



Ordinal Classification with Label-Dependent Loss

Rosario Delgado¹

Received: 3 November 2025 / Revised: 1 January 2026 / Accepted: 19 February 2026
© The Author(s) 2026

Abstract

Ordinal classification addresses prediction tasks where class labels have a natural order but are not necessarily equally spaced. While traditional approaches typically assume symmetric misclassification costs, many real-world applications exhibit asymmetric, label-dependent penalties. This paper extends previous work on Bayes-optimal decision rules for ordinal classification under symmetric loss (Delgado, 2025) to this more general, cost-sensitive setting. Within a unified decision-theoretic framework, we formalize the interplay between three fundamental components of classification: the *loss function*, which encodes misclassification severity; the *scoring rule*, used to evaluate probabilistic predictions and shown here to satisfy regularity and properness; and the *decision criterion* that maps predictive distributions to class labels. We prove that the proposed scoring rule coincides with the expected loss up to a change of sign –a result of independent interest– and we explicitly characterize sufficient structural conditions under which the resulting decision criterion is well defined and Bayes-optimal. Special attention is given to the interval-scale case, where class distances are explicitly incorporated into the loss, the score, and the decision rule. We show that, depending on the structure of the loss function, Bayes optimality may hold either globally or locally in the space of predictive probability distributions. Empirical results on two real-world datasets, covering both interval-scale and fully ordinal cost-sensitive scenarios and different classifier families, illustrate the practical implications of the proposed approach.

Keywords Ordinal classification · Ordinal regression · Score · Loss function · Bayes-optimal decision rule

Abbreviations

CES Convex escalation structure (Definition 2)

Editor: Ronaldo Prati.

✉ Rosario Delgado
Rosario.Delgado@uab.cat

¹ Department of Mathematics, Autonomous University of Barcelona UAB, Av. de l'Eix Central s/n., 08193 Cerdanyola del Vallès, Spain

CRPS	Continuous ranked probability score (Sect. 1)
DC	Discrete convexity (Definition 3)
MAE	Mean absolute error (Remark 4)
ω -MAE	Omega-weighted Mean Absolute Error (Remark 4, Equation (6))
MAP	Maximum a posteriori. Bayes-optimal decision under 0–1 loss (Sect. 1)
MTh	Multi-thresholding meta-algorithm (Delgado et al., 2024) (Sect. 1)
OCA	Order-compatibility assumption (Definition 5)
Ord-MAP	Ordinal Maximum A Posteriori rule (Delgado et al., 2024, Sect. 1)
ω -Ord-MAP	Omega-weighted Ordinal Maximum A Posteriori rules (Sect. 3)

1 Introduction

Ordinal classification (also known as *ordinal regression*) is a branch of supervised learning that addresses tasks where the target variable represents discrete categories with a natural order—unlike in the nominal setting, where no such ordering exists. This type of problem frequently arises in real-world applications such as medical diagnosis, socioeconomic analysis, and quality evaluation. A typical example involves Likert-type scales (Likert, 1932), which commonly comprise 3, 5, or 7 points that respondents use to indicate their degree of agreement or disagreement with a given statement. These scales are widely used in domains such as sentiment analysis, satisfaction surveys, opinion mining, and information retrieval, especially in recommender systems (Pang and Lee, 2008).

Likert-type scales typically assume uniform spacing between successive categories. However, this assumption can be overly restrictive—particularly in *interval scale* scenarios, where class labels remain ordered but are separated by uneven intervals (Binotto and Delgado, 2025), or whenever a cost-sensitive approximation to misclassification is required. A simple illustration is offered by many digital-content platforms that gauge user preference using just three options: “Not for me”, “I like it”, and “I love it”. In this case, the psychological gap between the first and second categories is clearly larger than that between the second and third, so the classes are not equally spaced.

To address such asymmetries, García-Lapresta and colleagues (García-Lapresta and Pérez-Román, 2015; García-Lapresta et al., 2018; García-Lapresta and Pérez-Román, 2018) introduced the Ordered Qualitative Scales framework, which incorporates expert-elicited notions of ordinal proximity between adjacent categories in the context of unbalanced qualitative scales. Since these proximities are based on informed yet inherently subjective judgments, the authors explore aggregation procedures to combine expert opinions on the relative distances between linguistic terms in ordered qualitative scales. In contrast, our work assumes that the distances between classes are already known—either objectively determined or subjectively assigned but treated as reliable. These distances may vary depending on the specific pair of classes, rendering standard ordinal classification approaches inadequate. It is in this setting that our framework is developed.

In a related vein, Delgado et al. (2024) proposed the Multi-Thresholding (MTh) meta-algorithm, a cost-sensitive strategy at the output level that adjusts predicted class probabilities according to pre-specified misclassification costs—typically provided by domain experts, as in their medical case study—and selects the class with the highest adjusted value. While MTh represents a promising step toward incorporating label-dependent misclassi-

fication costs in ordinal classification, it does not offer a fully integrated or theoretically grounded solution.

In this paper, we adopt a different and more comprehensive approach by extending the research introduced in Delgado (2025)—originally developed for ordinal classification with equally spaced classes—to a more general setting in which class labels are ordered but not necessarily equidistant. Specifically, we address **ordinal classification under label-dependent misclassification costs**: that is, when the penalties associated with classification errors vary according to the involved classes. Rather than proposing a heuristic adjustment of existing decision rules, we develop a fully decision-theoretic framework in which the loss function, the scoring rule, and the decision criterion are derived and aligned consistently.

Our aim is, thus, to make explicit the intrinsic relationship among several fundamental components of the classification process, which are often implicitly conflated but play conceptually distinct roles in a decision-theoretic framework. Specifically, we distinguish the following elements:

- (a) the **loss function**, which is part of the problem specification and encodes the misclassification cost structure, quantifying the severity of different types of errors;
- (b) the **scoring rule**, which evaluates the quality of the predicted probability distribution and determines which probabilistic forecasts are optimal during model training (we refer to its value as **score**); and
- (c) the **decision criterion**, which maps a predictive probability distribution to a single class label and is Bayes-optimal with respect to the chosen loss function.

In addition, we consider a fourth element:

- (d) the **performance metric**, used for empirical evaluation, which quantifies discrepancies between predicted and observed class labels and should remain consistent with the loss function underlying the decision rule.

The main contribution of this work is to characterize how these elements must be aligned in ordinal classification problems with label-dependent misclassification costs, and to derive the corresponding Bayes-optimal decision rule.

In ordinal classification, the relative positions of the classes determine the severity of misclassification errors: predicting a class close to the true label typically incurs a smaller cost than predicting a distant one. Such asymmetries cannot be represented under the 0–1 loss used in nominal classification and require loss functions that explicitly encode the ordering and spacing of the labels. These “distances” are not properties of the predictors but are encoded in the loss function, reflecting domain-specific misclassification costs.

In this setting, the Bayes decision rule that incorporates these costs is not a post-hoc adjustment but the statistically optimal criterion, as it minimizes the expected loss under the relevant cost structure. The predictive distribution produced by the classifier must thus be combined with an appropriate loss function to obtain a decision rule that fully respects the ordinal structure of the labels.

We emphasize that in this work we do not address the construction of the probabilistic classifier itself—a topic investigated, for instance, by Suárez et al. (2021) and Alcacer et al. (2024) in the context of standard ordinal classification. Rather, we take the probabilistic out-

put of a classifier as given and focus on identifying the decision rule that best suits ordinal classification when misclassification losses are label-dependent.

The decision criterion (component (c)) is typically derived as the **Bayes-optimal decision rule**, which minimizes the expected loss (component (a)) with respect to the predicted class distribution. Formally, we observe training pairs of (X, Y) , where X denotes the feature vector, and Y is the ordinal class label, taking values in $\{1, \dots, r\}$, with $r \geq 2$. Given that $X = x$, the classifier produces a predictive (a posterior) probability distribution $p(x) = (p_1(x), \dots, p_r(x))$ over the classes $\{1, \dots, r\}$, and the aim is to select a class label that minimizes the expected loss under the chosen misclassification cost structure. In general, given a loss function λ and a probability distribution $p = (p_1, \dots, p_r)$, the Bayes-optimal decision k^* is given as:

$$k^* = \arg \min_{k=1, \dots, r} \sum_{j=1}^r \lambda(j, k) p_j,$$

where $\lambda(j, k)$ is the loss associated to misclassify as belonging to class k an item which actually is of class j . In the nominal (unordered) setting, the most commonly used loss function is **0–1 loss**, which assigns a unit penalty to any misclassification, regardless of the classes involved. The corresponding Bayes-optimal decision rule is the well known **MAP** (maximum a posteriori) criterion, which selects the class with the highest posterior probability—that is, the **mode** of the distribution. This decision rule also maximizes the **Brier score**. The consistence among these components—loss function (0–1), Brier score, and MAP—has established a solid theoretical foundation for nominal classification (Section 2.2 Delgado (2025)).

In the ordinal setting with equally spaced classes, a natural loss function is the **absolute error**, which penalizes errors proportionally to the distance between true and predicted classes (component (a)). Under this loss, the Bayes-optimal decision rule is the **Ord-MAP** criterion introduced in Delgado (2025) (component (c)), which selects the **median** of the posterior distribution—specifically, the first class whose cumulative probability exceeds 0.5. This rule is tailored for ordinal data, as it also maximizes a discrete version of the **Continuous Ranked Probability Score (CRPS)**, a scoring rule (component (b)) that takes into account the inherent ordering of the classes by leveraging the cumulative distribution function and rewarding predictions that concentrate probability mass near the true label. By aligning the absolute error loss, a proper ordinal scoring function, and an appropriate decision rule, Ord-MAP (Ordinal Maximum A Posteriori) provides a coherent and informative alternative to approaches based on nominal classification.

By formalizing the interplay between the loss function, scoring rule, and decision criterion, this paper offers both theoretical insight and practical tools for optimal decision-making in ordinal classification under label-dependent loss. This generalization not only deepens the understanding of previous frameworks but also enables more accurate, application-aware models in scenarios where class distances are heterogeneous—such as in health risk levels, satisfaction surveys with uneven perception gaps, or socioeconomic indicators characterized by unequal transitions between categories.

Throughout the paper, probabilistic inference is treated separately from decision-making. We assume that a classifier provides an estimated posterior distribution over the ordered

classes, and we do not modify or constrain this inference step in any way. The contribution of this work concerns exclusively the decision rule that maps a given posterior distribution to a single class label once a label-dependent loss function is specified. Figure 1 provides a schematic overview of the classification pipeline, making explicit the distinct roles played by the probabilistic prediction stage and the decision-theoretic layer.

These considerations can be made more explicit by examining how the loss function interacts with the predictive distribution to shape the Bayes-optimal decision. The following example clarifies this relationship.

Example 1 (Motivating example) Consider an ordinal classification problem with three ordered classes $\{1, 2, 3\}$. Suppose the classifier outputs the predictive probability distribution $p = (p_1, p_2, p_3) = (0.4, 0.3, 0.3)$, and assume three different scenarios:

- (i) **Nominal setting with 0–1 loss** If we treat the classes as unordered and use the standard 0–1 loss $\lambda(j, k) = 1_{\{j \neq k\}}$, where 1_A denotes a function that equals 1 if A holds and 0 otherwise, the expected loss incurred when predicting class k is

$$R(p, k) = \sum_{j=1}^3 \lambda(j, k) p_j = \sum_{j \neq k} p_j = 1 - p_k, \quad k = 1, 2, 3.$$

The Bayes rule chooses the class that minimizes the expected loss, hence, the class with maximum probability (the **mode**). In this case, class 1 since $0.4 = \max\{0.4, 0.3, 0.3\}$. This is the usual MAP rule.

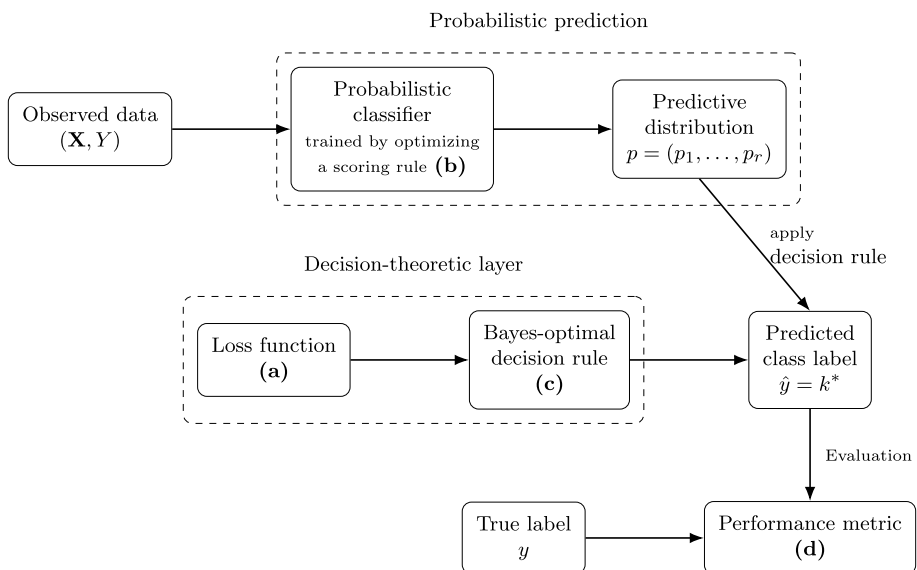


Fig. 1 Conceptual pipeline of ordinal classification: a probabilistic classifier (trained via a scoring rule) produces a predictive distribution p , which is converted into a point prediction $\hat{y} = k^*$ through the Bayes-optimal decision rule under a label-dependent loss. Performance is assessed through a metric aligned with the targeted loss structure

- (ii) **Ordinal setting with symmetric absolute-error loss** If we incorporate the order of the classes and use the absolute-error loss $\lambda(j, k) = |j - k|$, the expected loss becomes

$$R(p, 1) = 0 \times 0.4 + 1 \times 0.3 + 2 \times 0.3 = 0.9,$$

$$R(p, 2) = 1 \times 0.4 + 0 \times 0.3 + 1 \times 0.3 = 0.7,$$

$$R(p, 3) = 2 \times 0.4 + 1 \times 0.3 + 0 \times 0.3 = 1.1.$$

The Bayes-optimal decision is now class 2, corresponding to the **median** of the distribution, rather than the mode. This is precisely the Ord-MAP rule (Delgado (2025)).

