IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 24 October 2024, accepted 20 November 2024, date of publication 27 November 2024,
date of current version 10 December 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3507199

RESEARCH ARTICLE

Confusion Matrices: A Unified Theory

JOHAN ERBANI[1], PIERRE-ÉDOUARD PORTIER[2], ELŐD EGYED-ZSIGMOND[1], AND DIANA NURBAKOVA[1]
[1]INSA Lyon, CNRS, UCBL, LIRIS, UMR 5205, Université de Lyon, 69621 Villeurbanne, France
[2]Caisse d'Epargne Rhône-Alpes, 75201 Paris, France

Corresponding author: Johan Erbani (johan.leydet@insa-lyon.fr)

ABSTRACT The confusion matrix is a key tool for understanding and evaluating models in supervised classification problems. Various matrices are proposed depending on the problem framework: single-label, multi-label, or even soft-label restricted to probability distributions. However, most of these approaches are not compatible with each other and lack theoretical justification. Leveraging optimal transport theory and the principle of maximum entropy, we propose a unique confusion matrix applicable across single, multi, and soft-label contexts. The Transport-based Confusion Matrix (TCM) extends the classic Confusion Matrix (CM), being identical in the single-label context. TCM introduces a comprehensive, theory-supported description of previously inaccessible errors, thereby enhancing the consistency and scope of machine learning evaluation.

INDEX TERMS Classification, evaluation, machine learning, multi-label confusion matrix, optimal transport, single-label confusion matrix, soft-label confusion matrix.

I. INTRODUCTION

In this introduction, we first describe the significance of confusion matrices and explore the various frameworks that require generalizing the classic Confusion Matrix (CM). We then discuss the challenges that emerge from this requirement. Next, we present our initial intuition and introduce, through a practical example, the transport theory that underpins our approach. Finally, we outline the process for deriving transport-based matrices and emphasize the key contributions of our work.

All relevant notations and concepts are listed in Table 1, and the first occurrence of each concept is italicized.

A. CONFUSION MATRIX USEFULNESS

Confusion matrices provide a comprehensive understanding of model errors by revealing interdependencies between labeled and predicted classes. Understanding these error patterns is crucial for the design, training, and optimization of classifiers [1], [2]. Furthermore, CM allows the introduction of such useful performance measures as precision, recall, and the F1-score.

The associate editor coordinating the review of this manuscript and approving it for publication was Rossano Musca.

B. FRAMEWORKS

In supervised learning, single-label classification assigns each instance to one class from a predefined set of categories. In contrast, multi-label classification allows an observation to be associated with multiple classes at the same time. The soft-label context further extends this by assigning non-negative real values to each class, indicating the confidence, degree, or value of membership for each class. This framework encompasses both single and multi-label scenarios.

Multi-label and soft-label classification generate significant research interest due to their wide range of applications, including medical diagnosis [3], [4], [5], [6], emotion recognition [7], [8], [9], as well as image and text classification [10], [11], [12], [13], [14], [15], [16], [17], [18].

Notably, the soft-label framework has drawn attention for its ability to preserve annotator disagreements in tasks involving subjective judgments, such as sentiment analysis, sarcasm detection, and offensive language identification [15], [16], [17], [18]. Although learning with soft labels is less common, it has been shown to improve model generalization, robustness, and calibration, highlighting the importance of soft labels in machine learning [10], [19].

Despite the increasing focus on these frameworks, there remains a significant gap in developing a consistent confusion

TABLE 1. Notations and concepts.

(a) In general

Notation	Meaning
C	Number of classes
N	Test set size
y^n	n-th label, a vector in $\mathbb{R}^C_{\geq 0}$
$\hat{y}^n$	n-th prediction, a vector in $\mathbb{R}^C_{\geq 0}$
$\|\cdot\|_1$	ℓ_1 norm
$\mathrm{diag}(\cdot)$	$\mathrm{diag}(v)$ is the diagonal matrix whose diagonal is equal to the vector v.
$\min(\cdot,\cdot)$	Minimum of two elements applied element-wise.
$\otimes$	Outer product

(b) About confusion matrix

Concept	Meaning
Instance	A label and its corresponding prediction. The n-th instance refers to y^n and $\hat{y}^n$.
Contribution	Each confusion matrix considered here can be expressed as the sum of N matrices, where the n-th term derives solely from the n-th instance, referred to as the n-th contribution.

(c) About transportation

Concept	Meaning
Transference plan	With respect to two measures (see Subsection I-E), a transference plan represents a way of transforming one into the other. In our context, it is a matrix that describes how to transform a prediction into its corresponding label.
Optimal transference plan	A transference plan that minimizes the Kantorovich problem.
$T(y,\hat{y})$	Set of transference plans to solve the Kantorovich problem relative to instance y and $\hat{y}$, see (6) in Subsection III-B.
$T^{\mathrm{opt}}(y,\hat{y})$	Set of optimal transference plans of the Kantorovich problem relative to the instance $(y,\hat{y})$.

matrix that is suitable for both multi and soft-label scenarios. This gap represents a crucial challenge in enhancing model understanding and evaluation.

C. DIFFICULTIES AND LIMITATIONS

Understanding errors in a multi-label scenario can be difficult. For instance, if the model predicts `Apple, Lemon` while the actual label is `Banana, Pear`, it is unclear whether `Banana` has been confused with `Apple` and `Pear` with `Lemon`, or whether `Pear` has been confused with `Apple` and `Banana` with `Lemon`, or if the error is more complex. Additionally, the number of predicted classes can differ from the number of actual classes. For example, a prediction of `Apple, Lemon` might be associated with the label `Apple`. Capturing such an error in a confusion matrix is not straightforward: should `Apple` be considered correctly recognized, with `Lemon` treated as an unrelated error, or has the model only partially captured `Apple`, mistakenly associating it with `Lemon`? If the latter, to what extent does this confusion occur?

These difficulties are exacerbated in soft-label scenarios. Consider a tweet classification problem with four topics: `Politics`, `Science`, `People`, and `Fake News`. In this case, each label is represented by a vector of size four, with entries ranging between 0 and 1, indicating the degree to which each topic is related to the text. For example, the model might predict [1, 0, 0.5, 0.2] for a tweet predominantly about `Politics`, with some correspondence to `People` and low similarity to `Fake News`, while the label is [0.4, 1, 0, 0]. How can this pair label-prediction be interpreted to highlight any dependencies between predicted and labeled classes? So far, state-of-the-art confusion matrices cannot solve this situation.

D. FROM SINGLE TO SOFT-LABEL CONFUSION MATRIX: STARTING POINT

The single-label confusion matrix compares predictions with labels to reveal the model's behavior. To establish the average behavior, CM aggregates the model's behaviors across all *instances*, resulting in a sum of *contributions*,

$$\mathrm{CM} = \sum_{n=1}^{N} y^n \otimes \hat{y}^n, \tag{1}$$

where labels and predictions are vectors in $\{0,1\}^C$. Each contribution compares y^n with $\hat{y}^n$, recording errors through the matrix $y^n \otimes \hat{y}^n$.

Contributions can be regarded as instructions to transform $\hat{y}^n$ into y^n. The ij term represents the quantity in entry j to be shifted into entry i to obtain y from $\hat{y}$. In the case where $i = j$, the quantity in i is shifted to i; thus, nothing happens. Two simple examples are depicted in Fig. 1.

This view enables the generalization of the single-label confusion matrix to broader frameworks. Instructions to transform predictions into labels can be effectively formalized using *transference plans*, a key notion in optimal transport theory, which is illustrated with an example in subsection I-F. However, before delving into it, we have to introduce the concept of measure, a prerequisite for understanding transport theory.

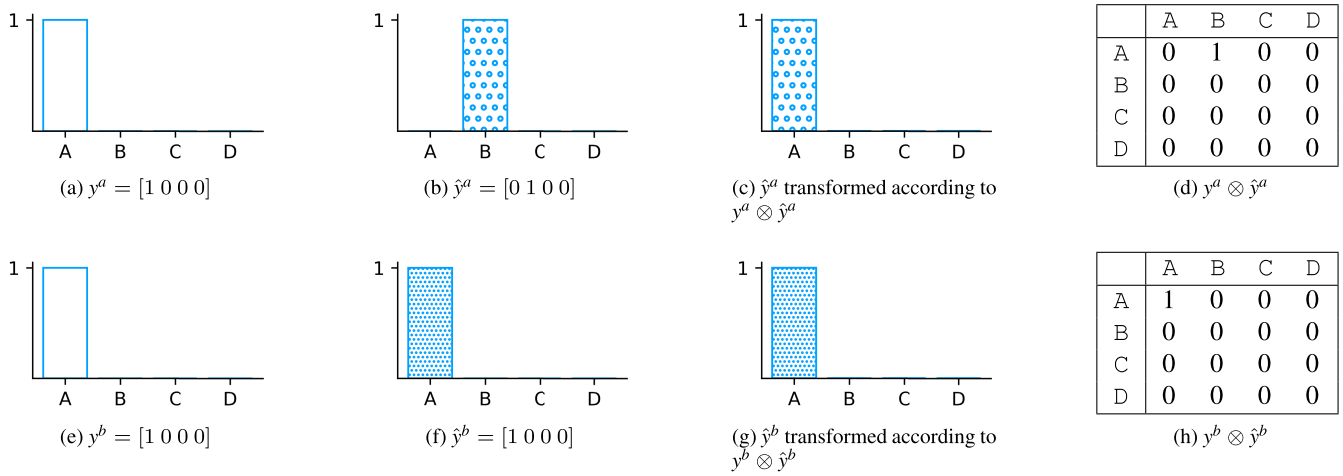
E. MEASURE

The concept of a measure is a generalization and formalization of length, area, volume, and other common notions such as mass and probability of events [20]. Informally, a measure μ is a function that assigns non-negative values to certain sets and satisfies basic properties, such as assigning 0 to the empty set (i.e., $\mu(\varnothing) = 0$) and ensuring that the measure of the union of disjoint sets is the sum of their individual measures (i.e., $\mu(A \cup B) = \mu(A) + \mu(B)$ with A and B two disjoint sets).

To apply transport theory, we interpret each label y (or prediction $\hat{y}$) as a measure representing a collection of point masses over the set of classes (see Susection III-B), which can be visualized as a bar chart indexed by classes.

F. OPTIMAL TRANSPORT EXAMPLE

We present a concrete use case of optimal transport theory, which can serve as a reference for subsequent discussions. This example is a discrete version of the introductory example provided by Villani [21].



	A	B	C	D
A	0	1	0	0
B	0	0	0	0
C	0	0	0	0
D	0	0	0	0

	A	B	C	D
A	1	0	0	0
B	0	0	0	0
C	0	0	0	0
D	0	0	0	0

FIGURE 1. Bar charts represent two instances with classes A, B, C, and D: $(y^a, \hat{y}^a)$ in (a) and (b), and $(y^b, \hat{y}^b)$ in (e) and (f). Quantities predicted in class A are shown with dotted bars, whereas those in B are circled. CM contributions of $y^a \otimes \hat{y}^a$ and $y^b \otimes \hat{y}^b$ appear in (d) and (h) and are interpreted as follows: (d) move a quantity of 1 from B to A, and (h) leave a quantity of 1 in A. Their corresponding transformation appears in (c) and (g).

Suppose we have K piles of sand with a total volume V_K and L holes with a total volume V_L. Additionally, let's assume V_K equals V_L. Our goal is to fill the holes with sand by transporting the sand over the shortest possible distance. A pile of sand can be spread over several holes.

Transportation theory formalizes this problem, aiming to determine how to transport a measure μ to a measure ν, defined respectively on some measure spaces X and Y while minimizing a given cost function $c : X \times Y \to \mathbb{R}_{\geq 0} \cup \{+\infty\}$. Denoting $s_1, s_2, \ldots s_K$ the volume of each pile of sand and $u_1, u_2, \ldots u_K$ their location, the measure μ on $\mathbb{R}^2$ is defined by $\mu(u_k) = s_k$ the volume of sand located in coordinates $u_k \in \mathbb{R}^2$ and μ null everywhere else. Similarly, denoting $h_1, h_2, \ldots h_L$ the volume of each hole and $v_1, v_2, \ldots v_L$ their location, $\nu(v_l) = h_l$ and ν null everywhere else. The cost function c corresponds to the Euclidean distance.

A transference plan is a measure on the product space $X \times Y$. In this case, it can be represented by a matrix $\pi \in \mathbb{R}_{\geq 0}^{K \times L}$ with π_{kl} the amount of sand coming from pile k and placed in hole l. More precisely, the problem is written as follows:

$$\arg\min_{T(\mu,\nu)} \sum_{k=1}^{K} \sum_{l=1}^{L} c(u_k, v_l)\pi_{kl} \tag{2}$$

where $T(\mu, \nu)$ is the set of admissible transference plans defined by:

$$T(\mu, \nu) = \{\pi \in \mathbb{R}_{\geq 0}^{K \times L} : \sum_{l=1}^{L} \pi_{kl} = \mu(u_k), \sum_{k=1}^{K} \pi_{kl} = \nu(v_l)\}. \tag{3}$$

Considering plans among $T(\mu, \nu)$ ensures that after the operation, all the sand piles are empty and all the holes are filled. In addition, the quantities of transported sand are non-negative.

This problem is called the Kantorovich problem. For a solution to exist, the total volume of sand V_K must equal the volume of the holes V_L. When a solution exists, it is not necessarily unique, for example, if there are some equidistances between piles and holes. Solutions are called *optimal transference plans*.

G. TRANSPORT-BASED CONFUSION MATRIX

The Transport-based Confusion Matrix extends the single-label confusion matrix to multi-label and soft-label frameworks by leveraging optimal transport theory. The procedure for deriving TCM is outlined below.

For each instance $(y^n, \hat{y}^n)$ in the set of instances $(y^1, \hat{y}^1), \ldots, (y^N, \hat{y}^N)$ do:

1) Interpret $(y^n, \hat{y}^n)$ as a pair of measures and solve the associated Kantorovich problem.
2) If the solution is not unique, select the optimal transference plan that maximizes entropy, denoted as $\pi^*(y^n, \hat{y}^n)$. This specific transference plan is a square matrix of size C which describes the model's behavior on the instance $(y^n, \hat{y}^n)$.

In line with the single-label confusion matrix approach, TCM captures the model's average behavior by aggregation:

$$\text{TCM} = \sum_{n=1}^{N} \lambda(y^n, \hat{y}^n)\pi^*(y^n, \hat{y}^n), \tag{4}$$

where $\lambda(y^n, \hat{y}^n) \in \mathbb{R}_{\geq 0}$ is the non-negative weighting factor of the n-th contribution. The complete definitions of $\pi^*(y^n, \hat{y}^n)$ and $\lambda(y^n, \hat{y}^n)$ can be found in Subsection III-D Proposition 3 and III-F, respectivly.

H. CONTRIBUTIONS

We introduce the Transport-based Confusion Matrix, with the following key properties:

- Universality: It offers a single method for analyzing model errors in single, multi, and soft-label frameworks. Moreover, to the best of our knowledge, it is the only

proposal in the soft-label framework that is not restricted to probability distributions.
- Unification: It extends the classic confusion matrix, remaining identical in the single-label context while also being partially consistent with existing proposals, thereby connecting apparently incompatible confusion matrix approaches.
- Reliability and Interpretability: The approach is grounded in well-established theories, specifically optimal transport and the principle of maximum entropy. Moreover, it relies on minimal assumptions, ensuring highly reliable and interpretable results.

The Appendix includes theoretical proofs for the claims made throughout the paper. The source code of TCM, as well as that of the experiments, is available on https://github.com/johan140391/TCM.

I. PAPER ORGANIZATION

The structure of the paper is as follows: Section II reviews existing approaches, their limitations, and the objectives of our proposal. Section III outlines the process for obtaining TCM, provides an interpretation of the mathematical formulas of the matrix, and includes a simple example for familiarization. Section IV establishes connections between TCM and state-of-the-art matrices and identifies key properties that TCM satisfies. Section V extends the traditional metrics of Recall, Precision, F1-score, and Accuracy. Section VI offers guidelines for reading confusion matrices in general, especially criteria for distinguishing different types of model errors. Section VII describes the experimental setup, while Section VIII presents a comparative analysis of TCM against the various confusion matrices in the literature. Section IX presents a case study with real-world data. Section X offers a critical discussion of our approach, and Section XI concludes the paper.

II. RELATED WORKS

This section reviews existing approaches and their limitations in both multi-label and soft-label frameworks, concluding with a summary of the limitations and the objectives of our proposal.

A. MULTI-LABEL FRAMEWORK

The Scikit-learn Python library [22] provides a method for evaluating errors in a multi-label framework. This algorithm calculates true positives, false positives, true negatives, and false negatives for each class, resulting in as many CMs as there are classes. However, as discussed in [2] and [23], this approach is insufficient to accurately describe the model's behavior, as it relies on class-wise analysis.

Several papers, including [1], [2], and [23], introduce multi-label confusion matrices. Krstinić, Braović, Šerić, et al. [1] proposes four formulas to compute contributions. Each formula addresses a specific situation (e.g., too many or too few predicted classes; see Section IV for more details). Similarly, [23] presents three formulas, yielding experimentally comparable results [24]. As pointed out in [2], using a single global formula would enhance consistency. Krstinić et al. [2] aim to recover the traditional precision and recall scores, resulting in two distinct confusion matrices, one for each score. However, the authors do not specify how to interpret their matrices. For instance, when evaluating confusion between class i and class j, it is unclear whether the ij entry should be taken from the recall or precision matrix, whereas these entries can differ significantly (see Section VIII). Moreover, recall and precision are class-wise metrics that do not consider all classes present in the labels and predictions. Placing them on the diagonal is questionable, as a multi-label confusion matrix should provide insights beyond class-wise scores, revealing relationships between classes even when considering only diagonal values. A key concern with the approaches proposed in [1], [2], and [23] is the limited justification behind their proposals.

To identify confusion in a hierarchical and multi-label scenario, [25] suggests a matrix based on multivariate probability distributions. The authors also use their matrix in non-hierarchical multi-label problems. Errors are analyzed on a group scale: confusing `Apple` with `Banana`; `Lemon` with `Pear`; and `Apple, Lemon` with `Banana, Pear`; are assumed to be three confusion types, whereas intuitively, two types are expected, `Apple` with `Banana` and `Lemon` with `Pear`. This results in an exponential increase in matrix size as the number of classes grows [25].

B. SOFT-LABEL FRAMEWORK RESTRICTED TO PROBABILITY

The authors of [26] and [27] propose matrices for soft-label scenarios restricted to probability distributions (i.e., the sum of the labels or predictions must be equal to 1). Binaghi et al. [26] introduce a matrix based on fuzzy set theory. Their method yields counterintuitive results, as a perfect prediction could produce a non-diagonal matrix [28].

The paper by Silván-Cárdenas and Wang [27] in the geographic information science field suggests a method for comparing terrestrial images. Each pixel in the images belongs to certain classes, with one image serving as the reference and the other as a comparison. Within the machine learning framework, one pixel in the reference image is considered the ground truth, while the corresponding pixel in the comparison image is the model's prediction. Various operators exist for soft pixel classification according to pixel ontology [28]. Some operators use fuzzy logic, while others rely on pixel-overlapping arguments. The authors introduce composite operators to determine the minimum and maximum sub-pixel class overlap, resulting in a matrix with interval entries. In each entry ij, the lower (respectively upper) bound corresponds to the minimum confusion of i with j (respectively maximum). To facilitate the reading, the authors propose to represent each entry by the interval center and the interval half-width, i.e., an entry of the form $[a,\ b]$ becomes $(a+b)/2 \pm (b-a)/2$.

C. CONCLUSION

Existing solutions appear incompatible. Some lack justification, while others lose the familiar structure of the confusion matrix, making it difficult to assess the model's behavior accurately. Furthermore, no method simultaneously integrates single, multi, and soft-label frameworks, resulting in inconsistencies in evaluation systems. The main state-of-the-art approaches discussed in this paper are summarized in Table 2.

To address these issues, we propose evaluating confusion using optimal transport and the principle of maximum entropy. This method aims to preserve as much model behavior information as possible from instances while avoiding uncertain assumptions. To the best of our knowledge, this approach is unexplored and promises to provide a more comprehensive and interpretable evaluation of model performance.

III. TRANSPORT BASED CONFUSION MATRIX

This section outlines the process for obtaining TCM. We start by describing the Kantorovich problem relevant to our context, along with the appropriate cost function, and then present the solution. Next, we offer a concrete interpretation of the mathematical formulas used in TCM, followed by an explanation of the weighting factor. Finally, we conclude with a simple example to help understanding.

To simplify the notation, when there is no ambiguity, we will omit the index n and refer to an instance as y and $\hat{y}$ rather than y^n and $\hat{y}^n$.

A. STARTING POINT

A set of instructions for transforming predictions into labels, such as "this portion of $\hat{y}$ is correctly classified, but this part assigned to class j should belong to class i," provides valuable insights for analyzing model behavior. This is also a way of interpreting the contributions of the single-label confusion matrix, as discussed in the introduction.

To derive these instructions, we employ optimal transport theory. This requires defining a relevant Kantorovich problem for our context.

B. KANTOROVICH PROBLEM

We interpret each instance $(y, \hat{y})$ as a pair of measures. Specifically, the label y (and similarly the prediction $\hat{y}$) is viewed as a weighted sum of Dirac measures over the set of classes.[1]

For a solution to exist, it is necessary that y and $\hat{y}$ possess the same norm, $\|y\|_1 = \|\hat{y}\|_1$. However, this condition is generally unmet in the multi or soft-label framework, as too many or too few classes can be predicted. Therefore, we propose comparing normalized instances, $y/\|y\|_1$ and $\hat{y}/\|\hat{y}\|_1$, resulting in a proportional comparison of labels and predictions. More details are provided in Subsection X-A.

The Kantorovich problem is defined as:

$$\arg\min_{T(y,\hat{y})} \sum_{i,j=1}^{C} c\,(i,j)\,\pi_{ij} \tag{5}$$

With $T(y, \hat{y})$ a set of matrices defined by:

$$T(y,\hat{y}) = \left\{ \pi \in \mathbb{R}_{\geq 0}^{C\times C} : \sum_{j=1}^{C} \pi_{ij} = \frac{y_i}{\|y\|_1}, \sum_{i=1}^{C} \pi_{ij} = \frac{\hat{y}_j}{\|\hat{y}\|_1} \right\} \tag{6}$$

To complete the description of this problem, we need to define the cost function.

C. COST FUNCTION

The cost function indicates the cost of transporting a unit of mass from one location to another [21]. A low cost between two locations facilitates mass transfer between them, whereas a high cost hinders it.

In the confusion matrix context, and for a given model, the cost function represents the interclass dissimilarities. For example, with a classifier that systematically confuses `A` and `B` but never `A` and `C`, a suitable cost function would assign $c(\texttt{A},\texttt{B}) = 0$ and $c(\texttt{A},\texttt{C}) = \infty$.

However, based on labels and predictions alone, this information is inaccessible with any certainty. Nevertheless, the following assumption is reliable: a given class `A` is very similar to itself (as it is the same class) but different from other classes. This qualitative information is captured by the *discrete metric*, which assigns a value of 0 when the classes are the same and 1 when the classes are different. Thus, for all classes i and j, $c(i, j) = 0$ if $i = j$ and $c(i, j) = 1$ if $i \neq j$.

D. SOLUTION

The following two propositions provide the solutions to our Kantorovich problem, identifying all optimal transference plans and demonstrating that this set typically contains an infinite number of solutions. Let $T^{\text{opt}}(y, \hat{y}) \subset T(y, \hat{y})$ denote the solution set.

Proposition 1. *Any matrix $\pi \in T(y, \hat{y})$ is an optimal transference plan if, and only if, its diagonal is defined by*:

$$\pi_{ii} = \min\left(\frac{y_i}{\|y\|_1}, \frac{\hat{y}_i}{\|\hat{y}\|_1}\right) \tag{7}$$

for $i = 1, 2, \ldots C$.

Proposition 2. *$T^{opt}(y, \hat{y})$ contains a single optimal transference plan if, and only if, there is at most one overestimated (i.e., $\frac{y_k}{\|y\|_1} < \frac{\hat{y}_k}{\|\hat{y}\|_1}$) or underestimated (i.e., $\frac{\hat{y}_k}{\|\hat{y}\|_1} < \frac{y_k}{\|y\|_1}$) class. Otherwise, there is an infinite number of optimal transference plans.*

Consequently, $T^{\text{opt}}(y, \hat{y})$ always contains a single plan when $C = 2$ or 3.

Generally, more than one optimal transference plan exist. We propose choosing the one that maximizes entropy. This

[1] Let $A = \{a_1, \ldots, a_C\}$ be the set of classes, $\mathcal{P}(A)$ the power set of A, and δ_{a_i} the Dirac measure concentrated on class a_i. Then, a vector v in $\mathbb{R}_{\geq 0}^{C}$ is mapped to the measure $\sum_{i=1}^{C} v_i \delta_{a_i}$ on the measurable space $(A, \mathcal{P}(A))$.

TABLE 2. Overview of state-of-the-art confusion matrices and TCM in single-label, multi-label, soft-label restricted to probability distributions, and soft-label frameworks.

Matrix	Single	Multi	Proba.	Soft	Description
Confusion Matrix (CM)	✓	✗	✗	✗	Each entry *ij* is the number of occurrences of event: "label *i* predicted *j*".
Multi-Label Confusion Matrix (MLCM) [1]	✓	✓	✗	✗	Four formulas are used to determine contributions, depending on the situation. No rigorous justification is given.
Multi-Label Confusion Tensor (MLCT) [2]	✓	✓	✗	✗	Two matrices with the property of matching the classical recall and precision scores on the diagonal. Interpretation is unclear.
Sup-pixel Confusion Matrix (SCM) [28]	✓	✗	✓	✗	Matrix based on composite operators to determine the minimum and maximum sub-pixel class overlap, resulting in a matrix with interval entries.
Transport-based Confusion Matrix (TCM)	✓	✓	✓	✓	A matrix based on minimal assumptions, supported by optimal transport theory and the maximum entropy principle, providing universal and interpretable results for all classification frameworks.

choice is justified because transference plans are probability measures, which allows us to apply the principle of maximum entropy [29]: when information is lacking but it is necessary to infer a distribution, we must choose the one that maximizes entropy while respecting the partial information available. This approach provides the least biased estimate possible given the incomplete information.

Proposition 3. *There is only one transference plan in* $T^{opt}(y, \hat{y})$ *that maximizes entropy, and it is defined by*:

$$
\begin{aligned}
D &= \operatorname{diag}\left(\min\left(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}\right)\right) \\
\overline{D} &= \frac{\left(\frac{y}{\|y\|_1} - \min\left(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}\right)\right) \otimes \left(\frac{\hat{y}}{\|\hat{y}\|_1} - \min\left(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}\right)\right)}{\| \frac{y}{\|y\|_1} - \min\left(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}\right)\|_1} \\
\pi^*(y, \hat{y}) &= D + \overline{D}. \qquad (8)
\end{aligned}
$$

Continuous extension makes $\overline{D}$ *a zero matrix when* $y = \hat{y}$.

There are several key points to consider regarding this result. Firstly, if a prediction is perfectly correct, meaning $y = \hat{y}$, the denominator of $\overline{D}$ becomes zero. In the Appendix, we demonstrate that as the prediction gets closer to the label, the entries of the matrix $\overline{D}$ decrease to 0. Specifically, $\|y - \hat{y}\|_1 \to 0$ implies $\overline{D} \to 0_C$, where 0_C denotes the zero matrix of size C. Thus,

$$
\pi^*(y, \hat{y}) = \begin{cases} D & \text{if } y = \hat{y} \\ D + \overline{D} & \text{otherwise.} \end{cases} \qquad (9)
$$

These equalities are summarized in Proposition 3 through the mathematical concept of continuous extension.

Secondly, when $T^{\text{opt}}(y, \hat{y})$ contains only one single plan, the constraints on diagonal terms set by Proposition 1 directly lead to the formula (8). Therefore, (8) can be used without addressing whether one or multiple optimal transference plans exist.

Thirdly, the plan $\pi^*(y, \hat{y})$ corresponds to the distribution that is closest to the uniform distribution, considering the constraints imposed by $T^{\text{opt}}(y, \hat{y})$. This follows directly from the choice to maximize entropy.

E. INTERPRETATION

This transformation plan $\pi^*(y, \hat{y})$ has a concrete interpretation. To explain it, we introduce the function f.

Definition 1. *Let* $\mathcal{D}$ *be the set of non-negative vectors of size* C *with the same norm*:

$$
\mathcal{D} = \{(u, v) : u \in \mathbb{R}^C_{\geq 0}, v \in \mathbb{R}^C_{\geq 0}, \text{ and } \|u\|_1 = \|v\|_1\}, \qquad (10)
$$

where the vector v *represents an approximation of the vector* u. *Let* $f : \mathcal{D} \to \mathcal{M}_C(\mathbb{R}_{\geq 0})$ *be a function defined as follow*:

$$
\begin{aligned}
f(u, v)_{ii} &= \Big(\textit{common quantities between } u_i \textit{ and } v_i\Big) \\
&= \min(u_i, v_i) \quad \textit{for } i = 1 \dots C \qquad (11)
\end{aligned}
$$

Additional vocabulary is required for off-diagonal entries. An entry i *is underestimated if* $v_i < u_i$; *in this case, the deficit is* $u_i - v_i$, *otherwise* 0. *Let* u'_i *be this missing quantity in* i. *Similarly, an entry* j *is overestimated if* $u_j < v_j$; *in this case, the excess quantity is* $v_j - u_j$, *otherwise* 0. *Let* v'_j *be this excess quantity in* j.

$$
\begin{aligned}
f(u, v)_{ij} &= \Big(\textit{deficit in } i\Big)\Big(\textit{share of excess quantity in } j\Big) \\
&= u'_i * \frac{v'_j}{\sum_{k=1}^C v'_k} \quad \textit{for } i, j = 1 \dots C, i \neq j \qquad (12)
\end{aligned}
$$

Since the equality $\sum_{k=1}^C u'_k = \sum_{k=1}^C v'_k$ *holds (according to Lemma 1 in Appendix Section XI), the entry* $f(u, v)_{ij}$ *is also*:

$$
\begin{aligned}
f(u, v)_{ij} &= \Big(\textit{share of deficit quantity in } i\Big)\Big(\textit{excess in } j\Big) \\
&= \frac{u'_i}{\sum_{k=1}^C u'_k} * v'_j \qquad (13)
\end{aligned}
$$

Each entry of the matrix $f(u, v)$ has a clear meaning. This interpretation aligns with that of the matrix $\pi^*(y, \hat{y})$.

Proposition 4. *The following equality holds*:

$$
\pi^*(y, \hat{y}) = f\left(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}\right). \qquad (14)
$$

The diagonal entries of $\pi^*(y, \hat{y})$ represent the common quantities between $\hat{y}/\|\hat{y}\|_1$ and $y/\|y\|_1$. In contrast, the off-diagonal entries ij are the product of the missing quantity in i and the excess quantity in j, divided by the sum of all excess or deficit quantities.

In conclusion, beyond its theoretical justification, $\pi^*(y, \hat{y})$ is easily interpretable.

F. WEIGHTING FACTOR

In a multi-label context, if an image contains both `Apple` and `Lemon`, it might be preferable for its contribution to count twice as much as an image containing only `Apple`. However, by design, all TCM contributions have the same importance since the entries of each contribution sum to 1: $\sum_{i,j=1}^{C} \pi^*(y,\hat{y})_{ij} = 1$.

To address this, we propose adding a weighting factor formalized by the function $\lambda : \mathbb{R}^C_{\geq 0} \times \mathbb{R}^C_{\geq 0} \to \mathbb{R}_{\geq 0}$. The coefficient $\lambda(y,\hat{y})$ represents the contribution weight associated with the instance y and $\hat{y}$. For example, the definition of λ could be:

- $\lambda : y, \hat{y} \mapsto 1$, all contributions are equally important.
- $\lambda : y, \hat{y} \mapsto \|y\|_1$, the more classes in a label, the more important its contribution is. An observation labeled `Apple, Lemon` contributes twice as much as an observation labeled `Apple`.
- $\lambda : y, \hat{y} \mapsto \|\hat{y}\|_1$, the more classes in a prediction, the more important its contribution is. An observation predicted as `Apple, Lemon` contributes twice as much as an observation predicted as `Apple`.

The direct interpretation of weighted TCM is that the confusion matrix results from a sum of contributions, with their importance varying based on the chosen weighting. A particular property of f (see Subsection III-E Definition 1) enables a better interpretation.

Proposition 5. *The function f is homogeneous, meaning $f(\alpha u, \alpha v) = \alpha f(u,v)$ for any positive real number α.*

According to Proposition 4 and Proposition 5, the following equalities hold:

$$\begin{aligned}\lambda(y,\hat{y})\pi^*(y,\hat{y}) &= \lambda(y,\hat{y}) f\left(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}\right) \\ &= f\left(\lambda(y,\hat{y})\frac{y}{\|y\|_1}, \lambda(y,\hat{y})\frac{\hat{y}}{\|\hat{y}\|_1}\right) \\ &= \begin{cases} f\left(y, \hat{y}\frac{\|y\|_1}{\|\hat{y}\|_1}\right) & \text{if } \lambda(y,\hat{y}) = \|y\|_1 \\ f\left(y\frac{\|\hat{y}\|_1}{\|y\|_1}, \hat{y}\right) & \text{if } \lambda(y,\hat{y}) = \|\hat{y}\|_1 \end{cases}\end{aligned} \tag{15}$$

Finally, TCM can be understood as a sum of unweighted contributions:

$$\sum_{n=1}^{N} f\left(\lambda(y^n,\hat{y}^n)\frac{y^n}{\|y^n\|_1}, \lambda(y^n,\hat{y}^n)\frac{\hat{y}^n}{\|\hat{y}^n\|_1}\right), \tag{16}$$

where each contribution is a matrix comparing a pair of vectors:

$$\left(\frac{y^n}{\|y^n\|_1}, \frac{\hat{y}^n}{\|\hat{y}^n\|_1}\right), \quad \left(y^n, \hat{y}^n\frac{\|y^n\|_1}{\|\hat{y}^n\|_1}\right), \quad \text{or } \left(y^n\frac{\|\hat{y}^n\|_1}{\|y^n\|_1}, \hat{y}^n\right) \tag{17}$$

depending on the chosen definition of λ. According to the definition of f, the diagonal terms correspond to the vector overlaps, and the non-diagonal entries ij to the product of the underestimation in i and the overestimation in j, normalized by the total of all the overestimations (or underestimations).

