

# Occam's Razor in Pooling of Probability Densities

Miroslav Kárný

**Abstract**—Geometric and linear poolings often serve for the fusion of the knowledge contained in a finite set of probability densities. Their pros and cons are relatively well understood. Many other ways have also been studied. A recent insightful survey paper by Koliander et al inspects a range of pooling ways based on various axioms, optimisation and supra-Bayesian handling. The gained extensive option set makes the proper choice of the pooling function harder. This paper reduces the extent of unjustified options. It provides the optimisation-based selection among available options. Its steps are justified by well-established, axiomatically supported, minimum relative entropy and approximation principles. The text applies Occam's razor to its theoretical tools, too. It simplifies the user's choice of the pooling function and its weights. This weakens the possibility of a bad choice and opens the way to a range of applications.

**Index Terms**—information entropy, minimum relative entropy principle, probability density function, forgetting

## I. INTRODUCTION

**A**N information fusion combines the knowledge expressed in various forms that originate from different sources. It includes, for instance, the fusion of data [1], the fusion of classifiers [2], multi-sensor fusion [3], online parameter estimation with forgetting [4], and many others [5]. Since the seminal work of [6], it is known that information is a characteristic of underlying probabilistic measures. They provide the most advanced knowledge quantification used, for instance, in agents' collective learning [7]. The popular large language models [8] represent a striking example of the probability strength. It stresses the importance of the information-fusion way that combines a finite collection of probability densities (pd<sup>1</sup>).

Geometric and linear poolings are standard, but competitive fusion ways of the knowledge quantified by pds. The papers of [12], [13], [14] represent permanent attempts to refine the pooling. The survey paper of [5] with an extensive list of references summarises such attempts. It added a range of methods on how to pool pds. The authors justified pooling functions using axiomatic, optimisation and supra-Bayesian ways. Their insightful study, however, makes the choice of the pooling method harder as it extends the wide option set and thus the extent of possible poor choices.

The current reductionist paper proposes a narrow and still sufficiently rich pooling basis. It uses only two axiomatic,

well-established principles: the minimum relative entropy principle, [15], [16], and the approximation principle, [17], [16]. This choice applies Occam's razor: "Do not multiply entities without necessity!" to the set of theoretical tools. This informally applied dictum is deeply discussed in [18].

The text complements the discussion of the pooling in [5]. Mainly, it decreases the user's effort spent on the choice of the proper pooling function and its parameters. The text extends the study on the pds' handling in [19]. Practically, a hybrid of arithmetic and geometric pooling arises (see Fig. 1 in the core Sec. II-B). It inherits the strong points of these popular techniques and suppresses their weaknesses. Methodologically, the combination results from the addressed problem formulation, not from an ad hoc design.

## A. Notation

*Throughout:* sans-serif fonts mark mappings;  $\mathcal{A} : \mathcal{B}$  : etc. label assumptions and various claims;  $:=$  defines the left-hand side by the assignment;  $\mathbf{v}$  means a set of  $v$ s, specified only if needed; decorated mnemonic labels are used, e.g.,  $c$  is a cover-set label in the set  $\mathcal{c}$  of the cardinality  $\mathfrak{c}_c < \infty$ ;  $^\circ$  marks the optimality;  $\text{supp}(\mathbf{p}) := \{v \in \mathbf{v} : \mathbf{p}(v) > 0\}$  is the support of the pd  $\mathbf{p}$ ;  $\propto$  is the implicitly normalised equality; the set indicator  $\chi_{\mathbf{v}}(v) := 1$  if  $v \in \mathbf{v}$ , otherwise  $\chi_{\mathbf{v}}(v) := 0$ ;  $D(\mathbf{p}||\mathbf{q}) := \int_{\mathbf{v}} \mathbf{p}(v) \ln \left( \frac{\mathbf{p}(v)}{\mathbf{q}(v)} \right) dv$  is the relative entropy, [20], of a pd pair  $\mathbf{p}, \mathbf{q}$ ; the dominating measure used in definitions of pds is formally marked as Lebesgue's one.

## B. Used Principles

Operationally, the used minimum relative entropy principle chooses a single pd serving to a subsequent optimising decision making. This aim singles out the optimising pooling category of [5] as the proper one. The pd choice respects the processed partial knowledge while keeping the constructed pd near to its prior approximation.

The quest for computational feasibility calls for the accompanying use of the approximation principle. It guides how to approximate a given pd. Both principles measure the similarity of pds. They single out the relative entropy with the proper argument order as the adequate divergence. The 1<sup>st</sup> principle is the main tool, the 2<sup>nd</sup> one supports its use.

