\;\;,$$

where $\tilde{\mu}_{j,k,n}$ is the posterior mean of x_k for prior $\tilde{\phi}_j$ conditioned on the n-th possible mode sequence $m_{1:k}^{(n)}$, $n = 1, \ldots, M^k$, and $w_{j,k,n}$ is the associated posterior weight.

Let L be the Kalman gain at equilibrium and denote the initial intra-cluster deviations by $\Delta\mu_{i,0} = \mu_i - \tilde{\mu}_j$, for $i \in \mathsf{C}_j$.

Then, for constants $\lambda_1, \lambda_2 > 0$, $\gamma \in (0,1)$, $\gamma < \beta_1 < \gamma^{-1}|\lambda_{\max}(A - LCA)|^{-2}$ and $\gamma < \beta_2 < (\gamma\bar{\alpha})^{-1}$, the discounted expected absolute estimation bias in a given Q-norm is upper-bounded by a function of the initial cluster deviations as follows:

$$\sum_{k=1}^{\infty}\gamma^{k-1}\,\mathrm{E}[\|x_{k|k} - \tilde{x}_{k|k}\|_Q] \leq$$

$$\sum_{j=1}^{\mathbf{N}_c}\tilde{w}_j\sum_{i\in\mathsf{C}_j}\frac{w_i}{\tilde{w}_j}\mathcal{W}_{2,H}^2\left(\phi_i, \tilde{\phi}_j\right) + \varrho\;\;, \tag{24}$$

$$\varrho = 2\frac{\lambda_2\bar{\alpha}}{1 - \gamma\beta_2\bar{\alpha}}\sigma_0^2 + \mathcal{O}\left(\|\Delta\mu_{i,0}\|_{\hat{\Sigma}^{-1}}^4\right)\;\;, \tag{25}$$

where $\sigma_0^2 := \operatorname{tr} Q\mathsf{Cov}_{\tilde{w}_j}(\tilde{\mu}_j)$, $\hat{\Sigma}^{-1} := A'(A\bar{\Sigma}A' + R_w)^{-1}A$ and

$$H = \left(\frac{\lambda_1}{2}H_{\beta_1} + \frac{1}{2}\left(\frac{\lambda_1^{-1}}{1 - \gamma\beta_1^{-1}} + \frac{\lambda_2^{-1}}{1 - \gamma\beta_2^{-1}}\right)\hat{\Sigma}^{-1}\right)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2020.2976274, IEEE Transactions on Automatic Control

and H_{β_1} satisfies the Lyapunov equation

$$\gamma\beta_1(A-LCA)'H_{\beta_1}(A-LCA)-H_{\beta_1}+(A-LCA)'Q(A-LCA)=0.$$

Proof. Define the likelihood function for the output sequence $y_{1:k}$ and the n-th possible mode sequence $m_{1:k}^{(n)}$ as

$$\ell_{i,k,n} := \int p\left(y_{1:k}, m_{1:k}^{(n)} | m_0, x_0\right) \phi_i(x_0) dx_0$$

and define $\tilde{\ell}_{j,k,n}$ analogously for $\tilde{\phi}$. Likewise, denote by $\mu_{i,k,n}$ and $\tilde{\mu}_{j,k,n}$ the posterior means at time k corresponding to the n-th mode sequence and to priors ϕ_i and $\tilde{\phi}_j$ respectively.

Then, by the hidden Markov structure of the process, the posterior means $x_{k|k}$ and $\tilde{x}_{k|k}$ are given by the sum of the means for the continuous filters weighted by the posterior probability for each component:

$$x_{k|k} - \tilde{x}_{k|k} = \sum_{i,n} \frac{w_i \ell_{i,k,n}}{\ell_k} \mu_{i,k,n} - \sum_{j,n} \frac{\tilde{w}_j \tilde{\ell}_{j,k,n}}{\tilde{\ell}_k} \tilde{\mu}_{j,k,n}$$

$$= \sum_{j,n} \sum_{i \in \mathsf{C}_j} \frac{w_i \ell_{i,k,n}}{\ell_k} (\mu_{i,k,n} - \tilde{\mu}_{j,k,n})$$

$$- \sum_{j,n} \left(\frac{\tilde{w}_j \tilde{\ell}_{j,k,n}}{\tilde{\ell}_k} - \sum_{i \in \mathsf{C}_j} \frac{w_i \ell_{i,k,n}}{\ell_k} \right) \tilde{\mu}_{j,k,n} \ ,$$

where $\ell_k = \sum_{i,n} w_i \ell_{i,k,n}$ and $\tilde{\ell}_k = \sum_{j,n} \tilde{w}_j \tilde{\ell}_{j,k,n}$ and where in the last equality we collected the intra- and inter-cluster deviations in separate terms.

