Mathematical Programming (2020) 182:103–140
https://doi.org/10.1007/s10107-019-01387-y

FULL LENGTH PAPER

Series A

Goal scoring, coherent loss and applications to machine learning

Wenzhuo Yang[1,2] · **Melvyn Sim**[3] · **Huan Xu**[4]

Received: 25 October 2016 / Accepted: 5 March 2019 / Published online: 25 March 2019
© Springer-Verlag GmbH Germany, part of Springer Nature and Mathematical Optimization Society 2019

Abstract

Motivated by the binary classification problem in machine learning, we study in this paper a class of decision problems where the decision maker has a list of goals, from which he aims to attain the maximal possible number of goals. In binary classification, this essentially means seeking a prediction rule to achieve the lowest probability of misclassification, and computationally it involves minimizing a (difficult) non-convex, 0–1 loss function. To address the intractability, previous methods consider minimizing the *cumulative loss*—the sum of convex surrogates of the 0–1 loss of each goal. We revisit this paradigm and develop instead an *axiomatic* framework by proposing a set of salient properties on functions for goal scoring and then propose the *coherent loss* approach, which is a tractable upper-bound of the loss over the *entire set* of goals. We show that the proposed approach yields a strictly tighter approximation to the total loss (i.e., the number of missed goals) than any convex cumulative loss approach while preserving the convexity of the underlying optimization problem. Moreover, this approach, applied to for binary classification, also has a robustness interpretation which builds a connection to robust SVMs.

Keywords Satisficing · Goal · Robust optimization · Classification · SVM · Coherent loss

✉ Wenzhuo Yang
yangwenzhuo08@gmail.com

Melvyn Sim
dscsimm@nus.edu.sg

Huan Xu
huan.xu@isye.gatech.edu

1 Department of Electrical and Computer Engineering, National University of Singapore, Singapore, Singapore

2 SAP Leonardo Machine Learning, Singapore, Singapore

3 Department of Decision Sciences, National University of Singapore, Singapore, Singapore

4 H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, USA

Mathematics Subject Classification 90C29

1 Introduction

Simon, in his seminal works [42,43], argued that rather than formulating and solving complicated optimization problems, in reality decision makers typically choose the actions that ensure certain aspiration levels being achieved. In other words, a natural decision making paradigm for real-world agent is to seek solutions to *satisfy* predefined goals. A phenomenon termed "satisficing" by Simon himself [43] has been extensively studied since then (e.g., [7,10,11,16,26]).

In practice, decision makers often face goals that either compete for limited resource, or are inherently contradicting. Consequently, some tradeoffs among the goals are sought after, since achieving all goals simultaneously would be impossible. In this paper, we focus on the case where the decision maker is interested in maximizing her goals attainment—a problem we call "goal scoring". There are many practical examples of goal scoring including, among others, the number of visitors to a web page who click the ads banners and the number of scenarios that a randomly perturbed constraint would remain feasible.

While goal scoring includes a wide variety of problems, this paper is mostly motivated by a specific example of goal scoring—the classification task in machine learning. Classification is one of the central aspects in supervised learning. The goal of supervised learning is to predict an unobserved output value y from an observed input $\mathbf{x}$. This is achieved by learning a function relationship $y \approx f(\mathbf{x})$ from a set of observed training examples $\{(y_i, \mathbf{x}_i)\}_{i=1}^m$. The quality of predictor $f(\cdot)$ is often measured by some loss function $\ell(f(\mathbf{x}), y)$. A typical statistical setup in machine learning assumes that all training and test samples are i.i.d. samples drawn from an unknown distribution μ, and the goal is to find a predictor $f(\cdot)$ such that the expected loss $\mathbb{E}_{(y,\mathbf{x})\sim\mu}\ell(f(\mathbf{x}), y)$ is minimized. Since μ is unknown, the expected loss is often replaced by the empirical loss

$$L_{emp}(f) \triangleq \frac{1}{m}\sum_{i=1}^{m} \ell(f(\mathbf{x}_i), y_i). \tag{1}$$

Minimizing $L_{emp}(f)$, as well as numerous regularization based variants of it, is one of the fundamental cornerstones of statistical machine learning, e.g., [34,44,45].

For binary classification problems, the labels $y \in \{-1, +1\}$. A point $(y, \mathbf{x})$ is correctly predicted if $\text{sign}(f(\mathbf{x})) = y$, and its classification error is given by the 0–1 loss:

$$\ell(f(\mathbf{x}_i),\ y_i) = \mathbf{1}(y \neq \text{sign}(f(\mathbf{x})) = \mathbf{1}(yf(\mathbf{x}) \leq 0).$$

Due to the non-convexity of the indicator function, minimizing the empirical classification error $\sum_i \mathbf{1}(y_i f(\mathbf{x}_i) \leq 0)$ is known to be NP-hard even to approximate [2,6]. A number of methods have been proposed to mitigate this computational difficulty

based on minimizing the "cumulative loss" that is the sum of individual losses given by,

$$L_\phi(f) \triangleq \frac{1}{m}\sum_{i=1}^{m}\phi(yf(\mathbf{x}))$$

where $\phi(\cdot)$ is a convex upper bound of the classification error $\mathbf{1}(yf(\mathbf{x}) \leq 0)$. For example, AdaBoost [21,22,38] employs the exponential loss function $\exp(-yf(\mathbf{x}))$, and Support Vector Machines (SVMs) [8,17] employ a hinge-loss function $\max\{1 - yf(\mathbf{x}), 0\}$.

The empirical loss of the classification naturally fits the goal scoring framework, where correctly predicting each data point is an individual goal, and the learner aims to maximize the number of goals attained. In this paper, we revisit the paradigm of *cumulative loss*. We take an axiomatic approach, eliciting salient properties for loss functions for quantifying the performance of a decision rule in terms of attaining multiple goals. We characterize functions satisfying these properties, which we term *coherent loss* as opposed to cumulative loss, by establishing a (dual) representation theorem. This dual representation result enables us to identify the *minimal coherent loss* function, which, loosely speaking, is the coherent loss function that *best approximates* the number of unattained goals, and in particular is a tighter bound than any convex cumulative loss. We show that optimizing this function is equivalent to a convex optimization problem that is computationally tractable.

An alternative perspective to understand the proposed coherent loss approach is that, instead of using an upper bound of the *individual* 0–1 loss on each missed goal, the coherent loss is indeed a tractable upper bound of the *total* number of missed goals. That is, we look for $\Phi : \mathbb{R}^m \mapsto \mathbb{R}$ such that

$$\Phi(c_1, \ldots, c_m) \geq \frac{1}{m}\sum_{i=1}^{m}\mathbf{1}(c_i \leq 0), \ \ \forall (c_1, \ldots, c_m) \in \mathbb{R}^m.$$

Thus, as coherent loss functions are more general than cumulative loss functions, one may expect to obtain a tighter and still tractable bound via the coherent loss function.

The rest of the paper is organized as follows. In Sect. 2, we formally define the goal scoring problem and list some examples, ranging from machine learning, to discrete optimization, to stochastic programming and beyond, that falls into this category. In Sect. 3, we present the salient properties that define coherent loss functions, and derive a dual-representation theorem that characterize the set of coherent loss function. Equipped with the representation theorem, we identify the minimal coherent loss, and show that optimizing such a loss function is tractable. We then turn to the application of the coherent loss framework to the classification task in Sect. 4, analyzing both the computational issue and the statistical implications for such a framework. We remark that a tighter approximation of the 0–1 loss due to the coherent loss framework can potentially reduce the impact of outliers on the classification accuracy, i.e., more robustness. Section 5 reports the experimental results which show that classification methods based on coherent loss function outperform the standard SVM. The proofs

to Theorem 1, 2, 3 and 5 are lengthy, and hence deferred to Sect. 6. Some concluding remarks are offered in Sect. 7.

Notations We use boldface letters to represent column vectors, and capital letters for matrices. We reserve $\mathbf{e}$ for special vectors: $\mathbf{e}_i$ is the vector whose i-th entry is 1, and the rest are 0; $\mathbf{e}_N$, where N is an index set, is the vector that for all $i \in N$, the corresponding entry equals 1, and zero otherwise; $\mathbf{e}^n \in \mathbb{R}^n$ is the vector with all entries equal to 1. The i-th entry of a vector $\mathbf{x}$ is denoted by x_i. We use $[c]_+$ to denote $\max\{0, c\}$ and $\mathbf{1}[\cdot]$ to denote the indicator function, and let $\mathcal{P}_n$ be the set of all $n \times n$ permutation matrices and $\mathbf{I}_n$ be the $n \times n$ identity matrix.

2 Goal scoring

We now formally define the notion of *goal scoring*. We are given a set of *m subtasks*, where each subtask is abstracted as a payoff $X_i \in \mathbb{R}$ compared against an *aspiration level* $\tau_i \in \mathbb{R}$. For example, X_i may be the the return of a portfolio for a given scenario and τ_i is the index to benchmark with; or X_i may be the the time for finish a specific part of a project, and τ_i is the deadline; alternatively in classification, X_i may be the margin of the i-th data point, and τ_i is the required margin ($\tau_i = 0$ if only the sign of classification matters). We define the subtask premium $u_i = X_i - \tau_i$, to denote the excess payoff above the aspiration level, which essentially represents whether the i-th subtask is accomplished and by how much.

A *goal* $\mathcal{G}_i$ consists of a set of subtasks, mathematically represented as a non-empty subset of $\{1, 2, \ldots, m\}$. Notice that different goals may share common subtasks. A goal $\mathcal{G}_i$ is considered "achieved", if all its subtasks are fulfilled, i.e.,

$$u_j \geq 0, \quad \forall j \in \mathcal{G}_i.$$

Given a collection of goals $\mathcal{G} = \{\mathcal{G}_1, \ldots, \mathcal{G}_n\}$, and denote by $\mathcal{S}$ all such collections, then the *goal score* $\vartheta : \mathbb{R}^m \times \mathcal{S} \mapsto \{1, \ldots, n\}$ counts the number of goals in $\mathcal{G}$ that are achieved, i.e.,

$$\vartheta(u_1, \ldots, u_m, \mathcal{G}) = \sum_{i=1}^{n} \mathbf{1}[u_j \geq 0, \forall j \in \mathcal{G}_i].$$

As a side remark, notice that the collection $\mathcal{G}$ essentially induces a bipartite graph between m subtasks and n goals.

Equivalently, we may define the following loss function $\varrho : \mathbb{R}^m \times \mathcal{S} \mapsto [0, 1]$, termed *goal miss*, which is the fraction of goals not achieved. And the decision goal is to maximize the goal score, or equivalently to minimize the goal miss.

$$\varrho(u_1, \ldots, u_m, \mathcal{G}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}[u_j < 0, \exists j \in \mathcal{G}_i]. \tag{2}$$

A closely related topic to goal scoring is goal programming which has been extensively studied for decades, e.g., [12,14,15,18]. Goal programming is a subset of multi-criteria decision analysis, usually being taken as an extension of linear programming to handle multiple objectives or goals. Typically, goal programming introduces positive and negative deviational variables for goals and minimizes the deviations in the objective function. Instead of minimizing deviations, our goal scoring framework tries to minimize the number of goals that are not achieved and allows a goal consisting of multiple subtasks. Our framework has a close relationship with the buffered probability of exceedance (bPOE) [33] which is an alternative measure of tail probability. This paper shows that minimizing the empirical bPOE is equivalent to minimizing a special case of the goal miss.

Our previous work [47] only considered the case where goals are singleton, namely, $\mathcal{G} = \{\{1\}, \{2\}, \ldots, \{m\}\}$, being limited to tackling binary classification problems. This paper studies goal scoring – a generalization of [47], leading to more general theoretical results and a wider range of practical applications. The goal scoring problem can model a variety of decision problems besides classification tasks, which we illustrate with the following list of examples including binary classification, resource allocation, join chance constraints and graph problems.

Example 1: binary classification

Binary classification is a well-motivated example of goal scoring. As we discussed in Sect. 1, the goal is to predict an unknown label $y \in \{-1, +1\}$ from an observed input $\mathbf{x}$, by learning a function relationship $y \approx \text{sign}\big(f(\mathbf{x})\big)$ from a set of labeled training examples $\{(y_i, \mathbf{x}_i)\}_{i=1}^m$, and the quality of the learned function $f(\cdot)$ is measured by

$$L_{emp}(f) \triangleq \frac{1}{m}\sum_{i=1}^{m} \ell(f(\mathbf{x}_i), y_i) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}(yf(\mathbf{x}) \leq 0). \tag{3}$$

To avoid degenerate solutions, $yf(x) = 0$ is considered not achieving the goal (otherwise a trivial prediction rule $f(x) = 0$ will always be optimal), which appears to be different from our general form. This can be addressed by introducing an arbitrarily small margin ϵ, and replacing $\mathbf{1}(yf(\mathbf{x}) \leq 0)$ by $\mathbf{1}(yf(\mathbf{x}) - \epsilon < 0)$. Then it is easy to see that the binary classification case is an example of goal scoring, where $\mathcal{G} = \{\mathcal{G}_1, \ldots, \mathcal{G}_m\}$, $\mathcal{G}_i = \{i\}$, $u_i = y_i \cdot f(x_i) - \epsilon$, and Eq. (3) is the goal miss to minimize. In plain language, each training data point (to classify correctly) is a goal, and we are concerned about the total numbers of points correctly classified. For notational simplicity, when $\mathcal{G}_i = \{i\}$, we drop input $\mathcal{G}$, and write $\varrho(u_1, \ldots, u_m)$ instead, i.e.,

$$\varrho(u_1, \ldots, u_m) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}[u_i < 0]. \tag{4}$$

Example 2: resource allocation

Suppose there are n projects, and d common resources. Let $R \in \mathbb{R}^{d\times n}$ be the resource requirement matrix, in the sense that R_{ij} is the amount of the i-th resource needed for project j. A project can be completed only if all its resource requirements are met. Let s_i be the available amount of the i-th resource to allocate among projects and x_{ij} be the amount of the i-th resource allocated to the j-th project, so as to maximize the number of projects to complete. This thus leads to the following optimization problem,

$$\begin{aligned} \text{Minimize:}_{u,x} \quad & \varrho(u_{11}, \ldots, u_{1n}, u_{21} \cdots, u_{dn}, \mathcal{G}) \\ \text{Subject to:} \quad & u_{ij} = x_{ij} - r_{ij}; \\ & \sum_{j=1}^{n} x_{ij} = s_i, \ \forall i \in \{1, \ldots, d\}; \\ & x_{ij} \geq 0, \end{aligned}$$

where $\mathcal{G} = \{\mathcal{G}_1, \ldots, \mathcal{G}_n\}$, and $\mathcal{G}_j = \{1j, 2j, \ldots, dj\}$.

The above example contains the *set packing* problem as a special case. Given a universe $\mathcal{D}$ containing d items, and a family $\mathcal{F}$ of n subsets of $\mathcal{U}$, a pack is a subfamily $\mathcal{C} \subseteq \mathcal{F}$ such that any two subsets in $\mathcal{C}$ contain no common elements of $\mathcal{D}$. The maximal set pack is to find a pack $\mathcal{C}$ with the largest cardinality. Let $r_{ij} = 1$ if the j-th subset contains i-th item, and equals zero otherwise; also let $s_i = 1$ for all i. Then the maximal set pack problem reduces to the resource allocation problem, where the set of completed projects are the subsets belonging to the maximal pack.

Example 3: joint chance constraint

Let $\mathbf{A}(\xi)\mathbf{x} \geq \mathbf{b}(\xi)$ be a set of d linear constraints on the variable $\mathbf{x}$, and the value of A and b depends on some uncertain parameter ξ that follows a distribution μ. One commonly used decision paradigm under this uncertainty is to require $\mathbf{x}$ to be feasible with certain probability w.r.t. μ, which leads to the below joint chance constraint

$$\Pr_{\xi\sim\mu}\left(\mathbf{A}(\xi)\mathbf{x} \geq \mathbf{b}(\xi)\right) \geq 1 - \beta,$$

for some pre-specified $\beta \in [0, 1]$ [13,35,36,40]. Since for a general μ, even evaluating whether a given $\mathbf{x}$ satisfies the joint chance constraint involves taking high-dimensional integral, which is typically computationally intractable, a popular practical solution [1,30,31], termed sample average approximation, is to replace μ by an empirical distribution of i.i.d. samples $\xi_1, \xi_2, \ldots, \xi_n$ drawn according μ, which leads to the following constraint:

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left(\mathbf{A}(\xi_i)\mathbf{x} \geq \mathbf{b}(\xi_i)\right) \geq 1 - \beta. \tag{5}$$

We notice that this is an example of goal scoring too. Let $A_j(\xi_i)$ and $b_j(\xi_i)$ be the j-th row of $\mathbf{A}(\xi_i)$, and the j-th entry of $\mathbf{b}(\xi_i)$, respectively, then Constraint (5) is equivalent to the following.

$$\varrho(u_{11}, \ldots, u_{1d}, u_{21} \cdots, u_{nd}, \mathcal{G}) \leq \beta$$
$$u_{ij} = A_j(\xi_i)x - b_j(\xi_i); \qquad i = 1, \ldots, n,\ j = 1, \ldots, d,$$

where $\mathcal{G} = \{\mathcal{G}_1, \ldots, \mathcal{G}_n\}$, and $\mathcal{G}_i = \{i1, \ldots, id\}$.