Finally, no weighting is neutral, each having a specific meaning. Therefore, this parameter must be considered when analyzing results given by TCM. Several matrices with different weightings can be plotted to gain a deeper model's understanding.

G. SIMPLE EXAMPLE

This subsection aims to familiarize readers with TCM through a soft-label example illustrated in Fig. 2.

Let $(x^1,y^1), (x^2,y^2), (x^3,y^3), (x^4,y^4)$ be a dataset where x_i are the observations and y_i the labels. The labels are vectors in $\mathbb{R}^4_{\geq 0}$ representing classes `A`, `B`, `C`, and `D`. Additionally, the model performs the predictions $\hat{y}^1, \hat{y}^2, \hat{y}^3$, and $\hat{y}^4$. In Fig. 2, the first row (subfigures (a) to (d)) is linked to instance one, the second row (subfigures (e) to (h)) to instance two, the third row (subfigures (i) to (l)) to instance three, and the fourth row (subfigures (m) to (p)) to instance four.

The following proposition highlights the conditions under which the entries of an optimal transformation plan are non-zero:

Proposition 6. *Given an instance $(y,\hat{y})$, its contribution $\pi^*(y,\hat{y})$ has a zero diagonal if, and only if, no common quantity exists. Moreover, entry ij is non-zero if, and only if, class i is underestimated (i.e., $\hat{y}_i/\|\hat{y}\|_1 < y_i/\|y\|_1$) and class j overestimated (i.e., $y_j/\|y\|_1 < \hat{y}_j/\|\hat{y}\|_1$).*

In the instance $(y^1,\hat{y}^1)$, `A` and `C` are predicted, whereas the expected class is `B`, so all predicted quantities are in excess. Since there is no common quantity between y_1 and $\hat{y}_1$, the diagonal of $\pi^*(y^1,\hat{y}^1)$ is zero. Class `B` is underestimated, whereas classes `A` and `C` are overestimated. As a result, only the entries `BA` and `BC` are non-zero, indicating that the predicted quantities must be shifted to `B`. Since `C` is predicted twice as much as `A`, the quantity coming from `C` is twice that coming from `A`. Formally, entry `BA` is equal to,

$$\begin{aligned}&\Big(\text{deficit in B}\Big)\Big(\text{share of excess quantity in A}\Big) \\ &\quad = 1\frac{1/3}{1/3+2/3} = 1/3.\end{aligned} \tag{18}$$

Similarly, entry `BC` is equal to,

$$\begin{aligned}&\Big(\text{deficit in B}\Big)\Big(\text{share of excess quantity in C}\Big) \\ &\quad = 1\frac{2/3}{1/3+2/3} = 2/3.\end{aligned} \tag{19}$$

In the instance $(y^2,\hat{y}^2)$, `A` and `B` are predicted, whereas the expected classes are `C` and `D`, so all predicted quantities are in excess. There is no common quantity, so the diagonal of $\pi^*(y^2,\hat{y}^2)$ is zero. Classes `C` and `D` are underestimated, whereas `A` and `B` are overestimated. As a result, only the entries `CA`, `CB`, `DA`, and `DB` are non-zero. In the label, `C` is twice as large as `D`. Therefore, the transference plan indicates that the predicted quantities to be shifted to `C` are twice those to be shifted to `D`. Additionally, the quantities predicted in `A` and `B` are identical, implying that the quantities coming from

A and B are identical. Formally, entry CA is equal to entry CB:

$$\begin{aligned}&\Big(\text{deficit in C}\Big)\Big(\text{share of excess quantity in A}\Big)\\&\quad=\Big(\text{deficit in C}\Big)\Big(\text{share of excess quantity in B}\Big)\\&\quad=2/3\frac{2/4}{2/4+2/4}=1/3.\end{aligned}\tag{20}$$

Entry DA is equal to entry DB:

$$\begin{aligned}&\Big(\text{deficit in D}\Big)\Big(\text{share of excess quantity in A}\Big)\\&\quad=\Big(\text{deficit in D}\Big)\Big(\text{share of excess quantity in B}\Big)\\&\quad=1/3\frac{2/4}{2/4+2/4}=1/6.\end{aligned}\tag{21}$$

In the instance $(y^3, \hat{y}^3)$, B, C, and D are predicted, whereas the expected classes are A and C. The common quantity in C implies that the diagonal entry C is non-zero. Class A is underestimated, whereas classes B, C, and D are overestimated. As a result, AB, AC, and AD are non-zero. The excess quantities in A, B, and C are identical. Consequently, the transference plan indicates that the predicted excess quantities to be shifted to A are equal. Formally, the diagonal entry for C equals:

$$\Big(\text{common quantities in C}\Big)=\min\left(\frac{1}{4},\frac{2}{4}\right)=\frac{1}{4}.\tag{22}$$

The entries AB, AC, and AD are equal to:

$$\begin{aligned}&\Big(\text{deficit in A}\Big)\Big(\text{share of excess quantity in B}\Big)\\&\quad=\Big(\text{deficit in A}\Big)\Big(\text{share of excess quantity in C}\Big)\\&\quad=\Big(\text{deficit in A}\Big)\Big(\text{share of excess quantity in D}\Big)\\&\quad=3/4\frac{1/4}{1/4+1/4+1/4}=1/4.\end{aligned}\tag{23}$$

In the instance $(y^4, \hat{y}^4)$, A, C, and D are predicted, whereas the expected classes are A and B. The common quantity in A implies that the diagonal entry for A is non-zero. A and B are underestimated, whereas C and D are overestimated. As a result, the entries AC, AD, BC, and BD are non-zero. The underestimation of A is greater than that of B. Consequently, the transference plan indicates that A receives more quantity from C and D than B. The excess quantity in D is greater than the one in C. Therefore, the quantities to be shifted from D are greater than those from C. Formally, the diagonal entry for A equals:

$$\Big(\text{common quantities in A}\Big)=\min\left(\frac{3}{4},\frac{3}{8}\right)=\frac{3}{8}.\tag{24}$$

Entry AC equals:

$$\begin{aligned}&\Big(\text{deficit in A}\Big)\Big(\text{share of excess quantity in C}\Big)\\&\quad=3/8\frac{2/8}{2/8+3/8}=3/20.\end{aligned}\tag{25}$$

Entry AD equals:

$$\begin{aligned}&\Big(\text{deficit in A}\Big)\Big(\text{share of excess quantity in D}\Big)\\&\quad=3/8\frac{3/8}{2/8+3/8}=9/40.\end{aligned}\tag{26}$$

Entry BC equals:

$$\begin{aligned}&\Big(\text{deficit in B}\Big)\Big(\text{share of excess quantity in C}\Big)\\&\quad=1/4\frac{2/8}{2/8+3/8}=1/10.\end{aligned}\tag{27}$$

Entry BD equals:

$$\begin{aligned}&\Big(\text{deficit in B}\Big)\Big(\text{share of excess quantity in D}\Big)\\&\quad=1/4\frac{3/8}{2/8+3/8}=3/20.\end{aligned}\tag{28}$$

According to (4) and Subsection III-F, to obtain the unweighted TCM, where $\lambda : y, \hat{y} \mapsto 1$, the contributions from these four instances are summed as

$$\pi^*(y^1, \hat{y}^1) + \pi^*(y^2, \hat{y}^2) + \pi^*(y^3, \hat{y}^3) + \pi^*(y^4, \hat{y}^4).\tag{29}$$

Considering the weighting $\lambda : y, \hat{y} \mapsto \|y\|_1$, the label-weighted TCM is

$$\begin{aligned}&1\pi^*(y^1, \hat{y}^1) + 3\pi^*(y^2, \hat{y}^2)\\&\quad+ 4\pi^*(y^3, \hat{y}^3) + 4\pi^*(y^4, \hat{y}^4)\end{aligned}\tag{30}$$

Finally, considering the weighting $\lambda : y, \hat{y} \mapsto \|\hat{y}\|_1$, the prediction-weighted TCM is

$$\begin{aligned}&3\pi^*(y^1, \hat{y}^1) + 4\pi^*(y^2, \hat{y}^2)\\&\quad+ 4\pi^*(y^3, \hat{y}^3) + 8\pi^*(y^4, \hat{y}^4).\end{aligned}\tag{31}$$

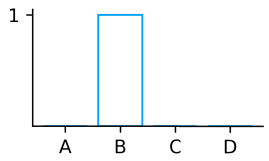
H. CONCLUSION

We regarded the contributions of the single-label confusion matrix as instructions for transforming predictions into labels. This view allowed us to generalize the confusion matrix to multi-label and soft-label frameworks using optimal transport theory.

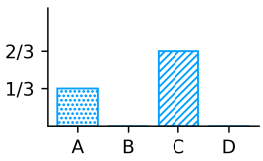
Specifically, the TCM contributions correspond to solving a Kantorovich problem characterized by a cost function that assumes intra-class similarity and inter-class dissimilarity. These broader frameworks often involve multiple optimal transference plans, and we propose selecting the one that maximizes entropy, following the principle of maximum entropy.

We provide an analytical solution ensuring scalability. For each instance, we compute the matrix that solves our optimization problem, and by summing these matrices, we capture the model's overall behavior. The sum of contributions can be weighted based, for example, on the quantity of classes in the label or prediction.

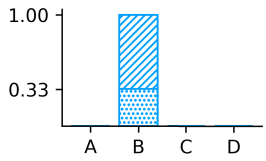
In addition to the theoretical results, we provide a concrete interpretation of the contributions. We also demonstrate that the weighted sum offers the same interpretation as the unweighted sum, differing only in the initial pair of vectors considered.



(a) $y^1 = [0\ 1\ 0\ 0]$ normalized



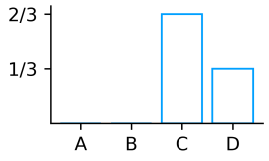
(b) $\hat{y}^1 = [1\ 0\ 2\ 0]$ normalized



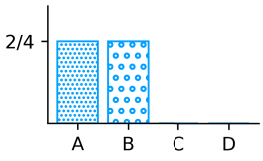
(c) Normalized $\hat{y}^1$ transformed according to $\pi^*(y^1, \hat{y}^1)$

	A	B	C	D
A	0	0	0	0
B	1/3	0	2/3	0
C	0	0	0	0
D	0	0	0	0

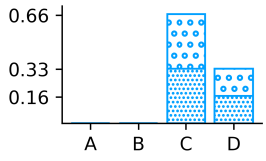
(d) $\pi^*(y^1, \hat{y}^1)$



(e) $y^2 = [0\ 0\ 2\ 1]$ normalized



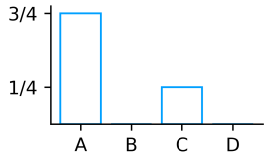
(f) $\hat{y}^2 = [2\ 2\ 0\ 0]$ normalized



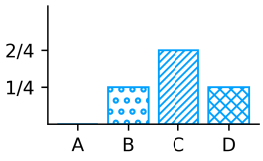
(g) Normalized $\hat{y}^2$ transformed according to $\pi^*(y^2, \hat{y}^2)$

	A	B	C	D
A	0	0	0	0
B	0	0	0	0
C	1/3	1/3	0	0
D	1/6	1/6	0	0

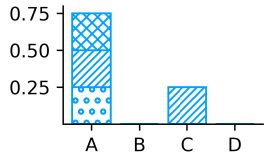
(h) $\pi^*(y^2, \hat{y}^2)$



(i) $y^3 = [3\ 0\ 1\ 0]$ normalized



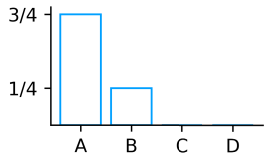
(j) $\hat{y}^3 = [0\ 1\ 2\ 1]$ normalized



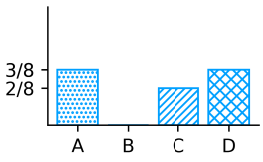
(k) Normalized $\hat{y}^3$ transformed according to $\pi^*(y^3, \hat{y}^3)$

	A	B	C	D
A	0	1/4	1/4	1/4
B	0	0	0	0
C	0	0	1/4	0
D	0	0	0	0

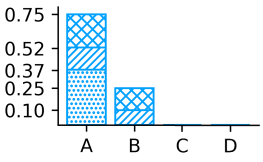
(l) $\pi^*(y^3, \hat{y}^3)$



(m) $y^4 = [3\ 1\ 0\ 0]$ normalized



(n) $\hat{y}^4 = [3\ 0\ 2\ 3]$ normalized



(o) Normalized $\hat{y}^4$ transformed according to $\pi^*(y^4, \hat{y}^4)$

	A	B	C	D
A	3/8	0	3/20	9/40
B	0	0	1/10	3/20
C	0	0	0	0
D	0	0	0	0

(p) $\pi^*(y^4, \hat{y}^4)$

FIGURE 2. **Four instances $(y^1, \hat{y}^1)$, $(y^2, \hat{y}^2)$, $(y^3, \hat{y}^3)$, and $(y^4, \hat{y}^4)$, and their contributions $\pi^*(y^1, \hat{y}^1)$, $\pi^*(y^2, \hat{y}^2)$, $\pi^*(y^3, \hat{y}^3)$, and $\pi^*(y^4, \hat{y}^4)$ are shown. The classes are A, B, C, and D. Labels and predictions are depicted using bar charts. Quantities predicted in class A are shown with dotted bars, circled in B, hatched in C, and crossed in D. The contribution (d) is interpreted as follows: move a quantity of 1/3 from A to B and of 2/3 from C to B. The contributions (h), (l), and (p) are interpreted in the same way.**

IV. STATE-OF-THE-ART CONNECTIONS

This section explores the connections between TCM, CM, MLCM [1], and SCM [27]. We first point out the common aspects of these approaches, then describe the key properties confusion matrices should have.

A. TCM: TOWARDS A UNIFIED FRAMEWORK

We will demonstrate how TCM unifies several propositions under a common theoretical framework.

The assertion that TCM extends CM is based on the following proposition:

Proposition 7. *In the single-label framework,* TCM *and* CM *are identical.*

Other state-of-the-art matrices, such as the MLCM and SCM, also exhibit this property.

To establish contributions of the multi-label confusion matrix, Krstinić, Braović, Šerić, et al. [1] propose MLCM and distinguish four cases, each associated with a specific formula. Considering an instance y and $\hat{y}$, let Y be the set of present classes in y and $\hat{Y}$ be the set of predicted classes in $\hat{y}$. The four cases are: (i) $Y = \hat{Y}$, (ii) $Y \subsetneq \hat{Y}$, (iii) $\hat{Y} \subsetneq Y$, and (iv) None of the previous cases.

Proposition 8. *In case (i), (ii), and (iii)* MLCM *and* TCM *weighted by* $\lambda(y, \hat{y}) = \|y\|_1$ *produce identical contributions.*

Our formula in case (iv) is preferable as it derives from a single, theoretically grounded expression. In contrast, the formula in [1] is specifically constructed for case (iv) without concrete theoretical justification.

In the soft-label framework restricted to probability, Silván-Cárdenas and Wang [27] presents SCM, a matrix with interval entries. While optimal transport isn't mentioned, the practical interpretation of the confusion phenomenon is quite similar. Our approach allows for the computation of SCM contributions, as outlined in the following proposition:

Proposition 9. *Contribution in* SCM *can be computed as follows: for each entry ij, select* $\underline{\pi}$ *and* $\overline{\pi}$ *in* $T^{opt}(y, \hat{y})$ *such that* $\underline{\pi}_{ij} \leq \pi_{ij} \leq \overline{\pi}_{ij}$ *for all* $\pi \in T^{opt}(y, \hat{y})$*, then define entry ij as* $[\underline{\pi}_{ij}, \overline{\pi}_{ij}]$.

Consequently, in a soft-label framework restricted to probability distributions, TCM entries are systematically included within SCM intervals. Each SCM entry ij represents the worst and best possible confusion scenario. This approach provides a pairwise description of errors, as the extreme elements $\underline{\pi}_{ij}$ and $\overline{\pi}_{ij}$ are selected for each pair of classes. In contrast, our proposal offers a more global description by selecting a single element from $T^{\text{opt}}(y, \hat{y})$, denoted as $\pi^*(y, \hat{y})$, which represents the least biased solution according to the principle of maximum entropy, while considering all classes simultaneously. We should also point out that $\pi^*(y, \hat{y})_{ij}$ is generally not the center of the interval $\left[\underline{\pi}_{ij}, \overline{\pi}_{ij}\right]$.

B. DESIRABLE PROPERTIES

We establish that TCM fulfills several important properties.

In the search for fundamental properties to extend the confusion matrix to soft classifications, it is suggested that contributions should fulfill two characteristics [27]:

- Diagonalization: The matrix should be diagonal if, and only if, the assessed data matches perfectly the reference data.
- Marginal sums: Row and column sums should align with the given instance. Specifically, for an instance $(y, \hat{y})$ and its corresponding contribution M, the conditions $\sum_{k=1}^{C} M_{ik} = y_i$ and $\sum_{k=1}^{C} M_{kj} = \hat{y}_j$ should hold for all i and j.

The first property is desirable to identify perfect match situations, while the second is desirable for deriving performance indicators consistent with label and prediction entries, increasing their interpretability [27].

The following proposition demonstrates that TCM satisfies these properties:

Proposition 10. *Considering normalized instance,* TCM *fulfills both the diagonalization and marginal sums properties.*

In contrast, other state-of-the-art matrices do not meet these two criteria. In particular, SCM, with its interval entries, does not satisfy the marginal sums, whether we consider the lower bound, the upper bound, or the center of the interval entries [27].

C. CONCLUSION

In summary, TCM serves as a unifying framework, aligning existing approaches from single-label, multi-label, and soft-label classifications within a single theoretical paradigm.

Furthermore, unlike the other confusion matrices in our comparison, TCM possesses key properties, such as diagonalization and marginal sums, which enhance its interpretability.

V. METRICS

This section proposes, in accordance with our approach, a generalization of Recall, Precision, F1-score, and Accuracy to multi-label and soft-label contexts. To achieve this, we must extend the definitions of True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). We first present these concepts in the single-label case for a given class k before introducing a broader generalization.

A. TRUE POSITIVE, FALSE POSITIVE, TRUE NEGATIVE, AND FALSE NEGATIVE

In the single-label context, the standard definitions for correct and incorrect classifications are as follows:

- TP: An observation that belongs to class k and is correctly classified as such by the model. It indicates that class k is present and has been detected as present.
- FP: An observation that belongs to a class $l \neq k$ but is mistakenly classified as class k. This reflects that class k is absent but has been incorrectly detected as present.
- TN: An observation belonging to class $l \neq k$ that is recognized as belonging to a class $m \neq k$ (regardless of whether $l = m$). It indicates that class k is absent and has been correctly detected as absent.
- FN: An observation belongs to class k but is incorrectly classified as class $m \neq k$. This indicates that class k is present but has been incorrectly detected as absent.

To generalize these concepts for multi-label and soft-label classification, we extend these concepts from binary class to continuous quantities reflecting class memberships.

Definition 2. *Let y and $\hat{y}$ be an instance, and let k be a class. In line with our approach, we consider normalized vectors.*

Let TP_k *be the quantity of class k present in both the label and the prediction*:

$$\text{TP}_k = \min\left(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\right) \tag{32}$$

Let FP_k *be the quantity of class k present in the prediction but absent from the label. In other words, the overestimation of class k:*

$$\text{FP}_k = \frac{\hat{y}_k}{\|\hat{y}\|_1} - \min\left(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\right) \tag{33}$$

Let TN_k *be the quantities predicted in the classes $l \neq k$ that correctly exclude class k. This is calculated by summing the predictions outside class k and subtracting any underestimation of class k (which may be zero). The resulting formula is*:

$$\begin{aligned}\text{TN}_k &= \sum_{l=1, l\neq k}^{C} \frac{\hat{y}_l}{\|\hat{y}\|_1} - \left(\frac{y_k}{\|y\|_1} - \min\left(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\right)\right) \\ &= 1 - \frac{\hat{y}_k}{\|\hat{y}\|_1} - \frac{y_k}{\|y\|_1} + \min\left(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\right)\end{aligned} \tag{34}$$

Let FN_k *be the quantity of class k present in the label but absent from the prediction. In other words, the underestimation of class k:*

$$\text{FN}_k = \frac{y_k}{\|y\|_1} - \min\left(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\right) \tag{35}$$

It is straightforward to see that previous values can be directly computed with $\pi^*(y, \hat{y})$:

$$\begin{aligned} \mathrm{TP}_k &= \pi^*(y,\hat{y})_{kk} \quad \mathrm{FN}_k = \sum_{j=1, j\neq k}^{C} \pi^*(y,\hat{y})_{kj} \\ \mathrm{TN}_k &= \sum_{i,j=1,\, i,j\neq k}^{C} \pi^*(y,\hat{y})_{ij} \quad \mathrm{FP}_k = \sum_{i=1, i\neq k}^{C} \pi^*(y,\hat{y})_{ik} \end{aligned} \tag{36}$$

Interestingly, these formulas align with the single-label confusion matrix: TP for class k corresponds to the diagonal entry k, FN to the sum of off-diagonal entries in row k, TN to the sum of all entries outside row and column k, and FP to the sum of off-diagonal entries in column k.

B. RECALL, PRECISION, F1-SCORE, AND ACCURACY

We use previous definitions to extend traditional metrics.

Definition 3. *Considering the weighting function λ, the Recall, Precision, F1-score, and Accuracy are defined as*:

$$\mathrm{Recall}_k = \frac{\sum_{n=1}^{N} \lambda(y^n, \hat{y}^n)\, \mathrm{TP}_k^n}{\sum_{n=1}^{N} \lambda(y^n, \hat{y}^n)\left(\mathrm{TP}_k^n + \mathrm{FN}_k^n\right)}, \tag{37}$$

$$\mathrm{Precision}_k = \frac{\sum_{n=1}^{N} \lambda(y^n, \hat{y}^n)\, \mathrm{TP}_k^n}{\sum_{n=1}^{N} \lambda(y^n, \hat{y}^n)\left(\mathrm{TP}_k^n + \mathrm{FP}_k^n\right)}, \tag{38}$$

$$\mathrm{F1}_k = 2\frac{\mathrm{Precision}_k * \mathrm{Recall}_k}{\mathrm{Precision}_k + \mathrm{Recall}_k}, \tag{39}$$

$$\mathrm{Accuracy} = \frac{\sum_{n=1}^{N} \lambda(y^n, \hat{y}^n) \sum_{k=1}^{C} \mathrm{TP}_k^n}{\sum_{n=1}^{N} \lambda(y^n, \hat{y}^n)}, \tag{40}$$

where the superscript n denotes quantities computed for the instance n.

We can immediately see that the above metrics can be computed with TCM, in particular,

$$\mathrm{Recall}_k = \frac{\mathrm{TCM}_{kk}}{\sum_{j=1}^{C} \mathrm{TCM}_{kj}}, \tag{41}$$

$$\mathrm{Precision}_k = \frac{\mathrm{TCM}_{kk}}{\sum_{i=1}^{C} \mathrm{TCM}_{ik}}, \tag{42}$$

$$\mathrm{Accuracy} = \frac{\sum_{k=1}^{C} \mathrm{TCM}_{kk}}{\sum_{i,j=1}^{C} \mathrm{TCM}_{ij}}. \tag{43}$$

Again, these formulas align with the single-label confusion matrix.

C. CONCLUSION

In summary, we propose a generalization of widely used metrics that can be computed directly from TCM. Moreover, the derived formulas are analogous to those in the single-label case, enhancing the consistency of our approach.

VI. GUIDELINES FOR READING

This section provides useful guidelines for interpreting the confusion matrix, especially when dealing with imbalanced datasets. These guidelines focus on ordering classes when plotting the matrix and normalization.

Typically, the confusion matrix is plotted using the test set [30] to evaluate the model's performance while the model is trained on the training set. The distribution of classes in both the training and test datasets significantly impacts the values in the confusion matrix. To illustrate this, we consider a simple scenario involving a standard neural network learning process. This process does not include any particular strategies for handling rare or specific classes in the training set, such as sampling techniques or loss weighting.

A. TEST AND TRAINING EFFECT

This subsection introduces the concepts of test and training effects, leading to recommendations for plotting confusion matrices.

The classes most prevalent in the test set generate rows with higher overall values, as the model is tested more frequently on these classes, making more errors and correct predictions. The extreme case occurs when only one class is present in the test set, leading to only one non-zero row in the confusion matrix. We refer to this phenomenon as the test effect. Table 3 (a) illustrates this phenomenon.

Additionally, the classes most prevalent in the training set generate columns with higher overall values, as errors from these prevalent classes contribute more to the loss value during the training process [31]. The extreme case occurs when only one class is predicted, leading to only one non-zero column in the confusion matrix. We refer to this phenomenon as the training effect. Table 3 (b) illustrates this phenomenon.

Based on the test set, we suggest ordering the vertical axis from the most common class at the top to the rarest class at the bottom. Based on the training set, we also propose ordering the horizontal axis from the most common class on the left to the rarest class on the right. Due to test and training effects, the values closer to the top-left corner are likely to be higher. Specifically, for diagonal entries, entry i is likely to be higher than entry $i+1$. For off-diagonal entries, ignoring diagonal ones, an entry ij is likely to be higher than the one to its right or below it and lower than the one to its left or above it. Table 3 (c) illustrates this order. We refer to this expected organization as test-training ranking.

When values deviate from this organization, it highlights specific types of errors, as these entries stand out despite the class distribution in the datasets. For the classifier, such errors might indicate classes that are more difficult or easier to recognize than others (e.g., in an image classification problem, the `Chameleon` class is likely more difficult to distinguish than the `Dog` class) or that one class closely resembles another (e.g., `Chameleon` often includes images of camouflaged chameleons on leaves, leading to a high value of the entry `Chameleon`,`Tree`).

The test-training ranking has certain limitations. First, the ranking can be preserved, while the model's errors may arise from factors beyond class distribution. Second, when an entry deviates from the test-training ranking, it is not always clear whether it is because the considered entry is too high or because the previous entry is too low.

TABLE 3. Illustration of the test and training effects on the confusion matrix with classes A, B, and C. A is the majority class, and C is the minority class in the test set, while B is the majority class and A is the minority class in the training set. The darker the color, the higher the entry.

(a) Test effect, no training effect

	B	C	A
A			
B			
C			

(b) No test effect, training effect

	B	C	A
A			
B			
C			

(c) Test effect, training effect

	B	C	A
A			
B			
C			

In the following subsection, we will see that the normalization process can provide a more precise understanding of the results.

B. NORMALIZATION

It could be interesting to capture errors that are not due to test or training effects. We begin by demonstrating how normalization can help to extract this information. Then, we explain the meaning of row and column normalization in the traditional confusion matrix and TCM.

Let M be a confusion matrix, and let i and j be two classes, which can possibly be the same. The row-normalized and column-normalized matrices are commonly defined as:

$$M_{ij}^{\text{row}} := \frac{M_{ij}}{\sum_{k=1}^{C} M_{ik}} \quad M_{ij}^{\text{col}} := \frac{M_{ij}}{\sum_{k=1}^{C} M_{kj}} \tag{44}$$

Let's show how row normalization can help capture errors that are not due to test effects. Adding more data to the test set labeled as i will increase the values in the i-th row of the confusion matrix. However, if the initial test set is a representative sample, this should not affect the distribution of values in that row. For instance, consider a classification problem with three classes: A, B, and C, and its corresponding confusion matrix. The first row, denoted r_A, shows that class A is well-recognized, slightly confused with B, and not confused with C. Doubling the amount of label A in the test set should lead to the same conclusions, with the new row approximately equal to $2r_A$. Thus, changes in label distribution can be modeled by multiplying the rows of the confusion matrix. Importantly, a row-normalized matrix remains invariant under this transformation:

$$\left(\begin{pmatrix} \alpha_1 & 0 & \cdots & 0 \\ 0 & \alpha_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \alpha_C \end{pmatrix} M\right)^{\text{row}} = M^{\text{row}}, \tag{45}$$

where α_i are strictly positive real values. M^{row} is likely to resemble Table 3 (b). In this sense, row normalization eliminates the test effect.

Let's show how column normalization can help to capture errors that are not due to the training effect. As with row normalization, changes in prediction distribution are modeled by multiplying the columns of the confusion matrix, and a column-normalized matrix remains invariant under this transformation:

$$\left(M \begin{pmatrix} \beta_1 & 0 & \cdots & 0 \\ 0 & \beta_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \beta_C \end{pmatrix}\right)^{\text{col}} = M^{\text{col}}, \tag{46}$$

where β_i are strictly positive real values. M^{col} is likely to resemble Table 3 (a). In this sense, column normalization eliminates the training effect.

In the case of the traditional single-label confusion matrix, the meaning of row and column normalization is the following: M_{ij} represents the number of predictions j with label i, and $\sum_{k=1}^{C} M_{ik}$ is the total number of instances with the label i. Therefore, among all instances labeled i, the row-normalized value M_{ij}^{row} gives the proportion of instances predicted as j:

$$M_{ij}^{\text{row}} = \frac{\text{number of predictions } j \text{ with label } i}{\text{number of labels } i} \tag{47}$$

Similarly, among all predictions j, the column-normalized value M_{ij}^{col} gives the proportion of instances with label i:

$$M_{ij}^{\text{col}} = \frac{\text{number of predictions } j \text{ with label } i}{\text{number of predictions } j} \tag{48}$$

In our generalized case, the core idea remains the same. Formally, the term ij of the matrix TCM^{row} is:

$$\begin{aligned} \text{TCM}_{ij}^{\text{row}} &= \frac{\text{TCM}_{ij}}{\sum_{k=1}^{C} \text{TCM}_{ik}} \\ &= \frac{\sum_{n=1}^{N} \lambda(y^n, \hat{y}^n) \pi^*(y^n, \hat{y}^n)_{ij}}{\sum_{k=1}^{C} \sum_{n=1}^{N} \lambda(y^n, \hat{y}^n) \pi^*(y^n, \hat{y}^n)_{ik}} \\ &= \frac{\sum_{n=1}^{N} \lambda(y^n, \hat{y}^n) \pi^*(y^n, \hat{y}^n)_{ij}}{\sum_{n=1}^{N} \lambda(y^n, \hat{y}^n) \sum_{k=1}^{C} \pi^*(y^n, \hat{y}^n)_{ik}} \\ &= \frac{\sum_{n=1}^{N} \lambda(y^n, \hat{y}^n) \pi^*(y^n, \hat{y}^n)_{ij}}{\sum_{n=1}^{N} \lambda(y^n, \hat{y}^n) y_i / \|y\|_1}. \end{aligned} \tag{49}$$

Considering normalized vectors $y^n/\|y^n\|$ and $\hat{y}^n/\|\hat{y}^n\|$, $\pi^*(y^n, \hat{y}^n)_{ij}$ estimates the quantity predicted in class j that should be in class i. Consequently, given normalized vectors and weighted aggregation, TCM_{ij} captures the predicted quantity in class j that should be in class i in the whole dataset. Due to marginal properties, $\sum_{k=1}^{C} \text{TCM}_{ik}$ gives the total quantity of label i across the dataset. Thus, the row-normalized value $\text{TCM}_{ij}^{\text{row}}$ represents the proportion of quantities predicted as j among all labels i:

$$\text{TCM}_{ij}^{\text{row}} = \frac{\text{quantities predicted in } j \text{ expected in } i}{\text{quantity of label } i} \tag{50}$$

Similarly, among all quantities predicted in class j, the column-normalized value $\text{TCM}_{ij}^{\text{col}}$ represents the proportion of quanties labeled i:

$$\text{TCM}_{ij}^{\text{col}} = \frac{\text{quantities predicted in } j \text{ expected in } i}{\text{quantity of prediction } j} \tag{51}$$

Finally, the interpretation of the row or column normalization of TCM is the continuous counterpart of the traditional confusion matrix.

C. CONCLUSION

We introduce the test-training ranking, a method for plotting the confusion matrix that helps differentiate between errors caused by the class distribution of the datasets and those due to other factors. Although this approach is straightforward to apply, it has some limitations.

We also demonstrate how normalization can eliminate the test or training effect, providing a clearer understanding of the results. Additionally, we show that the interpretation of row or column normalization of TCM is a generalization of the traditional single-label confusion matrix.

VII. EXPERIMENTAL SETUP

This section provides the essential details of the experimental setup. We start by outlining the baselines and the criteria for comparison, focusing on a method for comparing the F1-scores generated by each matrix. Next, we describe the datasets, models, and training processes used in the experiments.

We conduct three experiments: two comparative analyses to highlight the differences and similarities with state-of-the-art approaches and one case study that illustrates the real-world application of TCM.

