Distributionally robust optimization through the lens of submodularity

Karthik Natarajan* Divya Padmanabhan† Arjun Ramachandra‡

December 2023

Abstract

Distributionally robust optimization is used to solve decision making problems under adversarial uncertainty where the distribution of the uncertainty is itself ambiguous. In this paper, we identify a class of these instances that is solvable in polynomial time by viewing it through the lens of submodularity. We show that the sharpest upper bound on the expectation of the maximum of affine functions of a random vector is computable in polynomial time if each random variable is discrete with finite support and upper bounds (respectively lower bounds) on the expected values of a finite set of submodular (respectively supermodular) functions of the random vector are specified. This adds to the list of known polynomial time solvable instances of the multimarginal optimal transport problem and the generalized moment problem by bridging ideas from convexity in continuous optimization to submodularity in discrete optimization. In turn, we show that a class of distributionally robust optimization problems with discrete random variables is solvable in polynomial time using the ellipsoid method. When the submodular (respectively supermodular) functions are structured, the sharp bound is computable by solving a compact linear program. We illustrate this in two cases. The first is a multimarginal optimal transport problem where the univariate marginal distributions of the discrete random variables are given and the bivariate marginals satisfy specific positive dependence orders. We discuss an extension to incorporate higher order marginal information. Numerical experiments show that the bounds improve by 2 to 8 percent over bounds that use only univariate information. The second is a discrete moment problem where a set of marginal moments of the random variables are given along with lower bounds on the cross moments of pairs of random variables. Numerical experiments show that with higher order marginal moments, the bounds improve by 8 to 15 percent over bounds that use the first moment.

*Engineering Systems and Design, Singapore University of Technology and Design, 8 Somapah Road, Singapore 487372. Email: karthik_natarajan@sutd.edu.sg

†School of Mathematics and Computer Science, Indian Institute of Technology, Ponda-403401, Goa, India. Email: divya@iitgoa.ac.in

‡E207, Decision Sciences Area, Indian Institute of Management Bangalore-560076, India. Email: arjun.ramachandra@iimb.ac.in

1 Introduction

Consider a distributionally robust optimization problem of the form:

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}\left[f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}) := \max_{k\in[K]}\left(\boldsymbol{a}_k'(\boldsymbol{x})\tilde{\boldsymbol{\xi}} + b_k(\boldsymbol{x})\right)\right], \tag{1.1}$$

where the decision vector $\boldsymbol{x}$ is chosen from a set $\mathcal{X}$ before the true realization of the random vector $\tilde{\boldsymbol{\xi}}$ is revealed. The probability distribution of the N-dimensional random vector $\tilde{\boldsymbol{\xi}}$ is denoted by $\mathbb{P}$ and is itself ambiguous. The distribution $\mathbb{P}$ is however known to lie in a set of probability distributions denoted by $\mathcal{P}$, commonly referred to as an "ambiguity set". The cost incurred for a decision $\boldsymbol{x}$ and a realization of the random vector $\tilde{\boldsymbol{\xi}} = \boldsymbol{\xi}$ is given by $f(\boldsymbol{x},\boldsymbol{\xi}) = \max_{k\in[K]}(\boldsymbol{a}_k'(\boldsymbol{x})\boldsymbol{\xi} + b_k(\boldsymbol{x}))$ or equivalently $f(\boldsymbol{x},\boldsymbol{\xi}) = \max_{k\in[K]}(\sum_{i\in[N]} a_{i,k}(\boldsymbol{x})\xi_i + b_k(\boldsymbol{x}))$.

Throughout we assume that for each $k \in [K]$, the vector $\boldsymbol{a}_k(\boldsymbol{x})$ and the scalar $b_k(\boldsymbol{x})$ have an affine dependence on $\boldsymbol{x}$. The cost function $f(\boldsymbol{x},\boldsymbol{\xi})$ is piecewise affine and convex in $\boldsymbol{\xi}$ for a fixed $\boldsymbol{x}$ and likewise piecewise affine and convex in $\boldsymbol{x}$ for a fixed $\boldsymbol{\xi}$. In formulation (1.1), the decision $\boldsymbol{x} \in \mathcal{X}$ is selected to minimize the worst-case expected cost which is computed over all distributions $\mathbb{P} \in \mathcal{P}$ and hence is termed a "distributionally robust optimization" problem. When $\mathcal{P} = \{\mathbb{P}\}$ consists of a single distribution, (1.1) reduces to a stochastic optimization problem:

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}}\left[\max_{k\in[K]}\left(\boldsymbol{a}_k'(\boldsymbol{x})\tilde{\boldsymbol{\xi}} + b_k(\boldsymbol{x})\right)\right]. \tag{1.2}$$

When the set $\mathcal{X}$ is polyhedral and $\mathbb{P}$ is discrete ("scenario" representation of the uncertainty), it is straightforward to reformulate (1.2) as a linear program in size that grows linearly in the number of scenarios. The number of scenarios or joint realizations of the random variables can however grow exponentially in the dimension N; for example when the random variables are mutually independent. In fact, computing the expected cost in (1.2) for a fixed $\boldsymbol{x}$ is know to be #P-hard when the random variables are independent, both with continuous (see [31, 42]) and discrete (see [27]) random variables. In turn, the stochastic optimization problem (1.2) is #P-hard to solve for independent random variables. A popular solution methodology is to approximate the expected cost with a sample average and to optimize the sample average approximation (see [76]). On the other hand, with $\mathcal{P} = \mathcal{P}(\Xi)$ where $\mathcal{P}(\Xi)$ is the set of all probability distributions with support contained in a closed bounded set Ξ, (1.1) reduces to a robust optimization problem:

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{\xi}\in\Xi} \max_{k\in[K]} \left(\boldsymbol{a}_k'(\boldsymbol{x})\boldsymbol{\xi} + b_k(\boldsymbol{x})\right). \tag{1.3}$$

When both $\mathcal{X}$ and Ξ are polyhedral, (1.3) is solvable using linear optimization (see [8, 9]). Distributionally robust optimization lies between these two extremes.

This brings us to the main contributions of the current paper:

(a) We study a general ambiguity set $\mathcal{P}$ where the support of each random variable is specified along with upper bounds (respectively lower bounds) on the expected values of a finite set of submodular

(respectively supermodular) functions of the random vector. When the support of each random variable is discrete and finite, the worst-case expected cost in (1.1) is shown to be computable in polynomial time (Theorem 3.1 in the paper). This result forms the discrete counterpart of an existing result on polynomial time computability of the sharp bound when the support of the random vector lies in a convex set and upper bounds (respectively lower bounds) on the expected values of a finite set of convex (respectively concave) functions of the random vector are given (see Theorem 1.5 in [52]). In turn this helps us identify a class of polynomial time solvable distributionally robust optimization problems where the uncertainty is discrete by viewing such problems through the lens of submodularity. The result adds to the list of known polynomial time solvable instances of the multimarginal optimal transport problem and the generalized moment problem. The proof makes use of ideas from duality, submodular function minimization and the ellipsoid method.

(b) When the upper bounds (respectively lower bounds) are specified on the expected values of structured submodular (respectively supermodular) functions, it is possible to develop compact reformulations. As a first example, we consider the setting where univariate discrete marginal distributions with finite support are specified in $\mathcal{P}$ and the bivariate marginal distributions satisfy specific positive dependence orders. Computing the worst-case expected cost is then an instance of the multimarginal optimal transport problem where additional dependence information on bivariate marginals is provided. We develop a compact linear program to compute the sharp bound. We discuss an extension by incorporating dependence information on higher order marginals.

(c) As a second example, we assume that a set of marginal moments of random variables with discrete finite support are specified in $\mathcal{P}$ along with lower bounds on the cross moments of pairs of random variables. This is an instance of the moment problem. We also develop a compact linear program to compute the sharp bound in this case. The compact linear programs developed in both these examples have a natural probabilistic interpretation where the extremal distributions are constructed through a mixture of comonotonic random vectors.

The structure of the paper is as follows: We discuss the notations used in Section 1.1. In Section 2, we review prior work in distributionally robust optimization that is most related to this paper and key ideas from submodularity and comonotonicity. In Section 3, we identify a general ambiguity set where the sharp bound is polynomial time computable. We provide formulations for the multimarginal optimal transport problem using positive dependence orders in Section 4. We provide formulations for the moment problem in Section 5. Numerical experiments are reported in Section 6 before concluding in Section 7. All proofs not provided in the main paper are provided in the Appendix.

1.1 Notations

We use nonbold symbols (such as x, ξ) to denote scalars, bold symbols (such as $\boldsymbol{x}, \boldsymbol{\xi}$) to denote vectors and bold capital symbols (such as $\boldsymbol{X}, \boldsymbol{\Sigma}$) to denote matrices. Random numbers and random vectors are denoted with the tilde sign (examples are $\tilde{\xi}$ and $\tilde{\boldsymbol{\xi}}$). For a positive integer N, we use $[N]$ to

denote the set $\{1, 2, \ldots, N\}$ and $[0 \cup N]$ to denote the set $\{0, 1, 2, \ldots, N\}$. The cardinality of a set Ξ is denoted by $|\Xi|$ (possibly infinite). Given sets $\Xi_1, \Xi_2, \ldots, \Xi_N$, the Cartesian product set is given by $\prod_{i\in[N]} \Xi_i = \{(\xi_1, \xi_2, \ldots, \xi_N) | \xi_1 \in \Xi_1, \xi_2 \in \Xi_2, \ldots, \xi_N \in \Xi_N\}$. The indicator function of membership in a set Ξ is denoted by $\mathbb{1}_{\boldsymbol{\xi}\in\Xi}$ which takes a value of 1 if $\boldsymbol{\xi} \in \Xi$ and 0 if $\boldsymbol{\xi} \notin \Xi$. Let $\mathbb{R}^N$ and $\mathbb{Z}^N$ denote the sets of N-dimensional vectors with real entries and integer entries. Let $\mathbb{R}^N_+$ and $\mathbb{Z}^N_+$ denote the sets of N-dimensional vectors with nonnegative real entries and nonnegative integer entries. For any vector $\boldsymbol{x}$ (a column vector by default), we use $\boldsymbol{x}'$ to denote its transpose. The dot product of vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^N$ is given by $\boldsymbol{x}'\boldsymbol{y}$. We denote a vector of zeros by $\mathbf{0}$, a vector of ones by $\boldsymbol{e}$ and a vector with 1 in the ith position and 0 otherwise by $\boldsymbol{e}_i$. Given vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^N$, we write $\boldsymbol{x} = \boldsymbol{y}$ if $x_i = y_i$ for all $i \in [N]$, $\boldsymbol{x} \geq \boldsymbol{y}$ if $x_i \geq y_i$ for all $i \in [N]$, and $\boldsymbol{x} \leq \boldsymbol{y}$ if $x_i \leq y_i$ for all $i \in [N]$. Let $\mathcal{S}_N$ denote the set of $N \times N$ real symmetric matrices. Given a matrix $\boldsymbol{X} \in \mathcal{S}_N$, we use $\boldsymbol{X} \succeq 0$ (respectively $\boldsymbol{X} \succ 0$) to denote the matrix is positive semidefinite (respectively positive definite). The Frobenius inner product of matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ in $\mathcal{S}_N$ is given by $\boldsymbol{X} \cdot \boldsymbol{Y}$. Given matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ in $\mathcal{S}_N$, we write $\boldsymbol{X} \succeq \boldsymbol{Y}$ (respectively $\boldsymbol{X} \succ \boldsymbol{Y}$) if $\boldsymbol{X} - \boldsymbol{Y} \succeq 0$ (respectively $\boldsymbol{X} - \boldsymbol{Y} \succ 0$) and $\boldsymbol{X} \preceq \boldsymbol{Y}$ (respectively $\boldsymbol{X} \prec \boldsymbol{Y}$) if $\boldsymbol{Y} - \boldsymbol{X} \succeq 0$ (respectively $\boldsymbol{Y} - \boldsymbol{X} \succ 0$). Associated with a random vector $\tilde{\boldsymbol{\xi}}$ is a probability distribution $\mathbb{P}$ which we denote by $\tilde{\boldsymbol{\xi}} \sim \mathbb{P}$. We use $\mathbb{P}(\cdot)$ to denote the probability of an event and $\mathbb{E}_{\mathbb{P}}[\cdot]$ to denote the expectation with respect to $\mathbb{P}$. We denote the support of $\mathbb{P}$ by $\text{supp}(\mathbb{P})$. For a discrete random vector, the support is the set of realizations with strictly positive probabilities. More generally, the support is the smallest closed set with probability of the random vector lying in the set equal to one. The projection of a N-dimensional random vector $\tilde{\boldsymbol{\xi}} \sim \mathbb{P}$ on the set $I \subseteq [N]$ is given by $\tilde{\boldsymbol{\xi}}_I \sim \text{proj}_I(\mathbb{P})$. We use $\mathcal{P}(\Xi)$ to denote the set of all probability distributions with support contained in the set Ξ.

2 Background and literature review

2.1 Polynomial time solvable distributionally robust optimization

Over the past two decades, the distributionally robust optimization problem (1.1) has been extensively studied and polynomial time solvability has been shown for several ambiguity sets $\mathcal{P}$. We discuss some of these sets and the results associated with them next:

(a) *Marginal distribution ambiguity set*: When the marginal distribution $\mathbb{P}_i$ of the random variable $\tilde{\xi}_i$ is specified for each $i \in [N]$ where $\text{supp}(\mathbb{P}_i) = \Xi_i$, but the dependence structure among the random variables is unspecified, the ambiguity set is referred to as the "marginal distribution ambiguity set" or the "Fréchet set of distributions". When $\mathcal{X}$ is a polyhedron, problem (1.1) is solvable using a polynomial sized linear program when the marginals are discrete with finite support and using a convex program when the marginals are continuous (see [17]). The joint distribution with independent marginals is a feasible distribution in this ambiguity set. However unlike with the stochastic optimization problem (1.2), which is #P-hard to solve with independent random variables, the distributionally robust optimization problem (1.1) is solvable in polynomial time for this ambiguity set. Computing the worst-case expected cost in the marginal distribution ambiguity set

corresponds to solving a "multimarginal optimal transport problem" for which tractable instances have also been identified for other types of cost functions (see [68, 2]). This ambiguity set has been extended to allow for the specification of multivariate marginals of fixed subsets of the random variables. However verifying if the ambiguity set is nonempty is known to be NP-complete even if we restrict attention to Bernoulli random vectors where the distribution of pairs of Bernoulli random variables are specified (see [44, 38]). Efficient solvability of (1.1) requires further assumptions on the structure of the multivariate marginals. One such instance is when the random variables are partitioned into nonoverlapping subsets of small size with fixed multivariate marginal distributions of the subsets while allowing for arbitrary dependence among the random variables in different subsets (see [28]). When the subsets overlap, additional graph theoretic assumptions are required on the structure of the overlapping multivariate marginals to guarantee polynomial time solvability (see [29]). Such ambiguity sets have been particularly popular in the risk management community (see [73]). Most of these formulations have been extended to the setting where only limited information on the marginal distributions is available such as a few marginal moments or dispersion measures such as the mean absolute deviation or directional deviation (see [12, 13, 71, 18]).

(b) *Moment ambiguity set*: Assume the first moment vector $\mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{\xi}}] = \boldsymbol{\mu}$ and the second moment matrix $\mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{\xi}}\tilde{\boldsymbol{\xi}}'] = \boldsymbol{Q}$ are specified in the ambiguity set where the moments satisfy the feasibility condition $\boldsymbol{Q} \succeq \boldsymbol{\mu}\boldsymbol{\mu}'$. Given these moments and with the support of $\mathbb{P}$ contained in $\Xi = \mathbb{R}^N$ or in an ellipsoid, (1.1) is solvable in polynomial time with semidefinite optimization when $\mathcal{X}$ is polyhedral (see [11, 84, 39]). However given the first two moments and with support contained in a polyhedron, (1.1) is NP-hard to solve (see [14]). A related tractable moment ambiguity set was proposed in [23] where Ξ is a closed convex set with an efficient separation oracle, the first moment vector $\mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{\xi}}] = \boldsymbol{\mu}$ is specified or assumed to lie in an ellipsoid and the second moment matrix satisfies the condition $\mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{\xi}}\tilde{\boldsymbol{\xi}}'] \preceq \boldsymbol{Q}$ where $\boldsymbol{Q}$ is a fixed positive semidefinite matrix. For this ambiguity set, (1.1) is solvable in polynomial time (see [23]). Polynomial sized reformulations for these moment ambiguity sets are based on conic optimization that include semidefinite, second order cone and linear optimization as special cases. Polynomial sized conic programs for (1.1) have been developed when conic representable confidence sets for the support of the random vector satisfying a certain technical condition are specified in the ambiguity set and the first moment vector satisfies affine constraints (see [81]). Another tractable ambiguity set is the scenario wise moment ambiguity set (see [21]). Computing the worst-case expected cost for all these moment ambiguity sets requires solving a generalized moment problem (see [52]). Polynomial time computability of the sharp moment bound is known for a general ambiguity set $\mathcal{P}$ where Ξ is convex set and upper bounds (respectively lower bounds) on the expected values of a finite set of convex (respectively concave) functions of the random vector are given (see [70, 52]). We will revisit this in Section 3 since it is closely related to our work.

