



Research article**A reliability-weighted product-of-experts framework for multimodal Bayesian classification: convexity, conformal coverage, and an information-theoretic necessity criterion****Safia Benarbia^{1,*}, Fatimah Alshahrani², Mohammed Fethi Khalfi³ and Wahiba Bouabsa¹**

¹ Laboratory of Statistics and Stochastic Processes, Department of Probability and Statistics, Faculty of Exact Sciences, Djillali Liabes University, BP 89, Sidi Bel Abbès 22000, Algeria

² Department of Mathematical Sciences, College of Science, Princess Nourah bint Abdulrahman University, P. O. Box 84428, Riyadh 11671, Saudi Arabia

³ LabRI-SBA Lab (ESI-SBA), Department of Computer Science, Djillali Liabes University, Sidi Bel Abbès 22000, Algeria

* **Correspondence:** Email: safiabenarbia.2001@gmail.com.

Abstract: We present a complete mathematical treatment of the reliability-weighted product-of-experts (RW-PoE) framework for multimodal Bayesian classification with provable guarantees. The modality posteriors are fused via a log-linear pool with weights $\mathbf{w} \in \Delta_{M-1}$ fitted by maximum likelihood on the simplex. We prove the following: (i) Strict convexity of the weight objective if and only if the modality log-posteriors are independent in the simplex-tangent sense of Definition 4.2, which guarantees a unique global maximum likelihood estimate (MLE); (ii) Additive decomposition of $H(Y) - H(Y | X)$ across a hierarchical cascade; (iii) An explicit non-asymptotic contraction bound on an interior simplex region and the exact γ -dependent local linearization of the finite-step entropic mirror update; (iv) Finite-sample marginal conformal coverage at least $1 - \alpha$, with the upper discretization bound $1 - \alpha + 1/(n_{\text{cal}} + 1)$ under the no-ties condition. A Hellinger–Bayes criterion yields a distribution-specific floor $R_{a,\text{root}}^* \geq 0.235$ on the acoustic-only error at the root node for the observed distributions, showing that fusion is necessary in this setting (Remark 3.3 specifies the scope of this claim). A clinical study on $n = 1105$ infant cry episodes (CHU Sidi Bel-Abbès) reports $\hat{\mathbf{w}} = (0.42 \pm 0.03, 0.31 \pm 0.04, 0.27 \pm 0.03)$, realized coverage 0.910 (target 0.900; no-ties theoretical band $[0.900, 0.905]$), accuracy 0.916, $\kappa = 0.889$ (95 % CI: 0.844–0.934).

Keywords: product of experts; logarithmic opinion pool; Bayesian classification; strict convexity on the simplex; expectation-maximization; entropic mirror descent; conformal prediction; Bhattacharyya coefficient; Hellinger distance

Mathematics Subject Classification: 62F15, 62G08, 62H30, 68T05, 94A17

1. Introduction

1.1. Motivation

Many inference problems present the analyst with several streams of evidence about a common latent label, where the streams differ in quality and where the quality of each stream is not known in advance and may even depend on the value of the label. A canonical instance, which motivates and supplies the running case study of this paper, is multimodal classification: one observes an acoustic signal, a vector of contextual covariates, and a behavioral (video-derived) descriptor, and must infer a discrete cause from a fixed taxonomy. Concretely, the case study of Section 7 concerns inferring the cause of an infant's cry from its sound, the infant's recorded clinical context, and a short video of its movements; the causes call for different responses, and in the data studied here the sound alone separates them poorly, which is what makes the fusion problem non-trivial. The central statistical difficulty is not the construction of any one modality-specific classifier, but the principled *fusion* of the modality-specific posteriors into a single posterior that (a) respects the differing and unknown reliabilities of the modalities, (b) is determined by transparent axioms rather than by an ad hoc combination rule, (c) is fitted by an estimator with provable optimization guarantees, and (d) delivers a quantified, distribution-free statement of confidence for each individual decision. Each of these four requirements has a mature classical answer in isolation. Logarithmic opinion pooling is the unique coherent way to aggregate probability distributions under natural axioms [1]; weighted log-pooling coincides with a weighted Kullback–Leibler centroid [2]. Exact expectation-maximization (EM) has a classical missing-information convergence theory [3, 4]; the finite-step entropic mirror map used here instead has a step-size-dependent linearization. The Bhattacharyya coefficient and the Hellinger distance furnish sharp lower bounds on the Bayes error [5]; and split conformal prediction yields finite-sample, distribution-free coverage under exchangeability alone [6, 7]. What has been missing is a single, fully specified framework that assembles these classical tools into one coherent estimator for reliability-weighted multimodal Bayesian classification, together with the regularity conditions under which the assembled estimator inherits the guarantees of its components. Supplying that framework, and proving those conditions, is the purpose of this paper. Although our running example is multimodal classification, reliability-weighted log-linear pooling arises across applied mathematics: in sensor fusion, where heterogeneous sensors of differing accuracy are combined; in ensemble methods, where base learners are aggregated with learned weights; and in distributed or federated inference, where local posteriors are pooled across nodes. The convexity and identifiability geometry of the pooling weights, and the distribution-free coverage layer developed here, are therefore of methodological interest beyond the specific application, and the results are stated at that level of generality.

1.2. Contributions

The contributions of this paper are mathematical and are stated below, with care taken not to overstate their novelty relative to the classical results recalled above (see Section 1.3).

- (1) **A unified reliability-weighted product-of-experts estimator (*new as an integration; components classical*).** We give a complete specification of a reliability-weighted product-of-experts (RW-PoE) estimator that assembles the logarithmic opinion pool, finite-step

generalized expectation-maximization (GEM)/entropic-mirror weight estimation, observed-information standard errors, a Bhattacharyya–Hellinger analysis, and a conformal layer into one internally consistent object (Sections 2–6); novelty lies in the integration, not the classical components.

- (2) **Strict convexity and uniqueness on the simplex (*explicit specialization*)**. We make explicit and prove, in the reliability-weighted log-pool setting, that log-posterior independence in the sense of Definition 4.2 is necessary and sufficient for strict convexity of the weight objective on the simplex, with the Hessian identified as $\sum_i \text{Var}_{k \sim P_w(\cdot | x_i)}[\ell_i(k)]$; the condition is necessary and sufficient for strict convexity and, when it holds, guarantees a unique global minimizer.
- (3) **An entropy-decomposition theorem for the cascade (*new*)**. We prove (Theorem 2.2, full chain-rule proof) that the probabilistic hierarchical cascade decomposes $H(Y) - H(Y | X)$ additively across nodes and never increases residual-label entropy, with strict decrease characterized node-by-node.
- (4) **A data-driven necessity criterion for fusion (*new use of a classical bound*)**. We formulate (Theorem 3.2) a criterion under which one modality provably cannot reach a target accuracy; the Hellinger–Bayes bound is classical [5], and what is new is its use as a distribution-specific necessity test, a falsifiable statement about the observed distributions and the observed feature representation, from which nothing follows for other populations, other recording conditions, or other representations, and whose full scope is set out in Remark 3.3.
- (5) **A complete case study (*new*)**. We report (Section 7) a single-site trimodal study with a stated data split and protocol, including divergence estimates, node-level discrimination, learned weights with standard errors, conformal coverage, and hypothesis tests.

1.3. Positioning relative to prior work

We are deliberate about the division between classical results that we *use* and results that are *ours*, since this division is what a careful reader should check.

Classical, and cited as such. The uniqueness of the logarithmic opinion pool as the coherent aggregation operator on the simplex is due to Genest and Zidek [1]; the identification of the weighted log-pool with a weighted-KL (Bregman) centroid is due to Abbas [2]; Heskes [8] treated the distinct problem of selecting the pool weights. The missing-information rate for an *exact* EM operator is due to Dempster et al. [3] and Louis [4]. It is recalled only for contrast: the finite-step map implemented here is differentiated directly in Proposition 5.2. The Hellinger/Bhattacharyya lower bound on the Bayes error is classical and is taken from Devroye et al. [5]. The finite-sample, distribution-free coverage of split conformal prediction under exchangeability is due to Vovk et al. [6], with the modern treatment and adaptive scores in Angelopoulos and Bates [7]. None of these results is claimed here as new; each is invoked with attribution as part of the classical theory the framework builds on.

Ours. Against this backdrop the five contributions of Section 1.2 are extensions of the classical results just named: The novelty lies in the integration, in the identifiability conditions, and in the distribution-scoped reading, rather than in the classical components themselves. One scope note is worth isolating, as it is not restated with the contributions: Item (2) claims no priority over the classical strict convexity of the multinomial-logistic/exponential-family log-likelihood, of which the

weight objective is a structured instance; what is ours is the explicit necessary-and-sufficient identifiability condition and the covariance-Hessian identity specialized to the reliability-weighted log-pool on the simplex.

Related literature. Multiplicative pooling of expert distributions originates with the PoE of Hinton [9], where renormalized products of latent-variable models yield sharp posteriors when experts agree; the older mixture-of-experts architecture of Jacobs et al. [10] combines experts additively through a gating network, the additive counterpart of the multiplicative rule used here. Our weights act as exponents in a log-linear pool rather than as gating probabilities, which is what makes the simplex geometry of Section 4 available. On the conformal side, the adaptive prediction sets of Romano et al. [11] and the regularized variant of Angelopoulos et al. [12] supplied the non-conformity score we adapt to the RW-PoE posterior in Section 6, with the general split-conformal guarantee as in [6, 7]. The present contribution is to couple the log-linear pool, its identifiability geometry, and the conformal layer into one estimator with explicit regularity conditions, rather than to propose a new pooling or conformal primitive. Throughout, $(\Omega, \mathcal{F}, \mathbb{P})$ denotes the underlying probability space, Y the latent label taking values in

$$\mathcal{Y} = \{1, \dots, K\} \quad \text{and} \quad X = (X_m)_{m \in \mathcal{M}}$$

the collection of modality observations, indexed by a finite set $\mathcal{M}$. In the case study

$$|\mathcal{M}| = 3 \quad \text{and} \quad \mathcal{M} = \{a, b, v\},$$

standing for acoustic, biological, and video. For a modality m , $p_m(\cdot | k)$ denotes the class-conditional density of X_m with respect to a dominating measure μ_m , and $p_m(k | x_m)$ the corresponding posterior. The probability simplex on K outcomes is

$$\Delta^{K-1} = \{\mathbf{p} \in \mathbb{R}_{\geq 0}^K : \sum_k p_k = 1\},$$

and the weight simplex is $\Delta^{|\mathcal{M}|-1}$. The total-variation distance is TV, the Bhattacharyya coefficient is BC, the squared Hellinger distance is

$$H^2 = 1 - \text{BC} \quad \text{and} \quad D_B = -\ln \text{BC}$$

is the Bhattacharyya distance. We write $\ell_i(k)$ for the vector of modality log-posteriors at observation i and label k , and $P_{\mathbf{w}}(\cdot | x)$ for the RW-PoE posterior under weights $\mathbf{w} \in \Delta^{|\mathcal{M}|-1}$. Symbols are defined at first use and used consistently thereafter; vectors are set in bold and their scalar components in lightface, so that

$$\boldsymbol{\pi} = (\pi_1, \dots, \pi_K), \quad \mathbf{w} = (w_1, \dots, w_M) \quad \text{and} \quad \ell_i(k) = (\log p_m(k | x_{i,m}))_{m \in \mathcal{M}}$$

are bold while π_k , w_m , and $v^\top \ell_i(k)$ are not; the weight of the type therefore always distinguishes a vector from one of its entries. Table 1 collects the principal symbols of Sections 1–3 (setup and information geometry), and Table 2 those of Sections 4–6 (estimation and the conformal layer); the two tables are split at the Section 4 boundary due to length and share a single numbering scheme.

Table 1. Principal notation, Sections 1–3 (setup and information geometry), with the section, definition, or equation in which each symbol is introduced.

Symbol	Meaning	First appears
$(\Omega, \mathcal{F}, \mathbb{P})$	probability space	Section 1.3
$Y \in \mathcal{Y} = \{1, \dots, K\}$	latent label; K classes	Section 1.3
$k \in \mathcal{Y}$	generic realization (value) of the label Y	Section 1.3
$\mathcal{M}, M = \mathcal{M} $	modality index set; $\mathcal{M} = \{a, b, v\}$, $M = 3$	Section 1.3
Δ^d	probability simplex in $\mathbb{R}^{d+1}$	Section 1.3
TV, BC	total variation; Bhattacharyya coefficient	Section 1.3
$H^2=1-\text{BC}$, $D_B=-\ln \text{BC}$	squared Hellinger; Bhattacharyya distance	Section 1.3
X_m, μ_m	modality- m observation; dominating measure	Def. 2.1
$x_m \in \mathcal{X}_m$	generic realization (value) of X_m	Def. 2.1
$\pi = (\pi_k)$	class prior on Δ^{K-1}	Def. 2.1
$p_m(x_m k)$, $p_m(k x_m)$	class-conditional density; posterior	Def. 2.1
R^* , $R_{\mathcal{M}'}^*$	Bayes risk; Bayes risk from modalities $\mathcal{M}'$	Def. 2.2
$A_v, \text{ch}(v)$, D_v	node label set; children; node decision	Def. 2.3
$c_k(x)$, $\delta(x)$, TC	copula ratio; log-spread; total correlation	Def. 2.4
$\text{BC}_{k^*}^{\text{pool}}$	pooled one-vs-rest coefficient	Def. 3.2
$R_{a,\text{root}}^*$	Bayes-risk floor at the binary root node	Rem. 3.2

Table 2. Principal notation, Sections 4–6 (estimation and the conformal layer), with the section, definition, or equation in which each symbol is introduced.

Symbol	Meaning	First appears
$\ell_i(k)$	vector of modality log-posteriors at (i, k)	Eq (4.4)
$\mathbf{w} \in \Delta^{M-1}$	reliability weights; $\widehat{\mathbf{w}}$ the MLE	Section 4.1
$P_{\mathbf{w}}(\cdot x)$	RW-PoE posterior under weights $\mathbf{w}$	Def. 4.1
$\alpha_0 \geq 1$	prior-strength exponent; $\alpha_0 = 1$ in Section 7	Def. 4.1
$Z(x; \mathbf{w})$	RW-PoE normalizer	Def. 4.1
$q_m(k x_m)$	alias for $p_m(k x_m)$ as a pooling input	Section 4.1
$\eta_{i,k}(\mathbf{w})$, $b_{i,k}$	pooled logit; prior offset	Eq (4.4)
$L(\mathbf{w})$, $\ell_{\text{obs}}(\mathbf{w})$	negative log-likelihood; observed log-lik.	Eqs (4.5), (5.1)
T, Φ, Φ_T	tangent space; centered design matrix; proj.	Rem. 4.2
$t_m, s_m^{(i)}, r_i^{(i)}(k)$	observed statistic; model expectation; fitted class probability	Section 5.2
γ	finite-step entropic mirror-step size	Eq (5.6)
h, B_h	negative-entropy mirror map; Bregman div.	Def. 5.1
$S_{w_{\min}}, \bar{L}, \underline{c}, \mu$	restricted region; curvature constants	Thm. 5.3
Ξ_*, H_T, Ξ_T	simplex metric and tangent Hessian matrices	Prop. 5.2
q_γ	γ -dependent local spectral factor	Eq (5.11)
$\mathcal{I}_{\text{obs}, T}$	tangent observed-information matrix	Eq (5.13)
$s(x, y)$, $\lambda_r, k_{\text{reg}}$	PoE regularized adaptive prediction sets (RAPS) score; parameters	Def. 6.1
$n_{\text{cal}}, \hat{q}_\alpha$	calibration fold size; conformal threshold	Def. 6.2
$C_\alpha(x)$	conformal prediction set at level α	Def. 6.2

The remainder of the paper is organized as follows. Section 2 sets up the probabilistic model, the PoE factorization (Theorem 2.1), the cascade entropy decomposition (Theorem 2.2), and the dependence-robustness bound (Proposition 2.3); Section 3 develops the Hellinger–Bayes necessity criterion; Section 4 defines the reliability-weighted PoE and proves convexity and uniqueness on the simplex; Section 5 develops finite-step GEM/entropic mirror weight learning with a monotonicity guarantee, a region-wise bound, the γ -dependent local linearization, and observed-information standard errors; Section 6 presents the distribution-free conformal layer; Section 7 reports the case study in full; and Section 8 records the single-site scope and the natural multi-site and federated extensions.

2. Problem formulation

2.1. Probability space and the trimodal model

Definition 2.1. (*Probability space and label–modality model*) Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space carrying a latent label $Y: \Omega \rightarrow \mathcal{Y} = \{1, \dots, K\}$ and a finite family of modality observations $X = (X_m)_{m \in \mathcal{M}}$, $X_m: \Omega \rightarrow \mathcal{X}_m$, indexed by a finite set $\mathcal{M}$ with $|\mathcal{M}| = M$. Write

$$\mathcal{X} = \prod_{m \in \mathcal{M}} \mathcal{X}_m$$

and let μ_m be a σ -finite dominating measure on $\mathcal{X}_m$, with product measure

$$\mu = \bigotimes_m \mu_m.$$

The label law is the prior $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K) \in \Delta^{K-1}$, $\pi_k = \mathbb{P}(Y = k) > 0$, and each modality admits a class-conditional density

$$p_m(x_m | k) = \frac{d\mathbb{P}(X_m | Y = k)}{d\mu_m}(x_m), \quad x_m \in \mathcal{X}_m, \quad k \in \mathcal{Y}.$$

In the case study of Section 7, $M = 3$ and $\mathcal{M} = \{a, b, v\}$ (acoustic, biological, video).

Notation. The two meanings of the modality symbol are separated by the order of the arguments, never by context alone: $p_m(x_m | k)$ always denotes the class-conditional *density* in x_m , whereas $p_m(k | x_m)$ always denotes the label *posterior*. The subscript m is kept at first use because from Section 2.5 onward several modalities occur in the same expression, where an unsubscripted $p(x | k)$ would not distinguish $p_a(x_a | k)$ from $p_b(x_b | k)$; where a single modality is in play the subscript is dropped, and in Section 3 we accordingly write

$$p(x_a) := p_a(x_a | k^*) \quad \text{and} \quad q(x_a) := p_a(x_a | j).$$

2.2. Conditional independence as a modeling assumption

Assumption 2.1. (*Class-conditional independence of modalities*) The modality observations are mutually independent conditionally on the label:

$$\mathbb{P}((X_m)_{m \in \mathcal{M}} \in \prod_m B_m | Y = k) = \prod_{m \in \mathcal{M}} \mathbb{P}(X_m \in B_m | Y = k), \quad (2.1)$$

equivalently, the class-conditional joint density factorizes as

$$p(x | k) = \prod_{m \in \mathcal{M}} p_m(x_m | k)$$

for μ -a.e. x .

Remark 2.1. (Status and scope of Assumption 2.1) Assumption 2.1 is a tractable modeling approximation, not a physical law, and residual dependence after conditioning on Y is expected in applications. Two facts limit its impact here. First, it is testable: all class-conditional Spearman correlations between the acoustic log-posteriors and the biological/video log-posteriors satisfy $|\rho_s| < 0.20$ across the $5 \times 2 \times 5 = 50$ class-stratified, cross-validated pairs involving the acoustic modality (five classes, excluding cold/hot, whose rank correlation is undefined on these folds; two acoustic pairs; five folds), with maximum $|\rho_s| = 0.19$ (class hungry, acoustic vs. video). The biological–video pair itself is larger ($|\rho_s| \leq 0.54$, mean 0.24 over 25 cells) and is the regime Section 2.6 quantifies rather than excludes. Second, the effect of any residual violation is not merely asserted but bounded: Section 2.6 (Proposition 2.3) controls the excess Bayes risk by the class-conditional total correlation, which Section 7.7 measures on a controlled Gaussian model. What matters is the label-heterogeneity of the dependence, not its magnitude, and both are small here. The exact-PoE posterior below is thus the reference object that Section 4 generalizes and corrects.

2.3. The PoE factorization

Theorem 2.1. (PoE factorization) Under Assumption 2.1, for μ -almost every $x \in \mathcal{X}$ with positive marginal density the posterior over labels factorizes as

$$\mathbb{P}(Y = k | X = x) = \frac{\pi_k \prod_{m \in \mathcal{M}} p_m(x_m | k)}{\sum_{j=1}^K \pi_j \prod_{m \in \mathcal{M}} p_m(x_m | j)}, \quad k \in \mathcal{Y}. \quad (2.2)$$

Equivalently, in logarithmic form,

$$\log \mathbb{P}(Y = k | X = x) = \log \pi_k + \sum_{m \in \mathcal{M}} \log p_m(x_m | k) - \log Z(x), \quad (2.3)$$

with normalizer $Z(x) = \sum_j \pi_j \prod_m p_m(x_m | j)$.

Proof. By Bayes' theorem, for x with $\sum_j \pi_j \prod_m p_m(x_m | j) > 0$,

$$\mathbb{P}(Y = k | X = x) = \frac{\pi_k p(x | Y = k)}{\sum_j \pi_j p(x | Y = j)},$$

where $p(x | Y = k) = d\mathbb{P}(X | Y = k)/d\mu$ exists by the product structure of μ and the existence of the marginal densities p_m . Assumption 2.1 gives $p(x | Y = k) = \prod_m p_m(x_m | k)$ for μ -a.e. x . Substituting yields (2.2); taking logarithms and collecting the k -independent terms into $\log Z(x)$ gives (2.3). $\square$

Remark 2.2. (Reduction and robustness) Equation (2.3) and the reliability-weighted operator of Section 4 share a log-additive product structure, but they use different factors: Theorem 2.1 multiplies class-conditional likelihoods with unit exponents, whereas Definition 4.1 pools modality-specific label

posteriors with simplex-normalized exponents. Thus, the latter is a discriminative reliability-weighted generalization, not a literal simplex-parameter special case of (2.3). Setting one reliability weight to zero removes the corresponding posterior expert, so the pooled construction is well defined under arbitrary modality missingness. The learned weights (Section 5) accommodate heterogeneous reliability or residual dependence, as anticipated in Remark 2.1.

2.4. Bayes classifier and risk

Definition 2.2. (Bayes classifier and Bayes risk) For a measurable classifier $\phi: \mathcal{X} \rightarrow \mathcal{Y}$, the misclassification risk is $R(\phi) = \mathbb{P}(\phi(X) \neq Y)$. The Bayes classifier $\phi^*(x) \in \arg \max_{k \in \mathcal{Y}} \mathbb{P}(Y = k \mid X = x)$ minimizes R over all measurable classifiers, and the Bayes risk is

$$R^* = R(\phi^*) = 1 - \mathbb{E} \left[\max_{k \in \mathcal{Y}} \mathbb{P}(Y = k \mid X = x) \right] = \inf_{\phi} R(\phi). \quad (2.4)$$

For a sub-collection $\mathcal{M}' \subseteq \mathcal{M}$ of modalities we write $R_{\mathcal{M}'}^*$ for the Bayes risk of the classifier that observes only $(X_m)_{m \in \mathcal{M}'}$. Two instances recur: the single-modality acoustic risk $R_a^* := R_{\{a\}}^*$ and the full-fusion risk $R_{\text{fusion}}^* := R_{\mathcal{M}}^*$, instantiated along a nested chain in Table 6 of Section 7.4.

