Submitted to *Operations Research*
manuscript (Please, provide the manuscript number!)

Distributionally Robust Linear and Discrete Optimization with Marginals

Louis Chen
Naval Postgraduate School, Monterey, CA 93940, louis.chen@nps.edu

Will Ma
Decision, Risk, and Operations Division, Columbia University, New York, NY 10027, wm2428@columbia.edu

Karthik Natarajan
Engineering Systems and Design, Singapore University of Technology and Design, Singapore 487372, karthik_natarajan@sutd.edu.sg

David Simchi-Levi
Department of Civil and Environmental Engineering, Institute for Data, Systems, and Society and Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA 02139, dslevi@mit.edu

Zhenzhen Yan
School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore, 637371, yanzz@ntu.edu.sg

Abstract In this paper, we study linear and discrete optimization problems in which the objective coefficients are random, and the goal is to evaluate a robust bound on the expected optimal value, where the set of admissible joint distributions is assumed to be specified only up to the marginals. We study a primal-dual formulation for this problem, and in the process, unify existing results with new results. We establish NP-hardness of computing the bound for general polytopes and identify two sufficient conditions - one based on a dual formulation, and one based on sublattices which provide a class of polytopes where the robust bounds are efficiently computable. We discuss several examples and applications in areas such as scheduling.

Key words: Marginal Distribution Model; Linear Programming; Duality; Optimal Transport

1. Introduction

In optimization problems, decisions are often made in the face of uncertainty that might arise in the form of random costs or benefits. Traditionally, optimization problems under uncertainty have been modeled with stochastic optimization (see Shapiro et al. (2014)). Let $\mathcal{S} \subset \mathbb{R}^n$ denote the feasible region of the decision vector s and $\tilde{\xi}$ denote a random vector defined on the support set $\Xi \subset \mathbb{R}^m$

with a probability distribution p. The decision s is made before knowing the true realization of the uncertain data and the outcome is the random cost function $h(s,\tilde{\xi})$. The stochastic optimization problem is to choose a decision to minimize the expected cost by solving $\min_{s\in\mathcal{S}} \mathbb{E}_{\tilde{\xi}\sim p}\left[h(s,\tilde{\xi})\right]$. This formulation however makes the assumption that the joint distribution p is known or at the very least, a sufficient number of independent and identically distributed samples from the distribution are available. Recently, there has been a growing interest in the "distributionally robust optimization" paradigm where this assumption is relaxed. The distribution p is only assumed to lie in an "ambiguity" set of probability distributions denoted by $\mathcal{P}$, but the exact distribution is itself unknown. The distributionally robust optimization problem is to choose a decision to minimize the worst-case expected cost of the form $\min_{s\in\mathcal{S}} \sup_{p\in\mathcal{P}} \mathbb{E}_{\tilde{\xi}\sim p}\left[h(s,\tilde{\xi})\right]$. Such a set $\mathcal{P}$ has been constructed in a variety of ways so as to ensure it is suitable for practical applications. At the same time, the choice of this set has important implications on the computational tractability of the distributionally robust optimization problem. Examples of the types of sets $\mathcal{P}$ that have been analyzed in the literature include the set of distributions with information on the mean and covariance matrix (see Delage and Ye (2010), Bertsimas et al. (2010), Goh and Sim (2010)), the set of distributions with information on the marginal distributions (see Meilijson and Nadas (1979)) and marginal moments (see Bertsimas et al. (2004)), the set of distributions with confidence sets and mean values residing on an affine manifold (see Wiesemann et al. (2014)), the set of distributions that lie in a ball around a reference probability distribution where the distance is defined using the ϕ-divergence measure (see Ben-Tal et al. (2013)) or the Wasserstein distance measure (see Esfahani and Kuhn (2018), Gao and Kleywegt (2016), Blanchet et al. (2019)). This list is by no means comprehensive with an increasing number of applications of this technique to areas such as supply chains, finance, healthcare and machine learning. We refer the interested reader to the papers listed above and the references therein.

In this paper, we contribute to this literature by providing results for the case where the cost is the optimal value to a linear and discrete optimization problem, hereafter written $Z(s,\tilde{c})$, and the ambiguity set is the Fréchet class $\Gamma(\mu_1,\ldots,\mu_n)$ of multivariate distributions with fixed marginal measures $\{\mu_i\}_{i=1}^n$ (see Definition 1), i.e., $\min_{s\in\mathcal{S}}\sup_{\theta\in\Gamma}\mathbb{E}_{\tilde{c}\sim\theta}[Z(s,\tilde{c})]$. More precisely, we are motivated by those cases in which there exists a polytope $Q\subseteq\mathbb{R}^n$ so that for any $s\in\mathcal{S}$, $Z(s,\cdot)$ will be a parameterized support function to Q, typically of the form $Z(s,\tilde{c}):=\max_{x\in Q}\sum_{i=1}^n(\tilde{c}_i-s_i)x_i$ (see Section 5 and Appendix EC.2 for discussion of some other parametrized forms.)

We investigate the computational complexity of the distributionally robust optimization problem that combines this form of cost function with this class of distributions, and identify conditions under which the problem is computationally tractable. Due to the form of Z described above, it suffices to study the inner sup term for arbitrarily fixed first stage decisions s. Hence it is without loss of generality, we focus on the study of $Z(\tilde{c}):=\max\limits_{x\in\mathcal{X}}\sum\limits_{i=1}^n\tilde{c}_ix_i$, where $\mathcal{X}$ will denote a finite set. Consequently, the main theoretical focus of this paper is on finding a tight upper bound on the expected value of this quantity in which the objective coefficients are chosen randomly from a distribution in the Fréchet class $\Gamma(\mu_1,\ldots,\mu_n)$ as follows:

$$Z^* = \sup_{\theta\in\Gamma(\mu_1,\ldots,\mu_n)}\mathbb{E}_{\tilde{c}\sim\theta}[Z(\tilde{c})] = \sup_{\theta\in\Gamma(\mu_1,\ldots,\mu_n)}\mathbb{E}_{\tilde{c}\sim\theta}[\tilde{c}^\top x^*(\tilde{c})], \tag{1}$$

where $\tilde{c}$ is a random vector and $x^*(\cdot)$ is an optimal solution mapping. The convex hull of the finite set $\mathcal{X}$, denoted by $conv(\mathcal{X})$, is necessarily a polytope with $Z(c)=\max\{\sum_{i=1}^n c_ix_i : x\in conv(\mathcal{X})\}$. In studying the complexity of computing Z^*, the polytope will be assumed to be specified through a list of extreme points or a system of linear inequalities (see Section 3). Following the terminology in Natarajan et al. (2009), we refer to the model in (1) as the Marginal Distribution Model (MDM).

1.1. Main Contributions

We make the following contributions to the literature. In Sections 2.2 and 2.3, we use results from the field of optimal transport to establish for Problem (1), a primal-dual pairing of optimization problems which we use to unify existing and new results. For instance, an immediate consequence is a generalization of the existing Theorem 1 in Natarajan et al. (2009). In Section 2.4 we present additional structural results regarding the set of primal optimal solutions by showing: (1) marginal uniqueness; (2) separable transport maps under absolutely continuous marginals. These serve to justify the assumptions on the marginal distributions made in previous works, and discuss its relation to the "persistency" notion as considered in Bertsimas et al. (2006), Natarajan et al. (2009).

We analyze the computational complexity of MDM using its dual representation. In contrast to previous work, which has focused primarily on 0/1 polytopes, we provide extensions towards understanding the complexity for general polytopes. We establish that computing Z^* is NP-hard for parallelotopes and in Section 4, we identify two different sufficient conditions for a polytope to admit an efficiently computable robust bound Z^*. The first sufficiency condition is motivated from the analysis of the dual form and shows tractability for V, 0/1 H and $L^\natural$ convex polytopes as special cases. The second sufficiency condition exploits the relationship between a sublattice structure, supermodularity, and monotone couplings in optimal transport. The analysis extends known existing results on tractable instances provided in Meilijson and Nadas (1979) and Bertsimas et al. (2004) for 0/1 polytopes and Mak et al. (2015) for the appointment scheduling problem for a specific integer polytope. Our results show that just the existence of a compact 0/1 extended formulation is not sufficient for tractability, it also has to be of an "appropriate" form. Finally, we discuss the implications of the results in Section 5. Detailed proofs, several examples and computational experiments are provided in the Electronic Companion (EC).

1.2. Literature Review

To the best of our knowledge, the study of problem (1) in the context of combinatorial optimization problems was initiated by Meilijson and Nadas (1979) who developed an upper bound on the expected completion time in a PERT network assuming only the marginal distributions of the activity times on the network are known. The set $\mathcal{X}$ in this example is defined as the set of directed paths (i.e., their 0/1 incidence vectors) from the start to the end node in a directed acyclic graph representation of the PERT network. The key contribution of Meilijson and Nadas (1979) was to propose a convex optimization formulation to compute the tight upper bound on the expected completion time that is valid for the class of joint distributions with the specified marginals. In a follow-up paper, Nadas (1979) proposed a numerical solution to solve this problem efficiently by applying a network flow algorithm for project cost curves. A lower bound on the probability that a given path in this PERT network is critical in the worst-case distribution is obtained from the Lagrange multiplier associated with the constraint determined by the path. Haneveld (1986) derived several more results for this model using convex duality. Weiss (1986) furthered this study to obtain bounds for other combinatorial optimization problems such as the shortest path, maximum flow and network reliability problem by using the theory of clutters and blocking clutters. It should be noted that in all these instances, $\mathcal{X}$ is given explicitly as a finite collection of 0/1 vectors. There is lesser focus in these works on computational complexity issues of the model.

The study of the bound in (1) for general discrete optimization was initiated in Natarajan et al. (2009). The feasible region considered in their work was $\mathcal{X} = \{x \in \mathbb{Z}_+^n : Ax \leq b, x_i \in [\alpha_i, \beta_i], \forall i \in [n]\}$. Under the assumption that the optimization problem almost surely admits a unique solution with an absolute continuity assumption of the marginal distributions, Natarajan et al. (2009) showed that the bound Z^* is computable as the solution to a concave maximization over the convex hull of a binary reformulation of the original feasible region. While they showcased the strength of the formulation by finding an upper bound on Z^* for a stochastic knapsack problem, the complexity of the bound for general linear and discrete optimization problems beyond 0/1 polytopes is not discussed in their work. Agrawal et al. (2012) studied this model for more general cost functions. In particular, they consider the problem $\min_{s \in \mathcal{S}} \sup_{\theta \in \Gamma(\mu_1,\ldots,\mu_n)} \mathbb{E}_{\tilde{\xi} \sim \theta}\left[h(s, \tilde{\xi})\right]$, where $h(s, \tilde{\xi})$ is a general cost function. They showed that for a given decision s, the problem of computing the worst-case expected value with Bernoulli random variables is NP-hard even when the function $g(\xi) := h(s, \xi)$ is monotone and submodular in ξ. Towards studying this further, they defined the "price of correlation", a measure of the cost of model misspecification, by comparing the performance of operating under the assumption of independence as opposed to operating under the assumption of worst-case coupling. Using ideas from cost-sharing analysis in cooperative game theory, they identified sufficient conditions under which this price is bounded and provided applications to stochastic combinatorial problems. In contrast, in this paper we restrict our attention to support functions over polytopes.

Another line of distributionally robust optimization based on marginal information is the marginal moment model (MMM). Birge and Maddox (1995), Bertsimas et al. (2004) and Bertsimas et al. (2006) studied MMM assuming the knowledge of a finite set of marignal moments and derived their tractable reformulations. In particular, Bertsimas et al. (2006) proposed an important concept - "persistency" value of binary variables as the marginal distribution of the optimal solution, which can be obtained from a primal semidefinite programming formulation of MMM. Van Parys et al. (2016) discussed this model for cost functions of the form $Z(\tilde{c}) := l\left(\sum_{i=1}^{n} \tilde{c}_i\right)$. Mak et al. (2015) furthered this research to derive a tractable SDP reformulation of MMM for an appointment scheduling problem given the first k marginal moments. Their tractability result is derived using a compact 0/1 extended formulation.

Lastly, the field of optimal transport theory and the related Monge-Kantorovich problem provides several fundamental tools including a well-developed duality theory (see Villani (2003, 2009), Rachev and Rüschendorf (1998), Galichon (2016)). Broadly speaking, this field is concerned with transporting mass between two given probability measure spaces (X,μ) and (Y,ν) at optimal cost. The Monge-Kantorovich problem is formulated as: $\max_{\pi \in \Gamma(\mu,\nu)} \int_{X \times Y} c(x,y) d\pi(x,y)$, where the optimization is over the set of all transference plans $\pi \in \Gamma(\mu,\nu)$ and c is the profit obtained from moving one unit of mass from x to y. Traditionally, optimal transport has been mainly concerned with finding optimal couplings of two distributions. There is limited work in multi-marginal optimal transport (see Pass (2015), Bach (2019)). This stream of literature also has lesser focus on applications to linear and discrete optimization problems under uncertainty which is what we are interested in.

2. MDM: Primal-Dual Formulations

In this section, we will present a primal-dual formulation for (1).

2.1. Preliminaries

2.1.1. Convex and Measure-Theoretic Notation We first establish some facts, notations, and basic measure-theoretic concepts that will be useful for discussion. For any function $f:\mathbb{R}^n \to \mathbb{R}$, let $f^*:\mathbb{R}^n \to \mathbb{R}$ denote the Legendre Fenchel conjugate, defined by $f^*(c) := \sup_{x\in\mathbb{R}^n}\{c^\top x - f(x)\}$ for any $c \in \mathbb{R}^n$. Let $f^{**} := (f^*)^*$, which is equivalent to f in the case of lower semicontinuous, convex f, and also $f^{***} := (f^{**})^* = f^*$ (Rockafellar (1997)). Notice that these definitions are constructed for functions defined on a connected space, as in $\mathbb{R}^n$. However, throughout this text we will often encounter functions defined on a finite subset $\mathcal{X} \subset \mathbb{R}$. If $\psi:\mathcal{X} \to \mathbb{R}$ is such a univariate discrete function, then to discuss convexity as well as the Legendre Fenchel conjugate, we will identify ψ with its natural extension to all of $\mathbb{R}$. More precisely, we will consider ψ_{ext} to be the univariate linear interpolation of ψ over $conv(\mathcal{X})$, with $\psi_{ext} := +\infty$ outside of $conv(\mathcal{X})$. Notice that under this definition, $\max_{x\in\mathbb{R}^n}\{c^\top x - \psi_{ext}(x)\} = \max_{x\in\mathcal{X}}\{c^\top x - \psi(x)\}$ for any $c \in \mathbb{R}^n$. We thus identify a discrete function ψ with its extension and call it "convex" when ψ_{ext} is, and write ψ^* in place of ψ^*_{ext} for simplicity. Next, for any convex function $f:\mathbb{R}^n \to (-\infty,+\infty]$, we denote the set of subgradients to f at x by $\partial f(x) := \{s \in \mathbb{R}^n : f(y) \geq f(x) + s^\top(y-x), \forall y \in \mathbb{R}^n\}$. When $Z:\mathbb{R}^n \to \mathbb{R}(-\infty,+\infty]$ is given by $Z(c) := \max_{x\in Q} c^\top x$ for some polyhedron Q, then we say Z is the support function of Q.

For the following discussion, when $\mathcal{S}$ is a topological space, the Borel sigma algebra (denoted $\mathcal{B}$) will be the sigma algebra generated by the family of open sets, and we let the set of all probability measures defined on this Borel sigma algebra (called Borel probability measures) be denoted by $\mathcal{P}(\mathcal{S})$. We will always use standard topologies, unless otherwise stated. Whenever the space is a finite set $\mathcal{X}$, we will assume the discrete topology, in which all subsets of $\mathcal{X}$ are considered open. Additionally, for any set $\mathcal{X} \subset \mathbb{R}^n$ (finite or not) let $\mathcal{X}_i = \{x_i : x = (x_1, x_2, \ldots, x_n) \in \mathcal{X}\}$. Denote the projection operation: $Proj_i(x) := x_i$ for any $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, so that $\mathcal{X}_i = Proj_i(\mathcal{X})$. We will also use a notion for projection of measures:

Definition 1. Let γ be a Borel probability measure on $\mathbb{R}^n$. Then for any i, define the "i-th projection" $\Pi_i\gamma$ of the measure γ by:

$$\Pi_i\gamma(A) := \gamma(\underbrace{\mathbb{R}\times\ldots\times\mathbb{R}}_{\text{(i-1)- copies}}\times A\times\underbrace{\mathbb{R}\times\ldots\times\mathbb{R}}_{\text{(n-i)- copies}}) \text{ for all Borel measurable subsets } A \text{ of } \mathbb{R}.$$

Given a collection of Borel probability measures $\{\mu_i\}_i$, each defined on $\mathbb{R}$, let $\Gamma(\mu_1,\dots,\mu_n) := \{\gamma \in \mathcal{P}(\mathbb{R}^n) : \Pi_i \gamma = \mu_i, \forall i \in [n]\}$. Observe that this set is nonempty because the product measure $\mu_1 \otimes \mu_2 \otimes \dots \otimes \mu_n$ (the "independent coupling") is trivially a member.

Unlike Natarajan et al. (2009), wherein the results were derived under the assumption that the optimization problem almost surely admits a unique solution, we allow for a nonzero probability of the event of multiple optimal solutions occuring. To facilitate discussion, we associate with the random optimal objective function $Z(\tilde{c}) := \max\{\tilde{c}^\intercal x : x \in \mathcal{X}\}$, the optimal solution multifunction (see Shapiro et al. (2014)) $x^{OPT} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ defined as $x^{OPT}(c) = \arg\max_{x\in\mathcal{X}} c^\intercal x$ for all $c \in \mathbb{R}^n$. This multifunction from $(\mathbb{R}^n, \mathcal{B})$ into $(\mathbb{R}^n, \mathcal{B})$ is closed-valued and measurable. Namely, $x^{OPT}(c)$ is closed in $\mathbb{R}^n$ for all c, and for every closed set $A \subset \mathbb{R}^n$, $\{c : x^{OPT}(c) \cap A \neq \varnothing\}$ is measurable. Hence, there exists a measurable selection of x^{OPT}, which we will write as $x^* : \mathbb{R}^n \to \mathbb{R}^n$, where $x^*(c) \in x^{OPT}(c)$. The measurable mapping $x^*(\cdot)$ then returns a possible optimal solution to any given vector c.

2.1.2. Lattices and Submodular Functions Let $\vee$ and $\wedge$ be binary operators defined on $\mathbb{R}^n \times \mathbb{R}^n$ by $x \vee y = (max(x_1,y_1),\dots,max(x_n,y_n))$ and $x \wedge y = (min(x_1,y_1),\dots,min(x_n,y_n))$ for any $x,y \in \mathbb{R}^n$. We will refer to $\mathbb{R}^n$ as a lattice with join and meet operations in $\vee$ and $\wedge$ respectively. Further, if a subset $S \subseteq \mathbb{R}^n$ is closed under the join and meet operations, then it is a sublattice.

Definition 2.1 (Submodular Function) *Let $f : \mathbb{R}^n \to \mathbb{R}$. We call f submodular if $f(x \vee y) + f(x \wedge y) \le f(x) + f(y), \forall x,y \in \mathbb{R}^n$. If $-f$ is submodular, we say f is supermodular.*

2.1.3. Results from Optimal Transport We will make use of (1) a core duality result and (2) a monotone coupling result from optimal transport theory.

Lemma 1 **(Kantorovich Duality, Theorem 5.10 Villani (2009))**. *Let (X,μ) and (Y,ν) be Polish probability spaces. Let $h : X \times Y \to [-\infty,\infty)$ be an upper semicontinuous function and suppose that there exist real, lower semicontinuous functions $a \in \mathcal{L}^1(X,\mu)$ and $b \in \mathcal{L}^1(Y,\nu)$ such that $h(x,y) \le a(x) + b(y)$ for all $(x,y) \in X \times Y$. Then,*

$$\max_{\pi\in\Gamma(\mu,\nu)} \int h(x,y)d\pi(x,y) = \inf_{\psi\in\mathcal{L}^1(X,\mu)} \int \psi(x)d\mu(x) + \int \sup_{x\in X}\{h(x,y) - \psi(x)\}d\nu(y)$$
$$= \inf_{\phi\in\mathcal{L}^1(Y,\nu)} \int \sup_{y\in Y}\{h(x,y) - \phi(y)\}d\mu(x) + \int \phi(y)d\nu(y),$$

where if $\max_{\pi\in\Gamma(\mu,\nu)}\int hd\pi > -\infty$ *and for some* $f\in\mathcal{L}^1(X,\mu)$ *and* $g\in\mathcal{L}^1(Y,\nu)$, $f(x)+g(y)\le h(x,y)$ *for all* $(x,y)\in X\times Y$, *then the* inf *can be written as* min.

This duality is a classic result, which has an extension that handles any number of marginals two or more - the precise statement can be found in Haneveld (1986). Next, if the cost function is supermodular, then an explicit characterization of the optimal transference plan is known. The following result is well-known (see Haneveld (1986), Galichon (2016)).

