FEDERAL UNIVERSITY OF MINAS GERAIS

Department of Electronic Engineering

Technical Report

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

Alexandre Rodrigues Mesquita

December 31, 2018

1 Introduction

A major challenge in the state estimation of hybrid dynamical systems from a Bayesian approach lies in the exponential growth 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 continous filters with exponentially growing size. To cope with this problem, the multiple model multiple hypothesis filter (M^{3}H) was proposed in [1, 2]. Given that the Bayes posterior is given by 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]. By choosing different cluster sizes, one could obtain suitable combinations of estimation error and processing time.

If one considers the possibility of varying the cluster sizes with time, we see that, encrusted in the problem of Bayesian filtering of hybrid systems, there is a problem of precision control. More precisely, we have an optimal control problem in which one wants to minimize the time-averaged estimation error subject to a bounded computational time in each 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.

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 time is the control input. Our main contribution is in Section 3. By applying the framework in that section to information divergences in Section 4 and to the Wasserstein distance in Section 5, some new facts are also uncovered.

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

2 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 drop events may be modeled as 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. Let $x_k \in \mathbb{R}^d$ be the state of a linear system with dynamics given by

$$x_{k+1} = Ax_k + \epsilon_k B u_k + w_k \tag{1}$$

$$y_k = C x_k + v_k \ \ , \tag{2}$$

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

$$\Pr\{m_{k+1} = j | m_k = i\} = \pi_{j|i} \tag{3}$$

$$\Pr\{\epsilon_k = j | m_k = i\} = \varrho_{j|i} \tag{4}$$

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 M^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 $\phi_{m_0} = \mathcal{N}(\mu_{m_0}, \Sigma_{m_0})$. Define the likelihood function for the output sequence $y_{1:k}$ and the n-th possible mode sequence $\epsilon_{1:k}^{(n)}$, $n = 1, \ldots, 2M^k$, as

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

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

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

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

such that $m = 1$ always correspond 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.

3 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.

For a given space $\mathscr{P}$ of probability distributions, consider a generic divergence function $\mathscr{D} : \mathscr{P} \times \mathscr{P} \mapsto \mathbb{R}_{\geq 0} \cup \{\infty\}$ and assume that $\mathscr{D}$ is jointly convex.

Theorem 1. *Suppose that, for $t \in [0,1]$, ν_1 and $\nu_2 \in \mathscr{P}$, there exists a merging function $\gamma_t : \mathscr{P} \times \mathscr{P} \mapsto \mathscr{P}$ and a function $\bar{D}_t : \mathscr{P} \times \mathscr{P} \mapsto \mathbb{R}_{\geq 0} \cup \{\infty\}$ such that*

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

for all $\nu \in \mathscr{P}$. Now, consider a mixture probability distribution in $\mathscr{P}$ with components $(w_i, \nu_i), i = 1, \ldots, N$ and 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 $\bar{w}_n = \sum_{i=1}^n w_i$. Then, the total divergence resulting from such a merging is bounded as

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

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

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

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

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

Next, we note that

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

where the first inequality follows from the convexity of $\mathscr{D}$ 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 $\mathscr{D}$ in (5) by the Euclidean distance squared, $\gamma_t(x, y) = (1-t)x + ty$ and $\bar{\mathscr{D}}_t$ by $t(1-t)\|x-y\|^2$, 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 $\mathscr{D}$ admits a contraction rate α, i.e.,

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

Then, if we denote by $\mathscr{E}_k$ the bound on the truncation error accumulated from all times previous to k, we have that

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

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 with state $\mathcal{E}_k$, actions $N_{k,m}$ as the number of components of the reduced measure for mode m at time k and instantaneous cost $c(\mathcal{E}_k, N_k) = \mathcal{E}_k + \beta\tau(N_k)$, 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 time.

For a control policy with N_k and $\Delta^{(k)}$ constant in time, and for a discount factor $\gamma \in (0,1)$ for the overall cost, we would have a value function $V(\mathcal{E}_k) = \mathcal{E}_k/(1-\gamma\alpha)+\rho_0$ for some constant ρ_0. This value function leads to the suboptimal policy

$$N_k = \arg\min_{N_k} \frac{\gamma\alpha}{1-\gamma\alpha}\Delta^{(k)} + \beta\tau(N_k) \ . \tag{11}$$

Computing the minimum in (11) would by itself affect the computational time $\tau(N_K)$ if this is to be done online. Instead, 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. 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

$$\bar{w}_n \bar{\mathscr{D}}_{\frac{w_n}{\bar{w}_n}}(\bar{\nu}_{n-1}, \nu_n) > \frac{1-\gamma\alpha}{\gamma\alpha}\beta(\tau(N_k)-\tau(N_k-\delta_m)) \ . \tag{12}$$

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 is run within the M^{3}HR algorithm to replace the Runnalls' algorithm in the step of merging the posterior distributions at each time-step.

Algorithm 1 Suboptimal Gaussian Mixture Model Reduction for MJSs

```
Given the mixture $\sum_{m=1}^{M}\sum_{i=1}^{N_m} w_{i,m}\nu_{i,m}$, a constant $\kappa_0$ and an integer $N_{\max}$,
Compute $c_{i,j,m} = (w_{i,m}+w_{j,m})\bar{\mathscr{D}}_{w_{j,m}/\left(w_{i,m}+w_{j,m}\right)}(\nu_{i,m}, \nu_{j,m})$ for all $i < j$ and all $m$.
Set StopFlag=FALSE.
while $\sum_m N_m > M$ do
    Find the indices $i^* < j^*$ and $m^*$ that minimize $c_{i,j,m}$.
    if $c_{i^*,j^*,m^*} > \kappa_0[\tau([N_m]) - \tau([N_m]-\delta_{m^*})]$ then
        Set StopFlag=TRUE.
        if $\sum_m N_m \leq N_{\max}$ then
            break
        end if
    end if
    Set $\nu_{i^*,m^*} = \gamma_{w_{j^*,m^*}/(w_{i^*,m^*}+w_{j^*,m^*})}(\nu_{i^*,m^*}, \nu_{j^*,m^*})$.
    Set $w_{i^*,m^*} = w_{i^*,m^*} + w_{j^*,m^*}$.
    Remove component $j^*$ from the mixture of index $m^*$.
    Set $N_{m^*} = N_{m^*} - 1$.
    if $\sum_m N_m \leq N_{\max}$ & StopFlag then
        break
    end if
    Update $c_{i^*,j,m^*}$ for $j > i^*$.
end while
```

Note that 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 (12).