*Principle 1 (Minimum Relative Entropy Principle):* Let a random variable  $v \in \mathbf{v}$  have a partially known pd  $\mathbf{p} \in \mathbf{p} \neq \emptyset$ . Let the pd  $\mathbf{a}_0$  be the given prior approximation of the proper pd  $\mathbf{p}$ . Then,

$$\mathbf{p}^\circ \in \underset{\mathbf{p} \in \mathbf{p}}{\text{Argmin}} D(\mathbf{p}||\mathbf{a}_0) := \underset{\mathbf{p} \in \mathbf{p}}{\text{Argmin}} \int_{\mathbf{v}} \mathbf{p}(v) \ln \left( \frac{\mathbf{p}(v)}{\mathbf{a}_0(v)} \right) dv \quad (1)$$

optimally fuses the knowledge  $(\mathbf{v}, \mathbf{p}, \mathbf{a}_0)$ ; often,  $\mathbf{a}_0 \notin \mathbf{p}$ .  $\square$

This work was supported by the EU-COST Action CA21169.

M. Kárný is with The Czech Academy of Sciences, Institute of Information Theory and Automation, Prague 8,182 08, Czech Republic (e-mail: school@utia.cas.cz)

<sup>1</sup>They are Radon-Nikodým's derivatives, [9]. They are probability density/mass functions in continuous/discrete cases. The term "pd" covers both cases. Often, their characteristics are pooled [10], [11]. ©20205 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See <https://www.ieee.org/publications/rights/index.html> for more information

Principle 1, giving (1), is a proposition in [15] and in [16]. The axioms that imply its validity are weak.

*Principle 2 (Approximation Principle):* Let  $v \in \mathbf{v}$  be a random variable with the known pd  $\mathbf{p} \notin \mathbf{a}$ . Let its approximation  $\mathbf{a} \in \mathbf{a} \neq \emptyset$  be needed. The pd

$$\mathbf{a}^\circ \in \underset{\mathbf{a} \in \mathbf{a}}{\text{Argmin}} D(\mathbf{p}||\mathbf{a}) := \underset{\mathbf{a} \in \mathbf{a}}{\text{Argmin}} \int_{\mathbf{v}} \mathbf{p}(v) \ln \left( \frac{\mathbf{p}(v)}{\mathbf{a}(v)} \right) dv \quad (2)$$

serves as the optimal approximation of the pd  $\mathbf{p}$ .  $\square$

Principle 2, giving (2), is a proposition in [17] and in [16]. The axioms that imply its validity are weak.

*Remarks 1 (On Principles):*

- ▶ The principles serve to solve the pooling task, Prop. 1.
- ▶ Importantly, the principles are not selected ad hoc. They are propositions of an axiomatic decision-making theory called fully probabilistic design [21]. This implies that the problem is deductively solved. Weak explicit conditions imply the solution optimality.
- ▶ The principles delimit the order of the argument in the relative entropy. It fits the use of information theory in statistical inference [22].
- ▶ The principles give nontrivial solutions differing from the “reference pds”  $\mathbf{a}_0, \mathbf{p}$  if they are *not* in the sets over which the minimisations run.  $\square$

## II. POOLING PROBLEM AND ITS SOLUTION

The use of the above principles under assumptions spelt in this section leads to our main result, Prop. 1. It describes the optimal pooling of a given collection of pds  $(\mathbf{a}_{ck}(v))_{v \in \mathbf{v}, c \in \mathbf{c}, k \in \mathbf{k}}$  into a single  $\mathbf{p}^\circ(v)$  serving to decision making optimising under uncertainty. The text explains the indices’ meaning.

Fig. 1 (left) provides a prototype of the addressed pooling for which neither the geometric nor the arithmetic pooling suits. Weakly overlapping echo chambers [23] are a real example of such cases. Fig. 1 (right) shows the result implied by Prop. 1. Linear and geometric poolings serve comparison.