Example 4: graph problems

Many problems in graph theory can be cast as special cases of goal scoring. We provide below two examples to illustrate. In both examples we are given a graph $(\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is the set of vertex, and $\mathcal{E}$ is the set of edges, and we denote $|\mathcal{V}| = n$, and $|\mathcal{E}| = m$. Given the j-th edge, we denote by $V_1(j)$ and $V_2(j)$ the index of the two nodes connecting to it, with $V_1(j) \leq V_2(j)$.

– Maxcut. The max-cut problem seeks a subset S of the vertex set such that the number of edges between S and the complementary subset is the maximum. The max-cut problem can be formulated as the following target counting problem:

$$\begin{aligned} \text{Minimize:}_{x,u} \quad & \varrho(u_1^+, u_1^-, \cdots, u_n^+, u_n^-, \mathcal{G}) \\ \text{Subject to:} \quad & u_i^+ = x_i - 1; \\ & u_i^- = (1 - x_i) - 1; \\ & 0 \leq x_i \leq 1, \quad i = 1, \ldots, n, \end{aligned}$$

where $\mathcal{G}$ is defined as follows:

$$\mathcal{G} = \{\mathcal{G}_1^+, \mathcal{G}_1^-, \ldots, \mathcal{G}_m^+, \mathcal{G}_m^-\}, \mathcal{G}_j^+ = \{u_{V_1(j)}^+, u_{V_2(j)}^-\} \text{ and } \mathcal{G}_j^- = \{u_{V_1(j)}^-, u_{V_2(j)}^+\}.$$

In words, when $x_i = 1$ which corresponds to $u_i^+ \geq 0$, node i is assigned to S, and when $x_i = 0$ which corresponds to $u_i^- \geq 0$, node i is assigned to S^c. goal $\mathcal{G}_j^+$ corresponds to the first node of j-th edge is in S and the second node is in S^c, whereas goal $\mathcal{G}_j^-$ corresponds to the first node of j-th edge is in S^c and the second node is in S. Hence the allocation that maximizes the achieved goal is the maximal cut. Also notice that at the optimal solution, x will be binary, as otherwise both u_i^+ and u_j^+ are negative, and the number of attained goals cannot be larger.

– Maximum Independent Set. The maximum independent set problem seeks the largest subset of vertex, S, with no edges between any nodes in S. The maximum independent set problem is a special case of maximal set pack, where each edge is an item, and each node is a subset characterized by all its edges. Then, an independent set is a collection of subsets with no common elements, i.e., a set packing discussed above. Notice that when S is an independent set, then S^c is a vertex cover, and vice versa. Therefore, the minimum vertex cover problem is a special case of goal scoring.

3 Coherent loss function

The goal scoring problem includes as special cases some NP-hard problems (e.g., max-cut and maximum independent set), and is thus also NP-hard. While for problems of moderate size, discrete optimization based algorithms may be applied to solve the problem (either accurately, or approximately with guarantees, e.g., [5,29,46]), in general for large scale instances—for example in the case of binary classification where m can easily be of the order of millions—convex relaxation which replaces the 0–1 loss by a convex upper bound (i.e., the cumulative loss approach) appears to be the default method to apply thanks to its tractability and scalability.

In this paper we propose a different paradigm: instead of convexify each individual 0–1 loss, we convexify the sum to obtain a tighter approximation. More importantly, we derive this paradigm based on an axiomatic approach as opposed to in an ad hoc manner. In specific, we propose the notion of *coherent loss functions of goal miss* based on an axiomatic approach. Along the way, we show the existence of a "tight" coherent loss function which can achieve better approximation of the goal miss than *any* convex cumulative loss.

3.1 Salient properties and representation theorem

The notion of the coherent loss function is motivated from analyzing and axiomatizing the following five salient properties from the goal miss. Consider $\rho(\cdot, \cdot) : \mathbb{R}^m \times \mathcal{S} \to [0, 1]$.

Property 1 (Complete attainment) $\rho(\mathbf{u}, \mathcal{G}) = 0$ if and only if $\mathbf{u} \geq \mathbf{0}$.

The complete attainment property asserts that if all goals are achieved, then it is optimal.

Property 2 (Non attainment) If $\mathbf{u} < \mathbf{0}$, then $\rho(\mathbf{u}, \mathcal{G}) = 1$.

This property states that if no goal is achieved, then it is the worst and hence $\rho(\cdot, \cdot)$ achieves the maximal value.

Property 3 (Monotonicity) If $\min_{j \in \mathcal{G}_i} u_j \geq \min_{j \in \mathcal{G}_i} w_j$ for all $i = 1, \ldots, n$, then $\rho(\mathbf{u}, \mathcal{G}) \leq \rho(\mathbf{w}, \mathcal{G})$.

A decision is called "better" than other decisions under goal $\mathcal{G}_i$ if the smallest value of its subtasks under goal $\mathcal{G}_i$ is larger. Monotonicity requires that if a decision is better for every goal (but not necessarily for every subtask), then it is more desirable. When $G_i = \{i\}$ as in the binary classification case, this property reduces to $\mathbf{u} \geq \mathbf{v}$ implies $\rho(\mathbf{u}) \leq \rho(\mathbf{v})$.

Property 4 (Order invariance) For any permutations π of set $\{1, \ldots, m\}$ and τ of set $\{1, \ldots, n\}$, let $\tilde{\mathcal{G}}_{\tau(i)} = \{\pi(j) : j \in \mathcal{G}_i\}$ for $i = 1, \ldots, n$ and $\tilde{\mathbf{u}} = (u_{\pi(1)}, \ldots, u_{\pi(m)})^\top$, then we have $\rho(\mathbf{u}, \mathcal{G}) = \rho(\tilde{\mathbf{u}}, \tilde{\mathcal{G}})$.

Order invariance essentially states that neither the order of goals, nor the order of subtasks matter, as long as the relationship between the subtasks and goals remain the same. To put it another way, relabeling the nodes of bipartite graph induced by $\mathcal{G}$ does not change the loss. For all examples presented in Sect. 2, naturally the orders of subtasks and of goals are arbitrarily given, and only their mutual relationship matters. Take the classification problem as an example, since each data point is an i.i.d. draw, they should be treated equally.

Property 5 (Scale invariance) For all $\alpha > 0$, $\rho(\alpha\mathbf{u}, \mathcal{G}) = \rho(\mathbf{u}, \mathcal{G})$.

Scale invariance is a property that the goal miss function satisfies. It essentially means that changing the scale does not affect the preference between different decisions. While it may be debatable whether scale invariance is as necessary as other properties, indeed as we show later in this section, this property can be relaxed.

Based on the set of salient properties, we now define a set of functions called Coherent Loss Function (of Goal Miss), which are essentially tractable functions satisfying the set of salient properties.

Definition 1 *(Coherent Loss Function of Goal Miss)* A function $\rho(\mathbf{u}, \mathcal{G}) : \mathbb{R}^m \times \mathcal{S} \to [0, 1]$ is a *coherent loss function of goal miss* (CLF) if it satisfies Property 1 to 5, and is quasi-convex and lower semi-continuous w.r.t. $\mathbf{u}$.

Here, quasi-convexity and semi-continuity are introduced for tractability. These two properties are required for establishing connection between CLF and the coherent risk measure, leading to our first result—a (dual) representation theorem of any CLF (refer to Sect. 6.1 for more details).

Readers familiar with the risk measure concept [3,20] may recall that quasi-convexity is a critical property for risk measure motivated both from tractability and diversification. However, we want to stress that in goal achieving, although quasi-convexity is still related to diversification, diversification itself is no longer always desirable. Consider a case where half of the goals are just achieved (with no overshoot). If we diversify on all goals, then no goal is achieved. This is an example where diversification is indeed a bad choice. On the other hand, consider a case where half of the goal are over achieved, then diversification among all goals may lead to more goals being achieved, in which case it is desirable. In short, diversification is not necessarily desirable or undesriable property for goal achieving. Thus, quasi-convexity is mostly motivated by tractability in this paper, as the main motivation of the paper is to *approximate goal achieving in a tractable way.*

Before demonstrating the representation theorem of CLFs, we need the following definition first.

Definition 2 *(Admissible Class)* A class of sets $\mathbf{V}_k \subseteq \mathbb{R}^n$ parameterized by $k \in [0, 1]$ is called *admissible class*, if they satisfy the following properties:

1. For any $k \in [0, 1]$, $\mathbf{V}_k$ is a closed, convex cone, and is order invariant. Here, being order invariant means that $\mathbf{v} \in \mathbf{V}$ implies $\mathbf{Pv} \in \mathbf{V}$ for any $\mathbf{P} \in \mathcal{P}_n$;
2. $k \leq k'$ implies $\mathbf{V}_k \subseteq \mathbf{V}_{k'}$;
3. $\mathbf{V}_1 = \mathrm{cl}(\lim_{k\uparrow 1} \mathbf{V}_k)$ and $\mathbf{V}_0 = \lim_{k\downarrow 0} \mathbf{V}_k$.

4. $\mathbf{V}_1 = \mathbb{R}^n_+$;
5. For any $\lambda > 0$, we have $\lambda \mathbf{e}^n \in \mathbf{V}_0$.

Theorem 1 (Representation Theorem) *A function $\rho(\cdot, \cdot)$ is a CLF if and only if it can be written as*

$$\rho(\mathbf{u}, \mathcal{G}) = 1 - \sup\{k \in [0, 1] : \sup_{\mathbf{v} \in \mathbf{V}_k} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0, \tilde{\mathbf{u}} = (\min_{j \in \mathcal{G}_1} u_j, \cdots, \min_{j \in \mathcal{G}_n} u_j)^\top\}, \tag{6}$$

for an admissible class $\{\mathbf{V}_k\}$. Here sup *over an empty set is set as* 0.

3.2 Minimal coherent loss function

This section shows that among all CLF functions that *upper-bound the goal loss*, there exists a minimum (and hence best) one, which can be explicitly constructed.

Theorem 2 *Define $\overline{\rho}(\cdot, \cdot) : \mathbb{R}^m \times \mathcal{S} \mapsto [0, 1]$ as follows*

$$\overline{\rho}(\mathbf{u}, \mathcal{G}) = \frac{\max\{t : \sum_{i=1}^t \tilde{u}_{(i)} < 0\}}{n},$$

where $\tilde{u}_i = \min_{j \in \mathcal{G}_i} u_j$ for $i = 1, \ldots, n$, $\{\tilde{u}_{(i)}\}$ is a permutation of $\{\tilde{u}_i\}$ in a non-decreasing order, and max *over an empty set is taken as zero. Then the following holds.*

1. *$\overline{\rho}(\cdot, \cdot)$ is a CLF, and is an upper-bound of the* goal miss*, i.e., $\overline{\rho}(\mathbf{u}, \mathcal{G}) \geq \varrho(\mathbf{u}, \mathcal{G})$, for all $\mathbf{u} \in \mathbb{R}^m$ and $\mathcal{G} \in \mathcal{S}$.*
2. *Let $\overline{\mathbf{V}}_k \subset \mathbb{R}^n$ satisfy that if $k = 0$, then $\overline{\mathbf{V}}_k = \text{conv}\,\{\lambda \mathbf{e}^n | \lambda > 0\}$; and if $\frac{s}{n} < k \leq \frac{s+1}{n}$ for $s = 0, \ldots, n-1$, then*

$$\overline{\mathbf{V}}_k = \text{conv}\,\{\lambda \mathbf{e}_N | \, \forall \lambda > 0, \ \forall N : |N| = n - s\},$$

   where N is an index set. Then $\{\overline{\mathbf{V}}_k\}$ is an admissible class corresponding to $\overline{\rho}(\cdot, \cdot)$.
3. *$\overline{\rho}(\cdot, \cdot)$ is the* tightest *CLF bound. That is, if $\rho'(\cdot, \cdot)$ is a CLF function and satisfies $\rho'(\mathbf{u}, \mathcal{G}) \geq \varrho(\mathbf{u}, \mathcal{G})$ for all $\mathbf{u} \in \mathbb{R}^m$, then $\rho'(\mathbf{u}, \mathcal{G}) \geq \overline{\rho}(\mathbf{u}, \mathcal{G})$ for all $\mathbf{u} \in \mathbb{R}^m$ and $\mathcal{G} \in \mathcal{S}$.*

We next show that *scale invariance* can be relaxed. Indeed, compared to any quasi-convex upper bound of goal miss that satisfies the other properties, the minimal CLF is a tighter bound.

Theorem 3 *Let $\hat{\rho}(\mathbf{u}, \mathcal{G}) : \mathbb{R}^m \times \mathcal{S} \mapsto [0, 1]$ be a quasi-convex function w.r.t. $\mathbf{u}$ that satisfies complete attainment content, non attainment apathy, monotonicity, order invariance, and that $\hat{\rho}(\mathbf{u}, \mathcal{G}) \geq \varrho(\mathbf{u}, \mathcal{G})$. Then there exists a CLF $\rho(\cdot, \cdot)$ such that*

$$\varrho(\mathbf{u}, \mathcal{G}) \leq \overline{\rho}(\mathbf{u}, \mathcal{G}) \leq \rho(\mathbf{u}, \mathcal{G}) \leq \hat{\rho}(\mathbf{u}, \mathcal{G}), \quad \forall \mathbf{u} \in \mathbb{R}^m \text{ and } \mathcal{G} \in \mathcal{S}.$$

The detailed proof is deferred in Sect. 6 and here we provide a high-level sketch. The main idea is to construct such a function

$$\rho(\mathbf{u}, \mathcal{G}) \triangleq \lim_{\epsilon \downarrow 0} [\min_{\gamma > 0} \hat{\rho}((\mathbf{u} + \epsilon)/\gamma, \mathcal{G})],$$

and show that $\rho(\cdot, \cdot)$ is a CLF and $\hat{\rho}(\mathbf{u}, \mathcal{G}) \geq \rho(\mathbf{u}, \mathcal{G}) \geq \varrho(\mathbf{u}, \mathcal{G})$. Finally, since $\overline{\rho}(\cdot, \cdot)$ is the minimal CLF, this completes the proof.

One important property of $\overline{\rho}(\cdot, \cdot)$ is that it achieves better approximation of the goal miss function error than any *convex cumulative loss*.

Theorem 4 *If $f(\cdot)$ is a convex function and an upper bound of the 0–1 loss function, then for any $\mathbf{u} = (u_1, \ldots, u_m)$ and $\mathcal{G} \in \mathcal{S}$, we have $\varrho(\mathbf{u}, \mathcal{G}) \leq \overline{\rho}(\mathbf{u}, \mathcal{G}) \leq \frac{1}{n} \sum_{i=1}^{n} f(\tilde{u}_i)$ where $\tilde{u}_i = \min_{j \in \mathcal{G}_i} u_j$, for $i = 1, \ldots, n$.*

Proof Without loss of generality, assume $(\tilde{u}_1, \ldots, \tilde{u}_m)$ are in a non-decreasing order. Let $p \triangleq \max\{i : \tilde{u}_i < 0\}$ and $q \triangleq \max\{t : \sum_{i=1}^{t} \tilde{u}_i < 0\}$, then $\sum_{i=1}^{q} \tilde{u}_i = \sum_{i=1}^{p} \tilde{u}_i + \sum_{i=p+1}^{q} \tilde{u}_i < 0$. Since $f(\cdot)$ is convex and $f(x) \geq \mathbf{1}[x \leq 0]$, there exists $k \leq 0$ such that $f(x) \geq \max\{kx + 1, 0\}$ (this can be done for example by taking k as a subgradient of $f(x)$ at $x = 0$). If $k = 0$, then $\frac{1}{n} \sum_{i=1}^{n} f(\tilde{u}_i) \geq 1 \geq \overline{\rho}(\mathbf{u}, \mathcal{G})$, the theorem holds. Otherwise $k < 0$, we have

$$\begin{aligned} \sum_{i=1}^{n} f(\tilde{u}_i) &\geq \sum_{i=1}^{p} (k\tilde{u}_i + 1) + \sum_{i=p+1}^{n} f(\tilde{u}_i) = p + k \sum_{i=1}^{p} \tilde{u}_i + \sum_{i=p+1}^{n} f(\tilde{u}_i) \\ &> p - k \sum_{i=p+1}^{q} \tilde{u}_i + \sum_{i=p+1}^{n} f(\tilde{u}_i) \geq p + \sum_{i=p+1}^{q} (f(\tilde{u}_i) - k\tilde{u}_i). \end{aligned}$$

Note that $\tilde{u}_i \geq 0$ for $i = p + 1, \ldots, m$, then if $\tilde{u}_i \geq -\frac{1}{k}$, $f(\tilde{u}_i) - k\tilde{u}_i \geq -k\tilde{u}_i \geq 1$. Otherwise $f(\tilde{u}_i) - k\tilde{u}_i \geq k\tilde{u}_i + 1 - k\tilde{u}_i = 1$. Hence, $p + \sum_{i=p+1}^{q} (f(\tilde{u}_i) - k\tilde{u}_i) \geq p + (q - p) = q$. By the definition of $\overline{\rho}(\mathbf{u}, \mathcal{G})$, the theorem holds. □

3.3 Optimization with the coherent loss function

We now discuss the computational issue of optimization of the minimal CLF $\overline{\rho}(\cdot, \cdot)$. Indeed, we show that this can be converted to a tractable convex optimization problem. Specifically, for fixed $\mathcal{G}$, we consider the following problem on variables $(\mathbf{u}, \mathbf{w})$ where the minimal CLF is an objective function. the case where the minimal CLF appears in a constraint is essentially similar and hence omitted.

$$\begin{aligned} \min \quad & \overline{\rho}(\mathbf{u}, \mathcal{G}) \\ \text{s.t.} \quad & f_j(\mathbf{u}, \mathbf{w}) \leq 0, \quad j = 1, \ldots, k, \end{aligned} \tag{7}$$

where $f_j(\cdot, \cdot)$ are convex functions. We have the following theorem.