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 $M \sum_m N_m$ Kalman filters are run after we reduce each mixture to a size of N_m. Thus, the 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 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, step 2 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, step 19 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 $c_{i,j,m}$. This results in a computational time at time k proportional to:

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

where the constant τ_0 corresponds to the computational time of the Kalman filters. 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 obtain

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

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 bounds on error sizes and not the real errors. Secondly, $\mathscr{E}_k$ 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 roll-off policy and, on top of that, we have no guarantee that (12) gives global minimum.*

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

4 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 [8] for a definition in the general case and for further properties). Due to the convexity of the map $(x, y) \mapsto x f(y/x)$, 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, $D_f(\nu_1\|\nu_2) \geq 0$ if and only if $\nu_1 = \nu_2$. If P is a Markov transition operator, then $D_f(\nu_1\|\nu_2) \geq D_f(\nu_1\|\nu_2)$, 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 $\mathtt{TV}(\cdot,\cdot) := D_f(\cdot\|\cdot)$. For $f(t) = t\ln t$ we have the Kullback-Leibler divergence $\mathtt{KL}(\cdot,\cdot) := D_f(\cdot\|\cdot)$. For $f(t) = -\ln t$ we have the reverse Kullback-Leibler divergence $\mathtt{RKL}(\cdot,\cdot) := D_f(\cdot\|\cdot)$. For $f(t) = \frac{1}{2}(\sqrt{t}-1)^2$ we have the squared Hellinger distance $\mathscr{H}^2(\cdot,\cdot) := D_f(\cdot\|\cdot)$. 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 TV and $\mathscr{H}^2$ are symmetric. In addition, TV and $\mathscr{H}$ are true distances.

The optimal values for (γ_t, ν) in (5) can be defined by means of a min-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 $\bar{\mathscr{D}}$ is minimized by setting $\gamma_t = \nu^*$. For this reason, γ_t^* often coincides with the barycenter

$$\arg\min_{\nu} w_1\mathscr{D}(\nu_1, \nu) + w_2\mathscr{D}(\nu_2, \nu) \ .$$

When $\mathscr{D} = D_f$, [9] showed that the solutions to this problem, the so-called entropic means, are given by: the arithmetic mean of the pdfs in the case of the Kullback-Leibler divergence; the normalized geometric mean of the pdfs in the case of the reverse Kullback-Leibler divergence; the normalized mean square-roots of the pdfs in the case of the squared Hellinger divergence; and the normalized harmonic mean of the pdfs in the case of 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 $\bar{\mathscr{D}}_t$ that satisfy condition (5) when $\mathscr{P}$ is the space of multivariate normal distributions on $\mathbb{R}^d$, denoted here by $\mathscr{N}^d$.

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

1. *For the total variation distance:*

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

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

2. *for the Kullback-Leibler divergence:*

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

3. *for the reverse Kullback-Leibler divergence:*

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

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

4. *for the squared Hellinger distance:*

$$\begin{aligned}
\bar{\mu}_t &= \bar{\Sigma}_t \left((1-t)\Sigma_1^{-1}\mu_1 + t\Sigma_2^{-1}\mu_2\right)\\
\bar{\Sigma}_t &= \left((1-t)\Sigma_1^{-1} + t\Sigma_2^{-1}\right)^{-1}\\
\varphi_t &= \frac{1}{4}\left(t(1-t)(\mu_1-\mu_2)'\tilde{\Sigma}_t^{-1}(\mu_1-\mu_2) - \ln|\Sigma_t| + (1-t)\ln\Sigma_1 + t\ln\Sigma_2\right)\\
\bar{\mathscr{D}}_t &= 1-e^{-\varphi_t} \quad ,
\end{aligned}$$

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 [10, 11, 12]. 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 [13]). The expression given in the proposition can be checked by analyzing the Bhattacharyya distance $\ln(1-\mathscr{H}^2)$ in place of $\mathscr{H}^2$. By fixing the merging function as in the proposition, one can easily show that $\bar{\mathscr{D}}_t$ is maximized by $\nu = \mathscr{N}(\bar{\mu}_t, 0)$.

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

$$\bar{\mathscr{D}}_t \propto \frac{1}{2} t(1-t)(\mu_1-\mu_2)'\bar{\Sigma}^{-1}(\mu_1-\mu_2) \tag{13}$$

in the case of the last three divergences.

Notably, Runnals' 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 expressions for the Hellinger [13] and

χ^2-divergences [14] between multivariate normals:

$$\ln(1-\mathscr{H}^2(\nu_1,\nu_2)) = \frac{1}{4}(\mu_1-\mu_2)'(\Sigma_1+\Sigma_2)^{-1}(\mu_1-\mu_2)+\frac{1}{2}\ln\frac{|(\Sigma_1+\Sigma_2)/2|}{|\Sigma_1||\Sigma_2|} \quad (14)$$

and

$$\begin{aligned}\ln(\chi^2(\nu_1,\nu_2))) &= \frac{1}{2}(2\mu_2-\mu_1)'(2\Sigma_2-\Sigma_1)^{-1}(2\mu_2-\mu_1)\\ &+\frac{1}{2}\ln|2\Sigma_2-\Sigma_1|-\mu_2'\Sigma_2^{-1}\mu_2+\frac{1}{2}\mu_1'\Sigma_1^{-1}\mu_1-\ln|\Sigma_2|+\frac{1}{2}\ln|\Sigma_1| \ , \quad (15)\end{aligned}$$

for $2\Sigma_2 > \Sigma_1$.

5 Precision Control using the Wasserstein Distance

The limit behavior in (13) 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 te main facts about this distance and derive suitable merging functions and bounds for it.

In Section 5.3, 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 (13) by a combination of the form $\bar{\Sigma}^{-1}+f(Q)$.

5.1 The Wasserstein Distance

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

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

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

Endowed with the distance $\mathscr{W}_2(\nu_1, \nu_2)$, $\mathscr{P}_2(\mathbb{R}^d)$ is a metric space. Specifically, $\mathscr{W}_2(\nu_1, \nu_2)$ is a metrization of the weak topology in $\mathscr{P}_2(\mathbb{R}^d)$ [15, Thm. 6.9]. The space $(\mathscr{P}_2(\mathbb{R}^d), \mathscr{W}_2)$ is geodesic given that any two probability measures are connected by a minimizing geodesic and, moreover, if one the the measures is absolutely continuous with respect to the Lebesgue measure, this geodesic is unique [15, Cor. 7.22, Cor.7.23].