A. BASELINES

We compare TCM to three baselines: MLCM, MLCT, and SCM. There is no need to experimentally compare TCM with CM, as they are identical in the single-label context (according to Proposition 7).

MLCT consists of two matrices, one for recall and one for precision, denoted here as MLCT^{R} and MLCT^{P}, respectively.

To extend SCM for multi-label and soft-label frameworks, we normalize each instance (i.e., considering $y/\|y\|_1$ and $\hat{y}/\|\hat{y}\|_1$). Moreover, the original matrix entries are intervals represented by the interval center and half-width, i.e., $[a, b]$ is represented by $(a+b)/2 \pm (b-a)/2$. We omit the uncertainty $\pm(b-a)/2$ and consider only the interval center $(a+b)/2$ when a point value is required, such as for computing an F1-score. We refer to this extension as SCM^{ext}.

Finally, we denote the transport-based matrix weighted by $\lambda : y, \hat{y} \mapsto 1$ as TCM^{one}, the matrix weighted by $\lambda : y, \hat{y} \mapsto \|y\|_1$ as TCM^{lab}, and the matrix weighted by $\lambda : y, \hat{y} \mapsto \|\hat{y}\|_1$ as TCM^{pred}.

B. BASELINE COMPARISON CRITERIA

To emphasize the differences between confusion matrices, we select challenging datasets that encompass a large number of classes. While these numerous classes introduce complex errors, they also complicate the process of matrix comparison.

To address this challenge, we propose three criteria for comparison:

- A direct and partial display of the five most common classes,
- A complete heat map representation of each matrix,
- A comparison based on F1-scores.

The F1-score criterion allows easy comparison of matrices, even when faced with a large number of classes, as will be discussed in the following subsection.

C. F1-SCORE CRITERION

The greater the differences in computed F1-scores between matrices, the more their descriptions of model behavior differ.

In alignment with the single-label matrix framework, we propose the following definitions for any confusion matrix M and class k: Recall is defined as $M_{kk}/\sum_{j=1}^{C} M_{kj}$, and Precision is defined as $M_{kk}/\sum_{k=1}^{C} M_{ik}$. Therefore, the F1-score for class k can be calculated as the harmonic mean of Precision and Recall for any matrix M.

Comparing derived F1-scores is not straightforward: some matrices consistently yield high F1-scores, while others yield lower ones. Furthermore, the distribution of F1-scores varies across different matrices. As a result, a direct comparison between two sets of F1-scores is not meaningful.

A possibility is to rank classes by decreasing F1-scores for each matrix and then compare rankings. However, this approach has limitations. If F1-scores are identical or nearly identical, different rankings may not indicate significant differences in the descriptions of model behavior.

We propose the following process to effectively compare the F1-scores of matrices M_a and M_b, as used in Fig. 6 and Fig. 7:

1) Considering the matrix M_a and its F1-scores, rank the classes in descending order of F1-score. This order is called $Ranking_a$. Similarly, define $Ranking_b$ for M_b.
2) Draw a bar chart of M_a's F1-scores ordered by $Ranking_b$.
3) In a second figure, draw a bar chart of M_b's F1-scores following $Ranking_a$.

If M_a and M_b are similar, the bar charts will appear approximately ordered from highest to lowest scores. In contrast, if $Ranking_a$ differs from $Ranking_b$, the bar charts will be disordered, indicating different analyses.

Moreover, the visual representation highlights the extent of disparities between F1-scores. For instance, if class i's F1-score is very high for M_a and very low for M_b, in the first plot, a large bar will be surrounded by small bars (conversely in the second one), revealing a significant disparity. In contrast, if some scores are similar within their matrix but differently ranked, the diagram will remain relatively organized, indicating minimal disparity.

To quantify the resulting disorder, we introduce the function Δ. Let $F1^{a,a}$ be the vector of M^a's F1-scores ordered by $Ranking_a$ (resulting in entries arranged in decreasing order), and let $F1^{a,b}$ be M^a's F1-scores ordered by $Ranking_b$. The function Δ is defined as follows:

$$\Delta : M^a, M^b \mapsto 1000 * \frac{\|F1^{a,a} - F1^{a,b}\|_1}{C\|F1^{a,a}\|_1}, \qquad (52)$$

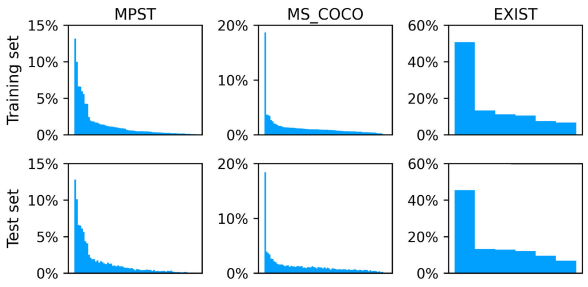


FIGURE 3. Bar plots showing the class distributions of the training and test datasets. The vertical axis represents the proportion of each class, while the horizontal axis lists all the classes shared between the train and test sets. The test sets appear to be representative of their training set, as the class distributions across them are very similar.

where the value 1000 is chosen for convenience in the plots. Since this quantity is normalized, it is comparable across all confusion matrices and datasets.

D. DATASETS

We use two datasets to compare confusion matrices: Movie Plot Synopses with Tags (MPST) and Microsoft Common Objects in Context (MS-COCO). We use a third dataset for a case study: the English version task 3 of the sEXism Identification in Social neTworks (EXIST) dataset from CLEF 2023. We utilized the splits provided with each dataset. Table 4 lists the training sets' statistics. All datasets are imbalanced, as depicted in Fig. 3.

MPST corpus [32] contains plot synopses for 14K movies, each associated with one or more of 71 tags. These tags are non-redundant and exclusively capture properties of the movie plots, avoiding any metadata or attributes unrelated to the plot. Each plot synopsis in the corpus is at least ten sentences long.

The Microsoft Common Objects in Context (MS-COCO) dataset [33] is a large-scale object detection, segmentation, key-point detection, and captioning dataset. The dataset consists of 328K images annotated with 80 objects. We use the MS-COCO dataset as a multi-label dataset, where each present class corresponds to an object within a bounding box, a previously seen procedure [34].

The scientific event sEXism Identification in Social neTworks (EXIST) at CLEF 2023 [18] involved categorizing several thousand tweets, each annotated by six annotators to avoid bias. The goal of the third task is to identify types of sexism. In its soft-label version, each label is a vector in $[0, 6]^6$, with one entry per class, resulting from the sum of six vectors in $[0, 1]^6$, one vector per annotator. An annotator can vote for one or more classes; see [35] for details. As a result, labels can have different norms.

E. MODELS

We select state-of-the-art models for each experiment: DeBERTa [37] for MPST, SqueezeNet 1.1 [38] for MS-COCO, and BERTweet for EXIST. Model performances are listed in Table 5. We use pre-trained models from Hugging Face and fine-tune them for specific tasks.

TABLE 4. Descriptive statistics of the training sets. The abbreviation Card. stands for cardinality, which is the average number of classes per label [36]. In line with this definition, the cardinality of EXIST is calculated as the sum of all annotators' votes divided by the number of annotators.

Dataset	Framework	Modality	Classes	Card.	Size
MPST	Multi-Label	Text	71	3	12K
MS-COCO	Multi-Label	Images	80	2.9	117K
EXIST	Soft-Label	Text	6	1.15	3K

F. TRAINING PROCEDURE

Multi-label deep learning typically uses a sigmoid activation function with binary cross-entropy applied to the final neural layer [34]. We follow this approach for multi-label tasks. For the soft-label dataset EXIST, the raw data provide the votes of each annotator. Instead of training the model on the sum of the votes, we trained it on a binary vector of size $6 * 6 = 36$, representing the concatenated votes of all annotators. This method allows for standard multi-label training. The aggregation of votes is only performed during the test phase, producing a vector of size 6, representing the sum of the votes for each class.

To help classifiers predict rare classes, we use two strategies. Firstly, PyTorch provides the `pos_weight` parameter, which modifies the binary cross-entropy formula by adjusting the weight of positive labels (see BCEWithLogitsLoss). We used the weighting recommended by PyTorch, which is the number of negative labels divided by the number of positive labels for each class, clipped at 200 for MPST and 100 for the other datasets.

Secondly, we adjust the output threshold, as the standard threshold of 0.5 is rarely optimal when dealing with imbalanced datasets [39], [40], [41]. We applied the method recommended by Johnson and Khoshgoftaar [39] label-wise: using the training data to find the classification threshold that maximizes the geometric mean of specificity and recall (i.e., $\sqrt{\text{Specificity} * \text{Recall}}$). As shown in Table 5, in our cases, this strategy had an overall neutral effect on F1-scores, increased recall, and decreased precision.

We train the models for 75 epochs, retaining the model weights that maximize the weighted F1-score on the test set. This procedure is uniformly applied across all experiments using standard hyperparameters without additional fine-tuning: the AdamW optimizer, a batch size of 64 for DeBERTa and BERTweet, 128 for SqueezeNet, and a learning rate of 1×10^{-5} for DeBERTa and BERTweet, and 1×10^{-4} for SqueezeNet with an image size fixed at 224×224. The seed is set to 42 across all experiments.

G. PERFORMANCE COMPARISONS

Despite differences between models and evaluation metrics, our model performance aligns with expectations.

TABLE 5. Performance achieved by different models on various datasets. Gmean refers to the threshold-moving method [39] used to predict rare classes (see Subsection VII-F). F1-score, Precision, and Recall for EXIST are computed using raw annotators' votes before vote aggregation.

Model	Dataset	Average	Threshold	F1	Precision	Recall
DeBERTa	MPST	Macro	0.5	0.18	0.14	0.28
			Gmean	0.19	0.14	0.31
		Weighted	0.5	0.34	0.30	0.43
			Gmean	0.35	0.18	0.50
SqueezeNet	MS-COCO	Macro	0.5	0.61	0.56	0.67
			Gmean	0.60	0.53	0.69
		Weighted	0.5	0.66	0.62	0.71
			Gmean	0.65	0.60	0.72
BERTweet	EXIST	Macro	0.5	0.48	0.40	0.63
			Gmean	0.46	0.38	0.67
		Weighted	0.5	0.58	0.51	0.69
			Gmean	0.58	0.52	0.70

For MPST, Rahman et al. [42] use lightweight Vanilla Neural Networks and RoBERTa language models in their pipeline to predict movie tags, achieving an F1-score of 0.38. Similarly, Rahman and Malik [43] achieve a top-3 F1-Micro of 0.37 with a model named CNN-FE + FastText. In our experiment, we achieve a weighted F1-score of 0.35.

For the multi-label classification problem with MS-COCO, an ImageNet-pretrained ResNet-50 [44] achieves an F1-example of 0.71, F1-Micro of 0.69, and F1-Macro of 0.64 [45]. Additionally, a TresNet [46] achieves a weighted F1-score of 0.79 [34]. In comparison, we achieve a weighted F1-score of 0.65 and an F1-Macro of 0.60. While these performances are lower than those reported in the cited papers, note that SqueezeNet is a much smaller model and that it was trained with smaller image dimensions. Moreover, we paid particular attention to rare classes (with weighted loss and threshold-moving), which may explain the lower performance.

Considering the EXIST challenge, the best team achieved an F1-score of 0.62 [18], while we achieved 0.58, placing us in the top 20% of participants.

VIII. EXPERIMENTAL COMPARISONS: TCM VS. STATE-OF-THE-ART MATRICES

The objective of our experiments is to compare MLCM, MLCT, and SCM with TCM in multi-label classification contexts. We evaluate the confusion matrices based on three key aspects: the display of the five most frequent classes, heat maps, and reorganized F1-scores. Each of these aspects is explored in detail in the subsections below.

In general, we observe that the differences between the matrices are more pronounced on the MPST dataset with DeBERTa compared to the MS-COCO dataset with SqueezeNet. This is likely due to the greater complexity of model errors, as indicated by the F1-scores in Table 5.

A. MATRIX DISPLAYS

Table 6 and 7 show the non-normalized values of the confusion matrices. All matrix entries discussed below are highlighted in bold.

The values vary significantly. For instance, in Table 6, the entry `C1C1` ranges from 121.6 (in TCM^{one}) to 757 (in MLCT^{R} or MLCT^{P}), while `C2C1` ranges from 22.1 (in TCM^{one}) to 98.2 (in MLCT^{P}). Similar patterns are observed in Table 7.

Some matrices provide more similar descriptions of model behavior than others. For instance, TCM^{pred} and MLCT^{P} have close values, as do TCM^{one} and SCM^{ext}, which aligns with Proposition 9. In contrast, MLCT^{R} has high diagonal values but low non-diagonal entries, setting it apart from the transport-based matrices.

Proposition 8 demonstrates that the contribution formulas of TCM^{lab} and MLCM are similar. However, certain values, such as `C5C5` and `C5C2` in Table 6 differ significantly (248.4 in MLCM versus 177.6 in TCM^{lab}, and 13.5 in MLCM versus 20.1 in TCM^{lab}). This indicates that formula differences can significantly impact the final values. Generally, MLCM has higher diagonal and lower non-diagonal values than TCM^{lab}, favoring agreement over error.

The uncertainties in SCM^{ext} can be significant and equal to the center of the interval (23.6 ± 14.6 in `C2C3` Table 6, 15.4 ± 12.5 in `C1C2`, or 3.4 ± 3.4 in `C5C2` Table 7), complicating the interpretation. Although when considering SCM^{ext}, we could focus only on central values, this approach would lack justification. In contrast, TCM is easier to interpret, with each entry corresponding to a single value, theoretically justified and fulfilling desirable properties.

Weighting impacts the values of TCM^{one}, TCM^{lab}, and TCM^{pred}. The weighting corresponds to 1 in TCM^{one}, $\|y\|_1$ in TCM^{lab}, and $\|\hat{y}\|_1$ in TCM^{pred}. Since $\|y\|_1$ and $\|\hat{y}\|_1$ are always at least 1, values in TCM^{lab} and TCM^{pred} are systematically greater than in TCM^{one}. Additionally, TCM^{pred} values are higher than those of TCM^{lab} because, in both experiments, Recall is higher than Precision, reflecting the model's over-prediction.

In summary, while the matrices yield different raw values, they generally align with a version of TCM, with the exception of MLCT^{R}. TCM^{lab} and MLCM may differ despite some shared characteristics outlined in Proposition 8. SCM^{ext} can be difficult to interpret, and weighting affects transport-based matrix results. However, a direct and partial comparison of

the matrices alone is insufficient to validate these findings. Additional insights can be gained by comparing heat maps and F1-scores.

B. HEAT MAPS

Fig. 4 and Fig. 5 display heat maps of matrices.

The heat maps appear similar. Luminous and non-luminous points tend to occupy comparable coordinates, indicating consistency in qualitative analysis. However, the variation in brightness suggests differences in quantitative findings.

Some matrices favor true positive entries across both experiments. This is particularly apparent in the row-normalized matrix. For instance, the diagonal of MLCT^{R} is notably bright, whereas the diagonal of MLCT^{P} is more subdued. Similarly, considering Fig. 4, this trend is also observed with MLCM compared to TCM^{one}, TCM^{lab}, TCM^{pred}, and SCM^{ext}.

Some matrices highlight errors more than others. In Fig. 5, the off-diagonal terms generally appear dark, whereas their brightness varies slightly in Fig. 4. We observe that MLCT^{P} and TCM^{pred} seem darker, while others, particularly TCM^{one}, are lighter.

In summary, the conclusions about model behavior are qualitatively similar but quantitatively different. Quantification is crucial, as model diagnostics rely on the relative comparison of matrix entries. Significant insights can be derived from comparing F1-scores, which will be further explored in the following subsection.

C. REORDERED F1-SCORES

Fig. 6 and Fig. 7 show the reorganized F1-scores, with the resulting disorder measured by the function Δ introduced in Section VII-C.

F1-score distributions can vary significantly. For example, diagonal plots in Fig. 6 show that MLCM and MLCT^{R} have high F1-scores, while MLCT^{P} and SCM^{ext} have lower ones.

No two matrices produce the same ranking, confirming the quantitative differences noted in Subsection VIII-B. Some are small, such as between TCM^{pred} and MLCT^{P} in Fig. 7, allowing for a relative decrease in the plot bars. Others are larger, leading to a loss of monotony, such as between TCM^{one} and MLCT^{R} in Fig. 6.

Some matrices yield similar rankings. For instance, TCM^{pred} and MLCT^{P} are closely aligned (in Fig. 6 $\Delta(\text{TCM}^{\text{pred}}, \text{MLCT}^{\text{P}}) = \Delta(\text{MLCT}^{\text{P}}, \text{TCM}^{\text{pred}}) = 2$, in Fig. 7 $\Delta(\text{TCM}^{\text{pred}}, \text{MLCT}^{\text{P}}) = \Delta(\text{MLCT}^{\text{P}}, \text{TCM}^{\text{pred}}) = 1$). Similarly, SCM^{ext} and TCM^{one} also show some alignment (in Fig. 6 $\Delta(\text{SCM}^{\text{ext}}, \text{TCM}^{\text{one}}) = 6$ and $\Delta(\text{TCM}^{\text{one}}, \text{SCM}^{\text{ext}}) = 7$, in Fig. 7 $\Delta(\text{SCM}^{\text{ext}}, \text{TCM}^{\text{one}}) = 3$ and $\Delta(\text{TCM}^{\text{one}}, \text{SCM}^{\text{ext}}) = 2$).

Proposition 8 demonstrates that TCM^{lab} and MLCM share some formulas. However, the remaining differences can lead to distinct analyses (for example, in Fig. 6 $\Delta(\text{TCM}^{\text{lab}}, \text{MLCM}) = 11$ and $\Delta(\text{MLCM}, \text{TCM}^{\text{lab}}) = 12$).

Weighting has a notable impact on transport-based matrices. TCM^{one}, TCM^{lab}, and TCM^{pred} show disorder scores between 12 and 24 in Fig. 6, and between 2 and 11 in Fig. 7.

In summary, the F1-score analysis highlights differences in the matrices, reflected in their varying F1-score distributions and rankings. This confirms that their quantitative analysis of the model's behavior varies. Some matrices, such as TCM^{pred} and MLCT^{P}, or TCM^{one} and SCM^{ext}, produce similar rankings. Proposition 8 suggests similarities between TCM^{lab} and MLCM, but this is not fully confirmed by the experiments. Finally, the distinct distributions and rankings in transport-based matrices emphasize the importance of considering weighting when assessing model performance.

D. CONCLUSION

The confusion matrices appear to be roughly similar in the two experiments. For instance, the heatmaps exhibit nearly identical qualitative analyses, and the rearranged F1-scores are not entirely disordered. This suggests a common intuition across the different approaches.

This observation is further reinforced by the fact that some state-of-the-art matrices closely resemble specific transport-based matrices (MLCT^{P} seems close to TCM^{pred}, SCM^{ext} to TCM^{one}, and to a lesser extent MLCM to TCM^{lab}). Specifically, transport matrices share a common understanding of agreement and error (as recorded for each instance in $\pi^*(y, \hat{y})$) and differ only in the weighting λ applied. In other words, they vary according to the vectors considered at each instance to establish the model's behavior:

$$\left(\frac{y^n}{\|y^n\|_1}, \frac{\hat{y}^n}{\|\hat{y}^n\|_1}\right), \quad \left(y^n, \hat{y}^n \frac{\|y^n\|_1}{\|\hat{y}^n\|_1}\right), \quad \text{or } \left(y^n \frac{\|\hat{y}^n\|_1}{\|y^n\|_1}, \hat{y}^n\right), \tag{53}$$

see Subsection III-F for details. Thus, the proximity of confusion matrices to specific transport matrices demonstrates that a similar understanding of agreement and error is shared in practice. The vectors considered to establish conclusions are likely the main source of difference. Our method provides a unifying framework by approximating state-of-the-art approaches with a unified theory.

Nevertheless, differences exist, as evidenced by the raw matrix displays, the varying heatmap intensities, and the remaining disorder in the F1-scores. On the one hand, some of these differences are related to the vectors considered to establish the model's behavior (the weighting in our approach). For instance, we observed that TCM^{one}, TCM^{pred}, and TCM^{lab} yield different conclusions. On the other hand, the shared understanding of agreement and error across the matrices is similar, reflecting a common intuition, but not identical. For each approach, the formalization of this intuition relies heavily on the definitions chosen to quantify agreement and error.

These experiments do not determine which definitions are the most appropriate, i.e., which matrix provides the best behavioral description of the model. They establish the existence of a shared intuition for describing this behavior.

TABLE 6. Partial display of the different confusion matrices of DeBERTa on MPST. Only the five most common classes in the training set are displayed. The matrices are not normalized. Entries cited in the body of the article are in bold.

(a) MLCM

	C1	C2	C3	C4	C5	⋯
C1	550.9	40.1	46.6	64.9	30.2	⋯
C2	32.4	437.9	39.7	43.1	22.3	⋯
C3	17.7	23.9	253.5	37.4	25.3	⋯
C4	14.6	12.1	24.5	299.9	11.3	⋯
C5	24.4	**13.5**	22.7	22.3	**248.4**	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋱

(b) $\mathrm{TCM}^{\mathrm{lab}}$

	C1	C2	C3	C4	C5	⋯
C1	412.7	49.8	56.6	69.4	37.5	⋯
C2	40.5	327.1	47.7	43.6	30.6	⋯
C3	21.6	27.7	189.1	40.7	26.8	⋯
C4	17.1	13.3	29	237.7	12.9	⋯
C5	29.8	**20.1**	26.4	22.8	**177.6**	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋱

(c) $\mathrm{SCM}^{\mathrm{ext}}$

	C1	C2	C3	C4	C5	⋯
C1	**121.6**	32.6±9.3	28.8±16.9	36.7±12.3	21.1±12.4	⋯
C2	24±8	86.9	**23.6±14.6**	18.4±8.2	17.3±10.5	⋯
C3	11.8±4.8	11.7±6.4	46.7	23.2±6.3	12±7.5	⋯
C4	9.7±4.1	5.5±3.2	19±7	104.1	6.8±3.4	⋯
C5	12.4±5.9	8.5±6.8	10.3±8.8	9.8±5.3	34.2	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋱

(d) $\mathrm{TCM}^{\mathrm{one}}$

	C1	C2	C3	C4	C5	⋯
C1	**121.6**	29.7	23.8	34.3	17	⋯
C2	**22.1**	86.9	19.4	16.7	13.7	⋯
C3	10.7	10	46.7	21.6	9.6	⋯
C4	8.7	4.8	16.9	104.1	5.7	⋯
C5	11.1	6.6	7.7	8.4	34.2	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋱

(e) $\mathrm{MLCT}^{\mathrm{R}}$

	C1	C2	C3	C4	C5	⋯
C1	**757**	31.3	34.1	61	18.6	⋯
C2	32.2	579	30.4	42.3	11.4	⋯
C3	21.6	25.1	333	33	20.2	⋯
C4	17	15.5	12.6	400	10.9	⋯
C5	32.9	17.7	18.8	24.4	301	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋱

(f) $\mathrm{MLCT}^{\mathrm{P}}$

	C1	C2	C3	C4	C5	⋯
C1	**757**	141.1	162.8	113.6	133.2	⋯
C2	**98.2**	579	118	49.5	108.1	⋯
C3	50.2	47	333	70.5	51.7	⋯
C4	42.8	22.9	91.9	400	37.1	⋯
C5	35.7	41.2	47.4	25.7	301	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋱

(g) $\mathrm{TCM}^{\mathrm{pred}}$

	C1	C2	C3	C4	C5	⋯
C1	712.1	137.6	154.1	113.3	129.4	⋯
C2	97	538.3	114.8	56.2	100.7	⋯
C3	53.2	51.9	311.8	67.8	56.5	⋯
C4	42.9	25.2	86.6	365	38.2	⋯
C5	40.8	41	48.6	28.7	286.4	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋱

What indicates that our definition is superior to pre-existing ones are sections III to VI. These sections demonstrate that we have formulated a very general definition based on minimal assumptions, which are intuitive and meet desirable properties not satisfied by other matrices.

IX. EXIST CASE STUDY

This section proposes interpreting the results of BERTweet on EXIST task 3 using transport-based matrices shown in Table 8. We start by examining non-normalized matrices before analyzing the results for normalized matrices.

According to Subsection VI-A, we order the vertical axis from the most common class to the rarest class, based on the test set, and the horizontal axis from the most common class to the rarest class, based on the training set. In this experiment, the class rankings are identical across both datasets.

A. NON-NORMALIZED MATRICES

This subsection analyses the non-normalized matrices (a), (b), and (c).

The class C1 is more frequently recognized, underestimated, and overestimated compared to the other classes, as indicated by the significant values of the row and column C1 across matrices (a), (b), and (c). These phenomenons primarily arise because C1 is largely dominant (constituting 47% of the training set and 43% of the test set).

In all three matrices, the test-training ranking shows that class C5 is poorly understood by the classifier (many entries in row or column C5 do not follow the test-training ranking). Specifically, while the model seems to distinguish C5 from C4 (since the C5C4 and C4C5 entries have low values), it tends to confuse C5 with other classes (C2, C3, and C6 have large values in both row and column C5). Additionally, we note that when C4 is underestimated, C2 and C3 are overestimated. We also observe that C3 is underestimated in favor of C6.

The classifier has correctly grasped the exclusivity of the non-sexist class. Specifically, we observe that the values of column C1 in matrix (c) are not as important as in the other non-normalized matrices. The weighting $\lambda : y, \hat{y} \mapsto \|\hat{y}\|_1$ reduces the importance of the C1 column in favor of the

TABLE 7. Partial display of the different confusion matrices of SqueezeNet on MS-COCO. Only the five most common classes in the training set are displayed. The matrices are not normalized. Entries cited in the body of the article are in bold.

(a) MLCM

	C1	C2	C3	C4	C5	⋯
C1	2114.1	23.7	30.1	12.9	12.6	⋯
C2	14.5	336.5	6.9	9.5	12.4	⋯
C3	14.8	4.3	331	2.5	3.2	⋯
C4	6.8	6.1	1.7	318	7.7	⋯
C5	5.1	9.9	2.1	10.4	216.6	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋱

(b) TCM^{lab}

	C1	C2	C3	C4	C5	⋯
C1	1913.6	34.6	38.9	19	20.5	⋯
C2	18.7	292.4	7.3	12.3	15.9	⋯
C3	18.5	4.5	302.7	2.8	3	⋯
C4	9	9.7	2.1	270.8	11	⋯
C5	8	11.9	1.9	11.7	183.3	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋱

(c) SCM^{ext}

	C1	C2	C3	C4	C5	⋯
C1	650.3	**15.4±12.5**	16.3±10.1	8.4±7.1	8.3±8	⋯
C2	5.5±3.3	59.1	2.3±1.6	4.4±3.6	5.3±4.3	⋯
C3	7.1±4.2	1.4±1.4	87.6	0.8±0.8	0.9±0.9	⋯
C4	2.5±2	3.1±2.5	0.6±0.6	51.6	3.9±3.8	⋯
C5	2.3±1.9	**3.4±3.4**	0.5±0.5	3.6±3.3	32.3	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋱

(d) TCM^{one}

	C1	C2	C3	C4	C5	⋯
C1	650.3	12.1	15	6	5.6	⋯
C2	4.8	59.1	2	3.2	3.6	⋯
C3	6.6	1	87.6	0.6	0.6	⋯
C4	1.9	2.3	0.4	51.6	2.5	⋯
C5	1.7	2.4	0.4	2.7	32.3	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋱

(e) MLCT^{R}

	C1	C2	C3	C4	C5	⋯
C1	2403	7.8	20.6	7.2	5.7	⋯
C2	33.7	376	4.3	8.2	7.5	⋯
C3	27.8	3.2	385	1.7	1.7	⋯
C4	15.3	10	1.9	352	5.5	⋯
C5	16	9.3	1.4	8.8	239	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋱

(f) MLCT^{P}

	C1	C2	C3	C4	C5	⋯
C1	2403	75.6	63	40.7	43.8	⋯
C2	6.4	376	6.7	17.4	20.4	⋯
C3	16.5	5.4	385	2.9	3	⋯
C4	3.3	11.1	1.9	352	20.9	⋯
C5	3.6	12.4	1.5	13.4	239	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋱

(g) TCM^{pred}

	C1	C2	C3	C4	C5	⋯
C1	2296.2	61.7	56.3	33.9	35.3	⋯
C2	14.4	358.8	8.3	18.3	21.6	⋯
C3	20	5.8	363.2	4.1	4.2	⋯
C4	6.8	11.2	2	339.5	16.6	⋯
C5	6.7	14.9	2	14.9	230.8	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋱

other columns. In other words, we likely observe $\|\hat{y}\|_1$ small when C1 is predicted and $\|\hat{y}\|_1$ large when it is not. This suggests that the model has understood the exclusivity of C1: for an annotator, a tweet is either non-sexist (i.e., C1) or sexist (i.e., one or more classes different from C1). When a tweet is classified as non-sexist, only C1 is predicted, leading to $\|\hat{y}\|_1 \approx 6$ (corresponding to an agreement of 6 annotators). In the opposite case, multiple non-exclusive classes are possible, potentially leading to $\|\hat{y}\|_1 \geq 6$.

The model tends to overpredict when not predicting C1. The C1 column in matrix (c) is roughly equal to that in (b), but the other columns are generally at least twice as large. This suggests that when C1 is predicted, $\|y\|_1 \approx \|\hat{y}\|_1$ occurs; otherwise, $\|y\|_1 \leq \|\hat{y}\|_1$ does.

B. NORMALIZED MATRICES

This subsection analyses the normalized matrices from (d) to (i). Since their results align with previous findings, we will focus on the new insights. We begin by discussing the row-normalized matrices (d), (f), and (h), followed by an analysis of the column-normalized matrices (e), (g), and (i).

Matrices (d), (f), and (h) reveal that underestimations in favor of C2 and C3 for C4 and C5 are at least as prominent as they are for C1. More precisely, C4C2 and C5C2 are greater than C1C2 in (d) and (f) and roughly equal in (h). The conclusions are similar if we consider C3 instead of C2.

Additionally, we observe that underestimations of C6 in favor of C5 are the most significant. Specifically, among all underestimates of C6, the predictions of C5 represent the larger portion compared to other classes (C6C5 is greater than CiC5 for $i \neq 5, 6$). The same applies if C5 is switched with C6.

Matrices (e), (g), and (i) do not reveal much new information, except that the overestimations of C2 originating from C1 are comparable to the ones of C5. Specifically, among all overestimations of C2 or C5, the C1 label accounts for a larger share than other classes (C1C2 and C1C5 are greater than C1Cj for $j \neq 1, 2, 5$).

C. CONCLUSION

The non-normalized matrices and the test-training ranking criterion were enough to capture the main behavioral patterns.

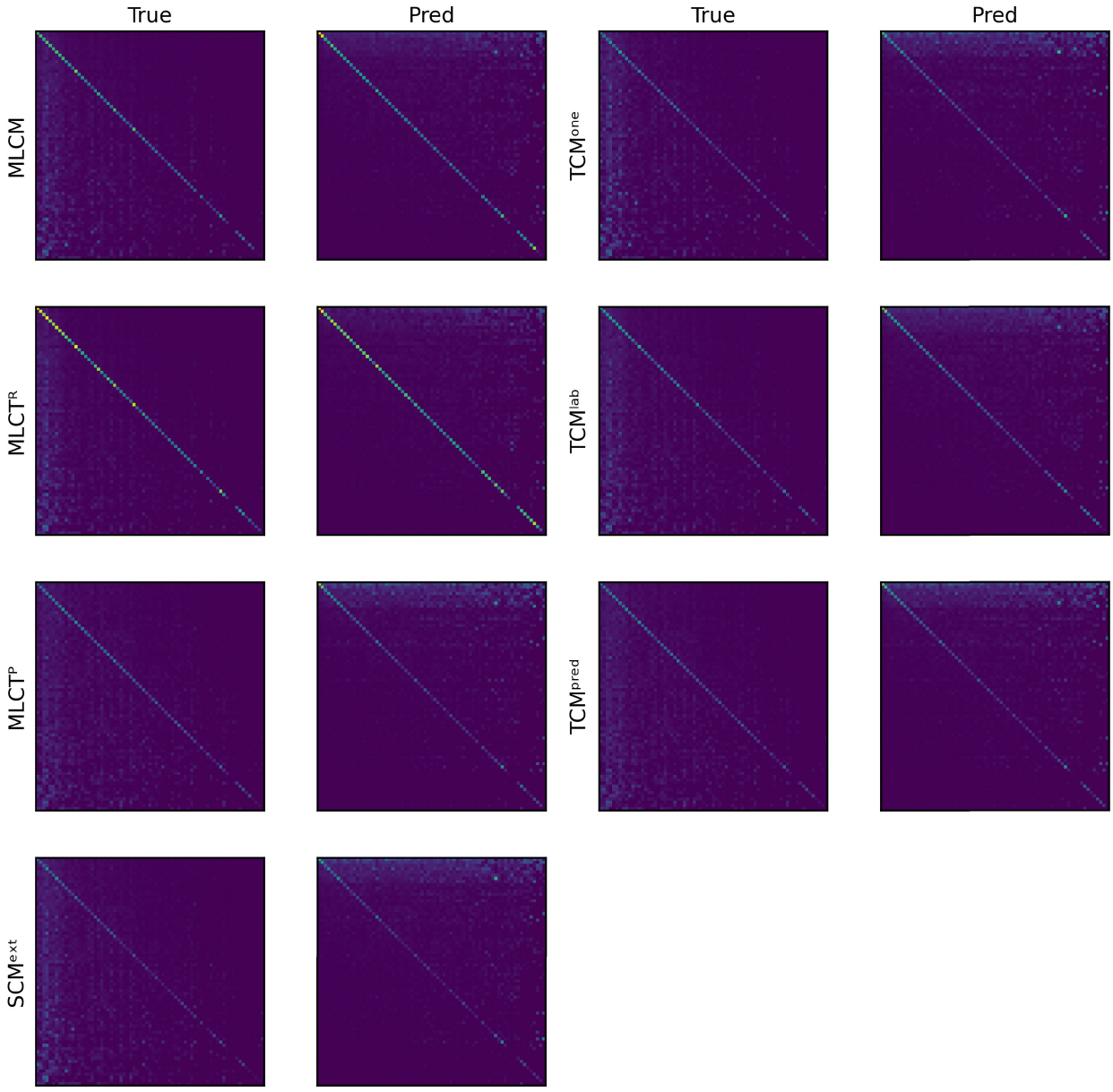


FIGURE 4. Heat maps of confusion matrices resulting from the MPST experiment with DeBERTa. Each matrix is plotted in two versions: row-normalized and column-normalized. The color scale is shared across all matrices.