Lemma 2. (**Theorem 7.2 Haneveld (1986)**) *Let* $c:\mathbb{R}^n\to\mathbb{R}$ *be a continuous, supermodular function. If* $\{\mu_i\}_{i=1\ldots n}$ *is any collection of Borel measures on* $\mathbb{R}$, *i.e.,* $\mu_i\in\mathcal{P}(\mathcal{X}_i)$ *for all* i, *then*

$$\sup_{\gamma\in\Gamma(\mu_1,\ldots,\mu_n)}\int_{\mathcal{X}:=\mathcal{X}_1\times\ldots\times\mathcal{X}_n} c(x)\ d\gamma(x)=\int_0^1 c(F_{\mu_1}^{-1}(t),\ldots,F_{\mu_n}^{-1}(t))\,dt,$$

where $F_{\mu_i}(x_i):=\mu_i(y_i\in(-\infty,x_i])$, *and* $F_{\mu_i}^{-1}(t_i):=inf\{x_i:F_{\mu_i}(x_i)\ge t_i\}$.

From a probabilistic perspective, the lemma says the monotone coupling on the probability space $((0,1),\mathcal{B},\lambda)$ with λ denoting Lebesgue measure given by $(X_1(t),\ldots X_n(t)):=(F_{\mu_1}^{-1}(t),\ldots,F_{\mu_n}^{-1}(t))$ yields the largest expectation when the transportation cost is supermodular.

2.2. A Lagrangian Primal-Dual Formulation

In this paper, we will largely focus on optimization over polytopes; in Sections 3 and 4 where we formalize the complexity of MDM, $\mathcal{X}$ will described using a list of extreme points or a system of linear inequalities. In the EC section (EC.1), we discuss extensions when $\mathcal{X}$ is: (1) an unbounded polyhedron, and (2) a compact subset of $\mathbb{R}^n$.

Given the one-dimensional marginal distributions $\mu_1,\ldots,\mu_n$ and set $\mathcal{X}\subset\mathbb{R}^n$, we define the convex-concave Lagrangian function $L:\left(\mathbb{R}^{\mathcal{X}_1}\times\ldots\times\mathbb{R}^{\mathcal{X}_n}\right)\times\mathcal{P}(\mathcal{X})$ by

$$L(\{\psi_i\}_{i=1}^n,\nu):=\sum_i\left(\int\psi_i d\Pi_i\nu+\int\max_{x_i\in\mathcal{X}_i}\{\tilde{c}_ix_i-\psi_i(x_i)\}d\mu_i\right),\qquad\text{(Lagrangian)}$$

The first key result is a minimax theorem.

Theorem 1 **(Primal-Dual Formulation of MDM).** *Let $\mathcal{X} \subset \mathbb{R}^n$ be finite, $Z(c) := \max_{x\in\mathcal{X}} c^\intercal x$ with n Borel probability measures $\{\mu_i\}_{i=1}^n$ over $\mathbb{R}$ given. Then*

$$\max_{\theta\in\Gamma(\mu_1,\ldots,\mu_n)} \mathbb{E}_{\tilde{c}\sim\theta}\left[Z(\tilde{c})\right] = \max_{\nu\in\mathcal{P}(\mathcal{X})} \sum_i \max_{\gamma_i\in\Gamma(\mu_i,\Pi_i\nu)} \mathbb{E}_{(\tilde{c}_i,\tilde{x}_i)\sim\gamma_i}\left[\tilde{c}_i\tilde{x}_i\right]. \tag{†}$$

If, in addition, $\mathbb{E}_{\tilde{c}_i\sim\mu_i}\left[|\tilde{c}_i|\right] < \infty$ for all i, then the bound is equal to

$$\max_{p\in\mathbb{R}_+^{\mathcal{X}},\sum_{x\in\mathcal{X}} p_x=1} \min_{\{\psi_i:\mathcal{X}_i\to\mathbb{R}\}_{i=1}^n} \sum_{x\in\mathcal{X}} p_x\Big(\sum_{i=1}^n \psi_i(x_i)\Big) + \sum_{i=1}^n \mathbb{E}_{\tilde{c}_i\sim\mu_i}\left[\max_{x_i\in\mathcal{X}_i} \tilde{c}_i x_i - \psi_i(x_i)\right] \tag{Primal}$$

$$= \min_{\{\psi_i:\mathcal{X}_i\to\mathbb{R}\}_{i=1}^n} \max_{p\in\mathbb{R}_+^{\mathcal{X}},\sum_{x\in\mathcal{X}} p_x=1} \sum_{x\in\mathcal{X}} p_x\Big(\sum_{i=1}^n \psi_i(x_i)\Big) + \sum_{i=1}^n \mathbb{E}_{\tilde{c}_i\sim\mu_i}\left[\max_{x_i\in\mathcal{X}_i} \tilde{c}_i x_i - \psi_i(x_i)\right]. \tag{Dual}$$

Finally, the ψ_i may be constrained to be convex.

We provide here only the key steps to a short, analytical proof using Lemma 1 - with justifications left to the full proof of a slightly more general result in EC.3.1.2. For the reader interested in actual construction of extremal couplings, we refer to Proposition 1 and its proof in EC.3.1.6. The proof is from the following set of equalities:

$$\sup_{\theta\in\Gamma(\mu_1,\ldots,\mu_n)} \mathbb{E}_{\tilde{c}\sim\theta}\left[Z(\tilde{c})\right] = \max_{\theta\in\Gamma(\mu_1,\ldots,\mu_n)} \mathbb{E}_{\tilde{c}\sim\theta}\left[\max_{x\in\mathcal{X}} \tilde{c}^\intercal x\right] = \max_{(\tilde{c},\tilde{x}):\tilde{c}\sim\theta\in\Gamma,\tilde{x}\sim\nu\in\mathcal{P}(\mathcal{X})} \mathbb{E}_{\tilde{c}\sim\theta}\left[\tilde{c}^\intercal\tilde{x}\right] \tag{2}$$

$$= \max_{\{(\tilde{c}_i,\tilde{x}_i)\}_{i=1}^n:\tilde{c}_i\sim\mu_i,\tilde{x}\sim\nu\in\mathcal{P}(\mathcal{X})} \sum_{i=1}^n \mathbb{E}_{(\tilde{c}_i,\tilde{x}_i)}\left[\tilde{c}_i\tilde{x}_i\right] \tag{3}$$

$$= \max_{\nu\in\mathcal{P}(\mathcal{X})} \sum_{i=1}^n \max_{\gamma_i\in\Gamma(\mu_i,\pi_i\nu)} \mathbb{E}_{(\tilde{c}_i,\tilde{x}_i)\sim\gamma_i}\left[\tilde{c}_i\tilde{x}_i\right] \tag{†}$$

$$= \max_{\nu\in\mathcal{P}(\mathcal{X})} \sum_{i=1}^n \min_{\psi_i:\mathcal{X}_i\to\mathbb{R}} \int \psi_i(x_i)\ d\pi_i\nu + \int \psi_i^*(\tilde{c}_i)\ d\mu_i \tag{4}$$

$$= \min_{\{\psi_i:\mathcal{X}_i\to\mathbb{R}\}_{i=1}^n} \max_{\nu\in\mathcal{P}(\mathcal{X})} \sum_{i=1}^n \int \psi_i(x_i)\ d\pi_i\nu + \int \psi_i^*(\tilde{c}_i)\ d\mu_i. \tag{5}$$

The abbreviated proof [1]illustrates two themes that will be investigated in more detail in Sections 2.4 and 4 respectively:

1. **Primal Solutions $\tilde{x}\sim\nu$ are Measurable Selections of x^{OPT}:** As (2) indicates, at optimality $\tilde{x}$ and $\tilde{c}$ satisfy $\tilde{x}\in x^{OPT}(\tilde{c})$, making $\tilde{x}$ a measurable selection. As well, (3) and (†) indicate that given any optimal $\tilde{x}\sim\nu$, the n separate $(\tilde{c}_i,\ \tilde{x}_i)$ transport problems are sufficient to fully specify an

[1] In reference to Section EC.1.2, we highlight that the finiteness of $\mathcal{X}$ is not needed to establish the minimax identity.

optimal $\tilde{c}$. This separation feature yields structural results to primal optimal $\tilde{x}$ (see Corollary 3, Proposition 1), with connections to the notion of "persistency" (Natarajan et al. (2009), Bertsimas et al. (2006)).

2. **The (Dual) MDM Separation Problem:** In (5), the inner maximization problem will be identified as a separation problem (see Proposition 2) which will be important for the analysis of the complexity. We note here that it only involves $\nu \in \mathcal{P}(\mathcal{X})$ by way of its marginals and being able to characterize $\{(\Pi_1\nu, \Pi_2\nu, \dots, \Pi_n\nu) : \nu \in \mathcal{P}(\mathcal{X})\}$ (see Theorem 3) will be one direction to identify "tractability" (to be defined in Section 3).

2.3. Primal and Dual Variables -"Saddle Point" Conditions

To facilitate the forthcoming discussions on the primal and dual problems in which we assume $\mathcal{X}$ is finite, we establish the following primal and dual terminology.

Definition 2.2 *A measure $\bar{\nu} \in \mathcal{P}(\mathcal{X})$ is called a primal optimal solution if it solves the maximization problem in (Primal). A collection of functions $(\psi_1, \dots, \psi_n)$ such that $\psi_i : \mathcal{X}_i \to \mathbb{R}$ for every i is called a dual optimal solution if it solves the minimization problem in (Dual).*

Corollary 1 **("Saddle-point" conditions)**. *Suppose $\mathcal{X} \subset \mathbb{R}^n$ is finite and let $(\bar{\nu}, \{\bar{\psi}_i\}_i)$ be a pair of optimal primal and dual solutions, with $\bar{\psi}_i$ convex on $\mathcal{X}_i$. Then the following conditions are necessary:*

1. $\Pi_i\bar{\nu}(x_i) \in \left[\mu_i\left(\partial\bar{\psi}_i(x_i) \setminus \cup_{x_i' \neq x_i} \partial\bar{\psi}_i(x_i')\right), \mu_i\left(\partial\bar{\psi}_i(x_i)\right)\right], \forall i \in [n], \forall x_i \in \mathcal{X}_i$
2. $\int_{\mathcal{X}} \sum_i \bar{\psi}_i(x_i) d\nu \leq \int_{\mathcal{X}} \sum_i \bar{\psi}_i(x_i) d\bar{\nu}, \forall \nu \in \mathcal{P}(\mathcal{X})$.

If the μ_is are absolutely continuous, then the two conditions are sufficient for $\bar{\nu}$ and $\{\bar{\psi}_i\}_i$ to be primal and dual optimal. In addition, condition 1. is equivalent to:

$$\Pi_i\bar{\nu}(x_i) = \mu_i\left[\partial\bar{\psi}_i(x_i)\right], \forall i \in [n], \forall x_i \in \mathcal{X}_i, \tag{6}$$

The second necessary condition in Corollary 1 is an important subproblem in the study of complexity which is examined in Section 4.1. The first necessary condition provides a way to estimate the primal optimal marginals $\Pi_i\bar{\nu}$ using the corresponding input marginals μ_i and dual optimal solutions $\bar{\psi}_i$. In (6), a dual optimal solution on hand can in fact exactly specify $\Pi_i\bar{\nu}$ (see Corollary 3) when the input marginals are absolutely continuous.

2.4. The MDM Primal

In this section, we examine (Primal) - both its objective and solutions. First, regarding the objective, we establish that (†) is a generalization of a form provided in Natarajan et al. (2009), as it relaxes the absolute continuity assumption made therein. Second, all primal optimal solutions are measurable selections from $x^{OPT}(\tilde{c})$, for any distribution optimal to (1). Consequently, (Primal) does not exhibit a unique optimal solution in general. Third, we demonstrate the manner in which absolutely continuous μ_i gives rise to a number of useful structural properties on the optimal solutions of (Primal) by effectively ensuring that $\arg\max_{x\in\mathcal{X}} \tilde{c}^\intercal x$ is a singleton. It yields "marginal uniqueness" as well as a justification for the computation of "persistency" pursued in the works Bertsimas et al. (2006) and Natarajan et al. (2009).

2.4.1. Primal Objective Form With (†), Lemma 2 immediately yields the following generalization of Natarajan et al. (2009)'s Theorem 1, which will facilitate discussions around upcoming examples. Towards presenting the generalized statement, we introduce the following notation. Let $x_i - e_i$ for $x_i \in \mathcal{X}_i$, denote the largest element in $\mathcal{X}_i$ that is less than but not equal to x_i; we let $x_i - e_i := -\infty$, if no such largest element exists.

COROLLARY 2 **(Theorem 1, Natarajan et al. (2009))**. *Let $\mathcal{X} \subset \mathbb{R}^n$ be finite and $\{\mu_i\}_{i=1}^n$ be n Borel probability measures over $\mathbb{R}$. Then,*

$$\max_{\theta\in\Gamma(\mu_1,\ldots,\mu_n)} \mathbb{E}_{\tilde{c}\sim\theta} Z(\tilde{c}) = \max_{\nu\in\mathcal{P}(\mathcal{X})} \sum_i \sum_{x_i\in\mathcal{X}_i} x_i \int_{\Pi_i\nu((-\infty,x_i-e_i])}^{\Pi_i\nu((-\infty,x_i])} F_{\mu_i}^{-1}(t)dt. \tag{7}$$

In particular, for any $\bar{\nu}$ solving the concave maximization problem (7), we can construct a maximizing joint distribution $\theta^*_{\bar{\nu}} \in \Gamma(\mu_1,\ldots,\mu_n)$.

2.4.2. Primal Solutions as Measurable Selections of $x^{OPT}(\tilde{c})$ From (2), any primal optimal solution $\bar{\nu}$ is the probability law to a random vector $\tilde{x}$ that is coupled to a $\tilde{c}$ whose own probability law θ^* solves (1); further, these two random vectors satisfy the relation $\tilde{x} \in x^{OPT}(\tilde{c})$ - making $\tilde{x}$ a measurable selection of $x^{OPT}(\tilde{c})$. We emphasize that the probability law $\bar{\nu}$, however, is distinct from the probability law of $x^{OPT}(\tilde{c})$. Crucially, the set $x^{OPT}(\tilde{c})$ is in general not a singleton and so for any $x \in \mathcal{X}$, it holds that $\bar{\nu}(x) \leq \theta^*(x \in x^{OPT}(\tilde{c}))$, without equality necessarily holding. We illustrate with the following example on finite-support input marginals.

Example 2.1 *Let* $\mathcal{X} = \{(0,0),(0,1),(1,0)\} \subset \mathbb{R}^2$*, and assume* $\tilde{c}_1$*,* $\tilde{c}_2$ *both take values 1 and 0 with probabilities* $1/3$ *and* $2/3$ *respectively. It can be verified that the unique* $\theta^* = \underset{\theta \in \Gamma(\mu_1,\ldots,\mu_n)}{\text{argmax}} \mathbb{E}_{\tilde{c}\sim\theta} Z(\tilde{c}) = 1/3 \cdot \mathbb{1}_{(1,0)} + 1/3 \cdot \mathbb{1}_{(0,1)} + 1/3 \cdot \mathbb{1}_{(0,0)}$*. Then the MDM Primal has an uncountable collection of optimal solutions* $\{\nu_{\epsilon_1,\epsilon_2} = (1/3+\epsilon_1)\mathbb{1}_{(1,0)} + (1/3+\epsilon_2)\mathbb{1}_{(0,1)} + (1/3-\epsilon_1-\epsilon_2)\mathbb{1}_{(0,0)} : \epsilon_1,\epsilon_2 \geq 0, \epsilon_1+\epsilon_2 \leq 1/3\}$*. Towards interpreting this family of solutions, we define on the probability space* $([0,1),\mathcal{B},\lambda)$*,* $\tilde{c}(\omega) = (1,0)\cdot\mathbb{1}_{[0,1/3)}(\omega) + (0,1)\cdot\mathbb{1}_{[1/3,2/3)}(\omega) + (0,0)\cdot\mathbb{1}_{[2/3,1)}(\omega)$*, implying* $x^{OPT}(\tilde{c}(\omega)) = \{(1,0)\}\cdot\mathbb{1}_{[0,1/3)}(\omega) + \{(0,1)\}\cdot\mathbb{1}_{[1/3,2/3)}(\omega) + \{(1,0),(0,0),(0,1)\}\cdot\mathbb{1}_{[2/3,1)}(\omega)$*. Then any measurable selection* x^* *of* $x^{OPT}(\tilde{c})$ *has a corresponding probability law* $\nu_{\epsilon_1,\epsilon_2}$ *for some* ϵ_1 *and* ϵ_2 *from the constructed* $\tilde{c}\sim\theta^*$*. For example, the measurable selection* $x^*(\omega) = (1,0)\cdot\mathbb{1}_{[0,1/3)\cup[2/3,2/3+\epsilon_1)}(\omega) + (0,1)\cdot\mathbb{1}_{[1/3,2/3)\cup[2/3+\epsilon_1,2/3+\epsilon_1+\epsilon_2)}(\omega) + (0,0)\cdot\mathbb{1}_{[2/3+\epsilon_1+\epsilon_2,1)}(\omega)$ *has* $\nu_{\epsilon_1,\epsilon_2}$ *as law when* $\epsilon_1,\epsilon_2 \geq 0, \epsilon_1+\epsilon_2 \leq 1/3$.

2.4.3. Primal Solutions under Absolutely Continuous Marginals When the input marginals $\{\mu_i\}_i$ are assumed to be absolutely continuous, we recover the setting of Natarajan et al. (2009). In this section, we establish two structural properties to the set of primal optimal solutions under this assumption: (1) marginal uniqueness; (2) separability. The motivation to study the behavior of the random vector $x^{OPT}(\tilde{c}) = (x_1^{OPT}(\tilde{c}), x_2^{OPT}(\tilde{c}), \ldots, x_n^{OPT}(\tilde{c}))$ has been extensively discussed in prior work. For example, it can encode a consumer's random choice as part of a random utility model (Natarajan et al. (2009), Mishra et al. (2014)) or it can represent a set of "critical" arcs to a network with stochastic arc lengths (Bertsimas et al. (2006)).

Marginal Uniqueness Recall that Example 2.1 establishes that there may be multiple (even uncountably many) primal optimal solutions in general, even in the case of a unique $\tilde{c}\sim\theta^*$ solving (1); after all, there are multiple measurable selections when x^{OPT} is a multi-valued mapping. The assumption of absolutely continuous marginals corrects this to some degree: while there may still be multiple primal optimal solutions (see example in EC.3.1.4), all their marginals agree.

Corollary 3 **(Marginal Uniqueness in Primal Optimality)**. *Suppose* $\mathcal{X} \subset \mathbb{R}^n$ *is finite and that the Borel probability measures* $\mu_1,\ldots,\mu_n$ *are all absolutely continuous. If* $\bar{\nu}$ *and* $\bar{\tau}$ *are both primal optimal solutions, then* $\Pi_i\bar{\nu} = \Pi_i\bar{\tau}$ *for all* i*.*

This result justifies the use of the primal formulation as a discrete choice model in Natarajan et al. (2009)'s, in which $\mathcal{X} \subseteq \{0,1\}^n$ denotes the possible assortments of items that can be selected and $\tilde{c}_i \sim \mu_i$ represents the random utility an agent has for item i, with $Z(\tilde{c}) := \max_{x \in \mathcal{X}} \tilde{c}^\top x$ being the utility an agent receives upon choosing the utilitiy-maximizing item. Indeed, as the $\Pi_i \bar{\nu}$ characterize the probability that option i is chosen, i.e., they form the choice probabilities. Corollary 3 indicates that all primal optimal probability measures over $\mathcal{X}$ yield the same choice probabilities (see Section EC.2.2 for a generalization to bundle selection).

Separability and Persistency The next result that follows from absolutely continuous marginals will establish the existence of a separable function that generates the unique primal optimal marginals. In the domain of choice modeling, this is interpreted as the existence of an (representative) agent whose decision to choose item i depends only on their utility for item i.

PROPOSITION 1 **(Transport Maps and Persistency)**. *Suppose $\mathcal{X} \subset \mathbb{R}^n$ is finite and that the Borel probability measures $\mu_1, \ldots, \mu_n$ are all absolutely continuous. Then there exists a measurable function $x^* : \mathbb{R}^n \to \mathcal{X}$ that takes the form $x^*(c) = (x_1^*(c_1), \ldots, x_n^*(c_n))$ and satisfies $x^*(c) \in x^{OPT}(c)$ for all $c \in \times_{i=1}^n supp(\mu_i)$, such that for any primal optimal solution $\bar{\nu} \in \mathcal{P}(\mathcal{X})$,*

$$P_{\tilde{c} \sim \theta^*}(x_i^{OPT}(\tilde{c}) = \xi_i) = P_{\tilde{c} \sim \theta^*}(x_i^*(\tilde{c}) = \xi_i) = P_{\tilde{c}_i \sim \mu_i}(x_i^*(\tilde{c}_i) = \xi_i) = \Pi_i \bar{\nu}(\xi_i), \forall \xi_i \in \mathcal{X}_i, \forall i \in [n], \tag{8}$$

where $\theta^* \in \Gamma(\mu_1, \ldots, \mu_n)$ is any optimal solution to MDM (1).

