TOWARDS BREGMAN LOSS WITH AITCHISON-INSPIRED POTENTIAL FOR LABEL DISTRIBUTION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Label Distribution Learning (LDL) represents label ambiguity using simplex-valued vectors over the label space, referred to as label distributions. In most applications, the observed training label distribution is aggregated from a finite number of annotations and can, under an unbiased sampling mechanism, be viewed as a finite-sample estimate of an underlying population label distribution rather than an error-free target. Under this assumption, the conditional arithmetic mean of label distributions equals the underlying population label distribution. Consequently, an LDL point prediction should approximate the conditional arithmetic mean, rather than an individual realization. This motivates strict arithmetic-mean properness of the training loss, requiring the conditional arithmetic mean to be its unique conditional Bayes prediction. Although the widely used Kullback-Leibler divergence and squared Euclidean loss satisfy this property, under restricted model capacity or data noise, small errors in probability space can still induce substantial distortions in relative label ratios, particularly near the simplex boundary. Such distortions can be detrimental when relative label proportions are critical to downstream decisions. We therefore investigate how to reconcile mean-properness with ratio-sensitivity in LDL. Since the squared Aitchison loss directly captures relative variation through log-ratios but is generally not arithmetic-mean proper, we seek a mean-proper loss with Aitchison local geometry as its reference. Specifically, we propose a Bregman loss with Aitchison-inspired potential, which is derived by matching the coordinatewise curvature scaling of its potential function to local Aitchison geometry. We establish its mean-properness, quantitatively compare its local geometry with Aitchison, and show its merits over conventional LDL losses. Finally, we conduct extensive experiments to show the effectiveness of our method.

1 INTRODUCTION

Label Distribution Learning (LDL) (Geng, 2016) is a supervised learning paradigm tailored for scenarios characterized by label ambiguity. In contrast to conventional hard-label learning that assigns deterministic labels to each instance, LDL characterizes an instance using a label distribution that is formally defined as a simplex-valued vector over the label space. Each element of the label distribution, referred to as the label mass[1], quantifies the degree to which the corresponding label describes the instance. Owing to the rich ambiguity information, LDL has been widely applied in tasks such as age estimation (Hou et al., 2017), facial expression (Zhou et al., 2022), emotion recognition (Jia et al., 2019), and other tasks involving label ambiguity.

A substantial body of research on LDL has focused on designing loss functions for comparing predicted and observed label distributions (Wang & Geng, 2023; Jia et al., 2023). We characterize an LDL loss along two complementary dimensions: its statistical target and its local geometry. The statistical target is the functional of the conditional law of observed label distributions whose value minimizes the conditional expected loss, and therefore specifies what population quantity the loss elicits, whereas the local geometry determines how the loss weights different deviations when a model approximates the statistical target. This distinction separates what an LDL model is intended to predict from how the model is guided toward that prediction.

[1]Although numerous previous studies typically refer to the elements within a label distribution as "label description degree", we employ the more concise term "label mass" throughout this paper.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

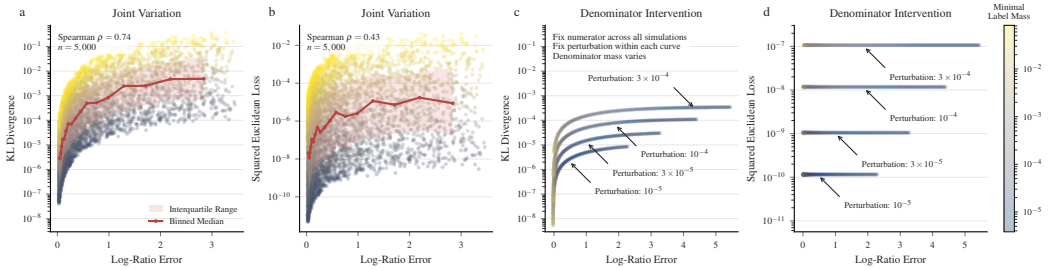

Figure 1: **Correlation between Log-Ratio Error and Conventional Losses**: (a) and (b) show the relationship of log-ratio error ($d_A(\boldsymbol{d}, \boldsymbol{p})$ from Equation (2)) with KL divergence and SE loss under controlled small target label masses. Red lines and bands denote medians and interquartile ranges, respectively. (c) and (d) show the same relationship under controlled denominator intervention.

The statistical target is particularly important when a training label distribution is aggregated from a finite number of annotations, measurements, or sampled events (Geng & Hou, 2015; Ren & Geng, 2017; Yang et al., 2017), which can, under an unbiased aggregation mechanism, be interpreted as a finite-sample estimate of an underlying population label distribution rather than as an error-free target. Formally, let $\boldsymbol{D}$ denote the random label distribution associated with an instance $\boldsymbol{x}$, then the training label distribution is one realization of a conditional law $\mathcal{P}_{\boldsymbol{x}}$ over observed label distributions of the instance $\boldsymbol{x}$. Under the same unbiased mechanism, $\mathbb{E}_{\boldsymbol{D}\sim\mathcal{P}_{\boldsymbol{x}}}[\boldsymbol{D}]$ equals the underlying population label distribution. Therefore, when LDL model needs to output a point in probability simplex, we desire the point prediction to approximate the conditional arithmetic mean $\boldsymbol{\pi}(\boldsymbol{x}) = \mathbb{E}[\boldsymbol{D} \mid \boldsymbol{x}]$ rather than a realization of $\mathcal{P}_{\boldsymbol{x}}$. Hence, the training loss ℓ is required to possess the strict arithmetic-mean properness (or briefly mean-properness): the unique conditional Bayes prediction $\arg\min_{\boldsymbol{p}} \mathbb{E}_{\boldsymbol{D}\sim\mathcal{P}_{\boldsymbol{x}}}[\ell(\boldsymbol{D}, \boldsymbol{p})]$ should be $\boldsymbol{\pi}(\boldsymbol{x})$, which identifies the pointwise Bayes target under unrestricted predictions.

Kullback-Leibler (KL) divergence and squared Euclidean (SE) loss, serving as the fundamental loss term of most existing LDL methods, possess the mean-properness. With unrestricted pointwise predictions, minimizing the population conditional risk of either loss recovers the conditional arithmetic mean. In practical tasks, however, the restricted model capacity or data noise generally precludes exact recovery. Under such constraints, we find that the local geometries of KL divergence and SE loss encourage a small aggregate discrepancy in probability coordinates while allowing substantial distortions in label-mass ratios. As shown in Figure 1, small KL divergence or squared-Euclidean error can coexist with large label-ratio error, particularly when the minimal label mass is small. However, in most practical applications, label distributions encode not only absolute label mass but also relative importance across labels (Geng, 2016), and the distortions in the relative label ratio (a measurement of relative label importance) can be detrimental, even when the affected labels have small masses. Breast-cancer tumour-microenvironment deconvolution provides a concrete example, where a model maps a bulk RNA-seq expression profile to a distribution over tumour, immune, and stromal cell types (Menden et al., 2020). Under the unbiased finite-cell sampling, the arithmetic conditional mean represents the expected cellular composition associated with the expression profile and is therefore relevant to estimating tumour purity and cell-type proportions. Simultaneously, biological interpretation can depend on relative abundances, such as the $CD8^+$/eTreg ratio associated with treatment response (Goda et al., 2022; Tran et al., 2023).

Therefore, this work investigates how to design a mean-proper loss function with ratio-sensitivity. The squared Aitchison loss is a natural candidate for capturing the relative importance of labels, as it measures discrepancies directly in centered log-ratio coordinates and thus explicitly characterizes relative variation in compositional data (Aitchison, 1982; Egozcue et al., 2003). However, its conditional Bayes prediction does not coincide with the arithmetic mean. We therefore utilize the squared Aitchison loss as a reference criterion for quantifying distortions in relative label importance and investigate how Aitchison local geometry can guide the design of a mean-proper loss. On the one hand, we analyze why the considered loss functions do not guarantee mean-properness and ratio-sensitivity simultaneously, which theoretically motivates this work. Specifically, we first theoretically characterize the mechanism by which KL divergence and SE loss control log-ratio error, demonstrating that such control requires the relevant label masses to be bounded away from zero and

is not uniform near the simplex boundary. We further prove that the conditional Bayes prediction of squared Aitchison loss generally differs from the conditional arithmetic mean. On the other hand, we propose a mean-proper LDL loss function with local ratio-sensitivity, named BAP (Bregman Loss with Aitchison-Inspired Potential). Specifically, we first employ existing characterization results to show that every regular mean-proper LDL loss admits a Bregman divergence representation. Within this representation, we then show a regular potential function that matches the coordinatewise curvature scaling of local Aitchison loss, and introduce a cap operation to prevent the curvature and derivative near zero from diverging, thereby generating a specific form of Bregman divergence. Besides, we establish the mean-properness of BAP, quantitatively compare the local geometry of BAP and Aitchison loss, and theoretically show the merits of BAP over KL divergence and SE loss. Finally, we conduct extensive experiments to demonstrate the effectiveness of our proposal.

2 Related Work

Our work primarily relates to label distribution learning. The original LDL formulation directly learns a mapping from instances to complete label distributions (Geng, 2016). Recent work has adapted this formulation to task-specific structures and evaluation requirements. Margin-based methods align distribution learning with downstream classification (Wang & Geng, 2021; Wang et al., 2022), ranking-aware methods preserve the order induced by label masses (Jia et al., 2023), and graph LDL jointly models instance-instance, label-label, and instance-label relations (Jin et al., 2024). More recent works address entropy-dependent performance bias or replace small numerical discrepancies with an approximately-correct criterion (Lu et al., 2025b; Li et al., 2025). These methods improve the model, regularizer, or evaluation criterion while retaining each observed distribution as the deterministic target. Another line of work explicitly relaxes the availability or reliability of numerical label distributions. Incomplete LDL reconstructs missing label masses using low-rank, neighborhood, or graph-regularized structure (Xu & Zhou, 2017; Xu et al., 2021). Other methods infer latent distributions from coarse-grained labels (Lu & Jia, 2024; Lu et al., 2023; Lu & Jia, 2022), separate instance-dependent corruption from latent clean distributions (Kou et al., 2025), or extract reliable coarse-grained supervision before refining it (Lu et al., 2025a). Besides, it is worth noting the connection between our work and previous studies on the Itakura-Saito (IS) divergence (Banerjee et al., 2005b), since the negative-log potential of BAP above the cap threshold coincides with the classical IS divergence. First, IS uses a negative-log potential that is singular at zero, whereas BAP replaces this singularity with a smooth quartic continuation below a prescribed threshold, thereby yielding a bounded loss over the closed simplex. Second, beyond previous studies on IS, our work explicitly analyzes the distortion of relative ratios between labels, and quantifies this sensitivity through a local comparison with Aitchison geometry. Third, we establish that the capped construction preserves mean-properness including for label distributions containing zero components.

3 Methodology

Prior to presenting the methodological details, we articulate the goal of our method. Let $\Delta^m = \{\boldsymbol{d} \in \mathbb{R}^m : \forall i, d_i \geq 0, \mathbf{1}^\top \boldsymbol{d} = 1\}$ denote the closed probability simplex. The training set $\{(\boldsymbol{x}_i, \boldsymbol{d}_i)\}_{i=1}^n$ consists of n independent realizations of a random vector pair $(\boldsymbol{X}, \boldsymbol{D})$ taking values in $\mathcal{X} \times \Delta^m$, where $\boldsymbol{X}$ and $\boldsymbol{D}$ denote the random vectors of features and label distribution, respectively. For each feature vector $\boldsymbol{x} \in \mathcal{X}$, let $\mathcal{P}_{\boldsymbol{x}}$ be a probability distribution over Δ^m such that $\boldsymbol{D}|\boldsymbol{x} \sim \mathcal{P}_{\boldsymbol{x}}$. The goal of our method is to train a mapping $f : \mathcal{X} \to \Delta^m$ from features to label distribution, thereby accurately predicting $\mathbb{E}_{\boldsymbol{D} \sim \mathcal{P}_{\boldsymbol{x}}}[\boldsymbol{D}]$ for unseen instance $\boldsymbol{x}$. In the subsequent subsections, we analyze the limitations of related loss functions, introduce the proposed BAP loss, and analyze its properties.

3.1 Limitation Analysis of Related Loss Functions

We begin by presenting Proposition 3.1 to elucidate the mechanisms by which the widely-used KL divergence and SE loss control relative abundance information.

Proposition 3.1 (Control of Log-Ratio by KL Divergence and SE Loss)**.** *Fix $i, j \in [m]$ with $i \neq j$. If $\boldsymbol{\pi}, \boldsymbol{p} \in \operatorname{int}(\Delta^m)$ and the label masses π_i, π_j, p_i, p_j are at least $\delta > 0$, then*

$$\left| \log \frac{\pi_i}{\pi_j} - \log \frac{p_i}{p_j} \right| \leq \frac{\sqrt{2}}{\delta} \|\boldsymbol{\pi} - \boldsymbol{p}\|_2, \qquad \left| \log \frac{\pi_i}{\pi_j} - \log \frac{p_i}{p_j} \right| \leq \frac{\sqrt{2}}{\delta} \sqrt{\mathcal{D}_{\mathrm{KL}}(\boldsymbol{\pi} \| \boldsymbol{p})}. \tag{1}$$

Proposition 3.1 establishes a lack of uniform control of KL divergence and SE loss: small probability-space error can coexist with severe relative ratio distortion when the corresponding label masses approach the simplex boundary, i.e., $\delta \to 0$. The proof is given in Section A.2.

In order to measure the relative abundance information, a natural reference is Aitchison geometry, which measures the squared Euclidean distance after mapping label distributions into the centered log-ratio (CLR) coordinates. The squared Aitchison loss is typically used to penalize log-ratio error:

$$d_A^2(\boldsymbol{d}, \boldsymbol{p}) = \frac{1}{m} \sum_{1 \le i < j \le m} \left(\log \frac{d_i/d_j}{p_i/p_j} \right)^2 = \|\psi(\boldsymbol{d}) - \psi(\boldsymbol{p})\|_2^2, \quad \psi(\boldsymbol{p}) = (\boldsymbol{I}_m - m^{-1}\mathbf{1}\mathbf{1}^\top) \log \boldsymbol{p}, \tag{2}$$

where $\boldsymbol{p}, \boldsymbol{d} \in \operatorname{int}(\Delta^m)$ denote the predicted and target label distributions, respectively. However, the squared Aitchison loss elicits the conditional Fréchet mean under the Aitchison metric (the Aitchison center), rather than the conditional arithmetic mean. More specifically, we first define the concept of strict arithmetic-mean properness in Definition 3.1, and elucidate the relationship between the mean-properness and the Bayes prediction of squared Aitchison loss in Proposition 3.2.

Definition 3.1 (Strict Arithmetic-Mean Properness)**.** *Let $\Delta^m = \{\boldsymbol{p} \in \mathbb{R}^m_{\ge 0} : \mathbf{1}^\top \boldsymbol{p} = 1\}$ denote the probability simplex, and let $\mathcal{X}$ denote the feature space. Let $\ell : \Delta^m \times \Delta^m \to \mathbb{R} \cup \{+\infty\}$ be a loss function, where the first argument is an observed label distribution and the second argument is a predicted label distribution. For each $\boldsymbol{x} \in \mathcal{X}$, let $\mathcal{P}_{\boldsymbol{x}}$ be an admissible distribution over Δ^m, interpreted as the conditional distribution of the random label distribution $\boldsymbol{D}$ given $\boldsymbol{x}$. Define the conditional risk by $\mathcal{R}_\ell(\boldsymbol{p}|\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{D} \sim \mathcal{P}_{\boldsymbol{x}}}[\ell(\boldsymbol{D}, \boldsymbol{p})]$. The loss ℓ is said to be strictly conditionally arithmetic-mean proper if $\arg\min_{\boldsymbol{p} \in \Delta^m} \mathcal{R}_\ell(\boldsymbol{p}|\boldsymbol{x}) = \{\mathbb{E}_{\boldsymbol{D} \sim \mathcal{P}_{\boldsymbol{x}}}[\boldsymbol{D}]\}$ holds for every admissible conditional distribution $\mathcal{P}_{\boldsymbol{x}}$ and every $\boldsymbol{x} \in \mathcal{X}$.*

Proposition 3.2 (Bayes Prediction of Squared Aitchison Loss)**.** *Fix $\boldsymbol{x} \in \mathcal{X}$, and let $\boldsymbol{D} \sim \mathcal{P}_{\boldsymbol{x}}$ be a random label distribution that belongs to the interior of Δ^m almost surely and whose CLR representation is square-integrable. For any strictly positive $\boldsymbol{p} \in \Delta^m$, define the conditional squared Aitchison risk as $\mathcal{R}_{\mathrm{A}}(\boldsymbol{p}|\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{D} \sim \mathcal{P}_{\boldsymbol{x}}}[\|\psi(\boldsymbol{D}) - \psi(\boldsymbol{p})\|_2^2]$. Then $\mathcal{R}_{\mathrm{A}}(\boldsymbol{p}|\boldsymbol{x})$ admits the unique Bayes predictor $\boldsymbol{p}^\star_{\mathrm{A}}(\boldsymbol{x}) = \varpi(\mathbb{E}_{\boldsymbol{D} \sim \mathcal{P}_{\boldsymbol{x}}}[\log \boldsymbol{D}])$, where $\varpi(\cdot)$ denotes the softmax function. Let $\boldsymbol{\pi}(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{D} \sim \mathcal{P}_{\boldsymbol{x}}}[\boldsymbol{D}]$ denote the conditional arithmetic mean. Then $\boldsymbol{p}^\star_{\mathrm{A}}(\boldsymbol{x}) = \boldsymbol{\pi}(\boldsymbol{x})$ if and only if $(\boldsymbol{I}_m - m^{-1}\mathbf{1}\mathbf{1}^\top)(\mathbb{E}_{\boldsymbol{D} \sim \mathcal{P}_{\boldsymbol{x}}}[\log \boldsymbol{D}] - \log \boldsymbol{\pi}(\boldsymbol{x})) = \mathbf{0}$. Equivalently, the coordinatewise Jensen gaps $\log \pi_i(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{D} \sim \mathcal{P}_{\boldsymbol{x}}}[\log D_i]$ are identical across all labels.*

Proposition 3.2 identifies a target shift: the Bayes prediction of squared Aitchison loss is the inverse CLR image of the conditional mean in log space, rather than the conditional arithmetic mean in probability space. The Bayes prediction equals the conditional arithmetic mean if and only if $\log \pi_i(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{D} \sim \mathcal{P}_{\boldsymbol{x}}}[\log D_i]$ is the same constant for all labels. See Section A.1 for the proof. Overall, Propositions 3.1 and 3.2 show that none of KL divergence, SE loss, or squared Aitchison loss simultaneously ensures mean-properness and uniform control of log-ratio distortion.

3.2 BAP: Bregman Loss with Aitchison-Inspired Potential

To derive a mean-proper loss with greater local sensitivity to relative prediction errors, we begin with a family of losses for which arithmetic-mean properness holds by construction. Within this family, we then seek a particular loss whose local curvature is guided by log-ratio errors.

Existing characterization results establish that losses eliciting a linear property of a distribution must be equivalent to a Bregman divergence under suitable regularity conditions (Banerjee et al., 2005a; Abernethy & Frongillo, 2012). We specialize these results to our problem settings, where the elicited property is the conditional arithmetic mean of a random label distribution finitely-supported on the simplex, which implies that every regular[2] mean-proper loss is generated by a strictly convex potential. Consequently, the local curvature of excess risk, i.e., the additional expected loss incurred by a prediction compared with the Bayes-optimal prediction, of any such loss must be determined by the Hessian of its generating potential. The formal specialization, including its precise conditions, is stated and proved in Proposition 3.3.

