A. Supplementary Material

Contents

A.1. List of Symbols

A.1.1 Mathematical Symbols

- $\mathbf{1}_U$: the 0-1 indicator function of subset U
- $\mathbb{R}$: the real numbers
- $\mathcal{R}$: a relation
- conv: the set of convex combinations
- $\vee$: the *inclusive disjunction* (*i.e.*, logical or)
- $\wedge$: the *conjunction* (*i.e.*, logical and)
- $\circ$: the composition of functions, *i.e.* $(g \circ f)(x) = g(f(x))$
- $\mathbf{E}$: the mathematical expectation

A.1.2 Symbols Related to Our Mathematical Framework

We organize these symbols according to the 6 pillars depicted in Fig. 1, which correspond to the 6 subsections of Sec. 3.

Symbols related to the 1st pillar (Sec. 3.1)
- Ω: the sample space (universe)
- ω: a sample (*i.e.*, an element of Ω)
- Σ: the event space (a σ-algebra on Ω, *e.g.* 2^Ω)
- E: an event (*i.e.*, an element of Σ)
- (Ω, Σ): the measurable space
- $\mathbb{P}_{(\Omega,\Sigma)}$: all performances on (Ω, Σ)
- Π: a set of performances ($\Pi \subseteq \mathbb{P}_{(\Omega,\Sigma)}$)
- P: a performance (*i.e.*, an element of $\mathbb{P}_{(\Omega,\Sigma)}$)

Symbols related to the 2nd pillar (Sec. 3.2)
- $\lesssim$: binary relation *worse or equivalent* on $\mathbb{P}_{(\Omega,\Sigma)}$
- $\gtrsim$: binary relation *better or equivalent* on $\mathbb{P}_{(\Omega,\Sigma)}$
- $\sim$: binary relation *equivalent* on $\mathbb{P}_{(\Omega,\Sigma)}$
- $>$: binary relation *better* on $\mathbb{P}_{(\Omega,\Sigma)}$
- $<$: binary relation *worse* on $\mathbb{P}_{(\Omega,\Sigma)}$
- $\lesssim \!\!\!\!\!\!\diagup \; \gtrsim$: binary relation *incomparable* on $\mathbb{P}_{(\Omega,\Sigma)}$

Symbols related to the 3rd pillar (Sec. 3.3)
- S: the random variable *Satisfaction*
- $s_{min,\Omega}$: the minimum satisfaction value
- $s_{max,\Omega}$: the maximum satisfaction value

Symbols related to the 4th pillar (Sec. 3.4)
- $\mathbb{E}$: the set of entities to rank
- ϵ: an entity, *i.e.* an element of $\mathbb{E}$
- eval: the performance *evaluation* function
- Φ: some performances that are for sure achievable

Symbols related to the 5th pillar (Sec. 3.5)
- X: a score
- $\mathrm{dom}(X)$: the domain of the score X
- X^E_V: the *expected value score* parameterized by the random variable V
- $X^P_{E_1|E_2}$: the *probabilistic score* parameterized by the events E_1 and E_2

Symbols related to the 6$^{\text{th}}$ pillar (Sec. 3.6)
- I: the random variable *Importance*

A.1.3 Symbols used for Operations on Performances

- filter_I: the *filtering* operation

A.1.4 Symbols used in the Performance Ordering and Performance-Based Ranking Theory

- $\text{rank}_{\mathbb{E}}$: the *ranking* function, w.r.t. the set of entities $\mathbb{E}$
- $\lesssim_X$: the ordering induced by the score X (*cf*. Theorem 1)
- R_I: the *ranking score* parameterized by the importance I
- τ: the rank correlation coefficient of Kendall [13]

A.1.5 Symbols used for the Particular Case of Two-Class Crisp Classifications

Particularization of the mathematical framework
- tn: the sample *true negative*
- fp: the sample *false positive*, *a.k.a.* type I error
- fn: the sample *false negative*, *a.k.a.* type II error
- tp: the sample *true positive*

Extensions to the mathematical framework
- ROC: the *Receiver Operating Characteristic* space, *i.e.* $FPR \times TPR$
- PR: the *Precision-Recall* space, *i.e.* $TPR \times PPV$
- Y: the random variable for the ground truth
- $\hat{Y}$: the random variable for the prediction
- $\mathbb{C}$: the set of classes
- c: a class (*i.e.*, an element of $\mathbb{C}$)
- c_-: the negative class
- c_+: the positive class

Scores
- A: the *accuracy*
- TNR: the *true negative rate*
- FPR: the *false positive rate*
- TPR: the *true positive rate*
- PPV: the *positive predictive value*
- F_β: the F-scores
- π_+: the *prior of the positive class*
- π_-: the *prior of the negative class*

A.2. How to Use our Framework: a Little Catalog of Problems

Throughout the paper, we have exemplified our theory with the problem of two-class classification. This section aims at showing the universality of our theory. It presents a little catalog of other problems, together with discussions on how to use our framework for them. These discussions are introductions. As shown by Sec. 5 and two recent works [12, 21], an in-depth analysis and particularization of our theory to the various problems (*e.g.*, to highlight their distinctive features, to review the current ranking practices from the literature and the consistency of popular scores, and to establish practical tools tailored to different user needs) may require substantial work that is out of the scope of this supplementary material.

In the following, we adopt a systematic approach: for each problem, we start by specifying a thought random experiment for the evaluation (this is an arbitrary choice since the random experiment is not unique for each problem; do not hesitate to use a different random experiment, the important is to specify it explicitly!), then we discuss the possible choices for the sample space Ω (and what the set of all performances is), for the modeling of tasks with the satisfaction S, for the modeling of the knowledge we have about the evaluation with the function Φ, and for the modeling of the application-specific preferences with the importance I.

A.2.1 Multi-Class Classification (with a Note on Micro- and Macro-Averaging)

Let us consider the following thought experiment to evaluate classifiers predicting classes in a finite and non-empty set $\mathbb{C}$.

Random Experiment 1. *(1) Draw a sample $s \in \mathbb{S}$ at random from a given* source $\mathcal{S}$*. (2) Apply the* oracle $\mathcal{O}$ *on s to obtain the ground-truth class $y(s) \in \mathbb{C}$. (3) Apply a* descriptor $\mathcal{D}$ *on s to obtain the features (*a.k.a. *attributes) $x(s) \in \mathbb{X}$. (4) Feed the* classifier $\mathcal{C}$ *with $x(s)$ to obtain the predicted class $\hat{y}(x(s)) \in \mathbb{C}$. (5) Set the outcome of the experiment to the pair $(y, \hat{y}) \in \mathbb{C}^2$.*

Choice for Ω and $\mathbb{P}_{(\Omega,\Sigma)}$. Our theory applies with $\Omega = \mathbb{C}^2$. By definition, the function eval gives, for any classifier $\mathcal{C} : \mathbb{X} \to \mathbb{C}$ (the evaluated entity, either deterministic or not), the distribution of outcomes resulting from this random experiment: $P_{\mathcal{C}} = \text{eval}(\mathcal{C})$. Note that it is implicit that the performances are specific for some given source $\mathcal{S}$ (*e.g.*, evaluation dataset), oracle $\mathcal{O}$, and descriptor $\mathcal{D}$. By convenience, one can manipulate the ground-truth and predicted classes with, respectively, the random variables Y and $\hat{Y}$ defined in such a way that $\omega = (Y(\omega), \hat{Y}(\omega))\, \forall \omega \in \Omega$.

Choice for S. Several classification tasks can be distinguished, as the following two examples show. (1) One can consider that all erroneous classifications are unsatisfactory and that correct classifications are satisfactory. For this task, the satisfaction is then binary and given by $S = \mathbf{1}_{Y=\hat{Y}}$. The expected value of the satisfaction, which is a particular case of ranking scores, is then equal to the multi-class accuracy. (2) One can also consider the similarity $\text{sim} : \mathbb{C}^2 \to \mathbb{R}$ between classes, and choose the satisfaction accordingly: $S(\omega) = \text{sim}(Y(\omega); \hat{Y}(\omega))$. A wide variety of tasks can be considered by tuning sim. In general, the expected value of the satisfaction is different from the multi-class accuracy.

Choice for Φ. As the classifier is used once and only once during the execution of the evaluation, we know that if the performances P_1 and P_2 are achievable by some classifiers $\mathcal{C}_1$ and $\mathcal{C}_2$, then any performance $\overline{P} = \lambda_1 P_1 + \lambda_2 P_2$ (with $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, and $\lambda_1 + \lambda_2 = 1$) is achievable by a classifier $\overline{\mathcal{C}}$ obtained by a non-deterministic combination of $\mathcal{C}_1$ and $\mathcal{C}_2$ that chooses them with respective probabilities λ_1 and λ_2. Thus, $\Phi = \text{conv}$ makes sense, and all ranking scores can be used to rank classifiers. However, it would be possible to go further, by considering other functions Φ that would include the knowledge that we can predict the performance achievable by composing the classifier with any of the $|\mathbb{C}|^{|\mathbb{C}|}$ functions $f : \mathbb{C} \to \mathbb{C}$. This would lead to other performance orderings suitable for ranking classifiers.

Choice for I. When $\Phi = \text{conv}$, we have demonstrated that all rankings induced by the ranking scores R_I satisfy all our three axioms. This leaves a great flexibility for the users to fine-tune the ranking w.r.t. their application-specific preferences through the random variable I. As we have seen, the only constraints are that $I \neq 0$ and $I(\omega) \geq 0\, \forall \omega \in \Omega$.

Note on micro- and macro-averaging. Micro- and macro-averaging are commonly used techniques to build scores for multi-class classification from scores for two-class classification [22]. We warn that they have pitfalls. In general, micro- and macro-averaging scores suitable for ranking two-class classifiers do not lead to scores suitable for ranking multi-class classifiers. The accuracy put aside, the performance orderings induced from the micro-averaged versions of the scores put in green in Tab. 1 are incompatible with $S = \mathbf{1}_{Y=\hat{Y}}$: our 2$^{\text{nd}}$ axiom is not satisfied. Moreover, the accuracy put again aside, the performance orderings induced from the macro-averaged versions of the scores put in green in Tab. 1 are incompatible with $\Phi = \text{conv}$: our 3$^{\text{rd}}$ axiom is not satisfied. A solution consists in using directly ranking scores defined for multi-class classification.

A.2.2 Regression (with a Note on the Mean Squared Error and the Mean Absolute Error)

Let us consider the following thought experiment to evaluate regressors.

Random Experiment 2. *(1) Draw a sample $s \in \mathbb{S}$ at random from a given* source $\mathcal{S}$*. (2) Apply the* oracle $\mathcal{O}$ *on s to obtain the ground-truth value $y(s) \in \mathbb{R}$. (3) Apply a* descriptor $\mathcal{D}$ *on s to obtain the features (*a.k.a. *attributes) $x(s) \in \mathbb{X}$. (4) Feed the* regressor $\mathcal{R}$ *with $x(s)$ to obtain the predicted value $\hat{y}(x(s)) \in \mathbb{R}$. (5) Set the outcome of the experiment to the pair $(y, \hat{y}) \in \mathbb{R}^2$.*

Choice for Ω and $\mathbb{P}_{(\Omega,\Sigma)}$. Our theory applies with $\Omega = \mathbb{R}^2$. By definition, the function eval gives, for any regressor $\mathcal{R} : \mathbb{X} \to \mathbb{R}$ (the evaluated entity, either deterministic or not), the distribution of outcomes resulting from this random experiment. It is the performance $P_{\mathcal{R}} = \mathrm{eval}(\mathcal{R})$ of $\mathcal{R}$. By convenience, one can manipulate the ground-truth and predicted values with, respectively, the random variables Y and $\hat{Y}$ defined in such a way that $\omega = (Y(\omega), \hat{Y}(\omega))\, \forall \omega \in \Omega$.

