2018 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 17–20, 2018, AALBORG, DENMARK

DYNAMIC BAYESIAN KNOWLEDGE TRANSFER BETWEEN A PAIR OF KALMAN FILTERS

Milan Papež [a] *and Anthony Quinn* [a,b]

[a] Institute of Information Theory and Automation, Czech Academy of Sciences, Czech Republic

[b] Department of Electronic and Electrical Engineering, Trinity College Dublin, the University of Dublin, Ireland

ABSTRACT

Transfer learning is a framework that includes—among other topics—the design of knowledge transfer mechanisms between Bayesian filters. Transfer learning strategies in this context typically rely on a complete stochastic dependence structure being specified between the participating learning procedures (filters). This paper proposes a method that does not require such a restrictive assumption. The solution in this *incomplete modelling* case is based on the fully probabilistic design of an unknown probability distribution which conditions on knowledge in the form of an externally supplied distribution. We are specifically interested in the situation where the external distribution accumulates knowledge dynamically via Kalman filtering. Simulations illustrate that the proposed algorithm outperforms alternative methods for transferring this dynamic knowledge from the external Kalman filter.

Index Terms— Bayesian transfer learning, fully probabilistic design, incomplete modelling, Kalman filtering

1. INTRODUCTION

Transfer learning [1] has become a key research direction in statistical machine learning [2]. The basic principle of transfer learning is to utilize the experience of an external learning agent (source task) to improve the learning of a primary agent (target task). Transfer learning has recently witnessed substantial attention in a multitude of theoretically and practically oriented scientific fields, such as reinforcement learning [3], deep learning [4], autonomous driving [5], computer vision [6], sensor networks [7], etc. This paper focuses on a specific transfer learning context referred to as Bayesian transfer learning and its deployment in statistical signal processing. We are specifically interested in developing a procedure for probabilistic knowledge transfer in sensor networks where each knowledge-bearing node constitutes a Bayesian filter acting on its associated state-space model.

The research has been supported by GAČR grant 18-15970S. Supplementary material for this paper can be downloaded from www.researchgate.net/profile/milan_papez

The conventional approach to Bayesian transfer learning involves replacing the prior distribution of standard Bayesian learning with a distribution conditioned on the transferred knowledge [8]. The methods based on this principle differ in the way the knowledge-conditioned prior is elicited [9]. An alternative principle is to define the joint posterior distribution of both source and target quantities of interest given source and target data, and then to compute the posterior distribution of the target quantity by marginalization [10]. Hierarchical Bayesian learning provides another formalization of Bayesian transfer learning [11], where the knowledge is transferred by means of a hyper-prior. However, it seems that a widely accepted consensus on Bayesian transfer learning is missing. This paper seeks to fill this gap.

The common aspect of the above approaches is that they assume existence of an explicit model of all unknown quantities of interest, enabling Bayes' rule to accommodate transfer learning, which we call here the *complete modelling* case. In the present paper, we are concerned with a scenario where there is not enough knowledge to construct such a model explicitly. We refer to this particular situation as the *incomplete modelling* case. The previous work in this respect [12] involved a static Bayesian knowledge transfer for a pair of Kalman filters, where the external knowledge is transferred in the form of a marginal distribution defined at a single time-step. The present paper extends this work by designing a mechanism for transferring distributions defined over multiple time-steps, thus achieving dynamic and on-line Bayesian knowledge transfer.

2. KNOWLEDGE TRANSFER BETWEEN A PAIR OF BAYESIAN FILTERS

2.1. Problem formulation

Let us consider a state-space model given by

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

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

where $x_i \in \mathsf{X} \subseteq \mathbb{R}^{n_x}$ and $z_i \in \mathsf{Z} \subseteq \mathbb{R}^{n_z}$ are respectively the state and observation variables defined at the discrete-time in-

978-1-5386-5477-4/18/$31.00

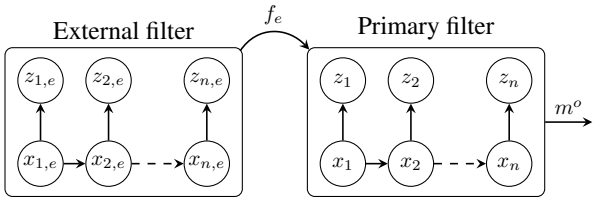

Fig. 1. A pair of Bayesian filters acting on their state-space models. The external filter provides the density f_e summarizing knowledge of the quantities (states or observations) gathered over the whole run of the filter. The primary filter makes use of this external knowledge to improve state inference over the corresponding time interval.

