# Online Supplementary Material

No Author Given

No Institute Given

## 1 Lebesgue measure estimation using Monte Carlo sampling

The Lebesgue measure $\lambda(H(F,R))$ integrates the area covered by a set of loss function vectors in a multi-dimensional objective space. This measure is comprised of three sets: $F$, $R$, and $Z$. $F$ denotes the set of representations of functions (which map the input data to a vector of loss function values). $R$ denotes the set of mutually non-dominating loss vectors. Initially, $R$ is set to the unit loss vector $\{1\}^3$, which denotes the worst possible performance for Hamming-loss, one minus the Micro-$F_1$, and one minus the label ranking average precision. Last, $Z$ denotes the set containing all possible loss function vectors in the applicable multi-dimensional loss objective space.

The Lebesgue contribution $\lambda(P(f))$ of a function $f$ measures the new marginal improvement of a function's loss vector over a set of previous loss vectors. In this paper, we use the Lebesgue contribution to quantify candidate functions found by CLML during the optimisation process. However, to efficiently calculate the Lebesgue contribution (especially when the set of functions $F$ and $R$ are sparsely populated during the early stages of the optimisation), we estimate the Lebesgue measure using Monte Carlo sampling. First, a sampling space $S \subseteq Z$ is defined that entirely contains $P(f)$, *i.e.,* $P(f) \subseteq S \subseteq Z$. The sampling space can be problem-specific, however, in this paper, it is defined to contain all possible loss vectors between $\{0\}^3$ and $\{1\}^3$. Following, $g$ samples are drawn from $s_i \in S$ randomly and with uniform probability. Given $\{s_1, \cdots, s_g\}$, the Lebesgue contribution is estimated via $\hat{\lambda}(P(f))$ via the following:

$$\hat{\lambda}(P(f)) = \lambda(S(f)) = \frac{|\{s_i | s_i \in P(f)\}|}{g} \tag{1}$$

where $|\{s_i|s_i \in P(f)\}|$ is denoted as the number of randomly sampled solutions that exist in $P(f)$, also known as *hits*. The probability $p$ of a sample being *hit* is i.i.d. Bernoulli distributed, therefore, $\hat{\lambda}(P(f))$ converges to $\lambda(P(f))$ with $\frac{1}{\sqrt{pg}}$ [3].

## 2 Proof of Theorem 4.5: A Consistent Lebesgue

$$\begin{aligned} \lim_{n\to\infty} \lambda(H(F^{(n)}, R)) \to \lambda(H(\mathbb{P}^B, R)) \quad \text{then} \\ R_{\mathcal{L}_1}(f^{(n)}) \to R^B_{\mathcal{L}_1}(f) \wedge \\ R_{\mathcal{L}_2}(f^{'(n)}) \to R^B_{\mathcal{L}_2}(f') \wedge \\ R_{\mathcal{L}_3}(f^{''(n)}) \to R^B_{\mathcal{L}_3}(f''). \end{aligned} \tag{2}$$

In other words, the maximisation of $\lambda(H(F^{(n)}, R))$ tends to the convergence toward the Bayes risk for each loss function $\mathcal{L}_i \; \forall i : 1 \leq i \leq 3$, $f^{(n)}, f^{'(n)}, f^{''(n)} \in F^{(n)}$ and that $f, f', f'' \in \mathbb{P}^B$.

*Proof.* We will use contradiction to prove that maximising $\lambda(H(F^{(n)}, R))$ ensures convergence to the Bayes predictors.

Assumption: suppose there exists a sequence $F^{(n)}$ such that maximising $\lambda(H(F^{(n)}, R)))$ does *not* converge to the Bayes predictors for $L_1$, $L_2$, or $L_3$. Specifically, assume there exists a function $f_\gamma \in F^{(n)}$ such that:

$$f_\gamma \notin \mathbb{P}^B \wedge R^B_{L_v}(f_\gamma) \quad \exists v \in \{1, 2, 3\}. \tag{3}$$

- Properties of $f_\gamma \notin \mathbb{P}^B$:
  If $f_\gamma \notin \mathbb{P}^B$, then by the definition of Pareto optimality, there exists another function $f_\beta \in \mathbb{P}^B$ such that:

$$\forall i : L_i(f_\beta) \leq L_i(f_\gamma) \wedge \exists k : L_k(f_\beta) < L_k(f_\gamma). \tag{4}$$