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

$$\mathscr{W}_2^2(w_1p_1 + w_2p_2, w_1q_1 + w_2q_2) \le w_1\mathscr{W}_2^2(p_1,q_1) + w_2\mathscr{W}_2^2(p_2,q_2) \ .$$

Proof. The bound on the right hand side is achieved by $\mathrm{E}[\|X-Y\|^2]$ when random variables (i,X,Y) are defined such that $\Pr\{X|i\} = p_i$, $\Pr\{Y|i\} = q_i$, $\Pr\{i\} = w_i$, and X and Y are independent given i. □

Proposition 4 (Sec. 6.2 in [16]; Sec. 2 in [17])**.** *Let $\gamma_{\nu_1,\nu_2}(t)$, $t \in [0,1]$ be a constant-speed geodesic curve from $\nu_1 \in \mathscr{P}_2(\mathbb{R}^d)$ to $\nu_2 \in \mathscr{P}_2(\mathbb{R}^d)$. Then, $\gamma_{\nu_1,\nu_2}(t)$ is also a barycenter of ν_1 and ν_2:*

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

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

The next result takes advantage of the fact that space $(\mathscr{P}_2(\mathbb{R}^d), \mathscr{W}_2)$ is a positively curved space 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 [18])**.** *For $w_1, w_2 \in [0,1]$, $w_2 = 1-w_1$, and probability measures ν_1, ν_2, $\nu \in \mathscr{P}_2(\mathbb{R}^d)$, the approximation error when the mixture $w_1\nu_1 + w_2\nu_2$ is replaced by ν is upper bounded as follows*

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

where γ is a geodesic curve as in Proposition 4. Moreover, the second inequality reduces to equality when the local curvature of $(\mathscr{P}_2(\mathbb{R}^d), \mathscr{W}_2)$ 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 $\bar{\mathscr{D}}_t(\nu_1, \nu_2) = t(1-t)\mathscr{W}^2(\nu_1, \nu_2)$.

Proposition 6 (Thm 2.2 [19],[20])**.** *The Wasserstein distance between two Gaussian distributions is given in closed-form by*

$$\mathscr{W}_2^2(\mathscr{N}(\mu_1, \Sigma_1), \mathscr{N}(\mu_2, \Sigma_2)) = \|\mu_1 - \mu_2\|^2 + \operatorname{tr}\Sigma_1 + \operatorname{tr}\Sigma_2 - 2\operatorname{tr}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2} \ .$$

Consider now the subspace $\mathscr{N}_0^d \subset \mathscr{P}_2(\mathbb{R}^d)$ 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 $\mathscr{N}_0^d$ is a totally geodesic submanifold of $\mathscr{P}_2(\mathbb{R}^d)$, i.e., any two points in $\mathscr{N}_0^d$ are connected by a geodesic that lies in $\mathscr{N}_0^d$.

Proposition 7 ([21], Example 1.7; [19]). *For $\mathscr{N}(0,V) \in \mathscr{N}_0^d$ and $\mathscr{N}(0,U) \in \mathscr{N}_0^d$, 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 $\mathscr{N}(0,\Gamma(t))$ is a geodesic from $\mathscr{N}(0,V)$ to $\mathscr{N}(0,U)$ in $(\mathscr{P}_2(\mathbb{R}^d), \mathscr{W}_2)$. In addition, $\Gamma(t)$ is itself a geodesic on the space $\mathbb{P}(d)$ endowed with the metric $\mathscr{W}_2(U,V) = \mathscr{W}_2(\mathscr{N}(0,U), \mathscr{N}(0,V))$.

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 $\mathscr{N}(\mu_1, U)$ to $\mathscr{N}(\mu_2, V)$ is given by $\mathscr{N}((1-t)\mu_1 + t\mu_2, \Gamma(t))$.

5.2 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. We investigate faster alternatives from approximations of the Wasserstein geodesic.

The next theorem shows that condition (5) is still satisfied when we replace the matrix geodesic mean by the matrix harmonic mean.

Theorem 8. *For positive definite matrices U and V, define their harmonic mean as*

$$H_{U,V}(t) := ((1-t)U^{-1} + tV^{-1})^{-1}, \ t \in [0,1]$$

and their arithmetic mean by

$$M_{U,V}(t) := (1-t)U + tV,\ t \in [0,1]\ .$$

Then,

$$(1-t)\mathscr{W}_2^2(U,X) + t\mathscr{W}_2^2(V,X) \leq \mathscr{W}_2^2(H_{U,V}(t),X) + \operatorname{tr} M_{U,V}(t) - \operatorname{tr} H_{U,V}(t)\ .$$

Proof. Using the fact that the derivative of the functional $X \mapsto \mathscr{W}_2^2(U,X)$ is $(I - U\#X^{-1})$ (see Section 6 of [20]), where the operator # denotes the matrix geometric mean, we have that the derivative of the functional

$$\varphi(X) := (1-t)\mathscr{W}_2^2(U,X) + t\mathscr{W}_2^2(V,X)$$

is given by

$$D_\varphi(X) = (1-t)(I - U\#X^{-1}) + t(I - V\#X^{-1})\ .$$

From the properties of matrix means given in [22], we have

$$\begin{aligned} D_\varphi(X) &= (1-t)(I - (U^{-1}\#X)^{-1}) + t(I - (V^{-1}\#X)^{-1}) \\ &\leq I - \left[(1-t)(U^{-1}\#X) + t(V^{-1}\#X)\right]^{-1} \\ &\leq I - \left[((1-t)U^{-1} + tV^{-1})\#X)\right]^{-1} \\ &= I - ((1-t)U^{-1} + tV^{-1})^{-1}\#X\ , \end{aligned} \tag{16}$$

where the inequalities (equality) follow, respectively, from the properties of self-duality of the geometric mean, minorization of the arithmetic mean by the harmonic mean, joint concavity of the geometric mean and self-duality of the geometric mean. The right-hand side of (16) is the derivative of the functional

$$\bar{\varphi}(X) := \mathscr{W}_2^2(H_{U,V}(t), X) \ .$$

This implies that the functional derivative $D_{\varphi-\bar{\varphi}}(X)$ is negative semidefinite. Therefore, the maximum of the functional $\varphi(X) - \bar{\varphi}(X)$ is attained for $X = 0$. It follows that

$$\begin{aligned} &(1-t)\mathscr{W}_2^2(U,X) + t\mathscr{W}_2^2(V,X) - \mathscr{W}_2^2(H_{U,V}(t),X) \\ &\leq (1-t)\mathscr{W}_2^2(U,0) + t\mathscr{W}_2^2(V,0) - \mathscr{W}_2^2(H_{U,V}(t),0) = (1-t)\operatorname{tr} U + t \operatorname{tr} V - \operatorname{tr} H_{U,V}(t) \ . \end{aligned}$$