stants $i = 1, \ldots, n$. The state-space model (1) is fully determined by the state transition (1a) and observation (1b) probability densities, with all their parameters being known. At the initial time step ($i = 1$), the state variable is distributed according to $x_1 \sim f(x_1)$. The time-evolution of the state-space model (1) is characterized by the joint augmented model

$$f(\mathbf{x}_n, \mathbf{z}_n) = f(\mathbf{z}_n|\mathbf{x}_n) f(\mathbf{x}_n) \equiv \prod_{i=1}^{n} f(z_i|x_i) f(x_i|x_{i-1}), \quad (2)$$

where $f(\mathbf{z}_n|\mathbf{x}_n)$ and $f(\mathbf{x}_n)$ define the joint observation model and joint state pre-prior model, respectively. In (2), we respect the convention $x_0 \equiv \varnothing$ and use the boldface notation $\mathbf{v}_n \equiv (v_1, \ldots, v_n)$ to denote a sequence of variables $v_i \in \mathsf{V}$, for $i = 1, \ldots, n$. Moreover, we use the symbols m and f to denote unspecified (variational form) and specified (fixed form) densities, respectively.

We are concerned with the problem of optimally transferring knowledge from an external Bayesian filter (source task) to a primary one (target task). The filters operate on their respective models, processing their local observations, and estimating their local states (Fig. 1). The conditional independence structure between the variables in each model is as specified in (2). However, an explicit conditioning mechanism describing dependence between $(\mathbf{x}_n, \mathbf{z}_n)$ and $(\mathbf{x}_{n,e}, \mathbf{z}_{n,e})$ is assumed missing. Note that there is no edge between these node sets in the graphical model in Fig. 1. The common modelling approach based on a joint density of the external and primary variables is therefore unavailable. This incomplete modelling scenario is addressed here as a problem of optimal design of an unknown probability density, processing the external (distributional) knowledge as a constraint. Specifically, we design a *dynamic* Bayesian knowledge transfer method, where knowledge is transferred in the form of a joint probability density, f_e, describing a sequence of external quantities, either $\mathbf{x}_{n,e}$ or $\mathbf{z}_{n,e}$.

2.2. Fully probabilistic design

A central concern of probabilistic inference is to design (i.e. infer) a stochastic model representing our beliefs about an unknown quantity of interest, $v \in \mathsf{V}$. The construction of such a model is naturally performed by processing our knowledge, k (from physical laws, empirical facts etc.), about the modelled quantity in some way. However, such knowledge is usually insufficient to determine the model completely. Thus, an explicit density, $m(k|v)$, quantifying our beliefs about k given v is unavailable, and we therefore cannot compute $m(v|k)$ directly by application of Bayes' rule. The model is then sought within a user-specified set of possible models, $m(v|k) \in \mathbf{M}$, that are compatible with k. To complete the decision-making set-up, we specify our preferences about the unknown model, $m(v|k)$, by defining its ideal prescription, $m_{\mathsf{I}}(v)$. Fully probabilistic design (FPD, [13]) is a principled and axiomatically justified [14] approach for optimally choosing $m \in \mathbf{M}$ while taking into account our knowledge and preferences. The optimal model (i.e. design) provides a compromise between the knowledge, k, and the ideal prescription, m_{I}. It is found as the density that is closest to $m_{\mathsf{I}}(v)$ in the minimum Kullback-Leibler divergence (KLD, [15]) sense, while respecting the set-based knowledge constraint, $m \in \mathbf{M}$:

$$m^o(v|k) \equiv \operatorname*{argmin}_{m \in \mathbf{M}} \mathcal{D}(m||m_{\mathsf{I}}),$$

where $\mathcal{D}(m||m_{\mathsf{I}})$ is the KLD from m to m_{I}, given as

$$\mathcal{D}(m||m_{\mathsf{I}}) = \mathsf{E}_m \left[\ln \left(\frac{m}{m_{\mathsf{I}}} \right) \right],$$

with E_m denoting the expected value with respect to m. The density $m^o(v|k) \in \mathbf{M}$ is also consistent with k and is referred to as the FPD-optimal design. Typically, $m_{\mathsf{I}} \notin \mathbf{M}$. The case where $m_{\mathsf{I}} \in \mathbf{M}$ implies that the knowledge constraint is inactive, leading to $m^o = m_{\mathsf{I}}$.

