Knowledge Transfer in a Pair of Uniformly Modelled Bayesian Filters

Ladislav Jirsa[1], Lenka Kuklišová Pavelková[1] [a], and Anthony Quinn[1,2] [b]
[1] *Institute of Information Theory and Automation, The Czech Academy of Sciences, Prague, Czech Republic*
[2] *Trinity College Dublin, The University of Dublin, Ireland*
{*jirsa, pavelkov*}*@utia.cas.cz, aquinn@tcd.ie*

Keywords: Fully Probabilistic Design, Bayesian Filtering, Uniform Noise, Knowledge Transfer, Predictor, Orthotopic Bounds.

Abstract: The paper presents an optimal Bayesian transfer learning technique applied to a pair of linear state-space processes driven by uniform state and observation noise processes. Contrary to conventional geometric approaches to boundedness in filtering problems, a fully Bayesian solution is adopted. This provides an approximate uniform filtering distribution and associated data predictor by processing the involved bounds via a local uniform approximation. This Bayesian handling of boundedness provides the opportunity to achieve optimal Bayesian knowledge transfer between bounded-error filtering nodes. The paper reports excellent rejection of knowledge below threshold, and positive transfer above threshold. In particular, an informal variant achieves strong transfer in this latter regime, and the paper discusses the factors which may influence the strength of this transfer.

1 INTRODUCTION

The problem of data (also known as measurement or information) fusion is now key in many areas of industry, driven by the IoT and Industry 4.0 agendas (Diez-Olivan et al., 2019). Data fusion systems are required in areas such as sensor networks, robotics, and video and image processing systems (Khaleghi et al., 2013). Applications range from health monitoring (Vitola et al., 2017) and environmental sensing (Zou et al., 2017), to cooperative systems for indoor tracking (Dardari et al., 2015) or urban vehicular localization (Nassreddine et al., 2010). In all cases, information about the same quantity of interest (or several quantities of interest, related in a specified way) is acquired by different types of detectors, under different conditions, via multiple experiments and measurement devices. (Lahat et al., 2015) provides an overview of the main challenges in multimodal data fusion across various disciplines, while the survey in (Gravina et al., 2017) discusses clear motivations and advantages of multi-sensor data fusion.

Diverse approaches to solving data fusion problems have been proposed in the literature, very often driven by the specifics of the application, and it is clear that no universal definition—much less a universal solution—has yet been accepted. Wiener-type criteria for measurement vector fusion (Willner et al., 1976) have long been available, while direct variants of the Kalman filter have also been proposed for fusion of measurements (Dou et al., 2016) and states (Yang et al., 2019). Artificial intelligence (Xiao, 2019), machine learning (Vitola et al., 2017; Shamshirband et al., 2015; Vapnik and Izmailov, 2017) and expert system approaches (Majumder and Pratihar, 2018) have all yielded progress in this area.

In the Bayesian framework which we adopt in this paper, concepts of data, knowledge and information are all unified via a probabilistic representation, being typically the distribution of the unknown quantities of interest, conditioned by data. Each measurement process (node or sensor) constructs its local distribution conditioned on its locally sensed data. The inferred quantity may be globally shared by all the nodes, or these may be distinct but related quantities (Azizi and Quinn, 2018). The task is then one of optimal merging of these local distributions, a problem closely related to distributional pooling (Abbas, 2009). Another closely related problem is that of knowledge transfer (Torrey and Shavlik, 2010). Again, in the Bayesian context of this paper, the requirement is to transfer probabilistic knowledge (a distribution) from secondary (i.e. external or source) nodes to a pri-

[a] https://orcid.org/0000-0001-5290-2389
[b] https://orcid.org/0000-0002-3873-7556

mary node, to support its inference of the quantity of interest. Optimal Bayesian transfer learning via Kullback-Leibler divergence (KLD) minimization—known as fully probabilistic design (FPD) (Quinn et al., 2016)—is developed in (Quinn et al., 2017), and specialized to the case of data-predictive transfer between Kalman filters in (Foley and Quinn, 2018), (Papež and Quinn, 2018). In this approach, the secondary data predictor is transferred to the primary filtering node via an optimal mean-field-type operator. This is shown to improve the primary state reconstruction in positive transfer cases, and rejects the secondary knowledge otherwise. In addition to the consistency and optimality properties of this Bayesian transfer learning framework, a key practical advantage here is that the relationship between the primary and secondary processes does not have to be explicitly modelled.

