Journal of Machine Learning Research 23 (2022) 1-30 Submitted 11/19; Revised 7/21; Published 7/22

# A Worst Case Analysis of Calibrated Label Ranking Multi-label Classification Method

**Lucas H. S. Mello** LUCASHSMELLO@GMAIL.COM
*Department of Informatics, Federal University of Espírito Santo, Vitória, Brazil*

**Flávio M. Varejão** FVAREJAO@INF.UFES.BR
*Department of Informatics, Federal University of Espírito Santo, Vitória, Brazil*

**Alexandre L. Rodrigues** ALEXANDRE.RODRIGUES@UFES.BR
*Department of Statistics, Federal University of Espírito Santo, Vitória, Brazil*

**Editor:** Francis Bach

## Abstract

Most multi-label classification methods are evaluated on real datasets, which is a good practice for comparing the performance among methods on the average scenario. Due to the large amount of factors to consider, this empirical approach does not explain, nor does show the factors impacting the performance. A reasonable way to understand some of the performance's factors of multi-label methods independently of the context is to find a mathematical proof about them. In this paper, mathematical proofs are given for the multi-label method ranking by pairwise comparison and its extension for classification named by calibrated label ranking, showing their performance on a worst case scenario for five multi-label metrics. The pairwise approach adopted by ranking by pairwise comparison enables the algorithm to achieve the optimal performance on Spearman rank correlation. However, the findings presented in this paper clearly show that the same pairwise approach adopted by the algorithm is also a crucial factor contributing to a very poor performance on other multi-label metrics.

**Keywords:** Multi-label learning, Loss minimization, Pairwise preference

## 1. Introduction

In multi-label classification (MLC) an instance can be associated with multiple classes or categories simultaneously. This dramatically affects complexity because the number of possible classifications increases exponentially with respect to the number of classes. A popular approach for developing multi-label methods is to transform the MLC problem into several binary classification problems, which can then be handled by single-label classifiers. Methods of this kind are called transformation-based multi-label methods, where the most common one is the Binary Relevance ($BR$) method (Tsoumakas and Katakis, 2007). More complex transformation-based methods are usually grounded on the main idea of exploiting dependencies among labels (Read et al., 2009; Montañés et al., 2014; Younes et al., 2011). The *calibrated label ranking by pairwise classification* (CLR) (Fürnkranz et al., 2008) is a method that exploits pairwise dependencies by using a composition of single-label classifiers, one for each distinct pair of labels.

Empirical evidence clearly shows CLR being good for optimizing ranking metrics in general, such as One-error and Coverage (Fürnkranz et al., 2008; Trohidis et al., 2011; Zhang and Schneider, 2012a; Tahir et al., 2016; Huang et al., 2019). On the other hand, it has also been shown empirically that CLR is not good for optimizing example-based metrics on average (Trohidis et al., 2011; Zhang and Schneider, 2012a,b; Wang et al., 2014; Tahir et al., 2016; Huang et al., 2019; Sun et al., 2019). This paper shows similar results, but extended to the worst-case scenario with the help of mathematical proofs. Although CLR has been used as a baseline for many multi-label methods (Fürnkranz et al., 2008; Trohidis et al., 2011; Zhang and Schneider, 2012a,b; Wang et al., 2014; He et al., 2019), yet they fail to provide a formal explanation to what makes CLR achieves a good/poor performance is still missing. Knowing such an explanation helps researchers choose and better understand multi-label methods. Therefore, the main objective in this paper is to explain what drives the CLR performance. To achieve this objective, the probabilistic framework of Dembczyński et al. (2012) is adopted and mathematical proofs are built upon it for the worst case scenario. With the results published in this paper, it is possible to conclude that the pairwise comparison approach adopted by CLR is the main factor behind a poor performance on some specific datasets. The main characteristic of these specific datasets is the presence of several pairs of labels that are mutually exclusive (i.e when one occurs, the other label does not). In datasets of this kind, the performance of the CLR is very sensible to the distribution of labels, i.e., a slight modification on the label distribution may lead to a great change on the CLR performance.

Mathematical results for different MLC methods are provided by Dembczyński et al. (2012) and Waegeman et al. (2014). They provide theoretical proofs with respect to the worst-case scenario of some MLC methods such as the Hamming loss optimizer, subset 0/1 loss optimizer and F-measure optimizer. They also analyze the performance of the optimal methods over other metrics such as Hamming loss, subset 0/1 loss, Jaccard distance, and F-measure. Motivated by those papers, this paper provides results for the same metrics but focusing on CLR. To the best of our knowledge, this is the first work to address this problem.

The paper is organized as follows. Section 2 provides a formal definition of multi-label classification where the probabilistic framework is presented. Section 3 presents the main results of this paper, where the worst-case scenario of CLR with respect to four metrics is analyzed. Section 4 shows final observations about this research.

## 2. Multi-label Learning

Let $\mathcal{X}$ denote a feature space and $\mathcal{L} = \{l_1, l_2, l_3, ...l_n\}$ be a set of labels with $n = |\mathcal{L}|$. An instance is defined as a pair of two vectors $(\mathbf{x}, \mathbf{y})$ where $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y}$ is a labelling (combination of labels) represented by a binary vector $\mathbf{y} = (y_1, y_2, ..., y_n)$ such that $y_i = 1$ only if the respective instance is associated to label $l_i$. A method $\mathbf{h}$ used to solve a multi-label task may be denoted as a multi-label classifier or a multi-label ranker. In the former, $\mathbf{h}$ is a function that maps the feature space $\mathcal{X}$ into the set of all possible labellings of $n$ labels. Let this set be denoted by $\mathcal{Y}$, therefore $\mathbf{h} : \mathcal{X} \to \mathcal{Y}$ and for a given instance $\mathbf{x} \in \mathcal{X}$ it returns a vector $\mathbf{h}(\mathbf{x}) = (h_1(\mathbf{x}), h_2(\mathbf{x}), ..., h_n(\mathbf{x}))$. In the latter, $\mathbf{h}$ is a function that maps the feature space $\mathcal{X}$ into the set of all possible rankings, i.e set of all possible permutations

of $\{1, 2, ..., n\}$. Let this set be denoted by $\mathfrak{S}_\mathbf{n}$, therefore $\mathbf{h} : \mathcal{X} \to \mathfrak{S}_n$ and, in this case, $h_i(\mathbf{x})$ denotes the rank of label $i$ for a given instance $\mathbf{x} \in \mathcal{X}$.

The framework proposed by Dembczyński et al. (2012) assumes that labellings are distributed according to a conditional probability distribution $\mathbf{P}(\mathbf{Y}|\mathbf{x})$ where $\mathbf{x} \in \mathcal{X}$ and $\mathbf{Y}$ is a random vector defined on $\mathcal{Y}$. This means that for a specific feature vector $\mathbf{x}$, each labelling $\mathbf{y} \in \mathcal{Y}$ occurs with a probability of $\mathbf{P}(\mathbf{Y} = \mathbf{y}|\mathbf{x})$ [1].

The risk of a multi-label method $\mathbf{h}$ and dataset features $\mathbf{x} \in \mathcal{X}$ is defined as the conditional expected loss

$$\begin{aligned} R_L(\mathbf{h}, \mathbf{x}) &= \mathbb{E}_{\mathbf{Y}|\mathbf{X}} L(\mathbf{Y}, \mathbf{h}(\mathbf{x})) \\ &= \sum_{\mathbf{y} \in \mathcal{Y}} L(\mathbf{y}, \mathbf{h}(\mathbf{x})) \cdot \mathbf{P}(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x}), \end{aligned} \tag{1}$$

where $\mathbf{P}(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x})$ represents the conditional probability of $\mathbf{Y}$ given feature vector $\mathbf{x}$ and $L(\cdot)$ is a loss function for multi-label predictions. The regret of a multi-label method $\mathbf{h}$ with respect to a loss function $L$ is defined as

$$r_L(\mathbf{h}, \mathbf{x}) = R_L(\mathbf{h}, \mathbf{x}) - R_L(\mathbf{h}^*, \mathbf{x}), \tag{2}$$

where $\mathbf{h}^*$ is a Bayes-optimal method that yields the minimum loss for $L$, defined as

$$\mathbf{h}^*(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{y}} \mathbb{E}_{\mathbf{Y}|\mathbf{X}}[L(\mathbf{Y}, \mathbf{y})]. \tag{3}$$

Since the feature vector $\mathbf{x}$ is always given and fixed at the start of all proofs or analysis in this paper, the given feature vector $\mathbf{x}$ will be omitted. In the same way, $\mathbf{x}$ will be omitted for any multi-label method $\mathbf{h}$, meaning that, $h_i = h_i(\mathbf{x})$ for any $i$.

For the rest of this paper, let the notation $\mathbf{P}^{(i)}$, for an arbitrary distribution $\mathbf{P}$ of $\mathbf{Y}$, be defined as the marginal distribution of label $i$:

$$\mathbf{P}^{(i)} = \mathbf{P}(Y_i = 1) = \sum_{\mathbf{y} \in \mathcal{Y}: y_i = 1} \mathbf{P}(\mathbf{y}).$$

The task of risk minimization is finding the optimal model $\mathbf{h}^*$ defined in Equation (3). Clearly this can be achieved by exhaustive search, which is testing all $2^n$ possible labellings for classification, or testing all $n!$ rankings for label ranking, but a more efficient way is desirable. In general, this is NP-hard as it contains a particular instance of risk minimization of the Jaccard distance, proved to be NP-complete (Chierichetti et al., 2010). Therefore, efficient algorithms are only designed for specific metrics where specific properties can be exploited, which is the case for Hamming loss, F-measure and rank loss (Dembczyński et al., 2012).

### 2.1 Multi-label metrics

Multi-label metrics are used to measure the quality of predictions or the cost for inaccuracy of predictions. When a metric quantifies the error, it is called loss function, otherwise it is

1. $\mathbf{P}(\mathbf{Y} = \mathbf{y}|\mathbf{x})$ stands for the probability that an instance has labelling $\mathbf{y}$ given its feature vector $\mathbf{x}$.

called an utility function. In multi-label learning, there are two main types of metrics: the instance-wise decomposable and label-wise decomposable. The instance-wise decomposable metrics are those which can be expressed as an average of losses computed individually for each instance. This one can be expressed as in Equation (1), where the loss function is a function $L(\cdot)$ of the target labelling $\mathbf{y}$ (the true labelling), and the predicted output of a multi-label method $\mathbf{h}$, associating a penalty to errors in multi-label prediction. All metrics analyzed in this paper are assumed to be instance-wise decomposable. The label-wise decomposable metrics are those that can be expressed as an average of losses computed individually for each label. They are generally an average over a metric for binary classification applied individually to each label. Note that a metric can be both instance-wise and label-wise decomposable (e.g., Hamming loss).

As discussed in the previous section, the predicted output of a multi-label method can be a labelling or a ranking. For the sake of easy reading, $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, ..., \hat{y}_n)$ will be used to denote a predicted labelling and $\hat{\mathbf{z}} = (\hat{z}_1, \hat{z}_2, ..., \hat{z}_n)$ to denote a predicted ranking. Hamming loss is a metric for classification, defined as the fraction of labels incorrectly predicted:

$$L_H(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{n} \sum_{i=1}^{n} [\![ y_i \neq \hat{y}_i ]\!], \tag{4}$$

where $[\![.]\!]$ is the Iverson bracket.

Other common loss function is the subset 0/1 loss, which detects a strict coincidence of the actual and estimated labels as

$$L_s(\mathbf{y}, \hat{\mathbf{y}}) = [\![ \mathbf{y} \neq \hat{\mathbf{y}} ]\!]. \tag{5}$$

More elaborate loss functions are the loss version of the F-measure and the Jaccard distance given respectively by

$$L_F(\mathbf{y}, \hat{\mathbf{y}}) = 1 - \frac{2 \sum_{i=1}^{n} y_i \hat{y}_i}{\sum_{i=1}^{n} (y_i + \hat{y}_i)}$$

and

$$L_J(\mathbf{y}, \hat{\mathbf{y}}) = 1 - \frac{\sum_{i=1}^{n} y_i \hat{y}_i}{\sum_{i=1}^{n} (y_i + \hat{y}_i) - \sum_{i=1}^{n} y_i \hat{y}_i}.$$

A simple metric that takes into account a rank is the rank loss, which is defined as

$$L_r(\mathbf{y}, \hat{\mathbf{z}}) = \sum_{(i,j): y_i > y_j} [\![ \hat{z}_i < \hat{z}_j ]\!].$$

The normalized rank loss is defined as

$$L_{\hat{r}}(\mathbf{y}, \hat{\mathbf{z}}) = \frac{L_r(\mathbf{y}, \hat{\mathbf{z}})}{s_{\mathbf{y}}(n - s_{\mathbf{y}})},$$

where $s_{\mathbf{y}} = \sum_{i=1}^{n} y_i$.

