Optimized Geometric Pooling of Probabilities for Information Fusion and Forgetting [1]

Miroslav Kárný

The Czech Academy of Sciences, Institute of Information Theory and Automation, POB 18, 182 08 Prague 8, Czech Republic

Abstract

Geometric pooling of probability densities (pd) is an old but basic technique of the fusion of probabilistic knowledge. Among its many justification, the use of the axiomatic minimum relative entropy principle (MREP) is the simplest one. Up to now, however, the common choice of the pooling weights is unavailable. It is done by a range of techniques. Mostly, they are of a heuristic nature and often interpret the weights as a relative trust. This paper shows that the full rigorous use of MREP enables quantitative choice of the weights, too. It quantifies the trust while using just the properly interpreted knowledge, which is deductively processed. The geometric pooling serves well adaptive estimation with forgetting that suits for illustration of our result. The paper presents an adaptive Bayesian estimator with the restricted stabilized forgetting.

Key words: minimum relative entropy principle; forgetting; probability; Bayes' rule;

1 INTRODUCTION

The geometric pooling of a finite collection of probability densities (pd) is one of basic techniques for combining knowledge quantified by the pooled pds. It is permanently refined, justified and used [2,4,11,18,32]. The rigorous choice of the pooling weights is still open. Practically, the problem has been solved by a range of (often heuristic) techniques, which provide a prior guess of the weights. For an insightful discussion see [8].

The *axiomatic minimum relative entropy principle* (MREP) [17,31] adopted here provides the well-grounded methodology how to extend a fragmental knowledge (represented by pooled pds) into a pd representing this knowledge. The paper [15] adopted MREP to the pooling problem. Among others, it confirmed the widely used rule of thumb [10] that geometric pooling should be used if the pooled pds have a common support. Neither that paper provides an algorithmic choice of the pooling weights. The solution gained here via the full exploitation of MREP is the main contribution of this communication. It may be immediately used for the choice of the forgetting factor used for parameter tracking [21,24] and in adaptive systems, e.g. [25].

[1] This paper was not presented at any IFAC meeting. The author corresponds at school@utia.cas.cz.

1.1 Layout

This section further introduces the notation and recalls the used principle. Core Sec. 2 formalizes the addressed problem and solves it. Sec. 3 illustratively equips the Bayesian estimation with a forgetting that makes the Bayesian estimator adaptive. Sec. 4 provides a numerical example. Sec. 5 offers concluding remarks.

1.2 Notation

Capital fonts mark mappings. Their discrete-valued arguments are at their subscript. $:=$ defines the left-hand side by the assignment ($=:$ swaps it). Decorated mnemonic labels are employed: v is variable in the **set** $\mathbf{v}$ of the cardinality $\mathfrak{c}_{\mathbf{v}}$. The superscript $^{\mathrm{opt}}$ marks optimality. $\mathrm{supp}(P) := \{v \in \mathbf{v} : P(v) > 0\}$ is the support of the pd P. $\propto$ is the equality up to the normalising factor. $R(P||P_{\mathfrak{o}}) := \int_{\mathbf{v}} P(v) \ln\left(\frac{P(v)}{P_{\mathfrak{o}}(v)}\right) \mathrm{d}v$ is the relative entropy [22] of the pd P to the pd $P_{\mathfrak{o}}$ on the set $\mathbf{v}$. The dominating measure [28] is either Lebesgue's or counting one reducing the integral to the sum. Time at subscripts is marked by $\mathfrak{t} \in \{\mathfrak{o}, \mathfrak{1}, \ldots, \mathfrak{c}_{\mathfrak{t}}\}$ while $\mathfrak{o}$ points to prior objects. Symbol $|$ separates the condition. Thus, $P_{\mathfrak{o}|h}(v)$ is the prior pd on $\mathbf{v}$ within (conditioned by) the hypothesis labelled by $h \in \mathbf{h} := \{1, \ldots, \mathfrak{c}_{\mathbf{h}}\}$.

1.3 Minimum relative entropy principle

The adopted minimum relative entropy principle chooses a pd P^{opt} acting on a set $\mathbf{v} \neq \emptyset$. Its choice complies with the knowledge that $P \in \mathbf{P} \neq \emptyset$ while keeping the opted pd near to its given prior guess $P_\circ$ (mostly, $P_\circ \notin \mathbf{P}$)

$$P^{\mathsf{opt}} \in \operatorname{Arg}\min_{P\in\mathbf{P}} R(P||P_\circ). \tag{1}$$

Properties of R imply that the value of the minimized functional is finite iff there is a pd P in the knowledge-expressing set $\mathbf{P} \subseteq \{P(v) \geq 0 : \int_{\mathbf{v}} P(v)\mathrm{d}v = 1\}$, such that $\mathrm{supp}(P) := \{v \in \mathbf{v} : P(v) > 0\} \subseteq \mathrm{supp}(P_\circ)\}$.