Problem (1.1) is also computationally tractable for statistical distance based ambiguity sets such as the phi-divergence ambiguity set (see [7, 6]) and the Wasserstein distance ambiguity set (see [34, 16, 37]). These ambiguity sets are characterized by distributions that lie within a certain statistical distance from

a reference distribution. The size of the convex reformulations for these ambiguity sets grow linearly with the number of support points of the reference distribution. The results in this paper are more closely related to existing results for the moment and marginal distribution ambiguity sets than the statistical distance based ambiguity sets.

2.2 Submodularity and comonotonicity

Consider a function $f : \prod_{i\in[N]} \Xi_i \to \mathbb{R}$ that maps a vector $\boldsymbol{\xi} \in \prod_{i\in[N]} \Xi_i$ to a number $f(\boldsymbol{\xi}) \in \mathbb{R}$. Given vectors $\boldsymbol{\xi}$ and $\boldsymbol{\chi}$ in $\prod_{i\in[N]} \Xi_i$, the meet vector is given by $\boldsymbol{\xi} \wedge \boldsymbol{\chi} = (\min(\xi_1, \chi_1), \ldots, \min(\xi_N, \chi_N))$ (componentwise minimum) and the join vector is given by $\boldsymbol{\xi} \vee \boldsymbol{\chi} = (\max(\xi_1, \chi_1), \ldots, \max(\xi_N, \chi_N))$ (componentwise maximum). Clearly, the meet vector and the join vector also lie in the set $\prod_{i\in[N]} \Xi_i$. We refer to such sets as lattices. Examples of such sets include $\Xi = \{0, 1\}^N$ (extreme points of the unit hypercube), $\Xi = [0, 1]^N$ (unit hypercube), $\Xi = [0 \cup B]^N$ (bounded integer lattice where B is a positive integer), $\Xi = [0, B]^N$ (hypercube of length B), $\Xi = \mathbb{Z}^N$ (integer lattice) and $\Xi = \mathbb{R}^N$ (Euclidean space). A function f is submodular if:

$$f(\boldsymbol{\xi}) + f(\boldsymbol{\chi}) \geq f(\boldsymbol{\xi} \wedge \boldsymbol{\chi}) + f(\boldsymbol{\xi} \vee \boldsymbol{\chi}), \forall \boldsymbol{\xi}, \boldsymbol{\chi} \in \prod_{i\in[N]} \Xi_i. \tag{2.1}$$

A function f is supermodular if $-f$ is submodular. When $\Xi = \{0, 1\}^N$, associate with each subset $S \subseteq [N]$, a realization of the N-dimensional binary vector $\boldsymbol{\xi} \in \Xi$ where $\xi_i = \mathbb{1}_{i\in S}$ for $i \in [N]$ and set $f(S) = f(\boldsymbol{\xi})$. The definition of submodularity in the set function case is then equivalent to:

$$f(S) + f(T) \geq f(S \cap T) + f(S \cap T), \forall S, T \subseteq N. \tag{2.2}$$

When $\Xi = \mathbb{R}^N$ and the function f is differentiable, the definition of submodularity is equivalent to:

$$\frac{\partial}{\partial \xi_i} f(\boldsymbol{\xi}) \geq \frac{\partial}{\partial \chi_i} f(\boldsymbol{\chi}), \forall \boldsymbol{\xi} \leq \boldsymbol{\chi}, \forall i \in [N] : \xi_i = \chi_i,$$

and when the function f is twice differentiable, the definition of submodularity is equivalent to:

$$\frac{\partial^2}{\partial \xi_i \partial \xi_j} f(\boldsymbol{\xi}) \leq 0, \forall \boldsymbol{\xi} \in \mathbb{R}^N, \forall i \neq j.$$

Three examples of submodular functions are:

(i) $f(\boldsymbol{\xi}) = h(\boldsymbol{a}'\boldsymbol{\xi})$ where $\boldsymbol{a} \geq \mathbf{0}$ and $h : \mathbb{R} \to \mathbb{R}$ is a concave function,

(ii) $f(\boldsymbol{\xi}) = \max(\xi_1, \ldots, \xi_N)$,

(iii) $f(\boldsymbol{\xi}) = -\prod_{i\in[N]} \xi_i$ where $\boldsymbol{\xi} \geq \mathbf{0}$.

In the set function case, these reduce to $f(S) = h(\sum_{i\in S} a_i)$, $f(S) = \mathbb{1}_{|S|\geq 1}$ and $f(S) = -\mathbb{1}_{|S|=N}$. As these examples show, submodular functions might be convex functions (example (ii)), concave functions (example (i)), neither convex or concave (example (iii)) or both convex and concave (the linear function $f(\boldsymbol{\xi}) = \boldsymbol{a}'\boldsymbol{\xi}$ and the modular set function $f(S) = \sum_{i\in S} a_i$). Other examples of submodular functions include the weighted cut function in a directed graph, the entropy of a random vector, the influence

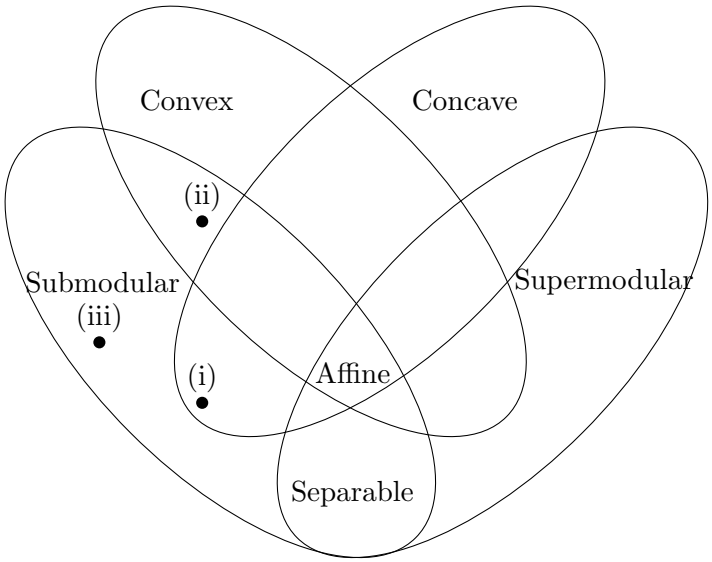


Figure 1: Venn diagram illustrating the set of convex, concave, submodular and supermodular functions defined over a continuous lattice. Affine functions lie at the intersection of all four sets.

function in a social network and the rank function of a matroid (see [57, 80, 5, 15] for examples of submodular and supermodular functions arising in graph theory, probability, operations research, game theory, machine learning and artificial intelligence). While submodularity is known to be preserved under certain operations such as taking a nonnegative weighted sum of submodular functions or the partial minimum of a submodular function, it is not preserved under operations such as taking the pointwise maximum or minimum of submodular functions. Univariate functions are both submodular and supermodular and hence all (additively) separable functions are both submodular and supermodular (see Figure 1 for an illustration).

Submodular functions defined over discrete domains behave somewhat similarly to convex functions defined over continuous domains, particularly in terms of the minimization of such functions (see [57]). Consider the submodular function minimization problem:

$$\inf_{\boldsymbol{\xi} \in \prod_{i \in [N]} \Xi_i} f(\boldsymbol{\xi}). \tag{2.3}$$

Assume the submodular function is given by an polynomial time evaluation oracle, namely given a $\boldsymbol{\xi} \in \prod_{i \in [N]} \Xi_i$, the oracle returns $f(\boldsymbol{\xi})$ in polynomial time[1]. A key result proved in [40, 41], building on the work of [32, 50], showed that (2.3) is solvable in time polynomial in the input size using the ellipsoid method when the sets Ξ_i are discrete and finite[2]. Since then a variety of algorithms have been

[1]It is common to assume the function is rational valued and the oracle returns it exactly in polynomial time.

[2]While many of the polynomial time algorithms for submodular function minimization in the literature are presented for $\Xi_i = \{0, 1\}$, these results extend to discrete finite sets Ξ_i by a transformation to a lattice or a ring family (see Section 49.3 in [74] or Section 4.4 in [4]) for details.

developed to solve the submodular function minimization problem ranging from convex optimization based methods (see [57, 41, 36, 5, 4, 54]) to combinatorial algorithms (see [46, 74, 67]). When the domain is discrete and finite (for example $\Xi = \{0,1\}^N$ or $\Xi = [0 \cup B]^N$), (2.3) is solvable in time polynomial in N (number of variables), $\max_{i\in[N]} |\Xi_i|$ (maximum number of values that a single variable can take) and the evaluation time of the oracle. The current state of art strongly polynomial time algorithm for $\Xi = \{0,1\}^N$ requires $O(N^3 \log\log(N)/\log(N))$ calls to the evaluation oracle (see [47]). When the domain is the hypercube $\Xi = [0,B]^N$ and the submodular function is L-Lipschitz continuous, an ϵ-additive approximation to the optimal value in (2.3) can be found in time polynomial in N, B, L, $1/\epsilon$ and the evaluation time of the oracle by solving a discretized version of the problem (see [4, 3]).

A key object that aids in the efficient minimization of submodular functions is the Lovász extension [57] (also known as the Choquet integral [22]) which is defined through the construction of a comonotonic random vector. Let $\mathcal{P}_F(\mathbb{P}_1,\ldots,\mathbb{P}_N)$ denote the Fréchet set of distributions where $\tilde{\xi}_i \sim \mathbb{P}_i$ with $\text{supp}(\mathbb{P}_i) \subseteq \Xi_i$ for each $i \in [N]$. Let $F_i(\cdot)$ denote the cumulative distribution function of $\tilde{\xi}_i$ for each $i \in [N]$. A comonotonic random vector has maximal positive dependence in the Fréchet set of distributions and is given by:

$$\tilde{\boldsymbol{\xi}}^c := (F_1^{-1}(\tilde{U}),\ldots,F_N^{-1}(\tilde{U})),$$

where $\tilde{U}$ is a uniform random variable on $[0,1]$ and $F_i^{-1}(\cdot)$ is the generalized inverse distribution function of the cumulative distribution function $F_i(\cdot)$. Let $\mathbb{P}^c$ denote the distribution of the comonotonic random vector. Clearly $\mathbb{P}^c \in \mathcal{P}_F(\mathbb{P}_1,\ldots,\mathbb{P}_N)$. The support of the comonotonic random vector is contained in completely ordered subsets of $\mathbb{R}^N$ with:

$$\mathbb{P}^c(\tilde{\boldsymbol{\xi}}^c > \boldsymbol{t}) = \min_{i\in[N]} \mathbb{P}_i(\tilde{\xi}_i > t_i) \text{ and } \mathbb{P}^c(\tilde{\boldsymbol{\xi}}^c \leq \boldsymbol{t}) = \min_{i\in[N]} \mathbb{P}_i(\tilde{\xi}_i \leq t_i), \forall \boldsymbol{t} \in \mathbb{R}^N. \tag{2.4}$$

The expected value of a function of the comonotonic random vector is computed as:

$$\mathbb{E}_{\mathbb{P}^c}\left[f(\tilde{\boldsymbol{\xi}}^c)\right] = \int_0^1 f\left(F_1^{-1}(t),\ldots,F_N^{-1}(t)\right) dt. \tag{2.5}$$

With discrete and finite Ξ_i, the cardinality of the support of the comonotonic random vector is at most $\sum_{i\in[N]} |\Xi_i|$ and the value in (2.5) is computable using a polynomial number of calls to the evaluation oracle. A well known extremal characterization of the comonotonic random vector (see [79, 73]) for general marginals $\mathbb{P}_1,\ldots,\mathbb{P}_N$ is given by:

$$\inf_{\mathbb{P}\in\mathcal{P}_F(\mathbb{P}_1,\ldots,\mathbb{P}_N)} \mathbb{E}_{\mathbb{P}}\left[f(\tilde{\boldsymbol{\xi}})\right] = \mathbb{E}_{\mathbb{P}^c}\left[f(\tilde{\boldsymbol{\xi}}^c)\right], \forall \text{submodular } f \text{ such that expectations exists}, \tag{2.6}$$

or equivalently:

$$\sup_{\mathbb{P}\in\mathcal{P}_F(\mathbb{P}_1,\ldots,\mathbb{P}_N)} \mathbb{E}_{\mathbb{P}}\left[f(\tilde{\boldsymbol{\xi}})\right] = \mathbb{E}_{\mathbb{P}^c}\left[f(\tilde{\boldsymbol{\xi}}^c)\right], \forall \text{supermodular } f \text{ such that expectations exists}. \tag{2.7}$$

For example, by considering the supermodular functions $f(\boldsymbol{\xi}) = 1_{\boldsymbol{\xi}>\boldsymbol{t}}$ and $f(\boldsymbol{\xi}) = 1_{\boldsymbol{\xi}\leq\boldsymbol{t}}$ for each $\boldsymbol{t}$, one

obtains the well known upper Fréchet bounds:

$$\sup_{\mathbb{P}\in\mathcal{P}_F(\mathbb{P}_1,\ldots,\mathbb{P}_N)} \mathbb{P}(\tilde{\boldsymbol{\xi}} > \boldsymbol{t}) = \min_{i\in[N]} \mathbb{P}_i(\tilde{\xi}_i > t_i) \text{ and } \sup_{\mathbb{P}\in\mathcal{P}_F(\mathbb{P}_1,\ldots,\mathbb{P}_N)} \mathbb{P}(\tilde{\boldsymbol{\xi}} \leq \boldsymbol{t}) = \min_{i\in[N]} \mathbb{P}_i(\tilde{\xi}_i \leq t_i), \forall \boldsymbol{t} \in \mathbb{R}^N.$$

The extremal characterization in (2.6)-(2.7) has been proved to be very useful in developing efficient algorithms for submodular function minimization. Specifically, the function $f : \prod_{i\in[N]} \Xi_i \to \mathbb{R}$ is submodular if and only if the functional $\inf_{\mathbb{P}\in\mathcal{P}_F(\mathbb{P}_1,\ldots,\mathbb{P}_N)} \mathbb{E}_{\mathbb{P}}[f(\tilde{\boldsymbol{\xi}})] : \mathbb{P}_1,\ldots,\mathbb{P}_N \to \mathbb{R}$ is convex (see [57, 4]). The domain of the functional is specified by the marginal probability measures $\mathbb{P}_1,\ldots,\mathbb{P}_N$ where $\text{supp}(\mathbb{P}_1) \subseteq \Xi_1,\ldots,\text{supp}(\mathbb{P}_N) \subseteq \Xi_N$. The corresponding functional value is exactly the Lovász extension or the Choquet integral. This lets one transform the submodular function minimization problem in (2.3) to a convex minimization problem over probability measures as follows:

$$\inf_{\boldsymbol{\xi}\in\prod_{i\in[N]}\Xi_i} f(\boldsymbol{\xi}) \overset{(a)}{=} \inf_{\text{supp}(\mathbb{P}_i)\subseteq\Xi_i,\forall i\in[N]} \inf_{\mathbb{P}\in\mathcal{P}_F(\mathbb{P}_1,\ldots,\mathbb{P}_N)} \mathbb{E}_{\mathbb{P}}\left[f(\tilde{\boldsymbol{\xi}})\right] \overset{(b)}{=} \inf_{\text{supp}(\mathbb{P}_i)\subseteq\Xi_i,\forall i\in[N]} \mathbb{E}_{\mathbb{P}^c}\left[f(\tilde{\boldsymbol{\xi}}^c)\right]. \tag{2.8}$$

The first equality holds since the optimal solution on the right hand side of (a) will be attained at a Dirac measure where the functional value and the function value coincide. Equality (b) comes from (2.6). When the domain is discrete and finite, the convex minimization problem on the right hand side of (b) is solvable in polynomial time using the ellipsoid method where both the function value and its subgradient are computable in polynomial time (see [57, 4]). Comonotonic random vectors have also been identified as the worst-case distributions for certain distributionally robust optimization problems which enable them to be solved efficiently (see [17, 56, 8]). Applications of comonotonicity are also found in risk management (see [25, 26]), appointment scheduling (see [59]), inventory pooling (see [60]) and dynamic robust optimization (see [45]). Distributionally robust optimization has also been explored with submodular and supermodular objective functions in [1] where the worst-case expected cost computed over all distributions with fixed marginals is compared to the expected cost with an independent distribution (see [78] for related work). In addition with submodular objective functions in the context of influence maximization, [85] show that worst-case expectation (minimization) over the Fréchet set of distributions preserves submodularity.

