Hierarchical Bayesian Transfer Learning Between a Pair of Kalman Filters

Milan Papež [a]

papez@utia.cas.cz
[a] *Institute of Information Theory and Automation*
Czech Academy of Sciences
Prague, Czech Republic

Anthony Quinn [a,b]

aquinn@tcd.ie
[b] *Department of Electronic and Electrical Engineering*
Trinity College Dublin, the University of Dublin
Dublin, Ireland

2021 32nd Irish Signals and Systems Conference (ISSC) | 978-1-6654-3429-4/20/$31.00 ©2021 IEEE | DOI: 10.1109/ISSC52156.2021.9467863

***Abstract*—Transfer learning strategies are typically designed in a deterministic manner, without processing uncertainty in the knowledge transfer mechanism. They also require the dependence between the participating learning procedures—Bayesian filters in this work—to be explicitly modelled. This letter develops an approach which relaxes both of these restrictive assumptions. We frame the proposed Bayesian transfer learning technique as fully probabilistic design of an unknown hierarchical probability distribution conditioned on knowledge in the form of an external probability distribution. This yields a randomized design around a base density for transfer learning which has been reported in previous work by the authors. In the Kalman filtering context, this hierarchical relaxation—which induces a knowledge-driven mixture state predictor—significantly improves tracking performance when compared to conventional transfer learning methods.**

***Index Terms*—Fully probabilistic design, hierarchical models, Bayesian transfer learning, randomized design, Kalman filters.**

I. Introduction

Transfer learning is a principled framework for exploiting knowledge of an external agent (source task) to improve learning of a primary agent (target task) [1]. Various signal processing applications rely on transfer learning to process language [2], brain activity [3], protein records [4], satellite images [5], etc. We are particularly motivated by developing a transfer learning strategy for a network of signal processing nodes which implement Bayesian filtering. These include Gaussian filters [6]—whose basic representative is the Kalman filter—and particle filters [7].

The standard way for Bayesian transfer learning strategies to incorporate external knowledge is through elicitation of a prior distribution [8]. To facilitate application of Bayes' rule, such methods require specification of a probabilistic model for conditioning on transferred knowledge, typically in the form of external data. We refer to this conventional setting as *complete modeling*. In contrast, this letter considers transfer of external knowledge in the form of a probability distribution, broadening the range of admissible external knowledge representations. However, a probabilistic model for conditioning on such a distribution is rarely available, obviating the use of Bayes' rule. In this *incomplete modeling* setting, fully probabilistic design (FPD) [9]—an axiomatically justified [10] generalization of the maximum entropy principle [11]—provides the optimal tool for design of the conditional distribution. FPD has recently been utilized to develop *static* [12] and *dynamic* [13] transfer learning strategies between a pair of Kalman filters. A recent extension of the FPD principle addresses incomplete modeling scenarios via hierarchical Bayesian model design [14], thereby quantifying the uncertainty in the unknown conditional distribution. This randomized design has been used to formulate a transfer learning framework which accounts for uncertainty in the knowledge transfer mechanism [15]. In this letter, we apply this hierarchical FPD approach to knowledge transfer between Bayesian filters. By accounting for the randomized nature of the transfer learning mechanism in this way, an infinite mixture of state predictors is induced, yielding an improved filtering performance in comparison with deterministic approaches, such as [12].

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

II. Hierarchical FPD Transfer Learning between a pair of Bayesian Filters

Let us consider a state-space model given by

$$x_i \sim \mathsf{F}(x_i|x_{i-1}), \tag{1a}$$

$$z_i \sim \mathsf{F}(z_i|x_i), \tag{1b}$$

where $i = 1, \dots, n$ denotes the discrete-time index. Here, we assume that the state variable $x_i \in \boldsymbol{x} \subseteq \mathbb{R}^{m_x}$ is measured only indirectly through the observation variable $z_i \in \boldsymbol{z} \subseteq \mathbb{R}^{m_z}$. The model (1) is fully characterized by the state-transition (1a) and observation (1b) probability densities, with the convention $x_0 \equiv \varnothing$. The essential object in formulating the basic inference tasks related to (1) is the joint predictive model

$$\mathsf{F}(z_i, x_i|\mathbf{z}_{i-1}), \tag{2}$$

where $\mathbf{z}_{i-1} \equiv (z_1, \dots, z_{i-1})$. Indeed, the conditional and marginal densities of (2) facilitate computation of the Bayesian filtering and predictive recursions [6].