In this paper, we address the important practical context where the involved uncertainties are bounded. Contemporary examples of sensor networks with bounded innovations and/or observation processes are studied in (Goudjil et al., 2017; He et al., 2017; Vargas-Melendez et al., 2017). Specifically, we extend the Bayesian transfer learning technique in (Quinn et al., 2017) to a pair of linear state-space processes driven by uniform state and observation noise processes. Boundedness in filtering problems is handled via several approaches. Within the Kalman filtering set-up, the state estimates are projected onto the constraint surface (Fletcher, 2000) or the Gaussian distribution is truncated (Simon and Simon, 2010). Using sequential Monte-Carlo sampling methods, the constraints are respected within the accept/reject steps of the algorithm, see e.g. (Lang et al., 2007). The non-probabilistic techniques dealing with "unknown-but-bounded error" provide an approximate set containing the estimates (Chisci et al., 1996), (Becis-Aubry et al., 2008). Recently, a fully Bayesian solution has yielded a sequentially uniform filtering distribution (Pavelková and Jirsa, 2018) and associated data predictor (Jirsa et al., 2019) by processing the involved bounds via a local uniform approximation of the state predictor at each filtering step, achieving a recursive, tractable algorithm. This Bayesian handling of boundedness now provides the opportunity for optimal Bayesian knowledge transfer between bounded-error filtering nodes, as set out below.

Throughout the paper, z_t is the value of a column vector z at a discrete time instant $t \in t^\star \equiv \{1,2,\ldots,\bar{t}\}$; $z_{t;i}$ is the i-th entry of z_t; $z(t) \equiv \{z_t, z_{t-1}, \ldots, z_1, z_0\}$; $\underline{z}$ and $\bar{z}$ are lower and upper bounds on z, respectively; $\equiv$ means equality by definition, $\propto$ means equality up to a constant factor. The symbol $f(\cdot|\cdot)$ denotes a conditional probability density function (pdf); names of arguments distinguish respective pdfs; no formal distinction is made between a random variable, its realisation and an argument of the pdf. $\mathcal{U}_z(\underline{z},\bar{z})$ denotes the uniform pdf of z with the orthotopic support $[\underline{z},\bar{z}]$. Also, $\chi(x \in x^\star)$ is the indicator function, i.e. if $x \in x^\star$, then $\chi(x \in x^\star) = 1$, otherwise $\chi(x \in x^\star) = 0$.

2 UNIFORMLY MODELLED BAYESIAN FILTERS: KNOWLEDGE REPRESENTATION AND TRANSFER

We consider a pair of Bayesian filters, and distinguish between the *primary* filter (called the target node in transfer learning (Torrey and Shavlik, 2010) and the *secondary* filter (sometimes called the source node), with all sequences and distributions subscripted by "s" in the latter case. Each filter processes its local observation sequence, y_t and $y_{s,t}$, $t \in t^\star$, respectively, informative of their local, hidden (state) process, x_t and $x_{s,t}$, respectively.

In this paper, we adopt linear stochastic state-space models with known parameters for all four processes, each driven by an independent, uniformly-distributed white noise process of known parameters. We summarize (isolated) Bayesian filtering for this class of model in Section 2.1. We briefly review probabilistic knowledge transfer between a pair of general Bayesian filters in Section 2.2, using the axiomatically optimal FPD principle. Then, in Section 2.3, we instantiate this FPD-optimal transfer in the current context of uniformly driven Bayesian filters, recovering a tractable, recursive flow via a local uniform approximation at each step.

2.1 Bayesian filtering with uniformly distributed noise processes

In the considered Bayesian set up (Kárný et al., 2005), the system is described by the following pdfs:

$$\begin{aligned} \text{prior pdf:} \quad & f(x_0) \\ \text{observation model:} \quad & f(y_t|x_t) \\ \text{time evolution model:} \quad & f(x_t|x_{t-1},u_{t-1}) \end{aligned} \tag{1}$$

where y_t is a scalar observable output, u_t is a known system input (optional, for generality), and x_t is an ℓ-dimensional unobservable system state, $t \in t^\star$.

We assume that (i) hidden process x_t satisfies the Markov property, (ii) no direct relationship between

input and output exists in the observation model, and (iii) the inputs consist of a known sequence $u_0, u_1, \ldots, u_{\bar{t}-1}$.

Bayesian sequential state estimation (known as filtering) consists of the evolution of the posterior pdf $f(x_t|d(t))$ where $d(t)$ is a sequence of observed data records $d_t = (y_t, u_t)$, $t \in t^\star$. The evolution of $f(x_t|d(t))$ is described by a two-steps recursion that starts from the prior pdf $f(x_0)$:

Time update $f(x_t|d(t-1)) =$

$$= \int_{x_{t-1}^\star} f(x_t|u_{t-1}, x_{t-1}) f(x_{t-1}|d(t-1))\,\mathrm{d}x_{t-1}, \tag{2}$$

Data update $f(x_t|d(t)) =$

$$\frac{f(y_t|x_t) f(x_t|d(t-1))}{\int_{x_t^\star} f(y_t|x_t) f(x_t|d(t-1))\mathrm{d}x_t} = \frac{f(y_t|x_t) f(x_t|d(t-1))}{f(y_t|d(t-1))}. \tag{3}$$

We introduce a linear state space model with a uniform noise (LSU model) in the form

$$\begin{aligned} f(x_t|u_{t-1}, x_{t-1}) &= \mathcal{U}_x(\tilde{x}_t - \rho, \tilde{x}_t + \rho) \\ f(y_t|x_t) &= \mathcal{U}_y(\tilde{y}_t - r, \tilde{y}_t + r). \end{aligned} \tag{4}$$

where $\tilde{x}_t = Ax_{t-1} + Bu_{t-1}$, $\tilde{y}_t = Cx_t$, A, B, C are the known model matrices/vectors of appropriate dimensions, $\nu_t \in (-\rho, \rho)$ is the uniform state noise with known parameter ρ, $n_t \in (-r, r)$ is the uniform observation noise with known parameter r.

