INDUSTRIAL AND SYSTEMS ENGINEERING

Stochastic Set-Valued Optimization and its Application to Robust Learning

T. GIOVANNELLI[1], J. TAN[2], AND L. N. VICENTE[2]

[1]University of Cincinnati

[2]Lehigh University

ISE Technical Report 26T-005

Stochastic set-valued optimization and its application to robust learning

T. Giovannelli* J. Tan† L. N. Vicente‡

March 19, 2026

Abstract

In this paper, we develop a stochastic set-valued optimization (SVO) framework tailored for robust machine learning. In the SVO setting, each decision variable is mapped to a set of objective values, and optimality is defined via set relations. We focus on SVO problems with hyperbox sets, which can be reformulated as multi-objective optimization (MOO) problems with finitely many objectives and serve as a foundation for representing or approximating more general mapped sets. Two special cases of hyperbox-valued optimization (HVO) are interval-valued (IVO) and rectangle-valued (RVO) optimization. We construct stochastic IVO/RVO formulations that incorporate subquantiles and superquantiles into the objective functions of the MOO reformulations, providing a new characterization for subquantiles. These formulations provide interpretable trade-offs by capturing both lower- and upper-tail behaviors of loss distributions, thereby going beyond standard empirical risk minimization and classical robust models. To solve the resulting multi-objective problems, we adopt stochastic multi-gradient algorithms and select a Pareto knee solution. In numerical experiments, the proposed algorithms with this selection strategy exhibit improved robustness and reduced variability across test replications under distributional shift compared with empirical risk minimization, while maintaining competitive accuracy.

1 Introduction

In set-valued optimization (SVO), the objective is to minimize a set-valued mapping S from $\mathbb{R}^n$ to $\mathbb{R}^m$, which assigns to each point in $\mathbb{R}^n$ a set of points in $\mathbb{R}^m$, rather than a single value. Two main solution concepts can be used to define optimality: the set approach and the vector approach. In the set approach, optimal solutions (referred to as minimal solutions in SVO contexts) are determined by comparing image sets of S, seeking a point $x_* \in \mathbb{R}^n$ whose image set $S(x_*) \subseteq \mathbb{R}^m$ is not dominated by others based on a set relation. In contrast, the vector approach focuses on comparing individual vectors within the image sets rather than the sets themselves, seeking a point $x_* \in \mathbb{R}^n$ for which there exists a point $y_* \in S(x_*) \subseteq \mathbb{R}^m$ that is non-dominated with respect to all points in $\cup_{x\in\mathbb{R}^n} S(x)$.

*Department of Mechanical and Materials Engineering, University of Cincinnati, Cincinnati, OH 45221, USA (giovanto@ucmail.uc.edu).

†Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, PA 18015-1582, USA (jit423@lehigh.edu).

‡Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, PA 18015-1582, USA (lnv@lehigh.edu).

Our paper adopts the set approach, as it aligns well with the robust learning applications we will propose. Among the set relations proposed for comparing the image sets of a set-valued mapping at different solutions, we will focus on the well-known set-less relations introduced by Kuroiwa et al. [27]. Intuitively, when using the first orthant as an ordering cone, an image set $S(\bar{x})$ dominates another image set $S(\tilde{x})$, with $\bar{x} \neq \tilde{x}$, based on the lower (upper) set-less relation if the points with the minimum (maximum) objective values in $S(\bar{x})$ are below the corresponding points in $S(\tilde{x})$. A feasible point x_* is referred to as a minimal solution if its image set $S(x_*)$ is not dominated by the image set of any other feasible point. This concept is analogous to efficient solutions in vector-valued optimization (VVO), where the objective is a vector function and the image of a feasible point under such a function is a vector in $\mathbb{R}^m$. Specifically, a feasible point is called an efficient solution if its image is not dominated by the image of any other feasible point. When the ordering cone is the non-negative orthant $\mathbb{R}^m_+$, which is the case of interest for our paper, VVO reduces to multi-objective optimization (MOO), and efficient solutions are called Pareto optimal solutions.

More formally, given a set-valued mapping $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$, a closed, convex, and pointed cone $K \subset \mathbb{R}^m$ (where convex means $K + K = K$ and pointed means $K \cap -K = \{0\}$), and a partial order relation between sets with respect to K, denoted by the set-less relation $\preceq_K$ (which will be rigorously introduced in Section 2.1), an SVO problem can be formulated as follows

$$\preceq_K \text{-} \min_{x \in X} S(x), \tag{1.1}$$

where $X \subseteq \mathbb{R}^n$. A class of set-valued optimization problems (1.1), considered in this paper, is hyperbox-valued optimization (HVO), in which each image set $S(x) \subset \mathbb{R}^m$ is a hyperbox specified by component-wise lower and upper bounds (see [1, 13]). HVO provides a foundation for representing or approximating more general set-valued mappings, and two special cases are interval-valued optimization (IVO), when $m = 1$ and $S(x)$ is an interval in $\mathbb{R}$, and rectangle-valued optimization (RVO), when $m = 2$ and $S(x)$ is a rectangle in $\mathbb{R}^2$. Low-dimensional hyperboxes, such as intervals and rectangles, lead to HVO formulations that are solvable and interpretable.

Robust learning motivation. Machine learning (ML) robustness refers to a model's ability to maintain its performance despite uncertainties or adversarial conditions [30, 45]. In practical applications, data is often noisy, incomplete, subject to adversarial attacks, or unrepresentative of the entire population [16, 26], leading to unreliable and inconsistent results in terms of accuracy and fairness across different datasets [14, 19, 33, 35]. In today's data-driven world, robust ML techniques can address such challenges by ensuring that models remain resilient in worst-case scenarios and enhancing their generalizability, allowing them to perform effectively on unseen data. Therefore, through robust learning, we can build ML models that are not only consistently accurate and fair but also resilient to unexpected future circumstances.

Optimization methods are instrumental components to achieve robust ML models and have been incorporated into ML application problems to achieve various forms of robustness. Adversarial robustness ensures a model withstands data perturbations designed to mislead it, and it is typically addressed through min-max or bilevel optimization, where the max/lower-level problem is posed on the variables that perturb the data in a worst-case fashion, while the min/upper-level problem minimizes the training loss function on the model parameters [46]. Robustness to imbalanced data adjusts weights in bilevel formulations to penalize misclassification of underrepresented classes [24, 36, 42]. Robustness to overfitting prevents the model

from excessively fitting training data, and it is achieved through L1/L2 regularization, cross-validation/hyperparameter tuning (which can be framed as a bilevel problem), or distributionally robust optimization (which can be framed as a min-max problem) [9, 29, 34], where the goal is to ensure that a model performs well across a range of possible data distributions.

Hence, the main optimization approaches for robust ML require tuning parameters, optimizing at different levels, or considering different data distributions. Additionally, traditional robust ML techniques, including bilevel, min-max, and regularization methods, fall short in capturing the full complexity of uncertainty in real-world data by reducing variability to a single measure: the empirical risk.

A brief SVO literature review. Research in SVO began in the late 1990s and early 2000s with Kuroiwa's work [27, 28], which introduced set relations to compare sets and establish preferences among them. A comprehensive resource on SVO is the book by Khan et al. [26], which covers foundational concepts for set-valued mappings derived from variational analysis [41], and also discusses optimality conditions, duality theory, and numerical methods for SVO. Another popular introduction to the topic is the review by Hamel et al. [20].

In 2015, Jahn made a fundamental contribution to the field of SVO by proposing vectorization, which transforms an SVO problem into a VVO problem by reducing the comparison of two sets to the pointwise comparison of functions in an infinite-dimensional vector space [22]. Vectorization for SVO plays the same role that scalarization plays for VVO, where a VVO problem is reduced to a real-valued optimization problem by weighting the components of the vector function into a single objective (weighted-sum method) or minimizing one objective subject to the others up to certain thresholds (ε-constrained method) [10]. In [12], Eichfelder and Rocktäschel provided error bounds on the approximation used to discretize the infinite-dimensional vector space resulting from vectorization. In [11], Eichfelder and Gerlach identified classes of SVO that can be formulated as MOO problems in a finite-dimensional objective space, including box-valued and ball-valued optimization.

To the best of our knowledge and according to the book by Khan et al. [25] and survey [20], no stochastic formulations for SVO have been proposed so far. The deterministic algorithms in [25] include a Newton method and an algorithm for solving polyhedral convex SVO problems. In 2015, Jahn proposed an algorithm that performs enumerative pairwise comparisons between sets [23]. In 2021, a deterministic gradient descent method for SVO problems with a set-valued mapping of finite cardinality was developed by Bouza et al. [3].

Our contributions. Motivated by the lack of large-scale, application-driven methodologies for set-valued optimization, this paper develops a stochastic SVO framework in which the stochasticity arises from an underlying data distribution. For this purpose, we consider hyperbox-valued optimization (HVO) models, in particular interval-valued optimization (IVO) and rectangle-valued optimization (RVO). HVO is known to admit an equivalent reformulation as a standard vector-valued optimization (VVO) problem with a finite number of vector components. The reformulated VVO problem can be viewed as a multi-objective optimization (MOO) problem with a finite number of objective functions (2 for IVO and 4 for RVO), obtained by treating each component of the objective vector as a separate objective.

Our stochastic set-valued objectives for IVO and RVO are defined through conditional expectation for super and subquantiles and are accessed in practice through random samples or

mini-batches. A superquantile is also known as Conditional Value-at-Risk or expected shortfall and enjoys the convex characterization of Rockafellar and Uryasev [40]. One contribution of this paper is the introduction and proof of a Rockafellar–Uryasev-type characterization for subquantiles, mirroring the classical representation of superquantiles.

In robust learning, IVO is appealing because it characterizes model performance via lower and upper assessments of the loss, rather than relying on a single average score, thereby retaining more information and potentially improving robustness on testing data. RVO extends this idea to settings where robustness must balance multiple criteria simultaneously, such as a predictive accuracy and a fairness metric. Building on the corresponding MOO reformulation, we design robust learning objectives from lower- and upper-tail risk measures of the loss via sub-/superquantile-type functionals. The resulting IVO and RVO formulations capture the relationship between uncertainty and solution quality, offering an alternative way to interpret loss-tail behavior and robustness trade-offs compared to traditional robust ML techniques such as bilevel, min–max, and regularization methods. We train the resulting models using stochastic mini-batch multi-gradient descent and then extract a single MOO solution via a curvature-based "knee" selection on the Pareto front. Across experiments under distribution shift, the proposed approach produces solutions that are more robust than single-objective empirical risk minimization while maintaining competitive predictive accuracy.

The remainder of the paper is organized as follows. Section 2 introduces the notation and background on partial orders for vectors and sets, set-valued mappings, and the set-less relations used to define optimality, along with a brief review of vectorization in set-valued optimization. Section 3 focuses on hyperbox-valued optimization and shows how the special hyperbox structure yields an equivalent MOO reformulation with a finite number of objectives. Section 4 develops stochastic HVO formulations based on subquantiles and superquantiles, including the resulting stochastic interval- and rectangle-valued models as two special cases. (The Rockafellar–Uryasev-type characterization for subquantiles is established in an appendix.) Section 5 applies these models to robust learning, describes the corresponding training and evaluation procedures, explains how we select a representative Pareto optimal solution, and reports numerical experiments under a distributional shift. Section 6 concludes and discusses directions for future work.