### A. Adopted Assumptions $\mathcal{A} : \mathcal{B} : \mathcal{C}$ :

The random variable  $v \in \mathbf{v}$  is modelled by the partially known pd  $\mathbf{p}(v)$  with the support  $\text{supp}(\mathbf{p}) = \mathbf{v}$ . The used knowledge about  $\mathbf{p}$  consists of:

$\mathcal{A}$  : given cover sets  $(\mathbf{v}_c)_{c \in \mathbf{c}}$  covering  $\mathbf{v} : \mathbf{v} = \cup_{c \in \mathbf{c}} \mathbf{v}_c$  with  $\mathbf{v}_{\tilde{c}} \cap \mathbf{v}_c$  of the zero dominating measure if  $\tilde{c} \neq c, \tilde{c}, c \in \mathbf{c} := \{1, \dots, \mathbf{c}_c\}, \mathbf{c}_c < \infty$ ;

$\mathcal{B}$  : a given pd  $\mathbf{a}_0$  a priori approximating  $\mathbf{p}$  with  $\text{supp}(\mathbf{a}_0) = \text{supp}(\mathbf{p}) = \mathbf{v}$ . It gets the form

$$\mathbf{a}_0(v) = \sum_{c \in \mathbf{c}} b_{c0} r_{c0}(v) \quad \text{with} \quad b_{c0} := \int_{\mathbf{v}_c} \mathbf{a}_0(v) dv$$

with the restrictions  $r_{c0}(v) := \frac{\mathbf{a}_0(v) \chi_{\mathbf{v}_c}(v)}{b_{c0}}$ ; (3)

$\mathcal{C}$  : given pooled pds provide<sup>2</sup>  $(\mathbf{a}_{ck})_{c \in \mathbf{c}, k \in \mathbf{k}}$   $k \in \mathbf{k} := \{1, \dots, \mathbf{c}_k\}, \mathbf{c}_k < \infty$ , with  $\text{supp}(\mathbf{a}_{ck}) \supseteq \mathbf{v}_c$  and

<sup>2</sup>The notation neglects dependencies of  $\mathbf{c}_k$  on  $c \in \mathbf{c}$ , cf. Fig. 1.

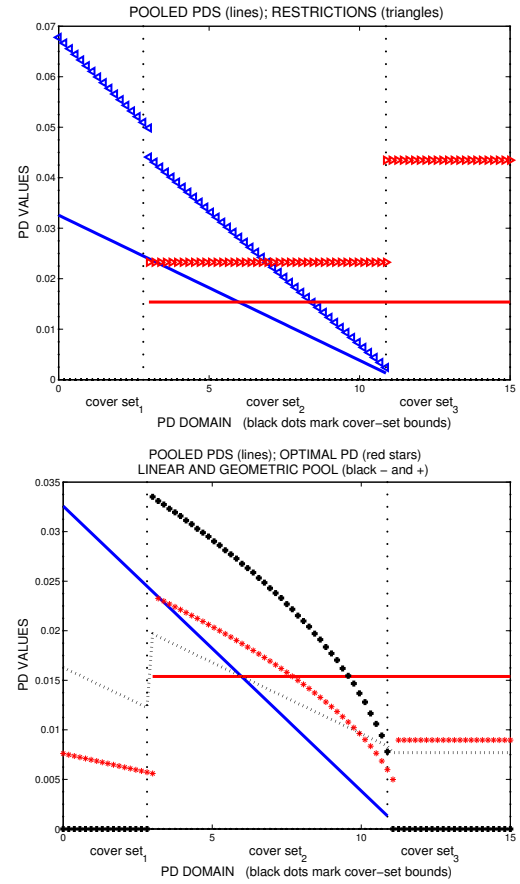


Fig. 1.

**Top:** The pooled are linear and constant pds (blue and red lines) (see  $\mathcal{C}$  :). They suit neither arithmetic nor geometric pooling. The non-empty intersection of their supports implies covering. Triangles, having the same colour as the pooled pds, mark the related restrictions.

**Bottom:** The pooled linear and constant pds (blue and red lines) from Fig. 1 repeat. Red stars mark the pd  $\mathbf{p}^\circ$ , see Proposition 1.

The optimal forgetting, (9) for uniform  $\mathbf{a}_{c=2,0}, f_{c=2,0} := 0$ , in the local geometric pooling is  $f_{c=2}^\circ = [0.5306, 0.4694]$ . The optimal beliefs into the restrictions  $r_c^\circ$  of the optimal  $\mathbf{p}^\circ$  (8) are  $b^\circ = [0.1122, 0.6815, 0.2063]$ .