State estimation of LSU model (4) according to (2) and (3) leads to a very complex form of posterior pdf. In (Pavelková and Jirsa, 2018)[1], an approximate Bayesian state estimation of this model (based on a minimising of Kullback-Leibler divergence of two pdfs) is proposed. The presented algorithm provides the evolution of the approximate posterior pdf $f(x_t|d(t))$ that is uniformly distributed on a parallelotopic support.

Approximate time update The time update starts at the time $t = 1$ with $f(x_{t-1}|d(t-1)) = f(x_0) = \mathcal{U}_{x_0}(\underline{x}_0, \overline{x}_0)$. Being χ the indicator function, it holds

$$f(x_t|d(t-1)) \approx \prod_{i=1}^{\ell} \frac{\chi(\underline{m}_{t;i} - \rho_i \le x_{t;i} \le \overline{m}_{t;i} + \rho_i)}{\overline{m}_{t;i} - \underline{m}_{t;i} + 2\rho_i} =$$

$$= \prod_{i=1}^{\ell} \mathcal{U}_{x_{t;i}}(\underline{m}_{t;i} - \rho_i, \overline{m}_{t;i} + \rho_i) = \mathcal{U}_{x_t}(\underline{m}_t - \rho, \overline{m}_t + \rho), \tag{5}$$

[1]Note that the paper (Pavelková and Jirsa, 2018) contains the typo in the formula (3): the relevant integration variable should be x_t instead of x_{t-1}.

where $\underline{m}_t = [\underline{m}_{t;1}, \ldots, \underline{m}_{t;\ell}]'$, $\overline{m}_t = [\overline{m}_{t;1}, \ldots, \overline{m}_{t;\ell}]'$,

$$\underline{m}_{t;i} = \sum_{j=1}^{\ell} \min(A_{ij}\underline{x}_{t-1;j} + B_i u_{t-1}, A_{ij}\overline{x}_{t-1;j} + B_i u_{t-1}), \tag{6}$$

$$\overline{m}_{t;i} = \sum_{j=1}^{\ell} \max(A_{ij}\underline{x}_{t-1;j} + B_i u_{t-1}, A_{ij}\overline{x}_{t-1;j} + B_i u_{t-1}),$$

A_{ij} means the term on the i-th row and the j-th column of A.

Approximate data update According to (3), we process the observation y_t as $y_t - r \le Cx_t \le y_t + r$ (see (4)) by the Bayes rule together with the prior (5) from the time update. The approximate posterior pdf is uniformly distributed on a parallelotopic support

$$f(x_t|d(t)) \approx K_t \chi(\underline{x}_t \le M_t x_t \le \overline{\overline{x}}_t), \tag{7}$$

where K_t is a normalising constant.

However, the time update (5) in the next step assumes pdf with an orthotopic support. Therefore, the parallelotope $\underline{x}_t \le M_t x_t \le \overline{\overline{x}}_t$ in (7) is circumscribed by an orthotope $\underline{x}_t \le x_t \le \overline{x}_t$. In this way, we obtain an approximate "orthotopic" posterior pdf

$$f(x_t|d(t)) \approx \mathcal{U}_{x_t}(\underline{x}_t, \overline{x}_t). \tag{8}$$

The orthotopic bounds $\underline{x}_t$ and $\overline{x}_t$ are used in the next time update step (5) for the computation of the terms $\underline{m}$ and $\overline{m}$ (6) and for the computation of point estimates of states $\hat{x}_t$

$$\hat{x}_t = \frac{\underline{x}_t + \overline{x}_t}{2}. \tag{9}$$

Predictive pdf for LSU model The data predictor of a linear state-space model is the denominator of (3), where $f(y_t|x_t)$ is given by (4) and $f(x_t|d(t-1))$ is the result of the approximate time update (5). The approximate uniform predictor as proposed in (Jirsa et al., 2019) is

$$f(y_t|d(t-1)) \approx \left(\overline{y}_t - \underline{y}_t\right)^{-1} \chi\left(\underline{y}_t \le y_t \le \overline{y}_t\right), \tag{10}$$

where

$$\underline{y}_t = C\underline{s} - r \tag{11}$$

$$\overline{y}_t = C\overline{s} + r \tag{12}$$

for $\underline{s}$ and $\overline{s}$ defined so that

$$\begin{aligned} \underline{s}_i &= \underline{m}_{t;i} - \rho_i, \quad \overline{s}_i = \overline{m}_{t;i} + \rho_i, \quad \text{if } C_i \ge 0, \\ \underline{s}_i &= \overline{m}_{t;i} + \rho_i, \quad \overline{s}_i = \underline{m}_{t;i} - \rho_i, \quad \text{if } C_i < 0, \end{aligned} \tag{13}$$

This predictive pdf is conditioned by $\underline{m}_t$ and $\overline{m}_t$ considered as statistics, provided the parameters A and B be known.