□

This result also holds in the case of non-zero-mean Gaussians since, for $\nu = \mathscr{N}(\mu, X)$,

$$\begin{aligned} (1-t)\mathscr{W}_2^2(\mathscr{N}(\mu_1, U), \nu) + t\mathscr{W}_2^2(\mathscr{N}(\mu_2, V), \nu) - \mathscr{W}_2^2(\mathscr{N}((1-t)\mu_1 + t\mu_2, H_{U,V}(t)), \nu) \\ = (1-t)\mathscr{W}_2^2(U,X) + t\mathscr{W}_2^2(V,X) - \mathscr{W}_2^2(H_{U,V}(t),X) \ , \end{aligned}$$

which is a consequence of the law of cosines and the planar geometry nature of the contribution of the means to the Wasserstein distance.

The following proposition shows that, on the other hand, condition (5) does not hold when the merging function is the arithmetic mean. Nevertheless, the arithmetic mean still gives better approximations in the Wasserstein sense than the moment preserving merge of the original Runnalls' method.

Proposition 9. *For positive definite matrices U and V, their arithmetic mean is such that*

$$M_{U,V}(t) \geq \Gamma(t), \ t \in [0,1] \ ,$$

where $\Gamma(t)$ is the geodesic matrix in Proposition 7. Moreover, any matrix $X \geq$

$M_{U,V}(t)$ is farther than $M_{U,V}(t)$ from the barycenter $\Gamma(t)$ in the sense that

$$(1-t)\mathscr{W}_2^2(U,M_{U,V}(t))+t\mathscr{W}_2^2(V,M_{U,V}(t))\leq(1-t)\mathscr{W}_2^2(U,X)+t\mathscr{W}_2^2(V,X)\ .$$

On the other hand, for positive semidefinite X, the positive valued map

$$X\mapsto(1-t)\mathscr{W}_2^2(U,X)+t\mathscr{W}_2^2(V,X)-\mathscr{W}_2^2(M_{U,V}(t),X)$$

is unbounded when $U\neq V$ and $t\in(0,1)$.

Proof. The first statement is found in [20, Theorem 6]. For the second part, we consider once again the derivative of the functional $\varphi(X)$:

$$D_\varphi(X)=(1-t)(I-U\#X^{-1})+t(I-V\#X^{-1})\geq I-((1-t)U+tV)\#X^{-1}\ ,$$

where the inequality follows from the concavity of the matrix geometric mean.

From the monotonicity of the geometric mean, if $X\geq(1-t)U+tV$, then

$$D_\varphi(X)\geq I-((1-t)U+tV)\#((1-t)U+tV)^{-1}=0\ .$$

For the third part, we use the strict concavity of $Z\mapsto\mathrm{tr}(Z^{1/2})$ [20, Theorem 7] to find that, for $\alpha>0$ and $t\in(0,1)$,

$$\begin{aligned}&(1-t)\mathscr{W}_2^2(U,\alpha X)+t\mathscr{W}_2^2(V,\alpha X)-\mathscr{W}_2^2(M_{U,V}(t),\alpha X)\\&=2\alpha^{1/2}\left(-(1-t)\,\mathrm{tr}(X^{1/2}UX^{1/2})^{1/2}-t\,\mathrm{tr}(X^{1/2}VX^{1/2})^{1/2}+\mathrm{tr}(X^{1/2}M_{U,V}(t)X^{1/2})^{1/2}\right)\\&\qquad>2\alpha^{1/2}\epsilon\end{aligned}$$

for some $\epsilon>0$. Taking α to infinity we see that the functional is unbounded. □

Despite the unboundedness of $\bar{\mathscr{D}}_t$ in the case of the merge with the arithmetic mean, it turns out that the arithmetic mean gives a tighter approximation of the Wasserstein geodesic for small distances. This assertion is related to the following theorem.

Theorem 10. *For positive definite matrices Σ_1 and Σ_2, the Wasserstein distance between them is bounded by*

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

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 $\mathscr{W}_2$ 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 that induces $\mathscr{W}_2$ and that is given in [20, 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 $4\sigma_i\sigma_j \leq (\sigma_i + \sigma_j)^2$, we have that

$$g_\Sigma(U,U) = \sum_{i=1}^{d}\sum_{j=1}^{d} \frac{\sigma_i\sigma_j}{(\sigma_i + \sigma_j)^2}\sigma_j^{-1}u_{ij}^2 \leq \sum_{i=1}^{d}\sum_{j=1}^{d} \frac{1}{4}\sigma_j^{-1}u_{ij}^2 = \frac{1}{4}\operatorname{tr} U\Sigma^{-1}U \ . \quad (17)$$

Since $\operatorname{tr} U\Sigma^{-1}U$ is invariant under similarity transformations, 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 (17) that the two metrics coincide in the scalar case.