The model has trouble with class C5, which is often over and underestimated in favor of C1, C2, and C3. However, in terms of proportions, the normalized matrices show that confusions involving C2 and C3 are at least as significant as those involving C1. The same is true for C4. This suggests that the errors between C4 or C5 and C2 or C3 are not just due to data distribution.

The normalized matrices also show that C6 is particularly underestimated in favor of C5, and vice versa, pointing to errors caused by factors beyond data distribution.

Finally, the different weightings reveal two key insights: first, the model recognizes the exclusivity of class C1, and second, when it does not predict C1, it tends to predict too many other classes.

X. DISCUSSION

This section addresses some limitations of TCM, specifically evaluating the advantages and drawbacks of normalization and the use of the discrete metric as the cost function.

A. NORMALIZATION

By forcing the labels and predictions to share the same norm, normalization ensures that any model error reflects confusion between two classes. Specifically, an overestimation of class j must correspond to an underestimation of class i, meaning errors in class j can not be viewed in isolation from class i. We expect the model's behavior to reveal actual class confusions through redundancy while other errors are spread

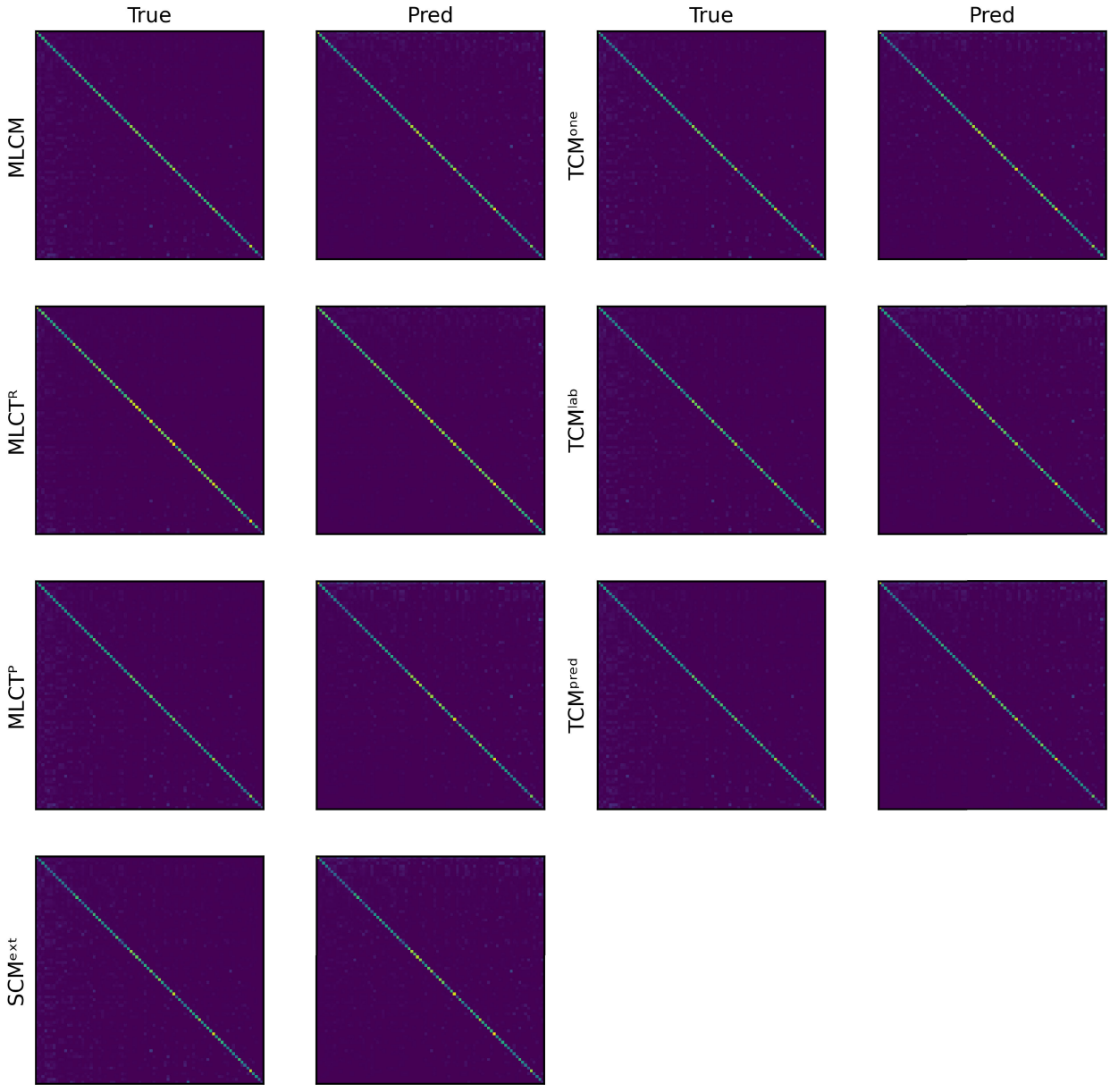


FIGURE 5. Heat maps of confusion matrices resulting from the MS-COCO experiment with SqueezeNet. Each matrix is plotted in two versions: row-normalized and column-normalized. The color scale is shared across all matrices.

evenly across affected classes. This issue also appears in the single-label confusion matrix.

Normalization can also lead to a loss of information. For example, consider two distinct instances, $(y, \hat{y})$ and $(y', \hat{y}')$, where $y' = \alpha y$ and $\hat{y}' = \beta\hat{y}$, with α and β as positive constants. In this case, the normalized contributions of $(y, \hat{y})$ and $(y', \hat{y}')$ would be identical despite representing different model behaviors. However, by applying appropriate weighting, we can recover some norm information, especially when using multiple matrices with different weightings.

The key advantage of normalization is that it enables the use of optimal transport theory, which supports all results. It also enhances the interpretability of transport-based matrices, for example, through properties like the marginal sum, leading to consistent metrics. Finally, as shown in this paper, normalization allows us to draw significant conclusions about the model's behavior.

B. DISCRETE METRIC

Better cost functions could be found. The set of labels and predictions provides insight into class dissimilarity within the model. By considering all instances, we can deduce more representative cost functions. However, assessing class proximity is exactly what a confusion matrix does, so if this information is already available, there is no need to go further. Moreover, the discrete metric offers simplicity, with explicit

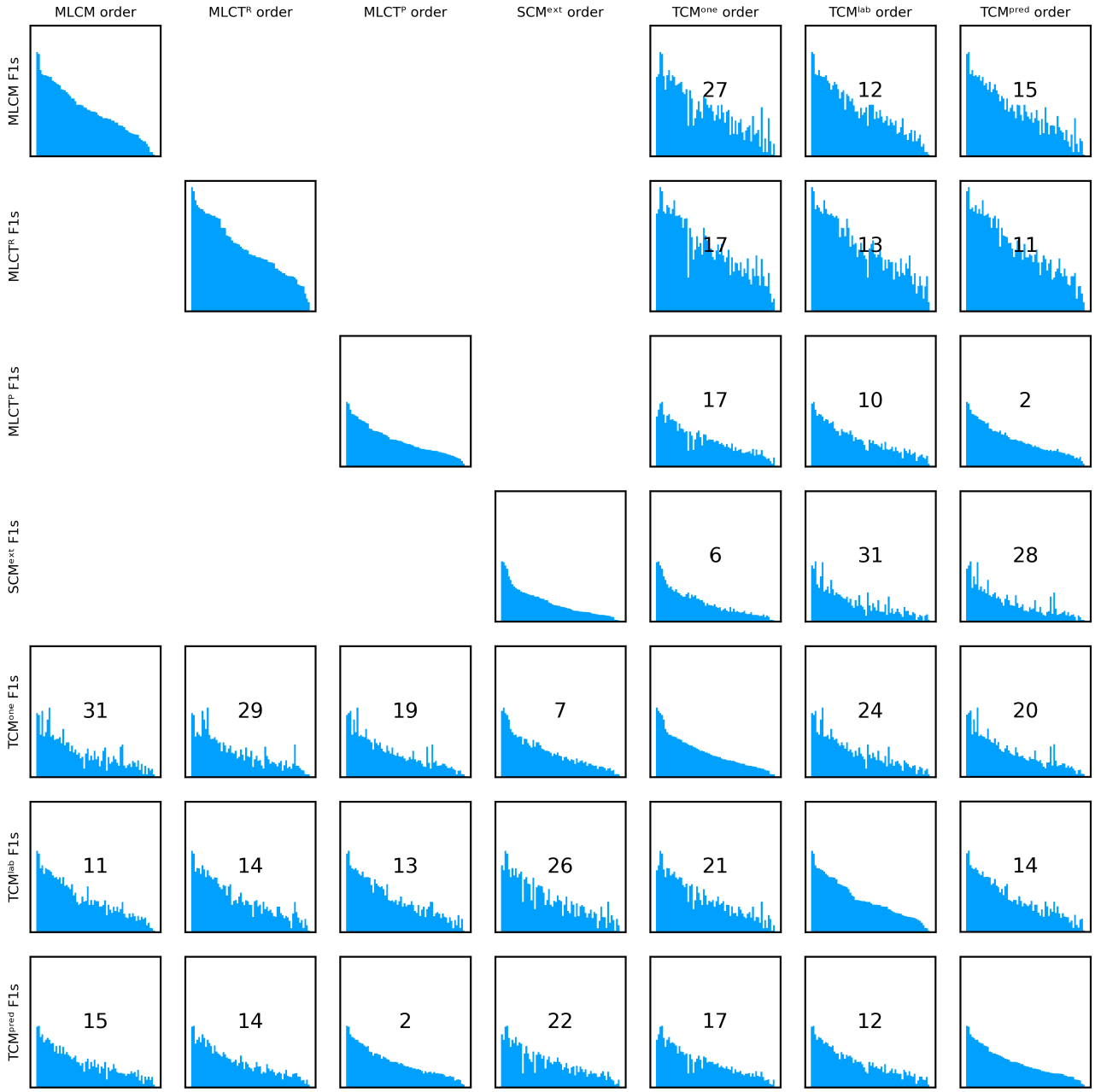


FIGURE 6. Bar plots display F1-scores derived from confusion matrices with different organizations on MPST with DeBERTa experiment. The bar plot *ij* represents F1-scores of matrix *i* organized by ranking derived from matrix *j*. Off-diagonal values represent the function Δ applied to matrices *i* and *j* (see Section VII-C).

solutions to the Kantorovich problem. It is cautious due to its minimal assumptions and highly interpretable, as the results are straightforward to understand.

XI. CONCLUSION

This paper addresses key limitations of existing confusion matrices that extend confusion matrices to broader frameworks, such as lack of justification, difficult interpretability, and limited frameworks.

To resolve these issues, we propose a novel approach using optimal transport theory combined with the principle of maximum entropy. The Trasport-based Confusion Matrix extends the traditional confusion matrix to multi and soft-label scenarios while being identical in single-label cases. We offer a very general definition based on minimal assumptions, which are intuitive and meet desirable properties not satisfied by other matrices. In addition, we provide an analytical solution that ensures scalability. While existing methods appear incompatible, our work demonstrates that a shared understanding of agreement and error exists, aligning different approaches under a unified theory. We also propose an extension of the Recall, Precision, F1-score, and Accuracy that can be computed directly from TCM.

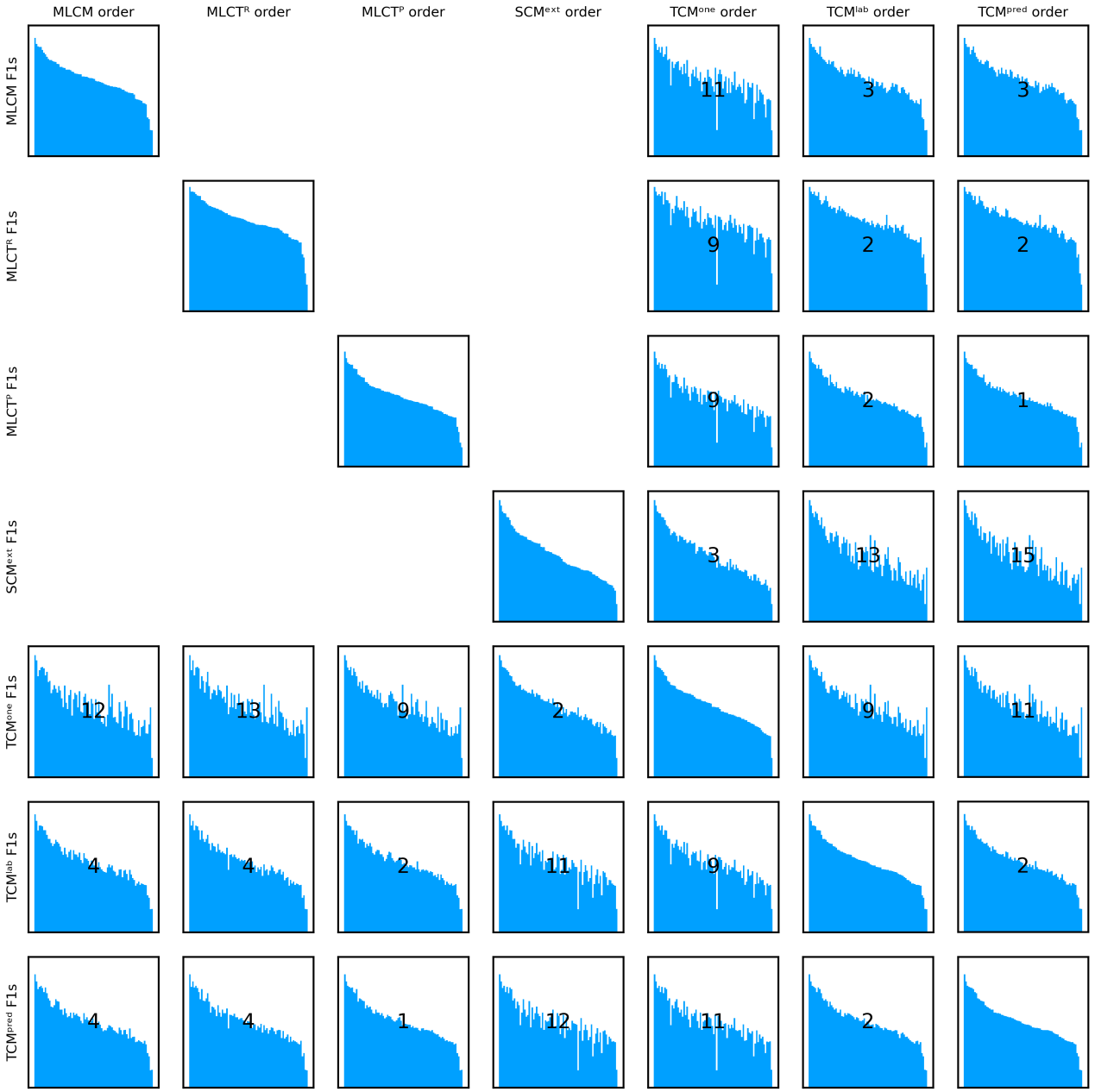


FIGURE 7. Bar plots display F1-scores derived from confusion matrices with different organizations on SqueezeNet with MS-COCO experiment. The bar plot *ij* represents F1-scores of matrix *i* organized by ranking derived from matrix *j*. Off-diagonal values represent the function Δ applied to matrices *i* and *j* (see Section VII-C).

Furthermore, we introduce the test-training ranking, a technique for plotting confusion matrices to distinguish between errors stemming from class distribution and those from other factors. We also show how normalization can help in this task.

Looking forward, future research could explore the direct use of raw model outputs without binarization in single or multi-label contexts for model analysis, employ TCM as a cost function or regularizer for model optimization, and develop active learning strategies to enhance model performance by selecting unlabeled examples based on TCM insights.

APPENDIX A
USEFUL LEMMAS FOR DEMONSTRATIONS

In this section, we provide some additional lemmas used in the demonstrations. Their proofs are given in the following appendix sections, except for Lemma 5 and Lemma 6, which are established analysis mathematical results. The reader is advised to pay attention to the lemmas when they are used in demonstrations, as their properties will become clearer in context.

Lemma 1 is a simple algebraic property.

TABLE 8. Transport-based matrices on EXIST with BERTweet. Classes are sorted in descending order of prevalence, C1 corresponds to Non-Sexist, C2 to Ideological-Inequality, C3 to Stereotyping-Dominance, C4 to Objectification, C5 to Sexual-Violence, and C6 to Misogyny-Non-Sexual-Violence. Color scales are specific to each matrix. For column-based normalization, colors compare entries within the same row. For row-based normalization, colors compare entries within the same column.

(a) Normalize: raw, Weighting: $\lambda(y, \hat{y}) = 1$

	C1	C2	C3	C4	C5	C6
C1	198	17	14	9	**10**	6
C2	15	**32**	3	3	3	2
C3	13	2	36	2	**3**	2
C4	13	**5**	**4**	25	2	2
C5	12	**3**	**4**	2	17	2
C6	5	2	2	1	**2**	16

(b) Normalize: raw, Weighting: $\lambda(y, \hat{y}) = \|y\|_1$

	C1	C2	C3	C4	C5	C6
C1	1238	115	98	62	**69**	44
C2	109	**276**	25	23	**24**	18
C3	94	20	289	16	**24**	**19**
C4	92	**39**	**36**	227	20	14
C5	85	**30**	**36**	16	158	**16**
C6	40	14	15	10	**18**	143

(c) Normalize: raw, Weighting: $\lambda(y, \hat{y}) = \|\hat{y}\|_1$

	C1	C2	C3	C4	C5	C6
C1	1231	243	208	152	**183**	105
C2	93	613	54	**58**	**61**	36
C3	82	40	597	41	**57**	**43**
C4	77	**81**	**71**	518	48	37
C5	77	**56**	54	35	369	**38**
C6	32	29	21	22	**43**	304

(d) Normalize: row, Weighting: $\lambda(y, \hat{y}) = 1$

	C1	C2	C3	C4	C5	C6
C1	78.3	6.6	5.5	3.5	3.8	2.4
C2	26	55.3	5.2	4.9	5.1	3.5
C3	22.5	4	61.5	3.5	4.7	3.7
C4	24.7	9.3	8.3	49.8	4.6	3.3
C5	29.1	8.5	10.3	4.8	42.4	5
C6	18.9	5.6	5.7	4.5	7.5	57.8

(e) Normalize: column, Weighting: $\lambda(y, \hat{y}) = 1$

	C1	C2	C3	C4	C5	C6
C1	77.4	27.3	22.3	20.7	26	20.1
C2	5.9	52.8	4.9	6.8	7.9	6.8
C3	5.1	3.8	56.9	4.8	7.4	7.3
C4	4.9	7.8	6.7	60.2	6.3	5.6
C5	4.6	5.7	6.7	4.6	46.8	6.8
C6	2	2.5	2.5	2.9	5.6	53.3

(f) Normalize: row, Weighting: $\lambda(y, \hat{y}) = \|y\|_1$

	C1	C2	C3	C4	C5	C6
C1	76.2	7.1	6	3.8	4.3	2.7
C2	23	58.3	5.2	4.8	5	3.7
C3	20.3	4.4	62.5	3.4	5.3	4.1
C4	21.4	9.2	8.5	52.9	4.7	3.3
C5	24.9	8.7	10.5	4.6	46.5	4.8
C6	16.6	6	6.1	4.1	7.7	59.5

(g) Normalize: column, Weighting: $\lambda(y, \hat{y}) = \|y\|_1$

	C1	C2	C3	C4	C5	C6
C1	74.7	23.3	19.6	17.5	22.1	17.1
C2	6.6	55.8	4.9	6.4	7.5	7
C3	5.7	4.1	58	4.5	7.8	7.5
C4	5.6	7.9	7.3	64.3	6.4	5.6
C5	5.1	6	7.2	4.4	50.4	6.4
C6	2.4	2.9	2.9	2.8	5.9	56.3

(h) Normalize: row, Weighting: $\lambda(y, \hat{y}) = \|\hat{y}\|_1$

	C1	C2	C3	C4	C5	C6
C1	58	11.4	9.8	7.1	8.6	5
C2	10.2	67	5.9	6.4	6.6	4
C3	9.5	4.7	69.4	4.8	6.6	5
C4	9.3	9.7	8.6	62.2	5.8	4.5
C5	12.2	8.9	8.6	5.5	58.7	6.1
C6	7.1	6.5	4.7	4.8	9.5	67.3

(i) Normalize: column, Weighting: $\lambda(y, \hat{y}) = \|\hat{y}\|_1$

	C1	C2	C3	C4	C5	C6
C1	77.3	22.9	20.7	18.4	24	18.7
C2	5.8	57.7	5.4	7	8	6.4
C3	5.1	3.8	59.3	5	7.5	7.7
C4	4.9	7.6	7.1	62.8	6.3	6.6
C5	4.8	5.3	5.4	4.2	48.5	6.8
C6	2	2.8	2.1	2.6	5.7	53.8

Lemma 1. *Let u and v be two vectors in $\mathbb{R}^C_{\geq 0}$ with the same norm, $\|u\|_1 = \|v\|_1$. Then,*

$$\|u - min(u, v)\|_1 = \|v - min(u, v)\|_1 \tag{54}$$

Given a matrix belonging to the set of optimal transference plans with classes sorted in a particular order, let's consider rearranging the order of the classes. Lemma 2 explains how to derive the matrix equivalent to the original one after the rearrangement.

Lemma 2. *Let σ be a permutation of the integer interval from 1 to C, and $(y, \hat{y})$ be an instance. Let $(s, \hat{s})$ be two vectors defined as*:

$$s_{\sigma(k)} = y_k, \quad \text{and} \quad \hat{s}_{\sigma(k)} = \hat{y}_k, \quad \text{for} \quad k = 1 \dots C. \tag{55}$$

If $\widetilde{M}$ is a matrix in $T^{opt}(s, \hat{s})$, then the matrix M defined as,

$$M_{ij} = \widetilde{M}_{\sigma(i)\sigma(j)}, \quad \text{for} \quad i, j = 1 \dots C, \tag{56}$$

is in $T^{opt}(y, \hat{y})$.

Lemma 3 is a property of the function f given in Definition 1.

Lemma 3. *Let u and v be two vectors in $\mathbb{R}^C_{\geq 0}$ with the same norm, and i and j two integers between 1 and C. The following equalities hold*:

$$\sum_{k=1}^{C} f(u, v)_{kj} = v_j \quad \text{and} \quad \sum_{k=1}^{C} f(u, v)_{ik} = u_i. \tag{57}$$

The following lemma exhibits an element of $T^{\text{opt}}(y, \hat{y})$ where the error is maximal in column C.

Lemma 4. *The following matrix, denoted as M, is in $T^{opt}(y, \hat{y})$,*

$$M := f\Big(\frac{y}{\|y\|_1} - m, \frac{\hat{y}}{\|\hat{y}\|_1} - E^C\big(\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})\big)\Big) + m \otimes E^C, \tag{58}$$

where E^C is a zero vector of size C such as its entry C equals 1, while others are zero. The vector m is in $\mathbb{R}^C_{\geq 0}$ and it is defined as follows: it exists an integer p such as

$$\sum_{k=1}^{p-1} \frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}) \leq \frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) \leq \sum_{k=1}^{p} \frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}). \tag{59}$$

If $p = 1$, then $p - 1 = 0$, and the first term is an empty sum equal to 0. Then, $m_k = \frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1})$ when $k < p$, $m_p = \frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) - \sum_{k=1}^{p-1} \frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1})$, and $m_k = 0$ if $k > p$.

Moreover, it holds

$$\begin{aligned} M_{1C} &= \min(\frac{y_i}{\|y\|_1} - \min(\frac{y_i}{\|y\|_1}, \frac{\hat{y}_i}{\|\hat{y}\|_1}), \frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})), \quad \text{and} \\ M_{C-1\,C} &= \max\Big(0, \frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) - \sum_{k=1:k\neq C-1}^{C} \frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1})\Big). \end{aligned} \tag{60}$$

Lemma 5 and Lemma 6 are established mathematical results from analysis; for more details, refer to [47].

Lemma 5. *Let f be differentiable function on $D \subset \mathbb{R}^n$. Then, f is concave on D if, and only if*:

$$\forall (x_1, x_2) \in D^2, \quad f(x_2) \leq f(x_1) + \langle \nabla f(x_1), x_2 - x_1 \rangle \tag{61}$$

where ∇ and $\langle \cdot, \cdot \rangle$ denote the gradient of a function and the usual scalar product, respectively.

Lemma 6. *Let f an affine function on $D \subset \mathbb{R}^n$. Then, it holds*:

$$\forall (x_1, x_2) \in D^2, \quad f(x_2) = f(x_1) + \langle \nabla f(x_1), x_2 - x_1 \rangle \tag{62}$$

where ∇ denotes the gradient of a function.

APPENDIX B
PROOF OF LEMMA 1

By assumption, $\|u\|_1 = \|v\|_1$, this leads to

$$\begin{aligned} \|u - min(u, v)\|_1 &= \sum_{k=1}^{C} |\underbrace{u_k - min(u_k, v_k)}_{\geq 0}| \\ &= \sum_{k=1}^{C} u_k - min(u_k, v_k) \\ &= \|u\|_1 - \sum_{k=1}^{C} min(u_k, v_k) \\ &= \|v\|_1 - \sum_{k=1}^{C} min(u_k, v_k) \\ &= \sum_{k=1}^{C} |\underbrace{v_k}_{\geq 0}| - \sum_{k=1}^{C} min(u_k, v_k) \\ &= \sum_{k=1}^{C} \underbrace{v_k - min(u_k, v_k)}_{\geq 0} \\ &= \sum_{k=1}^{C} |v_k - \sum_{k=1}^{C} min(u_k, v_k)| \\ &= \|v - min(u, v)\|_1, \end{aligned} \tag{63}$$

ending the proof.

APPENDIX C
PROOF OF LEMMA 2

Since the matrix $\widetilde{M}$ is in $T^{\text{opt}}(s, \hat{s})$, the following equalities hold:

$$\widetilde{M} \in \mathbb{R}^{C\times C}_{\geq 0}, \sum_{j=1}^{C} \widetilde{M}_{ij} = \frac{s_i}{\|s\|_1},$$

$$\sum_{i=1}^{C} \widetilde{M}_{ij} = \frac{\hat{s}_j}{\|\hat{s}\|_1}, \quad \text{and}$$

$$\widetilde{M}_{ii} = \min(\frac{s_i}{\|s\|_1}, \frac{\hat{s}_i}{\|\hat{s}\|_1}), \quad \text{for} \quad i = 1 \dots C. \tag{64}$$

Let M be the matrix defined as:

$$M_{ij} = \widetilde{M}_{\sigma(i)\sigma(j)}, \quad \text{for} \quad i,j = 1 \dots C. \tag{65}$$

Since the entries of M are a rearrangement the entries of $\widetilde{M} \in \mathbb{R}^{C\times C}_{\geq 0}$, it follows that $M \in \mathbb{R}^{C\times C}_{\geq 0}$. By design, and because $\|y\|_1 = \|s\|_1$ and $\|\hat{y}\|_1 = \|\hat{s}\|_1$, we have:

$$\begin{aligned}
\sum_{j=1}^{C} M_{ij} &= \sum_{j=1}^{C} \widetilde{M}_{\sigma(i)\sigma(j)} = \frac{s_{\sigma(i)}}{\|s\|_1} = \frac{y_i}{\|y\|_1}, \\
\sum_{i=1}^{C} M_{ij} &= \sum_{i=1}^{C} \widetilde{M}_{\sigma(i)\sigma(j)} = \frac{\hat{s}_{\sigma(j)}}{\|s\|_1} = \frac{\hat{y}_j}{\|\hat{y}\|_1}, \quad \text{and} \\
M_{ii} &= \widetilde{M}_{\sigma(i)\sigma(i)} = \min\left(\frac{s_{\sigma(i)}}{\|s\|_1}, \frac{\hat{s}_{\sigma(i)}}{\|\hat{s}\|_1}\right) \\
&= \min\left(\frac{y_i}{\|y\|_1}, \frac{\hat{y}_i}{\|\hat{y}\|_1}\right), \quad \text{for} \quad i = 1 \dots C,
\end{aligned} \tag{66}$$

achieving the proof.

APPENDIX D
PROOF OF LEMMA 3

Let u and v be two vectors in $\mathbb{R}^{C}_{\geq 0}$ with same norm, and i and j be two integers between 1 and C. It follows that

$$\begin{aligned}
&\sum_{k=1}^{C} f(u,v)_{kj} \\
&= f(u,v)_{jj} + \sum_{k=1:k\neq j}^{C} f(u,v)_{kj} \\
&= \min(u_j, v_j) \\
&\quad + \sum_{k=1:k\neq j}^{C} \frac{\Big(u_k - \min(u_k,v_k)\Big)\Big(v_j - \min(u_j,v_j)\Big)}{\|u - \min(u,v)\|_1}.
\end{aligned} \tag{67}$$

Since $\min(u_j, v_j)$ equals u_j or v_j, it follows that $\Big(u_j - \min(u_j,v_j)\Big)\Big(v_j - \min(u_j,v_j)\Big) = 0$, leading to

$$\begin{aligned}
&\sum_{k=1}^{C} f(u,v)_{kj} \\
&= \min(u_j,v_j) + \sum_{k=1}^{C} \frac{\Big(u_k - \min(u_k,v_k)\Big)\Big(v_j - \min(u_j,v_j)\Big)}{\|u - \min(u,v)\|_1} \\
&= \min(u_j,v_j) + \Big(v_j - \min(u_j,v_j)\Big) \underbrace{\sum_{k=1}^{C} \frac{u_k - \min(u_k,v_k)}{\|u - \min(u,v)\|_1}}_{=1} \\
&= \min(u_j,v_j) + v_j - \min(u_j,v_j) \\
&= v_j
\end{aligned} \tag{68}$$

In the same way,

$$\begin{aligned}
&\sum_{k=1}^{C} f(u,v)_{ik} \\
&= f(u,v)_{ii} + \sum_{k=1:k\neq i}^{C} f(u,v)_{ik} \\
&= \min(u_i,v_i) + \sum_{k=1}^{C} \frac{\Big(u_i - \min(u_i,v_i)\Big)\Big(v_k - \min(u_k,v_k)\Big)}{\underbrace{\|u - \min(u,v)\|_1}_{=\|v-\min(u,v)\|_1, \text{ according to Lemma 1}}} \\
&= \min(u_i,v_i) + \Big(u_i - \min(u_i,v_i)\Big) \underbrace{\sum_{k=1}^{C} \frac{v_k - \min(u_k,v_k)}{\|v - \min(u,v)\|_1}}_{=1} \\
&= u_i,
\end{aligned} \tag{69}$$

which completes the proof.

APPENDIX E
PROOF OF LEMMA 4

Considering

$$\begin{aligned}
M = f\Big(&\frac{y}{\|y\|_1} - m, \frac{\hat{y}}{\|\hat{y}\|_1} - E^C\Big(\frac{\hat{y}_C}{\|\hat{y}\|_1} \\
&- \min\Big(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}\Big)\Big)\Big) + m \otimes E^C,
\end{aligned} \tag{70}$$

to apply f, we must demonstrate that the variable vectors have non-negative entries and the same norm. Next, we must verify that M is in $T^{\text{opt}}(y,\hat{y})$, particularly by verifying the diagonal values and the sum constraints. Finally, we will compute the value of entry $M_{1\,C}$ and $M_{C-1\,C}$.

A. VECTORS WITH NON-NEGATIVE ENTRIES

Let's show that values in $\frac{y}{\|y\|_1} - m$ and $\frac{\hat{y}}{\|\hat{y}\|_1} - E^C\Big(\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min\Big(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}\Big)\Big)$ are non-negative. Let k be an integer between 1 and C.

Considering the first vector, $\frac{y}{\|y\|_1} - m$, if $k < p$, then, by definition, $m_k = \frac{y_k}{\|y\|_1} - \min\Big(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\Big)$, leading to $\frac{y_k}{\|y\|_1} - m_k = \min\Big(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\Big) \geq 0$. Otherwise, if $k = p$, then

$$\begin{aligned}
&\frac{y_p}{\|y\|_1} - m_p \\
&= \frac{y_p}{\|y\|_1} - \frac{\hat{y}_C}{\|\hat{y}\|_1} \\
&+ \min\Big(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}\Big) + \sum_{k=1}^{p-1} \frac{y_k}{\|y\|_1} - \min\Big(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\Big) \\
&= \min\Big(\frac{y_p}{\|y\|_1}, \frac{\hat{y}_p}{\|\hat{y}\|_1}\Big) \\
&+ \underbrace{\sum_{k=1}^{p} \frac{y_k}{\|y\|_1} - \min\Big(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\Big) - \frac{\hat{y}_C}{\|\hat{y}\|_1} + \min\Big(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}\Big)}_{\geq 0, \text{ by assumption about } p} \\
&\geq \min\Big(\frac{y_p}{\|y\|_1}, \frac{\hat{y}_p}{\|\hat{y}\|_1}\Big) \\
&\geq 0.
\end{aligned} \tag{71}$$

Finally if $k > p$, then, by definition, $m_k = 0$ which leads to $\frac{y_p}{\|y\|_1} - m_p = \frac{y_p}{\|y\|_1} \geq 0$. In conclusion, $\frac{y}{\|y\|_1} - m$ has non-negative entries.

