Distributionally Robust Linear and Discrete Optimization with Marginals

Louis Chen
Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA 02139, llchen@mit.edu

Will Ma
Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA 02139, willma@mit.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
Department of Decision Science, National University of Singapore a0109727@u.nus.edu

Abstract In this paper, we study the class of linear and discrete optimization problems in which the objective coefficients are chosen randomly from a distribution, and the goal is to evaluate robust bounds on the expected optimal value as well as the marginal distribution of the optimal solution. The set of joint distributions is assumed to be specified up to only the marginal distributions. We generalize the primal-dual formulations for this problem from the set of joint distributions with absolutely continuous marginal distributions to arbitrary marginal distributions using techniques from optimal transport theory. While the robust bound is shown to be NP-hard to compute for linear optimization problems, we identify a sufficient condition for polynomial time solvability using extended formulations. This generalizes the known tractability results under marginal information from 0-1 polytopes to a class of integral polytopes and has implications on the solvability of distributionally robust optimization problems in areas such as scheduling which we discuss.

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, Dentcheva and Ruszczyński (2014)) as follows. 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 as follows:

$$\min_{s \in \mathcal{S}} E_{\tilde{\xi} \sim p} \left[h(s, \tilde{\xi}) \right].$$

This formulation however makes the strong 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 a 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:

$$\min_{s \in \mathcal{S}} \sup_{p \in \mathcal{P}} E_{\tilde{\xi} \sim p} \left[h(s, \tilde{\xi}) \right].$$

Such a set $\mathcal{P}$ has been constructed in a wide 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 Scarf (1958), Bertsimas and Popescu (2005), Delage and Ye (2010), Bertsimas et al. (2010)), the set of distributions with information on the marginal distributions (see Meilijson and Nadas (1979)) and marginal moments (see Bertsimas, Natarajan and Teo (2004)), the set of distributions with confidence sets and mean values residing on an affine manifold (see Wiesemann, Kuhn and Sim (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 (2017), Gao and Kleywegt (2016), Blanchet et al. (2017a)). This list is by no means comprehensive with an increasing number of applications of this technique in areas such as supply chains, finance, healthcare and machine learning to name a few. We refer the interested reader to the papers listed above and the references therein.

In this paper, we contribute to this literature by providing new results for the case where $h(s,\xi)$ is defined as the optimal value to linear and discrete optimization problems and the set $\mathcal{P}$ is defined by the *Fréchet* class, the class of multivariate distributions with fixed marginal distributions. To motivate the problem, we use the appointment scheduling problem from healthcare as an example (see Gupta and Denton (2008)). 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 (see Kong et al. (2013), Mak, Rong and Zhang (2015) for examples). Assume that a set of n patients $\{1,2,\ldots,n\}$ who arrive in a fixed order need to be scheduled in a given time interval. We assume that for any patient i, the distribution μ_i of the possible service time $\tilde{c}_i$ with the doctor is known. Let $\Gamma(\mu_1,\mu_2,\ldots,\mu_n)$ denote the set of all possible joint distributions consistent with the given marginals μ_i. The decision variables are the amount of service times scheduled for each patient i, denoted by s_i, for $i\in[n]:=\{1,2,\ldots,n\}$. Patient 1 arrives at time 0 while we instruct patient 2 to arrive at time s_1, patient 3 to arrive at time s_1+s_2, and so on. The feasible region of the decision vector s is denoted by $\mathcal{S}$. An example of such a set is $\mathcal{S}=\{s\in\mathbb{R}^n:\sum_{i=1}^n s_i\leq T,\ s_i\geq 0\ \forall i\in[n]\}$, where we want to schedule all patients before time T. Assuming patient 1 has zero waiting time, and denoting the doctor by patient index n+1, the waiting time of patient $i+1$ is defined by the recursion $w_{i+1}=\max(0,w_i+c_i-s_i)$ for $i\in[n]$ where $w_1=0$ and w_{n+1} is the overtime of the doctor beyond time T. The distributionally robust appointment scheduling problem is to find a schedule to minimize the worst-case expected sum of the uncertain waiting times of the n patients and the overtime cost as follows:

$$\min_{s\in\mathcal{S}}\ \sup_{\theta\in\Gamma(\mu_1,\ldots,\mu_n)} E_{\tilde{c}\sim\theta}\left[Z(s,\tilde{c})\right], \tag{1}$$

where $Z(s,\tilde{c})$ is the optimal value to the random linear optimization problem:

$$\begin{aligned} Z(s,\tilde{c}) = \min\ & \sum_{i=1}^{n+1} w_i \\ \text{s.t. } & w_1 = 0, \\ & w_{i+1} \geq 0, \qquad \forall i = 1,\ldots,n, \\ & w_{i+1} \geq w_i + \tilde{c}_i - s_i,\ \forall i = 1,\ldots,n. \end{aligned} \tag{2}$$