Choice for S. Clearly, it is not a good idea to choose $S = \mathbf{1}_{Y=\hat{Y}}$, similarly as one can do in classification. In practice, a regressor as no chance to predict the same value as the oracle, so this unfortunate choice for S would lead to performances P such that $P(S = 0) = 1$. In other words, all performances observed about real regressors would belong to the set of the worst performances (see Corollary 2), and their ranking would be of little interest (see Corollary 1). A better option consists in specifying a tolerance $\epsilon > 0$ and choosing $S = \mathbf{1}_{|Y-\hat{Y}|\leq\epsilon}$. An even more flexible option, which takes advantage of the fact that satisfaction values do not necessarily have to be positive, is to choose $S = f(|Y - \hat{Y}|)$ with any arbitrarily chosen monotonically decreasing function f. The plethora of choices that can be made for S makes it clear that there is an infinity of tasks related to the regression problem.

Choice for Φ. As the regressor is used once and only once during the execution of the evaluation, we know that if the performances P_1 and P_2 are achievable by some regressors $\mathcal{R}_1$ and $\mathcal{R}_2$, then any performance $\overline{P} = \lambda_1 P_1 + \lambda_2 P_2$ (with $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, and $\lambda_1 + \lambda_2 = 1$) is achievable by a regressor $\overline{\mathcal{R}}$ obtained by a non-deterministic combination of $\mathcal{R}_1$ and $\mathcal{R}_2$ that chooses them with respective probabilities λ_1 and λ_2. Thus, $\Phi = \mathrm{conv}$ makes sense, and all ranking scores can be used to rank regressors. However, it would be possible to go further, by considering other functions Φ that would include the knowledge that we can predict the performance achievable by adding noise, or applying a transformation on the output of the regressor. This would lead to other performance orderings suitable for ranking regressors.

Choice for I. When $\Phi = \mathrm{conv}$, we have demonstrated that all rankings induced by the ranking scores R_I satisfy all our three axioms. This leaves a great flexibility for the users to fine-tune the ranking w.r.t. their application-specific preferences through the random variable I. As we have seen, the only constraints are that $I \neq 0$ and $I(\omega) \geq 0\, \forall \omega \in \Omega$.

Note on the mean squared error and the mean absolute error. If we choose $S = -|Y - \hat{Y}|^2$, the ranking score R_I corresponding to uniform importance values, *i.e.*, the expected satisfaction X_S^E, yields a ranking that minimizes the *mean squared error* (MSE). If we choose $S = -|Y - \hat{Y}|$, the ranking score R_I corresponding to uniform importance values, *i.e.*, the expected satisfaction X_S^E, yields a ranking that minimizes the *mean absolute error* (MAE).

A.2.3 Information Retrieval

Let us consider the following thought experiment to evaluate information retrieval systems. We denote by $\mathbb{Q}$ the set of all possible queries.

Random Experiment 3. *(1) Draw a query $q \in \mathbb{Q}$ at random from a given source $\mathcal{S}$. (2) Apply the oracle $\mathcal{O}$ on q to obtain the ground-truth set of results $\mathbb{Y}$. (3) Apply the evaluated information retrieval system $\mathcal{S}$ on q to obtain the predicted set of results $\hat{\mathbb{Y}}$. (4) If $\mathbb{Y} = \emptyset$ and $\hat{\mathbb{Y}} = \emptyset$, restart the experiment, otherwise draw a result r at random in $\mathbb{Y} \cup \hat{\mathbb{Y}}$. (5) Choose the outcome as follows: fp if $r \notin \mathbb{Y}$ and $r \in \hat{\mathbb{Y}}$, fn if $r \in \mathbb{Y}$ and $r \notin \hat{\mathbb{Y}}$, and tp if $r \in \mathbb{Y}$ and $r \in \hat{\mathbb{Y}}$.*

Choice for Ω and $\mathbb{P}_{(\Omega,\Sigma)}$. Our theory applies with $\Omega = \{fp, fn, tp\}$. By definition, the function eval gives, for any retrieval system $\mathcal{S}$ defined on $\mathbb{Q}$ (the evaluated entity, either deterministic or not), the distribution of outcomes resulting from this random experiment: $P_{\mathcal{S}} = \mathrm{eval}(\mathcal{S})$.

Choice for S. Intuitively, everyone certainly agrees that $S(fp) < S(tp)$ and $S(fn) < S(tp)$. But we expect different opinions regarding whether the outcome (sample) fp gives less, equal, or more satisfaction than fn.

Choice for Φ. This random experiment is very interesting as, during its execution, the evaluated entity (the information retrieval system $\mathcal{S}$) can be used multiple times. In such a case, we have to discuss whether $\Phi = \mathrm{conv}$ is adequate. Let us consider two systems $\mathcal{S}_1$, $\mathcal{S}_2$ and their respective performances $P_1 = \mathrm{eval}(\mathcal{S}_1)$, $P_2 = \mathrm{eval}(\mathcal{S}_2)$. It is possible to show that the performance of a retrieval system $\overline{\mathcal{S}}$ obtained by a non-deterministic combination of $\mathcal{S}_1$ and $\mathcal{S}_2$, that chooses them with respective probabilities λ_1 and λ_2, is some interpolated performance $\overline{P} = \mu_1 P_1 + \mu_2 P_2$ (with $\mu_1 \geq 0$, $\mu_2 \geq 0$, and $\mu_1 + \mu_2 = 1$). Unless being in very particular cases, $\mu \neq \lambda$. In other words, we know that the performances that are convex combinations of achievable performances are also achievable (for any λ, there exists μ), but we do not know in general how to achieve them (for most μ it is not possible to determine λ). This contrasts with the other kinds of problems discussed in this catalog. In fact, the question we raise here is not specific to the information retrieval problem: it is peculiar to the random experiment that we have chosen for it. By slightly modifying the thought experiment, the question vanishes: instead of restarting the experiment when $\mathbb{Y} \cup \hat{\mathbb{Y}} = \emptyset$, one could yield a fourth outcome (and add it to the sample space Ω). By doing so, the evaluated entity $\mathcal{S}$ is used only once and $\Phi = \mathrm{conv}$ makes sense for sure.

Choice for I. If $\Phi = \mathrm{conv}$ is considered as adequate, then we have demonstrated that all rankings induced by the ranking scores R_I satisfy all our three axioms. This leaves a great flexibility for the users to fine-tune the ranking w.r.t. their application-specific preferences through the random variable I. As we have seen, the only constraints are that $I \neq 0$ and $I(\omega) \geq 0 \, \forall \omega \in \Omega$.

A.2.4 Detection (with a Note about the Intersection-over-Union and the F-Score)

Different types of detections are present in the literature. An example of spatial detection aims at predicting the axis-aligned bounding boxes around all the objects that match some given properties (*i.e.*, a semantic class) in input images. Examples of temporal detections include the detection of events in video streams and in audio recordings. By definition, such detection problems are called *action spotting* when the temporal window is small, and *activity detection* otherwise. Let us consider the following, generic, thought experiment to evaluate detectors.

Random Experiment 4. *(1) Draw an input at random from a given* source $\mathcal{S}$ (e.g.*, dataset). (2) Apply the* oracle $\mathcal{O}$ *on it to obtain a set* $\mathbb{Y}$ *of ground-truth detections. (3) Also apply the* detector $\mathcal{D}$ *on it to obtain a set* $\hat{\mathbb{Y}}$ *of predicted detections. (4) If* $\mathbb{Y} = \emptyset$ *and* $\hat{\mathbb{Y}} = \emptyset$*, then end the experiment with the outcome* $\circledast$*. Otherwise: (5) Apply a matching criterion between* $\mathbb{Y}$ *and* $\hat{\mathbb{Y}}$ *such that to any detection in* $\mathbb{Y}$ *should be associated at most a detection in* $\hat{\mathbb{Y}}$ *and vice versa. (6) Draw a detection* d *at random in* $\mathbb{Y} \cup \hat{\mathbb{Y}}$*. (7) Give as outcome* fp*,* fn *or* tp *depending on whether* d *is a prediction, a ground truth, or both (*i.e.*, a match).*

Choice for Ω and $\mathbb{P}_{(\Omega,\Sigma)}$. Our theory applies with $\Omega = \{\circledast, fp, fn, tp\}$. By definition, the function eval gives, for any detector (the evaluated entity), the distribution of outcomes resulting from this random experiment. It is the performance $P_{\mathcal{D}} = \mathrm{eval}(\mathcal{D})$ of $\mathcal{D}$.

Choice for S. Intuitively, everyone certainly agrees that $\circledast$ and tp give entire satisfaction. Moreover, we expect agreement on $S(fp) < S(tp)$ and $S(fn) < S(tp)$. But we expect different opinions regarding whether fp gives less, equal, or more satisfaction than fn.

Choice for Φ. As the detector is used once and only once during the execution of the evaluation, we know that if the performances P_1 and P_2 are achievable by some detectors $\mathcal{D}_1$ and $\mathcal{D}_2$, then any performance $\overline{P} = \lambda_1 P_1 + \lambda_2 P_2$ (with $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, and $\lambda_1 + \lambda_2 = 1$) is achievable by a detector $\overline{\mathcal{D}}$ obtained by a non-deterministic combination of $\mathcal{D}_1$ and $\mathcal{D}_2$ that chooses them with respective probabilities λ_1 and λ_2. Thus, $\Phi = \mathrm{conv}$ makes sense, and all ranking scores can be used to rank detectors.

Choice for I. If $\Phi = \mathrm{conv}$ is considered as adequate, then we have demonstrated that all rankings induced by the ranking scores R_I satisfy all our three axioms. This leaves a great flexibility for the users to fine-tune the ranking w.r.t. their application-specific preferences through the random variable I. As we have seen, the only constraints are that $I \neq 0$ and $I(\omega) \geq 0 \, \forall \omega \in \Omega$.

Note about the Intersection-over-Union and the F-score. Traditionally, in the literature, as soon as one has symbols fp, fn, and tp, regardless of their very fine meaning, one defines quantities $IoU = \frac{P(tp)}{P(fp)+P(fn)+P(tp)}$ and $F_1 = \frac{2P(tp)}{P(fp)+P(fn)+2P(tp)}$, and name them *Intersection-over-Union* (or *Jaccard*) and *F-one*, respectively. The exact meaning

of these quantities is not well standardized. In particular, the random experiment supporting the evaluation, if it exists, is rarely specified explicitly. For this reason, we cannot give the guarantee that these quantities are suitable to rank detectors in all works in which they have been used. However, with the random experiment given here-above, with $\Phi = \text{conv}$, and with $S = \mathbf{1}_{\{\circledast, tp\}}$, we can guarantee that IoU and F_1 are suitable to rank detectors because they are equal to the ranking scores with, respectively, $I = \mathbf{1}_{\{fp,fn,tp\}}$ and $I = \mathbf{1}_{\{fp,tp\}} + \mathbf{1}_{\{fn,tp\}}$. Thus, the performance orderings induced by them fulfill our three axioms.