Proposition 3.3 (Bregman Representation of Regular Mean-Proper Losses)**.** *Let $m \ge 2$ and $r = m - 1$. Define $\mathcal{U} = \left\{\boldsymbol{\eta} \in \mathbb{R}^r : \forall i, \eta_i > 0, \mathbf{1}_r^\top \boldsymbol{\eta} < 1\right\}$, $q(\boldsymbol{\eta}) = [\boldsymbol{\eta}; 1 - \mathbf{1}_r^\top \boldsymbol{\eta}]$. Note that q is an affine*

[2]We call a loss regular if it satisfies the continuity and prediction-side differentiability stated in Proposition 3.3.

bijection from $\mathcal{U}$ to $\mathrm{int}(\Delta^m)$. *Let* $\ell : \mathrm{int}(\Delta^m) \times \mathrm{int}(\Delta^m) \to \mathbb{R}$ *be a loss whose first argument is an observed label distribution and whose second argument is a predicted label distribution. For* $\boldsymbol{\xi}, \boldsymbol{\eta} \in \mathcal{U}$, *define the normalized coordinate loss* $L(\boldsymbol{\xi}, \boldsymbol{\eta}) = \ell(q(\boldsymbol{\xi}), q(\boldsymbol{\eta})) - \ell(q(\boldsymbol{\xi}), q(\boldsymbol{\xi}))$. *Assume that* L *is jointly continuous, that* $\boldsymbol{\eta} \mapsto L(\boldsymbol{\xi}, \boldsymbol{\eta})$ *is twice continuously differentiable for every* $\boldsymbol{\xi}$, *and that its first and second derivatives with respect to* $\boldsymbol{\eta}$ *are jointly continuous in* $(\boldsymbol{\xi}, \boldsymbol{\eta})$. *For every integer* $S \geq 1$, *every collection* $\boldsymbol{d}_1, \ldots, \boldsymbol{d}_S \in \mathrm{int}(\Delta^m)$, *and all weights* $\lambda_1, \ldots, \lambda_S > 0$ *satisfying* $\sum_{s=1}^S \lambda_s = 1$, *define the probability law* $Q = \sum_{s=1}^S \lambda_s \delta(\boldsymbol{d}_s)$, *where* $\delta(\boldsymbol{d}_s)$ *is the point mass at* $\boldsymbol{d}_s$. *Let* $\boldsymbol{D} \sim Q$ *denote a random label distribution with law* Q. *For* $\boldsymbol{p} \in \mathrm{int}(\Delta^m)$, *define* $\mathcal{R}_Q(\boldsymbol{p}) = \mathbb{E}_Q[\ell(\boldsymbol{D}, \boldsymbol{p})]$ *and* $\boldsymbol{\pi}_Q = \mathbb{E}_Q[\boldsymbol{D}]$. *Assume* $\arg\min_{\boldsymbol{p} \in \mathrm{int}(\Delta^m)} \mathcal{R}_Q(\boldsymbol{p}) = \{\boldsymbol{\pi}_Q\}$, *then there exists a strictly convex potential* $\Phi \in \mathcal{C}^3(\mathcal{U})$ *such that* $L(\boldsymbol{\xi}, \boldsymbol{\eta}) = B_\Phi(\boldsymbol{\xi}, \boldsymbol{\eta})$ *holds for* $\forall \boldsymbol{\xi}, \boldsymbol{\eta} \in \mathcal{U}$, $B_\Phi(\boldsymbol{\xi}, \boldsymbol{\eta})$ *is Bregman divergence:*

$$B_\Phi(\boldsymbol{\xi}, \boldsymbol{\eta}) = \Phi(\boldsymbol{\xi}) - \Phi(\boldsymbol{\eta}) - \langle \nabla\Phi(\boldsymbol{\eta}), \boldsymbol{\xi} - \boldsymbol{\eta} \rangle, \tag{3}$$

where $\mathcal{C}^3(\mathcal{U})$ *is the class of real-valued functions on* $\mathcal{U}$ *with continuous partial derivatives up to order three. Hence, for every such probability law* Q, *let* $\boldsymbol{\eta}_Q = q^{-1}(\boldsymbol{\pi}_Q)$ *and* $\boldsymbol{p} = q(\boldsymbol{\eta})$ *for* $\boldsymbol{\eta} \in \mathcal{U}$, *then*

$$\mathcal{R}_Q(q(\boldsymbol{\eta})) - \mathcal{R}_Q(q(\boldsymbol{\eta}_Q)) = B_\Phi(\boldsymbol{\eta}_Q, \boldsymbol{\eta}). \tag{4}$$

Proposition 3.3 converts the design of a mean-proper loss into the design of its convex potential. Once this potential is obtained, we can construct a loss function using the Bregman divergence in Equation (3). We next derive the desired potential. Although the representation permits non-separable potentials, we consider a separable potential $\Phi(\boldsymbol{u}) = \sum_{i=1}^m \phi(u_i)$ for the sake of simplicity, where the function $\phi(\cdot)$ determines how different values of prediction errors are weighted. For an interior target distribution $\boldsymbol{\pi}$ and a sufficiently small perturbation $\boldsymbol{h}$ satisfying $\mathbf{1}_m^\top \boldsymbol{h} = 0$, the Bregman divergence generated by Φ has the local expansion

$$B_\Phi(\boldsymbol{\pi}, \boldsymbol{\pi} + \boldsymbol{h}) = \frac{1}{2} \sum_{i=1}^m \phi''(\pi_i) h_i^2 + o(\|\boldsymbol{h}\|_2^2), \quad \text{as } \boldsymbol{h} \to \mathbf{0}. \tag{5}$$

The detailed derivation of Equation (5) is provided in Section A.4. Equation (5) shows that the curvature $\phi''(\pi_i)$ determines the penalty assigned to a small error in label i. For two labels i and j, the squared log-ratio distortion induced by a sufficiently small perturbation $\boldsymbol{h}$ satisfies

$$\left(\log \frac{\pi_i + h_i}{\pi_j + h_j} - \log \frac{\pi_i}{\pi_j}\right)^2 = \left(\frac{h_i}{\pi_i} - \frac{h_j}{\pi_j}\right)^2 + o(\|\boldsymbol{h}\|_2^2) = \frac{h_i^2}{\pi_i^2} + \frac{h_j^2}{\pi_j^2} - \frac{2h_i h_j}{\pi_i \pi_j} + o(\|\boldsymbol{h}\|_2^2). \tag{6}$$

The expansion in Equation (6) contains both coordinatewise terms and a cross-label interaction term. Since the separable potential provides one curvature weight $\phi''(\pi_i)$ for each label, we match only the coordinatewise scaling shared by the pairwise log-ratio errors disregarding the cross-label terms. Comparing Equation (6) with Equation (5) yields the design criterion $\phi''(u) \propto u^{-2}$. We therefore adopt $\phi''(u) = u^{-2}$ to match the coordinatewise curvature scaling of local squared log-ratio errors.

Integrating $\phi''(u) = u^{-2}$ twice gives $\phi(u) = -\log u$, up to a term that does not affect the resulting Bregman divergence. Interestingly, the divergence $d_i/p_i - \log(d_i/p_i) - 1$ coincides with the classical Itakura-Saito Bregman form (Banerjee et al., 2005b). However, both the negative-log potential $-\log u$ and its curvature u^{-2} diverge as $u \to 0$. This behavior is problematic for finite-annotation label distributions. We therefore introduce a threshold τ and replace the negative-log potential below τ by a smoothly connected bounded function. The resulting capped potential preserves the desired u^{-2} curvature for $u \geq \tau$ while preventing the curvature and derivative for $0 \leq u < \tau$ from diverging. Below τ, the cap deliberately weakens this scaling to keep the curvature bounded. To ensure that the resulting loss remains within the class of regular mean-proper losses characterized by Proposition 3.3, the potential $\Phi(\cdot)$ must be strictly convex and belong to $\mathcal{C}^3(\mathcal{U})$. For our separable construction, a convenient sufficient condition is that ϕ_τ be strictly convex and belong to $\mathcal{C}^3([0, 1])$. Accordingly, we choose a quartic continuation below τ that maintains positive curvature and matches $-\log u$ and its first three derivatives at $u = \tau$, yielding the following potential:

$$\Phi_\tau(\boldsymbol{u}) = \sum_{i=1}^m \phi_\tau(u_i), \qquad \phi_\tau(u) = \begin{cases} -\log u, & u \geq \tau, \\ -\log \tau + \frac{7}{4} - \frac{8u}{3\tau} + \frac{u^2}{\tau^2} - \frac{u^4}{12\tau^4}, & 0 \leq u < \tau. \end{cases} \tag{7}$$

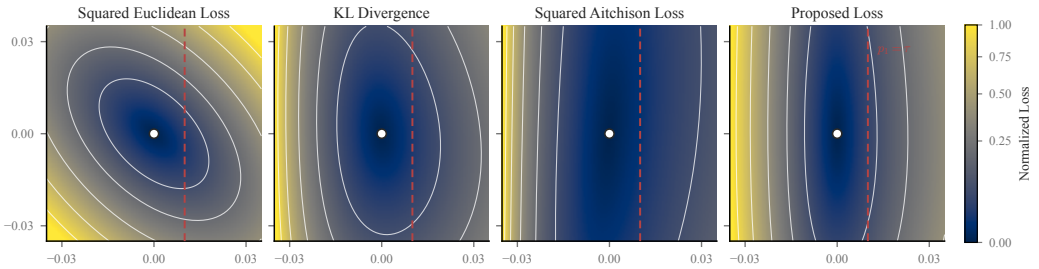

Figure 2: **Local Loss Geometry Comparison.** The target label distribution is $\boldsymbol{d} = [0.04; 0.20; 0.76]$ and parameterize predictions as $\boldsymbol{p} = \boldsymbol{d} + [h_1; h_2; -h_1 - h_2]$. The proposed loss is Bregman divergence generated by $\Phi_\tau(\cdot)$ with $\tau = 0.05$. The white marker denotes the Bayes target $\boldsymbol{p} = \boldsymbol{d}$. Each loss is normalized to $[0, 1]$. The x- and y-axis denote h_1 and h_2, respectively. The red line marks $p_1 = \tau$.

3.3 Theoretical Analysis of BAP

We give Proposition 3.4 to rigorously demonstrate the properties of $\phi_\tau(\cdot)$, thereby establishing that the Bregman divergence induced by the potential $\Phi_\tau(\cdot)$ is mean-proper.

Proposition 3.4 (Properties of Potential ϕ_τ)**.** *Given a threshold $0 < \tau < 1$, define a potential $\phi_\tau : [0, 1] \to \mathbb{R}$ by Equation (7). Then ϕ_τ is finite and strictly convex on $[0, 1]$, and $\phi_\tau \in \mathcal{C}^3([0, 1])$. Moreover, for every $u \in [0, 1]$, we have the following bounds:*

$$0 \le \phi_\tau(u) \le -\log \tau + \frac{7}{4}, \quad -\frac{8}{3\tau} \le \phi'_\tau(u) \le -1, \quad 1 \le \phi''_\tau(u) \le \frac{2}{\tau^2}, \quad -\frac{2}{\tau^3} \le \phi'''_\tau(u) \le 0. \tag{8}$$

The proof is provided in Section A.5. Finally, given a training set $\{(\boldsymbol{x}_i, \boldsymbol{d}_i)\}_{i=1}^n$ and an LDL model $f_{\boldsymbol{\theta}} : \mathcal{X} \to \Delta^m$, we define the overall loss function by combining Equation (3) and Equation (7):

$$\mathcal{L}_{\mathrm{BAP}} = \frac{1}{n} \sum_{i=1}^{n} B_{\Phi_\tau}(\boldsymbol{d}_i, f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)) = \frac{1}{n} \sum_{i=1}^{n} \Phi_\tau(\boldsymbol{d}_i) - \Phi_\tau(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)) - \langle \nabla \Phi_\tau(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)), \boldsymbol{d}_i - f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \rangle. \tag{9}$$

To gain further insight into the proposed BAP loss, Figure 2 compares the local geometry of the Bregman divergence generated by Φ_τ over the 2D simplex with those of SE loss, KL divergence, and squared Aitchison loss. For the displayed target label distribution and threshold, the proposed BAP has a more anisotropic local geometry, with increased sensitivity along the direction associated with the low-mass component, compared with SE loss and KL divergence.

Theorem 3.5 (Local Loss Geometry Comparison)**.** *Let $m \ge 2$ and $0 < \tau < 1$. Define $\Phi_\tau(\boldsymbol{u})$ by Equation (7), let $\ell_\tau(\boldsymbol{d}, \boldsymbol{p}) = B_{\Phi_\tau}(\boldsymbol{d}, \boldsymbol{p})$. Fix $\boldsymbol{d} \in \mathrm{int}(\Delta^m)$ and $\boldsymbol{h} \in T\Delta^m$, where $T\Delta^m = \{\boldsymbol{h} \in \mathbb{R}^m : \mathbf{1}^\top \boldsymbol{h} = 0\}$, such that $\boldsymbol{d} + \boldsymbol{h} \in \mathrm{int}(\Delta^m)$. Then we have $\ell_\tau(\boldsymbol{d}, \boldsymbol{d} + \boldsymbol{h}) = \mathcal{Q}_{\tau,\boldsymbol{d}}(\boldsymbol{h}) + r_{\tau,\boldsymbol{d}}(\boldsymbol{h})$, where $\mathcal{Q}_{\tau,\boldsymbol{d}}(\boldsymbol{h}) = \frac{1}{2}\sum_{i=1}^m \phi''_\tau(d_i) h_i^2$ and $|r_{\tau,\boldsymbol{d}}(\boldsymbol{h})| \le \frac{2}{3}\tau^{-3}\|\boldsymbol{h}\|_2^3$. Let $\boldsymbol{\kappa} \coloneqq (\boldsymbol{I}_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^\top)\mathrm{diag}(\boldsymbol{d})^{-1}\boldsymbol{h}$. For every i, we have $0 < d_i^2 \phi''_\tau(d_i) \le 1$, and*

$$\left(\min_{i\in[m]} d_i^2 \phi''_\tau(d_i)\right)\|\boldsymbol{\kappa}\|_2^2 \le 2\mathcal{Q}_{\tau,\boldsymbol{d}}(\boldsymbol{h}) \le \left(\max_{i\in[m]} d_i^2 \phi''_\tau(d_i)\right) m\|\boldsymbol{d}\|_2^2\|\boldsymbol{\kappa}\|_2^2 \le m\|\boldsymbol{\kappa}\|_2^2. \tag{10}$$

Moreover, the squared Aitchison loss satisfies $d_{\mathrm{A}}^2(\boldsymbol{d}, \boldsymbol{d} + \boldsymbol{h}) = \|\boldsymbol{\kappa}\|_2^2 + \mathcal{O}(\|\boldsymbol{h}\|_2^3)$ as $\boldsymbol{h} \to \mathbf{0}$. In particular, if $d_i \ge \tau$ for all i, then $d_i^2 \phi''_\tau(d_i) = 1$, $\frac{1}{2}\|\boldsymbol{\kappa}\|_2^2 \le \mathcal{Q}_{\tau,\boldsymbol{d}}(\boldsymbol{h}) \le \frac{1}{2}m\|\boldsymbol{d}\|_2^2\|\boldsymbol{\kappa}\|_2^2 \le \frac{1}{2}m\|\boldsymbol{\kappa}\|_2^2$. For comparing BAP with KL and SE, we define the coordinatewise local curvature of each loss as $c_\tau(u) = \phi''_\tau(u)$, $c_{\mathrm{SE}}(u) = 2$, and $c_{\mathrm{KL}}(u) = u^{-1}$, respectively. For $\tau \le d_i < d_j$, we have

$$c_\tau(d_j)^{-1} c_\tau(d_i) = \left(d_i^{-1} d_j\right)^2 > d_i^{-1} d_j = c_{\mathrm{KL}}(d_j)^{-1} c_{\mathrm{KL}}(d_i) > 1 = c_{\mathrm{SE}}(d_j)^{-1} c_{\mathrm{SE}}(d_i). \tag{11}$$

The proof is provided in Section A.6. For $d_i \ge \tau$, BAP retains the u^{-2} coordinatewise curvature scaling suggested by local log-ratio geometry. For $d_i < \tau$, the capped continuation smoothly moderates this weighting while maintaining bounded curvature. The threshold τ therefore controls the component scale over which the Aitchison-inspired curvature profile is retained, while the remainder bound quantifies the locality of the corresponding quadratic comparison. Besides, Equation (11)

shows that, above the capping threshold, the BAP curvature scales as u^{-2}, exactly matching the coordinatewise scaling of the local Aitchison geometry. In contrast, KL scales only as u^{-1}, while SE has constant curvature. Consequently, BAP preserves relative-coordinate sensitivity more strongly, particularly for smaller components. Finally, we prove that BAP is strictly arithmetic-mean proper in Proposition 3.6, and the proof is shown in Section A.7.

Proposition 3.6 (Mean-Properness of BAP)**.** *Let $0 < \tau < 1$, let Φ_τ be the potential defined in Equation* (7)*, and define $\ell_\tau(\boldsymbol{d}, \boldsymbol{p}) \coloneqq B_{\Phi_\tau}(\boldsymbol{d}, \boldsymbol{p})$, where $\boldsymbol{d}, \boldsymbol{p} \in \Delta^m$. For any $\boldsymbol{x} \in \mathcal{X}$, let $\boldsymbol{D} \sim \mathcal{P}_{\boldsymbol{x}}$ be an arbitrary random label distribution supported on Δ^m, and let $\boldsymbol{\pi}(\boldsymbol{x}) \coloneqq \mathbb{E}_{\boldsymbol{D}\sim\mathcal{P}_{\boldsymbol{x}}}[\boldsymbol{D}]$. Define the conditional risk as $\mathcal{R}_\tau(\boldsymbol{p}|\boldsymbol{x}) \coloneqq \mathbb{E}_{\boldsymbol{D}\sim\mathcal{P}_{\boldsymbol{x}}}\big[\ell_\tau(\boldsymbol{D}, \boldsymbol{p})\big]$. Then, for every $\boldsymbol{p} \in \Delta^m$,*

$$\mathcal{R}_\tau(\boldsymbol{p}|\boldsymbol{x}) - \mathcal{R}_\tau(\boldsymbol{\pi}(\boldsymbol{x})|\boldsymbol{x}) = B_{\Phi_\tau}(\boldsymbol{\pi}(\boldsymbol{x}), \boldsymbol{p}) \geq \|\boldsymbol{\pi}(\boldsymbol{x}) - \boldsymbol{p}\|_2^2/2. \tag{12}$$

Hence, $\arg\min_{\boldsymbol{p}\in\Delta^m} \mathcal{R}_\tau(\boldsymbol{p}|\boldsymbol{x}) = \{\boldsymbol{\pi}(\boldsymbol{x})\}$*, and ℓ_τ possesses the strict arithmetic-mean properness.*

4 Experiments

4.1 Experimental Configuration

Datasets and Evaluation Metrics. Experiments require a high-precision label distribution for each instance, which serves as the reference for assessing predictions learned from limited annotations. We hence select eight datasets in which the label distributions are aggregated from large numbers of annotators, including Jaffe Lyons et al. (1998), BU-3DFE Yin et al. (2006), RAF-ML Li & Deng (2019), SCUT-FBP Ren & Geng (2017), FBP5500 Liang et al. (2018), M2B Nguyen et al. (2012), Music Lee et al. (2021), Movie Geng & Hou (2015). See Section B.2 for more details. Following the standard LDL metrics Geng (2016), we report Chebyshev distance, Clark distance, Canberra metric, KL divergence, cosine coefficient, and intersection similarity. We additionally report mean absolute log-ratio (MALR) error: $\mathrm{MALR}(\boldsymbol{d}, \boldsymbol{p}) = \sum_{(i,j)\in S} |\log(d_i/d_j) - \log(p_i/p_j)|/|S|$, where $S = \{(i,j) : 1 \leq i < j \leq m,\ d_i d_j > 0\}$. For each instance, we compute the mean absolute log-ratio error over all label pairs whose reference label masses are both positive, since log-ratios are undefined for zero masses. When aggregating MALR across the test set, we exclude instances with only one reference label of positive mass, as no valid log-ratio pair can be formed for such instances.

Experimental Protocol. To simulate finite-sample estimates of the underlying population label distributions, we treat the original label distribution $\boldsymbol{d}_i$ of each instance as a high-precision reference and generate finite annotations under two levels of annotation counts, $n_a \in \{10, 30\}$. Specifically, we draw annotation counts via $\boldsymbol{c}_i \sim \mathrm{Multinomial}(n_a, \boldsymbol{d}_i)$, and construct the corresponding finite-sample training label distribution as $\boldsymbol{d}'_i = \boldsymbol{c}_i/n_a$. The dataset was split into a training set (70%) and a test set (30%). Models are trained on the instances with $\boldsymbol{d}'$ and are evaluated on the test instances with their original label distributions $\boldsymbol{d}$. The dataset is See Section B.3 for more details.

Comparison Methods. We compared BAP with two complementary groups of baselines. First, we considered the standard Kullback-Leibler (KL) divergence and squared Euclidean (SE) loss, the squared Aitchison loss, and two hybrid objectives, namely KL+λ·SA and SE+λ·SA. Second, we evaluated whether BAP can improve established LDL algorithms. Specifically, we included three recently proposed LDL algorithms: LDL-DVS (Lu et al., 2026), Delta-LDL (Li et al., 2025), and RkNN-LDL (Wang et al., 2025), together with their BAP-based counterparts: Delta†, DVS†, and RkNN†. For brevity, we omit "LDL" from the name of each method. Each counterpart retains the original algorithmic components of the corresponding method while replacing its KL- or SE-fitting term with BAP. More details are provided in Section B.4.

4.2 Results and Discussion

The main performance is shown in Table 1. Due to space limitations, we defer to the appendix the results for the cases where $n_a = 10$, along with additional evaluation metrics, and more result discussion is provided in Section B.5. A two-sided paired t-test is applied over the same ten random splits at $\alpha = 0.05$ without correction for multiple comparisons. The symbol • identifies the reference method within each comparison group; $\succ$, $\sim$, and $\prec$ indicate that the displayed method is significantly better than, statistically indistinguishable from, or significantly worse than that reference, respectively.

Table 1: **Performance under Finite-Sample Label Distributions with $n_b = 30$.** Each cell has the form "$s\ \mu_{(\pm\sigma)}$", where s is the significance outcome, μ is the mean, and σ is the standard deviation.