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:

$$\begin{aligned}\mathscr{W}_2^2(\Sigma_1,\Sigma_2) &= \left(\int_0^1 \sqrt{g_{\Gamma(t)}(\dot\Gamma(t),\dot\Gamma(t))}\,dt\right)^2\\ &\leq \left(\int_0^1 \sqrt{g_{\gamma(t)}(\dot\gamma(t),\dot\gamma(t))}\,dt\right)^2 \leq \int_0^1 g_{\gamma(t)}(\dot\gamma(t),\dot\gamma(t))\,dt \ ,\end{aligned}$$

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

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

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

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

□

From Theorem 10 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 approximation justifies the use of the arithmetic mean as a merging function with

$$\begin{aligned}&\bar{\mathscr{D}}_t(\Sigma_1,\Sigma_2) = t^2(1-t)\min\left\{\mathrm{tr}(\Sigma_1-\Sigma_2)\Sigma_1^{-1}(\Sigma_1-\Sigma_2), \mathrm{tr}(\Sigma_1-\Sigma_2)M_{\Sigma_1,\Sigma_2}(t)^{-1}(\Sigma_1-\Sigma_2)\right\}\\ &+t(1-t)^2\min\left\{\mathrm{tr}(\Sigma_1-\Sigma_2)\Sigma_2^{-1}(\Sigma_1-\Sigma_2), \mathrm{tr}(\Sigma_1-\Sigma_2)M_{\Sigma_1,\Sigma_2}(t)^{-1}(\Sigma_1-\Sigma_2)\right\} \ ,\end{aligned}$$

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

$$\min\{\Sigma_1^{-1}, M_{\Sigma_1,\Sigma_2}(t)^{-1}\} \approx \mathrm{diagm}(\max\{\mathrm{diag}(\Sigma_1)^{-1}, \mathrm{diag}(M_{\Sigma_1,\Sigma_2}(t))^{-1}\}) \ , \quad (18)$$

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

5.3 Controlling the Mean Absolute Estimation Error

In this section we show how a proper choice of Wasserstein distance may be used to control the mean absolute estimation error. To this purpose, we extend the definition of $\mathscr{W}_2$ above 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 2 to take advantage of its mode-independent dynamics in order to obtain formal bounds between the time-averaged mean absolute estimation error and the Wasserstein distance.

Consider the approximation of the N-component Gaussian mixture

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

by N_c clusters as given by the probability density

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

such that, for each cluster C_j, $\tilde{w}_j = \sum_{i \in C_j} w_i$, $\tilde{\mu}_j = \sum_{i \in C_j} w_i / \tilde{w}_j \; \mu_i$ 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. Define the likelihood function for the output sequence $y_{1:k}$ and the n-th possible mode sequence $m_{1:k}^{(n)}$, $n = 1, \ldots, M^k$, 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 $\tilde{\ell}_{i,k,n}$ 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:

$$\begin{aligned} x_{k|k} - \tilde{x}_{k|k} &= \sum_{i,n} \frac{w_i \ell_{i,k,n}}{\ell_k} \mu_{i,k,n} - \sum_{j,n} \frac{\tilde{w}_j \tilde{\ell}_{j,k,n}}{\tilde{\ell}_k} \tilde{\mu}_{j,k,n} \\ &= \sum_{j,n} \sum_{i \in C_j} \frac{w_i \ell_{i,k,n}}{\ell_k} (\mu_{i,k,n} - \tilde{\mu}_{j,k,n}) - \sum_{j,n} \left(\frac{\tilde{w}_j \tilde{\ell}_{j,k,n}}{\tilde{\ell}_k} - \sum_{i \in C_j} \frac{w_i \ell_{i,k,n}}{\ell_k} \right) \tilde{\mu}_{j,k,n} \ , \end{aligned}$$

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

Now, recall that the estimation error dynamics for a Kalman filter is given by

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

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, by denoting mean deviations by $\Delta\mu_{i,k} = (A - LCA)^k(\mu_i - \tilde{\mu}_j)$, $i \in C_j$, and defining $\bar{\ell}_{j,k,n} = \sum_{i \in C_j} w_i \ell_{i,k,n}/\ell_k$, we can write

$$\begin{aligned} x_{k|k} - \tilde{x}_{k|k} &= \sum_{j,n}\sum_{i \in C_j} \frac{w_i \ell_{i,k,n}}{\ell_k} \Delta\mu_{i,k} - \sum_{j,n} \tilde{w}_j \left(\frac{\tilde{\ell}_{j,k,n}}{\tilde{\ell}_k} - \frac{\bar{\ell}_{j,k,n}}{\ell_k} \right) \tilde{\mu}_{j,k,n} \\ &= \sum_{j,n}\sum_{i \in C_j} w_i \frac{\ell_{i,k,n} - \tilde{\ell}_{j,k,n}}{\ell_k} \Delta\mu_{i,k} - \sum_{j,n} \tilde{w}_j \left(\frac{\tilde{\ell}_{j,k,n}}{\tilde{\ell}_k} - \frac{\bar{\ell}_{j,k,n}}{\ell_k} \right) (\tilde{\mu}_{j,k,n} - \tilde{\mu}_k) \ \ , \end{aligned}$$

where in the last equality we used the fact that $\sum_{i \in C_j} w_i \Delta\mu_{i,k} = 0$ and introduced the full posterior mean $\tilde{\mu}_k = \sum_j \tilde{w}_{j,n} \tilde{\ell}_{j,k,n} \tilde{\mu}_{j,k,n}$.

Taking the matrix weighted norm and expectations on $y_{1:k}$ and $m_{1:k}$, we have

$$\begin{aligned} \mathrm{E}[\|x_{k|k} - \tilde{x}_{k|k}\|_Q] \leq & \sum_{j,n}\sum_{i \in C_j} w_i \int |\ell_{i,k,n} - \tilde{\ell}_{j,k,n}| dy_{1:n} \ \|\Delta\mu_{i,k}\|_Q \\ & + \int \left(\sum_{j,n} \tilde{w}_j \tilde{\ell}_{j,k,n} \left| \frac{\bar{\ell}_{j,k,n}}{\tilde{\ell}_{j,k,n}} - \frac{\ell_k}{\tilde{\ell}_k} \right| \|\tilde{\mu}_{j,k,n} - \tilde{\mu}_k\|_Q \right) dy_{1:n} \ \ . \qquad (19) \end{aligned}$$

Using the triangle inequality and the convexity of $|\cdot|$, the last term can be bounded by

$$\begin{aligned}\int \sum_{j,n} \tilde{w}_j \tilde{\ell}_{j,k,n} \left(\left| \frac{\bar{\ell}_{j,k,n}}{\tilde{\ell}_{j,k,n}} - 1 \right| + \left| \frac{\ell_k}{\tilde{\ell}_k} - 1 \right| \right) \|\tilde{\mu}_{j,k,n} - \tilde{\mu}_k\|_Q dy_{1:n} \\ \leq \int \sum_{j,n} \tilde{w}_j \sum_{i \in C_j} \frac{w_i}{\tilde{w}_j} \tilde{\ell}_{j,k,n} \left(1 + \frac{\tilde{w}_j \tilde{\ell}_{j,k,n}}{\tilde{\ell}_k} \right) \left| \frac{\ell_{i,k,n}}{\tilde{\ell}_{j,k,n}} - 1 \right| \|\tilde{\mu}_{j,k,n} - \tilde{\mu}_k\|_Q dy_{1:n} \\ \leq 2 \sum_j \tilde{w}_j \left(\sum_{i \in C_j} \frac{w_i}{\tilde{w}_j} \chi^2(\tilde{\ell}_{j,k,n}, \ell_{i,k,n}) \right)^{1/2} \mathsf{Var}_Q(\tilde{\mu}_k)^{1/2} \ ,\end{aligned}$$

where the last inequality follows from Hölder's inequality and where $\mathsf{Var}_Q(\tilde{\mu}_k)$ is the expected variance of the cluster centers at time k for the prior $\tilde{\phi}$.