These metrics are the most common in the multi-label scenario, and they will be used to analyze the performance of CLR. Not so common, but important for analyzing CLR, are two metrics from the preference learning field: the squared rank distance

$$L_{\text{srd}}(\mathbf{z}, \hat{\mathbf{z}}) = \sum_{i}^{n} (z_i - \hat{z}_i)^2$$

and the Spearman rank correlation[2] (Hüllermeier and Fürnkranz, 2004). The Spearman rank correlation is defined as the Pearson correlation between the ranked values of two variables. In the context of preference learning, the Spearman rank correlation can be obtained by the following formula

$$1 - \frac{6L_{\text{srd}}(\mathbf{z}, \hat{\mathbf{z}})}{n(n^2 - 1)}.$$

The Spearman rank correlation can be interpreted as a linear normalization of the squared rank distance to the interval $[-1, 1]$.

A review of optimal risk minimizers for the metrics defined in this section are presented in the next section.

### 2.2 Optimal Risk and Regret

This subsection presents a brief review and definitions on optimal solutions for the risk minimization and the regret of some metrics defined in Section 2.1. These definitions are useful for the proofs in Section 3.

In Dembczyński et al. (2010) the authors proved that the optimal labelling $\mathbf{y}^*$ for the risk of a Hamming loss can be obtained by just looking at the marginal distribution of labels, and it is given by

$$y_i^* = \begin{cases} 1, & \text{if } \mathbf{P}^{(i)} > \frac{1}{2}, \\ 0, & \text{if } \mathbf{P}^{(i)} \leq \frac{1}{2}. \end{cases} \tag{6}$$

The authors Dembczyński et al. (2012) have shown the optimal labelling $\mathbf{y}^*$ for the risk of subset 0/1 loss is given by the mode of the distribution:

$$\mathbf{y}^* = \operatorname*{argmax}_{\mathbf{y}} \mathbf{P}(\mathbf{y}).$$

Interestingly, they showed that the optimal expected Hamming loss may give the worst case regret of 1/2 in subset 0/1 loss. Furthermore, the optimal expected subset 0/1 loss solution may give a regret as closely as possible to 1, with respect to Hamming loss.

Dembczyński et al. (2012) have also shown that to achieve optimal ranking $\mathbf{z}^*$ in rank loss, it is sufficient to order the labels with respect to their marginal distribution:

$$z_i^* < z_j^* \quad \Longleftrightarrow \quad \mathbf{P}^{(i)} > \mathbf{P}^{(j)}.$$

2. Spearman rank correlation in the preference learning and multi-label ranking field is a utility function. The higher, the better.

Interestingly, the regret of rank loss can be obtained by just summing the difference $\mathbf{P}^{(i)} - \mathbf{P}^{(j)}$ of all pairs $(i, j)$ with misorder (Dembczyński et al., 2012):

$$r_r(\mathbf{z}) = \sum_{(i,j):z_i^* < z_j^*} [\![ z_i > z_j ]\!] \Big( \mathbf{P}^{(i)} - \mathbf{P}^{(j)} \Big). \tag{7}$$

Regarding the normalized rank loss, let $s_{\max}$ be defined as $s_{\max} = \max_{\mathbf{y}:\mathbf{P}(\mathbf{y})>0} s_{\mathbf{y}}(n - s_{\mathbf{y}})$, it is easy to see that

$$R_{\hat{r}}(\mathbf{z}) \geq \frac{R_r(\mathbf{z})}{s_{\max}} \tag{8}$$

The authors Hüllermeier and Fürnkranz (2004) proved that the ranking constructed by CLR is optimal for squared rank distance and, consequently, for Spearman rank correlation, two metrics commonly used for preference learning. As it will be seen in the next section, part of the pairwise approach adopted by CLR is essentially learning the preference of one label over another, which may be one of the reasons why this approach is optimal for Spearman rank correlation.

These formulas usually provide a much easier way of analyzing the risk and regret than using the general formula at Equation (1), since the general formula is an equation of $2^n$ parameters of the label distribution. For instance, the regret of rank loss can be computed by only using the marginal distributions. So, these formulas are widely used in the theorems presented in Section 3.

### 2.3 Ranking by pairwise comparison

Ranking by pairwise comparison (RPC) is a multi-label method composed of $\frac{n(n-1)}{2}$ binary classifiers, with the purpose of building a ranking for a given instance. The ranking is built by first giving a score $s_i$ for each label $i$. The score is computed by a pairwise preference scheme where there exists a binary classifier for each distinct pair of labels (say $i$ and $j$) whose task is to distinguish the occurrence of label $i$ and label $j$ when assuming that only one of both occurs. Therefore, each classifier outputs its preference towards one of the two labels. A pseudo code for training RPC is presented in Algorithm 1 and the computation of the score of a single label is presented in Algorithm 2.

Given this definition, let RPC be defined as a ranking method that prefers label $i$ to label $j$ if $s_i > s_j$, where $s_i$ is computed by

$$s_i = \sum_{k \neq i} [\![ \mathbf{P}(Y_i = 1, Y_k = 0 | Y_i = 1 \oplus Y_k = 1) > 0.5 ]\!],$$

where $(Y_i = 1 \oplus Y_k = 1)$ means $Y_i = 1$ or $Y_k = 1$ exclusively and $[\![p > 0.5]\!]$ evaluates to 1 if $p > \frac{1}{2}$, and 0 otherwise. The probability is conditioned on $Y_i = 1 \oplus Y_k = 1$, because in Algorithm 1, the binary classifier $c_{ij}$ is trained on $\mathbf{D}' \cup \mathbf{D}''$ (Line 7), which is equivalent to $\{(\mathbf{x}, \mathbf{y}) \in \mathbf{D} : y_i = 1 \oplus y_j = 1\}$, but replacing all labellings $\mathbf{y}$ with a 1 when $y_i = 1$, and with a 0 when $y_j = 1$. Therefore, the value $[\![\mathbf{P}(Y_i = 1, Y_k = 0 | Y_i = 1 \oplus Y_k = 1) > 0.5]\!]$ corresponds to the vote given by the binary classifier responsible for distinguishing the presence of label

**Algorithm 1:** Algorithm for training RPC.

**Data:** Training data set of $m$ samples $\mathbf{D} = \{(\mathbf{x}_1, \mathbf{y}_1), \cdots, (\mathbf{x}_m, \mathbf{y}_m)\}$
**Result:** Trained binary classifiers $c_{ij}$ for $1 \leq i \leq n$, $1 \leq j \leq n$ and $i \neq j$.

**for** *each pair of labels* $i, j$ **do**
  $\mathbf{D}' := \{(\mathbf{x}, \mathbf{y}) \in \mathbf{D} : y_i = 1 \text{ and } y_j = 0\}$
  $\mathbf{D}' := \{(\mathbf{x}, 1) : (\mathbf{x}, \mathbf{y}) \in \mathbf{D}'\}$ `// replace all labellings with a 1, representing the positive class.`
  $\mathbf{D}'' := \{(\mathbf{x}, \mathbf{y}) \in \mathbf{D} : y_i = 0 \text{ and } y_j = 1\}$
  $\mathbf{D}'' := \{(\mathbf{x}, 0) : (\mathbf{x}, \mathbf{y}) \in \mathbf{D}''\}$ `// replace all labellings with a 0, representing the negative class`
  $c_{ij} :=$ train_binary_classifier$(\mathbf{D}' \cup \mathbf{D}'')$ `// Binary classification problem.`
**end**

**Algorithm 2:** Scoring a single label $i$ in RPC.

**Input:** Trained binary classifiers $c_{ij}$ for all $j \neq i$.
**Result:** Score $s \in \mathbb{N}$

$s := 0$
**for** *each label* $j$ *different of* $i$ **do**
  $l =$ predict_label$(c_{ij}, \mathbf{x})$ `// Function` *`predict_label`* `returns 1 if` $i$ `is predicted positive, otherwise 0.`
  $s := s + l$ `// +1 if` $i$ `is predicted positive by` $c_{ij}$`.`
**end**

pair $(i, k)$. It is worth mentioning that $\mathbf{P}(Y_i = 1, Y_k = 0 | Y_i = 1 \oplus Y_k = 1)$ can be rewritten as:

$$\mathbf{P}(Y_i = 1, Y_k = 0 | Y_i = 1 \oplus Y_k = 1) = \frac{\mathbf{P}(Y_i = 1, Y_k = 0)}{\mathbf{P}(Y_i = 1, Y_k = 0) + \mathbf{P}(Y_i = 0, Y_k = 1)},$$

which is sometimes a more convenient form for calculating this conditional probability.

There may exist cases in which $Y_i = 1 \oplus Y_k = 1$ never occurs. In practice, this would mean that the binary classifier responsible for distinguishing label $i$ from $k$ would be trained on an empty dataset. In this case, usually a value from $\{0, \frac{1}{2}, 1\}$ ($\frac{1}{2}$ is the most frequent choice) is arbitrarily adopted for $[\![\mathbf{P}(Y_i = 1, Y_k = 0 | Y_i = 1 \oplus Y_k = 1) > 0.5]\!]$. Whatever the choice, as long as,

$$\mathbf{P}(Y_i = 1, Y_k = 0 | Y_i = 1 \oplus Y_k = 1) + \mathbf{P}(Y_i = 0, Y_k = 1 | Y_i = 1 \oplus Y_k = 1) = 1,$$

is satisfied, which is already true for $\mathbf{P}(Y_i = 1 \oplus Y_k = 1) \neq 0$, the proofs in this work are valid.

Calibrated label ranking (CLR) is an adaptation of RPC for multi-label classification. It adds an artificial label for constructing a bi-partition (a.k.a classification). The score of the artificial label is given by $n$ binary classifiers that are identical to the $n$ binary classifiers

of binary relevance method, as pointed out by Fürnkranz et al. (2008). The artificial label represents the "negative label" inside the one-against-all strategy of binary relevance. A label is said to be positive or relevant if the score $s_i$, as defined above, is greater than the score of the artificial label. Note that now the score $s_i$ should also include the artificial label. Therefore, CLR is a classifier that predicts label $i$ as positive only if

$$\sum_{k \neq i} [\![\mathbf{P}(Y_i = 1, Y_k = 0 | Y_i = 1 \oplus Y_k = 1) > 0.5]\!] + [\![\mathbf{P}(Y_i = 1) > 0.5]\!] > \sum_{k=1}^{n} [\![\mathbf{P}(Y_k = 0) > 0.5]\!].$$

The summation on the right-hand side of the inequality counts the number of votes in favor of the calibrated/artificial label and $[\![\mathbf{P}(Y_i = 1) > 0.5]\!]$ corresponds to the vote given by a one-against-all classifier (the classifier present in binary relevance method). Observe that, although CLR is trained on $Y_i = 1 \oplus Y_k = 1$, the algorithm can output multiple positive labels. It will usually output multiple positive labels if $\sum_{k=1}^{n} [\![\mathbf{P}(Y_k = 0) > 0.5]\!]$ is low, i.e, if the label cardinality is high.

Although the name CLR is often used in the literature to describe its ranking and/or classification components, the name RPC is used to emphasize the ranking component while CLR to emphasize its multi-label classification component in this paper. For the sake of simplicity, define function $f(\mathbf{P}, i, j)$ as

$$f(\mathbf{P}, i, j) = \begin{cases} \mathbf{P}(Y_i = 1, Y_j = 0 | Y_i = 1 \oplus Y_j = 1), & \text{if } i \neq j \\ 0, & \text{if } i = j, \end{cases}$$

so that, the CLR prediction of label $i$ can be redefined as:

$$\sum_{j=1}^{n} [\![f(\mathbf{P}, i, j) > 0.5]\!] \; + \; [\![\mathbf{P}(Y_i = 1) > 0.5]\!] > \sum_{j=1}^{n} [\![\mathbf{P}(Y_j = 0) > 0.5]\!],$$

and the RPC preference of $i$ over $j$ can be redefined as:

$$\sum_{k=1}^{n} [\![f(\mathbf{P}, i, k) > 0.5]\!] > \sum_{k=1}^{n} [\![f(\mathbf{P}, j, k) > 0.5]\!].$$

The versions where both CLR and RPC use the probability values as weights for voting are respectively expressed as

$$\sum_{j=1}^{n} f(\mathbf{P}, i, j) \; + \; \mathbf{P}(Y_i = 1) > \sum_{j=1}^{n} \mathbf{P}(Y_j = 0), \tag{9}$$

and

$$\sum_{k=1}^{n} f(\mathbf{P}, i, k) > \sum_{k=1}^{n} f(\mathbf{P}, j, k). \tag{10}$$

Note that the scores given by RPC and CLR to label $i$ are respectively defined as

$$s_i = \sum_{k=1}^{n} f(\mathbf{P}, i, k), \tag{11}$$

and

$$s_i = \sum_{j=1}^{n} f(\mathbf{P}, i, j) \ + \ \mathbf{P}(Y_i = 1). \tag{12}$$

This paper focuses on the weighted versions of RPC and CLR. We argue that the analysis presented in this paper for the weighted version is still valid for the original version, since the worst-case scenario distribution, named $\hat{P}$, has a special property where $f(\hat{\mathbf{P}}, i, j)$ is close to 0 or to 1, which means $f(\hat{\mathbf{P}}, i, j) \approx [\![ f(\hat{\mathbf{P}}, i, j) > 0.5 ]\!]$. Throughout the paper function $\mathbf{h}^{\text{clr}}$ is used to denote CLR and $\mathbf{h}^{\text{rpc}}$ to denote RPC.