2 PROBLEM AND ITS SOLUTION

We have a finite number of hypotheses $h \in \mathbf{h} := \{1, 2, \ldots, \mathfrak{c}_{\mathbf{h}}\}$, $\mathfrak{c}_{\mathbf{h}} > 1$, each stating that a given pd $P_{\circ|h}(v)$ is the best available prior description of the modelled variable $v \in \mathbf{v} \neq \emptyset$. The pds $P_{\circ|h}$, $h \in \mathbf{h}$, to be pooled differ but all have $\mathbf{v}$ as the support.

The prior pd (weight) $W_{\circ h} > 0$ on $h \in \mathbf{h}$ is assumed to be available. It is chosen either subjectively [29] or it results from a Bayesian learning [3], possibly using the extensions of the set $\mathbf{h}$ of hypotheses [14]. As the last resort, it is chosen as uniform when exploiting "insufficient-reasons arguments" [23].

A joint pd $J_h(v) := P(v)W_h$, acting on $(\mathbf{v}, \mathbf{h})$, is opted. The enforced independence of v and h expresses the wish to construct a single representative $P \in \mathbf{P} := \{\text{all pds on } \mathbf{v}\}$ of the pooled pds $(P_{\circ|h}(v))_{h\in\mathbf{h}, v\in\mathbf{v}}$. The weights (pds on $\mathbf{h}$) $W \in \mathbf{W} := \{W_h > 0 : \sum_{h\in\mathbf{h}} W_h = 1\}$ are opted, too. Symbolically, $J \in \mathbf{J} := \mathbf{PW}$. The pooled pds $P_{\circ|h}$ and the prior pd $W_{\circ h}$ provide the prior guess $J_\circ$ of the joint pd J via the chain rule for pds $J_{\circ h}(v) := P_{\circ|h}(v)W_{\circ h}$, $v \in \mathbf{v}, h \in \mathbf{h}$. Thus, all inputs to MREP are available. The next proposition applies it.

Proposition 1 (Geometric Pooling by MREP (1)) *For the given pooled pds $(P_{\circ|h}(v))_{h\in\mathbf{h}, v\in\mathbf{v}}$ and the prior weights $(W_{\circ h})_{h\in\mathbf{h}}$, giving the prior guess $J_{\circ h}(v) := P_{\circ|h}(v)W_{\circ h}$ of the joint pd $J_h(v) = P(v)W_h$, MREP provides*

$$J^{\mathsf{opt}} := P^{\mathsf{opt}}W^{\mathsf{opt}} \in \operatorname{Arg}\min_{J\in\mathbf{J}} R(J||J_\circ), \text{ with}$$

$$P^{\mathsf{opt}}(v) \propto \prod_{h\in\mathbf{h}} P_{\circ|h}^{w_h^{\mathsf{opt}}}(v). \text{ The optimal weight values are}$$

$$w^{\mathsf{opt}} \in \operatorname{Arg}\min_{w\in\mathbf{w}} F(w), \ \mathbf{w} := \Big\{w_h \geq 0, \sum_{h\in\mathbf{h}} w_h = 1\Big\}$$

$$F(w) := R(w||w_\circ) - \ln\Big(\int_{\mathbf{v}} \prod_{h\in\mathbf{h}} P_{\circ|h}^{w_h}(v)\mathrm{d}v\Big) \tag{2}$$

$$w_h := W_h, \ w_h^{\mathsf{opt}} := W_h^{\mathsf{opt}}, \ w_{\circ h} := W_{\circ h}, \ h \in \mathbf{h}.$$

The MREP-optimal weight values w^{opt} prescribed by (2) are unique and belong to the interior of the set of W-values.

Proof R to be minimized reads $R(PW||P_\circ W_\circ) =$

$$\begin{aligned}
&\sum_{h\in\mathbf{h}} \int_{\mathbf{v}} P(v)W_h \ln\left(\frac{P(v)W_h}{P_{\circ|h}(v)W_{\circ h}}\right)\mathrm{d}v \\
&= \sum_{h\in\mathbf{h}} W_h \left[\ln\left(\frac{W_h}{W_{\circ h}}\right) + \int_{\mathbf{v}} P(v)\left(\frac{P(v)}{P_{\circ|h}(v)}\right)\mathrm{d}v\right] \\
&= \sum_{h\in\mathbf{h}} W_h \ln\left(\frac{W_h}{W_{\circ h}\int_{\mathbf{v}}\prod_{h\in\mathbf{h}} P_{\circ|h}^{W_h}(\tilde{v})\mathrm{d}\tilde{v}}\right) \\
&\qquad + \int_{\mathbf{v}} P(v)\ln\left(\frac{P(v)}{\frac{\prod_{h\in\mathbf{h}} P_{\circ|h}^{W_h}(v)}{\int_{\mathbf{v}}\prod_{h\in\mathbf{h}} P_{\circ|h}^{W_h}(\tilde{v})\mathrm{d}\tilde{v}}}\right)\mathrm{d}v.
\end{aligned}$$

For values $w_h := W_h$, the last summand above reaches its smallest zero value for the claimed geometric mean. For the chosen $w_{\circ h} := W_{\circ h}$, the remainder to be minimized coincides with the function $F(w)$ (2).

The following induction proves that the optimal weights are in the interior of the set of possible W-values.