In order to compute the divergences between $\ell_{i,k,n}$ and $\tilde{\ell}_{j,k,n}$, 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 the likelihoods, first notice that

$$\begin{aligned}\ell_{i,k,n} &= \int p\left(y_{1:k}, m_{1:k}^{(n)} | m_0, x_0\right) \phi_i(x_0) dx_0 \\ &= \int \sum_{m_1=1}^{M} p\left(y_{1:k}, m_{2:k}^{(n)} | m_1^{(n)}, x_1\right) \pi_{m_1^{(n)}|m_0} \phi_i^+(x_1|m_1^{(n)}) dx_1 \ ,\end{aligned}$$

where $\phi_i^+(x_1|m_1)$ denotes the prior probability of x_1 given $x_0 \sim \phi_i$ and m_1. From the convexity of the map $(p,q) \mapsto qf(p/q)$, we may pull out the integration on $\pi_{m_1^{(n)}|m_0}\tilde{\phi}_i^+(x_1)dx_1$ to obtain

$$\int \sum_{m_k^{(n)}} \tilde{\ell}_{j,k,n} f\left(\frac{\ell_{i,k,n}}{\tilde{\ell}_{j,k,n}}\right) dy_{1:n} \leq \sum_{m_1^{(n)}} \pi_{m_1^{(n)}|m_0} \int \tilde{\phi}_j^+(x_1|m_1^{(n)}) f\left(\frac{\phi_i^+(x_1|m_1^{(n)})}{\tilde{\phi}_j^+(x_1|m_1^{(n)})}\right) dx_1$$

From this, 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 (15) to obtain

$$\begin{aligned}\chi^2(\tilde{\ell}_{j,k,n}, \ell_{i,k,n}) \leq \chi^2(\tilde{\phi}_j^+, \phi_i^+) &= \exp\left(\Delta\mu_{i,0}' A' \bar{\Sigma}_+^{-1} A \Delta\mu_{i,0}\right) - 1 \\ &= \Delta\mu_{i,0}' A' \bar{\Sigma}_+^{-1} A \Delta\mu_{i,0} + \mathscr{O}((\Delta\mu_{i,0}' A' \bar{\Sigma}_+^{-1} A \Delta\mu_{i,0})^2) \ . \quad (20)\end{aligned}$$

From (14) and the inequality between the Hellinger divergence and the total variation [23], we obtain:

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

Replacing these bounds on (19) and applying Hölder's inequality once again, we find that

$$\mathrm{E}[\|x_{k|k} - \tilde{x}_{k|k}\|_Q] \leq \sum_{j=1}^{N_c} \tilde{w}_j \left[\left(\sum_{i \in C_j} \frac{w_i}{\tilde{w}_j} \left\|\Delta\mu_{i,k}\right\|_Q^2\right)^{1/2} + 2\sigma_0\right] \left(\sum_{i \in C_j} \frac{w_i}{\tilde{w}_j} \left\|\Delta\mu_{i,0}\right\|_{Q_+}^2\right)^{1/2} ,$$

where $Q_+ := A'\bar{\Sigma}_+^{-1}A$ and where we assumed that $\mathsf{Var}_Q(\tilde{\mu}_k)^{1/2}$ is at the equilibrium value σ_0.

To compute the discounted cost, we apply Young's inequality with factors $\lambda_1\beta_1^{k-1}$ and λ_2 to turn the products of the square roots into sums and find that

$$\begin{aligned}
&\sum_{k=1}^{\infty}\gamma^{k-1}\,\mathrm{E}[\|x_{k|k}-\tilde{x}_{k|k}\|_Q] \\
&\quad\leq \sum_{j=1}^{N_c}\tilde{w}_j\sum_{i\in C_j}\frac{w_i}{\tilde{w}_j}\sum_{k=1}^{\infty}\gamma^{k-1}\left[\frac{\lambda_1\beta_1^{k-1}}{2}\left\|\Delta\mu_{i,k}\right\|_Q^2+\frac{\lambda_1^{-1}\beta_1^{1-k}}{2}\left\|\Delta\mu_{i,0}\right\|_{Q_+}^2\right] \\
&\qquad+\sum_{j=1}^{N_c}\tilde{w}_j\sum_{i\in C_j}\frac{w_i}{\tilde{w}_j}\sum_{k=1}^{\infty}\gamma^{k-1}\left[\frac{\lambda_2}{2}4\sigma_0^2+\frac{\lambda_2^{-1}}{2}\left\|\Delta\mu_{i,0}\right\|_{Q_+}^2\right] \\
&\qquad\qquad=2\frac{\lambda_2}{1-\gamma}\sigma_0^2+\sum_{j=1}^{N_c}\tilde{w}_j\left(\sum_{i\in C_j}\frac{w_i}{\tilde{w}_j}\Delta\mu_{i,0}'H\Delta\mu_{i,0}\right)\ ,
\end{aligned}\tag{21}$$

where

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

and where, since $\Delta\mu_{i,k}=(A-LCA)^k\Delta\mu_{i,0}$, H_β satisfies the Lyapunov equation

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

The last term in (21) corresponds to the bound $\mathscr{E}_k$ in Section 3 that would be obtained when, replacing the Euclidean norm by the norm $\|\cdot\|_H$, $\mathscr{W}_2^2$ is the divergence function. Therefore, by picking an appropriate H-norm for the Wasserstein distance, we are able to control the mean absolute error for a given Q-norm.

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, but it may not be satisfied by Q_+. Nevertheless, this contraction property is true for the posterior covariance $\bar{\Sigma}^{-1}$, which could have been used instead of Q_+ in the derivations starting in (20).