- Contradiction for the Bayes Predictor:
  If $f_\gamma$ is a Bayes predictor for $L_v$, then:

$$R_{L_v}(f_\gamma) = R^B_{L_v}(f_\gamma) \implies L_v(f_\gamma) \leq L_v(f_\beta). \tag{5}$$

  However, this contradicts $f_\beta$ strictly dominating $f_\gamma$ on $L_k$ (where $k \neq v$ or $k = v$ with strict inequality). Therefore, $f_\gamma \notin \mathbb{P}^B$ cannot be a Bayes predictor.
- Convergence of $F^{(n)}$ to $\mathbb{P}^B$:
  Maximising $\lambda(H(F^{(n)}, R))$ ensures that non-dominated solutions increasingly dominate the objective space $Z$ as $n \to \inf$. This implies:

$$\lambda(H(F^{(n)}, R)) \to \lambda(H(\mathbb{P}^B, R)) \quad as \quad n \to \inf. \tag{6}$$

  Since $\mathbb{P}^B$ contains only Pareto optimal functions, and every Bayes predictor belongs to $\mathbb{P}^B$, this convergence guarantees that the sequence $F^{(n)}$ minimises $L_1$, $L_2$, and $L_3$ asymptotically.

By contradiction, the assumption that $\lambda(H(F^{(n)}, R))$ does *not* converge to the Bayes predictors is false. Therefore:

$$\begin{aligned} \lim_{n\to\infty} \lambda(H(F^{(n)}, R)) \to \lambda(H(\mathbb{P}^B, R)) \quad \implies \\ R_{\mathcal{L}_1}(f^{(n)}) \to R^B_{\mathcal{L}_1}(f) \wedge \\ R_{\mathcal{L}_2}(f^{'(n)}) \to R^B_{\mathcal{L}_2}(f') \wedge \\ R_{\mathcal{L}_3}(f^{''(n)}) \to R^B_{\mathcal{L}_3}(f''). \end{aligned} \tag{7}$$

## 3 Experimental Protocol

We conduct the experiments on nine widely-used multi-label datasets. Several datasets such as tmc2007-500, enron, and IMDB-F originate from the text domain, although they have been processed into a tabular format. Flags and mediamill respectively similarly originate from image and video domains. The tabulated data in this paper encompasses a wide variety of domains, and CLML is applied to varying modalities that have been vectorised/flattened using pre-trained models or pre-specified methods. $K^\mu$ (the cardinality) of an instance measures the average number of associated class labels; $DK/K^\mu$, the theoretical maximum complexity of an instance, (*i.e.*, the instance-level average dispersion of feature to label interactions); and $DK^\mu$, the average feature to label interactions of an instance. There are two important cases to consider. First, if dispersion is less than the average interaction, *i.e.*, $DK/K^\mu < DK^\mu$, then the dataset contains high concentrations of rich instance-level feature-to-label interactions that are not apparent when examining the dataset as a whole. This can indicate that there are clusters of instances that share similar feature-to-label interactions, and therefore a less diverse dispersion of the possible feature-to-label interactions. Second, if dispersion is higher than the average interaction, *i.e.*, $DK/K^\mu > DK^\mu$, the dataset as a whole has a greater expression of feature-to-label interactions than a given individual instance. Put differently, the dataset's instances each contain a subset of the total dataset interactions. The latter case is particularly challenging as it indicates a high number and variability of potential patterns and interactions between features and labels. The first case occurs in both Flags and Yeast and the second case occurs in the remaining datasets.

For each dataset, 30% are partitioned to the test set [4]. The remaining 70% is further split such that 20% is used as a validation set, and the remaining is used for training. We apply normalisation to all numerical features before training.

## 4 Ablation Study

We trial the embedding dimension $C$ at eight separate values. It is important to note that the latent space does not need to express spatial relationships of tabulated data, hence the embedding dimension can be quite small (in contrast to computer vision in works such as [2]). In addition to $\mathcal{L}_1, \mathcal{L}_2$, and $\mathcal{L}_3$, we set $\mathcal{L}_4$ as the averaged binary cross-entropy loss and track its progress during optimisation. For each experiment, we set $\mathcal{O} = 750$ (the maximum number of epochs). Here, we present the results for each of the embedding dimensions.