First, the case $\mathfrak{c}_{\mathbf{h}} = 2$ is handled. W has values in $\{w, 1-w\}$ given by the optional $w \in [0, 1]$. The optimal w^{opt} minimizes the function $F(w) + (1-w)\gamma$ with a fixed $\gamma \geq 0$ ($\gamma = 0$ in this case) over $w \in [0, 1]$. The function and its derivatives read

$$\begin{aligned}
F(w) =& w\ln\left(\frac{w}{w_\circ}\right) + (1-w)\ln\left(\frac{1-w}{1-w_\circ}\right) \\
&- \ln\left(\int_{\mathbf{w}}\left[\frac{P_{\circ|1}(v)}{P_{\circ|2}(v)}\right]^w P_{\circ|2}(v)\mathrm{d}v\right) + (1-w)\gamma \\
\frac{\mathrm{d}F(w)}{\mathrm{d}w} =& \ln\left[\frac{w(1-w_\circ)}{w_\circ(1-w)}\right] - \int_{\mathbf{v}}\ln\left[\frac{P_{\circ|1}(v)}{P_{\circ|2}(v)}\right]P(v|w)\mathrm{d}v \\
&- \gamma, \text{ where } P(v|w) := \frac{P_{\circ|1}^w(v)P_{\circ|2}^{1-w}(v)}{\int_{\mathbf{v}} P_{\circ|1}^w(v)P_{\circ|2}^{1-w}(v)\mathrm{d}v}. \\
\frac{\mathrm{d}^2F(w)}{\mathrm{d}w^2} =& \frac{1}{w} + \frac{1}{1-w} - \text{variance}\left(\ln\left[\frac{P_{\circ|1}}{P_{\circ|2}}\right]\right),
\end{aligned}$$

the variance of the $\ln\left[\frac{P_{\circ|1}}{P_{\circ|2}}\right]$ is with respect to $P(v|w)$.

It remains to check the solvability of $\frac{\mathrm{d}F(w)}{\mathrm{d}w} = 0$ in $(0, 1)$. This equation gets the form

$$\ln\left[\frac{w(1-w_\circ)}{w_\circ(1-w)}\right] = \int_{\mathbf{v}}\ln\left[\frac{P_{\circ|1}}{P_{\circ|2}}\right]P(v|w)\mathrm{d}v + \gamma =: H(w). \tag{3}$$

The left-hand side of (3) increases in $w \in [0, 1]$ and covers the whole real line $[-\infty, \infty]$. The right-hand side $H(w)$ also increases in w. Its bounded range is $[\gamma -$

$R(P_{\mathfrak{o}|2}||P_{\mathfrak{o}|1}), \gamma + R(P_{\mathfrak{o}|1}||P_{\mathfrak{o}|2})]$. Thus, due to the continuity of both sides for $w \in (0,1)$, there is a finite number of solutions of (3). The strict convexity of R (2), which is unspoiled by the added term $(1-w)\gamma$, implies that the minimizer is uniquely one of them.

The induction step extends the properties of the solution to $\mathfrak{c}_\mathbf{h} > 2$ hypotheses. Let $\tilde{w}_h := \frac{w_h}{1-w_{\mathfrak{c}_\mathbf{h}}}$, $\tilde{w}_{\mathfrak{o}h} := \frac{w_{\mathfrak{o}h}}{1-w_{\mathfrak{o}\mathfrak{c}_\mathbf{h}}}$, $h \in \{1,\dots,\mathfrak{c}_\mathbf{h}-1\}$, and $\tilde{w}^{\mathsf{opt}}$ be the MREP solution, i.e. it is the best $\tilde{w}$. We minimize $F(w) + (1-w_{\mathfrak{c}_\mathbf{h}})\gamma$, $\gamma \geq 0$ for $\mathfrak{c}_\mathbf{h} > 2$. The minimization relies on the next re-arrangements of $F(w) + (1-w_{\mathfrak{c}_\mathbf{h}})\gamma :=$