Since enlarging the observed σ -algebra cannot increase the Bayes risk, $R_{\mathcal{M}_2}^* \leq R_{\mathcal{M}_1}^*$ whenever $\mathcal{M}_1 \subseteq \mathcal{M}_2 \subseteq \mathcal{M}$ (fresh index names, to avoid re-using $\mathcal{M}'$ of Definition 2.2); the strict-improvement direction is quantified information-theoretically in Section 3.

2.5. The probabilistic hierarchical cascade and an entropy-reduction theorem

The K -class decision is organized as a probabilistic hierarchical cascade (PHC): A rooted tree whose leaves are the classes and whose internal nodes are successive refinements of the label. This structure is convenient operationally and, as the next theorem shows, it yields an exact additive accounting of where the observed evidence reduces uncertainty about Y .

Definition 2.3. (PHC) Let T be a finite rooted tree with leaf set in bijection with $\mathcal{Y}$. For a node v let $A_v \subseteq \mathcal{Y}$ be the set of leaves of the subtree rooted at v , so that $A_{\text{root}} = \mathcal{Y}$, the sets $\{A_u : u \in \text{ch}(v)\}$ partition A_v for every internal node v (with $\text{ch}(v)$ its children), and each leaf is a singleton. A node v is reachable if $\mathbb{P}(Y \in A_v) > 0$. On the event $\{Y \in A_v\}$ define the node decision $D_v \in \text{ch}(v)$ to be the unique child u with $Y \in A_u$; thus D_v is a deterministic function of Y on $\{Y \in A_v\}$.

The tree is a modeling choice fixed before any data are observed. Because each leaf is a singleton, “leaf” and “class” are interchangeable; and because each D_v is a deterministic function of Y , the label law π of Definition 2.1 induces the law of every node decision. No symbol beyond those of Sections 2.1–2.4 is introduced.

Positioning. The decomposition of $H(Y)$ over a fixed label hierarchy follows from the classical entropy chain rule [13, Section 2.5]. Related information-decomposition constructions, such as Watanabe’s total correlation [14], allocate dependence among variables rather than across the nodes of a label hierarchy conditional on X . Theorem 2.2 instead decomposes $I(Y; X)$ across the reachable nodes of the hierarchy through the reach weights $\mathbb{P}(Y \in A_v)$.

Theorem 2.2. (PHC entropy decomposition and reduction) Let T be a probabilistic hierarchical cascade for Y and let X be any observation. Then, with all conditional entropies of D_v taken under the

conditional law $\mathbb{P}(\cdot \mid Y \in A_v)$ at each reachable internal node v ,

$$H(Y) = \sum_{v \text{ internal}} \mathbb{P}(Y \in A_v) H(D_v \mid Y \in A_v), \quad (2.5)$$

$$H(Y \mid X) = \sum_{v \text{ internal}} \mathbb{P}(Y \in A_v) H(D_v \mid X, Y \in A_v). \quad (2.6)$$

Consequently,

$$H(Y) - H(Y \mid X) = \sum_{v \text{ internal}} \mathbb{P}(Y \in A_v) I(D_v; X \mid Y \in A_v) \geq 0, \quad (2.7)$$

each summand being a nonnegative conditional mutual information. In particular $H(Y \mid X) \leq H(Y)$, with strict inequality if and only if $I(D_v; X \mid Y \in A_v) > 0$ for at least one reachable internal node v .

Proof sketch. A local entropy chain rule at each reachable internal node v unfolds recursively over the tree, the branch factors $\beta_u(x)$ telescoping to the reach probabilities $\mathbb{P}(Y \in A_v \mid X=x)$; a Bayes change-of-measure integrates out X , and subtracting the conditional from the unconditional decomposition yields Eq (2.7), each summand a nonnegative conditional mutual information. The complete proof is given in the supplementary material. $\square$

Remark 2.3. (Interpretation of Theorem 2.2) The reach weights in (2.7) localize the contribution of the observation to individual nodes of the hierarchy. Consequently, a modality may contribute little information at a high-reach node while remaining informative elsewhere. This distinction motivates the node-level necessity analysis of Section 3 and is examined empirically in Section 7.2.

2.6. How wrong may the independence assumption be?

Assumption 2.1 is an approximation. Its effect on the product posterior depends not only on the magnitude of class-conditional dependence, but on how that dependence varies across labels.

Definition 2.4. (Class-conditional copula spread) For $x = (x_m)_{m \in \mathcal{M}}$ write

$$c_k(x) = \frac{p(x \mid Y = k)}{\prod_{m \in \mathcal{M}} p_m(x_m \mid k)}, \quad \delta(x) = \log \max_{k \in \mathcal{Y}} c_k(x) - \log \min_{k \in \mathcal{Y}} c_k(x) \geq 0.$$

The ratio $c_k(x)$ is the class-conditional copula density: the factor by which the true within-class joint departs from the product of its margins. The spread $\delta(x)$ measures how much this factor varies across labels.

Writing the true and product posteriors side by side,

$$\frac{P(Y = k \mid x)}{P(Y = j \mid x)} = \frac{\pi_k \prod_m p_m(x_m \mid k)}{\pi_j \prod_m p_m(x_m \mid j)} \cdot \frac{c_k(x)}{c_j(x)},$$

so the entire effect of dependence on the decision is carried by the ratio $c_k(x)/c_j(x)$: If the copula does not depend on the label then every such ratio is 1 and the product posterior is *exact*.

Proposition 2.3. (Cost of ignoring label-dependent dependence) Let $\phi^\otimes$ be the plug-in classifier built from the product posterior $P^\otimes(\cdot \mid x) \propto \pi_k \prod_m p_m(x_m \mid k)$, and let $R^\star$ be the Bayes risk of the true (possibly dependent) law. Then:

- (i) Exactness. If $\delta \equiv 0$ then $P^\otimes(\cdot | x) = P(\cdot | x)$ for a.e. x and $R(\phi^\otimes) = R^*$.
- (ii) Pointwise. $\text{TV}(P(\cdot | x), P^\otimes(\cdot | x)) \leq \min\{1, \frac{1}{4}(e^{\delta(x)} - 1)\}$.
- (iii) Tail form. For every $t \geq 0$, $R(\phi^\otimes) - R^* \leq \frac{1}{2}(e^t - 1) + 2\mathbb{P}(\delta > t)$.
- (iv) Total-correlation form. $R(\phi^\otimes) - R^* \leq \sqrt{2\text{TC}}$, where $\text{TC} = \mathbb{E}_Y[\text{TC}(X | Y)]$ and $\text{TC}(X | Y = k) = \text{KL}(p(\cdot | k) \parallel \prod_m p_m(\cdot | k))$.

Proof. (i) If $\delta \equiv 0$ then $c_k(x)$ is constant in k ; the common factor cancels on normalization, so $P(\cdot | x) = P^\otimes(\cdot | x)$ a.e. and the induced Bayes rules, hence risks, coincide.

(ii) With

$$u_k = c_k(x)/Z(x), \quad Z(x) = \sum_j P^\otimes(j | x)c_j(x),$$

one has

$$P(k | x) = P^\otimes(k | x)u_k \quad \text{and} \quad \sum_k P^\otimes(k | x)u_k = 1;$$

since $c_k/c_j \in [e^{-\delta}, e^\delta]$ the multipliers satisfy $u_k \in [e^{-\delta}, e^\delta]$, and a short computation with the mean-one constraint bounds $\frac{1}{2} \sum_k P^\otimes(k | x)|u_k - 1|$ by $\min\{1, \frac{1}{4}(e^\delta - 1)\}$.

(iii) The plug-in regret satisfies $R(\phi^\otimes) - R^* \leq 2\mathbb{E}_x[\text{TV}(P, P^\otimes | x)]$ [5]. Splitting on $\{\delta \leq t\}$ and $\{\delta > t\}$ and using (ii) gives

$$\mathbb{E}[\text{TV}] \leq \frac{1}{4}(e^t - 1) + \mathbb{P}(\delta > t),$$

hence the claim.

(iv) By Pinsker,

$$\text{TV}(p(\cdot | k), \prod_m p_m(\cdot | k)) \leq \sqrt{\frac{1}{2}\text{TC}(X | Y = k)};$$

averaging over k , applying Jensen, and using the same plug-in inequality yields

$$R(\phi^\otimes) - R^* \leq 2\mathbb{E}_Y \sqrt{\frac{1}{2}\text{TC}(X | Y)} \leq \sqrt{2\text{TC}}.$$

This completes the proof. $\square$

Remark 2.4. (Why the total correlation, and not an exponential moment) Form (iv) is the certificate we adopt: the class-conditional total correlation is the Kullback–Leibler divergence between the true class-conditional law and the product of its margins, so it is the natural information-theoretic measure of how wrong Assumption 2.1 is, and it is estimable. The more elegant exponential-moment route is not used because $\mathbb{E}[e^\delta - 1]$ diverges for Gaussian modalities, so that certificate is vacuous in the model class in which Section 7.7 then measures.

3. Information-theoretic necessity of fusion

This section gives a formal sense in which a single modality can be provably insufficient. The development is entirely classical in its tools (the Bhattacharyya coefficient and the Hellinger–Bayes inequality of [5]); our contribution is to organize these tools into a *data-driven, distribution-scoped*

necessity criterion, stated with explicit limits of validity in Remark 3.3. Throughout, fix two labels $k^*, j \in \mathcal{Y}$, $k^* \neq j$, and write $p(x_a) := p_a(x_a | k^*)$, $q(x_a) := p_a(x_a | j)$ for the acoustic class-conditional densities.

3.1. Divergences and a Hellinger–Bayes lower bound

Definition 3.1. (*Bhattacharyya coefficient and Hellinger distance [15]*) For densities p, q with respect to μ_a ,

$$\text{BC}(p, q) = \int_{\mathcal{X}_a} \sqrt{p(x)q(x)} d\mu_a(x) \in [0, 1], \quad H^2(p, q) = 1 - \text{BC}(p, q), \quad (3.1)$$

and the Bhattacharyya distance is $D_B(p, q) = -\ln \text{BC}(p, q) \geq 0$. Total overlap ($\text{BC} = 1$) forces maximal confusion; disjoint supports ($\text{BC} = 0$) permit zero error.

We distinguish two acoustic overlap coefficients used below, as they are different quantities. The pairwise coefficient $\text{BC}(p_a(\cdot | k), p_a(\cdot | j))$ compares two single classes k, j (these are the entries reported in Figure 1). The pooled coefficient compares one index class against the mixture of all others.

Definition 3.2. (*Pooled one-vs-rest Bhattacharyya coefficient*) Fix an index class k^* with prior π_{k^*} . Form the pooled “rest” density as the prior-weighted mixture of the remaining class-conditional acoustic densities,

$$\bar{p}_{-k^*}(x) = \frac{1}{1 - \pi_{k^*}} \sum_{j \neq k^*} \pi_j p_a(x | j), \quad (3.2)$$

and define the pooled (one-vs-rest) coefficient $\text{BC}_{k^*}^{\text{pool}} = \text{BC}(p_a(\cdot | k^*), \bar{p}_{-k^*})$ and pooled distance $D_B^{\text{pool}} = -\ln \text{BC}_{k^*}^{\text{pool}}$. In general, $\text{BC}_{k^*}^{\text{pool}}$ is not a function of the pairwise coefficients $\{\text{BC}(p_a(\cdot | k^*), p_a(\cdot | j))\}_j$; it is the relevant quantity for the binary hunger node (index class versus the rest), whereas the pairwise coefficients describe the class-versus-class acoustic geometry. Two Bayes-risk floors derived from $\text{BC}_{k^*}^{\text{pool}}$ recur below and are named once here. The primary floor uses the actual empirical node prior $\alpha = \pi_{k^*}$ in the unequal-prior bound (3.4) and describes the problem as it is posed; the benchmark floor uses the conventional equal-prior value $\alpha = \frac{1}{2}$ in (3.5) and is reported only for comparability with the balanced classical statement of Lemma 3.1. For the hunger node ($\pi_{k^*} = 0.366$) these give $R_{a,\text{root}}^* \geq 0.235$ (primary) and $R_{a,\text{root}}^* \geq 0.262$ (benchmark), the two labeled rows of Table 3.

Lemma 3.1. (*Hellinger–Bayes lower bound [5, Chapter 3, Section 3.4, Theorem 3.1]*) For densities p, q with respect to μ_a ,

$$\text{TV}(p, q) \leq \sqrt{1 - \text{BC}(p, q)^2}. \quad (3.3)$$

More generally, for the two-class problem with priors α on p and $1 - \alpha$ on q ($\alpha \in (0, 1)$), the Bayes risk is

$$\begin{aligned} R_{\{p,q\}}^*(\alpha) &= \int_{\mathcal{X}_a} \min\{\alpha p, (1 - \alpha) q\} d\mu_a \\ &\geq \frac{1}{2} \left(1 - \sqrt{1 - 4\alpha(1 - \alpha) \text{BC}(p, q)^2} \right). \end{aligned} \quad (3.4)$$

The equal-prior case $\alpha = \frac{1}{2}$ recovers

$$R_{\{p,q\}}^*\left(\frac{1}{2}\right) = \frac{1}{2}(1 - \text{TV}(p, q)) \geq \frac{1}{2} \left(1 - \sqrt{1 - \text{BC}(p, q)^2} \right). \quad (3.5)$$

Proof sketch. Factor $p - q = (\sqrt{p} - \sqrt{q})(\sqrt{p} + \sqrt{q})$ and apply Cauchy–Schwarz, using

$$\int (\sqrt{p} \mp \sqrt{q})^2 d\mu_a = 2(1 \mp \text{BC});$$

the unequal-prior form follows from $\min\{u, v\} = \frac{1}{2}(u + v - |u - v|)$ with $u = \alpha p$, $v = (1 - \alpha)q$. The complete proof is given in the supplementary material. $\square$

In its equal-prior form the two-sided bound is classical [5, Chapter 3, Section 3.4, Theorem 3.1, pp. 23–24] (which the monograph attributes to earlier work of Kailath and of Hudimoto); the unequal-prior form (3.4) used in Definition 3.2, Theorem 3.2 and Table 3 is the elementary extension proved above, so no stronger claim is made for the source than it supports.

A worked closed-form instance (two Gaussians, showing what the floor does and does not say) is given in supplementary material, Example S1 and Figure S1.

3.2. The necessity theorem

The barrier established below is, by design, a statement about the *observed* acoustic distributions and the acoustic representation used to estimate them, and not a universal impossibility claim; the full scope is set out in Remark 3.3, stated immediately after the theorem to which it applies.

Theorem 3.2. (*Acoustic Bayes-risk barrier*) Let p, q be two acoustic densities with respect to μ_a (either two class-conditionals $p_a(\cdot | k^*)$, $p_a(\cdot | j)$, or an index class against the pooled mixture $\bar{p}_{-k^*}$ of Definition 3.2). Suppose $\text{BC}(p, q) \geq 1 - \varepsilon$ for some $\varepsilon \in (0, 1)$.

- (i) (**Two-point barrier, actual priors**) For the binary problem with priors α on p and $1 - \alpha$ on q , by Eq (3.4),

$$R_{\{p,q\}}^*(\alpha) \geq \frac{1}{2} \left(1 - \sqrt{1 - 4\alpha(1 - \alpha)(1 - \varepsilon)^2} \right). \quad (3.6)$$

Under the equal-prior convention $\alpha = \frac{1}{2}$, this reduces to the benchmark

$$R_{\{p,q\}}^*\left(\frac{1}{2}\right) \geq \frac{1}{2} (1 - \sqrt{\varepsilon(2 - \varepsilon)}) \rightarrow \frac{1}{2}$$

as $\varepsilon \rightarrow 0$.

- (ii) (**Acoustic K-class positivity**) The acoustic Bayes risk satisfies

$$R_a^* \geq (\pi_{k^*} + \pi_j) \tilde{R}_{\{k^*,j\}}^* > 0, \quad (3.7)$$

where $\tilde{R}_{\{k^*,j\}}^*$ is the Bayes risk of the binary problem conditional on $Y \in \{k^*, j\}$ (with renormalized priors). As $\varepsilon \rightarrow 0$, $\tilde{R}_{\{k^*,j\}}^* \rightarrow \min(\pi_{k^*}, \pi_j) / (\pi_{k^*} + \pi_j)$, so $R_a^* \rightarrow \min(\pi_{k^*}, \pi_j) > 0$.

Proof sketch. Part (i) substitutes $\text{BC} \geq 1 - \varepsilon$ into the unequal-prior bound (3.4); part (ii) restricts the Bayes classifier to $\{Y \in \{k^*, j\}\}$, where its conditional error is at least the binary Bayes risk $\tilde{R}_{\{k^*,j\}}^*$, and lets $\varepsilon \rightarrow 0$. The complete proof is given in the supplementary material. $\square$

3.3. Does fusion strictly help?

It is tempting to assert that any positive conditional mutual information $I(Y; (X_b, X_v) \mid X_a) > 0$ strictly lowers the Bayes risk. This is *not* correct in general: additional evidence can sharpen the posterior without ever moving its mode, leaving the Bayes risk unchanged. We therefore prove monotonicity and give the condition for strict improvement.

Proposition 3.3. (*Monotonicity and strict improvement*) Let $\mathcal{G}_1 = \sigma(X_a) \subseteq \mathcal{G}_2 = \sigma(X_a, X_b, X_v)$. Then $R_{\text{fusion}}^* \leq R_a^*$. Equality holds if and only if there exists a $\mathcal{G}_1$ -measurable label selection ψ such that

$$\psi \in \arg \max_{k \in \mathcal{Y}} \mathbb{P}(Y = k \mid \mathcal{G}_2) \quad \mathbb{P}\text{-almost surely.} \quad (3.8)$$

If both posterior modes are almost surely unique, this is equivalent to agreement of the $\mathcal{G}_1$ - and $\mathcal{G}_2$ -Bayes labels. Fusion strictly reduces Bayes risk precisely when no such common $\mathcal{G}_1$ -measurable maximizing selection exists.

Proof sketch. The tower property gives $\max_k \mathbb{P}(Y=k \mid \mathcal{G}_1) \leq \mathbb{E}[\max_k \mathbb{P}(Y=k \mid \mathcal{G}_2) \mid \mathcal{G}_1]$; taking expectations yields $R_{\text{fusion}}^* \leq R_a^*$, with equality precisely when a maximizer of $\mathbb{P}(Y=k \mid \mathcal{G}_2)$ admits a $\mathcal{G}_1$ -measurable selection a.s. — condition (3.8). The complete proof is given in the supplementary material. $\square$

Remark 3.1. (*Empirical confirmation of condition (3.8)*) Three observations from Section 7 show that the fitted fusion rule uses information beyond the acoustic channel and changes held-out predictions. They are consistent with failure of condition (3.8), but do not by themselves identify the unobserved population Bayes modes or prove strict Bayes-risk improvement. (i) The MLE $\widehat{\mathbf{w}} = (0.42, 0.31, 0.27)$ does not concentrate at the acoustic vertex $(1, 0, 0)$: $\widehat{\mathbf{w}}_b = 0.31 \pm 0.04$ and $\widehat{\mathbf{w}}_v = 0.27 \pm 0.03$ carry observed-information standard errors excluding zero (Proposition 5.2(iii)), supporting nonzero fitted contributions under the RW-PoE model. (ii) The acoustic mode is near-random at the root node ($\text{AUC}_{\text{audio, hunger}} = 0.560$, $p = 0.49$, DeLong [16]) while the trimodal Tier-1 selective accuracy is 95.8%. (iii) A McNemar test of the trimodal rule against the audio-only baseline gives $\chi^2_{\text{McNemar}} = 68.1$ ($p < 10^{-15}$), rejecting equality of the two fitted prediction rules on the test fold.

3.4. Numerical reading and scope

Remark 3.2. (*Numerical reading; an important distinction*) For the binary hunger node of the case study (Section 7), the index class is discriminated against the pooled mixture of the remaining classes, and the estimated pooled coefficient is $\text{BC}_{\text{hungry}}^{\text{pool}} \geq 0.88$, i.e., $\varepsilon \leq 0.12$ in Theorem 3.2. The hunger node is the binary problem {hungry, rest} with the actual renormalized priors $\alpha = \pi_{\text{hungry}} = 0.366$ and $W1 - \alpha = 0.634$. Substituting $\alpha = 0.366$ and $\text{BC}_{\text{hungry}}^{\text{pool}} = 0.88$ into the unequal-prior bound (3.6) gives, with $4\alpha(1 - \alpha) = 0.9282$,

$$R_{a, \text{root}}^* \geq \frac{1}{2} \left(1 - \sqrt{1 - 0.9282 \times 0.88^2} \right) = \frac{1}{2} (1 - 0.5303) = 0.2349 \approx 0.235. \quad (3.9)$$

Thus, 0.235 is a rigorous lower bound on the Bayes error of the optimal acoustic classifier at the hunger node under its true priors (3.6); the equal-prior benchmark is

$$\frac{1}{2} (1 - \sqrt{1 - 0.88^2}) \approx 0.262,$$

and we report the actual-prior value as primary. This bound is conceptually distinct from the error of any particular acoustic classifier; Remark 7.1 (Section 7.2) details that distinction once the empirical audio error is reported. Finally, the pooled coefficient $BC_{\text{hungry}}^{\text{pool}}$ of Definition 3.2 is a one-vs-rest quantity and is not determined by the pairwise class-versus-class coefficients of Figure 1 (whose smallest acoustic distance, hungry–discomfort, is $D_B = 0.29$, i.e., pairwise $BC \approx 0.75$); the two coefficients describe different comparisons.

Table 3 collects all four Bayes error lower bounds in one place for comparison. The primary bound uses the actual node prior $\alpha = 0.366$, whereas the benchmark uses the equal-prior convention $\alpha = \frac{1}{2}$. The bounds are not ordered by implication; they describe different aspects of the acoustic information geometry and must not be equated.

Table 3. Acoustic Bayes error lower bounds.

Comparison type	BC used	Prior convention	Formula	R^* lower bound
Pooled one-vs-rest (Def. 3.2, primary)	$BC^{\text{pool}} = 0.88$	Actual: $\alpha = 0.366$ $1 - \alpha = 0.634$	(3.6)	0.235 (primary)
Pooled one-vs-rest (Def. 3.2, benchmark)	$BC^{\text{pool}} = 0.88$	Equal: $\alpha = 0.5$	(3.5)	0.262 (benchmark)
Pairwise hungry–discomfort (Figure 1, $D_B = 0.29$)	$BC \approx 0.75$	Equal: $\alpha = 0.5$ $\varepsilon = 0.25$	(3.5) $\frac{1}{2}(1 - \sqrt{0.25 \times 1.75}) = \frac{1}{2}(1 - 0.661)$	≈ 0.170
Pairwise hungry–tired (Figure 1, $D_B = 0.31$)	$BC \approx 0.73$	Equal: $\alpha = 0.5$ $\varepsilon = 0.267$	(3.5) $\frac{1}{2}(1 - \sqrt{0.267 \times 1.733}) = \frac{1}{2}(1 - 0.680)$	≈ 0.160

Rows 3–4 are pairwise bounds, smaller than the pooled bound because the pairwise comparison isolates two classes while the pooled comparison aggregates all non-hungry classes; none of the four bounds implies any other.

Remark 3.3. (Scope: this is not a universal impossibility) Theorem 3.2 is deliberately and explicitly distribution-dependent.

(a) The bound is driven by the observed class-conditional acoustic distributions and the estimated pooled overlap $BC_{\text{hungry}}^{\text{pool}} \geq 0.88$; a population with different acoustic geometry (distinct recording conditions, vocal characteristics, or class taxonomy) could exhibit a different BC and hence a different floor.

