Pattern Recognition Letters
journal homepage: www.elsevier.com

On Assigning Probabilities to New Hypotheses

Miroslav Kárný**

[a]The Czech Academy of Sciences, Institute of Information Theory and Automation, Pod vodáreno věží 4, 182 08 Prague 8, Czech Republic

Article history:
Received July 12, 2021
Received in final form @@@@
Accepted @@@@
Available online @@@@

Communicated by @@@@

2000 MSC: 62C10, 62B99, 62H15, 62F03

Keywords: minimum relative-entropy principle, prior probability, hypothesis,

ABSTRACT

The paper proposes the way how to assign a proper prior probability to a new, generally compound, hypothesis. To this purpose, it uses the minimum relative-entropy principle and a forecaster-based knowledge transfer. Methodologically, it opens a way towards enriching the standard Bayesian framework by the possibility to extend the set of models during learning without the need to restart. The presented use scenarios concern: (a) creating new hypotheses, (b) learning problems with an insufficient amount of data, and (c) sequential Monte Carlo estimation. They indicate a strong application potential of the proposed technique. Related interesting open research problems are listed.

1. Introduction

The technical problem addressed in this brief paper is a building block of a broad, quite ambitious, research. This "environment" is outlined before focusing on the solved problem.

Motivation

The mentioned research tries to create a normative theory of dynamic decision making applicable by imperfect decision makers, Guy et al. (2012, 2013, 2015); Kárný (2020). Its aims are close to the quest for universal artificial intelligence, Hutter (2005). Both touch of the complex topic of scientific discovery, Langley et al. (1987). They fight with the world complexity reflected in no free lunch theorem, Wolpert and Macready (1997). The inherent complexity enforces local solutions, known as adaptive systems, Kárný (1998). They have to rationally drift within an infinitely complex world. This call for balancing exploitation with exploration, Črepinšek et al. (2013), transfer learning, Perrone et al. (2018), care about forgetting, Kulhavý and Zarrop (1993), etc. All this needs potentially unbounded sets and systematic ways how to move within them. The inspection concern not only parameters, He and Shao (2000), but also the sets of patterns, Höppner (2001), and models, Sec. 4, formally, the set of hypotheses. While human beings are well able to work in this way, artificial world, Simon (1996), is not matured in this respect. This paper tries to help.

Addressed Technical Problem

Bayes' rule deductively modifies probabilities of formulated hypotheses by data, Berger (1985). It does not guide what probability is to be assigned to a new hypothesis that has arisen during learning progress. A recent deep discussion of this problem with extensive references is in the paper of Wenmackers and Romeijn (2016). They, however, provide no constructive solution. It is offered here. The solution is based on exploitation of the minimum relative-entropy principle[1], Shore and Johnson (1980), and on a forecaster-based transfer of the knowledge collected under old hypotheses. It refines the knowledge transfer proposed by Kracík and Kárný (2005).

On Related Research

The paper provides a building block fitting to advanced Bayesian decision making paradigm having its roots in the seminal unification of Savage (1954). It naturally exploits achievements of Bayesianism and refers to them on the fly. In that sense the paper belongs to one of major research streams dealing with an information processing and its use. However, to our best knowledge nobody offered a constructive solution of the

**Corresponding author: Tel.: +420-2-6605-2274
e-mail: school@utia.cas.cz (Miroslav Kárný)

[1]These authors call it minimum cross-entropy principle. Now, the terms relative entropy or KL divergence, Kullback and Leibler (1951), are used.

addressed problem. Sec. 4.1 even recalls the classical refusal of the possibility to solve this problem. This attitude is generally (mostly implicitly) adopted. In that sense, the regular comparison with a standard methodology cannot be offered. The use scenarios, Sec. 4, however, show explicitly, and Sec. 4.2 even numerically, that the paper offers solutions, which go well beyond quality reachable by the state-of-the-art techniques.

Layout and Pattern Recognition Relevance

The paper starts with simple hypotheses and then it deals with compound ones. The forecaster-based knowledge transfer then serves us for constructing informative prior of a new compound hypothesis. The use scenarios follow. They, together with the list of open problems, indicate why the problem and its solution are relevant when data-based knowledge accumulation copes with a varying set of hypotheses. A varying number of classes given, say, by the varying number of mixture components, Roth et al. (2018), or the considered dimension of data space (memory length), Arunkumar et al. (2017), provide common application examples of this type.

2. Priors for New Hypothesis

The case of simple hypotheses is firstly dealt with, Subsection 2.1. It guides the reader through the addressed problem. Subsection 2.2 treats the general case of compound hypotheses. It reveals the need for a knowledge transfer presented in Sec. 3.

Throughout, $\equiv$ defines by assignment, boldface **fonts** mark sets and sanserif fonts denote mappings.

2.1. Simple Hypotheses

Let us consider hypotheses $h \in \underline{\mathbf{h}} \equiv \{1,\dots,\underline{\chi}\}$, $\underline{\chi} < \infty$, with given probabilities $\underline{\mathsf{p}} \equiv \{\underline{\mathsf{p}}(h)\}_{h\in\underline{\mathbf{h}}}$. A new hypothesis $\chi \equiv \underline{\chi}+1$ extends the set $\underline{\mathbf{h}}$. A further learning needs probabilities $\mathsf{p} \equiv \{\mathsf{p}(h)\}_{h\in\mathbf{h}}$, $\mathbf{h} \equiv \underline{\mathbf{h}} \cup \{\chi\}$. When selecting them, we require

$$\mathsf{p}(h) = \kappa\underline{\mathsf{p}}(h) \text{ with } \kappa \in (0,1) \text{ and } h \in \underline{\mathbf{h}} \subset \mathbf{h} = \underline{\mathbf{h}} \cup \{\chi\}. \quad (1)$$

This preserves the knowledge collected before handling the new hypothesis $h = \chi$. Other options modify Bayesian factors and violate the likelihood principle, Berger and Wolpert (1988).

