International Journal of Approximate Reasoning 84 (2017) 150–158

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Optimal design of priors constrained by external predictors

Anthony Quinn [a,*], Miroslav Kárný [b], Tatiana V. Guy [b]

[a] Trinity College Dublin, the University of Dublin, Dublin 2, Ireland
[b] The Institute of Information Theory and Automation, The Czech Academy of Sciences, Czechia

ARTICLE INFO

Article history:
Received 12 August 2016
Received in revised form 7 February 2017
Accepted 8 February 2017
Available online 16 February 2017

Keywords:
Fully probabilistic design
Parameter prior
External predictive distribution
Bayesian transfer learning
Kullback–Leibler divergence

ABSTRACT

This paper exploits knowledge made available by an external source in the form of a predictive distribution in order to elicit a parameter prior. It uses the terminology of Bayesian transfer learning, one of many domains dealing with reasoning as coherent knowledge processing. An empirical solution of the addressed problem was provided in [19], based on an interpretation of the external predictor as an empirical distribution constructed from fictitious data. In this paper, two main contributions are provided. First, the problem is solved using formal hierarchical Bayesian modeling [25], and the knowledge transfer is achieved optimally, i.e. in the minimum-KLD sense. Second, this hierarchical setting yields a distribution on the set of possible priors, with the choice [19] acting as the base distribution. This allows randomized choices of the prior to be generated, avoiding costly and/or intractable estimation of this prior. It also provides measures of uncertainty in the prior choice, allowing subsequent learning tasks to be assessed for robustness to this prior choice. The instantiation of the method in already published applications in knowledge elicitation, recursive learning and flat cooperation of adaptive controllers is recalled, and prospective application domains are also mentioned.

1. Bayesian transfer learning via prior elicitation

A fundamental concern in Bayesian transfer learning [32,34] is to exploit the probabilistic knowledge of an external agent (known as the source task), and use it to improve the Bayesian learning of a primary agent (known as the target task). A principal mechanism by which this knowledge transfer can take place is for the primary agent (target) to be parametric. Then, transfer learning involves elicitation of the parameter prior conditioned on the external (source) knowledge. Prior elicitation is a fundamental concern in Bayesian learning [15,22], with many principles adopted for the purpose. In particular, the minimum Kullback–Leibler divergence (KLD) principle [17] provides an axiomatically justified way of transferring source knowledge when eliciting a distribution. Maximum entropy [11] and minimum cross-entropy [29] prior elicitation are special cases. More recently, fully probabilistic design (FPD) [25] has provided an unrestricted minimum-KLD mechanism for eliciting Bayesian models. The current paper applies this general theory to a specific problem of Bayesian transfer learning. We elicit an optimal parameter prior conditioned by probabilistic source knowledge in the form of a predictor. This generalizes a result [19] available in the special case where the source knowledge is in the form of a random sample. It also equips the elicited prior with uncertainty quantifiers, something that is important in robustness studies [27].

* Corresponding author.
E-mail addresses: aquinn@tcd.ie (A. Quinn), school@utia.cas.cz (M. Kárný), guy@utia.cas.cz (T.V. Guy).

http://dx.doi.org/10.1016/j.ijar.2017.02.001
0888-613X/

The formulation of the transfer learning problem in this paper—and its solution via FPD—lies at an intersection between statistics, adaptive prediction and control, and decision making under uncertainty. The notation and vocabulary therefore differ in some respects from those usually adopted in machine learning, but the correspondences are explained where appropriate throughout the paper. In particular, the application contexts in machine learning [9,10,32,34,35] are reviewed.

2. Task definition and solution methodology

We address the fundamental task of optimally choosing (which we call *designing* in this paper) the prior distribution of a parameter, Θ, by transferring knowledge from an external source [34]. Specifically, this knowledge is transferred in the form of a predictive distribution, $\mathsf{F}_{\mathsf{E}}(x)$. It quantifies the beliefs of the external source in respect of an uncertain quantity of interest, x, and having processed a data set, survey or other statistical resources not directly available to the primary Θ-parameterized modeler of x (i.e. the target task). In the case where this predictor is degenerate at a point, x_{E} (i.e. the source knowledge is crisp), then Bayes' rule is the unique consistent mechanism for transferring x_{E} into a (conditional) distribution for Θ, once the primary modeler specifies a stochastic generative model, $\mathsf{F}_{\mathsf{I}}(x|\Theta)$. We interpret $\mathsf{F}_{\mathsf{E}}(x)$ as a constraint on the primary modeler's knowledge. Furthermore, we adopt a fully Bayesian hierarchical framework, furnishing the unknown parameter prior with its own hyper-prior.

In this paper, the optimal design of the hierarchy in the presence of the external predictive constraint, $\mathsf{F}_{\mathsf{E}}(x)$, is treated as a problem of randomized decision-making. We adopt fully probabilistic design (FPD), which is an axiomatic framework for optimal design of decision strategies. Its formal justification was provided in [17], and extended in [25] to the design of distributional hierarchies. FPD is closely related to KL control proposed by [12,33], and developed and applied in a range of papers, e.g. [8,13]. FPD is also consistent with Bernardo's theory on the utility-based ranking of distributional approximations [3]. The relations between FPD designs, and those emerging from classical maximum entropy [11] and minimum cross-entropy [29] principles, were explored in [25].

2.1. Layout

In Section 3, we formalize the Bayesian transfer learning problem in terms of FPD-based optimal parameter prior design, and we prove our main result by way of a theorem. This specializes the theory in [25] to the case where the source knowledge is a predictive distribution. In Section 4, we examine special cases of the prior design, based on the modeler's specific ideals. This provides formal justification for designs which have been proposed informally in [19] and applied in [15]. In Section 4, we relate the proposed methodology to published contexts in machine learning and related domains. In Section 5, we apply the theory specifically in a recursive signal processing context, in knowledge elicitation and in distributed adaptive control. We conclude the paper in Section 6, mainly by aligning our results with the general area of Bayesian knowledge transfer, and transfer learning, while pointing to promising future application domains.

2.2. Notational conventions

- M, A, S, F denote probability distributions. They are described by densities with respect to a dominating measure, either counting or Lebesgue [26].
- M and A denote unspecified (variational) distributions; F denotes a distribution whose functional form is known *a priori*.
- Subscripts, $_{\mathsf{I}}$ and $_{\mathsf{E}}$, denote *ideal* (specified by the modeler) and *external* quantities, respectively, as defined in Section 3.
- FPD-optimal objects (distributions, parameters, estimates, etc.) are distinguished by the superscript ^{o}, with definitions given in Section 3.
- $\mathsf{E}_{\mathsf{M}}[\cdot]$ is the expectation of the argument with respect to the distribution M.
- Bold symbols, $\mathbf{M}$, $\boldsymbol{x}$, etc., denote the set of possible values of an unknown quantity, M, x, etc.