A.2.5 Clustering (with a Note about Fowlkes-Mallows Index)

Let us consider the following thought experiment to evaluate clustering methods. These methods aim to place in different clusters (groups) dissimilar objects and in the same cluster (group) objects that are similar to each other. We denote by $\mathbb{E}$ the set of elements that these methods have to deal with. For the sake of simplicity, we do not consider hierarchical clustering.

Random Experiment 5. *(1) Apply both the* clustering method $\mathcal{C}$ *and the* oracle $\mathcal{O}$ *on* $\mathbb{E}$ *to obtain, respectively, the predicted and ground-truth clusterings. (2) Randomly draw two distinct elements,* ϵ_1 *and* ϵ_2*, from* $\mathbb{E}$*. (3) Consider that the pair* $(\epsilon_1; \epsilon_2)$ *is a negative or a positive, in a given clustering, when* ϵ_1 *and* ϵ_2 *are in different clusters or in the same cluster, respectively. (4) Choose the outcome as follows:* tn *when* $(\epsilon_1; \epsilon_2)$ *is negative in both the predicted and ground-truth clusterings,* fp *when* $(\epsilon_1; \epsilon_2)$ *is negative in the ground-truth clustering and positive in the predicted clustering,* fn *when* $(\epsilon_1; \epsilon_2)$ *is positive in the ground-truth clustering and negative in the predicted clustering, and* tp *when* $(\epsilon_1; \epsilon_2)$ *is positive in both the predicted and ground-truth clusterings.*

Choice for Ω and $\mathbb{P}_{(\Omega,\Sigma)}$. Our theory applies with $\Omega = \{tn, fp, fn, tp\}$. By definition, the function eval gives, for any clustering method (the evaluated entity), the distribution of outcomes resulting from this random experiment. It is the performance $P_{\mathcal{C}} = \text{eval}(\mathcal{C})$ of $\mathcal{C}$.

Choice for S. When S is chosen such that $S(fp) = S(fn) = 0$ and $S(tn) = S(tp) = 1$, we are in the same setting as the one we studied for the two-class classification in Sec. 5. This is indeed not because we use the same symbols for the elements of Ω —this is just a convention—, but because, in both settings, we have $|\Omega| = 4$, $|S = 0| = 2$, and $|S = 1| = 2$. This implies that the performance orderings that satisfy our three axioms for ranking two-class classifiers can also be used for ranking clustering methods, and vice versa.

Choice for Φ. As the clustering method is used once and only once during the execution of the evaluation, we know that if the performances P_1 and P_2 are achievable by some clustering methods $\mathcal{C}_1$ and $\mathcal{C}_2$, then any performance $\overline{P} = \lambda_1 P_1 + \lambda_2 P_2$ (with $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, and $\lambda_1 + \lambda_2 = 1$) is achievable by a clustering method $\overline{\mathcal{C}}$ obtained by a non-deterministic combination of $\mathcal{C}_1$ and $\mathcal{C}_2$ that chooses them with respective probabilities λ_1 and λ_2. Thus, $\Phi = \text{conv}$ makes sense, and all ranking scores can be used to rank clustering methods.

Choice for I. When $\Phi = \text{conv}$, we have demonstrated that all rankings induced by the ranking scores R_I satisfy all our three axioms. This leaves a great flexibility for the users to fine-tune the ranking w.r.t. their application-specific preferences through the random variable I. As we have seen, the only constraints are that $I \neq 0$ and $I(\omega) \geq 0 \, \forall \omega \in \Omega$.

Note about Fowlkes-Mallows index. The score FMI known as *Fowlkes-Mallows index* [9] and *cosine coefficient* [1], which is commonly used for clustering methods and defined as the geometric mean of the positive predictive value PPV (*a.k.a.* precision) and true positive rate TPR (*a.k.a.* sensitivity and recall), does not satisfy our 3rd axiom: the performance ordering induced by FMI is incompatible with $\Phi = \text{conv}$. More precisely, the clustering method $\mathcal{C}$ obtained by randomly choosing between some methods $\mathcal{C}_1$ or $\mathcal{C}_2$ can be such that $FMI(\text{eval}(\mathcal{C})) < \min(FMI(\text{eval}(\mathcal{C}_1)), FMI(\text{eval}(\mathcal{C}_2)))$, while it makes no sense to say that $\mathcal{C}$ can be worse than $\mathcal{C}_1$ or $\mathcal{C}_2$. From this perspective, it is advisable to use any ranking score instead of FMI, for example, those in green in Tab. 1.

A.2.6 Ranking (with a Note about Kendall's τ)

Let us consider the following thought experiment to evaluate ranking methods. We denote by $\mathbb{E}$ the set of elements that these methods have to rank. For the sake of simplicity, we prefer to deal only with the case with no tie hereafter.

Random Experiment 6. *(1) Apply both the* ranking method $\mathcal{R}$ *and the* oracle $\mathcal{O}$ *on* $\mathbb{E}$ *to obtain, respectively, the predicted and ground-truth sequences of elements. (2) Randomly draw two distinct elements,* ϵ_1 *and* ϵ_2*, from* $\mathbb{E}$*. (3) Four cases can*

occur depending on whether ϵ_1 is before or after ϵ_2 in the predicted sequence and whether ϵ_1 is before or after ϵ_2 in the ground-truth sequence. Nevertheless, two outcomes are enough: choose ☺ if ϵ_1 and ϵ_2 appear in the same order in both sequences, ☹ otherwise.

Choice for Ω and $\mathbb{P}_{(\Omega,\Sigma)}$. Our theory applies with $\Omega = \{☺, ☹\}$. By definition, the function eval gives, for any ranking method (the evaluated entity), the distribution of outcomes resulting from this random experiment. It is the performance $P_{\mathcal{R}} = \mathrm{eval}(\mathcal{R})$ of $\mathcal{R}$. The probability of drawing a *discordant pair* is given by $P(\{☹\})$, and the probability of drawing a *concordant pair* is given by $P(\{☺\})$.

Choice for S. Clearly, $S(☹) < S(☺)$ is wanted.

Choice for Φ. As the ranking method is used once and only once during the execution of the evaluation, we know that if the performances P_1 and P_2 are achievable by some ranking methods $\mathcal{R}_1$ and $\mathcal{R}_2$, then any performance $\overline{P} = \lambda_1 P_1 + \lambda_2 P_2$ (with $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, and $\lambda_1 + \lambda_2 = 1$) is achievable by a ranking method $\overline{\mathcal{R}}$ obtained by a non-deterministic combination of $\mathcal{R}_1$ and $\mathcal{R}_2$ that chooses them with respective probabilities λ_1 and λ_2. Thus, $\Phi = \mathrm{conv}$ makes sense, and all ranking scores can be used to rank ranking methods.

Choice for I. Because $|\Omega| = 2$ and $S(☹) \neq S(☺)$, we are in a particular case in which all ranking scores rank the ranking methods in the same way. From this point of view, fine-tuning I is useless.

Note about Kendall's τ. When $S(☹) = -1$ and $S(☺) = 1$, the expected value of the satisfaction is given by $X_S^E(P) = 1 - 2P(\{☹\}) = P(\{☺\}) - P(\{☹\}) = \tau(P)$. In other words, with the task corresponding to this choice for the satisfaction, Kendall's correlation coefficient τ [13] is the ranking score corresponding to uniform importance values.

A.3. Supplementary Material about Sec. 3.2

This section is devoted to reminders about the order theory.

A.3.1 Reminders of Classical Definitions.

Let $\mathcal{R}$ be a homogeneous binary relation on $\mathbb{P}_{(\Omega,\Sigma)}$. It is said:
- *reflexive iif* $P\mathcal{R}P\,\forall P$;
- *irreflexive iif* $\nexists P : P\mathcal{R}P$;
- *transitive iif* $P_1\mathcal{R}P_2 \wedge P_2\mathcal{R}P_3 \Rightarrow P_1\mathcal{R}P_3\,\forall P_1, P_2, P_3$;
- *symmetric iif* $P_1\mathcal{R}P_2 \Leftrightarrow P_2\mathcal{R}P_1\,\forall P_1, P_2$;
- *asymmetric iif* $\nexists (P_1, P_2) : P_1\mathcal{R}P_2 \wedge P_2\mathcal{R}P_1$;
- and *antisymmetric iif* $P_1\mathcal{R}P_2 \wedge P_2\mathcal{R}P_1 \Rightarrow P_1 = P_2$.

Two homogeneous binary relations $\mathcal{R}_a$ and $\mathcal{R}_b$ on $\mathbb{P}_{(\Omega,\Sigma)}$ are said *converse iif* $P_1\mathcal{R}_aP_2 \Leftrightarrow P_2\mathcal{R}_bP_1\,\forall P_1, P_2$.

A relation $\mathcal{R}$ is:
- an *equivalence iif* it is reflexive, transitive, and symmetric;
- a *preorder iif* it is reflexive and transitive;
- and an *order iif* it is reflexive, transitive, and antisymmetric.

An order $\mathcal{R}$ is said *total iif* $\nexists (P_1, P_2) : P_1 \not\mathcal{R}P_2 \wedge P_2 \not\mathcal{R}P_1$. It is said *partial* otherwise.

A.3.2 The 4 Cases in the Comparison of Two Performances with a Preorder $\lesssim$.

Let us now consider a preorder $\lesssim$ and derive the homogeneous binary relations $\sim$, $>$, $<$, $\lneq\!\!\gneq$ as follows:

$$P_1 \sim P_2 \Leftrightarrow P_1 \lesssim P_2 \wedge P_2 \lesssim P_1 \tag{1}$$
$$P_1 > P_2 \Leftrightarrow P_1 \not\lesssim P_2 \wedge P_2 \lesssim P_1 \tag{2}$$
$$P_1 < P_2 \Leftrightarrow P_1 \lesssim P_2 \wedge P_2 \not\lesssim P_1 \tag{3}$$
$$P_1 \lneq\!\!\gneq P_2 \Leftrightarrow P_1 \not\lesssim P_2 \wedge P_2 \not\lesssim P_1\,. \tag{4}$$

Indeed, we have:

$$P_1 \lesssim P_2 \Leftrightarrow P_1 < P_2 \vee P_1 \sim P_2\,. \tag{5}$$

Similarly, one can derive other binary relations taking unions of $\sim$, $>$, $<$, or $\lneq\!\!\gneq$. For example,

$$P_1 \gtrsim P_2 \Leftrightarrow P_1 > P_2 \vee P_1 \sim P_2\,. \tag{6}$$

A.3.3 Implications of the Transitivity of $\lesssim$.