2 Background

2.1 Notation and basic definitions

Throughout the paper, we denote the non-negative orthant of $\mathbb{R}^m$ as $\mathbb{R}^m_+$.

2.1.1 Partial order relations between vectors

Let a, b, and c be vectors in $A \subseteq \mathbb{R}^m$ (see Fig. 1). A closed, convex, and pointed cone $K \subset \mathbb{R}^m$ induces a partial order among such vectors. Specifically, we denote $a \leq_K b$ if $b - a \in K$, which means that *a (weakly) dominates b* with respect to (w.r.t.) K. If K has non-empty interior $\operatorname{int}(K)$, K also induces a strict partial order, in the sense that $a <_K c$ if $c - a \in \operatorname{int}(K)$, which means that *$a$ strictly dominates c* w.r.t. K. Note that when $K = \mathbb{R}^m_+$, which denotes the non-negative orthant of $\mathbb{R}^m$, $a \leq_K b$ implies $a_i \leq b_i$ and $a <_K c$ implies $a_i < c_i$, for all $i \in \{1, \ldots, m\}$.

We say that a point $y \in A$ is a *strictly minimal element* of A w.r.t K if one of the following equivalent relations is satisfied

$$\begin{aligned} &(y - K) \cap A = \{y\} \\ &\Updownarrow \\ &\nexists z \in A \text{ such that } z \leq_K y \text{ and } z \neq y \qquad (2.1) \\ &\Updownarrow \\ &\forall z \in A \text{ such that } z \leq_K y, \text{ we have } z = y \\ &\Updownarrow \\ &\forall z \in A \text{ such that } z \leq_K y, \text{ we have } y \leq_K z. \qquad (2.2) \end{aligned}$$

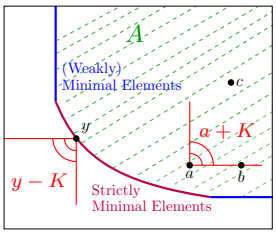


Figure 1: Partial order relations in $\mathbb{R}^2$ for $K = \mathbb{R}^2_+$.

Similarly, a point $y \in A$ is a *(weakly) minimal element* of A w.r.t. K if one of the following equivalent relations is satisfied

$$\begin{aligned} &(y - \text{int}(K)) \cap A = \varnothing \\ &\Updownarrow \\ &\nexists z \in A \text{ such that } z <_K y. \qquad (2.3) \end{aligned}$$

2.1.2 Partial order relation between sets

Let $\mathcal{F}$ be a family of subsets of $\mathbb{R}^m$ and let us consider non-empty sets $A \in \mathcal{F}$ and $B \in \mathcal{F}$. A closed, convex, and pointed cone K induces a partial order between such sets. Among the most popular set relations, one can include (see Fig. 2 (a)):

$$\begin{aligned} \text{upper set-less:} \quad & A \preceq^u_K B \\ & \forall a \in A, \exists b \in B \text{ such that } a \leq_K b \iff A \subset B - K, \\ \text{lower set-less:} \quad & A \preceq^\ell_K B \\ & \forall b \in B, \exists a \in A \text{ such that } a \leq_K b \iff B \subset A + K, \\ \text{set-less:} \quad & A \preceq_K B \iff A \preceq^u_K B \text{ and } A \preceq^\ell_K B. \end{aligned}$$

By replacing $\preceq_K$ with $\prec_K$, one can consider corresponding strict set relations, where K is replaced by its interior $\text{int}(K)$. The corresponding strict set relations are defined as follows:

$$\begin{aligned} \text{upper set-less:} \quad & A \prec^u_K B \\ & \forall a \in A, \exists b \in B \text{ such that } a <_K b \iff A \subset B - \text{int}(K), \\ \text{lower set-less:} \quad & A \prec^\ell_K B \\ & \forall b \in B, \exists a \in A \text{ such that } a <_K b \iff B \subset A + \text{int}(K), \\ \text{set-less:} \quad & A \prec_K B \iff A \prec^u_K B \text{ and } A \prec^\ell_K B. \end{aligned}$$

In this paper, we will use the set-less relation. However, all the definitions regarding set relations can also be formulated using the upper or lower set-less relations.

We say that $A \in \mathcal{F}$ is a *strictly minimal element* of $\mathcal{F}$ w.r.t. K if

$$\forall B \in \mathcal{F} \text{ such that } B \preceq_K A, \text{ we have } A \preceq_K B. \qquad (2.4)$$

Note that (2.4) corresponds to (2.2) in terms of set relations. We say that $A \in \mathcal{F}$ is a *(weakly) minimal element* of $\mathcal{F}$ w.r.t. K if

$$\nexists B \in \mathcal{F} \text{ such that } B \prec_K A. \tag{2.5}$$

Note that (2.5) corresponds to (2.3) in terms of set relations.

2.1.3 Vector-valued optimization and stochastic multi-objective optimization

Given a vector function $F : \mathbb{R}^n \to \mathbb{R}^m$ and a closed, convex, and pointed cone $K \subset \mathbb{R}^m$, let us consider the following vector-valued optimization (VVO) problem

$$\leq_K \text{-}\min_{x \in X} F(x) \;=\; (f_1(x), \ldots, f_m(x)), \tag{2.6}$$

where $X \subseteq \mathbb{R}^n$. When $K = \mathbb{R}^m_+$, a VVO problem can also be referred to as a multi-objective optimization problem. Denoting the image of the set X under the vector function F as $F(X) = \cup_{x \in X}\{F(x)\}$. We say that $x^* \in X$ is a *strictly efficient solution* (or strict Pareto optimal solution, if $K = \mathbb{R}^m_+$) to problem (2.6) if $F(x^*)$ is a strictly minimal element of $F(X)$ w.r.t. K. We say that $x^* \in X$ is a *weakly efficient solution* (or weak Pareto optimal solution, if $K = \mathbb{R}^m_+$) to problem (2.6) if $F(x^*)$ is a weakly minimal element of $F(X)$ w.r.t. K. When $K = \mathbb{R}_+$, the Pareto front of problem (2.6) is defined as the image of the set of Pareto optimal solutions under the vector function F. Each point on the Pareto front is referred to as a nondominated point.

Assume that $K = \mathbb{R}^m_+$ and each objective function $f_i(\cdot)$ is defined as the expected value of a stochastic function $h_i(\cdot, \xi)$, which depends on a random variable ξ within a probability space with probability measure independent of x. One can formulate problem (2.6) as the following stochastic MOO problem

$$\min_{x \in X} F(x) = (f_1(x), \ldots, f_m(x)) \;=\; (\mathbb{E}[h_1(x, \xi)], \ldots, \mathbb{E}[h_m(x, \xi)]), \tag{2.7}$$

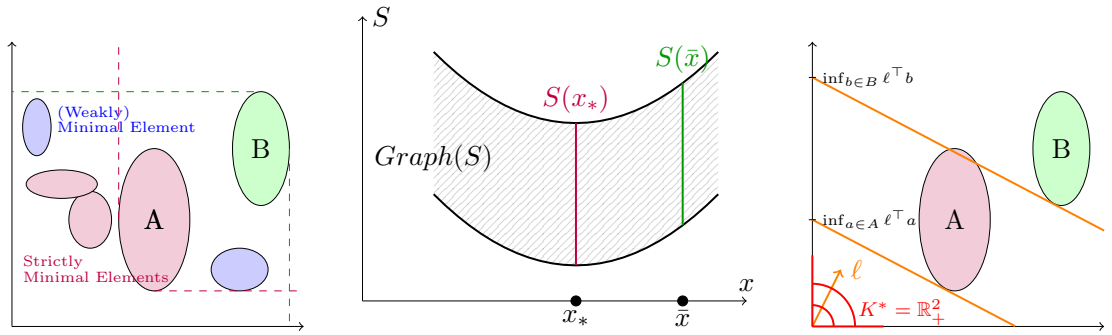
Problem (2.7) can be solved by applying stochastic approximation (SA) methods. The earliest prototypical SA algorithm is known as the *stochastic gradient (SG) algorithm* [6, 38, 43], designed for single-objective problems (i.e., $m = 1$). Its update step at the k-th iteration is defined by $x_{k+1} = x_k - \alpha_k g_i(x_k, \xi_k)$, where $g_i(x_k, \xi_k)$ is a stochastic gradient generated according to ξ_k (e.g., $g_i(x_k, \xi_k) = \nabla h_i(x_k, \xi_k)$) and α_k is a positive stepsize.

2.1.4 Set-valued mappings and set-valued optimization

Given $X \subseteq \mathbb{R}^n$, a set-valued mapping $S : X \rightrightarrows \mathbb{R}^m$ associates a set $S(x) \subset \mathbb{R}^m$ with each $x \in X$. The graph of S is denoted as $Graph(S) = \{(x, u) \mid u \in S(x)\} \subseteq X \times \mathbb{R}^m$. When $S(x)$ is a singleton for all $x \in X$, S is referred to as a single-valued function.

Let us consider the SVO problem in (1.1). Moreover, let us denote the image of the set X under the set-valued function S as $S(X) = \cup_{x \in X} S(x)$. In the literature, two approaches have been proposed to define the optimal solution of such a problem: the set approach and the vector approach. In the set approach, we say that $x^* \in X$ is a *strictly (weakly) minimal solution* to problem (1.1) if $S(x^*)$ is a strictly (weakly) minimal element of the family of subsets $\{S(x)\}_{x \in X}$ w.r.t. K (see Fig. 2 (b) for an example).

In the vector approach, we say that $x^* \in X$ is a strictly (weakly) minimal solution of problem (1.1) if $S(x^*)$ contains a strictly (weakly) minimal element of $S(X)$ w.r.t. K in the vector sense. In this paper, we focus on the set approach, which is the most popular [3].



(a) Partial order relations in $\mathbb{R}^2$ for $K = \mathbb{R}^2_+$ (note that $A \prec_K B$).

(b) When $m = 1$, x_* is a strictly minimal solution of S w.r.t. $K = \mathbb{R}_+$.

(c) Illustration of the first inequality in (2.10).

Figure 2: Key concepts in SVO.

2.2 Vectorization for set-valued optimization

Vectorization reduces an SVO problem to a VVO problem by reducing the comparison of two sets to the pointwise comparison of suitable vector functions [22]. Vectorization for SVO plays the same role that scalarization plays for VVO. Scalarization reduces a VVO problem to a real-valued optimization problem by weighting the components of the vector function into a single objective or minimizing one objective subject to the others up to certain thresholds [10].