3. FPD of the predictor-constrained parameter prior

Consider a Bayesian parametric modeler, $\mathcal{I}$, of an unknown quantity, x. Typically, x is a future observation, and $\mathcal{I}$ adopts the parametric distribution, $\mathsf{F}_{\mathsf{I}}(x|\Theta)$, which is known up to the value of the finite-dimensional parameter, $\Theta \in \boldsymbol{\Theta}$. In considered Bayesian modeling, $\mathcal{I}$ also specifies a prior distribution, $\mathsf{F}_{\mathsf{I}}(\Theta)$, for Θ, which quantifies their beliefs about Θ before x is observed. $\mathcal{I}$'s *ideal*[1] model of the augmented set, $(x, \Theta) \in (\boldsymbol{x} \times \boldsymbol{\Theta})$, is therefore factorized by the chain rule of probability functions:

$$\mathsf{M}_{\mathsf{I}}(x, \Theta|\mathsf{F}_{\mathsf{I}}) \equiv \mathsf{F}_{\mathsf{I}}(x, \Theta) = \mathsf{F}_{\mathsf{I}}(x|\Theta)\mathsf{F}_{\mathsf{I}}(\Theta). \tag{1}$$

The role of the ideal distribution in our work will be explored after (7).

[1] The term 'ideal' is used in papers dealing with fully probabilistic design [17,25]. In the related KL control literature [8,12], terms such as 'reference/desired/modeler-specified distribution' are used, but with a much narrower interpretation of its role.

We extend this usual setting of Bayesian inference to a broader transfer learning context, as follows:

- Additional knowledge about x is transferred to $\mathcal{I}$ via a known, externally sourced, predictive distribution, $\mathsf{F_E}(x)$. This is the external source's probabilistic knowledge of the *same* quantity, x, that is being modeled parametrically by $\mathcal{I}$, via $\mathsf{F_I}(x|\Theta)$.
- $\mathcal{I}$'s prior distribution of Θ, following the transfer of $\mathsf{F_E}(x)$, is denoted by $\mathsf{A}(\Theta|\mathsf{F_E})$. This prior is assumed to be *unknown* to $\mathcal{I}$. Therefore, following the Bayesian framework of this paper, it is modeled by $\mathcal{I}$ hierarchically, via the hyper-prior, $\mathsf{S}(\mathsf{A}|\mathsf{F_E})$.
  To avoid technicalities connected with the fact that $\mathsf{A} \sim \mathsf{S}$ is—in the most general case—a nonparametric process, we assume that the set $\boldsymbol{\Theta}$ has a finite cardinality. This allows us to treat S as a parametric density, as assumed in Section 2.2. The form of the resulting solution will not depend on this assumption, as further discussed in Section 4.3.

$\mathcal{I}$'s augmented collection of unknowns is therefore:

$$(x, \Theta, \mathsf{A}) \in \boldsymbol{x} \times \boldsymbol{\Theta} \times \mathbf{A}. \tag{2}$$

$\mathcal{I}$'s joint model, M, is now conditioned on the unspecified (i.e. variational) hyper-prior, $\mathsf{S}(\mathsf{A}|\mathsf{F_E})$, and on the specified transferred knowledge, $\mathsf{F_E}(x)$. It factorizes via the chain rule:

$$\mathsf{M}(x, \Theta, \mathsf{A}|\mathsf{S}, \mathsf{F_E}) \equiv \mathsf{M}(x|\Theta, \mathsf{A}, \mathsf{S}, \mathsf{F_E})\mathsf{M}(\Theta|\mathsf{A}, \mathsf{S}, \mathsf{F_E})\mathsf{M}(\mathsf{A}|\mathsf{S}, \mathsf{F_E}),$$

where the hyper-prior distribution, S, is defined in the second bullet point above. We adopt the following assumptions:

- $\mathcal{I}$ replaces their parametric knowledge of x with the externally sourced predictor, $\mathsf{F_E}(x)$. Technically, given $\mathsf{F_E}(x)$, $\mathcal{I}$ models x as conditionally independent of Θ:

  $$\mathsf{M}(x|\Theta, \mathsf{A}, \mathsf{S}, \mathsf{F_E}) \equiv \mathsf{F_E}(x).$$

  This key assumption allows a systematic exploitation of the knowledge accumulated by the external source irrespective of *how* that knowledge was accumulated. It can, for instance, arise as: (i) the probabilistic description of the whole population to which the modeled object, x, belongs; (ii) the predictor obtained via a completely different type of model from the one considered; (iii) the probabilistic expert's belief about x gained by informal reasoning; and (iv) the empirical distribution observed on a similar object (e.g. by simulating x).
- We invoke the usual conditional independence property of a distributional hierarchy:

  $$\mathsf{M}(\Theta|\mathsf{A}, \mathsf{S}, \mathsf{F_E}) \equiv \mathsf{A}(\Theta|\mathsf{F_E}). \tag{3}$$

  This reflects the fact that knowledge of the externally supplied distribution, $\mathsf{F_E}$, is sufficient for influencing the description of the unknown Θ, via A. Thus, the hyper-prior S generating it brings no additional information about Θ. This also reflects the usual Markov dependency structure between Θ, A and S in the hierarchy.
- By the definition of S:

  $$\mathsf{M}(\mathsf{A}|\mathsf{S}, \mathsf{F_E}) \equiv \mathsf{S}(\mathsf{A}|\mathsf{F_E}).$$

$\mathcal{I}$'s joint model for (2), conditioned on S and $\mathsf{F_E}$, therefore becomes:

$$\mathsf{M}(x, \Theta, \mathsf{A}|\mathsf{S}, \mathsf{F_E}) \equiv \mathsf{F_E}(x)\mathsf{A}(\Theta|\mathsf{F_E})\mathsf{S}(\mathsf{A}|\mathsf{F_E}). \tag{4}$$

We seek an optimal and principled choice for the unknown

$$\mathsf{M} \in \mathbf{M} \equiv \{\text{models of the form (4), } \mathsf{F_E} \text{ fixed, } \mathsf{A} \text{ generated by variational } \mathsf{S}\}. \tag{5}$$

In general, if: (i) $\mathcal{I}$'s knowledge confines M to a specific set, $\mathbf{M}$; and (ii) $\mathcal{I}$ specifies an ideal choice for M, denoted by the known distribution, $\mathsf{M_I}$; then the FPD optimal model—denoted by M^o—is defined as the following unique, axiomatically optimal choice, consistent with this knowledge and the ideal specification:

$$\mathsf{M}^o(x, \Theta, \mathsf{A}|\mathsf{S}^o, \mathsf{F_E}) \equiv \arg\min_{\mathsf{M}\in\mathbf{M}} \mathcal{D}(\mathsf{M}||\mathsf{M_I}), \quad \text{where} \tag{6}$$

$\mathcal{D}(\mathsf{M}||\mathsf{M_I})$ is the Kullback–Leibler divergence (KLD, [20]) from M to $\mathsf{M_I}$:

$$\mathcal{D}(\mathsf{M}||\mathsf{M_I}) \equiv \mathsf{E_M}\left[\ln\left(\frac{\mathsf{M}}{\mathsf{M_I}}\right)\right]. \tag{7}$$

We refer to M^o as the FPD-optimal choice of (augmented) model, and we refer to the Bayesian modeler, $\mathcal{I}$, who chooses M^o, as the FPD-optimal model. Conceptually, FPD chooses M^o as the one closest to $\mathsf{M_I}$ (in the KLD sense), subject to the knowledge constraint, $\mathsf{M} \in \mathbf{M}$.