(b) The bound applies to the specific acoustic feature representation used to estimate BC; a fundamentally different representation, or vastly more data, could in principle reduce the overlap.

(c) The result is a statement about the acoustic channel within the observed space; it neither asserts nor implies that acoustic discrimination of cry causes is impossible in general.

A population-level impossibility claim would require a minimax lower bound over a distributional class satisfying only the estimated constraints, which is a strictly stronger object and is not attempted here. What the theorem does establish is a rigorous, falsifiable, data-driven lower bound on the acoustic Bayes risk for the observed distributions, which, together with Proposition 3.3 and the

empirical gap of Section 7, explains why fusion is necessary here rather than merely convenient.

4. The RW-PoE

4.1. Definition and log-linear form

Write $q_m(k \mid x_m) = p_m(k \mid x_m)$ for the modality- m posterior over labels, viewed as a point of Δ^{K-1} , and let $\mathbf{w} = (w_m)_{m \in \mathcal{M}} \in \Delta^{M-1}$ be a vector of nonnegative *reliability weights* summing to one. The equality is an *alias*, not a second object: q_m and $p_m(\cdot \mid x_m)$ denote the same point of Δ^{K-1} . It is adopted only to match the pooling literature, which writes the operator inputs as $q_1, \dots, q_M$ and reserves p for the pooled output, so that $\mathcal{F}(q_1, \dots, q_M)$ in Proposition 4.1 does not collide with the p on the left of the centroid characterization (4.3).

Definition 4.1. (RW-PoE) For a prior $\pi \in \Delta^{K-1}$ and a prior-strength parameter $\alpha_0 \geq 1$, the RW-PoE posterior is

$$P_{\mathbf{w}}(Y = k \mid x) = \frac{1}{Z(x; \mathbf{w})} \pi_k^{\alpha_0-1} \prod_{m \in \mathcal{M}} p_m(k \mid x_m)^{w_m} \quad (4.1)$$

with normalizer

$$Z(x; \mathbf{w}) = \sum_{j=1}^K \pi_j^{\alpha_0-1} \prod_{m \in \mathcal{M}} p_m(j \mid x_m)^{w_m};$$

equivalently, in log-linear form,

$$\log P_{\mathbf{w}}(Y = k \mid x) = (\alpha_0 - 1) \log \pi_k + \sum_{m \in \mathcal{M}} w_m \log p_m(k \mid x_m) - \log Z(x; \mathbf{w}). \quad (4.2)$$

Setting $w_m = 0$ removes modality m (its factor becomes 1), so (4.1) remains a valid posterior under arbitrary modality missingness; on the observed set $\mathcal{O} \subseteq \mathcal{M}$ the deployed weights are $w_m^{(\mathcal{O})} = \widehat{w}_m / \sum_{m' \in \mathcal{O}} \widehat{w}_{m'}$, so the exponent budget is preserved and the pool is not implicitly cooled when a channel drops out. The symmetric choice $w_m \equiv 1/M$ recovers an unweighted logarithmic opinion pool. It should not be confused with the unit-exponent likelihood factorization of Theorem 2.1; the latter is the generative reference motivating the product structure. The parameter α_0 controls prior regularization: $\alpha_0 = 1$ (used throughout the case study of Section 7) removes the prior contribution entirely, recovering a pure log-linear pool; $\alpha_0 > 1$ introduces a prior-power factor that increases the influence of the fixed class prior π (and, as Proposition 4.3 (iii) shows, ultimately concentrates on its mode).

The role of α_0 , what it leaves invariant, what it costs, and how it behaves on the case study, is treated in full in Section 4.4 (Proposition 4.3) and Section 7.8.

4.2. Axiomatic grounding (classical)

The log-linear form (4.2) is not an arbitrary choice: it is the unique coherent way to pool the modality posteriors, and it admits a variational (weighted-KL centroid) characterization. Both facts are classical and are recalled here with attribution; neither is claimed as a contribution of this paper.

Proposition 4.1. (Coherence and weighted-KL centroid) Let $\mathcal{F}: (\Delta^{K-1})^M \rightarrow \Delta^{K-1}$ aggregate M label distributions into one, and assume $q_m(k) > 0$ for every $m \in \mathcal{M}$ and $k \in \mathcal{Y}$.

(a) (Genest–Zidek [1]) *Among operators satisfying the standard pooling axioms (in particular external Bayesianity), the unique form is the weighted logarithmic opinion pool $\mathcal{F}(q_1, \dots, q_M)(k) \propto \prod_m q_m(k)^{w_m}$, which coincides with (4.2).*

(b) (Abbas [2]) *This pool is the weighted Kullback–Leibler centroid of the inputs:*

$$\mathcal{F}(q_1, \dots, q_M) = \arg \min_{p \in \Delta^{K-1}} \sum_{m \in \mathcal{M}} w_m \text{KL}(p \parallel q_m). \quad (4.3)$$

Proof sketch. Part (a) is [1, Thm. 3]. For (b), the Karush–Kuhn–Tucker conditions of the strictly convex program $\min_{p \in \Delta^{K-1}} \sum_m w_m \text{KL}(p \parallel q_m)$ subject to $\sum_k p_k = 1$ give $\log p_k = \sum_m w_m \log q_m(k) + \text{const}$, i.e., $p_k \propto \prod_m q_m(k)^{w_m}$, the unique minimizer (4.3). The complete proof (the Lagrangian $\Lambda(p, \lambda)$, its stationarity condition, and the verification that the constraint $p_k \geq 0$ is inactive at the minimizer) is given in the supplementary material. $\square$

Attribution. The weighted-KL centroid identity (4.3) is due to Abbas [2]. Heskes [8] addressed the distinct problem of *selecting* the weights w_m by a Kullback–Leibler criterion and does not contain this centroid identity; it is therefore cited in Section 5.1, where weight estimation is treated, rather than here.

4.3. Strict convexity and uniqueness of the weights

We now turn to estimation. Given labeled data $\mathcal{D}_n = \{(x_i, y_i)\}_{i=1}^n$, the weights are fitted by maximum likelihood; we study the geometry of the associated objective. For each i and k , collect the modality log-posteriors into the vector

$$\ell_i(k) = (\log p_m(k \mid x_{i,m}))_{m \in \mathcal{M}} \in \mathbb{R}^M, \quad \eta_{i,k}(\mathbf{w}) = b_{i,k} + \mathbf{w}^\top \ell_i(k), \quad (4.4)$$

where $b_{i,k} = (\alpha_0 - 1) \log \pi_k$ gathers the $\mathbf{w}$ -independent prior offset, so that

$$P_{\mathbf{w}}(k \mid x_i) = e^{\eta_{i,k}(\mathbf{w})} / \sum_j e^{\eta_{i,j}(\mathbf{w})}.$$

The negative log-likelihood of the weights is

$$L(\mathbf{w}) = - \sum_{i=1}^n \log P_{\mathbf{w}}(y_i \mid x_i) = \sum_{i=1}^n \left[\log \sum_k e^{\eta_{i,k}(\mathbf{w})} - \eta_{i,y_i}(\mathbf{w}) \right]. \quad (4.5)$$

Definition 4.2. (*Log-posterior independence*) *The modalities are log-posterior independent on $\mathcal{D}_n$ if there is no nonzero $\mathbf{v} \in \mathbb{R}^M$ with $\sum_m v_m = 0$ such that, for every i , the scalar $\sum_m v_m \log p_m(k \mid x_{i,m}) = \mathbf{v}^\top \ell_i(k)$ is constant in $k \in \mathcal{Y}$.*

The restriction $\sum_m v_m = 0$ confines $\mathbf{v}$ to the tangent space of the weight simplex; a direction $\mathbf{v}$ violating Definition 4.2 is one along which moving the weights leaves every per-example posterior unchanged, hence a flat direction of L .

Theorem 4.2. (Strict convexity and unique minimizer) Assume $p_m(k \mid x_{i,m}) > 0$ for all i, k, m . Then, L in Eq (4.5) is convex on Δ^{M-1} , and its Hessian is the sum of posterior covariance matrices of the log-posterior vector:

$$\nabla^2 L(\mathbf{w}) = \sum_{i=1}^n \text{Var}_{k \sim P_{\mathbf{w}}(\cdot \mid x_i)}[\boldsymbol{\ell}_i(k)] \geq 0. \quad (4.6)$$

Here the covariance is taken over the random label $k \sim P_{\mathbf{w}}(\cdot \mid x_i)$, with $\boldsymbol{\ell}_i(k)$ the corresponding random vector; substituting the observed label y_i , whose $\boldsymbol{\ell}_i(y_i)$ is deterministic and has zero variance, is not intended. The summation over k hidden inside the variance operator is displayed explicitly in Eq (4.7) below. Moreover, L is strictly convex on Δ^{M-1} if and only if the modalities are log-posterior independent (Definition 4.2); in that case L has a unique global minimizer $\widehat{\mathbf{w}} \in \Delta^{M-1}$.

Proof. Convexity and Hessian. In (4.5) the terms $-\eta_{i,y_i}(\mathbf{w}) = -b_{i,y_i} - \mathbf{w}^\top \boldsymbol{\ell}_i(y_i)$ are affine in $\mathbf{w}$. Each $A_i(\mathbf{w}) = \log \sum_k e^{\eta_{i,k}(\mathbf{w})}$ is a log-sum-exp of functions affine in $\mathbf{w}$, hence convex; thus L is convex. Differentiating the log-sum-exp,

$$\nabla A_i(\mathbf{w}) = \sum_k P_{\mathbf{w}}(k \mid x_i) \boldsymbol{\ell}_i(k) = \mathbb{E}_{k \sim P_{\mathbf{w}}(\cdot \mid x_i)}[\boldsymbol{\ell}_i(k)].$$

Written out, so that the sum over $k \in \mathcal{Y}$ is explicit,

$$\begin{aligned} \nabla^2 A_i(\mathbf{w}) &= \sum_{k \in \mathcal{Y}} P_{\mathbf{w}}(k \mid x_i) \boldsymbol{\ell}_i(k) \boldsymbol{\ell}_i(k)^\top - \left(\sum_{k \in \mathcal{Y}} P_{\mathbf{w}}(k \mid x_i) \boldsymbol{\ell}_i(k) \right) \left(\sum_{k \in \mathcal{Y}} P_{\mathbf{w}}(k \mid x_i) \boldsymbol{\ell}_i(k) \right)^\top \\ &= \text{Var}_{k \sim P_{\mathbf{w}}(\cdot \mid x_i)}[\boldsymbol{\ell}_i(k)]. \end{aligned} \quad (4.7)$$

The affine terms have zero Hessian, so summing gives (4.6); each covariance matrix is positive semidefinite, hence $\nabla^2 L \geq 0$.

Strict convexity criterion. Movements within Δ^{M-1} are tangent directions $\mathbf{v} \in \mathbb{R}^M$ with $\mathbf{1}^\top \mathbf{v} = 0$. For such $\mathbf{v}$,

$$\mathbf{v}^\top \nabla^2 L(\mathbf{w}) \mathbf{v} = \sum_{i=1}^n \mathbf{v}^\top \text{Var}_{k \sim P_{\mathbf{w}}(\cdot \mid x_i)}[\boldsymbol{\ell}_i(k)] \mathbf{v} = \sum_{i=1}^n \text{Var}_{k \sim P_{\mathbf{w}}(\cdot \mid x_i)}[\mathbf{v}^\top \boldsymbol{\ell}_i(k)] \geq 0.$$

Because $p_m(k \mid x_{i,m}) > 0$, the responsibilities $P_{\mathbf{w}}(\cdot \mid x_i)$ have full support on $\mathcal{Y}$ for every $\mathbf{w}$ in the relative interior; thus $\text{Var}_{P_{\mathbf{w}}}[\mathbf{v}^\top \boldsymbol{\ell}_i(k)] = 0$ holds iff $k \mapsto \mathbf{v}^\top \boldsymbol{\ell}_i(k)$ is constant on all of $\mathcal{Y}$. Hence, $\mathbf{v}^\top \nabla^2 L(\mathbf{w}) \mathbf{v} = 0$ for some nonzero tangent $\mathbf{v}$ iff there is a nonzero $\mathbf{v}$ with $\mathbf{1}^\top \mathbf{v} = 0$ for which $\mathbf{v}^\top \boldsymbol{\ell}_i(k)$ is k -constant for every i , exactly the negation of Definition 4.2. Therefore, $\nabla^2 L(\mathbf{w}) > 0$ on the tangent space for all interior $\mathbf{w}$ (so L is strictly convex on Δ^{M-1}) if and only if the modalities are log-posterior independent. The criterion is uniform in $\mathbf{w}$ because full support holds throughout the interior.

Uniqueness. A strictly convex function on the nonempty compact convex set Δ^{M-1} attains its minimum (continuity and compactness) at a unique point $\widehat{\mathbf{w}}$ (strict convexity). This $\widehat{\mathbf{w}}$ is the maximum-likelihood weight vector computed by the finite-step recursion of Section 5. $\square$

Remark 4.1. (Interpretation and misspecification) By the Hessian identity (4.6), a modality whose log-posterior is nearly constant across labels (an uninformative channel, as for acoustics at the hunger node, Section 3) contributes little curvature and is weakly identified, whereas informative, non-redundant modalities sharpen the optimum. Log-posterior dependence, the failure mode of Definition 4.2, is the collinearity induced by redundant or perfectly correlated modalities. When

Definition 4.2 fails, there exists a nonzero $v \in T = \{v \in \mathbb{R}^M : \mathbf{1}^\top v = 0\}$ such that $L(w + tv) = L(w)$ whenever both points are feasible. Hence the tangent Hessian is singular and the likelihood cannot distinguish weights along feasible segments parallel to the corresponding kernel. If the minimizer lies on a nontrivial such segment, the MLE is nonunique and the finite-step recursion may select different points from different initializations. Boundary geometry can, however, leave an isolated minimizer even when strict convexity fails; failure of Definition 4.2 therefore establishes a flat, unidentified direction, not, by itself, the multiplicity of constrained minimizers. Thus, Theorem 4.2 guarantees a unique optimum under log-posterior independence and diagnoses loss of curvature under exact redundancy, the same issue anticipated in Remark 2.1.

Remark 4.2. (Sufficient condition for log-posterior independence) A verifiable algebraic sufficient condition for log-posterior independence (Definition 4.2) is as follows. Define the matrix $\Phi \in \mathbb{R}^{(nK) \times M}$ whose row indexed by (i, k) is the mean-centered log-posterior vector

$$\phi_{ik} = \ell_i(k) - \frac{1}{K} \sum_{j=1}^K \ell_i(j) \in \mathbb{R}^M.$$

Let Φ_T denote the projection of Φ onto the tangent space $T = \{v \in \mathbb{R}^M : \mathbf{1}^\top v = 0\}$, obtained by expressing each row in any orthonormal basis of T . If Φ_T has full column rank $M - 1$, then the modalities are log-posterior independent.

Rank deficiency of Φ_T is equivalent to the existence of a nonzero tangent $v \in T$ with $v^\top \ell_i(k)$ constant in k for every i , the negation of Definition 4.2, so the condition is checkable directly on the training data (a $663 \times 6 \times 2$ array projected to two tangent coordinates here). Rank deficiency can arise from any exact tangent linear dependence among the centered modality log-posteriors; pairwise proportional experts are only one special case. Pairwise Spearman correlations do not certify this rank condition, which is why Φ_T is checked directly in Section 7.3.

Remark 4.3. (Practical reading of Definition 4.2) Three points bear on whether log-posterior independence holds in practice. (i) The condition constrains fitted posteriors, not raw modalities. Definition 4.2 concerns the log-posterior vectors $\ell_i(k)$, not the observations X_m : correlated inputs need not produce log-posterior-dependent scores, and independent inputs fail the condition once their classifiers are functionally related (e.g., one distilled from the other), so the condition neither implies nor is implied by Assumption 2.1. (ii) Genericity. Φ_T has $nK = 663 \times 6 = 3978$ rows and $M - 1 = 2$ columns in the case study; rank deficiency of a 3978×2 matrix is a measure-zero event under any absolutely continuous score law, so failure signals a structural cause, not sampling noise; an audit of the pipeline, not a statistical test. (iii) Exact structural dependence is the relevant issue. Near-redundancy may make the curvature small without making it zero. As one illustrative, non-exhaustive mechanism, suppose $M = 3$ and modality 2 is a temperature rescaling of modality 1, $\log p_2(k | x_{i,2}) = \tau \log p_1(k | x_{i,1}) + c_i$ with c_i independent of k . For a tangent v with $\mathbf{1}^\top v = 0$,

$$v^\top \ell_i(k) = (v_1 + \tau v_2) \log p_1(k | x_{i,1}) + v_3 \log p_3(k | x_{i,3}) + v_2 c_i.$$

If modalities 1 and 3 are not themselves affinely related in k modulo k -constants, this is k -constant for every i only if $v_1 + \tau v_2 = 0$ and $v_3 = 0$; with $v_1 + v_2 + v_3 = 0$ this forces $v_2(\tau - 1) = 0$, hence $v = 0$ unless $\tau = 1$. Under the stated non-affine condition on modalities 1 and 3, this particular

temperature-rescaling mechanism creates a flat tangent direction only at exact duplication ($\tau = 1$). Other exact multi-expert linear dependencies are possible; the full-rank criterion for Φ_T , rather than a list of examples, is decisive.

4.4. The regularization parameter α_0

Proposition 4.3. (Role of α_0) Fix $\alpha_0 \geq 1$ and assume $p_m(k \mid x_{i,m}) > 0$ and $\pi_k > 0$ for all i, k, m . Then:

- (i) Identifiability is α_0 -free. $L(\cdot; \alpha_0)$ is convex on Δ^{M-1} for every α_0 , and strictly convex iff the modalities are log-posterior independent (Definition 4.2); the criterion does not involve α_0 .
- (ii) Conformal validity is α_0 -free. For every α_0 fixed in advance, the lower coverage guarantee (6.4) holds; the upper bound in (6.5) additionally requires no ties, since Theorem 6.1 constrains the provenance of the score, not its predictive quality.
- (iii) What α_0 costs. If the prior mode $k_\pi = \arg \max_k \pi_k$ is unique, then as $\alpha_0 \rightarrow \infty$, $P_{\mathbf{w}}(\cdot \mid x) \rightarrow \delta_{k_\pi}$ uniformly in $\mathbf{w} \in \Delta^{M-1}$; consequently

$$\nabla^2 L(\mathbf{w}) = \sum_i \text{Var}_{k \sim P_{\mathbf{w}}(\cdot \mid x_i)}[\ell_i(k)] \rightarrow 0$$

and the weights become unidentified in the limit.

Proof. (i) The offset $b_{i,k} = (\alpha_0 - 1) \log \pi_k$ of (4.4) does not depend on $\mathbf{w}$ and so contributes nothing to $\nabla^2 L$; and $\pi_k > 0$ ensures the responsibilities retain full support on $\mathcal{Y}$, the only property of $P_{\mathbf{w}}$ used in the proof of Theorem 4.2. (ii) Immediate from the hypotheses of Theorem 6.1, with its no-ties assumption retained when the upper bound is invoked. (iii) For $k \neq k_\pi$,

$$\frac{P_{\mathbf{w}}(k \mid x)}{P_{\mathbf{w}}(k_\pi \mid x)} = \left(\frac{\pi_k}{\pi_{k_\pi}} \right)^{\alpha_0 - 1} \prod_{m \in \mathcal{M}} \left(\frac{p_m(k \mid x_m)}{p_m(k_\pi \mid x_m)} \right)^{w_m}.$$

Since $\sum_m w_m = 1$, the second factor is a weighted geometric mean of the M ratios and lies in a positive compact range fixed by the data and independent of $\mathbf{w}$; the first factor tends to 0 geometrically. Boundedness of ℓ_i then yields the Hessian statement. $\square$

$\alpha_0 - 1$ is the coefficient of the fixed log-prior term: increasing it strengthens the prior ranking and, by (iii), simultaneously flattens the weight likelihood, so the two effects run in opposite directions and no regime makes raising α_0 free. It should therefore be selected by cross-validated log-loss on the training fold alone. Departing from $\alpha_0 = 1$ may be considered when prior regularization is substantively justified; it is not implied by a sample-size ratio alone. In the case study $n_{\text{train}}/K = 663/6 = 110.5$, and $\alpha_0 = 1$ was fixed in advance. The empirical sweep is Table 8 of Section 7.8.

5. Finite-step GEM learning of reliability weights

5.1. Objective and score equation

The reliability weights are estimated by maximum likelihood,

$$\hat{\mathbf{w}} = \arg \max_{\mathbf{w} \in \Delta^{M-1}} \ell_{\text{obs}}(\mathbf{w}), \quad \ell_{\text{obs}}(\mathbf{w}) = \sum_{i=1}^n \log P_{\mathbf{w}}(y_i \mid x_i) = -L(\mathbf{w}), \quad (5.1)$$

with L the negative log-likelihood (4.5). By Theorem 4.2, under log-posterior independence ℓ_{obs} is strictly concave on Δ^{M-1} , so $\widehat{\mathbf{w}}$ exists and is unique. Differentiating (5.1) gives the score

$$\nabla \ell_{\text{obs}}(\mathbf{w}) = \sum_{i=1}^n \left(\ell_i(y_i) - \mathbb{E}_{k \sim P_{\mathbf{w}}(\cdot | x_i)}[\ell_i(k)] \right), \quad (5.2)$$

so, for the interior MLE considered in Proposition 5.2, the maximum-likelihood condition must be imposed on the simplex tangent space rather than by setting the full score vector to zero. If $E \in \mathbb{R}^{M \times (M-1)}$ has orthonormal columns spanning $T = \{v \in \mathbb{R}^M : \mathbf{1}^\top v = 0\}$, the correct constrained moment-matching identity is

$$E^\top \sum_{i=1}^n \left\{ \ell_i(y_i) - \mathbb{E}_{k \sim P_{\widehat{\mathbf{w}}}(\cdot | x_i)}[\ell_i(k)] \right\} = 0. \quad (5.3)$$

Equivalently, $\nabla L(\widehat{\mathbf{w}}) = \kappa \mathbf{1}$ for some scalar κ . Thus, the MLE equalizes the empirical and model-expected log-posterior statistics in every feasible (tangent) direction; componentwise equality is neither required nor generally true under the simplex constraint. At a boundary optimum the corresponding statement is the usual KKT normal-cone condition, not Eq (5.3).

Heskes [8] studied the related but distinct problem of selecting the pooling weights by minimizing a Kullback–Leibler criterion; the maximum-likelihood route (5.1) adopted here differs in both objective and estimator.