Considering the second vector, $\frac{\hat{y}}{\|\hat{y}\|_1} - E^C\big(\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})\big)$, if $k = C$, then

$$\begin{aligned}\Big(\frac{\hat{y}}{\|\hat{y}\|_1} - E^C\big(\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})\big)\Big)_C \\ = \frac{\hat{y}_C}{\|\hat{y}\|_1} - \frac{\hat{y}_C}{\|\hat{y}\|_1} + \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) \\ = \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) \geq 0. \end{aligned} \tag{72}$$

Otherwise,

$$\Big(\frac{\hat{y}}{\|\hat{y}\|_1} - E^C\big(\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})\big)\Big)_k = \frac{\hat{y}_k}{\|\hat{y}\|_1} \geq 0, \tag{73}$$

this conclude to the non-negativity of $\frac{\hat{y}}{\|\hat{y}\|_1} - E^C\big(\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})\big)$.

B. SAME NORM

Let's check that vectors $\frac{y}{\|y\|_1} - m$ and $\frac{\hat{y}}{\|\hat{y}\|_1} - E^C\big(\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})\big)$ have the same norm.

Firstly, we observe that (74), as shown at the bottom of the next page.

Secondly, by adding $\|\min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})\|_1$ to both sides of the equality

$$\begin{aligned}\|\frac{y}{\|y\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}) - m\|_1 \\ = \|\frac{\hat{y}}{\|\hat{y}\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}) \\ - E^C\big(\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})\big)\|_1, \end{aligned} \tag{75}$$

it follows that

$$\begin{aligned}\|\frac{y}{\|y\|_1}\min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}) - m\|_1 \\ - + \|\min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})\|_1 \\ = \|\frac{\hat{y}}{\|\hat{y}\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}) \\ - E^C\big(\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})\big)\|_1 \\ + \|\min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})\|_1 \\ \Rightarrow \|\frac{y}{\|y\|_1} - m\|_1 = \|\frac{\hat{y}}{\|\hat{y}\|_1} \\ - E^C\big(\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})\big)\|_1 \end{aligned} \tag{76}$$

This holds because all the entries of $\frac{y}{\|y\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}) - m$ and $\frac{\hat{y}}{\|\hat{y}\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}) - E^C\big(\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})\big)$ are positive. In conclusion, $\frac{y}{\|y\|_1} - m$ and $\frac{\hat{y}}{\|\hat{y}\|_1} - E^C\big(\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})\big)$ have the same norm.

C. DIAGONAL VALUES

To show the property of diagonal values, we need to show two prerequisites: $m_C = 0$ and

$$\begin{aligned}\min\Big(\frac{y}{\|y\|_1} - m, \frac{\hat{y}}{\|\hat{y}\|_1} - E^C\big(\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})\big)\Big) \\ = \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}). \end{aligned} \tag{77}$$

Firstly, let's show that $m_C = 0$. By design, if $p \leq C-1$ then $m_C = 0$. Otherwise, if $p = C$, then (78), as shown at the bottom of the next page.

Since $\min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})$ equals to $\frac{y_C}{\|y\|_1}$ or $\frac{\hat{y}_C}{\|\hat{y}\|_1}$, either the first term $|\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})|$, or the second one $|\frac{y_C}{\|y\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})|$ is zero. We observe that

$$\begin{aligned}|\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})| \leq \|\frac{\hat{y}}{\|\hat{y}\|_1} \\ - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})\|_1, \text{ and } |\frac{y_C}{\|y\|_1} \\ - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})| \leq \|\frac{y}{\|y\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})\|_1, \end{aligned} \tag{79}$$

which leads to $m_C \leq 0$. By design, all entries of m are non-negative, particularly $0 \leq m_C$. The inequalities $0 \leq m_C \leq 0$ imply that $m_C = 0$.

Secondly, let's compute the vector $\min\Big(\frac{y}{\|y\|_1} - m, \frac{\hat{y}}{\|\hat{y}\|_1} - E^C\big(\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})\big)\Big)$ entry per entry. Let k be an integer between 1 and C.

If $k = C$, then $m_C = 0$ and $E^C_C = 1$, leading to (80), as shown at the bottom of page 28.

Otherwise, if $p < k < C$, then $m_k = 0$ and $E^C_k = 0$, leading to (81), as shown at the bottom of page 28.

Otherwise, if $k < C$ and $k < p$, then $m_k = \frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1})$ and $E^C_k = 0$, leading to (82), as shown at the bottom of page 28.

Finally, if $k < C$ and $k = p$, then $m_k = \frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) - \sum_{l=1}^{p-1} \frac{y_l}{\|y\|_1} - \min(\frac{y_l}{\|y\|_1}, \frac{\hat{y}_l}{\|\hat{y}\|_1})$ and $E^C_k = 0$, leading to (83), as shown at the bottom of page 28.

We will bound A, and this will allow us to conclude. By assumption about p, (84), as shown at the bottom of page 29. which leads to (85), as shown at the bottom of page 29.

In conclusion, we prove that

$$\begin{aligned}\min\Big(\frac{y}{\|y\|_1} - m, \frac{\hat{y}}{\|\hat{y}\|_1} - E^C\big(\frac{\hat{y}_C}{\|\hat{y}\|_1} \\ - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})\big)\Big) = \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}) \end{aligned} \tag{86}$$

We can now conclude that the definition of E^C implies that the non-zero entries in $m \otimes E^C$ are exclusively in column C.

Furthermore, since $m_C = 0$, the diagonal entry C of $m \otimes E^C$ is also zero. Consequently, the entire diagonal of $m \otimes E^C$ is zero. Moreover, based on the definition of f and the fact that $\min\left(\frac{y}{\|y\|_1} - m, \frac{\hat{y}}{\|\hat{y}\|_1} - E^C\left(\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})\right)\right) = \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})$, the following equality holds:

$$\left[f\left(\frac{y}{\|y\|_1} - m, \frac{\hat{y}}{\|\hat{y}\|_1} - E^C\left(\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})\right)\right)\right.$$

$$
\begin{aligned}
&\|\frac{y}{\|y\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}) - m\|_1 \\
&= \sum_{k=1}^{C} |\frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}) - m_k| \\
&= |\frac{y_p}{\|y\|_1} - \min(\frac{y_p}{\|y\|_1}, \frac{\hat{y}_p}{\|\hat{y}\|_1}) - m_p| + \sum_{k=1}^{p-1} |\frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}) - m_k| + \sum_{k=p+1}^{C} |\frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}) - m_k| \\
&= |\frac{y_p}{\|y\|_1} - \min(\frac{y_p}{\|y\|_1}, \frac{\hat{y}_p}{\|\hat{y}\|_1}) - \frac{\hat{y}_C}{\|\hat{y}\|_1} + \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) + \sum_{k=1}^{p-1} \frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1})| + \\
&\qquad \underbrace{\sum_{k=1}^{p-1} |\frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}) - \frac{y_k}{\|y\|_1} + \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1})|}_{=0} + \sum_{k=p+1}^{C} |\frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1})| \\
&= \underbrace{|-\frac{\hat{y}_C}{\|\hat{y}\|_1} + \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) + \sum_{k=1}^{p} \frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1})|}_{\geq 0, \text{ by assumption about } p} + \sum_{k=p+1}^{C} \underbrace{|\frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1})|}_{\geq 0} \\
&= -\frac{\hat{y}_C}{\|\hat{y}\|_1} + \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) + \underbrace{\sum_{k=1}^{p} \frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}) + \sum_{k=p+1}^{C} \frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1})}_{=\|\frac{y}{\|y\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})\|_1} \\
&= -\frac{\hat{y}_C}{\|\hat{y}\|_1} + \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) + \underbrace{\|\frac{y}{\|y\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})\|_1}_{=\|\frac{\hat{y}}{\|\hat{y}\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})\|_1, \text{ according to Lemma 1}} \\
&= -\frac{\hat{y}_C}{\|\hat{y}\|_1} + \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) + \|\frac{\hat{y}}{\|\hat{y}\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})\|_1 \\
&= \|\frac{\hat{y}}{\|\hat{y}\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}) - E^C\left(\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})\right)\|_1
\end{aligned}
\tag{74}
$$

$$
\begin{aligned}
m_C &= \frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) - \sum_{k=1}^{C-1} \frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}) \\
&= \underbrace{\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})}_{\geq 0} + \underbrace{\frac{y_C}{\|y\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})}_{\geq 0} - \|\frac{y}{\|y\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})\|_1 \\
&= |\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})| + |\frac{y_C}{\|y\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})| - \underbrace{\|\frac{y}{\|y\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})\|_1}_{=\|\frac{\hat{y}}{\|\hat{y}\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})\|_1, \text{ according to Lemma 1}} \\
&= |\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})| + |\frac{y_C}{\|y\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})| - \|\frac{\hat{y}}{\|\hat{y}\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})\|_1
\end{aligned}
\tag{78}
$$

$$+ m \otimes E^C \Big]_{kk}$$
$$= f\Big(\frac{y}{\|y\|_1} - m, \frac{\hat{y}}{\|\hat{y}\|_1} - E^C\big(\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})\big)\Big)_{kk}$$
$$= \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}) \tag{87}$$

We have proved that the diagonal values correspond to $T^{\text{opt}}(y, \hat{y})$.

$$\begin{aligned}
&= \min\Big(\frac{y}{\|y\|_1} - m, \frac{\hat{y}}{\|\hat{y}\|_1} - E^C\big(\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})\big)\Big)_C \\
&= \min\Big(\frac{y_C}{\|y\|_1} - m_C, \frac{\hat{y}_C}{\|\hat{y}\|_1} - \frac{\hat{y}_C}{\|\hat{y}\|_1} + \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})\Big) \\
&= \min\Big(\frac{y_C}{\|y\|_1}, \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})\Big) \\
&= \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) \\
&= \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1})
\end{aligned} \tag{80}$$

$$\begin{aligned}
&= \min\Big(\frac{y}{\|y\|_1} - m, \frac{\hat{y}}{\|\hat{y}\|_1} - E^C\big(\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})\big)\Big)_k \\
&= \min\Big(\frac{y_k}{\|y\|_1} - m_k, \frac{\hat{y}_k}{\|\hat{y}\|_1}\Big) \\
&= \min\Big(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\Big)
\end{aligned} \tag{81}$$

$$\begin{aligned}
&= \min\Big(\frac{y}{\|y\|_1} - m, \frac{\hat{y}}{\|\hat{y}\|_1} - E^C\big(\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})\big)\Big)_k \\
&= \min\Big(\frac{y_k}{\|y\|_1} - m_k, \frac{\hat{y}_k}{\|\hat{y}\|_1}\Big) \\
&= \min\Big(\frac{y_k}{\|y\|_1} - \frac{y_k}{\|y\|_1} + \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}), \frac{\hat{y}_k}{\|\hat{y}\|_1}\Big) \\
&= \min\Big(\min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}), \frac{\hat{y}_k}{\|\hat{y}\|_1}\Big) \\
&= \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1})
\end{aligned} \tag{82}$$

$$\begin{aligned}
&= \min\Big(\frac{y}{\|y\|_1} - m, \frac{\hat{y}}{\|\hat{y}\|_1} - E^C\big(\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})\big)\Big)_p \\
&= \min\Big(\frac{y_p}{\|y\|_1} - m_p, \frac{\hat{y}_p}{\|\hat{y}\|_1}\Big) \\
&= \min\Big(\frac{y_p}{\|y\|_1} - \frac{\hat{y}_C}{\|\hat{y}\|_1} + \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) + \sum_{l=1}^{p-1}\frac{y_l}{\|y\|_1} - \min(\frac{y_l}{\|y\|_1}, \frac{\hat{y}_l}{\|\hat{y}\|_1}), \frac{\hat{y}_p}{\|\hat{y}\|_1}\Big) \\
&= \min\Big(\underbrace{\min(\frac{y_p}{\|y\|_1}, \frac{\hat{y}_p}{\|\hat{y}\|_1}) - \frac{\hat{y}_C}{\|\hat{y}\|_1} + \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) + \sum_{l=1}^{p}\frac{y_l}{\|y\|_1} - \min(\frac{y_l}{\|y\|_1}, \frac{\hat{y}_l}{\|\hat{y}\|_1})}_{:=A}, \frac{\hat{y}_p}{\|\hat{y}\|_1}\Big) \\
&= \min\Big(A, \frac{\hat{y}_p}{\|\hat{y}\|_1}\Big).
\end{aligned} \tag{83}$$

D. MARGINAL SUM PROPERTY

Let's show marginal sums proprieties are also met.

Firstly, we show that $\|m\|_1 = \frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})$ (88), as shown at the bottom of the next page.

We begin with column sums. If $j < C$, then (89), as shown at the bottom of the next page. If $j = C$, then (90), as shown at the bottom of the next page.

We continue with the row sums, then (91), as shown at the bottom of page 31.

We show that the properties of marginal sums are met.

In conclusion, by design of f, all entries in

$$f\Big(\frac{y}{\|y\|_1} - m, \frac{\hat{y}}{\|\hat{y}\|_1} - E^C\big(\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})\big)\Big) \tag{92}$$

are non-negative, as are the entries in $m \otimes E^C$. Consequently, the M's entries are non-negative and M meets all constraints of $T^{\text{opt}}(y, \hat{y})$. Therefore, $M \in T^{\text{opt}}(y, \hat{y})$.

E. ENTRY 1C

Let's show that the value of M_{1C} is

$$\begin{aligned} M_{1C} = \min(&\frac{y_i}{\|y\|_1} - \min(\frac{y_i}{\|y\|_1}, \frac{\hat{y}_i}{\|\hat{y}\|_1}), \frac{\hat{y}_C}{\|\hat{y}\|_1} \\ &- \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})). \end{aligned} \tag{93}$$

We will use the fact that

$$\min\Big(\frac{y}{\|y\|_1} - m, \frac{\hat{y}}{\|\hat{y}\|_1} - E^C\big(\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})\big)\Big)$$

$$= \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}), \tag{94}$$

previously demonstrated in Subsection XI-C.

By definition, we have: (95), as shown at the bottom of page 31. The formula for m_1 depends on the value of p. Considering two cases, $p = 1$ and $p > 1$, we will use the following inequalities, which are satisfied by the assumptions on p: (96), as shown at the bottom of page 31.

If $1 = p$, then, $m_1 = \frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})$ by definition. Moreover, we can rewrite m_1 with a minimum: (97), as shown at the bottom of page 31.

If $1 < p$ then, $m_1 = \frac{y_1}{\|y\|_1} - \min(\frac{y_1}{\|y\|_1}, \frac{\hat{y}_1}{\|\hat{y}\|_1})$ by definition. Moreover, we can rewrite m_1 with a minimum: (98), as shown at the bottom of page 31.

In conclusion, we prove

$$\begin{aligned} &M_{1C} \\ &= \min\Big(\frac{y_1}{\|y\|_1} - \min(\frac{y_1}{\|y\|_1}, \frac{\hat{y}_1}{\|\hat{y}\|_1}), \frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})\Big). \end{aligned} \tag{99}$$

F. ENTRY C-1C

Let's show that the value of $M_{C-1\,C}$ is:

$$\begin{aligned} M_{C-1\,C} = \max\Big(0, &\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) \\ &- \sum_{k=1:k\neq C-1}^{C} \frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1})\Big). \end{aligned} \tag{100}$$

$$\sum_{l=1}^{p-1} \frac{y_l}{\|y\|_1} - \min(\frac{y_l}{\|y\|_1}, \frac{\hat{y}_l}{\|\hat{y}\|_1}) \le \frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) \le \underbrace{\sum_{l=1}^{p} \frac{y_l}{\|y\|_1} - \min(\frac{y_l}{\|y\|_1}, \frac{\hat{y}_l}{\|\hat{y}\|_1})}_{\text{subtracting this sum}}$$

$$\Rightarrow -\frac{y_p}{\|y\|_1} + \min(\frac{y_p}{\|y\|_1}, \frac{\hat{y}_p}{\|\hat{y}\|_1}) \le \frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) - \sum_{l=1}^{p} \frac{y_l}{\|y\|_1} - \min(\frac{y_l}{\|y\|_1}, \frac{\hat{y}_l}{\|\hat{y}\|_1}) \le 0$$

$$\underbrace{\Rightarrow}_{\text{multiplying by } -1} 0 \le -\frac{\hat{y}_C}{\|\hat{y}\|_1} + \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) + \sum_{l=1}^{p} \frac{y_l}{\|y\|_1} - \min(\frac{y_l}{\|y\|_1}, \frac{\hat{y}_l}{\|\hat{y}\|_1}) \le \frac{y_p}{\|y\|_1} - \underbrace{\min(\frac{y_p}{\|y\|_1}, \frac{\hat{y}_p}{\|\hat{y}\|_1})}_{\text{adding this quantity}}$$

$$\Rightarrow \min(\frac{y_p}{\|y\|_1}, \frac{\hat{y}_p}{\|\hat{y}\|_1}) \le \underbrace{\min(\frac{y_p}{\|y\|_1}, \frac{\hat{y}_p}{\|\hat{y}\|_1}) - \frac{\hat{y}_C}{\|\hat{y}\|_1} + \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) + \sum_{l=1}^{p} \frac{y_l}{\|y\|_1} - \min(\frac{y_l}{\|y\|_1}, \frac{\hat{y}_l}{\|\hat{y}\|_1})}_{=A} \le \frac{y_p}{\|y\|_1}, \tag{84}$$

$$\underbrace{\min\Big(\min(\frac{y_p}{\|y\|_1}, \frac{\hat{y}_p}{\|\hat{y}\|_1}), \frac{\hat{y}_p}{\|\hat{y}\|_1}\Big)}_{=\min(\frac{y_p}{\|y\|_1}, \frac{\hat{y}_p}{\|\hat{y}\|_1})} \le \min\Big(A, \frac{\hat{y}_p}{\|\hat{y}\|_1}\Big) \le \min\Big(\frac{y_p}{\|y\|_1}, \frac{\hat{y}_p}{\|\hat{y}\|_1}\Big)$$

$$\Rightarrow \min\Big(A, \frac{\hat{y}_p}{\|\hat{y}\|_1}\Big) = \min\Big(\frac{y_p}{\|y\|_1}, \frac{\hat{y}_p}{\|\hat{y}\|_1}\Big) = \min\Big(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\Big) \tag{85}$$

Similar to the previous calculus in (95), we show that $M_{C-1C} = m_{C-1}$.

If $p < C-1$, then $m_{C-1} = 0$ by definition. Moreover, we can rewrite m_{C-1} with a maximum: (101), as shown at the bottom of page 32.

If $p = C-1$ and $\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) = 0$, then by definition we have: (102), as shown at the bottom of page 32.

Since entries in m are non-negative, it follows $0 \leq m_{C-1} \leq 0$, leading to $m_{C-1} = 0$. It follows (103), as shown at the bottom of page 32.

Otherwise, if $p = C-1$ and $\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) > 0$, then $\frac{y_C}{\|y\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) = 0$, and by definition we have: (104), as shown at the bottom of page 32.

Otherwise, if $p = C$, then $m_{C-1} = \frac{y_{C-1}}{\|y\|_1} - \min(\frac{y_{C-1}}{\|y\|_1}, \frac{\hat{y}_{C-1}}{\|\hat{y}\|_1})$ by definition.

If $\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) > 0$, we have $\frac{y_C}{\|y\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) = 0$. Additionally, as

$$0 \leq \frac{y_{C-1}}{\|y\|_1} - \min(\frac{y_{C-1}}{\|y\|_1}, \frac{\hat{y}_{C-1}}{\|\hat{y}\|_1}), \tag{105}$$

$$\begin{aligned}
\|m\|_1 &= \sum_{k=1}^{C} m_k = m_p + \sum_{k=1}^{p-1} m_k + \underbrace{\sum_{k=p+1}^{C} m_k}_{=0, \text{ by definition of } p} \\
&= \frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) \underbrace{- \sum_{k=1}^{p-1} \frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}) + \sum_{k=1}^{p-1} \frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1})}_{=0} \\
&= \frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}).
\end{aligned} \tag{88}$$

$$\begin{aligned}
&\sum_{k=1}^{C} \Big[f\Big(\frac{y}{\|y\|_1} - m, \frac{\hat{y}}{\|\hat{y}\|_1} - E^C \big(\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) \big) \Big) + \underbrace{m \otimes E^C}_{\text{only column } C \text{ is non zero}} \Big]_{kj} \\
&= \sum_{k=1}^{C} f\Big(\frac{y}{\|y\|_1} - m, \frac{\hat{y}}{\|\hat{y}\|_1} - E^C \big(\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) \big) \Big)_{kj} \\
&= \underbrace{\frac{\hat{y}_j}{\|\hat{y}\|_1} - \overbrace{E^C_j}^{=0} \big(\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) \big)}_{\text{Lemma 3}} = \frac{\hat{y}_j}{\|\hat{y}\|_1}
\end{aligned} \tag{89}$$

$$\begin{aligned}
&\sum_{k=1}^{C} \Big[f\Big(\frac{y}{\|y\|_1} - m, \frac{\hat{y}}{\|\hat{y}\|_1} - E^C \big(\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) \big) \Big) + m \otimes E^C \Big]_{kC} \\
&= \sum_{k=1}^{C} f\Big(\frac{y}{\|y\|_1} - m, \frac{\hat{y}}{\|\hat{y}\|_1} - E^C \big(\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) \big) \Big)_{kC} + m_k \\
&= \underbrace{\frac{\hat{y}_C}{\|\hat{y}\|_1} - \overbrace{E^C_C}^{=1} \big(\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) \big) + \overbrace{\|m\|_1}^{= \frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})}}_{\text{Lemma 3}} \\
&= \frac{\hat{y}_C}{\|\hat{y}\|_1} \underbrace{- \frac{\hat{y}_C}{\|\hat{y}\|_1} + \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) + \frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})}_{=0} \\
&= \frac{\hat{y}_C}{\|\hat{y}\|_1}.
\end{aligned} \tag{90}$$

we can rewrite

$$m_{C-1} = \max\left(0, \frac{y_{C-1}}{\|y\|_1} - \min(\frac{y_{C-1}}{\|y\|_1}, \frac{\hat{y}_{C-1}}{\|\hat{y}\|_1})\right). \quad (106)$$

Finally, we have: (107), as shown at the bottom of page 33.

According to (106) and (107), it holds

$$m_{C-1} = \max\Big(0, \frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) - \sum_{k=1:k\neq C-1}^{C} \frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1})\Big). \quad (108)$$

If $\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) = 0$, we have: (109), as shown at the bottom of page 33.

Since all entries in m are non-negative, $0 \leq m_{C-1} \leq 0$, leading to $m_{C-1} = 0$. We can conclude (110), as shown at the bottom of page 33.

In conclusion, we prove (111), as shown at the bottom of page 33.

APPENDIX F
PROOF OF PROPOSITION 1

First, we establish a lower bound for the minimization problem:

$$\arg\min_{T(y,\hat{y})} \sum_{i,j=1}^{C} c(i,j)\pi_{ij} \quad (112)$$

$$\begin{aligned}
&\sum_{k=1}^{C}\left[f\Big(\frac{y}{\|y\|_1} - m, \frac{\hat{y}}{\|\hat{y}\|_1} - E^C\big(\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})\big)\Big) + m \otimes E^C\right]_{ik} \\
&= \sum_{k=1}^{C} f\Big(\frac{y}{\|y\|_1} - m, \frac{\hat{y}}{\|\hat{y}\|_1} - E^C\big(\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})\big)\Big)_{ik} + m_i \underbrace{E_k^C}_{\text{1 only if } k=C \text{ otherwise 0}} \\
&= \underbrace{\frac{y_i}{\|y\|_1}\overbrace{-m_i + m_i}^{=0}}_{\text{Lemma 3}} \\
&= \frac{y_i}{\|y\|_1}
\end{aligned} \quad (91)$$

$$\begin{aligned}
M_{1C} &= \left[f\left(\frac{y}{\|y\|_1} - m, \frac{\hat{y}}{\|\hat{y}\|_1} - E^C\left(\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})\right)\right) + m \otimes E^C\right]_{1C} \\
&= \frac{\left(\frac{y_1}{\|y\|_1} - m_1 - \min(\frac{y_1}{\|y\|_1}, \frac{\hat{y}_1}{\|\hat{y}\|_1})\right)\overbrace{\left(\frac{\hat{y}_C}{\|\hat{y}\|_1} - \frac{\hat{y}_C}{\|\hat{y}\|_1} + \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})\right)}^{=0}}{\|\frac{y}{\|y\|_1} - m - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})\|_1} + m_1 \\
&= m_1.
\end{aligned} \quad (95)$$

$$\sum_{k=1}^{p-1} \frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}) \leq \frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) \leq \sum_{k=1}^{p} \frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}). \quad (96)$$

$$\underbrace{\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) \leq \frac{y_1}{\|y\|_1} - \min(\frac{y_1}{\|y\|_1}, \frac{\hat{y}_1}{\|\hat{y}\|_1})}_{\text{according to (96)}} \Rightarrow m_1 = \underbrace{\min\left(\frac{y_1}{\|y\|_1} - \min(\frac{y_1}{\|y\|_1}, \frac{\hat{y}_1}{\|\hat{y}\|_1}), \frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})\right)}_{= \frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})}. \quad (97)$$

$$\underbrace{\frac{y_1}{\|y\|_1} - \min(\frac{y_1}{\|y\|_1}, \frac{\hat{y}_1}{\|\hat{y}\|_1}) \leq \frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})}_{\text{according to (96)}} \Rightarrow m_1 = \underbrace{\min\left(\frac{y_1}{\|y\|_1} - \min(\frac{y_1}{\|y\|_1}, \frac{\hat{y}_1}{\|\hat{y}\|_1}), \frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})\right)}_{= \frac{y_1}{\|y\|_1} - \min(\frac{y_1}{\|y\|_1}, \frac{\hat{y}_1}{\|\hat{y}\|_1})}. \quad (98)$$

Next, we show that certain elements of $T(y, \hat{y})$ achieve this bound and identify the properties of these minimizers. Finally, we demonstrate the converse: satisfying these properties for an element of $T(y, \hat{y})$ implies that it is a minimizer.

$$\underbrace{\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min\left(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}\right) \leq \sum_{k=1}^{p} \frac{y_k}{\|y\|_1} - \min\left(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\right)}_{\text{according to (96)}}$$

$$\Rightarrow \frac{\hat{y}_C}{\|\hat{y}\|_1} - \min\left(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}\right) - \sum_{k=1}^{p} \frac{y_k}{\|y\|_1} - \min\left(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\right) \leq 0$$

$$\Rightarrow \frac{\hat{y}_C}{\|\hat{y}\|_1} - \min\left(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}\right) - \sum_{k=1}^{p} \frac{y_k}{\|y\|_1} - \min\left(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\right) - \sum_{k=p+1:k\neq C-1}^{C} \frac{y_k}{\|y\|_1} - \min\left(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\right) \leq 0$$

$$\Rightarrow \frac{\hat{y}_C}{\|\hat{y}\|_1} - \min\left(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}\right) - \sum_{k=1:k\neq C-1}^{C} \frac{y_k}{\|y\|_1} - \min\left(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\right) \leq 0$$

$$\Rightarrow_{C-1} = \max\Bigg(0, \underbrace{\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min\left(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}\right) - \sum_{k=1:k\neq C-1}^{C} \frac{y_k}{\|y\|_1} - \min\left(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\right)}_{\leq 0}\Bigg). \tag{101}$$

$$\begin{aligned} m_{C-1} &= \underbrace{\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min\left(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}\right)}_{=0, \text{ by assumption}} - \sum_{k=1}^{C-2} \frac{y_k}{\|y\|_1} - \min\left(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\right) \\ &= -\sum_{k=1}^{C-2} \frac{y_k}{\|y\|_1} - \min\left(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\right) \\ &\leq 0. \end{aligned} \tag{102}$$

$$m_{C-1} = \max\Bigg(0, \underbrace{\overbrace{\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min\left(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}\right)}^{=0} - \sum_{k=1:k\neq C-1}^{C} \frac{y_k}{\|y\|_1} - \min\left(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\right)}_{\leq 0}\Bigg). \tag{103}$$

$$\begin{aligned} m_{C-1} &= \underbrace{\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min\left(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}\right) - \sum_{k=1}^{C-2} \frac{y_k}{\|y\|_1} - \min\left(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\right)}_{\geq 0, \text{ by design of } m} \\ &= \frac{\hat{y}_C}{\|\hat{y}\|_1} - \min\left(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}\right) - \sum_{k=1}^{C-2} \frac{y_k}{\|y\|_1} - \min\left(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\right) - \underbrace{\left(\frac{y_C}{\|y\|_1} - \min\left(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}\right)\right)}_{=0, \text{ by assumption}} \\ &= \underbrace{\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min\left(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}\right) - \sum_{k=1:k\neq C-1}^{C} \frac{y_k}{\|y\|_1} - \min\left(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\right)}_{\geq 0} \\ &= \max\left(0, \frac{\hat{y}_C}{\|\hat{y}\|_1} - \min\left(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}\right) - \sum_{k=1:k\neq C-1}^{C} \frac{y_k}{\|y\|_1} - \min\left(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\right)\right). \end{aligned} \tag{104}$$

G. MINIMAL BOUND

Let y and $\hat{y}$ be instances.

$$\begin{aligned}
&\min_{T(y,\hat{y})} \sum_{i,j=1}^{C} \underbrace{c(i,j)}_{0 \text{ if } i=j \text{ otherwise } 1} \\
&\pi_{ij} = \min_{T(y,\hat{y})} \sum_{i,j=1, i\neq j}^{C} \\
&\pi_{ij} = \min_{T(y,\hat{y})} \underbrace{\sum_{i,j=1}^{C} \pi_{ij}}_{=1} - \sum_{k=1}^{C} \\
&\pi_{kk} = \min_{T(y,\hat{y})} 1 - \sum_{k=1}^{C} \\
&\pi_{kk} = 1 - \max_{T(y,\hat{y})} \sum_{k=1}^{C} \pi_{kk}
\end{aligned} \tag{113}$$

Since π is $T(y,\hat{y})$, it holds $\sum_{k=1}^{C} \pi_{ik} = \frac{y_i}{\|y\|_1}$. All entries in π are non-negative, as a result $\pi_{ij} \leq \frac{y_i}{\|y\|_1}$. Similarly, we get $\pi_{ij} \leq \frac{\hat{y}_j}{\|\hat{y}\|_1}$. These observations allow us to conclude: (114), as shown at the bottom of the next page.

If there exist matrices $\pi \in T(y,\hat{y})$ with a diagonal defined by $\pi_{kk} = \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1})$, then these matrices achieve the lower bound. This set of matrices is denoted as $T^{\text{opt}}(y,\hat{y})$.