Let $K \subseteq \mathbb{R}^m$ be a closed, convex, and pointed cone. Let K^* be the dual cone of K, i.e., $K^* = \{\ell \in \mathbb{R}^m \mid \ell^\top y \geq 0,\ \forall y \in K\}$. Note that the dual cone of $K = \mathbb{R}^m_+$ is $K^* = \mathbb{R}^m_+$. Moreover, let $\mathcal{F}$ be a family of subsets of $\mathbb{R}^m$ and let us consider non-empty sets $A \in \mathcal{F}$ and $B \in \mathcal{F}$. Assuming that both $A + K$ and $B - K$ are closed and convex, we have from [22, Lemma 2.1 and Theorem 2.1] (see Fig. 2 (c)) that

$$A \preceq^{\ell}_K B \iff \inf_{a\in A} \ell^\top a \leq \inf_{b\in B} \ell^\top b,\ \forall \ell \in K^*\setminus\{0\}\,, \tag{2.8}$$

$$A \preceq^{u}_K B \iff \sup_{a\in A} \ell^\top a \leq \sup_{b\in B} \ell^\top b,\ \forall \ell \in K^*\setminus\{0\}\,, \tag{2.9}$$

$$A \preceq_K B \iff \inf_{a\in A} \ell^\top a \leq \inf_{b\in B} \ell^\top b \ \text{ and } \ \sup_{a\in A} \ell^\top a \leq \sup_{b\in B} \ell^\top b,\ \forall \ell \in K^*\setminus\{0\}\,. \tag{2.10}$$

Based on [22, Corollary 2.2], one can apply definition (2.4) with (2.10) to obtain a characterization for sets that are strictly minimal elements. In particular, we have that a set A is a strictly minimal element of $\mathcal{F}$ w.r.t. K if and only if there is no B such that

$$\begin{aligned} &\sup_{b\in B} \ell^\top b \leq \sup_{a\in A} \ell^\top a \ \text{ and } \ \inf_{b\in B} \ell^\top b \leq \inf_{a\in A} \ell^\top a,\ \text{for all } \ \ell \in K^*\setminus\{0\}\,, \text{ and} \\ &\sup_{b\in B} \ell^\top b < \sup_{a\in A} \ell^\top a \ \text{ or } \ \inf_{b\in B} \ell^\top b < \inf_{a\in A} \ell^\top a, \ \text{ for some } \ell \in K^*\setminus\{0\}\,. \end{aligned} \tag{2.11}$$

Instead of $K^*\setminus\{0\}$, one can consider $\tilde{K} = K^* \cap \{y \in \mathbb{R}^m \mid \|y\| = 1\}$, and then a discretization $\{\ell_1, \ldots, \ell_p\} \subset \tilde{K}$. Therefore, as a consequence of (2.10), one has to solve 4 optimization subproblems (possibly in parallel) for each ℓ_i, with $i \in \{1, \ldots, p\}$. When A and B are polyhedrons (which includes the case of intervals), the subproblems are LPs and the discretization $\{\ell_1, \ldots, \ell_p\}$ can be coarser than in the general case [23].

For any non-empty set $A \in \mathcal{F}$, let us define the following vector-valued functions

$$\begin{array}{rcl} v_\ell(A) & = & \left(\begin{array}{c} \sup_{a\in A} \ell^\top a \\ \inf_{a\in A} \ell^\top a \end{array} \right), \forall \ell \in K^* \backslash \{0\}, \quad \text{and} \\ v(A) & = & (v_\ell(A) \mid \ell \in K^* \backslash \{0\}). \end{array} \tag{2.12}$$

Note that $v(\cdot)$ is a vector-valued function with an infinite number of component functions. Let us now introduce the following cone

$$K_\Pi = \prod_{\ell \in K^* \backslash \{0\}} K_\ell, \text{ with } K_\ell = \mathbb{R}^2_+,$$

which is given by the Cartesian product of an infinite number of two-dimensional non-negative orthants. Based on (2.8)–(2.10) and (2.12), vectorization reduces the comparison of two sets to the pointwise comparison of vector functions as follows

$$A \preceq_K B \iff v_\ell(A) \leq_{\mathbb{R}^2_+} v_\ell(B), \ \forall \ell \in K^* \backslash \{0\} \iff v(A) \leq_{K_\Pi} v(B), \tag{2.13}$$

where $A \in \mathcal{F}$ and $B \in \mathcal{F}$ are non-empty sets.

We can now include Theorem 2.1 below, provided in [22, Theorem 3.1], which states that solving an SVO problem amounts to solving a corresponding VVO problem.

Theorem 2.1 (Vectorization theorem) *Let $S(x) + K$ and $S(x) - K$ be closed and convex for all $x \in X$. Then, $\bar{x} \in X$ is a strictly minimal solution to problem (1.1) if and only if $\bar{x}$ is a strictly efficient solution to the following VVO problem*

$$\leq_{K_\Pi} \text{-}\min_{x\in X} V(x), \quad \text{where } \ V(x) = v(S(x)), \text{ for all } \ x \in X. \tag{2.14}$$

In problem (2.14), the vector-valued objective function is given by an infinite number of component functions. In [23], the author shows that the number of component functions is finite if the sets $S(x)$ are polyhedral, which includes the case when the sets $S(x)$ are given by intervals or rectangles.

3 Hyperbox-valued optimization

In this section, we consider the general hyperbox-valued optimization (HVO) problem, where the dimension is $m = k \in \mathbb{N}^+$ and $k \geq 2$. Given mappings

$$F_{lb} : \mathbb{R}^n \to \mathbb{R}^k, \quad F_{ub} : \mathbb{R}^n \to \mathbb{R}^k,$$

satisfying $F_{lb}(x) \leq_K F_{ub}(x)$ for all $x \in X$ (where $K = \mathbb{R}^k_+$), we define a hyperbox as follows

$$[F_{lb}(x), F_{ub}(x)] = (\{F_{lb}(x)\} + \mathbb{R}^k_+) \cap (\{F_{ub}(x)\} - \mathbb{R}^k_+).$$

The HVO problem is then formulated as

$$\preceq_K \text{-}\min_{x\in X} S(x) \quad \text{with} \quad S(x) = [F_{lb}(x), F_{ub}(x)]. \tag{3.1}$$

The vectorization theorem for SVO stated in Theorem 2.1 says that a set-valued optimization problem is equivalent to a vector-valued optimization problem with a number of objectives that can be infinite. In particular cases like the HVO, this number can be shown to be finite. For that purpose, we first describe a result (first stated in [21] and later proved in [11]) explaining how to directly compare any two hyperboxes of the type defined above.

Proposition 3.1 (Comparison of hyperboxes) *Let $A, B \subseteq \mathbb{R}^k$ be hyperboxes,*

$$A \coloneqq [a^1, b^1] = (\{a^1\} + \mathbb{R}^k_+) \cap (\{b^1\} - \mathbb{R}^k_+), \quad B \coloneqq [a^2, b^2] = (\{a^2\} + \mathbb{R}^k_+) \cap (\{b^2\} - \mathbb{R}^k_+),$$

with $a^1, a^2, b^1, b^2 \in \mathbb{R}^k$ satisfying $a^1 \leq_K b^1$ and $a^2 \leq_K b^2$. Then:

$$A \preceq^l_K B \iff a^1 \leq_K a^2 \quad \text{and} \quad A \preceq^u_K B \iff b^1 \leq_K b^2.$$

The following result describes the equivalent vector-valued optimization problem for HVO. As noted in [11], its proof follows directly from Proposition 3.1.

Theorem 3.1 (Vectorization for HVO) *Consider the HVO problem (3.1) with hyperboxes $S(x) = [F_{lb}(x), F_{ub}(x)] = (\{F_{lb}(x)\} + \mathbb{R}^k_+) \cap (\{F_{ub}(x)\} - \mathbb{R}^k_+)$. Let $K = \mathbb{R}^k_+$. Then the following statements are equivalent:*

(i) $\bar{x} \in X$ is a strictly minimal solution to (3.1) with respect to K.

(ii) $\bar{x} \in X$ is a strictly efficient solution of the following vector-valued optimization problem

$$\leq_{K \times K} \operatorname{-min}_{x \in X} \begin{pmatrix} F_{ub}(x) \\ F_{lb}(x) \end{pmatrix}, \tag{3.2}$$

with respect to $K \times K$.

In particular, since we are considering $K \times K = \mathbb{R}^{2k}_+$, the vector-valued optimization problem (3.2) is further equivalent to

$$\leq_{\mathbb{R}^{2k}_+} \operatorname{-min}_{x \in X} \begin{pmatrix} F_{ub}(x) \\ F_{lb}(x) \end{pmatrix}. \tag{3.3}$$

We note that the general HVO framework naturally includes two known specific set-valued optimization problems. When $k = 1$, the hyperboxes reduce to real intervals, and problem (3.1) becomes the standard interval-valued optimization (IVO) problem. Given $f_{\ell b} : \mathbb{R}^n \to \mathbb{R}$, $f_{ub} : \mathbb{R}^n \to \mathbb{R}$, and $f_{\ell b}(x) \leq f_{ub}(x)$, for all $x \in X$, an IVO problem can be formulated as follows

$$\preceq_{\mathbb{R}_+} \operatorname{-min}_{x \in X} S(x) \quad \text{with} \quad S(x) = \{y \in \mathbb{R} \mid f_{\ell b}(x) \leq y \leq f_{ub}(x)\}. \tag{3.4}$$

For IVO problems, Theorem 3.1 simplifies to the following corollary.

Corollary 3.1 *Let $K = \mathbb{R}_+$. Then, $\bar{x} \in X$ is a strictly minimal solution to problem (3.4) w.r.t. K if and only if $\bar{x}$ is a strictly efficient solution of the following vector-valued optimization problem*

$$\leq_{\mathbb{R}^2_+} \operatorname{-min}_{x \in X} (f_{\ell b}(x), f_{ub}(x)). \tag{3.5}$$

When $k = 2$, the hyperboxes are rectangles in $\mathbb{R}^2$, so the HVO problem (3.1) reduces to the rectangular-valued optimization (RVO) problem. Given $F_{\ell b} : \mathbb{R}^n \to \mathbb{R}^2$, $F_{ub} : \mathbb{R}^n \to \mathbb{R}^2$, and $F_{\ell b}(x) \leq_K F_{ub}(x)$, for all $x \in X$, an RVO problem can be formulated as follows

$$\preceq_{\mathbb{R}^2_+} \operatorname{-min}_{x \in X} S(x) \quad \text{with} \quad S(x) = [F_{\ell b}(x), F_{ub}(x)]. \tag{3.6}$$

Similarly, for RVO problems, the vectorization theorem 3.1 for the HVO simplifies to the following corollary.

Corollary 3.2 *Let $K = \mathbb{R}^2_+$. Then $\bar{x} \in X$ is a strictly minimal solution to problem* (3.6) *w.r.t K if and only if $\bar{x}$ is a strictly efficient solution of the following vector-valued optimization problem*

$$\leq_{\mathbb{R}^4_+} \text{-}\min_{x \in X} \begin{pmatrix} F_{ub}(x) \\ F_{\ell b}(x) \end{pmatrix}. \tag{3.7}$$

4 Stochastic hyperbox-valued optimization

We will introduce new stochastic hyperbox-valued optimization formulations based on super and sub-quantiles. Such formulations will be used for ML robustness in Section 5.

