

Distributionally robust optimization through the lens of submodularity

Karthik Natarajan* Divya Padmanabhan† Arjun Ramachandra‡

First Draft: May 2025, Revision: July 2026

Abstract

In this paper, we propose submodular ambiguity sets for distributionally robust optimization and showcase its usefulness in modeling discrete and continuous uncertainty. From a modeling perspective, we propose two new tractable submodular ambiguity sets that combine marginal distributions and covariance information for discrete uncertainty and mean and covariance information for continuous uncertainty. From a computational perspective, for discrete uncertainty, we show that a class of distributionally robust optimization problems is solvable in polynomial time and in special cases, as a polynomial sized linear program. With continuous uncertainty, we show that it is solvable approximately up to an additive error in pseudo-polynomial time and in special cases, as a polynomial sized semidefinite program. We provide numerical evidence on the usefulness of incorporating covariance information through these ambiguity sets in resource allocation problems where fairness is of concern and limited data is available. In addition, we provide theoretical bounds in stylized settings to quantify the improvement provided by these ambiguity sets. The paper highlights that submodular ambiguity sets form the natural discrete counterpart of convex ambiguity sets and supplements it for continuous uncertainty, both in modeling and computation.

1 Introduction

Consider a distributionally robust optimization problem of the form:

$$\inf_{\mathbf{x} \in \mathcal{X}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}} \left[g(\mathbf{x}, \tilde{\boldsymbol{\xi}}) \right], \quad (1.1)$$

where the decision vector $\mathbf{x}$ is chosen from the set $\mathcal{X}$ before the true realization of the random vector $\tilde{\boldsymbol{\xi}}$ is known. The probability distribution of $\tilde{\boldsymbol{\xi}}$ is denoted by $\mathbb{P}$ and is itself ambiguous. The distribution $\mathbb{P}$ is however assumed 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 $\mathbf{x}$ and a realization of the random vector

*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

$\tilde{\xi} = \xi$ is given by the value $g(\mathbf{x}, \xi)$. In formulation (1.1), the decision $\mathbf{x} \in \mathcal{X}$ 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 the set $\mathcal{P} = \{\mathbb{P}\}$ consists of a single distribution, (1.1) reduces to a classical stochastic optimization problem and when $\mathcal{P} = \mathcal{P}(\Xi)$ 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 (see [5, 7]). Distributionally robust optimization lies between the two extremes and aims to control the conservatism of planning for a worst-case realization in robust optimization and planning for a distribution in stochastic optimization that might possibly be misspecified. Several ambiguity sets have now been proposed in the literature that include the moment ambiguity set [20, 8, 56], the marginal ambiguity set [16, 22, 23], the phi-divergence ambiguity set [6] and the Wasserstein ambiguity set [26, 12, 27]. The recent article of [36] provides an excellent review on the various ambiguity sets. Broadly, computational tractability is guaranteed when the support Ξ of the uncertainty is a convex set and upper bounds on the expected value of a finite set of convex functions of the random vector are specified in the ambiguity set. Furthermore, such convex ambiguity sets have powerful modeling capabilities. When the support Ξ is a discrete set however, the complexity of solving the distributionally robust optimization problem (1.1) typically scales with the size of the set Ξ . This in turn might be exponential in the number of random variables. An exception is the marginal or the Fréchet ambiguity set where only marginal information on the random vector is specified. The resulting formulations of the marginal ambiguity set for certain cost functions only grows linearly in the number of random variables; see chapter 2 in [44]. While the marginal ambiguity set incorporates distributional or moment information on each random variable, it does not explicitly model the dependence among the random variables. A natural question is whether there exist practically useful ambiguity sets for discrete uncertainty that allow for modeling both marginal and dependence information such as covariances while preserving computational tractability in solving distributionally optimization problems? Furthermore, can any of these ambiguity sets be extended to continuous uncertainty with modeling expressivity that goes beyond existing convex ambiguity sets? Lastly, for what types of distributionally robust optimization problems are such ambiguity sets useful and if so how much improvement does it provide over alternate approaches? We partly answer these questions in this paper. Towards this we discuss as a motivating example, a resource allocation problem next.

1.1 Resource allocation under ambiguity with fairness considerations

Consider a centralized decision-maker who needs to allocate limited resources across locations (or individuals or groups) while trying to balance fairness and efficiency considerations. Such problems have been analyzed in healthcare [32], air traffic management [9], humanitarian supply chains [43], multi-portfolio optimization [34] and artificial intelligence applications [31]. Here we consider a generic robust resource allocation where the outcomes of the allocations are random and ambiguous. Suppose there are N locations to which resources have to be allocated. Allocating x_i units to location i is associated with a random cost (or outcome) $g_i(x_i, \tilde{\xi}_i)$. Here $\tilde{\xi}_i$ is the uncertain random variable associated with location

i and $g_i(x_i, \xi_i)$ is the cost incurred at location i when allocated x_i units of the resource with a random realization $\tilde{\xi}_i = \xi_i$. For example, in a humanitarian disaster relief setting, x_i might represent the amount of resources prepositioned at location i before a disaster strikes, $\tilde{\xi}_i$ might represent the random demand for resources at location i which is known only after the disaster strikes and $g_i(x_i, \tilde{\xi}_i)$ is the random cost incurred at location i . A simple cost function is of the form $g_i(x_i, \tilde{\xi}_i) = |x_i - \tilde{\xi}_i|$. The allocations have to be chosen subject to budget constraints that are encoded in $\mathcal{X}$. In such settings, it is often difficult to have precise probability distributions that model the random demands for resources since data is limited and scarce (see [39]). We consider a distributionally robust approach to this problem where the exact probability distribution of the uncertainties affecting the outcomes is not known and modeled by an ambiguity set $\mathbb{P} \in \mathcal{P}$. A possible criterion to optimize is given by a convex combination of two objectives: (i) fairness which is measured using the Rawlsian criterion (worst cost across locations) and (ii) efficiency which is measured using the average cost across locations. The formulation is given by:

$$\inf_{\mathbf{x} \in \mathcal{X}} \left(\alpha \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}} \left[\max_{i \in [N]} g_i(x_i, \tilde{\xi}_i) \right] + (1 - \alpha) \frac{1}{N} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}} \left[\sum_{i \in [N]} g_i(x_i, \tilde{\xi}_i) \right] \right), \quad (1.2)$$

where $\alpha \in [0, 1]$ is a fixed parameter that provides a tradeoff between the worst-case expected fairness cost and the worst-case expected efficiency cost over all distributions in the ambiguity set. As α is increased from 0 to 1, the optimization problem (1.2) gives more importance to worst-case expected fairness and decreases importance to worst-case expected efficiency. By varying α , the decision-maker captures the tradeoff between fairness and efficiency and its effect on the resource allocation decisions. As we will see later, the submodular ambiguity sets we propose in this paper are useful for such applications where fairness is of concern.

1.2 Contributions

The main contributions of the paper are described next:

- (a) *Modeling uncertainty with submodular ambiguity sets:* We propose a class of submodular ambiguity sets where the support of each random variable is specified along with upper bounds (respectively lower bounds) on the expected value of a finite set of submodular (respectively supermodular) functions of the random vector. We discuss two submodular ambiguity sets that combine marginal distributions and covariance information for discrete uncertainty and mean and covariance information for continuous uncertainty and compare and contrast it with related information that is modeled by existing ambiguity sets in the literature.
- (b) *Computational tractability with discrete uncertainty:* When the support of each random variable is discrete and finite we show that the inner supremum in (1.1) is computable in polynomial time under appropriate assumptions on the cost function and the submodular ambiguity set. The result makes use of the ellipsoid method and the polynomial time solvability of submodular function minimization. In turn this helps us identify a class of polynomial time solvable distributionally robust optimization problems by viewing it through the lens of submodularity. We then focus on the

marginal-covariance submodular ambiguity set and develop a polynomial sized linear program for specific cost functions. This provides a tractable approach to combine marginal distributions and covariance information for discrete random variables in contrast to existing NP-hard formulations.

- (c) *Computational tractability with continuous uncertainty:* When the support of each random variable lies in a finite interval, under appropriate assumptions on the cost function and the submodular ambiguity set, we show that the problem is solvable approximately up to additive error in pseudo-polynomial time. We then develop an exact polynomial sized semidefinite program for the mean-covariance submodular ambiguity when the random variables are supported in the non-negative orthant. This provides a tractable approach to combine mean and covariance information for nonnegative random variables in contrast to exact NP-hard formulations when the covariance matrix is exactly specified. We compare and contrast the formulation with a popular semidefinite program that has been developed for a related tractable mean-covariance convex ambiguity set.
- (d) *Numerical experiments:* We numerically show the value of adding in covariance information with submodular ambiguity sets in resource allocation problems where fairness is of concern. The models we propose in this paper incorporate relevant covariance information for such objective functions in a computationally tractable manner and provide improvements over the marginal distribution and marginal moment ambiguity sets that neglect this information. We also compare the models with existing formulations that make use of similar information and highlight the pros and cons of the approaches.
- (e) *Value of incorporating covariance information:* We provide theoretical bounds that quantify the value of adding information on the covariance through the submodular ambiguity set in stylized settings. For the maximum of nonnegative random variables (that appears in the Rawlsian fairness criterion) with a common marginal distribution, we show that the ratio of the maximum expected value with the marginal-covariance submodular ambiguity set where the lower bounds on covariances are set to zero to the maximum expected value with the marginal distribution ambiguity set that ignores it is at least $3/4$. When the lower bounds on the covariances are arbitrarily chosen in the marginal-covariance submodular ambiguity set, this ratio can go down to zero. These results extend to the moment ambiguity sets.

The structure of the paper is discussed next. We list the mathematical notations used in the paper in Section 1.3. In Section 2, we review convex ambiguity sets. In Section 3, we define submodular ambiguity sets and discuss two specific examples with marginal-covariance and mean-covariance information. We discuss the computational tractability of the submodular ambiguity set with discrete uncertainty in Section 4 and with continuous uncertainty in Section 5. Numerical experiments are reported in Section 6. We provide theoretical bounds to quantify the value of incorporating covariance information in Section 7 before concluding in Section 8. All proofs are provided in the Appendix A with additional material provided in Appendix B.

1.3 Notations

We use the nonbold symbol such as x to denote a scalar, bold symbol such as $\mathbf{x}$ to denote a vector and bold capital symbol such as $\mathbf{X}$ to denote a matrix. Random numbers and random vectors are denoted with the tilde sign; examples are $\tilde{\xi}$ and $\tilde{\mathbf{x}}$. For a positive integer N , we use $[N]$ to denote the set $\{1, 2, \dots, N\}$ of size N and $\binom{[N]}{2}$ to denote the set $\{(1, 2), (1, 3), \dots, (N-1, N)\}$ of size $\binom{N}{2}$. Given $p \geq 1$, the p -norm of $\mathbf{x}$ is denoted by $\|\mathbf{x}\|_p$. The cardinality of a set Ξ is denoted by $|\Xi|$ (possibly infinite) and the closure by $\text{cl}(\Xi)$. Given sets $\Xi_i \subseteq \mathbb{R}$ for $i \in [N]$, the Cartesian product set is given by $\prod_{i \in [N]} \Xi_i = \{(\xi_1, \xi_2, \dots, \xi_N) \mid \xi_1 \in \Xi_1, \xi_2 \in \Xi_2, \dots, \xi_N \in \Xi_N\}$. Examples include the Boolean hypercube $\{0, 1\}^N$, the hypercube $[0, 1]^N$, the real space $\mathbb{R}^N$ and the nonnegative orthant $\mathbb{R}_+^N$. The indicator function of membership in a set Ξ is given by $\mathbb{1}_{\xi \in \Xi}$ which takes a value of 1 if $\xi \in \Xi$ and 0 if $\xi \notin \Xi$. The set of real valued symmetric square matrices of size $N \times N$ is denoted by $\mathbb{S}^N$. Given $\mathbf{X}, \mathbf{Y} \in \mathbb{S}^N$, we let $\mathbf{X} \bullet \mathbf{Y} = \text{trace}(\mathbf{X}\mathbf{Y})$. We use $\mathbf{X} \succeq 0$ to denote the matrix $\mathbf{X}$ is positive semidefinite and $\mathbf{X} \succeq \mathbf{Y}$ (respectively $\mathbf{X} \preceq \mathbf{Y}$) when $\mathbf{X} - \mathbf{Y} \succeq 0$ (respectively $\mathbf{Y} - \mathbf{X} \succeq 0$). The vector formed with the diagonal entries of the matrix $\mathbf{X}$ is denoted by $\text{diag}(\mathbf{X})$. Associated with a random vector $\tilde{\xi}$ is a probability distribution $\mathbb{P}$ which we denote by $\tilde{\xi} \sim \mathbb{P}$. We use $\mathbb{P}(\cdot)$ to denote the probability of an event, $\mathbb{E}_{\mathbb{P}}[\cdot]$ to denote the expectation with respect to $\mathbb{P}$ and $\text{supp}(\mathbb{P})$ to denote the support of $\mathbb{P}$. The covariance matrix of $\tilde{\xi} \sim \mathbb{P}$ is denoted by $\text{Cov}_{\mathbb{P}}(\tilde{\xi})$ and correlation matrix is denoted by $\text{Corr}_{\mathbb{P}}(\tilde{\xi})$. We use $\mathcal{P}(\Xi)$ to denote the set of all probability distributions with support contained in the set Ξ . A Dirac measure at ξ is denoted by δ_{ξ} . The projection of the probability distribution of $(\tilde{\xi}, \tilde{\chi}) \sim \mathbb{P}$ on $\tilde{\xi}$ is given $\text{proj}_{\tilde{\xi}}(\mathbb{P})$ and likewise the projection of the ambiguity set of $(\tilde{\xi}, \tilde{\chi}) \sim \mathbb{P} \in \mathcal{P}$ on $\tilde{\xi}$ is given by $\text{proj}_{\tilde{\xi}}(\mathcal{P})$.

2 Review on convex ambiguity sets

Convexity plays a key role in the design of tractable ambiguity sets. We review the class of convex ambiguity sets next that is also key to our work. Given a closed convex set $\Xi \subseteq \mathbb{R}^N$, a set of convex functions $f_l : \Xi \rightarrow \mathbb{R}$ for $l \in [L]$ and a set of scalars $\gamma_l \in \mathbb{R}$ for $l \in [L]$, define the ambiguity set $\mathcal{P}_{\text{cvx}}$ as follows:

$$\mathcal{P}_{\text{cvx}} := \{\mathbb{P} \in \mathcal{P}(\Xi) \mid \mathbb{E}_{\mathbb{P}}[f_l(\tilde{\xi})] \leq \gamma_l, \forall l \in [L]\}, \quad (2.1)$$

where all the expectations are assumed to be well-defined with respect to the distributions in the set. We refer to (2.1) as a *convex ambiguity set*. The corresponding distributionally robust optimization problem is given by:

$$\inf_{\mathbf{x} \in \mathcal{X}} \sup_{\mathbb{P} \in \mathcal{P}_{\text{cvx}}} \mathbb{E}_{\mathbb{P}}[g(\mathbf{x}, \tilde{\xi})]. \quad (2.2)$$

Under the assumptions: (i) the convex set Ξ has a polynomial time separation oracle, (ii) each of the convex functions $f_l : \Xi \rightarrow \mathbb{R}$ in the ambiguity set has a polynomial time subgradient oracle that returns the function value and its subgradient efficiently and (iii) the cost function is given by $g(\mathbf{x}, \xi) := \max_{k \in [K]} g_k(\mathbf{x}, \xi)$ where each function $g_k : \mathcal{X} \times \Xi \rightarrow \mathbb{R}$ is concave in ξ with a polynomial time supergradient oracle that returns the function value and its supergradient with respect to ξ efficiently,

the inner supremum in (2.2) for a fixed $\mathbf{x} \in \mathcal{X}$ is computable in polynomial time using the ellipsoid method (see theorem 1.5 in [38], chapter 3 in [48] and proposition 1 in [20]). This in turn has important ramifications on the solvability of the distributionally robust optimization problem. When each function $g_k : \mathcal{X} \times \Xi \rightarrow \mathbb{R}$ is also convex in $\mathbf{x}$ with a polynomial time subgradient oracle with respect to $\mathbf{x}$ and the feasible region $\mathcal{X}$ is a compact convex set with a polynomial time separation oracle, (2.2) is solvable in polynomial time using the ellipsoid method. In many applications, the functions and sets are structured and conic representable. In these cases, the distributionally robust optimization problem can be reformulated as a conic program that include linear, second order and semidefinite programs as special cases (see [5, 7]). The problem is solvable with optimization software such as MOSEK, Gurobi, CPLEX and tailored open source robust optimization packages such as RSOME [17]. When some of the decision variables are discrete, the problems are tackled using mixed integer conic programming solvers. Furthermore, for the more challenging class of two stage distributionally robust linear optimization problems with fixed recourse where the number of terms K in the cost function grow exponentially in the size of the problem, one can employ the rich literature on affine and piecewise affine decision rules to obtain tractable conic approximations (see [5, 7]).

In addition to computational tractability, convex ambiguity sets provide great modeling flexibility. We refer the reader to [56, 17, 11] for interesting examples where upper bounds on even degree moments, upper bounds on dispersion measures such as the mean absolute deviation and the directional deviation of linear combinations of random variables are modeled using convex ambiguity sets. In fact, certain types of covariance information can also be modeled using such an ambiguity set; we discuss this in detail in the next section. Another popular class of ambiguity sets that builds on the convex ambiguity set is the Wasserstein ambiguity set (see [26, 12, 27]). These ambiguity sets are characterized by distributions that lie in a Wasserstein ball of a fixed radius around a reference distribution where the reference distribution is typically assumed to be discrete and sample based. When the support of the uncertainty lies in a convex set, the Wasserstein ambiguity set is computationally tractable.

While continuous uncertainty has been well studied in distributionally robust optimization, in some applications, modeling the data as discrete random variables is natural. Discrete uncertainty is used to represent demand of indivisible goods, failures in power systems, influence relationship among individuals in social networks, labels of images in classification tasks and tokens in large language models. Discretization/binning is also a common technique used to convert any continuous random variable to a discrete random variable. Ambiguity sets that deal with discrete random variables are hence important. One such ambiguity set which allows the marginal distributions (discrete or continuous) to be specified is the marginal ambiguity set which contains all joint distributions with given marginal information. In the marginal ambiguity set with each random variable taking a finite number of values, the number of possible realizations of the random vector is exponential in the number of random variables. Regardless, under specific assumptions on the cost function and the set $\mathcal{X}$, (1.1) is solvable in polynomial time using linear optimization; see chapter 2 in [44]. However such an ambiguity does not explicitly model the dependence information among the random variables. These ambiguity sets have since been generalized to nonoverlapping and overlapping marginals and in certain cases (1.1) is still efficiently solvable, see

chapter 4 in [44]. Another tractable ambiguity set that models discrete uncertainty is the phi-divergence ambiguity set (see [6]). In the input specification of this ambiguity set, all the realizations of the discrete random vector are explicitly listed out and the probabilities of the realizations are allowed to vary. However if the number of realizations is potentially large, these formulations become computationally intractable. The Wasserstein ambiguity set also allows for tractable reformulations with discrete uncertainty only in specific instances; see the recent work of [4]. In the next section, we introduce a class of ambiguity sets that is tractable for discrete uncertainty by viewing such problems through the lens of submodularity. Furthermore, we show that submodular ambiguity sets also provide useful modeling capability for continuous uncertainty that goes beyond the expressivity of convex ambiguity sets.

3 Modeling with submodular ambiguity sets

In this section, we discuss modeling of uncertainty, discrete or continuous, with ambiguity sets defined using submodular functions.

3.1 Preliminaries

Given sets $\Xi_i \subseteq \mathbb{R}$ for all $i \in [N]$, a function $f : \prod_{i \in [N]} \Xi_i \rightarrow \mathbb{R}$ is submodular if:

$$f(\xi) + f(\chi) \geq f(\xi \wedge \chi) + f(\xi \vee \chi), \forall \xi, \chi \in \prod_{i \in [N]} \Xi_i, \quad (3.1)$$

where $\xi \wedge \chi = (\min(\xi_1, \chi_1), \dots, \min(\xi_N, \chi_N))$ and $\xi \vee \chi = (\max(\xi_1, \chi_1), \dots, \max(\xi_N, \chi_N))$. With $\xi, \chi \in \prod_{i \in [N]} \Xi_i$, we have $\xi \wedge \chi, \xi \vee \chi \in \prod_{i \in [N]} \Xi_i$. When the function is in addition twice differentiable over the continuous domain, this condition is equivalent to:

$$\frac{\partial^2}{\partial \xi_i \partial \xi_j} f(\xi) \leq 0, \forall \xi \in \prod_{i \in [N]} \Xi_i, \forall i \neq j. \quad (3.2)$$

A function $f : \prod_{i \in [N]} \Xi_i \rightarrow \mathbb{R}$ is supermodular if and only if $-f : \prod_{i \in [N]} \Xi_i \rightarrow \mathbb{R}$ is submodular where the sign of the inequalities in (3.1) and (3.2) are reversed. In Appendix B.1, we review submodularity in more detail. The material is standard and the reader can refer to excellent articles and books on this topic; see [41, 54, 3]. Two key properties of the functions that are relevant to this paper are discussed next:

- (i) The submodular function minimization problem:

$$\min_{\xi \in \prod_{i \in [N]} \Xi_i} f(\xi). \quad (3.3)$$

is solvable in polynomial time (see [41]) when the sets Ξ_i are discrete and finite for all $i \in [N]$ and the submodular function $f : \prod_{i \in [N]} \Xi_i \rightarrow \mathbb{R}$ has an efficient evaluation oracle. Equivalently, supermodular function maximization over this set is possible in polynomial time.

- (ii) The minimum expected value of a submodular function over the set of joint distributions of the input arguments with fixed marginals is attained by a perfectly positively dependent random

vector. Specifically, let $\mathcal{P}_F(\mathbb{P}_1, \dots, \mathbb{P}_N)$ denote the Fréchet set of joint distributions where the marginal distribution $\tilde{\xi}_i \sim \mathbb{P}_i$ is fixed with $\text{supp}(\mathbb{P}_i) \subseteq \Xi_i$ for each $i \in [N]$. Let $F_i : \Xi_i \rightarrow [0, 1]$ denote the cumulative distribution function of $\tilde{\xi}_i$ for each $i \in [N]$. A comonotonic random vector with perfect positive dependence is given by:

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

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 this comonotonic random vector. Then (see [51]):

$$\inf_{\mathbb{P} \in \mathcal{P}_F(\mathbb{P}_1, \dots, \mathbb{P}_N)} \mathbb{E}_{\mathbb{P}} [f(\tilde{\xi})] = \mathbb{E}_{\mathbb{P}^c} [f(\tilde{\xi}^c)], \forall \text{submodular } f \text{ such that expectations exists.} \quad (3.4)$$

or equivalently:

$$\sup_{\mathbb{P} \in \mathcal{P}_F(\mathbb{P}_1, \dots, \mathbb{P}_N)} \mathbb{E}_{\mathbb{P}} [f(\tilde{\xi})] = \mathbb{E}_{\mathbb{P}^c} [f(\tilde{\xi}^c)], \forall \text{supermodular } f \text{ such that expectations exists.} \quad (3.5)$$

We define a general class of submodular ambiguity sets next using these functions and discuss two specific examples.

3.2 Submodular ambiguity sets with examples

Given sets $\Xi_i \subseteq \mathbb{R}$ for all $i \in [N]$ (discrete or continuous), a finite set of submodular functions $f_l : \prod_{i \in [N]} \Xi_i \rightarrow \mathbb{R}$ for all $l \in [L]$ and a set of scalars $\gamma_l \in \mathbb{R}$ for all $l \in [L]$, define the submodular ambiguity set $\mathcal{P}_{\text{sub}}$ as follows:

$$\mathcal{P}_{\text{sub}} := \{\mathbb{P} \in \mathcal{P}(\prod_{i \in [N]} \Xi_i) \mid \mathbb{E}_{\mathbb{P}}[f_l(\tilde{\xi})] \leq \gamma_l, \forall l \in [L]\}, \quad (3.6)$$

where all the expectations are assumed to be well-defined with respect to the distributions in the set. We refer to (3.6) as a *submodular ambiguity set*. The similarity to the convex ambiguity set in (2.1) is obvious where convex functions are replaced with submodular functions and the convex set $\Xi \subseteq \mathbb{R}^N$ is replaced with the Cartesian product set $\prod_{i \in [N]} \Xi_i \subseteq \mathbb{R}^N$. While the ambiguity set in (3.6) specifies upper bounds on the expected value of submodular functions, equivalently, we can specify lower bounds on the expected value of supermodular functions. We next discuss two examples where univariate and bivariate information of the random vector is captured with a marginal-covariance submodular ambiguity set (for discrete uncertainty) and a mean-covariance submodular ambiguity set (for continuous uncertainty). For the interested reader, other examples of submodular ambiguity sets are provided in Appendix B.2.