The minimum relative-entropy principle, Shore and Johnson (1980), implies that the adequate probabilities p minimise their relative entropy $\mathsf{D}(\mathsf{p}\|\mathsf{g})$ to their prior guess $\mathsf{g} \equiv \{\mathsf{g}(h)\}_{h\in\mathbf{h}}$

$$\mathsf{p} \in \underset{\{\mathsf{p}\, meeting\, (1)\}}{\mathrm{Arg\,min}}\, \mathsf{D}(\mathsf{p}\|\mathsf{g}) \equiv \underset{\{\mathsf{p}\, meeting\, (1)\}}{\mathrm{Arg\,min}} \sum_{h\in\mathbf{h}} \mathsf{p}(h) \ln\left[\frac{\mathsf{p}(h)}{\mathsf{g}(h)}\right]. \quad (2)$$

Proposition 1 (A New Simple Hypothesis). *The solution* p *of (2) reads*

$$\begin{aligned} \mathsf{p}(h) &= \kappa\underline{\mathsf{p}}(h), \ h \in \underline{\mathbf{h}}, & \mathsf{p}(\chi) &= 1-\kappa \quad (3)\\ \kappa &= \frac{1}{1+\frac{\mathsf{g}(\chi)}{1-\mathsf{g}(\chi)}\exp\left[\mathsf{D}\left(\underline{\mathsf{p}}\middle\|\underline{\mathsf{g}}\right)\right]}, & \underline{\mathsf{g}} &\equiv \left(\frac{\mathsf{g}(h)}{1-\mathsf{g}(\chi)}\right)_{h\in\underline{\mathbf{h}}}. \end{aligned}$$

Proof The constraint (1) leaves the single free optimised parameter $\kappa \in (0,1)$. With $\underline{\mathsf{g}}(h) = \frac{\mathsf{g}(h)}{1-\mathsf{g}(\chi)}$ on $\underline{\mathbf{h}}$ (the restriction of g on $\underline{\mathbf{h}}$) the relative entropy depends on κ as follows

$$\begin{aligned} \mathsf{D}(\mathsf{p}\|\mathsf{g}) &= \kappa\ln(\kappa) + (1-\kappa)\ln(1-\kappa)\\ &+ \kappa\ln\left[\frac{\mathsf{g}(\chi)}{1-\mathsf{g}(\chi)}\right] + \kappa\mathsf{D}(\underline{\mathsf{p}}\|\underline{\mathsf{g}}) - \ln(\mathsf{g}(\chi)). \end{aligned}$$

Zeroing its derivative with respect to the optional κ and a simple algebra gives the claimed result (3). □

Commentary 1 (On Qualitative Properties of (3)).

- *If* $\underline{\mathsf{p}}$ *diverges from the prior guess* $\underline{\mathsf{g}}$ (g *restricted on* $\underline{\mathbf{h}}$*) then* $\mathsf{D}\left(\underline{\mathsf{p}}\middle\|\underline{\mathsf{g}}\right)$ *is large and makes the probability* $\mathsf{p}(\chi)$ *relatively large. Thus, the new hypothesis is a priori competitive with well-stratified older hypotheses.*
- *The gained probabilities* p *depend on their prior guess* g*. This manifests the common danger of infinite regress, faced, for instance, in games, Insua et al. (2016). It is resolved here operationally:*
  - ⋆ *knowledge elicitation, Garthwaite et al. (2005), and past data, Peterka (1981), are used if possible;*
  - ⋆ *the minimum relative-entropy principle, Shore and Johnson (1980), is then applied (done here);*
  - ⋆ *the prior probability,* g*, on the discrete-valued random variable,* $h \in \mathbf{h}$*, is chosen as uniform; this choice reflects the principle of insufficient reasons, attributed to P.S. Laplace. It leads to the special case of the minimum relative-entropy principle known as maximum entropy principle, Jaynes (1957).*

2.2. Compound Hypotheses

The hypotheses $h \in \underline{\mathbf{h}} = \{1,\dots,\underline{\chi}\}$, $\underline{\chi} < \infty$ are generally compound. Each includes an unknown parameter $\Theta_h \in \mathbf{\Theta}_h$. The set $\mathbf{\Theta}_h$ may contain an infinite amount of members. The existing hypotheses $h \in \underline{\mathbf{h}}$ are described by $\underline{\mathsf{P}}(\Theta_h)\mathsf{p}(h)$, where the first factor is probability density (pd). The pd $\underline{\mathsf{P}}(\Theta_h)$ depends on $h \in \mathbf{h}$, i.e. $\underline{\mathsf{P}}(\Theta_h) \equiv \underline{\mathsf{P}}(\Theta_h|h)$.

The further use of the term probability density also for probabilities $\mathsf{p}(h)$ takes them as Radon-Nikodým derivatives with respect to the counting measure (also marked $\mathrm{d}h$), Rao (1987).

A new hypothesis $\chi = \underline{\chi}+1$ with its parameter $\Theta_\chi \in \mathbf{\Theta}_\chi$ extends $\underline{\mathbf{h}}$ to $\mathbf{h} = \underline{\mathbf{h}} \cup \{\chi\}$. A further learning needs pd $\mathsf{P}(\Theta_h)\mathsf{p}(h)$ on the extended set of hypotheses, $\mathbf{h}$, and all parameters

$$\Theta \equiv (\Theta_1,\dots,\Theta_\chi) \in \mathbf{\Theta} \equiv (\mathbf{\Theta}_1,\dots,\mathbf{\Theta}_\chi) \quad (4)$$

within them. The analogy of (1) has to hold

$$\mathsf{p}(h) = \kappa\underline{\mathsf{p}}(h), \ \kappa \in (0,1), \text{ and } \mathsf{P}(\Theta_h) = \underline{\mathsf{P}}(\Theta_h) \text{ for } h \in \underline{\mathbf{h}}. \quad (5)$$

It preserves the learnt relations of hypotheses $h \in \underline{\mathbf{h}}$ and the knowledge about the parameter within each hypothesis. Due to (5), the prior guesses

$$\mathsf{g}(h) \text{ of } \mathsf{p}(h) \text{ on } \mathbf{h} \text{ and } \mathsf{G}(\Theta_\chi) \text{ of } \mathsf{P}(\Theta_\chi) \text{ on } \mathbf{\Theta}_\chi \quad (6)$$

suffice for the use of the minimum relative-entropy principle.