5.2. Finite-step GEM/entropic mirror-descent recursion

We minimize the observed negative log-likelihood $L = -\ell_{\text{obs}}$ directly on the simplex. Write $t_m = \sum_i \log p_m(y_i | x_{i,m})$ for the observed log-posterior statistic of modality m . At iteration t , evaluate the fitted class probabilities $r_i^{(t)}(k) = P_{\mathbf{w}^{(t)}}(k | x_i)$ and the corresponding model expectations

$$s_m^{(t)} = \sum_{i=1}^n \sum_{k=1}^K r_i^{(t)}(k) \log p_m(k | x_{i,m}) = \sum_{i=1}^n \mathbb{E}_{k \sim P_{\mathbf{w}^{(t)}}(\cdot | x_i)}[\log p_m(k | x_{i,m})]. \quad (5.4)$$

Equation (5.2) gives the component identity

$$t_m - s_m^{(t)} = [\nabla \ell_{\text{obs}}(\mathbf{w}^{(t)})]_m = -[\nabla L(\mathbf{w}^{(t)})]_m. \quad (5.5)$$

Let $h(\mathbf{w}) = \sum_m w_m \log w_m$ be the negative-entropy mirror map. Its Bregman divergence is $B_h(\mathbf{u}, \mathbf{v}) = \sum_m u_m \log(u_m/v_m)$. One mirror step is

$$\begin{aligned} \mathbf{w}^{(t+1)} &= \arg \min_{\mathbf{w} \in \Delta^{M-1}} \left\{ \langle \nabla L(\mathbf{w}^{(t)}), \mathbf{w} \rangle + \frac{1}{\gamma} B_h(\mathbf{w}, \mathbf{w}^{(t)}) \right\}, \\ w_m^{(t+1)} &= \frac{w_m^{(t)} \exp(\gamma(t_m - s_m^{(t)}))}{\sum_{j=1}^M w_j^{(t)} \exp(\gamma(t_j - s_j^{(t)}))}, \quad m = 1, \dots, M, \end{aligned} \quad (5.6)$$

where $\gamma > 0$ is the step size. Thus, Eq (5.6) is exactly an entropic mirror-descent, or exponentiated-gradient, update for the *observed* objective. It preserves positivity and the simplex constraint whenever $\mathbf{w}^{(0)}$ is interior.

Remark 5.1. (*GEM terminology and the map analyzed below*) The expectation $s_m^{(t)}$ gives the recursion an EM-like, or GEM-type, form. For finite γ , however, (5.6) is not an exact EM M -step. Every convergence statement in this section therefore concerns the explicit finite-step map (5.6); no exact-EM missing-information Jacobian is attributed to Algorithm 1. Proposition 5.1 proves likelihood ascent directly from the mirror step, and Proposition 5.2 differentiates that same map.

Algorithm 1 Finite-step GEM / entropic mirror descent for RW-PoE weight estimation.

Require: data $\mathcal{D}_n$; interior $\mathbf{w}^{(0)}$; fixed step $0 < \gamma \leq 1/L_\Delta$ (Proposition 5.1)

Require: tolerance $\varepsilon_{\text{tol}} = 10^{-6}$; cap $T_{\text{max}} = 200$

Ensure: $\widehat{\mathbf{w}}$

- 1: precompute $t_m \leftarrow \sum_i \log p_m(y_i | x_{i,m})$ for each m ▷ observed sufficient statistics
 - 2: **repeat**
 - 3: **Expectation/score evaluation:** $r_i^{(t)}(k) \leftarrow P_{\mathbf{w}^{(t)}}(k | x_i); \quad s_m^{(t)} \leftarrow \sum_i \sum_k r_i^{(t)}(k) \log p_m(k | x_{i,m})$ ▷ cost $O(nKM)$
 - 4: **Entropic mirror step:** $\mathbf{w}_m^{(t+1)} \leftarrow \mathbf{w}_m^{(t)} \exp(\gamma(t_m - s_m^{(t)}))$; renormalize over m ▷ cost $O(M)$
 - 5: **until** $\|\mathbf{w}^{(t+1)} - \mathbf{w}^{(t)}\|_2 < \varepsilon_{\text{tol}}$ **or** $t \geq T_{\text{max}}$
 - 6: **return** $\widehat{\mathbf{w}} \leftarrow \mathbf{w}^{(t+1)}$
-

5.3. Monotone ascent of the finite-step update

Proposition 5.1. (*Likelihood ascent and convergence of the finite-step map*) Assume $p_m(k | x_{i,m}) > 0$ for all i, k, m , and let

$$L_\Delta = \sup_{\mathbf{w} \in \Delta^{M-1}} \lambda_{\text{max}}^T(\nabla^2 L(\mathbf{w})) < \infty, \quad T = \{\mathbf{v} \in \mathbb{R}^M : \mathbf{1}^\top \mathbf{v} = 0\}. \quad (5.7)$$

For an interior initialization and a fixed step $0 < \gamma \leq 1/L_\Delta$, the update (5.6) remains interior and satisfies

$$L(\mathbf{w}^{(t+1)}) \leq L(\mathbf{w}^{(t)}) - \frac{1}{\gamma} B_h(\mathbf{w}^{(t)}, \mathbf{w}^{(t+1)}) - \left(\frac{1}{\gamma} - L_\Delta\right) B_h(\mathbf{w}^{(t+1)}, \mathbf{w}^{(t)}) \leq L(\mathbf{w}^{(t)}). \quad (5.8)$$

Equivalently, $\ell_{\text{obs}}(\mathbf{w}^{(t+1)}) \geq \ell_{\text{obs}}(\mathbf{w}^{(t)})$, with strict inequality unless $\mathbf{w}^{(t)}$ is stationary. If, in addition, log-posterior independence holds and the iterates stay in some $\mathcal{S}_{w_{\min}} = \{\mathbf{w} \in \Delta^{M-1} : w_m \geq w_{\min} \forall m\}$ with $w_{\min} > 0$, then $\mathbf{w}^{(t)} \rightarrow \widehat{\mathbf{w}}$, the unique MLE.

Proof. The complete proof, including the relative-smoothness inequality, strict-ascent criterion, and limit-point argument, is given in supplementary material, Section S1.6. □

5.4. Local convergence of the finite-step update; observed-information standard errors

Proposition 5.2. (*Local rate of the finite-step GEM/entropic-mirror map*) Assume log-posterior independence and that the unique MLE $\widehat{\mathbf{w}}$ is in the relative interior of Δ^{M-1} . Let

$$\mathcal{M}_\gamma(\mathbf{w}) = \frac{\mathbf{w} \odot \exp\{-\gamma \nabla L(\mathbf{w})\}}{\mathbf{1}^\top [\mathbf{w} \odot \exp\{-\gamma \nabla L(\mathbf{w})\}]}, \quad \gamma > 0, \quad (5.9)$$

where products and exponentials are componentwise. This is exactly the finite-step update (5.6). Let

$$T = \{\mathbf{v} \in \mathbb{R}^M : \mathbf{1}^\top \mathbf{v} = 0\}, \quad \Xi_\star = \text{Diag}(\widehat{\mathbf{w}}) - \widehat{\mathbf{w}}\widehat{\mathbf{w}}^\top,$$

and let $E \in \mathbb{R}^{M \times (M-1)}$ have orthonormal columns spanning T . Write

$$H_T = E^\top \nabla^2 L(\widehat{\mathbf{w}}) E, \quad \Xi_T = E^\top \Xi_\star E.$$

Then:

- (i) (Jacobian of the implemented update) *The constrained first-order condition gives $\nabla L(\widehat{\mathbf{w}}) = \kappa \mathbf{1}$ for some scalar κ , and hence $\mathcal{M}_\gamma(\widehat{\mathbf{w}}) = \widehat{\mathbf{w}}$. The Jacobian of the actual finite-step map, restricted to the simplex tangent space, is*

$$J_{\gamma,T} := E^\top D\mathcal{M}_\gamma(\widehat{\mathbf{w}}) E = I_{M-1} - \gamma \Xi_T H_T. \quad (5.10)$$

- (ii) (γ -dependent local rate) *Both H_T and Ξ_T are positive definite. If $0 < \nu_1 \leq \dots \leq \nu_{M-1}$ are the eigenvalues of the symmetric positive-definite matrix $B_\star = \Xi_T^{1/2} H_T \Xi_T^{1/2}$, then the eigenvalues of $J_{\gamma,T}$ are $1 - \gamma \nu_j$, and the local spectral factor is*

$$q_\gamma = \rho(J_{\gamma,T}) = \max_{1 \leq j \leq M-1} |1 - \gamma \nu_j|. \quad (5.11)$$

In particular, $q_\gamma < 1$ precisely when

$$0 < \gamma < \frac{2}{\nu_{M-1}}, \quad (5.12)$$

and under this condition $\widehat{\mathbf{w}}$ is locally attracting and the finite-step iteration converges locally linearly with spectral factor q_γ .

- (iii) (Observed-information standard-error approximation) *Direct differentiation of the observed likelihood gives the tangent observed information*

$$\mathcal{I}_{\text{obs},T}(\widehat{\mathbf{w}}) = H_T = E^\top \left\{ \sum_{i=1}^n \text{Var}_{k \sim P_{\widehat{\mathbf{w}}}(\cdot|x_i)} [\ell_i(k)] \right\} E > 0. \quad (5.13)$$

Under independent-observation likelihood regularity, the usual observed-information covariance estimate under the simplex constraint is $\widehat{\text{Cov}}(\widehat{\mathbf{w}}) = E H_T^{-1} E^\top$, and the associated Wald standard-error estimate is $\widehat{\sigma}_{w_m} = [E H_T^{-1} E^\top]_{mm}^{1/2}$. These are asymptotic plug-in estimates, not exact finite-sample covariances.

Proof. The complete derivation, KKT fixed-point identity, quotient-rule differentiation, tangent-space reduction, spectral analysis, and constrained covariance formula, is given in supplementary material, Section S1.7. $\square$

For comparison, the classical missing-information Jacobian concerns a *distinct exact-EM operator* whose M -step maximizes its surrogate [3]. It is not the Jacobian of the finite-step map $\mathcal{M}_\gamma$ in (5.9); its spectral radius is therefore not a convergence rate for Algorithm 1.

We do not infer a numerical convergence factor from an empirical stopping count. Numerical evaluation of the local factor requires the implemented step size and loss-normalization convention; the mathematically identified factor is the γ -dependent quantity (5.11).

5.5. A non-asymptotic linear rate for the entropic mirror update

Proposition 5.2 gives the exact local linearization of the finite-step map for any fixed γ , but it applies only near $\widehat{\mathbf{w}}$. For a certified finite-iteration guarantee from an arbitrary initialization in $\mathcal{S}_{w_{\min}}$, Theorem 5.3 analyzes the same entropic mirror construction on that restricted region using the relative-smoothness framework of Lu et al. [17]. The results are complementary: Proposition 5.2 yields the γ -dependent asymptotic factor q_γ , whereas Theorem 5.3 supplies a region-wise non-asymptotic bound for the certified step $\gamma = 1/L$.

Definition 5.1. (Negative-entropy mirror map and relative smoothness/convexity) Let $h(\mathbf{w}) = \sum_{m \in \mathcal{M}} w_m \log w_m$ on Δ^{M-1} , with Bregman divergence $B_h(\mathbf{w}', \mathbf{w}) = \sum_m w'_m \log(w'_m/w_m)$ (the Kullback–Leibler divergence on the simplex). For constants $0 < \mu \leq L$, the objective $L(\cdot)$ of (4.5) is L -smooth relative to h and μ -strongly convex relative to h on a convex region $\mathcal{S} \subseteq \Delta^{M-1}$ if, for all $\mathbf{w}, \mathbf{w}' \in \mathcal{S}$,

$$\mu B_h(\mathbf{w}', \mathbf{w}) \leq L(\mathbf{w}') - L(\mathbf{w}) - \langle \nabla L(\mathbf{w}), \mathbf{w}' - \mathbf{w} \rangle \leq L B_h(\mathbf{w}', \mathbf{w}). \quad (5.14)$$

Theorem 5.3. (Explicit non-asymptotic linear rate) Assume log-posterior independence (Definition 4.2) and $p_m(k | x_{i,m}) > 0$ for all i, k, m . Fix $w_{\min} \in (0, 1/M]$ and restrict to the compact region $\mathcal{S}_{w_{\min}} = \{\mathbf{w} \in \Delta^{M-1} : w_m \geq w_{\min} \ \forall m\}$, assumed to contain the minimizer $\widehat{\mathbf{w}}$. Define, on the simplex tangent space $T = \{\mathbf{v} \in \mathbb{R}^M : \mathbf{1}^\top \mathbf{v} = 0\}$,

$$\bar{L} = \sup_{\mathbf{w} \in \mathcal{S}_{w_{\min}}} \lambda_{\max}^T(\nabla^2 L(\mathbf{w})), \quad \underline{c} = \inf_{\mathbf{w} \in \mathcal{S}_{w_{\min}}} \lambda_{\min}^T(\nabla^2 L(\mathbf{w})), \quad (5.15)$$

where $\lambda_{\min}^T, \lambda_{\max}^T$ are the smallest and largest eigenvalues of the positive semidefinite Hessian

$$\nabla^2 L(\mathbf{w}) = \sum_i \text{Var}_{k \sim P_{\mathbf{w}}(\cdot | x_i)}[\ell_i(k)]$$

of Theorem 4.2 restricted to T . Then $L(\cdot)$ is L -smooth and μ -strongly convex relative to h on $\mathcal{S}_{w_{\min}}$ with the explicit constants

$$L = \bar{L}, \quad \mu = w_{\min} \underline{c}, \quad 0 < \mu \leq L, \quad (5.16)$$

and $\underline{c} > 0$. Consequently the entropic-mirror update with feasible set restricted to $\mathcal{S}_{w_{\min}}$ and step $\gamma = 1/L$ (that is, $\mathbf{w}^{(t+1)} = \arg \min_{\mathbf{w} \in \mathcal{S}_{w_{\min}}} \{\langle \nabla L(\mathbf{w}^{(t)}), \mathbf{w} \rangle + L B_h(\mathbf{w}, \mathbf{w}^{(t)})\}$) satisfies, for every $t \geq 1$,

$$L(\mathbf{w}^{(t)}) - L(\widehat{\mathbf{w}}) \leq L \left(1 - \frac{\mu}{L}\right)^t B_h(\widehat{\mathbf{w}}, \mathbf{w}^{(0)}), \quad 1 - \frac{\mu}{L} = 1 - \frac{w_{\min} \underline{c}}{\bar{L}} \in [0, 1). \quad (5.17)$$

Proof. The full five-step proof, including the Loewner–Bregman integration and the four-stage contraction argument, is given in supplementary material, Section S1.8. $\square$

Remark 5.2. (Local and non-asymptotic factors) The factor $1 - \mu/L$ in Theorem 5.3 is a region-wise finite-iteration certificate. By contrast, q_γ in Proposition 5.2 is the exact spectral factor of the fixed-point linearization at $\widehat{\mathbf{w}}$. Their empirical evaluation requires the same loss normalization and the recorded step size; the corresponding computation and certification protocols are given in supplementary material, Sections S1.7 and S1.8, respectively.

Remark 5.3. (*What the weights do and do not measure*) (a) Imbalanced quality. By the Hessian identity (4.6), the information the data carry about w_m is the posterior variance of that modality's log-posterior; a channel whose log-posterior varies little across labels contributes little curvature and is weakly identified. The symptom is a large observed-information standard error, not a systematically small weight. A modality can be both uninformative and heavily weighted if nothing in the likelihood pins its exponent down. (b) A weight is a global, not a node-level, statistic. The acoustic weight is the largest, $\widehat{w}_a = 0.42 \pm 0.03$, even though the acoustic channel is statistically indistinguishable from chance at the root hunger node (AUC = 0.560, $p = 0.49$): the weight is fitted on the whole training fold, and the acoustic channel is informative elsewhere in the cascade (node B: AUC = 0.798). A large $\widehat{w}_m$ certifies aggregate usefulness, never uniform reliability across the tree, which is the argument for per-node weights recorded in Section 8. (c) A label-constant channel does not break the fit: if $\ell_{i,m}(k)$ is constant in k its contribution cancels in the softmax, and w_m becomes a temperature rather than a reliability. (d) Missing modalities. By Definition 4.1, on the observed set $O \subseteq \mathcal{M}$ the deployed weights are $w_m^{(O)} = \widehat{w}_m / \sum_{m' \in O} \widehat{w}_{m'}$, so the exponent budget is preserved and the pool is not implicitly cooled when a channel drops out; the conformal guarantee survives provided the triples $(X, Y, \text{missingness pattern})$ are exchangeable between calibration and test.

6. A distribution-free conformal layer

The estimator of Sections 4 and 5 produces a fitted posterior $P_{\widehat{w}}(\cdot \mid x)$, but a point posterior carries no finite-sample guarantee on its own. We add a split-conformal layer [6, 7] that converts the posterior into a set-valued predictor with a finite-sample, distribution-free coverage guarantee under exchangeability. The construction and its guarantee are classical; we state them for the RW-PoE score and give the full proof via rank uniformity, so the framework is self-contained.

6.1. The PoE-RAPS non-conformity score

Partition the data into a training fold (used to fit $\widehat{w}$, Section 5), a calibration fold of size n_{cal} , and a test fold. The score below is the RAPS score of [12], building on the adaptive prediction sets of [11], applied to the RW-PoE posterior.

Definition 6.1. (*PoE-RAPS score*) Fix the fitted posterior $P_{\widehat{w}}(\cdot \mid x)$. For an input x , let $(k_{(1)}, k_{(2)}, \dots, k_{(K)})$ order the labels by decreasing posterior mass, $P_{\widehat{w}}(k_{(1)} \mid x) \geq \dots \geq P_{\widehat{w}}(k_{(K)} \mid x)$, and let $\text{rank}(k \mid x)$ be the position of label k in this order. With regularization parameters $\lambda_r \geq 0$ and $k_{\text{reg}} \in \{1, \dots, K\}$, the non-conformity score of the pair (x, y) is

$$s(x, y) = \sum_{j=1}^{\text{rank}(y \mid x)} P_{\widehat{w}}(k_{(j)} \mid x) + \lambda_r (\text{rank}(y \mid x) - k_{\text{reg}})_+, \quad (6.1)$$

where $(t)_+ = \max(t, 0)$. The unregularized case $\lambda_r = 0$ recovers the adaptive prediction set (APS) cumulative score; $\lambda_r > 0$ penalizes deep ranks and controls set size.

Definition 6.2. (*Conformal quantile and prediction set*) Let $s_1, \dots, s_{n_{\text{cal}}}$ be the scores (6.1) on the calibration fold, and for level $\alpha \in (0, 1)$ define the conformal threshold

$$\hat{q}_\alpha = \left\lceil (1 - \alpha)(n_{\text{cal}} + 1) \right\rceil\text{-th smallest of } \{s_1, \dots, s_{n_{\text{cal}}}, +\infty\}. \quad (6.2)$$

The appended $+\infty$ is the usual finite-sample convention; it makes the definition also valid when the requested rank is $n_{\text{cal}} + 1$, without changing the quantile at the levels used in the case study. The conformal prediction set for a test input x is

$$C_\alpha(x) = \{k \in \mathcal{Y} : s(x, k) \leq \hat{q}_\alpha\}. \quad (6.3)$$

6.2. Finite-sample coverage

Theorem 6.1. (*Distribution-free conformal coverage*) Suppose the calibration pairs $(X_1, Y_1), \dots, (X_{n_{\text{cal}}}, Y_{n_{\text{cal}}})$ and the test pair $(X_{\text{te}}, Y_{\text{te}})$ are exchangeable, and the scoring function s is fixed (it depends on the data only through the training fold, which is disjoint from calibration and test). Then the set C_α of (6.3) satisfies the marginal lower bound

$$\mathbb{P}(Y_{\text{te}} \in C_\alpha(X_{\text{te}})) \geq 1 - \alpha. \quad (6.4)$$

If, in addition, the joint calibration-and-test scores are almost surely distinct (no ties; e.g., the score has a continuous distribution), then

$$1 - \alpha \leq \mathbb{P}(Y_{\text{te}} \in C_\alpha(X_{\text{te}})) \leq 1 - \alpha + \frac{1}{n_{\text{cal}} + 1}. \quad (6.5)$$

Proof sketch. Exchangeability of the $n_{\text{cal}} + 1$ scores makes the rank of the test score (sub)uniform; the inclusion

$$\{s_{\text{te}} \leq \hat{q}_\alpha\} \supseteq \{\text{rank}(s_{\text{te}}) \leq \lceil (1 - \alpha)(n_{\text{cal}} + 1) \rceil\}$$

gives the lower bound (6.4), and almost surely distinct scores give the matching upper bound. The complete proof is given in the supplementary material. $\square$

Remark 6.1. (*Case-study instantiation and the coverage band*) With $n_{\text{cal}} = 220$ and $\alpha = 0.10$, the theoretical coverage band (6.5) is

$$\left[1 - \alpha, 1 - \alpha + \frac{1}{n_{\text{cal}} + 1} \right] = \left[0.900, 0.900 + \frac{1}{221} \right] = [0.900, 0.905]$$

(to three decimals). The empirical test-fold coverage is 0.910 ($n_{\text{test}} = 222$), within sampling error of the band; the fraction of singleton (decisive, “Tier-1”) prediction sets is 85.2%. These values are reported in Section 7. The exact upper bound requires the no-ties condition, so $[0.900, 0.905]$ is reported as a conditional theoretical band. If score ties or atoms are present, the lower guarantee remains valid; exact finite-sample randomization requires a randomized conformal construction.

Remark 6.2. (*Scope of the conformal coverage guarantee*) The lower guarantee (6.4), and the upper bound in (6.5) when its no-ties condition holds, are marginal over the test distribution and apply to any fixed score, however the posterior was fitted; misspecification of the RW-PoE model affects the size of C_α (sharper posteriors give smaller sets) but not the coverage level, which depends only on exchangeability. It does not by itself deliver conditional (per-input) coverage, nor does it transfer across a distribution shift between calibration and test folds, a limitation relevant to the single-site scope discussed in Section 8.

Remark 6.3. (*Coverage under distribution shift*) The guarantee does not collapse abruptly when exchangeability fails; it degrades linearly. Let C_α be constructed from training and calibration folds drawn from a law P on $\mathcal{X} \times \mathcal{Y}$, and let the test pair follow Q instead. Then

$$Q(Y_{\text{te}} \in C_\alpha(X_{\text{te}})) \geq 1 - \alpha - \text{TV}(P, Q).$$

Indeed, conditionally on the training and calibration folds, C_α is a fixed set-valued map, so $A = \{(x, y) : y \in C_\alpha(x)\}$ is a fixed measurable event. Hence

$$Q(A) \geq P(A) - \text{TV}(P, Q).$$

Taking expectations over the calibration fold and applying Theorem 6.1 to the P -term gives the claim. The bound therefore degrades linearly with the total-variation shift and becomes vacuous when $\text{TV}(P, Q) \geq \alpha$. This limitation motivates the site-stratified and covariate-shift remedies discussed in Section 8.

7. Numerical case study

We illustrate the framework on a single-site trimodal classification problem: The PIDD infant-distress dataset collected at CHU Sidi Bel-Abbès. The study is presented as a *case study*. It demonstrates the estimator end-to-end and supplies the numbers referenced in Sections 3–6, while the inferential scope is the observed single-site distribution, whose limitations are stated in Section 8. All reported quantities are produced by the procedures of Sections 4–6 with hyperparameters fixed in advance ($\alpha_0 = 1$, $\alpha = 0.10$, $\varepsilon_{\text{tol}} = 10^{-6}$, bootstrap replicates $B = 2000$). With $\alpha_0 = 1$, the prior log-weight term $(\alpha_0 - 1) \log \pi_k$ vanishes identically from Eq (4.2), so the RW-PoE posterior (4.1) reduces to the unregularized log-linear pool

$$\log P_{\mathbf{w}}(Y=k \mid x) = \sum_m w_m \log p_m(k \mid x_m) - \log Z(x; \mathbf{w}),$$

and all reported weights, coverage, and accuracy figures correspond to this unregularized form. The case $\alpha_0 > 1$ increases the influence of the fixed class prior π ; its effect is quantified in Section 7.8.