Linear & geometric poolings (black -,+) with weight 0.5 serve comparison.

restrictions  $r_{ck}(v) \propto \mathbf{a}_{ck}(v) \chi_{\mathbf{v}_c}(v)$  approximating restrictions

$$r_c(v) := \frac{\mathbf{p}(v) \chi_{\mathbf{v}_c}(v)}{b_c}, \quad b_c := \int_{\mathbf{v}_c} \mathbf{p}(v) dv, \quad (4)$$

which are given by the partially known  $\mathbf{p}$  on  $\mathbf{v}_c, \forall c \in \mathbf{c}$ .  $\square$

The definition (4) implies the identity

$$\mathbf{p}(v) = \sum_{c \in \mathbf{c}} b_c r_c(v). \quad (5)$$

The pds  $(r_c)_{c \in \mathbf{c}}$  and the beliefs  $(b_c)_{c \in \mathbf{c}}$  (4), i.e. probabilities of  $\mathbf{v}_c$ , depend on the partially known  $\mathbf{p} \in \mathbf{p}$ . The assumed  $\text{supp}(\mathbf{a}_0) = \text{supp}(\mathbf{p}) = \mathbf{v}$ , ( $\mathcal{B}$  :), allows to use Principle 1 for opting  $\mathbf{p}^\circ$  as a minimiser of the relative entropy of  $\mathbf{p} \in \mathbf{p}$  to its prior approximation  $\mathbf{a}_0$  (3). Principle 2 expresses the knowledge in the pooled pds,  $\mathcal{C}$  :, via the set

$$\mathbf{p} := \mathbf{p}_\beta := \left\{ \mathbf{p} : (D(r_c||r_{ck}) \leq \beta_c < \infty)_{c \in \mathbf{c}, k \in \mathbf{k}} \right\}$$

$$\beta := (\beta_c)_{c \in \mathbf{c}} \in \mathbf{B} := \{ \beta_c \in [0, \infty), c \in \mathbf{c} \}. \quad (6)$$

Choice (6) says that the restrictions  $(r_{ck})_{k \in \mathbf{k}}$  (4) of all pooled pds  $(a_{ck})_{k \in \mathbf{k}}$  ( $\mathcal{C}$  : ) approximate  $r_c$  locally on  $\mathbf{v}_c$  ( $r_c$  restricts the pd  $\mathbf{p}$  on  $\mathbf{v}_c$ ). The optional bounds  $\beta$  in (6) quantify how well  $(r_{ck})_{k \in \mathbf{k}}$  approximate  $r_c$ .

*Remarks 2 (On Assumptions & Existence of the Solution):*

- ▶ Let us stress that the pooled pds  $a_0, a_{ck}$  (with the restrictions  $r_{ck}$ ,  $k \in \mathbf{k}_0 := \{0\} \cup \mathbf{k}$ ) may arise from different knowledge sources exploiting even completely different content. They have to model the same variable  $v \in \mathbf{v}$ .
- ▶ For each  $c \in \mathbf{c}$ , pds  $(r_{ck})_{k \in \mathbf{k}_0}$  are not in a full contradiction as they have supports identical with  $\mathbf{v}_c = \text{supp}(r_c)$ . This guarantees the existence of minimisers implied by elementary properties of the relative entropy [20], [22]. Finite bounds  $\beta$  in (6) thus also exist.  $\square$

Low values of  $\beta$ , for which a solution of (1) on the set  $\mathbf{p} := \mathbf{p}_\beta$  (6) exists, well exploit the processed knowledge. The adopted choice seeks the smallest set  $\mathbf{p}_\beta$ . It leads to the highest but finite achieved minimum of  $D(\mathbf{p}||a_0)$ . This gives the optimal bounds (6)

$$\beta^0 \in \text{Arg max}_{\beta \in \mathbf{p}} \min_{\mathbf{p} \in \mathbf{p}_\beta} D(\mathbf{p}||a_0). \quad (7)$$

### B. Main Result: Pooling Design

This part constructively demonstrates the key methodological claim of the paper by designing the optimal pooling.

*Proposition 1 (Pooling):*

$\mathcal{D}$  : Assumptions  $\mathcal{A}$  :,  $\mathcal{B}$  :,  $\mathcal{C}$  :, and Principle 1 with the set (6) for  $\beta := \beta^0$  (7), approximately (see proof) provide

$$\begin{aligned} \mathbf{p}^0(v) &\propto \sum_{c \in \mathbf{c}} b_c^0 r_c^0(v), \quad c \in \mathbf{c}, \quad \text{with } b_c^0 \propto b_{c0} \exp[-D(r_c^0||r_{c0})] \\ r_c^0 &:= r_{f_c}(v) := \frac{\prod_{k \in \mathbf{k}_0} r_{ck}^{f_{ck}}(v)}{\int_{\mathbf{v}_c} \prod_{k \in \mathbf{k}_0} r_{ck}^{f_{ck}}(v) dv}, \quad \mathbf{k}_0 := \{0\} \cup \mathbf{k}, \\ f_c &:= [f_{c0}, f_{c1}, \dots, f_{c\mathbf{k}}], \quad f_{ck} \in [0, 1], \quad \sum_{k \in \mathbf{k}_0} f_{ck} = 1. \end{aligned} \quad (8)$$