The fact that the absolute minimum of the relative-entropy is reached for equal arguments gives immediately the result *almost* identical with that of Proposition 1.

Proposition 2 (A New Compound Hypothesis). *For the constraint (5) and the given pds (6), the minimum cross-entropy principle provides the same* p *as Proposition 1. The optimal choice of* p *is given by (3) and the remaining free pd* $\mathsf{P}(\Theta_\chi)$ *is*

$$\mathsf{P}(\Theta_\chi) = \mathsf{G}(\Theta_\chi), \quad \Theta_\chi \in \boldsymbol{\Theta}_\chi. \tag{7}$$

The result depends on the prior guess g acting on the extended set of hypotheses $\mathbf{h}$ and on the pd $\mathsf{G}(\Theta_\chi)$. Laplace's insufficient reasons prefer uniform g. The choice of the pd $\mathsf{G}(\Theta_\chi)$ (7) is more peculiar. It is well seen when the volume of $\boldsymbol{\Theta}_\chi$ is infinite and the uniform pd does not exist. It is always possible to select a flat proper pd. However, to make the new hypothesis competitive such a non-informative prior, cf. Berger and Pericchi (1996), is to be corrected by transferring the knowledge accumulated in connection with older hypotheses. This is done in Sec. 3. Note that the need to avoid non-informative prior applies also to transfer learning understood in a more narrow sense, Perrone et al. (2018).

3. Knowledge Transfer Between Hypotheses

This section uses data forecasters, gained under older hypotheses, to transfer the knowledge into the proper, possibly almost non-informative, prior pd $\mathsf{G}(\Theta_\chi)$ of the parameter $\Theta_\chi \in \boldsymbol{\Theta}_\chi$ within the new hypothesis χ.

The adopted type of the knowledge transfer was heuristically proposed by Kracík and Kárný (2005). It was refined in several papers up to the advanced deductive version, Quinn et al. (2017). The presentation below slightly extends these results but it primarily meets our key wish: to gain as informative description of the parameter Θ_χ as possible.

The transfer exploits that any testable hypothesis $h \in \mathbf{h}$ provides a parametric model $\mathsf{m}(d|\Theta_h)$, the parameterised pd of the potentially observable data $d \in \mathbf{d}$. The modelled data $d \in \mathbf{d}$ is assumed to be the *only common part* of models $\mathsf{m}(d|\Theta_h)$, $h \in \underline{\mathbf{h}}$, with the new model, $\mathsf{m}(d|\Theta_\chi)$. This is a generic case.

Data complements the treated random variables to $(d, \Theta, h) \in (\mathbf{d}, \boldsymbol{\Theta}, \mathbf{h})$, see (4). Their joint prior pd is

$$\begin{aligned} \underline{\mathsf{J}}(d, \Theta, h) &= \underline{\mathsf{J}}(d|\Theta, h)\underline{\mathsf{J}}(\Theta|h)\underline{\mathsf{J}}(h) \\ &= \mathsf{m}(d|\Theta_h)\mathsf{P}(\Theta_h) \prod_{\tilde{h} \in \mathbf{h}\setminus\{h\}} \mathsf{J}(\Theta_{\tilde{h}})\mathsf{p}(h) \\ &= \mathsf{m}(d|\Theta_h) \prod_{\tilde{h} \in \underline{\mathbf{h}}} \mathsf{P}(\Theta_{\tilde{h}})\mathsf{G}(\Theta_\chi)\mathsf{p}(h). \end{aligned} \tag{8}$$

There, the chain rule for pds, Peterka (1981), provides the 1^{st} equality in (8). The unambiguous 2^{nd} row reflects that:

(a) data d depends on Θ_h only via the model $\mathsf{J}(d|\Theta, h) = \mathsf{m}(d|\Theta_h)$, for each hypothesis $h \in \mathbf{h}$;

(b) the hypothesis h only considers the parameter Θ_h;

(c) the respective parameters Θ_h are unrelated and thus modelled as independent.

The final formula in 3^{rd} row of (8) just re-arranges 2^{nd} row and uses the already adopted names of respective pds on parameters.

Formula (8) shows that the "joint" parameter $\Theta \in \boldsymbol{\Theta}$ (4) is in fact treated as independent of the specific hypothesis,

$$\underline{\mathsf{J}}(\Theta|h) = \underline{\mathsf{J}}(\Theta) = \prod_{\tilde{h} \in \underline{\mathbf{h}}} \mathsf{P}(\Theta_{\tilde{h}})\mathsf{G}(\Theta_\chi).$$

As data is the only common part of models, just forecasters

$$f(d|h) \equiv \begin{cases} \int_{\boldsymbol{\Theta}_h} \mathsf{m}(d|\Theta_h)\mathsf{P}(\Theta_h)\mathrm{d}\Theta_h & \text{for} \quad h \in \underline{\mathbf{h}} \\ \int_{\boldsymbol{\Theta}_\chi} \mathsf{m}(d|\Theta_\chi)\mathsf{G}(\Theta_\chi)\mathrm{d}\Theta_\chi & \text{for} \quad h = \chi \end{cases} \tag{9}$$

provide a ground for a knowledge transfer. They can enrich the prior pd $\mathsf{G}(\Theta_\chi)$ of the unknown parameter $\Theta_\chi \in \boldsymbol{\Theta}_\chi$ to a pd $\tilde{\mathsf{P}}(\Theta_\chi)$. The restricted modelling of data by forecasters (9) induces an alternative joint pd $\tilde{\mathsf{J}}(d, \Theta, h)$ of the discussed random variables. For any opted $\tilde{\mathsf{P}}(\Theta_\chi)$, the joint pd has the form

$$\tilde{\mathsf{J}}(d, \Theta, h) = \tilde{\mathsf{J}}(d|\Theta, h)\tilde{\mathsf{J}}(\Theta|h)\tilde{\mathsf{J}}(h) = \mathsf{f}(d|h) \prod_{\tilde{h} \in \underline{\mathbf{h}}} \mathsf{P}(\Theta_{\tilde{h}})\tilde{\mathsf{P}}(\Theta_\chi)\mathsf{p}(h),$$

where the 1^{st} equality is again the chain rule for pds. The unambiguous 2^{nd} equality reflects that:

- only forecasters of common data, independent of the unknown parameter Θ, are used, i.e. $\tilde{\mathsf{J}}(d|\Theta, h) = \mathsf{f}(d|h)$;
- the above arguments (b), (c) apply with

$$\tilde{\mathsf{J}}(\Theta|h) = \tilde{\mathsf{J}}(\Theta) = \prod_{\tilde{h} \in \underline{\mathbf{h}}} \mathsf{P}(\Theta_{\tilde{h}})\tilde{\mathsf{P}}(\Theta_\chi).$$

The minimum relative-entropy principle recommends the optimal choice $\mathsf{P}(\Theta_\chi)$ of $\tilde{\mathsf{P}}(\Theta_\chi)$ on $\boldsymbol{\Theta}_\chi$.