$$
\begin{aligned}
&R(w||w_\mathfrak{o}) - \ln\left(\int_\mathbf{v} \prod_{h\in\mathbf{h}} P_{\mathfrak{o}|h}^{w_h}(v)\mathrm{d}v\right) + (1-w_{\mathfrak{c}_\mathbf{h}})\gamma \\
&= w_{\mathfrak{c}_\mathbf{h}} \ln\left(\frac{w_{\mathfrak{c}_\mathbf{h}}}{w_{\mathfrak{o}\mathfrak{c}_\mathbf{h}}}\right) + (1-w_{\mathfrak{c}_\mathbf{h}}) \ln\left(\frac{1-w_{\mathfrak{c}_\mathbf{h}}}{1-w_{\mathfrak{o}\mathfrak{c}_\mathbf{h}}}\right) \\
&+ (1-w_{\mathfrak{c}_\mathbf{h}})\left[\gamma + \sum_{h=1}^{\mathfrak{c}_\mathbf{h}-1} \frac{w_h}{1-w_{\mathfrak{c}_\mathbf{h}}} \ln\left(\frac{w_h(1-w_{\mathfrak{o}\mathfrak{c}_\mathbf{h}})}{(1-w_{\mathfrak{c}_\mathbf{h}})w_{\mathfrak{o}\mathfrak{c}_\mathbf{h}}}\right)\right] \\
&- \ln\left[\int_\mathbf{v} P_{\mathfrak{o}|\mathfrak{c}_\mathbf{h}}^{w_{\mathfrak{c}_\mathbf{h}}}(v)\Big(\prod_{h=1}^{\mathfrak{c}_\mathbf{h}-1} P_{\mathfrak{o}|h}^{\frac{w_h}{1-w_{\mathfrak{c}_\mathbf{h}}}}(v)\Big)^{1-w_{\mathfrak{c}_\mathbf{h}}}\mathrm{d}v\right] \\
&= w_{\mathfrak{c}_\mathbf{h}} \ln\left(\frac{w_{\mathfrak{c}_\mathbf{h}}}{w_{\mathfrak{o}\mathfrak{c}_\mathbf{h}}}\right) + (1-w_{\mathfrak{c}_\mathbf{h}}) \ln\left(\frac{1-w_{\mathfrak{c}_\mathbf{h}}}{1-w_{\mathfrak{o}\mathfrak{c}_\mathbf{h}}}\right) \\
&+ (1-w_{\mathfrak{c}_\mathbf{h}})\left[\gamma + \sum_{h=1}^{\mathfrak{c}_\mathbf{h}-1} \tilde{w}_h \ln\left(\frac{\tilde{w}_h}{\tilde{w}_{\mathfrak{o}h}}\right)\right] \\
&- \ln\Bigg[\int_\mathbf{v} P_{\mathfrak{o}|\mathfrak{c}_\mathbf{h}}^{w_{\mathfrak{c}_\mathbf{h}}}(v) \underbrace{\left(\frac{\prod_{h=1}^{\mathfrak{c}_\mathbf{h}-1} P_{\mathfrak{o}|h}^{\tilde{w}_h}(v)}{\int_\mathbf{v}\prod_{h=1}^{\mathfrak{c}_\mathbf{h}-1} P_{\mathfrak{o}|h}^{\tilde{w}_h}(v)\mathrm{d}v}\right)^{1-w_{\mathfrak{c}_\mathbf{h}}}}_{\tilde{P}(v):=} \mathrm{d}v\Bigg] \\
&- (1-w_{\mathfrak{c}_\mathbf{h}}) \ln\left(\int_\mathbf{v} \prod_{h=1}^{\mathfrak{c}_\mathbf{h}-1} P_{\mathfrak{o}|h}^{\tilde{w}_h}(v)\mathrm{d}v\right) \\
&\geq w_{\mathfrak{c}_\mathbf{h}} \ln\left(\frac{w_{\mathfrak{c}_\mathbf{h}}}{w_{\mathfrak{o}\mathfrak{c}_\mathbf{h}}}\right) + (1-w_{\mathfrak{c}_\mathbf{h}}) \ln\left(\frac{1-w_{\mathfrak{c}_\mathbf{h}}}{1-w_{\mathfrak{o}\mathfrak{c}_\mathbf{h}}}\right) \\
&- \ln\left(\int_\mathbf{v} P_{\mathfrak{o}\mathfrak{c}_\mathbf{h}}^{w_{\mathfrak{c}_\mathbf{h}}}(v)\tilde{P}^{1-w_{\mathfrak{c}_\mathbf{h}}}(v)\mathrm{d}v\right) + (1-w_{\mathfrak{c}_\mathbf{h}})\times \\
&\underbrace{\left\{\left[\gamma + \sum_{h=1}^{\mathfrak{c}_\mathbf{h}-1} \tilde{w}_h^{\mathsf{opt}} \ln\left(\frac{w_h^{\mathsf{opt}}}{\tilde{w}_{\mathfrak{o}h}}\right)\right] - \ln\left[\int_\mathbf{v} \prod_{h=1}^{\mathfrak{c}_\mathbf{h}-1} P_{\mathfrak{o}|h}^{\tilde{w}_h^{\mathsf{opt}}}(v)\mathrm{d}v\right]\right\}}_{\tilde{\gamma}:= \qquad\qquad (\tilde{\gamma}\geq\gamma\geq 0)}.
\end{aligned}
$$

The optimality of $\tilde{w}^{\mathsf{opt}}$ (inductively assumed for $\mathfrak{c}_\mathbf{h}-1$) implies the above inequality and the claimed properties of $\tilde{w}^{\mathsf{opt}}$. The made rearrangement reduces the lower bound to the two-dimensional case with $P_{\mathfrak{o}|1}(v) := P_{\mathfrak{o}|\mathfrak{c}_\mathbf{h}}(v)$, $w := w_{\mathfrak{c}_\mathbf{h}}$, $P_{\mathfrak{o}|2}(v) = \tilde{P}(v)$ and $\tilde{\gamma}$ replacing γ. This gives unique $w_{\mathfrak{c}_\mathbf{h}}^{\mathsf{opt}} \in (0,1)$ and re-normalizes the solution $w_h^{\mathsf{opt}} := (1-w_{\mathfrak{c}_\mathbf{h}}^{\mathsf{opt}})\tilde{w}_h^{\mathsf{opt}} \in (0,1)$ for $h \in \{1,\dots,\mathfrak{c}_\mathbf{h}-1\}$. □

3 ADAPTIVE BAYESIAN ESTIMATOR

The survey [18] and its references inform on the uses of geometric pooling in the knowledge fusion. This allows us to focus on its role in Bayesian adaptive learning. Our illustration considers one scenario. For others, see [7].