	Method	Jaffe	BU-3DFE	RAF-ML	SCUT-FBP	FBP5500	M2B	Music	Movie
Chebyshev Distance (↓)	BAP	$\bullet.102_{(\pm.005)}$	$\bullet.107_{(\pm.002)}$	$\bullet.162_{(\pm.002)}$	$\bullet.248_{(\pm.005)}$	$\bullet.135_{(\pm.002)}$	$\bullet.361_{(\pm.009)}$	$\bullet.078_{(\pm.003)}$	$\bullet.115_{(\pm.001)}$
	KL	$\prec.133_{(\pm.008)}$	$\prec.114_{(\pm.002)}$	$\succ.159_{(\pm.002)}$	$\prec.276_{(\pm.004)}$	$\prec.144_{(\pm.002)}$	$\prec.378_{(\pm.007)}$	$\prec.110_{(\pm.006)}$	$\prec.115_{(\pm.001)}$
	SE	$\prec.109_{(\pm.005)}$	$\prec.113_{(\pm.002)}$	$\sim.162_{(\pm.002)}$	$\prec.271_{(\pm.006)}$	$\prec.147_{(\pm.002)}$	$\prec.380_{(\pm.007)}$	$\prec.083_{(\pm.004)}$	$\prec.116_{(\pm.001)}$
	SA	$\prec.622_{(\pm.043)}$	$\prec.322_{(\pm.009)}$	$\prec.420_{(\pm.004)}$	$\prec.514_{(\pm.014)}$	$\prec.381_{(\pm.007)}$	$\prec.545_{(\pm.014)}$	$\prec.730_{(\pm.017)}$	$\prec.449_{(\pm.006)}$
	KL-SA	$\prec.612_{(\pm.039)}$	$\prec.122_{(\pm.003)}$	$\prec.330_{(\pm.003)}$	$\prec.324_{(\pm.009)}$	$\prec.265_{(\pm.003)}$	$\prec.445_{(\pm.015)}$	$\prec.582_{(\pm.026)}$	$\prec.208_{(\pm.004)}$
	SE-SA	$\prec.609_{(\pm.037)}$	$\prec.142_{(\pm.004)}$	$\prec.406_{(\pm.004)}$	$\prec.324_{(\pm.007)}$	$\prec.304_{(\pm.005)}$	$\prec.522_{(\pm.018)}$	$\prec.718_{(\pm.015)}$	$\prec.239_{(\pm.003)}$
	Delta	$\sim.097_{(\pm.006)}$	$\prec.118_{(\pm.003)}$	$\prec.330_{(\pm.015)}$	$\prec.248_{(\pm.005)}$	$\sim.139_{(\pm.002)}$	$\prec.371_{(\pm.009)}$	$\prec.075_{(\pm.003)}$	$\sim.120_{(\pm.002)}$
	Delta†	$\bullet.098_{(\pm.005)}$	$\bullet.117_{(\pm.003)}$	$\bullet.309_{(\pm.013)}$	$\bullet.246_{(\pm.006)}$	$\bullet.139_{(\pm.002)}$	$\bullet.357_{(\pm.012)}$	$\bullet.074_{(\pm.002)}$	$\bullet.121_{(\pm.002)}$
	DVS	$\sim.105_{(\pm.008)}$	$\prec.118_{(\pm.003)}$	$\sim.335_{(\pm.015)}$	$\sim.251_{(\pm.005)}$	$\succ.139_{(\pm.002)}$	$\prec.375_{(\pm.009)}$	$\sim.076_{(\pm.005)}$	$\prec.129_{(\pm.003)}$
	DVS†	$\bullet.102_{(\pm.005)}$	$\bullet.117_{(\pm.003)}$	$\bullet.330_{(\pm.012)}$	$\bullet.251_{(\pm.005)}$	$\bullet.143_{(\pm.002)}$	$\bullet.371_{(\pm.009)}$	$\bullet.076_{(\pm.002)}$	$\bullet.124_{(\pm.002)}$
	RkNN	$\prec.116_{(\pm.008)}$	$\prec.124_{(\pm.002)}$	$\prec.315_{(\pm.002)}$	$\prec.261_{(\pm.005)}$	$\prec.152_{(\pm.002)}$	$\sim.381_{(\pm.009)}$	$\prec.084_{(\pm.005)}$	$\prec.121_{(\pm.001)}$
	RkNN†	$\bullet.102_{(\pm.007)}$	$\bullet.117_{(\pm.002)}$	$\bullet.285_{(\pm.006)}$	$\bullet.252_{(\pm.005)}$	$\bullet.143_{(\pm.002)}$	$\bullet.375_{(\pm.018)}$	$\bullet.077_{(\pm.003)}$	$\bullet.117_{(\pm.001)}$
Cosine Coefficient (↑)	BAP	$\bullet.944_{(\pm.005)}$	$\bullet.944_{(\pm.002)}$	$\bullet.920_{(\pm.002)}$	$\bullet.844_{(\pm.007)}$	$\bullet.955_{(\pm.002)}$	$\bullet.757_{(\pm.011)}$	$\bullet.917_{(\pm.004)}$	$\bullet.935_{(\pm.001)}$
	KL	$\prec.905_{(\pm.010)}$	$\prec.933_{(\pm.002)}$	$\succ.923_{(\pm.002)}$	$\prec.813_{(\pm.006)}$	$\prec.949_{(\pm.002)}$	$\prec.742_{(\pm.009)}$	$\prec.867_{(\pm.009)}$	$\succ.935_{(\pm.001)}$
	SE	$\prec.934_{(\pm.006)}$	$\prec.935_{(\pm.002)}$	$\succ.921_{(\pm.002)}$	$\prec.820_{(\pm.006)}$	$\prec.947_{(\pm.002)}$	$\prec.744_{(\pm.009)}$	$\prec.910_{(\pm.005)}$	$\prec.935_{(\pm.001)}$
	SA	$\prec.497_{(\pm.027)}$	$\prec.705_{(\pm.010)}$	$\prec.792_{(\pm.004)}$	$\prec.635_{(\pm.015)}$	$\prec.798_{(\pm.007)}$	$\prec.557_{(\pm.017)}$	$\prec.352_{(\pm.012)}$	$\prec.634_{(\pm.005)}$
	KL-SA	$\prec.503_{(\pm.027)}$	$\prec.923_{(\pm.003)}$	$\prec.838_{(\pm.004)}$	$\prec.764_{(\pm.010)}$	$\prec.875_{(\pm.003)}$	$\prec.641_{(\pm.018)}$	$\prec.425_{(\pm.017)}$	$\prec.848_{(\pm.004)}$
	SE-SA	$\prec.505_{(\pm.027)}$	$\prec.899_{(\pm.004)}$	$\prec.805_{(\pm.004)}$	$\prec.778_{(\pm.009)}$	$\prec.865_{(\pm.004)}$	$\prec.580_{(\pm.020)}$	$\prec.361_{(\pm.008)}$	$\prec.827_{(\pm.003)}$
	Delta	$\sim.949_{(\pm.004)}$	$\prec.937_{(\pm.003)}$	$\prec.758_{(\pm.025)}$	$\succ.842_{(\pm.007)}$	$\sim.952_{(\pm.002)}$	$\succ.770_{(\pm.011)}$	$\sim.920_{(\pm.006)}$	$\succ.932_{(\pm.002)}$
	Delta†	$\bullet.949_{(\pm.005)}$	$\bullet.938_{(\pm.002)}$	$\bullet.791_{(\pm.020)}$	$\bullet.841_{(\pm.007)}$	$\bullet.952_{(\pm.002)}$	$\bullet.746_{(\pm.022)}$	$\bullet.922_{(\pm.004)}$	$\bullet.929_{(\pm.001)}$
	DVS	$\sim.940_{(\pm.009)}$	$\sim.937_{(\pm.003)}$	$\prec.751_{(\pm.026)}$	$\sim.842_{(\pm.007)}$	$\succ.952_{(\pm.002)}$	$\sim.771_{(\pm.010)}$	$\sim.919_{(\pm.009)}$	$\prec.924_{(\pm.003)}$
	DVS†	$\bullet.943_{(\pm.006)}$	$\bullet.937_{(\pm.002)}$	$\bullet.762_{(\pm.020)}$	$\bullet.842_{(\pm.007)}$	$\bullet.951_{(\pm.002)}$	$\bullet.768_{(\pm.013)}$	$\bullet.920_{(\pm.005)}$	$\bullet.928_{(\pm.001)}$
	RkNN	$\prec.933_{(\pm.008)}$	$\prec.931_{(\pm.001)}$	$\prec.786_{(\pm.002)}$	$\succ.840_{(\pm.007)}$	$\prec.944_{(\pm.002)}$	$\sim.762_{(\pm.011)}$	$\prec.910_{(\pm.006)}$	$\prec.932_{(\pm.001)}$
	RkNN†	$\bullet.945_{(\pm.006)}$	$\bullet.938_{(\pm.002)}$	$\bullet.807_{(\pm.010)}$	$\bullet.831_{(\pm.006)}$	$\bullet.950_{(\pm.002)}$	$\bullet.757_{(\pm.022)}$	$\bullet.920_{(\pm.005)}$	$\bullet.935_{(\pm.002)}$
Mean Absolute Log-Ratio Error (↓)	BAP	$\bullet.419_{(\pm.018)}$	$\bullet.370_{(\pm.004)}$	$\bullet4.74_{(\pm.018)}$	$\bullet5.53_{(\pm.075)}$	$\bullet.754_{(\pm.015)}$	$\bullet1.64_{(\pm.026)}$	$\bullet.639_{(\pm.024)}$	$\bullet.641_{(\pm.006)}$
	KL	$\prec.593_{(\pm.034)}$	$\prec.430_{(\pm.008)}$	$\prec4.77_{(\pm.019)}$	$\prec5.59_{(\pm.071)}$	$\prec.854_{(\pm.012)}$	$\prec1.73_{(\pm.026)}$	$\prec.778_{(\pm.028)}$	$\prec.644_{(\pm.006)}$
	SE	$\prec.464_{(\pm.022)}$	$\prec.431_{(\pm.009)}$	$\prec4.79_{(\pm.020)}$	$\sim5.39_{(\pm.588)}$	$\prec.901_{(\pm.014)}$	$\prec2.49_{(\pm.161)}$	$\prec.659_{(\pm.024)}$	$\prec.651_{(\pm.007)}$
	SA	$\prec5.14_{(\pm.992)}$	$\prec1.47_{(\pm.055)}$	$\prec7.61_{(\pm.080)}$	$\prec13.2_{(\pm.912)}$	$\prec7.19_{(\pm.098)}$	$\prec11.7_{(\pm.218)}$	$\prec9.11_{(\pm.479)}$	$\prec4.84_{(\pm.152)}$
	KL-SA	$\prec4.83_{(\pm.758)}$	$\prec.476_{(\pm.010)}$	$\prec5.49_{(\pm.060)}$	$\prec5.88_{(\pm.294)}$	$\prec4.66_{(\pm.089)}$	$\prec9.85_{(\pm.235)}$	$\prec4.93_{(\pm.307)}$	$\prec1.71_{(\pm.048)}$
	SE-SA	$\prec4.78_{(\pm.721)}$	$\prec.594_{(\pm.017)}$	$\prec7.23_{(\pm.083)}$	$\prec7.73_{(\pm.255)}$	$\prec6.47_{(\pm.105)}$	$\prec11.3_{(\pm.279)}$	$\prec8.50_{(\pm.352)}$	$\prec2.47_{(\pm.076)}$
	Delta	$\sim.397_{(\pm.020)}$	$\prec.378_{(\pm.008)}$	$\prec5.56_{(\pm.049)}$	$\prec5.52_{(\pm.082)}$	$\prec.762_{(\pm.019)}$	$\prec1.73_{(\pm.034)}$	$\prec.634_{(\pm.025)}$	$\sim.661_{(\pm.009)}$
	Delta†	$\bullet.392_{(\pm.020)}$	$\bullet.374_{(\pm.006)}$	$\bullet5.49_{(\pm.050)}$	$\bullet5.46_{(\pm.087)}$	$\bullet.749_{(\pm.018)}$	$\bullet1.61_{(\pm.061)}$	$\bullet.626_{(\pm.023)}$	$\bullet.660_{(\pm.007)}$
	DVS	$\sim.439_{(\pm.049)}$	$\sim.378_{(\pm.008)}$	$\prec5.57_{(\pm.050)}$	$\prec5.59_{(\pm.081)}$	$\prec.762_{(\pm.019)}$	$\prec1.80_{(\pm.033)}$	$\sim.637_{(\pm.028)}$	$\sim.684_{(\pm.010)}$
	DVS†	$\bullet.420_{(\pm.028)}$	$\bullet.378_{(\pm.007)}$	$\bullet5.55_{(\pm.039)}$	$\bullet5.58_{(\pm.076)}$	$\bullet.742_{(\pm.016)}$	$\bullet1.74_{(\pm.042)}$	$\bullet.634_{(\pm.023)}$	$\bullet.679_{(\pm.009)}$
	RkNN	$\prec.436_{(\pm.021)}$	$\prec.394_{(\pm.003)}$	$\prec5.46_{(\pm.024)}$	$\prec5.64_{(\pm.106)}$	$\prec.772_{(\pm.016)}$	$\prec1.79_{(\pm.035)}$	$\prec.679_{(\pm.029)}$	$\prec.687_{(\pm.007)}$
	RkNN†	$\bullet.399_{(\pm.017)}$	$\bullet.371_{(\pm.003)}$	$\bullet5.30_{(\pm.026)}$	$\bullet5.49_{(\pm.066)}$	$\bullet.761_{(\pm.016)}$	$\bullet1.74_{(\pm.055)}$	$\bullet.633_{(\pm.027)}$	$\bullet.656_{(\pm.006)}$

Across the eight datasets, BAP generally achieves lower distribution errors and higher similarity scores than the five alternative losses, with its clearest advantage appearing in mean absolute log-ratio error, which measures discrepancies in relative label mass ratios. Many improvements are statistically significant, indicating consistent gains across data partitions. Replacing the original fitting loss of Delta, DVS, and RkNN with BAP also frequently improves their performance, particularly in MALR, supporting BAP's use as a broadly applicable training objective for existing LDL methods.

4.3 Further Analysis

Ablation Study via IS Divergence Here, we compare the performance of BAP with that of the IS divergence. Since the IS divergence is singular at zero, we construct a modified divergence, denoted by IS*, using the shifted potential $-\log(u + \varepsilon_{\mathrm{IS}})$. The shift parameter $\varepsilon_{\mathrm{IS}}$ controls the degree of smoothing near zero and is varied over $\{0.001, 0.01, 0.025, 0.05, 0.1, 0.2, 0.3, 0.4\}$. We use the same model configuration as in the main experiments, with $n_a = 30$. Figure 3 reports the performance difference between IS* and BAP, computed as the performance of IS* minus that of BAP. For each dataset, the difference is first averaged over random seeds and then arithmetically averaged across datasets. Accordingly, for distance-based metrics, a positive difference indicates better performance by BAP, whereas for similarity-based metrics, a negative difference indicates better performance by BAP.

Across all values of $\varepsilon_{\mathrm{IS}}$, the differences for the Chebyshev, Clark, Canberra, KL, and MALR metrics are consistently positive, while those for cosine and intersection are consistently negative. Thus, throughout the tested range of $\varepsilon_{\mathrm{IS}}$, BAP achieves better average performance. The performance gap is most pronounced for small values of $\varepsilon_{\mathrm{IS}}$ and generally narrows as $\varepsilon_{\mathrm{IS}}$ increases, which indicates the severe performance degradation caused by the diverging curvature as $\varepsilon_{\mathrm{IS}} \to 0$. Moreover, applying a uniform shift parameter globally alters the curvature, making it less suitable for approximating the Aitchison geometry.

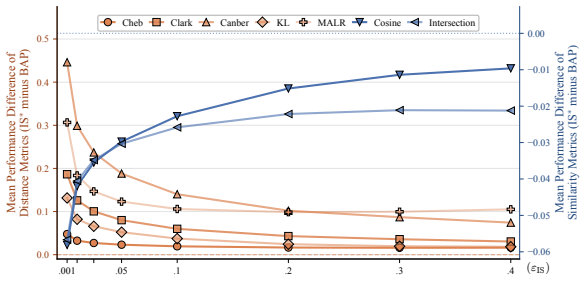

Figure 3: **Ablation Study via IS*.** Orange distance metrics use the left y-axis, whereas blue similarity metrics use the right y-axis. Positive distance differences and negative similarity differences favor BAP.

Hyperparameter Sensitivity. The cap threshold τ determines the label-mass scale below which BAP smoothly moderates the Aitchison-inspired coordinatewise curvature. We examine the empirical effect of τ by varying τ over $\{0.01, 0.025, 0.05, 0.075, 0.1, 0.15, 0.2\}$, using BAP as the sole loss. Because the metrics differ in numerical scale and semantics, we compare threshold values via rank-based aggregation. Specifically, we first average each metric value over ten random seeds, rank the seven threshold values separately for each metric, and then average the resulting metric-specific ranks. As shown in Figure 4 (a), the aggregate performance exhibits a non-monotonic dependence on τ. The lowest overall rank is achieved at $\tau = 0.05$, followed by $\tau = 0.075$ and $\tau = 0.10$. By contrast, the two endpoint settings, $\tau = 0.01$ and $\tau = 0.20$, yield higher mean ranks. The metric-specific results in Figure 4 (b) further reveal how this pattern arises. Clark, Canberra, and MALR attain their best average ranks at $\tau = 0.05$, while KL assigns the same best average rank to $\tau = 0.05$ and $\tau = 0.075$. These results indicate that the preferred threshold varies across evaluation metrics, whereas intermediate values of τ provide a more balanced performance. The n_a-specific results show the same qualitative pattern. At an annotation count of 10, $\tau \in \{0.05, 0.075, 0.10\}$ attain the same lowest combined rank of 3.232. At $n_a = 30$, $\tau = 0.05$ achieves the lowest combined rank, followed by $\tau = 0.075$. Thus, across both annotation counts, the best overall performance is concentrated in intermediate thresholds.

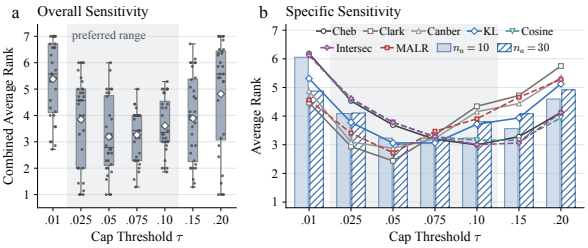

Figure 4: **Sensitivity of BAP to Cap Threshold τ.** (a) shows the distribution of the overall average rank of BAP for each value of τ, obtained by assigning equal weight to all corresponding evaluation results. (b) reports metric-specific average ranks aggregated across all datasets and annotation counts; the grouped bars show, for each annotation count, the average rank aggregated across datasets and metrics. The shaded region highlights the range of τ values with better performance.

5 Conclusion

This work investigated how to design a mean-proper loss function with ratio-sensitivity for label distribution learning. We showed that KL divergence and SE loss possess arithmetic-mean properness but do not provide uniform control of log-ratio distortion near the simplex boundary, while squared Aitchison loss directly captures relative variation but generally does not preserve the conditional arithmetic mean. To address these limitations, we proposed BAP, a Bregman divergence with Aitchison-inspired potential, whose coordinatewise curvature scaling is designed according to local Aitchison errors. We established the strict arithmetic-mean properness of BAP and quantitatively compared its local quadratic form with the local Aitchison geometry. Extensive experiments on eight LDL datasets further demonstrate the effectiveness of BAP. It generally achieves better performance than comparison methods, with the clearest merit in mean absolute log-ratio error, and frequently improves the state-of-the-art LDL methods when replacing their original distribution-fitting losses.

AI USE STATEMENT

In this work, we used generative AI tools for assisting in the writing of proofs, assisting with translation. We have not used generative AI tools for generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, supporting qualitative and thematic data analysis, interpreting results. Cleaning and reformatting dataset are not applicable to this work. Additionally, we used generative AI tools for brainstorming, editing a research paper to improve readability, identifying relevant literature. We have reviewed all AI-assisted work. For instance, we have verified the authenticity of the literature researched by the AI and checked whether the sentences polished by the AI align with our original intent. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Jacob D. Abernethy and Rafael M. Frongillo. A characterization of scoring rules for linear properties. In *Annual Conference on Learning Theory*, pp. 27.1–27.13, 2012.

J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982.

Arindam Banerjee, Xin Guo, and Hui Wang. On the optimality of conditional expectation as a Bregman predictor. *IEEE Transactions on Information Theory*, 51(7):2664–2669, 2005a.

Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6(58):1705–1749, 2005b.

J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35:279–300, 2003.

Xin Geng. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28 (7):1734–1748, 2016.

Xin Geng and Peng Hou. Pre-release prediction of crowd opinion on movies by label distribution learning. In *International Joint Conference on Artificial Intelligence*, pp. 3511–3517, 2015.

Noriko Goda, Shinsuke Sasada, Hideo Shigematsu, Norio Masumoto, Koji Arihiro, Hiroyoshi Nishikawa, Shimon Sakaguchi, Morihito Okada, and Takayuki Kadoya. The ratio of CD8+ lymphocytes to tumor-infiltrating suppressive FOXP3+ effector regulatory T cells is associated with treatment response in invasive breast cancer. *Discover Oncology*, 13:27, 2022.

Peng Hou, Xin Geng, Zeng-Wei Huo, and Jia-Qi Lv. Semi-supervised adaptive label distribution learning for facial age estimation. In *AAAI Conference on Artificial Intelligence*, pp. 2015–2021, 2017.

Xiuyi Jia, Xiang Zheng, Weiwei Li, Changqing Zhang, and Zechao Li. Facial emotion distribution learning by exploiting low-rank label correlations locally. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9833–9842, 2019.

Xiuyi Jia, Xiaoxia Shen, Weiwei Li, Yunan Lu, and Jihua Zhu. Label distribution learning by maintaining label ranking relation. *IEEE Transactions on Knowledge and Data Engineering*, 35 (2):1695–1707, 2023.

Yufei Jin, Richard Gao, Yi He, and Xingquan Zhu. GLDL: Graph label distribution learning. In *AAAI Conference on Artificial Intelligence*, pp. 12965–12974, 2024.

Zhiqiang Kou, Jing Wang, Yuheng Jia, Biao Liu, and Xin Geng. Instance-dependent inaccurate label distribution learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(1): 1425–1437, 2025.

Harin Lee, Frank Hoeger, Marc Schoenwiesner, Minsu Park, and Nori Jacoby. Cross-cultural mood perception in pop songs and its alignment with mood detection algorithms. In *International Society for Music Information Retrieval Conference*, 2021.

Shan Li and Weihong Deng. Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning. *International Journal of Computer Vision*, 127(6–7):884–906, 2019.

Weiwei Li, Haitao Wu, Yunan Lu, and Xiuyi Jia. Approximately correct label distribution learning. In *International Conference on Machine Learning*, pp. 36298–36309, 2025.

Lingyu Liang, Luojun Lin, Lianwen Jin, Duorui Xie, and Mengru Li. SCUT-FBP5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction. In *International Conference on Pattern Recognition*, pp. 1598–1603, 2018.