Proposition 3 (Correction of $\mathsf{G}(\Theta_\chi)$ by Forecasters). *Let us search for a joint pd* $\tilde{\mathsf{J}}(d, \Theta, h)$ *on* $(\mathbf{d}, \boldsymbol{\Theta}, \mathbf{h})$ *in the set* $\mathbf{J}$ *(10) below. Its members are determined by the given the fixed forecasters* $(\mathsf{f}(d|h))_{h \in \mathbf{h}}$ *(9), the pds of hypotheses* $(\mathsf{p}(h))_{h \in \mathbf{h}}$ *(3) and the pd* $\prod_{\tilde{h} \in \underline{\mathbf{h}}} \mathsf{P}(\Theta_{\tilde{h}})\tilde{\mathsf{P}}(\Theta_\chi)$ *with the optional* $\tilde{\mathsf{P}}(\Theta_\chi)$

$$\mathbf{J} \equiv \left\{ \tilde{\mathsf{J}}(d, \Theta, h) \equiv \mathsf{f}(d|h) \prod_{h \in \underline{\mathbf{h}}} \mathsf{P}(\Theta_h)\tilde{\mathsf{P}}(\Theta_\chi)\mathsf{p}(h),\ \tilde{\mathsf{P}}(\Theta_\chi)\ \textit{free} \right\}. \tag{10}$$

Let the prior guess of the optimal joint pd on $(\mathbf{d}, \boldsymbol{\Theta}, \mathbf{h})$ *be* $\underline{\mathsf{J}}(d, \Theta, h) = \mathsf{m}(d|\Theta_h) \prod_{h \in \underline{\mathbf{h}}} \mathsf{P}(\Theta_h)\mathsf{G}(\Theta_\chi)\mathsf{p}(h)$, *see (8).*

Then, the optimal factor $\mathsf{P}(\Theta_\chi) = \tilde{\mathsf{P}}(\Theta_\chi)$, *giving minimum of* $[\mathsf{D}(\tilde{\mathsf{J}}\|\underline{\mathsf{J}})]$ *over* $\mathbf{J}$ *(10), is*

$$\begin{aligned} \mathsf{P}(\Theta_\chi) &\propto \mathsf{G}(\Theta_\chi) \exp\Big[\int_{\mathbf{d}} \mathsf{f}(d) \ln[\mathsf{m}(d|\Theta_\chi)]\mathrm{d}d \Big] \\ \mathsf{f}(d) &\equiv \sum_{h \in \mathbf{h}} \mathsf{f}(d|h)\mathsf{p}(h), \ \textit{see (9).} \end{aligned} \tag{11}$$

Proof Using Fubini's theorem on multiple integration, Rao (1987), and forms of $\tilde{\mathsf{J}}$, $\underline{\mathsf{J}}$, the relative-entropy re-arranges

$$\begin{aligned} \mathsf{D}(\tilde{\mathsf{J}}\|\underline{\mathsf{J}}) &= \int_{\mathbf{d}, \boldsymbol{\Theta}_\chi, \mathbf{h}} \mathsf{f}(d|h)\tilde{\mathsf{P}}(\Theta_\chi)\mathsf{p}(h) \ln\left[\frac{\mathsf{f}(d|h)\tilde{\mathsf{P}}(\Theta_\chi)}{\mathsf{m}(d|\Theta_\chi)\mathsf{G}(\Theta_\chi)} \right] \mathrm{d}d\mathrm{d}\Theta_\chi \mathrm{d}h + \gamma \\ &= \int_{\boldsymbol{\Theta}_\chi} \tilde{\mathsf{P}}(\Theta_\chi) \left[\ln\left[\frac{\tilde{\mathsf{P}}(\Theta_\chi)}{\mathsf{G}(\Theta_\chi)} \right] - \int_{\mathbf{d}} \mathsf{f}(d) \ln[\mathsf{m}(d|\Theta_\chi)]\mathrm{d}d \right] \mathrm{d}\Theta_\chi + \tilde{\gamma} \\ &= \mathsf{D}(\tilde{\mathsf{P}}\|\mathsf{P}) + \tilde{\tilde{\gamma}}, \end{aligned}$$

where $\gamma, \tilde{\gamma}, \tilde{\tilde{\gamma}}$ are constants independent of $\tilde{\mathsf{P}}$ and the pd P is given by (11). The minimum of $\mathsf{D}(\tilde{\mathsf{J}}||\underline{\mathsf{J}})$ over the optional $\tilde{\mathsf{P}}$ is thus reached for $\tilde{\mathsf{P}} = \mathsf{P}$. □

Commentary 2 (On Proposition 3).