Mean value $\hat{y}_t \equiv \mathrm{E}[y_t|d(t-1)]$ of (10) is

$$\mathrm{E}[y_t|d(t-1)] = \frac{\overline{y}_t + \underline{y}_t}{2} = C \underbrace{\left(\frac{\overline{m}_t + \underline{m}_t}{2}\right)}_{\mathrm{E}[x_t|d(t-1)]}. \tag{14}$$

2.2 FPD-optimal knowledge transfer between Bayesian filters

If knowledge transfer from the secondary to the primary filter is to be effective (the case known as positive transfer (Torrey and Shavlik, 2010)), then we must assume that $y_s(t)$ is informative of $x(t)$. The core technical problems to be addressed are (i) that there is no complete model relating $x(t)$ and $x_{\mathrm{s}}(t)$, and, therefore, (ii) probability calculus (i.e. standard Bayesian conditioning) does not prescribe the required primary distribution conditioned on the transferred knowledge. Note that standard Bayesian conditioning yields the solution only in the case of complete modelling (Karbalayghareh et al., 2018). However, an insistence on specification of a complete model of inter-filter dependence may be highly restrictive, and incur model sensitivity in applications.

Instead, we acknowledge that the required conditional primary state predictor, $\breve{f}(x_t|d(t-1),f_{\mathrm{s}})$, after transfer of the secondary data predictor, $f_{\mathrm{s}}(y_{\mathrm{s},t}|d_{\mathrm{s}}(t-1))$ (10), is non-unique. Therefore, we solve the incurred decision problem optimally via the fully probabilistic design (FPD) principle (Quinn et al., 2016), choosing between all cases of $\breve{f}(x_t|d(t-1),f_{\mathrm{s}})$ consistent with the knowledge constraint introduced by the transfer of $f_{\mathrm{s}}(y_{\mathrm{s},t}|d_{\mathrm{s}}(t-1))$ (Quinn et al., 2017). FPD axiomatically prescribes an optimal choice, f^{o}, as a minimizer of the Kullback-Leibler divergence (KLD) from $\breve{f}$ to an ideal distribution, f^I, chosen by the designer (Kárný and Kroupa, 2012):

$$f^o(x_t|d(t-1),f_{\mathrm{s}}) \equiv \underset{\breve{f}\in\mathbf{F}}{\arg\min}\, D(\breve{f}||f_I). \tag{15}$$

Here, the KLD is defined as

$$D(\breve{f}||f_I) = E_{\breve{f}}\left[\ln\frac{\breve{f}}{f_I}\right], \tag{16}$$

and $\mathbf{F}$ denotes the set of $\breve{f}$ constrained by f_{s} (Quinn et al., 2016),(Quinn et al., 2017). The ideal distribution is consistently defined as

$$f_I(x_t|d(t-1)) \equiv f(x_t|d(t-1)),$$

being the state predictor of the *isolated* primary filter (5) (i.e. in the absence of any transfer from a secondary filter). In (Foley and Quinn, 2018), the following mean-field operator was shown to satisfy (15), and so to constitute the FPD-optimal primary state predictor after the (static) knowledge transfer specified above:

$$f^{\mathrm{o}}(x_t|d(t-1),f_{\mathrm{s}}) \propto f(x_t|d(t-1)) \times \\ \times \exp\left[\int_{y_t} f_{\mathrm{s}}(y_t|d_{\mathrm{s}}(t-1))\ln f(y_t|x_t)\,\mathrm{d}y_t\right]. \tag{17}$$

Note that the optional input u_t is *known* and constant. Here, it is *not* a part of FPD and plays only the role of a conditioning quantity in $d(t-1)$.

2.3 FPD-optimal uniform knowledge transfer

Here, we express the KLD minimiser (17) in the case of uniform pdfs. The functions appearing in (17) are

$$f(x_t|d(t-1)) \propto \chi\left(\underline{m}_t-\rho\le x_t\le\overline{m}_t+\rho\right) \tag{18}$$

$$f_{\mathrm{s}}(y_t|d_{\mathrm{s}}(t-1)) \propto \chi\left(\underline{y}_{\mathrm{s},t}\le y_t\le\overline{y}_{\mathrm{s},t}\right) \tag{19}$$

$$f(y_t|x_t) \propto \chi\left(Cx_t-r\le y_t\le Cx_t+r\right), \tag{20}$$

see (5), (10) and (4). Then, the form of (17) is

$$f^{\mathrm{o}}(x_t|d(t-1),f_{\mathrm{s}}) \propto \chi\left(\underline{m}_t-\rho\le x_t\le\overline{m}_t+\rho\right)\times \\ \times\exp\left[\int_{y_t}\chi\left(\underline{y}_{\mathrm{s},t}\le y_t\le\overline{y}_{\mathrm{s},t}\right)\times \right. \\ \left.\times\ln\chi\left(Cx_t-r\le y_t\le Cx_t+r\right)\,\mathrm{d}y_t\right] \tag{21}$$

The first term in the integral indicates the integration limits are $\underline{y}_{\mathrm{s},t}$ and $\overline{y}_{\mathrm{s},t}$. The characteristic function in the logarithm must equal one, which zeroes the exponent and generates an additional constraint for x_t.