Example 1 (Marginal-covariance submodular ambiguity set). A challenging problem in dependence modeling in areas such as risk management (see [25, 30]) is to construct joint distributions that are consistent with fixed marginal distributions and a prespecified covariance matrix. Typically fitting marginal distributions to data is much easier than fitting joint distributions, especially in high dimensions with limited data. Covariances are also a well understood linear measure of bivariate dependence that is simpler to

calibrate from data or provided through expert judgements in comparison to multivariate dependence structures such as copulas (see [51]). Unfortunately however, constructing a joint distribution or testing if there exists such a distribution is known to be a computationally hard problem. Specifically, consider the ambiguity set:

$$\{\mathbb{P} \in \mathcal{P}(\{0,1\}^N) \mid \mathbb{P}(\tilde{\xi}_i = 1) = p_i, \forall i \in [N], \mathbb{P}(\tilde{\xi}_i = 1, \tilde{\xi}_j = 1) = p_{i,j}, \forall (i,j) \in \binom{[N]}{2}\}, \quad (3.7)$$

where the univariate probabilities of Bernoulli random variables taking a value of 1 and the probabilities of pairwise Bernoulli random variables taking a value 1 are precisely specified. Verifying if (3.7) is nonempty is NP-complete and is equivalent to testing feasibility in a correlation polytope (see section 3 in [47]). Observe that since $\text{Cov}(\tilde{\xi}_i, \tilde{\xi}_j) = \mathbb{P}(\tilde{\xi}_i = 1, \tilde{\xi}_j = 1) - \mathbb{P}(\tilde{\xi}_i = 1)\mathbb{P}(\tilde{\xi}_j = 1) = p_{i,j} - p_i p_j$ and the univariate probabilities are known, we can equivalently specify the covariances in the ambiguity set instead of the pairwise probabilities. Our approach is to consider a relaxed ambiguity set of the form:

$$\{\mathbb{P} \in \mathcal{P}(\{0,1\}^N) \mid \mathbb{P}(\tilde{\xi}_i = 1) = p_i, \forall i \in [N], \mathbb{P}(\tilde{\xi}_i = 1, \tilde{\xi}_j = 1) \geq p_{i,j}, \forall (i,j) \in \binom{[N]}{2}\}. \quad (3.8)$$

Besides univariate probabilities, this ambiguity set specifies lower bounds on the probabilities of pairwise Bernoulli random variables taking a value of 1 or equivalently lower bounds on the covariances where $\text{Cov}(\tilde{\xi}_i, \tilde{\xi}_j) \geq p_{i,j} - p_i p_j$. The ambiguity set is defined using univariate functions of the form $\mathbb{1}_{\xi_i=1}$ (which is both submodular and supermodular) and bivariate supermodular functions of the form $\mathbb{1}_{\xi_i \geq 1, \xi_j \geq 1}$ where lower bounds on the expected value of the supermodular functions are given. Hence this is an example of a submodular ambiguity set. The classic Fréchet-Hoeffding bounds (see Theorems 2.12 and 2.14 in [51]) guarantees that for any distinct i, j :

$$\max(0, p_i + p_j - 1) \leq \mathbb{P}(\tilde{\xi}_i = 1, \tilde{\xi}_j = 1) \leq \min(p_i, p_j),$$

where the lower bound is attained by countermonotonic Bernoulli random variables (perfectly negatively dependent) with $(\tilde{\xi}_i, \tilde{\xi}_j) \sim (F_i^{-1}(\tilde{U}), F_j^{-1}(1 - \tilde{U}))$ and the upper bound is attained by comonotonic Bernoulli random variables (perfectly positively dependent) with $(\tilde{\xi}_i, \tilde{\xi}_j) \sim (F_i^{-1}(\tilde{U}), F_j^{-1}(\tilde{U}))$ where $\tilde{U} \sim \text{Uniform}[0, 1]$ and $F_i^{-1}(u) = 0$ if $u \in [0, 1 - p_i)$ and $F_i^{-1}(u) = 1$ if $u \in [1 - p_i, 1)$. If we define $p_{i,j} := \max(0, p_i + p_j - 1)$ in (3.8), we allow for all dependence structures that are possible among Bernoulli random vectors with the specified marginal probabilities. The ambiguity set then just reduces to specifying the univariate Bernoulli distributions which is an example of the marginal distribution ambiguity set. It is possible that the lower bounds on the respective bivariate probabilities $p_{i,j} = \max(0, p_i + p_j - 1)$ might not be jointly attained by any Bernoulli random vector (for example when $p_1 = p_2 = p_3 = 0.5$ and $p_{1,2} = p_{1,3} = p_{2,3} = 0$). On the other hand, if we set all $p_{i,j} := \min(p_i, p_j)$, the ambiguity set allows for only the comonotonic Bernoulli random vector where the bivariate probabilities $p_{i,j}$ across all pairs are jointly attained. As the input parameter $p_{i,j}$ is varied, the ambiguity set controls the minimum allowable covariance between the random variables $\tilde{\xi}_i$ and $\tilde{\xi}_j$ in the ambiguity set. The submodular ambiguity set (3.8) thus provides a relaxation of (3.7). This ambiguity set has been considered in [13] where the authors find polynomial time computable bounds on the probability

of the union of random events. More generally, when the uncertainty is discrete with finite support, we define the *marginal-covariance submodular ambiguity* set as:

$$\mathcal{P}_{\text{marg-cov-sub}} := \{\mathbb{P} \in \mathcal{P}(\prod_{i \in [N]} \Xi_i) \mid \mathbb{P}(\tilde{\xi}_i = \xi_i) = p_i(\xi_i), \forall \xi_i \in \Xi_i, \forall i \in [N], \\ \text{Cov}(\tilde{\xi}_i, \tilde{\xi}_j) \geq \Sigma_{i,j}, \forall (i, j) \in \binom{[N]}{2}\}. \quad (3.9)$$

In (3.9), the marginal distribution of each discrete random variable is fixed with given probabilities $p_i(\xi_i)$ and lower bounds on the covariance of pairs of random variables is specified as $\Sigma_{i,j}$. To see that it is a submodular ambiguity set, define the mean of each random variable as:

$$\mu_i := \sum_{\xi_i \in \Xi_i} p_i(\xi_i) \xi_i, \quad \forall i \in [N]. \quad (3.10)$$

We can then equivalently express (3.9) as:

$$\mathcal{P}_{\text{marg-cov-sub}} = \{\mathbb{P} \in \mathcal{P}(\prod_{i \in [N]} \Xi_i) \mid \mathbb{P}(\tilde{\xi}_i = \xi_i) = p_i(\xi_i), \forall \xi_i \in \Xi_i, \forall i \in [N], \\ \mathbb{E}[\tilde{\xi}_i \tilde{\xi}_j] \geq \mu_i \mu_j + \Sigma_{i,j}, \forall (i, j) \in \binom{[N]}{2}\}. \quad (3.11)$$

The ambiguity set is defined using univariate functions of the form $\mathbb{1}_{\xi_i=t}$ (which is both submodular and supermodular) and bivariate supermodular functions of the form $\xi_i \xi_j$ where lower bounds on the expected value of the bivariate supermodular functions are given. Hence this is an example of a submodular ambiguity set. As we will see later, this is a tractable ambiguity set that is useful for applications.

Example 2 (Mean-covariance submodular ambiguity set). Moment ambiguity sets are often used to model distributional information on the random vector through limited moments such as means, variances and covariances. A popular *mean-covariance convex ambiguity set* $\mathcal{P}_{\text{mean-cov-cvx}}$, first proposed in [20] (see [56] for related work) is:

$$\mathcal{P}_{\text{mean-cov-cvx}} := \{\mathbb{P} \in \mathcal{P}(\Xi) \mid \mathbb{E}_{\mathbb{P}}[\tilde{\xi}] = \mu, \text{Cov}_{\mathbb{P}}(\tilde{\xi}) \preceq \Sigma\}, \quad (3.12)$$

where $\Xi \subseteq \mathbb{R}^N$ is a convex set with the mean of the random vector fixed to μ and the covariance matrix bounded from above by Σ in a positive semidefinite order. This ambiguity set is given by the projection of a convex ambiguity of the form (2.1) (see example 2 in [56]). Specifically, $\mathcal{P}_{\text{mean-cov-cvx}} = \text{proj}_{\tilde{\xi}}(\mathcal{Q}_{\text{mean-cov-cvx}})$ where:

$$\mathcal{Q}_{\text{mean-cov-cvx}} := \{\mathbb{Q} \in \mathcal{P}(\mathbb{R}^N \times \mathbb{S}^N : \xi \in \Xi, U \succeq (\xi - \mu)(\xi - \mu)' \mid \mathbb{E}_{\mathbb{Q}}[\tilde{\xi}] = \mu, \mathbb{E}_{\mathbb{Q}}[\tilde{U}] = \Sigma\},$$

with $(\tilde{\xi}, \tilde{U}) \sim \mathbb{Q}$. $\mathcal{Q}_{\text{mean-cov-cvx}}$ is an example of a convex ambiguity set since the support of $\mathbb{Q}$ is a convex set where:

$$U \succeq (\xi - \mu)(\xi - \mu)' \Leftrightarrow \begin{pmatrix} 1 & \xi - \mu \\ \xi - \mu & U \end{pmatrix} \succeq 0,$$

and all the expectation constraints are defined on linear functions of the random variables. Incorporating

lower bounds on the covariance matrix in a positive semidefinite order is known to be much harder computationally; see remark 1 in [20]. Specifically, with $\Xi = \mathbb{R}_+^N$, consider the ambiguity set:

$$\{\mathbb{P} \in \mathcal{P}(\mathbb{R}_+^N) \mid \mathbb{E}_{\mathbb{P}}[\tilde{\xi}] = \boldsymbol{\mu}, \text{Cov}_{\mathbb{P}}(\tilde{\xi}) = \boldsymbol{\Sigma}\}, \quad (3.13)$$

where the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ of the nonnegative random vector are precisely specified. Verifying if (3.13) is nonempty is NP-hard and is equivalent to testing if a matrix is completely positive (see [48, 21]). The mean-covariance convex ambiguity set in (3.12) with $\Xi = \mathbb{R}_+^N$ provides a tractable relaxation to (3.13) where the equality of the covariance matrix is relaxed to an upper bound in a positive semidefinite order (see [20]). We now propose an alternate relaxation using a mean-covariance submodular ambiguity set. Specifically, given the mean vector $\boldsymbol{\mu}$ and a symmetric matrix $\boldsymbol{\Sigma} \in \mathbb{S}^N$ (not necessarily positive semidefinite) with nonnegative entries along the diagonal, define the *mean-covariance submodular ambiguity set* as:

$$\mathcal{P}_{\text{mean-cov-sub}} := \{\mathbb{P} \in \mathcal{P}(\mathbb{R}_+^N) \mid \mathbb{E}_{\mathbb{P}}[\tilde{\xi}] = \boldsymbol{\mu}, \text{diag}(\text{Cov}_{\mathbb{P}}(\tilde{\xi})) = \text{diag}(\boldsymbol{\Sigma}), \text{Cov}_{\mathbb{P}}(\tilde{\xi}) \geq \boldsymbol{\Sigma}\}. \quad (3.14)$$

In (3.14), the mean of the nonnegative random variable $\tilde{\xi}_i$ is fixed to μ_i and the variance is fixed to $\Sigma_{i,i}$ for each $i \in [N]$. In addition, the covariance of each distinct pair of random variables is bounded from below as $\text{Cov}(\tilde{\xi}_i, \tilde{\xi}_j) \geq \Sigma_{i,j}$. We can equivalently express (3.14) as:

$$\mathcal{P}_{\text{mean-cov-sub}} := \{\mathbb{P} \in \mathcal{P}(\mathbb{R}_+^N) \mid \mathbb{E}_{\mathbb{P}}[\tilde{\xi}] = \boldsymbol{\mu}, \mathbb{E}_{\mathbb{P}}[\text{diag}(\tilde{\xi}\tilde{\xi}')] = \text{diag}(\boldsymbol{\mu}\boldsymbol{\mu}' + \boldsymbol{\Sigma}), \mathbb{E}_{\mathbb{P}}[\tilde{\xi}\tilde{\xi}'] \geq \boldsymbol{\mu}\boldsymbol{\mu}' + \boldsymbol{\Sigma}\}. \quad (3.15)$$

The ambiguity set is defined using univariate functions of the form ξ_i and ξ_i^2 (these functions are both submodular and supermodular) and bivariate supermodular functions of the form $\xi_i\xi_j$ where lower bounds on the expected value of the bivariate supermodular functions are given. Hence this is an example of a submodular ambiguity set. As we will see later, this is a tractable ambiguity set that is useful for applications.

While submodular and supermodular functions have been previously used in robust and distributionally robust optimization problems; see [33, 1, 16, 40, 42], to the best of our knowledge, the use of such functions in explicitly constructing ambiguity sets has not been explored thus far. There is also related work on reducing the conservatism of the marginal distribution ambiguity set by restricting the associated copula to be close to a nominal copula in the Wasserstein distance; see [28]. In contrast, we reduce the conservatism through directly incorporating lower bounds on the termwise covariances and build on submodularity. In the following sections, we discuss the computational tractability of the submodular ambiguity set and its application to distributionally robust optimization.

4 Discrete uncertainty

In this section, we provide a computationally tractable model for the distributionally robust optimization problem with a submodular ambiguity set when the uncertainty is discrete. Specifically, we consider

the problem:

$$\inf_{\mathbf{x} \in \mathcal{X}} \sup_{\mathbb{P} \in \mathcal{P}_{\text{sub}}} \mathbb{E}_{\mathbb{P}} \left[g(\mathbf{x}, \tilde{\boldsymbol{\xi}}) \right], \quad (4.1)$$

and show that it is solvable in polynomial time under fairly general assumptions on the cost function and the submodular ambiguity set.

4.1 Polynomial time solvability with the ellipsoid method

Towards solving (4.1), we first focus on the problem of computing the inner supremum. Consider the problem:

$$\rho^* := \sup_{\mathbb{P} \in \mathcal{P}_{\text{sub}}} \mathbb{E}_{\mathbb{P}} \left[g(\tilde{\boldsymbol{\xi}}) \right] = \sup \left\{ \mathbb{E}_{\mathbb{P}}[g(\tilde{\boldsymbol{\xi}})] \mid \mathbb{P} \in \mathcal{P}(\prod_{i \in [N]} \Xi_i), \mathbb{E}_{\mathbb{P}}[f_l(\tilde{\boldsymbol{\xi}})] \leq \gamma_l, \forall l \in [L] \right\}, \quad (4.2)$$

where $g : \prod_{i \in [N]} \Xi_i \rightarrow \mathbb{R}$ is the cost function and ρ^* is the optimal value. We make the following assumptions on the ambiguity set and the cost function:

Assumption (A1): The set $\Xi_i \subset \mathbb{R}$ is a discrete, finite set for each $i \in [N]$.

Assumption (A2): For each $l \in [L]$, $f_l : \prod_{i \in [N]} \Xi_i \rightarrow \mathbb{R}$ is a submodular function with a polynomial time evaluation oracle.

Assumption (A3): The ambiguity set $\mathcal{P}_{\text{sub}}$ is nonempty.

Assumption (A4): The cost function is given by $g(\boldsymbol{\xi}) := \max_{k \in [K]} g_k(\boldsymbol{\xi})$ where for each $k \in [K]$, the function $g_k : \prod_{i \in [N]} \Xi_i \rightarrow \mathbb{R}$ is a supermodular function with a polynomial time evaluation oracle.

This brings us to the first theorem of the paper.

Theorem 4.1. *Suppose assumptions (A1)-(A4) hold. Then ρ^* is computable in polynomial time.*

Proof. See Appendix A. □

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

- (a) The optimal value ρ^* is computable in time polynomial in N (number of random variables), L (number of submodular functions defining the ambiguity set), K (number of piecewise terms in the objective function), $\max_{i \in [N]} |\Xi_i|$ (the maximum cardinality of any marginal support) and the function evaluation time of the oracles. It makes use of linear programming duality, the ellipsoid method and the polynomial time solvability of submodular function minimization. The result closely mirrors that of the convex ambiguity set discussed in Section 2 where the objective function is given by the maximum of concave functions of the random vector instead (see theorem 1.5 in [38], chapter 3 in [48] and proposition 1 in [20]). We believe the value of Theorem 4.1 lies in identifying new polynomial time solvable instances of (4.2). As an illustration, the result immediately implies that ρ^* is polynomial time computable for the marginal-covariance submodular ambiguity set introduced in Example 1.
- (b) The proof of Theorem 4.1 makes use of the ellipsoid method which helps with the theoretical analysis. However, the ellipsoid method is not easy to implement in practice. As we will show in the next section, for structured submodular ambiguity sets and cost functions, it is possible

to obtain compact linear programs which makes it practical to use. Similar developments have occurred over the past decade where conic reformulations have been developed for structured convex ambiguity sets.

- (c) The proof of Theorem 4.1 makes use of duality. A natural question is to characterize the extremal primal distribution which attains the bound ρ^* . When the submodular ambiguity set incorporates the full marginal distributions with $\tilde{\xi}_i \sim \mathbb{P}_i$ for each $i \in [N]$, a feasible distribution in the set is given by a comonotonic random vector with the given marginals. This distribution attains the supremum when the objective is supermodular (see (3.5)). However, the objective function we consider is given by the maximum of supermodular functions which need not be supermodular and the corresponding extremal distribution need not be comonotonic. The next example provides an illustration of this.

Example 3 (Extremal distribution for maximum of two Bernoulli random variables). Let $N = 2$ and $g(\xi_1, \xi_2) = \max(\xi_1, \xi_2)$. Consider the problem of computing the maximum expected value:

$$\max \left\{ \mathbb{E}_{\mathbb{P}} \left[\max(\tilde{\xi}_1, \tilde{\xi}_2) \right] \mid \mathbb{P} \in \mathcal{P}(\{0, 1\}^N), \mathbb{P}(\tilde{\xi}_1 = 1) = p_1, \mathbb{P}(\tilde{\xi}_2 = 1) = p_2, \mathbb{P}(\tilde{\xi}_1 = 1, \tilde{\xi}_2 = 1) \geq p_{1,2} \right\},$$

where we use the marginal-covariance submodular ambiguity set for Bernoulli random variables. This ambiguity set is nonempty when $p_1, p_2 \in [0, 1]$ and $p_{1,2} \leq \min(p_1, p_2)$. We can reformulate the problem by using a single decision variable for the probability $p = \mathbb{P}(\tilde{\xi}_1 = 1, \tilde{\xi}_2 = 1)$ as follows:

$$\max \left\{ p_1 + p_2 - p \mid p_1 - p \geq 0, p_2 - p \geq 0, 1 - p_1 - p_2 + p \geq 0, p \geq 0, p \geq p_{1,2} \right\},$$

where $\mathbb{P}(\tilde{\xi}_1 = 1, \tilde{\xi}_2 = 0) = p_1 - p$, $\mathbb{P}(\tilde{\xi}_1 = 0, \tilde{\xi}_2 = 1) = p_2 - p$ and $\mathbb{P}(\tilde{\xi}_1 = 0, \tilde{\xi}_2 = 0) = 1 - p_1 - p_2 + p$. The objective function is given by $\mathbb{E}_{\mathbb{P}}[\max(\tilde{\xi}_1, \tilde{\xi}_2)] = p_1 + p_2 - p$. The optimal value of p is given by $p^* = \max(0, p_1 + p_2 - 1, p_{1,2})$. This implies:

- (i) When $p_{1,2} \leq \max(0, p_1 + p_2 - 1)$, we have $p^* = \max(0, p_1 + p_2 - 1)$ and the extremal distribution is given by the countermonotonic distribution with perfectly negatively dependent random variables.
- (ii) When $p_{1,2} = p_1 p_2$, we have $p^* = p_1 p_2$ and the extremal distribution is given by the independent distribution.
- (iii) When $p_{1,2} = \min(p_1, p_2)$, we have $p^* = \min(p_1, p_2)$ and the extremal distribution is the comonotonic distribution with perfectly positively dependent random variables.

For every $p_{1,2} \in [\max(0, p_1 + p_2 - 1), \min(p_1, p_2)]$, the inequality constraint in the ambiguity set is tight at optimality. The extremal distribution hence ranges from perfectly negatively dependent to independent to perfectly positively dependent Bernoulli random variables and is based on the input parameter $p_{1,2}$. On the other hand, if we solve the problem over the marginal distribution ambiguity set which is also a submodular ambiguity set:

$$\max \left\{ \mathbb{E}_{\mathbb{P}} \left[\max(\tilde{\xi}_1, \tilde{\xi}_2) \right] \mid \mathbb{P} \in \mathcal{P}(\{0, 1\}^N), \mathbb{P}(\tilde{\xi}_1 = 1) = p_1, \mathbb{P}(\tilde{\xi}_2 = 1) = p_2 \right\},$$

the extremal distribution will always be the countermonotonic distribution (see Theorems 2.12 and 2.14 in [51]). The example thus illustrates the value of modeling covariance information, when available, using the submodular ambiguity set. In this example, the objective function is submodular and piecewise linear and the inequality constraint turns out to be tight for all $p_{1,2} \in [\max(0, p_1 + p_2 - 1), \min(p_1, p_2)]$. One might conjecture that the inequality constraints in the ambiguity set are tight for all submodular objective functions. However, this is not necessarily true (see Appendix B.3 for a counterexample).

The next theorem builds on Theorem 4.1 to identify conditions that guarantee polynomial time solvability of the distributionally robust optimization problem.

Theorem 4.2. *Consider the distributionally robust optimization problem in (4.1). Suppose assumptions (A1)-(A3) hold. Let $g(\mathbf{x}, \boldsymbol{\xi}) := \max_{k \in [K]} g_k(\mathbf{x}, \boldsymbol{\xi})$ where each function $g_k : \mathcal{X} \times \prod_{i \in [N]} \Xi_i \rightarrow \mathbb{R}$ is convex in $\mathbf{x}$ with a polynomial time subgradient oracle with respect to $\mathbf{x}$ and supermodular in $\boldsymbol{\xi}$ with a polynomial time evaluation oracle. When the feasible region $\mathcal{X}$ is a compact convex set with a polynomial time separation oracle, the distributionally robust optimization problem (4.1) is solvable in polynomial time.*

Proof. See Appendix A. □

A similar polynomial time complexity result exists for the convex ambiguity set that makes use of the ellipsoid method as discussed in Section 2. Theorem 4.2 extends this to discrete uncertainty.

4.2 Polynomial sized linear program

In this section, we develop a polynomial sized linear program for the marginal-covariance submodular ambiguity set for specific objective functions. Particularly, we focus on the problem of computing the tightest upper bound on the expected value of a piecewise quadratic function of the random vector given by:

$$\rho^* := \sup_{\mathbb{P} \in \mathcal{P}_{\text{marg-cov-sub}}} \mathbb{E}_{\mathbb{P}} \left[\max_{k \in [K]} \left(\tilde{\boldsymbol{\xi}}' \mathbf{A}_k \tilde{\boldsymbol{\xi}} + \mathbf{b}_k' \tilde{\boldsymbol{\xi}} + c_k \right) \right], \quad (4.3)$$

where the matrices $\mathbf{A}_1, \dots, \mathbf{A}_K$ in $\mathbb{S}^N$, the vectors $\mathbf{b}_1, \dots, \mathbf{b}_K$ in $\mathbb{R}^N$ and the scalars $c_1, \dots, c_K$ are given. The next theorem provides necessary and sufficient conditions to verify the ambiguity set $\mathcal{P}_{\text{marg-cov-sub}}$ is nonempty.

Theorem 4.3. *When Ξ_i is discrete and finite for each $i \in [N]$, the marginal-covariance submodular ambiguity set $\mathcal{P}_{\text{marg-cov-sub}}$ in (3.11) is nonempty if and only if there exists scalars $\lambda_{i,j}(\xi_i, \xi_j)$ for $\xi_i \in$*

$\Xi_i, \xi_j \in \Xi_j, (i, j) \in \binom{[N]}{2}$ so that the following set of linear conditions are satisfied:

$$\begin{aligned}
\sum_{\xi_i \in \Xi_i} p_i(\xi_i) &= 1, & \forall i \in [N], \\
\sum_{\xi_i \in \Xi_i} \sum_{\xi_j \in \Xi_j} \xi_i \xi_j \lambda_{i,j}(\xi_i, \xi_j) &\geq \mu_i \mu_j + \Sigma_{i,j}, & \forall (i, j) \in \binom{[N]}{2}, \\
\sum_{\xi_i \in \Xi_i} \lambda_{i,j}(\xi_i, \xi_j) &= p_j(\xi_j), & \forall \xi_j \in \Xi_j, \forall (i, j) \in \binom{[N]}{2}, \\
\sum_{\xi_j \in \Xi_j} \lambda_{i,j}(\xi_i, \xi_j) &= p_i(\xi_i), & \forall \xi_i \in \Xi_i, \forall (i, j) \in \binom{[N]}{2}, \\
p_i(\xi_i) &\geq 0, & \forall \xi_i \in \Xi_i, \forall i \in [N], \\
\lambda_{i,j}(\xi_i, \xi_j) &\geq 0, & \forall \xi_i \in \Xi_i, \forall \xi_j \in \Xi_j, \forall (i, j) \in \binom{[N]}{2},
\end{aligned} \tag{4.4}$$

where $p_i(\xi_i)$ for all $\xi_i \in \Xi_i, i \in [N]$ and $\Sigma_{i,j}$ for all $(i, j) \in \binom{[N]}{2}$ are specified with μ_i for all $i \in [N]$ defined in (3.10). This is verifiable in polynomial time.

Proof. See Appendix A. □

Theorem 4.3 avoids the use of the joint probabilities $\mathbb{P}(\tilde{\xi} = \xi)$ in characterizing the ambiguity set. In fact, as the proof shows, we require the use of the bivariate probabilities $\lambda_{i,j}(\xi_i, \xi_j) = \mathbb{P}(\tilde{\xi}_i = \xi_i, \tilde{\xi}_j = \xi_j)$ for a computationally tractable representation.

In the following theorem, we show that problem (4.3) is solvable efficiently as a compact linear program for this ambiguity set when the matrices in the objective function satisfy an additional condition.