Note, from the definition of FPD (6), that $\mathsf{M}^o = \mathsf{M}_\mathsf{I}$ if $\mathsf{M}_\mathsf{I} \in \mathbf{M}$. Therefore, the ideal can be interpreted as the $\mathcal{I}$-specified optimal choice *in the absence of any active knowledge constraint*. It acts as the datum to which knowledge-constrained choices, $\mathsf{M} \in \mathbf{M}$, are referenced, allowing these to be ranked in a KLD sense (6).

It remains to specify $\mathcal{I}$'s ideal distribution, M_I, in (6) and (7), on the augmented set (2). As already noted in (1), $\mathcal{I}$'s model for x—prior to the transfer of the externally sourced knowledge, $\mathsf{F}_\mathsf{E}(x)$—is parametric, being $\mathsf{F}_\mathsf{I}(x|\Theta)$. The parameter pre-prior is $\mathsf{F}_\mathsf{I}(\Theta)$, which, again, must be understood as $\mathcal{I}$'s knowledge of unknown Θ, not only prior to x being realized, but also prior to the transfer of the probabilistic knowledge of x, i.e. $\mathsf{F}_\mathsf{E}(x)$, from the external source. The optimal design of M (4) should therefore be close to this ideal, while also processing the supplied knowledge, $\mathsf{F}_\mathsf{E}(x)$. The appropriate specification of $\mathcal{I}$'s ideal augmented distribution for the purposes of FPD (6) is therefore:

$$\begin{aligned}\mathsf{M}_\mathsf{I}(x, \Theta, \mathsf{A}|\mathsf{F}_\mathsf{I}, \mathsf{S}_\mathsf{I}, \mathsf{F}_\mathsf{E}) &\equiv \mathsf{M}_\mathsf{I}(x, \Theta|\mathsf{F}_\mathsf{I})\mathsf{M}_\mathsf{I}(\mathsf{A}|\mathsf{S}_\mathsf{I}, \mathsf{F}_\mathsf{E}) \\ &\equiv \mathsf{F}_\mathsf{I}(x|\Theta)\mathsf{F}_\mathsf{I}(\Theta)\mathsf{S}_\mathsf{I}(\mathsf{A}|\mathsf{F}_\mathsf{E}); \end{aligned} \tag{8}$$

i.e. $\mathcal{I}$ specifies their ideal for $(x, \Theta) \in \boldsymbol{x} \times \boldsymbol{\Theta}$ and for A conditionally independently, given F_I and S_I, respectively. We will examine appropriate specifications of $\mathsf{S}_\mathsf{I}(\mathsf{A}|\mathsf{F}_\mathsf{E})$ in Section 4. For now, we derive an expression for the FPD-optimal choice of augmented prior model (6), in the transfer learning context specified above.

Theorem 1 *(FPD of predictor-constrained prior). Let the Bayesian modeler, $\mathcal{I}$: (i) adopt the ideal defined via the hierarchical model (8); (ii) constrain their knowledge via transfer of an externally sourced predictor, $\mathsf{F}_\mathsf{E}(x)$, in the way described in (4); and (iii) place no special constraints on the sets, $\mathbf{A} \neq \emptyset$ and $\mathbf{S} \neq \emptyset$, of A and S respectively. Then, the FPD-optimal hierarchical model, being the solution of (6), is:*

$$\mathsf{M}^o(x, \Theta, \mathsf{A}|\mathsf{S}^o, \mathsf{F}_\mathsf{E}) = \mathsf{F}_\mathsf{E}(x)\mathsf{A}(\Theta|\mathsf{F}_\mathsf{E})\mathsf{S}^o(\mathsf{A}|\mathsf{F}_\mathsf{E}), \quad \textit{where} \tag{9}$$

$$\mathsf{S}^o(\mathsf{A}|\mathsf{F}_\mathsf{E}) \propto \mathsf{S}_\mathsf{I}(\mathsf{A}|\mathsf{F}_\mathsf{E}) \exp\left(-\mathcal{D}(\mathsf{A}||\hat{\mathsf{A}})\right) \tag{10}$$

$$\hat{\mathsf{A}}(\Theta|\mathsf{F}_\mathsf{E}) \propto \mathsf{F}_\mathsf{I}(\Theta) \exp\left(\int_{\boldsymbol{x}} \ln(\mathsf{F}_\mathsf{I}(x|\Theta))\mathsf{F}_\mathsf{E}(x)\,\mathrm{d}x\right), \tag{11}$$

while it is assumed that (11) is a proper distribution. The consistent FPD-optimal design of A is

$$\mathsf{M}^o(\Theta|\mathsf{S}^o, \mathsf{F}_\mathsf{E}) \equiv \mathsf{A}^o(\Theta|\mathsf{F}_\mathsf{E}) = \mathsf{E}_{\mathsf{S}^o}[\mathsf{A}]. \tag{12}$$

Proof. Inserting (4) and (8) into (7):

$$\begin{aligned}\mathcal{D}(\mathsf{M}||\mathsf{M}_\mathsf{I}) &= \int_{\boldsymbol{x}\times\boldsymbol{\Theta}\times\mathbf{A}} \ln\left(\frac{\mathsf{F}_\mathsf{E}(x)\mathsf{A}(\Theta|\mathsf{F}_\mathsf{E})\mathsf{S}(\mathsf{A}|\mathsf{F}_\mathsf{E})}{\mathsf{F}_\mathsf{I}(x|\Theta)\mathsf{F}_\mathsf{I}(\Theta)\mathsf{S}_\mathsf{I}(\mathsf{A}|\mathsf{F}_\mathsf{E})}\right) \\ &\quad \times \mathsf{F}_\mathsf{E}(x)\mathsf{A}(\Theta|\mathsf{F}_\mathsf{E})\mathsf{S}(\mathsf{A}|\mathsf{F}_\mathsf{E})\,\mathrm{d}x\,\mathrm{d}\Theta\,\mathrm{d}\mathsf{A} \\ &= \int_{\boldsymbol{\Theta}\times\mathbf{A}} \left[\int_{\boldsymbol{x}} \mathsf{F}_\mathsf{E}(x) \ln\left(\frac{\mathsf{F}_\mathsf{E}(x)}{\mathsf{F}_\mathsf{I}(x|\Theta)}\right)\mathrm{d}x + \ln\left(\frac{\mathsf{A}(\Theta|\mathsf{F}_\mathsf{E})\mathsf{S}(\mathsf{A}|\mathsf{F}_\mathsf{E})}{\mathsf{F}_\mathsf{I}(\Theta)\mathsf{S}_\mathsf{I}(\mathsf{A}|\mathsf{F}_\mathsf{E})}\right)\right] \\ &\quad \times \mathsf{A}(\Theta|\mathsf{F}_\mathsf{E})\mathsf{S}(\mathsf{A}|\mathsf{F}_\mathsf{E})\,\mathrm{d}\Theta\,\mathrm{d}\mathsf{A} \\ &= \int_{\boldsymbol{\Theta}\times\mathbf{A}} \ln\left(\frac{\mathsf{A}(\Theta|\mathsf{F}_\mathsf{E})\mathsf{S}(\mathsf{A}|\mathsf{F}_\mathsf{E})}{\hat{\mathsf{A}}(\Theta|\mathsf{F}_\mathsf{E})\mathsf{S}_\mathsf{I}(\mathsf{A}|\mathsf{F}_\mathsf{E})}\right)\mathsf{A}(\Theta|\mathsf{F}_\mathsf{E})\mathsf{S}(\mathsf{A}|\mathsf{F}_\mathsf{E})\,\mathrm{d}\Theta\,\mathrm{d}\mathsf{A} - \mathcal{H}_{\mathsf{F}_\mathsf{E}} - \ln(c_{\mathsf{A}^o}),\end{aligned}$$