Applying linear programming duality, this is equivalent to:

$$\begin{aligned} Z(s,\tilde{c}) = \max\ & \sum_{i=1}^{n} (\tilde{c}_i - s_i) x_i \\ \text{s.t. } & x_i - x_{i-1} \geq -1,\ \forall i = 2,\ldots,n, \\ & x_n \leq 1, \\ & x_i \geq 0, \qquad \forall i = 1\ldots,n. \end{aligned} \tag{3}$$

The problem in (1) was first studied in Kong et al. (2013) under a different specification of the set of distributions. Under the assumption that only the mean and the covariance matrix of the service times are specified, Kong et al. (2013) showed that the distributionally robust appointment scheduling problem can be equivalently formulated as a copositive optimization problem. While such a problem is not solvable in polynomial time, the authors showed that a semidefinite relaxation to this problem provides good schedules. In a followup paper, Mak, Rong and Zhang (2015) studied problem (1) under the assumption that only the means and the variances of the service times are specified but the covariance matrix and more generally the dependence structure is completely unknown. Our work is closely related to this stream of research. Interestingly under this specification of the set of distributions, Mak, Rong and Zhang (2015) showed that the distributionally robust appointment scheduling problem is solvable in polynomial time with the use of second order conic optimization methods. This brings us to a natural question as to what is the underlying structure that makes the problem tractable for this set of distributions and how does the result generalize to other optimization problems? In this paper, we provide a partial answer to this question by identifying such a class of linear and discrete optimization problems for the Fréchet class of distributions.

A direct application of the dual linear program in (3) to the distributionally robust optimization problem (1), leads us to consider a maximization problem of the following form:

$$Z(\tilde{c}) := \max_{x \in \mathcal{X}} \sum_{i=1}^{n} \tilde{c}_i x_i, \tag{4}$$

where $\mathcal{X} \subset \mathbb{R}^n$ is an arbitrary finite set (typically very large), and the $\tilde{c}_i$'s are the random objective coefficients. We focus on finding a tight upper bound on the expected value of this maximization problem with a linear objective where the objective coefficients are chosen randomly from a distribution in the set $\Gamma(\mu_1, \ldots, \mu_n)$ as follows:

$$\begin{aligned} Z^* &= \sup_{\theta \in \Gamma(\mu_1,\ldots,\mu_n)} E_{\tilde{c} \sim \theta}\left[Z(\tilde{c})\right], \\ &= \sup_{\theta \in \Gamma(\mu_1,\ldots,\mu_n)} E_{\tilde{c} \sim \theta}\left[\tilde{c}^T x^*(\tilde{c})\right], \end{aligned} \tag{5}$$

where $\tilde{c}$ is a random vector and $x^*(\cdot)$ is an optimal solution mapping. Note that the convex hull of the set $\mathcal{X}$, denoted by $conv(\mathcal{X})$, is a polytope with $Z(c) = \max\{\sum_{i=1}^n c_i x_i : x \in conv(\mathcal{X})\}$. Our interest in this problem is primarily in cases where $\mathcal{X}$ is given implicitly by the extreme points to an H-polytope $\{x : Ax \le b\}$. Following the terminology in Natarajan, Song and Teo (2009), we refer to the model in (5) as the Marginal Distribution Model (MDM).

Another problem for which we show the MDM to be tractable is *max flow*, a fundamental problem in the application of operations research. Here, we are given a directed graph, and the capacity of each arc is random. We are given the marginal distribution of each arc capacity, but not the correlation between them. We would like to know: what is the expected flow that can be pushed from the source to the sink, under the *worst-case joint distribution* over arc capacities? In general, this worst-case distribution can be high-dimensional and thus intractable to represent. Nonetheless, we show that the value of Z^* in this case can still be evaluated as the optimal objective value of a polynomial-sized LP. Furthermore, this LP can be adapted to solve for the optimal locations to set aside excess arc capacity, to be robust against the worst-case joint distribution. For further details, as well as a proof of our main result in this special case which does not require any background in measure theory or convex duality, we refer to Subsection 5.2.