3 A polynomial time computable sharp bound

In this section, we discuss a general ambiguity set for which the worst-case expected cost in (1.1) for a fixed $\boldsymbol{x} \in \mathcal{X}$ is efficiently computable. Towards this, we focus on the problem of computing the sharpest upper bound on the expectation of the maximum of affine functions of a N-dimensional random vector $\tilde{\boldsymbol{\xi}}$. Consider the bound:

$$f^* = \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}\left[f(\tilde{\boldsymbol{\xi}}) := \max_{k\in[K]} \left(\boldsymbol{a}_k'\tilde{\boldsymbol{\xi}} + b_k\right)\right], \tag{3.1}$$

where $\{\boldsymbol{a}_1,\ldots,\boldsymbol{a}_K\} \subset \mathbb{R}^N$ is a given set of vectors and $\{b_1,\ldots,b_K\} \subset \mathbb{R}$ is a given set of scalars. Given a set $\Xi \subseteq \mathbb{R}^N$, a set of functions $f_j : \Xi \to \mathbb{R}$ for $j \in [J]$ and a set of scalars γ_j for $j \in [J]$, the ambiguity

set for the random vector $\tilde{\boldsymbol{\xi}}$ is defined as follows:

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\Xi) \mid \mathbb{E}_{\mathbb{P}}[f_j(\tilde{\boldsymbol{\xi}})] \leq \gamma_j, \forall j \in [J]\}, \tag{3.2}$$

where all the expectations are assumed to be well-defined with respect to the distributions in the set $\mathcal{P}$. We make two important assumptions on the ambiguity set $\mathcal{P}$.

`Assumption (A1):` $\Xi = \prod_{i\in[N]} \Xi_i$ where $\Xi_i \subset \mathbb{R}$ is a discrete finite set for each $i \in [N]$.

`Assumption (A2):` For each $j \in [J]$, $f_j : \prod_{i\in[N]} \Xi_i \to \mathbb{R}$ is a submodular function with a polynomial time evaluation oracle.

This brings us to the first theorem of the paper.

Theorem 3.1. *Suppose the ambiguity set $\mathcal{P}$ in* (3.2) *satisfies assumptions (A1)-(A2). Then f^* in* (3.1) *is computable in polynomial time.*

Proof. Under assumption (A1), we can reformulate (3.1) as a linear program:

$$\begin{aligned} f^* = \max \quad & \sum_{\boldsymbol{\xi}\in\Xi} \max_{k\in[K]} \left(\boldsymbol{a}_k'\boldsymbol{\xi} + b_k\right) p(\boldsymbol{\xi}) \\ \text{s.t.} \quad & \sum_{\boldsymbol{\xi}\in\Xi} f_j(\boldsymbol{\xi}) p(\boldsymbol{\xi}) \leq \gamma_j, && \forall j \in [J], \\ & \sum_{\boldsymbol{\xi}\in\Xi} p(\boldsymbol{\xi}) = 1, \\ & p(\boldsymbol{\xi}) \geq 0, && \forall \boldsymbol{\xi} \in \Xi, \end{aligned} \tag{3.3}$$

where the decision variables are $p(\boldsymbol{\xi}) = \mathbb{P}(\tilde{\boldsymbol{\xi}} = \boldsymbol{\xi})$ for $\boldsymbol{\xi} \in \Xi = \prod_{i\in[N]} \Xi_i$. We call this the primal linear program. It has a polynomial number of constraints (excluding the nonnegativity of the variables) and an exponential number of variables. The dual linear program is formulated as:

$$\begin{aligned} f_d^* = \min \quad & y_0 + \sum_{j\in[J]} y_j\gamma_j \\ \text{s.t.} \quad & y_0 + \sum_{j\in[J]} y_j f_j(\boldsymbol{\xi}) \geq \max_{k\in[K]} \left(\boldsymbol{a}_k'\boldsymbol{\xi} + b_k\right), && \forall \boldsymbol{\xi} \in \Xi, \\ & y_j \geq 0, && \forall j \in [J], \end{aligned} \tag{3.4}$$

where the decision variables are y_0 and y_j for $j \in [J]$. The dual linear program has a polynomial number of variables and an exponential number of constraints. The dual linear program is feasible (set $y_j = 0$ for $j \in [J]$ and $y_0 = \max_{\boldsymbol{\xi}\in\Xi} \max_{k\in[K]} (\boldsymbol{a}_k'\boldsymbol{\xi} + b_k)$). Strong duality of linear programming guarantees that $f^* = f_d^*$. The separation problem for the dual linear program is given by:

Given numbers y_0 and $y_j \geq 0$ for $j \in [J]$, decide whether

$$y_0 + \sum_{j\in[J]} y_j f_j(\boldsymbol{\xi}) \geq \max_{k\in[K]} \left(\boldsymbol{a}_k'\boldsymbol{\xi} + b_k\right), \forall \boldsymbol{\xi} \in \Xi,$$

and if the answer is no, return a violated inequality.

This reduces to checking if

$$y_0 + \sum_{j\in[J]} y_j f_j(\boldsymbol{\xi}) \geq \boldsymbol{a}_k'\boldsymbol{\xi} + b_k, \forall \boldsymbol{\xi} \in \Xi, \forall k \in [K],$$

which in turn reduces to checking if

$$y_0 - b_k + \min_{\boldsymbol{\xi}\in\Xi} \left(\sum_{j\in[J]} y_j f_j(\boldsymbol{\xi}) - \boldsymbol{a}_k'\boldsymbol{\xi} \right) \geq 0, \forall k \in [K].$$

Since $y_j \geq 0$ for all $j \in [J]$, f_j is a submodular function for all $j \in [J]$ (assumption (A2)) and $\boldsymbol{a}_k'\boldsymbol{\xi}$ is a linear function for all $k \in [K]$, the function $\sum_{j\in[J]} y_j f_j(\boldsymbol{\xi}) - \boldsymbol{a}_k'\boldsymbol{\xi}$ is submodular for each $k \in [K]$. Hence we need to solve a set of K submodular function minimization problems of the form in (2.3) to solve the separation problem. Since each of the problems is polynomial time solvable, the dual separation problem is solvable in polynomial time. From the equivalence of separation and optimization (see [40]), the dual linear program is solvable in polynomial time. Hence f^* is computable in polynomial time. □

We make several remarks about Theorem 3.1 next and connections to existing work.

(a) The formulation in (3.1)-(3.2) is an instance of the generalized moment problem (see [52]). Consider the following assumptions on the ambiguity set $\mathcal{P}$:
`Assumption (A1'):` Ξ is a closed, bounded, convex set with a polynomial time separation oracle.
`Assumption (A2'):` For each $j \in [J]$, $f_j : \Xi \to \mathbb{R}$ is a convex function with a polynomial time subgradient oracle that returns the function value and its subgradient efficiently.
Under assumptions (A1')-(A2'), the bound f^* for the generalized moment problem is known to be computable in polynomial time (see Theorem 1.5 in [52], Chapter 3 in [70] and Proposition 1 in [23]). Theorem 3.1 provides the discrete counterpart of this result for the generalized moment problem where submodularity is the natural analog of convexity. To the best of our knowledge, while prior work has developed numerical methods to solve univariate discrete moment problems (see [72, 19]) and multivariate discrete moment problems (see [58]), the tractability result for the general ambiguity set in (3.2) under assumptions (A1)-(A2) is new. This in turn implies that a new class of distributionally robust optimization problems with discrete uncertainty is solvable in polynomial time (see Theorem 3.2). In addition, the bound can be extended beyond maximum of affine functions to the maximum of supermodular functions (see Theorem 3.3).

Theorem 3.2. *Suppose $\mathcal{X}$ is a compact convex set with an efficient separation oracle and the ambiguity set $\mathcal{P}$ in* (3.2) *satisfies assumptions (A1)-(A2). Then the distributionally robust optimization in* (1.1) *is solvable in polynomial time.*

Proof. See Appendix. □

Theorem 3.3. *Consider the bound $\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}\left[\max_{k\in[K]} g_k(\tilde{\boldsymbol{\xi}})\right]$ where the ambiguity set $\mathcal{P}$ satisfies the assumptions (A1)-(A2). Suppose for each $k \in [K]$, the function $g_k : \prod_{i\in[N]} \Xi_i \to \mathbb{R}$ is a supermodular function with a polynomial time evaluation oracle, then the bound is efficiently computable.*

Proof. See Appendix. □

(b) The formulation in (3.1)-(3.2) is an instance of the multimarginal optimal transport problem (see [68, 2]). Consider the functions $\mathbb{1}_{\xi_i=t}$ and $-\mathbb{1}_{\xi_i=t}$ for each $t \in \Xi_i$, $i \in [N]$ (both of which are univariate functions and hence both submodular and supermodular). Then we can recreate the Fréchet ambiguity set with discrete marginals as follows:

$$\begin{aligned}\mathcal{P}_F(\mathbb{P}_1, \ldots, \mathbb{P}_N) &= \{\mathbb{P} \in \mathcal{P}(\prod_{i\in[N]} \Xi_i) \mid \mathbb{E}_{\mathbb{P}}[\mathbb{1}_{\tilde{\xi}_i=\xi_i}] \leq p_i(\xi_i), \mathbb{E}_{\mathbb{P}}[-\mathbb{1}_{\tilde{\xi}_i=\xi_i}] \leq -p_i(\xi_i), \forall \xi_i \in \Xi_i, \forall i \in [N]\big\}, \\ &= \{\mathbb{P} \in \mathcal{P}(\prod_{i\in[N]} \Xi_i) \mid \mathbb{P}(\tilde{\xi}_i = \xi_i) = p_i(\xi_i), \forall \xi_i \in \Xi_i, \forall i \in [N]\big\},\end{aligned} \tag{3.5}$$

where $p_i(\xi_i) = \mathbb{P}_i(\tilde{\xi}_i = \xi_i)$. While the sharpest upper bound on the expectation of the maximum of affine functions of a random vector for the Fréchet ambiguity set is known to be computable in polynomial time (see [17, 66]), Theorem 3.1 allows us to incorporate information on the dependence structure.

(c) Under assumptions (A1)-(A2), testing $\mathcal{P} \neq \emptyset$ is possible in polynomial time. To do so, we solve the dual problem with $f(\boldsymbol{\xi}) = 0$ and check if the dual optimal value is 0 or $-\infty$ (unbounded). When $\mathcal{P} \subseteq \mathcal{P}_F(\mathbb{P}_1, \ldots, \mathbb{P}_N)$, testing feasibility is done by computing $\mathbb{E}_{\mathbb{P}^c}[f_j(\tilde{\boldsymbol{\xi}}^c)]$ for each $j \in [J]$ where $\tilde{\boldsymbol{\xi}}^c \sim \mathbb{P}^c$ and checking if it less than or equal to γ_j. This arises from the extremal characterization of the comonotonic random vector in (2.6). However the comonotonic random vector does not have to be the extremal distribution which attains the bound in f^* unless $f(\boldsymbol{\xi})$ is supermodular (for example, $f(\boldsymbol{\xi}) = \max_i \xi_i$ which submodular, not supermodular). It is also straightforward to see that assumption (A2) which enforces upper bounds on the expected value of submodular functions is equivalent to enforcing lower bounds on the expected value of supermodular functions.

(d) Solving the dual separation problem in Theorem 3.1 requires solving decomposable submodular function minimization problems where the term "decomposable" refers to the sum of submodular functions. When each submodular function is structured, for example depending on only a few variables or having a specific functional form, specialized algorithms are available to minimize decomposable functions [51, 77, 48, 33]). We explore structured submodular functions in later sections of this paper for which we derive compact formulations.

(e) Theorem 3.1 is useful when the uncertainty arises as discrete random variables. Discrete random variables are used to represent random demand for indivisible goods such as cereals, houses and cars. It is used to represent count data such as the number of occurrences of an illness in a patient, the number of times a medication is taken or the number of commuters who choose a path in a network. It is used to represent multiple labels for images in classification tasks, the number of power outages or the number of failures of equipment in systems. Theorem 3.1 is also useful in settings where discretization is used to approximate continuous random variables.

4 Multimarginal optimal transport with positive dependence orders

In this section, we discuss the multimarginal optimal transport problem with positive dependence orders that are specified for the random vector. Dependence orders provide a natural way to compare random vectors in terms of the dependence among the underlying random variables. Such orders have be used in queueing, reliability, project management and risk management (see [65, 75, 73]). We will discuss specific positive dependence orders that can be modeled in a computationally tractable manner and derive sharp bounds.

4.1 Comparison of random vectors with positive dependence orders

Consider a N-dimensional random vector $\tilde{\boldsymbol{\xi}} \sim \mathbb{P}$ where $\mathbb{P}_i = \text{proj}_i(\mathbb{P})$ for all $i \in [N]$. Let $\tilde{\boldsymbol{\xi}}^{\perp} \sim \mathbb{P}^{\perp} := \mathbb{P}_1 \times \ldots \times \mathbb{P}_N$ denote the N-dimensional random vector with the same univariate marginals as $\tilde{\boldsymbol{\xi}}$ but with independent components. The random vector $\tilde{\boldsymbol{\xi}}$ is said to:

(a) Positive upper orthant dependent (PUOD) if:

$$\mathbb{P}\left(\tilde{\boldsymbol{\xi}} > \boldsymbol{t}\right) \geq \prod_{i \in [N]} \mathbb{P}\left(\tilde{\xi}_i > t_i\right), \forall \boldsymbol{t} \in \mathbb{R}^N, \tag{4.1}$$

(b) Positive lower orthant dependent (PLOD) if:

$$\mathbb{P}\left(\tilde{\boldsymbol{\xi}} \leq \boldsymbol{t}\right) \geq \prod_{i \in [N]} \mathbb{P}\left(\tilde{\xi}_i \leq t_i\right), \forall \boldsymbol{t} \in \mathbb{R}^N, \tag{4.2}$$

(c) Positive orthant dependent (POD) if it is both PUOD and PLOD.

Positive dependence in the random vector $\tilde{\boldsymbol{\xi}}$ defined using (a), (b) or (c) is based on comparison with $\tilde{\boldsymbol{\xi}}^{\perp}$ where the components are independent. One can similarly define a negative upper orthant dependent and negative lower orthant dependent random vector by reversing the inequalities in (4.1) and (4.2). For bivariate random vectors with $N = 2$, these dependence orders were first introduced in [55]. A closely related dependence order is defined using supermodular functions. The random vector $\tilde{\boldsymbol{\xi}} \sim \mathbb{P}$ is said to be:

(d) Positive supermodular dependent (PSMD) if:

$$\mathbb{E}_{\mathbb{P}}\left[f(\tilde{\boldsymbol{\xi}})\right] \geq \mathbb{E}_{\mathbb{P}^{\perp}}\left[f(\tilde{\boldsymbol{\xi}}^{\perp})\right], \forall \text{supermodular } f \text{ such that expectations exists.} \tag{4.3}$$

One can similarly define a negative supermodular dependent random vector by reversing the inequality in (4.3). When $N = 2$, all the four positive dependence orders are equivalent with PUOD $\iff$ PLOD $\iff$ PSMD $\iff$ POD (see [79]). All popular measures of association between pairs of random variables such as the Pearson correlation coefficient, the Kendall rank correlation coefficient, the Spearman rank correlation coefficient and the Blomqvist measure of dependence are nonnegative under these dependence orders (see [55]). However for $N \geq 3$, the positive supermodular dependent order is strictly stronger with the implications PSMD $\implies$ PUOD and PSMD $\implies$ PLOD (see [63]).

These definitions have been extended to dependence orders that compare general random vectors. A N-dimensional random vector $\tilde{\boldsymbol{\xi}} \sim \mathbb{P}$ is said to be larger than a N-dimensional random vector $\tilde{\boldsymbol{\chi}} \sim \mathbb{Q}$ in the:

(a) Upper orthant (UO) order if:

$$\mathbb{P}\left(\tilde{\boldsymbol{\xi}} > \boldsymbol{t}\right) \geq \mathbb{Q}\left(\tilde{\boldsymbol{\chi}} > \boldsymbol{t}\right), \forall \boldsymbol{t} \in \mathbb{R}^N, \tag{4.4}$$

(b) Lower orthant (LO) order if:

$$\mathbb{P}\left(\tilde{\boldsymbol{\xi}} \leq \boldsymbol{t}\right) \geq \mathbb{Q}\left(\tilde{\boldsymbol{\chi}} \leq \boldsymbol{t}\right), \forall \boldsymbol{t} \in \mathbb{R}^N, \tag{4.5}$$

(c) Concordance (or orthant) order if it is larger in both the UO and LO orders,

(d) Supermodular (SM) order if:

$$\mathbb{E}_{\mathbb{P}}\left[f(\tilde{\boldsymbol{\xi}})\right] \geq \mathbb{E}_{\mathbb{Q}}\left[f(\tilde{\boldsymbol{\chi}})\right], \forall \text{supermodular } f \text{ such that expectations exists.} \tag{4.6}$$

While the marginals of the random vectors $\tilde{\boldsymbol{\xi}}$ and $\tilde{\boldsymbol{\chi}}$ can be different under the UO and LO orders, it is easy to verify that the marginals of $\tilde{\boldsymbol{\xi}}$ and $\tilde{\boldsymbol{\chi}}$ have to be the same under the concordance and SM orders. For bivariate random vectors where $N = 2$, the UO and LO orders were first introduced in [83]. Again for $N = 2$, UO $\iff$ LO $\iff$ SM $\iff$ Concordance while for $N \geq 3$, SM $\implies$ UO and SM $\implies$ LO where the implications are strict (see [79, 63]). Testing for these dependence orders in the multivariate context is a challenging problem in general (see [30]). Distributionally robust optimization problems with comparison of random vectors using first order and second order stochastic dominance constraints have been studied in [24, 69]. While incorporating such constraints is not easy in general, relaxed notions of stochastic dominance have been considered recently (see [64]).