We can easily check, for each $\mathcal{R}_{ab} \in \{\lesssim, \not\lesssim\}$, each $\mathcal{R}_{ba} \in \{\lesssim, \not\lesssim\}$, each $\mathcal{R}_{bc} \in \{\lesssim, \not\lesssim\}$, each $\mathcal{R}_{cb} \in \{\lesssim, \not\lesssim\}$, each $\mathcal{R}_{ca} \in \{\lesssim, \not\lesssim\}$, and each $\mathcal{R}_{ac} \in \{\lesssim, \not\lesssim\}$, if there exists (P_a, P_b, P_c) such that $P_a\mathcal{R}_{ab}P_b$, $P_b\mathcal{R}_{ba}P_a$, $P_b\mathcal{R}_{bc}P_c$, $P_c\mathcal{R}_{cb}P_b$, $P_c\mathcal{R}_{ca}P_a$, and $P_a\mathcal{R}_{ac}P_c$. Because of the assumed transitivity of $\lesssim$, there are only 29 possible cases out of the 2^6:

1. $P_a \lneq\!\!\gneq P_b, P_b \lneq\!\!\gneq P_c, P_a \lneq\!\!\gneq P_c$
2. $P_a \lneq\!\!\gneq P_b, P_b \lneq\!\!\gneq P_c, P_a > P_c$
3. $P_a \lneq\!\!\gneq P_b, P_b \lneq\!\!\gneq P_c, P_a < P_c$
4. $P_a \lneq\!\!\gneq P_b, P_b \lneq\!\!\gneq P_c, P_a \sim P_c$
5. $P_a \lneq\!\!\gneq P_b, P_b > P_c, P_a \lneq\!\!\gneq P_c$
6. $P_a \lneq\!\!\gneq P_b, P_b > P_c, P_a > P_c$
7. $P_a \lneq\!\!\gneq P_b, P_b < P_c, P_a \lneq\!\!\gneq P_c$
8. $P_a \lneq\!\!\gneq P_b, P_b < P_c, P_a < P_c$
9. $P_a \lneq\!\!\gneq P_b, P_b \sim P_c, P_a \lneq\!\!\gneq P_c$
10. $P_a > P_b, P_b \lneq\!\!\gneq P_c, P_a \lneq\!\!\gneq P_c$
11. $P_a > P_b, P_b \lneq\!\!\gneq P_c, P_a > P_c$
12. $P_a > P_b, P_b > P_c, P_a > P_c$
13. $P_a > P_b, P_b < P_c, P_a \lneq\!\!\gneq P_c$
14. $P_a > P_b, P_b < P_c, P_a > P_c$
15. $P_a > P_b, P_b < P_c, P_a < P_c$

16. $P_a > P_b, P_b < P_c, P_a \sim P_c$
17. $P_a > P_b, P_b \sim P_c, P_a > P_c$
18. $P_a < P_b, P_b \lneqq P_c, P_a \lneqq P_c$
19. $P_a < P_b, P_b \lneqq P_c, P_a < P_c$
20. $P_a < P_b, P_b > P_c, P_a \lneqq P_c$
21. $P_a < P_b, P_b > P_c, P_a > P_c$
22. $P_a < P_b, P_b > P_c, P_a < P_c$
23. $P_a < P_b, P_b > P_c, P_a \sim P_c$
24. $P_a < P_b, P_b < P_c, P_a < P_c$
25. $P_a < P_b, P_b \sim P_c, P_a < P_c$
26. $P_a \sim P_b, P_b \lneqq P_c, P_a \lneqq P_c$
27. $P_a \sim P_b, P_b > P_c, P_a > P_c$
28. $P_a \sim P_b, P_b < P_c, P_a < P_c$
29. $P_a \sim P_b, P_b \sim P_c, P_a \sim P_c$

From this list, we can derive some rules for manipulating the binary relations $\sim$, $>$, $<$, and $\lneqq$. First, we can see that $\sim$, $>$, and $<$ are transitive:

$$P_1 \sim P_2 \wedge P_2 \sim P_3 \Rightarrow P_1 \sim P_3 \tag{7}$$
$$P_1 > P_2 \wedge P_2 > P_3 \Rightarrow P_1 > P_3 \tag{8}$$
$$P_1 < P_2 \wedge P_2 < P_3 \Rightarrow P_1 < P_3\,. \tag{9}$$

Second, we can also see how $\sim$ can be combined with the other 3 relations:

$$(P_1 \sim P_2 \wedge P_2 > P_3) \vee (P_1 > P_2 \wedge P_2 \sim P_3) \Rightarrow P_1 > P_3 \tag{10}$$
$$(P_1 \sim P_2 \wedge P_2 < P_3) \vee (P_1 < P_2 \wedge P_2 \sim P_3) \Rightarrow P_1 < P_3 \tag{11}$$
$$(P_1 \sim P_2 \wedge P_2 \lneqq P_3) \vee (P_1 \lneqq P_2 \wedge P_2 \sim P_3) \Rightarrow P_1 \lneqq P_3\,. \tag{12}$$

And, third, we can see how $\lneqq$ can be combined with $>$ and $<$:

$$(P_1 \lneqq P_2 \wedge P_2 > P_3) \vee (P_1 > P_2 \wedge P_2 \lneqq P_3) \Rightarrow P_1 > P_3 \vee P_1 \lneqq P_3 \tag{13}$$
$$(P_1 \lneqq P_2 \wedge P_2 < P_3) \vee (P_1 < P_2 \wedge P_2 \lneqq P_3) \Rightarrow P_1 < P_3 \vee P_1 \lneqq P_3\,. \tag{14}$$

A.3.4 Properties of $\sim$, $>$, $<$, $\lneqq$, $\lesssim$, and $\gtrsim$.

Lemma 1. *When $\lesssim$ is a preorder, $\sim$ is reflexive.*

Proof. This results from the reflexivity of $\lesssim$ and from Eq. (1): $P \sim P \Leftrightarrow P \lesssim P \wedge P \lesssim P \Leftrightarrow true$. □

Lemma 2. *When $\lesssim$ is a preorder, $\sim$ in transitive.*

Proof. This results from the transitivity of $\lesssim$ (*cf*. Eq. (7)). □

Lemma 3. *When $\lesssim$ is a preorder, $\sim$ is symmetric.*

Proof. This results from the fact that the conjunction is symmetric and from Eq. (1): $P_1 \sim P_2 \Leftrightarrow P_1 \lesssim P_2 \wedge P_2 \lesssim P_1 \Leftrightarrow P_2 \lesssim P_1 \wedge P_1 \lesssim P_2 \Leftrightarrow P_2 \sim P_1$. □

Lemma 4. *When $\lesssim$ is a preorder, $>$ and $<$ are converse.*

Proof. This results from the fact that the conjunction is symmetric and from Eqs. (2) and (3): $P_1 > P_2 \Leftrightarrow P_1 \not\lesssim P_2 \wedge P_2 \lesssim P_1 \Leftrightarrow P_2 \lesssim P_1 \wedge P_1 \not\lesssim P_2 \Leftrightarrow P_2 < P_1$. □

Lemma 5. *When $\lesssim$ is a preorder, $>$ and $<$ are irreflexive.*

Proof. For $>$, this results from the reflexivity of $\lesssim$ and from Eq. (2): $P > P \Leftrightarrow P \not\lesssim P \wedge P \lesssim P \Leftrightarrow false \wedge true = false$. For $<$, the proof is similar. □

Lemma 6. *When $\lesssim$ is a preorder, $>$ and $<$ are asymmetric.*

Proof. For $>$, this is because $P_1 > P_2 \wedge P_2 > P_1 \Leftrightarrow (P_1 \not\lesssim P_2 \wedge P_2 \lesssim P_1) \wedge (P_2 \not\lesssim P_1 \wedge P_1 \lesssim P_2) \Leftrightarrow (P_1 \not\lesssim P_2 \wedge P_1 \lesssim P_2) \wedge (P_2 \not\lesssim P_1 \wedge P_2 \lesssim P_1) \Leftrightarrow false \wedge false \Leftrightarrow false$. For $<$, the proof is similar. □

Lemma 7. *When $\lesssim$ is a preorder, $>$ and $<$ are transitive.*

Proof. This results from the transitivity of $\lesssim$ (*cf*. Eqs. (8) and (9)). □

Lemma 8. *When $\lesssim$ is a preorder, $\lneqq$ is irreflexive.*

Proof. This results from the reflexivity of $\lesssim$ and from Eq. (4): $P \lneqq P \Leftrightarrow P \not\lesssim P \wedge P \not\lesssim P \Leftrightarrow false \wedge false \Leftrightarrow false$. □

Lemma 9. *When $\lesssim$ is a preorder, $\lneqq$ is symmetric.*

Proof. This results from the fact that the conjunction is symmetric and from Eq. (4): $P_1 \lneqq P_2 \Leftrightarrow P_1 \not\lesssim P_2 \wedge P_2 \not\lesssim P_1 \Leftrightarrow P_2 \not\lesssim P_1 \wedge P_1 \not\lesssim P_2 \Leftrightarrow P_2 \lneqq P_1$. □

Lemma 10. *When $\lesssim$ is a preorder, $\lesssim$ and $\gtrsim$ are converse.*

Proof. From Eqs. (5) and (6), as $>$ and $<$ are converse, we have $P_1 \gtrsim P_2 \Leftrightarrow P_1 > P_2 \vee P_1 \sim P_2 \Leftrightarrow P_2 < P_1 \vee P_2 \sim P_1 \Leftrightarrow P_2 \lesssim P_1$. □

Lemma 11. *When $\lesssim$ is a preorder, $\lesssim$ and $\gtrsim$ are reflexive.*

Proof. For $\lesssim$, it is by definition of preorders. For $\gtrsim$, from Eqs. (1), (2) and (6), we have $P \gtrsim P \Leftrightarrow P > P \vee P \sim P \Leftrightarrow (P \not\lesssim P \wedge P \lesssim P) \vee (P \lesssim P \wedge P \lesssim P) \Leftrightarrow (false \wedge true) \vee (true \wedge true) \Leftrightarrow true$. □

Lemma 12. *When $\lesssim$ is a preorder, $\lesssim$ and $\gtrsim$ are transitive.*

Proof. For $\lesssim$, it is by definition of preorders. For $\gtrsim$, as $\lesssim$ and $\gtrsim$ are converse, $P_1 \gtrsim P_2 \wedge P_2 \gtrsim P_3 \Leftrightarrow P_3 \lesssim P_2 \wedge P_2 \lesssim P_1 \Rightarrow P_3 \lesssim P_1 \Leftrightarrow P_1 \gtrsim P_3$. □

A.4. Supplementary Material about Sec. 3.6

A.4.1 The Visual Inspection of Formulas, to Determine the Importance given by Scores, can be Misleading!

Let us consider the example of two-class classification, with $P(\{tn\})$, $P(\{fp\})$, $P(\{fn\})$, and $P(\{tp\})$ denoting, respectively, the probability (or proportion) of true negatives, false positives, false negatives, and true positives. Here are two classical scores, the accuracy and the true positive rate:

$$A = P(\{tn\}) + P(\{tp\}) \qquad\qquad TPR = \frac{P(\{tp\})}{P(\{fn\}) + P(\{tp\})}$$

The formula for the accuracy gives the illusion that the same importance is given to $\{tn\}$ and $\{tp\}$ and that no importance at all is given to $\{fp\}$ and $\{fn\}$. For the true positive rate, the formula might give the impression that the same importance is given to $\{fn\}$ and $\{tp\}$ and that no importance at all is given to $\{tn\}$ and $\{fp\}$. In fact, the visual inspection of formulas like these is not reliable at all to judge the importance given by a score to the various events. To see it, consider rewriting the previous equations as

$$\begin{aligned} A &= (1-\alpha)P(\{tn\}) - \alpha P(\{fp\}) - \alpha P(\{fn\}) + (1-\alpha)P(\{tp\}) + \alpha && \forall \alpha \\ TPR &= \frac{-\alpha P(\{tn\}) - \alpha P(\{fp\}) - \alpha P(\{fn\}) + (1-\alpha)P(\{tp\}) + \alpha}{-\beta P(\{tn\}) - \beta P(\{fp\}) + (1-\beta)P(\{fn\}) + (1-\beta)P(\{tp\}) + \beta} && \forall \alpha, \beta \end{aligned}$$

A visual inspection of such formulas would lead, indeed, to other illusions about the events that have no importance. This observation, however, should not stop us from thinking in terms of importance. In fact, we do it in this paper, but we do it in a mathematical framework that allows to do it rigorously.