1.1. Related Literature

To the best of our knowledge, the study of problem (5) 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 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 followup 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. Weiss (1986) generalized the result 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. Birge and Maddox (1995) relaxed the assumption on the knowledge of the entire marginal distribution and developed bounds when only the support and the first two moments of the activity durations are known in the PERT network. Bertsimas, Natarajan and Teo (2004) extended this approach to general combinatorial optimization problems with marginal moment information on the random coefficients and showed with the use of semidefinite optimization that the worst-case bound Z^* is computable in polynomial time when the original combinatorial optimization problem is solvable in polynomial time. In a followup paper, Bertsimas, Natarajan and Teo (2006) developed an alternative proof of this result in terms of a primal formulation that is directly defined in terms of the moments of the random variables and the optimal solution mapping instead of using duality techniques as in Bertsimas, Natarajan and Teo (2004). The advantage of the primal formulation is that it provides the “persistency” of the binary variables which is the marginal distribution of the optimal solution.

The study of the bound in (5) for general discrete optimization was initiated in Natarajan, Song and Teo (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 optimization problem almost surely admits a unique solution with absolutely continuous marginal distributions, Natarajan, Song and Teo (2009) showed that the bound Z^* and the persistency distribution 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 linear and discrete optimization problems is not discussed in their work. In the special case of the appointment scheduling problem with marginal moment information, Mak, Rong and Zhang (2015) considered an alternative binary reformulation to that proposed in Natarajan, Song and Teo (2009). Interestingly in this special case, the distributionally robust appointment scheduling problem was shown to be solvable in polynomial time. Natarajan, Shi and Toh (2017) extended these bounds to binary quadratic programs with random objective coefficients and showed that the complexity of computing this bound does not increase substantially with respect to the complexity of solving the corresponding deterministic problem. Natarajan, Song and Teo (2009) also showed that the persistency solution from such an optimization problem can be used in choice modeling applications. Indeed, in the context of choice modeling one can view the Marginal Distribution Model as providing a class of "semiparametric" discrete choice models, obtained through optimizing over a family of joint error distributions with prescribed marginal distributions. Mishra et al. (2014) showed that the family of generalized extreme value choice models in discrete choice can be recovered from such a scheme. Agrawal et al. (2012) provided results for the distributionally robust optimization problem $\min_{s \in \mathcal{S}} \sup_{\theta \in \Gamma(\mu_1,\ldots,\mu_n)} E_{\tilde{\xi} \sim \theta}\left[h(s,\tilde{\xi})\right]$ for general objective functions. 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 this, they defined the "price of correlation" as $POC :=$

$\min_s E_{\tilde{\xi}\sim\mu_1\otimes\ldots\otimes\mu_n}\left[h(s,\tilde{\xi})\right]/\min_s \sup_{\theta\in\Gamma(\mu_1,\ldots,\mu_n)} E_{\tilde{\xi}\sim\theta}\left[h(s,\tilde{\xi})\right]$, which is a measure of the difference in expectation computed under the independent coupling $\mu_1\otimes\mu_2\otimes\ldots\otimes\mu_n$ and the worst-case distribution. 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.

Lastly, the field of optimal transport theory and the related Monge-Kantorovich problem provides several fundamental tools including duality results to tackle such problems. This has been applied in a wide range of areas from physics to engineering to insurance and economics (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. Indeed, the product distribution $\mu\otimes\nu$ is always a feasible plan. In some cases, the optimal coupling is deterministic where each x is associated with a single y and this is referred to as the Monge coupling. Unfortunately, such a Monge coupling many not always exist. For instance, consider the case of μ having total mass at a single point while ν is a probability measure without atoms. Only if one is allowed to "split the mass", can such a transportation be done. While traditionally, optimal transport has been mainly concerned with finding optimal couplings of two distributions, there has lesser work in multi-marginal optimal transport (see Pass (2015), Bach (2018)). 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.