## 3. Theoretical insights

The objective of this section is to present interesting theoretical properties of CLR that show scenarios where CLR should not be used. The results show an issue in the way CLR makes its pairwise comparison, resulting in a poor performance for very particular probability label distributions types. As other authors already suggested, the issue lies mainly on how the probability $\mathbf{P}(Y_i = 1, Y_j = 0 | Y_i = 1 \oplus Y_j = 1)$ is used inside the CLR prediction. The results suggest CLR should be taken with caution when $\mathbf{P}(Y_i = 1 \oplus Y_j = 1)$ is close to zero for some labels $i$ and $j$. A special distribution where this occurs is defined as follows.

Denote $0_n$ as a vector of $n$ zeroes, $1_n$ as a vector of $n$ ones and $\mathbf{y}^{(i)}$ as a $n$-dimensional vector of zeros apart from a one at the $i$-th position. Let $\hat{\mathbf{P}}_m$ denote a special distribution of $\mathbf{Y}$ such that

$$\hat{\mathbf{P}}_m(\mathbf{y}) = \begin{cases} \frac{m+1}{2(n+1)}, & \text{if } \mathbf{y} = 1_n \\ \epsilon, & \text{if } \mathbf{y} = \mathbf{y}^{(i)} \ \text{ for any } 1 \leq i \leq m \\ 1 - \frac{m+1}{2(n+1)} - \epsilon \cdot m, & \text{if } \mathbf{y} = 0_n \\ 0, & \text{otherwise} \end{cases}$$

where $m$ is a positive integer such that $0 < m < n$ and $\epsilon$ is an arbitrary positive real number that is assumed to be "really close" to 0. An example of $\hat{\mathbf{P}}_2$ for $n = 4$:

$$\hat{\mathbf{P}}_2(0,0,0,0) = 70\% - 2\epsilon$$
$$\hat{\mathbf{P}}_2(1,1,1,1) = 30\%$$
$$\hat{\mathbf{P}}_2(1,0,0,0) = \hat{\mathbf{P}}_2(0,1,0,0) = \epsilon,$$

where null probabilities are omitted. The most important point to note about $\hat{\mathbf{P}}_m$ is the high probability of occurrence of labelling $0_n$, specially when $m$ is low. Also, note that $\hat{\mathbf{P}}_m$ has exactly $m + 2$ non-null values. The purpose of $\epsilon$ is to avoid undefined values when calculating $f$ (e.g $\frac{0}{0}$) and to conveniently manipulate the output of function $f$. Proposition 1 shows an important property of this distribution.

**Proposition 1** *When considering distribution $\hat{\mathbf{P}}_m$, CLR predicts ones for the first $m$ labels and zeroes for the other labels, i.e, $\sum_{i=1}^{m} h_i^{clr} = \sum_{i=1}^{n} h_i^{clr} = m$.*

**Proof** See Appendix A. ■

This proposition shows how much CLR is sensible to conditional probabilities. Just an arbitrarily small value $\epsilon$ in $\hat{\mathbf{P}}_m$ makes CLR predicts $m$ labels incorrectly. Next, it will be seen how much this impacts CLR performance where several theorems with respect to the regret of CLR and RPC are presented. Following each theorem, relevant observations are made.

**Theorem 2** *The following upper bound holds for the regret with respect to Hamming loss:*

$$\sup_{\mathbf{P}\in\mathcal{P}_n}(r_H(\mathbf{h}^{clr})) = \begin{cases} \frac{n}{4(n+1)}, & \text{if } n \text{ is even} \\ \frac{n-1}{4n}, & \text{if } n \text{ is odd,} \end{cases}$$

*where $\mathcal{P}_n$ denotes the set of all distributions over $n$ labels such that $\mathbf{P}^{(i)} \leq \frac{1}{2}$ for all $i$.*

**Proof** See Appendix B. ■

An interesting point to observe from Theorem 2 is that there exists at least one distribution in the family $\mathcal{P}_n$ such that $r_H(\mathbf{h}^{\text{clr}}) \leq \frac{1}{4}$. Empirically, CLR and BR (a.k.a one-against-all) has been shown to have a much closer performance on average with respect to Hamming loss, according to experiments in the literature (Fürnkranz et al., 2008; Trohidis et al., 2011; Zhang and Schneider, 2012a,b; Wang et al., 2014).

A more interesting result is presented with respect to subset 0/1 loss in Theorem 3.

**Theorem 3** *The following lower bound holds for the regret with respect to subset 0/1 loss:*

$$\sup_{\mathbf{P}} r_s(\mathbf{h}^{clr}) \geq \frac{n}{n+1}.$$

**Proof** Consider the regret $r_s(\mathbf{h}^{\text{clr}})$ on distribution $\hat{\mathbf{P}}_m$ for $m = 1$. If $\epsilon$ is sufficiently small, then the mode of $\hat{\mathbf{P}}_1$ is $0_n$, and $\hat{\mathbf{P}}_1(0_n) = 1 - \frac{1}{n+1} - \epsilon$. From Proposition 1, it has that $\hat{\mathbf{P}}_1(\mathbf{h}^{\text{clr}}) = \hat{\mathbf{P}}_1(\mathbf{y}^{(1)}) = \epsilon$. As the mode of distribution is an optimal labelling for subset 0/1 loss, the regret on distribution $\hat{\mathbf{P}}_m$ can be written as

$$\begin{aligned} r_s(\mathbf{h}^{\text{clr}}) &= \hat{\mathbf{P}}_1(0_n) - \hat{\mathbf{P}}_1(\mathbf{y}^{(1)}) \\ &= 1 - \frac{1}{n+1} - 2\epsilon \end{aligned}$$

The value of $\epsilon$ can be arbitrarily small, so the supremum of $r_s(\mathbf{h}^{\text{clr}})$ is at least $1 - \frac{1}{n+1} = \frac{n}{n+1}$. ■

Theorem 3 shows that when $n$ tends to infinity, the supremum of regret $r_s(\mathbf{h}^{\text{clr}})$ tends to 1, which is the highest regret possible for subset 0/1 loss. A high regret is already expected as seen in empirical evidence (Trohidis et al., 2011; Zhang and Schneider, 2012a,b; Wang et al., 2014; Tahir et al., 2016; Huang et al., 2019; Sun et al., 2019), but surely not of such magnitude. It is important to note that even for a small number of labels, the worst case regret is high, e.g. for $n = 4$ the highest regret is at least 0.8. Therefore, CLR is definitely not a good method for optimizing subset 0/1 loss if one is concerned about worst case scenarios.

**Theorem 4** *The following lower bound holds for the regret with respect to Jaccard distance:*

$$\sup_{\mathbf{P}} r_J(\mathbf{h}^{clr}) \geq 1 - \frac{1}{n}.$$

**Proof** See Appendix C. ■

Note that if $n \to \infty$, then $r_J(\mathbf{h}^{\text{clr}})$ tends to 1. Again, this is the highest regret possible for Jaccard distance and the regret is also high even for small $n$, e.g. for $n = 4$ the highest regret is at least 0.75. A high regret was already expected, but not this high. This is another metric researchers should be aware when considering the worst case.

**Theorem 5** *The following lower bound holds for the regret with respect to F-measure:*

$$\sup_{\mathbf{P}} r_F(\mathbf{h}^{clr}) \geq 1 - \frac{n+3}{(n+1)^2}.$$

**Proof** See Appendix D. ■

Note that if $n \to \infty$, then $r_F(\mathbf{h}^{\text{clr}})$ tends to 1, which is the highest possible regret for F-measure. For small values of $n$, the highest regret is still high, e.g. for $n = 4$ the highest regret is at least 0.72.

Another interesting result is shown for rank loss in Theorem 6, where RPC does not achieve optimal regret.

**Theorem 6** *For any $n$ divisible by 4, the following lower bound holds for the regret with respect to normalized rank loss:*

$$\sup_{\mathbf{P}} r_{\hat{r}}(\mathbf{h}^{rpc}) \geq \frac{1}{6}.$$

**Proof** See Appendix E. ■

Although Theorem 6 is not conclusive for stating that RPC performs poorly at worst case scenarios, it suggests that RPC does not optimize rank loss for $n \geq 4$, which is not the expected behaviour. The non-optimal performance for RPC does not occur for the same reason as the CLR: the function $f$ can be 1 even when the label cardinality is very low.

As it can be seen in the proofs, the poor performance of CLR in the worst case scenario comes from giving too much importance to conditional probabilities: an arbitrarily small value $\epsilon$ is enough to change the conditional probability at $f$ from zero to one and consequently changing rankings/classifications. The expected value of multi-label metrics does not give such importance to conditional probabilities, as it can be seen in their formulas at equations (6), (7) and others presented by Dembczyński et al. (2012).

It is natural to question how rare are the special distributions used in this work and if there are other distributions that yield similar results. Moreover, it is already expected that CLR achieves a non-optimal performance on distributions that yields more than pairwise dependencies, since CLR is specifically designed to exploit dependencies among pairs of labels. Despite this, it will be shown that CLR can achieve a poor performance on F-measure and subset 0/1 loss even when considering a distribution with only pairwise dependencies.

For this purpose, let us define a family of probability distributions $\mathcal{P}$ of $n$ labels such that for any $\bar{\mathbf{P}} \in \mathcal{P}$:

$$\bar{\mathbf{P}}(\mathbf{y}) = \bar{\mathbf{P}}(y_1, y_2) \cdot \bar{\mathbf{P}}(y_3) \cdot \bar{\mathbf{P}}(y_4) \ \cdots \ \bar{\mathbf{P}}(y_n) = \bar{\mathbf{P}}(y_1, y_2) \cdot \prod_{i=3}^{n} \bar{\mathbf{P}}(y_i),$$

where the probabilities $\bar{\mathbf{P}}(y_1, y_2)$ and $\bar{\mathbf{P}}(y_i)$ are abbreviations of $\bar{\mathbf{P}}(Y_1 = y_1, Y_2 = y_2)$ and $\bar{\mathbf{P}}(Y_i = y_i)$, respectively. Any $\bar{\mathbf{P}} \in \mathcal{P}$ is constructed such that it can be written as a function of only the joint distribution of two labels and the marginal distributions of the other labels. It is "almost" a distribution of independent variables. Not all probability distributions can be written in this form, because the joint distribution of three or more labels cannot be decomposed generically to the joint probability of only one or two labels. Readers are recommended to check the work of Teugels (1990), if interested in more details about decomposing and understanding the joint probability of a multivariate Bernoulli distribution. In order to show some properties of CLR, let a specific distribution $\bar{\mathbf{P}} \in \mathcal{P}$ be defined such that

$$\bar{\mathbf{P}}(y_1, y_2) = \begin{cases} 3\epsilon, & \text{if } y_1 = y_2 = 0, \\ \epsilon, & \text{if } y_1 = y_2 = 1, \\ \frac{1}{2} - 2\epsilon, & \text{if } y_1 = 1 \text{ and } y_2 = 0, \\ \frac{1}{2} - 2\epsilon, & \text{if } y_1 = 0 \text{ and } y_2 = 1 \end{cases}$$

and $\bar{\mathbf{P}}(Y_i = 1) = \phi_n$ for $i \geq 3$, where $\phi_n$ is a function of $n$ such that:

$$0 \leq \phi_n < \frac{\epsilon}{3n} \quad \text{and} \quad \lim_{n \to \infty} (1 - \phi_n)^n = 1,$$

for all $n \geq 3$. There are many functions satisfying these two conditions of $\phi_n$, for instance, $\phi_n = \epsilon / n^2$. It is crucial to note that if $\epsilon \approx 0$, then $\phi_n \approx 0$ and, consequently, $\bar{\mathbf{P}}$ will have only two labellings ($\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$) being with significant probabilities. Indeed,

$$\bar{\mathbf{P}}(\mathbf{y}^{(1)}) + \bar{\mathbf{P}}(\mathbf{y}^{(2)}) = (1 - 4\epsilon) \cdot (1 - \phi_n)^{(n-2)},$$

which tends to 1 as $\epsilon$ goes to 0. Before stating about the regret of CLR on distribution $\bar{\mathbf{P}}$, it is important to take a look at a property of $\bar{\mathbf{P}}$ stated in Proposition 7.

**Proposition 7** *For distribution $\bar{\mathbf{P}}$ of $n$ labels, CLR will predict $0_n$.*

**Proof** See Appendix F. ■

Theorem 8 shows the regret of CLR with respect to subset 0/1 loss in $\bar{\mathbf{P}}$.