- (iii) **Ordinal setting with label-dependent misclassification costs** Consider now a label-dependent loss structure, where errors of misclassifying extreme class 3 are penalized more heavily. For instance, let

$$\lambda(j, k) = \begin{cases} 6 & \text{if } j = 3, k = 1 \\ 3 & \text{if } j = 3, k = 2 \\ |j - k| & \text{otherwise} \end{cases}.$$

The corresponding expected loss is

$$R(p, 1) = 0 \times 0.4 + 1 \times 0.3 + 6 \times 0.3 = 2.1,$$

$$R(p, 2) = 1 \times 0.4 + 0 \times 0.3 + 3 \times 0.3 = 1.3,$$

$$R(p, 3) = 2 \times 0.4 + 1 \times 0.3 + 0 \times 0.3 = 1.1.$$

In this case, the Bayes-optimal decision is class 3.

Thus, for the same predictive distribution p , the optimal decision is: mode (class 1) under 0–1 loss, median (class 2) under symmetric absolute-error loss, and class 3 under the above label-dependent loss.

This simple example illustrates that once we move beyond 0–1 loss, the standard MAP rule is no longer appropriate: the Bayes decision depends crucially on the loss structure. The example anticipates the central question addressed in this paper: how to systematically derive a Bayes-optimal decision rule for ordinal classification when misclassification costs depend on the specific labels involved. Indeed, the aim of this paper is precisely to formalize the relationship between loss function and Bayes-optimal decision for ordinal problems with label-dependent misclassification costs, and to show how the Bayes-optimal rule leads to the ω -Ord-MAP (Omega-weighted Ordinal Maximum A Posteriori) criterion introduced in Sect. 3.

1.1 Organization of the Paper

Section 2 introduces the first two components of the fundamental classification triad: the **label-dependent loss function** and the **scoring rule**. In Sect. 2.3, we establish a key connection between them (Proposition 1), showing that the expected loss and the scoring rule coincide up to a change of sign. This section also discusses two essential properties of scoring rules—**regularity** and **properness**—with a formal proof of properness deferred to Appendix Appendix A (Theorem 2). Section 2.2 introduces the global structural properties

of the loss function that play a central role in the sequel: discrete convexity (DC) and convex escalation structure (CES).

Section 3 addresses the third element of the triad, namely the **decision criterion**. It formally introduces the ω -**Ord-MAP** rule and presents the main theoretical result of the paper, establishing its Bayes optimality (Theorem 1). The proof, given in Appendix B, is derived under a local order-compatibility assumption (OCA) between the loss function and the posterior distribution p . Proposition 2 shows that the global condition (CES) implies (OCA) for all distributions p . The ω -Ord-MAP criterion is illustrated through a toy example (Sect. 3.1) and further examined in the binary asymmetric setting (Sect. 3.2).

Section 4 focuses on a particularly relevant special case: classification on the interval scale, where classes are naturally interpreted as numerical intervals. A three-class scenario is analyzed using two alternative loss constructions: one based on the Hausdorff distance between intervals (Sect. 4.1) and another based on the cumulated cost of transition between adjacent classes (Sect. 4.2). Section 5 presents two empirical studies on real-world datasets, covering both interval-scale and fully ordinal cost-sensitive settings, and different classifier families, to assess the practical implications of the proposed decision rule. Appendix C provides complementary sensitivity and diagnostic analyses. Finally, Sect. 6 summarizes the main contributions and outlines directions for future research.

2 Label-Dependent Loss Function and Scoring Rule

In this section, we introduce a generalization of the absolute error loss function that accounts for label-dependent costs, denoting it by λ_ω (Sect. 2.1). In Sect. 2.3, we also propose a weighted version of the (discrete) Continuous Ranked Probability Score (CRPS), referred to as S_ω , which incorporates a class-dependent weighting scheme. We first establish the connection between these two fundamental components in Proposition 1, and then show that S_ω is a **proper** scoring rule.

2.1 Loss Function and Its Expectation

Consider an ordinal classification problem with $r \geq 1$ ordered classes, labeled as $1 < \dots < r$. The ordinal nature of the problem becomes meaningful only when $r \geq 3$. The loss function reflects the costs of misclassification, which are part of the decision problem. Through this work, we assume that the loss function satisfies a set of basic assumptions.

Definition 1 (*Basic Assumptions*) We assume that the loss incurred when the true class is j and the predicted class is k is defined by a loss function:

$$\lambda_\omega(j, k) = \omega_{jk}, \quad j, k = 1, \dots, r,$$

subject to the following **basic assumptions**:

1. *Non-negativity*: $\omega_{jk} \geq 0$ for all $j, k \in \{1, \dots, r\}$.
2. *Identity of indiscernibles*: $\omega_{jk} = 0 \iff j = k$ for all $j, k \in \{1, \dots, r\}$.

3. *Monotonicity*: $\omega_{jk} \leq \omega_{j\ell}$ and $\omega_{kj} \leq \omega_{\ell j}$ whenever $1 \leq j < k < \ell \leq r$, reflecting increasing penalty as the predicted class moves further away from the true one.

We also introduce the $r \times r$ loss matrix $W = (\omega_{jk})_{j,k=1,\dots,r}$.

For example, the absolute error loss, defined as $\omega_{jk} = |j - k|$, satisfies all these assumptions. In addition, it exhibits two properties that are not assumed in our general setting, *symmetry* ($\omega_{jk} = \omega_{kj}$) and *linearity*: $\omega_{j\ell} = \omega_{jk} + \omega_{k\ell}$ if $1 \leq j < k < \ell \leq r$. We also do not assume the *triangle inequality*: $\omega_{j\ell} \leq \omega_{jk} + \omega_{k\ell}$, which would be required if λ_ω were a proper distance metric—a property we do not require, though it may hold in specific cases.

Given a loss function λ_ω and a probability distribution $p = (p_1, \dots, p_r)$ over the classes with $p_i \geq 0$ and $\sum_{i=1}^r p_i = 1$, representing the posterior probabilities output by a classifier, the **expected loss** incurred when predicting class k is:

$$R_\omega(p, k) = \sum_{j=1}^r \lambda_\omega(j, k) p_j = \sum_{j=1}^r \omega_{jk} p_j = \sum_{j \neq k} \omega_{jk} p_j. \quad (1)$$

This quantity represents the expected penalty under λ_ω when the classifier predicts class k and the true class is j with probability p_j .

We emphasize that choosing a loss function is an integral part of statistical decision theory: once λ_ω is specified, the notion of error changes accordingly. The classification that minimizes the expected loss R_ω is therefore the Bayes-optimal rule for this problem, in exactly the same sense in which the MAP criterion is optimal for the 0–1 loss.

2.2 Global Structural Properties of the Loss Function

Throughout the paper, we make a clear distinction between global structural properties of the loss function, which depend solely on λ_ω , and local compatibility conditions, which involve both the loss function and the predictive probability distribution p , and determine whether the decision rule is well defined for a given p . We now introduce and formalize the former through two different structural properties.

Although both formalize the idea that misclassification errors become increasingly severe as the predicted class moves farther away from the true one, the key distinction lies in the level at which this escalation is enforced, as we will explain below.

1. **Convex escalation of misclassification costs** The first global structural property establishes that the marginal cost of moving one additional step away from the true class is nondecreasing with distance from this class, so that errors involving farther classes are penalized increasingly more severely.

Definition 2 (*Convex Escalation Structure (CES)*) A loss function λ_ω is said to satisfy the *Convex Escalation Structure (CES)* if, for each true class j , the incremental costs along each row of the loss matrix W , $\omega_{jk+1} - \omega_{jk}$ and $\omega_{jk} - \omega_{jk-1}$, are nondecreasing as k moves away from j on either side.

Note that (CES) is a row-wise property of the loss matrix: for each fixed true class j (row), it requires the marginal loss increments to be nondecreasing as the predicted class moves away from column j along the row. Importantly, these increments may depend on the specific true class, allowing different escalation patterns across rows of the loss matrix.

In contrast to (CES), which enforces convex escalation conditional on the true class, the stronger property below enforces convex escalation uniformly across classes by requiring the loss to depend only on the ordinal distance.

2. **Discrete convexity of distance-based losses** Misclassification severity can be modeled through an ordinal, additive global cost structure that generalizes the equally spaced classes setting viewing misclassifying from j to k as a sequence of unit steps across adjacent classes, in which the severity of each elementary step could be encoded by a nonnegative function of the ordinal displacement. A simple and principled choice is to let

$$\omega_{j\,k} = g(|j - k|)$$

where $g : \{0, \dots, r - 1\} \rightarrow \mathbb{R}_+$, the generating function, is an increasing and discretely convex function of the distance $|j - k|$. (Note that it depends on j and k only through the distance $|j - k|$).

Definition 3 (*Discrete convexity (DC)*) A function g defined on the integers is said to be **discretely convex (DC)** if $g(0) = 0$ and its second-order finite differences are nonnegative,

$$\Delta^2 g(h) = g(h + 2) - 2g(h + 1) + g(h) \geq 0,$$

which is equivalent to nondecreasing forward increments,

$$g(h + 2) - g(h + 1) \geq g(h + 1) - g(h).$$

Typical examples are

- Linear: $g(d) = d$, corresponding to the absolute error loss.
- Quadratic: $g(d) = \alpha d^2$ with $\alpha > 0$.
- Hybrid: $g(d) = d + \alpha d^2$ with $\alpha > 0$.
- Exponential (soft): $g(d) = \rho^d - 1$ with $\rho > 1$.
- Cumulated increments: $g(d) = \sum_{i=0}^{d-1} \delta_i$, with $\delta_0 \leq \delta_1 \leq \dots$, natural when each additional step adds at least as much cost as the previous one.

One may also allow for asymmetric costs by specifying two discretely convex functions g_L and g_R , and defining

$$\omega_{j\,k} = \begin{cases} g_L(j - k), & j > k, \\ g_R(k - j), & j < k, \end{cases}$$

so that g_L and g_R take as argument a nonnegative displacement or distance, and thereby accommodating situations where underestimation and overestimation have different severities, while preserving convex escalation of the loss on each side of the true class (e.g. $g_L(d) = 2d^2$ and $g_R(d) = d^2$, penalizing more underestimation).

Strictly speaking, discrete convexity (DC) is a property of the generating function. However, by a slight abuse of language, we will say that the loss function λ_ω satisfies (DC) whenever it is distance-based and admits a representation of the form

$$\omega_{jk} = \begin{cases} g(|j-k|) & \text{(symmetric case),} \\ g_L(j-k) & \text{if } j > k, \\ g_R(k-j) & \text{if } j < k, \end{cases}$$

where the generating function g (or g_L and g_R in the asymmetric case) is discretely convex.

Remark 1 In contrast to (CES), discrete convexity (DC) imposes a stronger global structural constraint: the loss is assumed to depend only on the ordinal distance $|j-k|$ through a single discretely convex function. As a consequence, all rows of the loss matrix share the same convex profile, up to translation, and the escalation of misclassification costs is uniform across classes.

Therefore, (DC) can be seen as a homogeneous, distance-based specialization of (CES). While every discretely convex loss satisfies (CES), the converse does not hold, as (CES) allows for class-dependent escalation patterns. Schematically,

$$(\text{DC}) \implies (\text{CES}) \quad \text{but} \quad (\text{CES}) \not\Rightarrow (\text{DC})$$

2.3 Scoring Rule

Savage (Savage (1971)) introduced the general concept of **scoring rule**, laying the theoretical groundwork for evaluating probabilistic forecasts. That work emphasized the importance of coherent probability assessments in both decision-making and statistical inference. A scoring rule assigns a numerical value (a score) to a probabilistic forecast, which typically takes the form of a distribution $p = (p_1, \dots, p_r)$, based on the observed (true) outcome. The goal is to assess the quality of the forecast: the better the forecast aligns with the actual outcome, the better the score.

In this context, we introduce a scoring rule S_ω , which is a weighted version of the (discrete) Continuous Ranked Probability Score (CRPS). For simplicity, we will sometimes refer to S_ω itself as “the score” when no ambiguity arises. This scoring rule incorporates a class-dependent weighting structure derived from a general loss function and is inspired by the **classical discrete CRPS score**, which is defined as:

$$S(p, k) = - \sum_{\ell=1}^r \left((p_1 + \dots + p_\ell) - 1_{\{k \leq \ell\}} \right)^2 \quad \text{where } 1_{\{k \leq \ell\}} = \begin{cases} 1 & \text{if } k \leq \ell \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

Then, we can rewrite

$$\begin{aligned}
 S(p, k) &= - \sum_{\ell=1}^{k-1} (p_1 + \cdots + p_\ell)^2 - \sum_{\ell=k}^r (p_{\ell+1} + \cdots + p_r)^2 \\
 &= - \sum_{\ell=1}^{k-1} \left(\sum_{j=1}^{\ell} p_j \right)^2 - \sum_{\ell=k}^r \left(\sum_{j=\ell+1}^r p_j \right)^2.
 \end{aligned} \tag{3}$$

In the above and all subsequent expressions, it is understood, as is customary, that any summation whose upper bound is smaller than its lower bound vanishes.

Building on this formulation, we define the weighted version as follows.

Definition 4 (The S_ω score) Given a probability distribution $p = (p_1, \dots, p_r)$, a loss function λ_ω , and a class $k = 1, \dots, r$, the S_ω score is defined by:

$$S_\omega(p, k) = - \sum_{\ell=1}^{k-1} \left(\sum_{j=1}^{\ell} (\omega_{j\ell+1} - \omega_{j\ell}) p_j \right) - \sum_{\ell=k}^r \left(\sum_{j=\ell+1}^r (\omega_{j\ell} - \omega_{j\ell+1}) p_j \right). \tag{4}$$

This quantity can be interpreted as the reward assigned to the quoted probability distribution p when the actual class is k , under the loss function λ_ω . It extends the CRPS by integrating label-dependent costs into the scoring rule, through weights applied to the non-true-class probabilities p_j ($j \neq k$), defined as local increments of the loss function.

Remark 2 If the loss function is the absolute error loss, i.e., $\omega_{ij} = |i - j|$, the score simplifies to:

$$S_\omega(p, k) = - \sum_{\ell=1}^{k-1} \left(\sum_{j=1}^{\ell} p_j \right) - \sum_{\ell=k}^r \left(\sum_{j=\ell+1}^r p_j \right) = - \sum_{\ell=1}^r \left| (p_1 + \cdots + p_\ell) - 1_{\{k \leq \ell\}} \right|.$$

This corresponds to a variant of the CRPS that uses absolute differences (ℓ_1 -norm) instead of the usual squared differences (ℓ_2 -norm). We refer to this variant as S^{ℓ_1} . Since the ℓ_1 and ℓ_2 norms are topologically equivalent in finite-dimensional spaces, we can relate this variant to the standard CRPS as follows:

$$-\sqrt{r} \left(-S(p, k) \right)^{1/2} \leq S^{\ell_1}(p, k) \leq - \left(-S(p, k) \right)^{1/2}.$$

The following result provides an alternative expression for the score S_ω and establishes its equivalence, up to a sign change, with the expected loss R_ω , introduced earlier in (1). This identity is particularly useful, as it facilitates the derivation of further results.