4.1 Superquantiles and subquantiles

Let us consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with probability measure $\mathbb{P}$ and σ-algebra $\mathcal{F}$ containing subsets of the sample space Ω. Let Z be a random variable defined in Ω. Denoting $F_Z : \mathbb{R} \to [0,1]$ as the cumulative distribution function of Z (i.e., $F_Z(z) = \mathbb{P}(Z \leq z)$ for all $z \in \mathcal{Z}$) and $Q_Z : (0,1) \to \mathbb{R}$ as the quantile function of Z (i.e., $Q_Z(p) = \min\{z \mid F_Z(z) \geq p\}$ for all $p \in (0,1)$), we can define the superquantile* of Z as a function $\bar{\mathbb{S}}_p$ given by

$$\bar{\mathbb{S}}_p[Z] \;=\; \frac{1}{1-p} \int_p^1 Q_Z(p')dp', \text{ for all } p \in (0,1) \tag{4.1}$$

and the subquantile of Z as a function $\underline{\mathbb{S}}_p[Z]$ given by

$$\underline{\mathbb{S}}_p[Z] \;=\; \frac{1}{p} \int_0^p Q_Z(p')dp', \text{ for all } p \in (0,1). \tag{4.2}$$

One can address the cases where $p = 0$ or $p = 1$ by extending the definition in (4.1) as follows (see [39])

$$\bar{\mathbb{S}}_0[Z] = \mathbb{E}[Z] \;\text{ and }\; \bar{\mathbb{S}}_1[Z] = \sup_{\omega \in \Omega} Z(\omega). \tag{4.3}$$

Similar to (4.3), one can extend the definition in (4.2) for $p = 0$ and $p = 1$ by setting

$$\underline{\mathbb{S}}_0[Z] = \inf_{\omega \in \Omega} Z(\omega) \;\text{ and }\; \underline{\mathbb{S}}_1[Z] = \mathbb{E}[Z]. \tag{4.4}$$

When Z has a continuous probability distribution, we can give characterizations equivalent to (4.1) for the superquantile and (4.2) for the subquantile. In particular, the superquantile of Z reduces to

$$\bar{\mathbb{S}}_p[Z] \;=\; \mathbb{E}[Z \mid Z \geq Q_Z(p)], \text{ for all } p \in (0,1), \tag{4.5}$$

which shows that $\bar{\mathbb{S}}_p[Z]$ is the mean of the upper p-tail distribution of Z. Analogously, the subquantile of Z reduces to

$$\underline{\mathbb{S}}_p[Z] \;=\; \mathbb{E}[Z \mid Z \leq Q_Z(p)], \text{ for all } p \in (0,1), \tag{4.6}$$

which shows that $\underline{\mathbb{S}}_p[Z]$ is the mean of the lower p-tail distribution of Z.

*In risk management and finance, the quantile and superquantile functions are also known as Value-at-Risk (VaR) and Conditional Value-at-Risk (CVaR), respectively [39].

An equivalent characterization of the superquantile function (4.1) that applies when Z has a general probability distribution is (see [40, Theorem 1])

$$\bar{\mathbb{S}}_p[Z] \;=\; \min_{\eta\in\mathbb{R}} \big\{\eta + \frac{1}{1-p}\mathbb{E}[\max(Z-\eta,0)]\big\}. \tag{4.7}$$

The objective function of the optimization problem in (4.7) is non-differentiable and convex in η [44, Section 3.2]. The optimal solution in η is equal to $Q_Z(p)$. Moreover, $\bar{\mathbb{S}}_p[Z]$ is convex in Z.

One can derive a similar characterization for the subquantile function $\underline{\mathbb{S}}_p[Z]$ (see Appendix A) as follows:

$$\underline{\mathbb{S}}_p[Z] \;=\; \max_{\eta\in\mathbb{R}} \big\{\eta - \frac{1}{p}\mathbb{E}[\max(\eta-Z,0)]\big\}. \tag{4.8}$$

Similarly, the optimal solution in η is equal to $Q_Z(p)$. However, $\underline{\mathbb{S}}_p[Z]$ is now a concave function in η and Z.

4.2 Stochastic interval-valued optimization

Let us consider a continuously differentiable function $\psi : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}$ and two scalars $\underline{p} \in [0,1]$ and $\bar{p} \in [0,1]$ (we include the boundaries in such intervals to account for the extended definitions provided in (4.3) and (4.4)). In this paper, we focus on stochastic IVO problems that can be written as follows

$$\text{problem (3.4) with } f_{\ell b}(x) = \underline{\mathbb{S}}_{\underline{p}}[\psi(x,\xi)] \text{ and } f_{ub}(x) = \bar{\mathbb{S}}_{\bar{p}}[\psi(x,\xi)], \tag{4.9}$$

where $\xi \in \mathbb{R}^d$ is a random variable having a probability distribution independent of x.

As a consequence of Corollary 3.1 in Section 3, the set of minimal solutions to the IVO problem (3.4) can be determined by solving the corresponding multi-objective problem (3.5), where $f_{\ell b}(x)$ and $f_{ub}(x)$ are given by (4.9). One can consider different cases based on the values of $\underline{p}$ and $\bar{p}$. In Sections 4.2.1, 4.2.2, and 4.2.3, we focus on the most significant ones, from which all the remaining cases can be trivially obtained.

4.2.1 Case $\underline{p} = 1$ and $\bar{p} = 0$

Recalling the extended definitions (4.3) and (4.4), the case $(\underline{p} = 1, \bar{p} = 0)$ results in a bi-objective optimization problem where both objective functions are given by equal expectations, i.e., $f_{\ell b}(x) = f_{ub}(x) = \mathbb{E}[\psi(x,\xi)]$. Such a problem can solved by applying the classical stochastic gradient method [2].

4.2.2 Case $\underline{p} = 0$ and $\bar{p} = 1$

When $\psi(x,\xi)$ is a continuous function of x and ξ, the case $(\underline{p} = 0, \bar{p} = 1)$ leads to IVO problems for which the equivalent bi-objective problem (3.5) can be written as

$$\leq_{\mathbb{R}^2_+} \text{-}\min_{x\in X} \;\; (f_{\ell b}(x) = \min_{\xi\in P} \psi(x,\xi),\;\; f_{ub}(x) = \max_{\xi\in P} \psi(x,\xi)), \tag{4.10}$$

where $P \subseteq \mathbb{R}^d$ represents the set of values taken on by ξ, assumed compact. Note that problem (4.10) is deterministic. Danskin's Theorem [7] provides a way to compute the subdifferentials

of f_{ub} and $f_{\ell b}$. In particular, if $\psi(x,\xi)$ is differentiable in x for all $\xi \in P$ and $\nabla_x \psi(x,\xi)$ is a continuous function of ξ for all $x \in X$, then the subdifferential of f_{ub} is

$$\partial f_{ub}(x) = \text{conv}\{\nabla_x \psi(x,\xi) \mid \xi \in P_{ub}(x)\}, \text{ where } P_{ub}(x) = \{\bar{\xi} \in P \mid \psi(x,\bar{\xi}) = \max_{\xi \in P} \psi(x,\xi)\}.$$

The subdifferential of $f_{\ell b}$ and the corresponding set $P_{\ell b}$ can be obtained from similar steps. To solve problem (4.10), one can apply a multi-subgradient method (and by that we mean the multi-gradient method using subgradients instead of gradients [31]) or a multi-gradient sampling method (and by that we mean the extension of gradient sampling [5] to MOO).

4.2.3 Case $\underline{p} \in (0,1)$ and $\bar{p} \in (0,1)$

We will first focus on f_{ub}. Assume that one has access to a set of N i.i.d. samples of ξ, denoted as $\{\xi_1, \ldots, \xi_N\}$. One can now approximate the superquantile $\bar{\mathbb{S}}_{\bar{p}}[\psi(x,\xi)]$ in $f_{ub}(x)$ using data samples, resulting in

$$f_{ub}^N(x) \;=\; \bar{\mathbb{S}}_{\bar{p}}^N[\psi(x,\xi)] = \min_{\eta \in \mathbb{R}} \big\{\eta + \frac{1}{N(1-\bar{p})} \sum_{i=1}^{N} [\max(\psi(x,\xi_i) - \eta, 0)]\big\}, \tag{4.11}$$

which is obtained from the empirical version of (4.7) by setting $Z = \psi(x,\xi)$ and $p = \bar{p}$. Similarly, for $f_{\ell b}$, the empirical approximation of the subquantile can be written as

$$f_{\ell b}^N(x) \;=\; \underline{\mathbb{S}}_{p}^N\big[\psi(x,\xi)\big] \;=\; \max_{\eta \in \mathbb{R}} \Big\{\eta \;-\; \frac{1}{Np} \sum_{i=1}^{N} [\max(\,\eta - \psi(x,\xi_i), 0)]\Big\}. \tag{4.12}$$

4.3 Stochastic rectangle-valued optimization

In this section, let us consider two continuously differentiable functions ψ_1 and ψ_2 (both from $\mathbb{R}^n \times \mathbb{R}^d$ to $\mathbb{R}$) with two scalars $\underline{p}$ and $\bar{p}$. In this paper, we are focusing on the following stochastic RVO problem

$$\text{problem (3.6) with } F_{\ell b}(x) = (\underline{\mathbb{S}}_{\underline{p}}[\psi_1(x,\xi)], \bar{\mathbb{S}}_{\bar{p}}[\psi_1(x,\xi)]) \text{ and } F_{ub}(x) = (\underline{\mathbb{S}}_{\underline{p}}[\psi_2(x,\xi)], \bar{\mathbb{S}}_{\bar{p}}[\psi_2(x,\xi)]), \tag{4.13}$$

where $\xi \in \mathbb{R}^d$ is a random variable having a probability distribution independent of x. Following Corollary 3.2 in Section 3, determining the minimal solutions of (4.13) reduces to solving the multi-objective problem (3.7), with the four objectives specified by (4.13). For the stochastic RVO problem, one can also consider different cases based on the values of $\underline{p}$ and $\bar{p}$, which we have discussed in Section 4.2.

5 Robust learning through stochastic hyperbox-valued optimization

To apply set-valued optimization in machine learning (ML), one can reformulate an ML problem as an SVO problem where a set-valued empirical risk function is minimized instead of a traditional finite-sum function. The set-valued empirical risk function, denoted as $L(x) \subseteq \mathbb{R}$, associates a set of loss function values (computed over all the points in the training set) with

each $x \in \mathbb{R}^n$, where x represents the vector of parameters of an ML model (e.g., weights of a neural network). A solution obtained through set-valued ML can result in lower errors on testing data compared to traditional ML because the set-valued empirical risk function considers both the expected value and variance of the loss function across all training data points, while traditional empirical risk focuses solely on the expected value.

The setting we consider in this paper is binary supervised classification. Let us denote a features/labels dataset used for training by $\mathcal{D}_{train} = \{(u_i, v_i),\, i \in \{1, \ldots, N\}\}$, which consists of N pairs of a real feature vector $u_i \in \mathbb{R}^d$ and a target label $v_i \in \{-1, +1\}$. We assume that for any data point $i \in \{1, \ldots, N\}$, the classification is deemed correct if the right label is predicted. To evaluate the loss incurred when using the prediction function $\phi(u; x)$, where $x \in \mathbb{R}^n$ is a vector of parameters, we use a loss function $\ell(\phi(u; x), v)$.

In traditional ML, we train the model by minimizing the single-valued empirical risk function $\mathcal{L}(x) = (1/N) \sum_{i=1}^{N} \ell(\phi(u_i; x), v_i)$, which computes the average loss over the samples in $\mathcal{D}_{train}$. A different perspective consists of minimizing a set-valued empirical risk mapping $L(x) = \{\ell(\phi(u_i; x), v_i) \mid i \in \{1, \ldots, N\}\}$, which computes a set of loss values over all training points. Note that $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}$ and $L : \mathbb{R}^n \rightrightarrows \mathbb{R}$. We will develop stochastic IVO and RVO formulations based on super and sub-quantiles, which provide flexibility in capturing likely scenarios while avoiding extreme ones that may not reflect practical outcomes.