- *The use of the implicitly defined forecaster* $\mathsf{f}(d|\chi) = \int_{\boldsymbol{\Theta}_\chi} \mathsf{m}(d|\Theta_\chi)\mathsf{P}(\Theta_\chi)\mathrm{d}\Theta$ *instead* $\mathsf{f}(d|\chi) = \int_{\boldsymbol{\Theta}_\chi} \mathsf{m}(d|\Theta_\chi)\mathsf{G}(\Theta_\chi)\mathrm{d}\Theta$ *offers a further improvement. Also, the knowledge transfer exploits only the jointly modelled d. The same treatment is possible if the parameter spaces* $\boldsymbol{\Theta}_h$, $h \in \mathbf{h}$, *have a common part. None of them is done to keep the text simple.*
- *In our context, it is important to stress that Bayesian estimation* accumulates knowledge *by conditioning or, when impossible, via the minimum relative-entropy principle, Campenhout and Cover (1981). The gained knowledge is then used to DM, e.g. to the point or interval estimation, acceptance or rejection hypotheses etc., Wald (1950).*
- *The described results give the overall processing algorithm bellow. It accumulates the knowledge and adds a single compound hypothesis and discards some. The reading ease motivates this. An extension to several hypotheses is straightforward.*

Algorithm 1 (Processing with New Compound Hypothesis).

(a) The pds of old hypotheses $\mathsf{P}(\Theta_h)$, $\underline{\mathsf{p}}(h)$ *are updated by observed data* $d \in \mathbf{d}$ *via Bayes' rule with the models* $\mathsf{m}(d|\Theta_h)$, $h \in \underline{\mathbf{h}}$. *The forecasting pds* $\mathsf{f}(d|h)$, $h \in \underline{\mathbf{h}}$, *(9) arise as a byproduct. Hypotheses with small values of the updated probabilities can be discarded. If no new hypothesis arises this step repeats. Otherwise, the next step applies.*

(b) A new hypothesis giving $\mathsf{m}(d|\Theta_\chi)$, $\chi = \underline{\chi} + 1$, *arises with a flat proper prior pd* $\mathsf{G}(\Theta_\chi)$. *The pd* $\mathsf{f}(d|\chi)$ *(9) is evaluated.*

(c) A prior pd $\mathsf{g}(h)$ *on* $\mathbf{h} = \underline{\mathbf{h}} \cup \{\chi\}$ *is chosen, generically, uniform one. The pd* $\mathsf{p}(h)$ *on* $\mathbf{h}$ *is computed according to (3). The pd* $\mathsf{f}(d)$ *on* $\mathbf{d}$ *is now at disposal and the prior* $\mathsf{G}(\Theta_\chi)$ *is corrected to* $\mathsf{P}(\Theta_\chi)$, *both see (11).*

(d) The pds $\mathsf{P}(\Theta_h), \mathsf{p}(h)$, $h \in \mathbf{h}$, *become the prior pds on the extended set* $\mathbf{h}$ *of compound hypotheses. The processing repeats from* (*a*) *while taking them as old hypotheses.*

4. Use Scenarios

This part samples tasks to which our solution may contribute.

4.1. Problem of Something Else

A classical view on novel hypotheses, explicitly expressed by Lindley (2006) on p. 188, states: "…it makes no sense to include […] another branch in your tree, corresponding to 'do something else'. Nor, when the uncertain events are listed, does it make sense to include 'something else happens'."

Our results shift this to: […] always keep a branch in your tree, corresponding to 'do something else'. When new hypotheses arise the theory guides you how to assign them probabilities using the knowledge collected before formulating them.

4.2. Problem of Initial Data

The title, adopted from Peterka (1981), concerns common cases in which the amount of available data is insufficient for evaluating the likelihood of the desirable complex model given by $h \in \mathbf{h} \setminus \underline{\mathbf{h}}$. The model falls into "something else" class, see Subsection 4.1. The belief that this model describes reality better than the already learnt ones can be assigned after collecting enough additional data. Often, this is impossible and Algorithm 1 helps. The next case of this type is widely met. It is simple enough to be presented in this brief paper.

A real-valued series $d^t \equiv (d_\tau)_{\tau=1}^t$, $t \in \mathbf{t} \equiv \{1, 2, \ldots\}$ are modelled. The hypothesis $h \in \underline{\mathbf{h}}$ uses Gaussian auto-regression of the order h with the unknown parameter Θ_h made of auto-regression coefficients θ_h and variance r_h

$$\mathsf{m}(d_t|d^{t-1}, \Theta_h) = (2\pi r_h)^{-0.5} \exp\left[-\frac{(d_t - \theta_h'\psi_{t-1;h})^2}{2r_h}\right] \quad (12)$$

$$\Theta_h \equiv (\theta_h, r_h), \quad \psi_{t-1;h} \equiv [d_{t-1} \ldots d_{t-h}]', \ ' \text{ is transposition.}$$

Each model (12) has a conjugated (self-reproducing) prior pd, Berger (1985). It is Gauss-inverse-gamma pd, Peterka (1981). Its sufficient statistic coincides with the well-know recursive least-squares (RLS) objects. At time $t-1$, it consists of the RLS point estimate $\hat{\theta}_{t-1;h}$ of θ_h, the RLS parameter covariance $C_{t-1;h} > 0$ ($C_{t-1;h}$ is positive definite), the RLS remainder $\lambda_{t-1;h} > 0$ giving, together with the degrees of freedom $\nu_{t-1;h} > 2$, the point estimate $\frac{\lambda_{t-1;h}}{\nu_{t-1;h}-2}$ of r_h. The self-reproducing posterior pd of θ_h, r_h conditioned on d^{t-1} is proportional to

$$r_h^{-\frac{\nu_{t-1;h}+h+2}{2}} \exp\left[-\frac{(\theta_h - \hat{\theta}_{t-1;h})' C_{t-1;h}^{-1} (\theta_h - \hat{\theta}_{t-1;h}) + \lambda_{t-1;h}}{2r_h}\right].$$