Theorem 5 *Assume that all the feasible solutions* $(\mathbf{u}, \mathbf{w})$ *to Problem* (7) *satisfy that* $\mathbf{u} > \mathbf{0}$ *or* $\mathbf{u} \ngeq \mathbf{0}$. *Let* $(\mathbf{s}^*, \mathbf{t}^*, h^*)$ *be an optimal solution to the following optimization problem:*

$$\begin{aligned} \min_{h,\mathbf{s},\mathbf{t}} \quad & \frac{1}{n}\sum_{i=1}^{n}\left[1 - \min_{j\in\mathcal{G}_i} s_j\right]_+ \\ s.t. \quad & hf_j(\mathbf{s}/h, \mathbf{t}/h) \le 0, \ \ j = 1, \ldots, k; \\ & h > 0. \end{aligned} \tag{8}$$

Then $(\mathbf{s}^*/h^*, \mathbf{t}^*/h^*)$ *is an optimal solution to Problem* (7).

Notice that $hf_j(\mathbf{s}/h, \mathbf{t}/h)$ is the perspective function of $f_j(\cdot, \cdot)$, and is hence jointly convex to $(h, \mathbf{s}, \mathbf{t})$ (Chapter 3.2.6 in [9]). Thus, Problem (8) is equivalent to a tractable convex optimization problem. Also notice that the assumptions in Theorem 5 are not restrictive: indeed if there exists a feasible $(\mathbf{u}^*, \mathbf{w}^*)$ with $\mathbf{u}^* \ge 0$, then solving Problem (7) is easy, as by *complete attainment content* such a $\mathbf{u}^*$ minimizes $\overline{\rho}(\mathbf{u}, \mathcal{G})$, and $\mathbf{u}^*$ can be obtained from the following convex problem:

$$\begin{aligned} \min \quad & 0 \\ \text{s.t.} \quad & \mathbf{u}_i \ge 0, \quad i = 1, \ldots, m \\ & f_j(\mathbf{u}, \mathbf{w}) \le 0, \quad j = 1, \ldots, k. \end{aligned}$$

From Theorem 5, when there is no $(\mathbf{u}, \mathbf{w})$ such that $\mathbf{u} \ge \mathbf{0}$ and $f_j(\mathbf{u}, \mathbf{w}) \le 0$ for $j = 1, \ldots, k$, Problem (7) is equivalent to minimizing the following optimization problem:

$$\begin{aligned} \min \quad & \Phi(\mathbf{u}, \mathcal{G}) \\ \text{s.t.} \quad & f_j(\mathbf{u}, \mathbf{w}) \le 0, \ \ j = 1, \ldots, k, \end{aligned} \tag{9}$$

where $\Phi(\mathbf{u}, \mathcal{G})$ is defined by

$$\Phi(\mathbf{u}, \mathcal{G}) \triangleq \min_{\gamma>0} \frac{1}{n}\sum_{i=1}^{n}\left[1 - \min_{j\in\mathcal{G}_i} u_j/\gamma\right]_+. \tag{10}$$

This leads to an alternative perspective of coherent loss function, stating that minimizing the minimal coherent loss function is equivalent to minimizing an *adaptive* "tighter" upper bound of the 0–1 loss function, and consequently the minimal coherent loss function achieves better approximation of the goal miss function than *any convex cumulative loss*.

Theorem 6 *Let* $\phi : \mathbb{R} \mapsto \mathbb{R}^+$ *be a non-increasing, convex function that satisfies*

$$\phi(c) \ge \mathbf{1}(c \le 0), \quad \forall c \in \mathbb{R}.$$

For all $\mathbf{u} \in \mathbb{R}^m$, *let* $\tilde{\mathbf{u}} = (\min_{j\in\mathcal{G}_1} u_j, \ldots, \min_{j\in\mathcal{G}_n} u_j)^\top$, *then we have*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(\tilde{u}_i \le 0) \le \Phi(\mathbf{u}, \mathcal{G}) \le \frac{1}{n}\sum_{i=1}^{n}\phi(\tilde{u}_i).$$

Proof Recall that the hinge-loss $\phi_1^*(c) \triangleq [1-c]_+$ is the tightest convex bound of 0–1 loss which has a derivative (or sub-gradient) -1 at $c=0$. That is, if a convex function $\phi(\cdot)$ satisfies $\phi(c) \geq \mathbf{1}(c \leq 0)$, for all c, and also satisfies $-1 \in \partial\phi(0)$, then $\phi_1^*(c) \leq \phi(c)$ for all c. Similarly, $\phi_\gamma^*(c) \triangleq \max[1-c/\gamma]_+$ is the tightest convex bound of 0–1 loss with a derivative $-1/\gamma$ at $x=0$. Since $\phi(\cdot)$ is non-increasing, it can not have positive derivative at $c=0$. Thus, $\Phi(\cdot,\cdot)$ is a tighter bound than any non-increasing, convex cumulative loss functions. □

4 Practical applications

Thus far we have investigated the proposed framework of coherent loss function for general goal scoring problem. In the remainder of the paper, we focus on applying this framework to the classification task and the resource allocation task discussed in Sect. 2. Notice that $\mathcal{G} = \{\{1\}, \{2\}, \ldots, \{m\}\}$ in classification (singleton goals), and $\mathcal{G} = \{\mathcal{G}_1, \ldots, \mathcal{G}_n\}$ in resource allocation where $\mathcal{G}_j = \{1j, 2j, \ldots, dj\}$ (non-singleton goals).

4.1 Example: linear SVM

We first consider the linear Support Vector Machines algorithm (SVMs) [8,17]. Given m training samples $(y_i, \mathbf{x}_i)_{i=1}^m$, the goal is to find a hyperplane that correctly classifies as many training samples as possible with a large margin, which leads to the following formulation:

$$\begin{aligned} \min \quad & \frac{1}{m}\sum_{i=1}^m \mathbf{1}[y_i(\mathbf{w}^\top \mathbf{x}_i + b) \leq 0] \\ \text{s.t.} \quad & \|\mathbf{w}\|_2 \leq C \end{aligned} \tag{11}$$

for a given $C > 0$. Since the objective function is non-convex, Problem (11) is an intractable problem. Hence, SVM uses the hinge-loss function $\phi_1^*(c) = [1-c]_+$ as a convex surrogate.

Following the proposed coherent loss function approach, we minimize the 0–1 loss function with margin $a \geq 0$: $\frac{1}{m}\sum_{i=1}^m \mathbf{1}[y_i(\mathbf{w}^\top \mathbf{x}_i + b) \leq a]$ and replace this objective function by the coherent loss function $\overline{\rho}(\mathbf{u})$ where $u_i = y_i(\mathbf{w}^\top \mathbf{x}_i + b) - a$.[1] Then we obtain the following formulation,

$$\begin{aligned} \min_{\mathbf{w},b,\gamma>0} \quad & \frac{1}{m}\sum_{i=1}^m [1 - (y_i(\mathbf{w}^\top \mathbf{x}_i + b) - a)/\gamma]_+ \\ \text{s.t.} \quad & \|\mathbf{w}\|_2 \leq C. \end{aligned} \tag{12}$$

[1] The margin a is introduced to ensure that Theorem 5 holds. Notice that the hinge-loss approximation with or without the margin leads to the same formulation of the standard SVM.

As discussed above, we can change variables $h = 1/\gamma$, $\hat{\mathbf{w}} = \mathbf{w}/\gamma$ and $\hat{b} = b/\gamma$, and simplify Formulation (12) as the following:

$$\begin{aligned} \min_{\hat{\mathbf{w}},\hat{b},h>0} \quad & \frac{1}{m}\sum_{i=1}^{m}[1 + ah - y_i(\hat{\mathbf{w}}^\top \mathbf{x}_i + \hat{b})]_+ \\ \text{s.t.} \quad & \|\hat{\mathbf{w}}\|_2 \le hC. \end{aligned}$$

An interesting observation is that this formulation is also equivalent to the robust formulation of SVM [41], which provides another interpretation for the coherent loss function.

Proposition 1 *Problem (12) is equivalent to the following optimization problem:*

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \sum_{i=1}^{m}\xi_i \\ s.t. \quad & \inf_{\tilde{\mathbf{x}}_i\sim(\mathbf{x}_i,\mathbf{I})} \mathbb{P}[y_i(\mathbf{w}^\top\tilde{\mathbf{x}}_i + b) \ge 1 - \xi_i] \ge 1 - \kappa,\ i = 1,\ldots,m, \end{aligned} \tag{13}$$

where $\tilde{\mathbf{x}}_i \sim (\mathbf{x}_i, \mathbf{I})$ *denotes a family of distributions which have a common mean* $\mathbf{x}_i$ *and covariance* $\mathbf{I}$*, and* $\kappa = a^2/(a^2 + C^2)$.

Proof Theorem 1 in [41] shows that Problem (13) is equivalent to

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^\top\tilde{\mathbf{x}}_i + b) \ge 1 - \xi_i + \gamma\|\mathbf{w}\|_2,\ i = 1,\ldots,m, \\ & \xi_i \ge 0,\ i = 1,\ldots,m, \end{aligned}$$

where $\gamma = \sqrt{\kappa/(1-\kappa)}$. When $\kappa = a^2/(a^2 + C^2)$, we have $\gamma = a/C$, which implies that the formulation above is equivalent to

$$\min_{\mathbf{w},b,\xi} \quad \sum_{i=1}^{m}[1 + \frac{a}{C}\|\mathbf{w}\|_2 - y_i(\mathbf{w}^\top\tilde{\mathbf{x}}_i + b)]_+.$$

Therefore, by moving $\frac{a}{C}\|\mathbf{w}\|_2$ into the constraint, we obtain this result. □

The approach can be extended to the case where one may like to impose additional constraints on $\mathbf{w}$. For instance, if the first feature is measured from a less reliable source, then an ideal classification rule should discount the importance of the first feature, by imposing a constraint like $|w_1| \le 0.001$. Thus, the linear classification problem becomes

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}[y_i(\mathbf{w}^\top \mathbf{x}_i + b) \le a] \\ \text{s.t.} \quad & \|\mathbf{w}\|_2 \le C \\ & A\mathbf{w} \le \mathbf{d}. \end{aligned}$$

Using the minimal coherent loss to replace the objective function, and simplifying the resulting formulation, we obtain the following second order cone program

$$\begin{aligned} \min_{\hat{\mathbf{w}},\hat{b},h>0} \quad & \frac{1}{m}\sum_{i=1}^{m} [1 + ah - y_i(\hat{\mathbf{w}}^\top \mathbf{x}_i + \hat{b})]_+ \\ \text{s.t.} \quad & \|\hat{\mathbf{w}}\|_2 - Ch \le 0 \\ & A\hat{\mathbf{w}} \le \mathbf{d}h. \end{aligned}$$

Finally, we remark that the coherent loss approach can be kernelized, since "the kernel trick" [39] still holds if the coherent loss function is used.

4.2 Example: multi-class SVM

Our second example is the multi-class classification problem. The main idea of previous approaches [19,27,28] of multi-class SVMs is solving one single regularization problem by imposing a penalty on the values of $f_y(\mathbf{x}) - f_z(\mathbf{x})$ for sample $(\mathbf{x}, y)$ where $f_y(\cdot)$ and $f_z(\cdot)$ are decision function for class y and z, respectively. Suppose that the training samples are drawn from k different classes and the decision function $f_y(\mathbf{x}) = \mathbf{w}_y^\top \mathbf{x} + b_y$ for each $y = 1, \ldots, k$. Consider the following 0–1 loss penalty formulation:

$$\begin{aligned} \min_{f_i} \quad & \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\left[\min_{z\in[k], z\neq y_i} \{f_{y_i}(\mathbf{x}_i) - f_z(\mathbf{x}_i)\} \le a\right] \\ \text{s.t.} \quad & G_i(\mathbf{w}_i) \le C,\ i = 1, \ldots, k \\ & \sum_{i=1}^{k} f_i = 0, \end{aligned}$$

where $\sum_{i=1}^{k} f_i = (\sum_{i=1}^{k} \mathbf{w}_i, \sum_{i=1}^{k} b_i)$, $G_i(\cdot)$ is convex (e.g., $G_i(\cdot) = \|\cdot\|_2$) and margin $a \ge 0$. Notice that the objective function is a goal miss function, and hence we can apply the coherent loss function approach to make an approximation:

$$\begin{aligned} \min_{f_i,\gamma>0} \quad & \frac{1}{m}\sum_{i=1}^{m} \left[1 - \frac{\min_{z\in[k], z\neq y_i}\{f_{y_i}(\mathbf{x}_i) - f_z(\mathbf{x}_i)\} - a}{\gamma}\right]_+ \\ \text{s.t.} \quad & G_i(\mathbf{w}_i) \le C,\ i = 1, \ldots, k \\ & \sum_{i=1}^{k} f_i = 0, \end{aligned}$$

which can be simplified as the following:

$$\begin{aligned}\min_{\hat{f}_i, h>0} \quad & \frac{1}{m}\sum_{i=1}^{m}\left[1+ah+\max_{z\in[k], z\neq y_i}\{\hat{f}_z(\mathbf{x}_i)-\hat{f}_{y_i}(\mathbf{x}_i)\}\right]_+ \\ \text{s.t.} \quad & hG_i(\hat{\mathbf{w}}_i/h)\leq hC,\ i=1,\ldots,k \\ & \sum_{i=1}^{k}\hat{f}_i=0,\end{aligned}$$

where $\hat{f}_i(\mathbf{x})=\hat{\mathbf{w}}_i^\top\mathbf{x}+\hat{b}_i$. Clearly, this is a convex optimization problem and can be solved efficiently.

4.3 Example: resource allocation

Our third example is the resource allocation problem. Given n projects, d common resources and a resource requirement matrix $R\in\mathbb{R}^{d\times n}$ where r_{ij} is the amount of the i-th resource needed for project j, the goal is to maximize the number of projects to complete. Suppose that s_i is the available amount of the i-th resource to allocate among projects and let x_{ij} be the amount of the i-th resource allocated to the j-th project, then this problem can be formulated by

$$\begin{aligned}\min_{x,u} \quad & \frac{1}{n}\sum_{j=1}^{n}\mathbf{1}\left[u_{ij}<0, \exists i\in\mathcal{G}_j=\{1,\ldots,d\}\right] \\ \text{s.t.} \quad & u_{ij}=x_{ij}-r_{ij}; \\ & \sum_{j=1}^{n}x_{ij}=s_i,\ \forall i\in\{1,\ldots,d\}; \\ & x_{ij}\geq 0,\ \forall i\in\{1,\ldots,d\}, j\in\{1,\ldots,n\}.\end{aligned}$$

The 0–1 loss can be approximated by the proposed coherent loss function:

$$\begin{aligned}\min_{x,u,\gamma>0} \quad & \frac{1}{n}\sum_{j=1}^{n}\left[1-\frac{\min_{i\in\mathcal{G}_j}u_{ij}}{\gamma}\right]_+ \\ \text{s.t.} \quad & u_{ij}=x_{ij}-r_{ij}; \\ & \sum_{j=1}^{n}x_{ij}=s_i,\ \forall i\in\{1,\ldots,d\}; \\ & x_{ij}\geq 0,\ \forall i\in\{1,\ldots,d\}, j\in\{1,\ldots,n\}.\end{aligned}$$

We can change variables $y = 1/\gamma$ and $\hat{u}_{ij} = u_{ij}/\gamma$, then the formulation above can be simplified as:

$$\begin{aligned} \min_{\hat{u}, y>0} \quad & \frac{1}{n}\sum_{j=1}^{n}\left[1 - \min_{i\in\mathcal{G}_j}\hat{u}_{ij}\right]_+ \\ \text{s.t.} \quad & \sum_{j=1}^{n}\hat{u}_{ij} = y\left(s_i - \sum_{j=1}^{n} r_{ij}\right), \ \forall i \in \{1,\ldots,d\}; \\ & \hat{u}_{ij} \geq -y r_{ij}, \ \forall i \in \{1,\ldots,d\}, j \in \{1,\ldots,n\}. \end{aligned}$$

Since s_i and r_{ij} are constants, this problem is a convex optimization problem.