Theorem 4.4. *Suppose the off-diagonal entries of the matrices $\mathbf{A}_k$ are nonnegative for each $k \in [K]$ and the marginal-covariance submodular ambiguity set satisfies the conditions in (4.4). Then ρ^* in (4.3)*

is given by the optimal value of the polynomial sized linear program:

$$\begin{aligned}
\max_{\lambda} \quad & \sum_{k \in [K]} \left(\sum_{(i,j) \in \binom{[N]}{2}} \sum_{\xi_i \in \Xi_i} \sum_{\xi_j \in \Xi_j} 2A_{i,j,k} \xi_i \xi_j \lambda_{i,j,k}(\xi_i, \xi_j) + \sum_{i \in [N]} \sum_{\xi_i \in \Xi_i} (A_{i,i,k} \xi_i^2 + b_{i,k} \xi_i) \lambda_{i,k}(\xi_i) + c_k \lambda_k \right) \\
\text{s.t.} \quad & \sum_{k \in [K]} \lambda_k = 1, \\
& \sum_{\xi_i \in \Xi_i} \lambda_{i,k}(\xi_i) = \lambda_k, \quad \forall i \in [N], \forall k \in [K], \\
& \sum_{k \in [K]} \lambda_{i,k}(\xi_i) = p_i(\xi_i), \quad \forall \xi_i \in \Xi_i, \forall i \in [N], \\
& \sum_{k \in [K]} \sum_{\xi_i \in \Xi_i} \sum_{\xi_j \in \Xi_j} \xi_i \xi_j \lambda_{i,j,k}(\xi_i, \xi_j) \geq \mu_i \mu_j + \Sigma_{i,j}, \quad \forall (i,j) \in \binom{[N]}{2}, \\
& \sum_{\xi_i \in \Xi_i} \lambda_{i,j,k}(\xi_i, \xi_j) = \lambda_{j,k}(\xi_j), \quad \forall \xi_j \in \Xi_j, \forall (i,j) \in \binom{[N]}{2}, \forall k \in [K], \\
& \sum_{\xi_j \in \Xi_j} \lambda_{i,j,k}(\xi_i, \xi_j) = \lambda_{i,k}(\xi_i), \quad \forall \xi_i \in \Xi_i, \forall (i,j) \in \binom{[N]}{2}, \forall k \in [K], \\
& \lambda_k \geq 0, \quad \forall k \in [K], \\
& \lambda_{i,k}(\xi_i) \geq 0, \quad \forall \xi_i \in \Xi_i, \forall i \in [N], \forall k \in [K], \\
& \lambda_{i,j,k}(\xi_i, \xi_j) \geq 0, \quad \forall \xi_i \in \Xi_i, \forall \xi_j \in \Xi_j, \forall (i,j) \in \binom{[N]}{2}, \forall k \in [K].
\end{aligned} \tag{4.5}$$

Proof. See Appendix A. $\square$

We make a few remarks about Theorem 4.4 and its implications next.

- (a) The proof is constructive where the extremal distribution is given by a mixture of comonotonic random vectors (see Figures 1 and 2 for an illustration of the construction). However the extremal distribution itself need not be comonotonic.

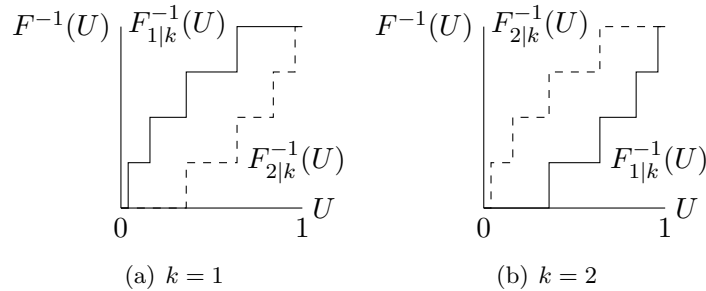


Figure 1: Comonotonic construction of the conditional distributions in step (ii) for $N = 2$ and $K = 2$. Here the solid line indicates $\tilde{\xi}_1 \sim F_{1|k}^{-1}(\tilde{U})$ and the dashed line indicates $\tilde{\xi}_2 \sim F_{2|k}^{-1}(\tilde{U})$ for $k = 1$ (left figure) and $k = 2$ (right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

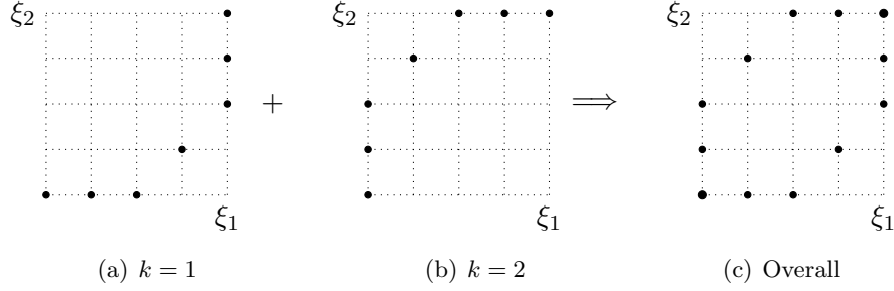


Figure 2: Subfigures (a) and (b) display the support of the conditional bivariate distributions for $k = 1$ and $k = 2$ while (c) shows that overall support of the extremal bivariate distribution using the weighted probabilities p_1^* and p_2^* . While the conditional bivariate distributions in (a) and (b) are comonotonic, the final bivariate distribution in (c) is not comonotonic but positively dependent here.

- (c) The size of the linear program in Theorem 4.4 grows quadratically in N (number of random variables) and $\max_{i \in [N]} |\Xi_i|$ (maximum number of values that a random variable can take) and linearly in K (number of pieces defining the objective function). The proof builds on the proof of Theorem 2.3.1 in [44]. However the linear program therein is developed for the marginal distribution ambiguity set. The key novelty in Theorem 4.4 is that by introducing additional decision variables for modeling bivariate probabilities and incorporating appropriate constraints, we show that it is possible to compute the sharp bound for the marginal-covariance submodular ambiguity set in polynomial time with linear programming.
- (d) When the lower bounds on the covariances in the ambiguity is specified only for a subset $E \subset \binom{[N]}{2}$, it is possible to be reduce the size of the linear program drastically. For example, if the matrices $\mathbf{A}_k$ have zero entries in the off-diagonal terms corresponding to the pairs $(i, j) \notin E$ (such as a piecewise linear objective), we can formulate the linear program in (4.5) using only the variables and constraints corresponding to the pairs $(i, j) \in E$ rather than $\binom{[N]}{2}$. This significantly reduces the quadratic dependence of the size of the linear program on the number of random variables when E is sparse.
- (e) It is straightforward to build on Theorem 4.4 to solve distributionally robust optimization problems. Specifically, consider:

$$\inf_{\mathbf{x} \in \mathcal{X}} \sup_{\mathbb{P} \in \mathcal{P}_{\text{marg-cov-sub}}} \mathbb{E}_{\mathbb{P}} \left[\max_{k \in [K]} \left(\tilde{\boldsymbol{\xi}}' \mathbf{A}_k(\mathbf{x}) \tilde{\boldsymbol{\xi}} + \mathbf{b}_k'(\mathbf{x}) \tilde{\boldsymbol{\xi}} + c_k(\mathbf{x}) \right) \right], \quad (4.6)$$

where for each $k \in [K]$, the matrix $\mathbf{A}_k(\mathbf{x})$, the vector $\mathbf{b}_k(\mathbf{x})$ and the scalar $c_k(\mathbf{x})$ have an affine dependence on $\mathbf{x}$. Assume the set $\mathcal{X}$ is given by the polyhedron $\{\mathbf{x} \mid \mathbf{D}\mathbf{x} \geq \mathbf{d}\}$ and the off-diagonal entries of the matrix $\mathbf{A}_k(\mathbf{x})$ are nonnegative for each $k \in [K]$ and $\mathbf{x} \in \mathcal{X}$. Then by using the dual of the linear program in (4.5), we can reformulate (4.6) as a linear program. Such a technique is commonly used in robust and distributionally robust optimization (see [5, 7]).

5 Continuous uncertainty

In this section, we discuss distributionally robust optimization with submodular ambiguity sets when the support of the uncertainty is allowed to be continuous.

5.1 Pseudo-polynomial time solvability with the ellipsoid method

We first focus on solving the problem (4.2) and make the following assumptions on the ambiguity set and the cost function:

Assumption (A1'): The set $\Xi_i = [0, B_i] \subset \mathbb{R}$ is a finite interval for each $i \in [N]$.

Assumption (A2'): For each $l \in [L]$, $f_l : \prod_{i \in [N]} \Xi_i \rightarrow \mathbb{R}$ is a M -Lipschitz submodular function with respect to the infinity norm with a polynomial time evaluation oracle.

Assumption (A3'): The ambiguity set $\mathcal{P}_{\text{sub}}$ is nonempty with $\sum_{r \in [R]} p_r \delta_{\xi_r} \in \mathcal{P}_{\text{sub}}$.

Assumption (A4'): The cost function is given by $g(\xi) := \max_{k \in [K]} g_k(\xi)$ where for each $k \in [K]$, the function $g_k : \prod_{i \in [N]} \Xi_i \rightarrow \mathbb{R}$ is a M -Lipschitz supermodular function with respect to the infinity norm with a polynomial time evaluation oracle.

This brings us to the following theorem.

Theorem 5.1. *Suppose assumptions (A1')-(A4') hold. Then for every $\epsilon > 0$, one can find a value $\rho_\epsilon^* \geq \rho^* - \epsilon$ in time polynomial in $N, L, K, \max_{i \in [N]} B_i, R, M, \frac{1}{\epsilon}$ and the evaluation time of the oracles such that there exists a distribution $\mathbb{P}_\epsilon^* \in \mathcal{P}(\prod_{i \in [N]} \Xi_i)$ with $\rho_\epsilon^* = \mathbb{E}_{\mathbb{P}_\epsilon^*}[g(\tilde{\xi})]$ and $\mathbb{E}_{\mathbb{P}_\epsilon^*}[f_l(\tilde{\xi})] \leq \gamma_l + \epsilon$ for all $l \in [L]$.*

Proof. See Appendix A. □

- (a) The proof technique in Theorem 5.1 builds on the discretization approach first proposed in [3] for deterministic Lipschitz continuous submodular function minimization. The usefulness of the theorem lies in that even when the uncertainty is defined on a continuous support, it is possible to obtain pseudo-polynomial time ϵ -additive guarantees with the submodular ambiguity set. Stronger polynomial time solution guarantees can be obtained for continuous uncertainty with a convex ambiguity set as discussed in Section 2. On the other hand, the theorem shows that it is possible to incorporate information on the uncertainty that might not be possible with a convex ambiguity set while still being solvable in pseudo-polynomial time. To the best of our knowledge, such a result is not covered by existing models in distributionally robust optimization
- (b) In fact, when the submodular ambiguity set is defined using only univariate and bivariate information with the support of each random variable given by an interval, we can use discretization of the marginal support and solve the linear program provided in Theorem 4.4 to obtain the additive approximation. The size of the linear program in this case is pseudo-polynomial in the size of the input. It is also straightforward to extend these results to the distributionally robust optimization problem through duality. In the next section, we provide stronger polynomial time guarantees using semidefinite programming for the mean-covariance submodular ambiguity set.

5.2 Polynomial sized semidefinite program

In this section, we develop a semidefinite program for the mean-covariance submodular ambiguity set $\mathcal{P}_{\text{mean-cov-sub}}$ defined in (3.15). Towards verifying non-emptiness of this ambiguity set, we start by defining a moment cone $\mathcal{M}$ as follows:

$$\mathcal{M} := \left\{ m_0(1, \mathbf{m}, \mathbf{M}) \mid \exists \mathbb{P} \in \mathcal{P}(\mathbb{R}_+^N) \text{ s.t. } \mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{\xi}}] = \mathbf{m}, \mathbb{E}_{\mathbb{P}}[\text{diag}(\tilde{\boldsymbol{\xi}}\tilde{\boldsymbol{\xi}}')] = \text{diag}(\mathbf{M}), \mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{\xi}}\tilde{\boldsymbol{\xi}}'] \geq \mathbf{M}, m_0 \geq 0 \right\}. \quad (5.1)$$

The ambiguity set $\mathcal{P}_{\text{mean-cov-sub}}$ is nonempty if and only if $(1, \boldsymbol{\mu}, \boldsymbol{\mu}\boldsymbol{\mu}' + \boldsymbol{\Sigma}) \in \mathcal{M}$. The set $\mathcal{M}$ is however not closed, as seen in the following example.

Example 4 (Moment cone $\mathcal{M}$ is not closed). Let $N = 2$ with $\mu_1 = \mu_2 = 0$, $\Sigma_{1,1} = \Sigma_{2,2} = \Sigma_{1,2} = 1$. Then, $(1, \boldsymbol{\mu}, \boldsymbol{\mu}\boldsymbol{\mu}' + \boldsymbol{\Sigma}) \notin \mathcal{M}$. This is because it not possible to have nonnegative random variables with zero means and strictly positive variances. However $(1, \boldsymbol{\mu}, \boldsymbol{\mu}\boldsymbol{\mu}' + \boldsymbol{\Sigma}) \in \text{cl}(\mathcal{M})$ (the closure of the moment cone). To see this, consider the limit of the moments of a sequence of probability measures of the form below with $\epsilon \in (0, 1]$ where $\epsilon \downarrow 0$:

$$\begin{aligned} \mathbb{P}_{\epsilon}(\tilde{\xi}_1 = 0, \tilde{\xi}_2 = 0) &= 1 - \epsilon, \\ \mathbb{P}_{\epsilon}\left(\tilde{\xi}_1 = \frac{1}{\sqrt{\epsilon}}, \tilde{\xi}_2 = \frac{1}{\sqrt{\epsilon}}\right) &= \epsilon. \end{aligned}$$

The means are given by $\mathbb{E}_{\mathbb{P}_{\epsilon}}[\tilde{\xi}_1] = \mathbb{E}_{\mathbb{P}_{\epsilon}}[\tilde{\xi}_2] = \sqrt{\epsilon}$ with $\mathbb{E}_{\mathbb{P}_{\epsilon}}[\tilde{\xi}_1^2] = \mathbb{E}_{\mathbb{P}_{\epsilon}}[\tilde{\xi}_2^2] = 1$ and $\mathbb{E}_{\mathbb{P}_{\epsilon}}[\tilde{\xi}_1\tilde{\xi}_2] = 1$. Then $(1, \boldsymbol{\mu}, \boldsymbol{\mu}\boldsymbol{\mu}' + \boldsymbol{\Sigma}) = \lim_{\epsilon \downarrow 0} (1, \boldsymbol{\mu}^{\epsilon}, \boldsymbol{\mu}^{\epsilon}\boldsymbol{\mu}^{\epsilon'} + \boldsymbol{\Sigma}^{\epsilon})$ where $(1, \boldsymbol{\mu}^{\epsilon}, \boldsymbol{\mu}^{\epsilon}\boldsymbol{\mu}^{\epsilon'} + \boldsymbol{\Sigma}^{\epsilon}) \in \mathcal{M}$ with $\boldsymbol{\mu}^{\epsilon} = \mathbb{E}_{\mathbb{P}_{\epsilon}}[\tilde{\boldsymbol{\xi}}]$ and $\boldsymbol{\Sigma}^{\epsilon} = \text{Cov}_{\mathbb{P}_{\epsilon}}(\tilde{\boldsymbol{\xi}})$ for $\epsilon \in (0, 1]$.

The next theorem provides a semidefinite characterization of the closure of the moment cone.

Theorem 5.2. *The point $(1, \boldsymbol{\mu}, \boldsymbol{\mu}\boldsymbol{\mu}' + \boldsymbol{\Sigma}) \in \text{cl}(\mathcal{M})$ if and only if there exists scalars $\lambda_{i,j}$ for $(i, j) \in \binom{[N]}{2}$ so that the following set of linear and semidefinite constraints are jointly satisfied:*

$$\begin{aligned} \lambda_{i,j} &\geq \mu_i\mu_j + \Sigma_{i,j}, & \forall (i, j) \in \binom{[N]}{2}, \\ \begin{pmatrix} 1 & \mu_i & \mu_j \\ \mu_i & \mu_i^2 + \Sigma_{i,i} & \lambda_{i,j} \\ \mu_j & \lambda_{i,j} & \mu_j^2 + \Sigma_{j,j} \end{pmatrix} &\succeq 0, & \forall (i, j) \in \binom{[N]}{2}, \\ \mu_i &\geq 0, & \forall i \in [N], \end{aligned} \quad (5.2)$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are specified. This is verifiable in polynomial time.

Proof. See Appendix A. □

Theorem 5.2 makes use of three by three positive semidefinite matrices to characterize the closure of the moment cone. A related moment cone is defined by:

$$\mathcal{M}_0 := \left\{ m_0(1, \mathbf{m}, \mathbf{M}) \mid \exists \mathbb{P} \in \mathcal{P}(\mathbb{R}_+^N) \text{ s.t. } \mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{\xi}}] = \mathbf{m}, \mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{\xi}}\tilde{\boldsymbol{\xi}}'] = \mathbf{M}, m_0 \geq 0 \right\}.$$

where we replace the term wise matrix inequality in (5.1) with an equality. The point $(1, \boldsymbol{\mu}, \boldsymbol{\mu}\boldsymbol{\mu}' + \boldsymbol{\Sigma}) \in$

$\text{cl}(\mathcal{M}_0)$ if and only if the following $(N+1) \times (N+1)$ moment matrix is completely positive:

$$\begin{pmatrix} 1 & \boldsymbol{\mu}' \\ \boldsymbol{\mu} & \boldsymbol{\mu}\boldsymbol{\mu}' + \boldsymbol{\Sigma} \end{pmatrix}.$$

Unfortunately, this is NP-hard to verify (see [21, 52]). Interestingly, Theorem 5.2 shows that when the equality is relaxed to an inequality, we obtain tractable conditions using multiple small positive semidefinite matrices rather than an intractable condition using a large completely positive matrix.

Now, consider the problem of computing the tightest upper bound on the expected value of a piecewise quadratic function of the random vector given the mean-covariance submodular ambiguity set as follows:

$$\rho^* := \sup_{\mathbb{P} \in \mathcal{P}_{\text{mean-cov-sub}}} \mathbb{E}_{\mathbb{P}} \left[\max_{k \in [K]} \left(\tilde{\boldsymbol{\xi}}' \mathbf{A}_k \tilde{\boldsymbol{\xi}} + \mathbf{b}_k' \tilde{\boldsymbol{\xi}} + b_k \right) \right]. \quad (5.3)$$

This brings us to the following theorem on computing the moment bound in (5.3) in polynomial time using semidefinite optimization.

Theorem 5.3. *Suppose the off-diagonal entries of the matrices $\mathbf{A}_k$ are nonnegative for each $k \in [K]$ and $(1, \boldsymbol{\mu}, \boldsymbol{\mu}\boldsymbol{\mu}' + \boldsymbol{\Sigma})$ satisfies the conditions in (5.2). Then ρ^* in (5.3) is given by the optimal value of the polynomial sized semidefinite program:*

$$\begin{aligned} \max_{\boldsymbol{\lambda}} \quad & \sum_{k \in [K]} \left(\sum_{(i,j) \in \binom{[N]}{2}} 2A_{i,j,k} \lambda_{i,j,k} + \sum_{i \in [N]} (A_{i,k} \lambda_{i,i,k} + b_{i,k} \lambda_{i,k}) + c_k \lambda_k \right) \\ \text{s.t.} \quad & \sum_{k \in [K]} \lambda_k = 1, \\ & \sum_{k \in [K]} \lambda_{i,k} = \mu_i, \quad \forall i \in [N], \\ & \sum_{k \in [K]} \lambda_{i,i,k} = \mu_i^2 + \Sigma_{i,i}, \quad \forall i \in [N], \\ & \sum_{k \in [K]} \lambda_{i,j,k} \geq \mu_i \mu_j + \Sigma_{i,j}, \quad \forall (i,j) \in \binom{[N]}{2}, \\ & \begin{pmatrix} \lambda_k & \lambda_{i,k} & \lambda_{j,k} \\ \lambda_{i,k} & \lambda_{i,i,k} & \lambda_{i,j,k} \\ \lambda_{j,k} & \lambda_{i,j,k} & \lambda_{j,j,k} \end{pmatrix} \succeq 0, \quad \forall (i,j) \in \binom{[N]}{2}, \forall k \in [K], \\ & \lambda_{i,k} \geq 0, \quad \forall i \in [N], \forall k \in [K]. \end{aligned} \quad (5.4)$$

Proof. See Appendix A. □

We make a few remarks about Theorem 5.3 and its implications next.

- (a) In comparison to Theorem 4.4 where probabilities are used as decision variables, the formulation in Theorem 5.3 makes use of moments as decision variables. The moment formulation is smaller in size and does not depend on the size of the marginal support. On the other hand, the formulation for the moment ambiguity set does not capture the marginal distribution information, when available.

The moment formulation can also exploit sparse information on the covariances in a manner similar to that discussed in remark (d) in Section 4.2 for the marginal-covariance submodular ambiguity set by enforcing the semidefinite constraints only for the specified pairs.

- (b) The ambiguity set $\mathcal{P}_{\text{mean-cov-sub}}$ in (3.14) can be equivalently expressed as:

$$\mathcal{P}_{\text{mean-cov-sub}} = \{\mathbb{P} \in \mathcal{P}(\mathbb{R}_+^N) \mid \mathbb{E}_{\mathbb{P}}[\tilde{\xi}_i] = \mu_i, \forall i \in [N], \mathbb{E}_{\mathbb{P}}[\tilde{\xi}_i^2] = \mu_i^2 + \Sigma_{i,i}, \forall i \in [N], \\ \mathbb{E}_{\mathbb{P}}[(\tilde{\xi}_i - \tilde{\xi}_j)^2] \leq \mu_i + \Sigma_{i,i} + \mu_j + \Sigma_{j,j} - 2\mu_i\mu_j - 2\Sigma_{i,j}, \forall (i,j) \in \binom{[N]}{2}\},$$

where besides a precise specification of the first and second moments of the random variables, upper bounds on pairwise Wasserstein distances are given. While it is a submodular ambiguity set, this is not a convex ambiguity set since the second moments of $\tilde{\xi}_i$ are precisely specified. If we relax this to $\mathbb{E}_{\mathbb{P}}[\tilde{\xi}_i^2] \leq \mu_i^2 + \Sigma_{i,i}$, only then, do we obtain a convex ambiguity set.

- (c) As in Section 4, we can use the previous theorem to solve the distributionally robust optimization problem described in (4.6) where the ambiguity set used is the mean-covariance submodular ambiguity set. In this case, by using standard duality arguments for semidefinite programs, we can solve the distributionally robust optimization problem.

5.2.1 Comparison with the mean covariance convex ambiguity set

In this section, we provide a comparison of the semidefinite program in (5.4) with another popular semidefinite programming formulation that has been developed for the ambiguity set $\mathcal{P}_{\text{mean-cov-cvx}}$ defined in (3.12) (see [20]). When the support $\Xi = \mathbb{R}_+^N$, the mean-covariance convex ambiguity set is nonempty if and only if the following conditions are satisfied:

$$\begin{pmatrix} 1 & \boldsymbol{\mu}' \\ \boldsymbol{\mu} & \boldsymbol{\mu}\boldsymbol{\mu}' + \boldsymbol{\Sigma} \end{pmatrix} \succeq 0 \\ \boldsymbol{\mu} \geq \mathbf{0}.$$

Feasibility can be verified using the Dirac measure $\delta_{\boldsymbol{\mu}}$ where $\boldsymbol{\mu} \geq \mathbf{0}$. Tightness of the moment bound in this case is given by:

$$\begin{aligned} \sup_{\mathbb{P} \in \mathcal{P}_{\text{mean-cov-cvx}}} \mathbb{E}_{\mathbb{P}} \left[\max_{k \in [K]} \left(\tilde{\boldsymbol{\xi}}' \mathbf{A}_k \tilde{\boldsymbol{\xi}} + \mathbf{b}_k' \tilde{\boldsymbol{\xi}} + c_k \right) \right] &= \max_{\boldsymbol{\lambda}, \boldsymbol{\Lambda}} \sum_{k \in [K]} \mathbf{A}_k \bullet \boldsymbol{\Lambda}_k + \sum_{k \in [K]} \mathbf{b}_k' \boldsymbol{\lambda}_k + \sum_{k \in [K]} c_k \lambda_k \\ \text{s.t.} \quad &\sum_{k \in [K]} \lambda_k = 1, \\ &\sum_{k \in [K]} \boldsymbol{\lambda}_k = \boldsymbol{\mu}, \\ &\sum_{k \in [K]} \boldsymbol{\Lambda}_k \preceq \boldsymbol{\mu}\boldsymbol{\mu}' + \boldsymbol{\Sigma}, \\ &\lambda_k \geq 0, \quad \forall k \in [K], \\ &\begin{pmatrix} \lambda_k & \boldsymbol{\lambda}_k' \\ \boldsymbol{\lambda}_k & \boldsymbol{\Lambda}_k \end{pmatrix} \succeq 0, \quad \forall k \in [K], \end{aligned} \tag{5.5}$$

when the matrices $\mathbf{A}_k$ are assumed to be negative semidefinite for equality to hold (see [20]). In comparing formulations (5.4) and (5.5), we observe that the newly proposed formulation uses $O(N^2K)$ positive semidefinite matrices of size 3×3 while the existing formulation uses $O(K)$ positive semidefinite matrices of size $(N+1) \times (N+1)$. The formulation (5.4) incorporates lower bounds on the covariances term by term and for tightness of the formulation, requires the matrices $\mathbf{A}_k$ to satisfy sign restrictions on the off-diagonal entries. On the other hand, formulation (5.5) specifies upper bounds in a positive semidefinite order and requires the matrices $\mathbf{A}_k$ to be negative semidefinite for tightness. In Table 1, we compare and contrast the two semidefinite programming formulations. Neither model nests the other as we show with an example in Appendix B.4. We provide numerical evidence in the next section comparing the two formulations.