A.4.2 So, How can we Determine the Importance given by Scores?

One cannot determine the application-specific preferences implicitly considered by any score X. However, one can determine those that are consistent, or have the best consistency, with X. We detail these notions hereafter.

For a given set of performances. Consider a score X and a ranking score R_I. If, for a given set $\Pi \subseteq \mathbb{P}_{(\Omega,\Sigma)}$, the scores X and R_I are linked by a strict monotonic relationship on $\Pi \cap \mathrm{dom}(X) \cap \mathrm{dom}(R_I)$ and if $\Pi \cap \mathrm{dom}(X) = \Pi \cap \mathrm{dom}(R_I)$, then the performance orderings $\lesssim_X$ and $\lesssim_{R_I}$ (induced by X and R_I in the way specified in Theorem 1) are identical on Π. We say that the score X *is consistent with* the application-specific preferences I, on this set. A score can be consistent with different importance values (*e.g.*, as consequence of Properties 3 and 4).

For a given distribution of performances. Let $\mathring{\mathbb{P}} = \{P \in \mathbb{P}_{(\Omega,\Sigma)} : P(\{\omega\}) > 0 \forall \omega \in \Omega)\}$. All ranking scores are defined on this set. Consider a score X and the set $\Pi = \mathrm{dom}(X) \cap \mathring{\mathbb{P}}$. We say that the score X *has the best consistency with* the application-specific preferences I when I maximizes the rank correlation between X and R_I, on Π, for the given distribution of performances. See Appendix A.7.2 for computational details.

A.5. Supplementary Material about Sec. 4.2

A.5.1 Proof of Theorem 1.

For convenience, we provide a reminder of Theorem 1 and Axiom 1 below.

Theorem 1 (Sufficient condition for Axiom 1)**.** *A binary relation $\lesssim_X$ on $\mathbb{P}_{(\Omega,\Sigma)}$ induced by a score X as $P_1 \lesssim_X P_2$ iif either $P_1 = P_2$ or $P_1 \in \mathrm{dom}(X)$ and $P_2 \in \mathrm{dom}(X)$ and $X(P_1) \leq X(P_2)$, is a preorder satisfying Axiom 1.*

Axiom 1. *The ranking function $\mathrm{rank}_{\mathbb{E}} : \mathbb{E} \to [1, |\mathbb{E}|] : \epsilon \mapsto \mathrm{rank}_{\mathbb{E}}(\epsilon)$ satisfies $|\{\epsilon' \in \mathbb{E} : \mathrm{eval}(\epsilon) < \mathrm{eval}(\epsilon')\}| + 1 \leq \mathrm{rank}_{\mathbb{E}}(\epsilon) \leq |\{\epsilon' \in \mathbb{E} : \mathrm{eval}(\epsilon) \lesssim \mathrm{eval}(\epsilon')\}|$, where $\lesssim$ is a preorder on $\mathbb{P}_{(\Omega,\Sigma)}$.*

Proof. To establish that $\lesssim_X$ is a preorder, we have to show that it is (1) reflexive and (2) transitive.

(1) The reflexivity of $\lesssim_X$ is trivial to establish, since $P_1 = P_2 \Rightarrow P_1 \lesssim_X P_2$.

(2) The transitivity of $\lesssim_X$ can be shown as follows. $P_1 \lesssim_X P_2 \wedge P_2 \lesssim_X P_3$ implies that:
- either $P_1 \in \mathrm{dom}(X)$, $P_2 \in \mathrm{dom}(X)$, $P_3 \in \mathrm{dom}(X)$, and $X(P_1) \leq X(P_2) \wedge X(P_2) \leq X(P_3) \Rightarrow X(P_1) \leq X(P_3) \Rightarrow P_1 \lesssim_X P_3$;
- or $P_1 \notin \mathrm{dom}(X)$, $P_2 \notin \mathrm{dom}(X)$, $P_3 \notin \mathrm{dom}(X)$, and $P_1 = P_2 \wedge P_2 = P_3 \Rightarrow P_1 = P_3 \Rightarrow P_1 \lesssim_X P_3$.

We conclude that, in all cases, $P \lesssim_X P$ and $P_1 \lesssim_X P_2 \wedge P_2 \lesssim_X P_3 \Rightarrow P_1 \lesssim_X P_3$. The orderings $\lesssim_X$ induced by scores X are thus preorders. □

Summary. If the homogeneous binary relations $\sim$, $>$, $<$, and $\lneq$ on $\mathbb{P}_{(\Omega,\Sigma)}$ are derived from the ordering $\lesssim_X$ as explained above, and if the $\lesssim_X$ is derived from the score X, then the comparison between performances P_1 and P_2 can be summarized as follows.

	$P_1 \in \mathrm{dom}(X)$	$P_1 \notin \mathrm{dom}(X)$
$P_2 \in \mathrm{dom}(X)$	$X(P_1) < X(P_2) \Leftrightarrow P_1 < P_2$ $X(P_1) = X(P_2) \Leftrightarrow P_1 \sim P_2$ $X(P_1) > X(P_2) \Leftrightarrow P_1 > P_2$	$P_1 \lneq P_2$
$P_2 \notin \mathrm{dom}(X)$	$P_1 \lneq P_2$	$P_1 = P_2 \Leftrightarrow P_1 \sim P_2$ $P_1 \neq P_2 \Leftrightarrow P_1 \lneq P_2$

A.5.2 Proof of Theorem 2

For convenience, we provide a reminder of Theorem 2 and Axiom 2 below.

Theorem 2 (Sufficient condition for Axiom 2)**.** *If a score X satisfies $\min_{\omega \in E} S(\omega) \leq X(P) \leq \max_{\omega \in E} S(\omega)$ for all events $E \in \Sigma$ and all performances $P \in \mathrm{dom}(X)$ such that $P(E) = 1$, then the ordering $\lesssim_X$ satisfies Axiom 2.*

Axiom 2. *For $P_1, P_2 \in \mathbb{P}_{(\Omega,\Sigma)}$ such that $P_1(S \leq s) = 1$ and $P_2(S \geq s) = 1$ for some s, then $P_1 \lesssim P_2$ or $P_1 \lneq P_2$.*

Proof. Axiom 2 is satisfied when $P_1 \notin \mathrm{dom}(X)$ or $P_2 \notin \mathrm{dom}(X)$.
- Either $P_1 = P_2 \Leftrightarrow P_1 \sim P_2 \Rightarrow P_1 \lesssim P_2$,
- or $P_1 \neq P_2 \Leftrightarrow P_1 \lneq P_2$.

Axiom 2 is also satisfied when $P_1 \in \mathrm{dom}(X)$ and $P_2 \in \mathrm{dom}(X)$.
- On the one hand, the axiom stipulates that the event $E_1 = \{\omega \in \Omega : S(\omega) \leq s\}$ and the performance P_1 are such that $P_1(E_1) = 1$. Trivially, we have $\max_{\omega \in E_1} S(\omega) \leq s$. On the other hand, the theorem stipulates that, as $P_1(E_1) = 1$, $X(P_1) \leq \max_{\omega \in E_1} S(\omega)$. Putting all together, we have $X(P_1) \leq s$.
- On the one hand, the axiom stipulates that the event $E_2 = \{\omega \in \Omega : S(\omega) \geq s\}$ and the performance P_2 are such that $P_2(E_2) = 1$. Trivially, we have $s \leq \min_{\omega \in E_2} S(\omega)$. On the other hand, the theorem stipulates that, as $P_2(E_2) = 1$, $\min_{\omega \in E_2} S(\omega) \leq X(P_2)$. Putting all together, we have $s \leq X(P_2)$.
- As we have established that $X(P_1) \leq s$ and $s \leq X(P_2)$, we have $X(P_1) \leq X(P_2) \Leftrightarrow P_1 \lesssim P_2$.

□

A.5.3 Proof of Theorem 3

For convenience, we provide a reminder of Theorem 3 and Axiom 3 below.

Theorem 3 (Sufficient condition for Axiom 3)**.** *If a score X is such that $\Pi \subseteq \mathrm{dom}(X) \Rightarrow \Phi(\Pi) \subseteq \mathrm{dom}(X)$ and $\min_{P\in\Pi} X(P) \leq X(\overline{P}) \leq \max_{P\in\Pi} X(P)$ for all $\Pi \subseteq \mathrm{dom}(X)$ and all $\overline{P} \in \Phi(\Pi)$, then the ordering $\lesssim_X$ satisfies Axiom 3.*

Axiom 3. *Let P be a performance, and Π be a set of performances on $\mathbb{P}_{(\Omega,\Sigma)}$ such that $P' \lesssim P \vee P \lesssim P' \forall P' \in \Pi$.*
- *$P' \lesssim P \forall P' \in \Pi \Rightarrow \overline{P} \lesssim P \forall \overline{P} \in \Phi(\Pi)$;*
- *$P' \not\lesssim P \forall P' \in \Pi \Rightarrow \overline{P} \not\lesssim P \forall \overline{P} \in \Phi(\Pi)$;*
- *$P \lesssim P' \forall P' \in \Pi \Rightarrow P \lesssim \overline{P} \forall \overline{P} \in \Phi(\Pi)$;*
- *and $P \not\lesssim P' \forall P' \in \Pi \Rightarrow P \not\lesssim \overline{P} \forall \overline{P} \in \Phi(\Pi)$.*

Proof. We take $\lesssim = \lesssim_X$ and $\Pi \neq \emptyset$.

Remainder of the conditions. The first condition of Theorem 3 is:

$$\Pi \subseteq \mathrm{dom}(X) \Rightarrow \Phi(\Pi) \subseteq \mathrm{dom}(X)\,. \tag{15}$$

The second condition of Theorem 3 is:

$$\min_{P\in\Pi} X(P) \leq X(\overline{P}) \leq \max_{P\in\Pi} X(P) \qquad \forall\, \Pi \subseteq \mathrm{dom}(X) \qquad \forall\, \overline{P} \in \Phi(\Pi)\,. \tag{16}$$

The condition of Axiom 3 is that P is comparable to all performances in the set Π:

$$P' \lesssim P \vee P \lesssim P' \qquad \forall\, P' \in \Pi\,. \tag{17}$$

On the domain of X. By Theorem 1, this last condition implies that

$$P \in \mathrm{dom}(X)\,, \tag{18}$$

and

$$\Pi \subseteq \mathrm{dom}(X)\,. \tag{19}$$

Taking Eq. (15) and Eq. (19) together, we have

$$\Phi(\Pi) \subseteq \mathrm{dom}(X) \qquad \Leftrightarrow \qquad \overline{P} \in \mathrm{dom}(X) \qquad \forall\, \overline{P} \in \Phi(\Pi)\,. \tag{20}$$

Proof that $P' \lesssim P \forall P' \in \Pi \Rightarrow \overline{P} \lesssim P \forall \overline{P} \in \Phi(\Pi)$. On the one hand, we have, by Theorem 1,

$$\begin{aligned} P' \lesssim P \forall P' \in \Pi \Leftrightarrow & X(P') \leq X(P) \qquad \forall\, P' \in \Pi \\ \Leftrightarrow & \max_{P'\in\Pi} X(P') \leq X(P)\,. \end{aligned}$$

On the other hand, Eq. (16) implies that

$$X(\overline{P}) \leq \max_{P'\in\Pi} X(P') \qquad \forall\, \overline{P} \in \Phi(\Pi)\,.$$

Considering the last two equations together, we obtain

$$\begin{aligned} & X(\overline{P}) \leq \max_{P'\in\Pi} X(P') \leq X(P) \qquad \forall\, \overline{P} \in \Phi(\Pi) \\ \Rightarrow & X(\overline{P}) \leq X(P) \qquad \forall\, \overline{P} \in \Phi(\Pi)\,. \end{aligned}$$

By Theorem 1, we have thus $\overline{P} \lesssim P$.