4.2 Bounds with positive dependence orders

In this section, we discuss compact linear programs to compute sharp bounds for discrete random vectors with positive dependence orders.

4.2.1 Bivariate marginals

We make the following assumptions on the univariate and bivariate marginal distributions for the rest of the section and later discuss how the assumptions can be relaxed.

`Assumption (B1):` Each random variable $\tilde{\xi}_i$ is discrete with probabilities given by $p_i(\xi_i) = \mathbb{P}(\tilde{\xi}_i = \xi_i)$ for $\xi_i \in \Xi_i$, where Ξ_i is a finite set of numbers. The marginal probabilities satisfy the conditions $p_i(\xi_i) > 0$ for all $\xi_i \in \Xi_i$ and $\sum_{\xi_i \in \Xi_i} p_i(\xi_i) = 1$ for all $i \in [N]$.

`Assumption (B2):` Each distinct pair of random variables $(\tilde{\xi}_i, \tilde{\xi}_j)$ is POD for $i \neq j$.

Under assumptions (B1)-(B2), the ambiguity set is given by:

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\textstyle\prod_{i\in[N]} \Xi_i) \mid \mathbb{P}(\tilde{\xi}_i = \xi_i) = p_i(\xi_i), \forall \xi_i \in \Xi_i, \forall i \in [N], \\ \mathbb{P}(\tilde{\xi}_i \geq \xi_i, \tilde{\xi}_j \geq \xi_j) \geq \textstyle\sum_{\xi\geq\xi_i} p_i(\xi) \sum_{\xi\geq\xi_j} p_j(\xi), \forall \xi_i \in \Xi_i, \forall \xi_j \in \Xi_j, \forall i < j \in [N]\}. \tag{4.7}$$

Clearly $\mathcal{P}$ is a subset of the Fréchet set of distributions where the bivariate marginals are also positive orthant dependent. Since in the bivariate case POD, PUOD, PLOD and PSMD are all equivalent, we can use any one of these definitions interchangeably. The strict inequalities in the original definition of PUOD in (4.1) are relaxed to inequalities in (4.7). This is without loss of generality, since the distributions are discrete. The set of distributions $\mathcal{P}$ in (4.7) is clearly nonempty. Both the independent random vector $\tilde{\boldsymbol{\xi}}^{\perp} \sim \mathbb{P}^{\perp}$ and the comonotonic random vector $\tilde{\boldsymbol{\xi}}^{c} \sim \mathbb{P}^{c}$ lie in $\mathcal{P}$. While positive dependence of pairs of random variables are enforced in $\mathcal{P}$, no assumption on the dependencies of higher order marginals (three or more) is made.

We show that the sharpest upper bound on the expectation of the maximum of affine functions of a random vector for this ambiguity set is computable using a polynomial sized linear program. Observe that since the functions $\mathbb{1}_{\xi_i > t_i, \xi_j > t_j}$ are supermodular in (ξ_i, ξ_j) for any t_i and t_j, the ambiguity set satisfies assumptions (A1)-(A2) and the bound f^* is computable in polynomial time from Theorem 3.1. We provide a direct probabilistic construction of the bound using a compact primal linear program without going through duality. One advantage of this approach is that it provides an interpretation of the extremal distribution as a mixture of comonotonic random vectors where the decision variables in the linear program can be viewed as conditional probabilities and the constraints as probabilistic conditions that the distribution must satisfy.

Theorem 4.1. *Under assumptions (B1)-(B2) on the ambiguity set $\mathcal{P}$, f^* in* (3.1) *is given by the optimal value of the polynomial sized linear program:*

$$\begin{aligned}
\max_{\boldsymbol{\lambda},\gamma} \quad & \sum_{k\in[K]} \sum_{i\in[N]} \sum_{\xi_i\in\Xi_i} a_{i,k}\xi_i\gamma_{i,k}(\xi_i) + \sum_{k\in[K]} b_k\lambda_k & & \\
s.t. \quad & \sum_{k\in[K]} \lambda_k = 1, & & \\
& \sum_{k\in[K]} \gamma_{i,k}(\xi_i) = p_i(\xi_i), & & \forall \xi_i \in \Xi_i, \forall i \in [N], \\
& \sum_{\xi_i\in\Xi_i} \gamma_{i,k}(\xi_i) = \lambda_k, & & \forall i \in [N], \forall k \in [K], \\
& \gamma_{i,j,k}(\xi_i, \xi_j) \leq \sum_{\xi\geq\xi_i} \gamma_{i,k}(\xi), & & \forall \xi_i \in \Xi_i, \forall \xi_j \in \Xi_j, \forall i < j \in [N], \forall k \in [K], \\
& \gamma_{i,j,k}(\xi_i, \xi_j) \leq \sum_{\xi\geq\xi_j} \gamma_{j,k}(\xi), & & \forall \xi_i \in \Xi_i, \forall \xi_j \in \Xi_j, \forall i < j \in [N], \forall k \in [K], \\
& \sum_{k\in[K]} \gamma_{i,j,k}(\xi_i, \xi_j) \geq \sum_{\xi\geq\xi_i} p_i(\xi) \sum_{\xi\geq\xi_j} p_j(\xi) & & \forall \xi_i \in \Xi_i, \forall \xi_j \in \Xi_j, \forall i < j \in [N], \\
& \lambda_k \geq 0, & & \forall k \in [K], \\
& \gamma_{i,k}(\xi_i) \geq 0, & & \forall \xi_i \in \Xi_i, \forall i \in [N], \forall k \in [K], \\
& \gamma_{i,j,k}(\xi_i, \xi_j) \geq 0, & & \forall \xi_i \in \Xi_i, \forall \xi_j \in \Xi_j, \forall i < j \in [N], \forall k \in [K].
\end{aligned} \tag{4.8}$$

Proof. Let f_u^* be optimal value of the linear program (4.8). We prove $f^* = f_u^*$ in two steps by showing $f^* \leq f_u^*$ and $f_u^* \leq f^*$.
Step (1): $f^* \leq f_u^*$
To show f_u^* is a valid upper bound on f^*, we start with a probabilistic construction of the formulation (4.8). Define for each possible realization $\boldsymbol{\xi} \in \prod_{i\in[N]} \Xi_i$, the set of indices that attain the maximum in $f(\boldsymbol{\xi})$ as follows:

$$K(\boldsymbol{\xi}) = \arg\max \left\{ \boldsymbol{a}_k' \boldsymbol{\xi} + b_k \mid k \in [K] \right\}.$$

For each $\boldsymbol{\xi}$, $K(\boldsymbol{\xi}) \subseteq [K]$ is a singleton where $|K(\boldsymbol{\xi})| = 1$ or contain multiple indices where $|K(\boldsymbol{\xi})| > 1$. Given a random vector $\tilde{\boldsymbol{\xi}}$ with distribution $\mathbb{P}$, let $k(\boldsymbol{\xi})$ be a measurable selection on the set $K(\boldsymbol{\xi})$ (for example, one can define $k(\boldsymbol{\xi})$ as the smallest index in $K(\boldsymbol{\xi})$). Define the decision variables as:

$$\begin{aligned}
\lambda_k = &\quad \mathbb{P}\left(k(\tilde{\boldsymbol{\xi}}) = k\right), && \forall k \in [K], \\
\gamma_{i,k}(\xi_i) = &\quad \mathbb{P}\left(\tilde{\xi}_i = \xi_i, k(\tilde{\boldsymbol{\xi}}) = k\right), && \forall \xi_i \in \Xi_i, \forall i \in [N], \forall k \in [K], \\
\gamma_{i,j,k}(\xi_i, \xi_j) = &\quad \mathbb{P}\left(\tilde{\xi}_i \geq \xi_i, \tilde{\xi}_j \geq \xi_j, k(\tilde{\boldsymbol{\xi}}) = k\right), && \forall \xi_i \in \Xi_i, \forall \xi_j \in \Xi_j, \forall i < j \in [N], \forall k \in [K].
\end{aligned}$$

Clearly the variables as defined must satisfy the nonnegativity constraints in (4.8). The first six constraints in the formulation are derived from necessary conditions that the variables by definition must satisfy:

1. Total sum of the probabilities of indices being optimal is one:

$$\sum_{k\in[K]} \mathbb{P}\left(k(\tilde{\boldsymbol{\xi}}) = k\right) = 1.$$

2. Law of total probability for the univariate marginal probabilities:

$$\sum_{k\in[K]} \mathbb{P}\left(\tilde{\xi}_i = \xi_i, k(\tilde{\boldsymbol{\xi}}) = k\right) = \mathbb{P}\left(\tilde{\xi}_i = \xi_i\right).$$

3. Law of total probability for the index being optimal:

$$\sum_{\xi_i\in\Xi_i} \mathbb{P}\left(\tilde{\xi}_i = \xi_i, k(\tilde{\boldsymbol{\xi}}) = k\right) = \mathbb{P}\left(k(\tilde{\boldsymbol{\xi}}) = k\right).$$

4. Probability of the event $\{\tilde{\xi}_i \geq \xi_i, \tilde{\xi}_j \geq \xi_j, k(\tilde{\boldsymbol{\xi}}) = k\}$ is upper bounded by the probability of the event $\{\tilde{\xi}_i \geq \xi_i, k(\tilde{\boldsymbol{\xi}}) = k\}$:

$$\mathbb{P}\left(\tilde{\xi}_i \geq \xi_i, \tilde{\xi}_j \geq \xi_j, k(\tilde{\boldsymbol{\xi}}) = k\right) \leq \sum_{\xi\geq\xi_i} \mathbb{P}\left(\tilde{\xi}_i = \xi, k(\tilde{\boldsymbol{\xi}}) = k\right).$$

5. Probability of the event $\{\tilde{\xi}_i \geq \xi_i, \tilde{\xi}_j \geq \xi_j, k(\tilde{\boldsymbol{\xi}}) = k\}$ is upper bounded by the probability of the

event $\{\tilde{\xi}_j \geq \xi_j, k(\tilde{\boldsymbol{\xi}}) = k\}$:

$$\mathbb{P}\left(\tilde{\xi}_i \geq \xi_i, \tilde{\xi}_j \geq \xi_j, k(\tilde{\boldsymbol{\xi}}) = k\right) \leq \sum_{\xi \geq \xi_j} \mathbb{P}\left(\tilde{\xi}_j = \xi, k(\tilde{\boldsymbol{\xi}}) = k\right).$$

6. Law of total probability for the PUOD condition:

$$\sum_{k\in[K]} \mathbb{P}\left(\tilde{\xi}_i \geq \xi_i, \tilde{\xi}_j \geq \xi_j, k(\tilde{\boldsymbol{\xi}}) = k\right) \geq \mathbb{P}\left(\tilde{\xi}_i \geq \xi_i\right) \mathbb{P}\left(\tilde{\xi}_j \geq \xi_j\right).$$

The objective in (4.8) is obtained by expressing the expected function value in terms of the decision variables:

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}}\left[\max_{k\in[K]}\left(\boldsymbol{a}_k'\tilde{\boldsymbol{\xi}} + b_k\right)\right] &= \sum_{k\in[K]} \mathbb{E}_{\mathbb{P}}\left[\left(\boldsymbol{a}_k'\tilde{\boldsymbol{\xi}} + b_k\right) | k(\tilde{\boldsymbol{\xi}}) = k\right] \mathbb{P}\left(k(\tilde{\boldsymbol{\xi}}) = k\right), \\
&= \sum_{k\in[K]} \sum_{i\in[N]} a_{i,k} \mathbb{E}_{\mathbb{P}}\left[\tilde{\xi}_i | k(\tilde{\boldsymbol{\xi}}) = k\right] \mathbb{P}\left(k(\tilde{\boldsymbol{\xi}}) = k\right) + \sum_{k\in[K]} b_k \mathbb{P}\left(k(\tilde{\boldsymbol{\xi}}) = k\right), \\
&= \sum_{k\in[K]} \sum_{i\in[N]} \sum_{\xi_i\in\Xi_i} a_{i,k}\xi_i \mathbb{P}\left(\tilde{\xi}_i = \xi_i, k(\tilde{\boldsymbol{\xi}}) = k\right) + \sum_{k\in[K]} b_k \mathbb{P}\left(k(\tilde{\boldsymbol{\xi}}) = k\right), \\
&= \sum_{k\in[K]} \sum_{i\in[N]} \sum_{\xi_i\in\Xi_i} a_{i,k}\xi_i \gamma_{i,k}(\xi_i) + \sum_{k\in[K]} b_k \lambda_k.
\end{aligned}$$

From the necessity of all the constraints, we have $f^* \leq f_u^*$. We next prove sufficiency.

Step (2): $f^* \geq f_u^*$

We construct a distribution $\mathbb{P}^* \in \mathcal{P}$ that attains the upper bound f_u^* using the optimal solution of the linear program. Consider an optimal solution of the linear program (4.8) denoted by $(\boldsymbol{\lambda}^*, \boldsymbol{\gamma}^*)$. Create a mixture distribution $\mathbb{P}^*$ as follows:

(i) Generate a discrete random variable $\tilde{z}$ that takes values in $[K]$ with probability $\mathbb{P}^*(\tilde{z} = k) = \lambda_k^*$.

(ii) Conditional on the realization of $\tilde{z}$, define the marginal distribution of each random variable $\tilde{\xi}_i$ as:

$$\mathbb{P}^*\left(\tilde{\xi}_i = \xi_i \middle| \tilde{z} = k\right) = \frac{\gamma_{i,k}^*(\xi_i)}{\sum_{\xi\in\Xi_i} \gamma_{i,k}^*(\xi)}, \forall \xi_i \in \Xi_i.$$

Generate in step (ii), a comonotonic random vector using these conditional marginal distributions. In the construction, $\sum_{\xi\in\Xi_i} \gamma_{i,k}^*(\xi) > 0$ if and only if $\mathbb{P}^*(\tilde{z} = k) = \lambda_k^* > 0$. The marginal distribution of $\tilde{\xi}_i$ in the mixture distribution $\mathbb{P}^*$ is given by:

$$\begin{aligned}
\mathbb{P}^*\left(\tilde{\xi}_i = \xi_i\right) &= \sum_{k\in[K]} \lambda_k^* \mathbb{P}^*\left(\tilde{\xi}_i = \xi_i \middle| \tilde{z} = k\right), \\
&= \sum_{k\in[K]} \lambda_k^* \left(\frac{\gamma_{i,k}^*(\xi_i)}{\sum_{\xi\in\Xi_i} \gamma_{i,k}^*(\xi)}\right), \\
&= p_i(\xi_i), \\
&\quad [\text{since } \textstyle\sum_{\xi\in\Xi_i} \gamma_{i,k}^*(\xi) = \lambda_k^*].
\end{aligned}$$

Hence the univariate marginal distributions of $\mathbb{P}^*$ match the univariate marginal distributions specified in $\mathcal{P}$. The bivariate upper orthant probability of $\tilde{\xi}_i$ and $\tilde{\xi}_j$ in $\mathbb{P}^*$ is given by:

$$\begin{aligned}
\mathbb{P}^*\left(\tilde{\xi}_i \geq \xi_i, \tilde{\xi}_j \geq \xi_j\right) &= \sum_{k\in[K]} \lambda_k^* \mathbb{P}^*\left(\tilde{\xi}_i \geq \xi_i, \tilde{\xi}_j \geq \xi_j | \tilde{z} = k\right), \\
&= \sum_{k\in[K]} \lambda_k^* \min\left(\mathbb{P}^*\left(\tilde{\xi}_i \geq \xi_i | \tilde{z} = k\right), \mathbb{P}^*\left(\tilde{\xi}_j \geq \xi_j | \tilde{z} = k\right)\right), \\
&\quad \text{[from the comonotonic construction in step (ii) and using property (2.4)]}, \\
&= \sum_{k\in[K]} \lambda_k^* \min\left(\frac{\sum_{\xi\geq\xi_i} \gamma_{i,k}^*(\xi)}{\sum_{\xi\in\Xi_i} \gamma_{i,k}^*(\xi)}, \frac{\sum_{\xi\geq\xi_j} \gamma_{j,k}^*(\xi)}{\sum_{\xi\in\Xi_j} \gamma_{j,k}^*(\xi)}\right), \\
&\quad \text{[from step (ii)]}, \\
&= \sum_{k\in[K]} \min\left(\sum_{\xi\geq\xi_i} \gamma_{i,k}^*(\xi), \sum_{\xi\geq\xi_j} \gamma_{j,k}^*(\xi)\right), \\
&\quad \text{[since } \textstyle\sum_{\xi\in\Xi_i} \gamma_{i,k}^*(\xi) = \sum_{\xi\in\Xi_j} \gamma_{j,k}^*(\xi) = \lambda_k^*\text{]}, \\
&\geq \sum_{k\in[K]} \gamma_{i,j,k}^*(\xi_i, \xi_j), \\
&\quad \text{[since } \min(\textstyle\sum_{\xi\geq\xi_i} \gamma_{i,k}^*(\xi), \sum_{\xi\geq\xi_j} \gamma_{j,k}^*(\xi)) \geq \gamma_{i,j,k}^*(\xi_i, \xi_j) \text{ from (4.8)]}, \\
&\geq \sum_{\xi\geq\xi_i} p_i(\xi) \sum_{\xi\geq\xi_j} p_j(\xi).
\end{aligned}$$