3.1 Estimation with the restricted stabilized forgetting

The modelled variable is the parameter $m \in \mathbf{m}$ entering the parametric model $M(d_\mathfrak{t}|\mathfrak{t}-\mathfrak{1},m)$. It is the pd relating the predicted data $d_\mathfrak{t} \in \mathbf{d}_\mathfrak{t}$ at discrete time, labelled by $\mathfrak{t} \in \{\mathfrak{1},\mathfrak{2},...,\mathfrak{c}_\mathfrak{t}\}$, to the available knowledge and the unknown values $m \in \mathbf{m}$. The pd $P_{\mathfrak{t}-\mathfrak{1}}(m)$ quantifies the knowledge about the values $m \in \mathbf{m}$. After getting $d_\mathfrak{t}$, Bayes' rule [26] updates this pd to the pd

$$
\begin{aligned}
\tilde{P}_\mathfrak{t}(m) &= \frac{M(d_\mathfrak{t}|\mathfrak{t}-\mathfrak{1},m)P_{\mathfrak{t}-\mathfrak{1}}(m)}{\int_\mathbf{m} M(d_\mathfrak{t}|\mathfrak{t}-\mathfrak{1},m)P_{\mathfrak{t}-\mathfrak{1}}(m)\mathrm{d}m} \\
&\propto M(d_\mathfrak{t}|\mathfrak{t}-\mathfrak{1},m)P_{\mathfrak{t}-\mathfrak{1}}(m), \quad \text{prior pd } P_\mathfrak{o} \text{ is given.}
\end{aligned} \tag{4}
$$

Notice that Bayes' rule (4) re-shapes $P_{\mathfrak{t}-\mathfrak{1}}(m)$ only on the set $\mathbf{m}_M \subset \mathbf{m}$ on which the *model-parameter-dependent likelihood* ($M(d_\mathfrak{t}|\mathfrak{t}-\mathfrak{1},m)$ with the inserted observed data) *varies*.

The gained pd $\tilde{P}_\mathfrak{t}(m)$ serves as the prior pd for time $\mathfrak{t}$ under the hypothesis, $h=1$, that the parameter is fixed between observations of data, i.e. $P_{\mathfrak{o}|1}(m) := \tilde{P}_\mathfrak{t}(m)$.

The estimation should model parameter changes and solve the filtering task [12] if the hypothesis $h=1$ is doubtful. Instead, the adaptive estimation considers the hypothesis, $h=2$, that the changes make another pd $P_{\mathfrak{o}|2}(m) := P_\mathfrak{o}(m)$ a better model of m than $P_{\mathfrak{o}|1}(m)$.

The used pessimistic alternative hypothesis, $h=2$, admits that all made observations of data are irrelevant for the description of the current parameter $P_{\mathfrak{o}|2} := P_\mathfrak{o}$. It ignores the observed data even on the set $\mathbf{m}_M$. Moreover, no modification is considered out of this set. The assumption that significant parameter changes occur on the set corrected by data underlies the adopted *restricted* forgetting [19,20].

The geometric pooling with the used $\mathfrak{c}_\mathbf{h}=2$ gives

$$
\begin{aligned}
&P_\mathfrak{t}(m) \propto M^w(d_\mathfrak{t}|\mathfrak{t}-\mathfrak{1},m)P_{\mathfrak{t}-\mathfrak{1}}^w(m)P_\mathfrak{o}^{1-w}(m) \text{ on } \mathbf{m}_M \\
&P_\mathfrak{t}(m) \propto P_{\mathfrak{t}-\mathfrak{1}}(m) \text{ out of the set } \mathbf{m}_M.
\end{aligned} \tag{5}
$$

The updating (5) on $\mathbf{m}_M$ is known as the *stabilized* forgetting [7]. It flattens the prior pd $P_{t-1}(m)$ in the way known as exponential forgetting but "returns" the forgotten part of the prior pd $P_\mathfrak{o}^{1-w}(m)$. Thus, $P_\mathfrak{o}(m)$ preserves. It is important when the processed data is non-informative. Prop. 1 guides how to choose $w = w^{\mathsf{opt}}$. It just needs a qualified guess $w_\mathfrak{o} := W_{\mathfrak{o}1}$, $W_{\mathfrak{o}2} = 1 - w_\mathfrak{o}$.