7.1. Data and protocol

What the problem is. The case study concerns classification of infant-cry causes from three complementary sources of information: acoustic features, biological context, and video-derived behavioral descriptors.

The dataset is the PIDD corpus assembled under ethical approval from the CHU Sidi Bel-Abbès ethics committee (reference: CHU-SBA-2025-017) in collaboration with the neonatology and pediatric units of the Centre Hospitalier Universitaire de Sidi Bel-Abbès, Algeria. It comprises $n = 1105$ labeled cry episodes, each carrying three modalities, an acoustic signal (a), a biological context vector (b), and a video-derived behavioral descriptor (v), and a ground-truth label in the $K = 6$ class taxonomy $\mathcal{Y} = \{\text{hungry, tired, discomfort, cold/hot, belly pain, burping}\}$. Accurate identification of cry cause is clinically consequential: misreading a pain or feeding cry as ordinary distress can delay intervention, while over-triage burdens neonatal units. An uncertainty-aware, set-valued output with a stated marginal coverage guarantee is therefore useful at the point of care; this is the role of the conformal layer in Section 6.

Modality descriptions. All three modalities are real clinical measurements:

- *Acoustic* ($X_a, p_a \approx 425$ features). A total of 1105 cry episodes (duration 2–45 s) were recorded in standard clinical rooms at 44.1 kHz using a calibrated Sony ECM-CS10 microphone at 30 cm from the infant’s face.
- *Biological context* ($X_b \in \mathbb{R}^{15}$). Caregiver-provided clinical metadata recorded via a structured questionnaire administered by trained nursing staff at the moment of each cry episode. The 15 features cover five clinical groups: feeding context (time since last feed, feeding type, caregiver-rated hunger score), thermal state (axillary and environmental temperature), activity (body-movement intensity, prior cry frequency, cry duration), medical history (reflux indicator, medication status, hydration status, recent illness), and developmental factors (corrected gestational age, weight percentile, time of day).
- *Video* ($X_v \in \mathbb{R}^{50}$). A total of 2,390 video clips (6.6 h at 25 fps) were captured with a Sony HDR-CX405 camcorder and processed offline: facial action units (AU4, AU6, AU12, AU17, AU20, AU23) were extracted via OpenFace 2.0, and pose kinematics (limb velocity, leg retraction frequency) via MediaPipe Pose, yielding a $p_v = 50$ -dimensional descriptor per episode.

Raw recordings are not publicly shared in accordance with clinical data governance and patient privacy requirements (CHU ethics approval CHU-SBA-2025-017); the feature matrices and fitted posteriors used in the reproducible case study are available from the corresponding author upon reasonable request.

Ground-truth labels. Labels were assigned independently by two pediatricians (inter-rater $\kappa = 0.81$); disagreements were resolved by a third senior pediatrician. This coefficient is reported as a measure of the reliability of the labeling process and is not used anywhere as a baseline for the classifier.

Data partition. The episodes are partitioned once, by label-stratified sampling, into

$$n_{\text{train}} = 663, \quad n_{\text{cal}} = 220, \quad n_{\text{test}} = 222, \quad (7.1)$$

so that $663 + 220 + 222 = 1105$. The training fold fits the modality-specific posteriors and the reliability weights (Section 5); the calibration fold sets the conformal threshold (Section 6); the test fold is used once, for the reported metrics. The prevalence of the index “hungry” class is $\pi_{\text{hungry}} = 0.366$. The class taxonomy is organized as the probabilistic hierarchical cascade of Definition 2.3: a root hunger node (hungry vs. the rest), a tired/active node, and a discomfort-subtype node.

7.2. Acoustic information geometry and the necessity of fusion

Figure 1 shows the pairwise acoustic Bhattacharyya distances $D_B = -\ln \text{BC}$ between classes. The “hungry” row carries the smallest distances (the acoustic *barrier*): The nearest neighbors are discomfort ($D_B = 0.29$, $\text{BC} \approx 0.75$) and tired ($D_B = 0.31$, $\text{BC} \approx 0.73$), confirming severe acoustic overlap of hunger with other states.

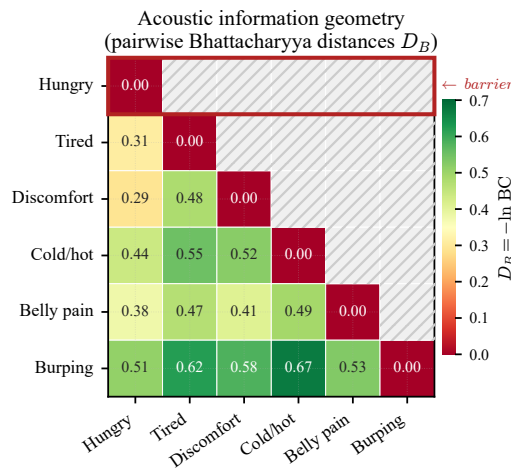


Figure 1. Pairwise acoustic Bhattacharyya distances $D_B = -\ln BC$ among the six classes. Larger D_B (greener) means more separable; the hungry row (boxed) has the smallest distances, the acoustic “barrier” that motivates fusion.

The barrier is confirmed empirically at the level of the audio-only classifier. Figure 2 (receiver operating characteristic (ROC) curves) and Figure 3 (node confusion matrices) show that at the root hunger node the acoustic channel is statistically indistinguishable from chance,

$$\text{AUC}_{\text{audio, hunger}} = 0.560, \quad p = 0.49 \text{ (DeLong test [16])}, \quad (7.2)$$

with node accuracy 0.568 (Table 4), whereas the tired/active node is acoustically easier ($\text{AUC} = 0.798$). The right panel of Figure 2 shows the finite-step GEM/entropic-mirror learned RW-PoE weights. Its annotations give the estimates and standard errors, and the whiskers give asymptotic observed-information Wald 95% intervals. The likelihood-ratio test rejects equal weights ($\Lambda = 47.3$, $p < 10^{-10}$).

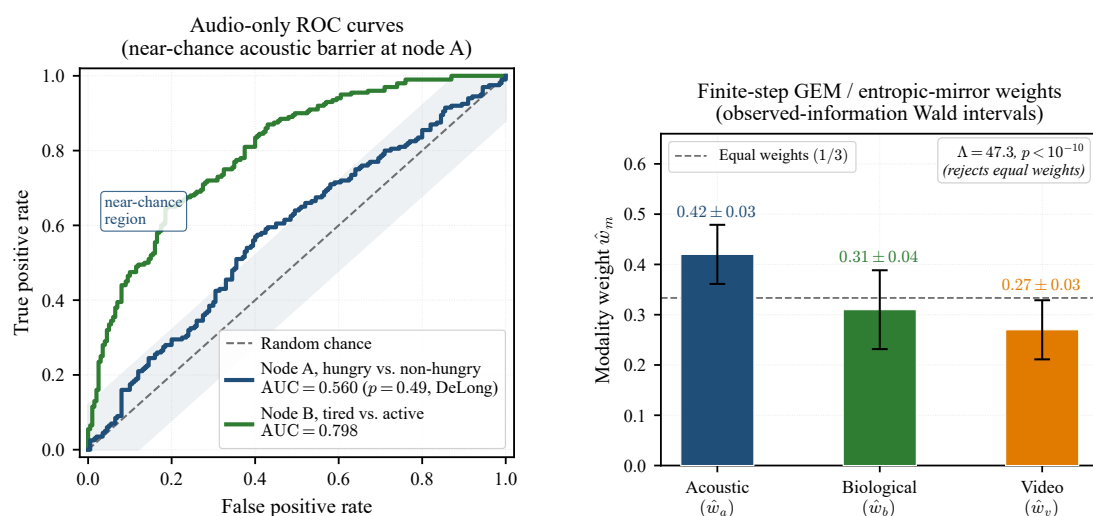


Figure 2. Audio-only ROC curves (left) and learned RW-PoE weights (right).

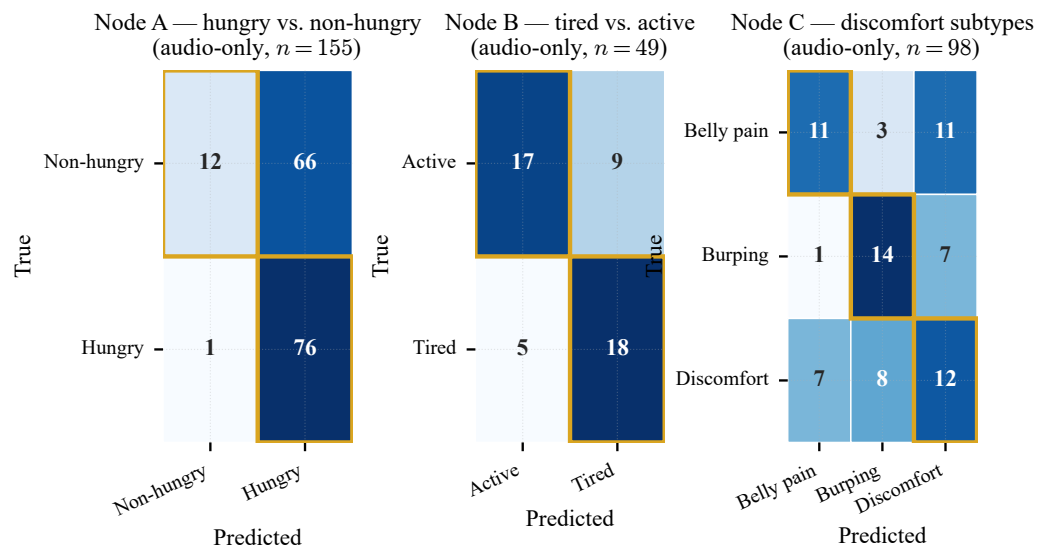


Figure 3. Audio-only confusion matrices at the cascade nodes: (a) hungry vs. non-hungry ($n = 155$; accuracy 0.568, near chance), (b) tired vs. active ($n = 49$), (c) discomfort subtypes ($n = 98$). Near-chance performance at node (a) is the empirical signature of the acoustic barrier.

Table 4. Audio-only discrimination by cascade node (test fold). AUC at the binary nodes with DeLong p -value against AUC = 0.5.

Node	n	Accuracy	AUC (p)
A: hungry vs. non-hungry	155	0.568	0.560 (0.49)
B: tired vs. active	49	0.714	0.798 (< 0.01)
C: discomfort subtypes	98	3-class; see Figure 3c	

Remark 7.1. (*The acoustic floor: theory versus observed error*) Two distinct quantities should be kept separate. The rigorous Hellinger–Bayes lower bound (3.6) on the best-possible acoustic error is, at the pooled one-vs-rest overlap $\text{BC}_{\text{hungry}}^{\text{pool}} = 0.88$ (Definition 3.2) with the actual hunger-node priors $(\alpha, 1 - \alpha) = (0.366, 0.634)$, $\frac{1}{2}(1 - \sqrt{1 - 4\alpha(1 - \alpha)0.88^2}) \approx 0.235$ (the equal-prior convention gives the benchmark ≈ 0.262); at the level of pairwise class comparisons (Figure 1) the analogous per-pair bounds are smaller still. The empirical audio error at the hunger node is $1 - 0.568 = 0.432$, i.e., near chance, consistent with AUC = 0.560. These are different objects: a lower bound on the error of the optimal acoustic classifier versus the observed error of a near-chance one. The empirical error lies above the rigorous floor because the audio classifier is far from Bayes-optimal at this node. The annotation $R_{\text{emp}}^* \approx 0.432$ in Figure 5 denotes this empirical near-chance audio error, not the Hellinger–Bayes floor. The primary rigorous floor is $R_{a,\text{root}}^* \geq 0.235$ (Remark 3.2, Table 3); it, the equal-prior benchmark 0.262, the smaller pairwise per-class bounds, and the empirical near-chance error $R_{\text{emp}}^* \approx 0.432$ measure different quantities and must not be equated.

7.3. Learned weights

The fitted reliability weights with observed-information standard errors (5.13) (Figure 2, right) are

$$\widehat{\mathbf{w}} = (\underbrace{0.42 \pm 0.03}_{\text{acoustic}}, \underbrace{0.31 \pm 0.04}_{\text{biological}}, \underbrace{0.27 \pm 0.03}_{\text{video}}). \quad (7.3)$$

The weights are recovered entirely from data; the corresponding two-sided 95% Wald intervals exclude 0, supporting nonzero fitted contributions from all three modalities. The acoustic weight remains the largest despite the near-chance audio performance at the hunger node ($\text{AUC} = 0.560$, Eq (7.2)) because acoustics is informative at *non-root* cascade nodes (Remark 2.3, Table 4), not at the hunger split.

Empirical identifiability check (Definition 4.2). On the training fold ($n_{\text{train}} = 663$), the centered 3978×2 tangent design matrix Φ_T of Remark 4.2 has full column rank, $\text{rank}(\Phi_T) = 2 = M - 1$. This directly verifies the log-posterior-independence condition for the reported fit and hence the uniqueness conclusion of Theorem 4.2. We deliberately do not reinterpret a pointwise Hessian eigenvalue as either the local factor q_γ or a region-wise constant: those quantities require, respectively, the recorded step size and the full calculation prescribed in Proposition 5.2, or certified bounds over $\mathcal{S}_{w_{\min}}$ as in Theorem 5.3.

7.4. Classification performance

Figures 4 (Cohen's κ forest plot) and 5 (accuracy and macro- F_1) compare modality configurations on the test fold; Table 5 collects the numerical values. The trimodal RW-PoE attains accuracy 0.916, Cohen's $\kappa = 0.889$ (95% CI 0.844–0.934), and macro- $F_1 = 0.910$, improving on every uni- and bimodal configuration. Relative to the audio-only baseline (0.568 accuracy), trimodal fusion adds 34.8 percentage points.

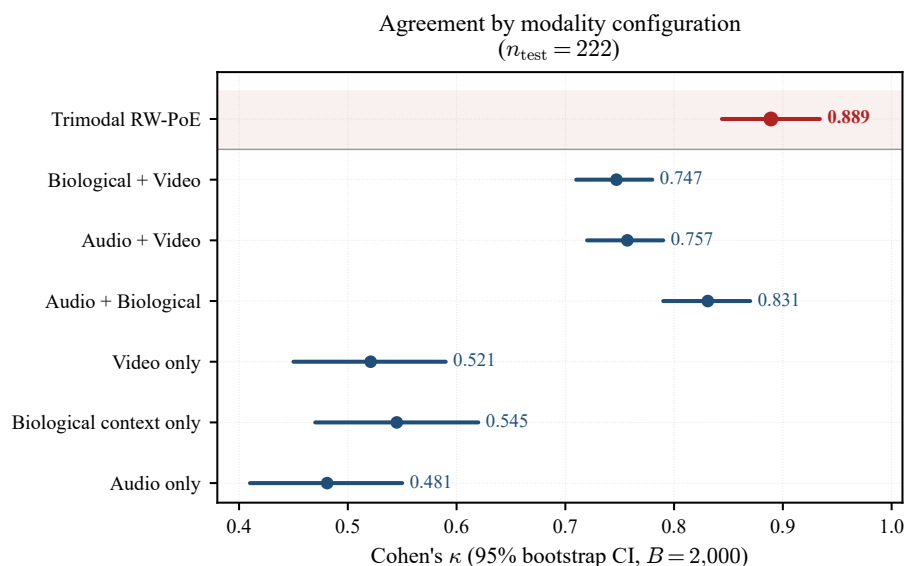


Figure 4. Cohen's κ by modality configuration with 95% bootstrap confidence intervals ($B = 2000$, $n_{\text{test}} = 222$). The trimodal RW-PoE reaches $\kappa = 0.889$.

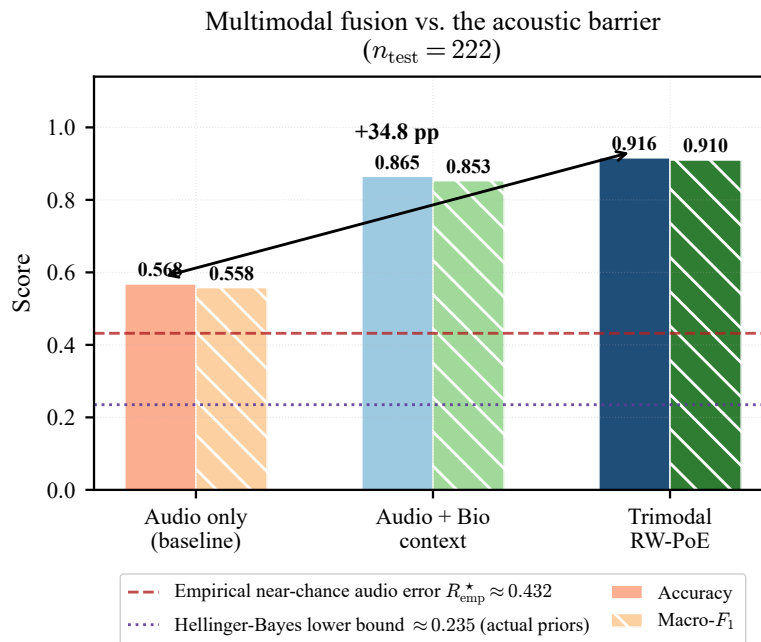


Figure 5. Accuracy and macro- F_1 for audio-only, audio+biological, and trimodal RW-PoE configurations. The marked level near 0.43 denotes the empirical near-chance audio error $R_{\text{emp}}^* \approx 0.432$ ($= 1 - 0.568$, the audio-only accuracy), which is the observed error of a near-chance classifier and is distinct from the rigorous Hellinger–Bayes lower bound (≈ 0.235 at the actual hunger-node priors); see Remark 7.1.

Table 5. Test-fold performance by modality configuration ($n_{\text{test}} = 222$). Cohen’s κ from Figure 4; accuracy and macro- F_1 are reported for the configurations documented in Figure 5 (others marked “–”). The trimodal κ carries a 95% bootstrap CI of 0.844–0.934.

Configuration	Accuracy	macro- F_1	κ
Audio only	0.568	0.558	0.481
Biological only	–	–	0.545
Video only	–	–	0.521
Audio + Video	–	–	0.757
Biological + Video	–	–	0.747
Audio + Biological	0.865	0.853	0.831
Trimodal RW-PoE	0.916	0.910	0.889

The nested-chain instantiation of the sub-collection risk $R_{\mathcal{M}}^*$ of Definition 2.2 (Table 6) gives the monotonicity of Section 2.4 an empirical counterpart. The entries are test-fold estimates of the risks of fitted plug-in classifiers $\widehat{R}(\phi_{\mathcal{M}})$. The corresponding population classifier risks, not their finite test estimates, are upper bounds on $R_{\mathcal{M}}^*$. That fitted classifiers on a nested chain respect the Bayes-risk ordering is a property of these data, not a theorem. The theoretical floor $R_{a,\text{root}}^* \geq 0.235$ is deliberately

kept out of this table because it is a binary root-node quantity, whereas the column holds six-class error rates (Remark 7.1).

Table 6. Fitted-classifier errors along a nested modality chain; $\widehat{R}(\phi_{\mathcal{M}'})$ is the six-class test error, standard errors binomial ($n_{\text{test}} = 222$).

$\mathcal{M}'$	Configuration	$\widehat{R}(\phi_{\mathcal{M}'})$	s.e.	κ
$\{a\}$	acoustic only	0.432	0.033	0.481
$\{a, b\}$	acoustic + biological	0.135	0.023	0.831
$\{a, b, v\}$	trimodal (full $\mathcal{M}$)	0.084	0.019	0.889

7.5. Conformal coverage and deployment tiers

With $\alpha = 0.10$ and $n_{\text{cal}} = 220$, Theorem 6.1 guarantees marginal coverage at least 0.900; under its additional no-ties condition, the coverage probability lies in

$$\left[1 - \alpha, 1 - \alpha + \frac{1}{n_{\text{cal}} + 1} \right] = [0.900, 0.905].$$

Figure 6 shows the empirical coverage tracking the band across α ; at $\alpha = 0.10$ the realized coverage is 0.910. The right panel of Figure 6 shows the induced three-tier deployment profile (Tier-1 singleton selective accuracy 95.8%), and Table 7 collects the values.

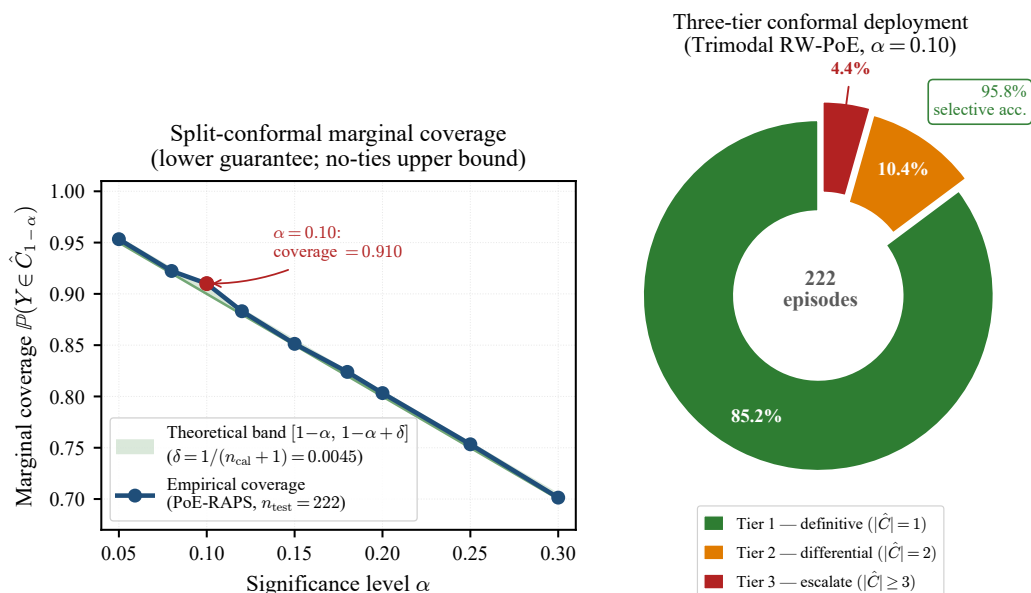


Figure 6. Left: empirical conformal coverage versus the lower guarantee $1 - \alpha$ and the no-ties upper discretization bound $1 - \alpha + \delta$, $\delta = 1/(n_{\text{cal}} + 1) = 0.0045$ (Theorem 6.1); at $\alpha = 0.10$ the coverage is 0.910. Right: three-tier prediction-set profile, singleton (Tier 1, 85.2%), pair (Tier 2, 10.4%), escalate (Tier 3, 4.4%).

Table 7. Split-conformal performance on the test fold ($\alpha = 0.10$, $n_{\text{cal}} = 220$, $n_{\text{test}} = 222$); lower guarantee 0.900 and no-ties theoretical band $[0.900, 0.905]$ from Theorem 6.1.

Quantity	Value
Theoretical coverage band	$[0.900, 0.905]$
Empirical test coverage	0.910
Tier 1 (singleton, $ \hat{C} = 1$)	85.2%
Tier 2 (pair, $ \hat{C} = 2$)	10.4%
Tier 3 (escalate, $ \hat{C} \geq 3$)	4.4%
Tier-1 selective accuracy	95.8%
Mean set size	1.19

The empirical coverage 0.910 sits just above the no-ties probability band, well within ordinary test-fold sampling variation of order $1/\sqrt{n_{\text{test}}}$ (Remark 6.2); the large singleton fraction shows the sets are typically decisive, and the Tier-1 selective accuracy of 95.8% indicates that decisive predictions are also accurate.