Proof that $P' \not\lesssim P \forall P' \in \Pi \Rightarrow \overline{P} \not\lesssim P \forall \overline{P} \in \Phi(\Pi)$. On the one hand, we have, by Eq. (17) and Theorem 1,

$$\begin{aligned} P' \not\lesssim P \forall\, P' \in \Pi \Rightarrow & P < P' \forall\, P' \in \Pi \\ \Leftrightarrow & X(P) < X(P') \qquad \forall\, P' \in \Pi \\ \Leftrightarrow & X(P) < \min_{P'\in\Pi} X(P') \end{aligned}$$

On the other hand, Eq. (16) implies that

$$\min_{P' \in \Pi} X(P') \leq X(\overline{P}) \qquad \forall \overline{P} \in \Phi(\Pi)$$

Considering the last two equations together, we obtain

$$\begin{aligned} X(P) < \min_{P' \in \Pi} X(P') \leq X(\overline{P}) \qquad & \forall \overline{P} \in \Phi(\Pi) \\ \Rightarrow X(P) < X(\overline{P}) \qquad & \forall \overline{P} \in \Phi(\Pi)\,. \end{aligned}$$

By Theorem 1, we have thus $P < \overline{P}$, and by Eq. (17), $\overline{P} \not\lesssim P$.

Proof that $P \lesssim P' \, \forall P' \in \Pi \Rightarrow P \lesssim \overline{P} \, \forall \overline{P} \in \Phi(\Pi)$. On the one hand, we have, by Theorem 1,

$$\begin{aligned} P \lesssim P' \, \forall P' \in \Pi \Leftrightarrow & X(P) \leq X(P') \qquad \forall P' \in \Pi \\ \Leftrightarrow & X(P) \leq \min_{P' \in \Pi} X(P')\,. \end{aligned}$$

On the other hand, Eq. (16) implies that

$$\min_{P' \in \Pi} X(P') \leq X(\overline{P}) \qquad \forall \overline{P} \in \Phi(\Pi)\,.$$

Considering the last two equations together, we obtain

$$\begin{aligned} X(P) \leq \min_{P' \in \Pi} X(P') \leq X(\overline{P}) \qquad & \forall \overline{P} \in \Phi(\Pi) \\ \Rightarrow X(P) \leq X(\overline{P}) \qquad & \forall \overline{P} \in \Phi(\Pi)\,. \end{aligned}$$

By Theorem 1, we have thus $P \lesssim \overline{P}$.

Proof that $P \not\lesssim P' \, \forall P' \in \Pi \Rightarrow P \not\lesssim \overline{P} \, \forall \overline{P} \in \Phi(\Pi)$. On the one hand, we have, by Eq. (17) and Theorem 1,

$$\begin{aligned} P \not\lesssim P' \, \forall \, P' \in \Pi \Rightarrow & P' < P \, \forall \, P' \in \Pi \\ \Leftrightarrow & X(P') < X(P) \qquad \forall \, P' \in \Pi \\ \Leftrightarrow & \max_{P' \in \Pi} X(P') < X(P) \end{aligned}$$

On the other hand, Eq. (16) implies that

$$X(\overline{P}) \leq \max_{P' \in \Pi} X(P') \qquad \forall \overline{P} \in \Phi(\Pi)$$

Considering the last two equations together, we obtain

$$\begin{aligned} X(\overline{P}) \leq \max_{P' \in \Pi} X(P') < X(P) \qquad & \forall \overline{P} \in \Phi(\Pi) \\ \Rightarrow X(\overline{P}) < X(P) \qquad & \forall \overline{P} \in \Phi(\Pi)\,. \end{aligned}$$

By Theorem 1, we have thus $\overline{P} < P$, and by Eq. (17), $P \not\lesssim \overline{P}$.

□

A.6. Supplementary Material about Sec. 4.3

A.6.1 All Ranking Scores can be Used to Rank Performances (for $\Phi = \mathrm{conv}$)

To show that all ranking scores can be used to rank performances, for $\Phi = \mathrm{conv}$, we show that these scores satisfy the conditions of Theorems 1, 2, and 3.

All ranking scores satisfy the conditions of Theorem 1, and thus Axiom 1. For convenience, we provide a reminder of Theorem 1 and Axiom 1 below.

Theorem 1 (Sufficient condition for Axiom 1)**.** *A binary relation $\lesssim_X$ on $\mathbb{P}_{(\Omega,\Sigma)}$ induced by a score X as $P_1 \lesssim_X P_2$ iif either $P_1 = P_2$ or $P_1 \in \mathrm{dom}(X)$ and $P_2 \in \mathrm{dom}(X)$ and $X(P_1) \leq X(P_2)$, is a preorder satisfying Axiom 1.*

Axiom 1. *The ranking function $\mathrm{rank}_{\mathbb{E}} : \mathbb{E} \to [1, |\mathbb{E}|] : \epsilon \mapsto \mathrm{rank}_{\mathbb{E}}(\epsilon)$ satisfies $|\{\epsilon' \in \mathbb{E} : \mathrm{eval}(\epsilon) < \mathrm{eval}(\epsilon')\}| + 1 \leq \mathrm{rank}_{\mathbb{E}}(\epsilon) \leq |\{\epsilon' \in \mathbb{E} : \mathrm{eval}(\epsilon) \lesssim \mathrm{eval}(\epsilon')\}|$, where $\lesssim$ is a preorder on $\mathbb{P}_{(\Omega,\Sigma)}$.*

Theorem 4. *All ranking scores satisfy the conditions of Theorem 1.*

Proof. For all ranking scores R_I, it is possible to induce an ordering $\lesssim_{R_I}$ satisfying the requirements of Theorem 1. □

All ranking scores satisfy the conditions of Theorem 2, and thus Axiom 2. For convenience, we provide a reminder of Theorem 2 and Axiom 2 below.

Theorem 2 (Sufficient condition for Axiom 2)**.** *If a score X satisfies $\min_{\omega\in E} S(\omega) \leq X(P) \leq \max_{\omega\in E} S(\omega)$ for all events $E \in \Sigma$ and all performances $P \in \mathrm{dom}(X)$ such that $P(E) = 1$, then the ordering $\lesssim_X$ satisfies Axiom 2.*

Axiom 2. *For $P_1, P_2 \in \mathbb{P}_{(\Omega,\Sigma)}$ such that $P_1(S \leq s) = 1$ and $P_2(S \geq s) = 1$ for some s, then $P_1 \lesssim P_2$ or $P_1 \lneqq P_2$.*

Theorem 5. *All ranking scores satisfy the conditions of Theorem 2.*

Proof. We take $X = R_I$. When $P(E) = 1$, we have

$$R_I(P) = \frac{\sum_{\omega\in\Omega} I(\omega)S(\omega)P(\{\omega\})}{\sum_{\omega\in\Omega} I(\omega)P(\{\omega\})} = \frac{\sum_{\omega\in E} I(\omega)S(\omega)P(\{\omega\})}{\sum_{\omega\in E} I(\omega)P(\{\omega\})}$$

with $\sum_{\omega\in\Omega} I(\omega)P(\{\omega\}) > 0$ when $P \in \mathrm{dom}(R_I)$.

- Let $M = \max_{\omega\in E} S(\omega)$. We have

$$\begin{aligned}
& S(\omega) \leq M \qquad \forall \omega \in E \\
\Leftrightarrow & S(\omega) - M \leq 0 \qquad \forall \omega \in E \\
\Rightarrow & \sum_{\omega\in E} I(\omega)\left[S(\omega) - M\right]P(\{\omega\}) \leq 0 \qquad \text{as } I(\omega) \geq 0 \text{ and } P(\{\omega\}) \geq 0 \\
\Leftrightarrow & \sum_{\omega\in E} I(\omega)S(\omega)P(\{\omega\}) \leq \sum_{\omega\in E} I(\omega)MP(\{\omega\}) \\
\Leftrightarrow & \sum_{\omega\in E} I(\omega)S(\omega)P(\{\omega\}) \leq M \underbrace{\sum_{\omega\in E} I(\omega)P(\{\omega\})}_{>0} \\
\Leftrightarrow & \frac{\sum_{\omega\in E} I(\omega)S(\omega)P(\{\omega\})}{\sum_{\omega\in E} I(\omega)P(\{\omega\})} \leq M \\
\Leftrightarrow & R_I(P) \leq M
\end{aligned}$$

- Let $m = \min_{\omega \in E} S(\omega)$. We have

$$
\begin{aligned}
& S(\omega) \geq m \qquad \forall \omega \in E \\
\Leftrightarrow & S(\omega) - m \geq 0 \qquad \forall \omega \in E \\
\Rightarrow & \sum_{\omega \in E} I(\omega) \left[S(\omega) - m\right] P(\{\omega\}) \geq 0 \qquad \text{as } I(\omega) \geq 0 \text{ and } P(\{\omega\}) \geq 0 \\
\Leftrightarrow & \sum_{\omega \in E} I(\omega) S(\omega) P(\{\omega\}) \geq \sum_{\omega \in E} I(\omega) m P(\{\omega\}) \\
\Leftrightarrow & \sum_{\omega \in E} I(\omega) S(\omega) P(\{\omega\}) \geq m \underbrace{\sum_{\omega \in E} I(\omega) P(\{\omega\})}_{>0} \\
\Leftrightarrow & \frac{\sum_{\omega \in E} I(\omega) S(\omega) P(\{\omega\})}{\sum_{\omega \in E} I(\omega) P(\{\omega\})} \geq m \\
\Leftrightarrow & R_I(P) \geq m
\end{aligned}
$$

- Putting all together, when $P(E) = 1$, we have $m \leq R_I(P) \leq M$, and so,

$$
\min_{\omega \in E} S(\omega) \leq R_I(P) \leq \max_{\omega \in E} S(\omega) \,. \tag{21}
$$

□

All ranking scores satisfy the conditions of Theorem 3, and thus Axiom 3 (for $\Phi = \text{conv}$). For convenience, we provide a reminder of Theorem 3 and Axiom 3 below.