Proposition 1 For any probability distribution $p = (p_1, \dots, p_r)$, any loss function λ_ω , and any class $k = 1, \dots, r$, the following identity holds:

$$S_\omega(p, k) = -R_\omega(p, k). \tag{5}$$

Proof From the definition of $S_\omega(p, k)$ (4), we can decompose it into two additive terms:

$$(a) = - \sum_{\ell=1}^{k-1} \left(\sum_{j=1}^{\ell} (\omega_{j\ell+1} - \omega_{j\ell}) p_j \right) \quad \text{and} \quad (b) = - \sum_{\ell=k}^r \left(\sum_{j=\ell+1}^r (\omega_{j\ell} - \omega_{j\ell+1}) p_j \right).$$

To simplify term (a), observe that the conditions $1 \leq \ell \leq k-1$, $1 \leq j \leq \ell$ are equivalent to $1 \leq j \leq k-1$, $j \leq \ell \leq k-1$. Therefore:

$$(a) = - \sum_{j=1}^{k-1} \left(\sum_{\ell=j}^{k-1} (\omega_{j\ell+1} - \omega_{j\ell}) \right) p_j = - \sum_{j=1}^{k-1} (\omega_{jk} - \omega_{jj}) p_j = - \sum_{j=1}^{k-1} \omega_{jk} p_j$$

since $\omega_{jj} = 0$ by definition of the loss function. For term (b), similarly, the conditions $k \leq \ell \leq r$, $\ell+1 \leq j \leq r$ are equivalent to $k+1 \leq j \leq r$, $k \leq \ell \leq j-1$. Hence:

$$(b) = - \sum_{j=k+1}^r \left(\sum_{\ell=k}^{j-1} (\omega_{j\ell} - \omega_{j\ell+1}) \right) p_j = - \sum_{j=k+1}^r (\omega_{jk} - \omega_{jj}) p_j = - \sum_{j=k+1}^r \omega_{jk} p_j.$$

Combining terms (a) and (b), we obtain:

$$S_\omega(p, k) = (a) + (b) = - \sum_{j \neq k} \omega_{jk} p_j,$$

which is precisely the negative of the expected loss $R_\omega(p, k)$, as defined earlier in (1). This completes the proof. $\square$

Remark 3 The form in which we define the scoring rule S_ω follows from the fact that, under label-dependent losses, the Bayes-optimal decision rule no longer depends merely on point-wise probabilities, but on how probability mass accumulates across ordered classes. The construction reflects this cumulative structure: each term of S_ω corresponds to the contribution of a possible misclassification displacement weighted by its incremental cost. Hence the score evaluates a predictive distribution by penalizing deviations in proportion to their severity under the loss λ_ω .

This mirrors the role played by the Continuous Ranked Probability Score (CRPS) in the case of ordinal classification with equally spaced classes, where CRPS is the strictly proper scoring rule whose maximization yields the Ord-MAP rule. The proposed score S_ω can be viewed as the natural extension of this idea to settings with label-dependent misclassification costs. It is strictly proper, and its maximization is equivalent to minimizing expected loss (Proposition 1), which explains both its form and its compatibility with the ω -Ord-MAP decision rule introduced in Section 3.

Remark 4 For a deterministic forecast where the probability distribution is given by $p = e_i$ for $e_i \in \mathbb{R}^r$ denoting the i -th unit vector, with all the probability mass concentrated on class i , expressions (5) and (1) imply that

$$S_\omega(p, k) = -\omega_{ik} = -\lambda_\omega(i, k).$$

This highlights the close connection between the score S_ω and the underlying loss function λ_ω . Such a relationship also arises in other classification settings. In nominal classification, the Brier score is directly linked to the 0-1 loss and the error rate (see Section 2.2 in Delgado (2025)). In ordinal classification with equally spaced classes, the classical discrete CRPS score given by Equation (2) is tied to the absolute error loss function (and the Mean Absolute Error, MAE). In particular, when the predictive distribution is degenerate, $p = e_i$, corresponding to a point prediction at class i , the score reduces to $S(p, k) = -|i - k|$, which coincides, up to a sign, with the absolute error loss (see Section 4.1 Delgado (2025)).

This interplay between scores and losses also extends to performance metrics, which evaluate prediction quality across a validation dataset. One of the most widely used metrics in ordinal classification is the *Mean Absolute Error* (MAE). It is defined as the average absolute difference between the true class labels y_i and the predicted labels $\hat{y}_i$, both taking values in $\{1, \dots, r\}$. Formally, $\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$, where n is the total number of instances. In a label-dependent setting, we introduce the ω -MAE (Omega-weighted Mean Absolute Error) metric, which is based on the loss function λ_ω . It generalizes MAE by replacing the absolute difference with the cost $\omega_{y_i, \hat{y}_i}$, as follows:

$$\omega\text{-MAE} = \frac{1}{n} \sum_{i=1}^n \lambda_\omega(y_i, \hat{y}_i) = \frac{1}{n} \sum_{i=1}^n \omega_{y_i, \hat{y}_i}, \quad (6)$$

thereby addressing component (d) of the classification process.

Following Savage (Savage, 1971), a scoring rule such as S_ω is expected to satisfy two key properties to be considered valid and useful, both in practice and in theoretical frameworks:

- **Regularity** According to Definition 1 in Gneiting and Raftery (2007), a scoring rule is regular if it is finite for all valid probability distributions. S_ω meets this condition because it is defined as a finite sum, and therefore always yields finite values.
- **Properness (or Propriety)** A scoring rule is said to be proper if the expected score is maximized when the forecast matches the true distribution (cf. (1) Gneiting and Raftery (2007)). Specifically, for any two probability distributions p and q over the set of classes, the score S_ω is proper if

$$S_\omega(p | p) \geq S_\omega(p | q),$$

where $S_\omega(p | q)$ denotes the expected score obtained when quoting forecast p while the true distribution is q , and is computed as

$$S_\omega(p | q) = \sum_{k=1}^r S_\omega(p, k) q_k \quad \text{if} \quad q = (q_1, \dots, q_r).$$

In other words, the expected reward is maximized when the predicted distribution p coincides with the true distribution. Proper scoring rules are thus designed to encourage honest and accurate probability estimates. A formal proof that S_ω satisfies the propriety property is provided in Appendix [Appendix A](#).

Because of these properties, proper scoring rules are widely used to evaluate probabilistic forecasts. For instance, Robert et al. (2024) introduces the Weighted Interval Score and the Ranked Probability Score, which are tailored for quantile and integer-valued predictions, respectively, and Vrugt (2024) provide a comprehensive review of scoring functions used in hydrology, covering applications in binary, categorical, and continuous density settings.

3 Decision Criterion: the ω -Ord-MAP

In this section, we introduce an extension of the Ord-MAP criterion proposed by Delgado (2025), referred to as the ω -Ord-MAP. This new criterion generalizes Ord-MAP to ordinal classification problems involving label-dependent loss functions, thereby allowing for greater flexibility in modeling varying class relationships and error costs.

We begin by formally defining the ω -Ord-MAP criterion. We then demonstrate that this rule constitutes a Bayes-optimal decision rule, meaning that it minimizes the expected loss R_ω . Moreover, since Proposition 1 establishes that this expected loss is equivalent (up to a change in sign) to the score S_ω , it follows that the ω -Ord-MAP criterion is also optimal in the sense that it maximizes the score.

Definition 5 (ω -Ord-MAP criterion and Order-Compatibility Assumption (OCA)) Given the probability distribution $p = (p_1, \dots, p_r)$ over the r ordered classes and a loss function λ_ω satisfying the basic assumptions, the ω -**Ord-MAP** criterion selects the class $k^* \in \{1, \dots, r\}$ given by:

$$k^* = \begin{cases} \min_r C_\omega(p) & \text{if } C_\omega(p) \neq \emptyset \\ r & \text{if } C_\omega(p) = \emptyset \end{cases}$$

where

$$C_\omega(p) = \left\{ k = 1, \dots, r-1 : \sum_{j=1}^k (\omega_{j,k+1} - \omega_{j,k}) p_j \geq \sum_{j=k+1}^r (\omega_{j,k} - \omega_{j,k+1}) p_j \right\},$$

imposing the following **Order-Compatibility Assumption (OCA)** between the loss function and p : if a class k satisfies the defining inequality of the ω -Ord-MAP rule, then all classes $k' > k$ also satisfy it. More formally,

$$(OCA) \quad \text{if } k \in C_\omega(p) \text{ and } k < r-1, \quad \text{then } k' \in C_\omega(p) \quad \forall k' = k+1, \dots, r-1.$$

This criterion is based on comparing two weighted cumulative sums of the probability mass: one from the left (up to class k) and one from the right (from class $k+1$ onward). The

weights used in the comparison are given by the differences of loss values between neighboring classes:

- $\omega_{j\ k+1} - \omega_{j\ k}$ for $j \leq k$, and
- $\omega_{j\ k} - \omega_{j\ k+1}$ for $j > k$.

Under the basic assumptions, these weight differences are always non-negative.

Note that since ω -Ord-MAP defines a deterministic mapping from the probability simplex to the label space, its stability and well-definedness depend solely on the structural properties of the loss function, and not on the convergence behavior of the underlying probabilistic learner.

Remark 5 (*Computational complexity and scalability*) The ω -Ord-MAP criterion is a closed-form decision rule that operates on the predictive probability distribution $p = (p_1, \dots, p_r)$. Its evaluation requires at most $r - 1$ cumulative sums and comparisons, yielding a computational cost of $O(r)$ per instance. Since r is typically small in ordinal classification, the cost of the decision rule is negligible relative to that of the underlying probabilistic classifier. Consequently, ω -Ord-MAP scales trivially to high-dimensional and large-sample settings.

Remark 6 When the loss function corresponds to equally spaced classes, that is, $\omega_{j\ k} = |j - k|$, (OCA) trivially holds with any probability distribution p and ω -Ord-MAP rule simplifies. In this case, the condition becomes:

$$k^* = \min \left\{ k = 1, \dots, r - 1 : \sum_{j=1}^k p_j \geq \sum_{j=k+1}^r p_j \right\}$$

if the set is non-empty, and $k^* = r$ otherwise. This expression coincides exactly with the median of the distribution p , which is the decision rule prescribed by the Ord-MAP criterion in Delgado (2025).

In general, the difference between Ord-MAP and ω -Ord-MAP does not lie in how probabilities are estimated, but in how they are used for decision-making. Both rules operate on the same posterior distribution; they differ only in the loss function with respect to which optimality is defined.

In Appendix [Appendix B](#) we prove the main theoretical result of the paper, which is Theorem [1](#) below. This result establishes that the ω -Ord-MAP decision criterion is Bayes-optimal with respect to the loss function λ_ω if basic assumptions and (OCA) hold. Specifically, it minimizes the expected loss R_ω and, equivalently, maximizes the associated score S_ω .

Theorem 1 Let $p = (p_1, \dots, p_r)$ be a probability distribution, and let λ_ω be a loss function satisfying the basic assumptions. Assume that hypothesis (OCA) holds. Then, the class k^* selected by the ω -Ord-MAP criterion (as given in Definition [5](#)) satisfies:

$$k^* = \arg \max_{k=1, \dots, r} S_\omega(p, k) = \arg \min_{k=1, \dots, r} R_\omega(p, k),$$

i.e., it maximizes the score and minimizes the expected loss and therefore constitutes the

Bayes-optimal decision rule.

It is important to emphasize that selecting a single class label from a posterior distribution necessarily involves a loss function. A so-called cost-free decision implicitly corresponds to assuming the 0–1 loss. In ordinal or cost-sensitive settings, this assumption is no longer appropriate, and adopting a loss-aware decision rule does not modify the inference itself, but rather aligns the decision with the problem-specific notion of error severity.

Remark 7 Assumption (OCA) reflects the monotonicity of the decision condition used to define the ω -Ord-MAP rule. That is, it ensures that the set of admissible classes forms an upper interval $\{k_0, k_0 + 1, \dots, r\}$, and the ω -Ord-MAP criterion is well defined as the minimum class in this interval, $k^* = k_0$.

This local compatibility condition (between the loss function λ_ω and the distribution of probability p), which is needed for the well definition of the Bayes-optimal decision rule, is implied by the global structural properties of the loss function introduced in Sect. 2.2, as the following proposition establishes.

Proposition 2 Assume that the loss function $\lambda_\omega(j, k)$ verifies (CES) (Definition 2). Therefore, λ_ω satisfies (OCA) with any probability distribution over the classes p .

Proof We must prove that if $k < r - 1$ belongs to the set $\mathcal{C}_\omega(p)$ for any p , necessarily also $k + 1 \in \mathcal{C}_\omega(p)$. That is, assuming

$$\sum_{j=1}^k (\omega_{j, k+1} - \omega_{j, k}) p_j \geq \sum_{j=k+1}^r (\omega_{j, k} - \omega_{j, k+1}) p_j \quad (7)$$

we must prove that

$$\sum_{j=1}^{k+1} (\omega_{j, k+2} - \omega_{j, k+1}) p_j \geq \sum_{j=k+2}^r (\omega_{j, k+1} - \omega_{j, k+2}) p_j,$$

but

$$\begin{aligned} \sum_{j=1}^{k+1} (\omega_{j, k+2} - \omega_{j, k+1}) p_j &= \sum_{j=1}^k (\omega_{j, k+2} - \omega_{j, k+1}) p_j + \omega_{k+1, k+2} p_{k+1} \\ &\stackrel{(\text{CES})}{\geq} \sum_{j=1}^k (\omega_{j, k+1} - \omega_{j, k}) p_j \stackrel{(7)}{\geq} \sum_{j=k+1}^r (\omega_{j, k} - \omega_{j, k+1}) p_j \\ &= \sum_{j=k+2}^r (\omega_{j, k} - \omega_{j, k+1}) p_j + \omega_{k+1, k} p_{k+1} \stackrel{(\text{CES})}{\geq} \sum_{j=k+2}^r (\omega_{j, k+1} - \omega_{j, k+2}) p_j, \end{aligned}$$

which finishes the proof. $\square$

Remark 8 When the loss function does not verifies (CES), (OCA) may still be satisfied for specific probability distributions p , as illustrated in Sect. 4. In such cases, the ω -Ord-MAP criterion remains well defined, but only locally in p . Schematically, the relationships between the three assumptions on λ_ω are:

$$(DC) \implies (CES) \implies (OCA) \quad \text{for any } p.$$

3.1 The Motivating Example Revisited

To illustrate the ω -Ord-MAP criterion –specifically, how the dynamic nature of the weight differences in Definition 5 and the choice of the loss function λ_ω can influence the selected class– we revisit the motivating Example 1, with $r = 3$ ordered classes. We assume that the classifier provides the probability distribution $p = (0.4, 0.3, 0.3)$. This example also illustrates the role of the Order-Compatibility Assumption (OCA) in ensuring that the criterion is well defined.