Now, recall that the estimation error e_k for each Kalman filter is such that

$$e_{k+1} = (I - LC)Ae_k + Lv_k - (I - LC)w_k \ .$$

Therefore, if two Kalman filters are initialized with means that differ by $\Delta\mu$, this difference will evolve with k according to $(A - LCA)^k \Delta\mu$ independently of the noise. This implies that the same evolution will apply for the hybrid system when we compare posterior means with equal mode sequences $m_{1:k}$. Therefore, denoting mean deviations by $\Delta\mu_{i,k} = (A - LCA)^k(\mu_i - \tilde{\mu}_j)$, $i \in \mathsf{C}_j$, and defining $\bar{\ell}_{j,k,n} = \sum_{i \in \mathsf{C}_j} w_i \ell_{i,k,n} / \tilde{w}_j$, we can write

$$x_{k|k} - \tilde{x}_{k|k} = \sum_{j,n} \sum_{i \in \mathsf{C}_j} \frac{w_i \ell_{i,k,n}}{\ell_k} \Delta\mu_{i,k} - \sum_{j,n} \tilde{w}_j \tilde{\mu}_{j,k,n} \tag{26}$$

$$\cdot \left(\frac{\tilde{\ell}_{j,k,n}}{\tilde{\ell}_k} - \frac{\bar{\ell}_{j,k,n}}{\ell_k} \right) = \sum_{j,n} \sum_{i \in \mathsf{C}_j} w_i \frac{\ell_{i,k,n} - \tilde{\ell}_{j,k,n}}{\ell_k} \Delta\mu_{i,k}$$

$$- \sum_{j,n} \tilde{w}_j \left(\frac{\tilde{\ell}_{j,k,n}}{\tilde{\ell}_k} - \frac{\bar{\ell}_{j,k,n}}{\ell_k} \right) (\tilde{\mu}_{j,k,n} - \tilde{\mu}_{k|k}) \ ,$$

where in the last equality we used the fact that $\sum_{i \in \mathsf{C}_j} w_i \Delta\mu_{i,k} = 0$ and introduced the full posteriors mean $\tilde{\mu}_{k|k} = \sum_j \tilde{w}_j \tilde{\ell}_{j,k,n} / \tilde{\ell}_k \ \tilde{\mu}_{j,k,n}$ multiplying a zero-sum term. Taking the matrix weighted norm and expectations on $y_{1:k}$ and $m_{1:k}$ (multiply by ℓ_k and integrate), we have

$$\mathrm{E}[\|x_{k|k} - \tilde{x}_{k|k}\|_Q]$$

$$\leq \sum_{j,n} \sum_{i \in \mathsf{C}_j} w_i \|\Delta\mu_{i,k}\|_Q \int |\ell_{i,k,n} - \tilde{\ell}_{j,k,n}| dy_{1:k}$$

$$+ \sum_{j,n} \tilde{w}_j \int \tilde{\ell}_{j,k,n} \left| \frac{\bar{\ell}_{j,k,n}}{\tilde{\ell}_{j,k,n}} - \frac{\ell_k}{\tilde{\ell}_k} \right| \ \|\tilde{\mu}_{j,k,n} - \tilde{\mu}_{k|k}\|_Q dy_{1:k} \ , \tag{27}$$

where we used the fact that the Kalman filter mean differences $\Delta\mu_{i,k}$ do not depend on y_k. Summing and subracting 1 in the argument of $|\cdot|$, the last term of (27) can be bounded by

$$\int \sum_{j,n} \tilde{w}_j \tilde{\ell}_{j,k,n} \left(\left| \frac{\bar{\ell}_{j,k,n}}{\tilde{\ell}_{j,k,n}} - 1 \right| + \left| \frac{\ell_k}{\tilde{\ell}_k} - 1 \right| \right) \|\tilde{\mu}_{j,k,n} - \tilde{\mu}_{k|k}\|_Q dy_{1:k}$$

$$\leq \left(\int \sum_{j,n} \tilde{w}_j \tilde{\ell}_{j,k,n} \left| \frac{\bar{\ell}_{j,k,n}}{\tilde{\ell}_{j,k,n}} - 1 \right|^2 dy_{1:k} \right)^{1/2} \mathsf{Var}_Q(\tilde{\mu}_{j,k,n})^{1/2}$$

$$+ \left(\int \sum_{j,n} \tilde{w}_j \tilde{\ell}_{j,k,n} \left| \frac{\ell_k}{\tilde{\ell}_k} - 1 \right|^2 dy_{1:k} \right)^{1/2} \mathsf{Var}_Q(\tilde{\mu}_{j,k,n})^{1/2}$$

$$= \left[\left(\sum_j \tilde{w}_j \chi^2(\bar{\ell}_{j,k,n}, \tilde{\ell}_{j,k,n}) \right)^{1/2} + \left(\chi^2(\ell_k, \tilde{\ell}_k) \right)^{1/2} \right] \mathsf{Var}_Q(\tilde{\mu}_{j,k,n})^{1/2}$$

$$\leq 2 \left(\sum_j \tilde{w}_j \sum_{i \in \mathsf{C}_j} \frac{w_i}{\tilde{w}_j} \chi^2(\ell_{i,k,n}, \tilde{\ell}_{j,k,n}) \right)^{1/2} \mathsf{Var}_Q(\tilde{\mu}_{j,k,n})^{1/2} \ ,$$

where the first inequality follows from Hölder's inequality and $\mathsf{Var}_Q(\tilde{\mu}_{j,k,n})$ is the expected variance of the cluster centers for prior $\tilde{\phi}$; the equality follows from the definition of χ^2 in Section IV and the last inequality follows from the convexity of f-divergences applied to the χ^2 functions.