We conclude the section by making several remarks on the relation to "persistency". The probability $P_{\tilde{c} \sim \theta^*}(x_i^{OPT}(\tilde{c}) = \xi_i)$, for any distribution θ^* solving (1), is referred to as a *persistence value* in Natarajan et al. (2009), Bertsimas et al. (2006). In (8), it is shown that these values are precisely the unique primal optimal marginals. More precisely, we see that the x_i^* component function is a transport map taking μ_i into $\Pi_i \bar{\nu}$, the unique i-th primal optimal marginal. We remark that this separability structure may not exist in the absence of absolutely continuous marginals - for example, consider $\tilde{c}_1 = 1/2$ $w.p.1$ and $\tilde{c}_2 \sim Unif(0,1)$. If $\mathcal{X} = \{(1,0),(0,1)\}$, then the optimal value of x_1 must depend on $\tilde{c}_2$. However, if both $\tilde{c}_1, \tilde{c}_2 \sim Unif(0,1)$, then by Corollary 2, the optimal

$\bar{\nu}$ places 1/2 probability on each of the two elements of $\mathcal{X}$, and it can be verified that $x_1(c_1) := \mathbb{1}_{c_1>1/2}$, $x_2(c_2) := \mathbb{1}_{c_2>1/2}$ satisfy equation (8) of Proposition 1. We refer the reader to the EC.3.1.6 for the proof of Proposition 1, which outlines explicitly how to construct x^* and EC.3.1.7 for examples illustrating persistency and separability.

3. The MDM Dual and Computational Complexity

In this section, we formalize the computational complexity of the MDM optimization problem (1). The problem is specified by the combination of a marginal and polytope class.

Marginal Class $\mathcal{M}$ We let $\mathcal{M} = (\mu_1, \mu_2, \ldots)$ denote the set of measures. Our results about computational complexity hold if the marginal measures are all finite-support and input in explicit form. More generally, we allow for each μ_n to be any Borel probability measure on $\mathbb{R}$, and assume the input of μ_i in the form of a (first-order) oracle $\mathcal{O}_i$ that upon given a function ψ_i, evaluates the expectation function as well as provides a subgradient, as formalized below.

Assumption 1. *For each measure $\mu_i \in \mathcal{M}$, let $\mathbb{E}_{\tilde{c}_i \sim \mu_i}[|\tilde{c}_i|] < \infty$. Given any $\psi_i : \mathcal{X}_i \to \mathbb{R}$, the function $F(\psi_i) := \int max_{x_i \in \mathcal{X}_i} \tilde{c}_i x_i - \psi_i(x_i) d\mu_i$ (as well as a subgradient) can be obtained efficiently - granted via oracle access, or computed efficiently (in the case of finite-support).*

Polytope Class Π_* For a polytope $Q = conv(\mathcal{X})$ we consider both the cases where:

1. $\mathcal{X}$ is explicitly input as a list of vectors $\{x^1, x^2, \ldots, x^m\} \subset \mathbb{R}^n$, with given integer m, in which case we say Q is given as a *V-polytope*;

2. We are given $conv(\mathcal{X})$ as a polyhedron, input in the form $\{x : Ax \le b\}$, where A is a matrix with m rows and n columns and b is a vector with m entries, in which case we say Q is given as an *H-polytope*.

If for each $n \in \mathbb{N}$, P_n is a family of polytopes in $\mathbb{R}^n$, then the collection of all these families forms a polytope class, $(P_1, P_2, \ldots)$, denoted generically by Π_*, and specifically by Π_V or Π_H if the entire collection of polytopes is in V or H representation respectively. For any polytope Q in n decision variables that is from this collection, it will be understood implicitly that $Q \in P_n$.

Problem MDM($\mathcal{M}\Pi_*$)

DEFINITION 2. Let a collection of measures $\mathcal{M} = (\mu_1, \mu_2, \ldots)$ and family of polytopes Π_* be given. Then **MDM**($\mathcal{M}\boldsymbol{\Pi}_*$) will denote the following optimization problem:

Instance: Positive integer n, a polytope Q from Π_* in n variables.

Question: Compute $Z^* = \sup_{\theta\in\Gamma(\mu_1,\ldots,\mu_n)} \mathbb{E}_{\tilde{c}\sim\theta}\left[Z(c) := \max_{x\in Q} \tilde{c}^\top x\right]$.

One example that we will consider is $\mathcal{M} = (\mu_1, \mu_2, \ldots)$ consisting of finite-support measures, where $\mu_n := (1-p_n)\mathbb{1}_{-1} + p_n\mathbb{1}_1$ for $p_n \in [0,1]$. In addition, let Π_H be the class of parallelotopes in $\mathbb{R}^n$, for all n. Analytically, the parallelotopes of $\mathbb{R}^n$ are of the form c + $\sum_{i=1}^m [-x^i, x^i]$, where $x^1, \ldots, x^m \in \mathbb{R}^n$ are linearly independent vectors ($m \le n$), and $c \in \mathbb{R}^n$ so that we call $(c, x^1, \ldots, x^m)$ the ρ - representation. In the case of parallelotopes, passage between $\rho-$ and H-representations can be done in polynomial time (Bodlaender et al. (1990)) so we may regard them as equivalent. As we will see later, for this specific class of problems, computing the bound, equiv., the expectation of Z with respect to an extremal θ, is NP-hard. But we emphasize that though this expectation can be hard, Z itself will be regarded as efficient, which will be later exploited in Section 4.2.

REMARK 1 ($Z(\cdot)$ IS EFFICIENT). If $\Pi_* = \Pi_V$, then for any Q, $Z(c) = \max(c^\top x^1, \ldots, c^\top x^m)$ for some finite collection of vectors $\{x^1, \ldots, x^m\}$ in some dimension n - clearly computable in polynomial (in inputs n and m) time. On the other hand, if $\Pi_* = \Pi_H$, then for any Q, $Z(\cdot)$ will be the support function, so there exists a polynomial (in inputs n and Q) time algorithm to compute $Z(c)$ for any c (Khachiyan (1980)).

MDM Dual and its Separation Problem We begin by identifying the separation problem for the MDM Dual, which will allow us to understand the complexity of $MDM(\mathcal{M}\Pi_*)$.

PROPOSITION 2. *Suppose $(\mu_1, \mu_2, \ldots, \mu_n)$ are Borel probability measures on $\mathbb{R}^n$ such that $\mathbb{E}_{\tilde{c}_i\sim\mu_i}[|\tilde{c}_i|] < \infty$ for all i, and let $\mathcal{X} \subset \mathbb{R}^n$ be finite. Then:*

$$Z^* = \min_{\psi_i:\mathcal{X}_i\to\mathbb{R}} \sum_{i=1}^n \int \psi_i^*(\tilde{c}_i) d\mu_i \tag{9}$$

$$s.t. \sum_{i=1}^n \sum_{x_i\in\mathcal{X}_i} \psi_i(x_i)\Pi_i\nu(x_i) \le 0, \forall \nu \in \mathcal{P}(\mathcal{X})$$

The proof can be found in EC.3.2.1. With (9) in view, it is revealed that we may solve the problem $MDM(\mathcal{M}\Pi_*)$ with the ellipsoid method, realizing its theoretically favorable complexity, if:

1. The subgradients to the objective function $\sum_{i=1}^n \int \psi_i^*(\tilde{c}_i)d\mu_i$ can be efficiently obtained for any dual solution $\{\psi_i\}_i$.

2. The **MDM separation problem:**

$$\max_{\nu\in\mathcal{P}(\mathcal{X})}\sum_{i=1}^n\sum_{x_i\in\mathcal{X}_i}\psi_i(x_i)\Pi_i\nu(x_i)=\max_{x\in\mathcal{X}}\sum_{i=1}^n\psi_i(x_i), \tag{10}$$

where $\mathcal{X}$ is the set of extreme points to Q, can be solved efficiently for any dual solution $\{\psi_i\}_i$.

While the first condition is ensured by Assumption 1, the second condition concerns efficiently solving the separation problem in the size of the input: n and Q. Under these considerations, by the equivalence of separation and optimization, we can analyze the computational complexity for $MDM(\mathcal{M}\Pi_*)$ by equivalently analyzing the computational complexity of the separation problem (10).

Definition 3. Given a family of polytopes Π_* and a collection $\mathcal{M}$ satisfying Assumption 1, we say $\mathbf{MDM(\mathcal{M}\Pi_*)}$ is "efficient", "efficiently computable", "tractable", or "poly-time solvable" if there exists an efficient algorithm to solve the separation problem (10) for any dual solution $\{\psi_i\}_i$.

Recalling from Theorem 1 that we may restrict the set of dual solutions ψ_i to be convex, and taking into account (10)'s right hand side, we see that the MDM separation problem amounts to maximizing additively separable convex functions over a polytope. The problem of maximizing the 1-norm of a vector over polytopes is known to be NP-hard (Mangasarian and Shiau (1986)). It remains so when restricted to the class of parallelotopes (Bodlaender et al. (1990)). By setting ψ_i to absolute value functions, we obtain a reduction that establishes NP-hardness in the following theorem.

Theorem 2. *Let $\mathcal{M}$ satisfy Assumption 1. Then:*

- *$MDM(\mathcal{M}\Pi_V)$ is polynomial-time solvable for any collection of V-polytopes Π_V.*

- $MDM(\mathcal{M}\Pi_H)$ *is NP-hard in general. More precisely, it is NP-hard when* $\mathcal{M}$ *consists of finite-support marginal measures and* Π_H *is the collection of all H-represented parallelotopes in* $\mathbb{R}^n$ *for arbitrary* n*, with* ρ *representation having* $c = 0$ *and* $x^1, \ldots, x^m \in \{-1, 0, 1\}^n$.

The statement in Theorem 2 formalizes results in Meilijson (1991) and Lai and Robbins (1976) to general V-polytopes. Theorem 2 complements the NP-hardness results for computing the worst-case bounds in the distributionally robust optimization literature. Closely related to the work, Agrawal et al. (2012) show that when $\mu_1, \ldots, \mu_n$ are measures concentrated on $\{0,1\}$, and Z is a function on $\{0,1\}^n$, then the problem of computing $\sup_{\theta \in \Gamma(\mu_1,\ldots,\mu_n)} \mathbb{E}_{\tilde{c}\sim\theta}[Z(\tilde{c})]$ is NP-hard, even when Z is monotone and submodular using a reduction from the MAX-CUT problem. Meanwhile, Bertsimas et al. (2010) proved that a distributionally robust linear optimization problem with a given mean and covariance matrix is NP-hard using a reduction from the 2-norm maximization over a polytope.

4. Sufficient Conditions for Tractability: The MDM Dual and Supermodularity of Z

As discussed in the previous section, $MDM(\mathcal{M}\Pi_*)$ is computationally challenging even when $\mathcal{M}$ consists of finite-support measures and when Π_* is the class of H-represented parallelotopes. It is natural to investigate conditions on Π_* for which $MDM(\mathcal{M}\Pi_*)$ is tractable - equivalently, for which the MDM Separation problem is tractable - assuming Assumption 1 on $\mathcal{M}$ holds. We next discuss two types of sufficiency conditions on Π_*: 1.) when the MDM Separation Problem can be solved efficiently in the size of the input of n and $Q \in \Pi_*$; 2.) the presence of supermodularity makes for a straightforward characterization of the extremal coupling. Following this, we will compare and contrast the conditions.

4.1. Sufficiency via $0/1$ Formulation

From the left hand side of (10), we observe that the decision space is the collection of marginals $\{\{\Pi_i \nu\}_{i=1}^n : \nu \in \mathcal{P}(\mathcal{X})\}$ to all possible joint distributions ν (and not necessarily the ν itself). Theorem 3 show that if this collection can be encoded with a particular polyhedral structure, then the separation problem reduces to linear programming.

THEOREM 3 (0/1 **Formulation of Separation Problem**). *Let $\mathcal{X}$ be a finite point set. Suppose there exists a 0/1 polytope $P \subseteq [0,1]^B$, where B is some finite set, such that the family of all possible collections of marginals $\{(\Pi_1\nu, \Pi_2\nu \dots \Pi_n\nu) : \nu \in \mathcal{P}(\mathcal{X})\}$ is in bijection with P such that for any $y \in P$, the corresponding marginals $(\Pi_1\nu, \Pi_2\nu \dots \Pi_n\nu)$ are given by*

$$\Pi_i\nu(\xi_i) = \sum_{S \in B_i(\xi_i)} y(S), \forall i \in [n], \forall \xi_i \in \mathcal{X}_i, \tag{11}$$

where the $B_i(\xi_i) \subseteq 2^B$, and $y(S) := \sum_{s \in S} y_s$. Then

$$\max_{\nu \in \mathcal{P}(\mathcal{X})} \sum_{i=1}^n \sum_{x_i \in \mathcal{X}_i} \psi_i(x_i)\Pi_i\nu(x_i) = \max_{y \in P} \sum_{i=1}^n \sum_{x_i \in \mathcal{X}_i} \psi_i(x_i) \sum_{S \in B_i(x_i)} y(S).$$

In particular, the separation problem (10) reduces to a linear program over a 0/1 polytope.

It is useful to interpret the identification of the 0/1 polytope P and set functions B_i as identifying a probability space with a structured collection of events such that for any $\mathcal{X}-$ valued random variable $\tilde{x}$ on this space, its marginals $\tilde{x}_i$ have their sigma algebras specified by B_i, i.e., $B_i(\xi_i)$ is essentially the collection of events for which $\tilde{x}_i = \xi_i \in \mathcal{X}_i$. Inherently, $\mathcal{X}_i$ will need to be identified, or in fact it suffices to efficiently identify, for each i, a superset $\mathcal{X}_i' \supset \mathcal{X}_i$ (see EC.2). We stress that the pieces P and B_i need not be unique, as we will see in Section 5.1. While these can be identified sometimes naturally from 0/1 extended formulations (when they exist), it is not always possible (more discussion on this connection is provided in Section 4.1.2). Theorem 3 and Proposition 2 indicate that $MDM(\mathcal{M}\Pi_*)$ is polynomial time solvable, if there exists an efficient routine to identify and formulate a compact (polynomial in size of input) P and $B_i(\cdot)$, given input from Π_*. This brings us to the following corollary.

COROLLARY 4. *Let Π_* be a collection of polytopes (all in V or H representation). Suppose there exists a routine such that for arbitrary input integer n and polytope Q (in n variables) from Π_* (with notation $\mathcal{X}$ denoting the extreme points), we can efficiently identify a compact polytope P and $B_i(\cdot) : \mathcal{X}_i \to 2^B$ described in Theorem 3. Then $MDM(\mathcal{M}\Pi_*)$ is efficient, assuming Assumption 1 on $\mathcal{M} = (\mu_1, \dots, \mu_n)$.*

We next highlight the use of Corollary 4 by briefly noting how three polytope families fit into its framework. The reader is referred to the EC on how the objects P and B_i can be constructed for these positive cases. We then explore some negative cases, as well as show how some common techniques like binary expansions and extended formulations fail to satisfy Corollary 4.

4.1.1. Polytope Families Satisfying Sufficiency in Corollary 4 In this section, we aim to illustrate the scope of the sufficiency condition of Corollary 4. More precisely, we will show that both the class of V-polytopes and 0/1 H-polytopes, for which MDM can be solved efficiently (see Meilijson and Nadas (1979), Lai and Robbins (1976), Haneveld (1986), Weiss (1986)) falls under the purview of Corollary 4. We will also prove that a particular class of H-polytopes (not necessarily 0/1) called $L^\natural$ convex also satisfies Corollary 4.

Corollary 5 **(V- and H-polytopes)**. *If* Π_* *is either a collection of V-polytopes or* 0/1 *H-polytopes, then* Π_* *satisfies the conditions of Corollary 4.*

In fact, in the case of V-polytopes, letting P be the probability simplex suffices. For previous work on bounds with marginals on specific V-polytopes, we refer the reader to Meilijson (1991) and Lai and Robbins (1976). As such, these results can be derived from Corollary 5. Many optimization problems on graphs where the decisions are commonly framed with 0/1 H-polytopes have a tractable MDM, just as discussed in previous distributionally robust work in PERT networks, shortest path, maximum flow and network reliability problems (Meilijson and Nadas (1979), Haneveld (1986), and Weiss (1986)). Owing to the fact that they use 0/1 H-polytopes, their results also follow from Corollary 5. Similar tractability results were provided for combinatorial optimization problems in the settings of Bertsimas et al. (2010, 2004, 2006) in which only the marginal moments are known.

We next examine a special class of integral polytopes that are not necessarily 0/1, called $L^\natural$ convex polytopes, and show that this class satisfies the conditions of Corollary 4. This class of polytopes can arise in dual formulations of network flow problems (Veinott Jr (1989)). We begin with an analytical characterization of $L^\natural$ convex polytopes. Consider a directed graph $(V = \{1,\dots,n\},\mathcal{A})$, an integer-valued distance function $d: V\times V \to \mathbb{Z}\cup\{+\infty\}$, and collections of bounds $\{l_i\}_{i=1}^n \subset \mathbb{Z}\cup\{-\infty\}$,

$\{u_i\}_{i=1}^n \subset \mathbb{Z} \cup \{+\infty\}$. Let $\mathcal{X} = \{x \in \mathbb{Z}^n : x_i - x_j \le d_{ij} \;\; \forall (i,j) \in \mathcal{A},\; l_i \le x_i \le u_i \;\; \forall i \in \{1,\ldots,n\}\}$. A set of this form is known as an $L^\natural$ convex set. It is nonempty if and only if d satisfies the triangle inequality: $d_{ik} \le d_{ij} + d_{jk}$, for all $i,j,k \in V$. Its convex hull is an integral polyhedron (e.g. see Murota (2003)) that when bounded we identify as an $L^\natural$ convex polytope. When assuming $conv(\mathcal{X})$ is bounded, observe that for any i, $\underline{l}_i := \max_{k:(k,i)\in\mathcal{A}} \{l_k - d_{ki}\} > -\infty$ and $\bar{u}_i := \min_{j:(i,j)\in\mathcal{A}} \{u_j + d_{ij}\} < +\infty$, so without loss of generality, let us assume that $\{l_i\}_{i=1}^n \subseteq \mathbb{Z}$, $\{u_i\}_{i=1}^n \subseteq \mathbb{Z}$.

Definition 4.1 ($L^\natural$ convex polytope) *Let $\mathcal{A} \subseteq [n]\times[n]$, $\{l_i\}_{i=1}^n \subseteq \mathbb{Z}$, $\{u_i\}_{i=1}^n \subseteq \mathbb{Z}$, $\{d_{ij}\}_{(i,j)\in\mathcal{A}} \subseteq \mathbb{Z}$. Then $Q = \{x \in \mathbb{R}^n : x_i - x_j \le d_{ij} \;\; \forall (i,j) \in \mathcal{A},\; l_i \le x_i \le u_i \;\; \forall i \in \{1,\ldots,n\}\}$ is an $L^\natural$ convex polytope.*

Note that we may efficiently obtain from any H-representation of a $L^\natural$ convex polytope the $\mathcal{A} \subseteq [n]\times[n]$, $\{l_i\}_{i=1}^n \subseteq \mathbb{Z}$, $\{u_i\}_{i=1}^n \subseteq \mathbb{Z}$, and$\{d_{ij}\}_{(ij)\in\mathcal{A}} \subseteq \mathbb{Z}$. The following corollary states that Π comprised of this type of input will satisfy the sufficiency condition described in Corollary 4.

Corollary 6 (**$L^\natural$ convex polytope**). *If Π_H is a collection of $L^\natural$ convex polytopes in H-representation, then Π_H satisfies the conditions of Corollary 4.*

Indeed, given $\mathcal{A} \subseteq [n]\times[n]$, $\{l_i\}_{i=1}^n \subseteq \mathbb{Z}$, $\{u_i\}_{i=1}^n \subseteq \mathbb{Z}$, and$\{d_{ij}\}_{(ij)\in\mathcal{A}} \subseteq \mathbb{Z}$, we may invoke Corollary 4 by setting $P = \{y \in \mathbb{R}_+^{n\times[l,u]_\mathbb{Z}} : \sum_{x_i \in [l_i,u_i]_\mathbb{Z}} y_{i,x_i} = 1 \;\; \forall i,\; \sum_{s=t}^{u} y_{i,s} + \sum_{s=l}^{t-d_{ij}-1} y_{js} \le 1 \;\; \forall (i,j) \in \mathcal{A}, t = l, l+1, \ldots, u\}$, and letting $B = \{1,\ldots,n\}\times[l,u]_\mathbb{Z}$, $B_i(\xi_i) = \{(i,\xi_i)\}$ for $\xi_i \in \mathcal{X}_i$. We refer the reader to EC.3.3.4 for a proof that follows from Möhring et al. (2001) and the integrality of the stable set polytope.