3.2 Parameter estimation of Markov-chain-type model

This part illustrates the theory on parameter estimation when both data $d_{\mathfrak{t}} \in \mathbf{d}$ and regressors $r_{\mathfrak{t}-\mathfrak{1}} \in \mathbf{r}$ (the knowledge part entering the parametric model) have finite amounts of possible values, $\mathbf{d} := \{1, 2, \ldots, \mathfrak{c}_{\mathbf{d}}\}$, $\mathfrak{c}_{\mathbf{d}} < \infty$, $\mathbf{r} := \{1, 2, \ldots, \mathfrak{c}_{\mathbf{r}}\}$, $\mathfrak{c}_{\mathbf{r}} < \infty$. In this case, the most general parametric model takes the transition probabilities $m := (m_{d|r})_{d,r\in\mathbf{d}}$ as unknowns $M(d_{\mathfrak{t}}|\mathfrak{t}-\mathfrak{1}, m) :=$

$$M_{d_{\mathfrak{t}}|r_{\mathfrak{t}-\mathfrak{1}}}(m) := m_{d_{\mathfrak{t}}|r_{\mathfrak{t}-\mathfrak{1}}} = \prod_{d\in\mathbf{d}}\prod_{r\in\mathbf{r}} m_{d|r}^{\Delta_{dd_{\mathfrak{t}}}\Delta_{rr_{\mathfrak{t}-\mathfrak{1}}}}$$

$$\Delta_{v\tilde{v}} := \begin{cases} 1 \text{ if } v = \tilde{v} \\ 0 \text{ otherwise} \end{cases} \tag{6}$$

$$m \in \mathbf{m} := \left\{ m_{d|r} \geq 0,\ d \in \mathbf{d},\ \sum_{d\in\mathbf{d}} m_{d|r} = 1,\ \forall r \in \mathbf{r} \right\}.$$

The last form of the parametric model in (6) helps to see that the next Dirichlet's pd of the parameter reproduces its form [9,16]

$$P_{\mathfrak{t}}(m) = P_{\mathfrak{t}|S_{\mathfrak{t}}}(m) := \prod_{r\in\mathbf{r}} \Gamma\Big(\sum_{\tilde{d}\in\mathbf{d}} S_{\mathfrak{t}\tilde{d}|r}\Big) \prod_{d\in\mathbf{d}} \frac{m_{d|r}^{S_{\mathfrak{t}d|r}-1}}{\Gamma(S_{\mathfrak{t}d|r})}$$

$$\Gamma(s) := \int_{\mathfrak{o}}^{\infty} v^{s-1}\exp(-v)\mathrm{d}v,\ s > 0, \text{ see [1]}. \tag{7}$$

It is given by the statistic $S_{\mathfrak{t}} := (S_{\mathfrak{t}d|r})_{d\in\mathbf{d},r\in\mathbf{r}}$ with positive entries. The statistic is sufficient one, i.e. it comprises all available information about the unknown m.

The predictive pd corresponding to (7) reads

$$P_{d|r,S} = \frac{S_{d|r}}{\sum_{\tilde{d}\in\mathbf{d}} S_{\tilde{d}|r}},\ d \in \mathbf{d},\ r \in \mathbf{r}. \tag{8}$$

Bayes' rule yields the statistic $\tilde{S}_{\mathfrak{t}d|r} := S_{(\mathfrak{t}-\mathfrak{1})d|r} + \Delta_{dd_{\mathfrak{t}}}\Delta_{rr_{\mathfrak{t}-\mathfrak{1}}}$. The restricted stabilized forgetting preserves Dirichlet's form with the updated statistic, $d \in \mathbf{d}$,

$$S_{\mathfrak{t}d|r_{\mathfrak{t}-\mathfrak{1}},w} = w(S_{(\mathfrak{t}-\mathfrak{1})d|r_{\mathfrak{t}-\mathfrak{1}}} + \Delta_{dd_{\mathfrak{t}}}) + (1-w)S_{\mathfrak{o}d|r_{\mathfrak{t}-\mathfrak{1}}}$$
$$S_{\mathfrak{t}d|r,w} = S_{\mathfrak{t}d|r} = S_{(\mathfrak{t}-\mathfrak{1})d|r} \text{ for } r \neq r_{\mathfrak{t}-\mathfrak{1}},\ \text{(out of } \mathbf{m}_M).$$

Properties of the non-linear function $F(w)$ (2) for $w \in (0,1)$ make its numerical minimisation simple. This allows us to use the MATLAB optimizer $FMINBND$ to find w^{opt}. The next explication of $F(w)$ omits time $\mathfrak{t}$ and the fixed condition $r_{\mathfrak{t}-\mathfrak{1}}$ and uses auxiliary sums

$$\tilde{S}_d := \tilde{S}_{\mathfrak{t}d|r_{\mathfrak{t}-\mathfrak{1}}},\ \tilde{C} := \sum_{d\in\mathbf{d}} \tilde{S}_d := \sum_{d\in\mathbf{d}} \tilde{S}_{\mathfrak{t}d|r_{\mathfrak{t}-\mathfrak{1}}}$$
$$S_d := S_{\mathfrak{o}d|r_{\mathfrak{t}-\mathfrak{1}}},\ C := \sum_{d\in\mathbf{d}} S_d := \sum_{d\in\mathbf{d}} S_{\mathfrak{o}d|r_{\mathfrak{t}-\mathfrak{1}}}.$$