According to the standard MAP criterion, the predicted class in the context of nominal classification would be 1, corresponding to the **mode**. In contrast, the Ord-MAP criterion for ordinal classification with equally spaced classes would assign class 2, the **median** of the distribution. What class does the ω -Ord-MAP criterion assign? The answer depends on the loss function λ_ω , which captures the label-dependent misclassification penalties. The selected class k^* will be:

$$\begin{aligned} k^* = 1 & \quad \text{if} & \quad \omega_{12} p_1 \geq \omega_{21} p_2 + (\omega_{31} - \omega_{32}) p_3 \\ k^* = 2 & \quad \text{if} & \quad \begin{cases} \omega_{12} p_1 < \omega_{21} p_2 + (\omega_{31} - \omega_{32}) p_3 & \text{and} \\ (\omega_{13} - \omega_{12}) p_1 + \omega_{23} p_2 \geq \omega_{32} p_3 \end{cases} \\ k^* = 3 & \quad \text{if} & \quad \begin{cases} \omega_{12} p_1 < \omega_{21} p_2 + (\omega_{31} - \omega_{32}) p_3 & \text{and} \\ (\omega_{13} - \omega_{12}) p_1 + \omega_{23} p_2 < \omega_{32} p_3 \end{cases} \end{aligned}$$

These conditions correspond directly to the defining inequalities in Definition 5 for $r = 3$. To illustrate how the choice of the loss function affects the prediction, consider the following loss matrices, corresponding to loss functions λ_ω that verify the basic assumptions and (CES):

$$W_1 = \begin{pmatrix} 0 & 2 & 4 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{pmatrix}, \quad W_2 = \begin{pmatrix} 0 & 1 & 3 \\ 1 & 0 & 1 \\ 3 & 1 & 0 \end{pmatrix}, \quad W_3 = \begin{pmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 6 & 3 & 0 \end{pmatrix}$$

Matrix W_2 also verifies (DC) with generating function g defined by: $g(0) = 0$, $g(1) = 1$ and $g(2) = 3$. This function is discretely convex since $2 = 3 - 1 = g(2) - g(1) \geq g(1) - g(0) = 1 - 0 = 1$.

Therefore, hypothesis (OCA) holds for any probability distribution p with any of the three loss functions, and the ω -Ord-MAP criterion is well defined. Under these loss functions:

- With matrix W_1 , the predicted class is 1. Indeed,

$$\omega_{12} p_1 = 2 \times 0.4 = 0.8 \geq 0.6 = 1 \times 0.3 + (2 - 1) \times 0.3 = \omega_{21} p_2 + (\omega_{31} - \omega_{32}) p_3$$

- With matrix W_2 , the predicted class is 2 since

$$\begin{aligned}\omega_{12} p_1 &= 1 \times 0.4 = 0.4 < 0.9 = 1 \times 0.3 + (3 - 1) \times 0.3 = \omega_{21} p_2 + (\omega_{31} - \omega_{32}) p_3, \\ (\omega_{13} - \omega_{12}) p_1 + \omega_{23} p_2 &= (3 - 1) \times 0.4 + 1 \times 0.3 = 1.1 \geq 0.3 = 1 \times 0.3 = \omega_{32} p_3.\end{aligned}$$

- With matrix W_3 , which corresponds to the particular case of Exemple 1, the predicted class should be 3, the Bayes-optimal decision obtained minimizing the expected loss. This is indeed the case since

$$\begin{aligned}\omega_{12} p_1 &= 1 \times 0.4 = 0.4 < 1.2 = 1 \times 0.3 + (6 - 3) \times 0.3 = \omega_{21} p_2 + (\omega_{31} - \omega_{32}) p_3, \\ (\omega_{13} - \omega_{12}) p_1 + \omega_{23} p_2 &= (2 - 1) \times 0.4 + 1 \times 0.3 = 0.7 < 0.9 = 3 \times 0.3 = \omega_{32} p_3.\end{aligned}$$

3.2 The Non-symmetric Binary Case

Although symmetric loss functions are common (as W_2 loss matrix in Sect. 3.1), there are situations in which the loss incurred when the true class is j and the predicted is k , ω_{jk} , differs from the loss incurred in the reverse case, ω_{kj} (loss matrices W_1 and W_3 in Sect. 3.1). The proposed ω -Ord-MAP criterion naturally accommodates such asymmetries.

The simplest scenario arises in the binary classification setting, where the two misclassification losses ω_{12} and ω_{21} are not necessarily equal. Under the usual symmetric assumption ($\omega_{12} = \omega_{21}$), the Bayes-optimal decision rule reduces to the standard MAP criterion, which assigns class 1 if $p_1 > 1/2$, and class 2 if $p_1 < 1/2$, with an explicit tie-breaking rule required when $p_1 = 1/2$.

What changes when the misclassification costs are unequal? In this case, the ω -Ord-MAP criterion selects the predicted class according to the rule:

$$\begin{aligned}k^* &= 1 & \text{if } (\omega_{12} - \omega_{11}) p_1 \geq (\omega_{21} - \omega_{22}) p_2 &\iff p_1 \geq \frac{\omega_{21}}{\omega_{12} + \omega_{21}}, \\ k^* &= 2 & \text{if } p_1 < \frac{\omega_{21}}{\omega_{12} + \omega_{21}},\end{aligned}$$

where we use the fact that $p_1 + p_2 = 1$. As a result, the decision threshold is no longer fixed at $1/2$, but instead shifts depending on the relative magnitudes of the misclassification losses. Specifically, the threshold becomes $\omega_{21} / (\omega_{12} + \omega_{21})$, which corresponds to the classic threshold found in cost-sensitive binary classification (see Elkan (2001)). Note that the MAP rule is recovered as a special case when $\omega_{12} = \omega_{21}$. Figure 2 provides a graphical illustration of the decision rule in the asymmetric binary case.

This binary case highlights that the proposed framework subsumes classical cost-sensitive classification as a special case.

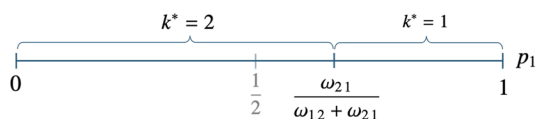


Fig. 2 Graphical representation of the ω -Ord-MAP decision rule in the non-symmetric binary case. In this example, it is assumed that $\omega_{21} \geq \omega_{12}$, which results in a decision threshold greater than $1/2$

4 A Particular Case: The Interval Scale

The *interval scale* is a quantitative measurement scale that enables meaningful comparisons of differences between values typically resulting from the discretization of a continuous variable. It is closely related to classification with label-dependent loss, as both rely on the principle that differences between labels —representing intervals in this context— are meaningful and should influence the penalization of classification errors.

We now introduce some notation, following (Binotto and Delgado, 2025). In the interval scale setting, the values $1, \dots, r$ represent the indices of the class-intervals $I_1, \dots, I_r$, where $I_j = [a_j, b_j)$ with $a_j < b_j$ for $j = 1, \dots, r$ and $a_j = b_{j-1}$ for $j = 2, \dots, r$. Their corresponding lengths are $\ell_j = b_j - a_j > 0$, $j = 1, \dots, r$. We assume the intervals are ordered: that is, for all $x \in I_i$, $y \in I_j$ if $i < j$, then $x < y$. We illustrate notations in Fig. 3, where

$$\delta_{ij} = \begin{cases} 0 & \text{if } j = i + 1, \\ \ell_{i+1} + \dots + \ell_{j-1} & \text{if } j > i + 1. \end{cases}$$

denotes the cumulative distance between intervals I_i and I_j .

Within this interval-scale framework, we consider two natural ways of defining label-dependent misclassification costs. The first one is based on a geometric distance between class intervals, while the second relies on an additive transition-cost interpretation. These two constructions lead to markedly different structural properties of the resulting loss function and, in particular, to different conditions under which the ω -Ord-MAP criterion is well defined and Bayes-optimal.

4.1 Loss Based on the Hausdorff Distance

In the interval scale context, a natural first idea is to define misclassification costs in terms of a geometric distance between the corresponding class intervals. However, standard metrics such as the Hausdorff distance d_H does not in general satisfy the convex escalation structure (CES). Therefore, for Hausdorff-based costs, Bayes-optimality ω -Ord-MAP can only be guaranteed on the subset of probability distributions p for which (OCA) holds, which we characterize below for $r = 3$.

In this subsection we define the loss function by

$$\lambda_\omega(i, j) = \omega_{ij} = \omega_{ji} = d_H(I_i, I_j) = d_H([a_i, b_i), [a_j, b_j)) = \max\{|a_i - a_j|, |b_i - b_j|\},$$

using the Hausdorff distance. Note that this function trivially satisfies the basic assumptions and also the *symmetry* property. To facilitate the analysis, observe that if $j > i$, then:

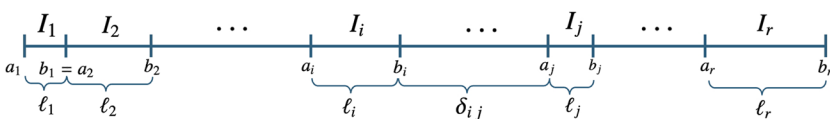


Fig. 3 Visual representation of the intervals $[a_i, b_i)$, their lengths ℓ_i , and the cumulative gaps δ_{ij} between intervals

$$\omega_{ij} = \max\{|a_i - a_j|, |b_i - b_j|\} = \max\{\ell_i + \delta_{ij}, \delta_{ij} + \ell_j\} = \delta_{ij} + \max\{\ell_i, \ell_j\}. \quad (8)$$

4.1.1 The Case $r = 3$

Note that for $r = 3$, $\omega_{12} = \max\{\ell_1, \ell_2\}$, $\omega_{13} = \ell_2 + \max\{\ell_1, \ell_3\}$, $\omega_{23} = \max\{\ell_2, \ell_3\}$. The associated loss matrix is:

$$W = \begin{pmatrix} 0 & \max\{\ell_1, \ell_2\} & \ell_2 + \max\{\ell_1, \ell_3\} \\ \max\{\ell_1, \ell_2\} & 0 & \max\{\ell_2, \ell_3\} \\ \ell_2 + \max\{\ell_1, \ell_3\} & \max\{\ell_2, \ell_3\} & 0 \end{pmatrix}$$

We can check easily that (CES) fails with this matrix, since the successive increments of the loss $\lambda_\omega(j, k)$ are decreasing as k moves away from j . Indeed, for the first row, in order to (CES) be accomplished, it is necessary that

$$\ell_2 + \max\{\ell_1, \ell_3\} - \max\{\ell_1, \ell_2\} \geq \max\{\ell_1, \ell_2\} \iff \ell_2 + \max\{\ell_1, \ell_3\} \geq 2 \max\{\ell_1, \ell_2\} \quad (*)$$

The second row implies no condition, and for the third row, it would be necessary that

$$\ell_2 + \max\{\ell_1, \ell_3\} - \max\{\ell_2, \ell_3\} \geq \max\{\ell_2, \ell_3\} \iff \ell_2 + \max\{\ell_1, \ell_3\} \geq 2 \max\{\ell_2, \ell_3\} \quad (**)$$

It is easy to see that no ordering of lengths ℓ_1 , ℓ_2 and ℓ_3 accomplishes simultaneously (*) and (**), since they together imply that

$$\ell_2 + \max\{\ell_1, \ell_3\} \geq 2 \max\{\ell_1, \ell_2, \ell_3\}$$

Therefore, although we cannot ensure (OCA) for every probability distribution p , it is possible to find the set of distributions for which this assumption holds and, consequently, the ω -Ord-MAP criterion is well defined (and Bayes-optimal with respect to the loss function defined from the Hausdorff distance), and can be written as:

$$\begin{aligned} k^* &= 1 \quad \text{if} \quad \max\{\ell_1, \ell_2\} p_1 \geq \max\{\ell_1, \ell_2\} p_2 + (\ell_2 + \max\{\ell_3, \ell_1\} - \max\{\ell_3, \ell_2\}) p_3. \\ k^* &= 2 \quad \text{if} \quad \begin{cases} \max\{\ell_1, \ell_2\} p_1 < \max\{\ell_1, \ell_2\} p_2 + (\ell_2 + \max\{\ell_3, \ell_1\} - \max\{\ell_3, \ell_2\}) p_3, \\ (\ell_2 + \max\{\ell_1, \ell_3\} - \max\{\ell_1, \ell_2\}) p_1 + \max\{\ell_2, \ell_3\} p_2 \geq \max\{\ell_2, \ell_3\} p_3. \end{cases} \\ k^* &= 3 \quad \text{if} \quad \begin{cases} \max\{\ell_1, \ell_2\} p_1 < \max\{\ell_1, \ell_2\} p_2 + (\ell_2 + \max\{\ell_3, \ell_1\} - \max\{\ell_3, \ell_2\}) p_3, \\ (\ell_2 + \max\{\ell_1, \ell_3\} - \max\{\ell_1, \ell_2\}) p_1 + \max\{\ell_2, \ell_3\} p_2 < \max\{\ell_2, \ell_3\} p_3. \end{cases} \end{aligned}$$

4.1.2 Case: $\ell_1, \ell_3 < \ell_2$

Just to illustrate the procedure, we consider the particular case: $\ell_1, \ell_3 < \ell_2$. In this situation, class 2 has a longer interval than both class 1 and class 3. Then, the ω -Ord-MAP criterion can be re-expressed as

$$\begin{aligned}
 k^* &= 1 & \text{if } p_1 &\geq p_2 + \mu p_3. \\
 k^* &= 2 & \text{if } p_1 < p_2 + \mu p_3 & \text{ and } \mu p_1 + p_2 \geq p_3. \\
 k^* &= 3 & \text{if } p_1 < p_2 + \mu p_3 & \text{ and } \mu p_1 + p_2 < p_3.
 \end{aligned}$$

with $\mu = \frac{\max\{\ell_1, \ell_3\}}{\ell_2} \in (0, 1)$. Equivalently, using that $p_3 = 1 - (p_1 + p_2)$,

$$\begin{aligned}
 k^* &= 1 & \text{if } p_1 &\geq a p_2 + b. \\
 k^* &= 2 & \text{if } p_1 < a p_2 + b & \text{ and } p_1 \geq c p_2 + d. \\
 k^* &= 3 & \text{if } p_1 < a p_2 + b & \text{ and } p_1 < c p_2 + d.
 \end{aligned}$$

with

$$a = \frac{1 - \mu}{1 + \mu}, \quad b = \frac{\mu}{1 + \mu}, \quad c = \frac{-2}{1 + \mu} \quad \text{and} \quad d = \frac{1}{1 + \mu}.$$

Observe that $a + b = d$ and $b < d$.