We address the task of knowledge transfer from an external to a primary Bayesian filter [12]. The objective is to design the prior of the primary Bayesian filter $\mathsf{A}(x_i|\mathsf{F}_\mathsf{E}, \mathbf{z}_{i-1})$—c.f., the marginal density of (2)—which accommodates additional

knowledge in the form of an observation predictor, F_E, provided by the external Bayesian filter. However, we assume that this prior is not only unknown but also *uncertain*, and we thus model it hierarchically by the hyper-prior $\mathsf{S}(\mathsf{A}|\mathsf{F}_\mathsf{E}, \mathbf{z}_{i-1})$. Therefore, we extend the basic setting (2) to form the hierarchical joint unknown model

$$\mathsf{M}(z_i, x_i, \mathsf{A}|\mathsf{S}, \mathsf{F}_\mathsf{E}, \mathbf{z}_{i-1}) \equiv \mathsf{F}_\mathsf{E}(z_i|\mathbf{z}_{\mathsf{E},i-1})\mathsf{A}(x_i|\mathsf{F}_\mathsf{E}, \mathbf{z}_{i-1})\mathsf{S}(\mathsf{A}|\mathsf{F}_\mathsf{E}, \mathbf{z}_{i-1}), \quad (3)$$

where we apply assumptions defined as

$$\mathsf{M}(z_i|x_i, \mathsf{A}, \mathsf{S}, \mathsf{F}_\mathsf{E}, \mathbf{z}_{i-1}) \equiv \mathsf{F}_\mathsf{E}(z_{\mathsf{E},i}|\mathbf{z}_{\mathsf{E},i-1})\big|_{z_{\mathsf{E},i}=z_i}, \quad (4a)$$

$$\mathsf{M}(x_i|\mathsf{A}, \mathsf{S}, \mathsf{F}_\mathsf{E}, \mathbf{z}_{i-1}) \equiv \mathsf{A}(x_i|\mathsf{F}_\mathsf{E}, \mathbf{z}_{i-1}), \quad (4b)$$

$$\mathsf{M}(\mathsf{A}|\mathsf{S}, \mathsf{F}_\mathsf{E}, \mathbf{z}_{i-1}) \equiv \mathsf{S}(\mathsf{A}|\mathsf{F}_\mathsf{E}, \mathbf{z}_{i-1}). \quad (4c)$$

Here, (4a) implements the knowledge transfer by constraining the F_E-conditioned model of $z_i \in \boldsymbol{z}$ to be the external observation predictor, F_E. We model z_i based on only the information accumulated in F_E and conditionally independently of $x_i \in \boldsymbol{x}$. (4b) considers that F_E and $\mathbf{z}_{i-1}$ are sufficient for influencing x_i, and that S provides no additional information about x_i. In (4c), we simply assign the hyper-prior to A. In this letter, M and A denote (unknown) variational-from densities and F denotes a (known) fixed-form density.

Assumptions (4) constrain the functional form of M and thus delineate the knowledge-constrained set of admissible models $\mathbf{M}$. With F_E being fixed and provided by the external filter, and A generated randomly by S, the only variational quantity to be optimized in (3) is S. In summary, we seek

$$\mathsf{M} \in \mathbf{M} \equiv \{\text{models (3) with } \mathsf{F}_\mathsf{E} \text{ fixed and } \mathsf{A} \text{ generated by variational } \mathsf{S}\}. \quad (5)$$

FPD is an approach for finding an optimal design M^o of an unknown model M while respecting the set-based knowledge constraint $\mathsf{M} \in \mathbf{M}$ (resulting from empirical facts, assumed density forms etc.) and taking into account preferences about M expressed by an ideal model M_I. The FPD-optimal design $\mathsf{M}^o \in \mathbf{M}$ is the unique density that is closest to M_I in the minimum Kullback-Leibler divergence (KLD, [16]) sense,

$$\mathsf{M}^o \equiv \underset{\mathsf{M}\in\mathbf{M}}{\operatorname{argmin}}\, \mathcal{D}(\mathsf{M}||\mathsf{M}_\mathsf{I}), \quad (6)$$

with the KLD from M to M_I being given by

$$\mathcal{D}(\mathsf{M}||\mathsf{M}_\mathsf{I}) \equiv \mathsf{E}_\mathsf{M}\left[\log\left(\frac{\mathsf{M}}{\mathsf{M}_\mathsf{I}}\right)\right],$$

where E_M denotes the expected value with respect to M.