7.6. Statistical tests

Two prediction-level tests on held-out data and one training-likelihood comparison are reported. A McNemar test comparing the trimodal classifier with the audio-only baseline on paired test predictions gives

$$\chi^2_{\text{McNemar}} = 68.1, \quad p < 10^{-15}; \quad (7.4)$$

a likelihood-ratio test of the equal-weight null $\mathbf{w} = (1/M, \dots, 1/M)$ against the free simplex-weight model (two degrees of freedom here), using the training likelihood, gives

$$\Lambda_{\text{LRT}} = 47.3, \quad p < 10^{-10}; \quad (7.5)$$

and the audio-only hunger-node discrimination is indistinguishable from chance, $\text{AUC} = 0.560$, and $p = 0.49$ (7.2). The bootstrap confidence intervals in Figure 4 use $B = 2000$ resamples.

7.7. Robustness to modality dependence

A controlled synthetic Gaussian stress test ($N = 40,000$, $K = 6$, $M = 3$) validates Proposition 2.3 numerically across homogeneous and heterogeneous dependence regimes; full results are given in supplementary material, Section S2.4 (Table S2). Up to $\rho = 0.2$ the cost of ignoring dependence is smooth and small (excess risk ≤ 0.0049); what matters is the *label-heterogeneity* of the dependence rather than its magnitude (a $2.77\times$ risk ratio at *lower* total dependence in the heterogeneous case); certificate (iv) of Proposition 2.3 holds throughout, with median slack $43\times$. In the case study the acoustic pairs satisfy $|\rho_s| \leq 0.19$, at or below the $\rho = 0.2$ row, while the biological–video pair is larger ($|\rho_s| \leq 0.54$) and is the regime this analysis quantifies.

7.8. Sensitivity to the regularization parameter α_0

We refit the weights and re-run the downstream pipeline at $\alpha_0 \in \{1, 2, 5, 10\}$, holding folds, conformal level $\alpha = 0.10$, and all else fixed (Table 8).

Table 8. Sensitivity to the prior-strength parameter α_0 (case-study test fold).

α_0	$\widehat{w}_a$	$\widehat{w}_b$	$\widehat{w}_v$	accuracy	macro- F_1	coverage
1	0.42 ± 0.03	0.31 ± 0.04	0.27 ± 0.03	0.916	0.910	0.910
2	0.54 ± 0.04	0.30 ± 0.04	0.16 ± 0.04	0.860	0.715	0.919
5	0.77 ± 0.05	0.23 ± 0.06	0.00 ± 0.06	0.446	0.222	0.901
10	1.00 ± 0.11	0.00 ± 0.17	0.00 ± 0.15	0.383	0.092	0.869

Classification degrades sharply (accuracy $0.916 \rightarrow 0.383$, macro- F_1 $0.910 \rightarrow 0.092$), whereas the realized conformal coverage remains within $[0.869, 0.919]$ across these four refits. This illustrates the distinction in Proposition 4.3: changing α_0 can affect predictive efficiency and accuracy, while the split-conformal validity statement is unchanged for every value fixed before calibration. No entry in this sensitivity table is used as a local or region-wise convergence factor; such a claim would require the separate calculations specified in Proposition 5.2 and Theorem 5.3.

7.9. Comparative weighting strategies

The finite-step likelihood-fitted weights are compared, on the identical split of (7.1), against three alternatives, uniform weights, held-out-accuracy-proportional weights, and an additive mixture-of-experts gate, in supplementary material, Section S2.5 (Table S3). These fitted weights are the unique maximizer of the observed likelihood (Theorem 4.2) and attain the highest descriptive test accuracy among the strategies in Table S3. Because the alternatives are not all nested likelihood models, no common likelihood-ratio interpretation is assigned to these comparisons.

7.10. A neutral remark on the κ comparison

Remark 7.2. (Why no comparison with human inter-rater agreement is offered) It would be natural to set the system-to-consensus agreement of Section 7.4 against the pediatrician inter-rater agreement of Section 7.1. We do not, because the two coefficients do not estimate the same quantity. The first compares one system against an aggregated consensus reference; the second compares two independent raters with each other. They are computed on different samples with different marginal class distributions. Moreover, the consensus reference is not an independent gold standard: It is constructed from the very raters whose mutual agreement the second coefficient measures, which biases any comparison in a direction that cannot be signed without a model of the adjudication process. No claim of parity with, or superiority over, human raters is made anywhere in this paper.

8. Conclusions

We have presented a RW-PoE framework for multimodal Bayesian classification and established its mathematical core: the conditional-independence factorization (Theorem 2.1); an exact entropy

decomposition for the hierarchical cascade (Theorem 2.2); a data-driven, distribution-scoped necessity criterion for fusion built from the Hellinger–Bayes bound (Lemma 3.1, Theorem 3.2, Proposition 3.3); the necessary-and-sufficient log-posterior-independence condition for strict convexity, which in turn guarantees a unique weight optimum on the simplex, with the covariance-Hessian identity (Theorem 4.2); a monotone finite-step GEM/entropic-mirror estimator with an explicit γ -dependent local Jacobian and observed-information standard-error estimates (Propositions 5.1 and 5.2); and a distribution-free conformal layer with a finite-sample lower coverage guarantee and a no-ties upper discretization bound (Theorem 6.1). The case study reports the fitted weights and downstream diagnostics on the prespecified data splits. Consistently with Proposition 5.2, no numerical local convergence factor is reported without the recorded step size and loss-normalization convention.

We are explicit about what the framework does and does not establish.

Single-site scope. The numerical results come from one site (the PIDD dataset, CHU Sidi Bel-Abbès). The conformal guarantee of Theorem 6.1 is marginal and assumes exchangeability of calibration and test data; it does not by itself transfer across a distribution shift to another site, nor does it deliver per-input (conditional) coverage. Restoring coverage under site or covariate shift is a natural next step: weighted or Mondrian conformal calibration and covariate-shift reweighting of the calibration scores are the standard routes, and adapting them to the RW-PoE score is left to future multi-site work. The necessity result of Theorem 3.2 is, by design, a statement about the *observed* acoustic distributions and feature representation (Remark 3.3); it is not a universal impossibility claim for acoustic discrimination.

Modeling assumptions. The exact factorization assumes class-conditional independence of modalities (Assumption 2.1); the cost of this approximation is controlled by the class-conditional total correlation (Proposition 2.3) and measured in Section 7.7: small, label-homogeneous dependence costs little, whereas strong or strongly label-heterogeneous dependence would require an explicit dependence model. Identifiability of the weights requires log-posterior independence (Definition 4.2); redundant modalities induce flat tangent directions. When such a direction intersects the optimum in a nontrivial feasible segment, it produces a ridge of constrained optima (Remark 4.1). Per-node reliability weights, suggested by Remark 5.3 (b), are a natural refinement.

Scope of the numerical estimates. The two acoustic overlap coefficients are kept separate throughout (Remark 7.1): The pooled one-vs-rest $BC_{\text{hungry}}^{\text{pool}} \geq 0.88$ governs the binary hunger node and yields the rigorous floor $R_{a,\text{root}}^* \geq 0.235$ (equal-prior benchmark ≈ 0.262), while the pairwise coefficients of Figure 1 ($D_B = 0.29$, $BC \approx 0.75$) describe class-versus-class geometry. The empirical near-chance audio error (≈ 0.432 , $AUC = 0.560$) is the observed error of a non-Bayes-optimal classifier and is reported separately from the lower bound, which it exceeds for that reason.

Future work. Three extensions are natural and preserve the present guarantees. First, a multi-site study with site-stratified (Mondrian) conformal calibration would convert the marginal coverage into group-conditional coverage and test cross-site transportability. Second, a federated mirror-descent implementation, in which the score contributions are aggregated across sites without sharing raw data, would fit the weights under privacy constraints while retaining the convexity and rate guarantees of Sections 4 and 5. Third, relaxing Assumption 2.1 to a structured (e.g., tree-dependent) modality-dependence model would extend the factorization while keeping the log-linear pooling interpretation. We make no efficacy or superiority claims beyond what the single-site case study supports.

Author contributions

Safia Benarbia designed the theoretical framework, proved the main theorems, conducted the numerical case study, and wrote the manuscript; Fatimah Alshahrani contributed to the information-theoretic sections and reviewed the manuscript; Mohammed Fethi Khalfi contributed to the algorithmic development and computational implementation; Wahiba Bouabsa supervised the research, provided guidance on clinical application, and critically revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Use of Generative-AI tools declaration

The authors declare Artificial Intelligence (AI) tools were used in the preparation of this article in a limited and specific capacity, as described below.

AI tools used: Claude (Anthropic, Sonnet).

How were the AI tools used? The AI assistant was used exclusively to improve the clarity, fluency, and readability of portions of the English prose in the manuscript. Specifically, the tool was used to refine sentence structure and academic writing style in selected passages. It was not used to generate, derive, or verify any mathematical content. All theorems, lemmas, propositions, proofs, definitions, algorithms, information-theoretic analyses, statistical methodology, data collection, numerical experiments, case study design, and scientific conclusions are entirely the original work of the authors and were developed independently of any AI tool.

The authors take full responsibility for the integrity, accuracy, and originality of all content presented in this manuscript.

Acknowledgments

The authors thank and express their sincere appreciation to the funders of this work: Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R358), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Conflict of interest

The authors declare no conflicts of interest.

References

1. C. Genest, J. V. Zidek, Combining probability distributions: a critique and an annotated bibliography, *Stat. Sci.*, **1** (1986), 114–135. <https://doi.org/10.1214/ss/1177013825>
2. A. E. Abbas, A Kullback–Leibler view of linear and log-linear pools, *Decis. Anal.*, **6** (2009), 25–37. <https://doi.org/10.1287/deca.1080.0133>
3. A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *J. R. Stat. Soc. Ser. B*, **39** (1977), 1–22. <https://doi.org/10.1111/j.2517-6161.1977.tb01600.x>

4. T. A. Louis, Finding the observed information matrix when using the EM algorithm, *J. R. Stat. Soc. Ser. B*, **44** (1982), 226–233. <https://doi.org/10.1111/j.2517-6161.1982.tb01203.x>
5. L. Devroye, L. Györfi, G. Lugosi, *A probabilistic theory of pattern recognition*, Springer, 1996. <https://doi.org/10.1007/978-1-4612-0711-5>
6. V. Vovk, A. Gammernan, G. Shafer, *Algorithmic learning in a random world*, Springer, 2005. <https://doi.org/10.1007/b106715>
7. A. N. Angelopoulos, S. Bates, Conformal prediction: a gentle introduction, *Found. Trends Mach. Learn.*, **16** (2023), 494–591. <https://doi.org/10.1561/22000000101>
8. T. Heskes, Selecting weighting factors in logarithmic opinion pools, *Proceedings of the 1997 Conference on Advances in Neural Information Processing Systems*, 1998, 266–272. <https://doi.org/10.5555/302528.302604>
9. G. E. Hinton, Training products of experts by minimizing contrastive divergence, *Neural Comput.*, **14** (2002), 1771–1800. <https://doi.org/10.1162/089976602760128018>
10. R. A. Jacobs, M. I. Jordan, S. J. Nowlan, G. E. Hinton, Adaptive mixtures of local experts, *Neural Comput.*, **3** (1991), 79–87. <https://doi.org/10.1162/neco.1991.3.1.79>
11. Y. Romano, M. Sesia, E. J. Candès, Classification with valid and adaptive coverage, *ArXiv*, 2020. <https://doi.org/10.48550/arXiv.2006.02544>
12. A. N. Angelopoulos, S. Bates, M. I. Jordan, J. Malik, Uncertainty sets for image classifiers using conformal prediction, *Int. Conf. Learn. Representations*, 2021.
13. T. M. Cover, J. A. Thomas, *Elements of information theory*, 2 Eds., John Wiley & Sons, Inc., 2006. <https://doi.org/10.1002/047174882X>
14. S. Watanabe, Information theoretical analysis of multivariate correlation, *IBM J. Res. Dev.*, **4** (1960), 66–82. <https://doi.org/10.1147/rd.41.0066>
15. A. Bhattacharyya, On a measure of divergence between two statistical populations defined by their probability distributions, *Bull. Calcutta Math. Soc.*, **35** (1943), 99–109.
16. E. R. DeLong, D. M. DeLong, D. L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, *Biometrics*, **44** (1988), 837–845. <https://doi.org/10.2307/2531595>
17. H. Lu, R. M. Freund, Y. Nesterov, Relatively smooth convex optimization by first-order methods, and applications, *SIAM J. Optim.*, **28** (2018), 333–354. <https://doi.org/10.1137/16M1099546>

Supplementary

This document collects the proofs deferred from the main article (Section S1), together with worked examples and secondary robustness checks moved here for length (Section S2). Equation, theorem, lemma and proposition numbers of the form Eq (4.6), Theorem 4.3, etc., refer to the main article; equations introduced here are numbered (S1.1), (S1.2), The main article retains the formal statements and conclusions, while the complete derivations are collected here to keep the exposition concise.

S1. Deferred proofs

S1.1. Proof of Theorem 2.2 (PHC entropy decomposition and reduction)

Proof. Step 1 (local chain rule at a node). Fix a reachable internal node v and condition throughout on $\{Y \in A_v\}$. Because D_v is a deterministic function of Y on this event, $H(Y, D_v \mid X=x, Y \in A_v) = H(Y \mid X=x, Y \in A_v)$, and the chain rule for entropy gives

$$H(Y \mid X=x, Y \in A_v) = H(D_v \mid X=x, Y \in A_v) + \sum_{u \in \text{ch}(v)} \beta_u(x) H(Y \mid X=x, Y \in A_u), \quad (\text{S1.1})$$

where $\beta_u(x) = \mathbb{P}(D_v=u \mid X=x, Y \in A_v) = \mathbb{P}(Y \in A_u \mid X=x, Y \in A_v)$, and we used that, given $D_v=u$, the event $\{Y \in A_v\}$ coincides with $\{Y \in A_u\}$.

Step 2 (unfolding the recursion over the tree). Apply (S1.1) at the root (where the conditioning event is all of Ω) and then recursively to each internal child. A leaf ℓ satisfies $H(Y \mid X=x, Y \in A_\ell) = 0$ since A_ℓ is a singleton. The branch factors $\beta_u(x)$ telescope along each root-to-node path: the cumulative weight multiplying $H(D_v \mid X=x, Y \in A_v)$ equals the product of branch probabilities from the root to v , which is exactly $\mathbb{P}(Y \in A_v \mid X=x)$. Hence,

$$H(Y \mid X=x) = \sum_{v \text{ internal}} \mathbb{P}(Y \in A_v \mid X=x) H(D_v \mid X=x, Y \in A_v). \quad (\text{S1.2})$$

Step 3 (integrating out X). Take expectations of (S1.2) over $X \sim \mathbb{P}$. For each internal node v , the change-of-measure identity

$$\mathbb{P}(Y \in A_v \mid X=x) d\mathbb{P}(X=x) = \mathbb{P}(Y \in A_v) d\mathbb{P}(X=x \mid Y \in A_v) \quad (\text{S1.3})$$

follows from Bayes' theorem applied to the event $\{Y \in A_v\}$, the last factor being the definition of the conditional distribution. Integrating both sides against a measurable f converts the left side to $\mathbb{E}_X[\mathbb{P}(Y \in A_v \mid X) f(X)]$ and the right side to $\mathbb{P}(Y \in A_v) \mathbb{E}_{X|Y \in A_v}[f(X)]$. Applying this with $f(x) = H_x(D_v)$, where $H_x(D_v)$ denotes the entropy of the conditional law of D_v given $X=x$ within $\{Y \in A_v\}$, gives

$$\mathbb{E}_X[\mathbb{P}(Y \in A_v \mid X) H(D_v \mid X, Y \in A_v)] = \mathbb{P}(Y \in A_v) H(D_v \mid X, Y \in A_v),$$

whose $\mathbb{P}(\cdot \mid Y \in A_v)$ -average of $H_x(D_v)$ is $H(D_v \mid X, Y \in A_v)$. This proves (2.6). Taking X to be degenerate (a constant) in (S1.2) removes the conditioning on X and yields (2.5).

Step 4 (reduction). Subtracting (2.6) from (2.5) and using, at each node, the identity $H(D_v \mid Y \in A_v) - H(D_v \mid X, Y \in A_v) = I(D_v; X \mid Y \in A_v)$ together with the nonnegativity of (conditional) mutual information gives (2.7). Equality $I(D_v; X \mid Y \in A_v) = 0$ holds iff D_v and X are independent under $\mathbb{P}(\cdot \mid Y \in A_v)$, so $H(Y \mid X) = H(Y)$ iff every reachable node carries no information about its split. $\square$

S1.2. Proof of Lemma 3.1 (Hellinger–Bayes lower bound)

Proof. Write $p - q = (\sqrt{p} - \sqrt{q})(\sqrt{p} + \sqrt{q})$ and apply the Cauchy–Schwarz inequality:

$$\int |p - q| d\mu_a \leq \left(\int (\sqrt{p} - \sqrt{q})^2 d\mu_a \right)^{1/2} \left(\int (\sqrt{p} + \sqrt{q})^2 d\mu_a \right)^{1/2}.$$

Since

$$\int (\sqrt{p} \mp \sqrt{q})^2 d\mu_a = 2(1 \mp \text{BC}(p, q)),$$

the right-hand side equals

$$\sqrt{2(1 - \text{BC})} \sqrt{2(1 + \text{BC})} = 2\sqrt{1 - \text{BC}^2},$$

whence

$$\text{TV}(p, q) = \frac{1}{2} \int |p - q| d\mu_a \leq \sqrt{1 - \text{BC}^2}$$

proving (3.3). For the general priors, use the pointwise identity $\min\{u, v\} = \frac{1}{2}(u + v - |u - v|)$ with $u = \alpha p$, $v = (1 - \alpha)q$: since $\int (\alpha p + (1 - \alpha)q) d\mu_a = 1$,

$$R_{\{p, q\}}^*(\alpha) = \frac{1}{2} \left(1 - \int |\alpha p - (1 - \alpha)q| d\mu_a \right).$$

Factoring $\alpha p - (1 - \alpha)q = (\sqrt{\alpha p} - \sqrt{(1 - \alpha)q})(\sqrt{\alpha p} + \sqrt{(1 - \alpha)q})$ and applying Cauchy–Schwarz gives

$$\begin{aligned} \left(\int |\alpha p - (1 - \alpha)q| d\mu_a \right)^2 &\leq (1 - 2\sqrt{\alpha(1 - \alpha)} \text{BC})(1 + 2\sqrt{\alpha(1 - \alpha)} \text{BC}) \\ &= 1 - 4\alpha(1 - \alpha) \text{BC}^2, \end{aligned}$$

where we used

$$\int (\sqrt{\alpha p} \mp \sqrt{(1 - \alpha)q})^2 d\mu_a = 1 \mp 2\sqrt{\alpha(1 - \alpha)} \text{BC}.$$

Substituting yields (3.4); setting $\alpha = \frac{1}{2}$ gives (3.5). $\square$

51.3. Proof of Theorem 3.2 (acoustic Bayes-risk barrier)

Proof. (i) By hypothesis $\text{BC}(p, q) \geq 1 - \varepsilon$. Substituting into the general unequal-prior bound (3.4) of Lemma 3.1 gives (3.6); the equal-prior special case $\alpha = \frac{1}{2}$ uses

$$\text{TV} \leq \sqrt{1 - (1 - \varepsilon)^2} = \sqrt{\varepsilon(2 - \varepsilon)}$$

in (3.5), and the limit follows as $\varepsilon \rightarrow 0$.

(ii) For the Bayes classifier ϕ^* ,

$$\begin{aligned} R_a^* &= \mathbb{P}(\phi^*(X_a) \neq Y) \geq \mathbb{P}(\phi^*(X_a) \neq Y, Y \in \{k^*, j\}) \\ &= (\pi_{k^*} + \pi_j) \mathbb{P}(\phi^*(X_a) \neq Y \mid Y \in \{k^*, j\}). \end{aligned}$$

Restricted to the event $\{Y \in \{k^*, j\}\}$, any classifier—in particular ϕ^* , which may even output a third label, always an error there—has conditional error at least the binary Bayes risk $\tilde{R}_{\{k^*, j\}}^*$ of that subpopulation. Hence (3.7). Strict positivity holds because $\text{BC} \geq 1 - \varepsilon > 0$ forces p and q to overlap on a set of positive μ_a -measure, so

$$\tilde{R}_{\{k^*, j\}}^* = \int \min\{\alpha p, (1 - \alpha)q\} d\mu_a > 0$$

with $\alpha = \pi_{k^*}/(\pi_{k^*} + \pi_j)$. As $\varepsilon \rightarrow 0$, $p \rightarrow q$ in L^1 and $\tilde{R}_{\{k^*, j\}}^* \rightarrow \int \min\{\alpha, 1 - \alpha\} p d\mu_a = \min(\alpha, 1 - \alpha)$, giving the stated limit. $\square$

S1.4. Proof of Proposition 3.3 (monotonicity and strict improvement)

Proof. For each k , $\mathbb{P}(Y = k \mid \mathcal{G}_1) = \mathbb{E}[\mathbb{P}(Y = k \mid \mathcal{G}_2) \mid \mathcal{G}_1]$ by the tower property. Since $\max_k \mathbb{E}[Z_k \mid \mathcal{G}_1] \leq \mathbb{E}[\max_k Z_k \mid \mathcal{G}_1]$ for any integrable family (Z_k) ,

$$\max_k \mathbb{P}(Y = k \mid \mathcal{G}_1) \leq \mathbb{E}[\max_k \mathbb{P}(Y = k \mid \mathcal{G}_2) \mid \mathcal{G}_1] \quad \text{a.s.}$$

Taking expectations gives $\mathbb{E}[\max_k \mathbb{P}(Y = k \mid \mathcal{G}_1)] \leq \mathbb{E}[\max_k \mathbb{P}(Y = k \mid \mathcal{G}_2)]$, i.e., $R_{\text{fusion}}^* \leq R_a^*$ by (2.4). Equality holds iff the displayed conditional Jensen inequality is an equality a.s. For the pointwise maximum of finitely many coordinates, this occurs precisely when there exists a $\mathcal{G}_1$ -measurable selection ψ that maximises $k \mapsto \mathbb{P}(Y = k \mid \mathcal{G}_2)$ a.s. This is condition (3.8); when the two posterior modes are unique a.s., it reduces to equality of their Bayes labels. $\square$

S1.5. Proof of Proposition 4.1 (coherence and weighted-KL centroid)

Proof. Part (a) is [1, Thm. 3]; we prove (b). The minimisation is the finite-dimensional convex program $\min_p \sum_m w_m \text{KL}(p \parallel q_m)$ subject to the single equality constraint $\sum_k p_k = 1$, and “stationarity” means its first-order (Karush–Kuhn–Tucker) condition. The objective $p \mapsto \sum_m w_m \text{KL}(p \parallel q_m)$ is strictly convex on Δ^{K-1} (a sum of strictly convex KL terms in p). The minimiser is interior whenever every q_m has full support, so the inequality constraints $p_k \geq 0$ are inactive and the stationarity condition below is both necessary and sufficient. With multiplier λ for $\sum_k p_k = 1$, the Lagrangian $\Lambda(p, \lambda) = \sum_m w_m \text{KL}(p \parallel q_m) + \lambda(\sum_k p_k - 1)$ has stationarity condition, for each k , $\partial\Lambda/\partial p_k = \sum_m w_m(\log(p_k/q_m(k)) + 1) + \lambda = 0$; using $\sum_m w_m = 1$ this rearranges to $\log p_k = \sum_m w_m \log q_m(k) + \text{const}$, i.e., $p_k \propto \prod_m q_m(k)^{w_m}$. Strict convexity makes this the unique minimiser, which is (4.2). $\square$