1.2. Our Contributions

We add to the previous literature in the following manner. In Section 2, we show that the problem of computing the tight bound for linear optimization is NP-hard. The primal formulation as a concave maximization problem in Natarajan, Song and Teo (2009) was developed for the case with absolutely continuous marginal distributions. In Section 3, we use results from the field of Optimal Transport to identify this primal problem as more generally part of a primal-dual pairing of optimization problems. In doing so, we generalize the primal formulation result to arbitrary marginal distributions, which admits the possibility of multiple optimal solutions in (4) when, say, the marginals are discrete, and hence causing the optimal solution mapping $x^*(\cdot)$ to become multi-valued. And in our argument for the generalization, we use to great effect the fact that monotone couplings are optimal for supermodular transport costs. Then in Section 4, we explore the dual problem form, which suggests the identification of a set of sufficiency conditions for polynomial time solvability using extended formulations. This generalizes the current known tractability results of Meilijson and Nadas (1979) and Bertsimas, Natarajan and Teo (2004) for 0-1 polytopes and the result of Mak, Rong and Zhang (2015) for the appointment scheduling problem to a class of integer polytopes. We discuss implications of the results on the solvability of distributionally robust optimization problems in areas such as scheduling in Section 5 and provide computational experiments in Section 6.

2. Hardness of the Marginal Distribution Model (MDM)

There are two key inputs to computing the bound Z^* in MDM: (a) the marginal probability measures $\mu_1, \ldots, \mu_n$, and (b) the feasible region of the optimization problem $\mathcal{X}$. We assume that the set $\mathcal{X}$ is defined implicitly by the set of extreme points in a H-polytope denoted as $\{x : Ax \leq b\}$, namely $\mathcal{X} = Extr(\{x : Ax \leq b\})$. While the number of elements in the set $\mathcal{X}$ is typically very large, the H-polytope often provides a compact representation. A natural question is to try and characterize the computational complexity of MDM given the input marginals and the H-polytope representation. While the deterministic problem in solvable in polynomial time as a linear program, we show that the distributionally robust bound is NP-hard to compute.

PROPOSITION 1. *Computing Z^* in MDM for the class of linear optimization problems given discrete marginal distributions and a 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:

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

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

$$\begin{aligned}
\max \ & \sum_{S\subseteq[n]} p_S \max\left\{\sum_{i\in S} x_i - \sum_{i\in S^c} x_i : Ax \leq b\right\} \\
\text{s.t.} \ & \sum_{S\subseteq[n]: S\ni i} p_S = p_i, && \forall i \in [n], \\
& \sum_{S\subseteq[n]} p_S = 1, \\
& p_S \geq 0, && \forall S \subseteq [n],
\end{aligned} \tag{6}$$

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:

$$\begin{aligned}
\min \ & y_0 + \sum_{i\in[n]} p_i y_i \\
\text{s.t.} \ & y_0 + \sum_{i\in S} y_i \geq \max\left\{\sum_{i\in S} x_i - \sum_{i\in S^c} x_i : Ax \leq b\right\}, \ \forall S \subseteq [n].
\end{aligned} \tag{7}$$

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

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

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

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

which is equivalent to

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

The right hand side corresponds to a 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. □

A related NP-hardness result for computing the worst-case bounds in distributionally robust linear optimization problems with a given mean and covariance matrix was shown in Bertsimas et al. (2010). The hardness result was shown by relating it to the NP-hard problem of 2-norm maximization over a polytope. Agrawal et al. (2012) showed that the problem of computing the expected value of a function of binary random variables with fixed marginal probabilities under the worst-case distribution is NP-hard, when the functions are monotone and submodular using a reduction from the MAX-CUT problem. Proposition 1 indicates that the worst-case bound is NP-hard even for the class of linear programs with the Fréchet class of distributions. In the next section, we identify a primal-dual formulation for this problem using which we identify conditions under which MDM is tractable.

3. MDM: Primal-Dual Formulations

In this section, we present a formulation for MDM using a convex-concave saddle function and develop an associated pair of primal and dual formulations.

3.1. Preliminaries

We recall some basic facts and notations related to measure-theoretic concepts that will be used later.