We choose the hierarchical joint ideal model as

$$\mathsf{M}_\mathsf{I}(z_i, x_i, \mathsf{A}|\mathsf{S}_\mathsf{I}, \mathsf{F}_\mathsf{E}, \mathbf{z}_{i-1}) \equiv \mathsf{F}(z_i, x_i|\mathbf{z}_{i-1})\mathsf{S}_\mathsf{I}(\mathsf{A}|\mathsf{F}_\mathsf{E}, \mathbf{z}_{i-1}), \quad (7)$$

where we apply assumptions given by

$$\mathsf{M}_\mathsf{I}(z_i, x_i|\mathsf{A}, \mathsf{S}_\mathsf{I}, \mathsf{F}_\mathsf{E}, \mathbf{z}_{i-1}) \equiv \mathsf{F}(z_i, x_i|\mathbf{z}_{i-1}), \quad (8a)$$

$$\mathsf{M}_\mathsf{I}(\mathsf{A}|\mathsf{S}_\mathsf{I}, \mathsf{F}_\mathsf{E}, \mathbf{z}_{i-1}) \equiv \mathsf{S}_\mathsf{I}(\mathsf{A}|\mathsf{F}_\mathsf{E}, \mathbf{z}_{i-1}). \quad (8b)$$

In (8a), we define the ideal for $(x_i, z_i) \in \boldsymbol{x} \times \boldsymbol{z}$ to be the joint model (2). This density is the key object in devising the primary filter and thus a reasonable reference for our design. (8b) simply defines a user-defined ideal hyper-prior for A.

Proposition 1. *The unknown joint augmented model belongs to the knowledge-constrained set,* $\mathsf{M} \in \mathbf{M}$ *(5), and the ideal model* M_I *is (7), then the FPD-optimal hierarchical model—i.e., the solution of (6)—is*

$$\mathsf{M}^o(z_i, x_i, \mathsf{A}|\mathsf{S}^o, \mathsf{F}_\mathsf{E}, \mathbf{z}_{i-1}) = \mathsf{F}_\mathsf{E}(z_i|\mathbf{z}_{\mathsf{E},i-1})\mathsf{A}(x_i|\mathsf{F}_\mathsf{E}, \mathbf{z}_{i-1})\mathsf{S}^o(\mathsf{A}|\mathsf{F}_\mathsf{E}, \mathbf{z}_{i-1}),$$

where

$$\mathsf{S}^o(\mathsf{A}|\mathsf{F}_\mathsf{E}, \mathbf{z}_{i-1}) \propto \mathsf{S}_\mathsf{I}(\mathsf{A}|\mathsf{F}_\mathsf{E}) \exp\left\{-\mathcal{D}(\mathsf{A}||\hat{\mathsf{A}})\right\}, \quad (9)$$

$$\hat{\mathsf{A}}(x_i|\mathsf{F}_\mathsf{E}, \mathbf{z}_{i-1}) \propto \mathsf{F}(x_i|\mathbf{z}_{i-1}) \times \exp\left\{-\int \ln\left(\mathsf{F}(z_i|x_i)\right)\mathsf{F}_\mathsf{E}(z_i|\mathbf{z}_{\mathsf{E},i-1})dz_i\right\}, \quad (10)$$

and the FPD-optimal design of A *becomes*

$$\mathsf{A}^o(x_i|\mathsf{F}_\mathsf{E}, \mathbf{z}_{i-1}) = \mathsf{E}_{\mathsf{S}^o}[\mathsf{A}] = \int \mathsf{A}(x_i|\mathsf{F}_\mathsf{E}, \mathbf{z}_{i-1})\mathsf{S}^o(\mathsf{A}|\mathsf{F}_\mathsf{E}, \mathbf{z}_{i-1})d\mathsf{A}. \quad (11)$$

Proof. The proof follows from Theorem 1 of [15]. □

$\hat{\mathsf{A}}$ in (10) fulfills the role of a base density [17] around which randomized choices of A are distributed via the FPD-optimal hyper-prior (9). It is a deterministic transformation of fixed-form distributions in the non-hierarchical setting [12], [13]. The expected prior (11) under (9) replaces the pre-prior, $\mathsf{F}(x_i|\mathbf{z}_{i-1})$, of standard Bayesian filtering. This ensures optimal transfer of F_E, while respecting uncertainty in the knowledge transfer mechanism.