Theorem 3 (Sufficient condition for Axiom 3)**.** *If a score X is such that $\Pi \subseteq \text{dom}(X) \Rightarrow \Phi(\Pi) \subseteq \text{dom}(X)$ and $\min_{P \in \Pi} X(P) \leq X(\overline{P}) \leq \max_{P \in \Pi} X(P)$ for all $\Pi \subseteq \text{dom}(X)$ and all $\overline{P} \in \Phi(\Pi)$, then the ordering $\lesssim_X$ satisfies Axiom 3.*

Axiom 3. *Let P be a performance, and Π be a set of performances on $\mathbb{P}_{(\Omega,\Sigma)}$ such that $P' \lesssim P \vee P \lesssim P' \, \forall P' \in \Pi$.*

- *$P' \lesssim P \, \forall P' \in \Pi \Rightarrow \overline{P} \lesssim P \, \forall \overline{P} \in \Phi(\Pi)$;*
- *$P' \not\lesssim P \, \forall P' \in \Pi \Rightarrow \overline{P} \not\lesssim P \, \forall \overline{P} \in \Phi(\Pi)$;*
- *$P \lesssim P' \, \forall P' \in \Pi \Rightarrow P \lesssim \overline{P} \, \forall \overline{P} \in \Phi(\Pi)$;*
- *and $P \not\lesssim P' \, \forall P' \in \Pi \Rightarrow P \not\lesssim \overline{P} \, \forall \overline{P} \in \Phi(\Pi)$.*

Theorem 6. *All ranking scores satisfy the conditions of Theorem 3 (for $\Phi = \text{conv}$).*

Proof. The proof is in two parts.

- First, let us show that $\Pi \subseteq \text{dom}(R_I) \Rightarrow \text{conv}(\Pi) \subseteq \text{dom}(R_I)$. For any $\overline{P} \in \text{conv}(\Pi)$ there exists a weighting function $\lambda_{\Pi,\overline{P}} : \Pi \to \mathbb{R}_{\geq 0} : P \mapsto \lambda_{\Pi,\overline{P}}(P)$ such that $\sum_{P \in \Pi} \lambda_{\Pi,\overline{P}}(P) = 1$ and $\sum_{P \in \Pi} \lambda_{\Pi,\overline{P}}(P) P = \overline{P}$. For all $\overline{P} \in \text{conv}(\Pi)$, we have:

$$
\begin{aligned}
\Pi \subseteq \text{dom}(R_I) & \Leftrightarrow \sum_{\omega \in \Omega} I(\omega) P(\{\omega\}) \neq 0 \qquad \forall P \in \Pi \\
& \Rightarrow \sum_{P \in \Pi} \lambda_{\Pi,\overline{P}}(P) \sum_{\omega \in \Omega} I(\omega) P(\{\omega\}) \neq 0 \\
& \Leftrightarrow \sum_{\omega \in \Omega} I(\omega) \sum_{P \in \Pi} \lambda_{\Pi,\overline{P}} P(\{\omega\}) \neq 0 \\
& \Leftrightarrow \sum_{\omega \in \Omega} I(\omega) \overline{P}(\{\omega\}) \neq 0 \\
& \Leftrightarrow \overline{P} \in \text{dom}(R_I) \,.
\end{aligned}
$$

- Second, let us show that, for all $\overline{P} \in \text{conv}(\Pi)$, $\min_{P\in\Pi} R_I(P) \leq R_I(\overline{P}) \leq \max_{P\in\Pi} R_I(P)$. Let us pose $l = \min_{P\in\Pi} R_I(P)$ and $u = \max_{P\in\Pi} R_I(P)$. We have:

$$
\begin{aligned}
& l \leq R_I(P) \leq u \qquad \forall P \in \Pi \\
\Leftrightarrow & l \leq \frac{\sum_{\omega\in\Omega} I(\omega)S(\omega)P(\{\omega\})}{\sum_{\omega\in\Omega} I(\omega)P(\{\omega\})} \leq u \qquad \forall P \in \Pi \\
\Leftrightarrow & l \sum_{\omega\in\Omega} I(\omega)P(\{\omega\}) \leq \sum_{\omega\in\Omega} I(\omega)S(\omega)P(\{\omega\}) \leq u \sum_{\omega\in\Omega} I(\omega)P(\{\omega\}) \qquad \forall P \in \Pi \\
\Rightarrow & l \sum_{\omega\in\Omega} I(\omega)\overline{P}(\{\omega\}) \leq \sum_{\omega\in\Omega} I(\omega)S(\omega)\overline{P}(\{\omega\}) \leq u \sum_{\omega\in\Omega} I(\omega)\overline{P}(\{\omega\}) \\
\Leftrightarrow & l \leq \frac{\sum_{\omega\in\Omega} I(\omega)S(\omega)\overline{P}(\{\omega\})}{\sum_{\omega\in\Omega} I(\omega)\overline{P}(\{\omega\})} \leq u \\
\Leftrightarrow & l \leq R_I(\overline{P}) \leq u
\end{aligned}
$$

□

A.6.2 On the Properties of Ranking Scores.

Proof of Property 1. Let us demonstrate that we have $R_I(P) = X_S^E(P')$ with $P' = \text{filter}_I(P)$. For all $\omega \in \Omega$, we have:

$$P'(\{\omega\}) = \frac{P(\{\omega\})I(\omega)}{\sum_{\omega'\in\Omega} P(\{\omega'\})I(\omega')}$$

Thus,

$$
\begin{aligned}
X_S^E(P') &= \sum_{\omega\in\Omega} P'(\{\omega\})S(\omega) \\
&= \sum_{\omega\in\Omega} \frac{P(\{\omega\})I(\omega)}{\sum_{\omega'\in\Omega} P(\{\omega'\})I(\omega')} S(\omega) \\
&= \frac{\sum_{\omega\in\Omega} P(\{\omega\})I(\omega)S(\omega)}{\sum_{\omega'\in\Omega} P(\{\omega'\})I(\omega')} \\
&= R_I(P)
\end{aligned}
$$

□

Proof of Property 2. Let $S' = \alpha S + \beta$ with $\alpha, \beta \in \mathbb{R}$.

$$\frac{\sum_{\omega\in\Omega} P(\{\omega\})S'(\omega)I(\omega)}{\sum_{\omega\in\Omega} P(\{\omega\})I(\omega)} = \alpha \frac{\sum_{\omega\in\Omega} P(\{\omega\})S(\omega)I(\omega)}{\sum_{\omega\in\Omega} P(\{\omega\})I(\omega)} + \beta$$

□

Proof of Property 3. $R_{kI} = \frac{\sum_{\omega\in\Omega} kI(\omega)S(\omega)P(\{\omega\})}{\sum_{\omega\in\Omega} kI(\omega)P(\{\omega\})} = \frac{\sum_{\omega\in\Omega} I(\omega)S(\omega)P(\{\omega\})}{\sum_{\omega\in\Omega} I(\omega)P(\{\omega\})} = R_I$ □

Proof of Property 4. Let us consider a binary satisfaction, that is $S(\omega) \in \{0, 1\}\, \forall \omega \in \Omega$. Let us define the events $E_0 = \{\omega \in \Omega : S(\omega) = 0\}$ and $E_1 = \{\omega \in \Omega : S(\omega) = 1\}$. If $I' = (\mathbf{1}_{S=0}\alpha_0 + \mathbf{1}_{S=1}\alpha_1)I$ with $\alpha_0 > 0$ and $\alpha_1 > 0$, then

$$
\begin{aligned}
R_I(P) &= \frac{\sum_{\omega\in\Omega} I(\omega)S(\omega)P(\{\omega\})}{\sum_{\omega\in\Omega} I(\omega)P(\{\omega\})} \\
&= \frac{\sum_{\omega\in E_1} I(\omega)P(\{\omega\})}{\sum_{\omega\in E_0} I(\omega)P(\{\omega\}) + \sum_{\omega\in E_1} I(\omega)P(\{\omega\})}
\end{aligned}
$$

and

$$\begin{aligned}
R_{I'}(P) &= \frac{\sum_{\omega\in\Omega} I'(\omega)S(\omega)P(\{\omega\})}{\sum_{\omega\in\Omega} I'(\omega)P(\{\omega\})} \\
&= \frac{\sum_{\omega\in E_1} I'(\omega)P(\{\omega\})}{\sum_{\omega\in E_0} I'(\omega)P(\{\omega\}) + \sum_{\omega\in E_1} I'(\omega)P(\{\omega\})} \\
&= \frac{\sum_{\omega\in E_1} \alpha_1 I(\omega)P(\{\omega\})}{\sum_{\omega\in E_0} \alpha_0 I(\omega)P(\{\omega\}) + \sum_{\omega\in E_1} \alpha_1 I(\omega)P(\{\omega\})} \\
&= \frac{\alpha_1 \sum_{\omega\in E_1} I(\omega)P(\{\omega\})}{\alpha_0 \sum_{\omega\in E_0} I(\omega)P(\{\omega\}) + \alpha_1 \sum_{\omega\in E_1} I(\omega)P(\{\omega\})}
\end{aligned}$$

Thus, $R_{I'} = \frac{\alpha_1 R_I}{\alpha_0(1-R_I)\alpha_1 R_I}$ and $\frac{\partial R_{I'}}{\partial R_I} = \frac{\alpha_0\alpha_1}{(\alpha_0(1-R_I)\alpha_1 R_I)^2} > 0$. This leads immediately to the conclusion that $\lesssim_{R_{I'}} = \lesssim_{R_I}$. □

Proof of Properties 5 and 6. Let us consider a binary satisfaction and the events $E_0 = \{\omega \in \Omega : S(\omega) = 0\}$ and $E_1 = \{\omega \in \Omega : S(\omega) = 1\}$. Let I_1 and I_2 be two random variables and $I = \lambda_1 I_1 + \lambda_2 I_2$ with $\lambda_1, \lambda_2 \in \mathbb{R}$ such that $\lambda_1 + \lambda_2 = 1$.