The forgetting factors  $(f_c)_{c \in \mathbf{c}}$  depend on: ▶ the prior belief-expressing  $(b_{c0} > 0)_{c \in \mathbf{c}}$  (3); ▶ the belief reflecting Kuhn-Tucker's multipliers  $(b_{ck} \geq 0)_{c \in \mathbf{c}, k \in \mathbf{k}}$ , and; ▶ the normalisations  $n_c := (\sum_{k \in \mathbf{k}_0} b_{ck}) > 0$   $c \in \mathbf{c}$ . They read

$$f_{ck} := \frac{b_{ck}}{n_c}, \quad c \in \mathbf{c}, \quad k \in \mathbf{k}_0.$$

The forgetting factors

$$f_{ck}^0 \text{ maximising } - \sum_{c \in \mathbf{c}} n_c \ln \left( \int_{\mathbf{v}_c} \prod_{k \in \mathbf{k}_0} r_{ck}^{f_{ck}}(v) dv \right) \quad (9)$$

make the bounds in (6) (approximately) the tightest ones  $\beta_c \approx \beta_c^0$ , see (7). The sole normalisation  $n_c > 0$  has no influence on the optimal  $\mathbf{p}^0$  (8).

$\mathcal{E}$  : The optimal forgetting factors  $(f_{ck}^0)_{c \in \mathbf{c}, k \in \mathbf{k}_0}$  (9) exist. The forgetting factors  $f_c^0$  are unique if  $(\ln(r_{ck}(v)))_{k \in \mathbf{k}_0}$  for the  $c \in \mathbf{c}$  are linearly independent. The optimal factors meet the well-solvable necessary conditions

$$\int_{\mathbf{v}_c} \ln(r_{ck}(v)) r_{f_c^0}(v) dv = \lambda_c, \quad c \in \mathbf{c}, \quad \forall k \in \mathbf{k}_0 \text{ with} \quad (10)$$

real Lagrange's multiplier  $\lambda_c$  giving  $\sum_{k \in \mathbf{k}_0} f_{ck}^0 = 1$ ,  $c \in \mathbf{c}$ .

**Proof** It deals with Kuhn-Tucker's functional (KT) [24] with multipliers  $(b_{ck} \geq 0)_{c \in \mathbf{c}, k \in \mathbf{k}}$  respecting the inequalities in (6). On  $\mathcal{D}$  : KT is an affine map of relative entropies, smallest for their equal arguments. Under (3)–(6), it holds

$$\begin{aligned} \text{KT} &:= D(\mathbf{p}||a_0) + \sum_{c \in \mathbf{c}, k \in \mathbf{k}} b_{ck} (D(r_c||r_{ck}) - \beta_c) \\ &= \sum_{c \in \mathbf{c}} \left[ \int_{\mathbf{v}_c} b_c r_c(v) \ln \left( \frac{b_c r_c(v)}{b_{c0} r_{c0}(v)} \right) dv \right. \\ &\quad \left. + \sum_{k \in \mathbf{k}} b_{ck} \left( \int_{\mathbf{v}_c} r_c(v) \ln \left( \frac{r_c(v)}{r_{ck}(v)} \right) dv - \beta_c \right) \right]. \end{aligned}$$

The covering  $\mathbf{v} = \cup_{c \in \mathbf{c}} \mathbf{v}_c$  and the relative entropy definition are employed. The unknown  $(b_c)_{c \in \mathbf{c}}$ -dependent part of KT can be directly rewritten as the relative entropy of the optimised  $(b_c)_{c \in \mathbf{c}}$  (4) to  $(\tilde{b}_c^0 \propto b_{c0} \exp(-D(r_c||r_{c0})))_{c \in \mathbf{c}}$ . It gives the optimal  $(b_c^0 := \tilde{b}_c^0)_{c \in \mathbf{c}}$  dependent on yet unspecified  $(r_c)_{c \in \mathbf{c}}$ . The minimum value of this part over  $(b_c)_{c \in \mathbf{c}}$  is

$$- \ln \left[ \sum_{c \in \mathbf{c}} b_{c0} \exp[-D(r_c||r_{c0})] \right]. \quad (11)$$