In order to compute the divergences between $\ell_{i,k,n}$ and $\tilde{\ell}_{j,k,n}$, we note that $p(y_{1:k}|m_{0:k}, x_0 \sim \phi_i)$ is a multivariate Gaussian distribution that may be computed in closed form offline. Since the likelihoods tend to grow apart with time, one may use $p(y_{1:\infty}|x_0 \sim \phi_i)$ to obtain an upper bound on their divergences. Alternatively, we provide in the sequence a looser bound that may be applied in more general situations.

To compute the f-divergence between likelihoods, note that

$$\ell_{i,k,n} = \int p\left(y_{1:k}, m_{1:k}^{(n)} | m_0, x_0\right) \phi_i(x_0) dx_0$$

$$= \int p\left(y_{1:k}, m_{2:k}^{(n)} | m_1^{(n)}, x_1\right) \pi_{m_1^{(n)}|m_0} \phi_i^+(x_1 | m_1^{(n)}) dx_1 \ ,$$

where $\phi_i^+(x_1|m_1)$ denotes the prior probability of x_1 given $x_0 \sim \phi_i$ and m_1. Then,

$$\begin{aligned}D_f(\ell_{i,k,n}, \tilde{\ell}_{j,k,n}) &= \int \sum_{m_k^{(n)}} \tilde{\ell}_{j,k,n} f\left(\frac{\ell_{i,k,n}}{\tilde{\ell}_{j,k,n}}\right) dy_{1\ :k} \\ &\leq \int \sum_{m_k^{(n)}} \int p\left(y_{1:k}, m_{2:k}^{(n)} | m_1^{(n)}, x_1\right) \pi_{m_1^{(n)}|m_0} \\ &\quad \cdot \tilde{\phi}_j^+(x_1|m_1^{(n)}) f\left(\frac{\phi_i^+(x_1|m_1^{(n)})}{\tilde{\phi}_j^+(x_1|m_1^{(n)})}\right) dx_1 dy_{1:k} \\ &= \sum_{m_1^{(n)}} \pi_{m_1^{(n)}|m_0} \int \tilde{\phi}_j^+(x_1|m_1^{(n)}) f\left(\frac{\phi_i^+(x_1|m_1^{(n)})}{\tilde{\phi}_j^+(x_1|m_1^{(n)})}\right) dx_1\end{aligned}$$

where the inequality follows from the convexity of the map $(p,q) \mapsto qf(p/q)$, which allows us to pull out from it the integration on dx_1.

Given that $\tilde{\phi}_j^+$ and ϕ_i^+ both have covariance $\bar{\Sigma}_+ := A\bar{\Sigma}A' + R_w$ and means that differ by $A\Delta\mu_{i,0}$ for all n and m_1, we can apply the previous equation and (19) to obtain

$$\begin{aligned}\chi^2(\ell_{i,k,n}, \tilde{\ell}_{j,k,n}) &\leq \chi^2(\phi_i^+, \tilde{\phi}_j^+) \\ &= \exp\left(\Delta\mu_{i,0}' A' \bar{\Sigma}_+^{-1} A \Delta\mu_{i,0}\right) - 1 \\ &= \Delta\mu_{i,0}' A' \bar{\Sigma}_+^{-1} A \Delta\mu_{i,0} \\ &\quad + \mathcal{O}((\Delta\mu_{i,0}' A' \bar{\Sigma}_+^{-1} A \Delta\mu_{i,0})^2)\ .\end{aligned} \tag{28}$$

From the inequality between the Hellinger divergence and the total variation given in [28], we obtain:

$$\begin{aligned}&\left(\sum_n \int |\ell_{i,k,n} - \tilde{\ell}_{j,k,n}| dy_{1:k}\right)^2 \leq 8\mathcal{H}(\ell_{i,k,n}, \tilde{\ell}_{j,k,n})^2 \\ &\leq 8\mathcal{H}(\phi_i^+, \tilde{\phi}_j^+)^2 = 8\left(1 - \exp\left(-\frac{1}{8}\Delta\mu_{i,0}' A' \bar{\Sigma}_+^{-1} A \Delta\mu_{i,0}\right)\right) \\ &\leq \Delta\mu_{i,0}' A' \bar{\Sigma}_+^{-1} A \Delta\mu_{i,0}\ \ ,\end{aligned}$$

where we used the expression for $\mathcal{H}^2$ derived from (16) and the fact that $1 - e^{-x} \leq x$.