4.1.2. On 0/1 Extended Formulations In this section, we briefly comment on the use of extended formulations (Conforti et al. (2010)) to invoke Corollary 4 - namely, in constructing a P and B_i. We call a polyhedron $P_{ext} = \{(x,y) : Ax + Bz \le b\}$ an extended formulation for the set $\mathcal{X}$ if $proj_x(P_{ext}) = conv(\mathcal{X}) =: Q$. If additionally the auxilliary variables lie in a 0/1 polytope, then we call the formulation 0/1. Indeed, in the case of the V and $L^\natural$ convex polytope families covered in Section 4.1.1, there exist natural 0/1 extended formulations whose projections onto the y-variables yield the respective 0/1 polytopes P we formulated. But in general the existence of 0/1 extended formulations may not outright provide a certificate of tractability. Particularly, we discuss how just using binary expansions is not sufficient and how parallelotopes that have compact 0/1 extended formulations are still intractable with MDM (Theorem 2).

In previous studies concerned with marginal information, like in Natarajan et al. (2009) and Mak et al. (2015), the technique of binary expansion was considered in the case of $\mathcal{X} \subset \mathbb{Z}^n$. The following example demonstrates that this does not in general provide a 0/1 polytope P of desired form outlined in Theorem 3. Consider the following $L^\natural$ convex polytope $Q = \{x : x_1 - x_2 \leq 1, 0 \leq x \leq 2\}$. Then the polytope $P = \{y \in [0,1]^{\{1,2\}\times\{0,1,2\}} \ |(\sum_j y_{1,j} \cdot j, \quad \sum_j y_{2,j} \cdot j) \in Q\} \supsetneq conv\Big(P \cap \mathbb{Z}^{\{1,2\}\times\{0,1,2\}}\Big)$, as y defined by $y_{1,0} = 1/2,\ y_{1,1} = 0,\ y_{1,2} = 1/2,\ y_{2,0} = 1,\ y_{2,1} = 0$, and $y_{2,2} = 0$, is a member of the former and not the latter. Hence, the binary expansion constraints in P do not lead to the desired form for Theorem 3.

Similarly, the existence of a compact 0/1 extended formulation does not guarantee the efficient construction of a P and $B_i(\cdot)$ for Corollary 4. For example, given parallelotope $Q = \sum_{i=1}^m [-x^i, x^i]$, where $x^1, \ldots, x^m \in \mathbb{R}^n$ are linearly independent vectors, we have $Q = Proj_x\Big(\{(x,y) : x = 2\Big[x^1\ x^2\ \ldots\ x^m\Big]y - \Big[x^1 + x^2 + \ldots + x^m\Big], y \in [0,1]^m\}\Big)$. While $P = [0,1]^m$ is a 0/1 polytope, the sufficiency condition of Corollary 4 cannot hold by Theorem 2.

4.2. Sufficiency via Sublattice Structure

We now discuss a second sufficiency condition on Π_*, motivated naturally by Lemma 2. Lemma 2 states that when $Z(c)$ is supermodular, the monotone coupling presents a worst-case coupling. The significance is that this reduces the multi-dimensional integral to a single dimensional integral calculation. In particular, when the marginals $\mu_1, \ldots, \mu_n$ are all finite-supported, and the support function $Z(c) := \max_{x \in \mathcal{X}} c^\intercal x$ over $\mathcal{X}$ is supermodular, then upon formulating the monotone coupling θ_{mon} of the marginals (poly-sized support), computation of Z^* amounts to simply computing $\mathbb{E}_{\tilde{c} \sim \theta_{mon}}[Z(\tilde{c})]$. This is efficiently computable by Remark 1. So, towards identifying a sufficiency condition on Π_*, we discuss conditions on $\mathcal{X}$ which yield a supermodular support function $Z(c)$. For a broader study on necessary and sufficient conditions for supermodularity in parametric optimization, we refer the interested reader to Chen et al. (2021). The next result follows from the known fact that polyhedral sublattices admit supermodular support functions (see Theorem 8.1, Murota (2003)).

COROLLARY 7. *Let Π_V and Π_H be a collection of V- and H-polytopes respectively.*

- *If every polytope in Π_V has input vertex set $\mathcal{X}$ forming a sublattice, then $MDM(\mathcal{M}\Pi_V)$ is efficient, assuming $\mathcal{M} = (\mu_1, \mu_2, \ldots)$ consists of finite-support marginals.*
- *If every polytope in Π_H forms a sublattice, then $MDM(\mathcal{M}\Pi_H)$ is efficient, assuming $\mathcal{M} = (\mu_1, \mu_2, \ldots)$ consists of finite-support marginals.*

4.2.1. On Identifying Polyhedral Sublattices In Corollary 7, we distinguish the two types of input because a finite set $\mathcal{X}$ forming a sublattice does not in general imply that $conv(\mathcal{X})$ is a sublattice (see Example 4.1). Checking the sublattice structure of $\mathcal{X}$ in the case of V-polytope input can be clearly done efficiently. As for the decision problem of whether a given H-polytope input forms a polyhedral sublattice or not, we refer the reader to Veinott Jr (1989) and Queyranne and Tardella (2006) for an efficient routine, as well as a number of convenient characterizations (e.g., bimonotone linear inequalities, generalized node-arc incidence matrices). Finally, we remark that though sublattice structure is sufficient, it is not necessary for a polytope's support function to be supermodular. The following provides such an example.

Example 4.1 *Consider the polytope in $\mathbb{R}^3$ generated by the set $\mathcal{X}$ comprised of: $(0,0,0)^\intercal, (0,1,1)^\intercal$, and $(1,2,1)^\intercal$. While $\mathcal{X}$ forms a sublattice, $conv(\mathcal{X})$ does not. Yet, the support function is supermodular (shown in EC.3.3.6, using a characterization by Shioura and Tamura (2015)).*

4.3. Comparison of Sufficiency Conditions

Broadly speaking, the two previous sections described two different approaches: (1) identify an efficient reduction of the MDM (dual) separation problem to a compact linear program - Corollary 4; (2) identify when the monotone coupling is optimal - Corollary 7. We conclude this section by comparing and contrasting the two classes of Π_* that are captured by the sufficiency conditions. In doing so, we will compare the value of the dual form with that of the prior work on supermodularity, with respect to MDM's complexity.

We emphasize that Corollary 7 does not subsume Corollary 4. For an illustrative summarization of the results of this section, we refer the reader to Figure 1. In this Venn diagram, we have a circle for each of the two corollaries that form the sufficiency conditions we have discussed. Further, we have added two additional circles for the class of V- and H-polytopes. Each dot labeled "X" signifies a Π_* consisting of polytopes of type "X". As well, we note that while Corollary 4 uses Assumption 1, Corollary 7 makes the stronger assumption of finite support marginals. In terms of similarities, the two conditions intersect at the class of $L^\natural$ convex polytopes. Indeed, $L^\natural$ convex polytopes are closed under $\wedge$ and $\vee$ (Murota (2003)). As well, by Theorem 2, when Π_H consists of all the parallelotopes of all dimensions, it will satisfy neither the sufficiency framework of Corollary 4 nor Corollary 7. Finally, we will see in the next section that the permutahedron presents a Π_H class satisfying Corollary 4 while simultaneously failing to provide supermodular support functions Z. It remains an open question as to whether or not there is a polytope class that satisfies Corollary 7 but not Corollary 4.

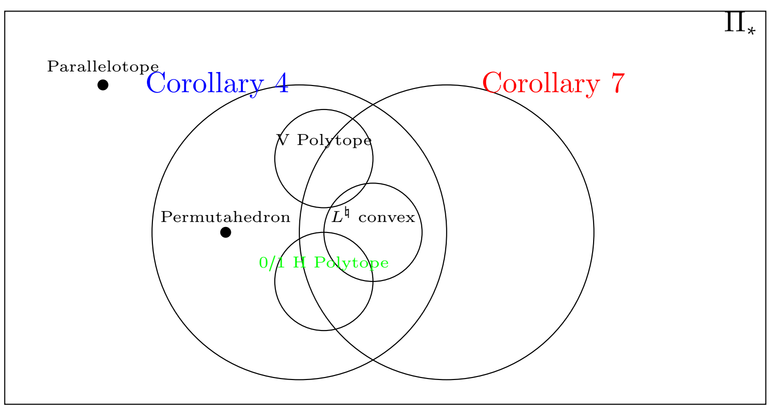


Figure 1 A Venn Diagram illustrating the family of polytopes Π_* that invoke Corollary 7 and/or Corollary 4.

5. Applications of Polynomial Time Solvable Instances

This section focuses on applications that involve polytope classes forming tractable $MDM(\mathcal{M}\Pi)$ problems. In a particularly illustrative application, the appointment scheduling problem can call upon many results of the previous section for its study. We will show in detail how the desired P of

Corollary 4 can be constructed and/or the identification of the desired supermodularity property for Corollary 7 to apply.

We refer the interested reader to EC.2 for other application problems: (1) resource allocation with scheduling; (2) choice bundling; and (3) project scheduling with random, irregular starting time costs. The resource allocation with scheduling problem explores the permutahedron, which poses as an example Π_* that satisfies Corollary 4 but not Corollary 7. In the choice bundling problem, we explore the application of $L^\natural$ convex polytopes (tractable by Corollary 6) in modeling item dependencies/complementarities, and we demonstrate how choice probabilities can be computed via Corollary 1's saddle point conditions. Finally, in the last problem of EC.2, we explore the application of $L^\natural$ convex polytopes (tractable by Corollary 6) in modeling precedence constraints among jobs.

5.1. Appointment Scheduling

The appointment scheduling problem (see Gupta and Denton (2008)) is an interesting application where the results apply. In the simplest version of this problem, a schedule is decided upon upfront with the goal of minimizing the total waiting time incurred by all the patients who see a doctor in a day and any possible overtime of the doctor. While the actual time that each patient spends with the doctor is uncertain at the time of the scheduling, it is common to have partial distributional information on the individual patient processing times, that might be leveraged on to develop an optimal appointment schedule (cf. Kong et al. (2013), Mak et al. (2015)). Consider n patients who arrive in a fixed order $\{1,2,\ldots,n\}$ and need to be scheduled in a given time interval $[0,T]$. We assume that for any patient i, the distribution μ_i of the service time $\tilde{c}_i$ with the doctor is known. The dependence among the distribution of the patients is however unknown. The decision variables are the time interval assigned to each patient i, denoted by s_i. Let $\mathcal{S} = \{s \in \mathbb{R}^n : \sum_{i=1}^n s_i \leq T,\ s_i \geq 0\ \forall i \in [n]\}$.

We are interested in the problem of minimizing the worst-case (over all distributions consistent with the marginals) the expected weighted (positive integer weights d_i) sum of wait times of the n

patients seen by one doctor, plus any overtime weighted by an integer γ:

The problem is formulated as: $Z^* := \min_{s\in\mathcal{S}} \sup_{\theta\in\Gamma(\mu_1,\ldots,\mu_n)} \mathbb{E}_{\tilde{c}\sim\theta}\left[Z(s,\tilde{c})\right]$, where:

$$\begin{aligned} Z(s,\tilde{c}) &= \min \sum_{i=1}^{n} d_i w_i + \gamma w_{n+1} \\ &\text{s.t. } w_1 = 0, \quad w_{i+1} \geq 0,\ \forall i = 1,\ldots,n, \\ &\quad w_{i+1} \geq w_i + \tilde{c}_i - s_i,\ \forall i = 1,\ldots,n, \end{aligned} \quad \text{or} \quad \begin{aligned} Z(s,\tilde{c}) &= \max \sum_{i=1}^{n} (\tilde{c}_i - s_i) x_i \\ &\text{s.t. } x_i - x_{i-1} \geq -d_i, \quad \forall i = 2,\ldots,n, \\ &\quad x_n \leq \gamma, \quad x_i \geq 0,\ \forall i = 1\ldots,n. \end{aligned} \tag{12}$$

5.1.1. Exploiting a 0/1 Formulation in MDM Appointment Scheduling To properly account for the effect of the scheduling decision variables s_i in the inner MDM problem, observe that $\max_{\theta\in\Gamma(\mu_1,\ldots,\mu_n)} \mathbb{E}_{\tilde{c}\sim\theta}[Z(s,\tilde{c})] = \min_{\{\psi_i:\mathcal{X}_i\to\mathbb{R}\}_{i=1}^n} \max_{\nu\in\mathcal{P}(\mathcal{X})} \sum_i \left(\int \psi_i d\Pi_i \nu + \int \psi_i^*(\tilde{c}_i - s_i) d\mu_i \right)$, where $\mu_i - s_i$ denotes the measure on $\mathbb{R}$ induced by the random variable $\tilde{c}_i - s_i$, when $\tilde{c}_i \sim \mu_i$. Define the set $\mathcal{X} := Extr\left(\{x \in \mathbb{R}_+^n : x_i - x_{i-1} \geq -d_i \ \ \forall i = 2,\ldots,n,\ x_n \leq \gamma\}\right)$. As $\mathcal{X}$ turns out to be a bounded $L^\natural$ convex set, Corollary 6 can provide the desired P representation of Corollary 4. Alternatively this set $\mathcal{X}$ has also been well-characterized in previous works (see Zangwill (1966), Zangwill (1969)), providing a different P to invoke Corollary 4 (see Mak et al. (2015) for this in MMM), demonstrating different approaches to solve this problem.

5.1.2. Exploiting Supermodularity in MDM Appointment Scheduling In this section we show that the "worst" possible realization of the patient processing times $(\tilde{c}_1,\ldots,\tilde{c}_n)$ occurs when all these values simultaneously take larger or smaller values. This follows from supermodularity which in turn follows from sublattice structure.

Proposition 3. *For any s, $Z(s,\tilde{c})$ is the support function of a polytope that forms a sublattice; hence, it is supermodular in the $\tilde{c}$ variables. Therefore, the monotone coupling provides a worst-case coupling.*

When the marginals are finite-support, the problem then reduces to solving the compact LP given by $\min_{s\in\mathcal{S}} \mathbb{E}_{\tilde{c}\sim\theta_{mon}}[Z(s,\tilde{c})]$, where θ_{mon} denotes the monotone coupling of the marginals and is independent of the decision s.

Acknowledgments

We would like to thank Professors James B. Orlin and Chung-Piaw Teo for several interesting discussions. The research of the third author was partially supported by the grants MOE2016-T2-2-148, On the Interplay of Choice, Robustness and Optimization in Transportation and MOE2019-T2-2-138, Enhancing Robustness of Networks to Dependence via Optimization.

References

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

Bach, F. 2019. Submodular functions: from discrete to continous domains. *Mathematical Programming* **175** 419–459.

Ben-Tal, A., De Waegenaere, B. Melenberg, G. Rennen. 2013. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science* **59** 341–357.

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

Bertsimas, D., K. Natarajan, C-P. Teo. 2004. Probabilistic combinatorial optimization: Moments, semidefinite programming and asymptotic bounds. *SIAM Journal on Optimization* **15**(1) 185–209.

Bertsimas, D., K. Natarajan, C-P. Teo. 2006. Persistence in discrete optimization under data uncertainty. *Mathematical Programming, Series B* **108**(2-3) 251–274.

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

Birge, J. R., M. J. Maddox. 1995. Bounds on expected project tardiness. *Operations Research* **43**(5) 838–850.

Blanchet, J., Y. Kang, K. Murthy. 2019. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability* **56**(3) 830–857.

Bodlaender, H. L., P. Gritzmann, V. Klee, J. Van Leeuwen. 1990. Computational complexity of norm-maximization. *Combinatorica* **10**(2) 203–225.

Chen, Xin, Daniel Zhuoyu Long, Jin Qi. 2021. Preservation of supermodularity in parametric optimization: Necessary and sufficient conditions on constraint structures. *Operations Research* **69**(1) 1–12.

Conforti, Michele, Gérard Cornuéjols, Giacomo Zambelli. 2010. Extended formulations in combinatorial optimization. *4OR* **8**(1) 1–48.

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

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

Galichon, A. 2016. *Optimal transport methods in economics*. Princeton University Press.

Gao, R., A. J. Kleywegt. 2016. Distributionally robust stochastic optimization with Wasserstein distance.

Goh, J., M. Sim. 2010. Distributionally robust optimization and its tractable approximations. *Operations Research* **58**(4) 902–917.

Gupta, D., B. Denton. 2008. Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions* **40**(9) 800–819.

Haneveld, Willem K. 1986. *Robustness against dependence in PERT: An application of duality and distributions with known marginals*. Springer Berlin Heidelberg.

Khachiyan, L. G. 1980. Polynomial algorithms in linear programming. *USSR Computational Mathematics and Mathematical Physics* **20**(1) 53–72.

Kong, Q., C-Y. Lee, C-P. Teo, Z. Zheng. 2013. Scheduling arrivals to a stochastic service delivery system using copositive cones. *Operations Research* **61**(3) 711–726.

Lai, T. L., H. Robbins. 1976. Maximally dependent random variables. *Proceedings of the National Academy of Sciences of the United States of America* **73**(2) 286–288.

Mak, Y. Rong, J. Zhang. 2015. Appointment scheduling with limited distributional information. *Management Science* **61**(2) 316–334.

Mangasarian, O. L., T. H. Shiau. 1986. A variable-complexity norm maximization problem. *SIAM Journal on Algebraic and Discrete Methods* **7**(3) 455–461.

Meilijson, I. 1991. Sharp bounds on the largest of some linear combinations of random variables with given marginal distributions. *Probability in the Engineering and Informational Sciences* **5** 1–14.

Meilijson, I., A. Nadas. 1979. Convex majorization with an application to the length of critical paths. *Journal of Applied Probability* **16**(3) 671–677.

Mishra, V. K., K. Natarajan, D. Padmanabhan, C-P. Teo, X. Li. 2014. On theoretical and empirical aspects of marginal distribution choice models. *Management Science* **60**(6) 1511–1531.

Möhring, R. H., A. S. Schulz, F. Stork, M. Uetz. 2001. On project scheduling with irregular starting time costs. *Operations Research Letters* **28** 149–154.

Murota, Kazuo. 2003. *Discrete Convex Analysis: Monographs on Discrete Mathematics and Applications 10*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

Nadas, A. 1979. Probabilistic PERT. *IBM Journal of Research and Development* **23**(3) 339–347.

Natarajan, K., M. Song, C-P. Teo. 2009. Persistency model and its applications in choice modeling. *Management Science* **55**(3) 453–469.

Pass, B. 2015. Multi-marginal optimal transport: Theory and applications. *ESAIM: Mathematical Modelling and Numerical Analysis* **49** 1771–1790.

Queyranne, Maurice, Fabio Tardella. 2006. Bimonotone linear inequalities and sublattices of $\mathbb{R}^n$. *Linear Algebra and Its Applications* **413**(1) 100–120.

Rachev, S. T., L. Rüschendorf. 1998. *Mass transportation problems: Volume I: Theory*. Springer.

Rockafellar, R. T. 1997. *Convex analysis*. Princeton University Press.

Shapiro, A., D. Dentcheva, A. Ruszczyński. 2014. *Lectures on stochastic programming: Modeling and theory*. 2nd ed. MOS-SIAM Series on Optimization.

Shioura, A., A. Tamura. 2015. Gross substitutes condition and discrete concavity for multi-unit valuations: a survey. *Journal of the Operations Research Society of Japan* **58**(1) 61–103.

Sion, M. 1958. On general minimax theorems. *Pacific Journal of mathematics* **8**(1) 171–176.

Van Parys, B. P. G, P. J. Goulart, P. Embrechts. 2016. Fréchet inequalities via convex optimization.

Veinott Jr, A. F. 1989. Representation of general and polyhedral subsemilattices and sublattices of product spaces. *Linear Algebra and its Applications* **114** 681–704.

Villani, C. 2003. *Topics in optimal transportation*. American Mathematical Society.

Villani, C. 2009. *Optimal transport: old and new*. Grundlehren der Mathematischen Wissenschaften, Fundamental Principles of Mathematical Sciences, 338, Springer-Verlag, Berlin.

Weiss, G. 1986. Stochastic bounds on distributions of optimal value functions with applications to pert, network flows and reliability. *Operations Research* **34**(4) 595–605.

Wiesemann, W., D. Kuhn, M. Sim. 2014. Distributionally robust convex optimization. *Operations Research* **62**(6) 1358–1376.

Zangwill, W. I. 1966. A deterministic multi-period production scheduling model with backlogging. *Management Science* **13**(1) 105–119.

Zangwill, W. I. 1969. A backlogging model and a multi-echelon model of a dynamic economic lot size production system - a network approach. *Management Science* **15**(9) 506–527.

EC.1. On Infinite $\mathcal{X}$

So that the reader may understand the motivation behind the focus on finite $\mathcal{X}$, as well as get a sense as to why assuming finiteness is without loss of generality, we comment here on cases when $\mathcal{X}$ is infinite: (1) an unbounded polyhedron, or (2) a compact subset of $\mathbb{R}^n$.

EC.1.1. $\mathcal{X}$ is an Unbounded Polyhedron

When $P \subseteq \mathbb{R}^n$ is a polyhedron, $P = conv(Extr(P)) + (0^+)P$, where $Extr(P)$ is the finite set of extreme points and $(0^+)P$ is the recession cone. The following result indicates that the study is nontrivial if and only if the marginal measures and the polyhedron satisfy a joint condition.