Figures 1 and 2 plot the Lebesgue measure of the sequence of functions obtained by CLML as $n \to \mathcal{O}$ (*i.e.*, the archive of non-dominated solutions obtained by CLML in $\mathcal{O}$ epoch). Smaller embedding dimensions (*i.e.*, $C \leq 80$) result in the best validation scores of $\lambda(H(F, R))$. To exemplify this, we tabulate the incumbent solution of the function sequence in terms of its $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$ scores on the validation set, against $\mathcal{L}_4$ according to each (non-normalised) value of $C$ in Table 1 and 2. When $C = 20$, we observe the lowest $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$ validation loss scores on the emotions dataset, and the lowest $\lambda(H(F, R))$ score on the CAL500 dataset. This observation indicates that CLML converges toward a better approximation of the Bayes predictors of $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$ on the emotions dataset, while on CAL500, CLML finds functions with more desirable trade-offs between the variant loss functions, hence the higher Lebesgue measure. These values motivate our recommendation to set the number of embedding dimensions to $C = 20$.

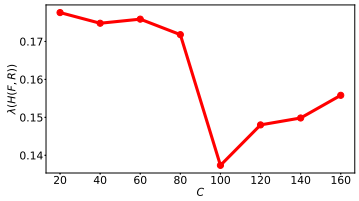



Fig. 1: The best Lebesgue measure obtained on CAL500 at each embedding dimension of the sequence of function sets $\lim_{n \to \mathcal{O}} \lambda(H(F^{(n)}, R))$.

Table 1: Best validation loss values of the incumbent solution for each embedding dimension on CAL500.

| $C$ | $\mathcal{L}_1$ | $\mathcal{L}_2$ | $\mathcal{L}_3$ | $\mathcal{L}_4$ |
|---|---|---|---|---|
| 20.0 | 0.169 | 0.523 | **0.509** | **138.068** |
| 40.0 | 0.171 | **0.522** | 0.518 | 143.809 |
| 60.0 | 0.169 | 0.523 | 0.520 | 144.201 |
| 80.0 | **0.161** | 0.527 | 0.525 | 149.604 |
| 100.0 | 0.196 | 0.529 | 0.534 | 157.877 |
| 120.0 | 0.171 | 0.529 | 0.539 | 155.555 |
| 140.0 | 0.167 | 0.528 | 0.534 | 151.250 |
| 160.0 | 0.168 | 0.526 | 0.533 | 153.533 |

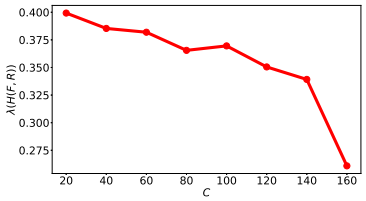



Fig. 2: The best Lebesgue measure obtained on emotions at each embedding dimension of the sequence of function sets $\lim_{n \to \mathcal{O}} \lambda(H(F^{(n)}, R))$.

Table 2: Best validation loss values of the incumbent solution for each embedding dimension on emotions.

| $C$ | $\mathcal{L}_1$ | $\mathcal{L}_2$ | $\mathcal{L}_3$ | $\mathcal{L}_4$ |
|---|---|---|---|---|
| 20.0 | **0.187** | **0.283** | **0.178** | 3.399 |
| 40.0 | 0.199 | 0.307 | 0.197 | **3.246** |
| 60.0 | 0.192 | 0.299 | 0.196 | 3.255 |
| 80.0 | 0.196 | 0.306 | 0.196 | 3.910 |
| 100.0 | 0.199 | 0.302 | 0.199 | 3.742 |
| 120.0 | 0.210 | 0.333 | 0.212 | 3.557 |
| 140.0 | 0.202 | 0.313 | 0.193 | 4.380 |
| 160.0 | 0.250 | 0.398 | 0.252 | 5.758 |

Table 3: Lebesgue measure contributions of each $f$ on datasets: CAL500 to genbase