**Theorem 8** *The following expression holds for the regret with respect to subset 0/1 loss*

$$r_s(\mathbf{h}^{clr}) = \left( \frac{1}{2} - 5\epsilon \right) \cdot (1 - \phi_n)^{n-2}, \text{ for distribution } \bar{\mathbf{P}},$$

*and, consequently*

$$\lim_{n \to \infty, \epsilon \to 0} r_s(\mathbf{h}^{clr}) = \frac{1}{2}.$$

**Proof** See Appendix G. ■

One half is a much better regret than 1, but is surely high considering such a simple distribution as $\bar{\mathbf{P}}$. This is further evidence that it is not enough to exploit dependencies to improve performance, even considering only pairwise dependencies. The same argument can be used for the regret with respect to Jaccard distance, as shown in Theorem 9.

**Theorem 9** *The following expression holds for the regret with respect to Jaccard distance*

$$\lim_{\epsilon \to 0} r_J(\mathbf{h}^{clr}) = \frac{1}{2}, \text{ for distribution } \bar{\mathbf{P}}.$$

**Proof** See Appendix H. ■

Theorem 10 shows the regret of CLR with respect to F-measure in $\bar{\mathbf{P}}$.

**Theorem 10** *The following expression holds for the regret with respect to F-measure loss*

$$\lim_{\epsilon \to 0} r_J(\mathbf{h}^{clr}) = \frac{2}{3}, \text{ for distribution } \bar{\mathbf{P}}.$$

**Proof** See Appendix I. ■

Similar to Theorem 8, CLR achieves a poor performance in such a simple distribution, therefore it is not enough to exploit dependencies to improve performance, since $\frac{2}{3}$ is too much for a regret.

Note the independence of both $Y_1$ and $Y_2$ with respect to all other variables $\bar{\mathbf{P}}$. This means that even assuming a low level of dependency among labels, CLR may present a poor performance. In fact, there is only a single dependence, which is between $Y_1$ and $Y_2$.

In addition to being a high regret distribution for CLR, $\bar{\mathbf{P}}$ is the worst case distribution for the regret of the optimal solution for Hamming loss with respect to subset 0/1 loss:

$$\sup_{\mathbf{P}} r_s(\mathbf{h}^*_H) = R_s(\mathbf{h}^*_H) - R_s(\mathbf{h}^*_S) = \frac{1}{2}, \text{ when } \epsilon \to 0$$

where $\mathbf{h}^*_H$ is the optimal solution for Hamming loss on distribution $\bar{\mathbf{P}}$ and $\mathbf{h}^*_s$ for subset 0/1 loss (Dembczyński et al., 2012). Hence, $\bar{\mathbf{P}}$ simultaneously gives poor regret for Hamming loss optimizer and $\mathbf{h}^{\text{clr}}$ with respect to subset 0/1 loss.

## 4. Conclusion

It has been revealed a single factor highly impacting a poor performance of RPC and CLR in a worst case scenario: The adopted pairwise approach. That is, what is supposed to increase performance in average, is the main cause for a poor performance in a worst case scenario. Interestingly, although there exists classifiers achieving a worst-case regret as high as CLR, until now, as far as we know, only CLR was found to have such a high worst-case regret for multiple metrics, and this in the same single distribution. Therefore, it is expected that the results presented in this paper help researchers to be aware of the consequences of using the pairwise comparison approach done by RPC and CLR multi-label problems. The analysis carried out in this paper takes one step closer on understanding the factors causing a good/bad performance in multi-label algorithms.

It is important to know that despite all results against RPC and CLR, RPC was proved to be a risk minimizer for a particular loss function called Spearman rank correlation (Hüllermeier and Fürnkranz, 2004). This loss function is for ranking problems, where the function receives the target ranking and a predicted rank as parameters, while rank loss receives the real labelling instead and a predicted ranking. Therefore, it can be concluded that when comparing multi-label algorithms using multiple multi-label metrics, researches should keep in mind that CLR may occasionally in some datasets present the worst possible regret with respect to subset 0/1 loss, F-measure and Jaccard distance, but this does not mean a useless performance, as it been shown that in the same scenario, RPC presents the optimal expected Spearman rank correlation. Since CLR is an adaptation of an approach which is essentially learning the preference among labels, it is expected that its performance is better in preference learning than in multi-label classification.

We hope that the idea used in this paper of constructing special distributions in which the conditional probabilities used by CLR and RPC are conveniently manipulated, can be extended for finding the supreme regret of other multi-label methods that also use conditional probabilities as a main part of its prediction. Future researches should address others factors that are not considered here, such as the average scenario, instead of the worst one. Further investigations can be done in order to improve the performance of CLR in the worst-case regret such as better choosing the calibration threshold and/or changing the pairwise approach.

## References

F. Chierichetti, R. Kumar, S. Pandey, and S. Vassilvitskii. Finding the Jaccard median. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, page 293–311. Society for Industrial and Applied Mathematics, 2010.

K. Dembczyński, W. Cheng, and E. Hüllermeier. Bayes optimal multilabel classification via probabilistic classifier chains. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, page 279–286. Omnipress, 2010.

K. Dembczyński, W. Kotłowski, and E. Hüllermeier. Consistent multilabel ranking through univariate loss minimization. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, page 1347–1354. Omnipress, 2012.

K. Dembczyński, W. Waegeman, W. Cheng, and E. Hüllermeier. On label dependence and loss minimization in multi-label classification. *Machine Learning*, 88(1):5–45, 2012.

J. Fürnkranz, E. Hüllermeier, E. Loza Mencía, and K. Brinker. Multilabel classification via calibrated label ranking. *Machine Learning*, 73(2):133–153, 2008.

Z.-F. He, M. Yang, H.-D. Liu, and L. Wang. Calibrated multi-label classification with label correlations. *Neural Processing Letters*, 50(2):1361–1380, 2019. doi: 10.1007/s11063-018-9925-2.

M. Huang, F. Zhuang, X. Zhang, X. Ao, Z. Niu, M.-L. Zhang, and Q. He. Supervised representation learning for multi-label classification. *Machine Learning*, 108(5):747–763, 2019.

E. Hüllermeier and J. Fürnkranz. Ranking by pairwise comparison a note on risk minimization. In *2004 IEEE International Conference on Fuzzy Systems*, volume 1, pages 97–102. IEEE, 2004.

E. Montañés, R. Senge, J. Barranquero, J. R. Quevedo, J. J. del Coz, and E. Hüllermeier. Dependent binary relevance models for multi-label classification. *Pattern Recognition*, 47 (3):1494–1508, 2014.

J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part II*, pages 254–269. Springer Berlin Heidelberg, 2009.

L. Sun, H. Ge, and W. Kang. Non-negative matrix factorization based modeling and training algorithm for multi-label learning. *Frontiers of Computer Science*, 13(6):1243–1254, 2019.

M. Tahir, J. Kittler, and A. Bouridane. Multi-label classification using stacked spectral kernel discriminant analysis. *Neurocomputing*, 171:127–137, 2016.

J. L. Teugels. Some representations of the multivariate Bernoulli and binomial distributions. *Journal of Multivariate Analysis*, 32(2):256–268, 1990.

K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas. Multi-label classification of music by emotion. *EURASIP Journal on Audio, Speech, and Music Processing*, 2011(4), 2011.

G. Tsoumakas and I. Katakis. Multi label classification: An overview. *International Journal of Data Warehouse and Mining*, 3:1–13, 2007.

W. Waegeman, K. Dembczyński, A. Jachnik, W. Cheng, and E. Hüllermeier. On the Bayes-optimality of F-measure maximizers. *Journal of Machine Learning Research*, 15 (103):3513–3568, 2014.

S. Wang, J. Wang, Z. Wang, and Q. Ji. Enhancing multi-label classification by modeling dependencies among labels. *Pattern Recognition*, 47(10):3405–3413, 2014.

Z. Younes, F. Abdallah, T. Denoeux, and H. Snoussi. A dependent multilabel classification method derived from the k-nearest neighbor rule. *EURASIP Journal on Advances in Signal Processing*, 2011(645964), 2011.

Y. Zhang and J. Schneider. A composite likelihood view for multi-label classification. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22, pages 1407–1415. PMLR, 2012a.

Y. Zhang and J. Schneider. Maximum margin output coding. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1575–1582. Omnipress, 2012b.

## Appendix A. Proof of Proposition 1

**Proposition 1** *When considering distribution $\hat{\mathbf{P}}_m$, CLR predicts ones for the first $m$ labels and zeroes for the other labels, i.e, $\sum_{i=1}^{m} h_i^{clr} = \sum_{i=1}^{n} h_i^{clr} = m$.*

**Proof** It will be shown that $\mathbf{h}^{\text{clr}}$ satisfies Inequality (9) on distribution $\hat{\mathbf{P}}_m$ if and only if $1 \leq i \leq m$. Firstly, it will be shown that (9) is not satisfied for $i > m$, that is

$$\sum_{j=1}^{n} f(\hat{\mathbf{P}}_m, i, j) + \hat{\mathbf{P}}_m^{(i)} < \sum_{j=1}^{n} \Big(1 - \hat{\mathbf{P}}_m^{(j)}\Big), \text{ for all } i > m. \tag{13}$$

Knowing that

$$\hat{\mathbf{P}}_m^{(i)} = \begin{cases} \hat{\mathbf{P}}_m(1_n) + \hat{\mathbf{P}}_m(\mathbf{y}^{(i)}) = \frac{m+1}{2(n+1)} + \epsilon, & \text{for } i \leq m, \\ \hat{\mathbf{P}}_m(1_n) = \frac{m+1}{2(n+1)}, & \text{for } i > m, \end{cases} \tag{14}$$

the right-hand side of (13) is equivalent to:

$$\sum_{j=1}^{n} \Big(1 - \hat{\mathbf{P}}_m^{(j)}\Big) = n - n \cdot \frac{m+1}{2(n+1)} - m\epsilon. \tag{15}$$

For the left-hand side of (13), and for $1 \leq j \leq m < i \leq n$, it can be observed that

$$\begin{aligned} f(\hat{\mathbf{P}}_m, i, j) &= \frac{\hat{\mathbf{P}}_m(Y_i = 1, Y_j = 0)}{\hat{\mathbf{P}}_m(Y_i = 1, Y_j = 0) + \hat{\mathbf{P}}_m(Y_i = 0, Y_j = 1)} \\ &= \frac{0}{0 + \hat{\mathbf{P}}_m(\mathbf{y}^{(j)})} = \frac{0}{\epsilon} = 0, \qquad \text{for } j \leq m < i. \end{aligned} \tag{16}$$

Using (16) and (15), (13) is equivalent to

$$\sum_{j=m+1}^{n} f(\hat{\mathbf{P}}_m, i, j) + \hat{\mathbf{P}}_m^{(i)} < n - n \cdot \frac{m+1}{2(n+1)} - m\epsilon, \quad \text{for all } i > m.$$

The last inequality is always satisfied, even if the left-hand side assumes an upper bound of $\sum_{j=m+1}^{n} f(\hat{\mathbf{P}}_m, i, j) \leq \sum_{j=m+1:j\neq i}^{n} 1 = n - m - 1$:

$$\begin{aligned} n - m - 1 + \frac{m+1}{2(n+1)} < n - n\frac{m+1}{2(n+1)} - m\epsilon &\iff -m - 1 < -(n+1)\frac{m+1}{2(n+1)} - m\epsilon \\ &\iff 2m + 2 > m + 1 + 2m\epsilon \\ &\iff m + 1 > 2m\epsilon. \end{aligned}$$

The last inequality is satisfied for a sufficiently small $\epsilon$. This concludes the proof for $i > m$.