Definition 1. Let $\{\mathcal{X}_i\}_{i=1,\ldots,n}$ be a finite collection of compact measure spaces, with $\mathcal{X}_1 \times \ldots \times \mathcal{X}_n$ denoting their product space with measure γ. Then for any i, define the "i-th projection" $\Pi_i\gamma$ of the measure γ over the product space by:

$$\Pi_i\gamma(A) := \gamma(\mathcal{X}_1 \times \ldots \times \mathcal{X}_{i-1} \times A \times \mathcal{X}_{i+1} \times \ldots \times \mathcal{X}_n) \text{ for all measurable subsets } A \text{ of } \mathcal{X}_i.$$

Let the set of all Borel probability measures on a compact set $\mathcal{X}$ be denoted by $\mathcal{P}(\mathcal{X})$. Given probability measures $\{\nu_i\}_i$, each defined on its own compact probability space $\{\mathcal{X}_i\}_i$, let:

$$\Gamma(\nu_1,\ldots,\nu_n) := \{\gamma \in \mathcal{P}(\mathcal{X}_1 \times \ldots \times \mathcal{X}_n) : \Pi_i \gamma = \nu_i, \ \forall i \in [n]\},$$
$$= \{\gamma \in \mathcal{P}(\mathcal{X}_1 \times \ldots \times \mathcal{X}_n) : \int \sum_{i=1}^n f_i d\gamma = \sum_{i=1}^n \int f_i d\nu_i \ \ f_i \in C^0(\mathcal{X}_i, \mathbb{R}), \ \forall i \in [n]\},$$

where the latter representation exploits the duality between $\mathcal{P}(\mathcal{X}_i)$ and $C^0(\mathcal{X}_i,\mathbb{R})$, and illustrates that $\Gamma(\nu_1,\ldots,\nu_n)$ is weak-* compact. Observe that this set is nonempty because the product measure $\nu_1 \otimes \nu_2 \otimes \ldots \otimes \nu_n$ is trivially a member. If f is a measurable function from $(\mathcal{X},\mathcal{F},\mu)$ into the reals, we will refer to $f\#\mu := \mu \circ f^{-1}$ as the "measure induced by f" on the reals where $E_{x\sim\mu}[f(x)] := \int f(x)\mu(dx)$ and the sigma algebras for all real measure spaces will be the Borel sigma algebra which we'll denote by $\mathcal{B}$. The set $\mathcal{X}$ will always denote a set that is compact, and we will take advantage of the fact that $C(\mathcal{X})$ and the set of radon measures on $\mathcal{X}$ are dual, with a scalar product $< f,\nu > := \int_{\mathcal{X}} f(x) d\mu(x)$. Further, the notion of weak-* topology will refer to the topology we endow upon the set of radon measures on the compact set $\mathcal{X}$, as per this duality relation (see Royden and Fitzpatrick (2010)). Let γ be a measure defined on the product set $\mathbb{R} \times \mathcal{X}$ and $\nu = \Pi_{\mathcal{X}} \# \gamma$. Then the Disintegration theorem tells us that there exists a family of measures on $\mathbb{R}$, indexed by members x in $\mathcal{X}$, denoted by $\{\mu^x\}_{x\in\mathcal{X}}$, such that:

$$\int_{\mathbb{R}\times\mathcal{X}} f \, d\gamma = \int_{\mathcal{X}} \int_{\mathbb{R}} f \, d\mu^x d\nu(x).$$

We will express this "disintegration" of the measure γ as $\gamma = \mu^x \otimes \nu$.

Unlike Natarajan, Song and Teo (2009), where the results were derived under the assumption that the optimization problem almost surely admits a unique solution, we allow for the possibility of multiple optimal solutions. Towards this, we associate with the random optimal objective function $Z(\tilde{c}) := \max\left\{\tilde{c}^T x : x \in \mathcal{X}\right\}$, the optimal solution multifunction (see Shapiro, Dentcheva and Ruszczyński (2014)) $x^{OPT} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ defined as follows:

$$x^{OPT}(\tilde{c}) = \arg\max_{x\in\mathcal{X}} \tilde{c}^T x.$$

This multifunction from $(\mathbb{R}^n, \mathcal{B})$ into $(\mathbb{R}^n, \mathcal{B})$ is closed-valued and measurable. In other words, $x^{OPT}(\tilde{c})$ is closed in $\mathbb{R}^n$ for all $\tilde{c}$, and for every closed set $A \subset \mathbb{R}^n$, $\{\tilde{c} : x^{OPT}(\tilde{c}) \cap A \neq \varnothing\}$ 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 $x^*(\tilde{c}) \in x^{OPT}(\tilde{c})$. The measurable mapping $x^*(\cdot)$ then returns a possible optimal solution to any given vector $\tilde{c}$. Note that the maximization problem indeed has a solution. This follows by the weak-* compactness of the set $\Gamma(\mu_1, \ldots, \mu_n)$ and the continuity of the function $Z(\cdot)$ (which follows by its convexity and it being real-valued everywhere).