The Bayesian updating, Algorithm 1.(*a*), applicable for all orders h smaller than the number of observations t, $h < t$, reads

$$\begin{aligned}
\hat{\theta}_{t;h} &\equiv \hat{\theta}_{t-1;h} + \frac{C_{t-1;h}\psi_{t-1;h}}{\omega_{t-1;h}}\hat{e}_{t;h}, \ \ \hat{e}_{t;h} \equiv d_t - \hat{\theta}_{t-1;h}'\psi_{t-1;h} \\
\lambda_{t;h} &\equiv \lambda_{t-1;h} + \frac{\hat{e}_{t;h}^2}{\omega_{t-1;h}}, \ \ \omega_{t-1;h} \equiv 1 + \psi_{t-1;h}' C_{t-1;h}\psi_{t-1;h}, \\
C_{t;h}^{-1} &\equiv C_{t-1;h}^{-1} + \psi_{t-1;h}\psi_{t-1;h}', \ \ \nu_{t;h} \equiv \nu_{t-1;h} + 1 \qquad (13) \\
\mathsf{f}(d_t|h) &\equiv \mathsf{f}(d_t|h, d^{t-1}) \propto \sqrt{\frac{\nu_{t-1;h}}{\lambda_{t-1;h}\omega_{t-1;h}}}\left(1 + \frac{\hat{e}_{t;h}^2}{\lambda_{t-1;h}\omega_{t-1;h}}\right)^{\frac{-\nu_{t-1;h}}{2}} \\
\underline{\mathsf{p}}(h) &\equiv \underline{\mathsf{p}}(h|d^t) \propto \mathsf{f}(d_t|h, d^{t-1})\underline{\mathsf{p}}(h|d^{t-1}).
\end{aligned}$$

The last two definitions relate the conditional forecasters and the conditional order probabilities to (9) and (5).

The prior values initiating this recursion are $\hat{\theta}_{0;h}$, $\lambda_{0;h} > 0$, $C_{0;h} > 0$, $\nu_{0;h} > 0$ and the order priors $\underline{\mathsf{p}}(h|d^0)$, $h \in \mathbf{h}$.

The *initial-data problem* arises if $t \leq h$ and the regression vector $\psi_{t-1;h}$ (12) is unavailable. This is critical if the usage does not allow to wait. Adaptive forecasters, classifiers and controllers are typical examples of this type.

The classical Bayesian solution requires the specification and use of the prior pd also over the unknown regression vector. Even if its reliable version is available, the estimation becomes a hard nonlinear task as the product of the unknown regression coefficients with the unknown regression vector is to be learnt.

Thus, it makes sense to start with low-order models and to extend them with time, cf. He and Shao (2000). Just before time $t+1$, the models of the orders $h \le \underline{\chi} = t$ are updated by RLS. The regression vectors $\psi_{t;h}$ are available and thus also pds $\mathsf{f}(d_{t+1}|h)$, $h \in \underline{\mathbf{h}}$, cf. Algorithm 1.($a$) and (13).

The model of the order $\chi = \underline{\chi}+1$ arises, Algorithm 1.(b). The prior pd $\mathsf{G}(\Theta_\chi)$ is gained by the extension $\lambda_{t;\chi} = \lambda_{t;\underline{\chi}}$, $\nu_{t;\chi} = \nu_{t;\underline{\chi}}$,

$$C^{-1}_{t;\chi} = \begin{bmatrix} C^{-1}_{t;\underline{\chi}} & 0 \\ 0 & \beta^{-1}_{t;\chi} \end{bmatrix}, \quad \hat{\theta}_{t;\chi} = \begin{bmatrix} \hat{\theta}_{t;\underline{\chi}} \\ 0 \end{bmatrix}$$

with $\beta_{t;\chi} > 0$ chosen so that the auto-regression remains stable with a high probability. This specifies the magnitude of $\beta_{t;\chi} > 0$.

The pd $\mathsf{f}(d_{t+1})$ (11) based on orders $h \le t = \chi$ has the expectation $\hat{d} = \sum_{h\in\mathbf{h}} \hat{\theta}'_{h;t}\psi_{t;h}\mathsf{p}(h|d^t)$ and the variance $\hat{r} = \sum_{h\in\mathbf{h}} \mathsf{p}(h|d^t)\left[\frac{\lambda_{t;h}}{\nu_{t;h}-2}\omega_{t;h} + (\hat{\theta}'_{h;t}\psi_{t;h})^2\right] - \hat{d}^2$. There, the pd $\mathsf{p}(h|d^t)$, $h \in \mathbf{h}$, evaluates according to Proposition 1 with the uniform g.

The correction of $\mathsf{G}(\Theta_\chi)$ to $\mathsf{P}(\Theta_\chi)$, Algorithm 1.(c), reduces to RLS processing $\hat{d}_{t+1}$ and ψ_t. The use of the regressand $\hat{d}_{t+1}$ instead of the unavailable d_{t+1} is paid by the increased influence of the forecast error on the RLS remainder. It must be set equal to $\lambda_{t;h=t} = \hat{r}(\nu_{t;h} - 2)$. This completes Algorithm 1.(c).

Numerical Experiment: Figure 1 illustrates the contribution of the proposed processing. It presents averaged results of 500 runs of the model-order estimation. The simulated system is the auto-regression of the 6th order with the multiple root equal to 0.8. The upper bound on the compared orders was 15 and 150 data items were processed in each Monte Carlo run. Thus, the usual waiting for enough data cannot use 10% of data. The results are presented as sample means of the cumulative order (hypothesis) probability and of individual order probabilities. They confirm that this omission is paid by a much higher uncertainty about the order. The situation is naturally much more critical for shorter runs (say 50 data items).

The example confirmed the observation that the well initiated Bayesian learning provides a good order estimate much faster than a parameter estimate. In our case, the point parameter estimate approached the true one after 20 000 samples.

Commentary 3 (On the Solution of Initial-Data Problem).

- *The processing generates the nested sufficient statistics so advantageous in efficient implementations including signal-processing hardware, Pohl et al. (2008).*
- *The controlled case is even more sensitive to the prior pd as an insufficiently exciting feedback, Ljung (1987), may create an almost unidentifiable situation. Good priors decrease demands on the inevitable exploration effort.*
- *The solution of the initial-data problem is urgent if the number of potential features, to be used as regressors, exceeds the number of available data records. This is typical in genomic, where an insufficient care about priors (often) makes estimates of the data-model structure quite unreliable, e.g. see Hlaváčková-Schindler et al. (2016).*

4.3. Sequential Monte Carlo Estimation

Monte Carlo (MC) methods exploiting the cheap computational power become standard tool in non-linear filtering,