The use of (7) gives the specific form of the minimized

$$F(w) := w\left[\ln\left(\frac{w}{w_{\mathfrak{o}}}\right) - \ln(\Gamma(\tilde{C})) + \sum_{d\in\mathbf{d}} \ln(\Gamma(\tilde{S}_d))\right]$$
$$+ (1-w)\left[\ln\left(\frac{1-w}{1-w_{\mathfrak{o}}}\right) - \ln(\Gamma(C)) + \sum_{d\in\mathbf{d}} \ln(\Gamma(S_d))\right]$$
$$+ \ln[\Gamma(w\tilde{C} + (1-w)C)] - \sum_{d\in\mathbf{d}} \ln[\Gamma(w\tilde{S}_d + (1-w)S_d)].$$

4 NUMERICAL EXAMPLE

This part numerically illustrates the estimation with forgetting. It simulates Markov chain with the observed states $d_{\mathfrak{t}} \in \mathbf{d} := \{1, 2, 3, 4\}$ and the regressor $r_{\mathfrak{t}-\mathfrak{1}} = d_{\mathfrak{t}-\mathfrak{1}}$, for $\mathfrak{t} \in \{\mathfrak{1}, \ldots, \mathfrak{c}_{\mathfrak{t}}\}$ with $\mathfrak{c}_{\mathfrak{t}} = 2000$. The simulated M has the transition probabilities $(M_{d_{\mathfrak{t}}|r_{\mathfrak{t}-\mathfrak{1}}})_{d_{\mathfrak{t}},r_{\mathfrak{t}-\mathfrak{1}}\in\mathbf{d}} :=$

$$\alpha_{\mathfrak{t}} \begin{bmatrix} 0.0 & 0.3 & 0.7 & 0.0 \\ 0.7 & 0.0 & 0.0 & 0.3 \\ 0.3 & 0.0 & 0.0 & 0.7 \\ 0.0 & 0.7 & 0.3 & 0.0 \end{bmatrix} + (1-\alpha_{\mathfrak{t}}) \begin{bmatrix} 0.5 & 0.7 & 0.0 & 0.0 \\ 0.5 & 0.3 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.2 & 0.6 \\ 0.0 & 0.0 & 0.8 & 0.4 \end{bmatrix}. \tag{9}$$

Its components, taken from [30] Fig. 13(b), Fig. 13(d), have uniform and [0.583 0.417 0 0] stationary probabilities, respectively. The simulated component weight $\alpha_{\mathfrak{t}} = |\sin(2\pi\mathfrak{t}/500)|$ makes the transition probabilities time-dependent. The prior statistic $S_{\mathfrak{o}}$ has constant entries 0.5 and $w_{\mathfrak{o}} = 0.5$ corresponds to the insufficient-reasons arguments.

The maximizer of the predictive pd (8) serves as the one-step-ahead point prediction of the state. The evaluation of prediction errors uses them. Fig. 1 illustrates the typical results. The values of the mean and standard deviation of absolute values of prediction errors were $[0.86, 1.41]$. They were smaller than [0.92,1.45] gained for the constant, non-optimized $w_{\mathfrak{o}} = 0.5$. The values are typical but exceptions exist.

5 CONCLUDING REMARKS

The optimized choice of the weights in the geometric pooling is the key paper result. It opens many opportunities due to its wide use. For instance, for: ▶ various versions of forgetting [6,7]; ▶ estimation of varying trusts into reliability of the external knowledge [27]; ▶ an objective, widely applicable way quantifying a unified group opinion [5,13], and generally ▶ knowledge fusion [18,33,34]. This communication indicates that the expended implementation effort will be rewarded.

Acknowledgement

EU-COST Action CA21169 supports this research.

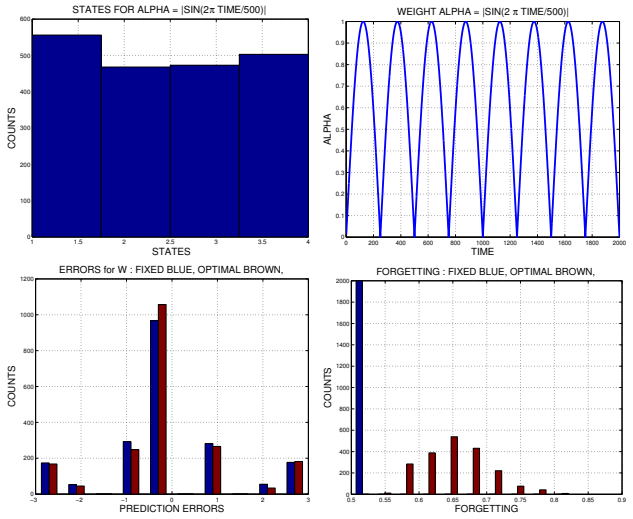

Fig. 1. Typical results for the random generator with the seed 13. They consist of the histogram of states, the time course of the simulated component weight α_t in (9), the histograms of the prediction errors and of the forgetting factors.

References

[1] M. Abramowitz and I.A. Stegun. *Handbook of Mathematical Functions*. Dover Publ., N.Y., 1972.