where $\hat{\mathsf{A}}(\Theta|\mathsf{F}_\mathsf{E})$ is defined in (11), with normalizing constant,

$$c_{\mathsf{A}^o} \equiv \int_{\boldsymbol{\Theta}} \mathsf{F}_\mathsf{I}(\Theta) \exp\left(\int_{\boldsymbol{x}} \mathsf{F}_\mathsf{E}(x) \ln(\mathsf{F}_\mathsf{I}(x|\Theta))\,\mathrm{d}x\right)\mathrm{d}\Theta,$$

and $c_{\mathsf{A}^o} < \infty$ by assumption. Also,

$$\mathcal{H}_{\mathsf{F}_\mathsf{E}} \equiv -\int_{\boldsymbol{x}} \mathsf{F}_\mathsf{E}(x) \ln(\mathsf{F}_\mathsf{E}(x))\,\mathrm{d}x$$

is the differential entropy of the external predictor, F_E. Hence:

$$\mathcal{D}(\mathsf{M}||\mathsf{M}_\mathsf{I}) = \int_{\mathbf{A}} \left[\mathcal{D}(\mathsf{A}||\hat{\mathsf{A}}) + \ln\left(\frac{\mathsf{S}(\mathsf{A}|\mathsf{F}_\mathsf{E})}{\mathsf{S}_\mathsf{I}(\mathsf{A}|\mathsf{F}_\mathsf{E})}\right)\right] \mathsf{S}(\mathsf{A}|\mathsf{F}_\mathsf{E})\,\mathrm{d}\mathsf{A} - \mathcal{H}_{\mathsf{F}_\mathsf{E}} - \ln(c_{\mathsf{A}^o})$$

$$= \mathcal{D}(\mathsf{S}||\mathsf{S}^o) - \mathcal{H}_{\mathsf{F}_\mathsf{E}} - \ln(c_{\mathsf{A}^o} c_{\mathsf{S}^o}), \tag{13}$$

where $\mathsf{S}^o(\mathsf{A}|\mathsf{F}_\mathsf{E})$ is defined in (10), with normalizing constant,

$$c_{\mathsf{S}^o} \equiv \int_{\mathbf{A}} \mathsf{S}_\mathsf{I}(\mathsf{A}|\mathsf{F}_\mathsf{E}) \exp\left(-\mathcal{D}(\mathsf{A}||\hat{\mathsf{A}})\right) \mathrm{d}\mathsf{A},$$

and $c_{\mathsf{S}^o} < \infty$ (i.e. S^o is proper). By the properties of the KLD, the first term on the right-hand side of (13) achieves its minimal value of 0 if $\mathsf{S} = \mathsf{S}^o$. Furthermore, the second and third terms are invariant with S. Inserting (13) into (6), and noting that $\mathsf{S}(\mathsf{A}|\mathsf{F}_\mathsf{E})$ is the only free (variational) factor in M (4), the FPD-optimal design of M is (9). Noting, from (9), that the FPD-optimal predictor of x is $\mathsf{M}^o(x|\mathsf{S}^o, \mathsf{F}_\mathsf{E}) = \mathsf{F}_\mathsf{E}(x)$ (the externally supplied predictor), by design, and that x is independent of (Θ, A) under M^o, it follows that the FPD-optimal distribution of A is $\mathsf{M}^o(\mathsf{A}|\mathsf{S}^o, \mathsf{F}_\mathsf{E}) = \mathsf{S}^o(\mathsf{A}|\mathsf{F}_\mathsf{E})$ (10). Furthermore, the FPD-optimal distribution of $\Theta \in \boldsymbol{\Theta}$ is, by marginalization,

$$\mathsf{M}^o(\Theta|\mathsf{S}^o, \mathsf{F}_\mathsf{E}) = \int_{\mathbf{A}} \mathsf{M}^o(\Theta, \mathsf{A}|\mathsf{S}^o, \mathsf{F}_\mathsf{E})\,\mathrm{d}\mathsf{A}$$

$$\overset{(9)}{=} \int_{\mathbf{A}} \mathsf{A}\, \mathsf{S}^o(\mathsf{A}|\mathsf{F}_\mathsf{E})\,\mathrm{d}\mathsf{A} \equiv \mathsf{E}_{\mathsf{S}^o}[\mathsf{A}]. \qquad \square$$

4. Discussion

Fully probabilistic design of hierarchical models—in quite general settings—has been published recently [25]. Hierarchical FPD allows many specifications of probabilistic knowledge—including hierarchical moment constraints, generalizing conventional entropy optimization settings [11,29]—to be transferred from an external source to the primary modeler, $\mathcal{I}$, i.e. to the target [34]. The crisp (estimated) prior elicitations, $\hat{\mathsf{A}}(\cdot)$, of conventional approaches are here replaced by an optimal stochastic model, $\mathsf{A} \sim \mathsf{S}^o$, allowing randomized and/or robust designs of A to be chosen. Theorem 1 above can be understood as a specialization of the general framework [25] to the problem of the FPD-optimal parameter prior design for a parametric hierarchy, M^o (9), under the constraint of an external predictive distribution, $\mathsf{F}_\mathsf{E}(x)$. It accomplishes an important Bayesian transfer learning task with wide applicability, as addressed in Sections 5 and 6.

The discussion below relates our solution to specific settings in the literature, providing additional insight into it.

4.1. On special cases of the FPD-optimal prior, $\mathsf{A}^o(\Theta|\mathsf{F}_\mathsf{E})$

The FPD-optimal design of $\mathsf{A}(\Theta|\mathsf{F}_\mathsf{E})$, having transferred the predictive distribution, $\mathsf{F}_\mathsf{E}(x)$, from the external source is given by $\mathsf{A}^o(\Theta|\mathsf{F}_\mathsf{E})$ (12), and its FPD-optimal stochastic model (hyper-prior) is given by (10). Note that $\hat{\mathsf{A}}(\Theta|\mathsf{F}_\mathsf{E})$ (11) enters $\mathsf{S}^o(\mathsf{A}|\mathsf{F}_\mathsf{E})$ as its base distribution [7], around which randomized choices of $\mathsf{A}(\Theta|\mathsf{F}_\mathsf{E})$ are distributed. Of course, $\hat{\mathsf{A}} \neq \mathsf{A}^o$ in general, see (11), (12). From (10), $\hat{\mathsf{A}}$ is the maximum *a posteriori* (MAP) distributional estimate of A in the case where S_I is a high-entropy (i.e. diffuse) distribution.