5.1 Robust learning through stochastic interval-valued optimization

Let us consider a continuously differentiable function $\psi : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}$ and two scalars $\underline{p} \in [0, 1]$ and $\bar{p} \in [0, 1]$ (we include the boundaries in such intervals to account for the extended definitions of superquantiles and subquantiles provided in (4.3)). Let ξ be a random vector taking values in $\Xi \subseteq \mathbb{R}^d$ with a probability distribution independent of x. We propose reformulating any ML problem involving a single objective (e.g., minimizing misclassification error) by using the stochastic IVO problem (4.9), where $\psi(x, \xi)$ plays the role of the loss function $\ell(\phi(u; x), v)$, with $\xi = (u, v)$. Figure 3 provides an illustration for $d = 1$. As part of this section, we will use the bi-objective optimization problem (3.5) to study the robustness of the optimal solution to this stochastic IVO problem in popular ML applications. Remark 5.1 motivates the proposed robust IVO approach by providing an example that interprets the geometric meaning of Corollary 3.1 and contrasts the resulting solution set with both ERM and the classical min–max robust solution.

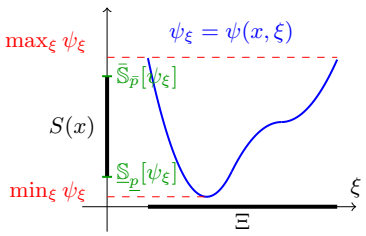
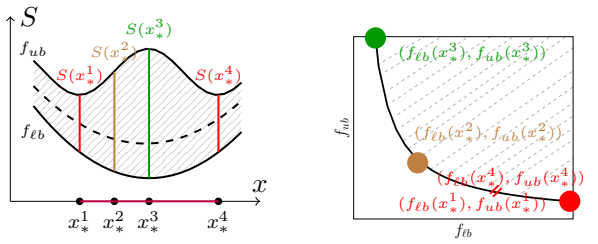


Figure 3: $S(x)$ for problem (4.9).



(a) Interval-valued mapping.

(b) Pareto front for (3.5).

Figure 4: Vectorization theorem for IVO.

Remark 5.1 *Figure 4 illustrates Corollary 3.1 and problem* (3.5) *when* $f_{\ell b}(x) = \underline{\mathbb{S}}_{\underline{p}}[\psi(x,\xi)]$ *and* $f_{ub}(x) = \bar{\mathbb{S}}_{\bar{p}}[\psi(x,\xi)]$*, where* $\psi(x,\xi)$ *represents the training loss function. In Figure 4(a), all points between* x_*^1 *and* x_*^4 *are minimal solutions of* S *w.r.t.* $\mathbb{R}_+$*. Such points correspond to the strict Pareto optimal solutions of problem* (3.5)*, as shown in Figure 4(b). Note that* x_*^3 *is the minimizer of the dashed line in Figure 4(a), which one can interpret as the single-valued expected training loss function* $\mathbb{E}[\psi(x,\xi)]$*. Such a minimizer results in training loss values that are either very low (risk of overfitting) or very high (risk of underfitting). Additionally,* x_*^1 *and* x_*^4 *are local minimizers of* f_{ub} *and represent the solutions from the traditional robust optimization approach (min-max formulation), which focuses on the worst-case scenario. In contrast, points like* x_*^2 *offer a better compromise and are expected to perform more effectively on validation data.*

5.2 Robust learning through stochastic rectangle-valued optimization

We claim that any ML application problem involving two objectives (e.g., minimizing both misclassification error and social bias) can be formulated as a stochastic RVO problem (4.13). In addition to the accuracy loss function ℓ, one can consider a second objective, such as a fairness loss function ℓ_F, resulting in an additional set-valued empirical risk function $L_F(x) = \{\ell_F(\phi(u_i;x), v_i) \mid i \in \{1,\dots,N\}\}$. Note that here we focus on robustness to biased datasets, but other loss functions can be used to address different forms of robustness or fairness. Given two continuously differentiable functions ψ_1 and ψ_2 (both from $\mathbb{R}^n \times \mathbb{R}^d$ to $\mathbb{R}$), $\underline{p}$, $\bar{p}$, and ξ, we propose reformulating a fair ML problem by using the four-objective optimization problem (3.7) with $F_{\ell b}(x) = (\underline{\mathbb{S}}_{\underline{p}}[\psi_1(x,\xi)], \bar{\mathbb{S}}_{\bar{p}}[\psi_1(x,\xi)])$ and $F_{ub}(x) = (\underline{\mathbb{S}}_{\underline{p}}[\psi_2(x,\xi)], \bar{\mathbb{S}}_{\bar{p}}[\psi_2(x,\xi)])$, where ψ_1 and ψ_2 represent the accuracy and fairness loss functions.

5.3 Numerical results

5.3.1 Solving the optimization/learning problems

First, as a baseline for comparison, we consider the empirical risk minimization (ERM) model obtained by minimizing the empirical risk $\mathcal{L}(x)$. We denote an ERM solution by x_{ERM}, defined as

$$x_{\text{ERM}} \in \arg\min_x \mathcal{L}(x) = \arg\min_x \frac{1}{N}\sum_{i=1}^{N} \ell\big(\phi(u_i;x),\, v_i\big).$$

We compute x_{ERM} using the standard stochastic gradient (SG) method.

Next, we need to select x_{IVO} as a certain Pareto optimal solution of the proposed bi-objective problem (3.5) for $f_{\ell b}(x) = \underline{\mathbb{S}}_{\underline{p}}[\psi(x,\xi)]$ and $f_{ub}(x) = \bar{\mathbb{S}}_{\bar{p}}[\psi(x,\xi)]$, where $\psi(x,\xi)$ is the loss function $\ell(\phi(u;x),v)$ with $\xi = (u,v)$. However, the set of Pareto optimal solutions typically contains

many candidate solutions, and it is therefore neither practical nor principled to choose an arbitrary point in each test instance. Instead, an automatic selection rule is needed to identify a representative solution in a consistent and reproducible manner, and Pareto knee solutions offer an interesting trade-off as they are verbally characterized as Pareto optimal solutions corresponding to nondominated points on the Pareto front where a small improvement in any objective causes a large deterioration on at least one other objective [4, 8, 17]. We select x_{IVO} as the knee solution using an angle-based criterion in the objective space. Specifically, after ordering the candidate nondominated points on the approximated front, we choose the knee solution as the Pareto optimal solution corresponding to the point with the smallest interior angle formed by its two adjacent neighbors, as this identifies the sharpest local bend of the Pareto front.

We also denote by x_{RVO} an approximate selected Pareto knee solution of the proposed four-objective problem (3.7) with $F_{\ell b}(x) = (\underline{\mathbb{S}}_{\underline{p}}[\psi_1(x,\xi)], \bar{\mathbb{S}}_{\bar{p}}[\psi_1(x,\xi)])$ and $F_{ub}(x) = (\underline{\mathbb{S}}_{\underline{p}}[\psi_2(x,\xi)], \bar{\mathbb{S}}_{\bar{p}}[\psi_2(x,\xi)])$, where ψ_1 and ψ_2 represent the accuracy and fairness loss functions. To compute x_{RVO}, we first orthogonally project all 4-dimensional points of the computed Pareto front onto the bi-dimensional space of accuracy and into the bi-dimensional space of fairness (resulting into 2 subsets of bi-dimensional approximated Pareto fronts). Then, for each of them, we calculate a knee point using the procedure described for x_{IVO}, resulting in 2 knees. For each knee, we then calculate a pair of weights (summing to one), by relating them to the corresponding extreme points. Finally, we sum the two knees in a weighted fashion (normalizing the weights to 1) to obtain x_{RVO}.

To compute the Pareto fronts needed for $x_{\text{IVO}}/x_{\text{RVO}}$, we use the Pareto-front stochastic multi-gradient (PF-SMG) method (see Appendix D). In doing so, to handle the inner non-smooth hinge term $\max(\cdot, 0)$ in the empirical sub/superquantile-based objectives (4.11) and (4.12), we apply the smoothing stochastic gradient (SSG) method (see Appendix C). Specifically, each objective of the bi-objective and four-objective MOO reformulations derived from (3.5) and (3.7) is smoothed, allowing the computation of stochastic gradients. This smoothing leads to stable gradient estimates in practice. For the outer minimization over the auxiliary variable η in (4.7) and (4.8), in the finite-sum setting with samples $z_1, \ldots, z_n$, we approximate the quantile $Q_z(p)$ by sorting $\{z_i\}_{i=1}^n$ and selecting the empirical p-quantile. For comparison, we also report results obtained with the single-solution stochastic multi-gradient (SMG) method (see Appendix B), which returns a single Pareto point rather than an entire front.

5.3.2 Robust evaluation

To compare ERM and stochastic IVO/RVO in terms of robustness, we evaluate each model on multiple independent test sets. Specifically, we consider N_{rep} test replications. For each replication $r \in \{1, \ldots, N_{\text{rep}}\}$, we draw an independent test set

$$\mathcal{D}_{\text{test}}^{(r)} \;=\; \big\{(u_i^{(r)}, v_i^{(r)})\big\}_{i=1}^{N_{\text{test}}}$$

from a possibly shifted (or class-imbalanced) test distribution. More specifically, starting from the original, unmodified test set, we construct class-imbalanced test sets in each replication by adjusting the class proportions via sub-sampling from either the positive or the negative class. This yields test sets in which the fraction of positive instances is either smaller or larger than that in $\mathcal{D}_{\text{train}}$, thereby inducing a controlled distributional shift. Such shifts are common in practice: models are often trained on historical data from a particular region or time period, but deployed on populations whose demographic or economic characteristics differ across regions or evolve

over time. Under this setting, our approach is expected to deliver more robust performance in the presence of significant distributional changes.

Let us now introduce the metrics used for the robust evaluation. Let the empirical test risk for the r-th test replication be defined as

$$\mathrm{Loss}^{(r)}(x) = \frac{1}{N_{\mathrm{test}}} \sum_{i=1}^{N_{\mathrm{test}}} \mathrm{loss}\big(\phi(u_i^{(r)}; x),\, v_i^{(r)}\big),$$

where $\mathrm{loss}(\cdot)$ denotes either the standard empirical loss function $\ell(\cdot)$ used in the ERM model, the weighted loss $\lambda \underline{\mathbb{S}}_{\underline{p}}[\ell(\cdot)] + (1-\lambda)\overline{\mathbb{S}}_{\bar{p}}[\ell(\cdot)]$ used in the IVO formulation (where the sub- and super-quantiles are represented as $f_{lb}(x)$ and $f_{ub}(x)$, respectively), or the fairness loss $\ell_F(\cdot)$ used in the RVO formulation. The weight λ is chosen according to the knee point of the Pareto front obtained from the IVO model.