Hence $\mathbb{P}^* \in \mathcal{P}$. The final step is to show the sharpness of the bound under this distribution as follows:

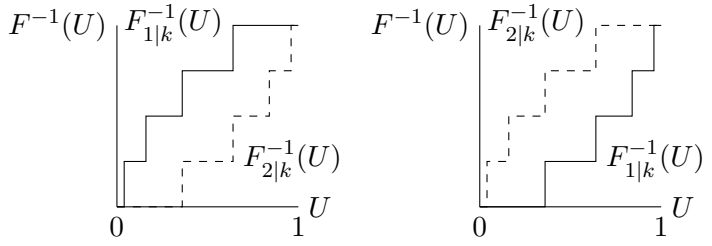
$$\begin{aligned}
f^* &\geq \mathbb{E}_{\mathbb{P}^*}\left[\max_{k\in[K]}\left(\boldsymbol{a}_k'\tilde{\boldsymbol{\xi}} + b_k\right)\right], \\
&\quad \text{[since } \mathbb{P}^* \in \mathcal{P}\text{]}, \\
&\geq \sum_{k\in[K]} \lambda_k^* \mathbb{E}_{\mathbb{P}^*}\left[\left(\boldsymbol{a}_k'\tilde{\boldsymbol{\xi}} + b_k\right) | \tilde{z} = k\right], \\
&\quad \text{[evaluating the expected value at the } k\text{th piece in step (ii) instead of the optimal piece]}, \\
&= \sum_{k\in[K]} \lambda_k^* \sum_{i\in[N]} \sum_{\xi_i\in\Xi_i} \left(\frac{a_{i,k}\xi_i \gamma_{i,k}^*(\xi_i)}{\sum_{\xi\in\Xi_i} \gamma_{i,k}^*(\xi)}\right) + \sum_{k\in[K]} b_k \lambda_k^*, \\
&= \sum_{k\in[K]} \sum_{i\in[N]} \sum_{\xi_i\in\Xi_i} \lambda_k^* \left(\frac{a_{i,k}\xi_i \gamma_{i,k}^*(\xi_i)}{\sum_{\xi\in\Xi_i} \gamma_{i,k}^*(\xi)}\right) + \sum_{k\in[K]} b_k \lambda_k^*, \\
&= \sum_{k\in[K]} \sum_{i\in[N]} \sum_{\xi_i\in\Xi_i} a_{i,k}\xi_i \gamma_{i,k}^*(\xi_i) + \sum_{k\in[K]} b_k \lambda_k^*, \\
&\quad \text{[since } \textstyle\sum_{\xi\in\Xi_i} \gamma_{i,k}^*(\xi) = \lambda_k^*\text{]}, \\
&= f_u^*.
\end{aligned}$$

From steps (1) and (2), $f^* = f_u^*$. □

We make a few remarks about Theorem 4.1 and its implications next.

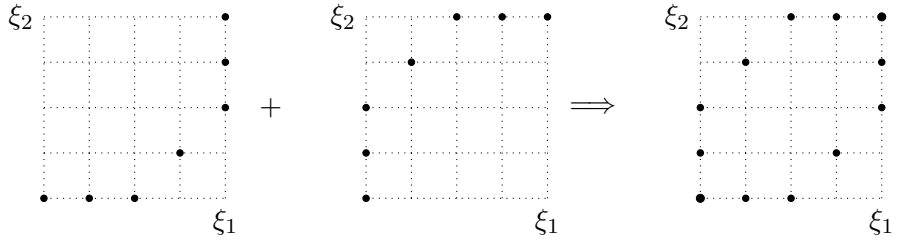
(a) In the proof of Theorem 4.1, the extremal distribution is constructed using a mixture of comonotonic random vectors (see Figures 2 and 3 for an illustration of the construction). However the

extremal distribution itself need not be comonotonic.



(a) $k = 1$ (b) $k = 2$

Figure 2: Comonotonic construction for the conditional distributions in step (ii) for $N = 2$ and $K = 2$. Here the solid line indicates $\tilde{\xi}_1 \sim F^{-1}_{1|k}(U)$ and the dashed line indicates $\tilde{\xi}_2 \sim F^{-1}_{2|k}(U)$ for $k = 1$ (left figure) and $k = 2$ (right figure) where $F_{i|k}$ is the conditional marginal distribution of $\tilde{\xi}_i$ for index k being optimal.



(a) $k = 1$ (b) $k = 2$ (c) Overall

Figure 3: Subfigures (a) and (b) display the support of the conditional bivariate distributions for $k = 1$ and $k = 2$ while (c) shows that overall support of the extremal bivariate distribution using the weighted probabilities λ^*_1 and λ^*_2. While the conditional bivariate distributions in (a) and (b) are comonotonic, the final bivariate distribution in (c) is not comonotonic. The distribution in (c) is however POD.

(b) The size of the linear program in Theorem 4.1 is polynomial in N (number of random variables), K (number of affine pieces defining the function f) and $\max_{i\in[N]} |\Xi_i|$ (maximum number of values that a variable can take). The proof builds on the proof of Theorem 2.3.1 in [66]. However the linear program therein is developed only under assumption (B1). The key novelty in Theorem 4.1 is that by introducing additional decision variables in terms of conditional bivariate probabilities and incorporating the correct constraints, it is possible to compute the sharp bound and preserve computational tractability. The structured form of the supermodular functions defining the ambiguity set in (4.7) helps us do so.

(c) While Theorem 4.1 is derived under the assumption of POD bivariate marginals, it is straightforward to see that the formulation can be extended to concordance orders for bivariate marginals. Specifically if the bivariate marginal distribution $(\tilde{\xi}_i, \tilde{\xi}_j) \sim \text{proj}_{i,j}(\mathbb{P})$ is assumed to be larger than $(\tilde{\chi}_i, \tilde{\chi}_j) \sim \mathbb{Q}_{i,j}$ in the concordance order, then the right hand side of the sixth constraint in (4.8) will be changed from $\mathbb{P}_i(\tilde{\xi}_i \geq \xi_i)\mathbb{P}_j(\tilde{\xi}_j \geq \xi_j)$ to $\mathbb{Q}_{i,j}(\tilde{\xi}_i \geq \xi_i, \tilde{\xi}_j \geq \xi_j)$. Another straightforward

extension of the formulation is when the lower bounds on the bivariate tail probabilities of $(\tilde{\xi}_i, \tilde{\xi}_j)$ are given for only a subset $\Xi_{i,j} \subseteq \Xi_i \times \Xi_j$. In this case, the fourth, fifth and sixth set of linear constraints in (4.8) have to be enforced only for the set $(\xi_i, \xi_j) \in \Xi_{i,j}$.

(d) A special case where a sharp closed form bound is known is when $\Xi = \{0,1\}^N$ and $f(\boldsymbol{\xi}) = \max_{i\in[N]} \xi_i$. In this case, the bound reduces to finding the maximum probability of the union of a set of dependent events where the marginal probability of each event and lower bounds on the probability of pairs of events occurring are known. Let $p_i = \mathbb{P}(\tilde{\xi}_i = 1) = 1 - \mathbb{P}(\tilde{\xi}_i = 0)$ for $i \in [N]$ and $\mathbb{P}(\tilde{\xi}_i = 1, \tilde{\xi}_j = 1) \geq p_{ij}$ for $i < j \in [N]$. The ambiguity set is given by:

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\{0,1\}^N) \mid \mathbb{P}(\tilde{\xi}_i = 1) = p_i, \forall i \in [N], \mathbb{P}(\tilde{\xi}_i = 1, \tilde{\xi}_j = 1) \geq p_{ij}, \forall i < j \in [N]\big\}.$$

Then, [61, 62] proved that the Hunter-Worsley bound (see [43, 82]) is a tight upper bound on the union probability for this ambiguity set:

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}\left[\max_{i\in[N]} \tilde{\xi}_i\right] = \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\left(\sum_{i\in[N]} \tilde{\xi}_i \geq 1\right) = \min\left(1, \sum_{i\in[N]} p_i - \max_{\mathbb{T}\in\mathcal{T}} \sum_{(i,j)\in\mathbb{T}} p_{ij}\right),$$

where $\mathcal{T}$ is the set of all spanning trees in a complete graph with nodes indexed by $[N]$ and the weight of the edge between distinct nodes i and j is given by p_{ij}. Theorem 4.1 provides a compact linear program and generalizes from Boolean random variables to discrete random variables with finite support and from the expected maximum of random variables to the maximum of affine functions of the random vector.

(e) The bound in Theorem 4.1 can be used to solve the distributionally robust optimization problem in (1.1). Specifically by using linear programming duality, we obtain the reformulation:

$$\begin{array}{ll}
\inf & t + \displaystyle\sum_{i\in[N]} \sum_{\xi_i\in\Xi_i} y_i(\xi_i) p_i(\xi_i) - \sum_{i\in[N]} \sum_{j\in[N]:j>i} \sum_{\xi_i\in\Xi_i} \sum_{\xi_j\in\Xi_j} l_{ij}(\xi_i,\xi_j) \left(\sum_{\xi\geq\xi_i} p_i(\xi) \sum_{\xi\geq\xi_j} p_j(\xi)\right) \\
\text{s.t.} & t - \displaystyle\sum_{i\in[N]} g_{ik} \geq b_k(\boldsymbol{x}), \forall k \in [K], \\
& h_{ijk}(\xi_i,\xi_j) + q_{ijk}(\xi_i,\xi_j) - l_{ij}(\xi_i,\xi_j) \geq 0, \forall \xi_i \in \Xi_i, \forall \xi_j \in \Xi_j, \forall i \in [N], \forall j \in [N], \forall k \in [K], \\
& y_i(\xi_i) + g_{ik} - \displaystyle\sum_{\xi\leq\xi_i} \sum_{j\in[N]:j>i} \sum_{\xi_j\in\Xi_j} h_{ijk}(\xi,\xi_j) - \sum_{j\in[N]:j<i} \sum_{\xi_j\in\Xi_j} \sum_{\xi\leq\xi_i} q_{ijk}(\xi_j,\xi) \geq a_{i,k}(\boldsymbol{x})\xi_i, \\
& \qquad\qquad \forall \xi_i \in \Xi_i, \forall i \in [N], \forall k \in [K], \\
& h_{ijk}(\xi_i,\xi_j) \geq 0, \forall \xi_i \in \Xi_i, \forall \xi_j \in \Xi_j, \forall i < j \in [N], \forall k \in [K], \\
& q_{ijk}(\xi_i,\xi_j) \geq 0, \forall \xi_i \in \Xi_i, \forall \xi_j \in \Xi_j, \forall i < j \in [N], \forall k \in [K], \\
& l_{ij}(\xi_i,\xi_j) \geq 0, \forall \xi_i \in \Xi_i, \forall \xi_j \in \Xi_j, \forall i < j \in [N], \\
& \boldsymbol{x} \in \mathcal{X}.
\end{array} \tag{4.9}$$

Here the decision variables are $\boldsymbol{x}$, t, $y_i(\xi_i)$, $l_{ij}(\xi_i,\xi_j)$, $h_{ijk}(\xi_i,\xi_j)$, $q_{ijk}(\xi_i,\xi_j)$ and g_{ik}.

4.2.2 Higher order marginals

We discuss a generalization to higher order marginals. For ease of exposition, we restrict attention to Boolean random variables in this section. The results can be extended in a straightforward manner to discrete random variables.

`Assumption (C1):` Each random variable $\tilde{\xi}_i$ is Boolean with probabilities given by $p_i = \mathbb{P}(\tilde{\xi}_i = 1) = 1 - \mathbb{P}(\tilde{\xi}_i = 0)$ where $p_i \in (0,1)$ for $i \in [N]$.

`Assumption (C2):` Every subset of random variables of size up to M, namely $(\tilde{\xi}_i; i \in I)$ for all $I \subseteq [N]$, $1 < |I| \leq M$ is PUOD.

Under assumptions (C1)-(C2), the ambiguity set is given by:

$$\mathcal{P} = \big\{\mathbb{P} \in \mathcal{P}(\{0,1\}^N) \mid \mathbb{P}(\tilde{\xi}_i = 1) = p_i, \forall i \in [N], \mathbb{P}(\textstyle\prod_{i\in I} \tilde{\xi}_i = 1) \geq \prod_{i\in I} p_i, \forall I \subseteq [N] : 1 < |I| \leq M\big\}. \tag{4.10}$$

This brings us to the following theorem.

Theorem 4.2. *Under assumptions (C1)-(C2) on the ambiguity set $\mathcal{P}$, f^* in* (3.1) *is given by the optimal value of the polynomial sized linear program:*

$$\begin{aligned}
\max_{\boldsymbol{\lambda},\boldsymbol{\gamma}} \quad & \sum_{k\in[K]} \sum_{i\in[N]} a_{i,k}\gamma_{i,k} + \sum_{k\in[K]} b_k\lambda_k & & \\
s.t. \quad & \sum_{k\in[K]} \lambda_k = 1, & & \\
& \sum_{k\in[K]} \gamma_{i,k} = p_i, & \forall i \in [N], & \\
& \gamma_{i,k} \leq \lambda_k, & \forall i \in [N], \forall k \in [K], & \\
& \gamma_{I,k} \leq \lambda_k, & \forall I \subseteq [N] : 1 < |I| \leq M, \forall k \in [K], & \\
& \gamma_{I,k} \leq \gamma_{i,k}, & \forall I \subseteq [N] : 1 < |I| \leq M, \forall i \in I, \forall k \in [K], & \\
& \sum_{k\in K} \gamma_{I,k} \geq \prod_{i\in I} p_i & \forall I \subseteq [N] : 1 < |I| \leq M, & \\
& \lambda_k \geq 0, & \forall k \in [K], & \\
& \gamma_{i,k} \geq 0, & \forall i \in [N], \forall k \in [K], & \\
& \gamma_{I,k} \geq 0, & \forall I \subseteq [N] : 1 < |I| \leq M, \forall k \in [K]. &
\end{aligned} \tag{4.11}$$

Proof. See Appendix. □

We make some remarks about Theorem 4.2 next.

(a) The polynomial time computability of the bound for $f(\boldsymbol{\xi}) = \max_{i\in[N]} \xi_i$ and the ambiguity set in (4.10) was shown in [62] (see Theorems 2.3 and 2.4 therein). While their result is based on the ellipsoid method, the key usefulness of Theorem 4.2 is that it provides a compact linear program and generalizes to the maximum of affine functions.

(b) An interesting aspect of formulation (4.11) is that somewhat surprisingly no constraints are required to link the variables $\gamma_{I,k}$ and $\gamma_{J,k}$ for sets I and J where $I \subseteq J$ with $|I| > 1$. This helps reduce the number of linear constraints while still guaranteeing sharpness of the bound.

(c) The result can be extended to find the sharp upper bound on the expected value of $f(\boldsymbol{\xi}) = \min(\sum_{i\in[N]} \xi_i, B)$ where $\boldsymbol{\xi}$ is Boolean and B is an integer between 1 and N. In this case, $f(\boldsymbol{\xi})$ reduces to the set function $f(S) = \min(|S|, B)$ which is exactly the rank function of a B-uniform matroid (see the next corollary).

(d) As with Theorem 4.1, the formulation in (4.11) can be extended to concordance orders for higher order marginals. Specifically, for a given $I \subseteq [N] : 1 < |I| \leq M$, the marginal distributions $\tilde{\xi}_I \sim \text{proj}_I(\mathbb{P})$ is assumed to be larger than $\tilde{\chi}_I \sim \mathbb{Q}_I$ and the right hand side of the sixth constraint in (4.11) can be modified from $\prod_{i\in I} p_i$ to $\mathbb{E}_{\mathbb{Q}}[\tilde{\chi}_I] = q_I$.