4.4 Robustness

One advantage of the proposed coherent loss function approach in classification, which we illustrate with an example, is that it can be more robust to outlying data. Let $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^{100}$ be the followings: $\mathbf{u}_1 = (-1000, 1000, \cdots, 1000)$, and $\mathbf{u}_2 = (+1, -1, +1, -1, \ldots, +1, -1)$. In this case, $\mathbf{u}_2$ appears to be a less favorable classification since 50% of samples are misclassified. It is easy to check that $\mathbf{u}_1$ incurs a much larger hinge-loss than $\mathbf{u}_2$ (hinge-loss for $\mathbf{u}_1$ is 10.01 while hinge-loss for $\mathbf{u}_2$ is 1), even though only one sample is misclassified. In contrast, the coherent loss of $\mathbf{u}_1$ is no more than 0.02 (take $\gamma = 1/1000$), and that of $\mathbf{u}_2$ is at least 0.5 (since 50% samples are misclassified, and the coherent loss is an upper bound). Thus, the coherent loss is more robust in this example, partly because it strictly better approximates the 0–1 loss, and hence is less affected by large outliers. See Fig. 1.

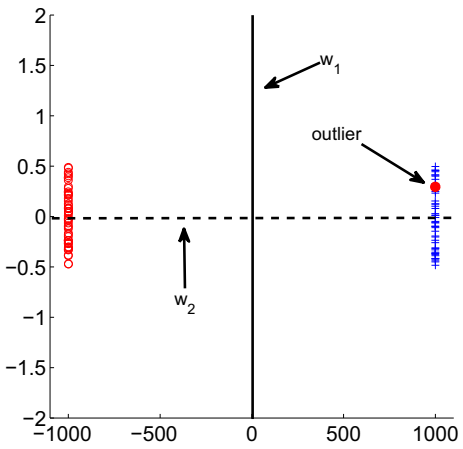

Fig. 1 Illustration of the effect of outliers to the cumulative loss versus the coherent loss. Here, $\mathbf{w}_1$ has a margin $\mathbf{u}_1$, and $\mathbf{w}_2$ has a margin $\mathbf{u}_2$. The cumulative loss approach will pick $\mathbf{w}_2$, where the proposed method will pick $\mathbf{w}_1$, which is a better classification

4.5 Connection with bPOE

Recently, Norton et al. [32,33] proposed a new concept for optimization of tail probabilities called buffered probability of exceedance (bPOE). bPOE is an alternative measure of tail probability which can be viewed as a generalization of the buffered probability of Failure introduced by Rockafellar [37]. The advantage of bPOE compared with POE and AUC is that minimizing bPOE is computationally tractable and often reduces to convex or even linear programming. With this computational benefit, the authors applied bPOE to classification problems and showed that the soft margin support vector machine is equivalent to minimization of bPOE.

More specifically, given a real valued random variable X with continuous distribution, let the corresponding quantile function at probability level α be $q_\alpha(X) = \min\{z : \mathbb{P}[X \leq z] \geq \alpha\}$ and conditional value at risk (CVaR) be $\text{CVaR}_\alpha(X) = \mathbb{E}[X|X \geq q_\alpha(X)]$. bPOE is defined as the inverse of CVaR, i.e., bPOE of X at threshold z is:

$$\bar{p}_z(X) = \begin{cases} \max\{1-\alpha : \text{CVaR}_\alpha(X) \geq z\}, & \text{if } z \leq \sup X \\ 0, & \text{otherwise.} \end{cases}$$

As shown in [33], $\bar{p}_z(X)$ is equivalent to

$$\bar{p}_z(X) = \min_{\gamma < z} \frac{\mathbb{E}\left[X - \gamma\right]_+}{z - \gamma}. \tag{14}$$

Given a set of training examples $\{x_1, x_2, \ldots, x_n\}$ independently sampled from the distribution of X, we consider the following empirical loss of (14) with threshold $z = 0$:

$$\bar{p}_{emp}(X) = \min_{\gamma > 0} \frac{1}{n} \sum_{i=1}^{n} [1 + x_i/\gamma]_+. \tag{15}$$

Clearly, the empirical bPOE (15) is the same as $\Phi(\mathbf{u}, \mathcal{G})$ defined by Eq. (10) with singleton goals $\mathcal{G} = \{\{1\}, \{2\}, \ldots, \{n\}\}$ and $u_i = -x_i$, implying that minimizing the empirical bPOE is equivalent to minimizing the coherent loss function with n singleton goals. Section 4.1 shows that the soft margin SVM can be derived by applying the coherent loss function in binary classification tasks, which reproves the equivalence between minimization of bPOE and the soft margin SVM from our CLF perspective.

4.6 Statistical interpretation

Finally, we develop some statistical understanding of minimizing the minimal coherent loss function for classification. As standard in learning theory, we assume that the training samples are drawn i.i.d. from an unknown distribution $\mathbb{P}$, and the goal is to find a predictor $f(\cdot)$ such that the classification error of f given below is as small as possible:

$$L(f(\cdot)) = \mathbb{E}_{(\tilde{\mathbf{x}}, \tilde{y}) \sim \mathbb{P}}[I(f(\tilde{\mathbf{x}}), \tilde{y})].$$

Here $(\tilde{\mathbf{x}}, \tilde{y}) \sim \mathbb{P}$ means sample $(\tilde{\mathbf{x}}, \tilde{y})$ follows the distribution $\mathbb{P}$, and $I(f(\tilde{\mathbf{x}}), \tilde{y}) = \mathbf{1}[\tilde{y} f(\tilde{\mathbf{x}}) \leq 0]$. Recall that in the binary classification case, $\mathcal{G} = \{\{1\}, \{2\}, \ldots, \{m\}\}$, and hence minimizing the minimal coherent loss function is equivalent to minimizing the following function

$$\Phi(\mathbf{u}) = \min_{\gamma > 0} \frac{1}{m} \sum_{i=1}^{m} \phi_\gamma(u_i),$$

where $\phi_\gamma(u) = [1 - u/\gamma]_+$. Let $\eta(\mathbf{x}) = \mathbb{P}[\tilde{y} = 1 | \tilde{\mathbf{x}} = \mathbf{x}]$ be the probability that sample $\mathbf{x}$ belongs to the first class, then the optimal Bayes error $L^* = L(2\eta(\cdot) - 1)$. We now develop an upper bound of the difference between $L(f(\cdot))$ and L^* by using similar techniques in [48].

For fixed γ, denote the expected loss of $f(\cdot)$ w.r.t $\phi_\gamma(\cdot)$ by

$$Q_\gamma(f(\cdot)) = \mathbb{E}_{(\tilde{\mathbf{x}}, \tilde{y}) \sim \mathbb{P}}[\phi_\gamma(\tilde{y} f(\tilde{\mathbf{x}}))],$$

and define two quantities

$$Q_\gamma(\eta, f) = \eta \phi_\gamma(f) + (1 - \eta)\phi_\gamma(-f), \ \Delta Q_\gamma(\eta, f) = Q_\gamma(\eta, f) - Q_\gamma(\eta, f_\gamma^*(\eta)),$$

where $f_\gamma^*(\eta) = \arg\min_f Q_\gamma(\eta, f)$. By simple calculation, we know that $f_\gamma^*(\eta) = \text{sign}(2\eta - 1)\gamma$ when $\phi_\gamma(u) = [1 - u/\gamma]_+$. Then we have the following lemma.

Lemma 1 *For $\gamma > 0$, we have $\Delta Q_\gamma(\eta, 0) = |2\eta - 1|$.*

Proof From the definition of $Q_\gamma(\eta, f)$ and $\Delta Q_\gamma(\eta, f)$, we have

$$\begin{aligned}
\Delta Q_\gamma(\eta, f) &= \eta(\phi_\gamma(f) - \phi_\gamma(f_\gamma^*(\eta))) + (1 - \eta)(\phi_\gamma(-f) - \phi_\gamma(-f_\gamma^*(\eta))) \\
&= \eta[1 - f/\gamma]_+ + (1 - \eta)[1 + f/\gamma]_+ - \eta[1 - \text{sign}(2\eta - 1)]_+ \\
&\quad - (1 - \eta)[1 + \text{sign}(2\eta - 1)]_+ \\
&= \eta[1 - f/\gamma]_+ + (1 - \eta)[1 + f/\gamma]_+ - 1 + |2\eta - 1|.
\end{aligned}$$

This implies that $\Delta Q_\gamma(\eta, 0) = |2\eta - 1|$. □

By applying Lemma 1, we can bound the classification error of $f(\cdot)$ w.r.t $\phi_\gamma(\cdot)$ in terms of $\mathbb{E}_{\tilde{\mathbf{x}}} \Delta Q_\gamma(\eta(\tilde{\mathbf{x}}), f(\tilde{\mathbf{x}}))$.

Theorem 7 *For any $\gamma > 0$ and any measurable function $f(x)$, we have*

$$L(f(\cdot)) - L^* \leq \mathbb{E}_{\tilde{\mathbf{x}}} \Delta Q_\gamma(\eta(\tilde{\mathbf{x}}), f(\tilde{\mathbf{x}})) = \mathbb{E}_{\tilde{\mathbf{x}}}[Q_\gamma(\eta(\tilde{\mathbf{x}}), f(\tilde{\mathbf{x}})) + |2\eta(\tilde{\mathbf{x}}) - 1| - 1].$$

Proof By definition of $L(\cdot)$, it is easy to verify that

$$\begin{aligned}
L(f(\cdot)) - L(2\eta(\cdot) - 1) &= \mathbb{E}_{\eta(X) \geq 0.5, f(X) < 0}(2\eta(X) - 1) \\
&\quad + \mathbb{E}_{\eta(X) < 0.5, f(X) \geq 0}(1 - 2\eta(X)) \\
&\leq \mathbb{E}_{(2\eta(X) - 1) f(X) \leq 0} |2\eta(X) - 1|.
\end{aligned}$$

From Lemma 1, i.e., $\Delta Q_\gamma(\eta, 0) = |2\eta - 1|$, we have

$$L(f(\cdot)) - L^* \leq \mathbb{E}_{(2\eta(\tilde{\mathbf{x}})-1)f(\tilde{\mathbf{x}})\leq 0} \Delta Q_\gamma(\eta(\tilde{\mathbf{x}}), 0).$$

To complete the proof, since $\Delta Q_\gamma(\eta, f) = Q_\gamma(\eta, f) - Q_\gamma(\eta, f_\gamma^*(\eta))$, it suffices to show that $Q_\gamma(\eta(\mathbf{x}), 0) \leq Q_\gamma(\eta(\mathbf{x}), f(\mathbf{x}))$ for all $\mathbf{x}$ such that $(2\eta(\mathbf{x}) - 1)f(\mathbf{x}) \leq 0$. To see this, we consider three scenarios:

- $\eta > 0.5$: We have $f_\gamma^*(\eta) = \text{sign}(2\eta - 1)\gamma > 0$. In addition, $(2\eta - 1)f \leq 0$ implies $f \leq 0$. Since $0 \in [f, f_\gamma^*(\eta)]$ and the convexity of $Q_\gamma(\eta, f)$ w.r.t. f, we have $Q_\gamma(\eta, 0) \leq \max\{Q_\gamma(\eta, f), Q_\gamma(\eta, f_\gamma^*(\eta))\} = Q_\gamma(\eta, f)$.
- $\eta < 0.5$: In this case we have $f_\gamma^*(\eta) < 0$ and $f \geq 0$, which leads to $0 \in [f_\gamma^*(\eta), f]$, which implies $Q_\gamma(\eta, 0) \leq \max\{Q_\gamma(\eta, f), Q_\gamma(\eta, f_\gamma^*(\eta))\} = Q_\gamma(\eta, f)$.
- $\eta = 0.5$: Note that $f_\gamma^* = 0$, which implies that $Q_\gamma(\eta, 0) \leq Q_\gamma(\eta, f)$ for all f.

From the proof of Lemma 1, we have $\Delta Q_\gamma(\eta, f) = Q_\gamma(\eta, f) + |2\eta - 1| - 1$. Hence the theorem holds. □

Corollary 1 *For any measurable function $f(x)$,*

$$L(f(\cdot)) - L^* \leq \min_{\gamma>0} \mathbb{E}_{\tilde{\mathbf{x}}}[Q_\gamma(\eta(\tilde{\mathbf{x}}), f(\tilde{\mathbf{x}})) + |2\eta(\tilde{\mathbf{x}}) - 1| - 1]. \tag{16}$$

Proof Since Theorem 7 holds for any $\gamma > 0$, we obtain this corollary. □

For samples $\{\mathbf{x}_i, y_i\}_{i=1}^m$, since $\eta(\mathbf{x}_i) = y_i \in \{1, -1\}$, the empirical estimation of the bound in (16) is $\Phi(\mathbf{u}) = \min_{\gamma>0} \frac{1}{m} \sum_{i=1}^m \phi_\gamma(u_i)$ where $u_i = y_i f(\mathbf{x}_i)$, which implies that minimizing the coherent loss function is equivalent to minimizing the empirical bound of the difference between $L(f(\cdot))$ and L^*. Note that the bound for the cumulative loss is essentially $\gamma = 1$, while the bound for the coherent loss is the minimum over γ which is much tighter.

5 Simulations

We first report some numerical simulation results on classification in this section to illustrate the proposed approach. Besides the regularization constraints (e.g., $\|\mathbf{w}\| \leq C$ for binary-class SVMs and $\|\mathbf{w}_i\| \leq C, i = 1, \ldots, k$ for multi-class SVMs), we consider the case where additional linear constraints are also imposed on the coefficient $\mathbf{w}$. For clarity, we choose a simple additional constraint $\|\mathbf{A}\mathbf{w}\|_\infty \leq T$ to compare the performance of the cumulative loss formulation (SVM) and our coherent loss formulation (CLF) for binary-class and multi-class classification, where $\mathbf{A} = [\mathbf{I}_k, \mathbf{0}] \in \mathbb{R}^{k\times n}$. In other words, the constraint ensures that the maximum of the first k elements of $\mathbf{w}$ is bounded by T. We now compare their performance under two cases: 1) k is fixed, T varies; 2) T is fixed, k varies.