	Convex ambiguity set	Submodular ambiguity set
Moments	$\mathbb{E}[\tilde{\xi}] = \boldsymbol{\mu}, \text{Cov}_{\mathbb{P}}(\tilde{\xi}) \preceq \boldsymbol{\Sigma}$	$\mathbb{E}[\tilde{\xi}] = \boldsymbol{\mu}, \text{diag}(\text{Cov}_{\mathbb{P}}(\tilde{\xi})) = \text{diag}(\boldsymbol{\Sigma}), \text{Cov}_{\mathbb{P}}(\tilde{\xi}) \geq \boldsymbol{\Sigma}$
Feasibility	1 SDP matrix of size $O(N \times N)$	$O(N^2)$ SDP matrices of size 3×3
Tightness	Each $\mathbf{A}_k$ is negative semidefinite	Each $\mathbf{A}_k$ has nonnegative off diagonal entries
Size	$O(K)$ SDP matrices of size $O(N \times N)$	$O(N^2K)$ SDP matrices of size 3×3
Moment bound	Can be tighter in some cases	Can be tighter in some cases
Solution time	Slower	Faster
Sparsity	Does not exploit	Can be exploited

Table 1: Comparison of mean-covariance ambiguity sets for $\Xi = \mathbb{R}_+^N$.

6 Numerical Experiments

In the numerical experiments, we focus on the resource allocation model introduced in Section 1.1. The distributionally robust resource allocation problem that we solve is:

$$\begin{aligned}
& \inf \quad \alpha \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}} \left[\max_{i \in [N]} |x_i - \tilde{\xi}_i| \right] + (1 - \alpha) \frac{1}{N} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}} \left[\sum_{i \in [N]} |x_i - \tilde{\xi}_i| \right] \\
& \text{s.t.} \quad \sum_{i \in [N]} x_i \leq \text{Budget} \\
& \quad \mathbf{x} \geq \mathbf{0},
\end{aligned} \tag{6.1}$$

where resource allocation vector is chosen in the budget constrained set $\mathcal{X} = \{\mathbf{x} \geq \mathbf{0} \mid \sum_{i \in [N]} x_i \leq \text{Budget}\}$. The demand at each location $i \in [N]$ is random and is denoted by $\tilde{\xi}_i$. In contrast to the existing literature in stochastic and distributionally robust optimization which primarily focuses on minimizing the worst-case expected profit that is captured by the second term in the objective in (6.1), we focus on the effect of incorporating the Rawlsian fairness criterion into the model (see [32]). We consider the following ambiguity sets in the numerical experiments:

- (i) Marginal distribution ambiguity set: Set of distributions of the demand vector where the marginal

distributions are fixed:

$$\mathcal{P}_{\text{marg}} := \{\mathbb{P} \in \mathcal{P}(\prod_{i \in [N]} \Xi_i) \mid \mathbb{P}(\tilde{\xi}_i = \xi_i) = p_i(\xi_i), \forall \xi_i \in \Xi_i, \forall i \in [N]\}. \quad (6.2)$$

- (ii) Marginal-covariance submodular ambiguity set: Set of distributions $\mathcal{P}_{\text{marg-cov-sub}}$ in (3.9) where the marginal distributions of the demands are fixed and in addition lower bounds on covariances are specified.
- (iii) Mean-variance ambiguity set: Set of distributions of the random demand where the means and the variances are fixed:

$$\mathcal{P}_{\text{mean-var}} := \{\mathbb{P} \in \mathcal{P}(\mathbb{R}_+^N) \mid \mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{\xi}}] = \boldsymbol{\mu}, \text{diag}(\text{Cov}_{\mathbb{P}}(\tilde{\boldsymbol{\xi}})) = \text{diag}(\boldsymbol{\Sigma})\}. \quad (6.3)$$

- (iv) Mean-covariance submodular ambiguity set: Set of distributions $\mathcal{P}_{\text{mean-cov-sub}}$ in (3.14) where the mean vector $\boldsymbol{\mu}$ and the variance of nonnegative demands are fixed and in addition lower bounds on covariances are specified.
- (v) Mean-covariance convex ambiguity set: $\mathcal{P}_{\text{mean-cov-cvx}}$ in (3.12) where the means of the nonnegative demands are fixed and an upper bound on the covariance matrix in the positive semidefinite order is specified.

6.1 Impact of covariance on fairness objective

In the first set of experiments, we compare the worst-case expected fairness objective obtained by solving the inner problem for $\alpha = 1$ for the ambiguity sets for fixed resource allocation vectors $\mathbf{x}$. We vary the number of locations N in $\{2, 5, 10\}$. The demand at the locations are assumed to have identical marginal distributions for the ambiguity sets in (i) and (ii) and identical means and identical standard deviations for the ambiguity sets in (iii), (iv) and (v). We generate 30 random instances for each N . Each instance is generated by setting the mean demand μ_i to a fixed value μ which is uniformly sampled in the range $[0, 20/N]$. The standard deviation σ_i across locations $i \in [N]$ is fixed to a common σ which is sampled uniformly in the range $[0, \mu]$. The marginal distributions are generated as two point distributions of the form

$$\tilde{\xi}_i = \begin{cases} \mu - \sigma \sqrt{\frac{p}{1-p}}, & \text{with probability } 1-p, \\ \mu + \sigma \sqrt{\frac{1-p}{p}}, & \text{with probability } p, \end{cases}$$

where p is sampled uniformly in the range $[0, \mu^2/(\mu^2 + \sigma^2)]$ to ensure that the demand is nonnegative. We set the lower bounds on the correlation to be identical across all demands for the ambiguity sets (ii) and (iv) and vary it in the set $\{0, 0.3, 0.5\}$. We generate instances such that the ambiguity sets are nonempty. We then compute the worst-case expected fairness objective for five different resource allocation vectors where all the x_i values are set to a fixed x and varied in the set $\{0, \mu/2, \mu, 3\mu/2, 2\mu\}$. We compute the ratio of the worst-case fairness objective for the marginal-covariance submodular ambiguity set to

the the marginal ambiguity set. Similarly, we compute the ratio of the worst-case fairness objective for the mean-covariance submodular ambiguity set to the the mean-variance ambiguity set. The ratios as defined are between 0 and 1. In Tables 2 and 3, we provide statistics on these ratios by reporting the minimum, average and maximum value across the 30 instances for every (N, Corr, x) tuple. The

N	Corr	$x = 0$	$x = \mu/2$	$x = \mu$	$x = 3\mu/2$	$x = 2\mu$
2	0	[0.827,0.945,1]	[0.793,0.925,1]	[0.933,0.962,1]	[0.783,0.932,1]	[0.827,0.950,1]
	0.3	[0.755,0.906,1]	[0.721,0.867,0.991]	[0.881,0.910,1]	[0.705,0.890,1]	[0.755,0.920,1]
	0.5	[0.707,0.881,1]	[0.673,0.831,1]	[0.832,0.872,1]	[0.653,0.863,1]	[0.707,0.901,1]
5	0	[0.900,0.987,1]	[0.876,0.982,1]	[0.900,0.976,1]	[0.937,0.992,1]	[0.960,0.998,1]
	0.3	[0.744,0.940,1]	[0.684,0.916,1]	[0.745,0.882,1]	[0.793,0.961,1]	[0.847,0.982,1]
	0.5	[0.640,0.897,1]	[0.556,0.855,1]	[0.642,0.802,1]	[0.667,0.922,1]	[0.754,0.959,1]
10	0	[0.974,0.997,1]	[0.963,0.995,1]	[0.944,0.994,1]	[0.976,0.998,1]	[0.982,0.999,1]
	0.3	[0.786,0.962,1]	[0.741,0.948,1]	[0.735,0.929,1]	[0.802,0.974,1]	[0.862,0.986,1]
	0.5	[0.659,0.921,1]	[0.591,0.896,1]	[0.581,0.858,1]	[0.685,0.947,1]	[0.783,0.974,1]

Table 2: Minimum, average and maximum of the ratio of the worst-case expected fairness objective with the marginal-covariance submodular ambiguity set to by the worst-case expected fairness objective with the marginal distribution ambiguity set.

N	Corr	$x = 0$	$x = \mu/2$	$x = \mu$	$x = 3\mu/2$	$x = 2\mu$
2	0	[0.855,0.913,0.990]	[0.941,0.947,0.983]	[1,1,1]	[0.943,0.965,0.996]	[0.943,0.954,0.991]
	0.3	[0.797,0.878,0.987]	[0.905,0.918,0.976]	[0.988,0.989,0.990]	[0.916,0.940,0.977]	[0.916,0.932,0.987]
	0.5	[0.752,0.851,0.984]	[0.873,0.893,0.970]	[0.966,0.967,0.973]	[0.892,0.917,0.970]	[0.892,0.912,0.984]
5	0	[0.868,0.947,0.993]	[0.921,0.949,0.988]	[0.988,0.998,1]	[0.958,0.975,1]	[0.935,0.978,1]
	0.3	[0.833,0.883,0.984]	[0.823,0.879,0.971]	[0.934,0.961,0.970]	[0.896,0.929,0.982]	[0.892,0.941,0.988]
	0.5	[0.756,0.832,0.977]	[0.753,0.824,0.957]	[0.881,0.909,0.924]	[0.842,0.886,0.957]	[0.852,0.908,0.977]
10	0	[0.962,0.973,0.995]	[0.956,0.967,0.991]	[0.966,0.991,1]	[0.974,0.990,1]	[0.978,0.993,1]
	0.3	[0.846,0.890,0.981]	[0.824,0.870,0.966]	[0.860,0.923,0.951]	[0.891,0.937,0.975]	[0.910,0.956,0.985]
	0.5	[0.754,0.824,0.970]	[0.719,0.794,0.946]	[0.777,0.850,0.883]	[0.821,0.883,0.946]	[0.855,0.918, 0.970]

Table 3: Minimum, average and maximum of the ratio of the worst-case expected fairness objective with the mean-covariance submodular ambiguity set to by the worst-case expected fairness objective with the mean-variance submodular ambiguity set.

results in the tables show that as we increase the lower bounds on the correlations in the ambiguity sets, the ratios tend to increases, the ratios tends to decrease. This is true across both the marginal distribution and moment ambiguity sets. This indicates that if we know demands are more positively correlated, the information can be captured by the submodular ambiguity sets in providing improved estimates on the fairness objective. We provide theoretical bounds in the next section which provides a characterization of this improvement. The results for the mean-covariance convex ambiguity set was almost identical to the mean-covariance submodular ambiguity set in these set of instances and so we do not report it here.

6.2 Fairness-efficiency tradeoff

In the second set of experiments, we plot the fairness-efficiency tradeoff for the different ambiguity sets as α is varied in $[0, 1]$ for $N = 10$. For the moment ambiguity sets, we set the means as $\mu_i = i/N$ and standard deviations as $\sigma_i = i/(2N)$ (standard deviation increases as mean increases) for all $i \in [N]$. The budget was set as $\text{Budget} = C \sum_i \mu_i$ with $C \in \{1, 0.8\}$ and the lower bound on the correlation was varied in $\{0, 0.3, 0.5\}$. For the ambiguity sets that make use of the marginal distributions, we fit a three point distribution to ensure that the means and standard deviation match as follows:

$$\tilde{\xi}_i = \begin{cases} \mu_i - \frac{\sigma_i}{5}, & \text{with probability } \frac{25}{41}, \\ \mu_i, & \text{with probability } \frac{3}{8}, \\ \mu_i + 8\sigma_i, & \text{with probability } \frac{5}{328}. \end{cases}$$

The fairness-efficiency tradeoff curves are plotted in Figure 3 and illustrate the value of incorporating the covariance information. The figure illustrate that by modeling correlation information in the submodular ambiguity sets, we shift the fairness-utility tradeoff curve to the bottom left; namely we obtain a better fairness-utility cost tradeoff by incorporating the correlation information as compared to not incorporating the correlation information. Secondly, as the lower bounds on the correlation increases (demands are known to be more positively correlated), the improvement provided by incorporating it into the ambiguity set is more significant. Lastly, as the budget reduces and we are in a more resource constrained setting, the fairness-efficiency tradeoff shifts to the top right as should be expected. These results are consistent across the marginal distribution and moment ambiguity sets.

6.3 Computational times

In the third set of experiments, we compare the computational times of the mean-covariance submodular and convex ambiguity sets with sparse information. All computations were run in MATLAB with SDPT3 and an YALMIP interface. We varied N in $\{30, 40, 50\}$ and generated 10 random instance for each by N . The data was generated by choosing μ_i uniformly and independently in the range $[0, 1]$ and likewise σ_i uniformly and independently in the range $[0, 1]$. We set the correlation information in the ambiguity sets to 0.3. As the results show the computational times grow most rapidly for the mean-covariance-convex ambiguity set, then less rapidly for the mean-covariance-submodular ambiguity with the least time taken for the mean-variance ambiguity sets. In these experiments, the bounds obtained by the submodular and convex ambiguity sets were again very close to each other. This table shows that the submodular ambiguity sets can exploit sparse covariance information structures in comparison to the convex ambiguity sets.

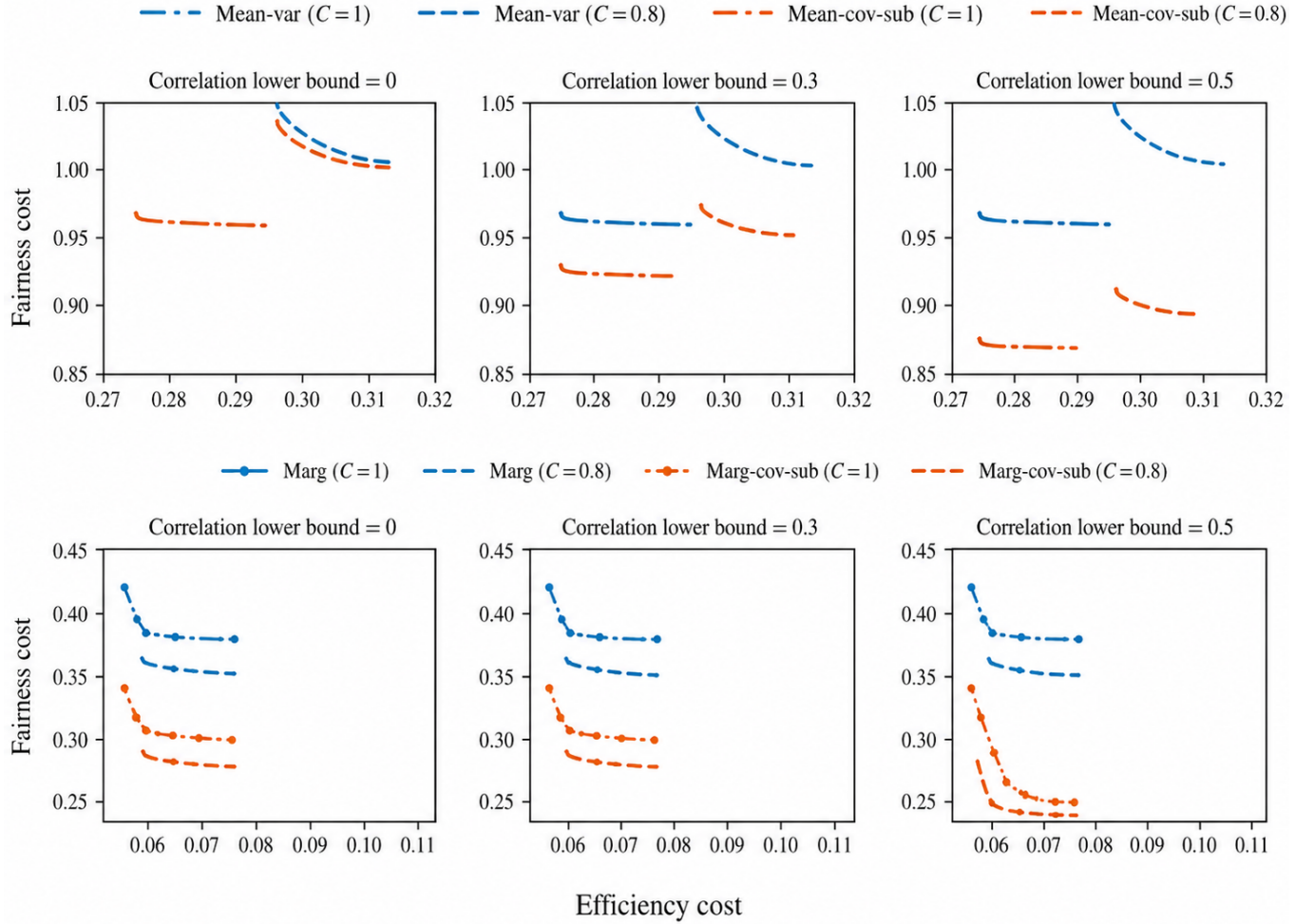


Figure 3: Fairness-efficiency tradeoff for all $\alpha \in [0, 1]$.

N	Mean-var	Mean-cov-sub	Mean-cov-cvx
30	[2.776, 4.283, 6.330]	[23.839, 36.318, 47.565]	[41.620, 56.828, 66.434]
40	[8.028, 9.806, 10.914]	[97.260, 115.329, 128.111]	[238.2958, 256.585, 269.831]
50	[9.057, 10.652, 17.863]	[123.502, 136.254, 148.483]	[515.449, 587.895, 661.349]

Table 4: Minimum, average and maximum computational run time in seconds.

7 Value of incorporating covariance information

In this section, we provide theoretical bounds in stylized settings to quantify the improvement provided by the submodular ambiguity set that incorporates covariance information. Our approach is inspired by prior work on the “correlation gap” analysis which compares the expected value of a function of a random vector with the independent distribution to the worst-case expected value computed over the marginal distribution ambiguity set (see [1]). Here we move from the stronger assumption of mutually independent random variables to the weaker assumption that the covariances of the random variables are nonnegative which can be modeled with the submodular ambiguity set (see [49, 50] for related work).

We focus on the objective given by the maximum of the random variables (that appears in the Rawlsian fairness criterion):

$$\max_{i \in [N]} \tilde{\xi}_i, \quad (7.1)$$

where $\tilde{\xi}_1, \dots, \tilde{\xi}_N$ is a set of nonnegative random variables with an arbitrary identical marginal distribution function F with a finite expected value. The asymptotics of the maximum of identically distributed random variables has been extensively studied under the assumption of independence in extreme value theory (see [19]) and under maximal dependence with the marginal distribution ambiguity set in the works of [37] and [53]. A key finding in [37, 53] is that under appropriate assumptions on the tail behavior of F , as $N \uparrow \infty$, the expected value of the maximum behaves similarly under independence and maximal dependence among the random variables (see theorem 3 in [37]). A related finding that compares the closeness of these values for any finite N is provided by the correlation gap bound in [1] (see [15] for a more refined analysis) where as a special case they compare the following two values:

- (i) Marginal distribution ambiguity set: With maximal dependence:

$$\rho_{\text{marg}}^* := \sup \left\{ \mathbb{E}_{\mathbb{P}} \left[\max_{i \in [N]} \tilde{\xi}_i \right] \mid \tilde{\xi}_i \sim F, \forall i \in [N] \right\}. \quad (7.2)$$

- (ii) Independence: With the mutually independent distribution $\mathbb{P}_{\text{ind}}$ where each $\tilde{\xi}_i \sim F$:

$$\rho_{\text{ind}}^* := \mathbb{E}_{\mathbb{P}_{\text{ind}}} \left[\max_{i \in [N]} \tilde{\xi}_i \right] \quad (7.3)$$

Clearly $0 \leq \rho_{\text{ind}}^* \leq \rho_{\text{marg}}^*$ since the independent distribution is a feasible distribution in (7.2) and the random variables are nonnegative. A key result¹ (see theorem 4 in [1], lemma 4.10 in [15]) is that the ratio of ρ_{ind}^* and ρ_{marg}^* is lower bounded by:

$$\frac{\rho_{\text{ind}}^*}{\rho_{\text{marg}}^*} \geq 1 - \left(1 - \frac{1}{N} \right)^N \geq 1 - \frac{1}{e}.$$

Furthermore, the bound is sharp for each N and is attained by identically distributed Bernoulli random

¹The results are developed in [1] and [15] for nonnegative, monotone submodular functions of which objective in (7.1) is a special case

variables with $\mathbb{P}(\tilde{\xi}_i = 1) = 1 - \mathbb{P}(\tilde{\xi}_i = 0) = \frac{1}{N}$ (see [14]). This result shows that the two models of distributional information (one with maximal dependence and the other with independence) are not too far apart for this specific objective function. Here we consider the value of incorporating covariance information in the submodular ambiguity set as follows:

- (iii) Marginal-covariance submodular ambiguity set: With given marginal F and lower bounds on covariances:

$$\rho_{\text{marg-cov-sub}}^* := \sup \left\{ \mathbb{E}_{\mathbb{P}} \left[\max_{i \in [N]} \tilde{\xi}_i \right] \mid \tilde{\xi}_i \sim F, \forall i \in [N], \text{Cov}(\tilde{\xi}_i, \tilde{\xi}_j) \geq \Sigma_{i,j}, \forall (i, j) \in \binom{[N]}{2} \right\}. \quad (7.4)$$

When the ambiguity set is nonempty, we have $0 \leq \rho_{\text{marg-cov-sub}}^* \leq \rho_{\text{marg}}^*$. This brings us to the following theorem which bounds the ratio.

Theorem 7.1. *Given nonnegative random variables with a finite expected value:*

- (i) *When the marginal distributions are identical and $\Sigma_{i,j} = 0$ for all $(i, j) \in \binom{[N]}{2}$, we have:*

$$\frac{\rho_{\text{marg-cov-sub}}^*}{\rho_{\text{marg}}^*} \geq 1 - \frac{1}{N} + \frac{1}{N^2} \geq \frac{3}{4}. \quad (7.5)$$

Furthermore the lower bound is sharp for each N .

- (ii) *When the marginals are identical and $\Sigma_{i,j}$ is chosen arbitrarily, there exists an instance where as $N \uparrow \infty$:*

$$\frac{\rho_{\text{marg-cov-sub}}^*}{\rho_{\text{marg}}^*} \downarrow 0. \quad (7.6)$$

- (iii) *With moment ambiguity sets, given identical means and identical variances, the lower bounds on the ratio in (7.5) and (7.6) continues to hold for $\Sigma_{i,j} = 0$ and arbitrary $\Sigma_{i,j}$ respectively.*

Proof. See Appendix A. □

8 Conclusion

In this paper, we have introduced a submodular ambiguity set and showcased the modeling of uncertainty using it. With discrete and continuous uncertainty, we have developed linear and semidefinite programs that can be used to incorporate univariate marginal and moment information and bivariate covariance information. The paper shows that the submodular ambiguity set is the natural discrete counterpart of the convex ambiguity set and supplements it for continuous uncertainty, both in modeling and computation. There are several research directions to pursue along this line of work, Firstly, it will be interesting to test the use of decision rules to solve two stage optimization problems under uncertainty with submodular ambiguity sets or to find new decision rules in this context. Secondly, can one identify instances of two stage distributionally robust optimization problems with a submodular ambiguity set that are solvable in polynomial time? Lastly exploring new algorithms for other structured submodular ambiguity sets in applications is a natural direction to consider. We leave this for future research.

Appendix

Appendix A

Proof of Theorem 4.1

Let $\Xi := \prod_{i \in [N]} \Xi_i$. Under assumptions (A1) and (A4), we can reformulate (4.2) as a finite sized linear program:

$$\begin{aligned} \rho^* = \max \quad & \sum_{\boldsymbol{\xi} \in \Xi} \left(\max_{k \in [K]} g_k(\boldsymbol{\xi}) \right) p(\boldsymbol{\xi}) \\ \text{s.t.} \quad & \sum_{\boldsymbol{\xi} \in \Xi} f_l(\boldsymbol{\xi}) p(\boldsymbol{\xi}) \leq \gamma_l, \quad \forall l \in [L], \\ & \sum_{\boldsymbol{\xi} \in \Xi} p(\boldsymbol{\xi}) = 1, \\ & p(\boldsymbol{\xi}) \geq 0, \quad \forall \boldsymbol{\xi} \in \Xi, \end{aligned}$$

where the decision variables are $p(\boldsymbol{\xi}) = \mathbb{P}(\tilde{\boldsymbol{\xi}} = \boldsymbol{\xi})$ for $\boldsymbol{\xi} \in \Xi$. We call this the primal linear program. In the input size, it has a polynomial number of constraints and an exponential number of nonnegative decision variables. The dual linear program is formulated as:

$$\begin{aligned} \rho_d^* := \min \quad & y_0 + \sum_{l \in [L]} y_l \gamma_l \\ \text{s.t.} \quad & y_0 + \sum_{l \in [L]} y_l f_l(\boldsymbol{\xi}) \geq \max_{k \in [K]} g_k(\boldsymbol{\xi}), \quad \forall \boldsymbol{\xi} \in \Xi, \\ & y_l \geq 0, \quad \forall l \in [L], \end{aligned}$$

where the decision variables are y_0, y_l for $l \in [L]$ and ρ_d^* is the optimal dual objective. The dual linear program has a polynomial number of variables and an exponential number of constraints. The primal linear program is feasible under assumption (A3) and the dual linear program is feasible (set $y_l = 0$ for $l \in [L]$ and $y_0 = \max_{\boldsymbol{\xi} \in \Xi} \max_{k \in [K]} g_k(\boldsymbol{\xi})$). Strong duality of linear programming guarantees that $\rho^* = \rho_d^*$ and the optimal value is finite since the primal and dual linear programs are feasible. The separation problem for the dual linear program is given by:

Given scalars $\bar{y}_0$ and $\bar{y}_l \geq 0$ for $l \in [L]$, decide whether

$$\bar{y}_0 + \sum_{l \in [L]} \bar{y}_l f_l(\boldsymbol{\xi}) \geq \max_{k \in [K]} g_k(\boldsymbol{\xi}), \quad \forall \boldsymbol{\xi} \in \Xi,$$

and if the answer is no, find a separating hyperplane.