Consider the special case where the transferred knowledge of x is certain, with x_E being a 'crisp' realization of x:

$$\mathsf{F}_\mathsf{E}(x) \equiv \delta(x - x_\mathsf{E}). \tag{14}$$

Here, $\delta(\cdot)$ is the distribution that is degenerate at x_E. Inserting this into (11), and using the standard integral property of $\delta(\cdot)$:

$$\hat{\mathsf{A}}(\Theta|\mathsf{F}_\mathsf{E}) \equiv \hat{\mathsf{A}}(\Theta|x_\mathsf{E}) \propto \mathsf{F}_\mathsf{I}(\Theta)\mathsf{F}_\mathsf{I}(x_\mathsf{E}|\Theta).$$

Therefore, in this case, the FPD-optimal transfer of external knowledge to $\mathcal{I}$'s base distribution (10) is consistent with Bayes' rule using the ideal model, M_I (1). However, this is not true in general. Consider the transfer of an external random sample, $\mathbf{x}_\mathsf{E} \equiv \{x_{\mathsf{E},1}, \ldots, x_{\mathsf{E},\nu_\mathsf{E}}\}$, $\nu_\mathsf{E} \geq 1$. The induced predictive distribution under a Dirichlet nonparametric process prior is the empirical distribution [7]:

$$\mathsf{F}_\mathsf{E}(x) = \nu_\mathsf{E}^{-1} \sum_i \delta(x - x_{\mathsf{E},i}). \tag{15}$$

For consistency with Bayes' rule in this $\nu_\mathsf{E} \neq 1$ case, the following heuristic adaptation of (11) is necessary:

$$\hat{\mathsf{A}}(\Theta|\mathsf{F}_\mathsf{E}, \nu_\mathsf{E}) \propto \mathsf{F}_\mathsf{I}(\Theta) \exp\left(\nu_\mathsf{E} \int_{\boldsymbol{x}} \ln \mathsf{F}_\mathsf{I}(x|\Theta)\mathsf{F}_\mathsf{E}(x)\,\mathrm{d}x\right). \tag{16}$$

Now, inserting (15) into (16) leads again to Bayes' rule:

$$\hat{\mathsf{A}}(\Theta|\mathsf{F}_\mathsf{E}, \nu_\mathsf{E}) \equiv \hat{\mathsf{A}}(\Theta|\mathbf{x}_\mathsf{E}) \propto \mathsf{F}_\mathsf{I}(\Theta) \prod_i \mathsf{F}_\mathsf{I}(x_{\mathsf{E},i}|\Theta). \tag{17}$$

The deterministic prior (16) for transferring an externally sourced predictive distribution, F_E, was informally proposed in [19], based on the progression (15)–(17). In Section 4.2, we will discuss how the adaptation (16) might be justified in the formal (FPD) context of the current paper.

Theorem 1 reveals that (11) is the MAP distributional estimate of $\mathsf{A}(\Theta|\mathsf{F}_\mathsf{E})$, if a high-entropy $\mathsf{S}_\mathsf{I}(\mathsf{A}|\mathsf{F}_\mathsf{E})$ is chosen by the designer of S. As already mentioned, the hierarchical setting of this paper quantifies the uncertainty in choices of unknown A, via the FPD-optimal hyper-prior, S^o (10). Among the advantages of this relaxation are: (i) that randomized choices, $\mathsf{A} \sim \mathsf{S}^o$, may be drawn, eliminating the need for (often expensive) computation of distributional estimates such as $\hat{\mathsf{A}}$ (11) or A^o (12); and (ii) that robustness to choices of A can be studied formally through the provision of S^o.

4.2. On degrees-of-freedom parameters

Fully probabilistic design allows free specification of the ideal factors in (8) by the designer.[2] The FPD design, (10) and (11), reveals the conjugate (invariant) choice [5] for $\mathsf{S}_\mathsf{I}(\mathsf{A}|\mathsf{F}_\mathsf{E})$ and $\mathsf{F}_\mathsf{I}(\Theta)$, respectively. Adopting these (with degrees-of-freedom parameters, $\nu_{\mathsf{S}_\mathsf{I}} > 0$ and $\nu_{\mathsf{F}_\mathsf{I}} > 0$, respectively), then—by construction—the following functionally invariant forms result:

$$\mathsf{S}^o(\mathsf{A}|\mathsf{F}_\mathsf{E}, \nu_{\mathsf{S}^o}) \propto \exp\left(-\nu_{\mathsf{S}^o}\mathcal{D}(\mathsf{A}||\hat{\mathsf{A}})\right), \quad \text{where}$$

$$\hat{\mathsf{A}}(\Theta|\mathsf{F}_\mathsf{E}, \nu_{\hat{\mathsf{A}}}) \propto \exp\left(\nu_{\hat{\mathsf{A}}} \int_{\boldsymbol{x}} \ln \mathsf{F}_\mathsf{I}(x|\Theta)\mathsf{F}_\mathsf{E}(x)\,\mathrm{d}x\right). \tag{18}$$

Here, $\nu_{\mathsf{S}^o} = \nu_{\mathsf{S}_\mathsf{I}} + 1$, and $\nu_{\hat{\mathsf{A}}} = \nu_{\mathsf{F}_\mathsf{I}} + 1$. This same adapted form of $\hat{\mathsf{A}}$ is recovered in the case where an annealed observation model [31] is adopted in (8):

$$\mathsf{M}_\mathsf{I}(x|\Theta, \mathsf{F}_\mathsf{I}, \nu_{\hat{\mathsf{A}}}) \propto [\mathsf{F}_\mathsf{I}(x|\Theta)]^{\nu_{\hat{\mathsf{A}}}}.$$

$T_{\hat{\mathsf{A}}} \equiv \nu_{\hat{\mathsf{A}}}^{-1}$ is the chosen annealing temperature [31], which may be set by the designer in order to control the influence of the knowledge transfer through $\hat{\mathsf{A}}$ (18), ranging from the uniform case when $T_{\hat{\mathsf{A}}} \to \infty$ (little knowledge transfer), to cases that are highly informed by $\mathsf{F}_\mathsf{E}(x)$, when $T_{\hat{\mathsf{A}}} \to 0$. Comparing (18) with (16), and assuming a high-entropy pre-prior, $\mathsf{F}_\mathsf{I}(\Theta)$, a related interpretation of $\nu_{\hat{\mathsf{A}}}$ (18) is therefore as a prior degrees-of-freedom parameter, counting the number of samples available for transfer from the external source via the empirical predictor, $\mathsf{F}_\mathsf{E}(x)$.

These proposals for engendering (non-unitary) ν_{S^o} and $\nu_{\hat{\mathsf{A}}}$ in (18) are unsatisfactory in one key respect: they depend on informative and highly structured prior distributions, without clearly specifying the knowledge thereby processed by the modeler, $\mathcal{I}$ (Section 3). A focus of our current research is on the formal specification of such knowledge in hierarchical FPD. This will allow Theorem 1 (i.e. (10) and (11)) to be adapted axiomatically, yielding the degrees-of-freedom terms in (18). A preliminary proposal along these lines has recently been published [14].