Three binary-class datasets "Breast cancer", "Ionosphere" and "Diabetes", and two multi-class datasets "Wine" and "Iris" from UCI [4] are used, where we randomly pick 50% of data as the training sample, 20% as the validation sample, and the rest as the

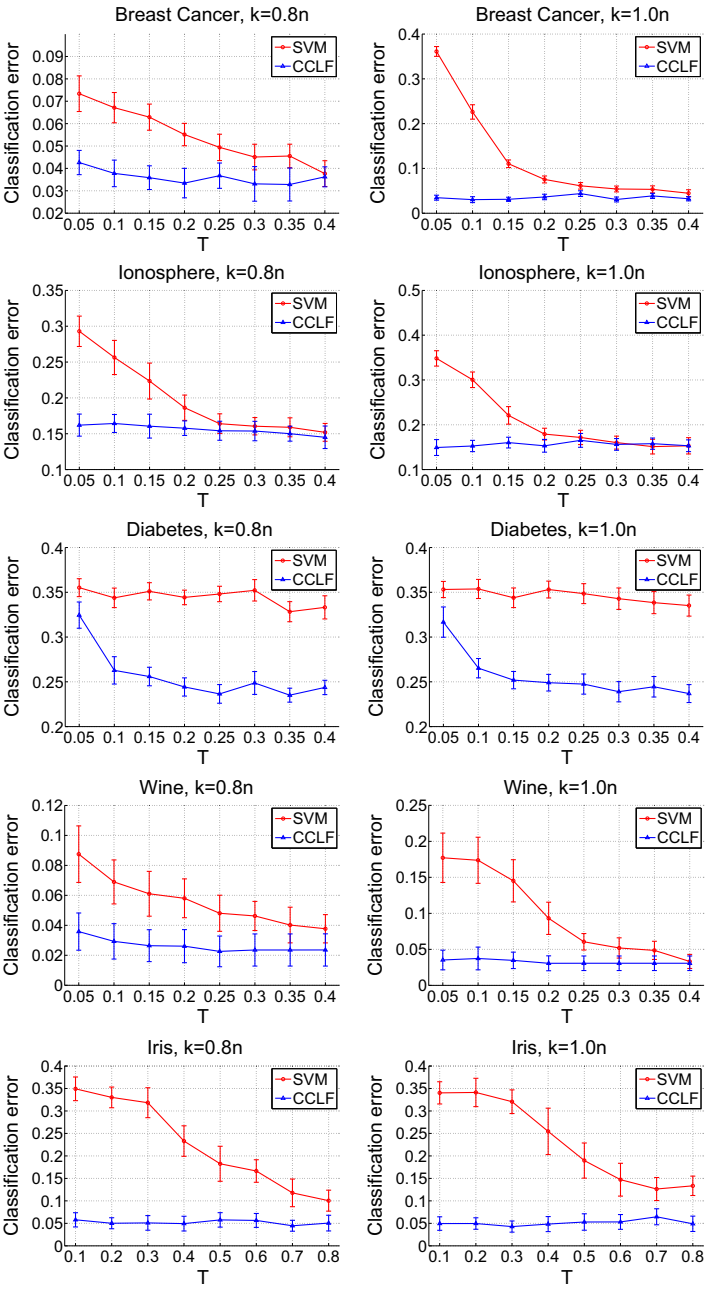

Fig. 2 Performance comparison of *cumulative loss approach* versus *coherent loss approach*. Left and right columns report the classification errors for the two cases $k = 0.8n$ and $k = n$ (recall that k and n are the numbers of the rows and columns of matrix $\mathbf{A}$, respectively). The five rows, from top to bottom, report results for *Breast Cancer*, *Ionosphere*, *Diabetes*, *Wine* and *Iris*, respectively

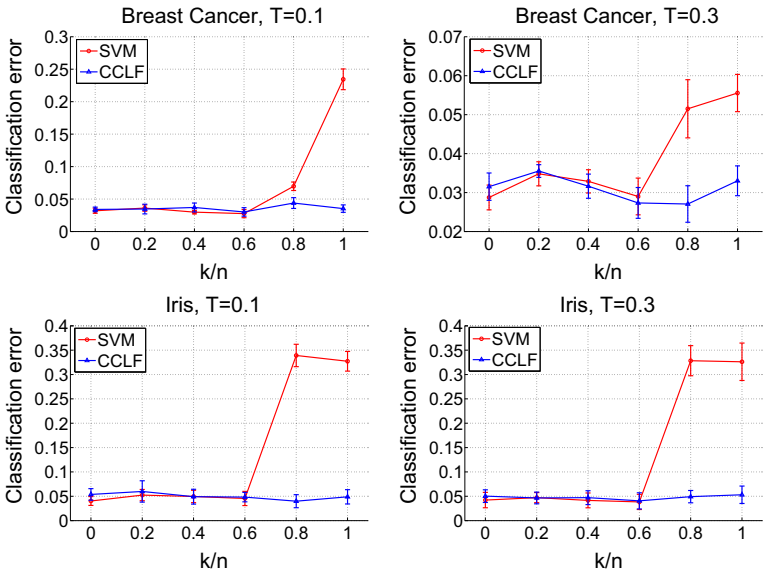

Fig. 3 Performance comparison of *cumulative loss approach* versus *coherent loss approach* where bound T is fixed and the fraction k/n varies from 0.0 to 1.0. Left and right columns report the classification errors for the two cases $T = 0.1$ and $T = 0.3$

testing sample. For the cumulative loss function approach, parameter C is determined by cross-validation. For the coherent loss function approach, parameter C is fixed while parameter a is determined by cross-validation. For each T, we repeated the experiments 20 times and computed the average classification errors. To solve the resulting optimization problems, we use CVX [23,24], and Gurobi [25] as the solver.

Figure 2 shows the simulation results under fixed k. Clearly, when additional constraints are imposed, it appears that the coherent loss approach consistently outperforms the cumulative loss approach. When T is small, the cumulative loss approach performs much worse. When T becomes large, its performance becomes closer to the coherent loss approach. Figure 3 provides the results under fixed T, which shows that the coherent loss and cumulative loss approaches have similar performance when k/n is small but the coherent loss approach outperforms the cumulative loss approach when k/n is large. We believe that these phenomena are due to the fact that the coherent loss is a better approximation for the empirical classification error.

We then report the simulation results on resource allocation. Suppose that we are given $n = 5$ projects, $d = 3$ resources and the available amount of the i-th resource $s_i = 10$ for $i = 1, 2, 3$. Recall that our coherent loss function approach tries to solve the following optimization problem:

$$\begin{aligned}\min_{\hat{u},y>0} \quad & \frac{1}{n}\sum_{j=1}^{n}\left[1-\min_{i\in\mathcal{G}_j}\hat{u}_{ij}\right]_+ \\ \text{s.t.} \quad & \sum_{j=1}^{n}\hat{u}_{ij} = y\left(s_i-\sum_{j=1}^{n}r_{ij}\right),\ \forall i\in\{1,\ldots,d\}; \\ & \hat{u}_{ij}\geq -yr_{ij},\ \forall i\in\{1,\ldots,d\},\ j\in\{1,\ldots,n\}.\end{aligned}$$

The traditional hinge loss function approach formulates this task as follows:

$$\begin{aligned}\min_{u} \quad & \frac{1}{n}\sum_{j=1}^{n}\left[1-\min_{i\in\mathcal{G}_j}u_{ij}\right]_+ \\ \text{s.t.} \quad & \sum_{j=1}^{n}u_{ij} = s_i-\sum_{j=1}^{n}r_{ij},\ \forall i\in\{1,\ldots,d\}; \\ & u_{ij}\geq -r_{ij},\ \forall i\in\{1,\ldots,d\},\ j\in\{1,\ldots,n\}.\end{aligned}$$

In order to compare the performance of these two approaches, we check whether the allocation result generated by our CLF approach allows more projects to complete. In other words, given a solution u, we compute the number of projects that can be completed: $\sum_{j=1}^{n}\mathbf{1}\left[u_{ij}>=0,\forall i\in\{1,\ldots,d\}\right]$. These two problems are solved via the CVX [23,24] solver.

In order to compare the two approaches, we first consider the following resource requirement matrix R:

	Project 1	Project 2	Project 3	Project 4	Project 5
Resource 1	2	3	2	3	2
Resource 2	2	2	2	2	3
Resource 3	2	3	3	2	3

Obviously, the optimal solution is to allocate the resources to the first four projects, which means the maximum number of the project that can be completed is 4. However, the hinge loss approach obtains the following solution:

	Project 1	Project 2	Project 3	Project 4	Project 5
Resource 1	1.6	2.6	1.6	2.6	1.6
Resource 2	1.8	1.8	1.8	1.8	2.8
Resource 3	1.4	2.4	2.4	1.4	2.4

Clearly, none of the projects can be completed with this allocation, which means that the hinge loss approach does not find a proper solution for this task. Instead, our CLF approach generates the following solution:

	Project 1	Project 2	Project 3	Project 4	Project 5
Resource 1	1.3349	3.3629	0.8201	2.0772	2.4037
Resource 2	1.5113	2.1678	0.9965	1.1879	4.1353
Resource 3	1.0922	3.0444	1.5076	0.7689	3.5854

With this allocation, Project 2 and Project 5 can be completed. Although this solution is still sub-optimal, this example shows that the coherent loss function approximates the 0–1 loss better than the hinge loss function.

To investigate whether the CLF approach consistently outperforms the hinge loss approach, we randomly generate the resource requirement matrix R (The entries of R are randomly sampled from a uniform distribution over $\{1, 2, 3\}$) and compute the number of the projects that can be completed. We repeat this procedure 100 times and report the average number of the projects to complete for each approach. For the CLF approach, this experiment considers a slightly different formulation:

$$\begin{aligned} \min_{\hat{u}, y \geq b} \quad & \frac{1}{n} \sum_{j=1}^{n} \left[1 - \min_{i \in \mathcal{G}_j} \hat{u}_{ij} \right]_+ \\ \text{s.t.} \quad & \sum_{j=1}^{n} \hat{u}_{ij} = y \left(s_i - \sum_{j=1}^{n} r_{ij} \right), \ \forall i \in \{1, \ldots, d\}; \\ & \hat{u}_{ij} \geq -y r_{ij}, \ \forall i \in \{1, \ldots, d\}, j \in \{1, \ldots, n\}. \end{aligned}$$

The only difference is the constraint imposed on the variable y, namely, $y \geq b$ where b is a constant. This lower bound on y controls how well the CLF approximates the 0–1 loss function, e.g., $b = 0$ leads to "full" CLF and $b = 1$ is close to the hinge loss. Table 1 shows the performance of the hinge loss approach and the different settings of the CLF approach, from which we can observe that when b goes to zero, the solutions generated by the CLF approach allow more projects to complete. This is consistent with our theoretical analysis.

This experiment shows that our CLF approach is able to provide a better approximation to the 0–1 loss than the hinge loss approach, which is able to solve real-world decision making tasks better.

6 Proofs of technical results

In this section we provide proofs to Theorem 1, 2, 3 and 5.

Table 1 The comparison between the CLF approach and the hinge loss approach

	Hinge	$b = 1$	$b = 0.1$	$b = 0.01$	$b = 0.001$	$b = 0.0001$
Average number	1.25	1.25	1.25	1.34	1.98	2.56

The average number of projects that can be completed is reported

6.1 Proof of Theorem 1

Proof **Step 1—the "if" part** Given a function $\rho(\mathbf{u}, \mathcal{G}) = 1 - \sup\{k \in [0,1] : \sup_{\mathbf{v}\in\mathbf{V}_k}(-\mathbf{v}^\top\tilde{\mathbf{u}}) \le 0, \tilde{\mathbf{u}} = (\min_{j\in\mathcal{G}_1} u_j, \ldots, \min_{j\in\mathcal{G}_n} u_j)^\top\}$ for some admissible class $\{\mathbf{V}_k\}$, we show that $\rho(\cdot,\cdot)$ satisfies all properties required for a CLF.

Step 1.1—Complete attainment content If $\mathbf{u} \ge 0$, then by $\mathbf{V}_1 = \mathbb{R}^n_+$ we have that $\mathbf{v}^\top\tilde{\mathbf{u}} \ge 0$ for all $\mathbf{v} \in \mathbf{V}_1$, which implies that $\sup_{\mathbf{v}\in\mathbf{V}_1}(-\mathbf{v}^\top\tilde{\mathbf{u}}) \le 0$. Hence $\rho(\mathbf{u}, \mathcal{G}) = 0$. Conversely, if $\mathbf{u} \not\ge 0$, without loss of generality we assume that there exists $j \in \mathcal{G}_1$ such that $u_j < 0$, then we have

$$\sup_{\mathbf{v}\in\mathbf{V}_1}(-\mathbf{v}^\top\tilde{\mathbf{u}}) = \sup_{\mathbf{v}\in\mathbb{R}^m_+}(-\mathbf{v}^\top\tilde{\mathbf{u}}) \ge -\mathbf{e}_1^\top\tilde{\mathbf{u}} > 0.$$

This inequality, combined with $\mathbf{V}_1 = \mathrm{cl}(\lim_{k\uparrow 1}\mathbf{V}_k)$, leads to that $\exists\delta > 0$ such that

$$\sup_{\mathbf{v}\in\mathbf{V}_{1-\delta}}(-\mathbf{v}^\top\tilde{\mathbf{u}}) > 0,$$

which implies that $\rho(\mathbf{u}, \mathcal{G}) > 0$. This shows that $\rho(\cdot,\cdot)$ satisfies *complete attainment content*.

Step 1.2—Non attainment apathy Fix $\mathbf{u}$ such that $\mathbf{u} < \mathbf{0}$ which implies $\tilde{\mathbf{u}} < \mathbf{0}$. Since $\mathbf{e} \in \mathbf{V}_0$, we have

$$\sup_{\mathbf{v}\in\mathbf{V}_0}(-\mathbf{v}^\top\tilde{\mathbf{u}}) \ge -\mathbf{e}^\top\tilde{\mathbf{u}} > 0.$$

Hence $\rho(\mathbf{u}, \mathcal{G}) = 1$. Thus, $\rho(\cdot,\cdot)$ satisfies *non attainment apathy*.

Step 1.3—Monotonicity Note that $\min_{j\in\mathcal{G}_i} u_j \ge \min_{j\in\mathcal{G}_i} w_j$ for all $i = 1, \ldots, n$ implies $\tilde{\mathbf{u}} \ge \tilde{\mathbf{w}}$. Then for any $k \in [0,1]$, since $\mathbf{V}_k \subseteq \mathbf{V}_1 = \mathbb{R}^n_+$, we have that $-\mathbf{v}^\top\tilde{\mathbf{u}} \le -\mathbf{v}^\top\tilde{\mathbf{w}}$ for any $\mathbf{v} \in \mathbf{V}_k$. Thus,

$$\sup_{\mathbf{v}\in\mathbf{V}_k}(-\mathbf{v}^\top\tilde{\mathbf{w}}) \le 0 \implies \sup_{\mathbf{v}\in\mathbf{V}_k}(-\mathbf{v}^\top\tilde{\mathbf{u}}) \le 0.$$

Hence $\rho(\mathbf{u}, \mathcal{G}) \le \rho(\mathbf{w}, \mathcal{G})$. Thus, $\rho(\cdot,\cdot)$ satisfies *monotonicity*.

Step 1.4—Order & scale invariance Order invariance follows directly from the fact that $\mathbf{V}_k$ is order invariant for all k. Scale invariant holds because for $\alpha > 0$ and $k \in [0,1]$,

$$\sup_{\mathbf{v}\in\mathbf{V}_k}(-\mathbf{v}^\top\mathbf{u}) \le 0 \iff \sup_{\mathbf{v}\in\mathbf{V}_k}(-\mathbf{v}^\top\alpha\mathbf{u}) \le 0.$$

Step 1.5—Quasi-convexity To show quasi-convexity in $\mathbf{u}$, let $c = \max(\rho(\mathbf{u}, \mathcal{G}), \rho(\mathbf{w}, \mathcal{G}))$ and without loss of generality assume $c < 1$ since otherwise the claim trivially holds. Thus we have that for any $\epsilon > 0$

$$\sup_{\mathbf{v}\in\mathbf{V}_{1-c-\epsilon}}(-\mathbf{v}^\top\tilde{\mathbf{u}}) \le 0 \text{ and } \sup_{\mathbf{v}\in\mathbf{V}_{1-c-\epsilon}}(-\mathbf{v}^\top\tilde{\mathbf{w}}) \le 0,$$

which implies that for $\alpha \in [0, 1]$

$$\sup_{\mathbf{v}\in\mathbf{V}_{1-c-\epsilon}} \{-\mathbf{v}^\top[\alpha\tilde{\mathbf{u}} + (1-\alpha)\tilde{\mathbf{w}}]\} \le 0.$$

Recall that $\tilde{\mathbf{u}} = (\min_{j\in\mathcal{G}_1} u_j, \cdots, \min_{j\in\mathcal{G}_n} u_j)^\top$ and $\tilde{\mathbf{w}} = (\min_{j\in\mathcal{G}_1} w_j, \cdots, \min_{j\in\mathcal{G}_n} w_j)^\top$. Let $\mathbf{y} = \alpha\mathbf{u} + (1-\alpha)\mathbf{w}$ and $\tilde{\mathbf{y}} = (\min_{j\in\mathcal{G}_1} y_j, \cdots, \min_{j\in\mathcal{G}_n} y_j)^\top$. Then we have $\tilde{\mathbf{y}} \ge \alpha\tilde{\mathbf{u}} + (1-\alpha)\tilde{\mathbf{w}}$ which implies that

$$\sup_{\mathbf{v}\in\mathbf{V}_{1-c-\epsilon}} \{-\mathbf{v}^\top\tilde{\mathbf{y}}\} \le \sup_{\mathbf{v}\in\mathbf{V}_{1-c-\epsilon}} \{-\mathbf{v}^\top[\alpha\tilde{\mathbf{u}} + (1-\alpha)\tilde{\mathbf{w}}]\} \le 0.$$

Thus, we have $\rho(\alpha\mathbf{u}_1 + (1-\alpha)\mathbf{u}_2, \mathcal{G}) \le c$ since ϵ can be arbitrarily close to 0. The quasi-convexity holds.