III. Hierarchical FPD Transfer Learning between a pair of Kalman Filters

This section specifies Proposition 1 to the case of (1) being the normal linear state-space model,

$$\mathsf{F}(x_i|x_{i-1}) \equiv \mathcal{N}_{x_i}(Ax_{i-1}, Q), \quad (12a)$$

$$\mathsf{F}(z_i|x_i) \equiv \mathcal{N}_{z_i}(Cx_i, R), \quad (12b)$$

where $\mathcal{N}_v(\mu, \Sigma)$ is the normal density of argument v, with the mean vector, μ, and the covariance matrix, Σ; and A and C are matrices of appropriate dimensions. If we adopt (12) and $\mathsf{F}(x_1) \equiv \mathcal{N}_{x_1}(\mu_{1|0}, \Sigma_{1|0})$, then the conditional and marginal densities of (2) become

$$\mathsf{F}(x_i|\mathbf{z}_i) = \mathcal{N}_{x_i}(\mu_{i|i}, \Sigma_{i|i}), \quad (13a)$$

$$\mathsf{F}(x_i|\mathbf{z}_{i-1}) = \mathcal{N}_{x_i}(\mu_{i|i-1}, \Sigma_{i|i-1}), \quad (13b)$$

$$\mathsf{F}(z_i|\mathbf{z}_{i-1}) = \mathcal{N}_{z_i}(z_{i|i-1}, R_{i|i-1}), \quad (13c)$$

with the shaping parameters being computed recursively as

$$\mu_{i|i} = \mu_{i|i-1} + K(z_i - z_{i|i-1}), \quad (14a)$$

$$\Sigma_{i|i} = \Sigma_{i|i-1} - KR_{i|i-1}K^\top, \quad (14b)$$

$$\mu_{i|i-1} = A\mu_{i-1|i-1}, \quad (14c)$$

$$\Sigma_{i|i-1} = A\Sigma_{i-1|i-1}A^\top + Q, \quad (14d)$$

$$z_{i|i-1} = C\mu_{i|i-1}, \quad (14e)$$

$$R_{i|i-1} = C\Sigma_{i|i-1}C^\top + R, \quad (14f)$$

where $K = \Sigma_{i-1|i} C^\top R^{-1}_{i|i-1}$, and $^\top$ denotes matrix transposition. Specifically, (13a) and (13b) form the data and time steps of the conventional Kalman filter.

To appropriately choose the form of the unknown randomized design, A, and the ideal hyper-prior, $\mathsf{S_I}$, in (9), we first need to investigate the form of the base density (10) in the context of (12). We do so in the next lemma.

Lemma 1. *The observation model and the state pre-prior are (12b) and (13b), respectively, and the external observation predictor is* $\mathsf{F_E}(z_i|\mathbf{z}_{\mathsf{E},i-1}) \equiv \mathcal{N}_{z_i}(z_{\mathsf{E},i|i-1}, R_{\mathsf{E},i|i-1})$. *Then, the base density (10) becomes*

$$\hat{\mathsf{A}}(x_i|\mathsf{F_E}, \mathbf{z}_{i-1}) = \mathcal{N}_{x_i}(\hat{\mu}_{i|i-1}, \hat{\Sigma}_{i|i-1}), \tag{15}$$

where the shaping parameters are computed according to

$$\hat{\mu}_{i|i-1} = \mu_{i|i-1} + L(z_{\mathsf{E},i|i-1} - z_{i|i-1}), \tag{16a}$$
$$\hat{\Sigma}_{i|i-1} = \Sigma_{i|i-1} - L R_{i|i-1} L^\top, \tag{16b}$$

with $L = \Sigma_{i|i-1} C^\top R^{-1}_{i|i-1}$.