Corollary 4.1. *Under assumptions (C1)-(C2) on the ambiguity set $\mathcal{P}$ with $f(\boldsymbol{\xi}) = \min(\sum_{i\in[N]} \xi_i, B)$ and $B = O(1)$ (constant that is independent of N), f^* in (3.1) is computable by solving a polynomial sized linear program.*

Proof. See Appendix. □

5 Moment problems

In this section, we discuss moment problems that are efficiently solvable in the discrete case. Throughout, we assume that lower bounds on the cross moments of pairs of random variables are given. Such moment conditions satisfy the condition discussed in Section 3, namely lower bounds on the expected value of supermodular functions of the random vector are specified. We develop compact linear to solve these problems in polynomial time. We make the following assumptions on the moments of the random vector for the rest of the section.

`Assumption (D1):` Each random variable $\tilde{\xi}_i$ is discrete with support contained in a finite set of numbers denoted by Ξ_i. For each random variable, the first L moments are specified where $m_{i,l} = \mathbb{E}[\tilde{\xi}_i^l]$ for $l \in [L]$.

`Assumption (D2):` Each distinct pair of random variables has a lower bound on the cross moment given by $\mathbb{E}[\tilde{\xi}_i\tilde{\xi}_j] \geq Q_{i,j}$ for $i \neq j$.

Under assumptions (D1)-(D2), the ambiguity set is given by:

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\textstyle\prod_{i\in[N]} \Xi_i) \mid \mathbb{E}_{\mathbb{P}}[\tilde{\xi}_i^l] = m_{i,l}, \forall l \in [L], \forall i \in [N], \mathbb{E}_{\mathbb{P}}[\tilde{\xi}_i\tilde{\xi}_j] \geq Q_{ij}, \forall i < j \in [N]\}. \tag{5.1}$$

Unlike the previous section, the marginal distributions are not specified but a set of marginal moments are. The next theorem provides a linear program to compute the sharp bound.

Theorem 5.1. *Under assumptions (D1)-(D2) on the ambiguity set $\mathcal{P}$, f^* in (3.1) is given by the*

optimal value of the polynomial sized linear program:

$$
\begin{aligned}
\max_{\boldsymbol{\lambda},\gamma} \quad & \sum_{k\in[K]}\sum_{i\in[N]}\sum_{\xi_i\in\Xi_i} a_{i,k}\xi_i\gamma_{i,k}(\xi_i) + \sum_{k\in[K]} b_k\lambda_k \\
s.t. \quad & \sum_{k\in[K]} \lambda_k = 1, \\
& \sum_{k\in[K]}\sum_{\xi_i\in\Xi_i} \xi_i^l\gamma_{i,k}(\xi_i) = m_{i,l}, && \forall l\in[L], \forall i\in[N], \\
& \sum_{\xi_i\in\Xi_i} \gamma_{i,k}(\xi_i) = \lambda_k, && \forall i\in[N], \forall k\in[K], \\
& \sum_{\xi_i\in\Xi_i} \gamma_{i,j,k}(\xi_i,\xi_j) = \gamma_{j,k}(\xi_j), && \forall \xi_j\in\Xi_j, \forall i<j\in[N], \forall k\in[K], \\
& \sum_{\xi_j\in\Xi_j} \gamma_{i,j,k}(\xi_i,\xi_j) = \gamma_{i,k}(\xi_i), && \forall \xi_i\in\Xi_i, \forall i<j\in[N], \forall k\in[K], \\
& \sum_{k\in[K]}\sum_{\xi_i\in\Xi_i}\sum_{\xi_j\in\Xi_j} \xi_i\xi_j\gamma_{i,j,k}(\xi_i,\xi_j) \geq Q_{ij}, && \forall i<j\in[N], \\
& \lambda_k \geq 0, && \forall k\in[K], \\
& \gamma_{i,k}(\xi_i) \geq 0, && \forall \xi_i\in\Xi_i, \forall i\in[N], \forall k\in[K], \\
& \gamma_{i,j,k}(\xi_i,\xi_j) \geq 0, && \forall \xi_i\in\Xi_i, \forall \xi_j\in\Xi_j, \forall i<j\in[N], \forall k\in[K].
\end{aligned}
\tag{5.2}
$$

Proof. See Appendix. □

We make a few remarks about Theorem 5.1 next.

(a) The definition of the variable $\gamma_{i,k}(\xi_i)$ in the formulation is identical to (4.8) where $\gamma_{i,k}(\xi_i) = \mathbb{P}(\tilde{\xi}_i = \xi_i, k(\tilde{\boldsymbol{\xi}}) = k)$. However the definition of the variable $\gamma_{i,j,k}(\xi_i,\xi_j)$ in the formulation is different. Here the variable $\gamma_{i,j,k}(\xi_i,\xi_j) = \mathbb{P}(\tilde{\xi}_i = \xi_i, \tilde{\xi}_j = \xi_j, k(\tilde{\boldsymbol{\xi}}) = k)$ while in (4.8) the decision variable is defined as $\gamma_{i,j,k}(\xi_i,\xi_j) = \mathbb{P}(\tilde{\xi}_i \geq \xi_i, \tilde{\xi}_j \geq \xi_j, k(\tilde{\boldsymbol{\xi}}) = k)$. The modified definition of the variable makes it easier to incorporate the cross moment constraints.

(b) The size of the linear program in Theorem 5.1 is polynomial in N (number of random variables), K (number of affine pieces defining the function f), $\max_{i\in[N]} |\Xi_i|$ (maximum number of values that a variable can take) and L (number of marginal moments). To the best of our knowledge, this is one of the few instances for the multivariate discrete moment problem where the sharp bound is computable in polynomial time with linear programming.

(c) Theorem 5.1 can also be extended to handle the case where moments of L univariate functions of $\tilde{\boldsymbol{\xi}}$ are available and lower bounds on the cross moments $\mathbb{E}[\tilde{\xi}_i\tilde{\xi}_j]$ are available. More specifically let the ambiguity set $\mathcal{P}$ contain all probability distributions satisfying: $\mathbb{E}_{\mathbb{P}}[h_l(\tilde{\xi}_i)] = m_{i,l} \forall i\in[N], \forall l\in[L]$ and $\mathbb{E}_{\mathbb{P}}[\tilde{\xi}_i\tilde{\xi}_j] \geq Q_{ij} \forall i<j\in[N]$. The f^* can be computed by the linear program in Theorem 5.1 with the constraint involving $m_{i,l}$ replaced by $\sum_{k\in[K]}\sum_{\xi_i\in\Xi_i} h_l(\xi_i)\gamma_{i,k}(\xi_i) = m_{i,l} \forall l\in[L], \forall i\in[N]$.

6 Numerical experiments

In this section, we provide numerical experiments to illustrate the quality of the bounds. Our numerical experiments in Section 6.1 show that the bounds with positive dependence orders improve by 2 to 8 percent over bounds that use no dependence information. In the moment case, with higher order marginal moments, the bounds improve by 8 to 15 percent over bounds that use only the first moment. All experiments were conducted on a MacBook with 16GB of RAM using Gurobi solver version 10 with Python.

6.1 Analysis of bounds with positive dependence orders

In this section, we present numerical experiments to showcase the benefits of using Theorem 4.2 in solving multimarginal optimal transport problems with higher order marginal information for Bernoulli random vectors. In particular, the compact formulation in (4.11) was implemented for $N = 8$ with random variables that satisfy the concordance order

$$\mathbb{E}_{\mathbb{P}}\left[\tilde{\xi}_I\right] \geq \mathbb{E}_{\mathbb{Q}}\left[\tilde{\chi}_I\right] = q_I,\ \forall I \subseteq [N] : 2 \leq |I| \leq 8.$$

It is straightforward to apply Theorem 4.2 to this case by modifying the the right hand side of the sixth constraint in (4.11) from $\prod_{i\in I} p_i$ to q_I. The distribution $\mathbb{Q}$ was generated such that it satisfies:

$$\begin{aligned} q_I &\geq \alpha_2 \min\ (p_i, p_j)\,, \quad \forall I = \{i,j\} \subseteq [N] : |I| = 2 \\ q_I &\geq \alpha_{|I|} \max_{J\subseteq I} q_J, \qquad \forall I \subseteq [N] : 3 \leq |I| \leq 8 \end{aligned}$$

where $\boldsymbol{\alpha} = [1, 0.165, 0.4, 1.2, 1.11, 1.13, 1.15, 1.16]$ is a suitably chosen scaling vector. The above con-

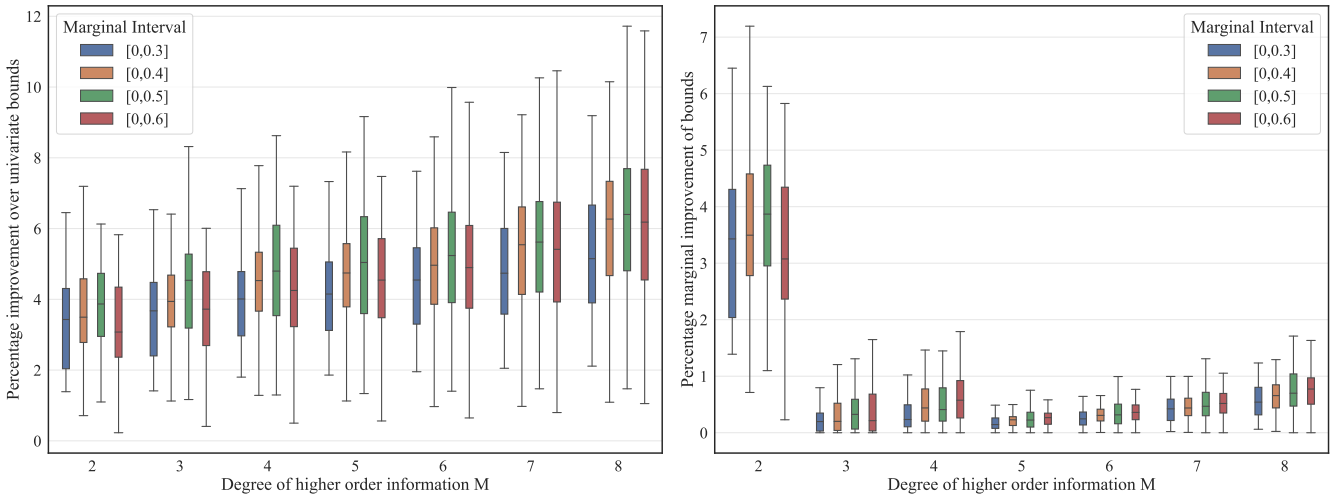


Figure 4: Box plots of univariate (left) and marginal (right) percentage improvements for $N = M = 8$.

ditions were incorporated into the linear program with the q_I as decision variables to induce new facet defining constraints (and thus preserve non-triviality of the bounds) every time M is increased

(progressively higher orders of marginal information are available). For the objective, we considered $f(\boldsymbol{\xi}) = \max_{k\in[K]} (\boldsymbol{a}_k'\boldsymbol{\xi} + b_k)$, where the number of pieces K was fixed at N and for each $k \in [K]$, $\boldsymbol{a}_k$ and b_k were randomly generated in $[-1, 1]^N$ and $[-1, 1]$. For each of 100 randomly generated instances of $\boldsymbol{a}_k$, b_k and the marginal probability vector $\boldsymbol{p}$ (in each of the hypercubes $[0, a]^N$, where $a \in \{0.3, 0.4, 0.5, 0.6\}$), M was varied from 1 to 8 and $f^\star$ was computed from (4.11). For each instance, the distribution $\mathbb{Q}$ was generated once for $M = 8$ and used for all other $M \in [7]$. Figure 4 (left) shows box plots of the percentage improvements over the univariate tight bound ($M = 1$). The best improvements in terms of the median and upper quartile are observed for $\boldsymbol{p} \in [0, 0.5]^N$. Figure 4 (right) shows box plots for the percentage marginal reduction in bounds *i.e.* $[(f^\star(M) - f^\star(M + 1)) \times 100] / f^\star(M)$. In this example, introducing bivariate marginal information adds most value in terms of percentage reduction from the univariate bound. The value addition diminishes beyond $M = 2$ with the introduction of higher order marginal information.

Under the setting described above, Table 1 showcases an instance with small marginal probabilities $\boldsymbol{p} \in [0, 0.1]^N$, where the tight bound $f^\star$ continuously improves up to $M = 8$ for three different objective functions. The three functions are structured submodular functions of the form $f(\boldsymbol{\xi}) = \min(\sum_{i\in[N]} \xi_i, B),\ B \in [3]$.

$f(\boldsymbol{\xi})$	Univariate	$M = 2$	$M = 3$	$M = 4$	$M = 5$	$M = 6$	$M = 7$	$M = 8$
$\min(\sum_{i\in[N]} \xi_i, 1)$	0.31199	0.30173	0.26809	0.26444	0.26277	0.26059	0.25689	0.25290
$\min(\sum_{i\in[N]} \xi_i, 2)$	0.31199	0.31067	0.29457	0.28172	0.27839	0.27402	0.26833	0.26134
$\min(\sum_{i\in[N]} \xi_i, 3)$	0.31199	0.31199	0.30208	0.28677	0.28399	0.28035	0.27561	0.26978

Table 1: Tight bounds for different objective functions and increasing order of marginal information with $N = 8$ random variables satisfying concordance order.

We next highlight the improved computational performance of the compact linear program in (4.11) with $f(\boldsymbol{\xi}) = \min(\sum_{i\in[N]} \xi_i, 1)$, by comparing our results with the large sized linear program in (3.3). We vary the number of random variables from $N = 5$ to $N = 13$ while keeping M fixed at 5. The box plots in Figure 5 (left) show the variation in execution time of both linear programs (in seconds shown on log scale) computed over 10 instances of randomly generated small marginal probabilities (such that $\sum_{i\in[N]} p_i \leq 1$). As N increases beyond 7, the compact linear program clearly runs much faster than the large-sized linear program. Given the computing power at our disposal and using a time limit of 700 seconds, we were able to solve the compact linear programs up to size $N = 23$ while for the large-sized linear program we were able to solve it up to $N = 14$ and for M up to 5.

6.2 Analysis of discrete moment bounds

We will now analyse the quality of bounds obtained by assuming higher order univariate moment information as provided in Theorem 5.1. We take 100 randomly generated instances of the maximum of affine functions $f(\boldsymbol{\xi})$ where the pieces are constructed by $\boldsymbol{a}_k \in [-5, 5]^N$ and $b_k \in [-2, 2]$. The number of random variables $N = 5$ and number of pieces $K = 3$. The support for the random variables was fixed as $\Xi_i = \{-5, -2, 0, 3, 6, 8, 11, 14, 17, 20\}$ for each i. The univariate probability distribution for each

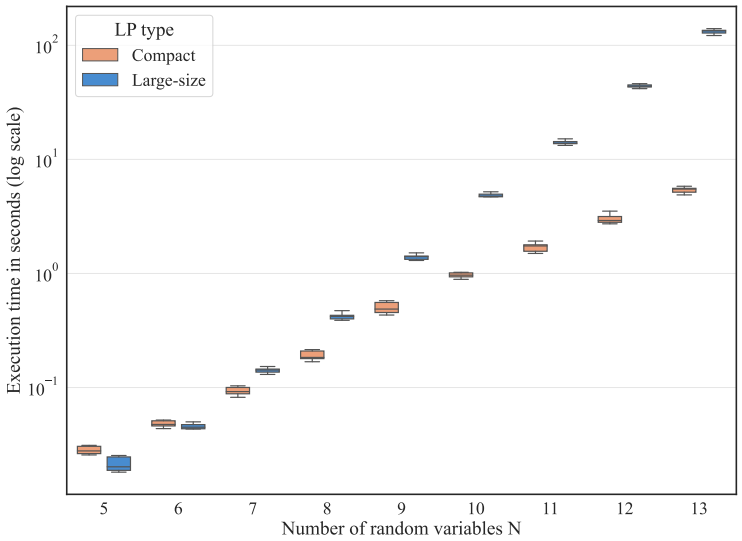


Figure 5: Computational time (of the compact and large-sized formulations for varying N (shown in log scale).

random variable $\tilde{\xi}_i$ was taken as a sample drawn from a Dirichlet distribution with all ten parameters set to 2. The distribution thus generated was used to set the univariate moments $m_{i,l}$ as $\mathbb{E}[\tilde{\xi}_i^l]$. The cross moment Q was fixed as $\mathbb{E}[\tilde{\boldsymbol{\xi}}\tilde{\boldsymbol{\xi}}']$ evaluated via pairwise independence on the generated univariate distributions. The number of univariate moments L used was varied from 1 to 15 and f^* was computed using the linear program (5.2) for each value of L.

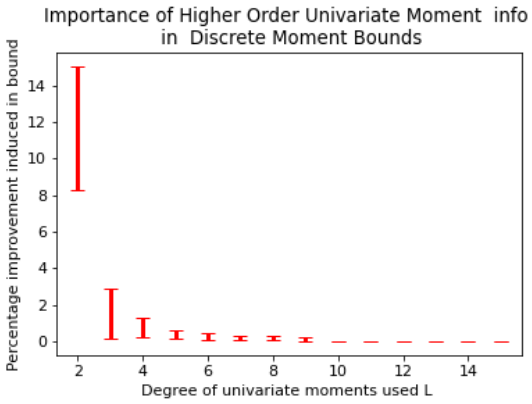


Figure 6: Value of higher order univariate moment.