Step 1.6—Lower semi-continuity We show that $\rho(\mathbf{u}^*, \mathcal{G}) \le \liminf_i \rho(\mathbf{u}_i, \mathcal{G})$ for $\mathbf{u}_i \xrightarrow{i} \mathbf{u}^*$. Let $c > \liminf_i \rho(\mathbf{u}_i, \mathcal{G})$, then there exists an infinite sub-sequence $\{\mathbf{u}_{i_j}\}$ such that $\rho(\mathbf{u}_{i_j}, \mathcal{G}) < c$. That is

$$-\mathbf{v}^\top\tilde{\mathbf{u}}_{i_j} \le 0; \quad \forall\mathbf{v} \in \mathbf{V}_{1-c}, \forall j.$$

Note that $\mathbf{u}_{i_j} \to \mathbf{u}^*$, hence

$$-\mathbf{v}^\top\tilde{\mathbf{u}}^* \le 0; \quad \forall\mathbf{v} \in \mathbf{V}_{1-c},$$

i.e., $\rho(\mathbf{u}^*, \mathcal{G}) \le c$. Since c can be arbitrarily close to $\liminf_i \rho(\mathbf{u}_i, \mathcal{G})$, the semi-continuity follows.

Step 2—the "only if" part Given a function $\rho(\cdot, \cdot)$ which is a CLF, we show that it can be represented as

$$\rho(\mathbf{u}, \mathcal{G}) = 1 - \sup\{k \in [0, 1] : \sup_{\mathbf{v}\in\mathbf{V}_k} (-\mathbf{v}^\top\tilde{\mathbf{u}}) \le 0, \tilde{\mathbf{u}} = (\min_{j\in\mathcal{G}_1} u_j, \ldots, \min_{j\in\mathcal{G}_n} u_j)^\top\},$$

for some admissible class $\{\mathbf{V}_k\}$. This consists of three steps. We first show that $\rho(\cdot, \cdot)$ can be represented as

$$\rho(\mathbf{u}, \mathcal{G}) = 1 - \sup\{k \in [0, 1] : \sup_{\mathbf{v}\in\overline{\mathbf{V}}_k} (-\mathbf{v}^\top\tilde{\mathbf{u}}) \le 0, \tilde{\mathbf{u}} = (\min_{j\in\mathcal{G}_1} u_j, \ldots, \min_{j\in\mathcal{G}_n} u_j)^\top\},$$

for some $\{\overline{\mathbf{V}}_k\}$. Here $\{\overline{\mathbf{V}}_k\}$ is not necessarily admissible, but satisfies $\overline{\mathbf{V}}_k \subseteq \overline{\mathbf{V}}_{k'}$ for all $k \le k'$. We then show that we can replace $\overline{\mathbf{V}}_k$ by a class of closed, convex, order-invariant, cones $\mathbf{V}_k$. Finally we show that $\{\mathbf{V}_k\}$ is admissible to complete the proof. □

Step 2.1 The representability of $\rho(\cdot, \cdot)$ follows from the following lemma which is a variant of Theorem 2 in [10].

Lemma 2 *Given a CLF $\rho(\cdot,\cdot)$, then there exists $\{\overline{\mathbf{V}}_k\}$ that satisfies $\overline{\mathbf{V}}_k \subseteq \overline{\mathbf{V}}_{k'}$ for all $k \le k'$, such that*

$$\rho(\mathbf{u},\mathcal{G}) = 1 - \sup\{k \in [0,1] : \sup_{\mathbf{v}\in\overline{\mathbf{V}}_k} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \le 0, \tilde{\mathbf{u}} = (\min_{j\in\mathcal{G}_1} u_j, \ldots, \min_{j\in\mathcal{G}_n} u_j)^\top\}.$$

Proof We recall the definition of the collective satisfying measure in [10]. □

Definition 3 Let $\mathcal{U}$ be the set of random variables on the probability space $(\Omega,\mathcal{F},\mathbb{P})$. A function $\overline{\rho}(\cdot) : \mathcal{U} \to [0,1]$ is a *collective satisfying measure* if the following holds for all $U, U' \in \mathcal{U}$.

1. If $U \ge 0$, then $\overline{\rho}(U) = 1$;
2. If $U < 0$, then $\overline{\rho}(U) = 0$;
3. If $U \ge U'$ then $\overline{\rho}(U) \ge \overline{\rho}(U')$;
4. $\lim_{\alpha\ge 0} \overline{\rho}(U+\alpha) = \overline{\rho}(U)$;
5. If $\lambda \in [0,1]$, then $\overline{\rho}(\lambda U + (1-\lambda)U') \ge \min(\overline{\rho}(U), \overline{\rho}(U'))$;
6. If $\lambda > 0$, then $\overline{\rho}(\lambda U) = \overline{\rho}(U)$.

We now consider $\mathcal{U}$ – a special set of random variables defined on the probability space $(\Omega,\mathcal{F},\mathbb{P})$ with $\Omega = \{1,\ldots,m\}$. Note that each random variable $U : \Omega \mapsto \mathbb{R}$ can be represented as a vector $\mathbf{u} \in \mathbb{R}^m$ where $u_i = U(i)$. Let $\mathcal{G}$ be a partition of the set $\{1,\ldots,m\}$, namely, $\mathcal{G} = \{\mathcal{G}_1,\ldots,\mathcal{G}_n\}$ satisfying that

$$\mathcal{G}_i \ne \emptyset,\ \mathcal{G}_i \subseteq \{1,\ldots,m\},\ \bigcup_{i=1}^n \mathcal{G}_i = \{1,\ldots,m\}.$$

We define another random variable $\tilde{U}$ on probability space $(\tilde{\Omega},\tilde{\mathcal{F}},\tilde{\mathbb{P}})$ with $\tilde{\Omega} = \{1,\ldots,n\}$ by taking $\tilde{U}(i) = \min_{j\in\mathcal{G}_i} u_i$. The mapping from U to $\tilde{U}$ is denoted by g, i.e., $\tilde{U} = g(U,\mathcal{G})$. Let $\hat{\rho}(\cdot)$ be a collective satisfying measure on $\mathcal{U}$ that satisfies all the properties given by Definition 3 and another property that for all $U, U' \in \mathcal{U}$, if $\tilde{U} \ge \tilde{U}'$, then $\hat{\rho}(U) \ge \hat{\rho}(U')$.

Theorem 8 *The collective satisfying measure $\hat{\rho}(\cdot)$ can be represented as*

$$\hat{\rho}(U) = \sup\{k \in [0,1] : \sup_{\mathbb{Q}\in\mathcal{Q}_k} \mathbb{E}_{\mathbb{Q}}(-\tilde{U}) \le 0, \tilde{U} = g(U,\mathcal{G})\},$$

for a class of sets of probability measures $\mathcal{Q}_k$ satisfying $\mathcal{Q}_k \subseteq \mathcal{Q}_{k'}$ for $k \le k'$.

Proof Let X be a random variable on probability space $(\tilde{\Omega},\tilde{\mathcal{F}},\tilde{\mathbb{P}})$. For $k \in [0,1]$, we define

$$\mu_k(X) = \inf\{a : \hat{\rho}(\hat{X} + a) \ge k \text{ for some r.v. } \hat{X} \text{ such that } X = g(\hat{X},\mathcal{G})\}. \tag{17}$$

Then we have

$$\hat{\rho}(U) = \sup\{k : \mu_k(\tilde{U}) \le 0, k \in [0,1]\}. \tag{18}$$

To verify this equality, note that

$$\begin{aligned}
&\sup\{k : \mu_k(\tilde{U}) \le 0, k \in [0,1]\} \\
&\quad = \sup\{k : \exists a \le 0, \hat{U} \text{ such that } \hat{\rho}(\hat{U}+a) \ge k \text{ and } \tilde{U} = g(\hat{U}, \mathcal{G})\} \\
&\quad = \sup\{\hat{\rho}(\hat{U}+a) : a \le 0, \tilde{U} = g(\hat{U}, \mathcal{G})\} \\
&\quad = \sup\{\hat{\rho}(\hat{U}) : \tilde{U} = g(\hat{U}, \mathcal{G})\}
\end{aligned}$$

where the last equality holds due to the monotonicity of $\hat{\rho}(\cdot)$. Since $\tilde{U} = g(\hat{U}, \mathcal{G})$ and $\tilde{U} = g(U, \mathcal{G})$, by the additional property of $\hat{\rho}(\cdot)$ given above, we have $\hat{\rho}(U) = \hat{\rho}(\hat{U})$. Hence (18) holds. We next verify that $\mu_k(\cdot)$ defined by (17) is a coherent risk measure. Recall the definition of coherent risk measure: □

Definition 4 Let $\mathcal{U}$ be the set of random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. A function $\mu(\cdot) : \mathcal{U} \to \mathbb{R}$ is a *coherent risk measure* if the following holds for all $X, Y \in \mathcal{U}$.

1. If $X \ge Y$ then $\mu(X) \le \mu(Y)$;
2. If $c \in \mathbb{R}$, then $\mu(X + c) = \mu(X) - c$;
3. If $\lambda \in [0, 1]$, then $\mu(\lambda X + (1-\lambda)Y) \le \lambda\mu(X) + (1-\lambda)\mu(Y)$;
4. If $\lambda > 0$, then $\mu(\lambda X) = \lambda\mu(X)$.

We now verify that μ_k satisfies these properties. For random variables X and Y, let $\hat{X}$ and $\hat{Y}$ be any random variables satisfying that $X = g(\hat{X}, \mathcal{G})$ and $Y = g(\hat{Y}, \mathcal{G})$. If $X \ge Y$, then by the property of $\hat{\rho}(\cdot)$, we have $\hat{\rho}(\hat{X}) \ge \hat{\rho}(\hat{Y})$, which implies $\mu_k(X) \le \mu_k(Y)$. Hence Property 1 holds. Property 2 can be easily seen from the definition of $\mu_k(\cdot)$. For Property 3, note that for all $\epsilon > 0$, we have

$$\hat{\rho}(\hat{X} + \mu_k(X) + \epsilon) \ge k, \ \hat{\rho}(\hat{Y} + \mu_k(Y) + \epsilon) \ge k.$$

On the other hand, since $\hat{\rho}$ is quasi-concave, we have

$$\begin{aligned}
&\hat{\rho}(\lambda(\hat{X} + \mu_k(X)) + (1-\lambda)(\hat{Y} + \mu_k(Y)) \\
&\quad + \epsilon) \ge \min\{\hat{\rho}(\hat{X} + \mu_k(X) + \epsilon), \hat{\rho}(\hat{Y} + \mu_k(Y) + \epsilon)\} \ge k. \qquad (19)
\end{aligned}$$

Now consider special $\hat{X}$ and $\hat{Y}$ such that $\hat{X}(i) = X(j)$ for $i \in \mathcal{G}_j$ and $\hat{Y}(i) = Y(j)$ for $i \in \mathcal{G}_j$. Clearly, these $\hat{X}$ and $\hat{Y}$ are the "smallest", namely, for all $\tilde{X}$ and $\tilde{Y}$ such that $X = g(\tilde{X}, \mathcal{G})$ and $X = g(\tilde{Y}, \mathcal{G})$, we have $\tilde{X} \ge \hat{X}$ and $\tilde{Y} \ge \hat{Y}$. This implies

$$\begin{aligned}
\mu_k(\lambda X + (1-\lambda)Y) &= \inf\{a : \hat{\rho}(\hat{Z} + a) \ge k, \exists \hat{Z} \text{ such that } \lambda X + (1-\lambda)Y = g(\hat{Z}, \mathcal{G})\} \\
&\le \inf\{a : \hat{\rho}(\lambda\hat{X} + (1-\lambda)\hat{Y} + a) \ge k\} \\
&\le \lambda\mu_k(X) + (1-\lambda)\mu_k(Y),
\end{aligned}$$

where the last inequality follows from (19). For the last property, note that for $\lambda > 0$,

$$\begin{aligned}\mu_k(\lambda X) =& \inf\{a : \hat{\rho}(\lambda\hat{X} + a) \geq k \text{ for some r.v. } \hat{X} \text{ such that } \lambda X = g(\lambda\hat{X}, \mathcal{G})\}\\ =& \inf\{\lambda a : \hat{\rho}(\lambda\hat{X} + \lambda a) \geq k \text{ for some r.v. } \hat{X} \text{ such that } \lambda X = g(\lambda\hat{X}, \mathcal{G})\}\\ =& \inf\{\lambda a : \hat{\rho}(\hat{X} + a) \geq k \text{ for some r.v. } \hat{X} \text{ such that } X = g(\hat{X}, \mathcal{G})\}\\ =& \lambda\mu_k(X).\end{aligned}$$

Hence $\mu_k(\cdot)$ is a coherent risk measure. It is known that coherent risk measure $\mu_k(\cdot)$ can be written in the form

$$\mu_k(X) = \sup_{\mathbb{Q}\in\mathcal{Q}_k} \mathbb{E}_{\mathbb{Q}}(-X)$$

for a family of generating measures $\mathcal{Q}_k$. By combining this formula with (18), Theorem 8 is established. □

We now turn to the proof of Lemma 2. Given a CLF $\rho(\cdot, \cdot)$, for fixed $\mathcal{G}$, we define $\overline{\rho} : \mathcal{U} \mapsto \mathbb{R}$ as following

$$\overline{\rho}(U) = 1 - \rho(\mathbf{u}, \mathcal{G}); \quad \text{where } u_i = U(i),\ i = 1, \ldots, m.$$

It is straightforward to check that $\overline{\rho}(\cdot)$ has all the properties of the collective satisfying measure $\hat{\rho}(\cdot)$. Thus, Theorem 8 states there exists a class of sets of probability measure $\mathcal{Q}_k$ such that

$$1 - \rho(\mathbf{u}) = \overline{\rho}(U) = \sup\{k \in [0, 1] : \sup_{\mathbb{Q}\in\mathcal{Q}_k} \mathbb{E}_{\mathbb{Q}}(-\tilde{U}) \leq 0, \tilde{U} = g(U, \mathcal{G})\}.$$

Note that any probability measure Q on $\tilde{\Omega} = \{1, \ldots, n\}$ can be represented by a vector $\mathbf{v} \in \mathbb{R}^n$ such that $v_i = Q(i)$. Thus $\mathbb{E}_Q(-\tilde{U}) = -\mathbf{v}^\top\tilde{\mathbf{u}}$ where $\mathbf{v}$ and $\tilde{\mathbf{u}}$ are the vector form for Q and $\tilde{U}$ respectively. Hence we have there exists $\overline{\mathbf{V}}_k$ such that

$$\rho(\mathbf{u}) = 1 - \sup\{k \in [0, 1] : \sup_{\mathbf{v}\in\overline{\mathbf{V}}_k} (-\mathbf{v}^\top\tilde{\mathbf{u}}) \leq 0, \tilde{\mathbf{u}} = (\min_{j\in\mathcal{G}_1} u_j, \cdots, \min_{j\in\mathcal{G}_n} u_j)^\top\}.$$

Note that for $k \leq k'$, $\overline{\mathbf{V}}_k \subseteq \overline{\mathbf{V}}_{k'}$ since $\mathcal{Q}_k \subseteq \mathcal{Q}_{k'}$. This concludes the proof of Lemma 2. □

Step 2.2: We construct the admissible class $\{\mathbf{V}_k\}$ as follows. Define $\hat{\mathbf{V}}_k \triangleq \text{cl}(\text{cc}(\text{or}(\overline{\mathbf{V}}_k)))$. Then we let $\mathbf{V}_k \triangleq \hat{\mathbf{V}}_k$ for $k \in (0, 1)$, and $\mathbf{V}_0 \triangleq \bigcap_{k\in(0,1)} \hat{\mathbf{V}}_k$, and $\mathbf{V}_1 \triangleq \text{cl}(\bigcup_{k\in(0,1)} \hat{\mathbf{V}}_k)$. Here $\text{or}(\cdot)$ (respectively $\text{cc}(\cdot)$) is the minimal **or**der invariant (respectively, **c**onvex **c**one) superset, defined as

$$\text{or}(S) = \{\mathbf{P}\mathbf{v} : \mathbf{P} \in \mathcal{P}_n, \mathbf{v} \in S\}, \quad \text{cc}(S) = \left\{\sum_{i=1}^{k} \lambda_i\mathbf{v}_i | k \in \mathbb{N}, \mathbf{v}_i \in S, \lambda_i \geq 0\right\},$$

where $\mathcal{P}_n$ is the set of all $n \times n$ permutation matrices. Let

$$\rho'(\mathbf{u}, \mathcal{G}) = 1 - \sup \left\{ k \in [0,1] : \sup_{\mathbf{v} \in \hat{\mathbf{V}}_k} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0, \tilde{\mathbf{u}} = (\min_{j \in \mathcal{G}_1} u_j, \cdots, \min_{j \in \mathcal{G}_n} u_j)^\top \right\},$$

and observe that $\overline{\mathbf{V}}_k \subseteq \hat{\mathbf{V}}_k$, hence $\rho(\mathbf{u}) \leq \rho'(\mathbf{u})$. To show that $\rho(\mathbf{u}) \geq \rho'(\mathbf{u})$, it suffices to show that for any k, ϵ and $\tilde{\mathbf{u}}$, the following holds,

$$\sup_{\mathbf{v} \in \overline{\mathbf{V}}_k} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0 \quad \Longrightarrow \quad \sup_{\mathbf{v} \in \hat{\mathbf{V}}_{k-\epsilon}} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0.$$