Yunan Lu and Xiuyi Jia. Predicting label distribution from multi-label ranking. In *Advances in Neural Information Processing Systems*, pp. 36931–36943, 2022.

Yunan Lu and Xiuyi Jia. Predicting label distribution from ternary labels. In *Advances in Neural Information Processing Systems*, pp. 70431–70452, 2024.

Yunan Lu, Weiwei Li, Huaxiong Li, and Xiuyi Jia. Predicting label distribution from tie-allowed multi-label ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12): 15364–15379, 2023.

Yunan Lu, Weiwei Li, Dun Liu, Huaxiong Li, and Xiuyi Jia. Adaptive-grained label distribution learning. In *AAAI Conference on Artificial Intelligence*, pp. 19161–19169, 2025a.

Yunan Lu, Bowen Xue, Xiuyi Jia, and Lei Yang. Entropy-calibrated label distribution learning. In *Advances in Neural Information Processing Systems*, 2025b.

Yunan Lu, Haitao Wu, Weiwei Li, Lei Yang, and Xiuyi Jia. Divisiveness-consistent label distribution learning. In *International Conference on Machine Learning*, 2026.

Michael Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with gabor wavelets. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 200–205, 1998.

Kevin Menden, Matteo Marouf, Sergio Oller, Anupriya Dalmia, Daniel Sumner Magruder, Karin Kloiber, Peter Heutink, and Stefan Bonn. Deep learning-based cell composition analysis from tissue expression profiles. *Science Advances*, 6(30):eaba2619, 2020.

Tam V. Nguyen, Si Liu, Bingbing Ni, Jun Tan, Yong Rui, and Shuicheng Yan. Sense beauty via face, dressing, and/or voice. In *ACM International Conference on Multimedia*, pp. 239–248, 2012.

Yi Ren and Xin Geng. Sense beauty by label distribution learning. In *International Joint Conference on Artificial Intelligence*, pp. 2648–2654, 2017.

Khoa A. Tran, Venkateswar Addala, Rebecca L. Johnston, David Lovell, Andrew Bradley, Lambros T. Koufariotis, Scott Wood, Sunny Z. Wu, Daniel Roden, Ghamdan Al-Eryani, Alexander Swarbrick, Elizabeth D. Williams, John V. Pearson, Olga Kondrashova, and Nicola Waddell. Performance of tumour microenvironment deconvolution methods in breast cancer using single-cell simulated bulk mixtures. *Nature Communications*, 14:5758, 2023.

Jing Wang and Xin Geng. Label distribution learning machine. In *International Conference on Machine Learning*, pp. 10749–10759, 2021.

Jing Wang and Xin Geng. Label distribution learning by exploiting label distribution manifold. *IEEE Transactions on Neural Networks and Learning Systems*, 34(2):839–852, 2023.

Jing Wang, Xin Geng, and Hui Xue. Re-weighting large margin label distribution learning for classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5445–5459, 2022.

Jing Wang, Fu Feng, Jianhui Lv, and Xin Geng. Residual k-nearest neighbors label distribution learning. *Pattern Recognition*, 158:111006, 2025.

Chao Xu, Shilin Gu, Hong Tao, and Chenping Hou. Fragmentary label distribution learning via graph regularized maximum entropy criteria. *Pattern Recognition Letters*, 145:147–156, 2021.

Miao Xu and Zhi-Hua Zhou. Incomplete label distribution learning. In *International Joint Conference on Artificial Intelligence*, pp. 3175–3181, 2017.

Jufeng Yang, Ming Sun, and Xiaoxiao Sun. Learning visual sentiment distributions via augmented conditional probability neural network. In *AAAI Conference on Artificial Intelligence*, pp. 224–230, 2017.

Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J. Rosato. A 3D facial expression database for facial behavior research. In *International Conference on Automatic Face and Gesture Recognition*, pp. 211–216, 2006.

Xiuzhuang Zhou, Zeqiang Wei, Min Xu, Shan Qu, and Guodong Guo. Facial depression recognition by deep joint label distribution and metric learning. *IEEE Transactions on Affective Computing*, 13(3):1605–1618, 2022.

A Proof of Theorems

A.1 Proof of Proposition 3.2

Proof. Fix $\boldsymbol{x} \in \mathcal{X}$ and abbreviate expectations with respect to $\boldsymbol{D} \sim \mathcal{P}_{\boldsymbol{x}}$ by $\mathbb{E}_{\boldsymbol{x}}[\cdot]$. Let $\boldsymbol{Y} = \psi(\boldsymbol{D})$ and $\boldsymbol{y} = \psi(\boldsymbol{p})$. By assumptions, $\mathbb{E}_{\boldsymbol{x}}[\|\boldsymbol{Y}\|_2^2] < \infty$, so $\overline{\boldsymbol{Y}} = \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{Y}]$ is well defined. Moreover, since $\mathbf{1}^\top \psi(\boldsymbol{D}) = 0$ for every strictly positive $\boldsymbol{D} \in \Delta^m$, both $\boldsymbol{Y}$ and $\overline{\boldsymbol{Y}}$ belong to the CLR subspace $\mathbf{1}^\perp := \{\boldsymbol{z} \in \mathbb{R}^m : \mathbf{1}^\top \boldsymbol{z} = 0\}$, which is the $(m-1)$-dimensional subspace orthogonal to the all-ones vector $\mathbf{1}$. The CLR transformation is a bijection from the interior of Δ^m to $\mathbf{1}^\perp$, with inverse $\psi^{-1}(\boldsymbol{y}) = \varpi(\boldsymbol{y})$. Therefore, minimizing the conditional squared Aitchison risk over strictly positive predictions $\boldsymbol{p}$ is equivalent to minimizing $\mathbb{E}_{\boldsymbol{x}}[\|\boldsymbol{Y} - \boldsymbol{y}\|_2^2]$ over $\boldsymbol{y} \in \mathbf{1}^\perp$. The Euclidean bias-variance decomposition gives

$$\mathbb{E}_{\boldsymbol{x}}[\|\boldsymbol{Y} - \boldsymbol{y}\|_2^2] = \mathbb{E}_{\boldsymbol{x}}[\|\boldsymbol{Y} - \overline{\boldsymbol{Y}}\|_2^2] + \|\boldsymbol{y} - \overline{\boldsymbol{Y}}\|_2^2.$$

The first term is independent of $\boldsymbol{y}$, whereas the second term is nonnegative and vanishes only when $\boldsymbol{y} = \overline{\boldsymbol{Y}}$. Hence, the unique optimal CLR coordinate is $\boldsymbol{y}_{\mathrm{A}}^\star(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{x}}[\psi(\boldsymbol{D})]$. The expectation $\mathbb{E}_{\boldsymbol{x}}[\log \boldsymbol{D}]$ is well defined under the stated square-integrability condition. Indeed, since $\boldsymbol{D} = \varpi(\boldsymbol{Y})$, we have $\log D_i = Y_i - \log \sum_{j=1}^m \exp(Y_j)$ and therefore $|\log D_i| \le \log m + 2\|\boldsymbol{Y}\|_2$. The right-hand side is integrable because square-integrability of $\boldsymbol{Y}$ implies its integrability. Define $\boldsymbol{P}_\perp = \boldsymbol{I}_m - m^{-1}\mathbf{1}\mathbf{1}^\top$. Using $\psi(\boldsymbol{D}) = \boldsymbol{P}_\perp \log \boldsymbol{D}$ and the linearity of expectation, we obtain $\boldsymbol{y}_{\mathrm{A}}^\star(\boldsymbol{x}) = \boldsymbol{P}_\perp \mathbb{E}_{\boldsymbol{x}}[\log \boldsymbol{D}]$. Applying the inverse CLR transformation yields

$$\boldsymbol{p}_{\mathrm{A}}^\star(\boldsymbol{x}) = \varpi(\boldsymbol{P}_\perp \mathbb{E}_{\boldsymbol{x}}[\log \boldsymbol{D}]) = \varpi(\mathbb{E}_{\boldsymbol{x}}[\log \boldsymbol{D}])\,.$$

The second equality follows from the translation invariance of softmax, since $\boldsymbol{P}_\perp \boldsymbol{z} = \boldsymbol{z} - m^{-1}(\mathbf{1}^\top \boldsymbol{z})\mathbf{1}$ differs from $\boldsymbol{z}$ only by a common coordinate shift. Let $\boldsymbol{\pi}(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{D}]$. Since $\boldsymbol{D}$ is strictly positive almost surely, $\boldsymbol{\pi}(\boldsymbol{x})$ is also strictly positive. By the injectivity of CLR,

$$\boldsymbol{p}_{\mathrm{A}}^\star(\boldsymbol{x}) = \boldsymbol{\pi}(\boldsymbol{x}) \quad \Longleftrightarrow \quad \boldsymbol{P}_\perp(\mathbb{E}_{\boldsymbol{x}}[\log \boldsymbol{D}] - \log \boldsymbol{\pi}(\boldsymbol{x})) = \mathbf{0}.$$

Because $\ker(\boldsymbol{P}_\perp) = \operatorname{span}\{\mathbf{1}\}$, the preceding condition holds if and only if the coordinates of $\mathbb{E}_{\boldsymbol{x}}[\log \boldsymbol{D}] - \log \boldsymbol{\pi}(\boldsymbol{x})$ are identical. Equivalently, the coordinatewise Jensen gaps $\log \pi_i(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x}}[\log D_i]$ are identical for all $i \in \{1, \dots, m\}$. □

A.2 Proof of Proposition 3.1

Proof. For positive $a, b \ge \delta$, the mean value theorem gives $|\log a - \log b| \le |a - b|/\delta$. Applying this inequality to coordinates i and j and then using Cauchy-Schwarz yields

$$\left|\log \frac{\pi_i}{\pi_j} - \log \frac{p_i}{p_j}\right| \le \frac{|\pi_i - p_i| + |\pi_j - p_j|}{\delta} \le \frac{\sqrt{2}}{\delta}\|\boldsymbol{\pi} - \boldsymbol{p}\|_2.$$

The same first inequality is at most $\|\boldsymbol{\pi} - \boldsymbol{p}\|_1/\delta$. Pinsker's inequality, in the convention used here, states that $\|\boldsymbol{\pi} - \boldsymbol{p}\|_1 \le \sqrt{2\mathcal{D}_{\mathrm{KL}}(\boldsymbol{\pi}\|\boldsymbol{p})}$, which proves the second bound in Equation (1). □

A.3 Proof of Proposition 3.3

Proof. We first explain how the Bregman representation follows from the classical characterization of scoring rules for linear properties (Abernethy & Frongillo, 2012, Theorems 11–12); see also the related characterization of conditional-mean losses in Banerjee et al. (2005a).

Let the class of all finitely supported probability laws on $\mathcal{U}$ be denoted by

$$\mathcal{P}_{\mathrm{fin}}(\mathcal{U}) = \Big\{\sum\nolimits_{s=1}^{S} \lambda_s \delta(\boldsymbol{\xi}_s) : S \ge 1,\ \boldsymbol{\xi}_s \in \mathcal{U},\ \lambda_s > 0,\ \sum\nolimits_{s=1}^{S} \lambda_s = 1\Big\}.$$

Define $\Gamma(\tilde{Q}) = \int_{\mathcal{U}} \boldsymbol{\xi}\, \mathrm{d}\tilde{Q}(\boldsymbol{\xi})$, and $\tilde{Q} \in \mathcal{P}_{\mathrm{fin}}(\mathcal{U})$. This is a linear property of $\tilde{Q}$. Because every point mass $\delta(\boldsymbol{\xi})$, with $\boldsymbol{\xi} \in \mathcal{U}$, belongs to $\mathcal{P}_{\mathrm{fin}}(\mathcal{U})$, we have $\mathrm{range}(\Gamma) = \mathcal{U}$. For every prediction $\boldsymbol{\eta} \in \mathcal{U}$, define a score on the outcome space $\mathcal{U}$ by $s[\boldsymbol{\eta}](\boldsymbol{\xi}) = -L(\boldsymbol{\xi}, \boldsymbol{\eta})$. The change of sign converts the loss-minimization convention used in this paper into the score-maximization convention used by Abernethy & Frongillo (2012). Consider an arbitrary $\tilde{Q} = \sum_{s=1}^{S} \lambda_s \delta(\boldsymbol{\xi}_s) \in \mathcal{P}_{\mathrm{fin}}(\mathcal{U})$ and define $Q = \sum_{s=1}^{S} \lambda_s \delta(q(\boldsymbol{\xi}_s))$. Since q is affine, $q(\Gamma(\tilde{Q})) = \sum_{s=1}^{S} \lambda_s q(\boldsymbol{\xi}_s) = \mathbb{E}_Q[\boldsymbol{D}] = \boldsymbol{\pi}_Q$. Moreover, subtracting $\ell(q(\boldsymbol{\xi}), q(\boldsymbol{\xi}))$ from the sample loss does not change its minimizer with respect to the prediction. Therefore, $\arg\min_{\boldsymbol{p}} \mathcal{R}_Q(\boldsymbol{p}) = \{\boldsymbol{\pi}_Q\}$ is equivalent to

$$\arg\max_{\boldsymbol{\eta} \in \mathcal{U}} \int_{\mathcal{U}} s[\boldsymbol{\eta}](\boldsymbol{\xi})\, \mathrm{d}\tilde{Q}(\boldsymbol{\xi}) = \{\Gamma(\tilde{Q})\}.$$

Thus, s is strictly proper for the linear property Γ. The assumed differentiability of $\boldsymbol{\eta} \mapsto L(\boldsymbol{\xi}, \boldsymbol{\eta})$, together with the finite support of $\tilde{Q}$, ensures that the expected score is directionally differentiable at every elicited value $\boldsymbol{\eta} = \Gamma(\tilde{Q})$. Hence s satisfies the Γ-differentiability condition of Abernethy & Frongillo (2012). Their characterization therefore implies that s, on $\mathrm{range}(\Gamma) = \mathcal{U}$, is equivalent to a Bregman score generated by a strictly convex potential. Equivalently, the normalized loss has a Bregman representation, up to a term that depends solely on the observations. The normalization $L(\boldsymbol{\xi}, \boldsymbol{\xi}) = 0$ eliminates this observation-only term because every Bregman divergence vanishes on the diagonal. This establishes the classical Bregman structure. The cited characterization is stated under weaker smoothness assumptions and does not directly provide the $\mathcal{C}^3$ regularity. We next establish this additional regularity under the assumptions of the present proposition.

Define a function $\boldsymbol{g}_{\boldsymbol{\eta}}(\boldsymbol{\xi}) = \nabla_{\boldsymbol{\eta}} L(\boldsymbol{\xi}, \boldsymbol{\eta}) \in \mathbb{R}^r$ for $\boldsymbol{\xi}, \boldsymbol{\eta} \in \mathcal{U}$. Fix $\boldsymbol{\eta} \in \mathcal{U}$. For any $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_s \in \mathcal{U}$ and positive weights $\lambda_1, \ldots, \lambda_S$ satisfying $\sum_{s=1}^{S} \lambda_s = 1$ and $\sum_{s=1}^{S} \lambda_s \boldsymbol{\xi}_s = \boldsymbol{\eta}$, the affine property of q gives $\sum_{s=1}^{S} \lambda_s q(\boldsymbol{\xi}_s) = q(\boldsymbol{\eta})$. Since $q(\boldsymbol{\eta})$ is the unique minimizer of the corresponding differentiable risk, its first-order condition is

$$\sum\nolimits_{s=1}^{S} \lambda_s \boldsymbol{g}_{\boldsymbol{\eta}}(\boldsymbol{\xi}_s) = \mathbf{0}_r. \tag{13}$$

Define $\mathcal{V}_{\boldsymbol{\eta}} = \mathcal{U} - \boldsymbol{\eta}$, $\tilde{\boldsymbol{g}}_{\boldsymbol{\eta}}(\boldsymbol{u}) = \boldsymbol{g}_{\boldsymbol{\eta}}(\boldsymbol{\eta} + \boldsymbol{u})$, $\boldsymbol{u} \in \mathcal{V}_{\boldsymbol{\eta}}$. Substituting $S = 1$ in Equation (13) yields $\tilde{\boldsymbol{g}}_{\boldsymbol{\eta}}(\mathbf{0}_r) = \mathbf{0}_r$. Because $\mathcal{V}_{\boldsymbol{\eta}}$ is open and contains $\mathbf{0}_r$, there exists $\rho > 0$ such that $\mathbb{B}(\mathbf{0}_r, 3\rho) \subset \mathcal{V}_{\boldsymbol{\eta}}$, where $\mathbb{B}(\boldsymbol{a}, t) = \{\boldsymbol{u} \in \mathbb{R}^r : \|\boldsymbol{u} - \boldsymbol{a}\|_2 < t\}$. For every $\boldsymbol{\omega} \in \mathbb{B}(\mathbf{0}_r, 3\rho)$, applying Equation (13) to the equally weighted law on $\boldsymbol{\eta} + \boldsymbol{\omega}$ and $\boldsymbol{\eta} - \boldsymbol{\omega}$ gives $\tilde{\boldsymbol{g}}_{\boldsymbol{\eta}}(-\boldsymbol{\omega}) = -\tilde{\boldsymbol{g}}_{\boldsymbol{\eta}}(\boldsymbol{\omega})$. For $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{B}(\mathbf{0}_r, \rho)$, applying the same condition to the equally weighted law on $\boldsymbol{\eta} + \boldsymbol{u}$, $\boldsymbol{\eta} + \boldsymbol{v}$, $\boldsymbol{\eta} - \boldsymbol{u} - \boldsymbol{v}$, then gives $\tilde{\boldsymbol{g}}_{\boldsymbol{\eta}}(\boldsymbol{u} + \boldsymbol{v}) = \tilde{\boldsymbol{g}}_{\boldsymbol{\eta}}(\boldsymbol{u}) + \tilde{\boldsymbol{g}}_{\boldsymbol{\eta}}(\boldsymbol{v})$. Thus, $\tilde{\boldsymbol{g}}_{\boldsymbol{\eta}}$ satisfies the Cauchy's functional equation on a neighborhood of the origin. Its continuity implies that there exists a matrix $\boldsymbol{A}(\boldsymbol{\eta}) \in \mathbb{R}^{r \times r}$ such that $\tilde{\boldsymbol{g}}_{\boldsymbol{\eta}}(\boldsymbol{u}) = \boldsymbol{A}(\boldsymbol{\eta})\boldsymbol{u}$ on that neighborhood. For any $\boldsymbol{u} \in \mathcal{V}_{\boldsymbol{\eta}}$, choose a sufficiently small $\epsilon > 0$ such that $-\epsilon \boldsymbol{u} \in \mathbb{B}(\mathbf{0}_r, \rho)$. The two-point law supported on $\boldsymbol{\eta} + \boldsymbol{u}$ and $\boldsymbol{\eta} - \epsilon\boldsymbol{u}$, with respective weights $\frac{\epsilon}{1+\epsilon}$ and $\frac{1}{1+\epsilon}$, has mean $\boldsymbol{\eta}$. Its first-order condition gives $\tilde{\boldsymbol{g}}_{\boldsymbol{\eta}}(\boldsymbol{u}) = \boldsymbol{A}(\boldsymbol{\eta})\boldsymbol{u}$. The identity therefore holds on all of $\mathcal{V}_{\boldsymbol{\eta}}$. Defining $\boldsymbol{H}(\boldsymbol{\eta}) = -\boldsymbol{A}(\boldsymbol{\eta})$, we obtain

$$\nabla_{\boldsymbol{\eta}} L(\boldsymbol{\xi}, \boldsymbol{\eta}) = \boldsymbol{H}(\boldsymbol{\eta})(\boldsymbol{\eta} - \boldsymbol{\xi}) \qquad \forall \boldsymbol{\xi}, \boldsymbol{\eta} \in \mathcal{U}. \tag{14}$$

Fix $\boldsymbol{\eta}_0 \in \mathcal{U}$. Since $\mathcal{U}$ is open, choose fixed points $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_r \in \mathcal{U}$ such that $\boldsymbol{\eta}_0 - \boldsymbol{\xi}_1, \ldots, \boldsymbol{\eta}_0 - \boldsymbol{\xi}_r$ are linearly independent. Define $\boldsymbol{G}(\boldsymbol{\eta}) = \big[\boldsymbol{g}_{\boldsymbol{\eta}}(\boldsymbol{\xi}_1), \ldots, \boldsymbol{g}_{\boldsymbol{\eta}}(\boldsymbol{\xi}_r)\big]$ and $\boldsymbol{V}(\boldsymbol{\eta}) = [\boldsymbol{\eta} - \boldsymbol{\xi}_1, \ldots, \boldsymbol{\eta} - \boldsymbol{\xi}_r]$. On

a neighborhood of $\boldsymbol{\eta}_0$, the matrix $\boldsymbol{V}(\boldsymbol{\eta})$ is invertible, and Equation (14) gives $\boldsymbol{H}(\boldsymbol{\eta}) = \boldsymbol{G}(\boldsymbol{\eta})\boldsymbol{V}(\boldsymbol{\eta})^{-1}$. The assumed joint continuity of the second prediction derivatives of L implies that $\boldsymbol{G}$ is continuously differentiable. Therefore, $\boldsymbol{H} \in \mathcal{C}^1(\mathcal{U})$. Differentiating Equation (14) with respect to $\boldsymbol{\eta}$, while holding $\boldsymbol{\xi}$ fixed, and then setting $\boldsymbol{\xi} = \boldsymbol{\eta}$, yields $\boldsymbol{H}(\boldsymbol{\eta}) = \nabla^2_{\boldsymbol{\eta}} L(\boldsymbol{\xi}, \boldsymbol{\eta})|_{\boldsymbol{\xi}=\boldsymbol{\eta}}$. Consequently, $\boldsymbol{H}(\boldsymbol{\eta})$ is symmetric. It is also positive semidefinite because $\boldsymbol{\eta}$ minimizes the point-mass risk $\boldsymbol{\zeta} \mapsto L(\boldsymbol{\eta}, \boldsymbol{\zeta})$.