The partially minimised KT should be optimised over  $(r_c)_{c \in \mathbf{c}}$ . It cannot be done explicitly. KT is bounded from above via Jensen's inequality. The bound<sup>3</sup> is minimised. This is the approximation stated in the proposition. Jensen's bound on (11) is  $-\ln \left[ \sum_{c \in \mathbf{c}} b_{c0} \exp[-D(r_c||r_{c0})] \right] \leq \sum_{c \in \mathbf{c}} b_{c0} D(r_c||r_{c0})$ . For  $\mathbf{k}_0 := \{0\} \cup \mathbf{k}$ ,  $n_c := \sum_{k \in \mathbf{k}_0} b_{ck} > 0$  ( $b_{c0} > 0$ ),

$$\text{KT} \leq \sum_{c \in \mathbf{c}} n_c \int_{\mathbf{v}_c} r_c(v) \ln \left( \frac{r_c(v)}{\prod_{k \in \mathbf{k}_0} r_{ck}^{n_c} (v)} \right) dv. \quad (12)$$

A simple algebra provides this upper bound. The summands are just multiplied by ratios  $n_c/n_c$ .  $n_c > 0$  normalises  $b_{ck}$  to forgetting factors  $(f_{ck})_{k \in \mathbf{k}_0}$ . The normalisation  $n_c > 0$  in (12) as  $b_{c0} > 0$  (3), and Kuhn-Tucker's multipliers  $(b_{ck} \geq 0)_{k \in \mathbf{k}}$ . The upper bound (12) is the weighted sum of relative entropies and it is minimised by

$$r_c^0 := r_{f_c}(v) \propto \prod_{k \in \mathbf{k}_0} r_{ck}^{f_{ck}}(v), \quad f_{ck} := \frac{b_{ck}}{\sum_{k \in \mathbf{k}_0} b_{ck}}, \quad k \in \mathbf{k}_0.$$

Clearly,  $f_{ck} \in [0, 1]$ ,  $k \in \mathbf{k}_0$ , and  $\sum_{k \in \mathbf{k}_0} f_{ck} = 1$ ,  $c \in \mathbf{c}$ .

On  $\mathcal{E}$  : The reached minimum (9) is the negative sum of  $n_c$  times logarithms of the normalisation integrals of the optimal  $r_c^0 = r_{f_c}$  minus  $f$ -independent  $\sum_{c \in \mathbf{c}} n_c \beta_c$ . The minimisers of the addends in (9) depend on the forgetting factors, not on the positive normalisation  $n_c$ . Thus, the sole  $n_c$ -values have no influence on the pd  $\mathbf{p}^0$  (8). The maximising forgetting values guarantee the tightest bounds (allowed by the adopted Jensen's approximation) and thus the (implicitly) smallest  $\beta$  in (6).

For each  $c \in \mathbf{c}$ , the optimising forgetting factors  $f_c \geq 0$ , constrained by  $\sum_{k \in \mathbf{k}_0} f_{ck} = 1$ , are to be the stationary points, i.e. have to meet (10). The real  $\lambda_c$  is Lagrange's multiplier (shifted by  $\beta_c$ ) making  $\sum_{k \in \mathbf{k}_0} f_{ck} = 1$ .

<sup>3</sup>It is the tightest bound unless additional assumptions are enforced [9].

The arguments of  $\ln(\cdot)$  in (9) have, for  $k, \tilde{k} \in \mathbf{k}_0$ ,

$$\begin{aligned} \text{Hessian}_{c\tilde{k}} &= - \int_{\mathbf{v}_c} \ln(r_{c\tilde{k}}(v)) \ln(r_{c\tilde{k}}(v)) r_{f_c}(v) dv \\ &+ \int_{\mathbf{v}_c} \ln(r_{c\tilde{k}}(v)) r_{f_c}(v) dv \int_{\mathbf{v}_c} \ln(r_{c\tilde{k}}(\tilde{v})) r_{f_c}(\tilde{v}) d\tilde{v}, \quad \forall c \in \mathbf{c}. \end{aligned}$$