- When the random variables I_1 and I_2 are such that $I_1(\omega) = I_2(\omega) = I(\omega)\,\forall\omega \in E_1$, if we take $f : x \mapsto x^{-1}$,

$$\begin{aligned}
&\lambda_1 f\left(R_{I_1}(P)\right) + \lambda_2 f\left(R_{I_1}(P)\right) \\
=&\lambda_1 \left(1 + \frac{\sum_{\omega\in E_0} I_1(\omega)P(\{\omega\})}{\sum_{\omega\in E_1} I_1(\omega)P(\{\omega\})}\right) + \lambda_2 \left(1 + \frac{\sum_{\omega\in E_0} I_2(\omega)P(\{\omega\})}{\sum_{\omega\in E_1} I_2(\omega)P(\{\omega\})}\right) \\
=&\lambda_1 \left(1 + \frac{\sum_{\omega\in E_0} I_1(\omega)P(\{\omega\})}{\sum_{\omega\in E_1} I(\omega)P(\{\omega\})}\right) + \lambda_2 \left(1 + \frac{\sum_{\omega\in E_0} I_2(\omega)P(\{\omega\})}{\sum_{\omega\in E_1} I(\omega)P(\{\omega\})}\right) \\
=&(\lambda_1 + \lambda_2) + \frac{\lambda_1 \sum_{\omega\in E_0} I_1(\omega)P(\{\omega\}) + \lambda_2 \sum_{\omega\in E_0} I_2(\omega)P(\{\omega\})}{\sum_{\omega\in E_1} I(\omega)P(\{\omega\})} \\
=&1 + \frac{\sum_{\omega\in E_0} (\lambda_1 I_1 + \lambda_2 I_2)(\omega)P(\{\omega\})}{\sum_{\omega\in E_1} I(\omega)P(\{\omega\})} \\
=&1 + \frac{\sum_{\omega\in E_0} I(\omega)P(\{\omega\})}{\sum_{\omega\in E_1} I(\omega)P(\{\omega\})} \\
=&f\left(R_I(P)\right)
\end{aligned}$$

- When the random variables I_1 and I_2 are such that $I_1(\omega) = I_2(\omega) = I(\omega)\,\forall\omega \in E_0$, if we take $f : x \mapsto (1-x)^{-1}$,

$$\begin{aligned}
&\lambda_1 f\left(R_{I_1}(P)\right) + \lambda_2 f\left(R_{I_1}(P)\right) \\
=&\lambda_1 \left(1 + \frac{\sum_{\omega\in E_1} I_1(\omega)P(\{\omega\})}{\sum_{\omega\in E_0} I_1(\omega)P(\{\omega\})}\right) + \lambda_2 \left(1 + \frac{\sum_{\omega\in E_1} I_2(\omega)P(\{\omega\})}{\sum_{\omega\in E_0} I_2(\omega)P(\{\omega\})}\right) \\
=&\lambda_1 \left(1 + \frac{\sum_{\omega\in E_1} I_1(\omega)P(\{\omega\})}{\sum_{\omega\in E_0} I(\omega)P(\{\omega\})}\right) + \lambda_2 \left(1 + \frac{\sum_{\omega\in E_1} I_2(\omega)P(\{\omega\})}{\sum_{\omega\in E_0} I(\omega)P(\{\omega\})}\right) \\
=&(\lambda_1 + \lambda_2) + \frac{\lambda_1 \sum_{\omega\in E_1} I_1(\omega)P(\{\omega\}) + \lambda_2 \sum_{\omega\in E_1} I_2(\omega)P(\{\omega\})}{\sum_{\omega\in E_0} I(\omega)P(\{\omega\})} \\
=&1 + \frac{\sum_{\omega\in E_1} (\lambda_1 I_1 + \lambda_2 I_2)(\omega)P(\{\omega\})}{\sum_{\omega\in E_0} I(\omega)P(\{\omega\})} \\
=&1 + \frac{\sum_{\omega\in E_1} I(\omega)P(\{\omega\})}{\sum_{\omega\in E_0} I(\omega)P(\{\omega\})} \\
=&f\left(R_I(P)\right)
\end{aligned}$$

□

Proof of Property 7. Let $\mathcal{R} \in \{<, \leq, =, \geq, >\}$ and $v = R_I(P)$. For all $P' \in \mathbb{P}_{(\Omega,\Sigma)}$, we have:

$$
\begin{aligned}
& R_I(P')\mathcal{R}R_I(P) && (22)\\
\Leftrightarrow & R_I(P')\mathcal{R}v && (23)\\
\Leftrightarrow & \frac{\sum_{\omega\in\Omega} I(\omega)S(\omega)P'(\{\omega\})}{\sum_{\omega\in\Omega} I(\omega)P'(\{\omega\})}\mathcal{R}v && (24)\\
\Leftrightarrow & \left[\sum_{\omega\in\Omega} I(\omega)S(\omega)P'(\{\omega\})\right] \mathcal{R} \left[v\sum_{\omega\in\Omega} I(\omega)P'(\{\omega\})\right] && (25)\\
\Leftrightarrow & \sum_{\omega\in\Omega} I(\omega)\left[S(\omega)-v\right]P'(\{\omega\})\mathcal{R}0 && (26)
\end{aligned}
$$

This is either a linear equality or a linear inequality constraint. Thus,

$$
\begin{aligned}
\phi_{\mathcal{R}}(P) &= \left\{P' \in \mathbb{P}_{(\Omega,\Sigma)} : R_I(P')\mathcal{R}R_I(P)\right\} && (27)\\
&= \left\{P' \in \mathbb{P}_{(\Omega,\Sigma)} : \sum_{\omega\in\Omega} I(\omega)\left[S(\omega)-v\right]P'(\{\omega\})\mathcal{R}0\right\} && (28)
\end{aligned}
$$

is a convex subset of $\mathbb{P}_{(\Omega,\Sigma)}$. □

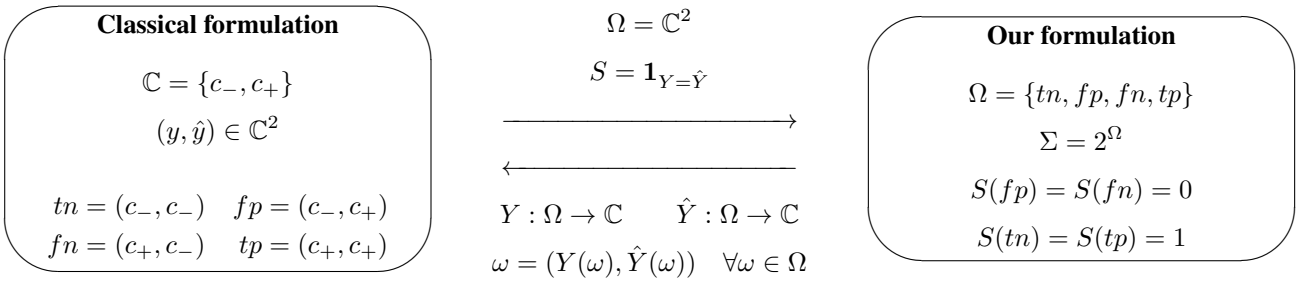

Figure A.7.1. Passages between two formulations (left: classical, right: ours) for the performance analysis of two-class classification problems.

A.7. Supplementary Material about Sec. 5.2

A.7.1 Link between Classical Formulation and Ours.

Fig. A.7.1 shows the connections between the classical formulation of the two-class classification task and our formulation, as explained in Sec. 5.

A.7.2 Custom Optimization Algorithm to Estimate Kendall's τ.

For any score X, our algorithm aims at determining the minimum and maximum values that a rank correlation between X and our ranking scores R_I can take over all possible importances I. Note that this algorithm is not specific to Kendall's τ [13] and could also be used with any other rank correlation, for example Spearman's ρ [23].

Variables. Leveraging Properties 3 and 4, we know that the rank-correlation between X and a ranking score R_{I_1} is equal to the rank-correlation between X and another ranking score R_{I_2} if $\frac{I_1(tp)}{I_1(tn)+I_1(tp)} = \frac{I_2(tp)}{I_2(tn)+I_2(tp)}$ and $\frac{I_1(fn)}{I_1(fp)+I_1(fn)} = \frac{I_2(fn)}{I_2(fp)+I_2(fn)}$. For this reason, we consider only two variables: $a = \frac{I(tp)}{I(tn)+I(tp)} \in [0,1]$ and $b = \frac{I(fn)}{I(fp)+I(fn)} \in [0,1]$.

Objective function. We optimize the function $\tau(a,b)$ that gives the rank correlation between X and R_{I^*} with $I^*(tp) = 1-a$, $I^*(tp) = 1-b$, $I^*(tp) = b$, and $I^*(tp) = a$. In practice, this is an estimation based on a finite set of performances on which X and R_{I^*} are applied. Note that $\tau(a,b)$ is not a continuous function when estimated on a finite set of performances. The chosen optimization technique circumvents the difficulties related to that.

Optimization technique. We implemented a custom coarse-to-fine grid-based direct search [5]: we compute $\tau(a,b)$ on a coarse grid over the unit square, locate the maximum on the grid, center a smaller square and a finer grid around that point, and iterate until the square is small enough.

A.7.3 Scores Perfectly Correlated with a Ranking Score, for all Performances

- **The accuracy:** $A = R_I$ with $I(tn) = 1/2$, $I(fp) = 1/2$, $I(fn) = 1/2$, and $I(tp) = 1/2$.
- **The F-score with $\beta = 0.5$:** $F_{0.5} = R_I$ with $I(tn) = 0$, $I(fp) = 4/5$, $I(fn) = 1/5$, and $I(tp) = 1$.
- **The F-score with $\beta = 1.0$:** $F_1 = R_I$ with $I(tn) = 0$, $I(fp) = 1/2$, $I(fn) = 1/2$, and $I(tp) = 1$.
- **The F-score with $\beta = 2.0$:** $F_2 = R_I$ with $I(tn) = 0$, $I(fp) = 1/5$, $I(fn) = 4/5$, and $I(tp) = 1$.
- **The negative predictive value:** $NPV = R_I$ with $I(tn) = 1$, $I(fp) = 0$, $I(fn) = 1$, and $I(tp) = 0$.
- **The positive predictive value:** $PPV = R_I$ with $I(tn) = 0$, $I(fp) = 1$, $I(fn) = 0$, and $I(tp) = 1$.
- **The true negative rate:** $TNR = R_I$ with $I(tn) = 1$, $I(fp) = 1$, $I(fn) = 0$, and $I(tp) = 0$.
- **The true positive rate:** $TPR = R_I$ with $I(tn) = 0$, $I(fp) = 0$, $I(fn) = 1$, and $I(tp) = 1$.

A.7.4 Scores Perfectly Correlated with a Ranking Score, for the Performances Corresponding to Given Class Priors $\pi_- \neq 0$ and $\pi_+ \neq 0$

- **The balanced accuracy:** $BA = R_I$ with $I(tn) = \pi_+$, $I(fp) = \pi_+$, $I(fn) = \pi_-$, and $I(tp) = \pi_-$.
- **Cohen's kappa:** $\kappa = \frac{R_I - 2\pi_-\pi_+}{\pi_-^2+\pi_+^2}$ with $I(tn) = \frac{\pi_+^2}{\pi_-^2+\pi_+^2}$, $I(fp) = \frac{1}{2}$, $I(fn) = \frac{1}{2}$, and $I(tp) = \frac{\pi_-^2}{\pi_-^2+\pi_+^2}$. Thus, $\frac{\partial \kappa}{R_I} > 0$.

- **The informedness (*a.k.a.* Youden's J):** $J_Y = 2R_I - 1$ with $I(tn) = \pi_+$, $I(fp) = \pi_+$, $I(fn) = \pi_-$, and $I(tp) = \pi_-$. Thus, $\frac{\partial J_Y}{R_I} > 0$.
- **The negative likelihood ratio:** $NLR = \frac{1-R_I}{R_I}$ with $I(tn) = 1$, $I(fp) = 0$, $I(fn) = 1$, and $I(tp) = 0$. Thus, $\frac{\partial NLR}{R_I} < 0$.
- **The positive likelihood ratio:** $PLR = \frac{R_I}{1-R_I}$ with $I(tn) = 0$, $I(fp) = 1$, $I(fn) = 0$, and $I(tp) = 1$. Thus, $\frac{\partial PLR}{R_I} > 0$.
- **The probability of the elementary event *true negative*:** $PTN = \pi_- R_I$ with $I(tn) = 1$, $I(fp) = 1$, $I(fn) = 0$, and $I(tp) = 0$. Thus, $\frac{\partial PTN}{R_I} > 0$.
- **The probability of the elementary event *true positive*:** $PTP = \pi_+ R_I$ with $I(tn) = 0$, $I(fp) = 0$, $I(fn) = 1$, and $I(tp) = 1$. Thus, $\frac{\partial PTP}{R_I} > 0$.