Figure 4 shows for three values of μ the regions of the simplex $\{(p_1, p_2) \in [0, 1]^2 : p_1 + p_2 \leq 1\}$ in which the ω -Ord-MAP rule predicts classes $k^* = 1, 2$ or 3 when hypothesis (OCA) holds (in colors). The white, black-striped regions correspond to probability distributions (p_1, p_2) in the simplex for which (OCA) fails. In these domains, the defining inequality is not monotone in k , so $\mathcal{C}_\omega(p)$ need not form an upper set; consequently the ω -Ord-MAP criterion is not well defined and Bayes-optimality cannot be guaranteed. Note that this later case occurs when $p_1 \geq a p_2 + b$ and at the same time, $p_1 < c p_2 + d$. The lines $p_1 = a p_2 + b$ and $p_1 = c p_2 + d$ cross at point $(\frac{d-b}{a-c}, \frac{a d - c b}{a-c}) = (\frac{1-\mu}{3-\mu}, \frac{1}{3-\mu})$.

From Fig. 4 we obtain the area of each decision region. The three areas are:

$$\text{Area}(k^* = 1) = \text{Area}(k^* = 3) : \frac{1}{4} \frac{1 + 3\mu - 2\mu^2}{(1 + \mu)(3 - \mu)}, \quad \text{Area}(k^* = 2) : \frac{1}{2} \frac{1}{3 - \mu},$$

which sum up to

$$\frac{1 + 2\mu - \mu^2}{(3 - \mu)(1 + \mu)} < \frac{1}{2} \quad \left(\text{this sum tends to } \frac{1}{2} \text{ if } \mu \rightarrow 1, \text{ and to } \frac{1}{3} \text{ if } \mu \rightarrow 0 \right)$$

The area of the white, black-striped region, which is the difference between $1/2$ and the sum, is:

$$\frac{1}{2} \frac{(1 - \mu)^2}{(1 + \mu)(3 - \mu)},$$

which is a monotonically decreasing function with its maximum and minimum values equal to $1/6$ and 0 for $\mu = 0$ and $\mu = 1$, respectively.

Dividing by the simplex area $1/2$ yields the corresponding probabilities under a uniform distribution over the simplex. These areas are doubled to yield probabilities: when (p_1, p_2) is sampled uniformly from the simplex $\{(p_1, p_2) \in [0, 1]^2 : p_1 + p_2 \leq 1\}$, the probability

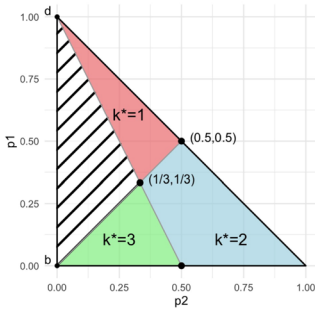
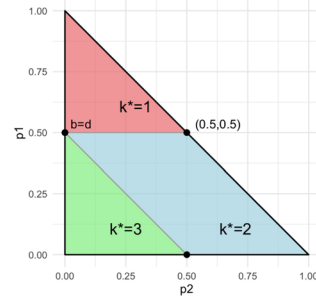
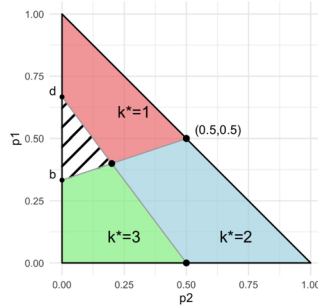
(a) Limit case $\mu \rightarrow 0$.(b) Limit case $\mu \rightarrow 1$, equivalent to Ord-MAP.(c) Intermediate case $\mu = 0.5$.

Fig. 4 Decision regions induced by the ω -Ord-MAP criterion for $r = 3$ under the assumption $\ell_1, \ell_3 < \ell_2$, with $\mu = \frac{\max\{\ell_1, \ell_3\}}{\ell_2} \in (0, 1)$. Each colored zone shows the class predicted, $k^* = 1, 2$, or 3 , as a function of the values of (p_1, p_2) . Striped regions indicate pairs (p_1, p_2) that do not satisfy (OCA)

that the rule outputs class k equals $2 \times \text{Area}(k)$. These probabilities are plotted against μ in Fig. 5, being that corresponding to $k^* = 2$ consistently the highest.

This analysis shows that, although the Hausdorff distance provides a natural geometric notion of discrepancy between class intervals, the induced loss function does not satisfy the Convex Escalation Structure (CES). As a consequence, the order-compatibility assumption may fail for a non-negligible subset of probability distributions, and the ω -Ord-MAP criterion is then only locally well defined.

This limitation motivates the search for an alternative loss construction that preserves the interval-scale interpretation while ensuring a more regular escalation of misclassification costs and, consequently, more robust conditions for order-compatibility. This alternative loss is given below.

4.2 Loss Based on Cumulated Cost of Transition Between Classes

As the predicted class k moves away from the true class j , the Hausdorff distance between the corresponding intervals may vary non-monotonically when interval lengths are heterogeneous. As shown in Sect. 4.1, this lack of monotonicity prevents the resulting loss

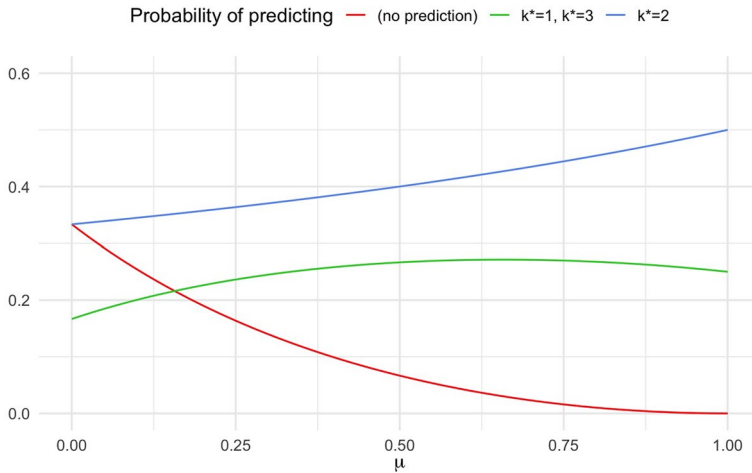


Fig. 5 Probability of each decision by an uninformed classifier using the ω -Ord-MAP criterion as a function of μ for $r = 3$ under the assumption $\ell_1, \ell_3 < \ell_2$, with $\mu = \frac{\max\{\ell_1, \ell_3\}}{\ell_2} \in (0, 1)$

structure from satisfying the Convex Escalation Structure (CES), and consequently the Order-Compatibility Assumption (OCA) may fail for some probability distributions p .

To overcome this limitation, rather than measuring the discrepancy between class intervals through a purely geometric distance, we now adopt a transition-based perspective, in which the cost of a misclassification is interpreted as the **cumulative cost of crossing** the successive interval boundaries separating the true and the predicted classes. Specifically, the cost of misclassifying an observation from interval I_j as belonging to interval I_k is defined as

$$\omega_{jk} = \sum_{m=j}^{k-1} \ell_m \text{ for } j < k, \text{ and } \sum_{m=k+1}^j \ell_m \text{ for } j > k,$$

where ℓ_m represents the cost of crossing the boundary between consecutive intervals I_m and I_{m+1} (or I_{m-1}). Equivalently,

$$\omega_{jk} = \ell_j + \delta_{jk} \text{ for } j < k, \text{ and } \ell_j + \delta_{kj} \text{ for } j > k.$$

Unlike geometric distances between intervals, the transition-cost loss explicitly accounts for the cumulative severity of crossing intermediate classes, which is precisely what the interval-scale discretization represents. This interpretation naturally leads to an additive structure: predicting class $k > j$ requires crossing the boundaries from I_j to I_{j+1} , from I_{j+1} to I_{j+2} , and so on, up to I_k .

This loss function satisfies the basic assumptions introduced in Definition 1, but is generally not symmetric unless all interval lengths coincide. The increments of the loss function are given by

$$\begin{aligned}\omega_{j\ k+1} - \omega_{j\ k} &= \delta_{j\ k+1} - \delta_{j\ k} = \ell_k \text{ if } j < k, \\ \omega_{j\ k} - \omega_{j\ k+1} &= \delta_{k\ j} - \delta_{k+1\ j} = \ell_{k+1} \text{ if } j > k.\end{aligned}$$

Whether (CES) holds under this transition-cost loss depends on the pattern of interval lengths $\ell_1, \dots, \ell_r$. In particular, (CES) is satisfied under certain monotonicity patterns of the lengths, while it may fail when they vary irregularly. On the other hand, by substituting the increments above into Definition 5, the set $\mathcal{C}_\omega$ can be expressed as follows for any probability distribution p :

$$\mathcal{C}_\omega(p) = \left\{ k = 1, \dots, r-1 : \ell_k \sum_{j=1}^k p_j \geq \ell_{k+1} \sum_{j=k+1}^r p_j \right\}.$$

This characterization allows us to determine the subsets of the probability simplex for which (OCA) holds and the ω -Ord-MAP criterion is well defined, as a function of the interval lengths.

4.2.1 The Case $r = 3$

We illustrate this analysis for $r = 3$, where the associated loss matrix is:

$$W = \begin{pmatrix} 0 & \ell_1 & \ell_1 + \ell_2 \\ \ell_2 & 0 & \ell_2 \\ \ell_2 + \ell_3 & \ell_3 & 0 \end{pmatrix}$$

Note that W is symmetric if and only if $\ell_1 = \ell_2 = \ell_3$, which corresponds to the equally spaced ordinal setting. We now analyze several representative configurations of interval lengths for $r = 3$, illustrating how the validity of (CES) and (OCA) depends on their relative magnitudes.

4.2.2 Case (i): $\ell_1, \ell_3 < \ell_2$

For $r = 3$, (CES) reduces to requiring $\ell_2 \geq \ell_1$ (row 1) and $\ell_2 \geq \ell_3$ (row 3), hence it holds whenever $\ell_1, \ell_3 < \ell_2$. As in this configuration (CES) is satisfied, (OCA) holds for any probability distribution p , and the ω -Ord-MAP criterion is globally well defined and Bayes-optimal. In this case, the criterion is expressed as:

$$\begin{aligned}k^* &= 1 & \text{if } \ell_1 p_1 &\geq \ell_2 p_2 + \ell_2 p_3 \\ k^* &= 2 & \text{if } \ell_1 p_1 < \ell_2 p_2 + \ell_2 p_3 & \text{ and } \ell_2 p_1 + \ell_2 p_2 \geq \ell_3 p_3 \\ k^* &= 3 & \text{if } \ell_1 p_1 < \ell_2 p_2 + \ell_2 p_3 & \text{ and } \ell_2 p_1 + \ell_2 p_2 < \ell_3 p_3,\end{aligned}$$

which using notations $\tau_1 = \frac{\ell_1}{\ell_2} < 1$ and $\tau_2 = \frac{\ell_2}{\ell_3} > 1$ is equivalent to

$$\begin{aligned}
 k^* &= 1 & \text{if } p_1 &\geq \frac{1}{1+\tau_1} & \left(> \frac{1}{2} \right) \\
 k^* &= 2 & \text{if } p_1 < \frac{1}{1+\tau_1} & \text{ and } p_1 + p_2 \geq \frac{1}{1+\tau_2} & \left(< \frac{1}{2} \right) \\
 k^* &= 3 & \text{if } p_1 < \frac{1}{1+\tau_1} & \text{ and } p_1 + p_2 < \frac{1}{1+\tau_2} .
 \end{aligned}$$

We use the notations $b = \frac{1}{1+\tau_1}$ and $d = \frac{1}{1+\tau_2}$ in Fig. 6 to denote some notable points. This figure shows the regions of the simplex $\{(p_1, p_2) \in [0, 1]^2 : p_1 + p_2 \leq 1\}$ in which the ω -Ord-MAP rule predicts classes $k^* = 1, 2$ or 3 , in two different scenarios for which $\ell_1, \ell_3 < \ell_2$.

4.2.3 Case (ii): $\ell_1, \ell_2 < \ell_3$

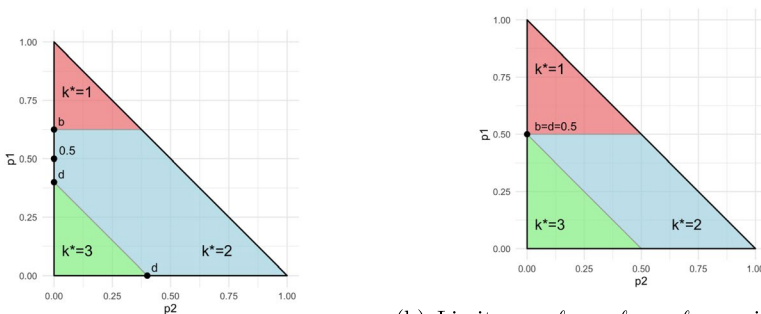
In contrast to Case (i), (CES) fails in this configuration. As a result, (OCA) is no longer guaranteed globally, but may still hold for specific regions of the probability simplex. The ω -Ord-MAP criterion, in this case, is expressed as in Case (i) but now τ_1 is not bounded, while $\tau_2 < 1$, except for pairs (p_1, p_2) for which

$$p_1 \geq \frac{1}{1+\tau_1} \quad \text{and} \quad p_1 + p_2 < \frac{1}{1+\tau_2}$$

as shown in Fig. 7 (a). Outside the stripped region, $\mathcal{C}_\omega(p)$ is an upper set and ω -Ord-MAP is well defined. This illustrates that, even when (CES) fails, the ω -Ord-MAP criterion can remain locally well defined and Bayes-optimal for subsets of probability distributions p .