Thus, the Hessian is the negative covariance of  $[\ln(r_{c0}(v)), \dots, \ln(r_{c\mathbf{k}_k}(v))]$  with respect to  $r_{f_c}(v)$  (8). For  $c \in \mathbf{c}$ , it is negative definite for linearly independent  $(\ln(r_{c\mathbf{k}}(v)))_{\mathbf{k} \in \mathbf{k}_0}$ . The maximised, continuous, unimodal function is thus strictly convex. Even the non-strict convexity makes the solution of (10) simple. The optimal  $\mathbf{p}^o(v) = \sum_{c \in \mathbf{c}} b_c^o \chi_{\mathbf{v}_c}(v) r_c^o(v)$  (8) is due to (5).  $\square$

### III. WRAP UP: TECHNIQUE, METHODOLOGY & PRACTICE

The addressed task and its solution have many links to other pooling ways. They are worth inspecting. The implementations also need extra effort. It mainly concerns data-based updating of inputs of the solution (8), and the covering  $(\mathbf{v}_c)_{c \in \mathbf{c}}$ .

*Technically:*

- ▶ The optimal pd  $\mathbf{p}^o$  (8) linearly pools the geometric means of compatible pds. The linear pooling is the special case for  $\mathbf{c}_k = 1$  and non-overlapping supports of the pooled pds. For  $\mathbf{c}_c = 1$ , it pools geometrically.
- ▶ The exponents  $f_{ck} \in [0, 1]$  flatten the respective pds, which is the main role of forgetting [25]. It acts as the stabilised forgetting: the more  $(r_{ck})_{\mathbf{k} \in \mathbf{k}}$ , are forgotten, the stronger is the influence of the pd  $r_{c0}$ .
- ▶ The knowledge in a pd  $r_{ck}$  is not incorporated if  $f_{ck} \approx 0$ . It happens if the multiplier  $b_{ck} \approx 0$  (the constraint in the set (6) are inactive). It indicates that the forgetting factors influence how seriously a knowledge piece is included in the pd  $\mathbf{p}^o$ : they quantify trust into pds  $(r_{ck})_{\mathbf{k} \in \mathbf{k}_0}$ .
- ▶ The pooled pds  $(a_{ck})_{\mathbf{k} \in \mathbf{k}_0}$  pds may (sequentially) incorporate knowledge contained in the observed data. Bayes' rule and its generalisation [26] serve this purpose.
- ▶ Principle 1 allows the (hard) optimisation of the user's options, including the choice of the covering.
- ▶ The belief  $b_c$  (4) says how probable (as per  $\mathbf{p}(v)$ ) is  $v \in \mathbf{v}_c$ . The desirable uniform  $(b_c)_{c \in \mathbf{c}}$  (principle of insufficient reasons [27]) could be sought by a covering redefinition. Its prior design groups pds  $(a_{ck})_{\mathbf{k} \in \mathbf{k}}$  with the supports' intersection  $\mathbf{v}_c$ , Fig. 1.
- ▶ Our result is valid even if the sets  $(\mathbf{v}_c)_{c \in \mathbf{c}}$  do not cover the whole  $\mathbf{v}$ . It only needs  $\text{supp}(a_0) = \mathbf{v} = \text{supp}(\mathbf{p})$ , see the assumption  $\mathcal{B}$  :. This way focuses on highly probable parts of  $\mathbf{v}$  and takes  $a_0$  as a "background".
- ▶ A use of a collection of pds potentially decreases sensing, modelling, processing and computational complexity. However, the latter is increased by the optimised pooling. A proper balance has to be sought.  $\square$

*Methodologically:*

- ▶ It is important that the result is the deductive outcome. It has clear, general, parsimonious and widely applicable conditions of its validity.
- ▶ Deductive results need no simulations to verify their validity. Real-life applications are, however, vital. The paper enables them.  $\square$

*Practically*

- ▶ The burden on the pooling users connected with the choice of an appropriate method is greatly reduced. This allows them to focus on their specific tasks.
- ▶ The tasks using probability pooling (fusion of experts' priors [14] and forgetting [4] for Bayesian parameter estimation, multi-sensor data fusion [3], collective learning [7] and many others [5]) get the balanced compromise between geometric and arithmetic pooling.