In common with the minimum cross-entropy (MXE) principle [16], the FPD framework is a *deterministic* approach for designing an unknown density. A recent extension of FPD leading to a stochastic design of the unknown density has been provided in [17], conferring measures of uncertainty on the designed density. The key feature that distinguishes FPD from the MXE principle is that FPD allows preferences about the unknown model to be processed. The MXE principle follows the same setting as presented above, but the ideal model, m_{I}, is replaced by a prior model, m_{P}.

3. DYNAMIC BAYESIAN KNOWLEDGE TRANSFER

This section formalizes dynamic Bayesian knowledge transfer as an FPD-based optimal design of an unknown density and shows its application in Bayesian filtering. A principal purpose of Bayesian filtering is to compute the marginal (filtering) density, $f(x_i|\mathbf{z}_i)$, of the joint state posterior density, $f(\mathbf{x}_i|\mathbf{z}_i)$. Under the conditional independence assumptions adopted in (2), this density becomes

$$f(x_i|\mathbf{z}_i) = \frac{f(z_i|x_i) f(x_i|\mathbf{z}_{i-1})}{f(z_i|\mathbf{z}_{i-1})}, \quad (3a)$$

with

$$f(x_i|\mathbf{z}_{i-1}) = \int f(x_i|x_{i-1}) f(x_{i-1}|\mathbf{z}_{i-1}) dx_{i-1}, \quad (3b)$$

$$f(z_i|\mathbf{z}_{i-1}) = \int f(z_i|x_i) f(x_i|\mathbf{z}_{i-1}) dx_i. \quad (3c)$$

(3b) and (3c) are the one-step-ahead state and observation predictors, respectively.

To solve the transfer learning problem (Fig. 1), we use FPD to choose optimally the unknown joint augmented model of the states and observations, $(\mathbf{x}_n, \mathbf{z}_n)$, conditioned on the external density, f_e. This factorizes as follows:

$$m(\mathbf{x}_n, \mathbf{z}_n|f_e) = m(\mathbf{z}_n|\mathbf{x}_n, f_e) m(\mathbf{x}_n|f_e). \quad (4)$$

We express our joint preferences about the quantities $(\mathbf{x}_n, \mathbf{z}_n)$ by defining the ideal joint augmented model as (2), that is,

$$m_{\mathsf{I}}(\mathbf{x}_n, \mathbf{z}_n) \equiv f(\mathbf{x}_n, \mathbf{z}_n). \quad (5)$$

The FPD-optimal choice, m^o, conditioned on the external knowledge, f_e, is found as the unique minimizer of the KLD from the unknown model (4) to the ideal model (5):

$$m^o(\mathbf{x}_n, \mathbf{z}_n|f_e) \in \underset{m \in \mathbf{M}}{\operatorname{argmin}} \mathcal{D}(m||m_{\mathsf{I}}). \quad (6)$$

The external knowledge—encoded as f_e—is transferred by constraining the set $\mathbf{M}$ in a specific way, as we now show.

3.1. Transferring an external joint observation predictor

We choose to transfer the external joint observation predictor, $f_e(\mathbf{z}_{n,e})$. To do so, we must specify exactly how the f_e condition constrains the functional form of m in (4). First, we consider the f_e-conditioned joint observation model, which factorizes as

$$m(\mathbf{z}_n|\mathbf{x}_n, f_e) = \prod_{i=1}^{n} m(z_i|\mathbf{x}_i, \mathbf{z}_{i-1}, f_e),$$

and we impose the following conditional independence assumption:

$$m(z_i|\mathbf{x}_i, \mathbf{z}_{i-1}, f_e) \equiv f_e(z_{i,e}|\mathbf{z}_{i-1,e}) \; |_{z_{i,e}=z_i}. \quad (7)$$