This reduces to checking if

$$\bar{y}_0 + \sum_{l \in [L]} \bar{y}_l f_l(\boldsymbol{\xi}) \geq g_k(\boldsymbol{\xi}), \quad \forall \boldsymbol{\xi} \in \Xi, \forall k \in [K],$$

which in turn reduces to checking if

$$\bar{y}_0 + \min_{\boldsymbol{\xi} \in \Xi} \left(\sum_{l \in [L]} \bar{y}_l f_l(\boldsymbol{\xi}) - g_k(\boldsymbol{\xi}) \right) \geq 0, \forall k \in [K].$$

Since $\bar{y}_l \geq 0$ for all $l \in [L]$ and f_l is a submodular function for each $l \in [L]$ (assumption (A2)) and $g_k(\boldsymbol{\xi})$ is a supermodular function (assumption (A4)), the function $\sum_{l \in [L]} \bar{y}_l f_l(\boldsymbol{\xi}) - g_k(\boldsymbol{\xi})$ is submodular for each $k \in [K]$; see Appendix B. Hence to solve the separation problem, we need to solve K submodular function minimization problems where all the functions have efficient evaluation oracles. If we find an index k^* and a vector $\boldsymbol{\xi}^*$ such that one of the conditions above is violated, namely:

$$\bar{y}_0 + \sum_{l \in [L]} \bar{y}_l f_l(\boldsymbol{\xi}^*) - g_{k^*}(\boldsymbol{\xi}^*) < 0,$$

then a separating hyperplane given by:

$$y_0 + \sum_{l \in [L]} y_l f_l(\boldsymbol{\xi}^*) - g_{k^*}(\boldsymbol{\xi}^*) \geq 0.$$

Since all these steps can be done in polynomial time, the dual separation problem is solvable in time polynomial in N , K , L , $\max_{i \in [N]} |\Xi_i|$ and the evaluation time of the oracles. From the equivalence of separation and optimization (see [29]), the dual linear program is solvable in polynomial time. Hence ρ^* is computable in polynomial time. $\square$

Proof of Theorem 4.2

Let $\Xi := \prod_{i \in [N]} \Xi_i$. The distributionally robust optimization problem is reformulated as:

$$\begin{aligned} \inf \quad & y_0 + \sum_{l \in [L]} y_l \gamma_l \\ \text{s.t.} \quad & y_0 + \sum_{l \in [L]} y_l f_l(\boldsymbol{\xi}) \geq \max_{k \in [K]} g_k(\boldsymbol{x}, \boldsymbol{\xi}), \quad \forall \boldsymbol{\xi} \in \Xi, \\ & y_l \geq 0, \quad \forall l \in [L], \\ & \boldsymbol{x} \in \mathcal{X}, \end{aligned}$$

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

Given scalars $\bar{y}_0$ and $\bar{y}_l \geq 0$ for $l \in [L]$ and vector $\bar{\boldsymbol{x}}$, decide whether

$$\bar{\boldsymbol{x}} \in \mathcal{X} \text{ and } \bar{y}_0 + \sum_{l \in [L]} \bar{y}_l f_l(\boldsymbol{\xi}) \geq \max_{k \in [K]} g_k(\bar{\boldsymbol{x}}, \boldsymbol{\xi}), \forall \boldsymbol{\xi} \in \Xi,$$

and if the answer is no, find a separating hyperplane.

Verifying the feasibility of $\bar{\boldsymbol{x}} \in \mathcal{X}$ and if it is not feasible, providing a separating hyperplane can be done in polynomial time. The separation problem hence reduces to:

Given scalars $\bar{y}_0, \bar{y}_l \geq 0$ for $l \in [L]$ and vector $\bar{\mathbf{x}} \in \mathcal{X}$, decide whether

$$\bar{y}_0 + \min_{\boldsymbol{\xi} \in \Xi} \left(\sum_{l \in [L]} \bar{y}_l f_l(\boldsymbol{\xi}) - g_k(\bar{\mathbf{x}}, \boldsymbol{\xi}) \right) \geq 0, \forall k \in [K].$$

We need to solve a set of K submodular function minimization problems to solve the separation problem. If all the left hand side values are nonnegative, we are done. Else we find an index $k^* \in [K]$ and a vector $\boldsymbol{\xi}^* \in \Xi$ such that:

$$\bar{y}_0 + \sum_{l \in [L]} \bar{y}_l f_l(\boldsymbol{\xi}^*) - g_{k^*}(\bar{\mathbf{x}}, \boldsymbol{\xi}^*) < 0.$$

Since $g_{k^*}(\mathbf{x}, \boldsymbol{\xi}^*)$ is convex in $\mathbf{x}$ for a fixed $\boldsymbol{\xi}^*$, a separating hyperplane is given by:

$$y_0 + \sum_{l \in [L]} y_l f_l(\boldsymbol{\xi}^*) - g_{k^*}(\bar{\mathbf{x}}, \boldsymbol{\xi}^*) - \nabla_{\bar{\mathbf{x}}} g_{k^*}(\bar{\mathbf{x}}, \boldsymbol{\xi}^*)'(\mathbf{x} - \bar{\mathbf{x}}) \geq 0,$$

where $\nabla_{\bar{\mathbf{x}}} g_{k^*}(\bar{\mathbf{x}}, \boldsymbol{\xi}^*)$ is the subgradient of $g_{k^*}(\mathbf{x}, \boldsymbol{\xi}^*)$ with respect to $\mathbf{x}$ evaluated at $\bar{\mathbf{x}}$. Since all steps can be done in polynomial time, the distributionally robust optimization problem is solvable in polynomial time. $\square$

Proof of Theorem 4.3

Given a distribution $\mathbb{P}$ in the ambiguity set, let the bivariate probabilities be defined as:

$$\lambda_{i,j}(\xi_i, \xi_j) = \mathbb{P} \left(\tilde{\xi}_i = \xi_i, \tilde{\xi}_j = \xi_j \right), \quad \forall \xi_i \in \Xi_i, \forall \xi_j \in \Xi_j, \forall (i, j) \in \binom{[N]}{2}.$$

We use these probability variables to characterize feasibility of the ambiguity set.

Step (1): Necessity of linear constraints in (4.4)

Clearly the bivariate probabilities variables as defined must be nonnegative for any distribution. In addition, the specified univariate probabilities $p_i(\xi_i)$ in the ambiguity set must also be nonnegative. The other linear constraints are derived as follows:

1. Total sum of the marginal probabilities is one:

$$\sum_{\xi_i \in \Xi_i} \mathbb{P} \left(\tilde{\xi}_i = \xi_i \right) = 1.$$

2. Law of total expectation for the lower bound on the bivariate supermodular function:

$$\sum_{\xi_i \in \Xi_i} \sum_{\xi_j \in \Xi_j} \xi_i \xi_j \mathbb{P} \left(\tilde{\xi}_i = \xi_i, \tilde{\xi}_j = \xi_j \right) = \mathbb{E} \left[\tilde{\xi}_i \tilde{\xi}_j \right] \geq \mu_i \mu_j + \Sigma_{i,j},$$

where μ_i is defined in (3.10).

3. Consistency of bivariate marginal of $(\tilde{\xi}_i, \tilde{\xi}_j)$ with univariate marginal of $\tilde{\xi}_j$:

$$\sum_{\xi_i \in \Xi_i} \mathbb{P}(\tilde{\xi}_i = \xi_i, \tilde{\xi}_j = \xi_j) = \mathbb{P}(\tilde{\xi}_j = \xi_j).$$

4. Consistency of bivariate marginal of $(\tilde{\xi}_i, \tilde{\xi}_j)$ with univariate marginal of $\tilde{\xi}_i$:

$$\sum_{\xi_j \in \Xi_j} \mathbb{P}(\tilde{\xi}_i = \xi_i, \tilde{\xi}_j = \xi_j) = \mathbb{P}(\tilde{\xi}_i = \xi_i).$$

Step (2): Sufficiency of linear constraints in (4.4)

To show sufficiency, let $\lambda_{i,j}(\xi_i, \xi_j)$ for all $\xi_i \in \Xi_i$, $\xi_j \in \Xi_j$, $(i, j) \in \binom{[N]}{2}$ satisfy the linear constraints in (4.4). We construct a comonotonic distribution $\mathbb{P}^c$ in the ambiguity set where the marginal probabilities are given by $p_i(\xi_i)$ for $\xi_i \in \Xi_i$, $i \in [N]$. Now let $\mathbb{P}_{i,j}$ denote the bivariate distribution of the random variables $(\tilde{\xi}_i, \tilde{\xi}_j)$ given by:

$$\mathbb{P}_{i,j}(\tilde{\xi}_i = \xi_i, \tilde{\xi}_j = \xi_j) = \lambda_{i,j}(\xi_i, \xi_j), \quad \forall \xi_i \in \Xi_i, \forall \xi_j \in \Xi_j.$$

Since $\sum_{\xi_i \in \Xi_i} \sum_{\xi_j \in \Xi_j} \lambda_{i,j}(\xi_i, \xi_j) = 1$ holds in (4.4) with $\lambda_{i,j}(\xi_i, \xi_j) \geq 0$ for all $\xi_i \in \Xi_i, \xi_j \in \Xi_j$, $\mathbb{P}_{i,j}$ is a valid bivariate distribution. Furthermore, the univariate marginal distributions of $\mathbb{P}^c$ and $\mathbb{P}_{i,j}$ match for the random variables $\tilde{\xi}_i$ and $\tilde{\xi}_j$ for each $(i, j) \in \binom{[N]}{2}$ from the consistency conditions. However $\mathbb{P}_{i,j}$ might not be a bivariate comonotonic distribution, namely it is possible that $\mathbb{P}_{i,j} \neq \text{proj}_{i,j}(\mathbb{P}^c)$. The expected value of the bivariate supermodular function $\mathbb{E}[\tilde{\xi}_i \tilde{\xi}_j]$ in the comonotonic distribution $\mathbb{P}^c$ satisfies:

$$\begin{aligned} \mathbb{E}_{\mathbb{P}^c}[\tilde{\xi}_i \tilde{\xi}_j] &= \mathbb{E}_{\text{proj}_{i,j}(\mathbb{P}^c)}[\tilde{\xi}_i \tilde{\xi}_j] \\ &\quad [\text{since the function is bivariate}] \\ &\geq \mathbb{E}_{\mathbb{P}_{i,j}}[\tilde{\xi}_i \tilde{\xi}_j] \\ &\quad [\text{from (3.5) since } \xi_i \xi_j \text{ is submodular and the univariate marginals of } \\ &\quad \text{proj}_{i,j}(\mathbb{P}^c) \text{ and } \mathbb{P}_{i,j} \text{ are the same}] \\ &\geq \mu_i \mu_j + \Sigma_{i,j} \\ &\quad [\text{from (4.4)}]. \end{aligned}$$

Hence $\mathbb{P}^c$ is a feasible distribution in the marginal-covariance submodular ambiguity set. $\square$

Proof of Theorem 4.4

The proof builds on the characterization in Theorem 4.3 using conditional probability measures. Let ρ_u^* be optimal value of the linear program (4.5). We prove $\rho^* = \rho_u^*$ in two steps by showing $\rho^* \leq \rho_u^*$ and $\rho^* \geq \rho_u^*$.

Step (1): $\rho^ \leq \rho_u^*$*

To show ρ_u^* is a valid upper bound on ρ^* , we start with a probabilistic construction of the formulation (4.5). Define for each possible realization $\xi \in \prod_{i \in [N]} \Xi_i$, the set of indices that attain the maximum in

the objective function as follows:

$$K(\boldsymbol{\xi}) = \arg \max \{ \boldsymbol{\xi}' \mathbf{A}_k \boldsymbol{\xi} + \mathbf{b}'_k \boldsymbol{\xi} + b_k \mid k \in [K] \}.$$

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 the probabilities:

$$\begin{aligned} \lambda_k &= \mathbb{P} \left(k(\tilde{\boldsymbol{\xi}}) = k \right), & \forall k \in [K], \\ \lambda_{i,k}(\xi_i) &= \mathbb{P} \left(\tilde{\xi}_i = \xi_i, k(\tilde{\boldsymbol{\xi}}) = k \right), & \forall \xi_i \in \Xi_i, \forall i \in [N], \forall k \in [K], \\ \lambda_{i,j,k}(\xi_i, \xi_j) &= \mathbb{P} \left(\tilde{\xi}_i = \xi_i, \tilde{\xi}_j = \xi_j, k(\tilde{\boldsymbol{\xi}}) = k \right), & \forall \xi_i \in \Xi_i, \forall \xi_j \in \Xi_j, \forall (i, j) \in \binom{[N]}{2}, \forall k \in [K]. \end{aligned}$$

Clearly the variables as defined must satisfy the nonnegativity constraints in (4.5). The first two sets of constraints in the formulation are derived from:

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 index being optimal:

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

The third and fourth set of constraints are obtained from the law of the total expectation for the univariate marginal probabilities and bivariate cross moment terms while the fifth and sixth set of constraints arise from the consistency conditions (see Theorem 4.3). The objective in (4.5) is obtained by expressing the expected function value in terms of the decision variables:

$$\begin{aligned} \mathbb{E}_{\mathbb{P}} \left[\max_{k \in [K]} \left(\tilde{\boldsymbol{\xi}}' \mathbf{A}_k \tilde{\boldsymbol{\xi}} + \mathbf{b}'_k \tilde{\boldsymbol{\xi}} + c_k \right) \right] &= \sum_{k \in [K]} \mathbb{E}_{\mathbb{P}} \left[\left(\tilde{\boldsymbol{\xi}}' \mathbf{A}_k \tilde{\boldsymbol{\xi}} + \mathbf{b}'_k \tilde{\boldsymbol{\xi}} + c_k \right) \mid k(\tilde{\boldsymbol{\xi}}) = k \right] \mathbb{P} \left(k(\tilde{\boldsymbol{\xi}}) = k \right) \\ &= \sum_{k \in [K]} \sum_{(i,j) \in \binom{[N]}{2}} \sum_{\xi_i \in \Xi_i} \sum_{\xi_j \in \Xi_j} 2A_{i,j,k} \xi_i \xi_j \lambda_{i,j,k}(\xi_i, \xi_j) \\ &\quad + \sum_{k \in [K]} \sum_{i \in [N]} \sum_{\xi_i \in \Xi_i} (A_{i,i,k} \xi_i^2 + b_{i,k} \xi_i) \lambda_{i,k}(\xi_i) + \sum_{k \in [K]} c_k \lambda_k. \end{aligned}$$

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

Step (2): $\rho^* \geq \rho_u^*$

We construct a distribution $\mathbb{P}^*$ in the marginal-covariance submodular ambiguity set that attains the upper bound ρ_u^* using the optimal solution of the linear program. Consider an optimal solution of the linear program (4.5) denoted by $\boldsymbol{\lambda}^*$. Let $\mathcal{K}_+ \subseteq [K]$ such that $\lambda_k^* > 0$ for all $k \in \mathcal{K}_+$ and $\lambda_k^* = 0$ for all $k \in \mathcal{K}_0 := [K] \setminus \mathcal{K}_+$. Create a mixture distribution $\mathbb{P}^*$ as follows:

- (i) Generate a discrete random variable $\tilde{z}$ that takes values in $\mathcal{K}_+$ 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:

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

Generate in step (ii), a comonotonic random vector using these conditional marginal distributions. The constraints guarantee that $\sum_{\xi \in \Xi_i} \lambda_{i,k}^*(\xi_i) > 0$ for all $i \in [N]$, $k \in \mathcal{K}_+$ and $\lambda_{i,k}^*(\xi_i) = 0$ for all $\xi \in \Xi_i$, $i \in [N]$, $k \in \mathcal{K}_0$. It is a valid probability measure since $\sum_{k \in \mathcal{K}_+} \lambda_k^* = 1$. The marginal distribution of $\tilde{\xi}_i$ in the mixture distribution $\mathbb{P}^*$ is given by:

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

Hence the univariate marginal information of $\mathbb{P}^*$ matches the univariate marginal information specified in the ambiguity set. Let $\mathbb{P}_{i,j}$ denote a bivariate distribution of the random variables $(\tilde{\xi}_i, \tilde{\xi}_j)$ where conditional on $\tilde{z} = k$ we define it as:

$$\mathbb{P}_{i,j}(\tilde{\xi}_i = \xi_i, \tilde{\xi}_j = \xi_j | \tilde{z} = k) = \frac{\lambda_{i,j,k}^*(\xi_i, \xi_j)}{\sum_{\xi \in \Xi_i} \sum_{\eta \in \Xi_j} \lambda_{i,j,k}^*(\xi, \eta)}, \forall \xi_i \in \Xi_i, \forall \xi_j \in \Xi_j, \forall k \in \mathcal{K}_+.$$

The constraints guarantee that $\sum_{\xi \in \Xi_i} \sum_{\eta \in \Xi_j} \lambda_{i,j,k}^*(\xi, \eta) > 0$ for all $(i, j) \in \binom{[N]}{2}$, $k \in \mathcal{K}_+$ and $\lambda_{i,j,k}^*(\xi, \eta) = 0$ for all $\xi \in \Xi_i$, $\eta \in \Xi_j$, $(i, j) \in \binom{[N]}{2}$, $k \in \mathcal{K}_0$. From the feasibility conditions in the linear program, we see that $\text{proj}_i(\mathbb{P}_{i,j}|k) = \text{proj}_i(\mathbb{P}^*|k)$ and $\text{proj}_j(\mathbb{P}_{i,j}|k) = \text{proj}_j(\mathbb{P}^*|k)$ where $|k$ denotes conditional on $\tilde{z} = k$. This implies the existence of conditional bivariate distributions of each $(\tilde{\xi}_i, \tilde{\xi}_j)$ consistent with the conditional marginal distributions of $\tilde{\xi}_i$ and $\tilde{\xi}_j$. However $\mathbb{P}_{i,j}|k$ does not have to be the conditional bivariate distribution for $(\tilde{\xi}_i, \tilde{\xi}_j)$ in $\mathbb{P}^*$, namely it is possible that $\mathbb{P}_{i,j}|k \neq \text{proj}_{i,j}(\mathbb{P}^*|k)$. The expected

value of $\mathbb{E}[\tilde{\xi}_i, \tilde{\xi}_j]$ in the mixture distribution $\mathbb{P}^*$ satisfies:

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}^*} [\tilde{\xi}_i \tilde{\xi}_j] &= \sum_{k \in \mathcal{K}_+} \lambda_k^* \mathbb{E}_{\mathbb{P}^*} [\tilde{\xi}_i \tilde{\xi}_j | \tilde{z} = k] \\
&\geq \sum_{k \in \mathcal{K}_+} \lambda_k^* \mathbb{E}_{\mathbb{P}_{i,j}} [\tilde{\xi}_i \tilde{\xi}_j | \tilde{z} = k] \\
&\quad [\text{from (3.5) since } \xi_i \xi_j \text{ is supermodular and } \mathbb{P}^*|k \text{ and } \mathbb{P}_{i,j}|k \text{ have the same marginals}] \\
&= \sum_{k \in \mathcal{K}_+} \lambda_k^* \sum_{\xi_i \in \Xi_i} \sum_{\xi_j \in \Xi_j} \left(\xi_i \xi_j \frac{\lambda_{i,j,k}^*(\xi_i, \xi_j)}{\sum_{\xi \in \Xi_i} \sum_{\eta \in \Xi_j} \lambda_{i,j,k}^*(\xi, \eta)} \right) \\
&= \sum_{k \in \mathcal{K}_+} \sum_{\xi_i \in \Xi_i} \sum_{\xi_j \in \Xi_j} \xi_i \xi_j \lambda_{i,j,k}^*(\xi_i, \xi_j), \\
&\quad [\text{since } \sum_{\xi \in \Xi_i} \sum_{\eta \in \Xi_j} \lambda_{i,j,k}^*(\xi, \eta) = \lambda_k^*] \\
&\geq \mu_i \mu_j + \Sigma_{i,j}.
\end{aligned}$$

Hence the distribution satisfies the bivariate moment conditions in the ambiguity set. The final step is to show the sharpness of the bound under this distribution as follows:

$$\begin{aligned}
\rho^* &\geq \mathbb{E}_{\mathbb{P}^*} \left[\max_{k \in [K]} \left(\tilde{\xi}' \mathbf{A}_k \tilde{\xi} + \mathbf{b}'_k \tilde{\xi} + c_k \right) \right] \\
&\quad [\text{since } \mathbb{P}^* \text{ is feasible}] \\
&\geq \sum_{k \in \mathcal{K}_+} \lambda_k^* \mathbb{E}_{\mathbb{P}^*} \left[\left(\tilde{\xi}' \mathbf{A}_k \tilde{\xi} + \mathbf{b}'_k \tilde{\xi} + c_k \right) | \tilde{z} = k \right] \\
&\quad [\text{evaluating the expected value at the } k\text{th piece in step (ii) instead of the optimal piece}] \\
&\geq \sum_{k \in \mathcal{K}_+} \lambda_k^* \sum_{(i,j) \in \binom{[N]}{2}} \sum_{\xi_i \in \Xi_i} \sum_{\xi_j \in \Xi_j} \left(\frac{2A_{i,j,k} \xi_i \xi_j \lambda_{i,j,k}^*(\xi_i, \xi_j)}{\sum_{\xi \in \Xi_i} \sum_{\eta \in \Xi_j} \lambda_{i,j,k}^*(\xi, \eta)} \right) \\
&\quad + \sum_{k \in \mathcal{K}_+} \sum_{i \in [N]} \sum_{\xi_i \in \Xi_i} \lambda_k^* \left(\frac{(A_{i,i,k} \xi_i^2 + b_{i,k} \xi_i) \lambda_{i,k}^*(\xi_i)}{\sum_{\xi \in \Xi_i} \lambda_{i,k}^*(\xi)} \right) + \sum_{k \in \mathcal{K}_+} c_k \lambda_k^* \\
&\quad [\text{applying (3.5) where } A_{i,j,k} \geq 0 \text{ implies the quadratic terms } \xi_i A_{i,j,k} \xi_j \text{ are supermodular and} \\
&\quad \text{since } \mathbb{P}_{|k}^* \text{ and } \mathbb{P}_{i,j|k} \text{ have the same univariate marginals for } \tilde{\xi}_i \text{ and } \tilde{\xi}_j \text{ with } \mathbb{P}_{|k}^* \text{ comonotonic}] \\
&= \sum_{k \in \mathcal{K}_+} \sum_{(i,j) \in \binom{[N]}{2}} \sum_{\xi_i \in \Xi_i} \sum_{\xi_j \in \Xi_j} 2A_{i,j,k} \xi_i \xi_j \lambda_{i,j,k}^*(\xi_i, \xi_j) \\
&\quad + \sum_{k \in \mathcal{K}_+} \sum_{i \in [N]} \sum_{\xi_i \in \Xi_i} (A_{i,i,k} \xi_i^2 + b_{i,k} \xi_i) \lambda_{i,k}^*(\xi_i) + \sum_{k \in \mathcal{K}_+} c_k \lambda_k^* \\
&\quad [\text{since } \sum_{\xi \in \Xi_i} \sum_{\eta \in \Xi_j} \lambda_{i,j,k}^*(\xi, \eta) = \lambda_k^* \text{ and } \sum_{\xi \in \Xi_i} \lambda_{i,k}^*(\xi) = \lambda_k^*] \\
&= \rho_u^*.
\end{aligned}$$

From steps (1) and (2), $\rho^* = \rho_u^*$. □

Proof of Theorem 5.1

Let $\Xi := \prod_{i \in [N]} [0, B_i]$. We can reformulate (4.2) as a semi-infinite linear program:

$$\begin{aligned} \rho^* = \sup & \int_{\xi \in \Xi} \max_{k \in [K]} g_k(\xi) d\mathbb{P}(\xi) \\ \text{s.t.} & \int_{\xi \in \Xi} f_l(\xi) d\mathbb{P}(\xi) \leq \gamma_l, \quad \forall l \in [L], \\ & \int_{\xi \in \Xi} d\mathbb{P}(\xi) = 1, \\ & d\mathbb{P}(\xi) \geq 0, \quad \forall \xi \in \Xi, \end{aligned}$$

where the decision variable is the probability measure $\mathbb{P}$ on the set Ξ . This is a semi-infinite linear program with a finite number of constraints and infinite number of nonnegative decision variables. The semi-infinite linear program is feasible by assumption (A3') and since the support of the random vector is bounded, ρ^* is finite. Given $\epsilon > 0$, let $\Xi_{i,\epsilon} = \{0, \epsilon, 2\epsilon, \dots, \lfloor \frac{B_i}{\epsilon} \rfloor \epsilon\}$ for each $i \in [N]$. Let $\Xi_\epsilon := \prod_{i \in [N]} \Xi_{i,\epsilon}$. Then Ξ_ϵ forms an ϵ -grid of Ξ . Define $\bar{\Xi}_\epsilon = \Xi_\epsilon \cup \{\xi_1, \dots, \xi_R\} \subseteq \Xi$ where $\{\xi_1, \dots, \xi_R\}$ is the support of the known probability measure in $\mathcal{P}_{\text{sub}}$. Consider a discretized problem:

$$\begin{aligned} \rho_\epsilon^* := \max & \sum_{\xi \in \bar{\Xi}_\epsilon} \max_{k \in [K]} g_k(\xi) p(\xi) \\ \text{s.t.} & \sum_{\xi \in \bar{\Xi}_\epsilon} f_l(\xi) p(\xi) \leq \gamma_l + M\epsilon, \quad \forall l \in [L], \\ & \sum_{\xi \in \bar{\Xi}_\epsilon} p(\xi) = 1, \\ & p(\xi) \geq 0, \quad \forall \xi \in \bar{\Xi}_\epsilon, \end{aligned}$$

where M is the Lipschitz constant such that $\max_{i \in [N]} |\xi_i - \chi_i| \leq z$ implies $|f_l(\xi) - f_l(\chi)| \leq Mz$ for all l and $|g_k(\xi) - g_k(\chi)| \leq Mz$ for all k . This is a linear program with a finite number of constraints and finite number of decision variables. The number of decision variables is however exponential in N . The linear program is feasible since $\sum_{r \in [R]} p_r \delta_{\xi_r}$ is a feasible solution of the linear program. Under the assumptions of submodularity of the functions f_l and supermodularity of the functions g_k with efficient evaluation oracles, computing ρ_ϵ^* is possible in time polynomial in N , K , L , $\max_i B_i$, R , $\frac{1}{\epsilon}$ and the evaluation time of the oracles (from the proof of Theorem 4.1 and the polynomial time solvability of submodular function minimization in (3.3)). Now from Carathéodory theorem, there exists an optimal probability measure of the semi-infinite linear program that is supported on at most $L + 1$ points in Ξ with optimal value ρ^* . Round each support point in this optimal probability measure to the closest point in $\bar{\Xi}_\epsilon$, namely within an ϵ distance, merging points if need be. We thus obtain a modified discrete distribution $\hat{\mathbb{P}}_\epsilon^*$ with support contained in $\bar{\Xi}_\epsilon$ of cardinality at most $L + 1$ that is feasible for the discretized problem. Let $\hat{\rho}^*$ denote the expected objective function value for this new distribution. Then from Lipschitz continuity, $\rho^* - M\epsilon \leq \hat{\rho}^* \leq \rho^* + M\epsilon$. Since $\hat{\rho}^* \leq \rho_\epsilon^*$, we get $\rho_\epsilon^* \geq \rho^* - M\epsilon$. The result follows by setting ϵ appropriately. $\square$