Let $g_{\boldsymbol{\eta},i}(\boldsymbol{\xi})$ and $H_{ij}(\boldsymbol{\eta})$ denote the components of $\boldsymbol{g}_{\boldsymbol{\eta}}(\boldsymbol{\xi})$ and $\boldsymbol{H}(\boldsymbol{\eta})$, respectively. Because $\boldsymbol{g}_{\boldsymbol{\eta}}(\boldsymbol{\xi})$ is the gradient of the scalar function $L(\boldsymbol{\xi}, \boldsymbol{\eta})$, equality of mixed partial derivatives gives $\partial_{\eta_j} g_{\boldsymbol{\eta},i}(\boldsymbol{\xi}) = \partial_{\eta_i} g_{\boldsymbol{\eta},j}(\boldsymbol{\xi})$. Substituting Equation (14) and using the symmetry of $\boldsymbol{H}$ gives $\sum_{k=1}^{r} \left\{\partial_{\eta_j} H_{ik}(\boldsymbol{\eta}) - \partial_{\eta_i} H_{jk}(\boldsymbol{\eta})\right\} (\eta_k - \xi_k) = 0$. This identity holds for every $\boldsymbol{\xi}$ in the open set $\mathcal{U}$. Hence every coefficient must vanish:

$$\partial_{\eta_j} H_{ik} = \partial_{\eta_i} H_{jk} \qquad \forall i, j, k \in \{1, \ldots, r\}. \tag{15}$$

Since $\boldsymbol{H} \in \mathcal{C}^1(\mathcal{U})$, $\boldsymbol{H}$ is symmetric, and $\mathcal{U}$ is open and convex, the integrability conditions in Equation (15) allow two successive applications of the Poincaré lemma. Therefore, there exists $\Phi \in \mathcal{C}^3(\mathcal{U})$ such that $\nabla^2\Phi(\boldsymbol{\eta}) = \boldsymbol{H}(\boldsymbol{\eta})$ holds for every $\boldsymbol{\eta} \in \mathcal{U}$. Because $\boldsymbol{H}(\boldsymbol{\eta})$ is positive semidefinite throughout the convex domain $\mathcal{U}$, the function Φ is convex.

For the Bregman divergence generated from Φ, we have $B_\Phi(\boldsymbol{\xi}, \boldsymbol{\eta}) = \Phi(\boldsymbol{\xi}) - \Phi(\boldsymbol{\eta}) - \langle \nabla\Phi(\boldsymbol{\eta}), \boldsymbol{\xi} - \boldsymbol{\eta} \rangle$, direct differentiation with respect to its second argument gives $\nabla_{\boldsymbol{\eta}} B_\Phi(\boldsymbol{\xi}, \boldsymbol{\eta}) = \nabla^2\Phi(\boldsymbol{\eta})(\boldsymbol{\eta} - \boldsymbol{\xi}) = \boldsymbol{H}(\boldsymbol{\eta})(\boldsymbol{\eta} - \boldsymbol{\xi})$. According to Equation (14), we have $\nabla_{\boldsymbol{\eta}} \{L(\boldsymbol{\xi}, \boldsymbol{\eta}) - B_\Phi(\boldsymbol{\xi}, \boldsymbol{\eta})\} = \mathbf{0}_r$. Because $\mathcal{U}$ is connected, the difference is constant in $\boldsymbol{\eta}$. Evaluating it at $\boldsymbol{\eta} = \boldsymbol{\xi}$ and using $L(\boldsymbol{\xi}, \boldsymbol{\xi}) = B_\Phi(\boldsymbol{\xi}, \boldsymbol{\xi}) = 0$ shows that this constant is zero. Hence $L(\boldsymbol{\xi}, \boldsymbol{\eta}) = B_\Phi(\boldsymbol{\xi}, \boldsymbol{\eta})$ holds for every $\boldsymbol{\xi}, \boldsymbol{\eta} \in \mathcal{U}$. For a point-mass law at $q(\boldsymbol{\xi})$, strict mean-properness implies $L(\boldsymbol{\xi}, \boldsymbol{\eta}) > 0$ whenever $\boldsymbol{\eta} \neq \boldsymbol{\xi}$. Therefore, $B_\Phi(\boldsymbol{\xi}, \boldsymbol{\eta}) > 0$ whenever $\boldsymbol{\eta} \neq \boldsymbol{\xi}$, which implies that Φ is strictly convex.

For the law Q in the lemma, write $\boldsymbol{\xi}_s = q^{-1}(\boldsymbol{d}_s)$, $\boldsymbol{\eta}_Q = q^{-1}(\boldsymbol{\pi}_Q)$. Since q is affine, $\boldsymbol{\eta}_Q = \sum_{s=1}^{S} \lambda_s \boldsymbol{\xi}_s$. The observation-dependent normalization term cancels from a risk difference. Hence

$$\mathcal{R}_Q(q(\boldsymbol{\eta})) - \mathcal{R}_Q(q(\boldsymbol{\eta}_Q)) = \sum\nolimits_{s=1}^{S} \lambda_s \left[B_\Phi(\boldsymbol{\xi}_s, \boldsymbol{\eta}) - B_\Phi(\boldsymbol{\xi}_s, \boldsymbol{\eta}_Q)\right] = B_\Phi(\boldsymbol{\eta}_Q, \boldsymbol{\eta}), \tag{16}$$

where the last equality follows by expanding the Bregman divergences and using $\sum_s \lambda_s \boldsymbol{\xi}_s = \boldsymbol{\eta}_Q$. This proves both $L(\boldsymbol{\xi}, \boldsymbol{\eta}) = B_\Phi(\boldsymbol{\xi}, \boldsymbol{\eta})$ and Equation (4). □

A.4 Derivation of Equation (5)

Let $\boldsymbol{\pi} \in \operatorname{int}(\Delta^m)$ be a target label distribution, and let $\boldsymbol{h} \in \mathbb{R}^m$ satisfy $\mathbf{1}_m^\top \boldsymbol{h} = 0$ and $\boldsymbol{\pi} + \boldsymbol{h} \in \operatorname{int}(\Delta^m)$. Using $\boldsymbol{\pi} - (\boldsymbol{\pi} + \boldsymbol{h}) = -\boldsymbol{h}$, the Bregman divergence can first be written as

$$\begin{aligned} B_\Phi(\boldsymbol{\pi}, \boldsymbol{\pi} + \boldsymbol{h}) &= \Phi(\boldsymbol{\pi}) - \Phi(\boldsymbol{\pi} + \boldsymbol{h}) - \langle \nabla\Phi(\boldsymbol{\pi} + \boldsymbol{h}), -\boldsymbol{h} \rangle \\ &= \Phi(\boldsymbol{\pi}) - \Phi(\boldsymbol{\pi} + \boldsymbol{h}) + \langle \nabla\Phi(\boldsymbol{\pi} + \boldsymbol{h}), \boldsymbol{h} \rangle . \end{aligned} \tag{17}$$

For a twice continuously differentiable Φ, Taylor expansion around $\boldsymbol{\pi}$ gives

$$\Phi(\boldsymbol{\pi} + \boldsymbol{h}) = \Phi(\boldsymbol{\pi}) + \langle \nabla\Phi(\boldsymbol{\pi}), \boldsymbol{h} \rangle + \frac{1}{2}\boldsymbol{h}^\top \nabla^2\Phi(\boldsymbol{\pi})\boldsymbol{h} + o(\|\boldsymbol{h}\|_2^2), \tag{18}$$

$$\nabla\Phi(\boldsymbol{\pi} + \boldsymbol{h}) = \nabla\Phi(\boldsymbol{\pi}) + \nabla^2\Phi(\boldsymbol{\pi})\boldsymbol{h} + o(\|\boldsymbol{h}\|_2). \tag{19}$$

Substituting Equations (18) and (19) into Equation (17) yields

$$\begin{aligned} B_\Phi(\boldsymbol{\pi}, \boldsymbol{\pi} + \boldsymbol{h}) &= -\langle \nabla\Phi(\boldsymbol{\pi}), \boldsymbol{h} \rangle - \frac{1}{2}\boldsymbol{h}^\top \nabla^2\Phi(\boldsymbol{\pi})\boldsymbol{h} + \langle \nabla\Phi(\boldsymbol{\pi}), \boldsymbol{h} \rangle + \boldsymbol{h}^\top \nabla^2\Phi(\boldsymbol{\pi})\boldsymbol{h} + o(\|\boldsymbol{h}\|_2^2) \\ &= \frac{1}{2}\boldsymbol{h}^\top \nabla^2\Phi(\boldsymbol{\pi})\boldsymbol{h} + o(\|\boldsymbol{h}\|_2^2). \end{aligned}$$

For the separable potential, $\nabla^2\Phi(\boldsymbol{\pi}) = \operatorname{diag}\{\phi''(\pi_1), \ldots, \phi''(\pi_m)\}$, and hence

$$B_\Phi(\boldsymbol{\pi}, \boldsymbol{\pi} + \boldsymbol{h}) = \frac{1}{2}\sum_{i=1}^{m} \phi''(\pi_i) h_i^2 + o(\|\boldsymbol{h}\|_2^2).$$

A.5 Proof of Proposition 3.4

Proof. We first derive the polynomial branch below τ. Above the threshold, the negative-log potential has curvature $w_+(u) = u^{-2}$, $w'_+(u) = -2u^{-3}$. For the complete potential to belong to $\mathcal{C}^3([0,1])$, its curvature must be continuously differentiable at $u = \tau$. Thus, the curvature w_- below the threshold must satisfy

$$w_-(\tau) = \tau^{-2}, \qquad w'_-(\tau) = -2\tau^{-3}. \tag{20}$$

We choose the lowest-degree polynomial curvature that satisfies these matching conditions while having zero slope at the boundary $u = 0$. Specifically, let $w_-(u) \coloneqq a + bu + cu^2$, and impose $w'_-(0) = 0$. The last condition gives $b = 0$. The two conditions in Equation (20) then become $a + c\tau^2 = \tau^{-2}$, and $2c\tau = -2\tau^{-3}$. Therefore, $c = -\tau^{-4}$, $a = 2\tau^{-2}$, and hence

$$w_-(u) = \tau^{-2}2 - \tau^{-4}u^2. \tag{21}$$

We next integrate this curvature twice. To match the negative-log branch at $u = \tau$, we impose $\phi_\tau(\tau) = -\log\tau$, and $\phi'_\tau(\tau) = -\frac{1}{\tau}$. For $0 \le u < \tau$, integrating Equation (21) from τ to u gives

$$\phi'_\tau(u) = -\frac{1}{\tau} + \int_\tau^u \left(\frac{2}{\tau^2} - \frac{s^2}{\tau^4}\right) \mathrm{d}s = -\frac{8}{3\tau} + \frac{2u}{\tau^2} - \frac{u^3}{3\tau^4}. \tag{22}$$

A second integration yields

$$\phi_\tau(u) = -\log\tau + \int_\tau^u \left(-\frac{8}{3\tau} + \frac{2s}{\tau^2} - \frac{s^3}{3\tau^4}\right) \mathrm{d}s = -\log\tau + \frac{7}{4} - \frac{8u}{3\tau} + \frac{u^2}{\tau^2} - \frac{u^4}{12\tau^4}. \tag{23}$$

This proves the stated form of the lower branch. We now verify smoothness at the threshold. Evaluating the first three derivatives at $u = \tau$ gives

$$\begin{aligned} &\phi_\tau(\tau^-) = -\log\tau, \quad \phi'_\tau(\tau^-) = -\tau^{-1}, \quad \phi''_\tau(\tau^-) = \tau^{-2}, \quad \phi'''_\tau(\tau^-) = -2\tau^{-3} \\ &\phi_\tau(\tau^+) = -\log\tau, \quad \phi'_\tau(\tau^+) = -\tau^{-1}, \quad \phi''_\tau(\tau^+) = \tau^{-2}, \quad \phi'''_\tau(\tau^+) = -2\tau^{-3}. \end{aligned} \tag{24}$$

Thus, the function and its first three derivatives are continuous at $u = \tau$, proving $\phi_\tau \in \mathcal{C}^3([0,1])$. For $0 \le u \le \tau$, Equation (21) gives $\frac{1}{\tau^2} \le \phi''_\tau(u) \le \frac{2}{\tau^2}$. For $\tau \le u \le 1$, we have $1 \le u^{-2} \le \tau^{-2}$. Combining these inequalities yields $1 \le \phi''_\tau(u) \le 2\tau^{-2}$. In particular, $\phi''_\tau(u) > 0$ everywhere, so ϕ_τ is strictly convex. Since $\phi''_\tau > 0$, its first derivative is strictly increasing. Its endpoint values are $\phi'_\tau(0) = -\frac{8}{3\tau}$ and $\phi'_\tau(1) = -1$, which proves $-\frac{8}{3\tau} \le \phi'_\tau(u) \le -1$. In particular, $\phi'_\tau(u) < 0$ on $[0,1]$, so ϕ_τ is strictly decreasing. Because $\phi_\tau(0) = -\log\tau + \frac{7}{4}$ and $\phi_\tau(1) = 0$, we obtain $0 \le \phi_\tau(u) \le -\log\tau + \frac{7}{4}$. The third derivative satisfies $-\frac{2}{\tau^3} \le -\frac{2u}{\tau^4} \le 0$ holds for $0 \le u \le \tau$, and $-\frac{2}{\tau^3} \le -\frac{2}{u^3} \le -2$ holds for $\tau \le u \le 1$. This proves $-\frac{2}{\tau^3} \le \phi'''_\tau(u) \le 0$. □

A.6 Proof of Theorem 3.5

Proof. Define $\boldsymbol{v} \coloneqq \mathrm{diag}(\boldsymbol{d})^{-1}\boldsymbol{h}$, $\boldsymbol{P}_\perp \coloneqq \boldsymbol{I}_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^\top$, $\boldsymbol{\kappa} \coloneqq (\boldsymbol{I}_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^\top)\mathrm{diag}(\boldsymbol{d})^{-1}\boldsymbol{h}$, $c_\tau(u) \coloneqq \phi''_\tau(u)$, $a_{\tau,i}(\boldsymbol{d}) \coloneqq d_i^2 \cdot c_\tau(d_i)$ $\underline{a}_\tau(\boldsymbol{d}) = \min_i a_{\tau,i}(\boldsymbol{d})$, $\overline{a}_\tau(\boldsymbol{d}) = \max_i a_{\tau,i}(\boldsymbol{d})$. Differentiating the two branches of ϕ_τ yields

$$\phi''_\tau(u) = \begin{cases} u^{-2}, & u \ge \tau, \\ 2\tau^{-2} - u^2\tau^{-4}, & 0 \le u < \tau, \end{cases} \qquad \phi'''_\tau(u) = \begin{cases} -2u^{-3}, & u \ge \tau, \\ -2u\tau^{-4}, & 0 \le u < \tau. \end{cases} \tag{25}$$

The two branches agree through the third derivative at $u = \tau$. Moreover, we have

$$\sup_{u\in[0,1]} |\phi'''_\tau(u)| \le 2\tau^{-3}. \tag{26}$$

Consequently, $c_\tau = \phi''_\tau$ is globally $2\tau^{-3}$-Lipschitz on $[0,1]$. Define $g(t) = \Phi_\tau(\boldsymbol{d} + t\boldsymbol{h})$, where $t \in [0,1]$. We have the Bregman divergence:

$$B_{\Phi_\tau}(\boldsymbol{d}, \boldsymbol{d} + \boldsymbol{h}) = g(0) - g(1) + g'(1) = \int_0^1 t\, g''(t)\, \mathrm{d}t = \int_0^1 t\, \boldsymbol{h}^\top \nabla^2\Phi_\tau(\boldsymbol{d} + t\boldsymbol{h})\boldsymbol{h}\, \mathrm{d}t. \tag{27}$$

Since Φ_τ is separable, $\nabla^2\Phi_\tau(\boldsymbol{u}) = \mathrm{diag}(c_\tau(u_1), \ldots, c_\tau(u_m))$. Subtracting $\boldsymbol{h}^\top\nabla^2\Phi_\tau(\boldsymbol{d})\boldsymbol{h}/2$ from Equation (27) gives $r_{\tau,\boldsymbol{d}}(\boldsymbol{h}) = \int_0^1 t\,\boldsymbol{h}^\top\left[\nabla^2\Phi_\tau(\boldsymbol{d}+t\boldsymbol{h}) - \nabla^2\Phi_\tau(\boldsymbol{d})\right]\boldsymbol{h}\,\mathrm{d}t$. By Equation (26), we have $\left\|\nabla^2\Phi_\tau(\boldsymbol{d}+t\boldsymbol{h}) - \nabla^2\Phi_\tau(\boldsymbol{d})\right\|_{\mathrm{op}} \le \frac{2t}{\tau^3}\|\boldsymbol{h}\|_2$. Therefore,

$$|r_{\tau,\boldsymbol{d}}(\boldsymbol{h})| \le \frac{2}{\tau^3}\int_0^1 t^2\,\mathrm{d}t\,\|\boldsymbol{h}\|_2^3 = \frac{2}{3\tau^3}\|\boldsymbol{h}\|_2^3,$$

We next compare the local quadratic forms. Since $h_i = d_i v_i$, we have $\mathcal{Q}_{\tau,\boldsymbol{d}}(\boldsymbol{h}) = \frac{1}{2}\sum_{i=1}^m a_{\tau,i}(\boldsymbol{d})v_i^2$. For $0 < d_i < \tau$, writing $r_i = d_i/\tau \in (0,1)$ gives $a_{\tau,i}(\boldsymbol{d}) = r_i^2(2-r_i^2) \in (0,1)$, whereas $a_{\tau,i}(\boldsymbol{d}) = 1$ for $d_i \ge \tau$. It follows that

$$\frac{\underline{a}_\tau(\boldsymbol{d})}{2}\|\boldsymbol{v}\|_2^2 \le \mathcal{Q}_{\tau,\boldsymbol{d}}(\boldsymbol{h}) \le \frac{\overline{a}_\tau(\boldsymbol{d})}{2}\|\boldsymbol{v}\|_2^2. \tag{28}$$

The tangent constraint implies $\boldsymbol{d}^\top\boldsymbol{v} = \mathbf{1}_m^\top\boldsymbol{h} = 0$. Let $\boldsymbol{u} = \boldsymbol{P}_\perp\boldsymbol{v}$, $\bar{v} = m^{-1}\mathbf{1}_m^\top\boldsymbol{v}$, so that $\boldsymbol{v} = \boldsymbol{u}+\bar{v}\mathbf{1}_m$. Because $\boldsymbol{d}^\top\mathbf{1}_m = 1$ and $\mathbf{1}_m^\top\boldsymbol{u} = 0$, the tangent constraint gives $\bar{v} = -\boldsymbol{d}^\top\boldsymbol{u} = -(\boldsymbol{P}_\perp\boldsymbol{d})^\top\boldsymbol{u}$. Therefore, Cauchy-Schwarz yields

$$\|\boldsymbol{v}\|_2^2 = \|\boldsymbol{u}\|_2^2 + m\bar{v}^2 \le \left(1 + m\|\boldsymbol{P}_\perp\boldsymbol{d}\|_2^2\right)\|\boldsymbol{u}\|_2^2 = m\|\boldsymbol{d}\|_2^2\|\boldsymbol{P}_\perp\boldsymbol{v}\|_2^2. \tag{29}$$

The reverse inequality $\|\boldsymbol{P}_\perp\boldsymbol{v}\|_2^2 \le \|\boldsymbol{v}\|_2^2$ follows from the non-expansiveness of orthogonal projection. Combining these inequalities with Equation (28) proves Equation (10). Finally, $\log(\boldsymbol{d}+\boldsymbol{h}) - \log\boldsymbol{d} = \log(\mathbf{1}_m + \boldsymbol{v}) = \boldsymbol{v} + \mathcal{O}(\|\boldsymbol{h}\|_2^2)$. Hence, $d_{\mathrm{A}}^2(\boldsymbol{d}, \boldsymbol{d}+\boldsymbol{h}) = \|\boldsymbol{P}_\perp\left[\log(\boldsymbol{d}+\boldsymbol{h}) - \log\boldsymbol{d}\right]\|_2^2 = \|\boldsymbol{P}_\perp\boldsymbol{v}\|_2^2 + \mathcal{O}(\|\boldsymbol{h}\|_2^3)$, which proves $d_{\mathrm{A}}^2(\boldsymbol{d}, \boldsymbol{d}+\boldsymbol{h}) = \|\boldsymbol{\kappa}\|_2^2 + \mathcal{O}(\|\boldsymbol{h}\|_2^3)$. Finally, if $\tau \le d_i < d_j$, substituting $c_\tau(u) = u^{-2}$, $c_{\mathrm{KL}}(u) = u^{-1}$, and $c_{\mathrm{SE}}(u) = 2$ gives Equation (11). □

A.7 Proof of Proposition 3.6

Proof. We first verify the curvature of the generating potential. If $\tau \le u \le 1$, then $u^{-2} \ge 1$. If $0 \le u < \tau$, then $\phi_\tau''(u) = \tau^{-4}2\tau^2 - u^2 \ge \tau^{-2} \ge 1$. Thus, Φ_τ is 1-strongly convex on the simplex. In particular, for any $\boldsymbol{a}, \boldsymbol{b} \in \Delta^m$, the second-order integral representation of the Bregman divergence gives

$$\begin{aligned} B_{\Phi_\tau}(\boldsymbol{a}, \boldsymbol{b}) &= \int_0^1 (1-t)(\boldsymbol{a}-\boldsymbol{b})^\top\nabla^2\Phi_\tau\big(\boldsymbol{b}+t(\boldsymbol{a}-\boldsymbol{b})\big)(\boldsymbol{a}-\boldsymbol{b})\,\mathrm{d}t \\ &\ge \int_0^1 (1-t)\,\mathrm{d}t\,\|\boldsymbol{a}-\boldsymbol{b}\|_2^2 = \frac{1}{2}\|\boldsymbol{a}-\boldsymbol{b}\|_2^2. \end{aligned} \tag{30}$$