Proof. The proof follows from the standard calculus. □

Recall that Proposition 1 furnishes an FPD-optimal relaxation around the normal base density design (15). Therefore, it is reasonable to specify the functional form of the randomized density A as normal,

$$\mathsf{A}(x_i|\mathsf{F_E}, \mathbf{z}_{i-1}) \equiv \mathcal{N}_{x_i}(\mu, \Sigma), \tag{17}$$

with unknown parameters defined as $\Theta \equiv (\mu, \Sigma)$. Only the normal-inverse-Wishart ideal hyper-prior $\mathsf{S_I}$ is invariant under the FPD-optimal hierarchical knowledge transfer mechanism in (9):

$$\mathsf{S_I}(\mu, \Sigma|\mathsf{F_E}, \mathbf{z}_{i-1}) = \mathcal{N}_\mu(\mu_\mathsf{I}, \beta_\mathsf{I}\Sigma) i\mathcal{W}_\Sigma(\nu_\mathsf{I}, \Lambda_\mathsf{I}). \tag{18}$$

The parametric constraint on A allows us to write $\mathsf{S}(\mathsf{A}|\mathsf{F_E}, \mathbf{z}_{i-1}) = \mathsf{S}(\Theta|\mathsf{F_E}, \mathbf{z}_{i-1})$.

Proposition 2. *The base density and the unknown prior are given by (15) and (17), respectively, and the ideal hyper-prior is (18). Then, the FPD-optimal hyper-prior (9) becomes*

$$\mathsf{S}^o(\mu, \Sigma|\mathsf{F_E}, \mathbf{z}_{i-1}) = \mathsf{S}^o(\mu|\Sigma, \mathsf{F_E}, \mathbf{z}_{i-1}) \mathsf{S}^o(\Sigma|\mathsf{F_E}, \mathbf{z}_{i-1}), \tag{19}$$

where the conditional factor is

$$\mathsf{S}^o(\mu|\Sigma, \mathsf{F_E}, \mathbf{z}_{i-1}) = \mathcal{N}_\mu(\bar{\mu}_{i|i-1}, \bar{\Sigma}_{i|i-1}), \tag{20}$$

with the shaping parameters

$$\bar{\mu}_{i|i-1} = \bar{\Sigma}_{i|i-1}\big(\tfrac{1}{\beta_\mathsf{I}}\Sigma^{-1}\mu_\mathsf{I} + \hat{\Sigma}^{-1}_{i|i-1}\hat{\mu}_{i|i-1}\big), \tag{21a}$$
$$\bar{\Sigma}_{i|i-1} = \big(\tfrac{1}{\beta_\mathsf{I}}\Sigma^{-1} + \hat{\Sigma}^{-1}_{i|i-1}\big)^{-1}, \tag{21b}$$

and the marginal factor is

$$\mathsf{S}^o(\Sigma|\mathsf{F_E}, \mathbf{z}_{i-1}) \propto i\mathcal{W}_\Sigma(\nu_\mathsf{I}, \Lambda_\mathsf{I})|\Sigma|^{\frac{1}{2}} \times \mathcal{N}_{\mu_\mathsf{I}}(\hat{\mu}_{i|i-1}, \beta_\mathsf{I}\Sigma + \hat{\Sigma}_{i|i-1}) \exp\{-\tfrac{1}{2}\operatorname{tr}(\Sigma\hat{\Sigma}^{-1}_{i|i-1}). \tag{22}$$

Proof. The proof follows from the standard calculus. □

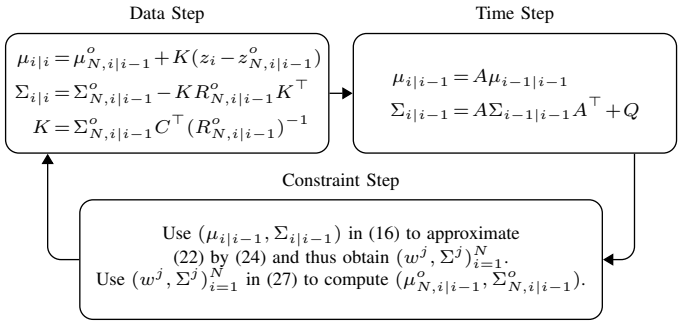

Fig. 1. FPD-optimal processing for hierarchical static knowledge transfer between Kalman filters

The exponential term in (9) prevents us from applying the conjugate analysis between the normal and normal-inverse-Wishart densities. Therefore, the marginal factor (22) is intractable, and we need to resort to approximate techniques. However, the conditional factor (20) is tractable, allowing us to integrate μ out analytically in the derivation of (11), as shown in the following proposition.