Now consider $i \leq m$. Let us show that

$$\sum_{j=1}^{n} f(\hat{\mathbf{P}}_m, i, j) + \hat{\mathbf{P}}_m^{(i)} > \sum_{j=1}^{n} \Big(1 - \hat{\mathbf{P}}_m^{(i)}\Big). \tag{17}$$

Firstly, note that, for if $1 \leq i \leq m < j \leq n \longrightarrow f(\hat{\mathbf{P}}_m, i, j) = \frac{\hat{\mathbf{P}}_m(\mathbf{y}^{(i)})}{\hat{\mathbf{P}}_m(\mathbf{y}^{(i)})+\hat{\mathbf{P}}_m(\mathbf{y}^{(j)})} = \frac{\epsilon}{\epsilon+0} = 1$. Moreover, note that $f(\hat{\mathbf{P}}_m, i, j) = \frac{1}{2}$ for any $1 \leq i \leq m$, $1 \leq j \leq m$ and $i \neq j$. Therefore, Inequality (17) is equivalent to

$$\sum_{j=1:j\neq i}^{m} \frac{1}{2} + \sum_{j=m+1}^{n} 1 \; + \; \hat{\mathbf{P}}_m^{(i)} > \sum_{j=1}^{n} \Big(1 - \hat{\mathbf{P}}_m^{(i)}\Big),$$

and then

$$\frac{m-1}{2} + (n-m) + \hat{\mathbf{P}}_m^{(i)} > \sum_{j=1}^{n} \Big(1 - \hat{\mathbf{P}}_m^{(i)}\Big).$$

From (14), it has that $\hat{\mathbf{P}}_m^{(i)} = \frac{m+1}{2(n+1)} + \epsilon$ for all $i \leq m$, so the above inequality is equivalent to

$$\frac{m-1}{2} + n - m + \frac{m+1}{2(n+1)} + \epsilon > \sum_{j=1}^{n} \left(1 - \frac{m+1}{2(n+1)}\right),$$

and then simplifying

$$2n - m - 1 + \frac{m+1}{n+1} + \epsilon > 2n - n \cdot \frac{m+1}{n+1},$$

and again

$$-m - 1 + (n+1)\frac{m+1}{(n+1)} + \epsilon > 0,$$

and finally

$$\epsilon > 0,$$

which is, by definition, always true. ■

## Appendix B. Proof of Theorem 2

**Theorem 2** *The following upper bound holds for the regret with respect to Hamming loss:*

$$\sup_{\mathbf{P}\in\mathcal{P}_n} (r_H(\mathbf{h}^{clr})) = \begin{cases} \frac{n}{4(n+1)}, & \text{if } n \text{ is even} \\ \frac{n-1}{4n}, & \text{if } n \text{ is odd,} \end{cases}$$

*where $\mathcal{P}_n$ denotes the set of all distributions over $n$ labels such that $\mathbf{P}^{(i)} \leq \frac{1}{2}$ for all $i$.*

**Proof** For an arbitrary distribution $\mathbf{P}$ of $\mathbf{Y}$, denote $\mathbf{y}^*$ as the optimal expected Hamming loss for $\mathbf{P}$. The risk of an arbitrary labelling $\hat{\mathbf{y}}$ with respect to Hamming loss can be written as

$$R_H(\hat{\mathbf{y}}) = \frac{1}{n}\sum_{i=1}^{n}(1 - \mathbf{P}(Y_i = \hat{y}_i)), \tag{18}$$

which can be derived from the definition:

$$\begin{aligned}
R_H(\hat{\mathbf{y}}) &= \sum_{\mathbf{y}\in\mathcal{Y}} L_H(\mathbf{y},\hat{\mathbf{y}})\cdot\mathbf{P}(\mathbf{Y}=\mathbf{y}) \\
&= \sum_{\mathbf{y}\in\mathcal{Y}}\left(\frac{1}{n}\sum_{i=1}^{n}[\![y_i\neq\hat{y}_i]\!]\right)\cdot\mathbf{P}(\mathbf{Y}=\mathbf{y}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{\mathbf{y}\in\mathcal{Y}}[\![y_i\neq\hat{y}_i]\!]\cdot\mathbf{P}(\mathbf{Y}=\mathbf{y}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{\mathbf{y}\in\mathcal{Y}:y_i\neq\hat{y}_i}\mathbf{P}(\mathbf{Y}=\mathbf{y}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbf{P}(Y_i\neq\hat{y}_i) = \frac{1}{n}\sum_{i=1}^{n}(1-\mathbf{P}(Y_i=\hat{y}_i)).
\end{aligned}$$

The regret with respect to Hamming loss can be expressed as

$$r_H(\hat{\mathbf{y}}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i^*)(1-2\mathbf{P}^{(i)}), \tag{19}$$

by using the definition of regret and (18):

$$\begin{aligned}
r_H(\hat{\mathbf{y}}) &= R_H(\hat{\mathbf{y}}) - R_H(\mathbf{y}^*) \\
&= \frac{1}{n}\sum_{i=1}^{n}(1-\mathbf{P}(Y_i=\hat{y}_i)) - \frac{1}{n}\sum_{i=1}^{n}(1-\mathbf{P}(Y_i=y_i^*)) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{P}(Y_i=y_i^*) - \mathbf{P}(Y_i=\hat{y}_i)).
\end{aligned}$$

Note that

$$\mathbf{P}(Y_i=y_i^*) - \mathbf{P}(Y_i=\hat{y}_i) = \begin{cases} 0, & \text{if } y_i^*=\hat{y}_i, \\ \mathbf{P}(Y_i=0)-\mathbf{P}(Y_i=1), & \text{if } y_i^*=1 \text{ and } \hat{y}_i=0, \\ \mathbf{P}(Y_i=1)-\mathbf{P}(Y_i=0), & \text{if } y_i^*=0 \text{ and } \hat{y}_i=1. \end{cases}$$

Therefore

$$\begin{aligned}
r_H(\hat{\mathbf{y}}) &= \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i^*)(\mathbf{P}(Y_i=0)-\mathbf{P}(Y_i=1)) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i^*)(1-2\mathbf{P}^{(i)}).
\end{aligned}$$

Said that, define $A = \{i : y_i^* = 0 \wedge h_i^{\text{clr}} = 1\}$, i.e the set of all false positive labels, and $a = |A|$. From Equation (6), it is easy to see that $\mathbf{y}^* = 0_n$ for all distributions in $\mathcal{P}_n$, so

$$\begin{aligned} r_H(\mathbf{h}^{\text{clr}}) &= \frac{1}{n}\sum_{i=1}^{n} h_i^{\text{clr}}(1 - 2\mathbf{P}^{(i)}) \\ &= \frac{1}{n}\left(a - 2\sum_{i\in A}\mathbf{P}^{(i)}\right) \end{aligned} \tag{20}$$

In the next steps, we will find a lower bound for $\sum_{i\in A}\mathbf{P}^{(i)}$, consequently giving an upper bound for $r_H(\mathbf{h}^{\text{clr}})$. Summing the scores $\sum_{i\in A} s_i$, defined in (12), results in:

$$\begin{aligned} \sum_{i\in A} s_i &= \sum_{i\in A}\sum_{j=1}^{n} f(\mathbf{P}, i, j) + \sum_{i\in A}\mathbf{P}^{(i)} \\ &= \sum_{i\in A}\sum_{j\in A} f(\mathbf{P}, i, j) + \sum_{i\in A}\sum_{j\notin A} f(\mathbf{P}, i, j) + \sum_{i\in A}\mathbf{P}^{(i)} \end{aligned}$$

Knowing that $f(\mathbf{P}, i, j) + f(\mathbf{P}, j, i) = 1$ for any $i \neq j$, we have that $\sum_{i\in A}\sum_{j\in A} f(\mathbf{P}, i, j) = \frac{a(a-1)}{2}$, therefore

$$\sum_{i\in A} s_i = \frac{a(a-1)}{2} + \sum_{i\in A}\sum_{j\notin A} f(\mathbf{P}, i, j) + \sum_{i\in A}\mathbf{P}^{(i)}. \tag{21}$$

Using the upper bound $f(\mathbf{P}, i, j) \leq 1 - \mathbf{P}(Y_i = 0, Y_j = 1)$ for any $i$ and $j$, it can be shown that

$$\begin{aligned} \sum_{i\in A}\sum_{j\notin A} f(\mathbf{P}, i, j) &\leq \sum_{i\in A}\sum_{j\notin A}(1 - \mathbf{P}(Y_i = 0, Y_j = 1)) \\ &= \sum_{i\in A}\sum_{j\notin A}(1 + \mathbf{P}(Y_i = 1, Y_j = 1) - \mathbf{P}^{(j)}) \\ &\leq \sum_{i\in A}\sum_{j\notin A}(1 + \mathbf{P}^{(i)} - \mathbf{P}^{(j)}) \\ &= a(n-a) + \sum_{i\in A}\sum_{j\notin A}(\mathbf{P}^{(i)} - \mathbf{P}^{(j)}) \\ &= a(n-a) + (n-a)\sum_{i\in A}\mathbf{P}^{(i)} \;-\; a\sum_{j\notin A}\mathbf{P}^{(j)}. \end{aligned} \tag{22}$$

Using the upper bound at (22) on Equation (21):

$$\begin{aligned} \sum_{i\in A} s_i &\leq \frac{a(a-1)}{2} + a(n-a) + (n-a)\sum_{i\in A}\mathbf{P}^{(i)} \;-\; a\sum_{j\notin A}\mathbf{P}^{(j)} + \sum_{i\in A}\mathbf{P}^{(i)} \\ &= \frac{a(a-1)}{2} + a(n-a) + (n-a+1)\sum_{i\in A}\mathbf{P}^{(i)} \;-\; a\sum_{j\notin A}\mathbf{P}^{(j)} \\ &= -\frac{a(a+1)}{2} + an + (n-a+1)\sum_{i\in A}\mathbf{P}^{(i)} \;-\; a\sum_{j\notin A}\mathbf{P}^{(j)}. \end{aligned} \tag{23}$$

By definition of CLR, it is true that $\sum_{i\in A} s_i \geq a\Big(n - \sum_{j=1}^{n} \mathbf{P}^{(j)}\Big)$, which is the sum of all conditions associated to false positive labels. Applying the upper bound in (23) on it:

$$-\frac{a(a+1)}{2} + an + (n-a+1)\sum_{i\in A}\mathbf{P}^{(i)} \;-\; a\sum_{j\notin A}\mathbf{P}^{(j)} \geq a\left(n - \sum_{j=1}^{n}\mathbf{P}^{(j)}\right).$$

Note that $an$ is present on both sides, so it can be simplified to

$$-\frac{a(a+1)}{2} + (n-a+1)\sum_{i\in A}\mathbf{P}^{(i)} \;-\; a\sum_{j\notin A}\mathbf{P}^{(j)} \geq -a\sum_{j=1}^{n}\mathbf{P}^{(j)}.$$

Also note that $\sum_{j=1}^{n}\mathbf{P}^{(j)} = \sum_{j\in A}^{n}\mathbf{P}^{(j)} + \sum_{j\notin A}^{n}\mathbf{P}^{(j)}$, so the inequality is again simplified to

$$-\frac{a(a+1)}{2} + (n-a+1)\sum_{i\in A}\mathbf{P}^{(i)} \geq -a\sum_{j\in A}^{n}\mathbf{P}^{(j)},$$

which is equivalent to

$$(n+1)\sum_{i\in A}\mathbf{P}^{(i)} \geq \frac{a(a+1)}{2},$$

and finally

$$\sum_{i\in A}\mathbf{P}^{(i)} \geq \frac{a(a+1)}{2(n+1)}.$$

Using this last lower bound for $\sum_{i\in A}\mathbf{P}^{(i)}$ at Inequality (20), it can be derived an upper bound for $r_H(\mathbf{h}^{\text{clr}})$:

$$\begin{aligned}
r_H(\mathbf{h}^{\text{clr}}) &\leq \frac{1}{n}\left(a - 2\sum_{i\in A}\mathbf{P}^{(i)}\right)\\
&\leq \frac{1}{n}\left(a - 2\frac{a(a+1)}{2(n+1)}\right)\\
&= \frac{a}{n}\left(1 - \frac{a+1}{n+1}\right)\\
&= \frac{a}{n}\left(\frac{n-1+a+1}{n+1}\right)\\
&= \frac{a(n-a)}{n(n+1)},
\end{aligned}$$

which is a quadratic polynomial with respect to $a$ that clearly has a maximum when $a = \frac{n}{2}$, if $n$ is even. Therefore

$$r_H(\mathbf{h}^{\text{clr}}) \leq \frac{n}{4(n+1)}.$$

If $n$ is odd, the maximum is given when $a = \frac{n-1}{2}$ or $a = \frac{n+1}{2}$.