Therefore, $B_{\Phi_\tau}(\boldsymbol{a}, \boldsymbol{b}) \ge 0$, with equality if and only if $\boldsymbol{a} = \boldsymbol{b}$. We now compute the conditional risk. By the definition of the Bregman divergence and the linearity of conditional expectation,

$$\begin{aligned} \mathcal{R}_\tau(\boldsymbol{p}\,|\,\boldsymbol{x}) &= \mathbb{E}[\Phi_\tau(\boldsymbol{D}) - \Phi_\tau(\boldsymbol{p}) - \langle\nabla\Phi_\tau(\boldsymbol{p}), \boldsymbol{D}-\boldsymbol{p}\rangle|\boldsymbol{x}] \\ &= \mathbb{E}[\Phi_\tau(\boldsymbol{D}) \mid \boldsymbol{x}] - \Phi_\tau(\boldsymbol{p}) - \langle\nabla\Phi_\tau(\boldsymbol{p}), \boldsymbol{\pi}(\boldsymbol{x}) - \boldsymbol{p}\rangle\,. \end{aligned} \tag{31}$$

At $\boldsymbol{p} = \boldsymbol{\pi}(\boldsymbol{x})$, this becomes $\mathcal{R}_\tau(\boldsymbol{\pi}(\boldsymbol{x})\,|\,\boldsymbol{x}) = \mathbb{E}[\Phi_\tau(\boldsymbol{D}) \mid \boldsymbol{x}] - \Phi_\tau(\boldsymbol{\pi}(\boldsymbol{x}))$. Subtracting the latter equality from Equation (31) yields

$$\mathcal{R}_\tau(\boldsymbol{p}\,|\,\boldsymbol{x}) - \mathcal{R}_\tau(\boldsymbol{\pi}(\boldsymbol{x})\,|\,\boldsymbol{x}) = \Phi_\tau(\boldsymbol{\pi}(\boldsymbol{x})) - \Phi_\tau(\boldsymbol{p}) - \langle\nabla\Phi_\tau(\boldsymbol{p}), \boldsymbol{\pi}(\boldsymbol{x}) - \boldsymbol{p}\rangle = B_{\Phi_\tau}(\boldsymbol{\pi}(\boldsymbol{x}), \boldsymbol{p}). \tag{32}$$

Applying Equation (30) proves Equation (12). The conditional excess risk is therefore zero if and only if $\boldsymbol{p} = \boldsymbol{\pi}(\boldsymbol{x})$, establishing the unique Bayes prediction. Finally, Proposition 3.4 ensures that ϕ_τ and ϕ_τ' are finite on $[0,1]$. Consequently, the loss and the preceding expectations remain well defined even when $\boldsymbol{D}$, $\boldsymbol{\pi}(\boldsymbol{x})$, or $\boldsymbol{p}$ contains zero coordinates. □

B Experimental Details

B.1 Generation of Figure 1

Small KL divergence or SE loss can coexist with denominator-amplified log-ratio distortions. Subfigures (a) and (b): Joint-variation experiment based on $n = 5{,}000$ pairs of strictly positive

$K = 10$ label distributions. For each pair, the total mass M of two rare labels was sampled log-uniformly from $[10^{-4}, 0.25]$, their initial ratio was sampled log-uniformly from $[0.5, 2]$, and the remaining mass was distributed uniformly across the other eight labels. The internal log-ratio of the rare-label pair was then perturbed by a randomly signed shift with magnitude sampled log-uniformly from $[0.05, 4]$, while preserving M and leaving all other labels unchanged. Panels show the resulting global log-ratio error, $E_{\mathrm{LR}} = \|\psi(\boldsymbol{p}) - \psi(\boldsymbol{d})\|_2$, against subfigure a, $\mathrm{KL}(\boldsymbol{d}\|\boldsymbol{p})$, and subfigure b, $\|\boldsymbol{d} - \boldsymbol{p}\|_2^2$. Red curves and shaded regions denote medians and interquartile ranges over 18 equal-count bins, respectively. Subfigures (c) and (d): Controlled denominator intervention. Targets were constructed as $\boldsymbol{d} = (a, b, c/m, \ldots, c/m)$ and predictions as $\boldsymbol{p} = (a, b + \epsilon, (c - \epsilon)/m, \ldots, (c - \epsilon)/m)$, where $a = 0.25$, $m = K - 2 = 8$, $c = 1 - a - b$, and the absolute perturbation $\epsilon \in \{10^{-5}, 3 \times 10^{-5}, 10^{-4}, 3 \times 10^{-4}\}$. The denominator mass b was the sole free intervention and was varied over 200 logarithmically spaced values from 10^{-6} to 10^{-2}; the numerator, perturbation direction, perturbation magnitude, and simplex-closure rule were held fixed. This construction yields the pairwise distortion $\log(1 + \epsilon/b)$ while keeping $\|\boldsymbol{d} - \boldsymbol{p}\|_2^2 = \epsilon^2(1 + 1/m)$ constant. Panels report the corresponding KL divergence (subfigure c) and squared Euclidean loss (subfigure d). Points are coloured by the minimum target mass, which equals b in the controlled experiment. All vertical axes are logarithmic.

B.2 Details of Datasets

Jaffe (Lyons et al., 1998) contains 213 facial-expression images, each evaluated by 60 annotators who independently rated the intensities of six basic emotions (happiness, sadness, surprise, fear, anger, and disgust) on a five-point scale ranging from 1 (lowest intensity) to 5 (highest intensity). The averaged emotion intensities were subsequently normalized to obtain the corresponding label distribution. BU-3DFE (Yin et al., 2006) follows a similar annotation protocol for 2,500 images, with each image evaluated by 23 annotators using the same six emotion categories and five-level intensity scale. RAF-ML (Li & Deng, 2019) comprises 4,908 images annotated by a pool of 315 trained annotators. Each image received approximately 40 independent annotations, with each annotator selecting one of the six basic emotions. The resulting categorical votes were aggregated across annotators and normalized to form a six-dimensional label distribution. SCUT-FBP (Ren & Geng, 2017) contains 1,500 facial images evaluated by a panel of 75 annotators, each of whom assigned an ordinal attractiveness score from 1 to 5. The proportions of ratings at the five levels constitute the corresponding five-dimensional label distribution. Similarly, each image in FBP5500 (Liang et al., 2018) (5,500 images) was evaluated by 60 volunteers using a single attractiveness score from 1 to 5, where 5 denotes the highest level of attractiveness; the empirical distribution of the 60 ratings over the five score levels serves as the label distribution. M2B (Nguyen et al., 2012) (1,240 images and five labels) adopts a comparative rather than absolute annotation protocol, in which 40 participants ranked randomly presented groups of ten faces from most to least attractive. The Music dataset (Lee et al., 2021) contains 360 songs evaluated by 166 participants from Brazil, South Korea, and the United States. Participants listened to 20-second excerpts and independently rated nine mood attributes on a four-point scale ranging from "not at all" to "very." Each song received approximately 31.9 participant–song evaluations on average across the three cultural groups, and the aggregated mood intensities were normalized to produce a nine-dimensional label distribution. Finally, Movie (Geng & Hou, 2015) contains 7,755 films and 54,242,292 ratings contributed by 478,656 Netflix users, corresponding to an average of 6,994 ratings per film. Each user assigned an integer rating from 1 to 5 stars, and the empirical frequencies of the five rating levels were used to construct the label distribution for each film.

B.3 Details of Experimental Protocol

To simulate finite-sample estimates of the underlying population label distributions, we treat the original label distribution $\boldsymbol{d}_i$ of each instance as a high-precision reference and generate finite annotations under two levels of annotation counts, $n_a \in \{10, 30\}$. Specifically, for each instance, we draw annotation counts according to $\boldsymbol{c}_i \sim \mathrm{Multinomial}(n_a, \boldsymbol{d}_i)$, and construct the corresponding finite-sample training label distribution as $\boldsymbol{d}_i' = \boldsymbol{c}_i / n_a$. To ensure comparability across annotation counts, we use a nested sampling scheme: a single sequence of 40 categorical draws is generated for each instance, and the first n_a draws are used to construct the finite-sample label distribution for count n_a. Models are trained on the training instances using only these finite-sample label distributions and are evaluated on the test instances against their original label distributions. Hyperparameters are

Table 2: Chebyshev Distance, Cosine Coefficient, and Mean Absolute Log-Ratio Error under Finite-Sample Label Distributions at $n_a = 10$.

	Method	Jaffe	BU-3DFE	RAF-ML	SCUT-FBP	FBP5500	M2B	Music	Movie
Chebyshev Distance (↓)	BAP	$\bullet.123_{(\pm.006)}$	$\bullet.110_{(\pm.002)}$	$\bullet.170_{(\pm.003)}$	$\bullet.248_{(\pm.005)}$	$\bullet.137_{(\pm.002)}$	$\bullet.365_{(\pm.010)}$	$\bullet.087_{(\pm.003)}$	$\bullet.117_{(\pm.001)}$
	KL	$\prec.204_{(\pm.011)}$	$\prec.129_{(\pm.003)}$	$\succ.165_{(\pm.003)}$	$\prec.280_{(\pm.004)}$	$\prec.157_{(\pm.003)}$	$\prec.381_{(\pm.008)}$	$\prec.152_{(\pm.010)}$	$\prec.119_{(\pm.001)}$
	SE	$\prec.141_{(\pm.006)}$	$\prec.128_{(\pm.003)}$	$\prec.172_{(\pm.003)}$	$\prec.275_{(\pm.005)}$	$\prec.161_{(\pm.002)}$	$\prec.383_{(\pm.008)}$	$\prec.101_{(\pm.004)}$	$\prec.118_{(\pm.001)}$
	SA	$\prec.756_{(\pm.015)}$	$\prec.618_{(\pm.007)}$	$\prec.434_{(\pm.003)}$	$\prec.531_{(\pm.012)}$	$\prec.396_{(\pm.005)}$	$\prec.549_{(\pm.015)}$	$\prec.806_{(\pm.011)}$	$\prec.603_{(\pm.007)}$
	KL-SA	$\prec.757_{(\pm.014)}$	$\prec.220_{(\pm.008)}$	$\prec.360_{(\pm.015)}$	$\prec.339_{(\pm.008)}$	$\prec.289_{(\pm.004)}$	$\prec.466_{(\pm.011)}$	$\prec.770_{(\pm.025)}$	$\prec.312_{(\pm.004)}$
	SE-SA	$\prec.757_{(\pm.014)}$	$\prec.480_{(\pm.011)}$	$\prec.423_{(\pm.002)}$	$\prec.335_{(\pm.006)}$	$\prec.365_{(\pm.004)}$	$\prec.535_{(\pm.012)}$	$\prec.800_{(\pm.013)}$	$\prec.382_{(\pm.006)}$
	Delta	$\prec.115_{(\pm.009)}$	$\prec.119_{(\pm.003)}$	$\prec.330_{(\pm.015)}$	$\prec.248_{(\pm.005)}$	$\succ.140_{(\pm.002)}$	$\prec.372_{(\pm.011)}$	$\prec.080_{(\pm.003)}$	$\sim.126_{(\pm.003)}$
	Delta†	$\bullet.113_{(\pm.007)}$	$\bullet.117_{(\pm.002)}$	$\bullet.306_{(\pm.013)}$	$\bullet.246_{(\pm.006)}$	$\bullet.142_{(\pm.002)}$	$\bullet.355_{(\pm.014)}$	$\bullet.076_{(\pm.003)}$	$\bullet.126_{(\pm.002)}$
	DVS	$\prec.132_{(\pm.016)}$	$\sim.119_{(\pm.003)}$	$\sim.334_{(\pm.016)}$	$\prec.252_{(\pm.005)}$	$\succ.140_{(\pm.002)}$	$\prec.376_{(\pm.009)}$	$\sim.083_{(\pm.011)}$	$\prec.141_{(\pm.005)}$
	DVS†	$\bullet.127_{(\pm.011)}$	$\bullet.118_{(\pm.003)}$	$\bullet.330_{(\pm.012)}$	$\bullet.251_{(\pm.004)}$	$\bullet.143_{(\pm.002)}$	$\bullet.370_{(\pm.010)}$	$\bullet.081_{(\pm.004)}$	$\bullet.132_{(\pm.003)}$
	RkNN	$\prec.119_{(\pm.008)}$	$\prec.126_{(\pm.002)}$	$\prec.316_{(\pm.003)}$	$\succ.261_{(\pm.005)}$	$\prec.153_{(\pm.002)}$	$\sim.382_{(\pm.008)}$	$\prec.085_{(\pm.005)}$	$\prec.122_{(\pm.001)}$
	RkNN†	$\bullet.111_{(\pm.005)}$	$\bullet.118_{(\pm.002)}$	$\bullet.302_{(\pm.011)}$	$\bullet.268_{(\pm.005)}$	$\bullet.144_{(\pm.001)}$	$\bullet.370_{(\pm.020)}$	$\bullet.079_{(\pm.002)}$	$\bullet.119_{(\pm.001)}$
Cosine Coefficient (↑)	BAP	$\bullet.917_{(\pm.007)}$	$\bullet.941_{(\pm.002)}$	$\bullet.915_{(\pm.003)}$	$\bullet.843_{(\pm.008)}$	$\bullet.953_{(\pm.002)}$	$\bullet.750_{(\pm.012)}$	$\bullet.906_{(\pm.005)}$	$\bullet.934_{(\pm.002)}$
	KL	$\prec.824_{(\pm.013)}$	$\prec.914_{(\pm.004)}$	$\succ.919_{(\pm.002)}$	$\prec.808_{(\pm.005)}$	$\prec.940_{(\pm.002)}$	$\prec.738_{(\pm.010)}$	$\prec.813_{(\pm.012)}$	$\prec.934_{(\pm.001)}$
	SE	$\prec.896_{(\pm.008)}$	$\prec.916_{(\pm.003)}$	$\prec.913_{(\pm.003)}$	$\prec.816_{(\pm.005)}$	$\prec.937_{(\pm.002)}$	$\prec.740_{(\pm.009)}$	$\prec.888_{(\pm.007)}$	$\sim.934_{(\pm.001)}$
	SA	$\prec.430_{(\pm.018)}$	$\prec.518_{(\pm.008)}$	$\prec.788_{(\pm.005)}$	$\prec.634_{(\pm.015)}$	$\prec.814_{(\pm.006)}$	$\prec.559_{(\pm.018)}$	$\prec.338_{(\pm.011)}$	$\prec.568_{(\pm.007)}$
	KL-SA	$\prec.431_{(\pm.017)}$	$\prec.814_{(\pm.008)}$	$\prec.824_{(\pm.009)}$	$\prec.757_{(\pm.009)}$	$\prec.867_{(\pm.004)}$	$\prec.621_{(\pm.013)}$	$\prec.343_{(\pm.013)}$	$\prec.767_{(\pm.004)}$
	SE-SA	$\prec.431_{(\pm.017)}$	$\prec.613_{(\pm.006)}$	$\prec.798_{(\pm.003)}$	$\prec.773_{(\pm.008)}$	$\prec.840_{(\pm.005)}$	$\prec.574_{(\pm.013)}$	$\prec.340_{(\pm.011)}$	$\prec.733_{(\pm.005)}$
	Delta	$\sim.929_{(\pm.010)}$	$\prec.936_{(\pm.003)}$	$\prec.759_{(\pm.025)}$	$\succ.842_{(\pm.007)}$	$\sim.951_{(\pm.002)}$	$\succ.768_{(\pm.014)}$	$\prec.916_{(\pm.006)}$	$\succ.927_{(\pm.002)}$
	Delta†	$\bullet.931_{(\pm.008)}$	$\bullet.937_{(\pm.002)}$	$\bullet.793_{(\pm.019)}$	$\bullet.840_{(\pm.007)}$	$\bullet.951_{(\pm.002)}$	$\bullet.745_{(\pm.019)}$	$\bullet.920_{(\pm.005)}$	$\bullet.925_{(\pm.002)}$
	DVS	$\prec.909_{(\pm.020)}$	$\sim.936_{(\pm.003)}$	$\prec.751_{(\pm.027)}$	$\sim.842_{(\pm.007)}$	$\succ.951_{(\pm.002)}$	$\sim.770_{(\pm.011)}$	$\sim.911_{(\pm.017)}$	$\prec.914_{(\pm.005)}$
	DVS†	$\bullet.914_{(\pm.014)}$	$\bullet.937_{(\pm.002)}$	$\bullet.762_{(\pm.020)}$	$\bullet.842_{(\pm.007)}$	$\bullet.951_{(\pm.002)}$	$\bullet.768_{(\pm.013)}$	$\bullet.915_{(\pm.007)}$	$\bullet.922_{(\pm.002)}$
	RkNN	$\prec.929_{(\pm.008)}$	$\prec.928_{(\pm.001)}$	$\sim.784_{(\pm.002)}$	$\succ.839_{(\pm.008)}$	$\prec.944_{(\pm.002)}$	$\sim.762_{(\pm.011)}$	$\prec.908_{(\pm.006)}$	$\prec.931_{(\pm.001)}$
	RkNN†	$\bullet.937_{(\pm.004)}$	$\bullet.936_{(\pm.001)}$	$\bullet.774_{(\pm.019)}$	$\bullet.816_{(\pm.006)}$	$\bullet.950_{(\pm.002)}$	$\bullet.754_{(\pm.023)}$	$\bullet.918_{(\pm.005)}$	$\bullet.933_{(\pm.001)}$
Mean Absolute Log-Ratio Error (↓)	BAP	$\bullet.527_{(\pm.020)}$	$\bullet.389_{(\pm.006)}$	$\bullet4.76_{(\pm.016)}$	$\bullet5.53_{(\pm.075)}$	$\bullet.741_{(\pm.017)}$	$\bullet1.68_{(\pm.033)}$	$\bullet.669_{(\pm.023)}$	$\bullet.646_{(\pm.007)}$
	KL	$\prec.917_{(\pm.037)}$	$\prec.513_{(\pm.013)}$	$\prec4.78_{(\pm.017)}$	$\prec5.61_{(\pm.072)}$	$\prec.933_{(\pm.017)}$	$\prec1.74_{(\pm.032)}$	$\prec.934_{(\pm.027)}$	$\prec.651_{(\pm.006)}$
	SE	$\prec.616_{(\pm.022)}$	$\prec.513_{(\pm.012)}$	$\prec4.81_{(\pm.019)}$	$\sim5.87_{(\pm.926)}$	$\prec1.01_{(\pm.022)}$	$\prec2.51_{(\pm.311)}$	$\prec.718_{(\pm.024)}$	$\prec.653_{(\pm.006)}$
	SA	$\prec17.4_{(\pm1.79)}$	$\prec5.08_{(\pm.116)}$	$\prec8.55_{(\pm.071)}$	$\prec14.4_{(\pm1.02)}$	$\prec9.64_{(\pm.151)}$	$\prec12.1_{(\pm.194)}$	$\prec15.5_{(\pm1.06)}$	$\prec9.76_{(\pm.118)}$
	KL-SA	$\prec16.4_{(\pm1.70)}$	$\prec.977_{(\pm.036)}$	$\prec6.08_{(\pm.343)}$	$\prec6.03_{(\pm.268)}$	$\prec5.78_{(\pm.129)}$	$\prec10.1_{(\pm.241)}$	$\prec12.3_{(\pm2.40)}$	$\prec3.47_{(\pm.062)}$
	SE-SA	$\prec16.2_{(\pm1.61)}$	$\prec2.86_{(\pm.126)}$	$\prec8.15_{(\pm.077)}$	$\prec8.51_{(\pm.246)}$	$\prec8.94_{(\pm.142)}$	$\prec11.8_{(\pm.278)}$	$\prec14.9_{(\pm.912)}$	$\prec5.74_{(\pm.081)}$
	Delta	$\sim.481_{(\pm.053)}$	$\sim.384_{(\pm.008)}$	$\prec5.55_{(\pm.051)}$	$\prec5.52_{(\pm.072)}$	$\prec.764_{(\pm.020)}$	$\prec1.73_{(\pm.045)}$	$\prec.646_{(\pm.022)}$	$\sim.679_{(\pm.009)}$
	Delta†	$\bullet.471_{(\pm.042)}$	$\bullet.383_{(\pm.006)}$	$\bullet5.48_{(\pm.050)}$	$\bullet5.45_{(\pm.090)}$	$\bullet.742_{(\pm.017)}$	$\bullet1.64_{(\pm.087)}$	$\bullet.632_{(\pm.023)}$	$\bullet.679_{(\pm.009)}$
	DVS	$\prec.566_{(\pm.090)}$	$\sim.385_{(\pm.009)}$	$\prec5.57_{(\pm.052)}$	$\prec5.59_{(\pm.073)}$	$\prec.762_{(\pm.020)}$	$\prec1.80_{(\pm.036)}$	$\sim.660_{(\pm.042)}$	$\prec.721_{(\pm.019)}$
	DVS†	$\bullet.543_{(\pm.062)}$	$\bullet.384_{(\pm.007)}$	$\bullet5.55_{(\pm.040)}$	$\bullet5.58_{(\pm.074)}$	$\bullet.743_{(\pm.016)}$	$\bullet1.73_{(\pm.045)}$	$\bullet.646_{(\pm.023)}$	$\bullet.702_{(\pm.009)}$
	RkNN	$\prec.452_{(\pm.024)}$	$\prec.410_{(\pm.007)}$	$\prec5.47_{(\pm.024)}$	$\prec5.64_{(\pm.081)}$	$\prec.778_{(\pm.017)}$	$\sim1.80_{(\pm.034)}$	$\prec.685_{(\pm.027)}$	$\prec.692_{(\pm.007)}$
	RkNN†	$\bullet.432_{(\pm.019)}$	$\bullet.384_{(\pm.006)}$	$\bullet5.36_{(\pm.033)}$	$\bullet5.54_{(\pm.103)}$	$\bullet.765_{(\pm.015)}$	$\bullet1.77_{(\pm.063)}$	$\bullet.641_{(\pm.025)}$	$\bullet.661_{(\pm.007)}$

tuned by five-fold cross-validation using only the training set. The entire procedure is repeated ten times for statistical reporting.