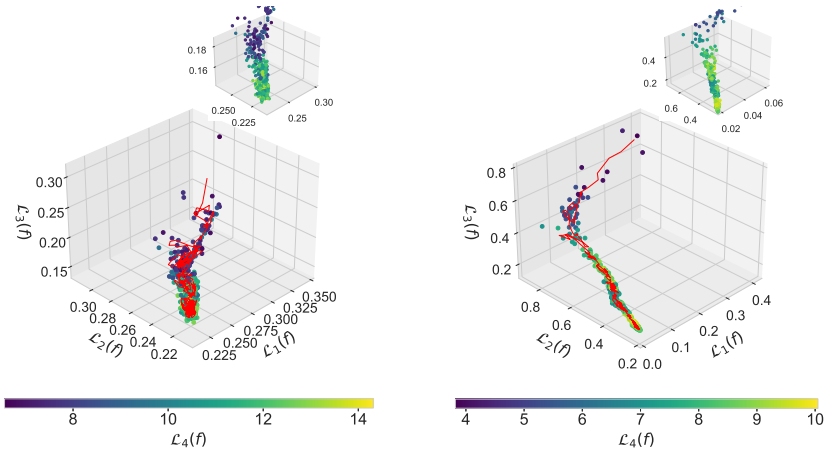
| Dataset | Method | Solution HV | Contribution | Normalized Contribution | Geometric Mean |
|---|---|---|---|---|---|
| CAL500 | GNB-BR | (0.547, 0.713, 0.647) | 0 | 0 | 0.631976 |
| CAL500 | GNB-CC | (0.255, 0.633, 0.741) | 0 | 0 | 0.492774 |
| CAL500 | MLKNN | (0.150, 0.637, 0.554) | 0.000351 | 0.020488 | 0.375585 |
| CAL500 | C2AE | (0.258, 0.536, 0.534) | 0 | 0 | 0.419409 |
| CAL500 | CLIF | (0.137, 0.681, 0.502) | 0.007068 | 0.412185 | 0.360752 |
| CAL500 | DELA | (0.171, 0.633, 0.596) | 0 | 0 | 0.400750 |
| CAL500 | CLML | (0.168, 0.526, 0.520) | 0.009729 | 0.567326 | 0.358231 |
| yeast | GNB-BR | (0.319, 0.472, 0.351) | 0 | 0 | 0.375014 |
| yeast | GNB-CC | (0.319, 0.481, 0.415) | 0 | 0 | 0.399048 |
| yeast | MLKNN | (0.213, 0.375, 0.298) | 0 | 0 | 0.287821 |
| yeast | C2AE | (0.221, 0.358, 0.272) | 0.003081 | 0.425447 | 0.278355 |
| yeast | CLIF | (0.227, 0.391, 0.275) | 0 | 0 | 0.290108 |
| yeast | DELA | (0.226, 0.391, 0.276) | 0 | 0 | 0.289957 |
| yeast | CLML | (0.211, 0.364, 0.266) | 0.004160 | 0.574553 | 0.273480 |
| enron | GNB-BR | (0.198, 0.725, 0.776) | 0 | 0 | 0.481206 |
| enron | GNB-CC | (0.125, 0.638, 0.742) | 0 | 0 | 0.389782 |
| enron | MLKNN | (0.056, 0.529, 0.436) | 0 | 0 | 0.234964 |
| enron | C2AE | (0.189, 0.665, 0.487) | 0 | 0 | 0.393941 |
| enron | CLIF | (0.053, 0.499, 0.381) | 0.002600 | 0.449837 | 0.216576 |
| enron | DELA | (0.054, 0.493, 0.386) | 0.000126 | 0.021742 | 0.218104 |
| enron | CLML | (0.054, 0.488, 0.411) | 0.003055 | 0.528421 | 0.220966 |
| genbase | GNB-BR | (0.052, 0.479, 0.538) | 0 | 0 | 0.237314 |
| genbase | GNB-CC | (0.008, 0.078, 0.091) | 0 | 0 | 0.037745 |
| genbase | MLKNN | (0.033, 0.454, 0.331) | 0 | 0 | 0.170749 |
| genbase | C2AE | (0.345, 0.823, 0.561) | 0 | 0 | 0.542500 |
| genbase | CLIF | (0.046, 0.793, 0.539) | 0 | 0 | 0.269161 |
| genbase | DELA | (0.002, 0.020, 0.001) | 0.144566 | 1.000000 | 0.003138 |
| genbase | CLML | (0.020, 0.239, 0.117) | 0 | 0 | 0.082065 |

Table 4: Lebesgue measure contributions of each $f$ on datasets: emotions to mediamill