PROPOSITION EC.1. *Let $\mu_1, \dots, \mu_n$ be marginal measures on $\mathbb{R}$, P a polyhedron not necessarily bounded, and $C = [(0^+)P]^*$ be the polar cone to the recession cone. Then*

$$Z^* = \sup_{\theta \in \Gamma(\mu_1,\dots,\mu_n)} \mathbb{E}_{\tilde{c}\sim\theta}\left[\sup_{x\in P} \tilde{c}^\top x\right] < +\infty \iff supp(\mu_1) \times supp(\mu_2) \times \dots \times supp(\mu_n) \subset C,$$

where $supp(\mu_i)$ denotes the support of measure μ_i.

We begin with the following observation:

$$\inf_{\theta\in\Gamma(\mu_1,\dots,\mu_n)} \theta(C) = 1 \iff \sup_{\theta\in\Gamma(\mu_1,\dots,\mu_n)} \mathbb{E}[\sup_{x\in\mathcal{X}} \tilde{c}^\top x] < +\infty,$$

which is merely stating the elementary fact that Z^* is finite if and only if all possible marginal couplings $\tilde{c}$ are concentrated in the polar of the recession cone.

$\Leftarrow$: If $supp(\mu_1) \times supp(\mu_2) \times \dots \times supp(\mu_n) \subset C$, then clearly $\inf_{\theta\in\Gamma(\mu_1,\dots,\mu_n)} \theta(C) = 1$.

$\Rightarrow$: Next suppose $Z^* = \sup_{\theta\in\Gamma(\mu_1,\dots,\mu_n)} \mathbb{E}[\sup_{x\in P} \tilde{c}^\top x] < +\infty$, equivalently, that $\inf_{\theta\in\Gamma(\mu_1,\dots,\mu_n)} \theta(C) = 1$. Suppose for the sake of a contradiction that there exists a point $x \in \times_{i=1}^n supp(\mu_i) \cap [\mathbb{R}^2 \setminus C]$. The complement of C, $\mathbb{R}^2 \setminus C$, is open, which means that by the product topology on $\mathbb{R}^n$ there exist open intervals $V_i \ni x_i$ such that $\times_{i=1}^n V_i \subset \mathbb{R}^n \setminus C$. By the definition of support, each V_i is of positive μ_i measure. Therefore, the independent coupling yields $(\mu_1 \otimes \mu_2 \otimes \dots \otimes \mu_n)(\times_{i=1}^n V_i) = \Pi_{i=1}^n \mu_i(V_i) > 0$, meaning $\inf_{\theta\in\Gamma(\mu_1,\dots,\mu_n)} \theta(C) = 1$, a contradiction.

As this result indicates, as long as the input marginal measures and polyhedron satisfy the necessary and sufficient geometrical condition, then Z^* is finite. This is precisely because then any choice of coupling $\tilde{c}$ will present a random linear cost vector such that $\sup_{x\in P}\tilde{c}^\intercal x = \sup_{x\in Extr(P)}\tilde{c}^\intercal x$ with probability one. In other words, assuming the condition, it is enough to study finite $\mathcal{X}$, if only because $Extr(P)$ is finite.

EC.1.2. $\mathcal{X}$ is a Compact set

In Section 2.2, Theorem 1 shows that when $\mathcal{X}$ is finite, a primal-dual formulation holds. In fact mere compactness is already sufficient to establish a minimax theorem (interchange of infsup and supinf). So we state and prove a slightly more general form of Theorem 1 in Section EC.3.1.2.

EC.2. Additional Applications

EC.2.1. Ranking with Applications to a Scheduling Problem

We now discuss an application of this ranking formulation to the problem of allocating resources to jobs to minimize the sum of completion times on a single machine. Assume that we are given a set of n jobs, each with random duration $\tilde{c}_i$ that is processed on a single machine. The objective function of interest is the sum of completion times. Consider the problem of allocating resources to these jobs which reduces the time to do the jobs. However this resource allocation has to be done before knowing the true realization of the job durations or the arrival (priority) sequence of the jobs. The optimization problem is to allocate the resources to minimize the expected sum of completion times allowing for the worst-case joint distribution of job times and a worst-case arrival sequence of jobs. This problem is formulated as: $Z^* := \min_{t\in\mathcal{T}} \sup_{\theta\in\Gamma(\mu_1,\dots,\mu_n)} \mathbb{E}_{\tilde{c}\sim\theta}[Z(t,\tilde{c})]$. Here $Z(t,\tilde{c})$ is the optimal value to the linear optimization problem: $Z(t,\tilde{c}) := \max_{x\in\mathcal{X}_{perm}(n)} \sum_{i=1}^{n} (\tilde{c}_i - t_i)^+ x_i$, where t is the reduction in the individual job times which is assumed to lie in a polyhedral set $\mathcal{T}$, and $\mathcal{X}_{perm}(n) := \{x : [n] \rightarrowtail [n]\}$ is the set of all permutations on $[n] := \{1,2,\dots,n\}$. Equivalently, $\mathcal{X}_{perm}(n)$ can be viewed as the set of extreme points to the permutahedron $P_{perm}(n)$ in $\mathbb{R}^n$ which is defined by the set of inequalities: $\sum_{i\in S} x_i \geq \frac{|S|(|S|+1)}{2}, \forall S \subset [n], S \neq \phi$ and $\sum_{i=1}^{n} x_i = \dfrac{n(n+1)}{2}$.

As a first observation, we note that in MDM for this example, we find a case of submodularity and not supermodularity as desired to invoke Corollary 7.

PROPOSITION EC.2. *For any* $t \in \mathbb{R}^n$, $\max_{x \in \mathcal{X}_{perm}} \sum_{i=1}^n (\tilde{c}_i - t_i) \cdot x_i$ *is monotone, submodular in the* $\tilde{c}$ *variables. Consequently,* $Z(t, \tilde{c})$ *is monotone, submodular in the* $\tilde{c}$ *variables.*

Proof: As the statement concerns any fixed t, for the sake of notation, we absorb $\mathbf{t}$ into $\mathbf{c}$, considering $c_i - t_i$ as a whole by replacing it with c_i. As well, we will write $Z(c)$ instead of $Z(t, c)$. So we will show that $F(c) := \max_{x \in \mathcal{X}_{perm}} \sum_{i=1}^n (c_i) \cdot x_i$ is monotone, submodular, before then going on to establish that $Z(c) := F(c^+)$ is monotone, submodular.

Regarding monotonicity of F, this is clear. As for submodularity on $\mathbb{R}^n$, this is equivalent to the satisfaction of the decreasing differences property. As a result, we need only seek to verify that for any pair of distinct indices $i, j \in \{1, 2, \ldots, n\}$ such that $i < j$, $c_i, c_i', c_j, c_j' \in \mathbb{R}$ with $c_i \le c_i'$ and $c_j \le c_j'$, and for any $\hat{c}_{ij} = (c_1, \ldots, c_{i-1}, c_{i+1} \ldots, c_{j-1}, c_{j+1}, \ldots, c_n) \in \mathbb{R}^{n-2}$, that

$$F_{\hat{c}_{ij}}(c_i, c_j') - F_{\hat{c}_{ij}}(c_i, c_j) \ge F_{\hat{c}_{ij}}(c_i', c_j') - F_{\hat{c}_{ij}}(c_i', c_j), \tag{EC.1}$$

where $F_{\hat{c}_{ij}} : \mathbb{R}^2 \to \mathbb{R}$ is the function F with all arguments other than the $i - th$ and $j - th$ argument held fixed to those values given in $\hat{c}_{ij}$.

We now establish (EC.1) to argue for the submodularity of F. Let $c \in \mathbb{R}^n$ be given, as well as a number $c_j' \ge c_j$, for some fixed $1 \le j \le n$. Without loss of generality, we will assume that $c_1 \le c_2 \le \ldots \le c_n$. Let $r := max\{k : c_k \le c_j'\}$. We will characterize the difference in F caused by the transition of vector c to $\bar{c} := (c_1, c_2, \ldots, c_{j-1}, c_j', c_{j+1}, c_{j+2}, \ldots, c_n)$ and then (in $r - j$ swaps) to the vector with increasing components $(c_1, c_2, \ldots, c_{j-1}, c_{j+1}, c_{j+2}, \ldots, c_r, c_j', c_{r+1}, \ldots, c_n)$. This total difference will reflect the change in F upon swapping out c_j for c_j'.

To facilitate the argument, it will be useful to let σ^0 denote the identity permutation, and then developing the collection $\{\sigma^l\}_{l=1}^{r-j}$ of permutations inductively by $\sigma^{l+1}(k) := \begin{cases} \sigma^l(j+l+1) & k = j+l \\ \sigma^l(j+l) & k = j+l+1 \\ \sigma^l(k) & o.t.w. \end{cases}$

In words, we are making one successive swap after another as we define the next permutation. The net effect of these successive swaps is to carry c'_j from the j-th position to the r-th position in $r-j$ steps. Importantly, $\bar{c}_{\sigma^l(j+l)} = c'_j$, for all l. We observe that

$$\begin{aligned}
F_{\hat{c}_{ij}}(c_i, c'_j) - F_{\hat{c}_{ij}}(c_i, c_j) &= \left[\sum_{k=1}^{n} \bar{c}_{\sigma^0(k)} \cdot k - \sum_{k=1}^{n} c_k \cdot k\right] + \sum_{l=0}^{r-j-1} \left[\sum_{k=1}^{n} \bar{c}_{\sigma^{l+1}(k)} \cdot k - \sum_{k=1}^{n} \bar{c}_{\sigma^l(k)} \cdot k\right] \\
&= (c'_j - c_j) \cdot j + \sum_{l=0}^{r-j-1} \left[\bar{c}_{\sigma^{l+1}(j+l+1)} \cdot (j+l+1) + \bar{c}_{\sigma^{l+1}(j+l)} \cdot (j+l) - \bar{c}_{\sigma^l(j+l+1)} \cdot (j+l+1) - \bar{c}_{\sigma^l(j+l)} \cdot (j+l)\right] \\
&= (c'_j - c_j) \cdot j + \sum_{l=0}^{r-j-1} \left[\bar{c}_{\sigma^l(j+l)} \cdot (j+l+1) + \bar{c}_{\sigma^l(j+l+1)} \cdot (j+l) - \bar{c}_{\sigma^l(j+l+1)} \cdot (j+l+1) - \bar{c}_{\sigma^l(j+l)} \cdot (j+l)\right] \\
&= (c'_j - c_j) \cdot j + \sum_{l=0}^{r-j-1} \left[\bar{c}_{\sigma^l(j+l)} - \bar{c}_{\sigma^l(j+l+1)}\right] \\
&= (c'_j - c_j) \cdot j + \sum_{l=0}^{r-j-1} \left[c'_j - c_{j+l+1}\right] = (c'_j - c_j) \cdot j + \sum_{k=j+1}^{n} max(c'_j - c_k, 0).
\end{aligned}$$

This last expression is nonincreasing as a function of c_i for any $i \neq j$, as desired, establishing the submodularity of F.

Finally, we show that F being submodular in c implies that $Z(c) = F(c^+)$ is submodular in c. This follows from

$$F([x \vee y]^+) + F([x \wedge y]^+) = F(x^+ \vee y^+) + F(x^+ \wedge y^+) \leq F(x^+) + F(y^+) \quad \forall x, y \in \mathbb{R}^n,$$

as $[x \vee y]^+ = x^+ \vee y^+$, $[x \wedge y]^+ = x^+ \wedge y^+$. □

The combination of monotonicity and submodularity is of note because this then means that Z^* can be approximated within $e/(e-1)$ by computing the expectation with respect to the independent coupling. Indeed, this is the subject of the study conducted in Agrawal et al. (2012), which has something to say about the practice in stochastic optimization in which stochastic quantities are assumed to be independently distributed.

But fortunately because the Permutahedron is well-studied and admits an extended formulation (see Conforti et al. (2010)) that informs the construction of a 0/1 polytope P as in Theorem 3, we can do better than approximate. We can derive an exact form for Z^*.

Proposition EC.3. *If $\mu_1, \ldots, \mu_n$ are Borel measures on $\mathbb{R}$ and $\mathcal{X}$ is the set of extreme points to the Permutahedron, then* $\min_{t \in \mathcal{T}} \max_{\theta \in \Gamma(\mu_1,\ldots,\mu_n)} \mathbb{E}_{\tilde{c} \sim \theta}\left[\max_{x \in \mathcal{X}_{perm}(n)} \sum_{i=1}^n (\tilde{c}_i - t_i)^+ x_i \right]$ *can be formulated as*

$$
\begin{aligned}
& min \sum_{j=1}^n \alpha_j + \sum_{i=1}^n \beta_i + \sum_{i=1}^n \mathbb{E}_{\tilde{c}_i \sim \mu_i} [\max_{j \in [n]} \{(\tilde{c}_i - t_i)^+ \cdot j - d_{ij}\}] \\
& s.t.\ t \in \mathcal{T}, \qquad \alpha_j + \beta_i \geq d_{ij}, \quad \forall i, j.
\end{aligned}
\tag{EC.2}
$$

In particular, if $\Pi = (P_1, P_2, \ldots)$ with $P_n = P_{perm}(n)$ then Π satisfies the conditions of Corollary 4.

Proof: Observe that if x is a permutation, then we can define an associated 0/1 permutation matrix $Y \in \{0,1\}^{n \times n}$ via $Y_{i,j} = 1 \iff x(j) = i$, i.e., $x = (1, \ldots, n) \cdot Y$. Let $\mathcal{X}'$ be the set of such 0/1 permutation matrices, which is clearly in bijection with $\mathcal{X}_{perm}$. Next, Birkhoff's theorem tells us that the Birkhoff polytope $\{Y \in [0,1]^{n \times n} : \sum_i Y_{ij} = 1 \quad \forall j, \quad \sum_j Y_{ij} = 1 \quad \forall i\}$ is in fact $conv(\mathcal{X}')$; in other words, $conv(\mathcal{X}')$ can be formulated as a 0/1 polytope in n^2 variables $\{y_{ij}\}_{i,j \in [n]}$. Thus, we set $B = [n] \times [n]$, $B_i(x_i) = \{(i, x_i)\}$ for all $i, x_i \in [n]$, and P as the Birkhoff polytope. This implies that the separation problem (10) to $\max_{\theta \in \Gamma(\mu_1,\ldots,\mu_n)} \mathbb{E}_{\tilde{c} \sim \theta}\left[\max_{x \in \mathcal{X}_{perm}(n)} \sum_{i=1}^n (\tilde{c}_i - t_i)^+ x_i \right]$ can be formulated as a set of primal and dual linear programs:

$$
\begin{aligned}
\max_{\{y_{ij}\}_{i,j=1}^n} \quad & \sum_{i=1}^n \sum_{j=1}^n \psi_i(j) y_{ij} & = \quad & \min_{\{\alpha_j\}_{j=1}^n, \{\beta_i\}_{i=1}^n} \sum_{j=1}^n \alpha_j + \sum_{i=1}^n \beta_i \\
\text{s.t.} \quad & \sum_i y_{ij} = 1, \quad \forall j, & & \text{s.t. } \alpha_j + \beta_i \geq \psi_i(j), \quad \forall i, j, \\
& \sum_j y_{ij} = 1, \quad \forall i, & & \\
& y_{ij} \geq 0, \quad \forall i, j. & &
\end{aligned}
$$

Combining this with Theorem 1, $\min_{t \in \mathcal{T}} \max_{\theta \in \Gamma(\mu_1,\ldots,\mu_n)} \mathbb{E}_{\tilde{c} \sim \theta}\left[\max_{x \in \mathcal{X}_{perm}(n)} \sum_{i=1}^n (\tilde{c}_i - t_i)^+ x_i \right]$ can be formulated as:

$$
\min_{t \in \mathcal{T}, d_{ij}} \max_{\sum_i y_{ij}=1, \sum_j y_{ij}=1, y \geq 0} \sum_{i=1}^n \sum_{j=1}^n d_{ij} y_{ij} + \sum_{i=1}^n E_{\tilde{c}_i \sim \mu_i} [\max_{j \in [n]} \{(\tilde{c}_i - t_i)^+ \cdot j - d_{ij}\}],
$$

with which we arrive at the form

$$
\begin{aligned}
\min \ & \sum_{j=1}^n \alpha_j + \sum_{i=1}^n \beta_i + \sum_{i=1}^n E_{\tilde{c}_i \sim \mu_i} [\max_{j \in [n]} \{(\tilde{c}_i - t_i)^+ \cdot j - d_{ij}\}] \\
\text{s.t. } & t \in \mathcal{T}, \\
& \alpha_j + \beta_i \geq d_{ij}, \quad \forall i, j.
\end{aligned}
$$

□

To conclude this example, we emphasize that according to Proposition EC.2 when Π_H consists of Permutahedra, it does not satisfy Corollary 7. Hence, demonstrating that Corollary 4 and 7 do indeed capture different classes of polytopes.

EC.2.2. Choice of Bundle

In Section 5.1, when the goal was to obtain a scheduling decision, neither the MDM's primal nor dual variables were of particular interest. Rather, it was enough to be able to characterize the MDM quantity that formed the inner problem. And we did this in two ways - one using the MDM dual form and one using supermodularity. But in this section, we consider a problem in which we will actually be interested in the primal and dual variables to MDM.

In prior works Natarajan et al. (2009) and Mishra et al. (2014), the MDM choice model took the form

$$\sup_{\theta\in\Theta(F_0,F_1,\ldots,F_n)} \mathbb{E}_{\tilde{\epsilon}\sim\theta}\left[\max_{x\in\mathcal{X}} \sum_{j=1}^{n} (v_j - p_j + \tilde{\epsilon}_j)x_j\right], \tag{EC.3}$$

where v_j, p_j, and ϵ_j denote, respectively, the nominal utility, price, and random shock of alternative j. And $\mathcal{X}$ consists of the possible decisions $x \in \{0,1\}^n$, which encode choices. In Natarajan et al (2009), $\mathcal{X} = \left\{x \in \{0,1\}^n \,\middle|\, \sum_{j=1}^{n} x_j \le 1\right\}$, for which it is shown that any primal solution's marginals $\{\pi_i\nu\}_{i=1}^n$ represents the choice probabilities by the consumer. But this $\mathcal{X}$ models a consumer that is only interested in at most one item.

Towards modeling a consumer that may choose more than one item in a setting where items may exhibit dependencies or complementarity, we can take inspiration from the $L^\natural$ convex polytope for $\mathcal{X}$. Specifically, consider $\mathcal{X} = Extr(\{x \,|\, 0 \le x_j \le 1, x_i - x_j \le 0, \forall (i,j) \in \mathcal{A}\})$, where $\mathcal{A}$ denotes the arc set in a dependence graph, which captures the dependence among items. For instance, if $(i,j) \in \mathcal{A}$, item j must be chosen if item i is selected. According to the tractability conditions in Section 4, under Assumption 1, (EC.3) can model the choice probability of bundles and it is tractable. From Theorem 3, we can reformulate $\max\limits_{\nu\in\mathcal{P}(\mathcal{X})} \sum_{i=1}^n \sum_{x_i\in\mathcal{X}_i} \psi_i(x_i)\Pi_i\nu(x_i)$ as

$$\max_{y\in P} \sum_{i=1}^n \psi_{i1}y_{i1} + \psi_{i0}y_{i0} \tag{EC.4}$$

where $P = \left\{ y \,\middle|\, \begin{array}{l} y_{i1} + y_{i0} = 1, \forall i = 1, \ldots, n \\ y_{i_1} + y_{j0} \le 1, \forall (i,j) \in \mathcal{A}. \end{array} \right\}$ Therefore, (EC.3) can be reformulated as

$$\min_{\{\psi_{ij}\}} \max_{y \in P} \sum_{i=1}^{n} \psi_{i1} y_{i1} + \psi_{i0} y_{i0} + \sum_{i=1}^{n} \mathbb{E}\left[\max\{\tilde{c}_i - \psi_{i1}, -\psi_{i0}\}\right] \tag{EC.5}$$

Note that the second term in (EC.5) can be formulated as

$$(\psi_{i1} - \psi_{i0}) F_i(\psi_{i1} - \psi_{i0}) - \psi_{i1} + \int_{\psi_{i1} - \psi_{i0}} c_i dF_i(c_i) \tag{EC.6}$$

Hence (EC.5) can be reformulated as

$$\min_{\{\psi_{ij}\}} \max_{y \in P} \sum_{i=1}^{n} -(\psi_{i1} - \psi_{i0}) y_{i0} + (\psi_{i1} - \psi_{i0}) F_i(\psi_{i1} - \psi_{i0}) + \int_{\psi_{i1} - \psi_{i0}}^{\infty} c_i dF_i(c_i) \tag{EC.7}$$

By replacing $\psi_{i1} - \psi_{i0}$ as η_i and taking dual of the inner maximization problem, we can reduce (EC.7) to the following convex optimization problem in η_i.

$$\begin{aligned} \min_{\{\eta_i\}_i, \{\lambda_i\}_i, \{\mu_{ij}\}_{(i,j) \in \mathcal{A}}} \quad & \sum_{i=1}^{n} \lambda_i + \sum_{(i,j) \in \mathcal{A}} \mu_{ij} + \sum_{i=1}^{n} \eta_{i1} F_i(\eta_i) + \int_{\eta_i}^{\infty} c_i dF_i(c_i) \\ s.t. \quad & \lambda_i + \sum_{j:(i,j) \in \mathcal{A}} \mu_{ij} \ge 0, \forall i = 1, \ldots, n \\ & \lambda_j + \sum_{i:(i,j) \in \mathcal{A}} \mu_{ij} \ge -\eta_j, \forall j = 1, \ldots, n \end{aligned} \tag{EC.8}$$

The dual formula provides a tractable formula for (EC.3) but loses the information of the persistence value, which is $\Pi_i \nu$ under the absolutely continuity assumption. Fortunately, according to the "Saddle-point" conditions in Corollary 3, we in fact can obtain the persistence value $\Pi_i \nu(x_i) = \mu_i \partial \eta_i(x_i)$. Specifically, $\Pi_i \nu(1) = \mu_i \partial \eta_i(1) = \mu_i([\eta_i, \infty)) = 1 - F_i(\eta_i)$. With the $\{\Pi_i \nu\}$ obtainable in this way, we have the likelihood that any alternative i will be selected amongst the bundle that the consumer selects.