Proof of Theorem 5.2

Step (1): Necessity of constraints in (5.2)

Define $\mathbf{S} := \boldsymbol{\mu}\boldsymbol{\mu}' + \boldsymbol{\Sigma}$. Let $(1, \boldsymbol{\mu}, \mathbf{S}) \in \text{cl}(\mathcal{M})$. First, consider the case where there exists a distribution $\mathbb{P}$ with support contained in $\mathbb{R}_+^N$ with $\mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{\xi}}] = \boldsymbol{\mu}$, $\mathbb{E}_{\mathbb{P}}[\text{diag}(\tilde{\boldsymbol{\xi}}\tilde{\boldsymbol{\xi}}')] = \text{diag}(\mathbf{S})$ and $\mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{\xi}}\tilde{\boldsymbol{\xi}}'] \geq \mathbf{S}$. Let the bivariate cross moments be defined as variables as follows:

$$\lambda_{i,j} = \mathbb{E}_{\mathbb{P}}[\tilde{\xi}_i \tilde{\xi}_j], \quad \forall (i,j) \in \binom{[N]}{2}.$$

The first set of constraints in (5.2) follows from the lower bound on the cross moment:

$$\mathbb{E}_{\mathbb{P}}[\tilde{\xi}_i \tilde{\xi}_j] \geq S_{i,j}.$$

The second set of constraints is due to the positive semidefiniteness of the three by three moment matrices for distinct i, j where:

$$\begin{aligned} \mathbb{E}_{\mathbb{P}} \left[\begin{pmatrix} 1 \\ \tilde{\xi}_i \\ \tilde{\xi}_j \end{pmatrix} \begin{pmatrix} 1 \\ \tilde{\xi}_i \\ \tilde{\xi}_j \end{pmatrix}' \right] &= \mathbb{E}_{\mathbb{P}} \left[\begin{pmatrix} 1 & \tilde{\xi}_i & \tilde{\xi}_j \\ \tilde{\xi}_i & \tilde{\xi}_i^2 & \tilde{\xi}_i \tilde{\xi}_j \\ \tilde{\xi}_j & \tilde{\xi}_i \tilde{\xi}_j & \tilde{\xi}_j^2 \end{pmatrix} \right] \\ &= \begin{pmatrix} 1 & \mu_i & \mu_j \\ \mu_i & S_{i,i} & \lambda_{i,j} \\ \mu_j & \lambda_{i,j} & S_{j,j} \end{pmatrix} \\ &\succeq 0. \end{aligned}$$

The last set of constraints is due to the nonnegativity of the means since the random variables are nonnegative. Next, consider the case where $(1, \boldsymbol{\mu}, \mathbf{S}) \in \text{cl}(\mathcal{M})$ but is not attained by a probability measure. Using the definition of the closure, there exists a sequence $m_0^\ell (1, \mathbf{m}^\ell, \mathbf{M}^\ell) \in \mathcal{M}$ such that $m_0^\ell (1, \mathbf{m}^\ell, \mathbf{M}^\ell) \rightarrow (1, \boldsymbol{\mu}, \mathbf{S})$. Since $m_0^\ell \rightarrow 1$, we have $m_0^\ell > 0$ for all sufficiently large ℓ . Thus, after normalization and relabeling the sequence appropriately, we may assume without loss of generality that $m_0^\ell = 1$ for every ℓ with $\boldsymbol{\mu}^\ell \rightarrow \boldsymbol{\mu}$ and $\mathbf{S}^\ell \rightarrow \mathbf{S}$ where $\mathbb{P}_\ell$ is a distribution generating the point $(1, \boldsymbol{\mu}^\ell, \mathbf{S}^\ell) \in \mathcal{M}$. Define:

$$\lambda_{i,j}^\ell := \mathbb{E}_{\mathbb{P}_\ell}[\tilde{\xi}_i \tilde{\xi}_j], \quad \forall (i,j) \in \binom{[N]}{2}.$$

By the Cauchy-Schwarz inequality,

$$0 \leq \lambda_{i,j}^\ell \leq \sqrt{\mathbb{E}_{\mathbb{P}_\ell}[\tilde{\xi}_i^2] \mathbb{E}_{\mathbb{P}_\ell}[\tilde{\xi}_j^2]} = \sqrt{S_{i,i}^\ell S_{j,j}^\ell}.$$

The right-hand side is bounded because $S_{i,i}^\ell \rightarrow S_{i,i}$ and $S_{j,j}^\ell \rightarrow S_{j,j}$. Since there are finitely many pairs (i, j) , we can pass to a common subsequence along which:

$$\lambda_{i,j}^\ell \rightarrow \lambda_{i,j}, \quad \forall (i,j) \in \binom{[N]}{2}.$$

For every such ℓ , $\lambda_{i,j}^\ell \geq S_{i,j}^\ell$ which taking limits gives $\lambda_{i,j} \geq S_{i,j}$. Moreover,

$$\begin{pmatrix} 1 & \mu_i^\ell & \mu_j^\ell \\ \mu_i^\ell & S_{i,i}^\ell & \lambda_{i,j}^\ell \\ \mu_j^\ell & \lambda_{i,j}^\ell & S_{j,j}^\ell \end{pmatrix} \succeq 0.$$

Since the positive semidefinite cone is closed, taking limits yields

$$\begin{pmatrix} 1 & \mu_i & \mu_j \\ \mu_i & S_{i,i} & \lambda_{i,j} \\ \mu_j & \lambda_{i,j} & S_{j,j} \end{pmatrix} \succeq 0.$$

Since $\mu_i^\ell \geq 0$ for all ℓ , we have $\mu_i \geq 0$. This establishes the necessity of (5.2).

Step (2): Sufficiency of constraints in (5.2)

Given $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\lambda_{i,j}$ for $(i,j) \in \binom{[N]}{2}$ that satisfy the conditions in (5.2), we construct a distribution (or a sequence of distributions) to show sufficiency. To do so, we first note that by taking the Schur complement of the upper-left entry of the moment matrix gives

$$\begin{pmatrix} \Sigma_{i,i} & \lambda_{i,j} - \mu_i \mu_j \\ \lambda_{i,j} - \mu_i \mu_j & \Sigma_{j,j} \end{pmatrix} \succeq 0,$$

which in turn gives,

$$\Sigma_{i,i} \Sigma_{j,j} \geq (\lambda_{i,j} - \mu_i \mu_j)^2.$$

This implies

$$\mu_i \mu_j + \sqrt{\Sigma_{i,i} \Sigma_{j,j}} \geq \lambda_{i,j}.$$

We start with the simpler case where $\boldsymbol{\mu}$ is a vector with all strictly positive components. In this case, we construct a comonotonic joint distribution $\mathbb{P}_\epsilon^c$ as follows:

$$\begin{aligned} \mathbb{P}_\epsilon^c \left(\tilde{\xi}_i = \mu_i - \sqrt{\Sigma_{i,i}} \sqrt{\frac{\epsilon}{1-\epsilon}}, \forall i \in [N] \right) &= 1 - \epsilon, \\ \mathbb{P}_\epsilon^c \left(\tilde{\xi}_i = \mu_i + \sqrt{\Sigma_{i,i}} \sqrt{\frac{1-\epsilon}{\epsilon}}, \forall i \in [N] \right) &= \epsilon, \end{aligned}$$

where $\epsilon \in (0, 1]$. Consider $\epsilon \in (0, \min_{i \in [N]} \mu_i^2 / (\mu_i^2 + \Sigma_{i,i})]$ where the random vector is guaranteed to be nonnegative. The mean of $\tilde{\xi}_i$ is given by:

$$\begin{aligned} \mathbb{E}_{\mathbb{P}_\epsilon^c} [\tilde{\xi}_i] &= (1 - \epsilon) \left(\mu_i - \sqrt{\Sigma_{i,i}} \sqrt{\frac{\epsilon}{1-\epsilon}} \right) + \epsilon \left(\mu_i + \sqrt{\Sigma_{i,i}} \sqrt{\frac{1-\epsilon}{\epsilon}} \right) \\ &= \mu_i. \end{aligned}$$

The second moment of $\tilde{\xi}_i$ is given by:

$$\begin{aligned}\mathbb{E}_{\mathbb{P}_\epsilon^c}[\tilde{\xi}_i^2] &= (1-\epsilon) \left(\mu_i - \sqrt{\Sigma_{i,i}} \sqrt{\frac{\epsilon}{1-\epsilon}} \right)^2 + \epsilon \left(\mu_i + \sqrt{\Sigma_{i,i}} \sqrt{\frac{1-\epsilon}{\epsilon}} \right)^2 \\ &= \mu_i^2 + \Sigma_{i,i}.\end{aligned}$$

The cross moment of $\tilde{\xi}_i$ and $\tilde{\xi}_j$ is given by:

$$\begin{aligned}\mathbb{E}_{\mathbb{P}_\epsilon^c}[\tilde{\xi}_i \tilde{\xi}_j] &= (1-\epsilon) \left(\mu_i - \sqrt{\Sigma_{i,i}} \sqrt{\frac{\epsilon}{1-\epsilon}} \right) \left(\mu_j - \sqrt{\Sigma_{j,j}} \sqrt{\frac{\epsilon}{1-\epsilon}} \right) \\ &\quad + \epsilon \left(\mu_i + \sqrt{\Sigma_{i,i}} \sqrt{\frac{1-\epsilon}{\epsilon}} \right) \left(\mu_j + \sqrt{\Sigma_{j,j}} \sqrt{\frac{1-\epsilon}{\epsilon}} \right) \\ &= \mu_i \mu_j + \sqrt{\Sigma_{i,i} \Sigma_{j,j}} \\ &\geq \lambda_{i,j} \\ &\quad \text{[from Schur's complement lemma]} \\ &\geq \mu_i \mu_j + \Sigma_{i,j} \\ &\quad \text{[from the first set of constraints in (5.2)].}\end{aligned}$$

Thus we construct a distribution in the ambiguity set for all $\epsilon \in (0, \min_{i \in [N]} \mu_i^2 / (\mu_i^2 + \Sigma_{i,i})]$. We next consider the case where some components of $\boldsymbol{\mu}$ are zero. Define

$$\mathcal{N}_+ := \{i \in [N] : \mu_i > 0\}, \quad \mathcal{N}_0 := \{i \in [N] : \mu_i = 0\}.$$

For each sufficiently small $\epsilon \in (0, 1)$, we use the same two point comonotonic distribution as before for all random variables in $\mathcal{N}_+$. We construct the marginal distributions for the random variables in $\mathcal{N}_0$ as follows:

$$\begin{aligned}\mathbb{P}_\epsilon^c(\tilde{\xi}_i = 0) &= 1 - \epsilon, \\ \mathbb{P}_\epsilon^c\left(\tilde{\xi}_i = \sqrt{\frac{\Sigma_{i,i}}{\epsilon}}\right) &= \epsilon.\end{aligned}$$

If $i \in \mathcal{N}_0$ and $\Sigma_{i,i} = 0$, this simply sets $\tilde{\xi}_i = 0$. Then for all $i \in \mathcal{N}_0$, $\mathbb{E}_{\mathbb{P}_\epsilon^c}[\tilde{\xi}_i] = \sqrt{\epsilon \Sigma_{i,i}}$ with $\lim_{\epsilon \downarrow 0} \mathbb{E}_{\mathbb{P}_\epsilon^c}[\tilde{\xi}_i] = 0$. Furthermore, $\mathbb{E}_{\mathbb{P}_\epsilon^c}[\tilde{\xi}_i^2] = \Sigma_{i,i}$. Let for any distinct i, j :

$$\lambda_{i,j}^\epsilon := \mathbb{E}_{\mathbb{P}_\epsilon^c}[\tilde{\xi}_i \tilde{\xi}_j].$$

Then for $i, j \in \mathcal{N}_0$, $\lambda_{i,j}^\epsilon = \sqrt{\Sigma_{i,i} \Sigma_{j,j}}$ and for $i \in \mathcal{N}_0$, $j \in \mathcal{N}_+$, $\lambda_{i,j}^\epsilon = \mu_j \sqrt{\epsilon \Sigma_{i,i}} + \sqrt{(1-\epsilon) \Sigma_{i,i} \Sigma_{j,j}}$. Furthermore, for $i, j \in \mathcal{N}_+$, $\lambda_{i,j}^\epsilon = \mu_i \mu_j + \sqrt{\Sigma_{i,i} \Sigma_{j,j}}$ from earlier. Define $\boldsymbol{\mu}^\epsilon := \mathbb{E}_{\mathbb{P}_\epsilon^c}[\tilde{\boldsymbol{\xi}}]$ and the symmetric matrix $\mathbf{S}^\epsilon$ with entries $S_{i,i}^\epsilon := \mu_i^2 + \Sigma_{i,i}$ for all $i \in [N]$ and $S_{i,j}^\epsilon := \min(\lambda_{i,j}, \lambda_{i,j}^\epsilon)$ for all $i \neq j$. By construction, $\mathbb{E}_{\mathbb{P}_\epsilon^c}[\text{diag}(\tilde{\boldsymbol{\xi}} \tilde{\boldsymbol{\xi}}')] = \text{diag}(\mathbf{S}^\epsilon)$ and $\mathbb{E}_{\mathbb{P}_\epsilon^c}[\tilde{\boldsymbol{\xi}} \tilde{\boldsymbol{\xi}}'] \geq \mathbf{S}^\epsilon$. It follows that

$$(1, \boldsymbol{\mu}^\epsilon, \mathbf{S}^\epsilon) \in \mathcal{M},$$

where $\boldsymbol{\mu}^\epsilon \rightarrow \boldsymbol{\mu}$. Also, for any distinct i, j ,

$$\begin{aligned} \lim_{\epsilon \downarrow 0} \lambda_{i,j}^\epsilon &= \mu_i \mu_j + \sqrt{\Sigma_{i,i} \Sigma_{j,j}} \\ &\geq \lambda_{i,j} \\ &\quad [\text{from Schur's complement lemma}]. \end{aligned}$$

We hence have

$$\lim_{\epsilon \downarrow 0} S_{i,j}^\epsilon = \lambda_{i,j} \geq \mu_i \mu_j + \Sigma_{i,j}.$$

Hence $(1, \boldsymbol{\mu}, \boldsymbol{\mu} \boldsymbol{\mu}' + \boldsymbol{\Sigma}) \in \text{cl}(\mathcal{M})$.

Proof of Theorem 5.3

Let ρ_u^* be optimal value of the semidefinite program (5.4). As before, we prove $\rho^* = \rho_u^*$ in two steps by showing $\rho^* \leq \rho_u^*$ and $\rho_u^* \leq \rho^*$. A key difference here is that we use moments as decision variables rather than probabilities and we need to deal with the issue of the support of the random vector being potentially unbounded.

Step (1): $\rho^ \leq \rho_u^*$*

To show that ρ_u^* is a valid upper bound on ρ^* , we start by providing a probabilistic interpretation of the formulation. Define the decision variables as:

$$\begin{aligned} \lambda_k &= \mathbb{P}\left(k(\tilde{\boldsymbol{\xi}}) = k\right), & \forall k \in [K], \\ \lambda_{i,k} &= \mathbb{E}_{\mathbb{P}}\left[\tilde{\xi}_i | k(\tilde{\boldsymbol{\xi}}) = k\right] \mathbb{P}\left(k(\tilde{\boldsymbol{\xi}}) = k\right), & \forall i \in [N], \forall k \in [K], \\ \lambda_{i,i,k} &= \mathbb{E}_{\mathbb{P}}\left[\tilde{\xi}_i^2 | k(\tilde{\boldsymbol{\xi}}) = k\right] \mathbb{P}\left(k(\tilde{\boldsymbol{\xi}}) = k\right), & \forall i \in [N], \forall k \in [K], \\ \lambda_{i,j,k} &= \mathbb{E}_{\mathbb{P}}\left[\tilde{\xi}_i \tilde{\xi}_j | k(\tilde{\boldsymbol{\xi}}) = k\right] \mathbb{P}\left(k(\tilde{\boldsymbol{\xi}}) = k\right), & \forall (i, j) \in \binom{[N]}{2}, \forall k \in [K]. \end{aligned}$$

The six set of constraints in the formulation are derived from necessary conditions that the variables must satisfy :

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

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

2. Law of total expectation for the first marginal moment:

$$\sum_{k \in [K]} \mathbb{E}_{\mathbb{P}}\left[\tilde{\xi}_i | k(\tilde{\boldsymbol{\xi}}) = k\right] \mathbb{P}\left(k(\tilde{\boldsymbol{\xi}}) = k\right) = \mathbb{E}_{\mathbb{P}}\left[\tilde{\xi}_i\right] = \mu_i.$$

3. Law of total expectation for the second marginal moment:

$$\sum_{k \in [K]} \mathbb{E}_{\mathbb{P}}\left[\tilde{\xi}_i^2 | k(\tilde{\boldsymbol{\xi}}) = k\right] \mathbb{P}\left(k(\tilde{\boldsymbol{\xi}}) = k\right) = \mathbb{E}_{\mathbb{P}}\left[\tilde{\xi}_i^2\right] = \mu_i^2 + \Sigma_{i,i}.$$

4. Law of total expectation for the cross moment of $\tilde{\xi}_i$ and $\tilde{\xi}_j$ for $i \neq j$:

$$\sum_{k \in [K]} \mathbb{E}_{\mathbb{P}} \left[\tilde{\xi}_i \tilde{\xi}_j | k(\tilde{\xi}) = k \right] \mathbb{P} \left(k(\tilde{\xi}) = k \right) = \mathbb{E}_{\mathbb{P}} \left[\tilde{\xi}_i \tilde{\xi}_j \right] \geq \mu_i \mu_j + \Sigma_{i,j}.$$

5. Positive semidefiniteness of the three by three moment matrices comes from:

$$\begin{aligned} & \mathbb{E}_{\mathbb{P}} \left[\mathbb{1}_{k(\tilde{\xi})=k} \begin{pmatrix} 1 & \tilde{\xi}_i & \tilde{\xi}_j \\ \tilde{\xi}_i & \tilde{\xi}_i^2 & \tilde{\xi}_i \tilde{\xi}_j \\ \tilde{\xi}_j & \tilde{\xi}_i \tilde{\xi}_j & \tilde{\xi}_j^2 \end{pmatrix} \right] \\ &= \mathbb{P} \left(k(\tilde{\xi}) = k \right) \begin{pmatrix} 1 & \mathbb{E}_{\mathbb{P}} \left[\tilde{\xi}_i | k(\tilde{\xi}) = k \right] & \mathbb{E}_{\mathbb{P}} \left[\tilde{\xi}_j | k(\tilde{\xi}) = k \right] \\ \mathbb{E}_{\mathbb{P}} \left[\tilde{\xi}_i | k(\tilde{\xi}) = k \right] & \mathbb{E}_{\mathbb{P}} \left[\tilde{\xi}_i^2 | k(\tilde{\xi}) = k \right] & \mathbb{E}_{\mathbb{P}} \left[\tilde{\xi}_i \tilde{\xi}_j | k(\tilde{\xi}) = k \right] \\ \mathbb{E}_{\mathbb{P}} \left[\tilde{\xi}_j | k(\tilde{\xi}) = k \right] & \mathbb{E}_{\mathbb{P}} \left[\tilde{\xi}_i \tilde{\xi}_j | k(\tilde{\xi}) = k \right] & \mathbb{E}_{\mathbb{P}} \left[\tilde{\xi}_j^2 | k(\tilde{\xi}) = k \right] \end{pmatrix} \\ &\succeq 0. \end{aligned}$$

6. Finally, since the support is in $\mathbb{R}_+^N$, we obtain the nonnegativity condition on the conditional marginal moment:

$$\mathbb{E}_{\mathbb{P}} \left[\tilde{\xi}_i | k(\tilde{\xi}) = k \right] \mathbb{P} \left(k(\tilde{\xi}) = k \right) \geq 0.$$

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

$$\begin{aligned} \mathbb{E}_{\mathbb{P}} \left[\max_{k \in [K]} \left(\tilde{\xi}' \mathbf{A}_k \tilde{\xi} + \mathbf{b}_k' \tilde{\xi} + c_k \right) \right] &= \sum_{k \in [K]} \mathbb{E}_{\mathbb{P}} \left[\left(\tilde{\xi}' \mathbf{A}_k \tilde{\xi} + \mathbf{b}_k' \tilde{\xi} + c_k \right) | k(\tilde{\xi}) = k \right] \mathbb{P} \left(k(\tilde{\xi}) = k \right) \\ &= \sum_{k \in [K]} \sum_{(i,j) \in \binom{[N]}{2}} 2A_{i,j,k} \lambda_{i,j,k} \\ &\quad + \sum_{k \in [K]} \sum_{i \in [N]} (A_{i,i,k} \lambda_{i,i,k} + b_{i,k} \lambda_{i,k}) + \sum_{k \in [K]} c_k \lambda_k. \end{aligned}$$

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

Step (2): $\rho^ \geq \rho_u^*$*

Consider an optimal solution of the semidefinite program (5.4) which is denoted by $\boldsymbol{\lambda}^*$. Partition $[K] = \mathcal{K}_+ \cup \mathcal{K}_0$ such that $\lambda_k^* > 0$ for all $k \in \mathcal{K}_+$ and $\lambda_k^* = 0$ for all $k \in \mathcal{K}_0$. It is possible $\mathcal{K}_0 = \emptyset$ but $\mathcal{K}_+ \neq \emptyset$ always. Also if $\lambda_k^* = 0$, then $\lambda_{i,k}^* = 0$ from the positive semidefiniteness condition of the matrices. We construct a distribution (case (2a)) or a sequence of distributions (case (2b)) that in the limit attains the upper bound ρ_u^* depending on the structure of the optimal solution $\boldsymbol{\lambda}^*$.

Case (2a): Values of $\boldsymbol{\lambda}^$ where $\lambda_{i,k}^* = 0 \Rightarrow \lambda_{i,i,k}^* = 0$ for any $k \in \mathcal{K}$.*

In this case $\lambda_{i,j,k}^* = 0$ for all $(i,j) \in \binom{[N]}{2}$, $k \in \mathcal{K}_0$ from the positive semidefiniteness condition of the matrices. Hence all the optimal decision variables indexed by $k \in \mathcal{K}_0$ are zero. Let $\epsilon \in (0, 1)$. Create a distribution $\mathbb{P}_{\epsilon}^*$ as follows:

(i) Generate a discrete random variable $\tilde{z}$ that takes each value $k \in \mathcal{K}_+$ with probability $\mathbb{P}_{\epsilon}^*(\tilde{z} = k) = \lambda_k^*$.