H. ELEMENTS ACHIEVING THE MINIMAL BOUND

Let's show that

$$\varnothing \neq T^{\text{opt}}(y,\hat{y}) := \left\{ \pi \in \mathbb{R}^{C\times C} : \pi_{kk} = \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}) \right\} \bigcap T(y,\hat{y}) \tag{115}$$

If $y = \hat{y}$, it is straightforward to see that the diagonal matrix with its diagonal entries equal to $\min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}) = \frac{y}{\|y\|_1} = \frac{\hat{y}}{\|\hat{y}\|_1}$ is in $T^{\text{opt}}(y,\hat{y})$. As a result $T^{\text{opt}}(y,\hat{y}) \neq \varnothing$. Now,

$$\underbrace{\sum_{k=1}^{C-1} \frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}) \leq \frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) \leq \sum_{k=1}^{C} \frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1})}_{\text{according to (96)}}$$

$$\Rightarrow \underbrace{\frac{y_C}{\|y\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})}_{=0} + \sum_{k=1}^{C-1} \frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}) \leq \frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) \leq \sum_{k=1}^{C} \frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1})$$

$$\Rightarrow \sum_{k=1}^{C} \frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}) \leq \frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) \leq \sum_{k=1}^{C} \frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1})$$

$$\Rightarrow \frac{y_C}{\|y\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) = \sum_{k=1}^{C} \frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1})$$

$$\Rightarrow \frac{y_C}{\|y\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) - \sum_{k=1:k\neq C-1}^{C} \frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}) = \frac{y_{C-1}}{\|y\|_1} - \min(\frac{y_{C-1}}{\|y\|_1}, \frac{\hat{y}_{C-1}}{\|\hat{y}\|_1}). \tag{107}$$

$$m_{C-1} = \frac{y_{C-1}}{\|y\|_1} - \min(\frac{y_{C-1}}{\|y\|_1}, \frac{\hat{y}_{C-1}}{\|\hat{y}\|_1}) \leq \underbrace{\sum_{k=1}^{C-1} \frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}) \leq \frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})}_{\text{according to (96)}} = 0 \tag{109}$$

$$m_{C-1} = \max\left(0, \overbrace{\frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1})}^{=0, \text{ by assumption}} - \underbrace{\sum_{k=1:k\neq C-1}^{C} \frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1})}_{\leq 0}\right). \tag{110}$$

$$M_{C-1\,C} = \max\left(0, \frac{\hat{y}_C}{\|\hat{y}\|_1} - \min(\frac{y_C}{\|y\|_1}, \frac{\hat{y}_C}{\|\hat{y}\|_1}) - \sum_{k=1:k\neq C-1}^{C} \frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1})\right). \tag{111}$$

we suppose $y \neq \hat{y}$. Let π^* be the matrix defined by:

$$\pi^*_{kk} = \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}),$$
$$\pi^*_{ij} = \Big(\frac{y_i}{\|y\|_1} - \min(\frac{y_i}{\|y\|_1}, \frac{\hat{y}_i}{\|\hat{y}\|_1})\Big)\Big(\frac{\hat{y}_j}{\|\hat{y}\|_1} - \min(\frac{y_j}{\|y\|_1}, \frac{\hat{y}_j}{\|\hat{y}\|_1})\Big)/\|\frac{y}{\|y\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})\|_1 \tag{116}$$

We will prove that $\pi^* \in T^{\text{opt}}(y, \hat{y})$. All entries of π^* are positive, and

$$\sum_{k=1:k\neq i}^{C} \pi^*_{ik} = \sum_{k=1:k\neq i}^{C} \Big(\frac{y_i}{\|y\|_1} - \min(\frac{y_i}{\|y\|_1}, \frac{\hat{y}_i}{\|\hat{y}\|_1})\Big)\Big(\frac{\hat{y}_k}{\|\hat{y}\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1})\Big)/\|\frac{y}{\|y\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})\|_1. \tag{117}$$

Since $\min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1})$ equals either $\frac{y_k}{\|y\|_1}$ or $\frac{\hat{y}_k}{\|\hat{y}\|_1}$, we have $\Big(\frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1})\Big)\Big(\frac{\hat{y}_k}{\|\hat{y}\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1})\Big) = 0$. It follows (118), as shown at the bottom of the next page.

which leads to $\sum_{k=1}^{C} \pi^*_{ik} = \frac{y_i}{\|y\|_1}$ because $\pi^*_{kk} = \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1})$. Similarly, we prove that $\sum_{k=1}^{C} \pi^*_{jk} = \frac{y_j}{\|y\|_1}$. In conclusion, $T^{\text{opt}}(y, \hat{y}) \neq \varnothing$, and all matrix π with its diagonal equals to $\min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})$ minimise our Kantorovitch problem.

I. ALL MINIMIZERS SATISFY THE DIAGONAL CONDITION

Now, let's prove the converse. Let π be a minimizer. We showed in Subsections XI-G and XI-H that

$$\min_{T(y,\hat{y})} \sum_{i,j=1}^{C} c(i,j)\pi_{ij} = 1 - \sum_{k=1}^{C} \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}), \tag{119}$$

which leads to

$$\underbrace{1 - \sum_{k=1}^{C} \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}) = \sum_{i,j=1:i\neq j}^{C} \pi_{ij}}_{\text{according to (119)}} = \underbrace{\sum_{i,j=1}^{C} \pi_{ij}}_{=1} - \sum_{k=1}^{C} \pi_{kk}$$
$$\Rightarrow \sum_{k=1}^{C} \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}) = \sum_{k=1}^{C} \pi_{kk} \tag{120}$$

We will demonstrate that

$$\sum_{k=1}^{C} \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}) = \sum_{k=1}^{C} \pi_{kk}$$
$$\Rightarrow \pi_{kk} = \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}) \quad \text{for} \quad k = 1 \dots C, \tag{121}$$

thus completing the proof. Let's proceed by contradiction. Assume $\sum_{k=1}^{C} \pi_{kk} = \sum_{k=1}^{C} \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1})$ and there is a index l such that $\pi_{ll} < \min(\frac{y_l}{\|y\|_1}, \frac{\hat{y}_l}{\|\hat{y}\|_1})$. Given that for all k, $\pi_{kk} \leq \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1})$, we have:

$$\sum_{k=1:k\neq l}^{C} \pi_{kk} \leq \sum_{k=1:k\neq l}^{C} \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}), \quad \text{and}$$
$$\pi_{ll} < \min(\frac{y_l}{\|y\|_1}, \frac{\hat{y}_l}{\|\hat{y}\|_1})$$
$$\Rightarrow \sum_{k=1}^{C} \pi_{kk} < \sum_{k=1}^{C} \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}), \tag{122}$$

contradicting our initial assumption. We just showed that if π is a minimizer, then its diagonal is defined by $\min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})$.

APPENDIX G
PROOF OF PROPOSITION 2

Let's define the assumption $\mathcal{H}$ as:

$$\mathcal{H} := \text{“}\textit{At most one class is underestimated or overestimated}\text{”}. \tag{123}$$

$$\pi_{ij} \leq \frac{y_i}{\|y\|_1}, \quad \text{and} \quad \pi_{ij} \leq \frac{\hat{y}_j}{\|\hat{y}\|_1} \Rightarrow \pi_{ij} \leq \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1})$$
$$\Rightarrow \sum_{k=1}^{C} \pi_{kk} \leq \sum_{k=1}^{C} \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1})$$
$$\Rightarrow \max_{T(y,\hat{y})} \sum_{k=1}^{C} \pi_{kk} \leq \sum_{k=1}^{C} \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}).$$
$$\Rightarrow 1 - \sum_{k=1}^{C} \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}) \leq 1$$
$$- \max_{T(y,\hat{y})} \sum_{k=1}^{C} \pi_{kk} = \min_{T(y,\hat{y})} \sum_{i,j=1}^{C} c(i,j)\pi_{ij}. \tag{114}$$

Firstly, we prove that under assumption $\mathcal{H}$, the set $T^{\text{opt}}(y, \hat{y})$ contains a single matrix. Secondly, we need to prove that if there is a unique solution, then the previous assumption is satisfied. We demonstrate the contrapositive, which is an equivalent implication: if $\mathcal{H}$ is not met, then $T^{\text{opt}}(y, \hat{y})$ contains an infinite number of matrices. To achieve this, we will show that $T^{\text{opt}}(y, \hat{y})$ is a convex set and then exhibit two different solutions within $T^{\text{opt}}(y, \hat{y})$. Consequently, any convex combination of these solutions will also belong to $T^{\text{opt}}(y, \hat{y})$, thereby establishing the existence of an infinite number of solutions.

J. $\mathcal{H}$ IMPLIES A UNIQUE ELEMENT IN $T^{OPT}(Y, \hat{Y})$

Let y and $\hat{y}$ be instances, and π be a matrix in $T^{\text{opt}}(y, \hat{y})$. We assume $\mathcal{H}$ is true.

The matrix π meets the following properties:

$$\pi \in \mathbb{R}_{\geq 0}^{C\times C}, \quad \sum_{j=1}^{C} \pi_{ij} = \frac{y_i}{\|y\|_1}, \quad \sum_{i=1}^{C} \pi_{ij} = \frac{\hat{y}_j}{\|\hat{y}\|_1}, \quad \text{and}$$
$$\pi_{ii} = \min\left(\frac{y_i}{\|y\|_1}, \frac{\hat{y}_i}{\|\hat{y}\|_1}\right). \tag{124}$$

In particular, we have:

$$\sum_{j=1, j\neq i}^{C} \pi_{ij} = \frac{y_i}{\|y\|_1} - \min\left(\frac{y_i}{\|y\|_1}, \frac{\hat{y}_i}{\|\hat{y}\|_1}\right), \quad \text{and}$$
$$\sum_{i=1, i\neq j}^{C} \pi_{ij} = \frac{\hat{y}_j}{\|\hat{y}\|_1} - \min\left(\frac{y_j}{\|y\|_1}, \frac{\hat{y}_j}{\|\hat{y}\|_1}\right). \tag{125}$$

Under $\mathcal{H}$, most of the sums are zero:

- If at most one class is underestimated, among all quantities $\frac{y_i}{\|y\|_1} - \min(\frac{y_i}{\|y\|_1}, \frac{\hat{y}_i}{\|\hat{y}\|_1})$ with i between 1 and C, at most one is not zero.
- If at most one class is overestimated, among all quantities $\frac{\hat{y}_j}{\|\hat{y}\|_1} - \min(\frac{y_j}{\|y\|_1}, \frac{\hat{y}_j}{\|\hat{y}\|_1})$ with j between 1 and C, at most one is not zero.

As a result, except for a row or a column, all off-diagonal terms are zero.

For example, let's suppose that only class j is overestimated:

$$\frac{\hat{y}_j}{\|\hat{y}\|_1} - \min\left(\frac{y_j}{\|y\|_1}, \frac{\hat{y}_j}{\|\hat{y}\|_1}\right) > 0, \quad \text{and}$$
$$\frac{\hat{y}_k}{\|\hat{y}\|_1} - \min\left(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\right) = 0, \quad \text{for} \quad k \neq j. \tag{126}$$

Consequently, when $l \neq j$,

$$\sum_{k=1:k\neq l}^{C} \pi_{kl} = 0 \quad \Rightarrow \quad \pi_{kl} = 0, \quad \text{for} \quad k \neq j, l \tag{127}$$

Moreover, by summing on rows different from row j, it follows:

$$\sum_{l=1, l\neq k}^{C} \pi_{kl} = \pi_{kj} = \frac{y_k}{\|y\|_1} - \min\left(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\right), \quad \text{for} \quad k \neq j. \tag{128}$$

We have demonstrated that all entries of π are determined by the assumption: *only one class is overestimated*. The diagonal entries are given by $T^{\text{opt}}(y, \hat{y})$'s constraints. Due to the constraints of row and column sums, the off-diagonal entries, except those in column j, are 0. The values in column j are $\frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1})$. Therefore, when only one class is overestimated, there is a unique matrix in $T^{\text{opt}}(y, \hat{y})$.

The scenario where, at most, one class is underestimated is similar. In conclusion, under $\mathcal{H}$, there is a unique matrix in $T^{\text{opt}}(y, \hat{y})$.

K. $T^{OPT}(Y, \hat{Y})$ IS A CONVEX SET

Let π and π' be two matrices in $T^{\text{opt}}(y, \hat{y})$. Let α be a real number in [0, 1].

Let's show that $\widetilde{\pi} := \alpha\pi + (1-\alpha)\pi'$ is in $T^{\text{opt}}(y, \hat{y})$, i.e.,

$$\widetilde{\pi} \in \mathbb{R}_{\geq 0}^{C\times C}, \quad \sum_{j=1}^{C} \widetilde{\pi}_{ij} = \frac{y_i}{\|y\|_1}, \quad \sum_{i=1}^{C} \widetilde{\pi}_{ij} = \frac{\hat{y}_j}{\|\hat{y}\|_1}, \quad \text{and}$$
$$\widetilde{\pi}_{ij} = \min\left(\frac{y_i}{\|y\|_1}, \frac{\hat{y}_i}{\|\hat{y}\|_1}\right). \tag{129}$$

As a convex combination of positive real value matrices, $\widetilde{\pi}$ is in $\mathbb{R}_{\geq 0}^{C\times C}$. Moreover, the following equalities hold:

$$\sum_{j=1}^{C} \widetilde{\pi}_{ij} = \sum_{j=1}^{C} \alpha\pi_{ij} + (1-\alpha)\pi'_{ij} = \alpha \sum_{j=1}^{C} \pi_{ij} + (1-\alpha) \sum_{j=1}^{C} \pi'_{ij}$$
$$= \alpha \frac{y_i}{\|y\|_1} + (1-\alpha)\frac{y_i}{\|y\|_1} = \frac{y_i}{\|y\|_1}. \tag{130}$$

$$\sum_{k=1:k\neq i}^{C} \pi^*_{ik} = \sum_{k=1}^{C} \left(\frac{y_i}{\|y\|_1} - \min\left(\frac{y_i}{\|y\|_1}, \frac{\hat{y}_i}{\|\hat{y}\|_1}\right)\right)\left(\frac{\hat{y}_k}{\|\hat{y}\|_1} - \min\left(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\right)\right) / \left\|\frac{y}{\|y\|_1} - \min\left(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}\right)\right\|_1$$
$$= \left(\frac{y_i}{\|y\|_1} - \min\left(\frac{y_i}{\|y\|_1}, \frac{\hat{y}_i}{\|\hat{y}\|_1}\right)\right) \underbrace{\sum_{k=1}^{C} \frac{\frac{\hat{y}_k}{\|\hat{y}\|_1} - \min\left(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\right)}{\left\|\frac{y}{\|y\|_1} - \min\left(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}\right)\right\|_1}}_{=1}$$
$$= \frac{y_i}{\|y\|_1} - \min\left(\frac{y_i}{\|y\|_1}, \frac{\hat{y}_i}{\|\hat{y}\|_1}\right), \tag{118}$$

Similarly, $\sum_{i=1}^{C} \widetilde{\pi}_{ij} = \frac{\hat{y}_j}{\|\hat{y}\|_1}$ holds. Finally,

$$\begin{aligned}\widetilde{\pi}_{kk} &= \alpha\pi_{kk} + (1-\alpha)\pi'_{kk} \\ &= \alpha \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}) + (1-\alpha) \\ &\min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}) = \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}),\end{aligned} \tag{131}$$

this leads to $\widetilde{\pi} \in T^{\text{opt}}(y, \hat{y})$

In conclusion, $T^{\text{opt}}(y, \hat{y})$ is a convex set.

L. NOT $\mathcal{H}$ IMPLIES A INFINITE NUMBER OF SOLUTIONS

We now assume that the hypothesis $\mathcal{H}$ is unmet. As a result, there are at least two overestimated classes, denoted as o and o', and two underestimated classes, denoted as u and u'. We will demonstrate the existence of two distinct matrices M^a and M^b in $T^{\text{opt}}(y, \hat{y})$, then conclude that there are an infinite number of matrices in $T^{\text{opt}}(y, \hat{y})$, given by the set of all convex combinations of M^a and M^b.

Let σ and τ be two permutations on the integer set from 1 to C. The permutation σ is defined by $\sigma = (o\ C)(u\ 1)(u'\ C-1)$, while the permutation τ is defined by $\tau = (o\ C)(u'\ 1)(u\ C-1)$. Let s, $\hat{s}$, t, and $\hat{t}$ be vectors of $\mathbb{R}^C_{\geq 0}$ such that $s_{\sigma(k)} = y_k$, $\hat{s}_{\sigma(k)} = \hat{y}_k$, and $t_{\tau(k)} = y_k$, $\hat{t}_{\tau(k)} = \hat{y}_k$ for all k from 1 to C (s stands for sigma, while t stands for tau).

According to Lemma 4, the matrices

$$\begin{aligned}\widetilde{M}^a &:= f\Big(\frac{s}{\|s\|_1} - m, \frac{\hat{s}}{\|\hat{s}\|_1} - E^C\big(\frac{\hat{s}_C}{\|\hat{s}\|_1} - \min(\frac{s_C}{\|s\|_1}, \frac{\hat{s}_C}{\|\hat{s}\|_1})\big)\Big) \\ &\quad + m \otimes E^C, \quad \text{and} \\ \widetilde{M}^b &:= f\Big(\frac{t}{\|t\|_1} - m, \frac{\hat{t}}{\|\hat{t}\|_1} - E^C\big(\frac{\hat{t}_C}{\|\hat{t}\|_1} - \min(\frac{t_C}{\|t\|_1}, \frac{\hat{t}_C}{\|\hat{t}\|_1})\big)\Big) \\ &\quad + m \otimes E^C,\end{aligned} \tag{132}$$

are in $T^{\text{opt}}(s, \hat{s})$, and $T^{\text{opt}}(t, \hat{t})$ respectively. According to Lemma 2, the matrices M^a, and M^b, defined as,

$$M^a_{ij} = \widetilde{M}^a_{\sigma(i)\sigma(j)}, \quad \text{and} \quad M^b_{ij} = \widetilde{M}^b_{\tau(i)\tau(j)}, \quad \text{for} \quad i, j = 1 \dots C, \tag{133}$$

are both in $T^{\text{opt}}(y, \hat{y})$.

Let's show that M^a and M^b are different. (134), as shown at the bottom of the next page.

The last equality holds by definition of s, and because $\|s\|_1 = \|y\|_1$ and $\|\hat{s}\|_1 = \|\hat{y}\|_1$. (135), as shown at the bottom of the next page.

The last equality holds by definition of s, and because $\|s\|_1 = \|y\|_1$ and $\|\hat{s}\|_1 = \|\hat{y}\|_1$.

In the same way, we show that

$$M^b_{u'o} = M^a_{uo}, \quad \text{and} \quad M^b_{uo} = M^a_{u'o}. \tag{136}$$

As a result, to demonstrate that M^a and M^b are different, we only have to prove that $M^a_{uo} \neq M^a_{u'o}$, implying different entry uo and different entry $u'o$ in M^a and M^b.

If $M^a_{uo} = \frac{y_u}{\|y\|_1} - \min(\frac{y_u}{\|y\|_1}, \frac{\hat{y}_u}{\|\hat{y}\|_1})$, then firstly $0 < M^a_{uo}$ by assumption. If $M^a_{u'o} = 0$, then these entries are different. Else, if

$$M^a_{u'o} = \frac{\hat{y}_o}{\|\hat{y}\|_1} - \min\left(\frac{y_o}{\|y\|_1}, \frac{\hat{y}_o}{\|\hat{y}\|_1}\right) - \sum_{k=1:k\neq u'}^{C} \frac{y_k}{\|y\|_1} - \min\left(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\right), \tag{137}$$

then we will prove that $M^a_{uo} \neq M^a_{u'o}$ by contradiction. Let's suppose that (138), as shown at the bottom of the next page.

The left term is positive, whereas the right term is negative, leading to a contradiction, proving that $M^a_{uo} \neq M^a_{u'o}$.

If $M^a_{uo} = \frac{\hat{y}_o}{\|\hat{y}\|_1} - \min(\frac{y_o}{\|y\|_1}, \frac{\hat{y}_o}{\|\hat{y}\|_1})$, then firstly $0 < M^a_{uo}$ by assumption. If $M^a_{u'o} = 0$, then these entries are different. Else, if

$$\begin{aligned}M^a_{u'o} &= \frac{\hat{y}_o}{\|\hat{y}\|_1} - \min\left(\frac{y_o}{\|y\|_1}, \frac{\hat{y}_o}{\|\hat{y}\|_1}\right) - \sum_{k=1:k\neq u'}^{C} \frac{y_k}{\|y\|_1} - \min\left(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\right) \\ &\leq \frac{\hat{y}_o}{\|\hat{y}\|_1} - \min\left(\frac{y_o}{\|y\|_1}, \frac{\hat{y}_o}{\|\hat{y}\|_1}\right) \\ &\quad - \underbrace{\left(\frac{y_u}{\|y\|_1} - \min(\frac{y_u}{\|y\|_1}, \frac{\hat{y}_u}{\|\hat{y}\|_1})\right)}_{>0, \text{ by assumption}} \\ &< \frac{\hat{y}_o}{\|\hat{y}\|_1} - \min\left(\frac{y_o}{\|y\|_1}, \frac{\hat{y}_o}{\|\hat{y}\|_1}\right) = M^a_{u'o},\end{aligned} \tag{139}$$

implying that these entries are different.

In conclusion, we prove that $M^a_{uo} \neq M^a_{u'o}$, which implies $M^a \neq M^b$. As $T^{\text{opt}}(y, \hat{y})$ is a convex set, any convex combination $\alpha M^a + (1-\alpha)M^b$ with $\alpha \in [0, 1]$ is in $T^{\text{opt}}(y, \hat{y})$. In conclusion, when two classes or more are underestimated, and two classes or more are overestimated, then there is an infinite number of solutions.

APPENDIX H
PROOF OF PROPOSITION 3

Let's show that the matrix

$$\begin{aligned}&\pi^*(y, \hat{y}) \\ &= \text{diag}\left(\min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})\right) \\ &+ \frac{\left(\frac{y}{\|y\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})\right) \otimes \left(\frac{\hat{y}}{\|\hat{y}\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})\right)}{\|\frac{y}{\|y\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})\|_1}\end{aligned} \tag{140}$$

is in $T^{\text{opt}}(y, \hat{y})$ and maximises the entropy. Moreover, let's demonstrate that $\pi^*(y, \hat{y})$ becomes

$$\pi^*(y, \hat{y}) = \text{diag}\left(\min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})\right), \tag{141}$$

when $y = \hat{y}$ by continuous extension.

M. CASE Y = $\hat{Y}$

Let $(y, \hat{y})$ be an instance such that $y \neq \hat{y}$. Let η be a real values such that $\eta \leq \| \frac{y}{\|y\|_1} - \frac{\hat{y}}{\|\hat{y}\|_1} \|_1 \leq 2\eta$. We will demonstrate that

$$\frac{\left(\frac{y}{\|y\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})\right) \otimes \left(\frac{\hat{y}}{\|\hat{y}\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})\right)}{\| \frac{y}{\|y\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}) \|_1} \underset{\eta \to 0}{\to} 0_{\mathcal{M}_C}, \tag{142}$$

$E :=$ where $0_{\mathcal{M}_C}$ is the zero matrix of size C.

$$\begin{aligned}
M^a_{uo} &= \widetilde{M}^a_{\sigma(u)\sigma(o)} \\
&= \widetilde{M}^a_{1C} \\
&= \left[f\left(\frac{s}{\|s\|_1} - m, \frac{\hat{s}}{\|\hat{s}\|_1} - E^C \left(\frac{\hat{s}_C}{\|\hat{s}\|_1} - \min(\frac{s_C}{\|s\|_1}, \frac{\hat{s}_C}{\|\hat{s}\|_1}) \right) \right) + m \otimes E^C \right]_{1C} \\
&= \underbrace{\min\left(\frac{s_1}{\|s\|_1} - \min(\frac{s_1}{\|s\|_1}, \frac{\hat{s}_1}{\|\hat{s}\|_1}), \frac{\hat{s}_C}{\|\hat{s}\|_1} - \min(\frac{s_C}{\|s\|_1}, \frac{\hat{s}_C}{\|\hat{s}\|_1}) \right)}_{\text{using Lemma 4}} \\
&= \min\left(\frac{y_u}{\|y\|_1} - \min(\frac{y_u}{\|y\|_1}, \frac{\hat{y}_u}{\|\hat{y}\|_1}), \frac{\hat{y}_o}{\|\hat{y}\|_1} - \min(\frac{y_o}{\|y\|_1}, \frac{\hat{y}_o}{\|\hat{y}\|_1}) \right)
\end{aligned} \tag{134}$$

$$\begin{aligned}
M^a_{u'o} &= \widetilde{M}^a_{\sigma(u')\sigma(o)} \\
&= \widetilde{M}^a_{C-1\,C} \\
&= \left[f\left(\frac{s}{\|s\|_1} - m, \frac{\hat{s}}{\|\hat{s}\|_1} - E^C \left(\frac{\hat{s}_C}{\|\hat{s}\|_1} - \min(\frac{s_C}{\|s\|_1}, \frac{\hat{s}_C}{\|\hat{s}\|_1}) \right) \right) + m \otimes E^C \right]_{C-1\,C} \\
&= \underbrace{\max\left(0, \frac{\hat{s}_C}{\|\hat{s}\|_1} - \min\left(\frac{s_C}{\|s\|_1}, \frac{\hat{s}_C}{\|\hat{s}\|_1} \right) - \sum_{k=1:k\neq C-1}^{C} \frac{s_k}{\|s\|_1} - \min\left(\frac{s_k}{\|s\|_1}, \frac{\hat{s}_k}{\|\hat{s}\|_1} \right) \right)}_{\text{using Lemma 4}} \\
&= \max\left(0, \frac{\hat{y}_o}{\|\hat{y}\|_1} - \min\left(\frac{y_o}{\|y\|_1}, \frac{\hat{y}_o}{\|\hat{y}\|_1} \right) - \sum_{k=1:k\neq u'}^{C} \frac{y_k}{\|y\|_1} - \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}) \right)
\end{aligned} \tag{135}$$

$$\begin{aligned}
& M^a_{uo} = M^a_{u'o} \\
& \Rightarrow \frac{y_u}{\|y\|_1} - \min(\frac{y_u}{\|y\|_1}, \frac{\hat{y}_u}{\|\hat{y}\|_1}) = \frac{\hat{y}_o}{\|\hat{y}\|_1} - \min\left(\frac{y_o}{\|y\|_1}, \frac{\hat{y}_o}{\|\hat{y}\|_1} \right) - \sum_{k=1:k\neq u'}^{C} \frac{y_k}{\|y\|_1} - \min\left(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1} \right) \\
& \Rightarrow \underbrace{\frac{y_u}{\|y\|_1} - \min(\frac{y_u}{\|y\|_1}, \frac{\hat{y}_u}{\|\hat{y}\|_1}) + \frac{y_{u'}}{\|y\|_1} - \min(\frac{y_{u'}}{\|y\|_1}, \frac{\hat{y}_{u'}}{\|\hat{y}\|_1})}_{>0, \text{ by assumption}} = \frac{\hat{y}_o}{\|\hat{y}\|_1} - \min\left(\frac{y_o}{\|y\|_1}, \frac{\hat{y}_o}{\|\hat{y}\|_1} \right) - \sum_{k=1}^{C} \frac{y_k}{\|y\|_1} - \min\left(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1} \right) \\
& \qquad = \frac{\hat{y}_o}{\|\hat{y}\|_1} - \min\left(\frac{y_o}{\|y\|_1}, \frac{\hat{y}_o}{\|\hat{y}\|_1} \right) - \underbrace{\| \frac{y}{\|y\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}) \|_1}_{= \| \frac{\hat{y}}{\|\hat{y}\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}) \|_1, \text{ according to Lemma 1}} \\
& \qquad = \frac{\hat{y}_o}{\|\hat{y}\|_1} - \min\left(\frac{y_o}{\|y\|_1}, \frac{\hat{y}_o}{\|\hat{y}\|_1} \right) - \| \frac{\hat{y}}{\|\hat{y}\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}) \|_1 \\
& \qquad \leq \frac{\hat{y}_o}{\|\hat{y}\|_1} - \min\left(\frac{y_o}{\|y\|_1}, \frac{\hat{y}_o}{\|\hat{y}\|_1} \right) - \left(\frac{\hat{y}_o}{\|\hat{y}\|_1} - \min(\frac{y_o}{\|y\|_1}, \frac{\hat{y}_o}{\|\hat{y}\|_1}) + \frac{\hat{y}'_o}{\|\hat{y}\|_1} - \min(\frac{y'_o}{\|y\|_1}, \frac{\hat{y}'_o}{\|\hat{y}\|_1}) \right) \\
& \qquad \leq \underbrace{-\frac{\hat{y}'_o}{\|\hat{y}\|_1} + \min(\frac{y'_o}{\|y\|_1}, \frac{\hat{y}'_o}{\|\hat{y}\|_1})}_{<0, \text{ by assumption}}
\end{aligned} \tag{138}$$

Let i, j be two integers. The entry E_{ij} is

$$E_{ij} = \frac{\left(\frac{y_i}{\|y\|_1} - \min(\frac{y_i}{\|y\|_1}, \frac{\hat{y}_i}{\|\hat{y}\|_1})\right)\left(\frac{\hat{y}_j}{\|\hat{y}\|_1} - \min(\frac{y_j}{\|y\|_1}, \frac{\hat{y}_j}{\|\hat{y}\|_1})\right)}{\|\frac{y}{\|y\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})\|_1}. \tag{143}$$

We will prove that $E_{ij} \underset{\eta\to 0}{\to} 0$. To do so, we will establish an upper bound for E_{ij} that depends on η. Firstly, we will find an upper bound for the numerator and, secondly, a lower bound for the denominator.

For the numerator:

$$\underbrace{\frac{y_i}{\|y\|_1} - \min(\frac{y_i}{\|y\|_1}, \frac{\hat{y}_i}{\|\hat{y}\|_1})}_{=0 \text{ if } \frac{y_i}{\|y\|_1} \le \frac{\hat{y}_i}{\|\hat{y}\|_1}, \text{ otherwise } \frac{y_i}{\|y\|_1} - \frac{\hat{y}_i}{\|\hat{y}\|_1}} \le |\frac{y_i}{\|y\|_1} - \frac{\hat{y}_i}{\|\hat{y}\|_1}| \le 2\eta. \tag{144}$$

Similarly, $\frac{\hat{y}_j}{\|\hat{y}\|_1} - \min(\frac{y_j}{\|y\|_1}, \frac{\hat{y}_j}{\|\hat{y}\|_1}) \le 2\eta$. As a result, an upper bound for the numerator is $4\eta^2$.

For the denominator: (145), as shown at the bottom of the next page. a lower bound for the denominator is $\eta/2$.

Since, by design, $0 \le E_{ij}$, it follows that $|E_{ij}| \le \frac{4\eta^2}{\eta/2} = 8\eta$ for all i and j between 1 and C, which leads to

$$E_{ij} \underset{\eta\to 0}{\to} 0 \quad \text{for} \quad i, j = 1 \dots C \quad \Rightarrow \quad E \underset{\eta\to 0}{\to} 0_{\mathcal{M}_C}. \tag{146}$$

We proved, by continuous extension in $y = \hat{y}$, that $\pi^*(y, \hat{y})$ equals to

$$\pi^*(y, \hat{y}) = \text{diag}\left(\min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})\right). \tag{147}$$

In this case, it is straightforward to see that $\pi^*(y, \hat{y}) \in T^{\text{opt}}(y, \hat{y})$. According to Proposition 2, there is only one transference plan in $T^{\text{opt}}(y, \hat{y})$, which implies that $T^{\text{opt}}(y, \hat{y}) = \{\pi^*(y, \hat{y})\}$. Consequently, since $T^{\text{opt}}(y, \hat{y})$ has only one element, it must be the one that maximizes the entropy.

N. FRAMEWORK TO SOLVE THE CASE $Y \neq \hat{Y}$

This subsection introduces the framework for addressing the $y \neq \hat{y}$ scenario. The major aim of the next subsections is to solve an optimization problem denoted as ($\mathfrak{A}$). To achieve this, we employ a series of strategies that lead to an easier optimization problem, denoted as ($\mathfrak{B}$). Firstly, we demonstrate that it is unnecessary to consider all the π_{ij} variables. Following this, we present a more favorable optimization problem where the entropy is differentiable over the set of constraints, specifically addressing the issue of $\pi_{ij} = 0$. Finally, we propose an alternative function to entropy that simplifies the calculation of derivatives. These strategies lead to ($\mathfrak{B}$), and we will see later that solving ($\mathfrak{B}$) allows us to solve ($\mathfrak{A}$).

Let H be the entropy of a random variable in information theory. We aim to solve the following problem denoted as ($\mathfrak{A}$):

$$(\mathfrak{A}) \quad \underset{T^{\text{opt}}(y,\hat{y})}{\arg\max}\, H(\pi), \quad \text{with } H(\pi) = -\sum_{i,j=1}^{C} \pi_{ij} \ln(\pi_{ij}), \tag{148}$$

where we set $0 \ln(0) = 0$ within the entropy.