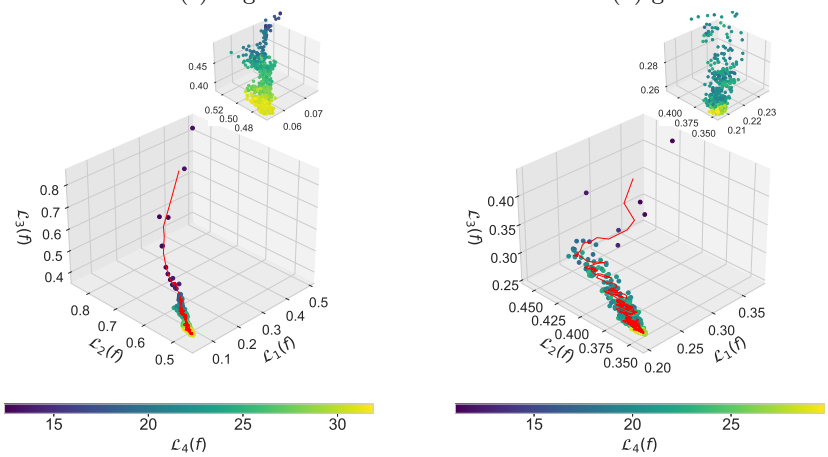
| Dataset | Method | Solution HV | Contribution | Normalized Contribution | Geometric Mean |
|---|---|---|---|---|---|
| emotions | GNB-BR | (0.410, 0.458, 0.271) | 0 | 0 | 0.370604 |
| emotions | GNB-CC | (0.265, 0.383, 0.256) | 0 | 0 | 0.295937 |
| emotions | MLKNN | (0.268, 0.497, 0.361) | 0 | 0 | 0.363696 |
| emotions | C2AE | (0.537, 0.556, 0.488) | 0 | 0 | 0.526199 |
| emotions | CLIF | (0.223, 0.412, 0.246) | 0 | 0 | 0.282547 |
| emotions | DELA | (0.216, 0.353, 0.214) | 0.005072 | 0.191264 | 0.253682 |
| emotions | CLML | (0.205, 0.328, 0.224) | 0.021444 | 0.808736 | 0.246669 |
| flags | GNB-BR | (0.443, 0.560, 0.439) | 0 | 0 | 0.477465 |
| flags | GNB-CC | (0.402, 0.496, 0.360) | 0 | 0 | 0.415483 |
| flags | MLKNN | (0.307, 0.302, 0.233) | 0 | 0 | 0.278388 |
| flags | C2AE | (1.000, 1.000, 1.000) | 0 | 0 | 1.000000 |
| flags | CLIF | (0.298, 0.316, 0.217) | 0 | 0 | 0.273610 |
| flags | DELA | (0.271, 0.284, 0.231) | 0.006058 | 0.463227 | 0.260929 |
| flags | CLML | (0.281, 0.285, 0.205) | 0.007020 | 0.536773 | 0.254035 |
| IMDB-F | GNB-BR | (0.276, 0.875, 0.489) | 0 | 0 | 0.490350 |
| IMDB-F | GNB-CC | (0.391, 0.892, 0.506) | 0 | 0 | 0.560657 |
| IMDB-F | MLKNN | (0.044, 0.929, 0.376) | 0.000180 | 0.003999 | 0.249118 |
| IMDB-F | C2AE | (0.052, 0.743, 0.324) | 0.044679 | 0.993734 | 0.231745 |
| IMDB-F | CLIF | (0.049, 0.825, 0.391) | 0 | 0 | 0.250780 |
| IMDB-F | DELA | (0.054, 0.831, 0.381) | 0 | 0 | 0.257251 |
| IMDB-F | CLML | (0.048, 0.802, 0.358) | 0.000102 | 0.002268 | 0.240283 |
| tmc2007-500 | GNB-BR | (0.598, 0.759, 0.822) | 0 | 0 | 0.719892 |
| tmc2007-500 | GNB-CC | (0.413, 0.697, 0.784) | 0 | 0 | 0.608986 |
| tmc2007-500 | MLKNN | (0.065, 0.351, 0.261) | 0 | 0 | 0.181723 |
| tmc2007-500 | C2AE | (0.051, 0.250, 0.145) | 0 | 0 | 0.122771 |
| tmc2007-500 | CLIF | (0.040, 0.203, 0.123) | 0.007761 | 1.000000 | 0.099828 |
| tmc2007-500 | DELA | (0.041, 0.207, 0.128) | 0 | 0 | 0.102543 |
| tmc2007-500 | CLML | (0.080, 0.427, 0.321) | 0 | 0 | 0.222227 |
| mediamill | GNB-BR | (0.338, 0.845, 0.787) | 0 | 0 | 0.608021 |
| mediamill | GNB-CC | (0.130, 0.708, 0.786) | 0 | 0 | 0.416542 |
| mediamill | MLKNN | (0.030, 0.412, 0.283) | 0 | 0 | 0.152028 |
| mediamill | C2AE | (0.042, 0.445, 0.285) | 0 | 0 | 0.174138 |
| mediamill | CLIF | (0.027, 0.364, 0.216) | 0.035504 | 1.000000 | 0.128975 |
| mediamill | DELA | (0.031, 0.380, 0.252) | 0 | 0 | 0.144388 |
| mediamill | CLML | (0.035, 0.464, 0.353) | 0 | 0 | 0.178885 |