Here, we have constrained the f_e-conditioned model for the *primary* observations to be the externally supplied one-step-ahead observation predictor. Next, we consider the f_e-conditioned joint state prior model in (4), which factorizes as

$$m(\mathbf{x}_n|f_e) = \prod_{i=1}^{n} m(x_i|\mathbf{x}_{i-1}, f_e),$$

and we impose the conventional Markov property:

$$m(x_i|\mathbf{x}_{i-1}, f_e) \equiv m(x_i|x_{i-1}, f_e).$$

Under these specified knowledge constraints, the unknown f_e-conditioned joint augmented model (4) becomes

$$m(\mathbf{x}_n, \mathbf{z}_n|f_e) \equiv f_e(\mathbf{z}_n) m(\mathbf{x}_n|f_e). \quad (8)$$

With $f_e(\mathbf{z}_n)$ fixed via the external filter, the f_e-conditioned joint state prior factor, $m(\mathbf{x}_n|f_e)$, is the only variational quantity which we can now choose via FPD for the purpose of optimal knowledge transfer. In summary, the f_e-constrained set of candidate models is

$$\mathbf{M} \equiv \big\{\text{models (8) with } f_e(\mathbf{z}_n) \text{ fixed and } m(\mathbf{x}_n|f_e) \text{ variational}\big\}. \quad (9)$$

The following proposition establishes the fact that $f_e(\mathbf{z}_{n,e})$ is *sequentially* processed into the FPD-optimal joint state prior of the primary filter. This will be key in securing a *recursive*, causal, dynamic Bayesian transfer learning algorithm between a pair of Kalman filters, as we will see in Section 3.2.

Proposition 1. *The unknown joint augmented model satisfies the knowledge constraint, $m \in \mathbf{M}$(9), imposed by transfer of the external joint observation predictor, $f_e(\mathbf{z}_{n,e})$. The ideal model is defined in (5), and $D(m||m_{\mathsf{I}}) < \infty, \forall m \in \mathbf{M}$. Then, an FPD-optimal design of m—i.e. a solution of (6)—is*

$$m^o(\mathbf{x}_n, \mathbf{z}_n|f_e) = f_e(\mathbf{z}_n) m^o(\mathbf{x}_n|f_e), \quad (10)$$

with

$$m^o(\mathbf{x}_n|f_e) = \prod_{i=1}^{n} m^o(x_i|x_{i-1}, f_e) \quad (11a)$$

$$\propto f(\mathbf{x}_n) \prod_{i=1}^{n} \exp\{-\mathcal{D}(f_e||f)\} \gamma(x_i). \quad (11b)$$

Here,

$$m^o(x_i|x_{i-1}, f_e) \equiv \frac{f(x_i|x_{i-1}) \exp\{-\mathcal{D}(f_e||f)\} \gamma(x_i)}{\gamma(x_{i-1})}, \quad (12)$$

$$\mathcal{D}(f_e||f) \equiv \int f_e(z_i|\mathbf{z}_{i-1,e}) \ln \frac{f_e(z_i|\mathbf{z}_{i-1,e})}{f(z_i|x_i)} dz_i, \quad (13)$$

$$\gamma(x_{i-1}) \equiv \int f(x_i|x_{i-1}) \times \exp\{-\mathcal{D}(f_e||f)\} \gamma(x_i) dx_i. \quad (14)$$

The normalization functions, $\gamma(x_i)$, need to be computed via a backward sweep through the recursions (14), for $i = n, \ldots, 1$, initialized with $\gamma(x_n) \equiv 1$.

Proof. See the supplementary material. □

Proposition 1 shows that FPD-optimal Bayesian transfer learning is achieved by updating the pre-prior, $f(\mathbf{x}_n)$, to the prior, $m^o(\mathbf{x}_n|f_e)$. This is achieved via modulation with a product term (11b) containing the external knowledge over

the full time horizon. Correspondingly, at each time instant, i, the update of the state transition model to the FPD-optimal state transition model is achieved via the modulation (12). This optimal joint prior, $m^o(\mathbf{x}_n|f_e)$, can therefore be sequentially processed by the primary filter, via (3), since it enjoys the recursive factorization form in (11b,12). In particular, (12) replaces (1a) in the standard Bayesian filtering setting (3), optimally transferring the external joint observation predictor, $f_e(\mathbf{z}_{n,e})$, in a sequential manner.