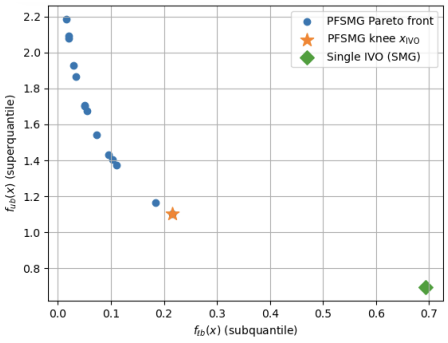
Since the testing procedure is repeated for N_{rep} replications, the robustness of the candidate models can be assessed by applying standard statistical measures to the operator $\mathrm{Loss}^{(r)}(\cdot)$ across replications, such as the sample mean and sample variance. We denote the mean and variance of $\mathrm{Loss}(\cdot)$ as

$$M(x) = \frac{1}{N_{\mathrm{rep}}} \sum_{i=1}^{N_{\mathrm{rep}}} \mathrm{Loss}^{(i)}(x), \qquad V(x) = \frac{1}{N_{\mathrm{rep}} - 1} \sum_{i=1}^{N_{\mathrm{rep}}} \big(\mathrm{Loss}^{(i)}(x) - M(x)\big)^2.$$

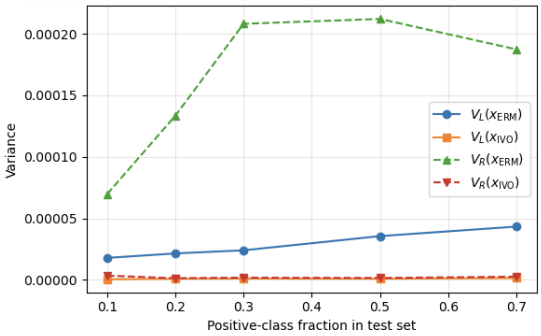
5.3.3 Adult dataset results

We test our methodology on the Adult dataset, a standard benchmark for binary classification. Each sample $u \in \mathbb{R}^d$ represents demographic and socio-economic attributes of an individual (e.g., age, education level, occupation, hours per week), and the binary label $v \in \{0, 1\}$ indicates whether the individual's annual income exceeds \$50,000. After standard preprocessing (one-hot encoding of categorical variables, normalization of continuous features, and removal of examples with missing values), we randomly split the data into a training set $\mathcal{D}_{\mathrm{train}}$ and an initial test set $\mathcal{D}_{\mathrm{test}}$. And we choose logistic regression as our prediction model. For this test and all other tests in the following sections, we use the same experimental settings. In particular, to ensure a fair comparison, all training procedures are performed with the same computational budget of 1500 iterations, the same initial stepsize of 1.3 (with the same discount rate). For the test phase, we use $N_{\mathrm{rep}} = 30$ replications of the test procedure.

Figure 5a reports the PF-SMG output in the $\big(f_{\ell b}(x), f_{ub}(x)\big)$ plane. The non-dominated set forms a Pareto front, and the chosen knee point x_{IVO} (orange star) provides a balanced compromise between the lower- and upper-tail criteria. For reference, the single-objective IVO solution produced by SMG (green diamond) achieves a much smaller $f_{ub}(x)$ but with a markedly larger $f_{\ell b}(x)$, illustrating that extreme behavior may arise from a fixed scalarization. Let us denote by $V_L(x) = V(x)$ the variance evaluated at $x = x_{\mathrm{ERM}}$ with $\mathrm{loss}(\cdot) = \ell(\cdot)$, and by $V_R(x) = V(x)$ the variance evaluated at $x = x_{\mathrm{IVO}}$ with $\mathrm{loss}(\cdot) = \lambda \underline{\mathbb{S}}_{\underline{p}}[\ell(\cdot)] + (1-\lambda)\overline{\mathbb{S}}_{\bar{p}}[\ell(\cdot)]$. Figure 5b summarizes stability under altered label proportions. Across all tested positive-class fractions, x_{IVO} yields uniformly smaller replication-to-replication variability than x_{ERM}.



(a) PF-SMG approximation of the non-dominated set and the selected knee point.

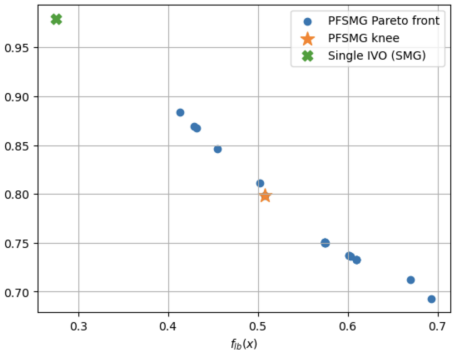


(b) Replication variability under different positive-class fractions.

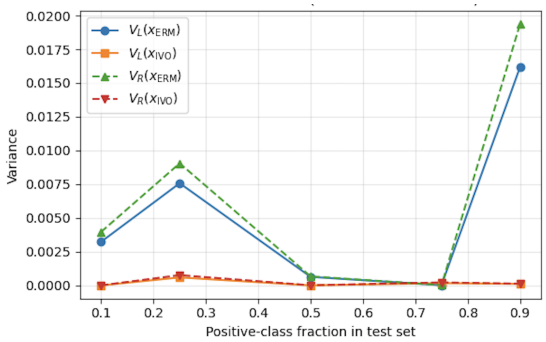
Figure 5: Adult dataset results for IVO.

5.3.4 German credit dataset results

We next test our methodology on the German credit dataset, also a popular benchmark for binary credit-risk classification. For this dataset, each sample $u \in \mathbb{R}^d$ represents applicant-related attributes (e.g., account status, credit history, employment and housing), and the binary label $v \in \{0, 1\}$ indicates good versus bad credit. We apply the same preprocessing pipeline as in the Adult experiments (one-hot encoding for categorical variables, normalization of continuous features, and removal of examples with missing values), randomly split the data into a training set $\mathcal{D}_{\text{train}}$ and an initial test set $\mathcal{D}_{\text{test}}$, and choose logistic regression as the prediction model.



(a) PF-SMG approximation of the non-dominated set and the selected knee point.



(b) Stability under different positive-class fractions.

Figure 6: German Credit dataset results for IVO.

Similar to the previous test results, Figure 6a visualizes the PF-SMG approximation of the Pareto front in the $\big(f_{\ell b}(x), f_{ub}(x)\big)$ space. Figure 6b summarizes stability under class-imbalance shift. Using the same evaluation protocol and robustness metrics as in the Adult experiments, we compare x_{IVO} and x_{ERM} across a range of positive-class fractions. Again, $V_L(x) = V(x)$ is the variance evaluated at $x = x_{\text{ERM}}$ with $\text{loss}(\cdot) = \ell(\cdot)$, and $V_R(x) = V(x)$ is the variance evaluated at $x = x_{\text{IVO}}$ with $\text{loss}(\cdot) = \lambda \underline{\mathbb{S}}_{\underline{p}}[\ell(\cdot)] + (1-\lambda)\overline{\mathbb{S}}_{\overline{p}}[\ell(\cdot)]$. The test results indicate that

x_{IVO} exhibits smaller variability than x_{ERM} on the German credit dataset as well.

5.3.5 RVO results

In this session, we test the stochastic RVO model and compare it with the baseline ERM model. As discussed in Section 5.2, beyond the standard accuracy loss ℓ, we also incorporate a fairness loss function ℓ_F. The motivation is that optimizing accuracy alone can yield classifiers that achieve strong overall performance while still exhibiting systematic disparities across sensitive groups. Thus, introducing a fairness objective makes this trade-off explicit and allows us to search for solutions that balance predictive performance and equity. Moreover, from a robust learning perspective, fairness constraints/objectives may mitigate sensitivity to distributional shift (e.g., changes in group proportions or group-conditional feature distributions) by discouraging overly group-dependent decision rules, thereby improving model stability and robustness. In particular, we target disparate impact, approximated by a covariance-based metric that penalizes statistical dependence between the sensitive attribute and the prediction score [32]. Concretely, for model parameters x, with prediction function $\phi(u_i; x)$, we define the fairness loss ℓ_F (a covariance proxy for disparate impact) by

$$\ell_F(x) \;:=\; \left(\frac{1}{N}\sum_{i=1}^{N}\big(a_i-\bar{a}\big)\,\phi(u_i;x)\right)^2, \qquad \bar{a} := \frac{1}{N}\sum_{i=1}^{N} a_i,$$

where a_i denotes the sensitive attribute associated with u_i.

We denote by $V_F(x)$ the variance evaluated at both x_{RVO} and x_{ERM} with $\text{loss}(\cdot) = \ell_F(\cdot)$. To summarize robustness with a single scalar that emphasizes fairness stability while accounting for predictive performance, we introduce the fairness-risk-per-accuracy (FRPA) metric

$$\text{FRPA}(x) = \frac{V_F(x)}{M_{\text{ACC}}(x)},$$

where $M_{\text{ACC}}(x) = \frac{1}{N_{\text{rep}}}\sum_{i=1}^{N_{\text{rep}}} \text{Accuracy}^{(i)}(x)$, and for the r-th test replication the accuracy is defined as

$$\text{Accuracy}^{(r)}(x) = \frac{1}{N_{\text{test}}}\sum_{i=1}^{N_{\text{test}}} \mathbf{1}\Big(\phi(u_i^{(r)};x) = v_i^{(r)}\Big).$$

The numerator $V_F(x)$ measures how sensitive the fairness loss is to test resampling under distributional shift, while the denominator $M_{\text{ACC}}(x)$ measures the average predictive performance across replications. Hence, smaller values of $\text{FRPA}(x)$ indicate that the model exhibits more stable fairness behavior across replications while maintaining competitive predictive accuracy.

We conducted an experiment on the Adult dataset, taking gender as the sensitive attribute. For comparison, we also solve a standard bi-objective problem involving fairness and predictive performance, without using the subquantile/superquantile criteria, and select its corresponding Pareto knee solution (denoted by BI knee). Figure 7 reports the performance of ERM, BI knee, and the RVO knee solution under class-proportion shifts using the FRPA metric. The results show that BI knee uniformly outperforms ERM, while the RVO knee solution attains the lowest FRPA values across all tested shift levels. Thus, although the bi-objective knee already improves fairness relative to ERM by accounting for both fairness and accuracy rather than accuracy alone, the RVO knee provides the strongest robustness, producing more stable fairness behavior while preserving competitive predictive performance under distributional shift.

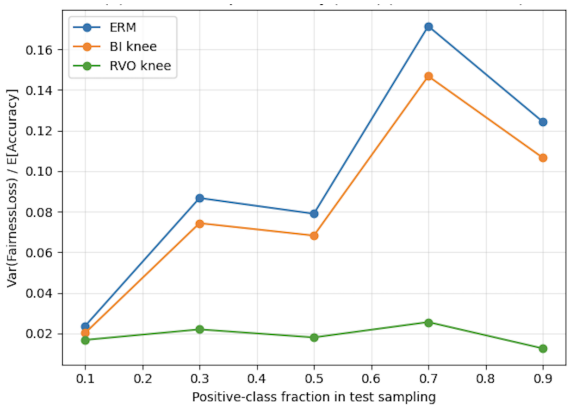


Figure 7: Adult dataset results for RVO.

6 Conclusions and future works

This work illustrates that set-valued modeling can be used as a practical tool for robust learning. Specifically, in the hyperbox setting, the set-less relation reduces the comparison of image sets to a multi-objective optimization problem with a finite number of objectives. Consequently, an effective modeling framework depends on constructing meaningful and informative lower and upper objectives.