It was shown in [25] that the FPD-optimal choice of the hyper-prior $\mathsf{S}(\mathsf{A}|\mathsf{F}_\mathsf{E})$ (4) concentrates at $\hat{\mathsf{A}}$ (11), i.e.

$$\mathsf{S}^o(\mathsf{A}|\mathsf{F}_\mathsf{E}) = \delta(\mathsf{A} - \hat{\mathsf{A}}),$$

if $\mathsf{S}_\mathsf{I} \equiv \mathsf{S}$. This is the case where no ideal is specified for S. It is therefore the formal condition which justifies the *deterministic* prior design in [19]. As discussed in Section 4.1, there are significant advantages in the *stochastic* relaxation achieved in Theorem 1, by way of $\mathsf{A} \sim \mathsf{S}^o$ (10).

4.3. On the nonparametric case

Recall that the adoption of a probability distribution (density), $\mathsf{S}(\mathsf{A}|\mathsf{F}_\mathsf{E})$, on $\mathsf{A}(\Theta|\mathsf{F}_\mathsf{E}) \in \mathbf{A}$ (4) formally makes A parametric. If this assumption is relaxed, then the fully probabilistic design of a nonparametric hierarchy [7] (4) is required. In order to avoid technicalities of measure theory, $\Theta \in \boldsymbol{\Theta}$ can be quantized into $m < \infty$ measurable states, ensuring that A is a

[2] Nevertheless, the parametric model for x, being $\mathsf{F}_\mathsf{I}(x|\Theta)$ (1), is often informed or imposed by the modeler's knowledge of the mechanism, often physical, for generation of x.

probability mass function in the $(m-1)$-dimensional open probability simplex [24]. Since no constraint is placed on the value and number, m, of the vertices of this quantizer, it follows that (10) is the FPD-optimal choice of S for any finite measurable partition of $\boldsymbol{\Theta}$. We conjecture that this result can be fully extended to a nonparametric process, A [7], and that (10) is, indeed, the FPD-optimal prior for arbitrary algebras on $\boldsymbol{\Theta}$.

5. Current applications of the proposed transfer learning mechanism

Bayesian transfer learning via prior elicitation, in the presence of probabilistic knowledge transferred from an external source, has many important precedents in the machine learning literature. As a way of encouraging the adoption of the axiomatic framework in this paper, we summarize: (i) a classic application in signal processing, Subsection 5.1; (ii) its use in prior elicitation, Subsection 5.2; (iii) an application in distributed adaptive control, Subsection 5.3. These illustrate the potential ease with which the FPD-optimal knowledge transfer can be accomplished in quite diverse domains.

5.1. Recursive learning of the normal mean and variance via transfer of a normal source

Consider: (i) Bayesian point estimation of the mean, θ, and variance, r—i.e. $\Theta \equiv (\theta, r)$ in (1)—of the normal model, $\mathcal{N}(\theta, r)$, for scalar x; (ii) an externally sourced *normal* predictor, $x \sim \mathsf{F}_{\mathsf{E}}(x) \equiv \mathcal{N}(x_{\mathsf{E}}, r_{\mathsf{E}})$; (iii) the conjugate, normal-inverse-gamma pre-prior distribution [2], $\mathsf{F}_{\mathsf{I}}(\Theta)$, with expected values, $\mathsf{E}_{\mathsf{F}_{\mathsf{I}}}[\theta] = \hat{\theta}_{\mathsf{I}}$ and $\mathsf{E}_{\mathsf{F}_{\mathsf{I}}}[r] = \hat{r}_{\mathsf{I}}$, and having $\nu_{\mathsf{F}_{\mathsf{I}}}$ degrees-of-freedom; and (iv) the FPD-optimal MAP distributional estimate, $\hat{\mathsf{A}}(\Theta|\mathsf{F}_{\mathsf{E}}, \nu_{\hat{\mathsf{A}}})$ (18), of the unknown prior, $\mathsf{A}(\Theta|\mathsf{F}_{\mathsf{E}})$, after transfer learning with F_{E}. Here, we are adopting the high-entropy specification of $\mathsf{S}_{\mathsf{I}}(\mathsf{A}|\mathsf{F}_{\mathsf{E}})$ in (10), and the variant (18) of (11), with $\nu_{\hat{\mathsf{A}}}$ degrees-of-freedom, which was discussed in Subsection 4.2. Under these assumptions, $\hat{\mathsf{A}}(\Theta|\mathsf{F}_{\mathsf{E}}, \nu_{\hat{\mathsf{A}}})$ (18) preserves the conjugate normal-inverse-gamma form, with moments

$$\begin{aligned} \hat{\theta}^o &= \hat{\theta}_{\mathsf{I}} + \frac{\nu_{\mathsf{F}_{\mathsf{I}}}}{\nu^o} e_{\mathsf{E}}, \quad e_{\mathsf{E}} = x_{\mathsf{E}} - \hat{\theta}_{\mathsf{I}}, \quad \text{where} \quad \nu^o = \nu_{\mathsf{F}_{\mathsf{I}}} + \nu_{\hat{\mathsf{A}}}, \\ \hat{r}^o &= \hat{r}_{\mathsf{I}} + \frac{\nu_{\mathsf{F}_{\mathsf{I}}}}{\nu^o}(e_{\mathsf{E}}^2 - \hat{r}_{\mathsf{I}} + r_{\mathsf{E}}). \end{aligned} \tag{19}$$

The evolution (19), whose general setting is provided by Proposition 6 in [16], resembles a step of recursive least squares [23] and reduces to it for 'crisp' external prediction (14) with zero variance of F_{E}, i.e. $r_{\mathsf{E}} = 0$. The weight, $\nu_{\hat{\mathsf{A}}} > 0$, set by the designer, $\mathcal{I}$, allows the influence of the processed external information to be controlled, even in this special case. In sequential processing of actual realizations of x, the performance of the predictor, $\int_{\boldsymbol{\Theta}} \mathsf{F}_{\mathsf{I}}(x|\Theta)\hat{\mathsf{A}}(\Theta|\mathsf{F}_{\mathsf{E}}, \nu_{\hat{\mathsf{A}}})\,\mathrm{d}\Theta$, can be optimized with respect to the weight $\nu_{\hat{\mathsf{A}}}$ [15].

This simple illustration points to the potential for optimal Bayesian transfer learning using FPD. Recursive computational flows of this kind are vital in computation-critical machine learning applications, such as those in real-time and/or high-dimensional contexts [28]. It illustrates how the uncertainty in the source knowledge (via $r_{\mathsf{E}} > 0$) is transferred to the target task in a principled way, and without disrupting the recursive flow. What this simple example does *not* illustrate is the practical impact of the theory in providing an optimal hyper-prior S^o for unknown A, via (10). A recursive structure will be preserved, if unknown A is assumed to have the parametric normal-inverse-gamma form.

The self-reproducing property of the distributions in this setting is not rare. It extends to the whole exponential family [1], being the prominent class of distributions that have self-reproducing [18] priors in the way shown above. This family includes multivariate Gaussian autoregressive–regressive normal models linear in the regression coefficients, the most generally parameterized Markov chain models, and many static models including the Poisson distribution, exponential distribution, etc.