B.4 Details of Comparison Algorithms

Each loss function is used to train a linear model with softmax activation function. The optimization algorithm is L-BFGS, where the maximum iteration is $5{,}000$, tolerance of gradient variation is 10^{-8}, tolerance of loss function variation is 10^{-8}. For SA, KL-SA, and SE-SA, a small value of 10^{-8} is added to the zero label mass to avoid singularities. Hyperparameters are selected separately for each dataset and annotation count. For BAP, the cap threshold is searched over $\tau \in \{0.01, 0.025, 0.05, 0.1\}$. For KL-SA and SE-SA, which denote $\mathrm{KL} + \lambda\mathrm{SA}$ and $\mathrm{SE} + \lambda\mathrm{SA}$, respectively, the mixing coefficient was chosen from $\lambda \in \{0.01, 0.025, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10\}$. For Delta-LDL and RkNN-LDL, the hyperparameter selecting method follows their respective literatures, and the BAP-based counterparts follows the same selecting method. In LDL-DVS, $\boldsymbol{\rho}$ and $\boldsymbol{\eta}$ denote the positive and negative semantic weight vectors, respectively, and are predefined for each dataset according to its label semantics. For facial-expression datasets (Jaffe, BU-3DFE, RAF-ML), happiness and surprise are assigned to the positive group, whereas sadness, anger, disgust, and fear are assigned to the negative

Table 3: Clark Distance, Canberra Metric, KL Divergence, and Intersection Similarity under Finite-Sample Label Distributions at $n_a = 30$.

	Method	Jaffe	BU-3DFE	RAF-ML	SCUT-FBP	FBP5500	M2B	Music	Movie
Clark Distance (↓)	BAP	$\bullet.392_{(\pm.015)}$	$\bullet.357_{(\pm.004)}$	$\bullet1.45_{(\pm.004)}$	$\bullet1.38_{(\pm.013)}$	$\bullet1.27_{(\pm.006)}$	$\bullet1.09_{(\pm.014)}$	$\bullet.734_{(\pm.025)}$	$\bullet.516_{(\pm.005)}$
	KL	$\prec.541_{(\pm.028)}$	$\prec.403_{(\pm.008)}$	$\succ1.42_{(\pm.003)}$	$\prec1.42_{(\pm.013)}$	$\prec1.30_{(\pm.005)}$	$\prec1.15_{(\pm.011)}$	$\prec.867_{(\pm.026)}$	$\prec.519_{(\pm.004)}$
	SE	$\prec.431_{(\pm.018)}$	$\prec.404_{(\pm.007)}$	$\succ1.43_{(\pm.003)}$	$\sim1.38_{(\pm.060)}$	$\prec1.31_{(\pm.005)}$	$\prec1.39_{(\pm.071)}$	$\prec.752_{(\pm.024)}$	$\prec.525_{(\pm.005)}$
	SA	$\prec1.97_{(\pm.077)}$	$\prec1.18_{(\pm.029)}$	$\prec2.06_{(\pm.004)}$	$\prec1.88_{(\pm.012)}$	$\prec1.85_{(\pm.005)}$	$\prec1.97_{(\pm.007)}$	$\prec2.72_{(\pm.027)}$	$\prec1.48_{(\pm.014)}$
	KL-SA	$\prec1.95_{(\pm.072)}$	$\prec.443_{(\pm.009)}$	$\prec1.86_{(\pm.006)}$	$\prec1.55_{(\pm.033)}$	$\prec1.68_{(\pm.003)}$	$\prec1.82_{(\pm.011)}$	$\prec2.37_{(\pm.046)}$	$\prec.893_{(\pm.012)}$
	SE-SA	$\prec1.95_{(\pm.070)}$	$\prec.544_{(\pm.014)}$	$\prec2.03_{(\pm.004)}$	$\prec1.76_{(\pm.007)}$	$\prec1.79_{(\pm.004)}$	$\prec1.94_{(\pm.017)}$	$\prec2.69_{(\pm.023)}$	$\prec1.12_{(\pm.017)}$
	Delta	$\sim.374_{(\pm.017)}$	$\prec.373_{(\pm.008)}$	$\prec1.54_{(\pm.016)}$	$\succ1.38_{(\pm.013)}$	$\prec1.27_{(\pm.006)}$	$\prec1.13_{(\pm.016)}$	$\sim.729_{(\pm.026)}$	$\sim.532_{(\pm.007)}$
	Delta†	$\bullet.371_{(\pm.017)}$	$\bullet.369_{(\pm.006)}$	$\bullet1.52_{(\pm.015)}$	$\bullet1.39_{(\pm.014)}$	$\bullet1.27_{(\pm.006)}$	$\bullet.999_{(\pm.029)}$	$\bullet.722_{(\pm.024)}$	$\bullet.529_{(\pm.005)}$
	DVS	$\sim.410_{(\pm.042)}$	$\sim.373_{(\pm.008)}$	$\prec1.54_{(\pm.016)}$	$\sim1.38_{(\pm.012)}$	$\prec1.27_{(\pm.007)}$	$\prec1.22_{(\pm.012)}$	$\sim.732_{(\pm.029)}$	$\sim.547_{(\pm.007)}$
	DVS†	$\bullet.394_{(\pm.022)}$	$\bullet.372_{(\pm.007)}$	$\bullet1.53_{(\pm.013)}$	$\bullet1.38_{(\pm.012)}$	$\bullet1.27_{(\pm.005)}$	$\bullet1.13_{(\pm.019)}$	$\bullet.728_{(\pm.025)}$	$\bullet.545_{(\pm.006)}$
	RkNN	$\prec.417_{(\pm.020)}$	$\prec.388_{(\pm.002)}$	$\prec1.51_{(\pm.004)}$	$\prec1.39_{(\pm.012)}$	$\sim1.27_{(\pm.006)}$	$\prec1.22_{(\pm.021)}$	$\prec.771_{(\pm.028)}$	$\prec.551_{(\pm.005)}$
	RkNN†	$\bullet.377_{(\pm.014)}$	$\bullet.367_{(\pm.004)}$	$\bullet1.48_{(\pm.007)}$	$\bullet1.38_{(\pm.015)}$	$\bullet1.27_{(\pm.006)}$	$\bullet1.14_{(\pm.035)}$	$\bullet.728_{(\pm.027)}$	$\bullet.529_{(\pm.004)}$
Canberra Metric (↓)	BAP	$\bullet.805_{(\pm.034)}$	$\bullet.748_{(\pm.008)}$	$\bullet2.84_{(\pm.013)}$	$\bullet2.56_{(\pm.032)}$	$\bullet2.17_{(\pm.014)}$	$\bullet2.08_{(\pm.034)}$	$\bullet1.86_{(\pm.066)}$	$\bullet.991_{(\pm.010)}$
	KL	$\prec1.12_{(\pm.063)}$	$\prec.837_{(\pm.017)}$	$\succ2.76_{(\pm.010)}$	$\prec2.67_{(\pm.033)}$	$\prec2.24_{(\pm.011)}$	$\prec2.20_{(\pm.025)}$	$\prec2.22_{(\pm.076)}$	$\prec.994_{(\pm.009)}$
	SE	$\prec.891_{(\pm.043)}$	$\prec.837_{(\pm.016)}$	$\succ2.78_{(\pm.012)}$	$\sim2.61_{(\pm.117)}$	$\prec2.27_{(\pm.012)}$	$\prec2.75_{(\pm.193)}$	$\prec1.91_{(\pm.063)}$	$\prec1.00_{(\pm.010)}$
	SA	$\prec4.58_{(\pm.218)}$	$\prec2.55_{(\pm.070)}$	$\prec4.70_{(\pm.012)}$	$\prec3.93_{(\pm.037)}$	$\prec3.63_{(\pm.018)}$	$\prec4.16_{(\pm.021)}$	$\prec7.93_{(\pm.102)}$	$\prec2.99_{(\pm.032)}$
	KL-SA	$\prec4.53_{(\pm.202)}$	$\prec.918_{(\pm.021)}$	$\prec4.10_{(\pm.018)}$	$\prec3.03_{(\pm.075)}$	$\prec3.07_{(\pm.012)}$	$\prec3.71_{(\pm.033)}$	$\prec6.69_{(\pm.161)}$	$\prec1.67_{(\pm.023)}$
	SE-SA	$\prec4.51_{(\pm.194)}$	$\prec1.12_{(\pm.031)}$	$\prec4.63_{(\pm.009)}$	$\prec3.47_{(\pm.021)}$	$\prec3.38_{(\pm.021)}$	$\prec4.08_{(\pm.052)}$	$\prec7.84_{(\pm.085)}$	$\prec2.06_{(\pm.030)}$
	Delta	$\sim.770_{(\pm.036)}$	$\prec.792_{(\pm.017)}$	$\prec3.26_{(\pm.061)}$	$\succ2.57_{(\pm.033)}$	$\prec2.17_{(\pm.015)}$	$\prec2.14_{(\pm.046)}$	$\sim1.84_{(\pm.077)}$	$\sim1.02_{(\pm.013)}$
	Delta†	$\bullet.760_{(\pm.040)}$	$\bullet.782_{(\pm.012)}$	$\bullet3.19_{(\pm.056)}$	$\bullet2.58_{(\pm.035)}$	$\bullet2.16_{(\pm.016)}$	$\bullet1.86_{(\pm.051)}$	$\bullet1.82_{(\pm.066)}$	$\bullet1.02_{(\pm.008)}$
	DVS	$\sim.843_{(\pm.082)}$	$\sim.792_{(\pm.018)}$	$\prec3.28_{(\pm.063)}$	$\sim2.57_{(\pm.030)}$	$\prec2.17_{(\pm.015)}$	$\prec2.33_{(\pm.034)}$	$\sim1.85_{(\pm.089)}$	$\sim1.05_{(\pm.013)}$
	DVS†	$\bullet.806_{(\pm.039)}$	$\bullet.789_{(\pm.014)}$	$\bullet3.26_{(\pm.052)}$	$\bullet2.57_{(\pm.031)}$	$\bullet2.16_{(\pm.012)}$	$\bullet2.13_{(\pm.039)}$	$\bullet1.84_{(\pm.070)}$	$\bullet1.04_{(\pm.010)}$
	RkNN	$\prec.863_{(\pm.046)}$	$\prec.828_{(\pm.005)}$	$\prec3.18_{(\pm.010)}$	$\sim2.60_{(\pm.032)}$	$\prec2.19_{(\pm.015)}$	$\prec2.38_{(\pm.052)}$	$\prec1.94_{(\pm.069)}$	$\prec1.05_{(\pm.011)}$
	RkNN†	$\bullet.777_{(\pm.031)}$	$\bullet.780_{(\pm.008)}$	$\bullet3.05_{(\pm.023)}$	$\bullet2.58_{(\pm.039)}$	$\bullet2.18_{(\pm.014)}$	$\bullet2.18_{(\pm.097)}$	$\bullet1.84_{(\pm.074)}$	$\bullet1.01_{(\pm.010)}$
KL Divergence (↓)	BAP	$\bullet.061_{(\pm.006)}$	$\bullet.057_{(\pm.001)}$	$\bullet.213_{(\pm.005)}$	$\bullet.365_{(\pm.015)}$	$\bullet.106_{(\pm.004)}$	$\bullet.538_{(\pm.020)}$	$\bullet.111_{(\pm.006)}$	$\bullet.098_{(\pm.002)}$
	KL	$\prec.114_{(\pm.014)}$	$\prec.070_{(\pm.003)}$	$\succ.203_{(\pm.004)}$	$\prec.565_{(\pm.040)}$	$\prec.117_{(\pm.003)}$	$\prec.563_{(\pm.017)}$	$\prec.175_{(\pm.012)}$	$\succ.098_{(\pm.002)}$
	SE	$\prec.072_{(\pm.008)}$	$\prec.070_{(\pm.003)}$	$\succ.210_{(\pm.004)}$	$\prec.853_{(\pm.166)}$	$\prec.123_{(\pm.003)}$	$\prec.571_{(\pm.015)}$	$\prec.119_{(\pm.006)}$	$\prec.099_{(\pm.002)}$
	SA	$\prec4.17_{(\pm1.03)}$	$\prec.622_{(\pm.041)}$	$\prec2.51_{(\pm.037)}$	$\prec4.66_{(\pm.266)}$	$\prec1.67_{(\pm.040)}$	$\prec4.34_{(\pm.197)}$	$\prec7.98_{(\pm.318)}$	$\prec2.45_{(\pm.109)}$
	KL-SA	$\prec3.87_{(\pm.835)}$	$\prec.083_{(\pm.004)}$	$\prec1.16_{(\pm.031)}$	$\prec1.51_{(\pm.100)}$	$\prec.696_{(\pm.018)}$	$\prec2.20_{(\pm.169)}$	$\prec3.83_{(\pm.276)}$	$\prec.426_{(\pm.014)}$
	SE-SA	$\prec3.80_{(\pm.802)}$	$\prec.120_{(\pm.007)}$	$\prec2.26_{(\pm.039)}$	$\prec2.59_{(\pm.150)}$	$\prec1.15_{(\pm.027)}$	$\prec3.66_{(\pm.344)}$	$\prec7.42_{(\pm.298)}$	$\prec.665_{(\pm.022)}$
	Delta	$\sim.055_{(\pm.005)}$	$\prec.064_{(\pm.003)}$	$\prec.544_{(\pm.040)}$	$\succ.375_{(\pm.016)}$	$\sim.114_{(\pm.004)}$	$\succ.507_{(\pm.023)}$	$\sim.109_{(\pm.008)}$	$\succ.102_{(\pm.002)}$
	Delta†	$\bullet.055_{(\pm.005)}$	$\bullet.063_{(\pm.002)}$	$\bullet.494_{(\pm.035)}$	$\bullet.395_{(\pm.025)}$	$\bullet.115_{(\pm.004)}$	$\bullet.663_{(\pm.103)}$	$\bullet.106_{(\pm.006)}$	$\bullet.106_{(\pm.002)}$
	DVS	$\sim.066_{(\pm.012)}$	$\sim.064_{(\pm.003)}$	$\prec.555_{(\pm.043)}$	$\sim.372_{(\pm.014)}$	$\succ.114_{(\pm.004)}$	$\succ.499_{(\pm.022)}$	$\sim.110_{(\pm.011)}$	$\prec.114_{(\pm.004)}$
	DVS†	$\bullet.062_{(\pm.008)}$	$\bullet.064_{(\pm.002)}$	$\bullet.538_{(\pm.033)}$	$\bullet.372_{(\pm.014)}$	$\bullet.120_{(\pm.003)}$	$\bullet.508_{(\pm.026)}$	$\bullet.108_{(\pm.007)}$	$\bullet.107_{(\pm.002)}$
	RkNN	$\prec.071_{(\pm.008)}$	$\prec.070_{(\pm.001)}$	$\prec.490_{(\pm.004)}$	$\succ.390_{(\pm.019)}$	$\prec.141_{(\pm.004)}$	$\sim.518_{(\pm.021)}$	$\prec.121_{(\pm.009)}$	$\prec.105_{(\pm.002)}$
	RkNN†	$\bullet.057_{(\pm.006)}$	$\bullet.062_{(\pm.001)}$	$\bullet.422_{(\pm.011)}$	$\bullet.397_{(\pm.019)}$	$\bullet.119_{(\pm.004)}$	$\bullet.527_{(\pm.041)}$	$\bullet.108_{(\pm.007)}$	$\bullet.099_{(\pm.002)}$
Intersection Similarity (↑)	BAP	$\bullet.864_{(\pm.007)}$	$\bullet.867_{(\pm.002)}$	$\bullet.802_{(\pm.002)}$	$\bullet.700_{(\pm.006)}$	$\bullet.850_{(\pm.003)}$	$\bullet.632_{(\pm.009)}$	$\bullet.815_{(\pm.006)}$	$\bullet.836_{(\pm.002)}$
	KL	$\prec.814_{(\pm.011)}$	$\prec.853_{(\pm.003)}$	$\succ.804_{(\pm.002)}$	$\prec.675_{(\pm.005)}$	$\prec.843_{(\pm.002)}$	$\prec.615_{(\pm.007)}$	$\prec.769_{(\pm.009)}$	$\prec.836_{(\pm.002)}$
	SE	$\prec.850_{(\pm.007)}$	$\prec.854_{(\pm.003)}$	$\prec.800_{(\pm.002)}$	$\prec.683_{(\pm.006)}$	$\prec.840_{(\pm.002)}$	$\prec.608_{(\pm.008)}$	$\prec.808_{(\pm.006)}$	$\prec.835_{(\pm.002)}$
	SA	$\prec.332_{(\pm.030)}$	$\prec.602_{(\pm.010)}$	$\prec.571_{(\pm.003)}$	$\prec.478_{(\pm.013)}$	$\prec.617_{(\pm.006)}$	$\prec.452_{(\pm.014)}$	$\prec.201_{(\pm.010)}$	$\prec.506_{(\pm.005)}$
	KL-SA	$\prec.339_{(\pm.028)}$	$\prec.841_{(\pm.004)}$	$\prec.654_{(\pm.003)}$	$\prec.647_{(\pm.008)}$	$\prec.731_{(\pm.003)}$	$\prec.548_{(\pm.015)}$	$\prec.317_{(\pm.018)}$	$\prec.741_{(\pm.004)}$
	SE-SA	$\prec.342_{(\pm.027)}$	$\prec.811_{(\pm.005)}$	$\prec.585_{(\pm.003)}$	$\prec.651_{(\pm.007)}$	$\prec.694_{(\pm.005)}$	$\prec.474_{(\pm.018)}$	$\prec.212_{(\pm.007)}$	$\prec.707_{(\pm.003)}$
	Delta	$\sim.870_{(\pm.006)}$	$\prec.858_{(\pm.003)}$	$\prec.577_{(\pm.020)}$	$\prec.699_{(\pm.006)}$	$\sim.844_{(\pm.003)}$	$\prec.622_{(\pm.009)}$	$\sim.817_{(\pm.008)}$	$\sim.831_{(\pm.002)}$
	Delta†	$\bullet.872_{(\pm.007)}$	$\bullet.860_{(\pm.002)}$	$\bullet.604_{(\pm.018)}$	$\bullet.702_{(\pm.007)}$	$\bullet.844_{(\pm.003)}$	$\bullet.636_{(\pm.013)}$	$\bullet.819_{(\pm.007)}$	$\bullet.830_{(\pm.002)}$
	DVS	$\sim.858_{(\pm.012)}$	$\sim.858_{(\pm.003)}$	$\prec.571_{(\pm.021)}$	$\sim.694_{(\pm.006)}$	$\succ.844_{(\pm.003)}$	$\prec.617_{(\pm.009)}$	$\sim.816_{(\pm.011)}$	$\prec.823_{(\pm.003)}$
	DVS†	$\bullet.864_{(\pm.006)}$	$\bullet.859_{(\pm.003)}$	$\bullet.578_{(\pm.017)}$	$\bullet.695_{(\pm.006)}$	$\bullet.839_{(\pm.003)}$	$\bullet.623_{(\pm.010)}$	$\bullet.817_{(\pm.007)}$	$\bullet.826_{(\pm.001)}$
	RkNN	$\prec.852_{(\pm.009)}$	$\prec.851_{(\pm.001)}$	$\prec.606_{(\pm.002)}$	$\prec.682_{(\pm.007)}$	$\prec.825_{(\pm.002)}$	$\sim.611_{(\pm.009)}$	$\prec.806_{(\pm.007)}$	$\prec.828_{(\pm.002)}$
	RkNN†	$\bullet.867_{(\pm.006)}$	$\bullet.860_{(\pm.002)}$	$\bullet.649_{(\pm.005)}$	$\bullet.696_{(\pm.005)}$	$\bullet.839_{(\pm.002)}$	$\bullet.618_{(\pm.018)}$	$\bullet.817_{(\pm.008)}$	$\bullet.834_{(\pm.002)}$

group. For five-level rating datasets (SCUT-FBP, FBP5500, M2B, Movie), scores 5 and 4 receive positive weights of 1 and 0.5, respectively, while scores 1 and 2 receive negative weights of 1 and 0.5, with score 3 treated as neutral. For Music dataset, clam, cheerful, danceable, love, and dreamy are assigned to the positive group with weight of 1, whereas tense and sad are assigned to the negative group with weight of 1. These semantic weight vectors are fixed across all experimental procedures and random seeds. The trade-off coefficient α, which balances the distribution-fitting term and

Table 4: Clark Distance, Canberra Metric, KL Divergence, and Intersection Similarity under Finite-Sample Label Distributions at $n_a = 10$.