Neglecting higher order terms, we replace these bounds in (27) and apply Hölder's inequality once again, to find that

$$\begin{aligned}\mathrm{E}[\|x_{k|k} - \tilde{x}_{k|k}\|_Q] &\leq \left(\sum_{j=1}^{\mathbf{N}_c} \tilde{w}_j \sum_{i \in \mathsf{C}_j} \frac{w_i}{\tilde{w}_j} \|\Delta\mu_{i,0}\|_{\hat{\Sigma}^{-1}}^2\right)^{1/2} \\ &\times \left[\left(\sum_{j=1}^{\mathbf{N}_c} \tilde{w}_j \sum_{i \in \mathsf{C}_j} \frac{w_i}{\tilde{w}_j} \|\Delta\mu_{i,k}\|_Q^2\right)^{1/2} + \ 2\bar{\alpha}^{k/2}\sigma_0\right]\ ,\end{aligned}$$

where $\hat{\Sigma}^{-1} := A'\bar{\Sigma}_+^{-1}A$ and where we used the assumption in the proposition to bound $\mathsf{Var}_Q(\tilde{\mu}_{j,k,n})^{1/2}$.

To compute the discounted cost, we apply Young's inequality twice with factors $\lambda_1\beta_1^{k-1}$ and $\lambda_2\beta_2^{k-1}$ to upper bound the products of the square roots by sums and find that

$$\begin{aligned}&\sum_{k\ =1}^{\infty} \gamma^{k-1}\,\mathrm{E}[\|x_{k|k} - \tilde{x}_{k|k}\|_Q] \\ &\leq \sum_{j=1}^{\mathbf{N}_c} \tilde{w}_j \sum_{i \in \mathsf{C}_j} \frac{w_i}{\tilde{w}_j} \sum_{k=1}^{\infty} \gamma^{k-1} \left[\frac{\lambda_1\beta_1^{k-1}}{2} \|\Delta\mu_{i,k}\|_Q^2 + \frac{\lambda_1^{-1}\beta_1^{1-k}}{2}\right. \\ &\left. \cdot \|\Delta\mu_{i,0}\|_{\hat{\Sigma}^{-1}}^2 + \frac{\lambda_2\beta_2^{k-1}}{2} 4\bar{\alpha}^k\sigma_0^2 + \frac{\lambda_2^{-1}\beta_2^{1-k}}{2} \|\Delta\mu_{i,0}\|_{\hat{\Sigma}^{-1}}^2\right] \\ &= 2\frac{\lambda_2\bar{\alpha}}{1 - \gamma\bar{\alpha}\beta_2}\sigma_0^2 + \sum_{j=1}^{\mathbf{N}_c} \tilde{w}_j \left(\sum_{i \in \mathsf{C}_j} \frac{w_i}{\tilde{w}_j} \Delta\mu_{i,0}' H \Delta\mu_{i,0}\right)\ \ ,\end{aligned} \tag{29}$$

where we used the fact that $H_{\beta_1} = \sum_{k=1}^{\infty}(\gamma\beta_1)^{k-1}[(A - LCA)']^k Q[A - LCA]^k$. From Proposition 6, the last term in (29) is the sum of the $\mathcal{W}_{2,H}^2$ distances as given in (24). □

Remark 4. *An important consequence of Proposition 9 is that, by picking an appropriate H-norm for the Wasserstein distance, we are able to control the mean absolute error for a given Q-norm within the framework of Section III. Nevertheless, this bound on the error cannot be made arbitrarily small due to the constant term depending on σ_0^2. Compared to (18), we see that the weight matrix H is a linear combinantion of a function of the weight matrix Q and of the inverse covariance $\hat{\Sigma}^{-1}$, which appeared in (18).*

The contraction with rate $\bar{\alpha}$ assumed in the proposition is a consequence of weight degeneration in Bayesian filtering, where, as time evolves and we take more process observations, one hypothesis will tend to have weight one whereas the other weights will tend to zero.

In choosing H, it would be interesting to enforce the contraction property $(A - LCA)'H(A - LCA) < \alpha H$ so that the filter would give a contraction in this particular Wasserstein space. By definition, this property is already satisfied by H_{β_1}, but it may not be satisfied by $\hat{\Sigma}^{-1}$. Nevertheless, this contraction property is true for the posterior covariance $\bar{\Sigma}^{-1}$, which could have been used instead of $\hat{\Sigma}^{-1}$ in the derivations preceding (28).

In practice, we expect the term associated to σ_0 to be much smaller. To see this, note that the last term in (26) is the covariance between the approximation error in cluster weights and the position of cluster centers. The correlation coefficient between these variables is expected to be small. Since this term takes an average over $\mathbf{N}_c$ weight errors at time 0, those of which are not strongly correlated, it would be reasonable to expect an asymptotic correlation coefficient $\rho_0/\sqrt{\mathbf{N}_c}$.