Imposing $y_t\in(Cx_t-r,Cx_t+r)\cap(\underline{y}_{\mathrm{s},t},\overline{y}_{\mathrm{s},t})$ with nonempty intersection, we get the constraint for x_t:

$$\underline{y}_{\mathrm{s},t}-r\le Cx_t\le\overline{y}_{\mathrm{s},t}+r. \tag{22}$$

The constraining inequalities in (22) represent a *strip* in the state space. Time update of the target filter yields an orthotope domain in the state space (Figure 2). The strip, containing information from the source filter, and the orthotope then intersect (left side of Figure 2). This constrained set is the form of the knowledge transfer between the filters.

However, the data update step requires the prior pdf to be on the orthotope domain. Therefore, the constrained set must be circumscribed by another orthotope. The orthotope is an intersection of ℓ perpendicular strips. According to (Vicino and Zappa, 1996), all the $\ell+1$ strips are *tightened*, i.e. shifted and/or narrowed, so that distance of their planes is minimal while their intersection is unchanged (removed redundancy). Then, the tightened orthotope is the domain of the approximate expression of $f^{\mathrm{o}}(x_t|d(t-1),f_{\mathrm{s}})$ as the constrained prior entering the data update.

Note: The similarity of (22) with the observation model (4) and their similar processing indicates

similarity of the constraint step and the data update step. The difference is that the knowledge transfer (22) accepts the interval, i.e. the uniform distribution $\mathcal{U}_y(\underline{y}_{s,t},\overline{y}_{s,t})$, whereas the data update is supplied with the observed point, y_t. This point can be interpreted as the Dirac distribution $\delta_y(y-y_t)$. In this distributional sense, both the steps are equivalent.

2.4 Algorithmic summary

The source and target filters run their state estimation tasks in parallel. The source filter is run in isolation, and the data predictor (10) is computed between the time update and data update steps. Processing of the source predictor is placed between time update and data update step in the target filter.

Initialisation:

- Choose final time $\bar{t} > 0$, set initial time $t = 0$
- Set values $\underline{x}_0$, $\bar{x}_0$, u_0, noises ρ, r, r_s

On-line:

1. Set $t = t+1$
2. Compute $\underline{m}_t$, $\overline{m}_t$ according to (6) for each filter
3. Time update: compute (5) for each filter
4. Compute data predictor of the source filter (10)
5. Knowledge transfer: get the constraining strip (22) from the source data predictor and process it to time updated pdf of the target filter ((22) and below) to transfer knowledge between the filters (alternatively, use the informal transfer described later)
6. Get new data u_t, y_t, $y_{s,t}$
7. Data update: add the data strip (4) and approximate the obtained support by a parallelotope (for details see Appendix A.2 in (Pavelková and Jirsa, 2018)) to obtain the resulting form (7) (alternatively, skip the parallelotope stage and only tighten the strips, as described in 5.) for each filter
8. Compute $\underline{x}_t$, $\bar{x}_t$ (8) for each filter
9. Compute the point estimate $\hat{x}_t$ (9) for the target filter
10. If $t < \bar{t}$, go to 1.

3 EXPERIMENTS

In this section, the simulation experiments demonstrate the proposed algorithm properties.

3.1 Experiment setup

We simulate the system studied in (Foley and Quinn, 2018), $x_t \in \mathbb{R}^2$, $y_t \in \mathbb{R}$. The known model parameters are $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$, $C = [1\ \ 0]$, $\rho = \mathbf{1}_{(2)} \times 10^{-5}$, where $\mathbf{1}_{(2)}$ is the unit vector of length 2, $r = 10^{-3}$. $B = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $u_t = 0$, i.e. a system without input/control. The estimation was run for $\bar{t} = 50$ time steps. We investigate the influence of the observation noise r_s of the source filter (20) on estimate precision of the target filter, quanitified by TNSE (total norm squared-error) defined as $\text{TNSE} = \sum_{t=1}^{\bar{t}} \|\hat{x}_t - x_t\|^2$, where $\|\cdot\|$ is the Euclidean norm. For each combination of the parameters, the computation was run 2000-times and the TNSEs were averaged.

Simulated states x_t, together with the state noise ρ, are common for both the source and target filters, whereas the observation noises r and r_s, contributing to observed values, differ. This simple setup is sufficient for illustration of the described method.