Proposition 3. *The unknown prior and the optimal hyper-prior are (17) and (19), respectively. Then, the FPD-optimal prior (11) becomes*

$$\mathsf{A}^o(x_i|\mathsf{F_E}, \mathbf{z}_{i-1}) = \int \mathcal{N}_{x_i}(\bar{\mu}_{i|i-1}, \Sigma + \bar{\Sigma}_{i|i-1}) \mathsf{S}^o(\Sigma|\mathsf{F_E}, \mathbf{z}_{i-1}) d\Sigma, \tag{23}$$

where $(\bar{\mu}_{i|i-1}, \bar{\Sigma}_{i|i-1})$ *are given by (21).*

Proof. The proof follows from the standard calculus. □

We approximate the intractable marginal density (22) by an appropriately chosen Monte Carlo method [18], allowing us to formulate the empirical approximation

$$\mathsf{S}^o_N(d\Sigma|\mathsf{F_E}, \mathbf{z}_{i-1}) = \sum_{j=1}^N w^j \delta_{\Sigma^j}(d\Sigma), \tag{24}$$

where w^j is the weight which assesses the contribution of the particle Σ^j to the approximation (24). After plugging (24) into (23), and integrating with respect to Σ, we obtain

$$\mathsf{A}^o_N(x_i|\mathsf{F_E}, \mathbf{z}_{i-1}) = \sum_{j=1}^N w^j \mathcal{N}_{x_i}(\bar{\mu}^j_{i|i-1}, \Sigma^j + \bar{\Sigma}^j_{i|i-1}), \tag{25}$$

being a finite mixture approximation of the infinite mixture induced in (11). If such a density were applied to replace (13b) in the Kalman filter recursive flow, the complexity of (13a) would grow exponentially with i. Therefore, we apply the moment matching [19] at each i to approximate (25) by

$$\mathsf{A}^o_N(x_i|\mathsf{F_E}, \mathbf{z}_{i-1}) \approx \mathcal{N}_{x_i}(\mu^o_{N,i|i-1}, \Sigma^o_{N,i|i-1}), \tag{26}$$

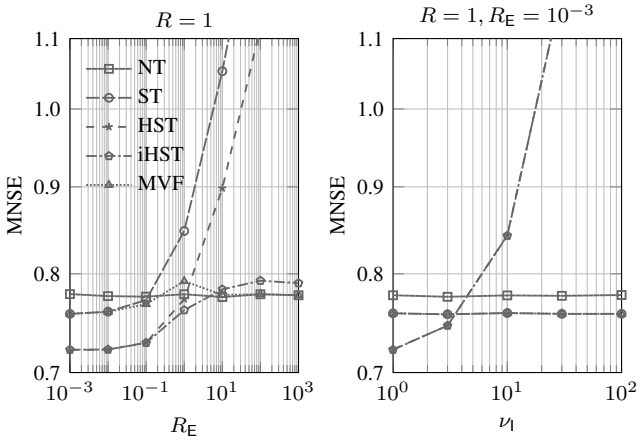

Fig. 2. Left: the mean norm squared-error (MNSE) of the primary filter versus the observation variance R_{E} of the external Kalman filter. Right: the MNSE of the primary filter versus the number of degrees of freedom ν_{I} of the ideal hyper-prior density (18). The results are averaged over 1000 independent simulation runs. We compare (i) the Kalman filter with *N*o *T*ransfer (*NT*), (ii) *S*tatic Bayesian knowledge *T*ransfer (*ST*) [12], (iii) *H*ierarchical *S*tatic Bayesian knowledge *T*ransfer (*HST*) given by Algorithm 1 (this paper), (iv) *i*nformally adapted version of *HST* (*iHST*) discussed in Section V (this paper); and (v) *M*easurement *V*ector *F*usion (*MVF*) [21].

with the shaping parameters computed as

$$\mu^{o}_{N,i|i-1} = \sum_{j=1}^{N} w^j \bar{\mu}^{j}_{i|i-1}, \tag{27a}$$

$$\Sigma^{o}_{N,i|i-1} = \sum_{j=1}^{N} w^j [\Sigma^j + \bar{\Sigma}^{j}_{i|i-1} + (\mu^{o}_{N,i|i-1} - \bar{\mu}^{j}_{i|i-1})(\mu^{o}_{N,i|i-1} - \bar{\mu}^{j}_{i|i-1})^{\top}]. \tag{27b}$$

The tractable substructure used to compute (23) allows us to decrease the variance of estimators associated with (25) compared to those we would eventually obtain when sampling μ, Σ jointly, the idea known as Rao-Blackwellization (RB) [20].