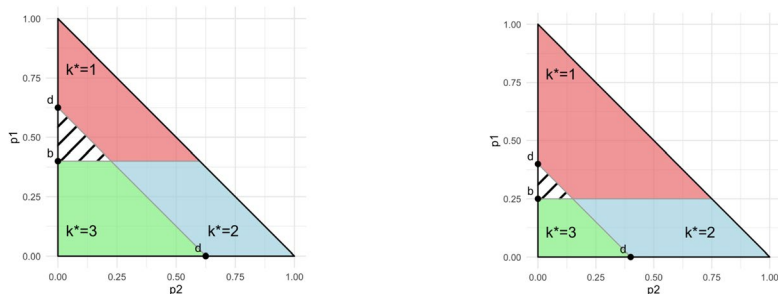
4.2.4 Case (iii): $\ell_2, \ell_3 < \ell_1$

In this case, as in Case (ii), in contrast to Case (i), assumption (CES) is not satisfied either, and although (OCA) does not hold globally, it may still hold locally for specific probability distributions. Figure 7 (b) shows the pairs (p_1, p_2) for which (OCA) holds and the ω -Ord-



(a) Case (i) $\ell_1/\ell_2 = 3/5$ and $\ell_2/\ell_3 = 3/2$. (b) Limit case $\ell_1 = \ell_2 = \ell_3$, equivalent to Ord-MAP.

Fig. 6 Case (i): decision regions induced by the ω -Ord-MAP criterion for $r = 3$ under the assumption $\ell_1, \ell_3 < \ell_2$ with the loss of the cumulated cost of transition. Each colored zone shows the class predicted, $k^* = 1, 2$, or 3 , as a function of the values of (p_1, p_2)



(a) Case (ii): $\ell_1/\ell_2 = 3/2$ and $\ell_2/\ell_3 = 3/5$. (b) Case (iii): $\ell_1/\ell_2 = 3$ and $\ell_2/\ell_3 = 3/2$.

Fig. 7 Cases (ii) and (iii): decision regions induced by the ω -Ord-MAP criterion for $r = 3$ under different scenarios, with the loss of the cumulated cost of transition. Each colored zone shows the class predicted, $k^* = 1, 2$, or 3 , as a function of the values of (p_1, p_2) . Striped regions indicate pairs (p_1, p_2) that do not satisfy (OCA)

MAP criterion is expressed as in the previous cases but with $\tau_1 > 1$ and τ_2 not bounded (in different colors indicating the resulting chosen class, and in striped white, pairs not satisfying (OCA)). Outside the stripped region, $\mathcal{C}_\omega(p)$ is an upper set and ω -Ord-MAP is well defined.

This section shows that interval-scale classification naturally leads to label-dependent loss functions whose structural properties critically determine the well-posedness of the ω -Ord-MAP criterion. While geometric distances such as the Hausdorff metric may violate global convex escalation, alternative constructions based on cumulated transition costs restore the required monotonicity, either globally or locally. These findings highlight the importance of explicitly modeling misclassification costs and motivate the empirical evaluation of ω -Ord-MAP under different loss specifications, which we address in the following section.

5 Experimental Study

This section evaluates the empirical impact of replacing the Ord-MAP decision rule (Delgado, 2025) with the proposed cost-sensitive ω -Ord-MAP criterion. Two complementary experiments are conducted:

- Experiment A (Sect. 5.2): a realistic interval-scale scenario based on facial age prediction using random forests.
- Experiment B (Sect. 5.3): a fully ordinal but non-interval scenario based on the Wine Quality dataset and cumulative link models (CLMs).

These two studies serve different purposes. Experiment A illustrates the natural setting where interval-scale distances directly induce a label-dependent loss. Experiment B demonstrates that the advantages of the ω -Ord-MAP decision rule persist in a different dataset and classifier family, even when the classes are not formed from contiguous numerical intervals.

Before describing the two experiments, we present the evaluation framework common to both.

5.1 Experimental Framework and Evaluation Methodology

We compare Ord-MAP and ω -Ord-MAP conditional on the same predictive probability distribution p . Thus, any performance difference arises solely from the decision rule and the underlying loss function, not from the classifier.

Performance is assessed using the label-dependent ω -MAE metric (Equation (6)), induced by the loss matrix used in each experiment. When tuning in Experiment A (Sect. 5.2), we use as error functions the error rate (equivalently, accuracy), the mean absolute error (MAE) and the ω -MAE metric. Both MAE and ω -MAE are non-negative but unbounded. To allow comparisons across folds and tuning settings, we follow (Binotto and Delgado, 2025) and use standardized versions of these metrics. Given a confusion matrix and a loss function, the metric is divided by its maximum possible value, which corresponds to the worst-case confusion matrix: one that assigns every instance of each true class to the predicted class incurring the largest loss. This yields standardized metrics in $[0, 1]$, with larger values indicating worse performance while preserving the ordinal and cost-sensitive structure of the measure.

For statistical comparison, we compute the standardized ω -MAE metric differences across folds and conduct paired one-sided tests (t-test and Wilcoxon signed-rank test) to assess whether

$$\text{Standardized } \omega\text{-MAE (using } \omega\text{-Ord-MAP)} < \text{Standardized } \omega\text{-MAE (using Ord-MAP)}.$$

Two-sided tests provide no additional interpretability here, as the scientific question concerns whether the cost-sensitive rule improves performance. Statistical significance is reported using the following notation: • for 10%, * for 5%, ** for 1% and *** for 1%₀₀. Box-plots of pairwise differences help visualize robustness across folds and tuning conditions (Experiment A) or model specifications (Experiment B).

5.2 Experiment A: Facial Age Dataset

5.2.1 Dataset and Interval-Based Loss

Experiment A uses a real-world **facial age** dataset from Kaggle,¹ where the goal is to classify individuals into age intervals from image-derived features. The original dataset contains 9,673 facial images labeled by age (0-98). Each image is resized to 32×32 pixels, converted to grayscale, and flattened into 1,024 predictors.

To balance computational feasibility with statistical precision in model tuning, we restrict the analysis to individuals aged ≤ 25 years, roughly the median of the dataset. Age is discretized into the four heterogeneous intervals $[0, 2)$, $[2, 11)$, $[11, 20)$, $[20, 25]$, which induce a non-equidistant ordinal target. The final sample consists of 4,945 observations distributed as: 1,105/ 1,963/ 1,067/810.

In this experiment, the loss function is constructed using the cumulated cost of transition between adjacent age intervals (Sect. 4.2), where $\omega_{j,k}$ equals the cumulative length

¹ <https://www.kaggle.com/datasets/frabbiw/facial-age>.

of the intervals crossed when moving from interval I_j to interval I_k . For interval lengths $\ell_1 = 2$, $\ell_2 = 9$, $\ell_3 = 9$, $\ell_4 = 5$, the loss matrix is:

$$W = \begin{pmatrix} 0 & \ell_1 & \ell_1 + \ell_2 & \ell_1 + \ell_2 + \ell_3 \\ \ell_2 & 0 & \ell_2 & \ell_2 + \ell_3 \\ \ell_2 + \ell_3 & \ell_3 & 0 & \ell_3 \\ \ell_2 + \ell_3 + \ell_4 & \ell_3 + \ell_4 & \ell_4 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 2 & 11 & 20 \\ 9 & 0 & 9 & 18 \\ 18 & 9 & 0 & 9 \\ 23 & 14 & 5 & 0 \end{pmatrix}.$$

This loss satisfies the Convex Escalation Structure (CES) because incremental penalties increase as one moves away from the true class. Consequently, the Order-Compatibility Assumption (OCA) holds for all predictive probability distributions p (Proposition 2), ensuring that the ω -Ord-MAP decision rule is well defined and Bayes-optimal under this loss.

5.2.2 Model Training, Tuning and Validation

Probabilistic predictions are obtained using random forest classifiers trained with the `train` function from the `caret` package². We tune the `mtry` hyper-parameter, which controls the number of features randomly sampled at each split. A small ensemble of three trees is used to reduce computational cost. Hyper-parameter tuning employs 3-fold cross-validation with 10 random search iterations, optimizing standardized MAE or standardized ω -MAE, depending on the setting, as custom performance metrics (accuracy is the default). This allows the tuning procedure to be aligned with ordinal and cost-sensitive loss structures.

Final performance is evaluated via 10-fold cross-validation. In each fold, we:

- 1) train and tune the probabilistic classifier using the error rate, the standardized MAE, and the standardized ω -MAE as custom performance metrics (error functions);
- 2) obtain predicted probability distributions on the held-out fold;
- 3) convert them into class predictions using Ord-MAP and ω -Ord-MAP rules;
- 4) compute the standardized ω -MAE for each rule. This yields paired observations across folds for statistical testing.

5.2.3 Results

Table 1 reports the p-values of one-sided paired tests comparing the standardized ω -MAE metric under both decision rules, across the three tuning metrics.

From Table 1 we see that across all tuning criteria, the ω -Ord-MAP rule significantly outperforms Ord-MAP. This improvement is systematic and robust, as illustrated in Fig. 8 through boxplots of paired metric differences, where boxes lying above zero indicate improved performance under the cost-sensitive rule.

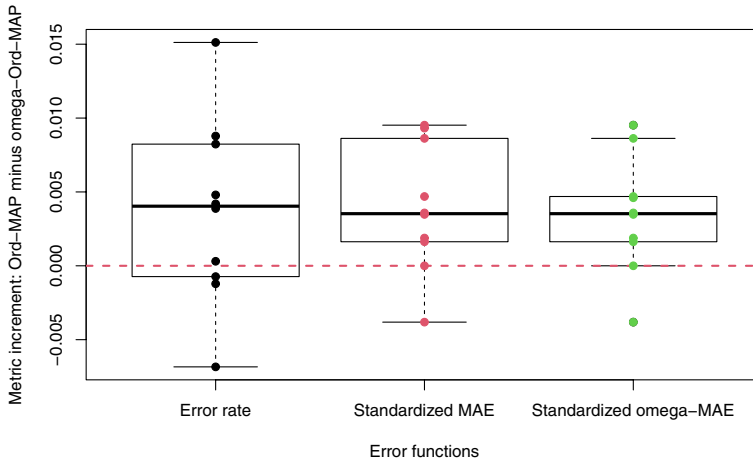
5.2.4 Discussion

The experiment demonstrates that incorporating interval-scale loss at the decision stage yields consistent and statistically significant gains, even when using simple probabilistic

²Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software, Vol. 28(5), pp. 1–26. <https://www.jstatsoft.org/index.php/jss/article/view/v028i05>.

Table 1 Experiment A: p-values comparing standardized ω -MAE for Ord-MAP vs. ω -Ord-MAP using one-sided paired Wilcoxon and t-tests

Statistical test	Error function		
	Error rate	Standardized MAE	Standardized ω -MAE
Wilcoxon	0.065430 [•]	0.022005*	0.022005*
t-test	0.047416*	0.009598**	0.010841*

**Fig. 8** Experiment A: boxplots of paired differences in standardized ω -MAE between Ord-MAP and ω -Ord-MAP. Positive values indicate better performance of ω -Ord-MAP

models such as small random forests. These improvements arise solely from the decision rule, confirming the practical relevance of the proposed ω -Ord-MAP criterion.

5.3 Experiment B: Wine Quality Dataset

5.3.1 Dataset and Asymmetric DC-Based Loss

Experiment B considers the **wine quality** dataset from the UCI repository³ (red wine subset), with 1,599 observations and 11 physicochemical predictors. The target quality score $q \in \{3, \dots, 8\}$ is collapsed into three ordered classes: *low* ($q \leq 5$), *medium* ($q = 6$), and *high* ($q \geq 7$). To induce asymmetric, label-dependent costs while ensuring global well-definedness of the ω -Ord-MAP decision rule, we use a distance-based loss with two different discretely convex functions for underestimation and overestimation:

$$\omega_{jk} = \begin{cases} g_L(j - k) & \text{if } j > k \\ g_R(j - k) & \text{if } j < k \end{cases} \quad g_L(d) = d + \alpha_L d^2, \quad g_R(d) = d + \alpha_R d^2, \quad \alpha_R > \alpha_L.$$

³<https://archive.ics.uci.edu/dataset/186/wine+quality>.

Choosing $\alpha_R > \alpha_L$ reflects that overestimating wine quality may be more costly (e.g., consumer dissatisfaction or reputational damage) than underestimating it. For $\alpha_L = 0.5$ and $\alpha_R = 2$, $g_L(1) = 1.5$, $g_L(2) = 4$, $g_R(1) = 3$, $g_R(2) = 10$, and the loss matrix becomes:

$$W = \begin{pmatrix} 0 & g_R(1) & g_R(2) \\ g_L(1) & 0 & g_R(1) \\ g_L(2) & g_L(1) & 0 \end{pmatrix} = \begin{pmatrix} 0 & 3 & 10 \\ 1.5 & 0 & 3 \\ 4 & 1.5 & 0 \end{pmatrix}.$$

The functions g_L and g_R satisfy discrete convexity (DC): $g(d) = d + \alpha d^2$ has increasing increments: $g(1) - g(0) = 1 + \alpha$, $g(2) - g(1) = 1 + 3\alpha$ and $1 + 3\alpha \geq 1 + \alpha$ if $\alpha \geq 0$. Hence the resulting loss satisfies (CES) (Remark 1), and therefore (OCA) holds globally (Proposition 2)). Consequently, the ω -Ord-MAP rule is well defined and Bayes-optimal for all predictive distributions.

5.3.2 Probabilistic Modeling Via Cumulative Link Models

To complement the tree-based models of Experiment A (Sect. 5.2), we use cumulative link models (CLMs) fitted using the `clm` function from the `ordinal` R package⁴ CLMs yield a predictive probability vector $p = (p_1, p_2, p_3)$, which is then converted into a class label using either Ord-MAP or ω -Ord-MAP.

As model specifications, we explore the 20 combinations of:

- five links: logistic, probit, loglog, complementary log-log (cloglog), Cauchy (cauchit);
- four threshold structures: flexible (no constraints on the threshold spacing), equidistant (equal spacing between thresholds), symmetric (thresholds symmetrically distributed around a central value), symmetric2 (similar to the previous one but with the latent mean of the reference group fixed at zero).

As expected in moderate-size ordinal datasets, many link–threshold combinations yield nearly identical predictive distributions, reflecting a well-known stability property of CLM estimators.

5.3.3 Results

Standardized ω -MAE values under both decision rules are compared using paired one-sided tests. Table 2 summarizes the results across all model specifications given by link–threshold structure combinations.

Figure 9 shows representative boxplots of metric differences.