Firstly, let's show that it is not necessary to consider all the π_{ij} variables to solve this problem. The entries of π are partially known. $T^{\text{opt}}(y, \hat{y})$ provides values for the diagonal entries, whereas for the off-diagonal entries, when

$$\frac{y_i}{\|y\|_1} - \min(\frac{y_i}{\|y\|_1}, \frac{\hat{y}_i}{\|\hat{y}\|_1}) = 0, \quad \text{or} \quad \frac{\hat{y}_j}{\|\hat{y}\|_1} - \min(\frac{y_j}{\|y\|_1}, \frac{\hat{y}_j}{\|\hat{y}\|_1}) = 0, \tag{149}$$

π_{ij} equals 0. For instance, if $\frac{y_i}{\|y\|_1} - \min(\frac{y_i}{\|y\|_1}, \frac{\hat{y}_i}{\|\hat{y}\|_1}) = 0$, then

$$\pi \in \mathbb{R}^{C\times C}_{\ge 0}, \quad \text{and} \quad \sum_{j=1}^{C} \pi_{ij} = \frac{y_i}{\|y\|_1} - \min(\frac{y_j}{\|y\|_1}, \frac{\hat{y}_j}{\|\hat{y}\|_1}) = 0 \Rightarrow 0 \le \pi_{ij} \le \sum_{j=1}^{C} \pi_{ij} \le 0 \Rightarrow \pi_{ij} = 0. \tag{150}$$

The same applies if $\frac{\hat{y}_j}{\|\hat{y}\|_1} - \min(\frac{y_j}{\|y\|_1}, \frac{\hat{y}_j}{\|\hat{y}\|_1}) = 0$. Consequently, we are interested in computing the off-diagonal values of π in the case where

$$\frac{y_i}{\|y\|_1} - \min(\frac{y_i}{\|y\|_1}, \frac{\hat{y}_i}{\|\hat{y}\|_1}) > 0, \quad \text{and} \quad \frac{\hat{y}_j}{\|\hat{y}\|_1} - \min(\frac{y_j}{\|y\|_1}, \frac{\hat{y}_j}{\|\hat{y}\|_1}) > 0. \tag{151}$$

We define three index sets $\mathcal{I}$, $\mathcal{J}$, and $\mathcal{K}$ as follows:

$$\mathcal{I} = \{i : \frac{y_i}{\|y\|_1} - \min(\frac{y_i}{\|y\|_1}, \frac{\hat{y}_i}{\|\hat{y}\|_1}) > 0\}, \quad \mathcal{J} = \{j : \frac{\hat{y}_j}{\|\hat{y}\|_1} - \min(\frac{y_j}{\|y\|_1}, \frac{\hat{y}_j}{\|\hat{y}\|_1}) > 0\}, \quad \text{and} \quad \mathcal{K} = \mathcal{I} \times \mathcal{J}. \tag{152}$$

Considering ($\mathfrak{A}$), all $\pi_{ij} \in T^{\text{opt}}(y, \hat{y})$ such that $(i, j) \notin \mathcal{K}$ can be considered as constants. More precisely, (153), as shown at the bottom of the next page.

where $\gamma(y, \hat{y})$ is a constant relative to $\pi \in T^{\text{opt}}(y, \hat{y})$, defined by

$$\gamma(y, \hat{y}) = -\sum_{k=1}^{C} \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}) \ln \min(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}). \tag{154}$$

To leverage this observation, we consider vectors of size $\#\mathcal{K}$ where each entry corresponds to an entry π_{ij} with $(i, j) \in \mathcal{K}$. We define a vector $p \in \mathbb{R}^{\#\mathcal{K}}_{\ge} 0$, which stacks all π_{ij} in

lexicographic order for $(i,j) \in \mathcal{K}$. Additionally, let ι be a function $\iota : \mathcal{K} \to \mathbb{N}$ such that $p_{\iota(i,j)} = \pi_{ij}$ for all $(i,j) \in \mathcal{K}$. We can now consider the entropy of entries π_{ij} with $(i,j) \in \mathcal{K}$ only. For simplicity, we use the same notation H:

$$H(p) = -\sum_{(i,j)\in\mathcal{K}} \pi_{ij} \ln(\pi_{ij}). \tag{155}$$

Given that $T^{\text{opt}}(y,\hat{y})$ contains matrices of size C, it is not appropriate for vectors of size $\#\mathcal{K}$. Instead, we introduce a similar set $\widetilde{T}(y,\hat{y})$ defined by:

$$\widetilde{T}(y,\hat{y}) = \bigcap_{i\in\mathcal{I}} \{g_i(p) = 0\} \bigcap_{j\in\mathcal{J}} \{\tilde{g}_j(p) = 0\} \bigcap \mathbb{R}^{\#\mathcal{K}}_{\geq 0} \tag{156}$$

where the functions g_i and $\tilde{g}_j$ are defined for $p \in \mathbb{R}^{\#\mathcal{K}}$ as follows:

$$\begin{aligned} g_i &: p \mapsto \min\left(\frac{y_i}{\|y\|_1}, \frac{\hat{y}_i}{\|\hat{y}\|_1}\right) - \frac{y_i}{\|y\|_1} + \sum_{j\in\mathcal{J}} \pi_{ij} \\ &\text{for all } i \in \mathcal{I}, \quad \text{and} \\ \tilde{g}_j &: p \mapsto \min\left(\frac{y_j}{\|y\|_1}, \frac{\hat{y}_j}{\|\hat{y}\|_1}\right) - \frac{\hat{y}_j}{\|\hat{y}\|_1} + \sum_{i\in\mathcal{I}} \pi_{ij} \quad \text{for all } j \in \mathcal{J} \end{aligned} \tag{157}$$

We observe that if we find $p^* \in \widetilde{T}(y,\hat{y})$ that maximizes the entropy, it will solve our optimization problem ($\mathfrak{A}$). Indeed, for all $p \in \widetilde{T}(y,\hat{y})$, we have

$$\begin{aligned} H(p) \leq H(p^*) \Rightarrow H(p) + \underbrace{\gamma(y,\hat{y})}_{\text{see (154)}} \leq H(p^*) \\ + \underbrace{\gamma(y,\hat{y})}_{\text{see (154)}} \Rightarrow H(\pi) \leq H(\pi^*) \end{aligned} \tag{158}$$

where π corresponds to the matrix

$$\begin{aligned} &\pi_{ij} = p_{\iota(i,j)} \text{ if } (i,j) \in \mathcal{K}, \quad \pi_{ij} = 0 \text{ if } (i,j) \notin \mathcal{K} \text{ and } i \neq j, \\ &\text{and } \pi_{kk} = \min\left(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\right) \text{ for } k = 1 \ldots C. \end{aligned} \tag{159}$$

The matrix π^* is defined similarly with p^*. Matrices π and π^* are in $T^{\text{opt}}(y,\hat{y})$.

Secondly, let's establish a more practical framework where the entropy is differentiable over the set of constraints. Some π_{ij} could be zero by design, making H non-differentiable at those points. To address this issue, we start by solving a different optimization problem. Let $\varepsilon > 0$ be a "sufficiently" small positive real number (the term "sufficiently" will be discussed in Subsection XI-P). The sum of a real number x and a vector v is defined as $(v+x)k = vk + x$. Let H_ε be the function such that $H_\varepsilon : p \mapsto H(p+\varepsilon)$ for all $p \in]-\varepsilon, +\infty[^{\#\mathcal{K}}$. We begin by solving:

$$\arg\max_{\widetilde{T}(y,\hat{y})} H_\varepsilon(p) = \arg\max_{\widetilde{T}(y,\hat{y})} H(p+\varepsilon) \tag{160}$$

$$\begin{aligned} \eta \leq \|\frac{y}{\|y\|_1} - \frac{\hat{y}}{\|\hat{y}\|_1}\|_1 &= \sum_{k=1}^{C} |\frac{y_k}{\|y\|_1} - \frac{\hat{y}_k}{\|\hat{y}\|_1}| = \sum_{k=1:\frac{\hat{y}_k}{\|\hat{y}\|_1} < \frac{y_k}{\|y\|_1}}^{C} \frac{y_k}{\|y\|_1} - \frac{\hat{y}_k}{\|\hat{y}\|_1} + \sum_{k=1:\frac{y_k}{\|y\|_1} < \frac{\hat{y}_k}{\|\hat{y}\|_1}}^{C} \frac{\hat{y}_k}{\|\hat{y}\|_1} - \frac{y_k}{\|y\|_1} \\ &= \sum_{k=1}^{C} \frac{y_k}{\|y\|_1} - \min\left(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\right) + \sum_{k=1}^{C} \frac{\hat{y}_k}{\|\hat{y}\|_1} - \min\left(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\right) \\ &= \|\frac{y}{\|y\|_1} - \min\left(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}\right)\|_1 + \underbrace{\|\frac{\hat{y}}{\|\hat{y}\|_1} - \min\left(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}\right)\|_1}_{= \|\frac{y}{\|y\|_1} - \min\left(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}\right)\|_1, \text{ by Lemma 1}} \\ &= 2\|\frac{y}{\|y\|_1} - \min\left(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}\right)\|_1, \end{aligned} \tag{145}$$

$$\begin{aligned} H(\pi) = -\sum_{i,j=1}^{C} \pi_{ij} \ln(\pi_{ij}) &= -\sum_{(i,j)\in\mathcal{K}} \pi_{ij} \ln(\pi_{ij}) - \sum_{(i,j)\notin\mathcal{K}} \underbrace{\pi_{ij} \ln(\pi_{ij})}_{\pi_{ij}=0, \text{ if } i\neq j, \text{ else } \pi_{kk} = \min\left(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\right)} \\ &= -\sum_{(i,j)\in\mathcal{K}} \pi_{ij} \ln(\pi_{ij}) - \underbrace{\sum_{k=1}^{C} \min\left(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\right) \ln \min\left(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\right)}_{\text{constant for all } \pi \in T^{\text{opt}}(y,\hat{y})} \\ &= -\sum_{(i,j)\in\mathcal{K}} \pi_{ij} \ln(\pi_{ij}) + \gamma(y,\hat{y}) \end{aligned} \tag{153}$$

The function $H_\varepsilon(p)$ is $\mathcal{C}^1(]-\varepsilon,+\infty[^{\#\mathcal{K}},\mathbb{R})$ because the entropy is $\mathcal{C}^1(]0,+\infty[^{\#\mathcal{K}},\mathbb{R})$. In particular, H_ε is $\mathcal{C}^1(\widetilde{T}(y,\hat{y}),\mathbb{R})$.

Thirdly, to simplify the calculation of derivatives, we introduce the function $\widetilde{H}_\varepsilon(p)$, defined by

$$\begin{aligned}\widetilde{H}_\varepsilon(p) &= -\sum_{(i,j)\in\mathcal{K}} (\pi_{ij}+\varepsilon)\left(\ln(\pi_{ij}+\varepsilon)-1\right)\\ &= -\sum_{(i,j)\in\mathcal{K}} (\pi_{ij}+\varepsilon)\ln(\pi_{ij}+\varepsilon) + \underbrace{\sum_{(i,j)\in\mathcal{K}} (\pi_{ij}+\varepsilon)}_{\text{constant relative to } p}\end{aligned} \tag{161}$$

More precisely, the last term equals to (162), as shown at the bottom of the next page.

Since adding a constant does not change the set of arguments that maximize a function, we have

$$\arg\max_{\widetilde{T}(y,\hat{y})} \widetilde{H}_\varepsilon(p) = \arg\max_{\widetilde{T}(y,\hat{y})} H_\varepsilon(p) = \arg\max_{\widetilde{T}(y,\hat{y})} H(p+\varepsilon) \tag{163}$$

Moreover, it does not change the class property of the function either. As a result, $\widetilde{H}_\varepsilon$ is in $\mathcal{C}^1(]-\varepsilon,+\infty[^{\#\mathcal{K}},\mathbb{R})$. We denote $(\mathfrak{B})$ the problem

$$(\mathfrak{B}) \quad \arg\max_{\widetilde{T}(y,\hat{y})} \widetilde{H}_\varepsilon(p). \tag{164}$$

In conclusion, in the following subsections, we will solve problem $(\mathfrak{B})$, which will enable us to solve problem $(\mathfrak{A})$.

O. $Y \neq \hat{Y}$, GRADIENT EQUALITY

We observe that $H_\varepsilon(p)$ is a strictly concave function because it is the composition of an affine function, $p \mapsto p+\varepsilon$, and a strictly concave function, the entropy. Since adding a constant to a function does not change its concavity, $\widetilde{H}_\varepsilon$ is also strictly concave.

Suppose that there exist $p^\varepsilon \in \widetilde{T}(y,\hat{y})$ such that

$$\frac{\partial \widetilde{H}_\varepsilon(p^\varepsilon)}{\partial \pi_{ij}} = \sum_{k\in\mathcal{I}} \alpha_k \frac{\partial g_k(p^\varepsilon)}{\partial \pi_{ij}} + \sum_{k\in\mathcal{J}} \beta_k \frac{\partial \tilde{g}_k(p^\varepsilon)}{\partial \pi_{ij}} \quad \text{for all } (i,j)\in\mathcal{K}, \tag{165}$$

then, according to Lemma 5and the concavity of $\widetilde{H}_\varepsilon$, we have

$$\begin{aligned}&\widetilde{H}_\varepsilon(p)\\ &\leq \widetilde{H}_\varepsilon(p^\varepsilon) + \left\langle \nabla\widetilde{H}_\varepsilon(p^\varepsilon), p-p^\varepsilon\right\rangle\\ &\leq \widetilde{H}_\varepsilon(p^\varepsilon) + \left\langle \sum_{k\in\mathcal{I}} \alpha_k \frac{\partial g_k(p^\varepsilon)}{\partial \pi_{ij}} + \sum_{k\in\mathcal{J}} \beta_k \frac{\partial \tilde{g}_k(p^\varepsilon)}{\partial \pi_{ij}}, p-p^\varepsilon\right\rangle\\ &\leq \widetilde{H}_\varepsilon(p^\varepsilon) + \sum_{k\in\mathcal{I}} \alpha_k \left\langle \nabla g_k\left(p^\varepsilon\right), p-p^\varepsilon\right\rangle\\ &+ \sum_{k\in\mathcal{J}} \beta_k \left\langle \nabla \tilde{g}_k\left(p^\varepsilon\right), p-p^\varepsilon\right\rangle\end{aligned} \tag{166}$$

We also observe that the functions g_i for $i\in\mathcal{I}$ and $\tilde{g}_j$ for $j\in\mathcal{J}$ are affine functions. As a result, using Lemma 6, and given that both p and p^ε are in $\widetilde{T}(y,\hat{y})$, we have

$$\begin{aligned}&\underbrace{g_k(p)}_{=0} =\\ &\underbrace{g_k(p^\varepsilon)}_{=0} - \left\langle \nabla g_k\left(p^\varepsilon\right), p-p^\varepsilon\right\rangle \Rightarrow \left\langle \nabla g_k\left(p^\varepsilon\right), p-p^\varepsilon\right\rangle = 0,\\ &\underbrace{\tilde{g}_k(p)}_{=0} =\\ &\underbrace{\tilde{g}_k(p^\varepsilon)}_{=0} - \left\langle \nabla \tilde{g}_k\left(p^\varepsilon\right), p-p^\varepsilon\right\rangle \Rightarrow \left\langle \nabla \tilde{g}_k\left(p^\varepsilon\right), p-p^\varepsilon\right\rangle = 0.\end{aligned} \tag{167}$$

Calculations (166) and (167) lead to the conclusion that $0 \leq \widetilde{H}_\varepsilon(p^\varepsilon) - \widetilde{H}_\varepsilon(p)$ for all p in $\widetilde{T}(y,\hat{y})$. In other words, p^ε is a solution of the optimization problem $\mathfrak{B}$. Moreover, since $\widetilde{H}_\varepsilon$ is strictly concave, p^ε is the unique solution. In the next two subsections, we will present a vector p^ε that meets these expected properties.

P. $Y \neq \hat{Y}$, P^ε IS IN $\widetilde{T}(Y,\hat{Y})$

We now define a vector belonging to $\widetilde{T}(y,\hat{y})$, which will be a candidate for satisfying equality (165).

Let p^ε be a vector of size $\#\mathcal{K}$ defined by (168), as shown at the bottom of the next page.

To prove that p^ε is in $\widetilde{T}(y,\hat{y})$, the vector p^ε must satisfy all the equalities $g_i(p^\varepsilon)=0$ and $\tilde{g}_j(p^\varepsilon)=0$. Additionally, these entries must be non-negative.

Firstly, let's show that for a well-chosen ε, its entries are non-negative. When $\varepsilon\to 0$, all entries of p^ε become positive: (169), as shown at the bottom of the next page.

Consequently, we set ε sufficiently small so that the entries of p^ε are non-negative.

Secondly, let's show that $\tilde{g}_j(p^\varepsilon)=0$ for all $j\in\mathcal{J}$. Let j be in $\mathcal{J}$. It follows, (170), as shown at the bottom of the next page. Additionally, by design, (171), as shown at the bottom of page 42.

Combining calculations (170) and (171), we have

$$\sum_{i=1:(i,j)\in\mathcal{K}}^{C} p^\varepsilon_{\iota(i,j)} = \frac{\hat{y}_j}{\|\hat{y}\|_1} - \min\left(\frac{y_j}{\|y\|_1}, \frac{\hat{y}_j}{\|\hat{y}\|_1}\right) \Rightarrow \tilde{g}_j(p^\varepsilon)=0. \tag{172}$$

Thus, $\tilde{g}_j(p^\varepsilon)=0$ for all $j\in\mathcal{J}$. Proceeding similarly, we show that $g_i(p^\varepsilon)=0$ for all $i\in\mathcal{I}$.

In conclusion, p^ε is in $\widetilde{T}(y,\hat{y})$.

Q. $Y \neq \hat{Y}$, P^ε MEETS GRADIENT EQUALITY

We will exhibit coefficients α_i and β_j such that (165) is satisfied.

The partial derivatives, for all $(i,j)\in\mathcal{K}$, are (173), as shown at the bottom of page 42.

For all $i\in\mathcal{I}$, for all $j\in\mathcal{J}$, we define (174), as shown at the bottom of page 42.

It is straightforward to see that (165) is met.

R. $Y \neq \hat{Y}$, SOLUTION

We will finally solve $\mathfrak{A}$.

By design of p^{ε}, for all p in $\widetilde{T}(y,\hat{y})$, we have: (175), as shown at the bottom of the next page.

$$
\begin{aligned}
\sum_{(i,j)\in\mathcal{K}}(\pi_{ij}+\varepsilon) &= \varepsilon\#\mathcal{K} + \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}\pi_{ij} \\
&= \varepsilon\#\mathcal{K} + \sum_{i\in\mathcal{I}}\Big(\sum_{j\in\mathcal{J}}\pi_{ij} + \underbrace{\sum_{j\notin\mathcal{J}:j\neq i}\pi_{ij}}_{=0,\text{ by assumption on }\mathcal{J}}\Big) \\
&= \varepsilon\#\mathcal{K} + \sum_{i\in\mathcal{I}}\sum_{j=1:j\neq i}^{C}\pi_{ij} \\
&= \varepsilon\#\mathcal{K} + \sum_{i\in\mathcal{I}}\frac{y_i}{\|y\|_1} - \min\Big(\frac{y_i}{\|y\|_1},\frac{\hat{y}_i}{\|\hat{y}\|_1}\Big) \\
&= \varepsilon\#\mathcal{K} + \underbrace{\sum_{i\in\mathcal{I}}\frac{y_i}{\|y\|_1} - \min\Big(\frac{y_i}{\|y\|_1},\frac{\hat{y}_i}{\|\hat{y}\|_1}\Big) + \overbrace{\sum_{i\notin\mathcal{I}}\frac{y_i}{\|y\|_1} - \min\Big(\frac{y_i}{\|y\|_1},\frac{\hat{y}_i}{\|\hat{y}\|_1}\Big)}^{=0,\text{ by assumption on }\mathcal{I}}}_{=\|\frac{y}{\|y\|_1}-\min(\frac{y}{\|y\|_1},\frac{\hat{y}}{\|\hat{y}\|_1})\|_1} \\
&= \varepsilon\#\mathcal{K} + \|\frac{y}{\|y\|_1} - \min\Big(\frac{y}{\|y\|_1},\frac{\hat{y}}{\|\hat{y}\|_1}\Big)\|_1.
\end{aligned}
\tag{162}
$$

$$
p^{\varepsilon}_{\iota(i,j)} = \frac{\Big(\frac{\hat{y}_j}{\|\hat{y}\|_1} - \min\Big(\frac{y_j}{\|y\|_1},\frac{\hat{y}_j}{\|\hat{y}\|_1}\Big) + \varepsilon\#\mathcal{K}\Big)\Big(\frac{y_i}{\|y\|_1} - \min\Big(\frac{y_i}{\|y\|_1},\frac{\hat{y}_i}{\|\hat{y}\|_1}\Big) + \varepsilon\#\mathcal{K}\Big)}{\varepsilon\#\mathcal{K}^2 + \|\frac{y}{\|y\|_1} - \min\Big(\frac{y}{\|y\|_1},\frac{\hat{y}}{\|\hat{y}\|_1}\Big)\|_1} - \varepsilon. \tag{168}
$$

$$
\begin{aligned}
\lim_{\varepsilon\to 0^+} p^{\varepsilon}_{\iota(i,j)} &= \lim_{\varepsilon\to 0^+}\frac{\Big(\frac{\hat{y}_j}{\|\hat{y}\|_1} - \min\Big(\frac{y_j}{\|y\|_1},\frac{\hat{y}_j}{\|\hat{y}\|_1}\Big) + \varepsilon\#\mathcal{K}\Big)\Big(\frac{y_i}{\|y\|_1} - \min\Big(\frac{y_i}{\|y\|_1},\frac{\hat{y}_i}{\|\hat{y}\|_1}\Big) + \varepsilon\#\mathcal{K}\Big)}{\varepsilon\#\mathcal{K}^2 + \|\frac{y}{\|y\|_1} - \min\Big(\frac{y}{\|y\|_1},\frac{\hat{y}}{\|\hat{y}\|_1}\Big)\|_1} - \varepsilon \\
&= \frac{\Big(\overbrace{\frac{\hat{y}_j}{\|\hat{y}\|_1} - \min\Big(\frac{y_j}{\|y\|_1},\frac{\hat{y}_j}{\|\hat{y}\|_1}\Big)}^{>0,\text{by assumption on }\mathcal{J}}\Big)\Big(\overbrace{\frac{y_i}{\|y\|_1} - \min\Big(\frac{y_i}{\|y\|_1},\frac{\hat{y}_i}{\|\hat{y}\|_1}\Big)}^{>0,\text{by assumption on }\mathcal{I}}\Big)}{\|\frac{y}{\|y\|_1} - \min\Big(\frac{y}{\|y\|_1},\frac{\hat{y}}{\|\hat{y}\|_1}\Big)\|_1} > 0.
\end{aligned}
\tag{169}
$$

$$
\begin{aligned}
&\sum_{i=1:(i,j)\in\mathcal{K}}^{C} p^{\varepsilon}_{\iota(i,j)} \\
&= \sum_{i=1:(i,j)\in\mathcal{K}}^{C}\frac{\Big(\frac{\hat{y}_j}{\|\hat{y}\|_1} - \min\Big(\frac{y_j}{\|y\|_1},\frac{\hat{y}_j}{\|\hat{y}\|_1}\Big) + \varepsilon\#\mathcal{K}\Big)\Big(\frac{y_i}{\|y\|_1} - \min\Big(\frac{y_i}{\|y\|_1},\frac{\hat{y}_i}{\|\hat{y}\|_1}\Big) + \varepsilon\#\mathcal{K}\Big)}{\varepsilon\#\mathcal{K}^2 + \|\frac{y}{\|y\|_1} - \min\Big(\frac{y}{\|y\|_1},\frac{\hat{y}}{\|\hat{y}\|_1}\Big)\|_1} - \varepsilon \\
&= -\varepsilon\#\mathcal{K} + \Big(\frac{\hat{y}_j}{\|\hat{y}\|_1} - \min\Big(\frac{y_j}{\|y\|_1},\frac{\hat{y}_j}{\|\hat{y}\|_1}\Big) + \varepsilon\#\mathcal{K}\Big)\sum_{i=1:(i,j)\in\mathcal{K}}^{C}\frac{\frac{y_i}{\|y\|_1} - \min\Big(\frac{y_i}{\|y\|_1},\frac{\hat{y}_i}{\|\hat{y}\|_1}\Big) + \varepsilon\#\mathcal{K}}{\varepsilon\#\mathcal{K}^2 + \|\frac{y}{\|y\|_1} - \min\Big(\frac{y}{\|y\|_1},\frac{\hat{y}}{\|\hat{y}\|_1}\Big)\|_1} \\
&= -\varepsilon\#\mathcal{K} + \Big(\frac{\hat{y}_j}{\|\hat{y}\|_1} - \min\Big(\frac{y_j}{\|y\|_1},\frac{\hat{y}_j}{\|\hat{y}\|_1}\Big) + \varepsilon\#\mathcal{K}\Big)\frac{\varepsilon\#\mathcal{K}^2 + \sum_{i=1:(i,j)\in\mathcal{K}}^{C}\frac{y_i}{\|y\|_1} - \min\Big(\frac{y_i}{\|y\|_1},\frac{\hat{y}_i}{\|\hat{y}\|_1}\Big)}{\varepsilon\#\mathcal{K}^2 + \|\frac{y}{\|y\|_1} - \min\Big(\frac{y}{\|y\|_1},\frac{\hat{y}}{\|\hat{y}\|_1}\Big)\|_1}.
\end{aligned}
\tag{170}
$$

The last implication holds because H is continuous on $]0, +\infty[^{\#\mathcal{K}}$. Moreover, (176), as shown at the bottom of the next page. It follows that the following inequalities holds for all $p \in \widetilde{T}(y, \hat{y})$:

$$H(p) \leq H(p^*) \Rightarrow \underbrace{H(p) + \gamma(y, \hat{y})}_{=H(\pi)} \leq \underbrace{H(p^*) + \gamma(y, \hat{y})}_{=H(\pi^*)}$$
$$\Rightarrow H(\pi) \leq H(\pi^*), \tag{177}$$

where π and π^* are defined as in (159). It is straightforward to verify that $\pi^* = \pi^*(y, \hat{y})$ and that $\pi^*(y, \hat{y})$ is in $T^{\text{opt}}(y, \hat{y})$ (as already proven in Subsection XI-H). This completes the proof.

APPENDIX I
PROOF OF PROPOSITION 4

By definition, the diagonal entry i is,

$$f(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})_{ii} = \min(\frac{y_i}{\|y\|_1}, \frac{\hat{y}_i}{\|\hat{y}\|_1}) \tag{178}$$

Considering off-diagonal entries ij, according to definition 1, the deficit in i, denoted as u'_i, is given by

$$u'_i = \begin{cases} \frac{y_i}{\|y\|_1} - \frac{\hat{y}_i}{\|\hat{y}\|_1} & \text{if } \frac{\hat{y}_i}{\|\hat{y}\|_1} < \frac{y_i}{\|y\|_1} \\ 0 & \text{otherwise} \end{cases} \tag{179}$$

Similarly, the excess quantity in j, denoted as v'_j, is given by

$$v'_j = \begin{cases} \frac{\hat{y}_j}{\|\hat{y}\|_1} - \frac{y_j}{\|y\|_1} & \text{if } \frac{y_j}{\|y\|_1} < \frac{\hat{y}_j}{\|\hat{y}\|_1} \\ 0 & \text{otherwise} \end{cases} \tag{180}$$

The following relationships are straightforward to verify:

$$u'_i = \frac{y_i}{\|y\|_1} - \min(\frac{y_i}{\|y\|_1}, \frac{\hat{y}_i}{\|\hat{y}\|_1}), \tag{181}$$
$$v'_j = \frac{\hat{y}_j}{\|\hat{y}\|_1} - \min(\frac{y_j}{\|y\|_1}, \frac{\hat{y}_j}{\|\hat{y}\|_1}), \tag{182}$$

and (183), as shown at the bottom of the next page.

$$\begin{aligned} \sum_{i=1:(i,j)\in\mathcal{K}}^{C} \frac{y_i}{\|y\|_1} - \min(\frac{y_i}{\|y\|_1}, \frac{\hat{y}_i}{\|\hat{y}\|_1}) &= \sum_{i\in\mathcal{I}}^{C} \frac{y_i}{\|y\|_1} - \min(\frac{y_i}{\|y\|_1}, \frac{\hat{y}_i}{\|\hat{y}\|_1}) \\ &= \sum_{i\in\mathcal{I}}^{C} \frac{y_i}{\|y\|_1} - \min(\frac{y_i}{\|y\|_1}, \frac{\hat{y}_i}{\|\hat{y}\|_1}) + \underbrace{\sum_{i\notin\mathcal{I}}^{C} \frac{y_i}{\|y\|_1} - \min(\frac{y_i}{\|y\|_1}, \frac{\hat{y}_i}{\|\hat{y}\|_1})}_{=0, \text{ by assumption on } \mathcal{I}} \\ &= \|\frac{y}{\|y\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})\|_1. \end{aligned} \tag{171}$$

$$\frac{\partial \widetilde{H}_\varepsilon(p^\varepsilon)}{\partial \pi_{ij}} = -\ln(p^\varepsilon_{\iota(i,j)} + \varepsilon) = -\ln \frac{\left(\frac{\hat{y}_j}{\|\hat{y}\|_1} - \min(\frac{y_j}{\|y\|_1}, \frac{\hat{y}_j}{\|\hat{y}\|_1}) + \varepsilon\#\mathcal{K}\right)\left(\frac{y_i}{\|y\|_1} - \min(\frac{y_i}{\|y\|_1}, \frac{\hat{y}_i}{\|\hat{y}\|_1}) + \varepsilon\#\mathcal{K}\right)}{\varepsilon\#\mathcal{K}^2 + \|\frac{y}{\|y\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})\|_1},$$
$$\text{and} \quad \sum_{k\in\mathcal{I}} \alpha_k \frac{\partial g_k(p^\varepsilon)}{\partial \pi_{ij}} + \sum_{k\in\mathcal{J}} \beta_k \frac{\partial \tilde{g}_k(p^\varepsilon)}{\partial \pi_{ij}} = \alpha_i + \beta_j \tag{173}$$

$$\alpha_i = -\ln \frac{\frac{y_i}{\|y\|_1} - \min(\frac{y_i}{\|y\|_1}, \frac{\hat{y}_i}{\|\hat{y}\|_1}) + \varepsilon\#\mathcal{K}}{\sqrt{\varepsilon\#\mathcal{K}^2 + \|\frac{y}{\|y\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})\|_1}}, \quad \text{and} \quad \beta_j = -\ln \frac{\frac{\hat{y}_j}{\|\hat{y}\|_1} - \min(\frac{y_j}{\|y\|_1}, \frac{\hat{y}_j}{\|\hat{y}\|_1}) + \varepsilon\#\mathcal{K}}{\sqrt{\varepsilon\#\mathcal{K}^2 + \|\frac{y}{\|y\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})\|_1}}. \tag{174}$$

$$\begin{aligned} & \widetilde{H}_\varepsilon(p) \leq \widetilde{H}_\varepsilon(p^\varepsilon) \\ \Rightarrow \quad & H(p+\varepsilon) + \underbrace{\varepsilon\#\mathcal{K} + \|\frac{y}{\|y\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})\|_1}_{\text{constant relative to } p, \text{ see (161) and (162)}} \leq H(p^\varepsilon + \varepsilon) + \underbrace{\varepsilon\#\mathcal{K} + \|\frac{y}{\|y\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})\|_1}_{\text{constant relative to } p, \text{ see (161) and (162)}} \\ \Rightarrow \quad & H(p+\varepsilon) \leq H(p^\varepsilon + \varepsilon) \\ \Rightarrow \quad & \lim_{\varepsilon\to 0^+} H(p+\varepsilon) \leq \lim_{\varepsilon\to 0^+} H(p^\varepsilon + \varepsilon) \\ \Rightarrow \quad & H(\lim_{\varepsilon\to 0^+} p + \varepsilon) \leq H(\lim_{\varepsilon\to 0^+} p^\varepsilon + \varepsilon) \end{aligned} \tag{175}$$

Consequently, (184), as shown at the bottom of the next page.

Considering the formula in proposition 3 entry by entry completes the proof.

APPENDIX J
PROOF OF PROPOSITION 5

Let u and v be two vectors in $\mathbb{R}^C_{\geq 0}$ such that $\|u\|_1 = \|v\|_1$, and let α be a positive real value.

By definition, the diagonal entry i is given by:

$$f(\alpha u, \alpha v)_{ii} = \min(\alpha u, \alpha v) = \alpha \min(u, v) = \alpha f(u, v)_{ii}. \tag{185}$$

Considering off-diagonal entries ij, we denote:

$$u'_i = \begin{cases} \frac{y_i}{\|y\|_1} - \frac{\hat{y}_i}{\|\hat{y}\|_1} & \text{if } \frac{\hat{y}_i}{\|\hat{y}\|_1} < \frac{y_i}{\|y\|_1} \\ 0 & \text{otherwise} \end{cases} \tag{186}$$

$$v'_j = \begin{cases} \frac{\hat{y}_j}{\|\hat{y}\|_1} - \frac{y_j}{\|y\|_1} & \text{if } \frac{y_j}{\|y\|_1} < \frac{\hat{y}_j}{\|\hat{y}\|_1} \\ 0 & \text{otherwise} \end{cases} \tag{187}$$

It is straightforward to verify that the deficit quantity in i according to the definition 1 is $\alpha u'_i$, and the excess quantity in j is $\alpha v'_j$. It follows that:

$$f(\alpha u, \alpha v)_{ij} = \frac{\alpha u'_i \alpha v'_j}{\sum_{k=1}^C \alpha u'_k} = \alpha \frac{u'_i v'_j}{\sum_{k=1}^C u'_k} = \alpha f(u, v)_{ij}, \tag{188}$$

which completes the proof.

APPENDIX K
PROOF OF PROPOSITION 6

Let $(y, \hat{y})$ be an instance.

If no common quantity exist between y and $\hat{y}$, then $\min\left(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\right) = 0$ for all k between 1 and C. Conversely, if $\min\left(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\right) = 0$ for all k between 1 and C, then there is no overlap between y and $\hat{y}$. In conclusion, $\pi^*(y, \hat{y})$ has a zero diagonal if, and only if, there is no overlap between y and $\hat{y}$.

If the class i is underestimated, i.e., $\frac{\hat{y}_i}{\|\hat{y}\|_1} < \frac{y_i}{\|y\|_1}$, and class j is overestimated, i.e., $\frac{y_j}{\|y\|_1} < \frac{\hat{y}_j}{\|\hat{y}\|_1}$, then

$$0 < \frac{y_i}{\|y\|_1} - \min\left(\frac{y_i}{\|y\|_1} \frac{\hat{y}_i}{\|\hat{y}\|_1}\right) \quad \text{and} \quad 0 < \frac{\hat{y}_j}{\|\hat{y}\|_1} - \min\left(\frac{y_j}{\|y\|_1}, \frac{\hat{y}_j}{\|\hat{y}\|_1}\right), \tag{189}$$

resulting in $0 < \pi^*(y, \hat{y})_{ij}$ according to (8).

Conversely, if $0 < \pi^*(y, \hat{y})_{ij}$, then

$$0 < \left(\frac{y_i}{\|y\|_1} - \min\left(\frac{y_i}{\|y\|_1}, \frac{\hat{y}_i}{\|\hat{y}\|_1}\right)\right) \left(\frac{\hat{y}_j}{\|\hat{y}\|_1} - \min\left(\frac{y_j}{\|y\|_1}, \frac{\hat{y}_j}{\|\hat{y}\|_1}\right)\right), \tag{190}$$

which implies that both terms are non-zero. Consequently, this means that i is underestimated, whereas j is overestimated. In conclusion, $0 < \pi^*(y, \hat{y})_{ij}$ if, and only if, i is underestimated, whereas j is overestimated.

APPENDIX L
PROOF OF PROPOSITION 7

Let y and $\hat{y}$ be two binary vectors such that $y_i = 1$ with all other entries as zero, and $\hat{y}_j = 1$ with all other entries as zero, representing a given instance.