The proposed algorithm can now be summarized in Fig. 1, where we use $z^{o}_{N,i|i-1}$ and $R^{o}_{N,i|i-1}$ to denote the shaping parameters of the observation predictor computed from $\mu^{o}_{N,i|i-1}$ and $\Sigma^{o}_{N,i|i-1}$, respectively.

IV. Numerical Illustration

We consider the scalar state-space model (12) with $A = 0.95$, $C = 1$, $Q = 1$, and $R = 1$, which illustrates the key features of our algorithm. The initial statistics are $\mu_{1|0} = 0$ and $\Sigma_{1|0} = 1$; the statistics of the ideal hyper-prior are $\nu_{\mathsf{I}} = 1$, $\beta_{\mathsf{I}} = 100$, $\mu_{\mathsf{I}} = 0$, and $\Lambda_{\mathsf{I}} = 0.3$; and the number of time steps is $n = 200$. We use importance sampling to compute (24), with the bootstrap proposal density [18] given by $i\mathcal{W}_{\Sigma}(\nu_{\mathsf{I}}, \Lambda_{\mathsf{I}})$ and $N = 25$. We calculate the mean norm squared-error, MNSE $= \frac{1}{n}\sum_{i=1}^{n} ||x_i - \mu_{i|i}||^2$, with $||\cdot||$ being the Euclidean norm, for various methods in Fig. 2.

In the left part of Fig. 2, we assess the primary filter with fixed observation variance, R, while varying the external observation variance, R_{E}, which quantifies the impact of the confidence of the external knowledge. The NT filter does not use any external knowledge and thus defines a reference MNSE level for evaluating whether the compared methods deliver positive or negative knowledge transfer. The MNSEs of the remaining filters change with the ratio of R to R_{E}. We see that the proposed HST filter offers improved performance over all the other algorithms for $R_{\mathsf{E}} < 1$ (positive transfer). This filter also surpasses the ST filter for all R_{E}. However, the HST filter is not robust above the threshold where $R_{\mathsf{E}} = R$ because it is not able to reject the external knowledge and recover the performance of the NT and MVF filters. We observe that the HST filter suffers from negative transfer when the external observations are more imprecise than the primary ones. The iHST filter is discussed in Section V.

In the right part of Fig. 2, we further investigate the positive transfer case where we hold $R = 1$ and $R_{\mathsf{E}} = 10^{-3}$ fixed. We vary the number of degrees of freedom ν_{I} in (18), which acts as the concentration parameter for (18) around its prior expected values of μ and Σ. The remaining hyper-parameters of (18) are the same as above. The MNSEs of the NT, ST, and MVF filters are obviously constant as these methods do not use the hyper-prior. As ν_{I} increases, the external knowledge is increasingly rejected, and the FPD-optimal predictor (23) becomes dominated by the hyper-parameters in (18). This prior domination greatly increases the MNSE, and points to the need for low values of ν_{I} in hierarchical FPD-optimal knowledge transfer.

V. Discussion

The ST filter developed in [12] considers unknown A as a non-random density, without introducing the top hierarchical level. The relation between the ST filter and its HST extension proposed in this letter is that the FPD-optimal design of A given in [12] coincides with the base density (10). Therefore, the ST filter can be seen as a certainty-equivalent type of Bayesian filter, thus unequipped with A-measure of uncertainty. The fact that this structure is involved in the HST filter carries over the disadvantage of the ST filter [12], [13], which lies in the insensitivity of transferring the covariance of the external observation predictor, $R_{\mathsf{E},i|i-1}$. In [12], this issue was informally resolved by replacing R by $R_{\mathsf{E},i|i-1}$ in the covariance of the primary observation predictor (14f), making the performance of the ST filter exactly equivalent to the MVF filter. We now apply the same substitution in the HST filter. This replacement forms the iHST filter introduced in Section IV. The left part of Fig. 2 shows that the iHST filter provides better performance than the HST filter for $R_{\mathsf{E}} > 10^{-1}$ and thus offers an increased robustness against the negative transfer. Nevertheless, the iHST filter still suffers from the negative transfer to a small degree. This issue maybe caused by accumulation of the approximation error brought by the importance sampling and moment matching.