## 5 Extended evaluation of multi-label classification performances

Tables 3 and 4 show the expanded view of the loss values ($\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$), the Lebesgue contribution ($\lambda(P(f))$), the normalised Lebesgue contribution, and geometric means of each comparative method on each dataset. A zero value on the Lebesgue contribution indicates that a given function is dominated by all other functions on the given dataset, *i.e.*, it does not contribute toward the improvement of the volume over $\boldsymbol{\mathcal{L}}(f(\mathbf{X}), \mathbf{Y})$.

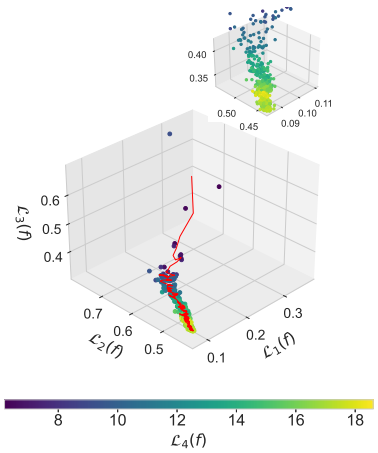



(c) flags (d) genbase

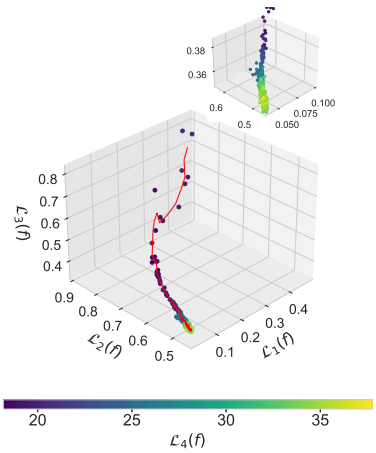



(e) enron (f) yeast

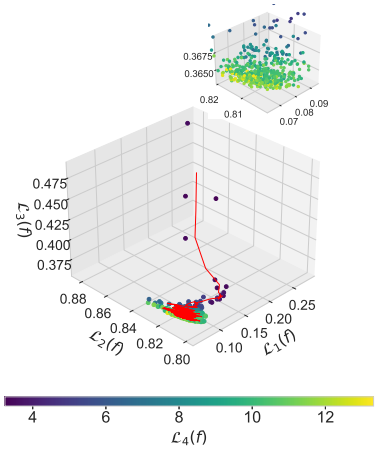
Fig. 3: The training curves of CLML on datasets flags through yeast (c-f).




(g) tmc2007-500




(h) mediamill




(i) IMDB-F

Fig. 4: The training curves of CLML on datasets tmc2007-500 through IMDB-F (g-i).

## 6 Extended results of training curves against surrogate loss

Figures 3 and 4 plot the training curves of CLML on datasets flags through IMDB-F (c-i).

## 7 Useful definitions, corollaries, and lemmas

**Definition 1 (Metric Risk).** *We define the conditional and Bayes risk of* $\mathcal{L}_1, \mathcal{L}_2$, *and* $\mathcal{L}_3$ *given* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *for* $i = 1, 2, 3$ *as follows:*

$$\begin{aligned} R_{\mathcal{L}_i}(f) &= \frac{1}{N}\sum_{j=1}^{N}\sum_{\boldsymbol{y}_j\in\mathcal{Y}} p(\boldsymbol{y}_j|\boldsymbol{x}_j)\mathcal{L}_i(f(\boldsymbol{x}_j),\boldsymbol{y}_j) \\ R^B_{\mathcal{L}_i}(f) &= \frac{1}{N}\sum_{j=1}^{N}\inf_{f'}[\sum_{\boldsymbol{y}_j\in\mathcal{Y}} p(\boldsymbol{y}_j|\boldsymbol{x}_j)\mathcal{L}_i(f'(\boldsymbol{x}_j),\boldsymbol{y}_j)] \end{aligned} \tag{8}$$