3.2 Results

Figure 1 shows the influence of the source filter observation noise r_s on precision of the target filter through the FPD knowledge transfer. The method rejects negative knowledge transfer (unlike in the Gaussian transfer case (Foley and Quinn, 2018)). For higher values of r_s, estimate precision coincides with the isolated target filter (despite fluctuations) and does not increase. However, decrease of estimation error for smaller r_s is not very significant, compared to (Foley and Quinn, 2018).

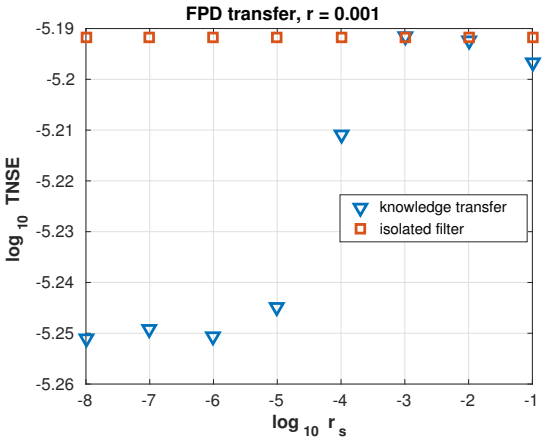

Figure 1: TNSE of the target LSU filter as a function of the observation noise variance r_s of the source filter. FPD transfer.

The probable reason for this last point can be illustrated by the left part of Figure 2. The grey area

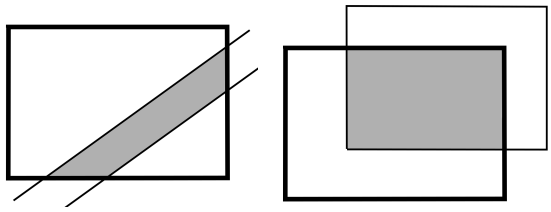

Figure 2: Domain of x_t. Thick line: target filter, thin line: transferred set. Left: FPD transfer. Right: informal transfer.

(the intersection) is to be circumscribed by an orthotope (5). If the strip is too close to the diagonal, the circumscribing orthotope is similar to the one given by the time update (if the diagonal is contained in the strip, the orthotopes are identical). This effect is probably caused by the fact that orthotopes are too conservative as approximators.

The knowledge transfer acts as a constraint of the target x_t set by the source filter. This inspires an informal transfer, illustrated in the right part of Figure 2. The idea is to use the intersection of the source and target time-updated state sets. Although this approach is not theoretically justified, it seems more suitable for orthotopic sets. The results of the prediction errors can be seen in Figure 3, in this informal case. The

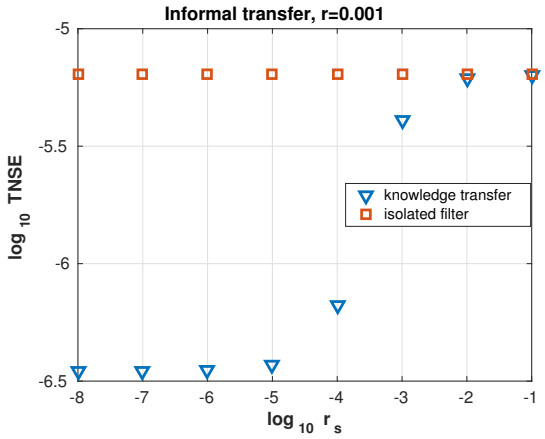

Figure 3: TNSE of the target LSU filter as a function of the observation noise variance r_s of the source filter. Informal transfer.

behaviour is qualitatively the same as in the previous case, but the influence on the target filter precision is now much more significant in the positive transfer regime.

It has also been observed that omitting the parallelotope stage in the data update, i.e. involving only strip tightening as in the FPD-based knowledge transfer, makes no difference in the target state precision.

3.3 Discussion

The isolated LSU filter performance serves as a reference against which to assess the effect of the knowledge transfer from a source LSU filter to a target one. A key advantage of the FPD-transfer—apart from its decision-theoretic optimality—is that the (stochastic) relationship between the filters does not have to be specified (in contrast to standard Bayesian transfer learning techniques (Karbalayghareh et al., 2018)). This confers robustness on the approach.

Both the FPD and informal transfers of Section 3 exhibit an important property: the ability to reject the source knowledge in the "below-threshold" case of inferior-quality data-predictive knowledge transfer (i.e. where $r_s \geq r$). This form of robustness is in contrast to FPD transfer between Kalman filters (Gaussian transfer) reported in (Foley and Quinn, 2018), where $r_s \geq r$ induces "negative transfer", decreasing the target precision in comparison with the isolated filter. The problem was overcome there by adjusting the transfer to take account of the external predictive variance. The avoidance of negative transfer in the LSU case is an important finding for the FPD knowledge transfer mechanism in general, since it supports a hypothesis—stated in (Foley and Quinn, 2018; Papež and Quinn, 2018)—that it is a moment-loss pathology of the Gaussian forms in Kalman filtering that incurred the negative FPD transfer in that case, and not an intrinsic limitation of the FPD transfer mechanism itself.