We also implemented the HST and iHST methods without the RB from Proposition 3. For fixed computational resources, we can achieve far greater precision with the proposed algorithms compared to the non-RB ones.

VI. Conclusion

This letter provides a framework for knowledge transfer between Bayesian filters. The methodology relies on the FPD-optimal, knowledge-constrained, design of a hierarchical Bayesian model, which is conditioned on an external probability density, and equips the transfer learning mechanism with measures of uncertainty. The specific instance of this generic approach, where the interacting signal processing nodes are represented by the Kalman filters, demonstrates that the top hierarchical level proposed in this letter utilizes the external knowledge in a more efficient way compared to the preceding approach presented in [12]. However, the adverse feature of the insensitivity to transferring the covariance of the external observation predictor—as originally reported in [12]—prevails in the algorithm developed here as well. The hyper-parameters of S_{l} (18) are chosen as fixed values for all time-steps of the current algorithm. Future work will focus on optimal, data-driven, adaptation of these hyper-parameters at each time step.

References

[1] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[2] D. Wang and T. F. Zheng, "Transfer learning for speech and language processing," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2015 Asia-Pacific*. IEEE, 2015, pp. 1225–1237.

[3] S. Makeig, C. Kothe, T. Mullen, N. Bigdely-Shamlo, Z. Zhang, and K. Kreutz-Delgado, "Evolving signal processing for brain–computer interfaces," *Proceedings of the IEEE*, vol. 100, pp. 1567–1584, 2012.

[4] S. Mei, W. Fei, and S. Zhou, "Gene ontology based transfer learning for protein subcellular localization," *BMC Bioinformatics*, vol. 12, no. 44, 2011.

[5] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon, "Combining satellite imagery and machine learning to predict poverty," *Science*, vol. 353, no. 6301, pp. 790–794, 2016.

[6] S. Särkkä, *Bayesian filtering and smoothing*. Cambridge University Press, 2013, vol. 3.

[7] A. Doucet and A. M. Johansen, "A tutorial on particle filtering and smoothing: Fifteen years later," in *The Oxford Handbook of Nonlinear Filtering*, D. Crisan and B. Rozovsky, Eds. Oxford University Press, 2009.

[8] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*. IGI Global, 2010, pp. 242–264.

[9] M. Kárný, "Towards fully probabilistic control design," *Automatica*, vol. 32, no. 12, pp. 1719–1722, 1996.

[10] M. Kárný and T. Kroupa, "Axiomatisation of fully probabilistic design," *Information Sciences*, vol. 186, no. 1, pp. 105–113, 2012.

[11] J. Shore and R. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Transactions on information theory*, vol. 26, no. 1, pp. 26–37, 1980.

[12] C. Foley and A. Quinn, "Fully probabilistic design for knowledge transfer in a pair of Kalman filters," *IEEE Signal Processing Letters*, vol. 25, no. 4, pp. 487–490, 2018.

[13] M. Papež and A. Quinn, "Dynamic Bayesian knowledge transfer between a pair of Kalman filters," in *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2018, pp. 1–6.

[14] A. Quinn, M. Kárnỳ, and T. V. Guy, "Fully probabilistic design of hierarchical Bayesian models," *Information Sciences*, vol. 369, pp. 532–547, 2016.

[15] ——, "Optimal design of priors constrained by external predictors," *International Journal of Approximate Reasoning*, vol. 84, pp. 150–158, 2017.

[16] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[17] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," *The annals of statistics*, pp. 209–230, 1973.

[18] J. S. Liu, *Monte Carlo strategies in scientific computing*. Springer Science & Business Media, 2008.

[19] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[20] G. Casella and C. P. Robert, "Rao-Blackwellisation of sampling schemes," *Biometrika*, vol. 83, no. 1, pp. 81–94, 1996.

[21] D. Willner, C. Chang, and K. Dunn, "Kalman filter algorithms for a multi-sensor system," in *Decision and Control including the 15th Symposium on Adaptive Processes, 1976 IEEE Conference on*, vol. 15. IEEE, 1976, pp. 570–574.