We therefore study and explain how subquantiles and superquantiles can serve as suitable lower and upper objectives for this robust learning framework because of their natural ability to capture lower- and upper-tail behavior, thereby providing informative measures of optimistic and pessimistic performance under uncertainty.

We also contributed on the theoretical side by establishing and proving a Rockafellar–Uryasev-type characterization for subquantiles. This result complements the classical optimization representation of superquantiles, integrates naturally into our stochastic robust learning framework, and fills a theoretical gap in the analysis of lower-tail risk measures.

For the numerical results, in the distributional-shift experiments generated by changing the class proportions in the test set, the selected IVO knee solutions typically exhibit smaller sample variances across test replications than ERM, indicating greater robustness under distributional shift. In the accuracy–fairness setting, the robust RVO model provides decision makers with a way to select solutions that account for the robustness of fairness while maintaining competitive predictive accuracy, rather than improving one objective at the expense of the other.

Several extensions follow naturally from the present framework. First, beyond the IVO and RVO constructions studied here, one could incorporate additional performance criteria, such as precision, recall, or other fairness measures, together with accuracy and fairness, thereby leading to higher-dimensional hyperbox-valued models. This would make it possible to address more complex robust learning scenarios and to improve several metrics that are important to stakeholders simultaneously. Second, although we adopt an angle-based knee selection rule on the Pareto front in this work, alternative selection criteria could be developed to target stability and robustness more directly [4, 8, 17]. Third, the tail levels in the subquantile and superquantile objectives act as hyperparameters. Therefore, developing adaptive or data-driven procedures for selecting and tuning these parameters could further improve the reliability of the resulting

decisions.

Acknowledgments

This work is partially supported by the U.S. Air Force Office of Scientific Research (AFOSR) award FA9550-23-1-0217 and the U.S. Office of Naval Research (ONR) award N000142412656.

References

[1] I. Ahmad, A. Jayswal, and J. Banerjee. On interval-valued optimization problems with generalized invex functions. *J. Inequal. Appl.*, 2013:313, 2013.

[2] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Rev.*, 60:223–311, 2018.

[3] G. Bouza, E. Quintana, and C. Tammer. A steepest descent method for set optimization problems with set-valued mappings of finite cardinality. *J. Optim. Theory Appl.*, 190: 711–743, September 2021.

[4] J. Branke, K. Deb, H. Dierolf, and M. Osswald. Finding knees in multi-objective optimization. In *Parallel Problem Solving from Nature - PPSN VIII*, pages 722–731, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[5] J. V. Burke, A. S. Lewis, and M. L. Overton. A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. *SIAM J. Optim.*, 15:751–779, 2005.

[6] K. L. Chung. On a stochastic approximation method. *Ann. Math. Statist.*, 25:463 – 483, 1954.

[7] J.M. Danskin. *The Theory of Max-Min and its Application to Weapons Allocation Problems*. Econometrics and Operations Research. Springer, 1967.

[8] K. Deb and S. Gupta. Understanding knee points in bicriteria problems and their implications as preferred solution principles. *Engineering Optimization*, 43, 08 2010.

[9] J. C. Duchi and H. Namkoong. Learning models with uniform performance via distributionally robust optimization. *Ann. Statist.*, 49:1378–1406, 2021.

[10] M. Ehrgott. *Multicriteria Optimization*, volume 491. Springer Science & Business Media, Berlin, 2005.

[11] G. Eichfelder and T. Gerlach. *On Classes of Set Optimization Problems which are Reducible to Vector Optimization Problems and its Impact on Numerical Test Instances*. CRC Press, 2019.

[12] G. Eichfelder and S. Rocktäschel. Solving set-valued optimization problems using a multiobjective approach. *Opt.*, 72:789–820, 2023.

[13] G. Eichfelder, T. Gerlach, E. Quintana, and S. Rocktäschel. On two vectorization schemes for set-valued optimization. 2025.

[14] H. Evans and D. Snead. Understanding the errors made by artificial intelligence algorithms in histopathology in terms of patient impact. *npj Digital Medicine*, 7:89, 2024.

[15] J. Fliege and B. F. Svaiter. Steepest descent methods for multicriteria optimization. *Math. Methods Oper. Res.*, 51:479–494, 2000.

[16] M. A. Gianfrancesco, S. Tamang, J. Yazdany, and G. Schmajuk. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern. Med.*, 178:1544–1547, Nov 2018.

[17] T .Giovannelli, M. M. Raimundo, and L. N. Vicente. Pareto sensitivity, most-changing sub-fronts, and knee solutions. Technical Report ISE Technical Report 25T-001, Industrial and Systems Engineering, Lehigh University, Bethlehem, PA, 2025.

[18] T. Giovannelli, J. Tan, and L. N. Vicente. Non-smooth stochastic gradient descent using smoothing functions. ISE Technical Report 25T-010, Lehigh University, 2025.

[19] S. Hajian, F. Bonchi, and C. Castillo. Algorithmic bias: From discrimination discovery to fairness-aware data mining. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[20] A. H. Hamel, F. Heyde, A. Löhne, B. Rudloff, and C. Schrage. Set optimization—A rather short introduction. In *Set Optimization and Applications - The State of the Art*, pages 65–141, Berlin, Heidelberg, 2015. Springer Berlin Heidelberg.

[21] E. Hernández and R. López. Some useful set-valued maps in set optimization. *Optimization*, 66:1273–1289, 2017.

[22] J. Jahn. Vectorization in set optimization. *J. Optim. Theory Appl.*, 167:783–795, December 2015.

[23] J. Jahn. A derivative-free descent method in set optimization. *Comput. Optim. Appl.*, 60: 393–411, 2015.

[24] M. M. Kamani, S. Farhang, M. Mahdavi, and J. Z. Wang. *Targeted Data-Driven Regularization for Out-of-Distribution Generalization*, page 882–891. Association for Computing Machinery, New York, NY, USA, 2020.

[25] A. A. Khan, C. Tammer, and C. Zalinescu. *Set-valued optimization*. Springer, 2016.

[26] M. A. Khan, M. F. Uddin, and N. Gupta. Seven V's of big data understanding big data to extract value. In *Proceedings of ASEE Zone 1 2014*, pages 1–5, 2014.

[27] D. Kuroiwa. The natural criteria in set-valued optimization. In *Sūrikaisekikenkyōroku*, volume 1031, pages 85–90, 1998. Research on nonlinear analysis and convex analysis (Japanese) (Kyoto, 1997).

[28] D. Kuroiwa. On set-valued optimization. *Nonlinear Anal. Theory Methods Appl.*, 47:1395–1400, 2001. Proc. Third World Congr. Nonlinear Anal.

[29] D. Levy, Y. Carmon, J. C. Duchi, and A. Sidford. Large-scale methods for distributionally robust optimization. In *Proceedings of NeurIPS 2020*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[30] J. Z. Li. *Principled Approaches to Robust Machine Learning and Beyond*. PhD thesis, Massachusetts Institute of Technology, 2018.

[31] S. Liu and L. N. Vicente. The stochastic multi-gradient algorithm for multi-objective optimization and its application to supervised machine learning. *Ann. Oper. Res.*, pages 1–30, 2021.

[32] S. Liu and L. N. Vicente. Accuracy and fairness trade-offs in machine learning: a stochastic multi-objective approach. *Comput. Manag. Sci.*, 19:513–537, 2022.

[33] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54, jul 2021.

[34] H. Namkoong and J. C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Adv. Neural Inf. Process. Syst.*, volume 29. Curran Associates, Inc., 2016.

[35] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M. E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, I. Kompatsiaris, K. Kinder-Kurlanda, C. Wagner, F. Karimi, M. Fernandez, H. Alani, B. Berendt, T. Kruegel, C. Heinze, K. Broelemann, G. Kasneci, T. Tiropanis, and S. Staab. Bias in data-driven artificial intelligence systems—an introductory survey. *WIREs Data Mining and Knowledge Discovery*, 10:e1356, 2020.

[36] M. S. Ozdayi, M. Kantarcioglu, and R. Iyer. BiFair: Training fair models with bilevel optimization. *arXiv e-prints*, art. arXiv:2106.04757, June 2021.

[37] M. Quentin, P. Fabrice, and J. A. Désidéri. A stochastic multiple gradient descent algorithm. *European J. Oper. Res.*, 271:808 – 817, 2018.

[38] H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Statist.*, 22: 400–407, 1951.

[39] R. T. Rockafellar and J. O. Royset. Superquantiles and their applications to risk, random variables, and regression. *TutORials in Operations Research*, 2013.

[40] R. T. Rockafellar and S. Uryasev. Optimization of conditional value-at risk. *J. Risk*, 3: 21–41, 2000.

[41] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*. Springer Verlag, Heidelberg, Berlin, New York, 1998.

[42] Y. Roh, K. Lee, S. Euijong Whang, and C. Suh. FairBatch: Batch selection for model fairness. *arXiv e-prints*, art. arXiv:2012.01696, December 2020.

[43] J. Sacks. Asymptotic distribution of stochastic approximation procedures. *Ann. Math. Statist.*, 29:373 – 405, 1958.

[44] S. Sarykalin, G. Serraino, and S. Uryasev. Value-at-risk vs conditional value-at-risk in risk management and optimization. 09 2008.

[45] J. Steinhardt. *Robust Learning: Information Theory and Algorithms.* PhD thesis, Stanford University, Stanford, CA, USA, 2018.

[46] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv e-prints*, art. arXiv:1312.6199, December 2013.

A Proof of the subquantile characterization

We now give a proof of the characterization (4.8) for subquantile functions, following the notation in Section 4.1. We begin with introducing a lemma that states the symmetry relationship between the quantiles of a random variable and those of its negation: at any level $u \in (0,1)$ where the quantile function Q_Z is continuous, the $(1-u)$-quantile of $-Z$ is exactly the negative of the u-quantile of Z, that is, $Q_{-Z}(1-u) = -Q_Z(u)$.

Lemma A.1 *Let Z be a real-valued random variable with cumulative distribution function $F_Z(t) = \mathbb{P}(Z \le t)$ and quantile function $Q_Z(u) = \inf\{t \in \mathbb{R} : F_Z(t) \ge u\}$ where $u \in (0,1)$. For every $u \in (0,1)$ that is a continuity point of the quantile function Q_Z, the following identity holds*

$$Q_{-Z}(1-u) = -Q_Z(u). \tag{A.1}$$

Proof. First, we write the CDF of $-Z$ (since F_Z is nondecreasing and bounded in $[0,1]$, the left limit $F_Z(t^-) = \lim_{s \uparrow t} F_Z(s)$ exists for every t)

$$F_{-Z}(t) = \mathbb{P}(-Z \le t) = \mathbb{P}(Z \ge -t) = 1 - \mathbb{P}(Z < -t) = 1 - F_Z((-t)^-). \tag{A.2}$$

For $u \in (0,1)$, by definition of the quantile function, we have

$$Q_{-Z}(1-u) = \inf\{t : F_{-Z}(t) \ge 1-u\} = \inf\{t : F_Z((-t)^-) \le u\}. \tag{A.3}$$

Let $x = Q_Z(u) = \inf\{s : F_Z(s) \ge u\}$. Combing (A.2) and (A.3), and using the fact that $F_Z(x) \ge u$ and $F_Z(x^-) \le u$ yields

$$F_{-Z}(-x) = 1 - F_Z(x^-) \ge 1 - u, \tag{A.4}$$

which implies $Q_{-Z}(1-u) \le -x = -Q_Z(u)$ (since from (A.4), we know that $-x$ belongs to the set $\{t : F_{-Z}(t) \ge 1-u\}$).