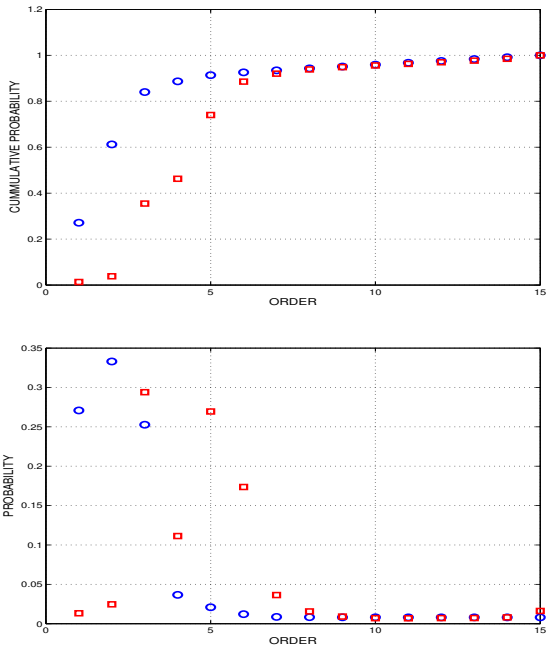

Fig. 1. Top: the mean of cumulative order probability for the usual RLS (blue circle) and our processing (red box). Bottom: the mean of order probability for the usual RLS (blue circle) and our processing (red box).

Doucet and Johansen (2011), and particularly in parameter estimation. They exploit that Bayes' rule on a fixed grid of unknowns is trivial. It well discards improbable values. The latter fact, however, makes the parameter-estimation problem susceptible to degeneracy: a few grid points get non-negligible probability. Then, new candidates (hypotheses) must be chosen and qualified by prior probabilities. The lack of a (relatively) universal way motivates various countermeasures, see e.g. Green and Maskell (2017); Song et al. (2019). The presented theory may help. In our terms, the sequential MC runs as follows.

- A finite number of hypotheses (samples) $h \in \underline{\mathbf{h}}$ and their pds $\mathsf{P}_h(\Theta)$, $\underline{\mathsf{p}}(h)$, $\Theta_h \in \boldsymbol{\Theta}$, $h \in \underline{\mathbf{h}}$, are given. They concern data models $\overline{\mathsf{m}}(d|\Theta)$ and provide forecasters $\mathsf{f}(d|h)$. The use of Dirac's delta $\delta(\Theta)$ embeds Θ-samples into our scheme.
- An easily sample-able prior (proposal) pd $\mathsf{G}(\Theta)$ is defined on $\boldsymbol{\Theta}$ and modified to $\mathsf{P}_\chi(\Theta)$, $\chi = \underline{\chi}+1$, according to Prop. 3 and possibly reduced by sampling it $\Theta_\chi \sim \mathsf{P}_\chi(\Theta) \to \mathsf{P}_\chi(\Theta) = \delta(\Theta - \Theta_\chi)$.
- Prop. 1 guides us how to modify the pds $(\underline{\mathsf{p}}(h))_{h\in\underline{\mathbf{h}}}$ and (uniform) $(\mathsf{g}(h))_{h\in\mathbf{h}}$ to $(\mathsf{p}(h))_{h\in\mathbf{h}}$ with $\mathbf{h} = \underline{\mathbf{h}} \cup \overline{\{\chi\}}$. As said, the extension to more samples is straightforward.
- Hypotheses with small $\mathsf{p}(h)$ are discarded. The new pd $\underline{\mathsf{p}}(h)$ on $\underline{\mathbf{h}}$ arises.
- New data and Bayes' rule are used to update pds $\underline{\mathsf{p}}(h)$, possibly together with non-Dirac $\mathsf{P}_h(\Theta)$ (Rao-Blackwell's version, Doucet and Johansen (2011)), and all repeats.

This outline indicates that the proposed way avoids the degeneracy, the bottleneck of the sequential MC estimation. The way surely helps by handling well pds p on $\underline{\mathbf{h}}$ and by making the choice of $\mathsf{G}(\Theta)$ less critical due to the knowledge brought by forecasters based on old samples.

5. Concluding Remarks

The formulation and solution of the addressed problem are simple. Their significance is primarily methodological as it enriches Bayesian paradigm by the technique for extending the set of learnt models. This ability is one of the key features, which distinguishes human and artificial worlds. The outlined use scenarios indicate immediate practical consequences to the order and feature selections as well as to non-linear estimation based on sequential Monte Carlo technique. Generally, it contributes to knowledge transfer and also to transfer learning, Perrone et al. (2018), when the methodology is applied to extending data set (as it is typical in processing of data streams).

Naturally, it calls for a range of research activities:

(a) the inspection of an alternative to (1) by formulating the preservation of the accumulated knowledge as the closeness of the involved pds in relative-entropy terms;

(b) addressing the generally non-trivial evaluation of $\int_{\mathbf{d}} \mathsf{f}(d) \ln[\mathsf{m}(d|\Theta_\chi)]\mathrm{d}d$, see Prop. 3;

(c) the projection of the constructed posterior pd to a more feasible class, Kárný (2014);

(d) selecting and re-solving other non-trivial and useful scenarios like Bayesian optimisation, Shahriari et al. (2016);

(e) elaborating options mentioned in Commentary 2;

(f) performing extensive simulation and real-life tests.

We shall take this paper as useful one if it will stimulate readers' interest in any of the open problems.

In the era dominated by "big data", we would especially like to turn the research attention to "small data" problems like in Ku and Fine (2006). In fact, always desirable refinements of the set of hypotheses, Harlé et al. (2016), and the possibility to widen the feature set can make almost any data set small.

Acknowledgement

This research is supported by MŠTM LTC18075 and EU-COST Action CA16228.

References

Arunkumar, A., Ramkumar, R., Venkatraman, V., Abdulhay, E., Lawrence, F., Kadr, S., Segal, S., 2017. Classification of focal and non focal EEG using entropies. Pattern Recognition Letters 94, 112–117.

Berger, J., 1985. Statistical Decision Theory and Bayesian Analysis. Springer.

Berger, J., Pericchi, L., 1996. The intrinsic Bayes factor for model selection and prediction. Journal of the American Statistical Association 91, 109–122.

Berger, J., Wolpert, R., 1988. The Likelihood Principle: A review, generalizations, and statistical implications. Institute of Mathematical Statistics.