S1.6. Proof of Proposition 5.1 (finite-step likelihood ascent and convergence)

Proof. Step 1: variational characterisation and the three-point inequality. Let $h(\mathbf{w}) = \sum_m w_m \log w_m$ and $B_h(\mathbf{u}, \mathbf{v}) = h(\mathbf{u}) - h(\mathbf{v}) - \langle \nabla h(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle$. The variational form in (5.6) is

$$\mathbf{w}^{(t+1)} = \arg \min_{\mathbf{w} \in \Delta^{M-1}} \left\{ \gamma \langle \nabla L(\mathbf{w}^{(t)}), \mathbf{w} \rangle + B_h(\mathbf{w}, \mathbf{w}^{(t)}) \right\}. \quad (\text{S1.4})$$

For interior $\mathbf{w}^{(t)}$, the objective in (S1.4) is strictly convex. Its Lagrange equations are

$$\gamma [\nabla L(\mathbf{w}^{(t)})]_m + \log \frac{w_m}{w_m^{(t)}} + 1 + \lambda = 0,$$

so the unique normalised solution is $w_m \propto w_m^{(t)} \exp\{-\gamma [\nabla L(\mathbf{w}^{(t)})]_m\}$. Every component is positive, proving both equivalence with (5.6) and preservation of the relative interior. The first-order optimality condition and the Bregman three-point identity give, for every $\mathbf{u} \in \Delta^{M-1}$,

$$\gamma \langle \nabla L(\mathbf{w}^{(t)}), \mathbf{w}^{(t+1)} - \mathbf{u} \rangle \leq B_h(\mathbf{u}, \mathbf{w}^{(t)}) - B_h(\mathbf{u}, \mathbf{w}^{(t+1)}) - B_h(\mathbf{w}^{(t+1)}, \mathbf{w}^{(t)}). \quad (\text{S1.5})$$

Step 2: relative smoothness on the simplex. We next verify the required relative-smoothness bound. Positivity of all modality posteriors makes every vector $\ell_i(k)$ finite, and Theorem 4.2 gives

$$\nabla^2 L(\mathbf{w}) = \sum_i \text{Var}_{k \sim P_{\mathbf{w}}(\cdot \mid x_i)}[\ell_i(k)].$$

This Hessian is continuous on the compact simplex, so the tangent-space constant L_Δ in (5.7) is finite. Moreover, for every interior $\mathbf{w}$ and $v \in T$,

$$v^\top \nabla^2 h(\mathbf{w}) v = \sum_m \frac{v_m^2}{w_m} \geq \|v\|_2^2.$$

Consequently $\nabla_T^2 L(\mathbf{w}) \leq L_\Delta \nabla_T^2 h(\mathbf{w})$. Integration along the line segment from $\mathbf{w}^{(t)}$ to $\mathbf{w}^{(t+1)}$ yields

$$L(\mathbf{w}^{(t+1)}) \leq L(\mathbf{w}^{(t)}) + \langle \nabla L(\mathbf{w}^{(t)}), \mathbf{w}^{(t+1)} - \mathbf{w}^{(t)} \rangle + L_\Delta B_h(\mathbf{w}^{(t+1)}, \mathbf{w}^{(t)}). \quad (\text{S1.6})$$

Step 3: monotonicity and the equality case. Taking $\mathbf{u} = \mathbf{w}^{(t)}$ in (S1.5) and substituting the result into (S1.6) gives

$$L(\mathbf{w}^{(t+1)}) \leq L(\mathbf{w}^{(t)}) - \frac{1}{\gamma} B_h(\mathbf{w}^{(t)}, \mathbf{w}^{(t+1)}) - \left(\frac{1}{\gamma} - L_\Delta \right) B_h(\mathbf{w}^{(t+1)}, \mathbf{w}^{(t)}),$$

which is (5.8). For $0 < \gamma \leq 1/L_\Delta$, both subtracted terms are nonnegative. Since KL divergence vanishes only when its arguments coincide, equality requires $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)}$. The first-order condition of (S1.4) then says that the tangent projection of $\nabla L(\mathbf{w}^{(t)})$ is zero, i.e. the iterate is stationary.

Step 4: convergence to the unique MLE. Finally, summing the sufficient-decrease inequality shows $\sum_i B_h(\mathbf{w}^{(i)}, \mathbf{w}^{(i+1)}) < \infty$, because L is bounded below on the simplex. Pinsker's inequality therefore gives $\|\mathbf{w}^{(i+1)} - \mathbf{w}^{(i)}\|_1 \rightarrow 0$. If the iterates stay in $\mathcal{S}_{w_{\min}}$, compactness supplies limit points in the relative interior. Along any convergent subsequence, continuity of the explicit map (5.6) and the vanishing successive differences show that the limit is a fixed point, hence stationary. Under log-posterior independence, L is strictly convex on the simplex, so its only stationary point is $\widehat{\mathbf{w}}$. Every limit point is thus $\widehat{\mathbf{w}}$, which proves convergence of the whole sequence. $\square$

S1.7. Proof of Proposition 5.2 (finite-step Jacobian and local rate)

Proof. Step 1: KKT identity and fixed-point property. We differentiate the map that is actually iterated. Put

$$g(\mathbf{w}) = \nabla L(\mathbf{w}), \quad a(\mathbf{w}) = \mathbf{w} \odot \exp\{-\gamma g(\mathbf{w})\}, \quad d(\mathbf{w}) = \mathbf{1}^\top a(\mathbf{w}),$$

so that $\mathcal{M}_\gamma(\mathbf{w}) = a(\mathbf{w})/d(\mathbf{w})$. Because $\widehat{\mathbf{w}}$ is an interior constrained minimiser, the KKT condition gives $g(\widehat{\mathbf{w}}) = \kappa \mathbf{1}$ for some $\kappa \in \mathbb{R}$. Consequently

$$a(\widehat{\mathbf{w}}) = e^{-\gamma\kappa} \widehat{\mathbf{w}}, \quad d(\widehat{\mathbf{w}}) = e^{-\gamma\kappa}, \quad \mathcal{M}_\gamma(\widehat{\mathbf{w}}) = \widehat{\mathbf{w}}.$$

Step 2: derivative of the normalised finite-step map. Let $H_\star = \nabla^2 L(\widehat{\mathbf{w}})$, $D_\star = \text{Diag}(\widehat{\mathbf{w}})$, and take an arbitrary perturbation $v \in \mathbb{R}^M$. Componentwise differentiation gives

$$Da(\widehat{\mathbf{w}})[v] = e^{-\gamma\kappa}(v - \gamma D_\star H_\star v)$$

and hence

$$Dd(\widehat{\mathbf{w}})[v] = e^{-\gamma\kappa}(\mathbf{1}^\top v - \gamma \widehat{\mathbf{w}}^\top H_\star v).$$

Applying the quotient rule at $\widehat{\mathbf{w}}$ therefore yields

$$D\mathcal{M}_\gamma(\widehat{\mathbf{w}})[v] = v - \widehat{\mathbf{w}} \mathbf{1}^\top v - \gamma(D_\star - \widehat{\mathbf{w}} \widehat{\mathbf{w}}^\top) H_\star v.$$

For $v \in T$, $\mathbf{1}^\top v = 0$, and with $\Xi_\star = D_\star - \widehat{\mathbf{w}}\widehat{\mathbf{w}}^\top$ this becomes

$$D\mathcal{M}_\gamma(\widehat{\mathbf{w}})[v] = v - \gamma\Xi_\star H_\star v, \quad v \in T. \quad (\text{S1.7})$$

The range is tangent because $\mathbf{1}^\top \Xi_\star = 0$. Let $P_T = EE^\top$ be the orthogonal projector onto T . Since $\Xi_\star \mathbf{1} = 0$, we have $\Xi_\star = \Xi_\star P_T = P_T \Xi_\star$; consequently

$$E^\top \Xi_\star H_\star E = (E^\top \Xi_\star E)(E^\top H_\star E) = \Xi_T H_T.$$

Multiplying (S1.7) on the left by $E^\top$ and setting $v = Eu$ proves

$$J_{\gamma,T} = E^\top D\mathcal{M}_\gamma(\widehat{\mathbf{w}})E = I_{M-1} - \gamma\Xi_T H_T.$$

Step 3: spectrum and local stability. To verify positive definiteness explicitly, take nonzero $u \in \mathbb{R}^{M-1}$ and put $v = Eu \in T$. Then

$$u^\top \Xi_T u = \sum_{m=1}^M \widehat{w}_m v_m^2 - \left(\sum_{m=1}^M \widehat{w}_m v_m \right)^2 = \text{Var}_{m \sim \widehat{\mathbf{w}}}(v_m) > 0.$$

Indeed, equality would force v to be constant on the full support of $\widehat{\mathbf{w}}$; because $v \in T$, that constant must be zero, contradicting $u \neq 0$. Thus $\Xi_T > 0$. Log-posterior independence gives $H_T > 0$ by Theorem 4.2. Moreover,

$$\Xi_T^{-1/2}(\Xi_T H_T)\Xi_T^{1/2} = \Xi_T^{1/2} H_T \Xi_T^{1/2} = B_\star,$$

so $\Xi_T H_T$ is similar to the symmetric positive-definite matrix $B_\star$. If v_j are its eigenvalues, the eigenvalues of $J_{\gamma,T}$ are exactly $1 - \gamma v_j$. Thus

$$\rho(J_{\gamma,T}) = \max_j |1 - \gamma v_j| = q_\gamma,$$

and $q_\gamma < 1$ is equivalent to $-1 < 1 - \gamma v_j < 1$ for every j , or $0 < \gamma < 2/v_{M-1}$. To make the local-convergence conclusion explicit, equip the tangent coordinates with the norm

$$\|u\|_{\Xi_T^{-1}} = (u^\top \Xi_T^{-1} u)^{1/2}.$$

Because

$$\Xi_T^{-1} J_{\gamma,T} = \Xi_T^{-1} - \gamma H_T$$

is symmetric, $J_{\gamma,T}$ is self-adjoint in the corresponding inner product. Consequently,

$$\|J_{\gamma,T}\|_{\Xi_T^{-1}} = \rho(J_{\gamma,T}) = q_\gamma.$$

Let $\widetilde{\mathcal{M}}_\gamma$ denote the tangent-coordinate representation of $\mathcal{M}_\gamma$ around $\widehat{\mathbf{w}}$. For every $r \in (q_\gamma, 1)$, continuity of $D\widetilde{\mathcal{M}}_\gamma$ provides a sufficiently small ball U around the origin such that

$$\sup_{u \in U} \|D\widetilde{\mathcal{M}}_\gamma(u)\|_{\Xi_T^{-1}} \leq r.$$

Since $\widetilde{\mathcal{M}}_\gamma(0) = 0$, the mean-value inequality gives

$$\|\widetilde{\mathcal{M}}_\gamma(u)\|_{\Xi_T^{-1}} \leq r\|u\|_{\Xi_T^{-1}}, \quad u \in U.$$

After reducing U if necessary, this inequality also shows that U is invariant under the iteration. Therefore,

$$\|\mathbf{w}^{(t)} - \widehat{\mathbf{w}}\|_2 \leq C_r r^t \|\mathbf{w}^{(0)} - \widehat{\mathbf{w}}\|_2, \quad t \geq 0,$$

for some $C_r < \infty$, by equivalence of norms in finite dimension. Thus $\widehat{\mathbf{w}}$ is locally attracting and q_γ is the local spectral factor of the implemented finite-step map.

Step 4: distinction from exact EM. The preceding derivative was obtained directly from $\mathcal{M}_\gamma$. The classical missing-information derivative belongs to a different exact-EM operator whose M-step maximises its surrogate [3, §3]. It contains no finite step-size parameter and cannot be substituted for $D\mathcal{M}_\gamma(\widehat{\mathbf{w}})$.

Step 5: observed information and constrained covariance. Direct differentiation of the observed likelihood gives

$$\mathcal{I}_{\text{obs},T}(\widehat{\mathbf{w}}) = E^\top \nabla^2 L(\widehat{\mathbf{w}}) E = H_T.$$

Under independent-observation likelihood regularity and for an interior maximum-likelihood estimator, the conventional asymptotic Wald covariance estimate in tangent coordinates is H_T^{-1} when L is the summed negative log-likelihood. Its ambient-space representation under the simplex constraint is therefore

$$\widehat{\text{Cov}}(\widehat{\mathbf{w}}) = E H_T^{-1} E^\top.$$

This identity specifies the covariance calculation. Any numerical standard errors must be obtained by evaluating H_T on the same training fold and under the same loss-normalisation convention used for estimation. If the mean negative log-likelihood is used instead, the corresponding factor of n must be restored before inversion. $\square$

Basis invariance. If $\widetilde{E} = EQ$ is another orthonormal tangent basis, where Q is orthogonal, then $\widetilde{H}_T = Q^\top H_T Q$ and $\widetilde{\Xi}_T = Q^\top \Xi_T Q$. Hence the resulting Jacobian is $Q^\top J_{\gamma,T} Q$, and the ambient covariance remains $\widetilde{E} \widetilde{H}_T^{-1} \widetilde{E}^\top = E H_T^{-1} E^\top$. Thus neither q_γ nor the standard errors depend on the chosen orthonormal representation of T .

Best scalar step for the local linearisation. Let $\nu_{\min} = \nu_1$ and $\nu_{\max} = \nu_{M-1}$. Minimising $q_\gamma = \max_j |1 - \gamma \nu_j|$ over scalar steps gives

$$\gamma_{\text{opt}} = \frac{2}{\nu_{\min} + \nu_{\max}}, \quad \inf_{\gamma > 0} q_\gamma = \frac{\nu_{\max} - \nu_{\min}}{\nu_{\max} + \nu_{\min}}. \quad (\text{S1.8})$$

Indeed, on the stable interval the maximum is minimised when the two endpoint errors have equal magnitude, $1 - \gamma \nu_{\min} = -(1 - \gamma \nu_{\max})$. Formula (S1.8) also supplies a direct consistency check for any proposed local rate.

Evaluation from the recorded fit. To evaluate the rate without ambiguity, record the loss convention (summed or averaged), the implemented γ , the full-precision $\widehat{\mathbf{w}}$, and H_T . Then compute

$$\Xi_T = E^\top [\text{Diag}(\widehat{\mathbf{w}}) - \widehat{\mathbf{w}}\widehat{\mathbf{w}}^\top] E, \quad \nu_j = \lambda_j \left(\Xi_T^{1/2} H_T \Xi_T^{1/2} \right), \quad q_\gamma = \max_j |1 - \gamma \nu_j|. \quad (\text{S1.9})$$

If $L_{\text{mean}} = L_{\text{sum}}/n$, the equivalent steps obey $\gamma_{\text{mean}} = n\gamma_{\text{sum}}$ and $H_{T,\text{mean}} = H_{T,\text{sum}}/n$; hence the product γH_T , and therefore q_γ , is invariant when the convention and step size are transformed consistently.

S1.8. Proof of Theorem 5.3 (explicit non-asymptotic linear rate)

Proof. Step 1: the finite-step update is entropic mirror descent. The mirror-descent step for the map h with step γ is $\mathbf{w}^{(t+1)} = \arg \min_{\mathbf{w} \in \Delta^{M-1}} \{\gamma \langle \nabla L(\mathbf{w}^{(t)}), \mathbf{w} \rangle + B_h(\mathbf{w}, \mathbf{w}^{(t)})\}$. Writing the Lagrangian for $\sum_m w_m = 1$ and differentiating in w_m gives $\gamma[\nabla L(\mathbf{w}^{(t)})]_m + \log(w_m/w_m^{(t)}) + 1 + \nu = 0$, i.e.,

$$w_m^{(t+1)} \propto w_m^{(t)} \exp(-\gamma[\nabla L(\mathbf{w}^{(t)})]_m),$$

renormalised. By the score identity (5.2), the m -th component of $\nabla L = -\nabla \ell_{\text{obs}}$ is $-(t_m - s_m^{(t)})$, so $-\gamma[\nabla L(\mathbf{w}^{(t)})]_m = \gamma(t_m - s_m^{(t)})$. Substituting reproduces the exponentiated-gradient form (5.6). For the non-asymptotic analysis of Theorem 5.3, the corresponding entropic mirror step is considered on the restricted convex region $\mathcal{S}_{w_{\min}}$. Thus, with $\gamma = 1/L$, the update analysed in Theorem 5.3 is

$$\mathbf{w}^{(t+1)} = \arg \min_{\mathbf{w} \in \mathcal{S}_{w_{\min}}} \{\gamma \langle \nabla L(\mathbf{w}^{(t)}), \mathbf{w} \rangle + B_h(\mathbf{w}, \mathbf{w}^{(t)})\}.$$

Equivalently, after multiplying the objective by L ,

$$\mathbf{w}^{(t+1)} = \arg \min_{\mathbf{w} \in \mathcal{S}_{w_{\min}}} \{\langle \nabla L(\mathbf{w}^{(t)}), \mathbf{w} \rangle + L B_h(\mathbf{w}, \mathbf{w}^{(t)})\},$$

which is the restricted entropic-mirror step appearing in Theorem 5.3.

Step 2: Hessian of the mirror map. On the relative interior of Δ^{M-1} , h is twice differentiable with $\nabla^2 h(\mathbf{w}) = \text{diag}(1/w_m)$. For any tangent $\mathbf{v} \in T$, $\mathbf{v}^\top \nabla^2 h(\mathbf{w}) \mathbf{v} = \sum_m v_m^2 / w_m$. Since $w_m \leq 1$ implies that $1/w_m \geq 1$, and $w_m \geq w_{\min}$ implies that $1/w_m \leq 1/w_{\min}$, we have the two-sided Loewner bound, on T and uniformly on $\mathcal{S}_{w_{\min}}$,

$$\|\mathbf{v}\|^2 \leq \mathbf{v}^\top \nabla^2 h(\mathbf{w}) \mathbf{v} \leq \frac{1}{w_{\min}} \|\mathbf{v}\|^2, \quad \text{i.e.,} \quad I_T \leq \nabla^2 h(\mathbf{w}) \leq \frac{1}{w_{\min}} I_T, \quad (\text{S1.10})$$

where I_T is the identity on T .

Step 3: two-sided Loewner bounds on $\nabla^2 L$. By Theorem 4.2, for $\mathbf{v} \in T$, $\mathbf{v}^\top \nabla^2 L(\mathbf{w}) \mathbf{v} = \sum_i \text{Var}_{k \sim P_{\mathbf{w}}(\cdot | x_i)}[\mathbf{v}^\top \boldsymbol{\ell}_i(k)] \geq 0$. Each summand is the variance of the bounded random variable $k \mapsto \mathbf{v}^\top \boldsymbol{\ell}_i(k)$ under the probability vector $P_{\mathbf{w}}(\cdot | x_i)$, which is a continuous function of $\mathbf{w}$ on the compact set $\mathcal{S}_{w_{\min}}$; hence $\mathbf{w} \mapsto \nabla^2 L(\mathbf{w})$ is continuous and its tangent eigenvalues attain their extrema on $\mathcal{S}_{w_{\min}}$. Thus $\bar{L} < \infty$ and, because log-posterior independence makes $\nabla^2 L(\mathbf{w})$ positive definite on T at every $\mathbf{w} \in \mathcal{S}_{w_{\min}}$ (Theorem 4.2) and the infimum of a positive continuous function over a compact set is attained and positive, $\underline{c} > 0$. By definition of $\bar{L}, \underline{c}$,

$$\underline{c} I_T \leq \nabla^2 L(\mathbf{w}) \leq \bar{L} I_T \quad \text{for all } \mathbf{w} \in \mathcal{S}_{w_{\min}}. \quad (\text{S1.11})$$

Step 4: relative smoothness and strong convexity. Combining (S1.10) and (S1.11) on T : for the upper bound, $\nabla^2 h(\mathbf{w}) \succeq I_T$ gives $\nabla^2 L(\mathbf{w}) \leq \bar{L} I_T \leq \bar{L} \nabla^2 h(\mathbf{w})$; for the lower bound, $\nabla^2 h(\mathbf{w}) \preceq (1/w_{\min}) I_T$ gives $w_{\min} \nabla^2 h(\mathbf{w}) \leq I_T$, hence $\nabla^2 L(\mathbf{w}) \geq \underline{c} I_T \geq \underline{c} w_{\min} \nabla^2 h(\mathbf{w})$. Therefore, on T and uniformly on $\mathcal{S}_{w_{\min}}$,

$$w_{\min} \underline{c} \nabla^2 h(\mathbf{w}) \leq \nabla^2 L(\mathbf{w}) \leq \bar{L} \nabla^2 h(\mathbf{w}). \quad (\text{S1.12})$$

We pass from the Hessian bounds (S1.12) to the Bregman inequalities (5.14) by carrying out the integration explicitly. Fix $\mathbf{w}, \mathbf{w}' \in \mathcal{S}_{w_{\min}}$, set $\mathbf{w}_s = \mathbf{w} + s(\mathbf{w}' - \mathbf{w})$ and $d = \mathbf{w}' - \mathbf{w} \in T$; the segment lies in $\mathcal{S}_{w_{\min}}$ by convexity and d is tangent. Taylor's theorem with integral remainder applied to a C^2 function f gives

$$f(\mathbf{w}') - f(\mathbf{w}) - \langle \nabla f(\mathbf{w}), d \rangle = \int_0^1 (1-s) d^T \nabla^2 f(\mathbf{w}_s) d \, ds,$$

in which $f = h$ makes the right-hand side $B_h(\mathbf{w}', \mathbf{w})$ and $f = L$ makes it the middle term of (5.14). Contracting (S1.12) at $\mathbf{w}_s$ with d gives, for every $s \in [0, 1]$,

$$w_{\min} \underline{c} d^T \nabla^2 h(\mathbf{w}_s) d \leq d^T \nabla^2 L(\mathbf{w}_s) d \leq \bar{L} d^T \nabla^2 h(\mathbf{w}_s) d;$$

multiplying by $(1-s) \geq 0$ and integrating over $s \in [0, 1]$ (legitimate because all three integrands are continuous on the compact segment and $\nabla^2 h > 0$ on T) yields

$$w_{\min} \underline{c} B_h(\mathbf{w}', \mathbf{w}) \leq L(\mathbf{w}') - L(\mathbf{w}) - \langle \nabla L(\mathbf{w}), \mathbf{w}' - \mathbf{w} \rangle \leq \bar{L} B_h(\mathbf{w}', \mathbf{w}),$$

which is (5.14) with $L = \bar{L}$ and $\mu = w_{\min} \underline{c}$. This is the standard Loewner–Bregman equivalence [17, Prop. 1.1, Section 2], here made explicit. Clearly

$$0 < \mu = w_{\min} \underline{c} \leq \bar{L} = L$$

because $w_{\min} \leq 1$ and $\underline{c} \leq \bar{L}$.