	Method	Jaffe	BU-3DFE	RAF-ML	SCUT-FBP	FBP5500	M2B	Music	Movie
Clark Distance (↓)	BAP	$\bullet.484_{(\pm.015)}$	$\bullet.372_{(\pm.005)}$	$\bullet1.46_{(\pm.004)}$	$\bullet1.38_{(\pm.013)}$	$\bullet1.27_{(\pm.006)}$	$\bullet1.11_{(\pm.018)}$	$\bullet.762_{(\pm.024)}$	$\bullet.520_{(\pm.005)}$
	KL	$\prec.808_{(\pm.026)}$	$\prec.473_{(\pm.011)}$	$\succ1.43_{(\pm.003)}$	$\prec1.42_{(\pm.012)}$	$\prec1.32_{(\pm.006)}$	$\prec1.15_{(\pm.014)}$	$\prec1.02_{(\pm.026)}$	$\prec.525_{(\pm.004)}$
	SE	$\prec.560_{(\pm.016)}$	$\prec.474_{(\pm.010)}$	$\succ1.45_{(\pm.005)}$	$\sim1.35_{(\pm.052)}$	$\prec1.34_{(\pm.006)}$	$\prec1.38_{(\pm.110)}$	$\prec.809_{(\pm.024)}$	$\prec.527_{(\pm.005)}$
	SA	$\prec2.25_{(\pm.015)}$	$\prec2.02_{(\pm.013)}$	$\prec2.10_{(\pm.005)}$	$\prec1.93_{(\pm.011)}$	$\prec1.93_{(\pm.002)}$	$\prec1.99_{(\pm.007)}$	$\prec2.84_{(\pm.014)}$	$\prec1.86_{(\pm.010)}$
	KL-SA	$\prec2.25_{(\pm.016)}$	$\prec.856_{(\pm.025)}$	$\prec1.93_{(\pm.033)}$	$\prec1.63_{(\pm.020)}$	$\prec1.75_{(\pm.005)}$	$\prec1.85_{(\pm.009)}$	$\prec2.79_{(\pm.037)}$	$\prec1.24_{(\pm.009)}$
	SE-SA	$\prec2.25_{(\pm.015)}$	$\prec1.70_{(\pm.027)}$	$\prec2.09_{(\pm.004)}$	$\prec1.73_{(\pm.007)}$	$\prec1.90_{(\pm.003)}$	$\prec1.98_{(\pm.007)}$	$\prec2.83_{(\pm.014)}$	$\prec1.61_{(\pm.014)}$
	Delta	$\sim.446_{(\pm.044)}$	$\prec.377_{(\pm.008)}$	$\prec1.54_{(\pm.016)}$	$\succ1.38_{(\pm.013)}$	$\prec1.28_{(\pm.007)}$	$\prec1.13_{(\pm.025)}$	$\prec.740_{(\pm.023)}$	$\sim.544_{(\pm.006)}$
	Delta†	$\bullet.438_{(\pm.035)}$	$\bullet.374_{(\pm.006)}$	$\bullet1.51_{(\pm.015)}$	$\bullet1.39_{(\pm.013)}$	$\bullet1.27_{(\pm.006)}$	$\bullet1.01_{(\pm.035)}$	$\bullet.725_{(\pm.024)}$	$\bullet.541_{(\pm.006)}$
	DVS	$\prec.519_{(\pm.076)}$	$\sim.378_{(\pm.009)}$	$\prec1.54_{(\pm.017)}$	$\prec1.38_{(\pm.012)}$	$\prec1.27_{(\pm.007)}$	$\prec1.22_{(\pm.013)}$	$\sim.753_{(\pm.042)}$	$\prec.573_{(\pm.012)}$
	DVS†	$\bullet.497_{(\pm.052)}$	$\bullet.376_{(\pm.007)}$	$\bullet1.53_{(\pm.013)}$	$\bullet1.38_{(\pm.012)}$	$\bullet1.27_{(\pm.005)}$	$\bullet1.12_{(\pm.023)}$	$\bullet.739_{(\pm.025)}$	$\bullet.561_{(\pm.006)}$
	RkNN	$\prec.429_{(\pm.022)}$	$\prec.401_{(\pm.005)}$	$\prec1.52_{(\pm.004)}$	$\sim1.39_{(\pm.013)}$	$\sim1.27_{(\pm.006)}$	$\prec1.22_{(\pm.017)}$	$\prec.776_{(\pm.027)}$	$\prec.555_{(\pm.005)}$
	RkNN†	$\bullet.408_{(\pm.015)}$	$\bullet.375_{(\pm.004)}$	$\bullet1.50_{(\pm.010)}$	$\bullet1.39_{(\pm.012)}$	$\bullet1.28_{(\pm.006)}$	$\bullet1.15_{(\pm.036)}$	$\bullet.735_{(\pm.026)}$	$\bullet.532_{(\pm.005)}$
Canberra Metric (↓)	BAP	$\bullet.999_{(\pm.032)}$	$\bullet.777_{(\pm.010)}$	$\bullet2.89_{(\pm.016)}$	$\bullet2.56_{(\pm.032)}$	$\bullet2.16_{(\pm.014)}$	$\bullet2.12_{(\pm.040)}$	$\bullet1.93_{(\pm.064)}$	$\bullet.996_{(\pm.010)}$
	KL	$\prec1.69_{(\pm.059)}$	$\prec.982_{(\pm.025)}$	$\succ2.79_{(\pm.014)}$	$\prec2.69_{(\pm.031)}$	$\prec2.31_{(\pm.014)}$	$\prec2.21_{(\pm.033)}$	$\prec2.61_{(\pm.081)}$	$\prec1.00_{(\pm.009)}$
	SE	$\prec1.16_{(\pm.032)}$	$\prec.982_{(\pm.022)}$	$\succ2.85_{(\pm.016)}$	$\sim2.55_{(\pm.118)}$	$\prec2.36_{(\pm.013)}$	$\prec2.75_{(\pm.266)}$	$\prec2.05_{(\pm.066)}$	$\prec1.01_{(\pm.010)}$
	SA	$\prec5.40_{(\pm.045)}$	$\prec4.71_{(\pm.038)}$	$\prec4.84_{(\pm.015)}$	$\prec4.09_{(\pm.032)}$	$\prec3.85_{(\pm.011)}$	$\prec4.23_{(\pm.024)}$	$\prec8.40_{(\pm.056)}$	$\prec3.93_{(\pm.029)}$
	KL-SA	$\prec5.39_{(\pm.051)}$	$\prec1.80_{(\pm.053)}$	$\prec4.30_{(\pm.100)}$	$\prec3.21_{(\pm.042)}$	$\prec3.22_{(\pm.017)}$	$\prec3.82_{(\pm.028)}$	$\prec8.18_{(\pm.138)}$	$\prec2.41_{(\pm.019)}$
	SE-SA	$\prec5.39_{(\pm.046)}$	$\prec3.83_{(\pm.073)}$	$\prec4.78_{(\pm.013)}$	$\prec3.41_{(\pm.027)}$	$\prec3.74_{(\pm.013)}$	$\prec4.18_{(\pm.024)}$	$\prec8.37_{(\pm.059)}$	$\prec3.18_{(\pm.038)}$
	Delta	$\sim.916_{(\pm.091)}$	$\prec.799_{(\pm.018)}$	$\prec3.26_{(\pm.064)}$	$\succ2.57_{(\pm.033)}$	$\prec2.18_{(\pm.016)}$	$\prec2.15_{(\pm.058)}$	$\prec1.87_{(\pm.064)}$	$\sim1.04_{(\pm.011)}$
	Delta†	$\bullet.897_{(\pm.072)}$	$\bullet.789_{(\pm.013)}$	$\bullet3.18_{(\pm.055)}$	$\bullet2.59_{(\pm.032)}$	$\bullet2.16_{(\pm.014)}$	$\bullet1.89_{(\pm.058)}$	$\bullet1.83_{(\pm.064)}$	$\bullet1.04_{(\pm.010)}$
	DVS	$\prec1.07_{(\pm.164)}$	$\sim.800_{(\pm.020)}$	$\sim3.28_{(\pm.065)}$	$\prec2.57_{(\pm.031)}$	$\prec2.18_{(\pm.016)}$	$\prec2.35_{(\pm.032)}$	$\sim1.91_{(\pm.119)}$	$\prec1.10_{(\pm.024)}$
	DVS†	$\bullet1.02_{(\pm.109)}$	$\bullet.796_{(\pm.014)}$	$\bullet3.26_{(\pm.052)}$	$\bullet2.57_{(\pm.031)}$	$\bullet2.16_{(\pm.012)}$	$\bullet2.10_{(\pm.051)}$	$\bullet1.87_{(\pm.070)}$	$\bullet1.07_{(\pm.011)}$
	RkNN	$\prec.891_{(\pm.046)}$	$\prec.854_{(\pm.010)}$	$\prec3.18_{(\pm.011)}$	$\succ2.60_{(\pm.033)}$	$\prec2.19_{(\pm.014)}$	$\prec2.38_{(\pm.044)}$	$\prec1.96_{(\pm.062)}$	$\prec1.05_{(\pm.009)}$
	RkNN†	$\bullet.844_{(\pm.028)}$	$\bullet.795_{(\pm.008)}$	$\bullet3.11_{(\pm.032)}$	$\bullet2.63_{(\pm.027)}$	$\bullet2.18_{(\pm.014)}$	$\bullet2.18_{(\pm.096)}$	$\bullet1.86_{(\pm.071)}$	$\bullet1.01_{(\pm.010)}$
KL Divergence (↓)	BAP	$\bullet.092_{(\pm.008)}$	$\bullet.061_{(\pm.002)}$	$\bullet.227_{(\pm.005)}$	$\bullet.366_{(\pm.015)}$	$\bullet.110_{(\pm.004)}$	$\bullet.560_{(\pm.021)}$	$\bullet.123_{(\pm.007)}$	$\bullet.100_{(\pm.002)}$
	KL	$\prec.257_{(\pm.020)}$	$\prec.094_{(\pm.005)}$	$\succ.213_{(\pm.004)}$	$\prec.592_{(\pm.052)}$	$\prec.137_{(\pm.004)}$	$\prec.573_{(\pm.017)}$	$\prec.262_{(\pm.018)}$	$\prec.100_{(\pm.002)}$
	SE	$\prec.123_{(\pm.011)}$	$\prec.094_{(\pm.005)}$	$\sim.228_{(\pm.006)}$	$\prec.945_{(\pm.162)}$	$\prec.146_{(\pm.004)}$	$\prec.583_{(\pm.017)}$	$\prec.144_{(\pm.009)}$	$\sim.100_{(\pm.002)}$
	SA	$\prec11.3_{(\pm.472)}$	$\prec3.92_{(\pm.112)}$	$\prec3.00_{(\pm.037)}$	$\prec5.38_{(\pm.196)}$	$\prec2.51_{(\pm.053)}$	$\prec5.06_{(\pm.182)}$	$\prec11.7_{(\pm.328)}$	$\prec5.06_{(\pm.079)}$
	KL-SA	$\prec11.0_{(\pm.555)}$	$\prec.292_{(\pm.021)}$	$\prec1.52_{(\pm.217)}$	$\prec1.77_{(\pm.104)}$	$\prec.955_{(\pm.028)}$	$\prec2.63_{(\pm.180)}$	$\prec9.93_{(\pm1.30)}$	$\prec1.18_{(\pm.024)}$
	SE-SA	$\prec11.0_{(\pm.476)}$	$\prec1.70_{(\pm.108)}$	$\prec2.75_{(\pm.036)}$	$\prec3.10_{(\pm.155)}$	$\prec2.04_{(\pm.038)}$	$\prec4.62_{(\pm.152)}$	$\prec11.5_{(\pm.353)}$	$\prec2.20_{(\pm.050)}$
	Delta	$\sim.079_{(\pm.014)}$	$\prec.065_{(\pm.003)}$	$\prec.543_{(\pm.041)}$	$\succ.376_{(\pm.016)}$	$\succ.114_{(\pm.004)}$	$\succ.509_{(\pm.028)}$	$\prec.113_{(\pm.007)}$	$\succ.109_{(\pm.002)}$
	Delta†	$\bullet.076_{(\pm.011)}$	$\bullet.064_{(\pm.002)}$	$\bullet.489_{(\pm.033)}$	$\bullet.398_{(\pm.023)}$	$\bullet.118_{(\pm.003)}$	$\bullet.683_{(\pm.081)}$	$\bullet.108_{(\pm.006)}$	$\bullet.113_{(\pm.002)}$
	DVS	$\prec.108_{(\pm.030)}$	$\sim.065_{(\pm.003)}$	$\prec.555_{(\pm.044)}$	$\prec.372_{(\pm.014)}$	$\succ.115_{(\pm.004)}$	$\succ.500_{(\pm.022)}$	$\sim.119_{(\pm.021)}$	$\prec.128_{(\pm.008)}$
	DVS†	$\bullet.099_{(\pm.020)}$	$\bullet.065_{(\pm.002)}$	$\bullet.538_{(\pm.033)}$	$\bullet.372_{(\pm.014)}$	$\bullet.120_{(\pm.003)}$	$\bullet.510_{(\pm.026)}$	$\bullet.113_{(\pm.008)}$	$\bullet.115_{(\pm.003)}$
	RkNN	$\prec.075_{(\pm.008)}$	$\prec.074_{(\pm.001)}$	$\prec.493_{(\pm.005)}$	$\succ.393_{(\pm.022)}$	$\prec.142_{(\pm.004)}$	$\sim.518_{(\pm.023)}$	$\prec.124_{(\pm.009)}$	$\prec.106_{(\pm.002)}$
	RkNN†	$\bullet.066_{(\pm.004)}$	$\bullet.065_{(\pm.001)}$	$\bullet.467_{(\pm.022)}$	$\bullet.433_{(\pm.019)}$	$\bullet.119_{(\pm.004)}$	$\bullet.537_{(\pm.038)}$	$\bullet.110_{(\pm.007)}$	$\bullet.102_{(\pm.002)}$
Intersection Similarity (↑)	BAP	$\bullet.831_{(\pm.006)}$	$\bullet.862_{(\pm.002)}$	$\bullet.793_{(\pm.003)}$	$\bullet.700_{(\pm.006)}$	$\bullet.847_{(\pm.002)}$	$\bullet.628_{(\pm.010)}$	$\bullet.805_{(\pm.006)}$	$\bullet.835_{(\pm.002)}$
	KL	$\prec.726_{(\pm.010)}$	$\prec.830_{(\pm.004)}$	$\succ.797_{(\pm.003)}$	$\prec.671_{(\pm.005)}$	$\prec.829_{(\pm.002)}$	$\prec.612_{(\pm.008)}$	$\prec.721_{(\pm.011)}$	$\prec.834_{(\pm.002)}$
	SE	$\prec.806_{(\pm.006)}$	$\prec.831_{(\pm.004)}$	$\prec.790_{(\pm.003)}$	$\prec.681_{(\pm.005)}$	$\prec.825_{(\pm.002)}$	$\prec.605_{(\pm.008)}$	$\prec.788_{(\pm.007)}$	$\prec.833_{(\pm.002)}$
	SA	$\prec.222_{(\pm.011)}$	$\prec.332_{(\pm.006)}$	$\prec.559_{(\pm.003)}$	$\prec.464_{(\pm.012)}$	$\prec.603_{(\pm.005)}$	$\prec.449_{(\pm.015)}$	$\prec.158_{(\pm.008)}$	$\prec.374_{(\pm.006)}$
	KL-SA	$\prec.223_{(\pm.011)}$	$\prec.713_{(\pm.008)}$	$\prec.627_{(\pm.014)}$	$\prec.636_{(\pm.008)}$	$\prec.708_{(\pm.004)}$	$\prec.528_{(\pm.011)}$	$\prec.178_{(\pm.013)}$	$\prec.637_{(\pm.004)}$
	SE-SA	$\prec.223_{(\pm.009)}$	$\prec.453_{(\pm.008)}$	$\prec.571_{(\pm.002)}$	$\prec.642_{(\pm.007)}$	$\prec.634_{(\pm.004)}$	$\prec.463_{(\pm.012)}$	$\prec.161_{(\pm.008)}$	$\prec.558_{(\pm.007)}$
	Delta	$\sim.845_{(\pm.014)}$	$\prec.857_{(\pm.003)}$	$\prec.578_{(\pm.021)}$	$\prec.699_{(\pm.006)}$	$\succ.844_{(\pm.003)}$	$\prec.621_{(\pm.011)}$	$\prec.813_{(\pm.007)}$	$\sim.825_{(\pm.002)}$
	Delta†	$\bullet.848_{(\pm.011)}$	$\bullet.858_{(\pm.002)}$	$\bullet.608_{(\pm.017)}$	$\bullet.702_{(\pm.007)}$	$\bullet.840_{(\pm.002)}$	$\bullet.638_{(\pm.014)}$	$\bullet.818_{(\pm.006)}$	$\bullet.825_{(\pm.002)}$
	DVS	$\prec.822_{(\pm.024)}$	$\sim.856_{(\pm.004)}$	$\sim.571_{(\pm.021)}$	$\prec.694_{(\pm.005)}$	$\succ.844_{(\pm.003)}$	$\prec.616_{(\pm.009)}$	$\sim.809_{(\pm.016)}$	$\prec.811_{(\pm.005)}$
	DVS†	$\bullet.828_{(\pm.015)}$	$\bullet.857_{(\pm.003)}$	$\bullet.578_{(\pm.017)}$	$\bullet.695_{(\pm.005)}$	$\bullet.838_{(\pm.003)}$	$\bullet.623_{(\pm.010)}$	$\bullet.813_{(\pm.008)}$	$\bullet.819_{(\pm.003)}$
	RkNN	$\prec.847_{(\pm.009)}$	$\prec.847_{(\pm.002)}$	$\prec.605_{(\pm.003)}$	$\sim.682_{(\pm.007)}$	$\prec.825_{(\pm.002)}$	$\sim.611_{(\pm.009)}$	$\prec.803_{(\pm.006)}$	$\prec.826_{(\pm.001)}$
	RkNN†	$\bullet.855_{(\pm.005)}$	$\bullet.857_{(\pm.002)}$	$\bullet.629_{(\pm.009)}$	$\bullet.682_{(\pm.006)}$	$\bullet.839_{(\pm.002)}$	$\bullet.623_{(\pm.020)}$	$\bullet.815_{(\pm.007)}$	$\bullet.832_{(\pm.002)}$

the divisiveness regularization term, is selected from $\{0.01, 0.1\}$ independently using an internal validation sets.

B.5 Further Discussion on Experimental Results

Tables 1 to 4 provide consistent empirical support for the central claim that BAP enhances sensitivity to relative label proportions while preserving the arithmetic mean as its Bayes target. Across

conventional LDL metrics, BAP generally yields lower distributional errors and higher similarity scores than KL and SE loss under both annotation counts. Specifically, Tables 1 and 2 demonstrate this pattern for Chebyshev distance and cosine coefficient, while Tables 3 and 4 corroborate it using Clark distance, Canberra metric, KL divergence, and intersection similarity. The consistency of these gains across different metrics shows that lower errors of label ratio prediction can accompany improved the distributional prediction performance. The clearest performance difference is provided by MALR, which directly quantifies discrepancies in pairwise label-mass ratios. At $n_a = 30$, BAP achieves significantly lower MALR than KL loss on all eight datasets; for example, MALR decreases from 0.593 to 0.419 on JAFFE, from 0.854 to 0.754 on FBP5500, and from 0.778 to 0.639 on Music. This statistically significant advantage over KL is retained on all eight datasets at $n_a = 10$, while BAP also significantly outperforms SE loss on seven of the eight datasets at each annotation count. Moreover, the advantage becomes more pronounced as supervision becomes more limited. On JAFFE, for instance, the relative reduction in MALR over KL increases from approximately 29% at $n_a = 30$ to 43% at $n_a = 10$, while on Music it increases from approximately 18% to 28%. This trend is consistent with the motivation underlying BAP: with finite annotations, low-mass labels are subject to greater sampling variability, and even modest absolute deviations can induce substantial distortions in relative proportions. The improvements in MALR therefore provide direct empirical evidence that the local curvature induced by BAP more effectively controls such ratio distortions. The comparison with SA loss further shows that ratio sensitivity alone is insufficient. Although SA directly penalizes log-ratio discrepancies, it performs substantially worse than BAP on both conventional distribution metrics and MALR. On JAFFE, for example, its MALR is 5.14 at $n_a = 30$ and increases to 17.4 at $n_a = 10$, compared with 0.419 and 0.527 for BAP, respectively. This behavior is consistent with the theoretical distinction established earlier: SA loss generally elicits an Aitchison center rather than the conditional arithmetic mean, whereas the finite-sample prediction problem considered here requires recovery of the latter. Consequently, directly optimizing log-ratio geometry does not necessarily yield an accurate estimate of the population label distribution. BAP avoids this target mismatch by embedding Aitchison-inspired local curvature within a mean-proper Bregman loss, thereby combining sensitivity to relative proportions with the appropriate statistical target. Results for BAP-based variants of existing LDL algorithms further suggest that these gains arise from the loss function rather than from a particular learning algorithm. Replacing the original fitting objectives of Delta, DVS, and RkNN with BAP frequently improves performance, with the most systematic gains again observed in MALR. Across the three methods, eight datasets, and two annotation counts, the BAP-based variants yield either statistically significant improvements or statistically indistinguishable MALR values, with no significant degradation. For example, at $n_a = 30$, RkNN† significantly reduces MALR relative to RkNN on all eight datasets, including reductions from 0.436 to 0.399 on JAFFE, from 5.46 to 5.30 on RAF-ML, and from 0.679 to 0.633 on Music. Similar improvements persist at $n_a = 10$. These results support BAP as a broadly applicable distribution-fitting objective rather than a method-specific component. The few exceptions in Tables 1 to 4 further delineate the scope of this empirical claim. On RAF-ML, for example, KL slightly outperforms BAP on several conventional metrics, including Chebyshev distance and cosine coefficient at $n_a = 30$, while BAP still achieves significantly lower MALR on the same dataset. The results support that BAP achieves a favorable trade-off between probability-space fidelity and preservation of relative label proportions. Overall, the empirical evidence is consistent with its theoretical design objective: BAP preserves accurate recovery of the arithmetic-mean population distribution while providing greater sensitivity to distortions in relative label mass, with particularly pronounced benefits under limited finite-sample supervision.