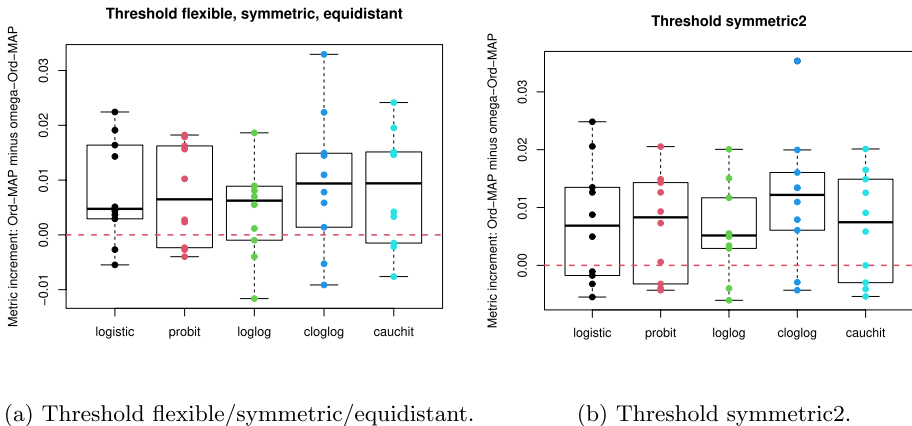
5.3.4 Discussion

Although different link and threshold specifications yielded nearly identical predictive distributions—a well-known stability property of CLM estimators—the ω -Ord-MAP rule con-

⁴Christensen, R.H.B. (2019). Ordinal – Regression Models for Ordinal Data. R package. <https://CRAN.R-project.org/package=ordinal>.

Table 2 Experiment B: p-values comparing standardized ω -MAE for Ord-MAP vs. ω -Ord-MAP using one-sided paired Wilcoxon and t-tests

Threshold	Test	Link function				
		logistic	probit	loglog	cloglog	cauchit
flexible/symmetric/ equidistant	Wilcoxon	0.01855*	0.04199*	0.08008 [•]	0.01855*	0.02441*
	t-test	0.01254*	0.01522*	0.07531 [•]	0.01915*	0.01612*
symmetric2	Wilcoxon	0.05273 [•]	0.03223*	0.04199*	0.00488**	0.02901*
	t-test	0.02635*	0.02026*	0.02227*	0.00527**	0.02540*

**Fig. 9** Experiment B: boxplots of paired differences in standardized ω -MAE between Ord-MAP and ω -Ord-MAP. Positive values indicate better performance of ω -Ord-MAP

sistently outperformed Ord-MAP, confirming that the improvements arise from the decision mechanism rather than the base model.

Comparing with Experiment A, this confirms that:

- gains originate from the decision rule, not the model architecture;
- improvements extend to fully ordinal, non-interval settings;
- the method generalizes across classifier families (tree-based vs. parametric).

This second experiment is not meant to replicate the interval-scale structure but to demonstrate that the ω -Ord-MAP decision rule provides systematic improvements also in fully ordinal, non-interval-scale scenarios, and across classifier families (tree-based RF vs parametric CLM). This experiment thus strengthens the external validity of the proposed approach.

A sensitivity analysis with respect to the asymmetry parameter α_R (with α_L fixed) was conducted to further characterize the behavior of the decision rule; the results, reported in Appendix C.1, are fully consistent with the theoretical role of cost asymmetry and discrete convexity.

Additional diagnostic rates, including disagreement rate between Ord-MAP and ω -Ord-MAP decisions, as well as the frequency of large-distance errors, are reported in Appendix C.2 to further illustrate how ω -Ord-MAP redistributes errors under an asymmetric loss.

6 Conclusions

Recent work by Delgado (2025) addresses the limitations of the traditional Maximum A Posteriori (MAP) criterion in ordinal classification, showing that its direct extension from the nominal setting often leads to suboptimal decisions due to its disregard for class ordering. To overcome this limitation, Delgado (2025) proposes **Ord-MAP**, a decision rule that selects the first class whose cumulative probability exceeds 0.5 –i.e., the **median**– providing a consistent and theoretically grounded alternative when class labels are equally spaced. That contribution implicitly highlights the interplay among three fundamental components of classification: the scoring rule, the loss function quantifying misclassification severity, and the resulting Bayes-optimal decision rule. However, the analysis is restricted to symmetric loss scenarios, where misclassification penalties depend only on class distance and not on the specific labels involved.

In this paper, we extend this framework to the more general and practically relevant setting in which class labels are ordered but not necessarily equidistant. This allows the severity of classification errors to vary across class transitions. Building on the foundation laid by Delgado (2025), we generalize the aforementioned triad –scoring rule, loss function, and decision criterion– to incorporate label-dependent (cost-sensitive) misclassification, thereby introducing formal versions of each component that accommodate asymmetric and heterogeneous penalties.

This generalization yields a unified decision-theoretic framework for designing proper scoring rules and Bayes-optimal decision criteria under label-dependent misclassification costs. By explicitly characterizing the structural conditions under which the proposed ω -**Ord-MAP** criterion is well defined and optimal, the approach places cost-sensitive ordinal classification on solid theoretical foundations, while retaining flexibility in the modeling of class relationships. In particular, it naturally applies to interval-scale settings, where class distances are meaningful but non-uniform.

To illustrate the expressiveness and applicability of the framework, we examine two representative cases in detail. First, we study the binary asymmetric setting, showing how our formulation recovers and extends classical cost-sensitive decision rules. Second, we focus on the interval-scale case, exemplified by three ordered intervals of unequal width, where the framework enables principled handling of non-equidistant labels and label-dependent penalties, both globally and locally in the space of probability distributions.

We further assess the practical relevance of the proposed decision rule through empirical studies comparing Ord-MAP and ω -Ord-MAP on real-world datasets, using probabilistic predictions obtained from different classifier families. The results show statistically significant improvements in favor of ω -Ord-MAP under the standardized ω -Ord-MAP, the label-dependent evaluation metric, supporting the practical usefulness of the proposed generalization in the considered experimental settings.

It is important to emphasize that ω -Ord-MAP is not designed to minimize ordinal distance per se, but rather to minimize expected loss under a specified cost structure, as dictated by statistical decision theory. Consequently, differences in distance-based error profiles between Ord-MAP and ω -Ord-MAP should be interpreted as reflecting different decision objectives rather than differences in inferential quality. The proposed rule is optimal with respect to the chosen loss, not universally superior under all evaluation criteria.

This work opens several avenues for future research. A particularly promising direction is the extension of the ω -Ord-MAP criterion to more complex output spaces, such as structured or hierarchical ordinal settings, where misclassification costs may depend both on ordinal distance and on position within the hierarchy. Another relevant direction concerns robustness with respect to probability calibration quality. While ω -Ord-MAP operates purely at the decision stage and does not alter the posterior distribution provided by the classifier, the interaction between cost-sensitive decision rules and calibration methods deserves further investigation and depends strongly on the learning algorithm employed.

The Properness of the score S_ω

Theorem 2 *For any loss function λ_ω , the score S_ω is **proper**.*

Proof To prove this result, as for proving Theorem 2 (Appendix A) in Delgado (2025), we use Theorem 1 Gneiting and Raftery (2007) with adapted notations, which guarantees that a score $\check{S}$ is proper if and only if there exists a convex, real-valued function G such that for all $k \in \{1, \dots, r\}$ and probability distribution p over $\{1, \dots, r\}$, the following holds:

$$\check{S}(p, k) = G(p) - \langle G'(p), p \rangle + G'_k(p), \quad (\text{A1})$$

where $G'(p)$ denotes the sub-gradient of G at point p . We will apply this result with $\check{S} = S_\omega$ by determining the suitable function G .

Note that by (A1),

$$\begin{aligned} \check{S}(p | p) &= \sum_{k=1}^r \check{S}(p, k) p_k = \sum_{k=1}^r (G(p) - \langle G'(p), p \rangle + G'_k(p)) p_k \\ &= (G(p) - \langle G'(p), p \rangle) \sum_{k=1}^r p_k + \sum_{k=1}^r G'_k(p) p_k \\ &= (G(p) - \langle G'(p), p \rangle) + \langle G'(p), p \rangle = G(p). \end{aligned}$$

Therefore, our candidate is

$$G(p) = \check{S}(p | p) = \sum_{k=1}^r \check{S}(p, k) p_k = \sum_{k=1}^r S_\omega(p, k) p_k = - \sum_{k=1}^r \left(\sum_{j \neq k} \omega_{j,k} p_j \right) p_k. \quad (\text{A2})$$

The rest of the proof is divided into two steps: verifying that (A1) holds, and showing that G defined by (A2) is convex. Then, a direct application of Theorem 1 Gneiting and Raftery (2007) completes the proof.

Step 1: To verify that (A1) holds, we observe that for any k , $G'_k(p) = - \sum_{j \neq k} \omega_{j,k} p_j$ and then,

$$\langle G'(p), p \rangle = - \sum_{k=1}^r \left(\sum_{j \neq k} \omega_{jk} p_j \right) p_k = G(p).$$

Therefore,

$$G(p) - \langle G'(p), p \rangle + G'_k(p) = G(p) - G(p) - \sum_{j \neq k} \omega_{jk} p_j = - \sum_{j \neq k} \omega_{jk} p_j = S_\omega(p, k),$$

which proves (A1) for $\check{S} = S_\omega$.

Step 2: Finally, we prove that function G is convex. Let $\alpha \in (0, 1)$ and p, q , probability distributions. Then, $(1 - \alpha)p + \alpha q$ is also a probability distribution, and

$$\begin{aligned} G((1 - \alpha)p + \alpha q) &= - \sum_{k=1}^r \left(\sum_{j \neq k} \omega_{jk} ((1 - \alpha)p_j + \alpha q_j) \right) ((1 - \alpha)p_k + \alpha q_k) \\ &= -(1 - \alpha) \sum_{k=1}^r \left(\sum_{j \neq k} \omega_{jk} ((1 - \alpha)p_j + \alpha q_j) \right) p_k - \alpha \sum_{k=1}^r \left(\sum_{j \neq k} \omega_{jk} ((1 - \alpha)p_j + \alpha q_j) \right) q_k. \end{aligned} \quad (\text{A3})$$

On the other hand,

$$\sum_{j \neq k} \omega_{jk} ((1 - \alpha)p_j + \alpha q_j) = (1 - \alpha) \sum_{j \neq k} \omega_{jk} p_j + \alpha \sum_{j \neq k} \omega_{jk} q_j$$

and then, by replacing this expression into (A3), we obtain that

$$\begin{aligned} G((1 - \alpha)p + \alpha q) &= -(1 - \alpha) \left((1 - \alpha) \sum_{k=1}^r \left(\sum_{j \neq k} \omega_{jk} p_j \right) p_k + \alpha \sum_{k=1}^r \left(\sum_{j \neq k} \omega_{jk} q_j \right) p_k \right) \\ &\quad - \alpha \left((1 - \alpha) \sum_{k=1}^r \left(\sum_{j \neq k} \omega_{jk} p_j \right) q_k + \alpha \sum_{k=1}^r \left(\sum_{j \neq k} \omega_{jk} q_j \right) q_k \right) \\ &= (1 - \alpha)^2 G(p) + \alpha^2 G(q) - \alpha(1 - \alpha) \left(\sum_{k=1}^r \left(\sum_{j \neq k} \omega_{jk} q_j \right) p_k + \sum_{k=1}^r \left(\sum_{j \neq k} \omega_{jk} p_j \right) q_k \right). \end{aligned} \quad (\text{A4})$$

Noting that

$$\sum_{k=1}^r \left(\sum_{j \neq k} \omega_{jk} q_j \right) p_k = \sum_{k=1}^r \left(\sum_{j \neq k} \omega_{jk} p_j \right) q_k \geq 0,$$

we finally obtain by (A4) that

$$\begin{aligned}
 G((1-\alpha)p + \alpha q) &= (1-\alpha)^2 G(p) + \alpha^2 G(q) - 2\alpha(1-\alpha) \sum_{k=1}^r \left(\sum_{j \neq k} \omega_{jk} q_j \right) p_k \\
 &\leq (1-\alpha)^2 G(p) + \alpha^2 G(q) \leq (1-\alpha) G(p) + \alpha G(q). \quad \square
 \end{aligned}$$

□

The optimality of the ω -Ord-MAP criterion

In this appendix we prove Theorem 1. This is the main theoretical result of the paper and establishes that the ω -Ord-MAP decision criterion is Bayes-optimal with respect to the loss function λ_ω , since it minimizes the expected loss R_ω (equivalently, it maximizes the associated score S_ω).

Its proof relies on Lemma 1, previously introduced in Delgado (2025), which we recall here for completeness.

Lemma 1 [Lemma 1 Delgado (2025)] *Given fixed a probability distribution $p = (p_1, \dots, p_r)$, let $\varphi(p, s)$ be a non-positive function of $s = 1, \dots, r$ depending on p . For any $k = 1, \dots, r-1$ and $\delta = 1, \dots, r-k$, define*

$$\Delta(p, k, \delta) = \varphi(p, k + \delta) - \varphi(p, k).$$

Assume that there exists $h = 1, \dots, r$ such that the following hypotheses hold:

- $h_1)$ If $h < r$, $\Delta(p, h, \delta) \leq 0$ for all $\delta = 1, \dots, r-h$.
- $h_2)$ If $h > 1$, $\Delta(p, \ell, 1) > 0$ for all $\ell = 1, \dots, h-1$.

Therefore,

$$\arg \max_{\ell=1, \dots, r} \varphi(p, \ell) = h.$$

(With the convention that we take the minimum of the points where the maximum is achieved, if there are multiple such points.)

Theorem 1 *Let $p = (p_1, \dots, p_r)$ be a probability distribution, and let λ_ω be a loss function satisfying the basic assumptions and (OCA). Then, the class k^* selected by the ω -Ord-MAP criterion (as given in Definition 5) satisfies:*

$$k^* = \arg \max_{k=1, \dots, r} S_\omega(p, k) = \arg \min_{k=1, \dots, r} R_\omega(p, k),$$

i.e., it maximizes the score and minimizes the expected loss.

Proof of Theorem 1 From Proposition 1, we know that maximizing S_ω is equivalent to minimizing R_ω . It is therefore enough to show that k^* maximizes the score S_ω . To do this, we apply Lemma 1, with φ equal to S_ω , which is always non-positive by construction. Fix $h = 1, \dots, r-1$ and $\delta = 1, \dots, r-h$. Then, we compute the difference:

$$\begin{aligned}\Delta(p, h, \delta) &= \varphi(p, h + \delta) - \varphi(p, h) = S_\omega(p, h + \delta) - S_\omega(p, h) \\ &= - \sum_{i=h}^{h+\delta-1} \left(\sum_{j=1}^i (\omega_{j,i+1} - \omega_{j,i}) p_j \right) + \sum_{i=h}^{h+\delta-1} \left(\sum_{j=i+1}^r (\omega_{j,i} - \omega_{j,i+1}) p_j \right).\end{aligned}$$

By the construction of the ω -Ord-MAP criterion, taking $h = k^*$, since $i \geq h$, this expression is non-positive, so condition h_1) of Lemma 1 is satisfied.