To show that this bound is tight, it just needs to be shown the existence of a distribution that yields a regret arbitrarily as close to the value above. Distribution $\hat{\mathbf{P}}_m$ for $m = \frac{n}{2}$ satisfies this condition. Given that $\hat{\mathbf{P}}^{(i)}_{n/2} = \frac{n+2}{4(n+1)} + \epsilon < \frac{1}{2}$ for any $i$, the optimal labelling for Hamming loss is $0_n$. Therefore, the value of $r_H(\mathbf{h}^{\text{clr}})$ on distribution $\hat{\mathbf{P}}_{n/2}$ is given by

$$\begin{aligned} r_H(\mathbf{h}^{\text{clr}}) =& \frac{1}{n}\sum_{i\in A}\left(1 - 2\hat{\mathbf{P}}^{(i)}_{n/2}\right) \\ =& \frac{1}{n}\sum_{i\in A}\left(1 - \frac{n+2}{2(n+1)} - \epsilon\right), \end{aligned}$$

and then, using Proposition 1,

$$\begin{aligned} r_H(\mathbf{h}^{\text{clr}}) =& \frac{n}{2n}\left(1 - \frac{n+2}{2(n+1)} - \epsilon\right) \\ =& \frac{1}{2}\left(\frac{2n+2-n-2}{2(n+1)} - \epsilon\right) \\ =& \frac{n}{4(n+1)} - \frac{\epsilon}{2}. \end{aligned}$$

■

## Appendix C. Proof of Theorem 4

**Theorem 4** *The following lower bound holds for the regret with respect to Jaccard distance:*

$$\sup_{\mathbf{P}} r_J(\mathbf{h}^{clr}) \geq 1 - \frac{1}{n}.$$

**Proof** Consider distribution $\hat{\mathbf{P}}_1$ and note that there are only three labellings with a non-null probability: $0_n$, $1_n$ and $\mathbf{y}^{(1)}$. From Proposition 1, it has that $\mathbf{h}^{\text{clr}} = \mathbf{y}^{(1)}$ for distribution $\hat{\mathbf{P}}_1$. Given that the loss $L_J(1_n, \mathbf{h}^{\text{clr}})$ can be calculated as the following

$$L_J(1_n, \mathbf{h}^{\text{clr}}) = L_J(1_n, \mathbf{y}^{(1)}) = 1 - \frac{\sum_{i=1}^n h_i^{\text{clr}}}{n + \sum_{i=1}^n h_i^{\text{clr}} - \sum_{i=1}^n h_i^{\text{clr}}} = 1 - \frac{1}{n},$$

the risk of CLR is

$$\begin{aligned} R_J(\mathbf{h}^{\text{clr}}) = R_J(\mathbf{y}^{(1)}) =& \sum_{\mathbf{y}\in\{0_n, 1_n, \mathbf{y}^{(1)}\}} L_J(\mathbf{y}, \mathbf{y}^{(1)})\hat{\mathbf{P}}_1(\mathbf{y}) \\ =& \underbrace{L_J(0_n, \mathbf{y}^{(1)})}_{1}\ \underbrace{\hat{\mathbf{P}}_1(0_n)}_{1-1/(n+1)-\epsilon} + \underbrace{L_J(1_n, \mathbf{y}^{(1)})}_{1-1/n}\underbrace{\hat{\mathbf{P}}_1(1_n)}_{1/(n+1)} + \underbrace{L_J(\mathbf{y}^{(1)}, \mathbf{y}^{(1)})}_{0}\hat{\mathbf{P}}_1(\mathbf{y}^{(1)}) \\ =& 1 - \frac{1}{n+1} - \epsilon + \frac{1}{n+1} - \frac{1}{n(n+1)} \\ =& 1 - \epsilon - \frac{1}{n(n+1)}. \end{aligned}$$

The risk of the optimal solution $\mathbf{y}^*$ is upper bounded by

$$\begin{aligned} R_J(\mathbf{y}^*) \leq R_J(0_n) &= \sum_{\mathbf{y}\in\{0_n,1_n,\mathbf{y}^{(1)}\}} L_J(\mathbf{y},0_n)\hat{\mathbf{P}}_1(\mathbf{y}) \\ &= \underbrace{L_J(0_n,0_n)}_{0}\hat{\mathbf{P}}_1(0_n) + \underbrace{L_J(1_n,0_n)}_{1}\underbrace{\hat{\mathbf{P}}_1(1_n)}_{1/(n+1)} + \underbrace{L_J(\mathbf{y}^{(1)},0_n)}_{\leq 1}\underbrace{\hat{\mathbf{P}}_1(\mathbf{y}^{(1)})}_{\epsilon} \\ &\leq \frac{1}{n+1} + \epsilon. \end{aligned}$$

The regret is lower bounded by

$$\begin{aligned} r_J(\mathbf{h}^{\text{clr}}) &\geq \overbrace{1-\epsilon-\frac{1}{n(n+1)}}^{R_J(\mathbf{h}^{\text{clr}})} - \overbrace{\left(\frac{1}{n+1}+\epsilon\right)}^{R_J(0_n)} \\ &= 1-2\epsilon-\frac{n+1}{n(n+1)} \\ &= 1-2\epsilon-\frac{1}{n}. \end{aligned}$$

The value of $\epsilon$ can be made arbitrarily small, so

$$\sup r_J(\mathbf{h}^{\text{clr}}) \geq 1-\frac{1}{n}.$$

■

## Appendix D. Proof of Theorem 5

**Theorem 5** *The following lower bound holds for the regret with respect to F-measure:*

$$\sup_{\mathbf{P}} r_F(\mathbf{h}^{clr}) \geq 1-\frac{n+3}{(n+1)^2}.$$

**Proof** Consider distribution $\hat{\mathbf{P}}_1$ and note that there are only three labellings with a non-null probability: $0_n$, $1_n$ and $\mathbf{y}^{(1)}$. This proof is very similar to Theorem 4. From Proposition 1, it has that $\mathbf{h}^{\text{clr}} = \mathbf{y}^{(1)}$ for distribution $\hat{\mathbf{P}}_1$. Given that the loss $L_F(1_n, \mathbf{h}^{\text{clr}})$ can be calculated as the following

$$L_F(1_n,\mathbf{h}^{\text{clr}}) = L_F(1_n,\mathbf{y}^{(1)}) = 1-\frac{2\sum_{i=1}^n h_i^{\text{clr}}}{n+\sum_{i=1}^n h_i^{\text{clr}}} = 1-\frac{2}{n+1}.$$

the risk of CLR is

$$\begin{aligned}
R_F(\mathbf{h}^{\mathrm{clr}}) = R_F(\mathbf{y}^{(1)}) &= \sum_{\mathbf{y}\in\{0_n,1_n,\mathbf{y}^{(1)}\}} L_F(\mathbf{y},\mathbf{y}^{(1)})\hat{\mathbf{P}}_1(\mathbf{y}) \\
&= \underbrace{L_F(0_n,\mathbf{y}^{(1)})}_{1}\ \underbrace{\hat{\mathbf{P}}_1(0_n)}_{1-1/(n+1)-\epsilon} + \underbrace{L_F(1_n,\mathbf{y}^{(1)})}_{1-2/(n+1)}\underbrace{\hat{\mathbf{P}}_1(1_n)}_{1/(n+1)} + \underbrace{L_F(\mathbf{y}^{(1)},\mathbf{y}^{(1)})}_{0}\hat{\mathbf{P}}_1(\mathbf{y}^{(1)}) \\
&= 1 - \frac{1}{n+1} - \epsilon + \frac{1}{n+1} - \frac{2}{(n+1)^2} \\
&= 1 - \epsilon - \frac{2}{(n+1)^2}.
\end{aligned}$$

The risk of the optimal solution $\mathbf{y}^*$ is upper bounded by

$$\begin{aligned}
R_F(\mathbf{y}^*) \leq R_F(0_n) &= \sum_{\mathbf{y}\in\{0_n,1_n,\mathbf{y}^{(1)}\}} L_F(\mathbf{y},0_n)\hat{\mathbf{P}}_1(\mathbf{y}) \\
&= \underbrace{L_F(0_n,0_n)}_{0}\hat{\mathbf{P}}_1(0_n) + \underbrace{L_F(1_n,0_n)}_{1}\underbrace{\hat{\mathbf{P}}_1(1_n)}_{1/(n+1)} + \underbrace{L_F(\mathbf{y}^{(1)},0_n)}_{\leq 1}\underbrace{\hat{\mathbf{P}}_1(\mathbf{y}^{(1)})}_{\epsilon} \\
&\leq \frac{1}{n+1} + \epsilon.
\end{aligned}$$

The regret is lower bounded by

$$\begin{aligned}
r_F(\mathbf{h}^{\mathrm{clr}}) &\geq \overbrace{1 - \epsilon - \frac{2}{(n+1)^2}}^{R_F(\mathbf{h}^{\mathrm{clr}})} - \overbrace{\left(\frac{1}{n+1} + \epsilon\right)}^{R_F(0_n)} \\
&= 1 - 2\epsilon - \frac{n+3}{(n+1)^2} \\
&= 1 - 2\epsilon - \frac{n+3}{(n+1)^2}.
\end{aligned}$$

The value of $\epsilon$ can be made arbitrarily small, so

$$\sup r_F(\mathbf{h}^{\mathrm{clr}}) \geq 1 - \frac{n+3}{(n+1)^2}.$$

■

## Appendix E. Proof of Theorem 6

**Theorem 6** *For any $n$ divisible by 4, the following lower bound holds for the regret with respect to normalized rank loss:*

$$\sup_{\mathbf{P}} r_{\hat{r}}(\mathbf{h}^{rpc}) \geq \frac{1}{6}.$$

**Proof** The proof is given by showing that a specific probability label distribution $\widetilde{\mathbf{P}}$ gives a regret of exactly $\frac{1}{6}$ for any $n$ divisible by 4. Before defining $\widetilde{\mathbf{P}}$, let three disjoint sets of labels, $A$, $B$ and $C$ be defined as following (note that we are using integers to represent labels):

$$A = \{i \in \mathbb{Z} \mid 1 \leq i \leq \frac{n}{4}\},$$
$$B = \{i \in \mathbb{Z} \mid \frac{n}{4} < i \leq \frac{n}{2}\},$$
$$C = \{i \in \mathbb{Z} \mid \frac{n}{2} < i \leq n\}.$$

Distribution $\widetilde{\mathbf{P}}$ is defined as

$$\widetilde{\mathbf{P}}(\mathbf{y}) = \begin{cases} \frac{3}{4} - n \cdot \epsilon, & \text{if all labels in } A \text{ are positive and all other labels are negative,} \\ \frac{1}{4}, & \text{if all labels in } A \text{ are negative and all other labels are positive,} \\ 2\epsilon, & \text{if exactly one label in } A \text{ is positive and all other labels are negative} \\ 2\epsilon, & \text{if exactly one label in } B \text{ is positive and all other labels are negative} \\ 0, & \text{otherwise} \end{cases}$$

where $\epsilon$ is an arbitrary positive real number that is assumed to be "really close" to 0. The purpose of $\epsilon$ in $\widetilde{\mathbf{P}}$ is identical to the purpose of $\epsilon$ in distribution $\hat{\mathbf{P}}_m$, which is to avoid undefined value for $f(\mathbf{P}, i, j)$ when the numerator and denominator are both null and to make $f(\mathbf{P}, i, j)$ be convenient values such as 1 or $\frac{1}{2}$.

It will be shown that RPC prefers any label in $B$ to any label in $A$. Consider an arbitrary pair of labels $(i, j)$ where $i \in A$ and $j \in B$. Let's check that RPC prefers label $j$ to $i$ by checking which score $s_i$ or $s_j$ is higher:

$$\sum_{k=1}^{n} f(\widetilde{\mathbf{P}}, j, k) - \sum_{k=1}^{n} f(\widetilde{\mathbf{P}}, i, k) > 0 \ ? \tag{24}$$

If the difference above ($s_j$-$s_i$) is positive, then RPC prefers label $j$ to label $i$. The distribution $\widetilde{\mathbf{P}}$ has so few non-null values that it is easy to check, for all $i \in A$, that:

$$f(\widetilde{\mathbf{P}}, i, k) = \begin{cases} \frac{3/4 - n\epsilon + 2\epsilon}{1 - (n-4)\epsilon}, & \text{if } k \in B, \\ \frac{3/4 - n\epsilon + 2\epsilon}{1 - (n-2)\epsilon}, & \text{if } k \in C, \\ \frac{1}{2}, & \text{if } k \in A \ \wedge \ k \neq i. \end{cases}$$

For all $j \in B$, it can be also checked that

$$f(\widetilde{\mathbf{P}}, j, k) = \begin{cases} \frac{1}{2}, & \text{if } k \in B \ \wedge \ k \neq j, \\ 1, & \text{if } k \in C, \\ \frac{2\epsilon + 1/4}{1 - (n-4)\epsilon}, & \text{if } k \in A \end{cases}$$

Therefore, the score $s_i$ is rewritten as:

$$s_i = \sum_{k=1}^{n} f(\widetilde{\mathbf{P}}, i, k) = \frac{|A| - 1}{2} + |B| \frac{3/4 - n\epsilon + 2\epsilon}{1 - (n-4)\epsilon} + |C| \frac{3/4 - n\epsilon + 2\epsilon}{1 - (n-2)\epsilon}$$

Analogously, the score $s_j$ is rewritten as:

$$s_j = \sum_{k=1}^{n} f(\widetilde{\mathbf{P}}, j, k) = |A|\frac{2\epsilon + 1/4}{1-(n-4)\epsilon} + \frac{|B|-1}{2} + |C|.$$

Note that $|A| = |B|$ so $\frac{|B|-1}{2}$ cancels out with $\frac{|A|-1}{2}$ on the difference $s_j - s_i$. Therefore the difference can be simplified to:

$$s_j - s_i = \left(|A|\frac{2\epsilon + 1/4}{1-(n-4)\epsilon} + |C|\right) - \left(|B|\frac{3/4 - n\epsilon + 2\epsilon}{1-(n-4)\epsilon} + |C|\frac{3/4 - n\epsilon + 2\epsilon}{1-(n-2)\epsilon}\right)$$

Given that $\frac{3/4-n\epsilon+2\epsilon}{1-(n-2)\epsilon} \leq \frac{3}{4}$ and $\frac{2\epsilon+1/4}{1-(n-4)\epsilon} \geq 2\epsilon + \frac{1}{4}$, a lower bound for $s_j - s_i$ can be found:

$$s_j - s_i \geq |A|\left(2\epsilon + \frac{1}{4}\right) + |C| - \left(\frac{3|B|}{4} + \frac{3|C|}{4}\right).$$

It will be shown that this lower bound is positive. Given that $2|A| = |C| = 2|B|$, it follows that

$$\begin{aligned} s_j - s_i &\geq |A|\left(2\epsilon + \frac{1}{4}\right) + 2|A| - \frac{3}{4}\cdot 3|A| \\ &= |A| \cdot 2\epsilon. \end{aligned}$$

The value $|A|\cdot 2\epsilon$ is always positive since $\epsilon > 0$ by definition. Therefore, it can be concluded that RPC prefers any label $j \in B$ to any label $i \in A$.