In order to optimize H, we may take the expected value on $\Delta\mu_{i,0}$ in (24) and assume that, given some (steady-state) covariance matrix Σ_0 for mean vectors, variance is homogeneously distributed among clusters so that

$$\mathrm{E}\left[\sum_{j=1}^{\mathbf{N}_c} \tilde{w}_j \left(\sum_{i \in \mathsf{C}_j} \frac{w_i}{\tilde{w}_j} \Delta\mu_{i,0}' H \Delta\mu_{i,0}\right)\right] \approx \frac{\operatorname{tr} H\Sigma_0}{\mathbf{N}_c} \tag{30}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2020.2976274, IEEE Transactions on Automatic Control

and $\sigma_0^2 \leq \operatorname{tr} Q\Sigma_0$. Then, we search for the parameters that minimize such an expected value: the factors λ_1, β_1 that minimize

$$\lambda_1 \operatorname{tr} H_{\beta_1}\Sigma_0 + \frac{\lambda_1^{-1}}{1-\gamma\beta_1^{-1}} \operatorname{tr} \hat{\Sigma}^{-1}\Sigma_0$$

and the factors λ_2, β_2 that minimize

$$\frac{4\rho_0^2\lambda_2\bar{\alpha}}{1-\gamma\beta_2\bar{\alpha}} \operatorname{tr} Q\Sigma_0 + \frac{\lambda_2^{-1}}{1-\gamma\beta_2^{-1}} \operatorname{tr} \hat{\Sigma}^{-1}\Sigma_0 \ .$$

Note that the actual values of $\mathbf{N}_c$ and (the amplitude of) Σ_0 do not play a role in this optimization as they contribute to both terms in (24).

Also noteworthy, the expected value of $\mathcal{W}_{2,H}^2$ given in (30) is consistent with the assumption in Section III that $\mathrm{E}[\Delta^{(k)}]$ is constant under policies where N_k is constant, given that $\mathbf{N}_c$ would be a constant over time in this case.

VI. Numerical Experiments

We conducted numerical experiments for the system described in Section II considering the discretized motion of a point on the line with linear friction coefficient ζ and with noisy position and acceleration measurements:

$$A = \begin{bmatrix} 1 & T_s & \frac{T_s^2}{2} \\ 0 & 1-\zeta\frac{T_s^2}{2} & T_s \\ 0 & -\zeta T_s & 1-\zeta\frac{T_s^2}{2} \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

$$C = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

where $T_s = 0.1$. The noise covariances were given by $R_v = \operatorname{diagm}(2\cdot 10^4, 0.1)$ and $R_w = \frac{1}{3}\cdot\operatorname{diagm}(10^{-8}, 10^{-5}, 10^{-5})$. The packet drop probability was $p_0 = 0.4$. The input u was given by the signal $u_k = 5\cdot 10^{-4}\sin(2\pi k\ 20/T) + 4.6\cdot 10^{-2}\tilde{w}$, where $\tilde{w}$ is a unit-variance white Gaussian noise and where $T = 3000$ is the total simulation time.

A. First scenario: two modes and no friction

For the first set of experiments, we defined $\zeta = 0$ and run 125 different realizations (the same set of 125 realizations was used in each filter) so as to obtain at least 4% precision in the time cost estimates and at least 2% precision in error cost estimates. We considered average costs (discount factor $\gamma = 1.0$) instead of discounted costs.

As a common metric for all filters, we compare the absolute estimation error weighted by a matrix $Q = \operatorname{diagm}(500, 20, 1)\bar{\Sigma}^{-1}\operatorname{diagm}(500, 20, 1)$, for $\bar{\Sigma}$ being the posterior covariance at equilibrium. This norm indicates that we give 500 and 20 times more importance to the estimation error of the position and of the velocity respectively. As in the asymptotic formula (18), f-divergences are more efficient at minimizing estimation errors weighted by $\bar{\Sigma}^{-1}$. In this sense, our choice of Q is to highlight that the Wasserstein distance can adapt to user-desired metrics whereas the f-divergences cannot.

The value of H was obtained as described in the previous section setting $\rho_0 = 0.14$ and $\bar{\alpha} = M^{-1} = 0.5$. For the sake of comparison, we give the Cholesky factors of the computed (normalized) H and $\bar{\Sigma}^{-1}$ matrices

$$\mathtt{chol}(H) \propto \begin{bmatrix} 1.0 & -5.67 & -69.88 \\ 0 & 22.11 & -5.72 \\ 0 & 0 & 170.97 \end{bmatrix}$$

and

$$\mathtt{chol}(\bar{\Sigma}^{-1}) \propto \begin{bmatrix} 1.0 & -21.48 & 177.58 \\ 0 & 16.92 & -227.61 \\ 0 & 0 & 288.45 \end{bmatrix}.$$

The angle between these two matrices as given by the trace inner product is of 59.5°, which demonstrates a substantial deviation from the behavior of the information divergences as given by (18). In addition, we have that $\Delta\mu' H \Delta\mu$ contracts under the action of $(A - LCA)$ with rate $\alpha \leq 0.99842$.