3.2. Transfer of an external Kalman filter observation predictor

Here, we describe a specific application of Proposition 1 to the case of transferring the external Kalman filter joint observation predictor. The Kalman filter is one of the very restricted instances which ensure that the Bayesian filtering equations (3) are tractable. Specifically, (1) is specialized to the linear-Gaussian case:

$$f(x_i|x_{i-1}) \equiv \mathcal{N}_{x_i}(Ax_{i-1}, Q), \tag{15a}$$
$$f(z_i|x_i) \equiv \mathcal{N}_{z_i}(Cx_i, R), \tag{15b}$$

and the marginal state pre-prior density has to be chosen as the Gaussian density $f(x_1) \equiv \mathcal{N}_{x_1}(\mu_{1|0}, \Sigma_{1|0})$. Here, $\mathcal{N}_v(\mu, \Sigma)$ denotes the Gaussian density of a (vector) random variable, v, with the mean, μ, and covariance matrix, Σ; and A and C are matrices of appropriate dimensions. Under these assumptions, the densities (3) preserve the Gaussian form across all iterations, $i = 1, \ldots, n$,

$$f(x_i|\mathbf{z}_i) = \mathcal{N}_{x_i}(\mu_{i|i}, \Sigma_{i|i}), \tag{16a}$$
$$f(x_i|\mathbf{z}_{i-1}) = \mathcal{N}_{x_i}(\mu_{i|i-1}, \Sigma_{i|i-1}), \tag{16b}$$
$$f(z_i|\mathbf{z}_{i-1}) = \mathcal{N}_{z_i}(z_{i|i-1}, R_{i|i-1}), \tag{16c}$$

with the shaping parameters being computed explicitly and recursively as follows:

$$\mu_{i|i} = \mu_{i|i-1} + K(z_i - z_{i|i-1}), \tag{17a}$$
$$\Sigma_{i|i} = \Sigma_{i|i-1} - KR_{i|i-1}K^\top, \tag{17b}$$
$$\mu_{i|i-1} = A\mu_{i-1|i-1}, \tag{18a}$$
$$\Sigma_{i|i-1} = A\Sigma_{i-1|i-1}A^\top + Q, \tag{18b}$$
$$z_{i|i-1} = C\mu_{i|i-1}, \tag{19a}$$
$$R_{i|i-1} = C\Sigma_{i|i-1}C^\top + R. \tag{19b}$$

Here, $K \equiv \Sigma_{i|i-1}C^\top R_{i|i-1}^{-1}$ and $^\top$ denotes matrix transposition. These formulae follow directly from application of the conditioning and marginalization rules for Gaussian densities containing affine transformations [18].

To support our next proposition, we present the following lemma, which specifies the computation of the normalization function (14) in this Kalman context.

Lemma 1. *Let the state-space model be defined by (15), and the external one-step-ahead observation predictor by (16c), i.e. $f_e(z_{i,e}|\mathbf{z}_{i-1,e}) \equiv \mathcal{N}_{z_{i,e}}(z_{i|i-1,e}, R_{i|i-1,e})$, $i = n, \ldots, 2$. Then, (14) preserves the form*

$$\gamma(x_{i-1}) \propto \exp\left[-\tfrac{1}{2}(x_{i-1}^\top S_{i-1|i}x_{i-1} - 2x_{i-1}^\top r_{i-1|i})\right], \tag{20}$$

and its explicit computation reduces to the recursion

$$r_{i-1|i} = A^\top(I_{n_x} - L)r_{i|i}, \tag{21a}$$
$$S_{i-1|i} = A^\top(I_{n_x} - L)S_{i|i}A, \tag{21b}$$

where, for $i = n-1, \ldots, 2$,

$$r_{i|i} = r_{i|i+1} + C^\top R^{-1} z_{i|i-1,e}, \tag{22a}$$
$$S_{i|i} = S_{i|i+1} + C^\top R^{-1} C, \tag{22b}$$

and, for $i = n$,

$$r_{n|n} = C^\top R^{-1} z_{n|n-1,e}, \tag{23a}$$
$$S_{n|n} = C^\top R^{-1} C. \tag{23b}$$

Here, $L \equiv S_{i|i}Q^{\frac{1}{2}}(Q^{\frac{\top}{2}}S_{i|i}Q^{\frac{1}{2}} + I_{n_x})^{-1}Q^{\frac{\top}{2}}$, I_{n_x} is the $n_x \times n_x$ identity matrix, and $Q^{\frac{1}{2}}$ is the Cholesky factor of Q.