On the "above threshold" side (i.e. $r_s < r$), FPD knowledge transfer between LSU filters increases the target precision (Fig. 1)—effecting "positive transfer"—but this is insignificant compared to the improvement reported in (Foley and Quinn, 2018). In contrast, the informal transfer, based on the intersection of orthotopes, improves the state estimate precision strongly (Fig. 3). At threshold ($r_s = r$), note that positive transfer is still achieved in this informal case. This may be caused by a displacement between source and target orthotopic domains of the same size (see the right part of Figure 2).

The weakness of the FPD (i.e. formal) transfer in the LSU case is probably caused by the fact that orthotopes are too conservative as set approximators, i.e. they approximate the complex polytopic sets too coarsely (see the diagonal problem described in Figure 2 (left)). The informal transfer approach (Section 3.2) can be suitable for orthotopic approximations but is not optimal in a Bayesian (FPD) sense. It may be that this solution does decrease the Kullback-Leibler divergence (15), (16), but does not give the minimum (being the FPD solution (17)). However,

this requires theoretical validation.

In the orthotopically approximated data update step, the parallelotope stage (7) can be omitted. If so, the FPD knowledge transfer (Section 2.3) and the data update involve the same algorithmic flows, as was the case in (Foley and Quinn, 2018).

4 CONCLUDING REMARKS

The proposed FPD knowledge transfer between LSU filters has achieved excellent performance, rejecting the source data predictor below threshold, while achieving positive transfer above threshold (particularly in the case of the informal variant). We have noted in Section 3.3 that the technique is not conditional on an explicit model of interaction between the filters. It will be interesting in future work to validate further the robustness reported here, by simulating a richer set of contexts than the one reported in Section 3, particularly cases of distinct but correlated state processes.

Future research will also focus on a search for less conservative approximating sets for the geometric supports of the uniform approximations adopted in the time and data update steps of LSU filtering. The current variant starts with an orthotopic support within the time update step that is transformed into a parallelotopic one during the data update step. This is then circumscribed by an orthotope, to achieve functional closure—and, so, recursion—for the next time update. It would be less conservative to avoid this circumscribing approximation, and to recurse within a parallelotopic support only. However, the computational overhead of this more complicated setting needs to be assessed.

In the knowledge transfer step, again, it appears that it is the conservative nature of the orthotopic approximation which induces the weakly positive FPD transfer reported in this paper. Once again, more flexible and so tighter approximating sets—such as parallelotopes and zonotopes—will be explored in future work. Meanwhile, it will be interesting to find a formal (FPD) interpretation of the informal proposal which has achieved excellent positive transfer between the LSU filters (Section 3.2).

In both contexts above, the uniform approximations on adapted geometric supports are applied locally at each step of the algorithm, and so it is theoretically difficult to assess it convergence after many steps. In addition, the knowledge transfer is static in nature (Foley and Quinn, 2018), involving the transfer of the marginal source data predictor at each step. Progress in this area will require the transfer of dynamic (i.e. joint) source knowledge, and a search for a global approximation of the exact transfer-conditional target filtering distribution.

ACKNOWLEDGEMENTS

This research has been supported by GAČR grant 18-15970S.

REFERENCES

Abbas, A. E. (2009). A Kullback-Leibler view of linear and log-linear pools. *Decision Analysis*, 6(1):25–37.

Azizi, S. and Quinn, A. (2018). Hierarchical fully probabilistic design for deliberator-based merging in multiple participant systems. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 48(4):565–573.

Becis-Aubry, Y., Boutayeb, M., and Darouach, M. (2008). State estimation in the presence of bounded disturbances. *Automatica*, 44:1867–1873.

Chisci, L., Garulli, A., and Zappa, G. (1996). Recursive state bounding by parallelotopes. *Automatica*, 32(7):1049–1055.

Dardari, D., Closas, P., and Djuric, P. M. (2015). Indoor tracking: theory, methods, and technologies. *IEEE Transactions on Vehicular Technology*, 64(4):1263–1278.

Diez-Olivan, A., Ser, J. D., Galar, D., and Sierra, B. (2019). Data fusion and machine learning for industrial prognosis: trends and perspectives towards industry 4.0. *Information Fusion*, 50:92 – 111.

Dou, Y., Ran, C., and Gao, Y. (2016). Weighted measurement fusion Kalman estimator for multisensor descriptor system. *International Journal of Sytems Science*, 47(11):2722–2732.

Fletcher, R. (2000). *Practical Methods of Optimization*. John Wiley & Sons.

Foley, C. and Quinn, A. (2018). Fully probabilistic design for knowledge transfer in a pair of Kalman filters. *IEEE Signal Processing Letters*, 25:487–490.

Goudjil, A., Pouliquen, M., Pigeon, E., Gehan, O., and Targui, B. (2017). Recursive output error identification algorithm for switched linear systems with bounded noise. *IFAC-PapersOnLine*, 50(1):14112–14117.