EC.2.3. Bounds for Project Scheduling with Random, Irregular Starting Time Costs

In this section, we discuss an extension of the polynomial time complexity results in Möhring et al. (2001) for project scheduling problems with irregular starting time costs to the case where randomness is incorporated in the cost function. Consider a set of jobs denoted by $N = \{1, \ldots, n\}$

with a fixed time horizon $\bar{T} = \{0, 1, \ldots, T\}$ in which all job starting times need to be scheduled. A job $j \in N$ is assumed to incur a random cost $\tilde{c}_j(t)$ if it is started at time t. Let S_j denote the start time of the job j. For example, the random cost might be defined as $\tilde{c}_j(S_j) = c_j^0(S_j)\tilde{\epsilon}_j$ where $c_j^0(S_j)$ is a deterministic cost function of the start time and $\tilde{\epsilon}_j$ is a random cost term for job j. The precedence constraints among two jobs i and j are denoted by the constraints $S_j \geq S_i - d_{ij}$ where $d_{ij} \in (-\infty, \infty)$ is an integer number imposing a time lag between the jobs. Assume the integer processing time of each job j is denoted by p_j. Then this can capture an ordinary precedence constraint that job j is started only after job i by incorporating the constraint $S_j \geq S_i + p_i$. The precedence among the jobs is denoted by the digraph $G = (N, A)$ where $A = \{(i, j) | d_{ij} > -\infty\}$. We assume that there is no directed cycle of positive length in the graph, to prevent conflicts in scheduling. Assume that $\Gamma(\mu_1, \ldots, \mu_n)$ is the set of distributions for the random terms $\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_n$. The optimization problem is to solve:

$$\inf_{\theta \in \Gamma(\mu_1, \ldots, \mu_n)} \mathbb{E}\left[\min_{S \in \mathcal{S}} \sum_{j=1}^{n} \tilde{c}_j(S_j)\right], \tag{EC.9}$$

where $\mathcal{S}$ is the set of feasible starting time vectors. We remark that because the c_j^0 are arbitrary functions, the monotone coupling can indeed be suboptimal here.

Remark EC.1. Consider

$$\begin{aligned} \min_{S_1, S_2} \quad & \frac{\epsilon_1}{\sqrt{S_1}} + \epsilon_2 S_2^{1/3} \\ \text{s.t.} \quad & 1 \leq S_1 \leq 3 \\ & 1 \leq S_2 \leq 2 \\ & S_1 \leq S_2, \end{aligned}$$

where the deterministic cost functions are defined by

$$c_1^0(S_1) := \begin{cases} \frac{1}{\sqrt{S_1}} & S_1 \in [1, 3] \\ +\infty & otw \end{cases} \quad \text{and} \quad c_2^0(S_2) := \begin{cases} S_2^{1/3} & S_2 \in [1, 2] \\ +\infty & otw \end{cases}$$

Then it is easy to verify that the monotone coupling yields an expected value of ≈ 0.980, while the anti-monotone yields a smaller expected value at ≈ 0.853. $\triangle$

With the above remark suggesting that there is no standard analytical characterization of the optimal coupling, we proceed with using the dual formulation of Theorem 1 and the extended formulation technique we illustrated in Section 4.

For a feasible schedule F, let x^F be defined as

$$x_{jt}^F := \begin{cases} 1, \text{if job j is started in period t}, \\ 0, \text{otherwise}. \end{cases}$$

Let $\mathcal{X} := \{x^F : F \text{ is a feasible schedule}\}$ be the extreme points to the "time-indexed polytope" $Q = conv\{x^F : F \text{ is a feasible schedule}\} = conv(\mathcal{X})$. The start time for a job j is then defined as $S_j = \sum_{t=0}^T t x_{jt}$. Then the inner scheduling problem of minimum cost in this case can be formulated as an integer program:

$$\begin{aligned} \min \quad & \sum_{j=1}^n \tilde{\epsilon}_j c_j^0 \\ \text{s.t.} \quad & c_j^0 = \textstyle\sum_{t=0}^T c_j^0(t) x_{jt} && \forall j = 1, \dots, n \\ & \sum_{t=0}^T x_{jt} = 1, && \forall j = 1, \dots, n, \\ & \sum_{s=t}^T x_{is} + \sum_{s=0}^{t+d_{ij}-1} x_{js} \le 1, && \forall (i,j) \in A, t = 0, \dots, T, \\ & x_{jt} \in \{0,1\}, && \forall j = 1, \dots, n, t = 0, \dots, T. \end{aligned}$$

The linear programming relaxation of this integer program provides integral x solutions (see the proof of Corollary 6). In other words, $conv(\mathcal{X})$ is a 0/1 polytope; hence, by Corollary 5, $MDM(\mathcal{M}\Pi)$ is efficient, assuming Assumption 1 on the marginals and Π is comprised of the 0/1 time-indexed polytopes.

EC.2.3.1. Tractable Reformulation for Project Scheduling with Random, Irregular Starting Time Costs To illustrate the tractable computational form for the project scheduling problem:

$$\begin{aligned} & \inf_{\theta \in \Gamma(\mu_1, \dots, \mu_n)} \mathbb{E}\left[\min_{S \in \mathcal{S}} \sum_{j=1}^n \tilde{c}_j(S_j)\right] \\ = & -\sup_{\theta \in \Gamma(\mu_1, \dots, \mu_n)} \mathbb{E}\left[-\min_{S \in \mathcal{S}} \sum_{j=1}^n \tilde{c}_j(S_j)\right] \end{aligned}$$

$$= -\sup_{\theta\in\Gamma(\mu_1,\dots,\mu_n)} \mathbb{E}\left[\max_{S\in\mathcal{S}} \sum_{j=1}^{n} -\tilde{c}_j(S_j)\right]$$
$$= -\min_{\psi_{jt}} \max_{x\in conv(\mathcal{X})} \sum_{j=1}^{n}\sum_{t=0}^{T} \psi_{jt}x_{jt} + \sum_{j=1}^{n} \mathbb{E}_{\tilde{\epsilon}_j\sim\mu_j}[\max_{\hat{x}_j\in\mathcal{X}_j}\{-\tilde{\epsilon}_j\cdot\sum_{t=0}^{T} c_j^0(t)\hat{x}_{jt} - \sum_{t=0}^{T}\psi_{jt}\hat{x}_{jt}\}]$$
$$= -\min_{(\psi_{jt},\lambda_j,\gamma_{ij}^t)\in Q} \sum_{j=1}^{n}\lambda_j + \sum_{(i,j)\in A, t=0:T} \gamma_{ij}^t + \sum_{j=1}^{n} \mathbb{E}_{\tilde{\epsilon}_j\sim\mu_j}[\max_{\hat{x}_j\in\mathcal{X}_j}\{-\tilde{\epsilon}_j\cdot\sum_{t=0}^{T} c_j^0(t)\hat{x}_{jt} - \sum_{t=0}^{T}\psi_{jt}\hat{x}_{jt}\}],$$

where

$$Q := \{\psi,\lambda,\gamma:\ \lambda_j + \sum_{(i,t):(j,i)\in A, t=0,\dots s} \gamma_{ji}^t + \sum_{(i,t):(i,j)\in A, t\in[(s-d_{ij}+1)\vee 0, T]_{\mathbb{Z}}} \gamma_{ij}^t \ge \psi_{js}, \forall j\in\{1,\dots,n\}, \forall s\in\{0,\dots,T\},$$
$$\gamma_{ij}^t \ge 0, \forall (i,j)\in A, t\in\{0,\dots,T\}\},$$

and for each j, the integrand term

$\max_{\hat{x}_j\in\mathcal{X}_j}\{-\tilde{\epsilon}_j\cdot\sum_{t=0}^{T} c_j^0(t)\hat{x}_{jt} - \sum_{t=0}^{T}\psi_{jt}\hat{x}_{jt}\} =$

$$\begin{aligned}
\max\quad & \sum_{t=0}^{T}(-\tilde{\epsilon}_j c_j^0(t) - \psi_{jt})\hat{x}_{jt} \\
\text{s.t.}\quad & \sum_{t=0}^{T}\hat{x}_{jt} = 1, && \forall j = 1,\dots,n,\\
& \sum_{s=t}^{T}\hat{x}_{is} + \sum_{s=0}^{t+d_{ij}-1}\hat{x}_{js} \le 1, && \forall (i,j)\in A, t = 0,\dots,T,\\
& \hat{x}_{jt} \ge 0, && \forall j = 1,\dots,n, t = 0,\dots,T.
\end{aligned}$$

has an equivalent dual minimization form:

$$\begin{aligned}
\min_{\{\alpha_{j'}\}_{j'=1}^n, \{\beta_{ij'}^t\}_{(i,j')\in A, t=0:T}}\quad & \sum_{j'=1}^{n}\alpha_{j'} + \sum_{(i,j',t):(i,j')\in A, t\in\{0,\dots,T\}} \beta_{ij'}^t \\
\text{s.t.}\quad & \alpha_{j'} + \sum_{(i,j',t):(j',i)\in A, t\in\{0,\dots,s\}} \beta_{j'i}^t + \sum_{(i,j',t):(i,j')\in A, t\in[(s-d_{ij'}+1)\vee 0, T]_{\mathbb{Z}}} \beta_{ij'}^t \ge 0,\ \forall j'\ne j, s\in\{0,\dots,T\}\\
& \alpha_j + \sum_{(i,t):(j,i)\in A, t\in\{0,\dots,s\}} \beta_{ji}^t + \sum_{(i,t):(i,j)\in A, t\in[(s-d_{ij}+1)\vee 0, T]_{\mathbb{Z}}} \beta_{ij}^t \ge -\tilde{\epsilon}_j c_j^0(s) - \psi_{js},\ \forall s\in\{0,\dots,T\}\\
& \beta_{ij'}^t \ge 0\ \forall (i,j')\in A, t\in\{0,\dots,T\}
\end{aligned}$$

□

EC.3. Proofs and Examples

EC.3.1. Section 2 Proofs and Examples

EC.3.1.1. Example: Monotone Coupling Not Always Optimal In the absence of supermodularity, the monotone coupling may not be an extremal coupling. Consider letting $\mathcal{X}$ =

$\{(-1,0,0),(0,-1,0),(0,0,-1)\}$ and $\mu_1,\mu_2,\mu_3 = 1/3\cdot\mathbb{1}_1 + 2/3\cdot\mathbb{1}_0$. Then the optimal coupling is found in: $(\tilde{c}_1,\tilde{c}_2,\tilde{c}_3) = (1,0,0)$, $(0,1,0)$ and $(0,0,1)$ each with probability 1/3 with expected value of 0. By contrast, the monotone coupling yields an expected value of $-1/3$, indicating its suboptimality.

EC.3.1.2. Proof of Theorem 1

Theorem 1: (Primal-Dual Formulation of MDM) *Let $\mathcal{X}\subset\mathbb{R}^n$ be compact, $Z(c) := \max_{x\in\mathcal{X}} c^\top x$. Let there be given n Borel probability measures $\{\mu_i\}_{i=1}^n$ over $\mathbb{R}$. Then*

$$\max_{\theta\in\Gamma(\mu_1,\ldots,\mu_n)} \mathbb{E}_{\tilde{c}\sim\theta}\left[Z(\tilde{c})\right] = \max_{\nu\in\mathcal{P}(\mathcal{X})} \sum_i \max_{\gamma_i\in\Gamma(\mu_i,\Pi_i\nu)} \mathbb{E}_{(\tilde{c}_i,\tilde{x}_i)\sim\gamma_i}\left[\tilde{c}_i\tilde{x}_i\right]. \tag{†}$$

If, in addition, $\mathbb{E}_{\tilde{c}_i\sim\mu_i}\left[|\tilde{c}_i|\right]<\infty$ for all i, then

$$= \max_{\nu\in\mathcal{P}(\mathcal{X})} \min_{\{\phi_i\in\mathcal{L}^1(\mathbb{R},\mu_i)\}_{i=1}^n} \sum_{i=1}^n \int \phi_i^*(x_i)\ d\pi_i\nu(x_i) + \int \phi_i(c_i)\ d\mu_i(c_i)$$

$$= \inf_{\{\phi_i\in\mathcal{L}^1(\mathbb{R},\mu_i)\}_{i=1}^n} \max_{\nu\in\mathcal{P}(\mathcal{X})} \sum_{i=1}^n \int \phi_i^*(x_i)\ d\pi_i\nu(x_i) + \int \phi_i(c_i)\ d\mu_i(c_i).$$

And if $\mathcal{X}$ is finite, then

$$= \max_{p\in\mathbb{R}_+^{\mathcal{X}},\sum_{x\in\mathcal{X}}p_x=1} \min_{\{\psi_i:\mathcal{X}_i\to\mathbb{R}\}_{i=1}^n} \sum_{x\in\mathcal{X}} p_x\Big(\sum_{i=1}^n \psi_i(x_i)\Big) + \sum_{i=1}^n \mathbb{E}_{\tilde{c}_i\sim\mu_i}\left[\max_{x_i\in\mathcal{X}_i}\tilde{c}_i x_i - \psi_i(x_i)\right] \tag{Primal}$$

$$= \min_{\{\psi_i:\mathcal{X}_i\to\mathbb{R}\}_{i=1}^n} \max_{p\in\mathbb{R}_+^{\mathcal{X}},\sum_{x\in\mathcal{X}}p_x=1} \sum_{x\in\mathcal{X}} p_x\Big(\sum_{i=1}^n \psi_i(x_i)\Big) + \sum_{i=1}^n \mathbb{E}_{\tilde{c}_i\sim\mu_i}\left[\max_{x_i\in\mathcal{X}_i}\tilde{c}_i x_i - \psi_i(x_i)\right], \tag{Dual}$$

where, in particular, the infimum in (Dual) is attained ("inf" to "min"). Finally, in all the infimum (/or minimum) statements, the ψ_i and ϕ_i may be imposed to be convex (as defined in Section 2.1.1).

Proof:

$$\sup_{\theta\in\Gamma(\mu_1,\ldots,\mu_n)} \mathbb{E}_{\tilde{c}\sim\theta}\left[Z(\tilde{c})\right] = \max_{\theta\in\Gamma} \mathbb{E}_{\tilde{c}\sim\theta}\left[\max_{x\in\mathcal{X}}\tilde{c}^\top x\right] = \max_{(\tilde{c},\tilde{x}):\tilde{c}\sim\theta\in\Gamma,\tilde{x}\sim\nu\in\mathcal{P}(\mathcal{X})} \mathbb{E}_{\tilde{c}\sim\theta}\left[\tilde{c}^\top\tilde{x}\right] \tag{2}$$

$$= \max_{\{(\tilde{c}_i,\tilde{x}_i)\}_{i=1}^n:\tilde{c}_i\sim\mu_i,\tilde{x}\sim\nu\in\mathcal{P}(\mathcal{X})} \sum_{i=1}^n \mathbb{E}_{(\tilde{c}_i,\tilde{x}_i)}\left[\tilde{c}_i\tilde{x}_i\right] \tag{3}$$

$$= \max_{\nu\in\mathcal{P}(\mathcal{X})} \sum_{i=1}^n \max_{\gamma_i\in\Gamma(\mu_i,\pi_i\nu)} \mathbb{E}_{(\tilde{c}_i,\tilde{x}_i)\sim\gamma_i}\left[\tilde{c}_i\tilde{x}_i\right] \tag{†}$$

$$= \max_{\nu\in\mathcal{P}(\mathcal{X})} \sum_{i=1}^n \min_{\phi_i\in\mathcal{L}^1(\mathbb{R},\mu_i)} \int \sup_{c_i\in\mathbb{R}}\{c_i x_i - \phi_i(c_i)\}\ d\pi_i\nu + \int \phi_i\ d\mu_i \tag{4}$$

$$= \inf_{\{\phi_i\in\mathcal{L}^1(\mathbb{R},\mu_i)\}_{i=1}^n} \max_{\nu\in\mathcal{P}(\mathcal{X})} \sum_{i=1}^n \int \sup_{c_i\in\mathbb{R}}\{c_i x_i - \phi_i(c_i)\}\ d\pi_i\nu + \int \phi_i\ d\mu_i, \tag{5}$$

The attainment of the supremum in the left hand side of (2) is due to the weak-* compactness of $\Gamma(\mu_1,\dots,\mu_n)$, along with Z being a bounded, continuous function. Equation (4) follows from Lemma 1 and the integrability assumption (note $-|c_i|\cdot|\max_{\xi\in\mathcal{X}}\xi_i|\le c_ix_i\le|c_i|\cdot|\max_{\xi\in\mathcal{X}}\xi_i|$). Finally, Sion's Minimax Theorem (Corollary 3.3 of Sion (1958)) grants equation (5). When $\mathcal{X}$ is finite, the infimum in (Dual) is attained, because $L(\cdot,\cdot)$ is "upper closed concave-convex" and hence is necessarily the Lagrangian of a convex program (see Theorem 36.5 in Rockafellar (1997)), whose dual program has a feasible solution in the relative interior (i.e., Slater's Condition / dual "strong consistency") (Corollary 30.5.2 in Rockafellar (1997)). □

EC.3.1.3. Proof of Corollary 1

Corollary 1 ("Saddle-point" conditions) *Suppose $\mathcal{X}\subset\mathbb{R}^n$ is finite and let $(\bar\nu,\{\bar\psi_i\}_i)$ be a pair of optimal primal and dual solutions, with $\bar\psi_i$ convex on $\mathcal{X}_i$. Then:*

1. $\Pi_i\bar\nu(x_i)\in\left[\mu_i\left(\partial\bar\psi_i(x_i)\setminus\cup_{x_i'\neq x_i}\partial\bar\psi_i(x_i')\right),\mu_i\left(\partial\bar\psi_i(x_i)\right)\right]\quad\forall i\in[n],\forall x_i\in\mathcal{X}_i$
2. $\int_{\mathcal{X}}\sum_i\bar\psi_i(x_i)d\nu\le\int_{\mathcal{X}}\sum_i\bar\psi_i(x_i)d\bar\nu,\quad\forall\nu\in\mathcal{P}(\mathcal{X})$.