Instead of calculating the regret of the prediction of RPC ($\mathbf{h}^{\text{rpc}}$) on distribution $\widetilde{\mathbf{P}}$, let us calculate the regret of the same prediction $\mathbf{h}^{\text{rpc}}$, but on a new distribution $\widetilde{\mathbf{P}}_0$, which is defined in the same way as $\widetilde{\mathbf{P}}$, but with $\epsilon$ being zero. It will be shown that $|r_{\hat{r}}(\mathbf{h}^{\text{rpc}}, \widetilde{\mathbf{P}}_0) - r_{\hat{r}}(\mathbf{h}^{\text{rpc}}, \widetilde{\mathbf{P}})| \leq n\epsilon 2^{n+1}$, where we are now using the notation where the probability distribution is an explicit parameter of the regret to avoid any confusion later. Although this upper bound seems a bit high, it is a multiple of $\epsilon$, which can be arbitrarily made small. So when $\epsilon$ tends to zero, this difference also tends to zero. Note that we do not use $\widetilde{\mathbf{P}}_0$ from the beginning, because RPC prediction on $\widetilde{\mathbf{P}}_0$ is undefined. Observe that these two distributions slightly differ: $|\widetilde{\mathbf{P}}_0(\mathbf{y}) - \widetilde{\mathbf{P}}(\mathbf{y})| \leq n\epsilon$ for all $\mathbf{y}$. For any arbitrary ranking $\mathbf{z}$,

$$\begin{aligned} R_{\hat{r}}(\mathbf{z}, \widetilde{\mathbf{P}}_0) - R_{\hat{r}}(\mathbf{z}, \widetilde{\mathbf{P}}) &= \sum_{\mathbf{y}} L_{\hat{r}}(\mathbf{y}, \mathbf{z}) \underbrace{(\widetilde{\mathbf{P}}_0(\mathbf{y}) - \widetilde{\mathbf{P}}(\mathbf{y}))}_{\leq n\epsilon} \\ &\leq n\epsilon \sum_{\mathbf{y}} L_{\hat{r}}(\mathbf{y}, \mathbf{z}) = n\epsilon 2^n. \end{aligned}$$

The difference $|r_{\hat{r}}(\mathbf{h}^{\text{rpc}}, \widetilde{\mathbf{P}}_0) - r_{\hat{r}}(\mathbf{h}^{\text{rpc}}, \widetilde{\mathbf{P}})|$ can not differ by twice of the above amount, since the regret is the difference of two risks.

$$\begin{aligned} r_{\hat{r}}(\mathbf{h}^{\text{rpc}}, \widetilde{\mathbf{P}}_0) - r_{\hat{r}}(\mathbf{h}^{\text{rpc}}, \widetilde{\mathbf{P}}) &= R_{\hat{r}}(\mathbf{h}^{\text{rpc}}, \widetilde{\mathbf{P}}_0) - R_{\hat{r}}(\mathbf{z}_0^*, \widetilde{\mathbf{P}}_0) - \left(R_{\hat{r}}(\mathbf{h}^{\text{rpc}}, \widetilde{\mathbf{P}}) - R_{\hat{r}}(\mathbf{z}^*, \widetilde{\mathbf{P}})\right) \\ &\leq R_{\hat{r}}(\mathbf{h}^{\text{rpc}}, \widetilde{\mathbf{P}}_0) - R_{\hat{r}}(\mathbf{z}_0^*, \widetilde{\mathbf{P}}_0) - \left(R_{\hat{r}}(\mathbf{h}^{\text{rpc}}, \widetilde{\mathbf{P}}) - R_{\hat{r}}(\mathbf{z}_0^*, \widetilde{\mathbf{P}})\right) \\ &= \underbrace{R_{\hat{r}}(\mathbf{h}^{\text{rpc}}, \widetilde{\mathbf{P}}_0) - R_{\hat{r}}(\mathbf{h}^{\text{rpc}}, \widetilde{\mathbf{P}})}_{\leq n\epsilon 2^n} + \underbrace{R_{\hat{r}}(\mathbf{h}^{\text{rpc}}, \widetilde{\mathbf{P}}) - R_{\hat{r}}(\mathbf{z}_0^*, \widetilde{\mathbf{P}}_0)}_{\leq n\epsilon 2^n} \\ &\leq n\epsilon 2^{n+1}. \end{aligned}$$

This can be done similarly with $r_{\hat{r}}(\mathbf{h}^{\text{rpc}}, \widetilde{\mathbf{P}}) - r_{\hat{r}}(\mathbf{h}^{\text{rpc}}, \widetilde{\mathbf{P}}_0)$, so

$$|r_{\hat{r}}(\mathbf{h}^{\text{rpc}}, \widetilde{\mathbf{P}}_0) - r_{\hat{r}}(\mathbf{h}^{\text{rpc}}, \widetilde{\mathbf{P}})| \leq n\epsilon 2^{n+1}. \tag{25}$$

To calculate the regret, it necessary to know what is the optimal solution for $\widetilde{\mathbf{P}}_0$. Observe that

$$s_{\mathbf{y}}(n - s_{\mathbf{y}}) = \frac{n}{4} \cdot \frac{3n}{4} = \frac{3n^2}{16}, \tag{26}$$

where $s_{\mathbf{y}} = \sum y_i$, for all $\mathbf{y}$ such that $\widetilde{\mathbf{P}}_0(\mathbf{y}) > 0$. Hence, the optimal solution for normalized rank loss in this distribution is exactly the same of rank loss, as observed in Equation (8). To show the optimizer for rank loss prefers labels in $A$ to labels in $B$, it just has to be shown that $\widetilde{\mathbf{P}}(Y_i = 1) - \widetilde{\mathbf{P}}(Y_j = 1) > 0$, for all $i \in A$ and all $j \in B$:

$$\widetilde{\mathbf{P}}_0(Y_i = 1) - \widetilde{\mathbf{P}}_0(Y_j = 1) = \frac{3}{4} - \frac{1}{4} = \frac{1}{2}. \tag{27}$$

Hence, it can be concluded that the optimizer for rank loss prefers labels from $A$.

So RPC makes at least $|A| \cdot |B| = \frac{n^2}{16}$ misorder. The regret given by each of these mistakes, as defined in Equation (7), are all equal and given by $\widetilde{\mathbf{P}}_0(Y_i = 1) - \widetilde{\mathbf{P}}_0(Y_j = 1)$ for $i \in A$ and $j \in B$. From Equation (27), we have that $\widetilde{\mathbf{P}}_0(Y_i = 1) - \widetilde{\mathbf{P}}_0(Y_j = 1) = \frac{1}{2}$. From Equation (7), the regret $r_{\hat{R}}(\mathbf{h}^{RPC})$ on $\widetilde{\mathbf{P}}_0$ is given by multiplying the number of misorder ($\frac{n^2}{16}$) by $\frac{1}{2}$ and dividing by the constant normalization factor of Equation (26):

$$r_{\hat{r}}(\mathbf{h}^{\text{rpc}}, \widetilde{\mathbf{P}}_0) = \frac{n^2}{16} \cdot \frac{1}{2} \cdot \frac{16}{3n^2} = \frac{1}{6}.$$

From the equation above and from (25), the regret $r_{\hat{r}}(\mathbf{h}^{\text{rpc}}, \widetilde{\mathbf{P}})$ differs from 1/6 only by a multiple of $\epsilon$. Since $\epsilon$ can be arbitrarily small, the supreme of $r_{\hat{r}}(\mathbf{h}^{\text{rpc}}, \widetilde{\mathbf{P}})$ is at least $\frac{1}{6}$. ■

## Appendix F. Proof of Proposition 7

**Proposition 7** *For distribution $\bar{\mathbf{P}}$ of $n$ labels, CLR will predict $0_n$.*

**Proof** It will be shown that $s_1 = s_2 < \sum_i (1 - \bar{\mathbf{P}}^{(i)})$ and $s_3 = s_4 = ... = s_n < \sum_i (1 - \bar{\mathbf{P}}^{(i)})$ (see (12)). Firstly, calculate $\sum_i (1 - \bar{\mathbf{P}}^{(i)})$. Knowing that

$$\bar{\mathbf{P}}^{(1)} = \underbrace{\bar{\mathbf{P}}(Y_1 = 1, Y_2 = 0)}_{1/2 - 2\epsilon} + \underbrace{\bar{\mathbf{P}}(Y_1 = 1, Y_2 = 1)}_{\epsilon} = \frac{1}{2} - \epsilon$$

$$\bar{\mathbf{P}}^{(2)} = \underbrace{\bar{\mathbf{P}}(Y_1 = 0, Y_2 = 1)}_{1/2 - 2\epsilon} + \underbrace{\bar{\mathbf{P}}(Y_1 = 1, Y_2 = 1)}_{\epsilon} = \frac{1}{2} - \epsilon,$$

it has that

$$\begin{aligned}\sum_{i=1}^{n}(1-\bar{\mathbf{P}}^{(i)}) &= n - \underbrace{\bar{\mathbf{P}}^{(1)}}_{1/2-\epsilon} - \underbrace{\bar{\mathbf{P}}^{(2)}}_{1/2-\epsilon} - \sum_{i=3}^{n}\underbrace{\bar{\mathbf{P}}^{(i)}}_{\phi_n} \\ &= n - 1 + 2\epsilon - (n-2)\phi_n.\end{aligned} \tag{28}$$

Now, it will be shown that $s_1 \leq n - 1 - \epsilon < \sum_i(1-\bar{\mathbf{P}}^{(i)})$. Before that, note that $\bar{\mathbf{P}}(Y_1 = 1, Y_2 = 0) = \bar{\mathbf{P}}(Y_1 = 0, Y_2 = 1)$, implying that $f(\bar{\mathbf{P}}, 1, 2) = f(\bar{\mathbf{P}}, 2, 1) = 1/2$. Analogously, for any pair of labels $i, j \geq 3$ and $i \neq j$, it has that $f(\bar{\mathbf{P}}, i, j) = f(\bar{\mathbf{P}}, j, i) = 1/2$. Said that, an upper bound for $s_1$ is

$$\begin{aligned}s_1 &= f(\bar{\mathbf{P}}, 1, 2) + \underbrace{\bar{\mathbf{P}}^{(1)}}_{1/2-\epsilon} + \sum_{j=3}^{n} f(\bar{\mathbf{P}}, 1, j) \\ &= \frac{1}{2} + \frac{1}{2} - \epsilon + \sum_{j=3}^{n}\underbrace{f(\bar{\mathbf{P}}, 1, j)}_{\leq 1} \\ &\leq 1 - \epsilon + n - 2 \;=\; n - 1 - \epsilon,\end{aligned}$$

and $n - 1 - \epsilon$ is lesser than $\sum_i(1-\bar{\mathbf{P}}^{(i)})$, because their difference is negative:

$$\begin{aligned}(n-1-\epsilon) - \sum_i(1-\bar{\mathbf{P}}^{(i)}) &= (n-1-\epsilon) - (n-1+2\epsilon-(n-2)\phi_n) && \text{From Equation (28)} \\ &= -3\epsilon + (n-2)\phi_n < 0. && \text{By definition } \phi_n < \frac{\epsilon}{n}.\end{aligned}$$

It will be shown that $s_3 = s_4 = ... = s_n \leq \sum_i(1-\bar{\mathbf{P}}^{(i)})$. An upper bound for $s_3$ is given by

$$\begin{aligned}s_3 &= f(\bar{\mathbf{P}}, 3, 1) + f(\bar{\mathbf{P}}, 3, 2) + \underbrace{\bar{\mathbf{P}}^{(3)}}_{\phi_n} + \sum_{j=4}^{n}\underbrace{f(\bar{\mathbf{P}}, 3, j)}_{1/2} \\ &= f(\bar{\mathbf{P}}, 3, 1) + f(\bar{\mathbf{P}}, 3, 2) + \phi_n + \frac{n-3}{2} \\ &\leq 2 + \phi_n + \frac{n-3}{2} \;=\; \phi_n + \frac{n+1}{2}.\end{aligned}$$

It is easy to see that $s_3 \leq \phi_n + \frac{n+1}{2} \leq n-1+2\epsilon-(n-2)\phi_n$, for a sufficiently large $n$ ($n \geq 3$). ■