Gravina, R., Alinia, P., Ghasemzadeh, H., and Fortino, G. (2017). Multi-sensor fusion in body sensor networks: state-of-the-art and research challenges. *Information Fusion*, 35:68–80.

He, J., Duan, X., Cheng, P., Shi, L., and Cai, L. (2017). Accurate clock synchronization in wireless sensor networks with bounded noise. *Automatica*, 81:350–358.

Jirsa, L., Pavelková, L., and Quinn, A. (2019). Approximate Bayesian prediction using state space model with uniform noise. In Gusikhin, O. and Madani, K., editors,

Informatics in Control Automation and Robotics, Lecture Notes in Electrical Engineering. Springer. Under review.

Karbalayghareh, A., Qian, X., and Dougherty, E. R. (2018). Optimal Bayesian transfer learning. *IEEE Transactions on Signal Processing*, 66(14):3724–3739.

Kárný, M., Böhm, J., Guy, T. V., Jirsa, L., Nagy, I., Nedoma, P., and Tesař, L. (2005). *Optimized Bayesian Dynamic Advising: Theory and Algorithms*. Springer, London.

Kárný, M. and Kroupa, T. (2012). Axiomatisation of fully probabilistic design. *Information Sciences*, 186(1):105–113.

Khaleghi, B., Khamis, A., Karray, F. O., and Razavi, S. N. (2013). Multisensor data fusion: a review of the state-of-the-art. *Information Fusion*, 14(1):28 – 44.

Lahat, D., Adali, T., and Jutten, C. (2015). Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477.

Lang, L., Chen, W., Bakshi, B. R., Goel, P. K., and Ungarala, S. (2007). Bayesian estimation via sequential Monte Carlo sampling — constrained dynamic systems. *Automatica*, 43(9):1615–1622.

Majumder, S. and Pratihar, D. K. (2018). Multi-sensors data fusion through fuzzy clustering and predictive tools. *Expert Systems with Applications*, 107:165 – 172.

Nassreddine, G., Abdallah, F., and Denoux, T. (2010). State estimation using interval analysis and belief-function theory: application to dynamic vehicle localization. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 40(5):1205–1218.

Papež, M. and Quinn, A. (2018). Dynamic Bayesian knowledge transfer between a pair of Kalman filters. In *2018 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, Aalborg, Denmark. IEEE.

Pavelková, L. and Jirsa, L. (2018). Approximate recursive Bayesian estimation of state space model with uniform noise. In *Proceedings of the 15th International Conference on Informatics in Control, Automation and Robotics (ICINCO)*, pages 388–394, Porto, Portugal.

Quinn, A., Kárný, M., and Guy, T. (2016). Fully probabilistic design of hierarchical Bayesian models. *Information Sciences*, 369(1):532–547.

Quinn, A., Kárný, M., and Guy, T. V. (2017). Optimal design of priors constrained by external predictors. *International Journal of Approximate Reasoning*, 84:150–158.

Shamshirband, S., Petkovic, D., Javidnia, H., and Gani, A. (2015). Sensor data fusion by support vector regression methodology-a comparative study. *IEEE Sensors Journal*, 15(2):850–854.

Simon, D. and Simon, D. L. (2010). Constrained Kalman filtering via density function truncation for turbofan engine health estimation. *International Journal of Systems Science*, 41:159–171.

Torrey, L. and Shavlik, J. (2010). Transfer learning. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, pages 242–264. IGI Global.

Vapnik, V. and Izmailov, R. (2017). Knowledge transfer in SVM and neural networks. *Annals of Mathematics and Artificial Intelligence*, 81(1-2):3–19.

Vargas-Melendez, L., Boada, B. L., Boada, M. J. L., Gauchia, A., and Diaz, V. (2017). Sensor fusion based on an integrated neural network and probability density function (PDF) dual Kalman filter for on-line estimation of vehicle parameters and states. *Sensors*, 17(5).

Vicino, A. and Zappa, G. (1996). Sequential approximation of feasible parameter sets for identification with set membership uncertainty. *IEEE Transactions on Automatic Control*, 41(6):774–785.

Vitola, J., Pozo, F., Tibaduiza, D. A., and Anaya, M. (2017). A sensor data fusion system based on *k*-nearest neighbor pattern classification for structural health monitoring applications. *Sensors*, 17(2).

Willner, D., Chang, C., and Dunn, K. (1976). Kalman filter algorithms for a multi-sensor system. In *1976 IEEE Conference on Decision and Control including the 15th Symposium on Adaptive Processes*, pages 570–574.

Xiao, F. (2019). Multi-sensor data fusion based on the belief divergence measure of evidences and the belief entropy. *Information Fusion*, 46:23 – 32.

Yang, C., Yang, Z., and Deng, Z. (2019). Robust weighted state fusion Kalman estimators for networked systems with mixed uncertainties. *Information Fusion*, 45:246 – 265.

Zou, T., Wang, Y., Wang, M., and Lin, S. (2017). A real-time smooth weighted data fusion algorithm for greenhouse sensing based on wireless sensor networks. *Sensors*, 17(11).