5.2. Prior elicitation

In the classical Bayesian sequential learning context, the Bayesian modeler, $\mathcal{I}$, observes a sequence of data, $x_1, \ldots, x_t$, $t = 1, 2, \ldots$, and adopts the identical generative model, $\mathsf{F}_{\mathsf{I}}(x|\Theta)$ (1) independently for each $x_t \in \boldsymbol{x}$. $\mathcal{I}$ learns about their unknown, finite-dimensional parameter, $\Theta \in \boldsymbol{\Theta}$, by evolving their posterior distribution, $\mathsf{F}_t(\Theta)$, (uniquely) using Bayes' rule: $\mathsf{F}_t(\Theta) \propto \mathsf{F}_{\mathsf{I}}(x_t|\Theta)\mathsf{F}_{t-1}(\Theta)$. The choice of the prior, $\mathsf{F}(\Theta) \equiv \mathsf{F}_0(\Theta)$ (1), is critical, particularly when short-horizon (small t) decisions must be constructed on the basis of this learning. The sensitivity to prior choice is especially critical in model selection, an issue of central concern in machine learning [9]. $\mathsf{F}(\Theta) \equiv \mathsf{F}(\Theta|K)$ acts as the transfer/encoding/projection of $\mathcal{I}$'s (expert) knowledge, K, into a stochastic model for $\Theta \in \boldsymbol{\Theta}$, *prior* to observing the x_t's. K is typically in the form of opinions, obsolete data, data from related but distinct (e.g. simulated) systems, data expressed on inconsistent scales and in inconsistent units. The transfer of these into $\mathsf{F}(\Theta)$—i.e. the prior elicitation task—is therefore a demanding technical challenge [22]. Even when a transfer mechanism is agreed among several experts, based on agreed expert knowledge, K, there remains sensitivity to the pre-prior (reference prior) [4] used by the designer to initiate the prior elicitation.

The contributions of the FPD prior elicitation framework in this paper are, among others, that: (i) It is formulated under the assumption that the heterogeneous knowledge, K, has been transformed into the unit-less and universal $\mathsf{F}_{\mathsf{E}}(x) \equiv \mathsf{F}_{\mathsf{E}}(x|K)$, being the predictive distribution of the sequence, x_t, above. This motivated the work in [15], which proposed a machine transformation of heterogeneous K into $\mathsf{F}_{\mathsf{E}}(x)$. (ii) This universal representation of external knowledge

is then transferred to $\mathcal{I}$'s prior by merging with their pre-prior, $\mathsf{F}_{\mathsf{I}}(\Theta)$, yielding the axiomatically supported and optimal choice, $\mathsf{F}(\Theta) \equiv \mathsf{A}^{o}(\Theta|\mathsf{F}_{\mathsf{E}})$ (12). As already discussed, [15] achieved the transfer of $\mathsf{F}_{\mathsf{E}}(x)$ via the heuristically motivated base distribution (18) only. They elaborated the solution for the flexible Gaussian linear regressive and autoregressive generative model. (iii) Theorem 1 furnishes $\mathsf{A}^{o}(\Theta|\mathsf{F}_{\mathsf{E}})$ with the optimized pre-prior, $\mathsf{S}^{o}(\mathsf{A}|\mathsf{F}_{\mathsf{E}})$. This will be an important component in studying robustness to particular prior elicitation choices.

5.3. Distributed adaptive control and learning

Distributed adaptive control [16] addresses a potentially unlimited network of relatively simple control nodes. Any particular node controls a small number of variables, $y_1, \ldots, y_t$, at time moments, $t = 1, 2, \ldots$, and considers other accessible variables, $x_1, \ldots, x_t$, as external. These external variables provide knowledge about the state of the network interacting with the node in question at time t, but their control is handled by other nodes. This is a classical approach to the control of complex industrial, transportation and social systems, and is an area of active research [30]. The Bayesian transfer learning paradigm has a profound impact in this distributed adaptive control context. If we consider any one node in the network, it recursively learns a model of all its variables (x_t, y_t, along with Θ, specified below), and uses this model in order to design an optimal controller. The forecasted time evolution of the node's external variables, x_t, is modeled by the node using a simple parametric dynamic model, $\mathsf{F}_{\mathsf{I}}(x_t|x_{t-1}, \Theta)$. The node learns this parametrically, by evolving the posterior distribution, $\mathsf{F}_{\mathsf{I}}(\Theta|x_{t-1}, \ldots, x_1)$ [23]. As usual, it designs a controller for its controlled variables, y_t, by optimizing an appropriate criterion. In the Bayesian context, this optimization seeks to exploit the predictive distribution of the node's controlled variables, i.e. $\mathsf{F}_{\mathsf{E}}(y_t|x_{t-1}, \ldots, x_1)$, which is readily available at each step. The node's controlled variables, y_t, are treated as external variables ($x_t \equiv y_t$) by other nodes. Therefore, these other nodes can transfer F_{E} from nodes they interact with, in order to improve their $\mathsf{F}_{\mathsf{I}}(\Theta|x_{t-1}, \ldots, x_1)$ to $\mathsf{F}_{\mathsf{I}}(\Theta|x_{t-1}, \ldots, x_1, \mathsf{F}_{\mathsf{E}})$. It is precisely this external predictive knowledge transfer that is accomplished in an FPD-optimal way in this paper (Theorem 1).

This approach was adopted in [16] using the heuristic distributional estimate in (18), but without the uncertainty elicitation in this estimate provided by Theorem 1 (10). Also, this framework is readily applicable to contexts such as distributed filtering [6], exploitation of crowd wisdom [21], and many other distributed learning tasks.

6. Conclusions

Methodologically, this brief paper casts the heuristically motivated transfer of external probabilistic knowledge [19] into a wider axiomatically-supported framework of fully probabilistic design [17,25]. It achieves this through prior elicitation, and, in so doing, adopts a hierarchical setting, which furnishes the elicited prior with uncertainty. It is therefore a rather general tool for use in a wide range of Bayesian transfer learning tasks [34]. In addition to those applications already considered in Section 5, we can point to others which should benefit from a systematic application of the general tool presented here. These include the reinforcement learning problem addressed in [32], the combination of population and instance-specific predictors examined in [35], and a transfer learning paradigm for enhancing brain–computer interfaces adopted in [10]. This tiny reference sample illustrates the commonality of the problem that our paper addresses:

(i) a source of knowledge—external to the modeler, $\mathcal{I}$—is projected into a probabilistic predictor, $\mathsf{F}_{\mathsf{E}}(x)$, of some quantity, x;
(ii) a stochastic relation, $\mathsf{F}_{\mathsf{I}}(x|\Theta)$, of potentially observable x to unknown quantities, Θ, is modeled by $\mathcal{I}$;
(iii) $\mathcal{I}$'s knowledge about Θ, before transfer learning is accomplished, is quantified by a pre-prior, $\mathsf{F}_{\mathsf{I}}(\Theta)$; and
(iv) the external probabilistic knowledge, $\mathsf{F}_{\mathsf{E}}(x)$, is to be transferred to $\mathcal{I}$, improving $\mathsf{F}_{\mathsf{I}}(\Theta)$ to $\mathsf{A}(\Theta|\mathsf{F}_{\mathsf{E}})$.