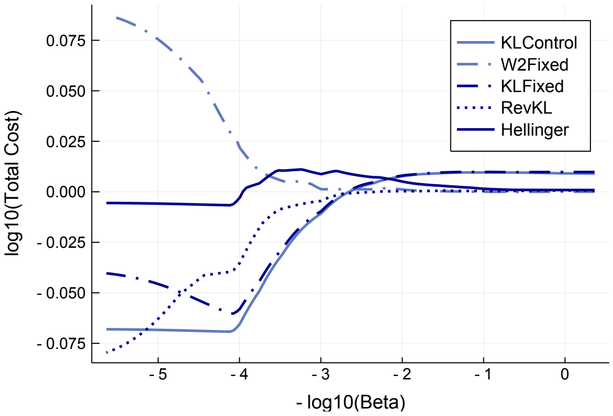

Fig. 1. Average total cost for $c_k = \|e_k\|_Q + \beta\tau_k$ computed from time and realization averages of the simulated estimation errors e_k and computation times τ_k in the first scenario. Each line corresponds to the optimal tuning of the respective algorithm for a given β. Labels refer to Algorithm 1 using Kullback-Leibler (`KL`), Wasserstein (`W2`), Hellinger and Reverse Kullback-Leibler (`RevKL`) as divergences. The labels `W2Fixed` and `KLFixed` refer to setting $\kappa_0 = 0$ (no control) in Algorithm 1 and varying $N_{\max}$ from 2 to 17 components. The remaining curves refer to Algorithm 1 with $N_{\max} = 30$, the best value of κ_0 for each β, and τ_0 in (14) given by 6.63 for $\mathcal{W}_2$, 3.9 for `KL`, 4.45 for $\mathcal{H}^2$ and 3.28 for `RKL`. Costs are normalized by the cost achieved by the Wasserstein distance (`W2Control`, shown as 0 in the plot).

We have performed experiments with all the proposed divergences. For the sake of comparison, we have also tested the case in which the number of reduced components is fixed as in the Runnalls' approach of [3] so that there is no closed-loop precision control. For the case of the Wasserstein distance, we only present here the results for the approximation given by (23) as they are significantly faster.

The results are summarized in Figures 1 and 2. A first conclusion is that controlling the number of components provides an improvement as compared to using a fixed number of components. A second conclusion is that the `KL` divergence is the most error efficient when we require smaller processing times but it is the worst when we allow larger processing times. We believe this performance degradation is due to the moment preserving merge of (15) that, even when the merged covariance matrices are at equilibrium and are equal, gives a different covariance matrix. The Wasserstein distance gives the best results when smaller errors are required. This result is expected given that the bounds in Proposition 9 are tighter for smaller errors.

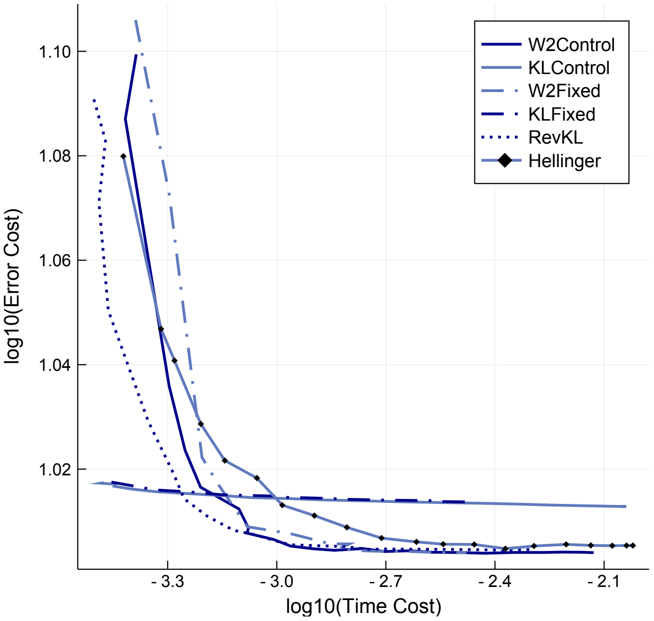

Fig. 2. Estimation error cost versus processing time cost for different parameters κ_0 (and hence different user-defined preferences β) and $N_{\max}$ in Algorithm 1 as described for the curves in Figure 1.

B. 2nd scenario: switched friction with a total of four modes

In a second set of experiments, we considered the scenario where the friction coefficient ζ is an unknown switching process. At each time-step, ζ assumes the value 1 with probability 0.8 and the value 10 with probability 0.2. As a consequence, we have a total of 4 unknown modes and the matrix A may change with the mode. In this case, however, there is no equilibrium value for the covariance matrices and the analysis from Section V-C does not hold.

The matrices Q and H were chosen as in the previous section, but considering the approximation that $c = 1$ would hold long enough for the covariances to reach an equilibrium $\bar{\Sigma}$. Again, we run 125 different realizations so as to obtain at least 5% precision in the time cost estimates and at least 7% precision in error cost estimates.