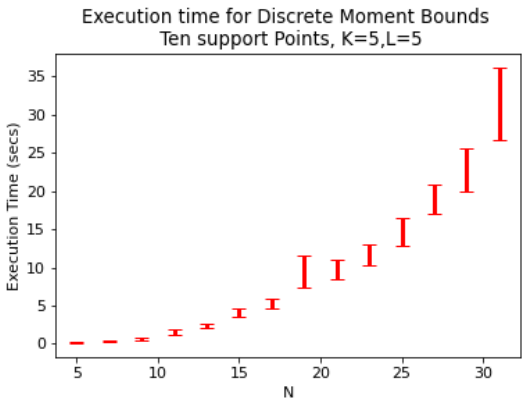


Figure 7: Execution times of (5.2).

Figure 6 shows the range of percentage marginal reduction in the bound f^* for various values of L starting from $L = 2$. For example at $L = 3$, we see the improvement obtained by using the third univariate moments over using the first two univariate moments and the cross-moment Q. The largest marginal improvement is obtained when the second univariate information is incorporated and the

improvement gradually reduces. After $L = 10$, we see negligible improvement thus showing that the first nine univariate moments aid in improvement of the bound. Similar trend was observed for other cases of support as well.

We also noted the execution times of computing higher order moments for various values of N, for the case of each random variable taking support in a set of ten discrete values. We fixed $L = 5$ and the number of pieces $K = 5$. The execution times over 10 random instances for each N is provided in Figure 7. For $N = 15$ random variables, the linear program is solved in about 4 seconds while for $N = 30$ random variables, the time taken is about 35 seconds. Clearly the proposed approach is tractable and scales well with an increase in the number of random variables.

7 Conclusion

In this paper, we provide an unified approach that helps solve instances of the multimarginal optimal transport problem, the generalized moment problem and distributionally robust optimization in polynomial time. While these problems are in general hard to solve with discrete random variables, by viewing the problems through the lens of submodularity, we gain tractability. Using ideas from submodular function minimization, we identify a general ambiguity set with discrete uncertainty which helps extend the tractability results beyond uncertainty defined over convex supports. To the best of our knowledge, this has not been exploited as yet in these settings. As we show, there are new instances of sharp bounds that can be computed efficiently using compact linear programs. Three natural follow up research ideas that arise from our work are: (i) to use specialized submodular function minimization techniques for solving the problems, (ii) to solve the problems with submodular functions defined over continuous uncertainty and (iii) to explore the use of these bounds in specific applications. We leave this for future research.

Appendix

Proof of Theorem 3.2

The distributionally robust optimization problem in (1.1) can be reformulated as:

$$\begin{array}{lll} \inf & y_0 + \sum\limits_{j\in[J]} y_j\gamma_j & \\ \text{s.t.} & y_0 + \sum\limits_{j\in[J]} y_j f_j(\boldsymbol{\xi}) \geq \max\limits_{k\in[K]} \left(\boldsymbol{a}_k(\boldsymbol{x})'\boldsymbol{\xi} + b_k(\boldsymbol{x})\right), & \forall \boldsymbol{\xi} \in \Xi, \\ & y_j \geq 0, & \forall j \in [J], \\ & \boldsymbol{x} \in \mathcal{X}, & \end{array} \tag{7.1}$$

where the decision variables are y_0, y_j for $j \in [J]$ and $\boldsymbol{x}$. The separation problem for this formulation program is given by:

Given y_0, $y_j \geq 0$ for $j \in [J]$ and $\boldsymbol{x}$, decide whether

$$y_0 + \sum_{j\in[J]} y_j f_j(\boldsymbol{\xi}) \geq \max_{k\in[K]} \left(\boldsymbol{a}_k'(\boldsymbol{x})\boldsymbol{\xi} + b_k(\boldsymbol{x})\right), \forall \boldsymbol{\xi} \in \Xi, \text{ and } \boldsymbol{x} \in \mathcal{X},$$

and if the answer is no, return a violated inequality.
Verifying $\boldsymbol{x} \in \mathcal{X}$ and if it is not, providing a separating hyperplane can be done efficiently for $\mathcal{X}$ with an efficient separation oracle. The separation problem hence reduces to:
Given y_0, $y_j \geq 0$ for $j \in [J]$ and $\boldsymbol{x} \in \mathcal{X}$, decide whether

$$y_0 - b_k(\boldsymbol{x}) + \min_{\boldsymbol{\xi}\in\Xi} \left(\sum_{j\in[J]} y_j f_j(\boldsymbol{\xi}) - \boldsymbol{a}_k'(\boldsymbol{x})\boldsymbol{\xi} \right) \geq 0, \forall k \in [K].$$

Again, we need to solve a set of K submodular function minimization problems of the form in (2.3) to solve the separation problem. Specifically, if all left hand side values above are nonnegative, we are done. Else we find a $k^* \in [K]$ and $\boldsymbol{\xi}^* \in \Xi$ such that:

$$y_0 - b_{k^*}(\boldsymbol{x}) + \sum_{j\in[J]} y_j f_j(\boldsymbol{\xi}^*) - \boldsymbol{a}_{k^*}'(\boldsymbol{x})\boldsymbol{\xi}^* < 0.$$

Since $\boldsymbol{a}_k(\boldsymbol{x})$ and $b_k(\boldsymbol{x})$ are affine in $\boldsymbol{x}$, this identifies a violated linear inequality. Since all steps can be done in polynomial time, the distributionally robust optimization problem is solvable in polynomial time.

Proof of Theorem 3.3
The sharp bound is given the optimal value of the following dual linear program:

$$\begin{array}{lll} \min & y_0 + \sum_{j\in[J]} y_j\gamma_j & \\ \text{s.t.} & y_0 + \sum_{j\in[J]} y_j f_j(\boldsymbol{\xi}) \geq \max_{k\in[K]} g_k(\boldsymbol{\xi}), & \forall \boldsymbol{\xi} \in \Xi, \\ & y_j \geq 0, & \forall j \in [J], \end{array}$$

The separation problem for the dual linear program is given by:
Given numbers y_0 and $y_j \geq 0$ for all $j \in [J]$, decide whether

$$y_0 + \sum_{j\in[J]} y_j f_j(\boldsymbol{\xi}) \geq \max_{k\in[K]} g_k(\boldsymbol{\xi}), \forall \boldsymbol{\xi} \in \Xi,$$

and if the answer is no, return a violated inequality.
This reduces to checking if

$$y_0 - b_k + \min_{\boldsymbol{\xi}\in\Xi} \left(\sum_{j\in[J]} y_j f_j(\boldsymbol{\xi}) - g_k(\boldsymbol{\xi}) \right) \geq 0, \forall k \in [K].$$

Since $y_j \geq 0$ for all $j \in [J]$, f_j is a submodular function for each $j \in [J]$ and $g_k(\boldsymbol{\xi})$ is a supermodular

function for each $k \in [K]$, the function $\sum_{j\in[J]} y_j f_j(\boldsymbol{\xi}) - \boldsymbol{g}_k'\boldsymbol{\xi}$ is submodular for each $k \in [K]$. Since we need to solve a set of K submodular function minimization problems of the form in (2.3) to solve the separation problem, the bound is computable in polynomial time.

Proof of Theorem 4.2

Given a Boolean random vector $\tilde{\boldsymbol{\xi}}$ with distribution $\mathbb{P}$, let $k(\boldsymbol{\xi})$ be a measurable selection on the optimal index set $K(\boldsymbol{\xi})$. Define the decision variables as:

$$\begin{aligned}
\lambda_k = &\quad \mathbb{P}\left(k(\tilde{\boldsymbol{\xi}}) = k\right), && \forall k \in [K],\\
\gamma_{i,k} = &\quad \mathbb{P}\left(\tilde{\xi}_i = 1, k(\tilde{\boldsymbol{\xi}}) = k\right), && \forall i \in [N], \forall k \in [K],\\
\gamma_{I,k} = &\quad \mathbb{P}\left(\prod_{i\in I} \tilde{\xi}_i = 1, k(\tilde{\boldsymbol{\xi}}) = k\right), && \forall I \subseteq [N] : 1 < |I| \leq M, \forall k \in [K].
\end{aligned}$$

The first three constraints are identical to the first three constraints in the formulation in Theorem 4.1. The fourth constraint $\gamma_{I,k} \leq \lambda_k$ arises from:

$$\mathbb{P}\left(\prod_{i\in I} \tilde{\xi}_i = 1, k(\tilde{\boldsymbol{\xi}}) = k\right) \leq \mathbb{P}\left(k(\tilde{\boldsymbol{\xi}}) = k\right),$$

and the fifth constraint $\gamma_{I,k} \leq \gamma_{i,k}$ comes from:

$$\mathbb{P}\left(\prod_{i\in I} \tilde{\xi}_i = 1, k(\tilde{\boldsymbol{\xi}}) = k\right) \leq \mathbb{P}\left(\tilde{\xi}_i = 1, k(\tilde{\boldsymbol{\xi}}) = k\right), \text{ for } i \in I.$$

The sixth constraint is from the PUOD condition since:

$$\sum_{k\in[K]} \mathbb{P}\left(\prod_{i\in I} \tilde{\xi}_i = 1, k(\tilde{\boldsymbol{\xi}}) = k\right) \geq \prod_{i\in I} \mathbb{P}\left(\tilde{\xi}_i = 1\right).$$

Necessity of the formulation then follows. For sufficiency, create a mixture distribution $\mathbb{P}^*$ as follows:
(i) Generate a discrete random variable $\tilde{z}$ that takes values in $[K]$ with $\mathbb{P}^*(\tilde{z} = k) = \lambda_k^*$.
(ii) Conditional on the realization of $\tilde{z}$, define the marginal distribution of each Bernoulli random variable $\tilde{\xi}_i$ as:

$$\mathbb{P}^*\left(\tilde{\xi}_i = 1 \middle| \tilde{z} = k\right) = \frac{\gamma_{i,k}^*}{\lambda_k^*} = 1 - \mathbb{P}^*\left(\tilde{\xi}_i = 0 \middle| \tilde{z} = k\right).$$

Generate in step (ii), a comonotonic random vector using these conditional marginal distributions. The proof of tightness follows from steps similar to Theorem 4.1 and we leave the reader to verify the steps.

Proof of Corollary 4.1

We can express the function value with $\boldsymbol{\xi} \in \{0,1\}^N$ as:

$$f(\boldsymbol{\xi}) = \min\left(\sum_{i\in[N]} \xi_i, B\right) = \max_{I\subseteq N:|I|\leq B} \sum_{i\in I} \xi_i.$$

The expression on the right hand side is the maximum of affine functions in $\boldsymbol{\xi}$. When B is a constant that is independent of N, the number of linear pieces is fixed and the linear program is of polynomial size from Theorem 4.2.

Proof of Theorem 5.1

Let f_u^* be optimal value of the linear program (5.2).

Step (1): $f^* \leq f_u^*$

Let $k(\boldsymbol{\xi})$ be a measurable selection on the set $K(\boldsymbol{\xi}) = \arg\max\{\boldsymbol{a}_k'\boldsymbol{\xi} + b_k \mid k \in [K]\}$. Define the decision variables as:

$$\begin{aligned}
\lambda_k = &\quad \mathbb{P}\left(k(\tilde{\boldsymbol{\xi}}) = k\right), && \forall k \in [K],\\
\gamma_{i,k}(\xi_i) = &\quad \mathbb{P}\left(\tilde{\xi}_i = \xi_i, k(\tilde{\boldsymbol{\xi}}) = k\right), && \forall \xi_i \in \Xi_i, \forall i \in [N], \forall k \in [K],\\
\gamma_{i,j,k}(\xi_i,\xi_j) = &\quad \mathbb{P}\left(\tilde{\xi}_i = \xi_i, \tilde{\xi}_j = \xi_j, k(\tilde{\boldsymbol{\xi}}) = k\right), && \forall \xi_i \in \Xi_i, \forall \xi_j \in \Xi_j, \forall i < j \in [N], \forall k \in [K].
\end{aligned}$$

The variables must satisfy the nonnegativity constraints. The other constraints in the formulation are obtained from necessary conditions that the variables must satisfy:

1. Total sum of the probabilities of indices being optimal is one:

$$\sum_{k\in[K]} \mathbb{P}\left(k(\tilde{\boldsymbol{\xi}}) = k\right) = 1.$$

2. Law of total expectation for the marginal moments:

$$\sum_{k\in[K]} \sum_{\xi_i\in\Xi_i} \xi_i^l \mathbb{P}\left(\tilde{\xi}_i = \xi_i, k(\tilde{\boldsymbol{\xi}}) = k\right) = \mathbb{E}_{\mathbb{P}}\left[\tilde{\xi}_i^l\right].$$

3. Law of total probability for the index being optimal:

$$\sum_{\xi_i\in\Xi_i} \mathbb{P}\left(\tilde{\xi}_i = \xi_i, k(\tilde{\boldsymbol{\xi}}) = k\right) = \mathbb{P}\left(k(\tilde{\boldsymbol{\xi}}) = k\right).$$

4. Consistency of the conditional bivariate marginals with the conditional univariate marginals:

$$\begin{aligned}
\sum_{\xi_i\in\Xi_i} \mathbb{P}\left(\tilde{\xi}_i = \xi_i, \tilde{\xi}_j = \xi_j, k(\tilde{\boldsymbol{\xi}}) = k\right) &= \mathbb{P}\left(\tilde{\xi}_j = \xi_j, k(\tilde{\boldsymbol{\xi}}) = k\right),\\
\sum_{\xi_j\in\Xi_j} \mathbb{P}\left(\tilde{\xi}_i = \xi_i, \tilde{\xi}_j = \xi_j, k(\tilde{\boldsymbol{\xi}}) = k\right) &= \mathbb{P}\left(\tilde{\xi}_i = \xi_i, k(\tilde{\boldsymbol{\xi}}) = k\right).
\end{aligned}$$

5. Law of total expectation for the lower bound on the cross moment:

$$\sum_{k\in[K]}\sum_{\xi_i\in\Xi_i}\sum_{\xi_j\in\Xi_i}\xi_i\xi_j\mathbb{P}\left(\tilde{\xi}_i=\xi_i,\tilde{\xi}_j=\xi_j,k(\tilde{\boldsymbol{\xi}})=k\right)\geq\mathbb{E}_{\mathbb{P}}\left[\tilde{\xi}_i\tilde{\xi}_j\right].$$

The objective function is obtained as in the proof of Theorem 4.1. From the necessity of all the constraints, we have $f^*\leq f_u^*$.

Step (2): $f^*\geq f_u^*$

We construct a distribution $\mathbb{P}^*\in\mathcal{P}$ that attains the upper bound f_u^* using the optimal solution $(\boldsymbol{\lambda}^*,\boldsymbol{\gamma}^*)$. Create a mixture distribution $\mathbb{P}^*$ as follows:

(i) Generate a discrete random variable $\tilde{z}$ that takes values in $[K]$ with probability $\mathbb{P}^*(\tilde{z}=k)=\lambda_k^*$.