Note that $\sup_{\mathbf{v} \in \overline{\mathbf{V}}_k} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0$ implies $\rho(\mathbf{u}) \leq 1 - k$. Hence by order invariance of $\rho(\cdot, \cdot)$, we have

$$\sup_{\mathbf{v} \in \overline{\mathbf{V}}_{k-\epsilon}} \sup_{\mathbf{P} \in \mathcal{P}_n} (-\mathbf{v}^\top \mathbf{P} \tilde{\mathbf{u}}) \leq 0,$$

which is equivalent to

$$\sup_{\mathbf{v} \in \mathrm{or}(\overline{\mathbf{V}}_{k-\epsilon})} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0.$$

By definition of $\mathrm{cc}(\cdot)$ and the continuity of $-\mathbf{v}^\top \tilde{\mathbf{u}}$ w.r.t. $\mathbf{v}$, this leads to

$$\sup_{\mathbf{v} \in \mathrm{cl}(\mathrm{cc}(\mathrm{or}(\overline{\mathbf{V}}_{k-\epsilon})))} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0.$$

Therefore we have $\rho(\mathbf{u}, \mathcal{G}) = \rho'(\mathbf{u}, \mathcal{G})$. Finally note that $\hat{\mathbf{V}}_k \subseteq \hat{\mathbf{V}}_{k'}$ for $k \leq k'$, which leads to the following

$$\begin{aligned}
&\sup_{\mathbf{v} \in \hat{\mathbf{V}}_0} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq \sup_{\mathbf{v} \in \bigcap_{k \in (0,1)} \hat{\mathbf{V}}_k} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq \sup_{\mathbf{v} \in \hat{\mathbf{V}}_\epsilon} (-\mathbf{v}^\top \tilde{\mathbf{u}}), \\
&\sup_{\mathbf{v} \in \hat{\mathbf{V}}_{1-\epsilon}} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq \sup_{\mathbf{v} \in \bigcup_{k \in (0,1)} \hat{\mathbf{V}}_k} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq \sup_{\mathbf{v} \in \hat{\mathbf{V}}_1} (-\mathbf{v}^\top \tilde{\mathbf{u}}).
\end{aligned}$$

By definitions of $\mathbf{V}_0$ and $\mathbf{V}_1$, together with the fact (due to continuity)

$$\sup_{\mathbf{v} \in \mathrm{cl}(\bigcup_{k \in (0,1)} \hat{\mathbf{V}}_k)} (-\mathbf{v}^\top \tilde{\mathbf{u}}) = \sup_{\mathbf{v} \in \bigcup_{k \in (0,1)} \hat{\mathbf{V}}_k} (-\mathbf{v}^\top \tilde{\mathbf{u}}),$$

we conclude that

$$\rho(\mathbf{u}, \mathcal{G}) = 1 - \sup\{k \in [0,1] : \sup_{\mathbf{v} \in \mathbf{V}_k} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0, \tilde{\mathbf{u}} = (\min_{j \in \mathcal{G}_1} u_j, \cdots, \min_{j \in \mathcal{G}_n} u_j)^\top\}.$$

Step 2.3: Now we check that $\{\mathbf{V}_k\}$ is indeed admissible. Property 1–3 are straightforward from the definition of $\mathbf{V}_k$. To see that $\mathbf{V}_0$ is closed, recall that the intersection of a class of closes sets is close.

We next show Property 4: $\mathbf{V}_1 = \mathbb{R}^n_+$. By definition of $\mathbf{V}_1$, we have

$$\lim_{k\to 1} \sup_{\mathbf{v}\in\mathbf{V}_k} (-\mathbf{v}^\top \tilde{\mathbf{u}}) = \sup_{\mathbf{v}\in\mathbf{V}_1} (-\mathbf{v}^\top \tilde{\mathbf{u}}).$$

Hence $\rho(\mathbf{u}) = 0$ if and only if $\sup_{\mathbf{v}\in\mathbf{V}_1}(-\mathbf{v}^\top\tilde{\mathbf{u}}) \le 0$. Therefore, by the property of *complete attainment content* we have the following

$$\sup_{\mathbf{v}\in\mathbf{V}_1} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \le 0 \iff \tilde{\mathbf{u}} \ge 0 \iff \mathbf{u} \ge 0. \tag{20}$$

Denote the dual cone of a cone $\mathbf{C}$ by $\mathbf{C}^*$ and recall that for any k, $\mathbf{V}_k$ is a closed convex cone, hence we have

$$(\mathbf{V}_1^*)^* = \mathbf{V}_1.$$

The definition of dual cone states that

$$\mathbf{V}_1^* = \{\mathbf{u} : \mathbf{u}^\top \mathbf{v} \ge 0, \forall \mathbf{v} \in \mathbf{V}_1\},$$

which combined with Eq. (20) implies that $\mathbf{V}_1^* = \mathbb{R}^n_+$. Since $\mathbb{R}^m_+$ is self-dual, we have $\mathbf{V}_1 = \mathbb{R}^n_+$.

We now turn to Property 5. Fix $k > 0$. Consider $\mathbf{u} = -\mathbf{e}^m$, which means $\tilde{\mathbf{u}} = -\mathbf{e}^n$. By misclassification avoidance, $\rho(\mathbf{u}, \mathcal{G}) = 1$, which means there exists $\mathbf{v} \in \mathbf{V}_k$ such that $\mathbf{v}^\top\tilde{\mathbf{u}} < 0$, i.e., $\sum_{i=1}^n v_i > 0$. Define a permutation matrix $\mathbf{P} \in \mathcal{P}_n$:

$$\mathbf{P} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 0 \end{pmatrix}$$

Thus, by order invariance of $\mathbf{V}_k$, $\mathbf{P}^t\mathbf{v} \in \mathbf{V}_k$ for $t = 0, \ldots, n-1$. By convexity, this implies $\frac{1}{n}\sum_{t=0}^{n-1} \mathbf{P}^t\mathbf{v} \in \mathbf{V}_k$. Note that $\frac{1}{n}\sum_{t=0}^{n-1}\mathbf{P}^t\mathbf{v} = [\frac{1}{n}\sum_{i=1}^n v_i]\mathbf{e}^n$, thus

$$\frac{\sum_{i=1}^n v_i}{n}\mathbf{e}^n \in \mathbf{V}_k.$$

Since $\sum_{i=1}^n v_i > 0$ and $\mathbf{V}_k$ is a cone, we have $\lambda\mathbf{e}^n \in \mathbf{V}_k$ for all $\lambda \ge 0$ and $k > 0$. By definition of $\mathbf{V}_0$, this implies $\lambda\mathbf{e}^n \in \mathbf{V}_0$. □

6.2 Proof of Theorem 2

Proof **Claim 1** We check that all conditions of Definition 1 are satisfied by $\overline{\rho}(\cdot)$. The only condition needs a proof is the semi-continuity. Consider a sequence $\mathbf{u}^j \to \mathbf{u}^0$, and let $t^0 = \max\{t : \sum_{i=1}^{t} \tilde{u}^0_{(i)} < 0\}$. Without loss of generality we let $\tilde{u}^0_1 \le \tilde{u}^0_2 \le \cdots \le \tilde{u}^0_n$. Thus we have that $\sum_{i=1}^{t^0} \tilde{u}^0_i < 0$. This implies that $\limsup_j \sum_{i=1}^{t^0} \tilde{u}^j_i < 0$, which further leads to $\liminf_j (\max\{t : \sum_{i=1}^{t} \tilde{u}^j_{(i)} < 0\}) \ge t^0$. Hence $\liminf_j \overline{\rho}(\mathbf{u}^j, \mathcal{G}) \ge \overline{\rho}(\mathbf{u}^0, \mathcal{G})$, which established the semi-continuity. Thus, we conclude that $\overline{\rho}(\cdot)$ is a CLF. Further, observe that $\max\{t : \sum_{i=1}^{t} \tilde{u}_{(i)} < 0\} \ge \sum_{i=1}^{n} \mathbf{1}(u_j < 0, \exists j \in \mathcal{G}_i)$, which established the first claim.

Claim 2 It is straightforward to check that $\overline{\mathbf{V}}_k$ satisfies all conditions of Definition 2, and hence is an admissible set. Thus, we proceed to show that $\overline{\mathbf{V}}_k$ is an admissible set *corresponding to* $\overline{\rho}(\cdot)$, i.e., to show

$$\overline{\rho}(\mathbf{u}) = 1 - \sup\{k \in [0,1] : \sup_{\mathbf{v} \in \overline{\mathbf{V}}_k} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \le 0, \tilde{\mathbf{u}} = (\min_{j \in \mathcal{G}_1} u_j, \ldots, \min_{j \in \mathcal{G}_n} u_j)^\top\}.$$

Fix a $\mathbf{u} \in \mathbb{R}^m$. If $\mathbf{u} \ge 0$, then we have $\overline{\rho}(\mathbf{u}) = 0$, as well as $\sup_{\mathbf{v} \in \overline{\mathbf{V}}_1} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \le 0$, and hence the equivalence holds trivially. Thus we assume $\mathbf{u} \not\ge 0$, and let $t^0 = \max\{t : \sum_{i=1}^{t} \tilde{u}_{(i)} < 0\}$. By definition we have

$$\overline{\mathbf{V}}_{1-t^0/n} = \operatorname{conv}\left\{\lambda \mathbf{e}_{N'} : \lambda > 0, |N'| = t^0 + 1\right\}.$$

Note that by definition of t^0

$$\min_{|N'| = t^0 + 1} \sum_{i \in N'} \tilde{u}_i \ge 0,$$

which implies that

$$\sup_{\mathbf{v} \in \{\mathbf{e}_{N'} : |N'| = t^0 + 1\}} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \le 0.$$

This leads to

$$\sup_{\mathbf{v} \in \overline{\mathbf{V}}_{1-t^0/n}} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \le 0. \tag{21}$$

On the other hand for arbitrarily small $\epsilon > 0$, by definition

$$\overline{\mathbf{V}}_{1-t^0/n+\epsilon} = \operatorname{conv}\left\{\lambda \mathbf{e}_N : \lambda > 0, |N| = t^0\right\}.$$

Because $\min_{N:|N|=t^0} \sum_{i\in N} \tilde{u}_i < 0$, we have

$$\sup_{\mathbf{v}\in\overline{\mathbf{V}}_{1-t^0/n+\epsilon}} (-\mathbf{v}^\top \tilde{\mathbf{u}}) > 0.$$

Combining with Eq. (21) we established the second claim.

Claim 3 Let $\rho'(\cdot)$ be a CLF satisfying that $\rho'(\mathbf{u}, \mathcal{G}) \geq \varrho(\mathbf{u}, \mathcal{G})$ for all $\mathbf{u} \in \mathbb{R}^m$, and let $\{\mathbf{V}'_k\}$ be its corresponding admissible set. Thus, it suffices to show that $\overline{\mathbf{V}}_k \subseteq \mathbf{V}'_k$ for all k. This holds trivially for $k = 0$, since $\rho'(\mathbf{u}, \mathcal{G}) = 1$ for all $\mathbf{u} < \mathbf{0}$ implies that $\lambda \mathbf{e}^n \in \mathbf{V}'_0$. When $k > 0$, let $s/n < k \leq (s+1)/n$ for some integer s. Then, since $\mathbf{V}'_k$ is an order-invariant convex cone, it suffices to show that $\mathbf{e}_{[1:n-s]} \in \mathbf{V}'_k$ to establish the third claim. Consider $\mathbf{u}^*$ such that $\tilde{\mathbf{u}}^* = -\mathbf{e}_{[1:n-s]}$. Then, by $\rho'(\mathbf{u}^*, \mathcal{G}) \geq \sum_i \mathbf{1}(\tilde{u}^*_i < 0)/n = 1 - s/n > 1 - k$, we have

$$\sup_{\mathbf{v}\in\mathbf{V}'_k} (-\mathbf{v}^\top \tilde{\mathbf{u}}^*) > 0 \implies \exists \mathbf{v}^* \in \mathbf{V}'_k : \sum_{i=1}^{n-s} v^*_i > 0.$$

Define a permutation matrix $\mathbf{P}$:

$$\mathbf{P} = \begin{pmatrix} \mathbf{P}_1 & 0_{(n-s)\times s} \\ 0_{(n-s)\times s} & 0_{s\times s}, \end{pmatrix}$$

where $\mathbf{P}_1$ is a $(n-s) \times (n-s)$ matrix:

$$\mathbf{P}_1 = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 0 \end{pmatrix}$$

Thus, by order invariance of $\mathbf{V}'_k$, $\mathbf{P}^t\mathbf{v}^* \in \mathbf{V}'_k$ for $t = 0, \ldots, n-s-1$. By convexity, this implies $\frac{1}{n-s}\sum_{t=0}^{n-s-1} \mathbf{P}^t\mathbf{v}^* \in \mathbf{V}'_k$. Note that $\frac{1}{n-s}\sum_{t=0}^{n-s-1} \mathbf{P}^t\mathbf{v}^* = \frac{1}{n-s}[\sum_{i\in[1:n-s]} v^*_i]\mathbf{e}_{[1:n-s]}$, thus

$$\frac{\sum_{i=1}^{n-s} v^*_i}{n-s}\mathbf{e}_{[1:n-s]} \in \mathbf{V}'_k.$$

Since $\frac{\sum_{i=1}^{n-s} v^*_i}{n-s}$ is positive, and $\mathbf{V}'_k$ is a cone, we have $\mathbf{e}_{[1:n-s]} \in \mathbf{V}'_k$, which completes the proof. □

6.3 Proof of Theorem 3

Proof We prove the theorem by constructing such a function $\rho(\cdot,\cdot)$. To do this, first consider $\tilde{\rho}:\mathcal{R}^m\times\mathcal{S}\mapsto[0,1]$ defined as

$$\tilde{\rho}(\mathbf{u},\mathcal{G})=\min_{\gamma>0}\hat{\rho}(\mathbf{u}/\gamma,\mathcal{G}).$$

Then it is easy to check that $\tilde{\rho}(\cdot)$ satisfies complete attainment content, non attainment apathy, monotonicity, order invariance, and scale invariance. To see that $\tilde{\rho}(\mathbf{u},\mathcal{G})\geq\varrho(\mathbf{u},\mathcal{G})$, note that if $\tilde{\mathbf{u}}$, where $\tilde{\mathbf{u}}=(\min_{j\in\mathcal{G}_1}u_j,\ldots,\min_{j\in\mathcal{G}_n}u_j)^\top$, has t negative coefficients, then for any $\gamma>0$, $\tilde{\mathbf{u}}/\gamma$ also has t negative coefficients, which means

$$\hat{\rho}(\mathbf{u}/\gamma,\mathcal{G})\geq t/n.$$

Taking minimization over γ, we have $\tilde{\rho}(\mathbf{u},\mathcal{G})\geq\varrho(\mathbf{u},\mathcal{G})$ holds. Finally, we show quasi-convexity of $\tilde{\rho}(\cdot)$. Fix $\mathbf{u}_1$, $\mathbf{u}_2$, and $\alpha\in[0,1]$, let γ_1, γ_2 be ϵ-optimal, i.e.,

$$\hat{\rho}(\mathbf{u}_i/\gamma_i,\mathcal{G})\leq\tilde{\rho}(\mathbf{u}_i,\mathcal{G})+\epsilon,\quad i=1,2.$$