The CM contribution is a matrix of size C, such as the entry ij equals 1, and all other entries are zero. Consequently, its contribution equals $y \otimes \hat{y}$.

The TCM contribution is given by:

$$\pi^*(y, \hat{y}) = \operatorname{diag}\left(\min\left(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}\right)\right) + \frac{\left(\frac{y}{\|y\|_1} - \min\left(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}\right)\right) \otimes \left(\frac{\hat{y}}{\|\hat{y}\|_1} - \min\left(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}\right)\right)}{\left\| \frac{y}{\|y\|_1} - \min\left(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}\right)\right\|_1}. \tag{191}$$

Moreover, in the single-label context, the following equality holds:

$$\min\left(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}\right) = y * \hat{y}, \tag{192}$$

$$\lim_{\varepsilon \to 0^+} p + \varepsilon = p$$

$$\lim_{\varepsilon \to 0^+} (p^\varepsilon + \varepsilon)_{\iota(i,j)} = \frac{\left(\frac{\hat{y}_j}{\|\hat{y}\|_1} - \min(\frac{y_j}{\|y\|_1}, \frac{\hat{y}_j}{\|\hat{y}\|_1})\right)\left(\frac{y_i}{\|y\|_1} - \min(\frac{y_i}{\|y\|_1}, \frac{\hat{y}_i}{\|\hat{y}\|_1})\right)}{\| \frac{y}{\|y\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})\|_1} := p^*_{\iota(i,j)} \tag{176}$$

$$\sum_{k=1}^C u'_k = \underbrace{\| \frac{y}{\|y\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})\|_1 = \| \frac{\hat{y}}{\|\hat{y}\|_1} - \min(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1})\|_1}_{\text{according to Lemma 1}}. \tag{183}$$

where $*$ denotes the Hadamard product (i.e., element-wise product).

If $i = j$, then $y = \hat{y}$ and the second term of the formula is considered to be 0 by continuous extension. It follows that,

$$\pi^*(y,\hat{y})_{ii} = \operatorname{diag}\left(\min\left(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}\right)\right)_{ii} = \min\left(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}\right)_i = (y * \hat{y})_i = (y \otimes \hat{y})_{ii}. \tag{193}$$

If $i \neq j$, then $y_k \hat{y}_k = 0$ for all k, leading to $y * \hat{y}$ being the null vector. As a result, we have:

$$\pi^*(y,\hat{y}) = \operatorname{diag}\left(\min\left(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}\right)\right) + \frac{\left(\frac{y}{\|y\|_1} - \min\left(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}\right)\right) \otimes \left(\frac{\hat{y}}{\|\hat{y}\|_1} - \min\left(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}\right)\right)}{\left\|\frac{y}{\|y\|_1} - \min\left(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}\right)\right\|_1} = \frac{\frac{y}{\|y\|_1} \otimes \frac{\hat{y}}{\|\hat{y}\|_1}}{\|\frac{y}{\|y\|_1}\|_1} = y \otimes \hat{y}, \tag{194}$$

because $\|y\|_1 = \|\hat{y}\|_1 = 1$.

We have shown that in the single-label context, $\pi^*(y,\hat{y}) = y \otimes \hat{y}$, which is the expression of the CM contribution. Since both the CM and TCM produce the same contribution, they are equal (whatever the weighting, because in the single-label context, $\lambda(y,\hat{y})$ is always equal to 1).

APPENDIX M
PROOF OF PROPOSITION 8

Let y and $\hat{y}$ be an instance. Let Y and $\hat{Y}$ be the sets of present classes in y and $\hat{y}$, respectively. The four formulas presented in [2] are presented in set notation; we translate them in vector form:

(i) In the case where $Y = \hat{Y}$, contribution formula is:

$$\operatorname{diag}(y) \tag{195}$$

(ii) In the case where $Y \subsetneq \hat{Y}$, contribution formula is:

$$\frac{y \otimes \left(\hat{y} - \min(y,\hat{y})\right)}{\|\hat{y}\|_1} + \operatorname{diag}(y)\frac{\|y\|_1}{\|\hat{y}\|_1} \tag{196}$$

(iii) In the case where $\hat{Y} \subsetneq Y$, contribution formula is:

$$\frac{\left(y - \min(y,\hat{y})\right) \otimes \hat{y}}{\|\hat{y}\|_1} + \operatorname{diag}(\hat{y}) \tag{197}$$

(iv) In none of the previous cases, the contribution formula is:

$$\frac{\left(y - \min(y,\hat{y})\right) \otimes \left(\hat{y} - \min(y,\hat{y})\right)}{\|\hat{y} - \min(y,\hat{y})\|_1} + \operatorname{diag}\left(\min(y,\hat{y})\right) \tag{198}$$

Combining Proposition 4 and 5, it holds:

$$\|y\|_1 \pi^*(y,\hat{y}) = \|y\|_1 f\left(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}\right) = f\left(y, \hat{y}\frac{\|y\|_1}{\|\hat{y}\|_1}\right). \tag{199}$$

Considering Definition 1, we set $u = y$ and $v = \hat{y}\frac{\|y\|_1}{\|\hat{y}\|_1}$. The following relationships are straightforward to verify:

$$u'_i = y_i - \min\left(y_i, \hat{y}_i\frac{\|y\|_1}{\|\hat{y}\|_1}\right), \quad v'_j = \frac{\hat{y}_j}{\|\hat{y}\|_1} - \min\left(y_j, \hat{y}_j\frac{\|y\|_1}{\|\hat{y}\|_1}\right), \tag{200}$$

and

$$\sum_{k=1}^{C} u'_k = \|y - \min\left(y, \hat{y}\frac{\|y\|_1}{\|\hat{y}\|_1}\right)\|_1. \tag{201}$$

Consequently, contributions weighted by $\|y\|_1$ are:

$$\|y\|_1 \pi^*(y,\hat{y}) = \operatorname{diag}\left(\min\left(y, \hat{y}\frac{\|y\|_1}{\|\hat{y}\|_1}\right)\right) + \frac{\left(y - \min\left(y, \hat{y}\frac{\|y\|_1}{\|\hat{y}\|_1}\right)\right) \otimes \left(\hat{y}\frac{\|y\|_1}{\|\hat{y}\|_1} - \min\left(y, \hat{y}\frac{\|y\|_1}{\|\hat{y}\|_1}\right)\right)}{\left\|y - \min\left(y, \frac{\hat{y}}{\|\hat{y}\|_1}\right)\right\|_1}. \tag{202}$$

Moreover, when $y = \hat{y}$, the second term is zero by continuous extension.

In case (i), $y = \hat{y}$, it follows that:

$$\|y\|_1 \pi^*(y,\hat{y}) = \operatorname{diag}\left(\min\left(y, \hat{y}\frac{\|y\|_1}{\|\hat{y}\|_1}\right)\right) = \operatorname{diag}(y), \tag{203}$$

leading to the same formula as MLCM.

In case (ii), $\|y\|_1 < \|\hat{y}\|_1$ leading to $\|y\|_1/\|\hat{y}\|_1 < 1$. Moreover, as $y_k = 1$ implies $\hat{y}_k = 1$, we have $\min\left(y, \hat{y}\frac{\|y\|_1}{\|\hat{y}\|_1}\right) = y\frac{\|y\|_1}{\|\hat{y}\|_1}$. It follows that:

$$\begin{aligned}\|y\|_1 \pi^*(y,\hat{y}) &= \operatorname{diag}\left(y\frac{\|y\|_1}{\|\hat{y}\|_1}\right) + \frac{\left(y - y\frac{\|y\|_1}{\|\hat{y}\|_1}\right) \otimes \left(\hat{y}\frac{\|y\|_1}{\|\hat{y}\|_1} - y\frac{\|y\|_1}{\|\hat{y}\|_1}\right)}{\left\|y - y\frac{\|y\|_1}{\|\hat{y}\|_1}\right\|_1} \\ &= \operatorname{diag}(y)\frac{\|y\|_1}{\|\hat{y}\|_1} + \frac{\left(y(1 - \frac{\|y\|_1}{\|\hat{y}\|_1})\right) \otimes \left((\hat{y} - y)\frac{\|y\|_1}{\|\hat{y}\|_1}\right)}{\left\|y(1 - \frac{\|y\|_1}{\|\hat{y}\|_1})\right\|_1}\end{aligned}$$

$$f\left(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}\right)_{ij} = \frac{u'_i v'_j}{\sum_{k=1}^{C} u'_k} = \frac{\left(\frac{y_i}{\|y\|_1} - \min\left(\frac{y_i}{\|y\|_1}, \frac{\hat{y}_i}{\|\hat{y}\|_1}\right)\right)\left(\frac{\hat{y}_j}{\|\hat{y}\|_1} - \min\left(\frac{y_j}{\|y\|_1}, \frac{\hat{y}_j}{\|\hat{y}\|_1}\right)\right)}{\|\frac{y}{\|y\|_1} - \min\left(\frac{y}{\|y\|_1}, \frac{\hat{y}}{\|\hat{y}\|_1}\right)\|_1}. \tag{184}$$

$$= \text{diag}(y)\frac{\|y\|_1}{\|\hat{y}\|_1} + \frac{y \otimes (\hat{y} - y)}{\|\hat{y}\|_1}, \tag{204}$$

leading to the same formula as MLCM. The equality in case (iii) can be demonstrated similarly.

In the case (iv), in TCM contribution, $\hat{y}$ is weighted by $\frac{\|y\|_1}{\|\hat{y}\|_1}$ whereas it is not the case in MLCM contribution, leading to different formula.

APPENDIX N
PROOF OF PROPOSITION 9

The framework of paper [27] is the soft-label framework restricted to probability distributions. Consequently, let y and $\hat{y}$ be an instance such as $\|y\|_1 = \|\hat{y}\|_1 = 1$. Silván-Cárdenas and Wang [27] introduced the MIN-LEAST and MIN-MIN operators to produce their Sub-pixel confusion matrix. The

$$I(i,j) := \left[\max\left(0, \hat{y}_j - \min\left(y_j, \hat{y}_j\right) - \sum_{k=1:k\neq i}^{C} y_k - \min\left(y_k, \hat{y}_k\right)\right), \quad \min\left(y_i - \min\left(y_i, \hat{y}_i\right), \hat{y}_j - \min\left(y_j, \hat{y}_j\right)\right)\right] \tag{205}$$

$$\begin{aligned}
\sum_{k=1:k\neq j}^{C} \pi_{kj} &= \pi_{ij} + \sum_{k=1:k\neq i,j}^{C} \underbrace{\pi_{kj}}_{\leq \min\left(y_k - \min\left(y_k,\hat{y}_k\right), \hat{y}_j - \min\left(y_j,\hat{y}_j\right)\right)} \\
&\leq \pi_{ij} + \sum_{k=1:k\neq i,j}^{C} \underbrace{\min\left(y_k - \min\left(y_k,\hat{y}_k\right), \hat{y}_j - \min\left(y_j,\hat{y}_j\right)\right)}_{\leq y_k - \min\left(y_k,\hat{y}_k\right)} \\
&\leq \underbrace{\pi_{ij}}_{< \hat{y}_j - \min\left(y_j,\hat{y}_j\right) - \sum_{k=1:k\neq i}^{C} y_k - \min\left(y_k,\hat{y}_k\right), \text{ by contradiction assumption}} + \sum_{k=1:k\neq i,j}^{C} y_k - \min\left(y_k, \hat{y}_k\right) \\
&< \hat{y}_j - \min\left(y_j, \hat{y}_j\right) \underbrace{- \sum_{k=1:k\neq i}^{C} y_k - \min\left(y_k, \hat{y}_k\right) + \sum_{k=1:k\neq i,j}^{C} y_k - \min\left(y_k, \hat{y}_k\right)}_{=0} \\
&= \hat{y}_j - \min\left(y_j, \hat{y}_j\right),
\end{aligned} \tag{208}$$

$$\begin{aligned}
M^a_{ij} &= \widetilde{M}^a_{\sigma(i)\sigma(j)} \\
&= \widetilde{M}^a_{1C} \\
&= \left[f\left(\frac{s}{\|s\|_1} - m, \frac{\hat{s}}{\|\hat{s}\|_1} - E^C\left(\frac{\hat{s}_C}{\|\hat{s}\|_1} - \min\left(\frac{s_C}{\|s\|_1}, \frac{\hat{s}_C}{\|\hat{s}\|_1}\right)\right)\right) + m \otimes E^C\right]_{1C} \\
&= \underbrace{\min\left(\frac{s_1}{\|s\|_1} - \min\left(\frac{s_1}{\|s\|_1}, \frac{\hat{s}_1}{\|\hat{s}\|_1}\right), \frac{\hat{s}_C}{\|\hat{s}\|_1} - \min\left(\frac{s_C}{\|s\|_1}, \frac{\hat{s}_C}{\|\hat{s}\|_1}\right)\right)}_{\text{using Lemma 4}} \\
&= \min\left(\frac{y_i}{\|y\|_1} - \min\left(\frac{y_i}{\|y\|_1}, \frac{\hat{y}_i}{\|\hat{y}\|_1}\right), \frac{\hat{y}_j}{\|\hat{y}\|_1} - \min\left(\frac{y_j}{\|y\|_1}, \frac{\hat{y}_j}{\|\hat{y}\|_1}\right)\right)
\end{aligned} \tag{211}$$

$$\begin{aligned}
M^b_{ij} &= \widetilde{M}^b_{\tau(i)\tau(j)} \\
&= \widetilde{M}^b_{C-1\ C} \\
&= \left[f\left(\frac{s}{\|s\|_1} - m, \frac{\hat{s}}{\|\hat{s}\|_1} - E^C\left(\frac{\hat{s}_C}{\|\hat{s}\|_1} - \min\left(\frac{s_C}{\|s\|_1}, \frac{\hat{s}_C}{\|\hat{s}\|_1}\right)\right)\right) + m \otimes E^C\right]_{C-1\ C} \\
&= \underbrace{\max\left(0, \frac{\hat{t}_C}{\|t\|_1} - \min\left(\frac{t_C}{\|t\|_1}, \frac{\hat{t}_C}{\|\hat{t}\|_1}\right) - \sum_{k=1:k\neq C-1}^{C} \frac{t_k}{\|t\|_1} - \min\left(\frac{t_k}{\|t\|_1}, \frac{\hat{t}_k}{\|\hat{t}\|_1}\right)\right)}_{\text{using Lemma 4}} \\
&= \max\left(0, \frac{\hat{y}_j}{\|\hat{y}\|_1} - \min\left(\frac{y_j}{\|y\|_1}, \frac{\hat{y}_j}{\|\hat{y}\|_1}\right) - \sum_{k=1:k\neq i}^{C} \frac{y_k}{\|y\|_1} - \min\left(\frac{y_k}{\|y\|_1}, \frac{\hat{y}_k}{\|\hat{y}\|_1}\right)\right)
\end{aligned} \tag{212}$$

diagonal of SCM and TCM are the same. The off-diagonal ij entries are intervals equal to: (205), as shown at the bottom of the previous page.

We begin to show that for all $\pi \in T^{\text{opt}}(y, \hat{y})$, for all i and j between 1 and C with $i \neq j$, $\pi_{ij} \in I(i, j)$. Once this is demonstrated, we will show that, for all i and j between 1 and C with $i \neq j$, it exists $\underline{\pi}$ and $\overline{\pi}$ in $T^{\text{opt}}(y, \hat{y})$ such as

$$\underline{\pi}_{ij} = \max\left(0, \hat{y}_j - \min\left(y_j, \hat{y}_j\right) - \sum_{k=1:k\neq i}^{C} y_k - \min\left(y_k, \hat{y}_k\right)\right)$$
$$\overline{\pi}_{ij} = \min\left(y_i - \min\left(y_i, \hat{y}_i\right), \hat{y}_j - \min\left(y_j, \hat{y}_j\right)\right), \tag{206}$$

thus ending the proof.

S. INSIDE THE INTERVAL

Let π be in $T^{\text{opt}}(y, \hat{y})$, let i and j be two different integers between 1 and C.

We begin by showing that π_{ij} is inferior or equal to the upper bound of $I(i, j)$. Since $\pi \in T^{\text{opt}}(y, \hat{y})$, all entries of π are positive, and the marginal sum property is met. Moreover, because $\|y\|_1 = 1$, it follows that:

$$\pi_{ij} \leq \sum_{k=1:k\neq i}^{C} \pi_{ik} = y_i - \min\left(y_i, \hat{y}_i\right),$$
$$\text{and} \quad \pi_{ij} \leq \sum_{k=1:k\neq j}^{C} \pi_{kj} = \hat{y}_j - \min\left(y_j, \hat{y}_j\right). \tag{207}$$

Consequently, $\pi_{ij} \leq \min\left(y_i - \min\left(y_i, \hat{y}_i\right), \hat{y}_j - \min\left(y_j, \hat{y}_j\right)\right)$. In conclusion, π_{ij} is inferior or equal to the upper bound of $I(i, j)$.

We will show that π_{ij} is superior or equal to the lower bound of $I(i, j)$. We reasoning by contradiction. Assuming that π is such as $\pi_{ij} < \max\left(0, \hat{y}_j - \min\left(y_j, \hat{y}_j\right) - \sum_{k=1:k\neq i}^{C} y_k - \min\left(y_k, \hat{y}_k\right)\right)$, it follows (208), as shown at the bottom of the previous page.

leading to a contradiction, because $\sum_{k=1:k\neq j}^{C} \pi_{kj} = \hat{y}_j - \min\left(y_j, \hat{y}_j\right)$. In conclusion, for all $\pi \in T^{\text{opt}}(y, \hat{y})$, for all i and j between 1 and C with $i \neq j$, $\pi_{ij} \in I(i, j)$.

T. EXISTENCE OF ELEMENTS REACHING THE BOUNDS

Let i and j be two different integers between 1 and C. Let σ and τ be two permutations on the integer set from 1 to C. The permutation σ is defined by $\sigma = (i\ 1)(j\ C)$, while the permutation τ is defined by $\tau = (i\ C-1)(j\ C)$. Let s, $\hat{s}$, t, and $\hat{t}$ be vectors of $\mathbb{R}^C_{\geq 0}$ such that $s_{\sigma(k)} = y_k$, $\hat{s}_{\sigma(k)} = \hat{y}_k$, and $t_{\tau(k)} = y_k$, $\hat{t}_{\tau(k)} = \hat{y}_k$ for all k from 1 to C (s stands for sigma, while t stands for tau).

According to Lemma 4, the matrices

$$\widetilde{M}^a := f\left(\frac{s}{\|s\|_1} - m, \frac{\hat{s}}{\|\hat{s}\|_1} - E^C\left(\frac{\hat{s}_C}{\|\hat{s}\|_1} - \min\left(\frac{s_C}{\|s\|_1}, \frac{\hat{s}_C}{\|\hat{s}\|_1}\right)\right)\right)$$
$$+ m \otimes E^C, \quad \text{and}$$
$$\widetilde{M}^b := f\left(\frac{t}{\|t\|_1} - m, \frac{\hat{t}}{\|\hat{t}\|_1} - E^C\left(\frac{\hat{t}_C}{\|\hat{t}\|_1} - \min\left(\frac{t_C}{\|t\|_1}, \frac{\hat{t}_C}{\|\hat{t}\|_1}\right)\right)\right)$$
$$+ m \otimes E^C, \tag{209}$$

are in $T^{\text{opt}}(s, \hat{s})$, and $T^{\text{opt}}(t, \hat{t})$ respectively. According to Lemma 2, the matrices M^a, and M^b, defined as,

$$M^a_{ij} = \widetilde{M}^a_{\sigma(i)\sigma(j)}, \quad \text{and} \quad M^b_{ij} = \widetilde{M}^b_{\tau(i)\tau(j)}, \quad \text{for} \quad i, j = 1 \ldots C, \tag{210}$$

are both in $T^{\text{opt}}(y, \hat{y})$.

Let's show that M^a_{ij} achieved the upper bound of $I(i, j)$: (211), as shown at the bottom of the previous page.

In conclusion, M^a_{ij} achieved the upper bound of $I(i, j)$.

Let's show that M^b_{ij} achieved the lower bound of $I(i, j)$: (212), as shown at the bottom of the previous page.

In conclusion, M^b_{ij} achieved the lower bound of $I(i, j)$.

Since M^a and M^b are in $T^{\text{opt}}(y, \hat{y})$, this achieved the proof.

APPENDIX O
PROOF OF PROPOSITION 10

Considering the diagonal property, if $\frac{y}{\|y\|_1} = \frac{\hat{y}}{\|\hat{y}\|_1}$, then $\pi^*(y, \hat{y})$ is a diagonal matrix by design. Conversely, if $\frac{y}{\|y\|_1} \neq \frac{\hat{y}}{\|\hat{y}\|_1}$, then there exist classes i and j, with $i \neq j$, such that

$$\frac{y_i}{\|y\|_1} - \min\left(\frac{y_i}{\|y\|_1}, \frac{\hat{y}_i}{\|\hat{y}\|_1}\right) > 0,$$
$$\text{and} \frac{\hat{y}_j}{\|\hat{y}\|_1} - \min\left(\frac{y_j}{\|y\|_1}, \frac{\hat{y}_j}{\|\hat{y}\|_1}\right) > 0. \tag{213}$$

According to Proposition 6, $\pi^*(y, \hat{y}) \in T(y, \hat{y})$ satisfies $\pi^*(y, \hat{y})_{ij} > 0$. In conclusion, $\pi^*(y, \hat{y}) \in T(y, \hat{y})$ meets the diagonal property.

Since $\pi^*(y, \hat{y}) \in T(y, \hat{y})$, the following equalities hold:

$$\sum_{j=1}^{C} \pi^*(y, \hat{y})_{ij} = \frac{y_i}{\|y\|_1}, \quad \sum_{i=1}^{C} \pi^*(y, \hat{y})_{ij} = \frac{\hat{y}_j}{\|\hat{y}\|_1}, \tag{214}$$

proving that $\pi^*(y, \hat{y}) \in T(y, \hat{y})$ meets marginal sum properties.

REFERENCES

[1] D. Krstinić, M. Braović, L. Šerić, and D. Božić-Štulić, "Multi-label classifier performance evaluation with confusion matrix," *Comput. Sci. Inf. Technol.*, vol. 1, pp. 1–14, Jun. 2020.

[2] D. Krstinić, A. K. Skelin, I. Slapnič ar, and M. Braović, "Multi-label confusion tensor," *IEEE Access*, vol. 12, pp. 9860–9870, 2024.

[3] J. Zhang, Y. Zheng, and Y. Shi, "A soft label method for medical image segmentation with multirater annotations," *Comput. Intell. Neurosci.*, vol. 2023, no. 1, Jan. 2023, Art. no. 1883597.

[4] C.-W. Wang, K.-Y. Lin, Y.-J. Lin, M.-A. Khalil, K.-L. Chu, and T.-K. Chao, "A soft label deep learning to assist breast cancer target therapy and thyroid cancer diagnosis," *Cancers*, vol. 14, no. 21, p. 5312, Oct. 2022.

[5] G. Li, R. Togo, T. Ogawa, and M. Haseyama, "Compressed gastric image generation based on soft-label dataset distillation for medical data sharing," *Comput. Methods Programs Biomed.*, vol. 227, Dec. 2022, Art. no. 107189.

[6] V. Yogarajan, J. Montiel, T. Smith, and B. Pfahringer, "Transformers for multi-label classification of medical text: An empirical comparison," in *Proc. Int. Conf. Artif. Intell. Med.* Cham, Switzerland: Springer, 2021, pp. 114–123.

[7] R. Kawamura, H. Hayashi, N. Takemura, and H. Nagahara, "MIDAS: Mixing ambiguous data with soft labels for dynamic facial expression recognition," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 6538–6548.

[8] T. Lukov, N. Zhao, G. H. Lee, and S.-N. Lim, "Teaching with soft label smoothing for mitigating noisy labels in facial expressions," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 648–665.
[9] Y. Zhou, H. Xue, and X. Geng, "Emotion distribution recognition from facial expressions," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 1247–1250.
[10] K. M. Collins, U. Bhatt, and A. Weller, "Eliciting and learning with soft labels from every annotator," in *Proc. AAAI Conf. Hum. Comput. Crowdsourcing*, Oct. 2022, vol. 10, no. 1, pp. 40–52. [Online]. Available: https://ojs.aaai.org/index.php/HCOMP/article/view/21986
[11] J. Lanchantin, T. Wang, V. Ordonez, and Y. Qi, "General multi-label image classification with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16473–16483.
[12] S. Coulibaly, B. Kamsu-Foguem, D. Kamissoko, and D. Traore, "Deep convolution neural network sharing for the multi-label images classification," *Mach. Learn. Appl.*, vol. 10, Dec. 2022, Art. no. 100422.
[13] H. Liu, G. Chen, P. Li, P. Zhao, and X. Wu, "Multi-label text classification via joint learning from label embedding and label correlation," *Neurocomputing*, vol. 460, pp. 385–398, Oct. 2021.
[14] X. Su, R. Wang, and X. Dai, "Contrastive learning-enhanced nearest neighbor mechanism for multi-label text classification," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 672–679.
[15] E. Leonardelli, A. Uma, G. Abercrombie, D. Almanea, V. Basile, T. Fornaciari, B. Plank, V. Rieser, and M. Poesio, "SemEval-2023 task 11: Learning with disagreements (LeWiDi)," 2023, *arXiv:2304.14803*.
[16] V. Basile, M. Fell, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, and A. Uma, "We need to consider disagreement in evaluation," in *Proc. 1st Workshop Benchmarking, Past, Present Future*, 2021, pp. 15–21.
[17] T. Fornaciari, A. Uma, S. Paun, B. Plank, D. Hovy, and M. Poesio, "Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2021, pp. 1–7.
[18] L. Plaza, J. Carrillo-de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, and P. Rosso, "Overview of exist 2023: Sexism identification in social networks," in *Proc. Eur. Conf. Inf. Retr. (ECIR)*, Dublin, Ireland. Cham, Switzerland: Springer, 2023, pp. 593–599.
[19] A. N. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, and M. Poesio, "Learning from disagreement: A survey," *J. Artif. Intell. Res.*, vol. 72, pp. 1385–1470, Dec. 2021.
[20] S.-M. Jung, "Measure theory," in *Ulam's Conjecture on Invariance of Measure in the Hilbert Cube*. Cham, Switzerland: Springer, 2023, pp. 55–90.
[21] C. Villani, *Topics in Optimal Transportation*, vol. 58. American Mathematical Society, 2021.
[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.
[23] M. Heydarian, T. E. Doyle, and R. Samavi, "MLCM: Multi-label confusion matrix," *IEEE Access*, vol. 10, pp. 19083–19095, 2022.
[24] D. Krstinić, L. Šerić, and I. Slapničar, "Comments on 'MLCM: Multi-label confusion matrix,'" *IEEE Access*, vol. 11, pp. 40692–40697, 2023.
[25] J. Görtler, F. Hohman, D. Moritz, K. Wongsuphasawat, D. Ren, R. Nair, M. Kirchner, and K. Patel, "NEO: Generalizing confusion matrix visualization to hierarchical and multi-output labels," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2022, doi: 10.1145/3491102.3501823.
[26] E. Binaghi, P. A. Brivio, P. Ghezzi, and A. Rampini, "A fuzzy set-based accuracy assessment of soft classification," *Pattern Recognit. Lett.*, vol. 20, no. 9, pp. 935–948, Sep. 1999.
[27] J. L. Silván-Cárdenas and L. Wang, "Sub-pixel confusion–uncertainty matrix for assessing soft classifications," *Remote Sens. Environ.*, vol. 112, no. 3, pp. 1081–1095, Mar. 2008.
[28] R. G. Pontius and M. L. Cheuk, "A generalized cross-tabulation matrix to compare soft-classified maps at multiple resolutions," *Int. J. Geographical Inf. Sci.*, vol. 20, no. 1, pp. 1–30, Jan. 2006.
[29] S. Guiasu and A. Shenitzer, "The principle of maximum entropy," *Math. Intell.*, vol. 7, no. 1, pp. 42–48, 1985.
[30] C. Sammut and G. I. Webb, *Encyclopedia of Machine Learning*. Cham, Switzerland: Springer, 2011.
[31] S. Wang, W. Liu, J. Wu, L. Cao, Q. Meng, and P. J. Kennedy, "Training deep neural networks on imbalanced data sets," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 4368–4374.
[32] S. Kar, S. Maharjan, A. Pastor López-Monroy, and T. Solorio, "MPST: A corpus of movie plot synopses with tags," 2018, *arXiv:1802.07858*.
[33] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Zurich, Switzerland: Springer, 2014, pp. 740–755.
[34] G. Bénédict, V. Koops, D. Odijk, and M. de Rijke, "SigmoidF1: A smooth F1 score surrogate loss for multilabel classification," 2021, *arXiv:2108.10566*.
[35] J. Erbani, E. Egyed-Zsigmond, D. Nurbakova, and P.-E. Portier, "When multiple perspectives and an optimization process lead to better performance, an automatic sexism identification on social media with pretrained transformers in a soft label context," in *Proc. Work. Notes CLEF*, 2023, pp. 1–11.
[36] F. C. Bernardini, R. B. D. Silva, E. Meza, and R. das Ostras-RJ-Brazil, "Analyzing the influence of cardinality and density characteristics on multi-label learning," in *Proc. X Encontro Nacional de Inteligencia Artificial e Computacional-ENIAC*, vol. 2013, 2013, pp. 1–11.
[37] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with disentangled attention," 2020, *arXiv:2006.03654*.
[38] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," 2016, *arXiv:1602.07360*.
[39] J. M. Johnson and T. M. Khoshgoftaar, "Thresholding strategies for deep learning with highly imbalanced big data," *Deep Learn. Appl.*, vol. 2, pp. 199–227, Jul. 2021.
[40] F. Provost, "Machine learning from imbalanced data sets 101," in *Proc. AAAI Workshop Imbalanced Data Sets*, vol. 68, 2000, pp. 1–3.
[41] Z.-H. Zhou and X.-Y. Liu, "Training cost-sensitive neural networks with methods addressing the class imbalance problem," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 1, pp. 63–77, Jan. 2006.
[42] M. M. Rahman, S. Malik, M. S. Islam, F. Saad, M. A. Hossain, and A. R. M. Kamal, "An efficient approach to automatic tag prediction from movie plot synopses using transformer-based language model," in *Proc. 25th Int. Conf. Comput. Inf. Technol. (ICCIT)*, Dec. 2022, pp. 501–505.
[43] M. M. Rahman and S. Malik, "Predicting tags for movies from plot synopses using deep learning techniques," Ph.D. dissertation, Dept. of Computer Science and Engineering, Islamic Univ. Technol., Gazipur, Bangladesh, 2019.
[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[45] P. Zhang and M. Wu, "Multi-label supervised contrastive learning," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, no. 15, pp. 16786–16793.
[46] T. Ridnik, H. Lawen, A. Noy, E. Ben, B. G. Sharir, and I. Friedman, "TResNet: High performance GPU-dedicated architecture," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1399–1408.
[47] P. Tan, "Calcul différentiel & optimisation," Tech. Rep. lU3MA261, 2024.

JOHAN ERBANI received the bachelor's degree (hourly equivalent of a double bachelor's degree) in intensive mathematics and the M.Sc. degree in apprentissage et algorithmes (M2A), a double degree in mathematics and informatics from Sorbonne University. He is currently pursuing the Ph.D. degree in computer science and mathematics with INSA Lyon, France, under the supervision of Eld Egyed-Zsigmond, Pierre-Édouard Portier, Diana Nurbakova, and Sonia Ben Mokhtar. His research interests include model evaluation and federated learning.

PIERRE-ÉDOUARD PORTIER received the M.Sc. and Ph.D. degrees in computer science from INSA Lyon, France, in 2010. From 2011 to 2023, he was an Associate Professor with the Computer Science and Information Technologies Department and the LIRIS Laboratory, INSA Lyon. Since September 2023, he has been the Lead Data Scientist at Caisse d'Epargne Rhône-Alpes, France. In addition, he works on natural language processing, including relation extraction and knowledge graph completion, as well as semantic web search and document engineering. His research focuses on machine learning and soft computing techniques applied to big data analytics, such as anomaly detection, XAI for crash prediction, and traffic forecasting. His research interests include the discovery, modeling, and representation of knowledge to be integrated into a data analytics process.

ELŐD EGYED-ZSIGMOND received the master's and Ph.D. degrees in computer science engineering from the National Institute of Applied Sciences of Lyon, INSA Lyon, Lyon, France, in 1999 and 2003, respectively. He is currently a Habilitated Associate Professor with the Computer Science Department, INSA Lyon, and a Researcher at the LIRIS, UMR 5202, CNRS Research Laboratory. His research interests include structured information extraction from documents, information modeling, and natural language processing.

DIANA NURBAKOVA received the Ph.D. degree in computer science from the National Institute of Applied Sciences of Lyon, INSA Lyon, Lyon, France, and the LIRIS, UMR 5205, CNRS Research Laboratory, in December 2018. Then, she occupied a temporary teaching and research fellow (ATER) position with the Preparatory Cycle, INSA Lyon, from 2018 to 2019. After that, she has made a switch to industry. From 2019 to 2020, she was the Research and Development Project Manager at Tilkee SAS, Lyon. She is currently an Associate Professor with the Department of Computer Science and the Preparatory Cycle, INSA Lyon, and the LIRIS Research Laboratory. Her research interests include applied machine learning, recommender systems, user modeling, and persuasive systems. Recently, she has started to work on natural language processing for automatic detection of persuasive techniques within text.