Campenhout, J.V., Cover, T., 1981. Maximum entropy and conditional probability. IEEE Tran. on Inf. Theory 27, 483–489.

Črepinšek, M., Liu, S., Mernik, M., 2013. Exploration and exploitation in evolutionary algorithms: A survey. ACM Computing Survey 45, 37–44.

Doucet, A., Johansen, A., 2011. A tutorial on particle filtering and smoothing: 15 years later, in: Handbook of Nonlinear Filtering. Oxford Univ. Press, UK.

Garthwaite, P., Kadane, J., O´Hagan, A., 2005. Statistical methods for eliciting probability distributions. J. of Amer. Stat. Association 100, 680–700.

Green, P., Maskell, S., 2017. Estimating the parameters of dynamical systems from big data using sequential Monte Carlo sampler. Mechanical Systems and Signal Processing 93, 379–396.

Guy, T., Kárný, M., Wolpert, D., 2012. Decision Making with Imperfect Decision Makers. volume 28. Springer, Berlin.

Guy, T., Kárný, M., Wolpert, D., 2013. Decision Making and Imperfection. volume 474. Springer, Berlin.

Guy, T., Kárný, M., Wolpert, D., 2015. Decision Making: Uncertainty, Imperfection, Deliberation and Scalability. volume 538. Springer, Switzerland.

Harlé, F., Chatelain, F., Gouy-Pailler, C., Achard, S., 2016. Bayesian model for multiple change-points detection in multivariate time series. IEEE Tran. on Signal Processing 64, 4351–4362.

He, X., Shao, Q., 2000. On parameters of increasing dimensions. Journal of Multivariate Analysis 73, 120–135.

Hlaváčková-Schindler, K., Naumova, V., Pereverzyev, S., 2016. Granger causality for ill-posed problems: Ideas, methods, and application in life sciences, in: Statistics and Causality: Methods for Applied Empirical Research. John Wiley & Sons, pp. 249–276.

Höppner, F., 2001. Discovery of temporal patterns, in: Raedt, L.D., Siebes, A. (Eds.), Principles of Data Mining and Knowledge Discovery, Springer Berlin Heidelberg. pp. 192–203.

Hutter, M., 2005. Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability. Springer, Berlin, Heidelberg, N.Y.

Insua, D.R., Banks, D., Rios, J., 2016. Modeling opponents in adversarial risk analysis. Risk Analysis 36, 742–755.

Jaynes, E., 1957. Information theory and statistical mechanics. Physical Review Series II 106, 620–630.

Kárný, M., 1998. Adaptive systems: Local approximators?, in: Workshop on Adaptive Systems in Control and Signal Processing, IFAC. pp. 129–134.

Kárný, M., 2014. Approximate Bayesian recursive estimation. Inf. Sci. 285, 100–111.

Kárný, M., 2020. Fully probabilistic design unifies and supports dynamic decision making under uncertainty. Inf. Sci. , 104–118.

Kracík, J., Kárný, M., 2005. Merging of data knowledge in Bayesian estimation, in: Filipe, J., et al (Eds.), Proc. of the 2nd Int. Conf. on Informatics in Control, Automation and Robotics, Barcelona. pp. 229–232.

Ku, C., Fine, T., 2006. A Bayesian independence test for small datasets. IEEE Tran. on Signal Processing 54, 4026–4031.

Kulhavý, R., Zarrop, M.B., 1993. On a general concept of forgetting. Int. J. of Control 58, 905–924.

Kullback, S., Leibler, R., 1951. On information and sufficiency. Ann Math Stat 22, 79–87.

Langley, P., Simon, H., Bradshaw, G., Zytkov, J., 1987. Scientific Discovery. The MIT Press, Cambridge, Massachusetts.

Lindley, D., 2006. Understanding Uncertainty. Wiley Interscience.

Ljung, L., 1987. System Identification: Theory for the User. Prentice-Hall, London.

Perrone, V., Jenatton, R., Seeger, M., Archambeau, C., 2018. Scalable hyperparameter transfer learning, in: Bengio, S., et al (Eds.), Advances in Neural Information Processing Systems 31. NIPS Foundation, pp. 6846–6856.

Peterka, V., 1981. Bayesian system identification, in: Eykhoff, P. (Ed.), Trends and Progress in System Identification. Perg. Press, pp. 239–304.

Pohl, Z., Tichý, M., Kadlec, J., 2008. Implementation of the least-squares lattice with order and forgetting factor estimation for FPGA. EURASIP Journal on Advances in Signal Processing 2008, 1–11.

Quinn, A., Kárný, M., Guy, T., 2017. Optimal design of priors constrained by external predictors. Int. J. Approximate Reasoning 84, 150–158.

Rao, M., 1987. Measure Theory and Integration. J. Wiley.

Roth, W., Peharz, R., Tschiatschek, S., Pernkopf, F., 2018. Hybrid generative-discriminative training of Gaussian mixture models. Pattern Recognition Letters 112, 131–137.

Savage, L., 1954. Foundations of Statistics. Wiley.

Shahriari, B., Swersky, K., Wang, Z., Adams, R., de Freitas, N., 2016. Taking the human out of the loop: A review of Bayesian optimization. Proc. of the IEEE 104.

Shore, J., Johnson, R., 1980. Axiomatic derivation of the principle of maximum entropy & the principle of minimum cross-entropy. IEEE Tran. on Inf. Th. 26, 26–37.

Simon, H.A., 1996. The Science of Artificial. MIT Press, Massachusets.

Song, D., Tharmarasa, R., Florea, M., Duclos-Hindie, N., Fernando, X., Kirubarajan, T., 2019. Multi-vehicle tracking with microscopic traffic flow model-based particle filtering. Automatica 105, 28 – 35.

Wald, A., 1950. Statistical Decision Functions. J. Wiley, N.Y., London.

Wenmackers, S., Romeijn, J., 2016. New theory about old evidence: A framework for open-minded Bayesianism. Synthese 193, 1225–1250.

Wolpert, D., Macready, W., 1997. No free lunch theorems for optimization. IEEE Trans. on Evolutionary Computation 1, 67–82.