Since $\hat{\rho}(\mathbf{u},,\mathcal{G})$ is quasi-convex w.r.t. $\mathbf{u}$, we have

$$\begin{aligned}\hat{\rho}\left(\frac{\alpha\mathbf{u}_1+(1-\alpha)\mathbf{u}_2}{\alpha\gamma_1+(1-\alpha)\gamma_2},\mathcal{G}\right)&=\hat{\rho}\left(\frac{\alpha\gamma_1}{\alpha\gamma_1+(1-\alpha)\gamma_2}\cdot\frac{\mathbf{u}_1}{\gamma_1}+\frac{(1-\alpha)\gamma_2}{\alpha\gamma_1+(1-\alpha)\gamma_2}\cdot\frac{\mathbf{u}_2}{\gamma_2},\mathcal{G}\right)\\&\leq\max\left\{\hat{\rho}\left(\frac{\mathbf{u}_1}{\gamma_1},\mathcal{G}\right),\hat{\rho}\left(\frac{\mathbf{u}_2}{\gamma_2},\mathcal{G}\right)\right\}\end{aligned}$$

which implies

$$\begin{aligned}\tilde{\rho}(\alpha\mathbf{u}_1+(1-\alpha)\mathbf{u}_2,\mathcal{G})\leq\hat{\rho}\left(\frac{\alpha\mathbf{u}_1+(1-\alpha)\mathbf{u}_2}{\alpha\gamma_1+(1-\alpha)\gamma_2},\mathcal{G}\right)&\leq\max\left\{\hat{\rho}\left(\frac{\mathbf{u}_1}{\gamma_1},\mathcal{G}\right),\hat{\rho}\left(\frac{\mathbf{u}_2}{\gamma_2},\mathcal{G}\right)\right\}\\&\leq\max\{\tilde{\rho}(\mathbf{u}_1,\mathcal{G}),\tilde{\rho}(\mathbf{u}_2,\mathcal{G})\}+\epsilon.\end{aligned}$$

Hence $\tilde{\rho}(\cdot)(\mathbf{u},\mathcal{G})$ is quasi-convex w.r.t. $\mathbf{u}$. Note that the only property that is not satisfied is the semi-continuity. To handle this, define $\rho:\mathcal{R}^m\times\mathcal{S}\mapsto[0,1]$ as

$$\rho(\mathbf{u},\mathcal{G})=\lim_{\epsilon\downarrow0}\tilde{\rho}(\mathbf{u}+\epsilon\mathbf{e}^m,\mathcal{G})$$

Because of monotonicity of $\tilde{\rho}(\cdot)$, $\rho(\cdot,\cdot)$ is well-defined. In addition, it can be shown that $\rho(\cdot,\cdot)$ is lower-semicontinuous. Complete attainment content, non attainment apathy, monotonicity, order invariance, scale invariance, and quasi-convexity all follows easily from the fact that same properties hold for $\tilde{\rho}(\cdot)$. Thus, $\rho(\cdot,\cdot)$ is a CLF w.r.t. m. Next, we show that

$$\hat{\rho}(\mathbf{u},\mathcal{G})\geq\rho(\mathbf{u},\mathcal{G})\geq\varrho(\mathbf{u},\mathcal{G}).$$

The first inequality holds due to $\hat{\rho}(\mathbf{u}, \mathcal{G}) \geq \tilde{\rho}(\mathbf{u}, \mathcal{G}) \geq \tilde{\rho}(\mathbf{u} + \epsilon \mathbf{e}^m, \mathcal{G})$. The second inequality holds because for any $\mathbf{u}$, there exists $\epsilon > 0$ small enough such that $\varrho(\mathbf{u} + \epsilon \mathbf{e}^m, \mathcal{G}) = \varrho(\mathbf{u}, \mathcal{G})$. Thus, taking limit over $\tilde{\rho}(\mathbf{u} + \epsilon \mathbf{e}^m, \mathcal{G}) \geq \varrho(\mathbf{u} + \epsilon \mathbf{e}^m, \mathcal{G})$ establishes the second inequality. Recall that $\bar{\rho}(\mathbf{u}, \mathcal{G})$ is the minimal CLF, we establish the lemma by

$$\varrho(\mathbf{u}, \mathcal{G}) \leq \bar{\rho}(\mathbf{u}, \mathcal{G}) \leq \rho(\mathbf{u}, \mathcal{G}).$$

□

6.4 Proof of Theorem 5

Proof To prove Theorem 5, we start with establishing the following lemma. Observe that $\overline{\rho}(\mathbf{u}, \mathcal{G})$ only takes value in $\{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\}$. □

Lemma 3 *The level set of Problem (7), i.e.,* $\mathcal{U}_i \triangleq \{(\mathbf{u}, \mathbf{w}) : \overline{\rho}(\mathbf{u}, \mathcal{G}) \leq 1 - \frac{i}{n};\ f_j(\mathbf{u}, \mathbf{w}) \leq 0, \forall j\}$ *for* $i = 1, \ldots, n$, *equals the following*

$$\left\{(\mathbf{u}, \mathbf{w}) : \exists d \text{ such that } \sum_{i=1}^{n} [d - \min_{j \in \mathcal{G}_i} u_j]_+ \leq (n - i + 1)d;\ f_j(\mathbf{u}, \mathbf{w}) \leq 0,\ \forall j.\right\}$$

Proof Let $\tilde{\mathbf{u}} = (\min_{j \in \mathcal{G}_1} u_j, \ldots, \min_{j \in \mathcal{G}_n} u_j)^\top$. From Property 2 of Theorem 2, we have that $\mathcal{U}_i$ equals to the feasible set of the following program

$$\sup_{\mathbf{v} \in \overline{\mathbf{V}}_{i/n}} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0;\ f_j(\mathbf{u}, \mathbf{w}) \leq 0,\ j = 1, \ldots, k.$$

Recall that $\overline{\mathbf{V}}_{i/n} = \text{conv}\,\{\lambda \mathbf{e}_N | \lambda > 0, |N| = n - i + 1\}$ we have that $\sup_{\mathbf{v} \in \overline{\mathbf{V}}_{i/n}} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0$ is equivalent to

$$\inf_{\mathbf{v}: \mathbf{0} \leq \mathbf{v} \leq \mathbf{e}, \mathbf{e}^\top \mathbf{v} = n - i + 1} \mathbf{v}^\top \tilde{\mathbf{u}} \geq 0,$$

which left-hand-side by duality theorem is equivalent to the following optimization problem on $(\mathbf{c}, d)$

$$\begin{aligned} \text{Maximize: } & \sum_{i=1}^{n} c_i + (n - i + 1)d \\ \text{Subject to: } & c_i + d \leq \tilde{u}_i,\ c_i \leq 0,\ i = 1, \ldots, n. \end{aligned}$$

Thus we have $\mathbf{u} \in \mathcal{U}_i$ if and only if there exists $\mathbf{c}$, d, and $\mathbf{w}$ such that

$$\begin{aligned} & \mathbf{e}^\top \mathbf{c} + (n - i + 1)d \geq 0; \\ & \mathbf{c} + d\mathbf{e} \leq \tilde{\mathbf{u}}; \end{aligned}$$

$$\begin{aligned} &\mathbf{c} \leq \mathbf{0}; \\ &f_j(\mathbf{u}, \mathbf{w}) \leq 0, \ \ j = 1, \ldots, k. \end{aligned}$$

Note that this can be further simplified, since optimal $c_i = -[d - \tilde{u}_i]_+$, as

$$\begin{aligned} &\sum_{i=1}^{n} [d - \tilde{u}_i]_+ \leq (n - i + 1)d \\ &f_j(\mathbf{u}, \mathbf{w}) \leq 0, \ \ j = 1, \ldots, k. \end{aligned} \tag{22}$$

This establishes the lemma. □

Now we turn to prove Theorem 5. When all feasible solutions $\mathbf{u}, \mathbf{w}$, i.e., $f_j(\mathbf{u}, \mathbf{w}) \leq 0$ for all $j = 1, \ldots, k$, satisfy that $\mathbf{u} > 0$ or $\mathbf{u} \ngeq \mathbf{0}$, we only need to consider the feasible solutions to (22) with $d > 0$. Hence the feasible set to Problem (22) is equivalent to that of

$$\begin{aligned} &\sum_{i=1}^{n} [1 - \tilde{u}_i/d]_+ \leq (n - i + 1) \\ &f_j(\mathbf{u}, \mathbf{w}) \leq 0, \ \ j = 1, \ldots, k. \end{aligned}$$

Thus, finding the optimal solution to Problem (7) is equivalent to solve the following

$$\begin{aligned} \text{Minimize:} \ \ & \sum_{i=1}^{n} [1 - \tilde{u}_i/d]_+ \\ \text{Subject to:} \ \ & f_j(\mathbf{u}, \mathbf{w}) \leq 0, \ \ j = 1, \ldots, k; \\ & d > 0. \end{aligned} \tag{23}$$

By a change of variable where we let $h = 1/d$, $\mathbf{s} = h\mathbf{u}$, $\mathbf{t} = h\mathbf{w}$, this is equivalent to

$$\begin{aligned} \text{Minimize:} \ \ & \sum_{i=1}^{n} [1 - \min_{j \in \mathcal{G}_i} s_j]_+ \\ \text{Subject to:} \ \ & h f_j(\mathbf{s}/h, \mathbf{t}/h) \leq 0, \ \ j = 1, \ldots, k; \\ & h > 0. \end{aligned}$$

Hence Theorem 5 is established. □

7 Conclusion

In this paper, motivated from binary classification in machine learning, we study a class of decision problems that aim at maximizing the number of goals attained. We develop an axiomatic approach, eliciting desirable properties that loss functions for

such problems naturally exhibit, which we call "coherent loss", and provide dual representation that characterizes the set of coherent loss functions. The dual representation enables us to identify the *minimal coherent loss* function which is the coherent loss that best approximates the number of goal missed, and is a tighter bound than any convex upper bounds. In the context of classification, the coherent loss is essentially a tractable upper-bound of the total classification error for the entire training set, as opposed to the standard cumulative-loss approach that minimizes the sum of convex surrogates for each data point. The coherent loss approach applied in classification thus yields a strictly tighter approximation to the classification than any cumulative loss, while preserving the tractability of the resulting optimization problem. The formulation also has a robustness interpretation, which builds a strong connection between the coherent loss and robust SVMs. Finally, we remark that the coherent loss approach has favorable statistical properties and the simulation results show that it outperforms the standard SVM.

References

1. Ahmed, S., Shapiro, A.: Solving chance-constrained stochastic programs via sampling and integer programming. In: Tutorial in Operations Research, pp. 261–269. Informs (2008)
2. Arora, S., Babai, L., Stern, J., Sweedyk, Z.: The hardness of approximate optima in lattices, codes, and systems of linear equations. J. Comput. Syst. Sci. **54**, 317–331 (1997)
3. Artzner, P., Delbaen, F., Eber, J., Heath, D.: Coherent measures of risk. Math. Finance **9**, 203–228 (1999)
4. Asuncion, A., Newman, D.J.: UCI machine learning repository (2007). http://www.ics.uci.edu/~mlearn/MLRepository.html
5. Atamtürk, A., Nemhauser, G.L., Savelsbergh, M.W.P.: The mixed vertex packing problems. Math. Program. **99**, 35–53 (2000)
6. Ben-David, S., Eiron, N., Long, P.M.: On the difficulty of approximately maximizing agreements. J. Comput. Syst. Sci. **66**, 496–513 (2003)
7. Bordley, R., LiCalzi, M.: Decision analysis using targets instead of utility functions. Decis. Econ. Finance **23**, 53–74 (2000)
8. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory, New York, NY, pp. 144–152 (1992)
9. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
10. Brown, D., Sim, M.: Satisficing measures for analysis of risky positions. Manag. Sci. **55**(1), 71–84 (2009)
11. Castagnoli, E., LiCalzi, M.: Expected utility without utility. Theory Decis. **41**, 281–301 (1996)
12. Charnes, A., Cooper, W.W.: Management models and industrial applications of linear programming. Manag. Sci. **4**(1), 38–91 (1957)
13. Charnes, A., Cooper, W.W.: Chance constrained programming. Manag. Sci. **6**, 73–79 (1959)
14. Charnes, A., Cooper, W.W., Ferguson, R.: Optimal estimation of executive compensation by linear programming. Manag. Sci. **1**, 138–151 (1955)
15. Charnes, A., Haynes, K.E., Hazleton, J.E., Ryan, M.J.: An hierarchical goal programming approach to environmental-land use management. In: Mathematical Analysis of Decision Problems in Ecology, pp. 2–13 (1975)
16. Chen, W., Sim, M.: Goal driven optimization. Oper. Res. **57**(2), 342–357 (2009)
17. Cortes, C., Vapnik, V.N.: Support vector networks. Mach. Learn. **20**, 1–25 (1995)
18. Courtney, J.F., Klastorin, T.D., Ruefli, T.W.: A goal programming approach to urban-suburban location preferences. Manag. Sci. **18**(6), 258–268 (1972)
19. Crammer, K., Singer, Y.: On the algorithmic implementation of multiclass kernel-based vector machines. J. Mach. Learn. Res. **2**, 265–292 (2002)

20. Delbaen, F.: Coherent Risk Measures on General Probability Spaces, pp. 1–37. Springer, Berlin (2002)
21. Freund, Y., Schapire, R.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. **55**(1), 119–139 (1997)
22. Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: a statistical view of boosting. Ann. Stat. **28**, 337–407 (2000)
23. Grant, M., Boyd, S.: Graph implementations for nonsmooth convex programs. In: Blondel, V., Boyd, S., Kimura, H. (eds.) Recent Advances in Learning and Control, pp. 95–110. Springer, Berlin (2008)
24. Grant, M., Boyd, S.: CVX: Matlab Software for Disciplined Convex Programming, Version 1.21 (2011). http://cvxr.com/cvx
25. Gurobi Optimization, I.: Gurobi Optimizer Reference Manual (2013). http://www.gurobi.com
26. Lam, S., Ng, T., Sim, M., Song, J.: Multiple objectives satisficing under uncertainty. Oper. Res. **61**(1), 214–227 (2013)
27. Lee, Y., Lin, Y., Wahba, G.: Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data. J. Am. Stat. Assoc. **99**, 67–81 (2004)
28. Liu, Y., Shen, X.: Multicategory φ-learning. J. Am. Stat. Assoc. **101**(474), 500–509 (2006)
29. Luedtke, J., Ahmed, S.: A sample approximation approach for optimization with probabilistic constraints. SIAM J. Optim. **19**, 674–699 (2008)
30. Nemirovski, A., Shapiro, A.: Scenario approximation of chance constraints. In: Calafiore, G., Dabbene, F. (eds.) Probabilistic and Randomized Methods for Design Under Uncertainty, pp. 3–48. Springer, London (2005)
31. Nemirovski, A., Shapiro, A.: Convex approximations of chance constrained programs. SIAM J. Optim. **17**, 969–996 (2006)
32. Norton, M., Mafusalov, A., Uryasev, S.: Soft margin support vector classification as buffered probability minimization. J. Mach. Learn. Res. **18**, 1–43 (2017)
33. Norton, M., Uryasev, S.: Maximization of AUC and buffered AUC in classification. Research Report (2015)
34. Poggio, T., Rifkin, R., Mukherjee, S., Niyogi, P.: General conditions for predictivity in learning theory. Nature **428**(6981), 419–422 (2004)
35. Prékopa, A.: On probabilistic constrained programming. In: Proceedings of the Princeton Symposium on Mathematical Programming, pp. 113–138 (1970)
36. Prékopa, A.: Stochastic Programming, pp. 319–371. Kluwer, Dordrecht (1995)
37. Rockafellar, R., Royset, J.: On buffered failure probability in design and optimization of structures. Reliabil. Eng. Syst. Safety **95**(5), 499–510 (2010)
38. Schapire, E., Singer, Y.: Improved boosting algorithms using confidence-rated predictions. Mach. Learn. **37**, 297–336 (1999)
39. Schölkopf, B., Smola, A.J.: Learning with Kernels, pp. 407–423. MIT Press, Cambridge (2002)
40. Shapiro, A., Dentcheva, D., Ruszczyński, A.: Lectures on Stochastic Programming: Modeling and Theory, 2nd edn. Society for Industrial and Applied Mathematics, Philadelphia (2014)
41. Shivaswamy, P.K., Bhattacharyya, C., Smola, A.J.: Second order cone programming approaches for handling missing and uncertain data. J. Mach. Learn. Res. **7**, 1283–1314 (2006)
42. Simon, H.: A behavior model for rational choice. Q. J. Econ. **69**, 99–118 (1955)
43. Simon, H.: Theories of decision-making in economics and behavioral science. Am. Econ. Rev. **49**(3), 253–283 (1959)
44. Vapnik, V.N., Chervonenkis, A.: The necessary and sufficient conditions for consistency in the empirical risk minimization method. Pattern Recognit. Image Anal. **1**(3), 260–284 (1991)
45. Vapnik, V.N., Lerner, A.: Pattern recognition using generalized portrait method. Autom. Remote Control **24**, 744–780 (1963)
46. Vazirani, V.: Approximation Algorithms. Springer, Berlin (2001)
47. Yang, W., Xu, H.: The Coherent Loss Function for Classification. ICML, Stockholm (2014)
48. Zhang, T.: Statistical behavior and consistency of classification methods based on convex risk minimization. Ann. Stat. **32**, 56–85 (2004)

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