If the μ_i are absolutely continuous, then the two conditions are sufficient for $\bar\nu$ and $\{\bar\psi_i\}_i$ to be primal and dual optimal; in addition, 1. becomes equivalent to:

$$\Pi_i\bar\nu(x_i)=\mu_i\partial\bar\psi_i(x_i),\quad\forall i\in[n],\forall x_i\in\mathcal{X}_i$$

Proof: Observe that $L(\{\bar\psi_i\}_{i=1}^n,\bar\nu)\le L(\{\psi_i\}_{i=1}^n,\bar\nu)$ if and only if

$$\sum_{x_i\in\mathcal{X}_i}\psi_i(x_i)\cdot\Pi_i\bar\nu(x_i)+\int\max_{x_i\in\mathcal{X}_i}\{\tilde c_i-\psi_i(x_i)\}d\mu_i$$

is minimized in ψ_i at $\bar\psi_i$. For notation's sake, let $F_{\tilde c_i}(\psi_i):=\max_{x_i\in\mathcal{X}_i}\{\tilde c_i-\psi_i(x_i)\}$, and $f(\psi_i):=\int\max_{x_i\in\mathcal{X}_i}\{\tilde c_i-\psi_i(x_i)\}d\mu_i$. Then, letting $\mathbb{1}_{x_i}$ denote the unit vector in $\mathbb{R}^{\mathcal{X}_i}$ with 1 at x_i,

$$\partial F_{\tilde c_i}(\psi_i)=conv(-\mathbb{1}_{x_i}:\tilde c_i\in\partial\psi_i(x_i))\subset\mathbb{R}^{\mathcal{X}_i}$$

so that by Theorem 7.47 of Shapiro et al. (2014) yields

$$\partial f(\psi_i)=\int\partial F_{\tilde c_i}(\psi_i)d\mu_i$$

implying

$$Proj_{x_i}\partial f(\psi_i) = [-1,0]\cdot \mu_i\left(\partial\bar{\psi}_i(x_i)\cap\left[\cup_{x_i'\neq x_i}\partial\bar{\psi}_i(x_i')\right]\right) - \mu_i\left(\partial\bar{\psi}_i(x_i)\setminus\cup_{x_i'\neq x_i}\partial\bar{\psi}_i(x_i')\right).$$

Then the optimality of $\bar{\psi}_i$ being equivalent to

$$0\in\Pi_i\bar{\nu}+\partial f(\bar{\psi}_i)$$

implies that for any $x_i\in\mathcal{X}_i$, $\Pi_i\bar{\nu}(x_i)\in -Proj_{x_i}\partial f(\psi_i)$, as desired. Condition 2 is equivalent to $L(\{\bar{\psi}_i\}_{i=1}^n,\bar{\nu})\geq L(\{\bar{\psi}_i\}_{i=1}^n,\nu)$. □

EC.3.1.4. Example: Nonunique θ^* with Absolutely Continuous Marginals Let $\mathcal{X}=\{(-1,1),(1,1),(-1.-1),(1,-1)\}$, $\tilde{c}_1\sim Unif[-1,1]$, $\tilde{c}_2\sim Unif[-1,1]$. Then on the probability space $((-1,1),\mathcal{B},\lambda)$, with λ denoting Lebesgue measure, we can define two mappings

$$\left(c_1(\omega),c_2(\omega)\right)=(\omega,-\omega)$$

and

$$\left(c_1(\omega),c_2(\omega)\right)=(\omega,\omega).$$

It is easy to see that both mappings induce measures on $\mathbb{R}^2$ that are consistent with the given marginals. As well, they both solve (1). In the first coupling, x^{OPT} is the northwest and southeast points with equal probability. In the second coupling, x^{OPT} is the northeast and southwest points with equal probability. But the marginal probability laws to both x^{OPT} are indeed equal, as guaranteed by Corollary 3.

EC.3.1.5. Proof of Corollary 3

COROLLARY 3 (MARGINAL UNIQUENESS IN PRIMAL OPTIMALITY) *Suppose $\mathcal{X}$ is finite and that the Borel probability measures $\mu_1,\ldots,\mu_n$ are all absolutely continuous. If $\bar{\nu}$ and $\bar{\tau}$ are both primal optimal solutions, then $\Pi_i\bar{\nu}=\Pi_i\bar{\tau}$ for all i.*

Proof: Let $\{\bar{\psi}_i\}_i$ be a dual optimal solution. Then, with $(\bar{\nu},\{\bar{\psi}_i\}_i)$ and $(\bar{\tau},\{\bar{\psi}_i\}_i)$ both being pairs of primal-dual optimal solutions, Corollary 3 follows from Corollary 1. □

EC.3.1.6. Proof of Proposition 1

PROPOSITION 1. *Suppose $\mathcal{X} \subset \mathbb{R}^n$ is finite and that the Borel probability measures $\mu_1, \ldots, \mu_n$ are all absolutely continuous. Then there exists a measurable function $x^* : \mathbb{R}^n \to \mathcal{X}$ that takes the form $x^*(c) = (x_1^*(c_1), \ldots, x_n^*(c_n))$ and satisfies $x^*(c) \in x^{OPT}(c)$ for all $c \in \times_{i=1}^n supp(\mu_i)$, such that for any primal optimal solution $\bar{\nu} \in \mathcal{P}(\mathcal{X})$,*

$$P_{\tilde{c} \sim \theta^*}(x_i^*(\tilde{c}) = \xi_i) = P_{\tilde{c}_i \sim \mu_i}(x_i^*(\tilde{c}_i) = \xi_i) = \Pi_i \bar{\nu}(\xi_i), \;\; \forall \xi_i \in \mathcal{X}_i, \;\; \forall i \in [n], \tag{8}$$

where $\theta^* \in \Gamma(\mu_1, \ldots, \mu_n)$ is any optimal solution to MDM (1).

Proof: Let ν and γ_i be optimal to the problem

$$\max_{\nu \in \mathcal{P}(\mathcal{X})} \sum_i \max_{\gamma_i \in \Gamma(\mu_i, \Pi_i \nu)} \mathbb{E}_{(\tilde{c}_i, \tilde{x}_i) \sim \gamma_i}[\tilde{c}_i \tilde{x}_i]. \tag{†}$$

Then let $\tilde{x} \sim \nu$ be a random vector. Next, with each component random variable $\tilde{x}_i$, and using the fact that $\gamma_i \in \Gamma(\mu_i, \Pi_i \nu)$, we form the random vectors $(\tilde{c}_i, \tilde{x}_i) \sim \gamma_i$. Each of these yields a conditional distribution $\tilde{c}_i \|_{\tilde{x_i}}$. Finally, we couple these independently (any coupling works in fact) to form the conditional random vector $\tilde{c}\|_{\tilde{x}} := \tilde{c}_1 \|_{\tilde{x_1}} \otimes \tilde{c}_2 \|_{\tilde{x_2}} \otimes \ldots \otimes \tilde{c}_n \|_{\tilde{x_n}}$. With $\tilde{x}$ and $\tilde{c}\|_{\tilde{x}}$, we have a random vector $\tilde{c}$, whose probability law we will denote with θ^*. It is clear that $\theta^* \in \Gamma(\mu_1, \ldots, \mu_n)$. Therefore,

$$\max_{\theta \in \Gamma(\mu_1, \ldots, \mu_n)} \mathbb{E}_{\tilde{c} \sim \theta}[Z(\tilde{c})] \geq \mathbb{E}_{\tilde{c} \sim \theta^*}[Z(\tilde{c})] \geq \mathbb{E}_{(\tilde{c}, \tilde{x})}[\tilde{c}^\top \tilde{x}] = \sum_{i=1}^n \mathbb{E}_{(\tilde{c}_i, \tilde{x}_i)}[\tilde{c}_i \tilde{x}_i] = \max_{\nu \in \mathcal{P}(\mathcal{X})} \sum_i \max_{\gamma_i \in \Gamma(\mu_i, \Pi_i \nu)} \mathbb{E}_{(\tilde{c}_i, \tilde{x}_i) \sim \gamma_i}[\tilde{c}_i \tilde{x}_i].$$

By Theorem 1, we establish tightness of the above inequalities so that not only is θ^* optimal to $\max_{\theta \in \Gamma(\mu_1, \ldots, \mu_n)} \mathbb{E}_{\tilde{c} \sim \theta}[Z(\tilde{c})]$ but also $\mathbb{E}_{\tilde{c} \sim \theta^*}[Z(\tilde{c})] = \mathbb{E}_{(\tilde{c}, \tilde{x})}[\tilde{c}^\top \tilde{x}]$. Hence, $\tilde{x} \in argmax_{x \in \mathcal{X}} \tilde{c}^\top x$ with probability 1.

We define, for each i, $x_i^*(c_i) := x_i$ when $c_i \in supp(\tilde{c}_i \|_{\tilde{x}_i = x_i})$, so that by construction,

$$P_{\tilde{c} \sim \theta^*}(x_i^*(\tilde{c}) = \xi_i) = \int P(x_i^*(\tilde{c}_i) = \xi_i \| \tilde{x}_i) \, d\Pi_i \nu = \int \mathbb{1}_{\tilde{x}_i = \xi_i} \, d\Pi_i \nu = \Pi_i \nu(\xi_i).$$

Recalling that by Corollary 3, all marginals to primal optimal solutions are the same, we have established (8) for all primal optimal solutions. To show that $x^*(\tilde{c}) \in x^{OPT}(\tilde{c})$, we note that by the construction of $\tilde{c}\|_{\tilde{x}}$,

$$P(x^*(\tilde{c}) = \tilde{x}) = P(x_i^*(\tilde{c}_i) = \tilde{x}_i \;\; \forall i) = \int P(x_i^*(\tilde{c}_i) = \tilde{x}_i \;\; \forall i \| \tilde{x}) d\nu = \int \prod_{i=1}^n P(x_i^*(\tilde{c}_i) = \tilde{x}_i \| \tilde{x}_i) d\nu = \int 1 d\nu = 1,$$

so that $x^*(\tilde{c}) \in \arg\max_{x \in \mathcal{X}} \tilde{c}^\top x$ with probability 1.

□

EC.3.1.7. Example: On Primal Optimal Solutions and Persistency Consider now when $\mu_1, \ldots, \mu_n$ are not absolutely continuous but in fact finite-support measures. For a first example, consider n = 1, $\mu_1 = \mathbb{1}_0$, and $\mathcal{X} = \{0,1\}$. Then any distribution over $\mathcal{X}$ can solve (7). Let us consider one in particular in $\bar{\nu}$ defined by $\bar{\nu}(0) = \bar{\nu}(1) = 1/2$. Then it is clear that no solution mapping (i.e., a measurable selection of x^{OPT}) x^* exists such that

$$P_{\tilde{c} \sim \theta^*}(x^*(\tilde{c}) = 0) = \bar{\nu}(0), \quad P_{\tilde{c} \sim \theta^*}(x^*(\tilde{c}) = 1) = \bar{\nu}(1).$$

However, if we were to consider $\bar{\nu} = \mathbb{1}_0$, then there does exist such a mapping x^*, defined by $x^* \equiv 0$.

This might lead one to ask, in the case of $\mu_1, \ldots, \mu_n$ that are not absolutely continuous, will there always exist a primal optimal $\bar{\nu}$ such that a measurable selection of x^{OPT}, call it x^*, of form $x^*(c) = (x_1^*(c_1), \ldots, x_n^*(c_n))$ can be defined to yield the "persistence value" statements? The following example indicates that the answer is no. Consider n = 2, $\mathcal{X} = \{(0,1), (1,0)\}$,

$$\tilde{c}_1 = \begin{cases} 1 & w.p. 1/4 \\ 2 & w.p. 1/4 \\ 3 & w.p. 1/2 \end{cases}, \qquad \tilde{c}_2 = \begin{cases} 1.5 & w.p. 1/5 \\ 2.5 & w.p. 4/5 \end{cases}$$

Then the unique optimal $\bar{\nu}$ to (7) is $\bar{\nu} = 1/2 \cdot \mathbb{1}_{(0,1)} + 1/2 \cdot \mathbb{1}_{(1,0)}$. By inspection, we find that the unique θ^* is given by: $(\tilde{c}_1, \tilde{c}_2) = \begin{cases} (1, 2.5) & w.p. 1/4 \\ (2, 2.5) & w.p. 1/4 \\ (3, 1.5) & w.p. 1/5 \\ (3, 2.5) & w.p. 3/10 \end{cases}$

Any solution mapping x^* maps all $(c_1, c_2) : c_1 < c_2$ to $(0,1)$ and all $(c_1 c_2) : c_1 > c_2$ to $(1,0)$, with arbitrary decision for the case of $c_1 = c_2$. So an example solution mapping x^* would involve: $x^*(c_1, c_2) = \begin{cases} (0,1); & (c_1, c_2) \in \{(1, 2.5), (2, 2.5)\} \\ (1,0); & (c_1, c_2) \in \{(3, 1.5), (3, 2.5)\} \end{cases}$, so that $x^* \# \theta^* = \bar{\nu}$. But, x^* is not of the form $x^*(c) = (x_1^*(c_1), \ldots, x_n^*(c_n))$.

△

EC.3.2. Section 3 Proofs and Examples

EC.3.2.1. Proof of Proposition 2

PROPOSITION 2. *The dual formula* $\underset{\{\psi_i:\mathcal{X}_i\to\mathbb{R}\}_{i=1}^n}{min} \ \underset{\nu\in\mathcal{P}(\mathcal{X})}{sup} L(\{\psi_i\}_i,\nu)$ *is equivalent to*

$$\begin{aligned}\min_{\psi_i:\mathcal{X}_i\to\mathbb{R}} & \sum_{i=1}^n \int \psi_i^*(\tilde{c}_i)d\mu_i \\ s.t. & \sum_{i=1}^n \psi_i(x_i) \le 0, \forall x \in \mathcal{X}.\end{aligned}$$

Proof: From Theorem 1, we have

$$\begin{aligned}\max_{\theta\in\Gamma(\mu_1,\dots,\mu_n)} \mathbb{E}_{\tilde{c}\sim\theta}[Z(\tilde{c})] &= \min_{\psi_i:\mathcal{X}_i\to\mathbb{R}} \max_{\nu\in\mathcal{P}(\mathcal{X})} \sum_{i=1}^n \int \psi_i^*(\tilde{c}_i)d\mu_i + \int \left[\sum_{i=1}^n \psi_i(x_i)\right] d\nu(x) \\ &= \min_{\psi_i:\mathcal{X}_i\to\mathbb{R}} \sum_{i=1}^n \int \psi_i^*(\tilde{c}_i)d\mu_i + \max_{\nu\in\mathcal{P}(\mathcal{X})} \sum_{x\in\mathcal{X}} \left[\sum_{i=1}^n \psi_i(x_i)\right]\nu(x) \\ &= \min_{\psi_i:\mathcal{X}_i\to\mathbb{R}} \sum_{i=1}^n \int \psi_i^*(\tilde{c}_i)d\mu_i + \max_{\nu\in\mathcal{P}(\mathcal{X})} \sum_{i=1}^n \sum_{x_i\in\mathcal{X}_i} \psi_i(x_i)\Pi_i\nu(x_i),\end{aligned}$$

which can be reformulated as

$$\begin{aligned}\min_{\theta,\psi_i:\mathcal{X}_i\to\mathbb{R}} & \sum_{i=1}^n \int \psi_i^*(\tilde{c}_i)d\mu_i + \theta \\ s.t. & \sum_{i=1}^n \sum_{x_i\in\mathcal{X}_i} \psi_i(x_i)\Pi_i\nu(x_i) \le \theta, \forall \nu \in \mathcal{P}(\mathcal{X})\end{aligned} \tag{EC.10}$$

What's more, we can further modify (EC.10) as

$$\begin{aligned}\min_{\psi_i:\mathcal{X}_i\to\mathbb{R}} & \sum_{i=1}^n \int \psi_i^*(\tilde{c}_i)d\mu_i \\ s.t. & \sum_{i=1}^n \sum_{x_i\in\mathcal{X}_i} \psi_i(x_i)\Pi_i\nu(x_i) \le 0, \forall \nu \in \mathcal{P}(\mathcal{X})\end{aligned}$$

To see it, denote an arbitrary feasible solution to (EC.10) as $((\bar{\Psi}_i)_i,\bar{\theta})$. Then $((\bar{\Psi}_i)_i - \bar{\theta}/n, 0)$ is also feasible to (EC.10) and generates the same objective value. To see it, note $\sum_{i=1}^n \sum_{x_i\in\mathcal{X}_i} \left(\bar{\psi}_i(x_i) - \bar{\theta}/n\right)\Pi_i\nu(x_i) \le 0$ holds. Hence it is feasible to (EC.10). In addition, it generates an objective value $\sum_{i=1}^n \int \max_{x_i}\left\{c_i x_i - (\bar{\Psi}_i(x_i) - \bar{\theta}/n)\right\} d\mu_i$, which is equal to $\sum_{i=1}^n \int \bar{\Psi}_i^*(\tilde{c}_i)d\mu_i + \bar{\theta}$. Therefore, it is equivalent to solve (9). □

EC.3.2.2. Proof of Theorem 2

THEOREM 2 *Computing Z^* in MDM for the class of linear optimization problems given discrete marginal distributions and an H-polytope is NP-hard.*

Proof: Assume that each $\tilde{c}_i$ is a random variable taking values in the set $\{-1,1\}$ where $\tilde{c}_i = 1$ with probability p_i and -1 with probability $1-p_i$. Given a realization of the vector c, we associate with it the set $S = \{i \in [n] : c_i = 1\}$ and $S^c = \{i \in [n] : c_i = -1\}$. The corresponding objective function of the linear program is given as:

$$Z(S) = \max\left\{\sum_{i\in S} x_i - \sum_{i\in S^c} x_i : Ax \le b\right\}.$$

Given the input probabilities $p_1, \dots, p_n \in [0,1]$ and an H-polytope, MDM is formulated as:

$$\begin{aligned}
\max \ & \sum_{S\subseteq[n]} p_S \max\left\{\sum_{i\in S} x_i - \sum_{i\in S^c} x_i : Ax \le b\right\} \\
\text{s.t.} \ & \sum_{S\subseteq[n]: S\ni i} p_S = p_i, && \forall i \in [n], \\
& \sum_{S\subseteq[n]} p_S = 1, \\
& p_S \ge 0, && \forall S \subseteq [n],
\end{aligned} \tag{EC.11}$$

where the decision variables are the probabilities of the scenarios denoted by p_S for $S \subseteq [n]$. The dual of this linear program is given as:

$$\begin{aligned}
\min \ & y_0 + \sum_{i\in[n]} p_i y_i \\
\text{s.t.} \ & y_0 + \sum_{i\in S} y_i \ge \max\left\{\sum_{i\in S} x_i - \sum_{i\in S^c} x_i : Ax \le b\right\}, \ \forall S \subseteq [n].
\end{aligned} \tag{EC.12}$$

The separation problem for the dual linear program is as follows: Given a set of values $y_0, y_1, \dots, y_n$, verify if:

$$y_0 + \sum_{i\in S} y_i \ge \max\left\{\sum_{i\in S} x_i - \sum_{i\in S^c} x_i : Ax \le b\right\}, \ \forall S \subseteq [n], \tag{Separation}$$

else find a violated inequality. Given the equivalence of separation and optimization, it suffices to show that the separation problem is NP-hard. Towards this end, let $y_i = 0$ for all $i \in [n]$. Then the separation problem is to verify that

$$y_0 \ge \max\left\{\sum_{i\in S} x_i - \sum_{i\in S^c} x_i : Ax \le b\right\}, \ \forall S \subseteq [n],$$

which is equivalent to

$$y_0 \geq \max_{x: Ax \leq b} \max_{S \subseteq [n]} \sum_{i \in S} x_i - \sum_{i \in S^c} x_i,$$
$$= \max_{x: Ax \leq b} \|x\|_1.$$

The right hand side corresponds to 1-norm maximization over polytopes which is known to be NP-hard (see Mangasarian and Shiau (1986)), implying that the problem of computing Z^* is NP-hard.

□

EC.3.3. Section 4 Proofs and Examples

EC.3.3.1. Proof of Theorem 3

THEOREM 3: SUFFICIENCY VIA 0/1 FORMULATION *Let $\mathcal{X}$ be a finite point set. Suppose there exists a 0/1 polytope $P \subseteq [0,1]^B$, where B is some finite set, such that the family of all possible collections of marginals $\{(\Pi_1\nu, \Pi_2\nu \dots \Pi_n\nu) : \nu \in \mathcal{P}(\mathcal{X})\}$ is in bijection with P in which to any $y \in P$, the corresponding marginals $(\Pi_1\nu, \Pi_2\nu \dots \Pi_n\nu)$ are given by*

$$\Pi_i\nu(\xi_i) = \sum_{S \in B_i(\xi_i)} y(S) \quad \forall i \in [n], \forall \xi_i \in \mathcal{X}_i, \tag{11}$$

where the $B_i(\xi_i) \subseteq 2^B$. Then

$$\max_{\nu \in \mathcal{P}(\mathcal{X})} \sum_{i=1}^n \sum_{x_i \in \mathcal{X}_i} \psi_i(x_i) \Pi_i\nu(x_i) = \max_{y \in P} \sum_{i=1}^n \sum_{x_i \in \mathcal{X}_i} \psi_i(x_i) \sum_{S \in B_i(x_i)} y(S).$$

In particular, the separation problem (10) reduces to a linear program over a 0/1 polytope.

Proof: Clearly (11) implies that

$$\max_{\nu \in \mathcal{P}(\mathcal{X})} \sum_{i=1}^n \sum_{x_i \in \mathcal{X}_i} \psi_i(x_i) \Pi_i\nu(x_i) = \max_{y \in P} \sum_{i=1}^n \sum_{x_i \in \mathcal{X}_i} \psi_i(x_i) \sum_{S \in B_i(x_i)} y(S).$$

□

REMARK EC.2 (INEXACT PROJECTIONS $\mathcal{X}_i$). In the dual formulation of Theorem 1, for any i the potential function ψ_i is defined on $\mathcal{X}_i$. In practice, we may not be able to identify $\mathcal{X}_i$ precisely.