Proof. See the supplementary material. □

Lemma 1 demonstrates the connection between the computation of (14) and the backward information filter [19] which takes the mean value of the external predictor $z_{i|i-1,e}$ as the observation input. Based on this result, the next proposition furnishes the explicit recursive computation of the FPD-optimal state transition model (12).

Proposition 2. *Under the conditions of Lemma 1, the FPD-optimal state transition model (12) is given by*

$$m^o(x_i|x_{i-1}, f_e) = \mathcal{N}_{x_i}(\mu_i^o, \Sigma_i^o), \tag{24}$$

with the shaping parameters calculated according to

$$\mu_i^o = (I_{n_x} - \Sigma_i^o S_{i|i})Ax_{i-1} + \Sigma_i^o r_{i|i}, \tag{25}$$
$$\Sigma_i^o = Q^{\frac{1}{2}}(Q^{\frac{\top}{2}}S_{i|i}Q^{\frac{1}{2}} + I_{n_x})^{-1}Q^{\frac{\top}{2}}. \tag{26}$$

Here, $r_{i|i}$ and $S_{i|i}$ are given by (22a) and (22b), respectively.

Proof. See the supplementary material. □

Proposition 2 specifies the optimal adaptation of the primary (i.e. target) Kalman filter flow, in order to process transferred knowledge in the form of the external joint observation predictor. If we apply (24) in (3b), then the one-step-ahead state predictor preserves the Gaussian form of (16b). However, the difference is that, now, the shaping parameters (18a,18b) are replaced with

$$\mu_{i|i-1} = (I_{n_x} - \Sigma_i^o S_{i|i})A\mu_{i-1|i-1} + \Sigma_i^o r_{i|i}, \tag{27a}$$
$$\Sigma_{i|i-1} = (I_{n_x} - \Sigma_i^o S_{i|i})A\Sigma_{i-1|i-1}A^\top(I_{n_x} - \Sigma_i^o S_{i|i})^\top + \Sigma_i^o, \tag{27b}$$

respectively. The resulting filter with FPD-optimal dynamic transfer is summarized in Algorithm 1.

Algorithm 1 FPD-optimal processing for dynamic transfer between Kalman filters

A. **Backward sweep:**
1. For $i = n$,
   * use $z_{n|n-1,e}$ in (23) to compute $(r_{n|n}, S_{n|n})$.
   * use $(r_{n|n}, S_{n|n})$ in (21) to compute $(r_{n-1|n}, S_{n-1|n})$.
2. For $i = n-1, \ldots, 2$;
   * use $z_{i|i-1,e}$ and $(r_{i|i+1}, S_{i|i+1})$ in (22) to compute $(r_{i|i}, S_{i|i})$.
   * use $(r_{i|i}, S_{i|i})$ in (21) to compute $(r_{i-1|i}, S_{i-1|i})$.

B. **Forward sweep:**
1. For $i = 1$, set $\mu_{1|0}, \Sigma_{1|0}$ and use it in (17) to compute $(\mu_{1|1}, \Sigma_{1|1})$.
2. For $i = 2, \ldots, n$;
   * use $(\mu_{i-1|i-1}, \Sigma_{i-1|i-1})$ in (27) to compute $(\mu_{i|i-1}, \Sigma_{i|i-1})$.
   * use $(\mu_{i|i-1}, \Sigma_{i|i-1})$ in (17) to compute $(\mu_{i|i}, \Sigma_{i|i})$.