To verify condition h_2), we consider $h = k^*$, if $h > 1$, and $\ell = 1, \dots, h - 1$. For each such ℓ ,

$$\Delta(p, \ell, 1) = - \sum_{j=1}^{\ell} (\omega_{j,\ell+1} - \omega_{j,\ell}) p_j + \sum_{j=\ell+1}^r (\omega_{j,\ell} - \omega_{j,\ell+1}) p_j.$$

Again by the definition of k^* , this expression is strictly positive for all $\ell < k^*$, satisfying condition h_2). Therefore, by Lemma 1, the score S_ω is maximized at k^* , completing the proof. $\square$

Sensitivity to Asymmetry and Diagnostic Error Analysis

Sensitivity Analysis with Respect to α_L and α_R

We perform a sensitivity analysis in Experiment B (Sect. 5.3) to assess how the advantage of ω -Ord-MAP varies with the degree of cost asymmetry, and to evaluate the robustness of the empirical findings with respect to the choice of the asymmetry parameters in the loss function. Specifically, we fixed $\alpha_L = 2$ and varied $\alpha_R \in \{0.5, 1, 2, \dots, 10\}$, thereby moving from a nearly symmetric loss ($\alpha_L \approx \alpha_R$) to increasingly asymmetric scenarios in which overestimating wine quality becomes progressively more costly than underestimating it ($\alpha_R > \alpha_L$) or vice versa ($\alpha_R < \alpha_L$). For each value of α_R , we recomputed the label-dependent loss matrix, the corresponding standardization of ω -MAE, and the decisions induced by the ω -Ord-MAP rule, but maintaining the same fitted CLM models and predictive probability distributions.

To further characterize the role of loss asymmetry, we re-parametrized the quadratic loss in terms of the asymmetry ratio $\rho_R = \alpha_R/\alpha_L$, fixing $\alpha_L (=2)$. When $\rho_R = 1$, the loss is symmetric and quadratic, and both Ord-MAP and ω -Ord-MAP yield nearly identical decisions. As ρ_R departs from unity, in either sense, the advantage of the cost-sensitive ω -Ord-MAP rule becomes increasingly pronounced. This behavior is fully consistent with the theoretical motivation of the proposed rule, which is able to exploit label-dependent cost asymmetries.

Figure 10 reports the sensitivity analysis with respect to the asymmetry ratio ρ_R , keeping $\alpha_L = 2$ fixed. Boxplots with superimposed fold-level differences illustrate both central tendency and variability across cross-validation splits, and show that the improvement of the ω -Ord-MAP criterion is stable across cross-validation folds and increases monotonically with the severity of overestimating/underestimating costs.

As we can observe in Fig. 10, when $\rho_R \approx 1$ the performance difference between Ord-MAP and ω -Ord-MAP is negligible, reflecting the fact that in this experiment, under an almost symmetric distance-based loss the two rules behave similarly. As ρ_R increases/decreases, the average improvement in standardized ω -MAE achieved by ω -Ord-MAP over Ord-MAP grows monotonically. That is, increasing asymmetry leads to a systematic improvement of the cost-sensitive decision rule. The proportion of test cases in which ω -Ord-MAP yields a lower loss than Ord-MAP as a function of the asymmetry ratio ρ_R is illustrated in Fig. 11.

Fig. 11 shows that the advantage of ω -Ord-MAP vanishes near the symmetric case $\rho_R \approx 1$ and increases rapidly as the loss function becomes more asymmetric, confirming that the proposed rule is especially beneficial under non-linear and asymmetric cost structures. These results of this experiment indicate that the benefit of ω -Ord-MAP is driven by genuinely asymmetric or strongly non-linear loss structures. This behavior is partly attributable to the low number of ordinal classes considered in this experiment and to the concentration of predictive mass around the central category, a common feature of well-calibrated ordinal models with three classes. In such settings, symmetric loss structures naturally favor median-based decisions, leaving limited room for improvement. That is, potential gains from incorporating non-linear but symmetric costs may remain empirically small in this setting, despite being theoretically distinct.

This analysis isolates the effect of asymmetry in the loss function, disentangling it from non-linearity alone. It confirms that the proposed decision rule not only remains stable under moderate changes in the loss parameters, but it also exploits stronger asymmetries in a systematic way, yielding larger gains when the misclassification costs for overestimation are more pronounced.

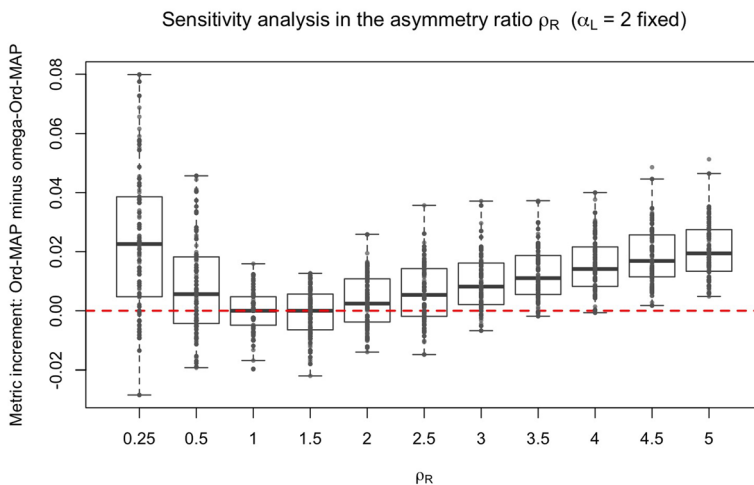


Fig. 10 Sensitivity analysis with respect to ρ_R . Boxplots for fold-level differences in standardized ω -MAE between Ord-MAP and ω -Ord-MAP. Superimposed points correspond to individual cross-validation folds

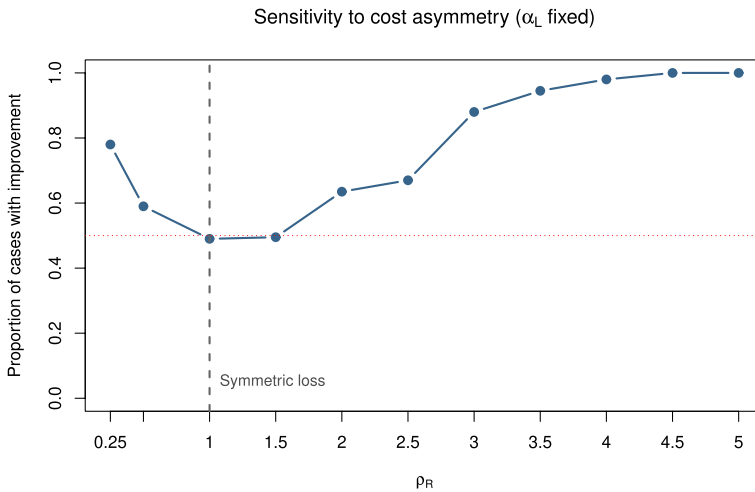


Fig. 11 Sensitivity analysis with respect to ρ_R . Proportion of test instances for which ω -Ord-MAP outperforms Ord-MAP as a function of ρ_R

Diagnostic Analysis of Decision Differences

While Appendix C.1 focuses on aggregate performance differences, this diagnostic analysis examines how asymmetry translates into qualitative changes in the decision behavior of the classifiers. For that, we focus on a representative configuration, $\alpha_L = 0.5$ and $\alpha_R = 2$ (used in Sect. 5.3).

Table 3 reports auxiliary decision-level diagnostics comparing Ord-MAP and ω -Ord-MAP decision rules under the asymmetric loss configuration given by parameters $\alpha_L = 0.5$ and $\alpha_R = 2$, providing insight into how the two decision criteria differ in terms of their induced classification behavior. In the present experiment, all threshold structures except symmetric2 lead to identical fitted probabilities for a given link function. For conciseness, Table 3 therefore reports only the distinct configurations, grouping the remaining threshold structures under the flexible case.

First, the disagreement rate measures the proportion of test instances for which the two criteria select different classes when applied to the same probabilistic output. The observed disagreement rate varies between 12.6 and 23.5%, confirming that incorporating label-dependent loss at the decision stage can materially alter the classifier's output, even when the underlying probabilistic model is held fixed. Second, the distance-2 error rate focuses on more severe ordinal misclassifications, namely predictions that differ from the true class by two ordinal levels. These errors typically incur substantially higher loss under convex cost structures. The results indicate that ω -Ord-MAP exhibits a higher proportion of distance-2 errors than Ord-MAP across all model specifications. This behavior is not contradictory to the observed improvement in expected cost: the proposed rule minimizes expected loss under a convex, label-dependent cost structure, rather than directly controlling ordinal distance.

Taken together, these diagnosis measures clarify the mechanism underlying the performance gains observed for ω -Ord-MAP. Improvements are not driven by a uniform reduction

Table 3 Diagnostic decision-level rates comparing Ord-MAP and ω -Ord-MAP under the asymmetric loss configuration with $\alpha_L = 0.5$ and $\alpha_R = 2$. Reported measures include the disagreement rate and the proportion of distance-2 errors under each decision rule. Except for the symmetric2 specification, all threshold structures lead to identical results and are therefore collapsed under the flexible threshold

Setting	Diagnosis measures		
	Disagreement	distance-2 error rate	distance-2 error rate
	rate	with Ord-MAP	with ω -Ord-MAP
logistic-flexible	0.1769856	0.007504690	0.014383990
logistic-symmetric2	0.1769856	0.006253909	0.013758599
probit-flexible	0.1882427	0.007504690	0.015009381
probit-symmetric2	0.1844903	0.006879300	0.015009381
loglog-flexible	0.1463415	0.009380863	0.018136335
loglog-symmetric2	0.1444653	0.008755472	0.018136335
cloglog-flexible	0.2351470	0.006879300	0.015009381
cloglog-symmetric2	0.2295184	0.008130081	0.015634772
cauchit-flexible	0.1313321	0.005628518	0.008755472
cauchit-symmetric2	0.1263290	0.005628518	0.008755472

in ordinal distance errors, but rather by a targeted reallocation of predictions that reduces the overall expected cost. In particular, ω -Ord-MAP may accept fewer but more severe misclassifications in order to avoid many moderately costly errors. This highlights the distinction between optimizing ordinal distance and optimizing cost-sensitive risk, and illustrates how the choice of loss function fundamentally shapes the induced decision behavior.

Acknowledgments The author would like to thank the Editors and the reviewers who, with their comments, helped make the final improved version of the paper possible.

Author Contributions The unique author made all the contribution.

Funding Open Access Funding provided by Universitat Autònoma de Barcelona. The author is supported by Ministerio de Ciencia e Innovación, Gobierno de España, project ref. PID2021-123733NB-I00.

Data Availability No datasets were generated or analysed during the current study.

Code Availability The R scripts used to implement the experimental phase detailed in Sect. 5 can be found at <https://github.com/RosDelgado/omega-Ord-MAP>.

Declarations

Conflict of interest The authors declare no Conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Alcacer, A., Martinez-Garcia, M., & Epifanio, I. (2024). Ordinal classification for interval-valued data and interval-valued functional data, *Expert Systems with Applications*, Volume 238, Part F. *ISSN*, 122277, 0957–4174. <https://doi.org/10.1016/j.eswa.2023.122277>
- Binotto, G., & Delgado, R. (2025). Adapting performance metrics for ordinal classification to interval scale: Length matters. *Machine Learning*. <https://doi.org/10.1007/s10994-024-06654-4>
- García-Lapresta, J. L., & Pérez-Román, D. (2015). Ordinal proximity measures in the context of unbalanced qualitative scales and some applications to consensus and clustering. *Applied Soft Computing*, 35, 864–872. <https://doi.org/10.1016/j.asoc.2015.02.035>
- Delgado, R. (2025). Ord-MAP criterion: extending MAP for Ordinal Classification. *Knowledge-Based Systems*, Volume 324, 113837, *ISSN* 0950–7051, <https://doi.org/10.1016/j.knosys.2025.113837>.
- Delgado, R., Fernández-Peláez, F., Pallarés, N., et al. (2024). Predictive risk models for COVID-19 patients using the multi-thresholding meta-algorithm. *Scientific Reports*, 14, 28453. <https://doi.org/10.1038/s41598-024-77386-7>
- Elkan, C. (2001). The Foundations of Cost-Sensitive Learning. In *Proceedings of the 17th International Joint Conference of Artificial Intelligence (IJCAI'01)* 2, 973–978. Seattle, Washington: Morgan Kaufmann <https://doi.org/10.5555/1642194.1642224>
- García-Lapresta, J. L., González del Pozo, R., & Pérez-Román, D. (2018). Metrizable ordinal proximity measures and their aggregation. *Information Sciences*, 448–449, 149–163. <https://doi.org/10.1016/j.ins.2018.03.034>
- García-Lapresta, J. L., & Pérez-Román, D. (2018). Aggregating opinions in non-uniform ordered qualitative scales. *Applied Soft Computing*, 67, 652–657. <https://doi.org/10.1016/j.asoc.2017.05.064>
- Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association*. <https://doi.org/10.1198/016214506000001437>
- Likert, R. (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology*, 22(140), 5–55.
- Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1), 1–135. <https://doi.org/10.1561/15000000011>
- Robert, A., Chapman, L. A. C., Grah, R., Niehus, R., Sandmann, F., Prasse, B., Funk, S., & Kucharski, A. J. (2024). Predicting subnational incidence of COVID-19 cases and deaths in EU countries. *BMC Infectious Diseases*, 24, 204. <https://doi.org/10.1186/s12879-024-08986-x>
- Savage, L. J. (1971). Elicitation of Personal Probabilities and Expectations. *Journal of the American Statistical Association*, 66(336), 783–801. <https://doi.org/10.1080/01621459.1971.10482346>
- Suárez, J. L., García, S., & Herrera, F. (2021). Ordinal regression with explainable distance metric learning based on ordered sequences. *Machine Learning*, 110, 2729–2762. <https://doi.org/10.1007/s10994-021-06010-w>
- Vrugt, J. A. (2024). Distribution-based model evaluation and diagnostics: Elicitability, propriety, and scoring rules for hydrograph functionals. *Water Resource Research*. <https://doi.org/10.1029/2023WR036710>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.