The results shown in Figures 3 and 4 indicate that the closed-loop control of the number of components has become more advantageous as the number of modes increased. On the other hand, we notice that in the new scenario the $\texttt{KL}$ divergence gave the best performance for all time preferences. To interpret such a change in performance, recall that the covariance matrices are no longer at equilibrium and consider the notion of entropic means discussed in Section IV. Notice that $\texttt{KL}$ is the only divergence whose entropic mean is the arithmetic mean of the pdfs. Since the pdf of a mixture is itself an arithmetic mean, the $\texttt{KL}$ divergence is the only information divergence with the correct target. Regarding the drawback associated with the moment preserving merge that we discussed earlier, it seems to not be as important given that covariances no longer are at equilibrium.

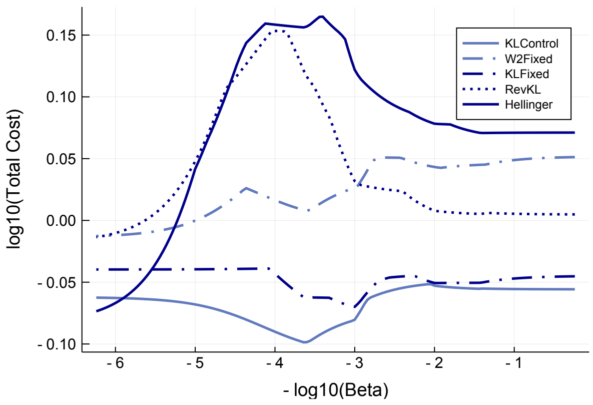

Fig. 3. Best average cost achieved in the second scenario by each divergence as a function of the the processing time weight β. The remaining settings were as in Figure 1. Costs are normalized by the cost achieved by the Wasserstein distance (shown as 0 in the plot).

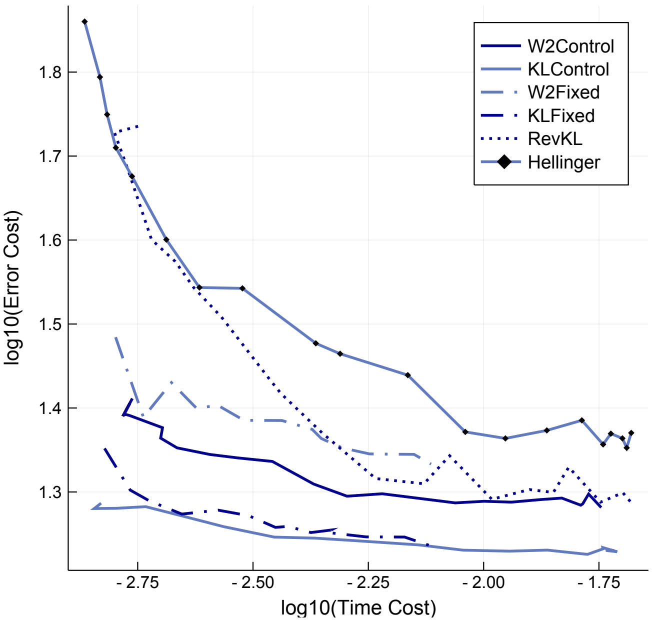

Fig. 4. Estimation error cost versus processing time cost for different parameters κ_0 and $N_{\max}$ in Algorithm 1 for the second scenario as described for the curves in Figure 3.

When comparing the performance of $\mathcal{H}^2$ and $\texttt{RKL}$, we see that the latter gives better results in most scenarios. This might be due to the fact that the error bound given in Proposition 2 for $\texttt{RKL}$ is tight whereas the error bound for $\mathcal{H}^2$ is overly conservative. At the same time, the proposition gives the same merging function for both divergences.

Lastly, one must acknowledge that these algorithms were designed to minimize the probability divergences between the true posterior and its approximation and that this goal might not be directly related to the minimization of the estimation error. For this reason, the lack of monotonicity or convexity of the graphs in Figures 2 and 4 should not come as a surprise.

VII. Concluding Remarks

This paper presented an optimal control formulation to the problem of controlling the precision of Bayes filters where the number of filter hypotheses grows exponentially and truncation is needed. A new approach to probability divergences was introduced in order to keep track of the aggregated approximation error due to successive truncations. Our numerical results show that there is an improvement coming from the closed-loop control of the precision. On the other hand, results were not conclusive in the sense of demonstrating that one divergence measure gives superior performance in all cases.

The derivations in Section V-C suggest that the present framework may be successfully adapted to the case of a bank of $\mathcal{H}_2$ filters such as those given in [29]. The LMI-based procedure to derive the $\mathcal{H}_2$ filters provides, at the same time, a Lyapunov function that may be used to generate a Wasserstein distance that contracts along filtering operations.

References

[1] H. Driessen and Y. Boers, "Multiple-model multiple-hypothesis filter for tracking maneuvering targets," in *Signal and Data Processing of Small Targets 2001*, vol. 4473. International Society for Optics and Photonics, 2001, pp. 279–289.

[2] Y. Boers and H. Driessen, "A multiple model multiple hypothesis filter for Markovian switching systems," *Automatica*, vol. 41, no. 4, pp. 709–716, 2005.