(ii) Conditional on the realization of $\tilde{z}$, define the marginal distribution of each random variable $\tilde{\xi}_i$ as:

$$\mathbb{P}^*\left(\tilde{\xi}_i=\xi_i\big|\tilde{z}=k\right)=\frac{\gamma_{i,k}^*(\xi_i)}{\sum_{\xi\in\Xi_i}\gamma_{i,k}^*(\xi)},\forall\xi_i\in\Xi_i.$$

Generate in step (ii), a comonotonic random vector using these conditional marginal distributions. The marginal moments of $\tilde{\xi}_i$ in the mixture distribution $\mathbb{P}^*$ is given by:

$$\begin{aligned}\mathbb{E}_{\mathbb{P}^*}\left[\tilde{\xi}_i^l\right] &= \sum_{k\in[K]}\lambda_k^*\mathbb{E}_{\mathbb{P}^*}\left[\tilde{\xi}_i^l\big|\tilde{z}=k\right],\\ &= \sum_{k\in[K]}\lambda_k^*\sum_{\xi_i\in\Xi_i}\left(\xi_i^l\frac{\gamma_{i,k}^*(\xi_i)}{\sum_{\xi\in\Xi_i}\gamma_{i,k}^*(\xi)}\right),\\ &= m_{i,l},\\ &\quad[\text{since }\textstyle\sum_{\xi\in\Xi_i}\gamma_{i,k}^*(\xi)=\lambda_k^*\text{ and }\sum_{k\in[K]}\sum_{\xi_i\in\Xi_i}\xi_i^l\gamma_{i,k}^*(\xi_i)=m_{i,l}].\end{aligned}$$

Hence the marginal moments of $\mathbb{P}^*$ match the marginal moments specified in $\mathcal{P}$. Let $\mathbb{Q}_{i,j}^*$ denote the distribution of the random variables $(\tilde{\xi}_i,\tilde{\xi}_j)$ defined by:

$$\mathbb{Q}_{i,j}^*\left(\tilde{\xi}_i=\xi_i,\tilde{\xi}_j=\xi_j\big|\tilde{z}=k\right)=\frac{\gamma_{i,j,k}^*(\xi_i,\xi_j)}{\sum_{\xi\in\Xi_i}\sum_{\eta\in\Xi_j}\gamma_{i,j,k}^*(\xi,\eta)},\forall\xi_i\in\Xi_i,\forall\xi_j\in\Xi_j,\forall k\in[K].$$

From the feasibility conditions, we see that $\text{proj}_i(\mathbb{Q}_{i,j|k}^*)=\text{proj}_i(\mathbb{P}_{|k}^*)$ and $\text{proj}_j(\mathbb{Q}_{i,j|k}^*)=\text{proj}_j(\mathbb{P}_{|k}^*)$ where $|k$ denotes conditional on $\tilde{z}=k$. This implies the existence of a conditional bivariate distribution for $(\tilde{\xi}_i,\tilde{\xi}_j)$ consistent with the conditional marginal distributions of $\tilde{\xi}_i$ and $\tilde{\xi}_j$ for each k. The cross

moment of $\tilde{\xi}_i$ and $\tilde{\xi}_j$ in $\mathbb{P}^*$ is then given by:

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}^*}\left[\tilde{\xi}_i\tilde{\xi}_j\right] &= \sum_{k\in[K]} \lambda_k^* \mathbb{E}_{\mathbb{P}^*}\left[\tilde{\xi}_i\tilde{\xi}_j \middle| \tilde{z}=k\right], \\
&\geq \sum_{k\in[K]} \lambda_k^* \mathbb{E}_{\mathbb{Q}_{i,j}^*}\left[\tilde{\xi}_i\tilde{\xi}_j \middle| \tilde{z}=k\right], \\
&\quad \text{[from (2.7) since } \xi_i\xi_j \text{ is supermodular with } \mathbb{P}_{|k}^* \text{ and } \mathbb{Q}_{i,j|k}^* \text{ having the same marginals]}, \\
&= \sum_{k\in[K]} \lambda_k^* \sum_{\xi_i\in\Xi_i}\sum_{\xi_j\in\Xi_j} \left(\xi_i\xi_j \frac{\gamma_{i,j,k}^*(\xi_i,\xi_j)}{\sum_{\xi\in\Xi_i}\sum_{\eta\in\Xi_j}\gamma_{i,j,k}^*(\xi,\eta)}\right), \\
&= \sum_{k\in[K]}\sum_{\xi_i\in\Xi_i}\sum_{\xi_j\in\Xi_j} \xi_i\xi_j\gamma_{i,j,k}^*(\xi_i,\xi_j), \\
&\quad \text{[since } \sum_{\xi\in\Xi_i}\sum_{\eta\in\Xi_j}\gamma_{i,j,k}^*(\xi,\eta) = \lambda_k^*\text{]}, \\
&\geq Q_{i,j}.
\end{aligned}$$

Hence $\mathbb{P}^* \in \mathcal{P}$. The final step is to show the sharpness of the bound under this distribution. This follows from steps identical to the proof of 4.1. Hence $f^* = f_u^*$.

References

[1] S. Agrawal, Y. Ding, A. Saberi, Y. Ye. Price of correlations in stochastic optimization. Operations Research, 60(1), 150-162, 2012.

[2] J. M. Altschuler, E. Boix-Adsera. Polynomial-time algorithms for multimarginal optimal transport problems with structure. Mathematical Programming, 199, 1107–1178, 2023.

[3] B. Axelrod, Y. P. Liu, A. Sidford. Near-optimal approximate discrete and continuous submodular function minimization. Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms, 837–853, 2020.

[4] F. Bach. Submodular functions: from discrete to continuous domains. Mathematical Programming, 175, 419-459, 2019.

[5] F. Bach. Learning with Submodular Functions: A Convex Optimization Perspective. Foundations and Trends in Machine Learning, 6(2-3), 145-373, 2013.

[6] G. Bayraksan, D. K. Love. Data-driven stochastic programming using phi-divergences. Tutorials in Operations Research. The Operations Research Revolution, 1-19, 2015.

[7] A. Ben-Tal, D. den Hertog, A. De Waegenaere, B. Melenberg, G. Rennen. Robust solutions of optimization problems affected by uncertain probabilities. Management Science, 59(2), 341-357, 2013.

[8] A. Ben-Tal, L. El Ghaoui, A. Nemirovski. Robust Optimization. Princeton University Press, 2009.

[9] D. Bertsimas, D. Brown, C. Caramanis. Theory and applications of robust optimization. SIAM Review, 53(3), 464-501, 2011.

[10] D. Bertsimas, D. Den Hertog. Robust and Adaptive Optimization. Dynamic Ideas, Belmont Massachusetts, 2022.

[11] D. Bertsimas, X. V. Doan, K. Natarajan, C-P. Teo. Models for minimax stochastic linear optimization problems with risk aversion. Mathematics of Operations Research, 35(3), 580-602, 2010.

[12] D. Bertsimas, K. Natarajan, C-P. Teo. Probabilistic combinatorial optimization: Moments, semidefinite programming, and asymptotic bounds. SIAM Journal on Optimization, 15(1), 185-209.

[13] D. Bertsimas, K Natarajan, C-P. Teo. Persistence in discrete optimization under data uncertainty. Mathematical Programming, 108, 251-274, 2006.

[14] D. Bertsimas, I. Popescu. Optimal inequalities in probability theory: A convex optimization approach. SIAM Journal on Optimization, 15(3), 780-804, 2005.

[15] J. Bilmes. Submodularity in machine learning and artificial intelligence. Arxiv, abs/2202.00132, 2022.

[16] J Blanchet, K. Murthy. Quantifying distributional model risk via optimal transport. Mathematics of Operations Research, 44(2), 565-600, 2019.

[17] L. Chen, W. Ma, K. Natarajan, D Simchi-Levi, Z Yan. Distributionally robust linear and discrete optimization with marginals. Operations Research, 70(3), 1822-1834, 2022.

[18] W. Chen, M. Sim, J. Sun, C-P. Teo. From CVaR to uncertainty set: Implications in joint chance-constrained optimization. Operations Research, 58(2), 470-485, 2010.

[19] X. Chen, S. He, B. Jiang, C. T. Ryan, T. Zhang. The discrete moment problem with nonconvex shape constraints. Operations Research, 69(1), 279-296, 2021.

[20] X. Chen, M. Li. Discrete convex analysis and its applications in operations: A survey. Production and Operations Management, 30(6), 1904-1926, 2021.

[21] Z. Chen, M. Sim and P. Xiong. Robust stochastic optimization made easy with ROME. Operations Research, 66(8), 3329-3339, 2020.

[22] G. Choquet. Theory of capacities. Annales de l'Institut Fourier, 5, 131–295, 1953.

[23] E. Delage, Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. Operations Research, 58(3), 595–612, 2010.

[24] D. Dentcheva, A. Ruszczynski. Optimization with multivariate stochastic dominance constraints. Mathematical Programming, 117, 111–127, 2009.

[25] J. Dhaene, M. Denuit, M. J. Goovaertsm R. Kaasm D. Vyncke. The concept of comonotonicity in actuarial science and finance: theory. Insurance: Mathematics and Economics, 31(1), 3-33, 2002.

[26] J. Dhaene, M. Denuit, M. J. Goovaerts, R. Kaas, D. Vyncke. The concept of comonotonicity in actuarial science and finance: applications. Insurance: Mathematics and Economics, 31(2), 133-161, 2002.

[27] A. Dhara, B. Das, K. Natarajan. Worst-case expected shortfall with univariate and bivariate marginals. INFORMS Journal on Computing, 33(1), 370-389, 2021.

[28] X. V. Doan, K. Natarajan. On the complexity of nonoverlapping multivariate marginal bounds for probabilistic combinatorial optimization problems. Operations Research, 60(1), 138-149, 2012.

[29] X. V. Doan, X. Li, K. Natarajan. Robustness to dependency in portfolio optimization using overlapping marginals. Operations Research, 63(6), 1468-1488, 2015.

[30] R. Dyckerhoff, K. Mosler. Orthant orderings of discrete random vectors. Journal of Statistical Planning and Inference, 62(2), 193-205, 1997.

[31] M. Dyer, L. Stougie. Computational complexity of stochastic programming problems. Mathematical Programming 106, 423–432, 2006.

[32] J. Edmonds. Submodular functions, matroids, and certain polyhedra in: Combinatorial structures and their applications (R. Guy, H. Hanani, N. Sauer, J. Schönmein eds.), Gordon and Breach, New York, 69-87, 1970.

[33] A. Ene, H. Nguyen, L. A Vegh. Decomposable submodular function minimization: discrete and continuous. In Advances in Neural Information Processing Systems 17, 2870–2880, 2017.

[34] P. M. Esfahani, D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. Mathematical Programming, 171, 115–166, 2018.

[35] S. Fujishige. Submodular Functions and Optimization, Elsevier, 2005.

[36] S. Fujishige. Lexicographically optimal base of a polymatroid with respect to a weight vector. Mathematics of Operations Research, 5, 186–196, 1980.

[37] R Gao, A Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. Mathematics of Operations Research. To appear, 2023.

[38] G. Georgakopoulos, D. Kavvadias, C. H. Papadimitriou. Probabilistic satisfiability. Journal of Complexity, 4(1), 1-11, 1988.

[39] J. Goh, M. Sim. Distributionally robust optimization and its tractable approximations, Operations Research, 58(4), 902-917, 2010.

[40] M. Grötschel, M., L. Lovász, A. Schrijver. The ellipsoid method and its consequences in combinatorial optimization. Combinatorica, 1(2), 169–197, 1981.

[41] M. Grötschel, M., L. Lovász, A. Schrijver. Geometric Algorithms and Combinatorial Optimization. Springer Heidelberg, 1988.

[42] G. A. Hanasusanto, D. Kuhn, W. Wiesemann, A comment on "computational complexity of stochastic programming problems". Mathematical Programming, 159, 557–569, 2016.

[43] D. Hunter. An upper bound for the probability of a union. Journal of Applied Probability. 13(3), 597-603, 1976.

[44] P. Honeyman, R. E. Lander, M. Yannakakis. Testing the universal instance assumption. Information Processing Letters, 10(1), 14-19, 1980.

[45] D. Iancu, M. Sharma, M. Sviridenko. Supermodularity and affine policies in dynamic robust optimization. Operations Research, 61(4), 941-956.

[46] S. Iwata, L. Fleischer, S. Fujishige. A combinatorial strongly polynomial algorithm for minimizing submodular functions. Journal of the ACM, 48(4), 761-777, 2001.

[47] H. Jiang. Minimizing convex functions with rational minimizers. Journal of the ACM, 70(1), 1–27, 2022.

[48] S. Jegelka, F. Bach, S. Sra. Reflection methods for user-friendly submodular optimization. In Advances in Neural Information Processing Systems, 13, 1313–1321, 2013

[49] H. Joe, Multivariate concordance. Journal of Multivariate Analysis, 35, 12-30, 1990.

[50] L. G. Khachiyan. A polynomial algorithm in linear programming, Doklady Akademii Nauk SSSR 244, 1093-1096 (English translation: Soviet Math. Dokl. 20, 191-194), 1979.

[51] V. Kolmogorov. Minimizing a sum of submodular functions. Discrete Applied Mathematics, 160(15), 2246-2258, 2012.

[52] J. B Lasserre. Moments, Positive Polynomials and Their Applications. World Scientific, 2009.

[53] Y. T. Lee, A. Sidford, S. S. Vempala. Efficient convex optimization with membership oracles. Proceedings of the 31st Conference On Learning Theory, PMLR 75, 1292-1294, 2018.

[54] Y. T. Lee, A. Sidford, S-C. Wong. A faster cutting plane method and its implications for combinatorial and convex optimization. IEEE 56th Annual Symposium on Foundations of Computer Science, 299, 1049-1065, 2015.

[55] E. L. Lehman. Some concepts of dependence. The Annals of Mathematical Statistics, 37(5), 1137-1153, 1966.

[56] D. Z. Long, J. Qin, A. Zhang. Supermodularity in two-stage distributional robust optimization, To appear in Operations Research, 2023.

[57] L. Lovász, Submodular functions and convexity. Mathematical Programming The State of the Art, edited by A. Bachem and B. Korte and M. Grötschel, 235-257, 1983.

[58] G. Mádi-Nagy, A. A. Prékopa. On multivariate discrete moment problems and their applications to bounding expectations and probabilities. Mathematics of Operations Research, 29(2), 229-258, 2004.

[59] H-Y. Mak, Y. Rong, J. Zhang. Appointment scheduling with limited distributional information. Management Science, 61(2), 316–334, 2015.

[60] H-Y. Mak, Z-J. M. Shen. Pooling and dependence of demand and yield in multiple-location inventory systems. Manufacturing and Service Operations Management, 16(2): 263–269, 2014.

[61] W. Maurer. Bivalent trees and forests or upper bounds for the probability of a union revisited. Discrete Applied Mathematics, 6 (2), 157-171, 1983.

[62] E. Boros, A. Scozzari, F. Tardella, P. Veneziani. Polynomially computable bounds for the probability of the union of events. Mathematics of Operations Research, 39(4), 1311–1329, 2014.

[63] A. Müller, M. Scarsini. Some remarks on the supermodular order. Journal of Multivariate Analysis, 73, 107-119, 2000.

[64] A. Muller, M. Scarsini, I. Tsetlin, R. L. Winkler. Multivariate almost stochastic dominance: Transfer characterizations and sufficient conditions under dependence uncertainty. To appear in Operations Research, 2023.

[65] A. Müller, D. Stoyan. Comparison Methods for Stochastic Models and Risks, Wiley, Chichester, 2002.

[66] K. Natarajan. Optimization with Marginals and Moments. Dynamic Ideas, Belmont Massachusetts, 2021.

[67] J. B. Orlin. A faster strongly polynomial time algorithm for submodular function minimization. Mathematical Programming Series A, 118, 237–251, 2009.

[68] B. Pass. Multi-marginal optimal transport: theory and applications. ESAIM: Mathematical Modelling and Numerical Analysis, 49, 1771-1790, 2015.

[69] C. Peng, E. Delage. Data-driven optimization with distributionally robust second order stochastic dominance constraints, Operations Research, Article in Advance, 2022.

[70] I. Popescu. Applications of optimization in probability, finance and revenue management, PhD Thesis, Massachusetts Institute of Technology, 1999.

[71] K. Postek , A. Ben-Tal, D. den Hertog, B. Melenberg. Robust optimization with ambiguous stochastic constraints under mean and dispersion information. Operations Research 66(3), 814–833, 2018.

[72] A. Prékopa. The discrete moment problem and linear programming. Discrete Applied Mathematics. 27(3), 235-254, 1990.

[73] L. Rüschendorf. Mathematical Risk Analysis: Dependence, Risk Bounds, Optimal Allocations and Portfolios, Springer Berlin, Heidelberg, 2013.

[74] A. Schrijver. A combinatorial algorithm minimizing submodular functions in strongly polynomial time. Journal of Combinatorial Theory, Series B, 80(2), 346-355, 2000

[75] M. Shaked, J. G. Shanthikumar. Stochastic Orders, Springer New York, 2006.

[76] A. Shapiro, D. Dentcheva, A. Ruszczyński. Lectures on Stochastic Programming: Modeling and Theory. MOS-SIAM Series on Optimization, SIAM, 2nd edition, 2014.

[77] P. Stobbe, A. Krause. Efficient minimization of decomposable submodular functions. In Advances in Neural Information Processing Systems, 23, 2208–2216, 2010.

[78] M. Staib, B. Wilder, S. Jegelka. Distributionally robust submodular maximization. Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics, PMLR 89, 506-516, 2019.

[79] A. H. Tchen. Inequalities for distributions with given marginals. The Annals of Probability, 8(4), 814-827, 1980.

[80] D. M. Topkis. Supermodularity and Complementarity. Princeton University Press, Princeton, New Jersey, 1998.

[81] W. Wiesemann, D. Kuhn, M. Sim. Distributionally robust convex optimization. Operations Research, 62(6), 1358-1376, 2014.

[82] K. J. Worsley. An improved Bonferroni inequality and applications. Biometrika, 69(2), 297-302, 1982.

[83] T. Yanagimoto, M. Okamoto. Partial orderings of permutations and monotonicity of a rank correlation statistic. Annals of the Institute of Statistical Mathematics, 21, 489–506, 1969.

[84] S. Zymler, D. Kuhn, B. Rustem. Distributionally robust joint chance constraints with second-order moment information. Mathematical Programming, 137, 167–198, 2013.

[85] L. Chen, D. Padmanabhan, C. Lim and K. Natarajan. Correlation robust influence maximization. Advances in Neural Information Processing Systems, 33, 7078-7089, 2020.