In practice, we note that the last term in (19) is a bound on the covariance between the approximation error in cluster weights and the position of cluster centers. We expect the correlation coefficient between these variables to be small. Since this term takes an average over multiple weight errors that are not strongly correlated, it would be reasonable to expect a time-varying correlation coefficient $\rho_k = \rho_0/\sqrt{N_c M^k}$. This is in line with the known problem of weight degeneration in Bayesian filtering. Indeed, 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. This implies that ρ_k should become small very fast.

We may therefore replace σ_0 in (21) by $\rho_k\sigma_0$ and introduce a new factor β_2^{k-1} when applying Young's inequality. Then, in order to optimize the factors λ_1, β_1 and λ_2, β_2, we may take the expected value on $\Delta\mu_{i,0}$ in (21) and assume that

$$\sum_{j=1}^{N_c} \tilde{w}_j \left(\sum_{i \in C_j} \frac{w_i}{\tilde{w}_j} \Delta\mu'_{i,0} H \Delta\mu_{i,0} \right) \approx \frac{\operatorname{tr} H\Sigma_0}{N_c}$$

and

$$\sigma_0^2 = \operatorname{tr} Q\Sigma_0 \ ,$$

and then search for the factors that minimize such an expected value (note that the actual value of Σ_0 does not play an important role in this optimization nor does the value of N_c).

6 Numerical Experiments

We conducted numerical experiments for the system described in Section 2 for a marginally stable dynamics given by

$$A = \begin{bmatrix} 1 & T_s & T_s^2/2 \\ 0 & 1 & T_s \\ 0 & 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix},$$

where $T_s = 0.3$. The noise covariances were given by $R_v = 1/3 \cdot 10^{-4}$ and $R_w = \text{diagm}(10^{-6}, 10^{-4}, 10^{-5})$. The packet drop probability was $p_0 = 0.4$. The input u was given by the signal $u_k = 5 \cdot 10^{-4} \sin(2\pi k \ 20/T) + 7.5 \cdot 10^{-2} \tilde{w}$, where $\tilde{w}$ is a unit variance white Gaussian noise and where $T = 2000$ is the total simulation time. We run as many realizations as needed to attain 1% precision in the cost estimates. We considered average costs (discount factor $\gamma = 1.0$) instead of discounted costs.

We chose a matrix norm with $Q = \text{diagm}(1, 5, 10)\bar{\Sigma}^{-1}\text{diagm}(1, 5, 10)$, for $\bar{\Sigma}$ being the posterior covariance at equilibrium. This norm indicates that we give 5 and 10 times more importance to the estimation error of the velocity and the acceleration respectively.

For the sake of comparison, we give the Cholesky factors of the computed (normalized) H and $\bar{\Sigma}_+^{-1}$ matrices

$$\texttt{chol}(H) = \begin{bmatrix} 4.33 & -0.22 & -0.24 \\ 0 & 1.97 & -2.97 \\ 0 & 0 & 3.9 \end{bmatrix}, \quad \texttt{chol}(\bar{\Sigma}_+^{-1}) = \begin{bmatrix} 1.0 & -0.81 & -0.12 \\ 0 & 1.38 & -1.19 \\ 0 & 0 & 3.57 \end{bmatrix}.$$

It is noticeable that H gives more weight to the position variable. The angle $\angle(H, \bar{\Sigma}^{-1})$ as given by the trace inner product is of 45.5°, which demonstrates a substantial deviation from the behavior of the information divergences. On the other hand, $\angle(H,Q) = 24.7^{\circ}$. In addition, we have that $\Delta\mu' H \Delta\mu$ contracts with rate at most 0.87.

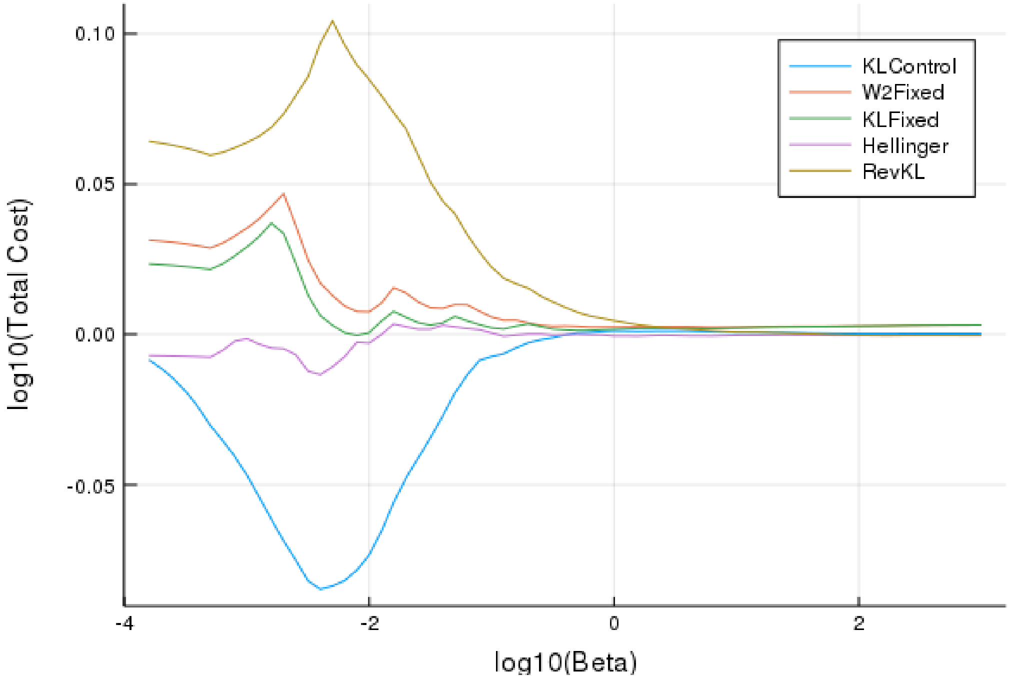

Figure 1: Best average cost achieved by each divergence as a function of the the processing time weight β. The labels W2Fixed and KLFixed refer to setting $\kappa_0 = 0$ in Algorithm 1 and varying $N_{\max}$ from to 2 to 15 components. The remaining curves refer to Algorithm 1 with $N_{\max} = 30$ and the best value of κ_0 for each β. Costs are normalized by the cost achieved by the Wasserstein distance (shows 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 so that there is no closed-loop precision control. For the case of the Wasserstein distance, we only present here the results for approximation given by (18) as they are significantly faster.