## Appendix G. Proof of Theorem 8

**Theorem 8** *The following expression holds for the regret with respect to subset 0/1 loss*

$$r_s(\mathbf{h}^{clr}) = \left(\frac{1}{2} - 5\epsilon\right) \cdot (1-\phi_n)^{n-2}, \text{ for distribution } \bar{\mathbf{P}},$$

*and, consequently*

$$\lim_{n\to\infty, \epsilon\to 0} r_s(\mathbf{h}^{clr}) = \frac{1}{2}.$$

**Proof** Clearly, the mode of $\bar{\mathbf{P}}$ is either $\mathbf{y}^{(1)}$ or $\mathbf{y}^{(2)}$. In both cases, the risk is the same:

$$\begin{aligned} R_s(\mathbf{y}^{(1)}) = 1 - \bar{\mathbf{P}}(\mathbf{y}^{(1)}) &= 1 - \bar{\mathbf{P}}(1,0) \cdot \left(1 - \bar{\mathbf{P}}^{(3)}\right)^{n-2} \\ &= 1 - \left(\frac{1}{2} - 2\epsilon\right) \cdot (1 - \phi_n)^{n-2}. \end{aligned}$$

The risk of CLR is given by

$$\begin{aligned} R_s(\mathbf{h}^{\text{clr}}) = R_s(0_n) &= 1 - \bar{\mathbf{P}}(0,0) \cdot \left(1 - \bar{\mathbf{P}}^{(3)}\right)^{n-2} \quad \text{From Proposition 7} \\ &= 1 - 3\epsilon \cdot (1 - \phi_n)^{n-2}. \end{aligned}$$

And finally, the regret is

$$\begin{aligned} r_s(\mathbf{h}^{\text{clr}}) &= R_s(\mathbf{h}^{\text{clr}}) - R_s(\mathbf{y}^{(1)}) \\ &= \left(\frac{1}{2} - 2\epsilon\right) \cdot (1 - \phi_n)^{n-2} - 3\epsilon \cdot (1 - \phi_n)^{n-2} \\ &= \left(\frac{1}{2} - 5\epsilon\right) \cdot (1 - \phi_n)^{n-2}. \end{aligned}$$

By definition, $\lim_{n\to\infty}(1 - \phi_n)^{n-2} = 1$, so

$$\lim_{n\to\infty,\epsilon\to 0} r_s(\mathbf{h}^{\text{clr}}) = \frac{1}{2}.$$

■

## Appendix H. Proof of Theorem 9

**Theorem 9** *The following expression holds for the regret with respect to Jaccard distance*

$$\lim_{\epsilon\to 0} r_J(\mathbf{h}^{clr}) = \frac{1}{2}, \text{ for distribution } \bar{\mathbf{P}}.$$

**Proof** Let $A$ be a set of labellings of $n$ labels defined as $A = \{0_n, \mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{y}^{(1,2)}\}$, and $A' = \mathcal{Y} \backslash A$ its complement of labellings of $n$ labels. Let the risk be expressed as the following

$$R_L(\hat{\mathbf{y}}) = \sum_{\mathbf{y}\in\mathcal{Y}} L(\mathbf{y}, \hat{\mathbf{y}})\bar{\mathbf{P}}(\mathbf{y}) = \sum_{\mathbf{y}\in A} L(\mathbf{y}, \hat{\mathbf{y}})\bar{\mathbf{P}}(\mathbf{y}) + \sum_{\mathbf{y}\in A'} L(\mathbf{y}, \hat{\mathbf{y}})\bar{\mathbf{P}}(\mathbf{y}).$$

It will be shown that $\sum_{\mathbf{y}\in A'} L_J(\mathbf{y}, \hat{\mathbf{y}})\bar{\mathbf{P}}(\mathbf{y}) \leq \epsilon/2$:

$$\begin{aligned} \sum_{\mathbf{y}\in A'} L_J(\mathbf{y}, \hat{\mathbf{y}})\bar{\mathbf{P}}(\mathbf{y}) &\leq \sum_{\mathbf{y}\in A'} \bar{\mathbf{P}}(\mathbf{y}) = 1 - \sum_{\mathbf{y}\in A} \bar{\mathbf{P}}(\mathbf{y}) \\ &= 1 - \left(\bar{\mathbf{P}}(0_n) + \bar{\mathbf{P}}(\mathbf{y}^{(1)}) + \bar{\mathbf{P}}(\mathbf{y}^{(2)}) + \bar{\mathbf{P}}(\mathbf{y}^{(1,2)})\right) \\ &= 1 - \left(\underbrace{\bar{\mathbf{P}}(0,0) + \bar{\mathbf{P}}(1,0) + \bar{\mathbf{P}}(0,1) + \bar{\mathbf{P}}(1,1)}_{1}\right) \cdot \underbrace{\bar{\mathbf{P}}(Y_3 = 0) \cdots \bar{\mathbf{P}}(Y_n = 0)}_{(1-\phi_n)^{n-2}} \\ &= 1 - (1 - \phi_n)^{n-2} \leq (n-2)\phi_n, \end{aligned}$$

where the last inequality comes from the Bernoulli inequality. By definition of $\phi_n$ it has that $(n-2)\phi_n < \epsilon/2$, which can be made arbitrarily small. Therefore,

$$\sum_{\mathbf{y}\in A} L_J(\mathbf{y},\hat{\mathbf{y}})\bar{\mathbf{P}}(\mathbf{y}) \leq R_J(\hat{\mathbf{y}}) \leq \epsilon + \sum_{\mathbf{y}\in A} L_J(\mathbf{y},\hat{\mathbf{y}})\bar{\mathbf{P}}(\mathbf{y}),$$

which implies that

$$\lim_{\epsilon\to 0} R_J(\hat{\mathbf{y}}) = \sum_{\mathbf{y}\in A} L_J(\mathbf{y},\hat{\mathbf{y}})\bar{\mathbf{P}}(\mathbf{y}). \tag{29}$$

Our objective is to calculate $\lim_{\epsilon\to 0} r_J(\mathbf{h}^{\text{clr}}) = \lim_{\epsilon\to 0} R_J(\mathbf{h}^{\text{clr}}) - \lim_{\epsilon\to 0} R_J(\mathbf{y}^*)$. When $\epsilon$ tends to zero, $\phi_n$ tends to zero and $\bar{\mathbf{P}}$ will have only 2 non-null probabilities, $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$. Thus, calculating the $\lim_{\epsilon\to 0} r_J(\mathbf{h}^{\text{clr}})$ is easy since there will be only 2 non-null probabilities to sum up. Hence, Equation (29) can be reduced to

$$\lim_{\epsilon\to 0} R_J(\hat{\mathbf{y}}) = L_J(\mathbf{y}^{(1)},\hat{\mathbf{y}})\bar{\mathbf{P}}(\mathbf{y}^{(1)}) + L_J(\mathbf{y}^{(2)},\hat{\mathbf{y}})\bar{\mathbf{P}}(\mathbf{y}^{(2)}). \tag{30}$$

Firstly, let us determine the optimal risk. An optimal solution for Jaccard distance on $\bar{\mathbf{P}}$ is clearly either $\mathbf{y}^{(1)}, \mathbf{y}^{(2)}$ or $\mathbf{y}^{(1,2)}$. This can be easily solved by checking all three values.

$$\begin{aligned}
\lim_{\epsilon\to 0} R_J(\mathbf{y}^{(1)}) &= \underbrace{L_J(\mathbf{y}^{(1)},\mathbf{y}^{(1)})}_{0}\bar{\mathbf{P}}(\mathbf{y}^{(1)}) + \underbrace{L_J(\mathbf{y}^{(2)},\mathbf{y}^{(1)})}_{1}\cdot\underbrace{\bar{\mathbf{P}}(\mathbf{y}^{(2)})}_{1/2} && \text{From (30)}\\
&= \frac{1}{2}.\\
\lim_{\epsilon\to 0} R_J(\mathbf{y}^{(1,2)}) &= \underbrace{L_J(\mathbf{y}^{(1)},\mathbf{y}^{(1,2)})}_{1/2}\cdot\underbrace{\bar{\mathbf{P}}(\mathbf{y}^{(1)})}_{1/2} + \underbrace{L_J(\mathbf{y}^{(2)},\mathbf{y}^{(1,2)})}_{1/2}\cdot\underbrace{\bar{\mathbf{P}}(\mathbf{y}^{(2)})}_{1/2} && \text{From (30)}\\
&= \frac{1}{2}.
\end{aligned} \tag{31}$$

The optimal risk is $\frac{1}{2}$, when $\epsilon \to 0$. The risk of $\mathbf{h}^{\text{clr}}$ is given by

$$\begin{aligned}
\lim_{\epsilon\to 0} R_J(\mathbf{h}^{\text{clr}}) &= \lim_{\epsilon\to 0} R_J(0_n) && \text{From Proposition 7}\\
&= \underbrace{L_J(\mathbf{y}^{(1)},0_n)}_{1}\cdot\bar{\mathbf{P}}(\mathbf{y}^{(1)}) + \underbrace{L_J(\mathbf{y}^{(2)},0_n)}_{1}\cdot\bar{\mathbf{P}}(\mathbf{y}^{(2)})\\
&= 1.
\end{aligned} \tag{32}$$

The regret is given by:

$$\begin{aligned}
\lim_{\epsilon\to 0} r_J(\mathbf{h}^{\text{clr}}) &= \lim_{\epsilon\to 0} R_J(\mathbf{h}^{\text{clr}}) - \lim_{\epsilon\to 0} R_J(\mathbf{y}^*)\\
&= \frac{1}{2} && \text{From (31) and (32)}
\end{aligned}$$

■

## Appendix I. Proof of Theorem 10

**Theorem 10** *The following expression holds for the regret with respect to F-measure loss*

$$\lim_{\epsilon \to 0} r_J(\mathbf{h}^{clr}) = \frac{2}{3}, \text{ for distribution } \bar{\mathbf{P}}.$$

**Proof** This proof is similar to the proof of Theorem 9. Let $A$ be a set of labellings of $n$ labels defined as $A = \{0_n, \mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{y}^{(1,2)}\}$. Like in Theorem 9, the risk on distribution $\bar{\mathbf{P}}$ can be expressed as (see Equation (30)):

$$\lim_{\epsilon \to 0} R_F(\hat{\mathbf{y}}) = L_F(\mathbf{y}^{(1)}, \hat{\mathbf{y}})\mathbf{P}(\mathbf{y}^{(1)}) + L_F(\mathbf{y}^{(2)}, \hat{\mathbf{y}})\mathbf{P}(\mathbf{y}^{(2)}).$$

An optimal solution for F-measure on $\bar{\mathbf{P}}$ is clearly either $\mathbf{y}^{(1)}, \mathbf{y}^{(2)}$ or $\mathbf{y}^{(1,2)}$. This can be easily solved by checking all three values:

$$\begin{aligned}
\lim_{\epsilon \to 0} R_F(\mathbf{y}^{(1)}) =& \underbrace{L_F(\mathbf{y}^{(1)}, \mathbf{y}^{(1)})}_{0} \cdot \bar{\mathbf{P}}(\mathbf{y}^{(1)}) + \underbrace{L_F(\mathbf{y}^{(2)}, \mathbf{y}^{(1)})}_{1} \cdot \underbrace{\bar{\mathbf{P}}(\mathbf{y}^{(2)})}_{1/2} \\
=& \frac{1}{2}. \\
R_F(\mathbf{y}^{(1,2)}) =& \underbrace{L_F(\mathbf{y}^{(1)}, \mathbf{y}^{(1,2)})}_{1/3} \cdot \underbrace{\bar{\mathbf{P}}(\mathbf{y}^{(1)})}_{1/2} + \underbrace{L_F(\mathbf{y}^{(2)}, \mathbf{y}^{(1,2)})}_{1/3} \cdot \underbrace{\bar{\mathbf{P}}(\mathbf{y}^{(2)})}_{1/2} \\
=& \frac{1}{3}.
\end{aligned} \tag{33}$$

The risk of $\mathbf{h}^{\text{clr}}$ is given by

$$\begin{aligned}
\lim_{\epsilon \to 0} R_F(\mathbf{h}^{\text{clr}}) &= \lim_{\epsilon \to 0} R_J(0_n) && \text{From Proposition 7} \\
&= \underbrace{L_F(\mathbf{y}^{(1)}, 0_n)}_{1} \cdot \bar{\mathbf{P}}(\mathbf{y}^{(1)}) + \underbrace{L_J(\mathbf{y}^{(2)}, 0_n)}_{1} \cdot \bar{\mathbf{P}}(\mathbf{y}^{(2)}) \\
&= 1.
\end{aligned} \tag{34}$$

The regret is given by:

$$\begin{aligned}
\lim_{\epsilon \to 0} r_F(\mathbf{h}^{\text{clr}}) =& \lim_{\epsilon \to 0} R_F(\mathbf{h}^{\text{clr}}) - \underbrace{\lim_{\epsilon \to 0} R_F(\mathbf{y}^*)}_{1/3} \\
=& \frac{2}{3} && \text{From (33) and (34)}
\end{aligned}$$

■