The overall risk and Bayes risk is given by:

$$\begin{aligned} R_{\boldsymbol{\mathcal{L}}}(f) &= (R_{\mathcal{L}_1}(f), R_{\mathcal{L}_2}(f), R_{\mathcal{L}_3}(f)) \\ R^B_{\boldsymbol{\mathcal{L}}}(f) &= (R^B_{\mathcal{L}_1}(f), R^B_{\mathcal{L}_2}(f), R^B_{\mathcal{L}_3}(f)) \end{aligned} \tag{9}$$

**Corollary 1 (Below-bounded and Interval).** *The Lebesgue measure is naturally below-bounded and interval, i.e., for any* $F, F'$ *and* $R, R' \subset Z$, $\lambda(H(F,R)) = \lambda(H(F',R'))$ *or* $|\lambda(H(F,R)) - \lambda(H(F',R'))| > 0$, *which is naturally inherited from the underlying below-bounded and interval properties of* $\mathcal{L}_1, \mathcal{L}_2$ *and* $\mathcal{L}_3$ *following [1].*

**Lemma 1 (The Lebesgue Contribution Equals Lebesgue Improvement).** *Let* $\lambda(H(F,R))$ *denote the Lebesgue measure over a set* $F$. *The overall improvement toward the minimisation of* $\mathcal{L}_1, \mathcal{L}_2$, *and* $\mathcal{L}_3$, *is prescribed by the volume of* $\lambda(H(F,R))$, *which can be expressed as the sum of contributions of losses for each function representation* $f \in F$*:*

$$\begin{aligned} \lambda(H(F,R)) &= \sum_{f\in F}\lambda(P(f)) = \\ &\sum_{f\in F}\int_{\mathbb{R}^o} \mathbf{1}_{H(\{f\},R)\setminus H(F\setminus\{f\},R)}(\boldsymbol{z})d\boldsymbol{z} \end{aligned} \tag{10}$$

*Proof.* Consider a redefined Lebesgue measure as the union of non-overlapping (disjoint) contribution regions for each $f \in F$. By substitution:

$$\begin{aligned} \lambda(H(F,R)) &= \int_{\mathbb{R}^o} \mathbf{1}_{H(F,R)}(\mathbf{z})d\mathbf{z} = \\ &\int_{\mathbb{R}^o} \mathbf{1}_{\cup_{f\in F} H(\{f\},R)\setminus H(F\setminus\{f\},R)}(\mathbf{z})d\mathbf{z} \end{aligned} \tag{11}$$

The integral can be re-written to express the sum over disjoint contribution regions:

$$\int_{\mathbb{R}^o} \mathbf{1}_{\cup_{f\in F} H(\{f\},R)\setminus H(F\setminus\{f\},R)}(\mathbf{z})d\mathbf{z} = \sum_{f\in F}\int_{\mathbb{R}^o} \mathbf{1}_{H(\{f\},R)\setminus H(F\setminus\{f\},R)}(\mathbf{z})d\mathbf{z} = \sum_{f\in F}\lambda(P(f)). \tag{12}$$

## References


1. Gao, W., Zhou, Z.H.: On the consistency of multi-label learning. In: Kakade, S.M., von Luxburg, U. (eds.) Proceedings of the 24th Annual Conference on Learning Theory. Proceedings of Machine Learning Research, vol. 19, pp. 341–358. PMLR, Budapest, Hungary (09–11 Jun 2011)
2. Gong, Y., Rouditchenko, A., Liu, A.H., Harwath, D., Karlinsky, L., Kuehne, H., Glass, J.R.: Contrastive audio-visual masked autoencoder. In: The Eleventh International Conference on Learning Representations (2022)
3. Laplace, P.S.: Théorie analytique des probabilités. Courcier (1814)
4. Sechidis, K., Tsoumakas, G., Vlahavas, I.: On the stratification of multi-label data. In: European Conference on Machine Learning and Knowledge Discovery in Database. pp. 145–158. Springer (2011)