(ii) Conditional on the realization of $\tilde{z} = k$, define the distribution of the random vector as:

$$\mathbb{P}_\epsilon \left(\tilde{\xi}_i = \frac{\lambda_{i,k}^*}{\lambda_k^*} - \sqrt{\frac{\lambda_{i,i,k}^*}{\lambda_k^*} - \left(\frac{\lambda_{i,k}^*}{\lambda_k^*}\right)^2} \sqrt{\frac{\epsilon}{1-\epsilon}}, \forall i \in [N] \mid \tilde{z} = k \right) = 1 - \epsilon,$$

$$\mathbb{P}_\epsilon \left(\tilde{\xi}_i = \frac{\lambda_{i,k}^*}{\lambda_k^*} + \sqrt{\frac{\lambda_{i,i,k}^*}{\lambda_k^*} - \left(\frac{\lambda_{i,k}^*}{\lambda_k^*}\right)^2} \sqrt{\frac{1-\epsilon}{\epsilon}}, \forall i \in [N] \mid \tilde{z} = k \right) = \epsilon.$$

Clearly, this is a valid probability measure since $\sum_{k \in \mathcal{K}_+} \lambda_k^* = 1$. Observe that in step (ii), we generate a two point comonotonic random vector with perfect positive dependence. The conditional moments for $k \in \mathcal{K}_+$ are given by:

$$\begin{aligned} \mathbb{E}_{\mathbb{P}_\epsilon^*} [\tilde{\xi}_i \mid \tilde{z} = k] &= \frac{\lambda_{i,k}^*}{\lambda_k^*}, & \forall i \in [N], \\ \mathbb{E}_{\mathbb{P}_\epsilon^*} [\tilde{\xi}_i^2 \mid \tilde{z} = k] &= \frac{\lambda_{i,i,k}^*}{\lambda_k^*}, & \forall i \in [N], \\ \mathbb{E}_{\mathbb{P}_\epsilon^*} [\tilde{\xi}_i \tilde{\xi}_j \mid \tilde{z} = k] &= \left(\frac{\lambda_{i,k}^*}{\lambda_k^*}\right) \left(\frac{\lambda_{j,k}^*}{\lambda_k^*}\right) + \sqrt{\frac{\lambda_{i,i,k}^*}{\lambda_k^*} - \left(\frac{\lambda_{i,k}^*}{\lambda_k^*}\right)^2} \sqrt{\frac{\lambda_{j,j,k}^*}{\lambda_k^*} - \left(\frac{\lambda_{j,k}^*}{\lambda_k^*}\right)^2} & \forall (i, j) \in \binom{[N]}{2}. \end{aligned}$$

The mean of $\tilde{\xi}_i$ in the distribution $\mathbb{P}_\epsilon^*$ is given by:

$$\mathbb{E}_{\mathbb{P}_\epsilon^*} [\tilde{\xi}_i] = \sum_{k \in \mathcal{K}_+} \lambda_k^* \mathbb{E}_{\mathbb{P}_\epsilon^*} [\tilde{\xi}_i \mid \tilde{z} = k] = \sum_{k \in \mathcal{K}_+} \lambda_{i,k}^* = \mu_i.$$

The second moment of $\tilde{\xi}_i$ in the distribution $\mathbb{P}_\epsilon^*$ is given by:

$$\mathbb{E}_{\mathbb{P}_\epsilon^*} [\tilde{\xi}_i^2] = \sum_{k \in \mathcal{K}_+} \lambda_k^* \mathbb{E}_{\mathbb{P}_\epsilon^*} [\tilde{\xi}_i^2 \mid \tilde{z} = k] = \sum_{k \in \mathcal{K}_+} \lambda_{i,i,k}^* = \mu_i^2 + \Sigma_{i,i}.$$

The cross moment of $\tilde{\xi}_i$ and $\tilde{\xi}_j$ in $\mathbb{P}^*$ for any $\epsilon \in (0, 1)$ is given by:

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}_\epsilon^*} [\tilde{\xi}_i \tilde{\xi}_j] &= \sum_{k \in \mathcal{K}_+} \lambda_k^* \mathbb{E}_{\mathbb{P}_\epsilon^*} [\tilde{\xi}_i \tilde{\xi}_j | \tilde{z} = k] \\
&= \sum_{k \in \mathcal{K}_+} \frac{1}{\lambda_k^*} \left(\lambda_{i,k}^* \lambda_{j,k}^* + \sqrt{\lambda_k^* \lambda_{i,i,k}^* - \lambda_{i,k}^{*2}} \sqrt{\lambda_k^* \lambda_{j,j,k}^* - \lambda_{j,k}^{*2}} \right) \\
&\geq \sum_{k \in \mathcal{K}_+} \frac{1}{\lambda_k^*} \left(\lambda_{i,k}^* \lambda_{j,k}^* + \sqrt{\left(\lambda_k^* \lambda_{i,j,k}^* - \lambda_{i,k}^* \lambda_{j,k}^* \right)^2} \right) \\
&\quad [\text{from Schur's complement lemma, the moment matrix is positive semidefinite} \\
&\quad \text{implies } (\lambda_k^* \lambda_{i,i,k}^* - \lambda_{i,k}^{*2})(\lambda_k^* \lambda_{j,j,k}^* - \lambda_{j,k}^{*2}) \geq (\lambda_k^* \lambda_{i,j,k}^* - \lambda_{i,k}^* \lambda_{j,k}^*)^2] \\
&\geq \mu_i \mu_j + \Sigma_{i,j} \\
&= \sum_{k \in \mathcal{K}_+} \frac{1}{\lambda_k^*} (\lambda_{i,k}^* \lambda_{j,k}^* + |\lambda_k^* \lambda_{i,j,k}^* - \lambda_{i,k}^* \lambda_{j,k}^*|) \\
&\geq \sum_{k \in \mathcal{K}_+} \frac{1}{\lambda_k^*} (\lambda_{i,k}^* \lambda_{j,k}^* + \lambda_k^* \lambda_{i,j,k}^* - \lambda_{i,k}^* \lambda_{j,k}^*) \\
&= \sum_{k \in \mathcal{K}_+} \lambda_{i,j,k}^* \\
&\geq \mu_i \mu_j + \Sigma_{i,j}.
\end{aligned}$$

The realizations of the random vector are nonnegative for any $\epsilon \in (0, \min_{k \in \mathcal{K}_+} \lambda_{i,k}^{*2} / (\lambda_k^* \lambda_{i,i,k}^*))$ where $0/0 := 1$. The final step is to show the bound is attained by the distribution for any ϵ in this range. This follows from:

$$\begin{aligned}
\rho^* &\geq \mathbb{E}_{\mathbb{P}_\epsilon^*} \left[\max_{k \in [K]} \left(\tilde{\xi}' \mathbf{A}_k \tilde{\xi} + \mathbf{b}'_k \tilde{\xi} + c_k \right) \right] \\
&\quad [\text{since } \mathbb{P}_\epsilon^* \in \mathcal{P}_{\text{mean-cov-sub}} \text{ for } \epsilon \in (0, \min_{k \in \mathcal{K}_+} \lambda_{i,k}^{*2} / (\lambda_k^* \lambda_{i,i,k}^*))] \\
&\geq \sum_{k \in \mathcal{K}_+} \lambda_k^* \mathbb{E}_{\mathbb{P}_\epsilon^*} \left[\left(\tilde{\xi}' \mathbf{A}_k \tilde{\xi} + \mathbf{b}'_k \tilde{\xi} + c_k \right) | \tilde{z} = k \right], \\
&\quad [\text{evaluating the expected value at the } k\text{th piece in step (ii)}] \\
&\geq \sum_{k \in \mathcal{K}_+} \sum_{(i,j) \in \binom{[N]}{2}} 2A_{i,j,k} \lambda_{i,j,k}^* + \sum_{k \in \mathcal{K}_+} \sum_{i \in [N]} (A_{i,i,k} \lambda_{i,i,k}^* + b_{i,k} \lambda_{i,k}^*) + \sum_{k \in \mathcal{K}_+} c_k \lambda_k^* \\
&\quad [\text{applying (3.5) where } A_{i,j,k} \geq 0 \text{ implies the quadratic terms } \xi_i A_{i,j,k} \xi_j \text{ are supermodular and} \\
&\quad \text{since } \mathbb{P}_{|k}^* \text{ and } \mathbb{P}_{i,j|k} \text{ have the same univariate marginals for } \tilde{\xi}_i \text{ and } \tilde{\xi}_j] \\
&= \rho_u^*.
\end{aligned}$$

*Case (2b): All other values of λ^**

Let $\epsilon \in (0, 1)$. Here, we create a set of distributions that satisfy the moment constraints and attains the bound in the limit. Create a distribution $\mathbb{P}_\epsilon^*$ as follows:

- (i) Generate a discrete random variable $\tilde{z}$ that takes each value $k \in \mathcal{K}_+$ with probability $\mathbb{P}_\epsilon^*(\tilde{z} = k) = \lambda_k^* (1 - \epsilon)$ and each value $k \in \mathcal{K}_0$ with probability $\mathbb{P}_\epsilon^*(\tilde{z} = k) = \epsilon / |\mathcal{K}_0|$.
- (ii) Conditional on the realization of $\tilde{z} = k$, define the distribution of the random vector as follows:

If $k \in \mathcal{K}_+$, then:

$$\begin{aligned}\mathbb{P}_\epsilon \left(\tilde{\xi}_i = \frac{\lambda_{i,k}^*}{\lambda_k^*}, \forall i \in [N] \mid \tilde{z} = k \right) &= 1 - \epsilon, \\ \mathbb{P}_\epsilon \left(\tilde{\xi}_i = \frac{\lambda_{i,k}^*}{\lambda_k^*} + \frac{1}{\sqrt{\epsilon}} \sqrt{\frac{\lambda_{i,i,k}^*}{\lambda_k^*} - \left(\frac{\lambda_{i,k}^*}{\lambda_k^*} \right)^2}, \forall i \in [N] \mid \tilde{z} = k \right) &= \epsilon.\end{aligned}$$

If $k \in \mathcal{K}_0$, then:

$$\mathbb{P}_\epsilon \left(\tilde{\xi}_i = \sqrt{\frac{\lambda_{i,i,k}^* |\mathcal{K}_0|}{\epsilon}}, \forall i \in [N] \mid \tilde{z} = k \right) = 1.$$

Clearly, it is a valid probability measure since $\sum_{k \in \mathcal{K}_+ \cup \mathcal{K}_0} \lambda_k^* = 1$. Observe that in step (ii), we generate a two point comonotonic random vector when $\tilde{z} = k \in \mathcal{K}_+$ and a one point random vector when $\tilde{z} = k \in \mathcal{K}_0$. It is straightforward to verify that the conditional moments in the distribution are given for each $k \in \mathcal{K}_+$ by:

$$\begin{aligned}\mathbb{E}_{\mathbb{P}_\epsilon^*} [\tilde{\xi}_i \mid \tilde{z} = k] &= \frac{\lambda_{i,k}^*}{\lambda_k^*} + \sqrt{\epsilon} \sqrt{\frac{\lambda_{i,i,k}^*}{\lambda_k^*} - \left(\frac{\lambda_{i,k}^*}{\lambda_k^*} \right)^2}, & \forall i \in [N], \\ \mathbb{E}_{\mathbb{P}_\epsilon^*} [\tilde{\xi}_i^2 \mid \tilde{z} = k] &= \frac{\lambda_{i,i,k}^*}{\lambda_k^*} + 2\sqrt{\epsilon} \left(\frac{\lambda_{i,k}^*}{\lambda_k^*} \right) \sqrt{\frac{\lambda_{i,i,k}^*}{\lambda_k^*} - \left(\frac{\lambda_{i,k}^*}{\lambda_k^*} \right)^2}, & \forall i \in [N], \\ \mathbb{E}_{\mathbb{P}_\epsilon^*} [\tilde{\xi}_i \tilde{\xi}_j \mid \tilde{z} = k] &= \left(\frac{\lambda_{i,k}^*}{\lambda_k^*} \right) \left(\frac{\lambda_{j,k}^*}{\lambda_k^*} \right) + \sqrt{\frac{\lambda_{i,i,k}^*}{\lambda_k^*} - \left(\frac{\lambda_{i,k}^*}{\lambda_k^*} \right)^2} \sqrt{\frac{\lambda_{j,j,k}^*}{\lambda_k^*} - \left(\frac{\lambda_{j,k}^*}{\lambda_k^*} \right)^2} \\ &\quad + \sqrt{\epsilon} \left(\left(\frac{\lambda_{i,k}^*}{\lambda_k^*} \right) \sqrt{\frac{\lambda_{j,j,k}^*}{\lambda_k^*} - \left(\frac{\lambda_{j,k}^*}{\lambda_k^*} \right)^2} + \left(\frac{\lambda_{j,k}^*}{\lambda_k^*} \right) \sqrt{\frac{\lambda_{i,i,k}^*}{\lambda_k^*} - \left(\frac{\lambda_{i,k}^*}{\lambda_k^*} \right)^2} \right), & \forall (i, j) \in \binom{[N]}{2},\end{aligned}$$

and for each $k \in \mathcal{K}_0$ by:

$$\begin{aligned}\mathbb{E}_{\mathbb{P}_\epsilon^*} [\tilde{\xi}_i \mid \tilde{z} = k] &= \sqrt{\frac{\lambda_{i,i,k}^* |\mathcal{K}_0|}{\epsilon}}, & \forall i \in [N], \\ \mathbb{E}_{\mathbb{P}_\epsilon^*} [\tilde{\xi}_i^2 \mid \tilde{z} = k] &= \frac{\lambda_{i,i,k}^* |\mathcal{K}_0|}{\epsilon}, & \forall i \in [N], \\ \mathbb{E}_{\mathbb{P}_\epsilon^*} [\tilde{\xi}_i \tilde{\xi}_j \mid \tilde{z} = k] &= \sqrt{\lambda_{i,i,k}^* \lambda_{j,j,k}^* \frac{|\mathcal{K}_0|}{\epsilon}} & \forall (i, j) \in \binom{[N]}{2}.\end{aligned}$$

As $\epsilon \downarrow 0$, the mean of $\tilde{\xi}_i$ converges to:

$$\begin{aligned}\lim_{\epsilon \downarrow 0} \mathbb{E}_{\mathbb{P}_\epsilon^*} [\tilde{\xi}_i] &= \lim_{\epsilon \downarrow 0} \left(\sum_{k \in \mathcal{K}_+} \lambda_k^* (1 - \epsilon) \mathbb{E}_{\mathbb{P}_\epsilon^*} [\tilde{\xi}_i \mid \tilde{z} = k] + \sum_{k \in \mathcal{K}_0} \frac{\epsilon}{|\mathcal{K}_0|} \mathbb{E}_{\mathbb{P}_\epsilon^*} [\tilde{\xi}_i \mid \tilde{z} = k] \right) \\ &= \sum_{k \in \mathcal{K}_+} \lambda_{i,k}^* \\ &= \mu_i.\end{aligned}$$

As $\epsilon \downarrow 0$, the second moment of $\tilde{\xi}_i$ converges to:

$$\begin{aligned}
\lim_{\epsilon \downarrow 0} \mathbb{E}_{\mathbb{P}_\epsilon^*} [\tilde{\xi}_i^2] &= \lim_{\epsilon \downarrow 0} \left(\sum_{k \in \mathcal{K}_+} \lambda_k^* (1 - \epsilon) \mathbb{E}_{\mathbb{P}_\epsilon^*} [\tilde{\xi}_i^2 | \tilde{z} = k] + \sum_{k \in \mathcal{K}_0} \frac{\epsilon}{|\mathcal{K}_0|} \mathbb{E}_{\mathbb{P}_\epsilon^*} [\tilde{\xi}_i^2 | \tilde{z} = k] \right) \\
&= \sum_{k \in \mathcal{K}_+} \lambda_{i,i,k}^* + \sum_{k \in \mathcal{K}_0} \lambda_{i,i,k}^* \\
&= \mu_i^2 + \Sigma_{i,i}.
\end{aligned}$$

As $\epsilon \downarrow 0$, the cross moment of $\tilde{\xi}_i$ and $\tilde{\xi}_j$ satisfies the condition:

$$\begin{aligned}
\lim_{\epsilon \downarrow 0} \mathbb{E}_{\mathbb{P}_\epsilon^*} [\tilde{\xi}_i \tilde{\xi}_j] &= \lim_{\epsilon \downarrow 0} \left(\sum_{k \in \mathcal{K}_+} \lambda_k^* (1 - \epsilon) \mathbb{E}_{\mathbb{P}_\epsilon^*} [\tilde{\xi}_i \tilde{\xi}_j | \tilde{z} = k] + \sum_{k \in \mathcal{K}_0} \frac{\epsilon}{|\mathcal{K}_0|} \mathbb{E}_{\mathbb{P}_\epsilon^*} [\tilde{\xi}_i \tilde{\xi}_j | \tilde{z} = k] \right) \\
&= \sum_{k \in \mathcal{K}_+} \frac{1}{\lambda_k^*} \left(\lambda_{i,k}^* \lambda_{j,k}^* + \sqrt{\lambda_k^* \lambda_{i,i,k}^* - \lambda_{j,k}^{*2}} \sqrt{\lambda_k^* \lambda_{j,j,k}^* - \lambda_{j,k}^{*2}} \right) + \sum_{k \in \mathcal{K}_0} \sqrt{\lambda_{i,i,k}^* \lambda_{j,j,k}^*} \\
&\geq \sum_{k \in \mathcal{K}_+} \frac{1}{\lambda_k^*} \left(\lambda_{i,k}^* \lambda_{j,k}^* + \sqrt{\left(\lambda_k^* \lambda_{i,j,k}^* - \lambda_{i,k}^* \lambda_{j,k}^* \right)^2} \right) + \sum_{k \in \mathcal{K}_0} \sqrt{\lambda_{i,j,k}^{*2}} \\
&\quad \text{[from the positive semidefiniteness of the 3 by 3 matrices]} \\
&\geq \sum_{k \in \mathcal{K}} \lambda_{i,j,k}^* \\
&\geq \mu_i \mu_j + \Sigma_{i,j}.
\end{aligned}$$

The support of the random vector is nonnegative by construction. The final step is to show the sharpness of the bound. As $\epsilon \downarrow 0$, we get:

$$\begin{aligned}
\rho^* &\geq \lim_{\epsilon \downarrow 0} \mathbb{E}_{\mathbb{P}_\epsilon^*} \left[\max_{k \in [K]} \left(\tilde{\xi}' \mathbf{A}_k \tilde{\xi} + \mathbf{b}'_k \tilde{\xi} + c_k \right) \right] \\
&\quad \text{[using the limit]} \\
&\geq \lim_{\epsilon \downarrow 0} \sum_{k \in \mathcal{K}_+} \lambda_k^* (1 - \epsilon) \mathbb{E}_{\mathbb{P}_\epsilon^*} \left[\left(\tilde{\xi}' \mathbf{A}_k \tilde{\xi} + \mathbf{b}'_k \tilde{\xi} + c_k \right) | \tilde{z} = k \right] + \lim_{\epsilon \downarrow 0} \sum_{k \in \mathcal{K}_0} \frac{\epsilon}{|\mathcal{K}_0|} \mathbb{E}_{\mathbb{P}_\epsilon^*} \left[\left(\tilde{\xi}' \mathbf{A}_k \tilde{\xi} + \mathbf{b}'_k \tilde{\xi} + c_k \right) | \tilde{z} = k \right], \\
&\quad \text{[evaluating the expected value at the } k\text{th piece in step (ii)]} \\
&\geq \sum_{k \in \mathcal{K}} \sum_{(i,j) \in \binom{[N]}{2}} 2A_{i,j,k} \lambda_{i,j,k}^* + \sum_{k \in \mathcal{K}} \sum_{i \in [N]} (A_{i,i,k} \lambda_{i,i,k}^* + b_{i,k} \lambda_{i,k}^*) + \sum_{k \in \mathcal{K}} c_k \lambda_k^* \\
&\quad \text{[since } A_{i,j,k} \geq 0\text{]} \\
&= \rho_u^*.
\end{aligned}$$

From steps (1) and (2), $\rho^* = \rho_u^*$. □

Proof of Theorem 7.1

- (i) First assume that F is atomless. The worst-case expected value in (7.2) over the marginal ambiguity set has been characterized in earlier works (see chapter 2 in [44] and the constructions in

theorem 2 in [37] and theorem 4 in [53]). Particularly:

$$\rho_{\text{marg}}^* = N \int_{1-1/N}^1 F^{-1}(u) du.$$

The corresponding extremal distribution which attains this worst-case expected value is given by:

(i) Generate a discrete random variable $\tilde{z}$ that takes each value $i \in [N]$ with probability $\mathbb{P}^*(\tilde{z} = i) = \frac{1}{N}$.

(ii) Conditional on the realization of $\tilde{z} = i$, define the distribution of the random variables as follows:

$$\begin{aligned} \tilde{\xi}_i &\sim F^{-1}(1 - 1/N + \tilde{U}_i/N), \\ \tilde{\xi}_j &\sim F^{-1}((1 - 1/N)\tilde{U}_j), \quad \forall j \neq i. \end{aligned}$$

where $\tilde{U}_j$ is a uniform random variable on $[0, 1]$ for each $j \in [N]$ generated independently in step (ii).

We build on this worst-case distribution to construct a feasible distribution $\mathbb{P}$ to the marginal-covariance submodular ambiguity set (7.4) where all $\Sigma_{i,j} = 0$:

(i) Generate a discrete random variable $\tilde{z}$ that takes values $i \in [N + 2]$ where the probabilities are given by $\mathbb{P}(\tilde{z} = i) = \frac{1}{N}(1 - \frac{1}{N})$ for all $i \in [N]$, $\mathbb{P}(\tilde{z} = N + 1) = \frac{1}{N^2}$ and $\mathbb{P}(\tilde{z} = N + 2) = \frac{1}{N} - \frac{1}{N^2}$.

(ii) Conditional on the realization of $\tilde{z} = i$, define the distribution of the random variables as follows:

If $i \in [N]$, then:

$$\begin{aligned} \tilde{\xi}_i &\sim F^{-1}(1 - 1/N + \tilde{U}_i/N), \\ \tilde{\xi}_j &\sim F^{-1}((1 - 1/N)\tilde{U}_j), \quad \forall j \neq i. \end{aligned}$$

If $i = N + 1$, then:

$$\tilde{\xi}_i \sim F^{-1}(1 - 1/N + \tilde{U}_i/N), \quad \forall i.$$

If $i = N + 2$, then:

$$\tilde{\xi}_i \sim F^{-1}((1 - 1/N)\tilde{U}_i), \quad \forall i,$$

where the uniform random variables in $[0, 1]$ in step (ii) are generated using independence.

It is straightforward to see that this is a valid distribution with each $\tilde{\xi}_i \sim F$ where the random variables are pairwise independent in $\mathbb{P}$. Hence the random variables are uncorrelated. The expected value of the maximum of the random variables in this distribution satisfies the lower bound:

$$\begin{aligned} \mathbb{E}_{\mathbb{P}} \left[\max_{i \in [N]} \tilde{\xi}_i \right] &\geq \left(1 - \frac{1}{N} + \frac{1}{N^2} \right) N \int_{1-1/N}^1 F^{-1}(u) du \\ &\quad [\text{since } \mathbb{E}[\max_j \tilde{\xi}_j | \tilde{z} = i] \geq \mathbb{E}[\tilde{\xi}_i | \tilde{z} = i] \text{ for each } i \in [N], \\ &\quad \mathbb{E}[\max_j \tilde{\xi}_j | \tilde{z} = N + 1] \geq \mathbb{E}[\tilde{\xi}_i | \tilde{z} = N + 1] \text{ for any } i \in [N] \text{ and} \\ &\quad \text{and } \mathbb{E}[\max_j \tilde{\xi}_j | \tilde{z} = N + 2] \geq 0 \text{ from nonnegativity of random variables}] \\ &= \left(1 - \frac{1}{N} + \frac{1}{N^2} \right) \rho_{\text{marg}}^*. \end{aligned}$$

Since this is a feasible distribution, we have:

$$\begin{aligned}\rho_{\text{marg-cov-sub}}^* &\geq \mathbb{E}_{\mathbb{P}} \left[\max_{i \in [N]} \tilde{\xi}_i \right] \\ &\geq \left(1 - \frac{1}{N} + \frac{1}{N^2} \right) \rho_{\text{marg}}^*,\end{aligned}$$

which gives the desired result. Now suppose F has an atom at $q = F^{-1}(1 - 1/N)$. Let $r = \mathbb{P}(\tilde{\xi} > q)$ and $a = \mathbb{P}(\tilde{\xi} = q)$. Then by definition of the quantile $r \leq 1/N \leq r + a$. Assign the amount $1/N - r$ of the atom at q to the upper-tail distribution. Define the distributions:

$$F_H(A) := \frac{F(A \cap (q, \infty)) + (1/N - r)\mathbb{1}_{\{q \in A\}}}{1/N} \text{ and } F_L(A) := \frac{F(A \cap [0, q)) + (r + a - 1/N)\mathbb{1}_{\{q \in A\}}}{1 - 1/N}.$$

Then

$$F = (1/N)F_H + (1 - 1/N)F_L \text{ with } \text{supp}(F_H) \subseteq [q, \infty), \text{supp}(F_L) \subseteq [0, q].$$

Now use the same construction as in the atomless case with F_H and F_L respectively for the upper and lower tails and the result follows. Furthermore, this lower bound is sharp. To see this, one can set $\mathbb{P}(\tilde{\xi}_i = 1) = 1 - \mathbb{P}(\tilde{\xi}_i = 0) = 1/N$. In this case, the univariate bound is given by $\rho_{\text{marg}}^* = 1$ (the union bound). Using the pairwise independent Bernoulli random variable construction in [49] which attains the Hunter-Worsley bound (see [13]), we get $\rho_{\text{marg-cov-sub}}^* = 1 - 1/N + 1/N^2$.

- (ii) In the case when $\Sigma_{i,j}$ is arbitrary, consider the marginal distribution $\mathbb{P}(\tilde{\xi}_i = 1) = 1 - \mathbb{P}(\tilde{\xi}_i = 0) = 1/N$. Set $\Sigma_{i,j} = 1/N - 1/N^2$ for all $(i, j) \in \binom{[N]}{2}$. In this case, the only feasible distribution in the marginal-covariance submodular ambiguity set is given by the two point distribution where all random variables jointly take a value of 0 with probability $1 - 1/N$ and take a value of 1 with probability $1/N$ (comonotonic). Then we have $\rho_{\text{marg}}^* = 1$ and $\rho_{\text{marg-cov-sub}}^* = 1/N$ which gives:

$$\frac{\rho_{\text{marg-cov-sub}}^*}{\rho_{\text{marg}}^*} = \frac{1}{N}.$$

The result follows.