### REFERENCES

- [1] C. Cong, S. Liu, P. Rana, M. Pagnucco, A. Di Ieva, S. Berkovsky, and Y. Song, "Adaptive unified contrastive learning with graph-based feature aggregator for imbalanced medical image classification," *Expert Systems with Applications*, vol. 251, p. 123783, 2024.
- [2] P. Ksieniewicz *et al.*, "Fusion of linear base classifiers in geometric space," *Knowledge-Based Systems*, vol. 227, p. 107231, 2021.
- [3] C. Fantacci, B. Vo, B. Vo, G. Battistelli, and L. Chisci, "Robust fusion for multisensor multiobject tracking," *IEEE Signal Process. Letter*, vol. 25, no. 5, pp. 640–644, 2018.
- [4] M. Kárný, "Optimized geometric pooling of probabilities for information fusion and forgetting," *Automatica*, vol. 177, p. 112337, 2025.
- [5] G. Koliander and *et al.*, "Fusion of probability density functions," *Proc. of the IEEE*, vol. 110, no. 4, pp. 404–453, 2022.
- [6] C. Shannon, "A mathematical theory of communication," *Bell System Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.
- [7] J. Lawry and C. Lee, "Probability pooling for dependent agents in collective learning," *Artificial Intelligence*, vol. 288, p. 103371, 2020.
- [8] Y. Chang, X. Wang, J. Wang, and *et al.*, "A survey on evaluation of large language models," *ACM Tran. on Intel' Syst. and Technology*, vol. 15, no. 3, pp. 1–45, 2024.
- [9] M. Rao, *Measure Theory and Integration*. J. Wiley, 1987.
- [10] M. Rouvier, P. Bousquet, and J. Duret, "Study on the temporal pooling used in deep neural networks for speaker verification," in *29th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 501–505.
- [11] E. Raninen *et al.*, "Linear Pooling of Sample Covariance Matrices," *IEEE Transactions on Signal Processing*, vol. 70, pp. 659–672, 2022.
- [12] S. Azizi and A. Quinn, "Hierarchical fully probabilistic design for deliberator-based merging in multiple participant systems," *IEEE Tran. on SMC*, vol. 48, no. 4, pp. 565–573, 2018.
- [13] C. Taylor and A. Bishop, "Homogeneous functionals & Bayesian data fusion with unknown correlation," *Inf. Fus.*, vol. 45, pp. 179–189, 2019.
- [14] D. Hartley and S. French, "A Bayesian method for calibration and aggregation of expert judgement," *Int. J. of Appr. Reasoning*, vol. 130, pp. 192–225, 2021.
- [15] J. Shore and R. Johnson, "Axiomatic derivation of the principle of maximum entropy & the principle of minimum cross-entropy," *IEEE Tran. on Inf. Th.*, vol. 26, no. 1, pp. 26–37, 1980.
- [16] M. Kárný and T. Guy, "On support of imperfect Bayesian participants," in *Decision Making with Imperfect Decision Makers*, T. Guy, M. Kárný, and D. Wolpert, Eds. Springer, 2012, pp. 29–56.
- [17] J. Bernardo, "Expected information as expected utility," *The An. of Stat.*, vol. 7, pp. 686–90, 1979.
- [18] J. Schaffer, "What not to multiply without necessity," *Australasian Journal of Philosophy*, vol. 93, no. 4, pp. 644–664, 2015.
- [19] M. Kárný, "Prescriptive inductive operations on probabilities serving to decision-making agents," *IEEE Tran. on SMC: Systems*, vol. 52, no. 4, pp. 2110–20, 2022.
- [20] S. Kullback and R. Leibler, "On information and sufficiency," *Ann. Math. Stat.*, vol. 22, pp. 79–87, 1951.
- [21] M. Kárný, "Axiomatisation of fully probabilistic design revisited," *SCL*, vol. 141, p. 104719, 2020.
- [22] I. Vajda, *Theory of Statistical Inference and Information*. Dordrecht: Kluwer Academic Publishers, 1989.
- [23] W. Cota, S. C. Ferreira, R. Pastor-Satorras, and M. Starnini, "Quantifying echo chamber effects in information spreading over political communication networks," *EPJ Data Science*, vol. 8, no. 1, pp. 1–13, 2019.
- [24] H. Kuhn and A. Tucker, "Nonlinear programming," in *Proc. of 2nd Berkeley Symp.* Univ. of California Press, 1951, pp. 481–492.
- [25] R. Kulhavý and M. B. Zarrop, "On a general concept of forgetting," *Int. J. of Control*, vol. 58, no. 4, pp. 905–924, 1993.
- [26] A. Quinn, M. Kárný, and T. Guy, "Optimal design of priors constrained by external predictors," *IJAR*, vol. 84, pp. 150–58, 2017.
- [27] P. Laplace, *Theorie Analytique des Probabilités*. Courcier, 1812.