Assume now that u is a continuity point of the monotone function Q_Z. We claim that for every $y > x$,

$$F_Z(y^-) > u. \tag{A.5}$$

Indeed, if $F_Z(y^-) \le u$ for some $y > x$, then for every $v \in (u, 1)$ we still have $F_Z(y^-) \le v$. Hence $F_Z(t) < v$ for all $t < y$, which yields $Q_Z(v) \ge y$. Letting $v \downarrow u$ gives $\lim_{v\downarrow u} Q_Z(v) \ge y > x = Q_Z(u)$, contradicting continuity at u. Thus (A.5) holds.

Now take any $t < -x$ and set $y = -t > x$. By (A.5) we have

$$F_{-Z}(t) = 1 - F_Z(y^-) < 1 - u,$$

so no $t < -x$ can satisfy $F_{-Z}(t) \ge 1 - u$. Then by using (A.3) again, one has $Q_{-Z}(1 - u) \ge -x = -Q_Z(u)$.

Combining both $Q_{-Z}(1-u) \le -x = -Q_Z(u)$ and $Q_{-Z}(1-u) \ge -x = -Q_Z(u)$ yields $Q_{-Z}(1-u) = -Q_Z(u)$ at every continuity point u of Q_Z. □

This identity (A.1) allows us to rewrite integrals involving Q_Z in terms of Q_{-Z}. In particular, it provides a direct way to express the subquantile $\underline{\mathbb{S}}_p[Z]$ in terms of the superquantile $\bar{\mathbb{S}}_{1-p}[-Z]$, which is the key step in proving the characterization below.

Theorem A.2 *Let Z be a real-valued random variable. For $p \in (0, 1)$, the subquantile of Z at level p admits the characterization in* (4.8)*, repeated here for convenience:*

$$\underline{\mathbb{S}}_p[Z] = \max_{\eta\in\mathbb{R}} \left\{ \eta - \frac{1}{p}\,\mathbb{E}\big[(\eta - Z)_+\big] \right\}. \tag{A.6}$$

Proof. By Lemma A.1, the quantile functions of Z and $-Z$ satisfy $Q_{-Z}(1-u) = -Q_Z(u)$. It is easy to see that $Q_Z(\cdot)$ is nondecreasing in u. Therefore, the set of discontinuity points of Q_Z is at most countable (hence has Lebesgue measure zero), so the identity may fail only on a null set of u, which does not affect the value of the integral. Thus, recalling the quantile representation of subquantiles (4.2) and using the relation above gives

$$\underline{\mathbb{S}}_p[Z] = -\frac{1}{p}\int_0^p Q_{-Z}(1-u)\,du = -\frac{1}{p}\int_{1-p}^1 Q_{-Z}(v)\,dv = -\bar{\mathbb{S}}_{1-p}[-Z], \tag{A.7}$$

where we use the change of variables $v = 1 - u$ and the definition of $\bar{\mathbb{S}}_{1-p}$ in (4.1).

Now apply Rockafellar–Uryasev's characterization of superquantiles (4.7) to the random variable $-Z$ at level $1 - p$, we have

$$\bar{\mathbb{S}}_{1-p}[-Z] = \min_{\alpha\in\mathbb{R}} \left\{ \alpha + \frac{1}{p}\,\mathbb{E}\big[(-Z - \alpha)_+\big] \right\}. \tag{A.8}$$

Negating both sides of (A.8) and using the relation of (A.7) yields

$$\underline{\mathbb{S}}_p[Z] = \max_{\alpha\in\mathbb{R}} \left\{ -\alpha - \frac{1}{p}\,\mathbb{E}\big[(-Z - \alpha)_+\big] \right\}.$$

Finally, by substituting $\eta := -\alpha$, we obtain

$$\underline{\mathbb{S}}_p[Z] = \max_{\eta\in\mathbb{R}} \left\{ \eta - \frac{1}{p}\,\mathbb{E}\big[(\eta - Z)_+\big] \right\},$$

which is exactly (A.6).

□

B Stochastic multi-gradient (SMG) method

In this section of the appendix, we briefly describe the stochastic multi-gradient (SMG) method, which extends the traditional single-objective stochastic gradient (SG) method to the multi-objective setting. Let us first consider a deterministic MOO problem

$$\min_{x\in X} F(x) = (f_1(x), \ldots, f_m(x)).$$

In the smooth (continuously differentiable) case, the steepest descent direction or negative multi-gradient is given by the solution of the subproblem (see [15])

$$g(x) \in \operatorname*{argmin}_{u\in\mathbb{R}^n} \max_{i\in\{1,\ldots,m\}} \nabla f_i(x)^\top u + (1/2)\|u\|^2,$$

or equivalently, by taking the dual of the subproblem, seen as the minimal-norm convex combination of the $\nabla f_i(x)$'s,

$$g(x) = \sum_{i=1}^m \lambda^i \nabla f_i(x) \quad \text{with } \lambda \text{ the solution of} \quad \min_\lambda \Big\| \sum_{i=1}^m \lambda^i \nabla f_i(x) \Big\|^2 \quad \text{s.t.} \quad \lambda \in \Delta,$$

where $\Delta = \{\lambda \in \mathbb{R}^m : \sum_{i=1}^m \lambda^i = 1, \lambda^i \geq 0, i = 1, \ldots, m\}$ is the simplex set. Given a step-size $\alpha_k > 0$, the corresponding multi-gradient descent update at each iteration k is

$$x_{k+1} = x_k - \alpha_k\, g(x_k). \tag{B.1}$$

For the stochastic setting, a stochastic MOO problem can be written as

$$\min_{x\in X} F(x) = (f_1(x), \ldots, f_m(x)) \;=\; (\mathbb{E}[h_1(x,\xi)], \ldots, \mathbb{E}[h_m(x,\xi)]), \tag{B.2}$$

where each objective function $f_i(\cdot)$ is defined as the expected value of a stochastic function $h_i(\cdot, \xi)$, which depends on a random variable ξ within a probability space with probability measure independent of x. Solving the stochastic MOO problem (B.2) at each iteration k, the stochastic multi-gradient (SMG) algorithm [31, 37] takes a step along a stochastic multi-gradient $g(x_k, \xi_k)$ (obtained by replacing the gradients in (B.1) by their stochastic estimates)

$$g(x_k, \xi_k) = \sum_{i=1}^m \lambda_k^i g_i(x_k, \xi_k) \quad \text{with } \lambda_k \text{ the solution of} \quad \min_\lambda \Big\| \sum_{i=1}^m \lambda^i g_i(x_k, \xi_k) \Big\|^2 \quad \text{s.t.} \quad \lambda \in \Delta.$$

C Smoothing stochastic gradient (SSG) method

Notice that the super- and sub-quantile objectives considered in this paper can be expressed in the compositional form

$$\max\ \{0, \mathbb{E}[\psi(x,\xi)])\}.$$

In particular, the outer mapping $t \mapsto \max\{0, t\}$ is non-smooth, which makes the resulting objective non-differentiable at the kink point even when the inner function $\psi(\cdot, \xi)$ is smooth. To handle this outer non-smoothness, we will use smoothing techniques that approximate $\phi(t) = \max\{0, t\}$ by a family of continuously differentiable smoothing functions $\tilde{\phi}(t, \mu)$ which converge

to the original function as the smoothing parameter μ goes to zero (at the cost of increasingly larger Lipschitz constants). This approximation motivated the smoothing stochastic gradient (SSG) method introduced in [18], which performs stochastic gradient steps on the smoothed objectives while driving the smoothing parameter to zero at a designated rate.

Next, we briefly present the idea of the SSG algorithm to solve the expected risk problem written in the general form

$$\min_{x\in\mathbb{R}^n} \; f(x) \;\; (f \text{ involves } \xi,\ \psi \text{ and } \phi), \tag{C.1}$$

where x is the vector of decision variable, ξ is a random vector, ψ is a differentiable function, and ϕ is a function which is not necessarily smooth. Problem (C.1) is intended to cover, as a special case, objectives such as $\max\{0, \mathbb{E}[\psi(x,\xi)]\}$ discussed above. We consider that a smoothing function $\tilde{f}(x,\mu)$ is available for $f(x)$ involving a smoothing function $\tilde{\phi}$ of ϕ and also including the randomness given by ξ. We also assume that one can draw gradient estimates $\nabla_x \tilde{f}(x,\xi,\mu)$ for the smoothing function $\tilde{f}(x,\mu)$. An SSG method can then be stated in Algorithm 1 and it only differs from the traditional stochastic gradient algorithm in the update of the smoothing parameter μ_k.

Algorithm 1 Smoothing stochastic gradient (**SSG**) method

Input: x_1, $\{\alpha_k\}$, and $\{\mu_k\}$.

For $k = 1, 2, \ldots$ **do**
  Step 1. Generate a realization ξ_k of the random variable ξ. Then compute $\nabla \tilde{f}_x(x_k, \xi_k, \mu_k)$.
  Step 2. Update iterate $x_{k+1} = x_k - \alpha_k \nabla \tilde{f}_x(x_k, \xi_k, \mu_k)$.
  Step 3. Update smoothing parameter $\mu_{k+1} \leq \mu_k$.
End do

When moving from the single-objective SSG setting to the multi-objective framework considered in this paper, the main modification to the SMG method is that each non-smooth stochastic objective $h_i(x,\xi)$ is replaced by a smoothed approximation $\tilde{h}_i(x,\xi,\mu)$ before forming the stochastic multi-gradient step. Accordingly, the expected objective $f_i(x) = \mathbb{E}[h_i(x,\xi)]$ is replaced by its smoothed counterpart $\tilde{f}_i(x,\mu) = \mathbb{E}[\tilde{h}_i(x,\xi,\mu)]$, and instead of using the original stochastic gradient estimates $g_i(x_k,\xi_k)$, one uses the smoothed stochastic gradients

$$\tilde{g}_i(x_k,\xi_k,\mu_k) = \nabla_x \tilde{h}_i(x_k,\xi_k,\mu_k), \qquad i = 1,\ldots,m.$$

The usual SMG procedure is then applied to these smoothed gradient estimates, that is, the stochastic multi-gradient is computed as

$$\tilde{g}(x_k,\xi_k,\mu_k) = \sum_{i=1}^{m} \lambda_k^i \tilde{g}_i(x_k,\xi_k,\mu_k),$$

where λ_k solves the same simplex-constrained minimum-norm problem as in the original SMG method.

D Pareto-front stochastic multi-gradient (PF-SMG) method

In the numerical tests of this paper, to compute good approximations of the entire Pareto front in a single run, we used the Pareto Front SMG (PF-SMG) algorithm, developed in [31]. The key idea of PF-SMG is to maintain and iteratively refine a list of candidate points that approximates the Pareto front. At each outer iteration, the algorithm first expands the current candidate list by adding perturbed points around each candidate point. Then, the algorithm proceeds by performing multiple short SMG runs from each candidate point and collecting the resulting endpoints. After that, it prunes the list by removing all dominated points. By repeating the procedure described above, PF-SMG yields a set of non-dominated points that approximates the Pareto front.