[2] S. Azizi and A. Quinn. Hierarchical fully probabilistic design for deliberator-based merging in multiple participant systems. *IEEE Tran. on SMC*, 48(4):565–573, 2018.

[3] J.O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, 1985.

[4] A.N. Bishop. Information fusion via the Wasserstein barycenter in the space of probability measures: Direct fusion of empirical measures and Gaussian fusion with unknown correlation. In *17th International Conference on Information Fusion*, pages 1–7. IEEE, 2014.

[5] J. Curtice and N. Sparrow. How accurate are traditional quota opinion polls? *J. of The Market Research Society*, 39(3):433–448, 1997.

[6] K. Dedecius, I. Nagy, and M.Kárný. Parameter tracking with partial forgetting method. *Int. J. ACASP.*, 26(1):1–12, 2012.

[7] J. Dokoupil and P. Václavek. Regularized estimation with variable exponential forgetting. In *59th IEEE Conference on Decision and Control*, pages 312–318, 2020.

[8] C.J. Feldbacher-Escamilla and G. Schurz. Meta-inductive probability aggregation. *Theory and Decision*, 95:663—689, 2023.

[9] T.S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1:209–230, 1973.

[10] C. Genest and J.V. Zidek. Combining probability distributions: A critique and annotated bibliography. *Stat. Sci.*, 1(1):114–148, 1986.

[11] D. Hartley and S. French. A Bayesian method for calibration and aggregation of expert judgement. *Int. J. of Appr. Reasoning*, 130:192–225, 2021.

[12] A.M. Jazwinski. *Stochastic Processes and Filtering Theory*. 1970.

[13] J.B. Kadane, J.M. Dickey, R.L. Winkler, W.S. Smith, and S.C. Peters. Interactive elicitation of opinions for normal linear models. *J. of the American Statistical Association*, 75(372):845–854, 1980.

[14] M. Kárný. On assigning probabilities to new hypotheses. *Pattern Recognition Letters*, 150:170–75, 2021.

[15] M. Kárný. Occam's razor in pooling of probability densities. *Information Sciences*, 2024. under review.

[16] M. Kárný, J. Böhm, T.V. Guy, L. Jirsa, I. Nagy, P. Nedoma, and L. Tesař. *Optimized Bayesian Dynamic Advising: Theory and Algorithms*. Springer, London, UK, 2006.

[17] M. Kárný and T.V. Guy. On support of imperfect Bayesian participants. In T.V. Guy, M. Kárný, and D.H. Wolpert, editors, *Decision Making with Imperfect Decision Makers*, pages 29–56. Springer, 2012.

[18] G. Koliander and et al. Fusion of probability density functions. *Proc. of the IEEE*, 110(4):404–453, 2022.

[19] R. Kulhavý. Restricted exponential forgetting in real-time identification. *Automatica*, 23(5):589–600, 1987.

[20] R. Kulhavý and M. Kárný. Tracking of slowly varying parameters by directional forgetting. In *Prepr. of the 9th IFAC World Congr.*, volume X, pages 178–183. IFAC, Budapest, 1984.

[21] R. Kulhavý and M. B. Zarrop. On a general concept of forgetting. *Int. J. of Control*, 58(4):905–924, 1993.

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

[23] P. Laplace. *Theorie Analytique des Probabilités*. Courcier, 1812.

[24] R. Ortega, J.G. Romero, and S. Aranovskiy. A new least squares parameter estimator for nonlinear regression equations with relaxed excitation conditions and forgetting factor. *Systems & Control Letters*, 169:105377, 2022.

[25] Y. Pan and T. Shi. Adaptive estimation and control with online data memory: A historical perspective. *IEEE Control Systems Letters*, 2024.

[26] V. Peterka. Bayesian system identification. In P. Eykhoff, editor, *Trends & Progress in System Identification*, pages 239–304. 1981.

[27] A. Quinn, M. Kárný, and T.V. Guy. Optimal design of priors constrained by external predictors. *Int. J. Appr. Reasoning*, 84:150–58, 2017.

[28] M.M. Rao. *Measure Theory and Integration*. J. Wiley, 1987.

[29] L.J. Savage. *Foundations of Statistics*. Wiley, 1954.

[30] E. Seabrook and L. Wiskott. A tutorial on the spectral theory of Markov chains, 2023.

[31] J. Shore and R. Johnson. Axiomatic derivation of the principle of maximum entropy & the principle of minimum cross-entropy. *IEEE Tran. on Inf. Th.*, 26(1):26–37, 1980.

[32] C.N. Taylor and A.N. Bishop. Homogeneous functionals and Bayesian data fusion with unknown correlation. *Inf. Fusion*, 45:179–189, 2019.

[33] P. Wang, L.T. Yang, J. Li, J. Chen, and S. Hu. Data fusion in cyber-physical-social systems: State-of-the-art and perspectives. *Inf. Fusion*, 51:42 – 57, 2019.

[34] Y.M. Zhu and X.R. Li. Unified fusion rules for multisensor multihypothesis network decision systems. *IEEE Tran. on Systems Man and Cybernetics, Part A – Systems and Humans*, 33(4):502–513, 2003.