[3] W. Y. Eras-Herrera, A. R. Mesquita, and B. O. S. Teixeira, "Multiple-model multiple-hypothesis filter with Gaussian mixture reduction," *International Journal of Adaptive Control and Signal Processing*, vol. 32, no. 2, pp. 286–300, 2018.

[4] D. F. Crouse, P. Willett, K. Pattipati, and L. Svensson, "A look at Gaussian mixture reduction algorithms," in *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*. IEEE, 2011, pp. 1–8.

[5] A. R. Runnalls, "Kullback-Leibler approach to Gaussian mixture reduction," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 43, no. 3, 2007.

[6] A. R. Mesquita, J. P. Hespanha, and G. N. Nair, "Redundant data transmission in control/estimation over lossy networks," *Automatica*, vol. 48, no. 8, pp. 1612–1620, 2012.

[7] J. P. Hespanha and A. R. Mesquita, "Networked control systems: estimation and control over lossy networks," *Encyclopedia of Systems and Control*, pp. 842–849, 2015.

[8] O. Cappé, E. Moulines, and T. Rydén, *Inference in hidden Markov models*. Springer, 2005.

[9] J. V. Candy, *Bayesian signal processing: classical, modern, and particle filtering methods*. John Wiley & Sons, 2016, vol. 54.

[10] D. P. Bertsekas, *Dynamic programming and optimal control*. Athena scientific Belmont, MA, 2005, vol. 2.

[11] I. Vajda, "On metric divergences of probability measures," *Kybernetika*, vol. 45, no. 6, pp. 885–900, 2009.

[12] A. Ben-Tal, A. Charnes, and M. Teboulle, "Entropic means," *Journal of Mathematical Analysis and Applications*, vol. 139, pp. 537–551, 1989.

[13] B. Pelletier, "Informative barycentres in statistics," *Annals of the Institute of Statistical Mathematics*, vol. 57, no. 4, pp. 767–780, 2005.

[14] F. Nielsen and R. Nock, "Sided and symmetrized Bregman centroids," *IEEE Transactions on Information Theory*, vol. 55, no. 6, pp. 2882–2904, 2009.

[15] ——, "The entropic centers of multivariate normal distributions," *Collection of Abstracts*, vol. 221, 2008.

[16] F. Nielsen and S. Boltz, "The Burbea-Rao and Bhattacharyya centroids," *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 5455–5466, 2011.

[17] R. Bhatia, *Positive definite matrices*. Princeton university press, 2009, vol. 16.

[18] F. Nielsen and R. Nock, "On the chi square and higher-order chi distances for approximating f-divergences," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 10–13, 2014.

[19] Y. Chen, T. T. Georgiou, and A. Tannenbaum, "Optimal transport for Gaussian mixture models," *IEEE Access*, vol. 7, pp. 6269–6278, 2018.

[20] A. Assa and K. N. Plataniotis, "Wasserstein-distance-based Gaussian mixture reduction," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1465–1469, Oct 2018.

[21] C. Villani, *Optimal transport: old and new*. Springer Science & Business Media, 2008, vol. 338.

[22] M. Agueh and G. Carlier, "Barycenters in the Wasserstein space," *SIAM Journal on Mathematical Analysis*, vol. 43, no. 2, pp. 904–924, 2011.

[23] P. C. Álvarez-Esteban, E. del Barrio, J. Cuesta-Albertos, and C. Matrán, "A fixed-point approach to barycenters in Wasserstein space," *Journal of Mathematical Analysis and Applications*, vol. 441, no. 2, pp. 744–762, 2016.

[24] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[25] A. Takatsu, "Wasserstein geometry of Gaussian measures," *Osaka Journal of Mathematics*, vol. 48, no. 4, pp. 1005–1026, 2011.

[26] R. Bhatia, T. Jain, and Y. Lim, "On the Bures–Wasserstein distance between positive definite matrices," *Expositiones Mathematicae*, 2018.

[27] R. J. McCann, "A convexity principle for interacting gases," *Advances in Mathematics*, vol. 128, no. 1, pp. 153–179, 1997.

[28] I. Sason and S. Verdu, "f-divergence inequalities," *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 5973–6006, 2016.

[29] A. R. Fioravanti, A. P. Goncalves, and J. C. Geromel, "$\mathcal{H}_2$ filtering of discrete-time Markov Jump Linear Systems through linear matrix inequalities," *International Journal of Control*, vol. 81, no. 8, pp. 1221–1231, 2008.

Alexandre Mesquita was born in Santo Antônio do Monte, Brazil, in 1982. He received the BSc and MSc degrees in Electronics Engineering from Instituto Tecnológico de Aeronáutica, São José dos Campos, Brazil, in 2004 and 2006 respectively, and the Ph.D. degree in Electrical Engineering from the University of California, Santa Barbara, in 2010. In 2012, he joined the Federal University of Minas Gerais, where he currently holds a position of Professor Adjunto in the Department of Electronics Engineering. His research interests include multi-agent systems, networked control systems and stochastic hybrid systems.