What if the best we have is an approximation in the form of a superset $\mathcal{X}_i' \supseteq \mathcal{X}_i$? This actually does not cause complications, as it remains true that

$$\min_{\{\psi_i:\mathcal{X}_i\to\mathbb{R}\}_{i=1}^n} \sup_{\nu\in\mathcal{P}(\mathcal{X})} \sum_i \int \psi_i d\Pi_i\nu + \int max_{x_i}\{\tilde{c}_i x_i - \psi_i(x_i)\} d\mu_i$$
$$= \min_{\{\psi_i:\mathcal{X}_i'\to\mathbb{R}\}_{i=1}^n} \sup_{\nu\in\mathcal{P}(\mathcal{X})} \sum_i \int \psi_i d\Pi_i\nu + \int max_{x_i}\{\tilde{c}_i x_i - \psi_i(x_i)\} d\mu_i.$$

This follows because the potential functions ψ_i in both formulations above may be assumed to be convex; further, if $x_i' \notin \mathcal{X}_i$, then $\Pi_i\nu(x_i') = 0$ for all primal optimal $\nu \in \mathcal{P}(\mathcal{X})$. More precisely, extending the definition of ψ_i from $\mathcal{X}_i$ to $\mathcal{X}_i'$ - in a manner that preserves convexity - can only result in increasing ψ_i^* (if any change at all) pointwise. And with $\psi_i(x_i')\cdot\Pi_i\nu(x_i') = 0$, for any $x_i' \in \mathcal{X}_i' \setminus \mathcal{X}_i$, we see that no such extension could help with minimization.

△

EC.3.3.2. Proof of Corollary 4

Corollary 4 *Let* $\Pi = (P_1, P_2, \ldots)$ *be a collection of polytopes. Suppose there exists a routine such that for arbitrary inputs* n *and* $Q \in P_n$ *(with* $\mathcal{X} := Extr(Q)$*), we may efficiently identify the polytope* P *as well as the* $B_i(\cdot) : \mathcal{X}_i \to 2^B$ *described in Theorem 3. Then* $MDM(\mathcal{M}\Pi)$ *is efficient, assuming Assumption 1 on* $\mathcal{M} = (\mu_1, \ldots, \mu_n)$*.*

Proof: With the hypothesized routine on hand, the separation problem always reduces to a linear program. □

EC.3.3.3. Proof of Corollary 5

V- and H-Polytopes *If* Π *is either a collection of V-polytopes or* $0/1$ *H-polytopes, then* Π *satisfies the conditions of Corollary 4.*

Proof: We begin by assuming Π is a collection of V-polytopes. Let n and V-polytope $Q \in P_n$ be given in the form of the list of points $\mathcal{X} = Extr(Q) = \{x^1, \ldots, x^m\}$. If we let $B := \{1, \ldots, m\}$, $\mathcal{F} := \{S \subset B : |S| = 1\}$, and $\varphi : \mathcal{F} \to \mathcal{X}$ defined by $\varphi(S) := G \cdot \chi_S$, where $\chi_S \in \{0,1\}^B$ denotes the 0/1 characteristic vector of the set $S \subset B$, and G is the $n \times m$ matrix formed by appending the elements of $\mathcal{X}$ together as columns, then clearly to any $\nu \in \mathcal{P}(\mathcal{X})$ there is a unique $\mathbb{P} \in \mathcal{P}(\mathcal{F})$ - and vice versa. And to every $\mathbb{P} \in \mathcal{P}(\mathcal{F})$ there is a unique $y \in P = \Delta_{m-1} := \{y \in \mathbb{R}^m : y \geq 0, \sum_i y_i = 1\}$ - and vice versa. Further,

$$\Pi_i \nu(\xi_i) = \sum_{j : x_i^j = \xi_i} y_j \quad \forall i \in [n], \forall \xi_i \in \mathcal{X}_i.$$

So upon identifying $B_i(\xi_i) := \{S \in \mathcal{F} : S = \{j\}, x_i^j = \xi_i\}$ for all $\xi_i \in \mathcal{X}_i$ (which is a matter of sorting the $i-th$ row of G), we establish satisfaction of the hypothesis of Theorem 3. This identification of P and $B_i(\cdot)$ being efficient in the input list $\mathcal{X}$ and integer n, we complete the argument by appealing to Corollary 4.

Next, we assume Π be a collection of 0/1 H-polytopes. Let n and H-polytope $Q \in P_n$ be given in the form of a linear system of inequalities $Q = \{x : Ax \leq b\}$, where by assumption $\mathcal{X} := Extr(Q) \subseteq \{0,1\}^n$. For the sake of notation, let $[n] := \{1, \ldots, n\}$. If we let $B := [2n]$, then to every collection $(\Pi_1 \nu, \Pi_2 \nu, \ldots, \Pi_n \nu)$ of marginals for some $\nu \in \mathcal{P}(\mathcal{X})$ is a unique member of $P = \{y = (y^+, y^0) \in [0,1]^n \times [0,1]^n : y^+ \in Q, y^0 = \mathbb{1}_n - y^+\}$ - and vice versa. Indeed, if $\nu \in \mathcal{P}(\mathcal{X})$, then $(y^+, y^0) := (\sum_{x \in \mathcal{X}} \nu(x) \cdot x, \mathbb{1}_n - \sum_{x \in \mathcal{X}} \nu(x) \cdot x)$ is the unique member of P that satisfies

$$\Pi_i \nu(1) = \sum_{x \in \mathcal{X} : x_i = 1} \nu(x) = y_i^+, \qquad \Pi_i \nu(0) = \sum_{x \in \mathcal{X} : x_i = 0} \nu(x) = 1 - y_i^+ = y_i^0 = y_{n+i} \quad i = 1, \ldots, n.$$

Conversely, if $y = (y^+, y^0) \in P$, then $y^+ = \sum_{x \in \mathcal{X}} \nu(x) \cdot x$, for some $\nu \in \mathcal{P}(\mathcal{X})$; while there may be more than one such ν, the collection of summands $\sum_{x \in \mathcal{X} : x_i = 1} \nu(x) =: \Pi_i \nu(1)$ is unique across all such ν, as desired. To conclude, upon identifying $B_i(1) := \{i\}$, $B_i(0) = \{n+i\}$, we establish that the hypothesis of Theorem 3 holds. With the identification of the $B_i(\cdot)$ and P being obviously efficient and compact in the input polytope Q and integer n, we complete the argument by appealing to Corollary 4. □

EC.3.3.4. Proof of Corollary 6

Corollary 6 *$L^\natural$ convex polytopes can be represented as an H-polytope that satisfies the sufficiency condition described in Theorem 3, hence Z^* is efficiently computable with any collection of marginals $\mu_1, \ldots, \mu_n$ that satisfy Assumption 1.*

Proof: Let $G=(V,\mathcal{A})$, $d: V\times V\to\mathbb{Z}\cup\{+\infty\}$ (satisfying the triangle inequality) and $\{l_i\}_{i=1}^n\subset\mathbb{Z}\cup\{-\infty\}$, $\{u_i\}_{i=1}^n\subset\mathbb{Z}\cup\{+\infty\}$ be given as in the above for the characterization of $\mathcal{X}$. Define $u:=\max_i u_i$ and $l:=\min_i l_i$. We claim that

$$x_i-x_j\le d_{ij}\ \ \forall(i,j)\in\mathcal{A}\iff[x_i\in\{t,\ldots,u\}\implies x_j\notin\{l,\ldots,t-d_{ij}-1\}]\ \ \forall(i,j)\in\mathcal{A},\forall t\in[l,u]\cap\mathbb{Z}.$$

The "only if" direction is clear. The "if" direction is also clear, as given an $(i,j)\in\mathcal{A}$, $x_i\in\{x_i,\ldots,u\}\implies x_j\notin\{l,\ldots,x_i-d_{ij}-1\}$, by hypothesis. Hence, $x_j\ge x_i-d_{ij}$, as desired.

Using indicator functions, we can view the equivalence just as well in the form:

$$x_i-x_j\le d_{ij}\quad\forall(i,j)\in\mathcal{A}\iff\sum_{s=t}^{u}\mathbb{1}_{x_i=s}+\sum_{s=l}^{t-d_{ij}-1}\mathbb{1}_{x_j=s}\le 1\ \ \forall(i,j)\in\mathcal{A},\forall t\in[l,u]\cap\mathbb{Z}$$

Additionally, since $x_i=\sum_{\bar{x}_i=l_i}^{u_i}\bar{x}_i\cdot y_{i,\bar{x}_i}$, we have found an extended form of the kind in Theorem 3, with

$$P:=\{y\in\mathbb{R}_+^{n\times[l,u]_{\mathbb{Z}}}:\sum_{x_i\in[l_i,u_i]_{\mathbb{Z}}}y_{i,x_i}=1\ \ \forall i,\ \sum_{s=t}^{u}y_{i,s}+\sum_{s=l}^{t-d_{ij}-1}y_{js}\le 1\ \ \forall(i,j)\in\mathcal{A},t=l,l+1,\ldots,u\},$$

where we have set $\mathrm{B}=\{1,\ldots,n\}\times[l,u]_{\mathbb{Z}}$, with u and l efficiently identifiable. Further, using this form, the projections $\mathcal{X}_i$ can be identified efficiently using linear programming, and $B_i(\bar{x}_i)$ can be identified with the singleton $\{(i,\bar{x}_i)\}$.

What remains is to show that P is an integral polytope. The basis of the following line of argument is described succinctly in Möhring et al. (2001), and we take the steps to elaborate on it here for the sake of completeness. Towards this end, we will construct an undirected graph $\bar{G}=(\bar{V},\bar{\mathcal{A}})$ for analysis. For each $i\in V$ and $t\in\{l,\ldots,u\}$, there is a node $x_{i,t}$. For every $(i,j)\in\mathcal{A}$ and $t\in\{l,\ldots,u\}$, draw an edge between node $x_{i,t}$ and each of the nodes $x_{j,s}$, for $l\le s\le t-d_{ij}-1$. Also, for each $i\in V$, draw edges to make the collection of nodes $\{x_{i,s}\}_{s\in[l,u]\cap\mathbb{Z}}$ a clique. This completes the construction of the undirected graph.

Next, we will assign an orientation to the edges to show that this graph has a transitive orientation. To do this, for $(i,j)\in\mathcal{A}$, let the edge between nodes $x_{i,t}$ and $x_{j,t'}$ (where $t'\le t-d_{ij}-1$) be drawn from $x_{i,t}$ towards $x_{j,t'}$. Let the edges among the clique $\{x_{i,s}\}_{s\in[l,u]\cap\mathbb{Z}}$ be drawn from $x_{i,s}$ towards

$x_{i,t}$ if and only if $s \geq t$. Now, we verify the transitivity. Observe that by the triangle inequality of d, for $(i,j),(j,k) \in \mathcal{A}$,

$$[x_{i,t} \to x_{j,t'}] \wedge [x_{j,t'} \to x_{k,t''}] \Longrightarrow [t' \leq t - d_{ij} - 1] \wedge [t'' \leq t' - d_{jk} - 1] \Longrightarrow t'' \leq t - d_{ik} - 2 < t - d_{ik} - 1 \Longrightarrow x_{i,t} \to x_{k,t''},$$

where $u \to v$ denotes the existence of a directed edge from node u towards node v in the oriented graph. Further, for $(i,j) \in \mathcal{A}$,

$$x_{i,t} \to x_{i,t'} \wedge x_{i,t'} \to x_{j,t''} \Longrightarrow t' \leq t \wedge t'' \leq t' - d_{ij} - 1 \Longrightarrow t'' \leq t - d_{ij} - 1 \Longrightarrow x_{i,t} \to x_{j,t''}.$$

This suffices to conclude that the orientation exhibits transitivity in this directed graph, making G a comparable graph. Then, by Mirsky's theorem, $\bar{G}$ is a perfect graph. By an established result, perfectness yields that the 0/1 stable set polytope for $\bar{G}$ is characterized by the following collection of facet-defining inequalities: $\{z \in \mathbb{R}_+^{\bar{V}} : \sum_{v \in K} z_v \leq 1\}$, where $K \subset \bar{V}$ is some maximal clique in $\bar{G}$. Observe that for arbitrary $(i,j) \in \mathcal{A}$ and $t \in \{l \dots, u\}$, the collection of nodes $\{x_{i,s}\}_{s=t}^{u} \cup \{x_{j,s}\}_{s=l}^{t-d_{ij}-1}$ forms precisely a maximal clique- further, it is clear that all maximal cliques of $\bar{G}$ are of this form. Hence, we can view P as a face of the stable set polytope of $\bar{G}$, concluding it is integral (as the stable set polytope is integral), as desired. □

EC.3.3.5. Proof of Corollary 7

Corollary 7 *Let* $\Pi = (P_1, P_2, \dots)$ *either be a collection of V-polytopes for which any member* $Q \in P_n$ *has its input vertex set* $\mathcal{X}$ *be a sublattice, or let* Π *be a collection of H-polytopes for which any member* $Q \in P_n$ *is a sublattice. Further, suppose the function* $Z(c)$ *is efficiently computable over all* $Q \in P_n$. *Then* $MDM(\mathcal{M}\Pi)$ *is efficient, assuming finite-support marginals in* $\mathcal{M} = (\mu_1, \dots, \mu_n)$.

Proof: For the V-polytope case:

Define the function $\delta_{\mathcal{X}}(x) := \begin{cases} 0 & x \in \mathcal{X} \\ +\infty & x \notin \mathcal{X} \end{cases}$. It follows that

$$\delta_{\mathcal{X}}(x \vee y) + \delta_{\mathcal{X}}(x \wedge y) \leq \delta_{\mathcal{X}}(x) + \delta_{\mathcal{X}}(y) \qquad \forall x, y \in \mathcal{X}$$

is equivalent to $\mathcal{X}$ being closed under the lattice operations $\vee$ and $\wedge$. So, $\delta_{\mathcal{X}}$ is submodular. And since Z is the Legendre-Fenchel conjugate of $\delta_{\mathcal{X}}$, Z is consequently supermodular (see Murota (2003)), as desired.

For the H-polytope case:

Let $\mathcal{X}$ denote the set of extreme points to an arbitrary H-polytope. Observe that $Z = \delta^*_{\mathcal{X}} = \delta^{***}_{\mathcal{X}} = \delta^*_{conv(\mathcal{X})}$. By hypothesis, $conv(\mathcal{X})$ is a sublattice, so as similarly argued already, $\delta_{conv(\mathcal{X})}$ is submodular, meaning its conjugate, Z, is supermodular, as desired. □

EC.3.3.6. Example 4.1 Consider the polytope in $\mathbb{R}^3$ generated by the set $\mathcal{X}$ comprised of: $(0,0,0)^\top, (0,1,1)^\top$, and $(1,2,1)^\top$. If Z is the support function to this polytope, then to verify that it is supermodular, by Theorem 3.14 of Shioura and Tamura (2015), it suffices to check that for any x, $\arg\max_a x^\top a - Z(a) = \{a : Aa \le \beta\}$, where $A = \begin{bmatrix} \alpha_1^\top \\ \dots \\ \alpha_m^\top \end{bmatrix}$ is composed of row vectors $\alpha_i \in \mathbb{R}^3_+ \cup \mathbb{R}^3_-$.

The case of $x \notin conv(\mathcal{X})$ is trivially true, so we focus on the case of $x \in conv(\mathcal{X}) =: Q$, for which $\arg\max_a x^\top a - Z(a) = \{a : x^\top a = Z(a)\}$.

- $x \in ri(Q) \implies \{a : x^\top a = Z(a)\} = \{a : \begin{bmatrix} 1 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix} a = \begin{bmatrix} 0 \\ 0 \end{bmatrix}\}$
- $x = \lambda(0,0,0)^\top + (1-\lambda)(0,1,1)^\top$ for some $\lambda \in (0,1)$

  $\implies$ the tangent cone $T_Q(x) = cone((0,1,1)^\top, -(0,1,1)^\top, (1,1,0)^\top)$

  $\iff \{a : x^\top a = Z(a)\} = \{a : \begin{bmatrix} 0 & 1 & 1 \\ 0 & -1 & -1 \\ 1 & 1 & 0 \end{bmatrix} a \le \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}\}$
- $x = \lambda(0,1,1)^\top + (1-\lambda)(1,2,1)^\top$ for some $\lambda \in (0,1)$

  $\implies$ the tangent cone $T_Q(x) = cone((1,1,0)^\top, -(1,1,0)^\top, -(1,2,1)^\top)$

  $\iff \{a : x^\top a = Z(a)\} = \{a : \begin{bmatrix} 1 & 1 & 0 \\ -1 & -1 & 0 \\ -1 & -2 & -1 \end{bmatrix} a \le \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}\}$

- $x = \lambda(0,0,0)^\top + (1-\lambda)(1,2,1)^\top$ for some $\lambda \in (0,1)$

$\Longrightarrow$ the tangent cone $T_Q(x) = cone((1,2,1)^\top, -(1,2,1)^\top, (0,1,1)^\top)$

$$\Longleftrightarrow \{a : x^\top a = Z(a)\} = \{a : \begin{bmatrix} 1 & 2 & 1 \\ -1 & -2 & -1 \\ 0 & 1 & 1 \end{bmatrix} a \leq \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}\}$$

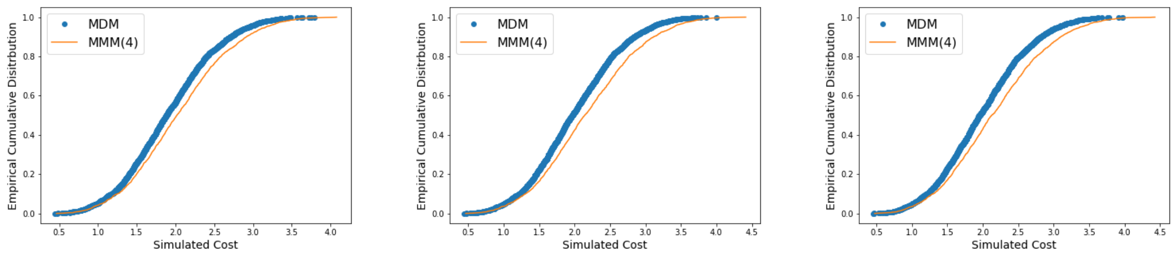
- At $x = (0,0,0)^\top$, $T_Q(x) = cone\Big((0,1,1)^\top, (1,2,1)^\top\Big)$
- At $x = (0,1,1)^\top$, $T_Q(x) = cone\Big(-(0,1,1)^\top, (1,1,0)^\top\Big)$
- At $x = (1,2,1)^\top$, $T_Q(x) = cone\Big(-(1,2,1)^\top, -(1,1,0)^\top\Big)$

Therefore, with the (necessary and) sufficient condition satisfied for all $x \in Q$, we conclude that the support function Z to Q is indeed supermodular by Theorem 3.14 of Shioura and Tamura (2015). However, $(1/2,1,1/2)^\top, (0,1,1)^\top \in Q$ and yet $(1/2,1,1/2)^\top \vee (0,1,1)^\top = (1/2,1,1)^\top \notin Q$, meaning Q is not a sublattice.

EC.4. Computational Experiments

In this section, we use appointment scheduling problem as an example to illustrate the performance of MDM. Mak et al. (2015) have studied the appointment scheduling problem considered in this paper with marginal moments. We use the MMM(k) model to denote their model, where the first k moments are provided in the ambiguity set. They show that MMM for any $k \geq 2$ can be formulated as an SDP using the univariate polynomial optimization technique proposed by Bertsimas and Popescu (2005). Compared to their model, MDM assumes more information on the uncertainties. We would like to examine the value of incorporating more information on uncertainties by comparing MDM with MMM. In particular, we evaluate and compare the performance of the robust solutions from MDM and MMM, respectively when the distributions of the service durations are known and independent.

Our experimental setup is based on Mak et al. (2015). Namely, we consider the case of $n = 5$ jobs, and assume that the duration of each job independently follows a probability distribution. Three types of probability distributions including normal, gamma, and log-normal are tested. Under each type of probability assumption, we generate our random problem instances as follows:



(a) Normal (b) Log-Normal (c) Gamma

Figure EC.1 Performance Comparison Between MDM Solution and MMM Solution

- Randomly generate 2000 instances each by sampling:

  —Mean[1] $m_i \sim U[3,6]$ and standard deviation $\sigma_i = m_i \cdot \epsilon$, with parameter $\epsilon \sim U[0,0.3]$,

  —Planning horizon $T = \sum_i m_i + (0.5) \cdot \sqrt{\sum_i \sigma_i^2}$.

For each instance, we generate 5000 samples for each job duration from a pre-specified probability distribution (normal/ gamma/ log-normal). Based on the samples, we can approximate (**??**) using the sample mean for the uni-variate expectation in the second term. Besides, we can get the empirical moments (up to the fourth moment in this study) from the samples, which serve as inputs of the MMM(4) model as the benchmark. For each robust solution (from either MDM or MMM), we evaluate its sample-average performance by re-generating 5000 samples from the aforementioned distribution. The comparison is based on this simulated cost.

We plot the empirical cumulative distribution function of the simulated costs of the solutions from the MDM and MMM(4) models for the 2000 randomly generated problem instances in Figure EC.1. As the figures illustrate, MDM can present a sizeable reduction of the expected waiting time over MMM(4). The mean of the relative reduction for the three cases are 6.61% (Normal), 8.07% (Log-Normal), 8.03% (Gamma). The results show that incorporating the whole marginal distribution indeed helps to mitigate the conservativeness of the robust solution compared to the robust model considering only marginal moments information.