- (iii) When the lower bound on the covariances is set to zero, the results for the moment ambiguity sets follows from using the extremal marginal distribution F^* that is identified for the marginal ambiguity set and using the construction in (i). For the case of general lower bounds on covariances, assume the mean $\mu_i = \mu > 0$, variance $\Sigma_{i,i} = \Sigma > 0$ where $N \geq 1 + \Sigma/\mu^2$. In this case, the marginal moment bound on the expected maximum is given by $\mu + \sqrt{\Sigma}\sqrt{N-1}$ (see section 2.2.1 in [10] for the construction of the extremal distribution). If we set the lower bound on the covariances to Σ (or the lower bound on correlations to 1), the expected maximum has to be μ (perfect linear dependence). In this case again the ratio of the bound will go to zero as $N \uparrow \infty$. $\square$

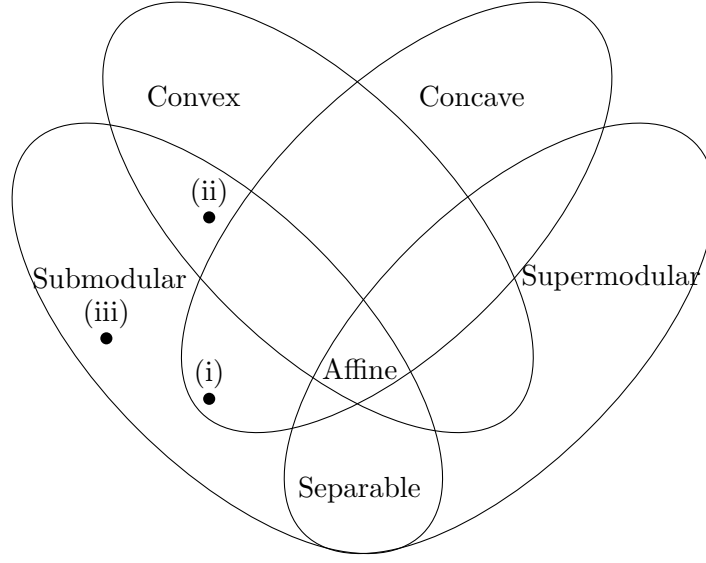


Figure 4: Venn diagram illustrating the set of convex, concave, submodular and supermodular functions defined over $\mathbb{R}_+^N$. Affine functions lie at the intersection of all four sets.

Appendix B

B.1 Review of submodularity

We list out three examples of submodular functions that satisfy the definition in (3.1):

- (i) $f(\boldsymbol{\xi}) = h(\mathbf{a}'\boldsymbol{\xi})$ where $\mathbf{a} \geq \mathbf{0}$ and $h : \mathbb{R} \rightarrow \mathbb{R}$ is a concave function,
- (ii) $f(\boldsymbol{\xi}) = \max(\xi_1, \dots, \xi_N)$,
- (iii) $f(\boldsymbol{\xi}) = -\prod_{i \in [N]} \xi_i$ where $\boldsymbol{\xi} \geq \mathbf{0}$.

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 $f(\boldsymbol{\xi}) = \mathbf{a}'\boldsymbol{\xi}$). Other examples of submodular functions include the weighted cut function in a directed graph, the entropy of a random vector, the influence function in a social network and the rank function of a matroid (see [41, 54, 3] 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 taking a 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 4 for an illustration). A bivariate function of the form $f(\xi_i, \xi_j) = h(\xi_i - \xi_j)$ where g is an univariate convex function is also submodular.

Submodular functions defined over discrete domains behave somewhat similarly to convex functions

defined over continuous domains, in terms of minimization of such functions (see [41]). A key result proved in [29], building on the work of [24, 35], showed that the submodular function minimization problem (3.3) is solvable in time polynomial in the input size using the ellipsoid method when the sets Ξ_i are discrete and finite. An object that aids in the efficient minimization of submodular functions is the Lovász extension [41] (also known as the Choquet integral [18]) which is defined through the expected value of a submodular function of a comonotonic random vector. The support of the comonotonic random vector $\tilde{\xi}^c$ is contained in completely ordered subsets of $\mathbb{R}^N$ with:

$$\mathbb{P}^c(\tilde{\xi}^c > \mathbf{t}) = \min_{i \in [N]} \mathbb{P}_i(\tilde{\xi}_i > t_i) \text{ and } \mathbb{P}^c(\tilde{\xi}^c \leq \mathbf{t}) = \min_{i \in [N]} \mathbb{P}_i(\tilde{\xi}_i \leq t_i), \forall \mathbf{t} \in \mathbb{R}^N. \quad (8.1)$$

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

$$\mathbb{E}_{\mathbb{P}^c} [f(\tilde{\xi}^c)] = \int_0^1 f(F_1^{-1}(t), \dots, F_N^{-1}(t)) dt. \quad (8.2)$$

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 (8.2) is computable using a polynomial number of calls to the evaluation oracle. The extremal characterization in (3.4) is useful in developing efficient algorithms for submodular function minimization. Specifically, the function $f : \prod_{i \in [N]} \Xi_i \rightarrow \mathbb{R}$ is submodular if and only if the functional $\inf_{\mathbb{P} \in \mathcal{P}_F(\mathbb{P}_1, \dots, \mathbb{P}_N)} \mathbb{E}_{\mathbb{P}}[f(\tilde{\xi})] : \mathbb{P}_1, \dots, \mathbb{P}_N \rightarrow \mathbb{R}$ is convex (see [41, 3]). The domain of the functional is specified by the marginal probability measures $\mathbb{P}_1, \dots, \mathbb{P}_N$ where $\text{supp}(\mathbb{P}_1) \subseteq \Xi_1, \dots, \text{supp}(\mathbb{P}_N) \subseteq \Xi_N$. The corresponding functional value is exactly the Lovász extension or the Choquet integral. This lets one transform the submodular function minimization problem in (3.3) to a convex minimization problem over probability measures as follows:

$$\inf_{\xi \in \prod_{i \in [N]} \Xi_i} f(\xi) \stackrel{(a)}{=} \inf_{\text{supp}(\mathbb{P}_i) \subseteq \Xi_i, \forall i \in [N]} \inf_{\mathbb{P} \in \mathcal{P}_F(\mathbb{P}_1, \dots, \mathbb{P}_N)} \mathbb{E}_{\mathbb{P}} [f(\tilde{\xi})] \stackrel{(b)}{=} \inf_{\text{supp}(\mathbb{P}_i) \subseteq \Xi_i, \forall i \in [N]} \mathbb{E}_{\mathbb{P}^c} [f(\tilde{\xi}^c)]. \quad (8.3)$$

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 (3.4). 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 [41, 3]).

B.2 Examples of submodular ambiguity sets

Here, we discuss further examples of submodular ambiguity sets and information that can be modeled using such sets:

- (i) *Univariate*: Give a finite set of arbitrary univariate functions $f_{i,l} : \Xi_i \rightarrow \mathbb{R}$ and scalars $\underline{\gamma}_{i,l}, \bar{\gamma}_{i,l} \in \mathbb{R}$ that satisfy $\underline{\gamma}_{i,l} \leq \bar{\gamma}_{i,l}$ for each $l \in [L_i], i \in [N]$, consider the ambiguity set:

$$\{\mathbb{P} \in \mathcal{P}(\prod_{i \in [N]} \Xi_i) \mid \mathbb{E}_{\mathbb{P}}[f_{i,l}(\tilde{\xi}_i)] \in [\underline{\gamma}_{i,l}, \bar{\gamma}_{i,l}], \forall l \in [L_i], \forall i \in [N]\}.$$

Since all univariate functions are submodular, by rewriting the constraints as $\mathbb{E}_{\mathbb{P}}[f_{i,l}(\tilde{\xi}_i)] \leq \bar{\gamma}_{i,l}$ and $\mathbb{E}_{\mathbb{P}}[-f_{i,l}(\tilde{\xi}_i)] \leq -\underline{\gamma}_{i,l}$, we see that this is an example of a submodular ambiguity set. By setting $\underline{\gamma}_{i,l} = -\infty$ (likewise $\bar{\gamma}_{i,l} = +\infty$), we allow for specifying only an upper bound (likewise lower bound) on $\mathbb{E}_{\mathbb{P}}[f_{i,l}(\tilde{\xi}_i)]$ and by setting $\underline{\gamma}_{i,l} = \bar{\gamma}_{i,l}$, we allow for precisely fixing the value in the ambiguity set. This ambiguity set allows for exactly specifying or bounding terms such as the marginal probability $\mathbb{P}(\tilde{\xi}_i \leq t)$ for a given t , marginal moment $\mathbb{E}_{\mathbb{P}}[\tilde{\xi}_i^l]$ for a given l , marginal dispersion measure $\mathbb{E}_{\mathbb{P}}[|\tilde{\xi}_i - m_{i,l}|^l]$ for a given l and $m_{i,l}$ or the moment generating function $\mathbb{E}_{\mathbb{P}}[e^{t\tilde{\xi}_i}]$ for a given t . While the ambiguity set allows for many such terms, only finitely many are allowed for the results in this paper to go through. When the support of each random variable is discrete and finite, this ambiguity set can incorporate the full marginal distribution if required and is equivalent to the Fréchet set of joint discrete distributions with finite marginal support. Solving the inner problem in (1.1) in this case then reduces to solving an instance of a multimarginal optimal transport problem (see [45, 2]).

- (ii) *Bivariate:* Give a finite set of bivariate submodular functions $f_{i,j,l} : \Xi_i \times \Xi_j \rightarrow \mathbb{R}$ and scalars $\gamma_{i,j,l} \in \mathbb{R}$ for $l \in [L_{i,j}]$, $(i,j) \in \binom{[N]}{2}$, consider the ambiguity set:

$$\{\mathbb{P} \in \mathcal{P}(\prod_{i \in [N]} \Xi_i) \mid \mathbb{E}[f_{i,j,l}(\tilde{\xi}_i, \tilde{\xi}_j)] \leq \gamma_{i,j,l}, \forall l \in [L_{i,j}], \forall (i,j) \in \binom{[N]}{2}\}.$$

Unlike the submodular ambiguity set with univariate information only, the sign of the inequalities matters here since not all bivariate functions are submodular. This ambiguity set allows to upper bound pairwise Wasserstein distance such as $\mathbb{E}_{\mathbb{P}}[|\tilde{\xi}_i - \tilde{\xi}_j|^p]$ for $p \geq 1$ or lower bound terms such as the bivariate tail probabilities $\mathbb{P}(\tilde{\xi}_i > t_i, \tilde{\xi}_i > t_j)$ for a given t_i, t_j and cross moments of order two such as $\mathbb{E}_{\mathbb{P}}[\tilde{\xi}_i \tilde{\xi}_j]$.

- (iii) *Norm of a nonnegative random vector:* Given $\prod_{i \in [N]} \Xi_i \subseteq \mathbb{R}_+^N$ and upper bounds γ_l on the expected p_l -norms of the random vector, consider the ambiguity set:

$$\{\mathbb{P} \in \mathcal{P}(\prod_{i \in [N]} \Xi_i \subseteq \mathbb{R}_+^N) \mid \mathbb{E}[\|\tilde{\boldsymbol{\xi}}\|_{p_l}] \leq \gamma_l, \forall l \in [L]\}.$$

When $p_l \geq 1$, the norm function $\|\boldsymbol{\xi}\|_{p_l}$ is submodular over the nonnegative orthant; see [46]. Hence this is an example of a submodular ambiguity set. When the support of the random vector is a convex set, this is also a convex ambiguity set. It should be noted however that an upper bound constraint on the expected norm of a translated random vector of the form $\mathbb{E}[\|\tilde{\boldsymbol{\xi}} - \mathbf{m}_l\|_{p_l}] \leq \gamma_l$ can be modeled using a convex ambiguity set but not with a submodular ambiguity set since the function $\|\boldsymbol{\xi} - \mathbf{m}_l\|_{p_l}$ is not submodular.

- (iv) *Concave function of sum of random variables:* Given univariate concave functions $f_l : \mathbb{R} \rightarrow \mathbb{R}$ for $l \in [L]$ and scalars γ_l for $l \in [L]$, consider an ambiguity set given by:

$$\{\mathbb{P} \in \mathcal{P}(\prod_{i \in [N]} \Xi_i) \mid \mathbb{E}_{\mathbb{P}} \left[f_l \left(\sum_{i \in [N]} \tilde{\xi}_i \right) \right] \leq \gamma_l, \forall l \in [L]\}.$$

This is an example of a submodular ambiguity set. Such functions include $\min_{r \in [R]} (a_r \sum_{i \in [N]} \xi_i +$

b_r) and $\ln(\sum_{i \in [N]} \xi_i)$ defined on the appropriate domain.

- (v) *Orthant probabilities:* Given a set of vectors $\xi_l \in \prod_{i \in [N]} \Xi_i$ and scalars $\gamma_l \in [0, 1]$ for $l \in [L]$, consider the ambiguity set:

$$\{\mathbb{P} \in \mathcal{P}(\prod_{i \in [N]} \Xi_i) \mid \mathbb{P}(\tilde{\xi} \geq \xi_l) \geq \gamma_l, \forall l \in [L]\}.$$

This ambiguity set gives lower bounds on the orthant (tail) probabilities and is an example of a submodular ambiguity set defined using lower bound on the expected value of supermodular functions of the form $\mathbb{1}_{\xi \geq \xi_l}$.

B.3 Additional example

Example 5 (Inequality constraints are not always tight for submodular objective functions). We consider the Bernoulli ambiguity sets in (3.8) and (3.7) with $p_{i,j} = p_i p_j$ where each $p_i \in [0, 1]$. Both the ambiguity sets are nonempty since (3.7) is the set of pairwise independent Bernoulli random vectors with fixed marginals which is feasible and (3.8) is a relaxation of this set. Let $g(\xi) = \max_{(i,j) \in \binom{[N]}{2}} (\xi_i + \xi_j)$ which is a submodular function. Set $N = 5$ and $p_1 = p_2 = \alpha$ and $p_3 = p_4 = p_5 = \beta$ and compute the bounds by solving the linear programs by varying α and β . Let ρ_{eq}^* be the maximum expected value obtained using the ambiguity set (3.7) (equality constraints) and ρ_{rel}^* be the maximum expected value obtained using the submodular ambiguity set (3.8) (relaxed inequality constraints). Clearly $\rho_{\text{rel}}^* \geq \rho_{\text{eq}}^*$. In Figure 5, we plot the percentage difference for several values of α and β , clearly highlighting that the bounds can be strictly different with $\rho_{\text{rel}}^* > \rho_{\text{eq}}^*$. This also illustrates that the inequalities in the ambiguity set need not be tight at optimality.

B.4 Comparison of convex and submodular ambiguity sets

Here we provide more details on the comparison of the two ambiguity sets. The next example shows that neither ambiguity set nests the other.

Example 6. [Neither moment ambiguity set nests the other] Consider $N = 1$ with $\mu > 0$ and $\Sigma > 0$. Let $a < 0$ and $b \in \mathbb{R}$ be fixed with $g(\tilde{\xi}) = a\tilde{\xi}^2 + b\tilde{\xi}$. Then it is easy to see that:

$$\sup_{\mathbb{P} \in \mathcal{P}_{\text{mean-cov-sub}}} \mathbb{E}_{\mathbb{P}} [a\tilde{\xi}^2 + b\tilde{\xi}] = a(\mu^2 + \Sigma) + b\mu \text{ and } \sup_{\mathbb{P} \in \mathcal{P}_{\text{mean-cov-cvx}}} \mathbb{E}_{\mathbb{P}} [a\tilde{\xi}^2 + b\tilde{\xi}] = a\mu^2 + b\mu.$$

In this case, the ambiguity set $\mathcal{P}_{\text{mean-cov-sub}}$ gives a tighter bound than $\mathcal{P}_{\text{mean-cov-cvx}}$ for each $a < 0$. This is since the second moment cannot be exactly specified in the convex ambiguity set and the worst-case distribution is attained by a Dirac measure at μ . On the other hand, it is precisely specified in the submodular ambiguity set. On the other hand, consider $N = 2$ with $\mu_1 = \mu_2 = 1$, $\Sigma_{1,1} = \Sigma_{2,2} = 1$ and $\Sigma_{1,2} = 0$. Let $g(\tilde{\xi}_1, \tilde{\xi}_2) = \max(\tilde{\xi}_1 + \tilde{\xi}_2 - 1, 0)$. Solving the semidefinite programs with SDPT3 gives:

$$\sup_{\mathbb{P} \in \mathcal{P}_{\text{mean-cov-sub}}} \mathbb{E}_{\mathbb{P}} [\max(\tilde{\xi}_1 + \tilde{\xi}_2 - 1, 0)] \approx 0.8536 \text{ and } \sup_{\mathbb{P} \in \mathcal{P}_{\text{mean-cov-cvx}}} \mathbb{E}_{\mathbb{P}} [\max(\tilde{\xi}_1 + \tilde{\xi}_2 - 1, 0)] \approx 0.7887.$$

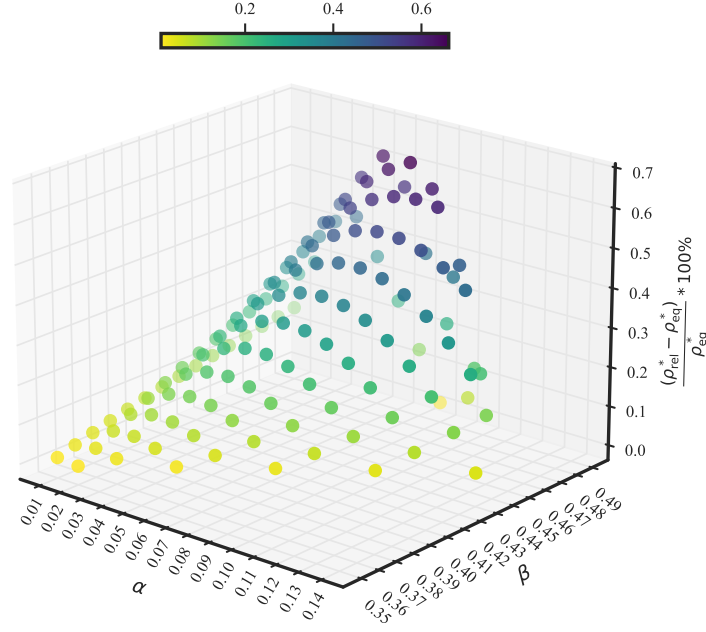


Figure 5: Percentage relative difference in objective values given by $\frac{(\rho_{\text{rel}}^* - \rho_{\text{eq}}^*)}{\rho_{\text{eq}}^*} * 100\%$ for varying values of α, β .

Here $\mathcal{P}_{\text{mean-cov-sub}}$ gives a weaker bound than $\mathcal{P}_{\text{mean-cov-cvx}}$ since the objective function is supermodular and the extremal distribution for the former will be a comonotonic random vector. On the other hand, the extremal distribution for the latter need not be comonotonic. In fact it is shown in [20] that for piecewise linear objective functions, the second moment constraint in the semidefinite program (5.5) for the convex ambiguity set can be made tight without losing optimality.

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] F. Bach. Submodular functions: from discrete to continuous domains. *Mathematical Programming*, 175, 419-459, 2019.
- [4] R. Belbasi, A. Selvi, W. Wiesemann. It’s all in the mix: Wasserstein machine learning with mixed features. To appear in *Manufacturing and Service Operations Management*, 2026.
- [5] A. Ben-Tal, L. El Ghaoui, A. Nemirovski. *Robust Optimization*. Princeton University Press, 2009.
- [6] 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.
- [7] D. Bertsimas, D. den Hertog. *Robust and Adaptive Optimization*. Dynamic Ideas, 2022.
- [8] 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.

- [9] D. Bertsimas, V. Farias, N. Trichakis. On the efficiency-fairness trade-off. *Management Science*, 58(12), 2234-2250, 2012
- [10] D. Bertsimas, K. Natarajan, C-P. Teo. Tight bounds on expected order statistics. *Probability in Engineering and Informational Sciences*, 20(4), 667-686, 2006.
- [11] D. Bertsimas, M. Sim, M. Zhang. Adaptive distributionally robust optimization. *Management Science*, 65(2), 604-618, 2019.
- [12] J. Blanchet, K. Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2), 565-600, 2019.
- [13] E. Boros, A. Scozzari. Polynomially computable bounds for the probability of the union of events. *Mathematical Programming*, 148, 77-113, 2014.
- [14] G. Calinescu, C. Chekuri, M. Pál, J. Vondrák. Maximizing a submodular set function subject to a matroid constraint. In: Fischetti, M., Williamson, D.P. (eds) *Integer Programming and Combinatorial Optimization. IPCO 2007. Lecture Notes in Computer Science*, vol 4513, Springer, Berlin.
- [15] C Chekuri, J Vondrák, R Zenklusen. Submodular function maximization via the multilinear relaxation and contention resolution schemes. *SIAM Journal on Computing* 43(6), 1831-1879, 2014.
- [16] 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.
- [17] Z. Chen, M. Sim, P. Xiong. Robust stochastic optimization made easy with RSOME. *Management Science*, 66(8), 3329-3339, 2020.
- [18] G. Choquet. Theory of capacities. *Annales de l'Institut Fourier*, 5, 131-295, 1953.
- [19] L. de Haan, A. Ferreira. *Extreme Value Theory: An Introduction*. Springer, New York, 2006.
- [20] E. Delage, Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3), 595-612, 2010.
- [21] P. J. C. Dickinson, L. Gijben. On the computational complexity of membership problems for the completely positive cone and its dual. *Computational Optimization and Applications*, 57, 403-415, 2014.
- [22] 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.
- [23] X. V. Doan, X. Li, K. Natarajan. Robustness to dependency in portfolio optimization using overlapping marginals. *Operations Research*, 63(6), 1468-1488, 2015.
- [24] J. Edmonds. Submodular functions, matroids, and certain polyhedra in: *Combinatorial structures and their applications* (R. Guy, H. Hanani, N. Sauer, J. Schönmeier eds.), Gordon and Breach, 69-87, 1970.
- [25] P. Embrechts, A. McNeil, D. Straumann. Correlation and dependence in risk management: properties and pitfalls. *Risk Management: Value at Risk and Beyond*. 1st Edition, Cambridge University Press, Cambridge, 176-223, 2002.
- [26] 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.
- [27] R. Gao, A. Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research*, 48(2), 603-655, 2023.
- [28] R. Gao, A. Kleywegt. Data-driven robust optimization with known marginal distributions. Working Paper, 2017.
- [29] M. Grötschel, L. Lovász, A. Schrijver. *Geometric Algorithms and Combinatorial Optimization*. Springer, 1988.
- [30] S. Ghosh, S. G. Henderson. Chessboard distributions and random vectors with specified marginals and covariance matrix. *Operations Research*, 50(5), 820-834, 2002.

- [31] H. Heidari, C. Ferrari, K. P. Gummadi, A. Krause. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. NIPS 18: Proceedings of the 32nd International Conference on Neural Information Processing Systems, 1273-1283, 2018.
- [32] J. N. Hooker, H. P. Williams. Combining equity and utilitarianism in a mathematical programming model. Management Science, 58(9), 1682–1693, 2012.
- [33] D. Iancu, M. Sharma, M. Sviridenko. Supermodularity and affine policies in dynamic robust optimization. Operations Research, 61(4), 941-956, 2013.
- [34] D. Iancu, N. Trichakis. Fairness and efficiency in multiportfolio optimization. Operations Research, 62(6), 1283-1301, 2014.
- [35] 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.
- [36] D. Kuhn, S. Shafiee, W. Wiesemann. Distributionally robust optimization. Acta Numerica, 34, 579-804, 2025.
- [37] T. L. Lai, H. Robbins. Maximally dependent random variables. Proceedings of the National Academy of Sciences of the United States of America, 73(2), 286-288, 1976.
- [38] J. B Lasserre. Moments, Positive Polynomials and Their Applications. World Scientific, 2009.
- [39] H. Li , E. Delage, N. Zhu, M. Pinedo, S. Ma. Distributional robustness and inequity mitigation in disaster preparedness of humanitarian operations. Manufacturing and Service Operations Management, 26(1), 197–214, 2023
- [40] D. Z. Long, J. Qin, A. Zhang. Supermodularity in two-stage distributional robust optimization. Operations Research, 70(3), 1394–1409, 2023.
- [41] 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.
- [42] H-Y. Mak, Y. Rong, J. Zhang. Appointment scheduling with limited distributional information. Management Science, 61(2), 316–334, 2015.
- [43] J. H. McCoy, H. L. Lee. Using fairness models to improve equity in health delivery fleet management. Production and Operations Management, 23, 965-977, 2014.
- [44] K. Natarajan. Optimization with Marginals and Moments. Dynamic Ideas, 2021.
- [45] B. Pass. Multi-marginal optimal transport: theory and applications. ESAIM: Mathematical Modelling and Numerical Analysis, 49, 1771-1790, 2015.
- [46] K. Patton, M. Russo, S. Singla. Submodular norms with applications to online facility location and stochastic probing. In Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2023). Leibniz International Proceedings in Informatics (LIPIcs), Volume 275, pp. 23:1-23:22, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023.
- [47] I. Pitowsky. Correlation polytopes: Their geometry and complexity. Mathematical Programming, 50, 395–414, 1991.
- [48] I. Popescu. Applications of optimization in probability, finance and revenue management, PhD Thesis, Massachusetts Institute of Technology, 1999.
- [49] A. K. Ramachandra, K. Natarajan. Tight probability bounds with pairwise independence. SIAM Journal on Discrete Mathematics, 37(2), 516–555, 2023.
- [50] A. K. Ramachandra, K. Natarajan. Pairwise independent correlation gap. Operations Research Letters, 60, 107255, 2025.
- [51] L. Rüschendorf. Mathematical Risk Analysis: Dependence, Risk bounds, Optimal Allocations and Portfolios. Springer, Berlin, 2013.

- [52] J. S. Sturm, S. Zhang. On cones of nonnegative quadratic functions. *Mathematics of Operation Research*, 28(2), 246-267, 2003.
- [53] A. H. Tchen. Inequalities for distributions with given marginals. *The Annals of Probability*, 8(4), 814-827, 1980.
- [54] D. M. Topkis. *Supermodularity and Complementarity*. Princeton University Press, 1998.
- [55] R. M. Van Slyke. Letter to the Editor—Monte Carlo methods and the PERT problem. *Operations Research*, 11(5), 839–860, 1963.
- [56] W. Wiesemann, D. Kuhn, M. Sim. Distributionally robust convex optimization. *Operations Research*, 62(6), 1358-1376, 2014.