4. EXPERIMENTS

The purpose of this section is to compare the proposed method against alternative approaches. We evaluate the performance of the primary filter when keeping its observation variance R fixed but changing the observation variance of the external filter R_e, which quantifies the confidence of the external knowledge. To assess the resulting state estimates, we use the mean norm squared-error, MNSE $= \frac{1}{n}\sum_{i=1}^{n} ||x_i - \mu_{i|i}||^2$, with $||\cdot||$ denoting the Euclidean norm. We are concerned with a simple position-velocity state-space model [20] specified by

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 \end{bmatrix}, \quad Q = 10^{-5} I_2, \quad R = 10^{-3}.$$

The number of time steps is $n = 50$. The results of the compared algorithms are illustrated in Fig. 2.

The MNSE of the NT filter defines a reference level against which the remaining filters are compared. This level is obviously constant as the external observation variance does not enter the standard Kalman filter via (19b). The error in the remaining filters varies according to the ratio of the primary and external observation variances. We can observe that the proposed DT filter achieves positive knowledge transfer for $R_e < 3 \times 10^{-3}$, which is evidenced by the fact that the error of the DT filter is lower than that of the NT filter in this range. Moreover, the DT filter outperforms the MVF filter in the same interval, and it also outperforms the ST filter for $R_e < 2 \times 10^{-2}$. The important observation is that the ST and MVF filters meet the performance of the NT filter close to the intersection where $R_e = R$, but the proposed DT filter passes this point with a markedly lower error and meets the NT filter later (i.e. for higher external observation variance). This increased robustness of the DT filter, which now benefits even from external observations that are of a lower quality than the primary ones, is achieved because of its ability to accumulate the external knowledge over multiple time steps via the *dynamic* transfer which is the focus of this paper. The ST and MVF filters do not have this property, as is evidenced by the fact that their error is, respectively, worse and very similar to the NT filter, above $R_e = R$. However, accumulating external knowledge of increasingly poor quality does lead to a more quickly decreasing performance of the DT filter for $R_e > 2 \times 10^{-2}$.

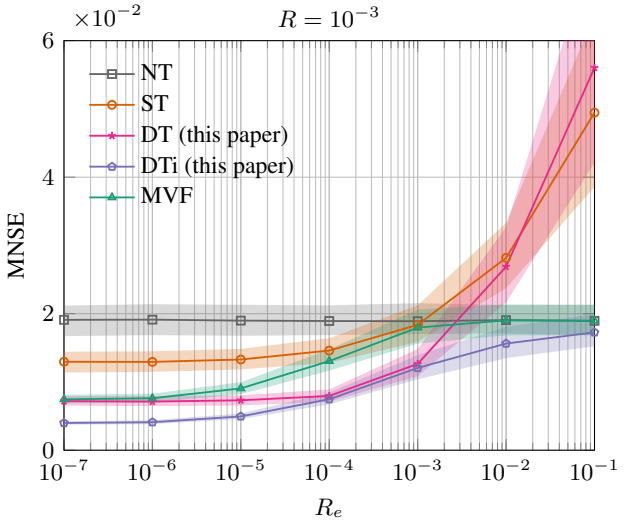

Fig. 2. The mean norm squared-error (MNSE) of the primary filter versus the observation variance R_e of the external Kalman filter. The results are averaged over 1000 independent simulation runs, with the solid line being the median and the shaded area delineating the interquartile range. The procedures that are compared are (i) the Kalman filter with *N*o *T*ransfer (*NT*), (ii) *S*tatic Bayesian knowledge *T*ransfer (*ST*) [12], (iii) *D*ynamic Bayesian knowledge *T*ransfer (*DT*) given by Algorithm 1 (this paper), (iv) an *i*nformally adapted version of DT (*DTi*) which we mention in Section 5 (this paper); and (v) *M*easurement *V*ector *F*usion (*MVF*) [21].