The results are summarized in Figures 1 and 2. A first conclusion is that controlling the number of components gives noticeable 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. The Wasserstein distance only improves over KL when very small errors are required. Still, the total cost for KL was not 0.2% larger than that of the Wasserstein distance. Finally, the reverse KL divergence gave the worst results.

To interpret such hierarchy of performances, recall the notion of entropic means discussed in Section 4. Notice that 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 KL divergence is the only information divergence with the correct target. Since the mean of the square roots is closer to the arithmetic mean than the geometric mean, we have that the Hellinger divergence provides better results than the reverse KL. Following the same reasoning, we may conjecture that the results for the χ^2-divergence should be even worse, since the harmonic mean is the smallest of the means above.

References

[1] Hans Driessen and Yvo Boers. Multiple-model multiple-hypothesis filter for tracking maneuvering targets. In *Signal and Data Processing of Small Targets 2001*, volume 4473, pages 279–289. International Society for Optics and Photonics, 2001.

[2] Yvo Boers and Hans Driessen. A multiple model multiple hypothesis filter for Markovian switching systems. *Automatica*, 41(4):709–716, 2005.

[3] WY Eras-Herrera, AR Mesquita, and BOS Teixeira. Multiple-model multiple-hypothesis filter with Gaussian mixture reduction. *International Journal of Adaptive Control and Signal Processing*, 32(2):286–300, 2018.

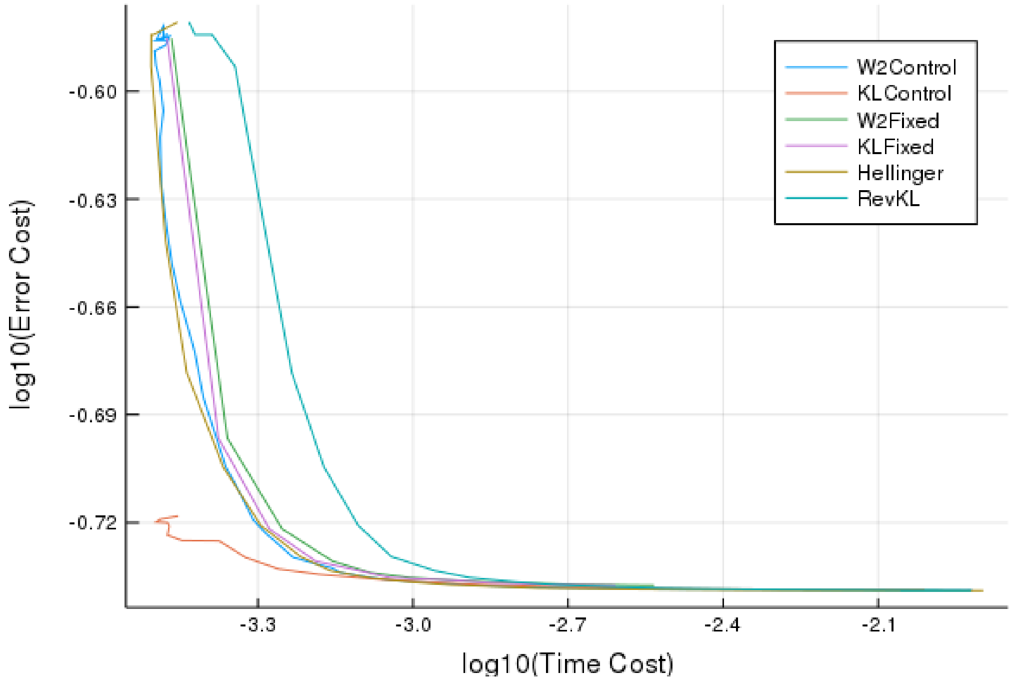

Figure 2: Estimation error cost versus processing time cost for different parameters κ_0 and $N_{\max}$ in Algorithm 1 as described for the curves in Figure 1.

[4] David F Crouse, Peter Willett, Krishna Pattipati, and Lennart Svensson. A look at Gaussian mixture reduction algorithms. In *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*, pages 1–8. IEEE, 2011.

[5] Andrew R Runnalls. Kullback-Leibler approach to Gaussian mixture reduction. *IEEE Transactions on Aerospace and Electronic Systems*, 43(3), 2007.

[6] Alexandre R Mesquita, João P Hespanha, and Girish N Nair. Redundant data transmission in control/estimation over lossy networks. *Automatica*, 48(8):1612–1620, 2012.

[7] Joao P Hespanha and Alexandre R Mesquita. Networked control systems: estimation and control over lossy networks. *Encyclopedia of Systems and Control*, pages 842–849, 2015.

[8] Igor Vajda. On metric divergences of probability measures. *Kybernetika*, 45(6):885–900, 2009.

[9] Aharon Ben-Tal, Abraham Charnes, and Marc Teboulle. Entropic means. 1989.

[10] Bruno Pelletier. Informative barycentres in statistics. *Annals of the Institute of Statistical Mathematics*, 57(4):767–780, 2005.

[11] Frank Nielsen and Richard Nock. Sided and symmetrized Bregman centroids. *IEEE transactions on Information Theory*, 55(6):2882–2904, 2009.

[12] Frank Nielsen and Richard Nock. The entropic centers of multivariate normal distributions. *Collection of Abstracts*, 221, 2008.

[13] Frank Nielsen and Sylvain Boltz. The Burbea-Rao and Bhattacharyya centroids. *IEEE Transactions on Information Theory*, 57(8):5455–5466, 2011.

[14] Frank Nielsen and Richard Nock. On the chi square and higher-order chi distances for approximating f-divergences. *IEEE Signal Processing Letters*, 21(1):10–13, 2014.

[15] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[16] Martial Agueh and Guillaume Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[17] Pedro C Álvarez-Esteban, E del Barrio, JA Cuesta-Albertos, and C Matrán. A fixed-point approach to barycenters in Wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016.

[18] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[19] Asuka Takatsu et al. Wasserstein geometry of Gaussian measures. *Osaka Journal of Mathematics*, 48(4):1005–1026, 2011.

[20] Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 2018.

[21] Robert J McCann. A convexity principle for interacting gases. *Advances in Mathematics*, 128(1):153–179, 1997.

[22] Tsuyoshi Ando, Chi-Kwong Li, and Roy Mathias. Geometric means. *Linear algebra and its applications*, 385:305–334, 2004.

[23] Igal Sason and Sergio Verdu. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.