As such, this paper is an invitation to the machine-learning, statistics, signal processing, adaptive systems and decision-making communities to deploy this fully probabilistic design framework in the many Bayesian transfer learning contexts which are amenable to it. Its promise is not only in its axiomatic justification and optimality. It is also in the construction of an optimal stochastic model for *choosing* $\mathsf{A}(\Theta|\mathsf{F}_{\mathsf{E}})$, establishing a Bayesian nonparametric (BNP) setting for the approach. This, in turn, brings the advantage of computing a randomized choice in place of (often expensive) computation of a distributional estimate. It also provides the basis for studying the robustness of down-stream learning to this prior choice. In particular, it can quantify the cost in replacing a complex choice of $\mathsf{A}(\Theta|\mathsf{F}_{\mathsf{E}})$ with a simpler one, and for balancing subsequent exploitation and exploration efforts. The hierarchical nature of this transfer learning solution is distinctive. Its full exploitation along these lines—as well as its formalization in a BNP setting—will be a worthwhile focus for our future research. A priority will also be to engender the degrees-of-freedom parameters in (18) via formal knowledge specification and FPD-based transfer.

Acknowledgement

This research has been supported by SFI grants 10/RFP/MTH2877 and 10/CE/11855 (Lero), and by GAČR grant GA16-09848S.

References

[1] O. Barndorff-Nielsen, Information and Exponential Families in Statistical Theory, Wiley, New York, 1978.
[2] J.O. Berger, Statistical Decision Theory and Bayesian Analysis, Springer, New York, 1985.
[3] J.M. Bernardo, Expected information as expected utility, Ann. Stat. 7 (3) (1979) 686–690.
[4] J.M. Bernardo, Reference posterior distributions for Bayesian inference, J. R. Stat. Soc. B 41 (1979) 113–147.
[5] J.M. Bernardo, A.F.M. Smith, Bayesian Theory, Wiley, 1994.
[6] F.S. Cattivelli, A.H. Sayed, Diffusion strategies for distributed Kalman filtering and smoothing, IEEE Trans. Autom. Control 55 (9) (2010) 2069–2084.
[7] T.S. Ferguson, A Bayesian analysis of some nonparametric problems, Ann. Stat. 1 (1973) 209–230.
[8] P. Guan, M. Raginsky, R.M. Willett, Online Markov decision processes with Kullback Leibler control cost, IEEE Trans. Autom. Control 59 (6) (2014) 1423–1438.
[9] I. Guyon, A. Saffari, G. Dror, G. Cawley, Model selection: beyond the Bayesian/frequentist divide, J. Mach. Learn. Res. 11 (2010) 61–87.
[10] V. Jayaram, M. Alamgir, Y. Altun, B. Schölkopf, M. Grosse-Wentrup, Transfer learning in brain–computer interfaces, IEEE Comput. Intell. Mag. 11 (1) (2016) 20–31.
[11] E.T. Jaynes, Probability Theory: The Logic of Science, Cambridge University Press, 2003.
[12] H.J. Kappen, Linear theory for control of nonlinear stochastic systems, Phys. Rev. Lett. 95 (20) (2005) 200201.
[13] H.J. Kappen, V. Gómez, M. Opper, Optimal control as a graphical model inference problem, Mach. Learn. 87 (2012) 159–182.
[14] M. Kárný, Towards implementable prescriptive decision making, in: T.V. Guy, et al. (Eds.), Proc. Workshop Imperfect Decision Makers: Admitting Real-World Rationality, in: Workshop and Conference Proceedings Series Journal of Machine Learning Research, vol. 58, 2017.
[15] M. Kárný, A. Bodini, T.V. Guy, J. Kracík, P. Nedoma, F. Ruggeri, Fully probabilistic knowledge expression and incorporation, Stat. Interface 7 (4) (2014) 503–515.
[16] M. Kárný, R. Herzallah, Scalable harmonization of complex networks with local adaptive controllers, IEEE Trans. Syst. Man Cybern. Syst. 47 (3) (2017) 394–404.
[17] M. Kárný, T. Kroupa, Axiomatisation of fully probabilistic design, Inf. Sci. 186 (1) (2012) 105–113.
[18] R. Koopman, On distributions admitting a sufficient statistic, Trans. Am. Math. Soc. 39 (1936) 399.
[19] J. Kracík, M. Kárný, Merging of data knowledge in Bayesian estimation, in: J. Filipe, J.A. Cetto, J.L. Ferrier (Eds.), Proc. of the Second Int. Conference on Informatics in Control, Automation and Robotics, Barcelona, INSTICC, 2005, pp. 229–232.
[20] S. Kullback, R. Leibler, On information and sufficiency, Ann. Math. Stat. 22 (1951) 79–87.
[21] W. Mason, J.W. Vaughan, H. Wallach, Special issue: Computational social science and social computing, Mach. Learn. 96 (2014) 257–469.
[22] A. O'Hagan, C.E. Buck, A. Daneshkhah, J.R. Eiser, P.H. Garthwaite, D.J. Jenkinson, J. Oakley, T. Rakow, Uncertain Judgement: Eliciting Experts' Probabilities, John Wiley & Sons, 2006.
[23] V. Peterka, Bayesian system identification, in: P. Eykhoff (Ed.), Trends and Progress in System Identification, Pergamon Press, Oxford, 1981, pp. 239–304.
[24] A. Quinn, M. Kárný, Learning for nonstationary Dirichlet processes, Int. J. Adapt. Control Signal Process. 21 (10) (2007) 827–855.
[25] A. Quinn, M. Kárný, T.V. Guy, Fully probabilistic design of hierarchical Bayesian models, Inf. Sci. 369 (2016) 532–547.
[26] M.M. Rao, Measure Theory and Integration, John Wiley, New York, 1987.
[27] D. Rios-Insua, F. Ruggeri (Eds.), Robust Bayesian Analysis, Springer Verlag, New York, 2000.
[28] Y.H. Shaoa, N.Y. Dengb, Z.M. Yanga, Least squares recursive projection twin support vector machine for classification, Pattern Recognit. 45 (6) (2012) 2299–2307.
[29] J. Shore, R. Johnson, Axiomatic derivation of the principle of maximum entropy & the principle of minimum cross-entropy, IEEE Trans. Inf. Theory 26 (1) (1980) 26–37.
[30] Y. Tang, F. Qian, H. Gao, J. Kurths, Synchronization in complex networks and its application — a survey of recent advances and challenges, Annu. Rev. Control 38 (2014) 184–198.
[31] M. Tanner, Tools for Statistical Inference, Springer Verlag, New York, 1993.
[32] M.E. Taylor, P. Stone, Transfer learning for reinforcement learning domains: a survey, J. Mach. Learn. Res. 10 (2009) 1633–1685.
[33] E. Todorov, Linearly-solvable Markov decision problems, in: B. Schölkopf, et al. (Eds.), Advances in Neural Inf. Processing, MIT Press, NY, 2006, pp. 1369–1376.
[34] L. Torrey, J. Shavlik, Transfer learning, in: E.S. Olivas, J.D.M. Guerrero, M.M. Sober, J.R.M. Benedito, A.J.S. López (Eds.), Handbook of Research on Machine Learning, Applications and Trends: Algorithms, Methods and Techniques, Hersey, New York, 2009, pp. 242–261.
[35] S. Visweswaran, Learning instance-specific predictive models, J. Mach. Learn. Res. 11 (2010) 3333–3369.