Step 5: linear convergence via the four stages of [17, Thm. 3.1]. With $L = \bar{L}$ -relative smoothness and $\mu = w_{\min} \underline{c}$ -relative strong convexity established in Step 4, the mirror objective underlying (5.6) is considered with feasible set $\mathcal{S}_{w_{\min}}$, step $\gamma = 1/L$, and reference function h of Definition 5.1. We make the four stages of [17, Thm. 3.1] explicit below.

Stage 1 (three-point majorisation). By Step 1, the restricted counterpart of (5.6) used in Theorem 5.3 is the Bregman-proximal update

$$\mathbf{w}^{(t+1)} = \arg \min_{\mathbf{w} \in \mathcal{S}_{w_{\min}}} \{ \gamma \langle \nabla L(\mathbf{w}^{(t)}), \mathbf{w} \rangle + B_h(\mathbf{w}, \mathbf{w}^{(t)}) \}$$

with $\gamma = 1/L$. The three-point property of this step [17, Lem. 3.1] therefore gives, for every $\mathbf{w} \in \mathcal{S}_{w_{\min}}$,

$$\langle \nabla L(\mathbf{w}^{(t)}), \mathbf{w}^{(t+1)} - \mathbf{w} \rangle \leq L [B_h(\mathbf{w}, \mathbf{w}^{(t)}) - B_h(\mathbf{w}, \mathbf{w}^{(t+1)}) - B_h(\mathbf{w}^{(t+1)}, \mathbf{w}^{(t)})]. \quad (\text{S1.13})$$

L -relative smoothness (the upper bound in (5.14), applied at $(\mathbf{w}^{(t)}, \mathbf{w}^{(t+1)})$) gives

$$L(\mathbf{w}^{(t+1)}) \leq L(\mathbf{w}^{(t)}) + \langle \nabla L(\mathbf{w}^{(t)}), \mathbf{w}^{(t+1)} - \mathbf{w}^{(t)} \rangle + L B_h(\mathbf{w}^{(t+1)}, \mathbf{w}^{(t)}), \quad (\text{S1.14})$$

and μ -relative strong convexity (the lower bound in (5.14), applied at $(\mathbf{w}, \mathbf{w}^{(t)})$) gives

$$\langle \nabla L(\mathbf{w}^{(t)}), \mathbf{w} - \mathbf{w}^{(t)} \rangle \leq L(\mathbf{w}) - L(\mathbf{w}^{(t)}) - \mu B_h(\mathbf{w}, \mathbf{w}^{(t)}). \quad (\text{S1.15})$$

Writing

$$\langle \nabla L(\mathbf{w}^{(t)}), \mathbf{w}^{(t+1)} - \mathbf{w}^{(t)} \rangle = \langle \nabla L(\mathbf{w}^{(t)}), \mathbf{w}^{(t+1)} - \mathbf{w} \rangle + \langle \nabla L(\mathbf{w}^{(t)}), \mathbf{w} - \mathbf{w}^{(t)} \rangle$$

and bounding the two terms on the right by (S1.13) and (S1.15) respectively, then substituting into (S1.14), the two occurrences of $B_h(\mathbf{w}^{(t+1)}, \mathbf{w}^{(t)})$ cancel, leaving, for every $\mathbf{w} \in \mathcal{S}_{w_{\min}}$,

$$L(\mathbf{w}^{(t+1)}) \leq L(\mathbf{w}) + (L - \mu) B_h(\mathbf{w}, \mathbf{w}^{(t)}) - L B_h(\mathbf{w}, \mathbf{w}^{(t+1)}). \quad (\text{S1.16})$$

Stage 2 (single-step contraction to $\widehat{\mathbf{w}}$). Setting $\mathbf{w} = \widehat{\mathbf{w}}$ in (S1.16) and using global optimality $L(\mathbf{w}^{(t+1)}) \geq L(\widehat{\mathbf{w}})$ to drop the left-hand gap yields

$$B_h(\widehat{\mathbf{w}}, \mathbf{w}^{(t+1)}) \leq \left(1 - \frac{\mu}{L}\right) B_h(\widehat{\mathbf{w}}, \mathbf{w}^{(t)}). \quad (\text{S1.17})$$

Stage 3 (telescoping). Iterating (S1.17) from $t = 0$,

$$B_h(\widehat{\mathbf{w}}, \mathbf{w}^{(t)}) \leq \left(1 - \frac{\mu}{L}\right)^t B_h(\widehat{\mathbf{w}}, \mathbf{w}^{(0)}). \quad (\text{S1.18})$$

Stage 4 (objective-value bound). Inequality (S1.16) holds for every consecutive pair of iterates; applying it with index $t - 1$ in place of t and $\mathbf{w} = \widehat{\mathbf{w}}$, this time retaining the nonnegative gap $L(\mathbf{w}^{(t)}) - L(\widehat{\mathbf{w}}) \geq 0$ instead of dropping it, gives

$$L(\mathbf{w}^{(t)}) - L(\widehat{\mathbf{w}}) \leq (L - \mu) B_h(\widehat{\mathbf{w}}, \mathbf{w}^{(t-1)}). \quad (\text{S1.19})$$

Substituting the telescoped bound (S1.18) at index $t - 1$, $B_h(\widehat{\mathbf{w}}, \mathbf{w}^{(t-1)}) \leq (1 - \mu/L)^{t-1} B_h(\widehat{\mathbf{w}}, \mathbf{w}^{(0)})$, and using $(L - \mu) = L(1 - \mu/L)$, gives

$$L(\mathbf{w}^{(t)}) - L(\widehat{\mathbf{w}}) \leq L \left(1 - \frac{\mu}{L}\right)^t B_h(\widehat{\mathbf{w}}, \mathbf{w}^{(0)}), \quad t \geq 1, \quad (\text{S1.20})$$

which is (5.17) with $L = \bar{L}$ and $\mu = w_{\min} \underline{c}$. Since $0 < \mu \leq L$, the contraction factor $1 - \mu/L$ belongs to $[0, 1)$; if $\mu < L$, it belongs to $(0, 1)$. $\square$

Certification of the region-wise constants. For a fixed orthonormal tangent basis E , define $H_T(\mathbf{w}) = E^\top \nabla^2 L(\mathbf{w}) E$. The constants used in the theorem must bound the entire region $\mathcal{S}_{w_{\min}}$:

$$\bar{L} \geq \sup_{\mathbf{w} \in \mathcal{S}_{w_{\min}}} \lambda_{\max}(H_T(\mathbf{w})), \quad 0 < \underline{c} \leq \inf_{\mathbf{w} \in \mathcal{S}_{w_{\min}}} \lambda_{\min}(H_T(\mathbf{w})).$$

They may be obtained analytically or by a certified global numerical bound. Evaluating H_T only at $\widehat{\mathbf{w}}$ gives pointwise curvature and does not, by itself, certify these two region-wise extrema. Once valid bounds are available, the guaranteed factor is $1 - w_{\min} \underline{c} / \bar{L}$; using conservative bounds preserves the theorem, although it may weaken the numerical certificate.

S1.9. Proof of Theorem 6.1 (distribution-free conformal coverage)

Proof. Write $s_i = s(X_i, Y_i)$ for the calibration scores and $s_{\text{te}} = s(X_{\text{te}}, Y_{\text{te}})$ for the test score. Because s is fixed given the training fold and the $n_{\text{cal}} + 1$ pairs are exchangeable, the scores $s_1, \dots, s_{n_{\text{cal}}}, s_{\text{te}}$ are exchangeable. For the rank argument, attach independent continuous uniforms to break score ties lexicographically; the resulting test rank is exactly uniform on $\{1, \dots, n_{\text{cal}} + 1\}$.

Lower bound. By construction $Y_{\text{te}} \in C_\alpha(X_{\text{te}})$ iff $s_{\text{te}} \leq \hat{q}_\alpha$, and by (6.2) $\hat{q}_\alpha$ is the $\lceil (1 - \alpha)(n_{\text{cal}} + 1) \rceil$ -th element of the calibration scores augmented by $+\infty$. Thus

$$\{s_{\text{te}} \leq \hat{q}_\alpha\} \supseteq \{\text{rank}(s_{\text{te}}) \leq \lceil (1 - \alpha)(n_{\text{cal}} + 1) \rceil\},$$

where $\text{rank}(s_{\text{te}}) \in \{1, \dots, n_{\text{cal}} + 1\}$ is the randomized rank just described. By exchangeability, $\mathbb{P}(\text{rank}(s_{\text{te}}) \leq r) = r/(n_{\text{cal}} + 1)$ for every integer r . Taking $r = \lceil (1 - \alpha)(n_{\text{cal}} + 1) \rceil \geq (1 - \alpha)(n_{\text{cal}} + 1)$ gives

$$\mathbb{P}(Y_{\text{te}} \in C_\alpha(X_{\text{te}})) \geq \frac{\lceil (1 - \alpha)(n_{\text{cal}} + 1) \rceil}{n_{\text{cal}} + 1} \geq 1 - \alpha,$$

which is (6.4).

Upper bound under no ties. If the scores are almost surely distinct, the rank of s_{te} is exactly uniform on $\{1, \dots, n_{\text{cal}} + 1\}$, so $\mathbb{P}(\text{rank}(s_{\text{te}}) \leq r) = r/(n_{\text{cal}} + 1)$ and, with no ties, $\{s_{\text{te}} \leq \hat{q}_\alpha\} = \{\text{rank}(s_{\text{te}}) \leq r\}$ for $r = \lceil (1 - \alpha)(n_{\text{cal}} + 1) \rceil$. Hence the coverage equals $r/(n_{\text{cal}} + 1)$. Using $r < (1 - \alpha)(n_{\text{cal}} + 1) + 1$,

$$\mathbb{P}(Y_{\text{te}} \in C_\alpha(X_{\text{te}})) = \frac{r}{n_{\text{cal}} + 1} < \frac{(1 - \alpha)(n_{\text{cal}} + 1) + 1}{n_{\text{cal}} + 1} = 1 - \alpha + \frac{1}{n_{\text{cal}} + 1},$$

which together with the lower bound gives (6.5). $\square$

S2. Supplementary examples and robustness checks

This section collects worked examples and secondary numerical robustness checks that accompany the main text but are not needed to follow its theorems or proofs. Each subsection is pointed to explicitly from the corresponding location in the main article.

S2.1. Worked example: two Gaussians (companion to Section 3.1)

Example S2.1. (Two Gaussians: what the bound says) Let $p = \mathcal{N}(0, 1)$ and $q = \mathcal{N}(d, 1)$ on $\mathbb{R}$. Then $\text{BC}(p, q) = e^{-d^2/8}$, so for $d = 1$,

$$\text{BC} = e^{-1/8} = 0.8825, \quad D_B = -\ln \text{BC} = 0.125.$$

Under equal priors, Eq (3.5) gives the floor $\frac{1}{2}(1 - \sqrt{1 - \text{BC}^2}) = 0.2648$, while the exact Bayes error is $\Phi(-d/2) = 0.3085$. Under the hunger-node priors $(\alpha, 1 - \alpha) = (0.366, 0.634)$, (3.4) gives 0.2368 against an exact Bayes error of 0.2834.

Reading. The bound is a floor on the error of any acoustic-only classifier (here 0.2648, respectively, 0.2368), not a prediction of the observed error, and an observed error above it is not evidence against it. At the root node it certifies $R_{a,\text{root}}^* \geq 0.235$ against a trimodal test error of 0.084, which is the sense in which fusion is necessary here.

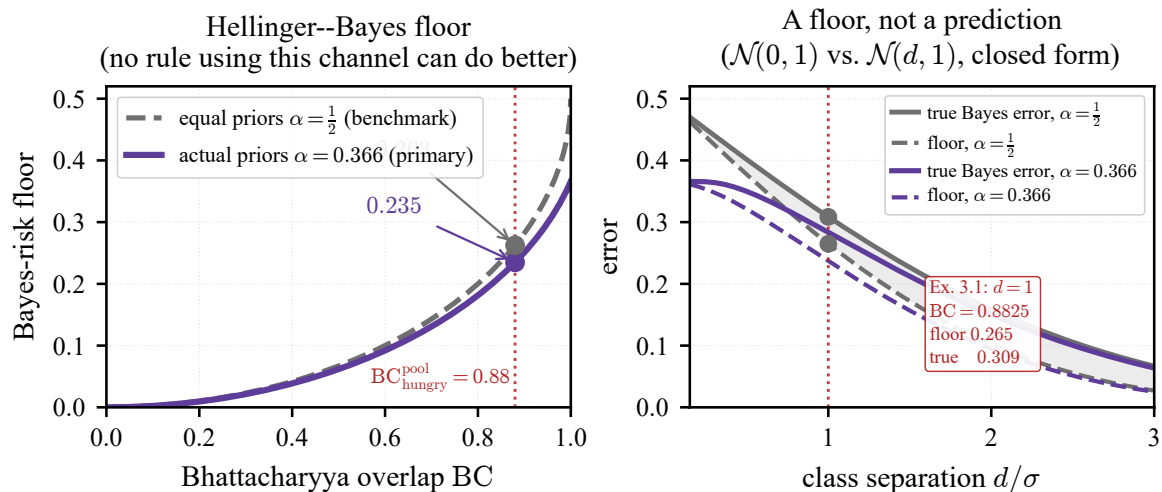


Figure S1. Left: the Hellinger–Bayes error floor $\frac{1}{2}(1 - \sqrt{1 - 4\alpha(1 - \alpha)BC^2})$ versus the Bhattacharyya coefficient BC, at the actual hunger-node prior ($\alpha = 0.366$, primary) and the equal-prior benchmark ($\alpha = \frac{1}{2}$), with $BC_{\text{hungry}}^{\text{pool}} = 0.88$ marked (floors 0.235 and 0.263, Table 3). Right: the closed-form Gaussian instance of Example S2.1 ($\mathcal{N}(0, 1)$ vs. $\mathcal{N}(d, 1)$), showing the floor against the true Bayes error across separations d/σ ; at $d = 1$, BC = 0.8825, floor 0.265, true 0.309.

S2.2. Worked example: finite-step learning on six observations (companion to §5.2)

Example S2.2. (Entropic mirror learning on six observations) Take $K = 2$, $M = 2$, $\alpha_0 = 1$, $n = 6$, with class-1 modality posteriors

$$\begin{array}{lcl} p_1(1 | x_i) : & 0.88 & 0.82 & 0.78 & 0.18 & 0.22 & 0.60 \\ p_2(1 | x_i) : & 0.72 & 0.68 & 0.62 & 0.38 & 0.34 & 0.15 \\ y_i : & 1 & 1 & 1 & 2 & 2 & 2 \end{array}$$

The observed sufficient statistics are $t_1 = -1.9379$, $t_2 = -2.2483$.

On $\mathcal{S}_{0.25}$ the curvature constants of Theorem 5.3 are $\bar{L} = 0.8727$, $\underline{c} = 0.7842$, so $\mu = w_{\min}\underline{c} = 0.1960$ and the certified contraction is

$$1 - \mu/\bar{L} = 0.775;$$

the update runs with

$$\gamma = 1/\bar{L} = 1.1459$$

from $\mathbf{w}^{(0)} = (1/2, 1/2)$. Table S1 tracks the iteration explicitly: at each t it lists $w_1^{(t)}, w_2^{(t)}$, the model-expected sufficient statistics $s_1^{(t)}, s_2^{(t)}$, the observed log-likelihood $\ell_{\text{obs}}(\mathbf{w}^{(t)})$, and the remaining gap $\ell_{\text{obs}}(\hat{\mathbf{w}}) - \ell_{\text{obs}}(\mathbf{w}^{(t)})$.

Table S1. Finite-step iterations for Example S2.2 (hand-checkable). The gap column is $\ell_{\text{obs}}(\widehat{\mathbf{w}}) - \ell_{\text{obs}}(\mathbf{w}^{(t)})$ with $\ell_{\text{obs}}(\widehat{\mathbf{w}}) = -1.853617$.

t	$w_1^{(t)}$	$w_2^{(t)}$	$s_1^{(t)}$	$s_2^{(t)}$	$\ell_{\text{obs}}(\mathbf{w}^{(t)})$	(gap)
0	0.5000	0.5000	-3.6699	-3.6559	-1.884108	(0.030491)
1	0.5919	0.4081	-3.5443	-3.6884	-1.861555	(0.007938)
2	0.6369	0.3631	-3.4840	-3.7065	-1.855828	(0.002211)
3	0.6599	0.3401	-3.4537	-3.7161	-1.854272	(0.000656)
4	0.6721	0.3279	-3.4376	-3.7214	-1.853819	(0.000202)
5	0.6787	0.3213	-3.4289	-3.7243	-1.853681	(0.000064)
6	0.6825	0.3175	-3.4241	-3.7260	-1.853637	(0.000020)

The iterates converge to $\widehat{\mathbf{w}} = (0.6873, 0.3127)$ with $\ell_{\text{obs}}(\widehat{\mathbf{w}}) = -1.853617$, and four features of §5 are visible on hand-checkable numbers:

(1) Monotone ascent (Proposition 5.1): the sixth column of Table S1 increases at every step and the gap decreases monotonically to zero; (2) An interior maximiser with a strict gain over either modality, since $\ell_{\text{obs}}(1, 0) = t_1$ and $\ell_{\text{obs}}(0, 1) = t_2$ are both below -1.8536 ; (3) The fixed point requires the tangential component of $t - s$ to vanish, not $t - s$ itself (here $t - s^{(6)} = (1.4862, 1.4777)$, still converging); (4) The empirical objective-gap ratios over the displayed transient iterates (0.26–0.32) sit comfortably inside the certified 0.775.

These finite-iterate ratios are not the fixed-point spectral factor q_γ of Proposition 5.2. They illustrate the structural conservatism of Remark 5.2. With $n = 6$ the tangent observed information is 0.8727, so the observed-information standard error is far too large to be useful; the same computation on the case-study fold ($n_{\text{train}} = 663$) produces the standard errors of §7.3. The example displays the mechanics, not the precision.

S2.3. Derivation: near-redundancy versus exact duplication (companion to Remark 4.3)

This derivation supports Remark 4.3 (iii) of the main article, which states that near-redundancy does not break log-posterior independence whereas exact duplication does.

Suppose $M = 3$ and modality 2 is a temperature rescaling of modality 1, $\log p_2(k | x_{i,2}) = \tau \log p_1(k | x_{i,1}) + c_i$ with c_i independent of k . For a tangent $\mathbf{v}$ with $\mathbf{1}^\top \mathbf{v} = 0$,

$$\mathbf{v}^\top \boldsymbol{\ell}_i(k) = (v_1 + \tau v_2) \log p_1(k | x_{i,1}) + v_3 \log p_3(k | x_{i,3}) + v_2 c_i.$$

If modalities 1 and 3 are not themselves affinely related in k modulo k -constants, this is k -constant for every i only if $v_1 + \tau v_2 = 0$ and $v_3 = 0$; with $v_1 + v_2 + v_3 = 0$ this forces $v_2(\tau - 1) = 0$, hence $\mathbf{v} = 0$ unless $\tau = 1$. Under the stated non-affine condition on modalities 1 and 3, this specific mechanism therefore creates a flat tangent direction only at $\tau = 1$. This is an illustration, not an exhaustive characterisation: other exact multi-expert linear dependencies are detected by the full-rank condition on Φ_T .

S2.4. Robustness to modality dependence: full results (companion to §7.7)

A controlled synthetic stress test quantifies the cost of ignoring class-conditional dependence (Proposition 2.3). We draw $N = 40,000$ observations from a $K = 6$, $M = 3$ Gaussian model with equal

class priors and unit marginal variances; six class-mean vectors are fixed once (scaled by 0.8) and held constant so that only the dependence varies. Each class carries an equicorrelation matrix, and we record the excess risk $R(\phi^{\otimes}) - R^*$ of fusing as though the modalities were independent. Two regimes are swept: *homogeneous* (all classes share one ρ) and *heterogeneous* (the ρ_k spread linearly about 0.2, holding total dependence roughly fixed while its variation across classes grows).

Table S2. Dependence stress test ($N = 40,000$; synthetic Gaussian model). TC is the mean class-conditional total correlation; $\delta_{0.5}, \delta_{0.95}$ its copula-spread quantiles; the last two columns are the certificates of Proposition 2.3(iv),(iii).

Regime	Level	TC	$\delta_{0.5}$	$\delta_{0.95}$	$\mathbb{E}[\text{TV}]$	excess risk	$\sqrt{2\text{TC}}$	tail bd.
homogeneous	$\rho = 0.0$	0.0000	0.000	0.000	0.0000	0.0000	0.000	0.000
homogeneous	$\rho = 0.1$	0.0142	0.281	0.551	0.0313	0.0013	0.169	0.465
homogeneous	$\rho = 0.2$	0.0549	0.560	1.081	0.0619	0.0049	0.331	0.990
homogeneous	$\rho = 0.3$	0.1217	0.853	1.645	0.0930	0.0108	0.493	1.000
homogeneous	$\rho = 0.4$	0.2169	1.183	2.315	0.1258	0.0195	0.659	1.000
heterog	spr. 0.2	0.0610	0.567	1.535	0.0683	0.0081	0.349	1.000
heterog	spr. 0.4	0.0797	0.739	2.496	0.0857	0.0165	0.399	1.000
heterog	spr. 0.6	0.1130	1.052	4.106	0.1111	0.0300	0.475	1.000

Four conclusions follow:

(1) Up to $\rho = 0.2$ the cost of ignoring dependence is at most 0.0049 in risk; the degradation is smooth, with no threshold over this range.

(2) Heterogeneity, not magnitude, is what costs: at lower total dependence (TC = 0.1130) the heterogeneous regime costs 0.0300 against 0.0108 at the higher TC = 0.1217 homogeneous value — a ratio of 2.77, the empirical face of Proposition 2.3.

(3) The adopted certificate (iv) holds in every cell with a median slack of 43×; we report the slack rather than suppress it.

(4) The tail certificate (iii) saturates at 1 for $\rho \geq 0.3$ and is vacuous there, which is why (iv) is primary. In the case study the acoustic pairs satisfy $|\rho_s| \leq 0.19$ and sit at or below the $\rho = 0.2$ row; the biological–video pair is larger ($|\rho_s| \leq 0.54$, mean 0.24 over the 25 admissible cells) and is the regime this analysis was built to quantify.

S2.5. Comparative weighting strategies: full results (companion to Section 7.9)

Table S3 descriptively compares the finite-step likelihood-fitted weights against three alternatives on the identical split of (7.1): uniform weights $w_m \equiv 1/M$; held-out-accuracy-proportional weights; and an additive mixture-of-experts gate. The fitted weights are the unique maximiser of the observed likelihood under this model (Theorem 4.2); the three alternatives carry no corresponding optimality property, so the comparison is between the likelihood-optimal weighting and three weightings without a corresponding guarantee. Because the additive gate is not a nested submodel of RW-PoE, a common likelihood-ratio test across these strategies would not be valid; no such inferential claim is made.

Table S3. Comparative weighting strategies on the identical train/calibration/test split of (7.1) ($n_{\text{test}} = 222$). Values are descriptive.

Strategy	Accuracy	κ
Likelihood-fitted (RW-PoE)	0.916	0.889
Uniform $w_m \equiv 1/M$	≈ 0.878	≈ 0.842
Accuracy-proportional	≈ 0.889	≈ 0.858
MoE additive gate	≈ 0.865	≈ 0.824



AIMS Press

© 2026 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0>)