5. DISCUSSION

In common with the ST filter of [12], the DT filter is also insensitive to the transfer of the covariance of the external observation predictor $R_{i|i-1,e}$. The loss of this moment information occurs when evaluating the KLD (13) and can informally be resolved by replacing R with $R_{i|i-1,e}$ in (22) and (23). This simple substitution defines the DTi filter introduced in Section 4. The experiments demonstrate that the DTi filter surpasses all the other filters across the full range of values of R_e. This outcome is remarkable as it proves that improved estimation accuracy is achieved by implementing this FPD-optimal Bayesian transfer learning, obviating the need—usually prohibitive—to specify an explicit stochastic dependence structure between the external and primary quantities. It is also important to note that the DT filter offers the same advantage, albeit over a slightly shorter range of values of R_e. However, it seems that the fragile dependence assumptions inherited by the MVF filter undermine its performance. The fact that we do not require these dependence assumptions is a markedly simplifying feature of this FPD-based transfer learning framework, and should ensure its consistency in a wide range of applications. In the supplementary material, we provide evidence that the proposed method also offers more robustness against higher values of the state covariance Q.

6. CONCLUSION

This paper has proposed an FPD-based optimal dynamic Bayesian transfer learning approach and showed its application to probabilistic knowledge transfer between a pair of Kalman filters. The resulting experiments demonstrate that FPD offers a potential for building a versatile framework for Bayesian transfer learning. However, there is still the question of dealing with the aforementioned insensitivity to the second moment transfer, as discussed in Section 5. A possible answer to this problem may lie in the recently proposed hierarchical FPD-based Bayesian transfer learning [22], which will be the primary aim of future work. We have focused thusfar on the basic scenario of one-directional knowledge transfer between two nodes. The natural extension of the proposed approach therefore consists of (i) facilitating the knowledge transfer among a greater number of nodes and (ii) making the transfer bi-directional. Specifically, the former point will require us to introduce an optimal weighting mechanism to assess knowledge in a network of nodes. Another possible extension is to replace the Kalman filters with different forms of Gaussian filters [18], leaving the derivations presented in Section 3.2 mostly intact. Although the application of sequential Monte Carlo methods [23] may be feasible, the recursive computation of (14) may present problems. Finally, one can change the transferred knowledge and conditional independence assumptions specified in (8) in order to propose other FPD-based transfer learning options, such transfer of the external joint state predictor.

A universal Bayesian transfer learning framework has been elusive so far. However, the practical evidence of this paper—along with the axiomatically driven optimality it provides—supports the assertion that FPD-optimal Bayesian transfer learning can become such a universal framework.

7. REFERENCES

[1] S. J. Pan, "Transfer learning," in *Data Classification: Algorithms and Applications*, pp. 537–558. Chapman and Hall/CRC, 2015.

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

[3] M. E. Taylor and P. Stone, "Transfer learning for reinforcement learning domains: A survey," *Journal of Machine Learning Research*, vol. 10, pp. 1633–1685, 2009.

[4] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," in *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, 2012, pp. 17–36.

[5] D. Isele and A. Cosgun, "Transferring autonomous driving knowledge on simulated and real intersections," *arXiv preprint arXiv:1712.01106*, 2017.

[6] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, "Visual domain adaptation: A survey of recent advances," *IEEE signal processing magazine*, vol. 32, no. 3, pp. 53–69, 2015.

[7] T. L. M. Van Kasteren, G. Englebienne, and B. J. A. Kröse, "Transferring knowledge of activity recognition across sensor networks," in *International Conference on Pervasive Computing*. Springer, 2010, pp. 283–300.

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

[9] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*, Wiley, 1994.

[10] A. Karbalayghareh, X. Qian, and E. R. Dougherty, "Optimal Bayesian transfer learning," *arXiv preprint arXiv:1801.00857*, 2018.

[11] A. Wilson, A. Fern, and P. Tadepalli, "Transfer learning in sequential decision problems: A hierarchical Bayesian approach," in *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, 2012, pp. 217–227.

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

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

[15] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[16] 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.

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

[18] S. Särkkä, *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013.

[19] D. Q. Mayne, "A solution of the smoothing problem for linear dynamic systems," *Automatica*, vol. 4, no. 2, pp. 73–92, 1966.

[20] R. Faragher, "Understanding the basis of the Kalman filter via a simple and intuitive derivation," *IEEE Signal processing magazine*, vol. 29, no. 5, pp. 128–132, 2012.

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

[22] A. Quinn, M. Kárný, and T. V. Guy, "Optimal design of priors constrained by external predictors," *International Journal of Approximate Reasoning*, vol. 84, pp. 150–158, 2017.

[23] 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.