3.2. A Lagrangian Formulation

In this section, we derive a Lagrangian formulation of MDM. Towards this, we make use of the following core duality result from optimal transport theory for the Monge-Kantorovich problem.

THEOREM 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 semi-continuous functions $a \in \mathcal{L}^1(\mu)$ and $b \in \mathcal{L}^1(\nu)$ such that $h(x,y) \leq a(x) + b(y)$ for all $(x,y) \in X \times Y$. Then,*

$$\max_{\pi \in \Gamma(\mu,\nu)} \int h d\pi = \inf_{(u,v) \in \Phi(h)} \int u d\mu + \int v d\nu,$$

where $\Phi(h)$ is the set of pairs (u,v) of Borel functions $u : X \to (-\infty, \infty]$ and $v : Y \to (-\infty, \infty]$ such that $u \in \mathcal{L}^1(\mu)$ and $v \in \mathcal{L}^1(\nu)$ satisfy:

$$h(x,y) \leq u(x) + v(y), \text{ for all } (x,y) \in X \times Y.$$

If in addition, the cost function is supermodular, then an explicit and intuitive characterization of the optimal transference plan is known.

LEMMA 1 **(Monotone Coupling Lemma, Theorem 3.1.2 in Rachev and Rüschendorf (1998))**. *Let h: $\mathbb{R}^2 \to \mathbb{R}$ be a right-continuous supermodular function. Then,*

$$\sup_{\gamma \in \Gamma(\mu,\nu)} E_{(x,y)\sim\gamma}[h(x,y)] = \int_0^1 h(F_\mu^{-1}(t), F_\nu^{-1}(t)) dt,$$

$$\inf_{\gamma\in\Gamma(\mu,\nu)} E_{(x,y)\sim\gamma}[h(x,y)] = \int_0^1 h(F_\mu^{-1}(t), F_\nu^{-1}(1-t))dt$$

where $F_\mu(x) := \mu(y \in (-\infty, x])$, *and* $F_\mu^{-1}(t) := inf\{x : F_\mu(x) > t\}$, *with* $F_\mu^{-1}(1) = +\infty$.

From a probabilistic perspective, the lemma says the monotone coupling on the probability space $((0,1), \mathcal{B}, \lambda)$ given by $(X_1(t), \dots X_n(t)) := (F_{\mu_1}^{-1}(t), \dots, F_{\mu_n}^{-1}(t))$ yields the largest expectation.

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

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

This brings us to the first key result.

Theorem 2 **(Lagrange Form)**. *Let* $Z(c) := \max_{x\in\mathcal{X}} c^T x$. *Let there be given n probability measures* $\{\mu_i\}_{i=1}^n$ *over* $\mathbb{R}$, *each with finite variance. And let* $\mathcal{X} \subset \mathbb{R}^n$ *be an arbitrary finite point set. Then*

$$\max_{\theta\in\Gamma(\mu_1,\dots,\mu_n)} E_{\tilde{c}\sim\theta} Z(\tilde{c}) = \max_{\nu\in\mathcal{P}(\mathcal{X})} \sum_i \max_{\gamma_i\in\Gamma(\mu_i,\Pi_i\nu)} E_{(\tilde{c}_i,x_i)\sim\gamma_i} \tilde{c}_i x_i = \max_{\nu\in\mathcal{P}(\mathcal{X})} \inf_{\{\psi_i:\mathcal{X}_i\to\mathbb{R}\}_{i=1}^n} L(\{\psi_i\}_i, \nu) \qquad \text{(Primal)}$$

$$= \min_{\{\psi_i:\mathcal{X}_i\to\mathbb{R}\}_{i=1}^n} \sup_{\nu\in\mathcal{P}(\mathcal{X})} L(\{\psi_i\}_i, \nu) \qquad \text{(Dual)}$$

For an economic interpretation of this max-min, min-max duality, the reader is referred to Appendix section EC.1.

Proof: We refer the reader to the Appendix section EC.2.1

□

Remark 1. All infimum in the theorem can be taken equivalently over all functions $\psi_i : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ that are univariate polyhedrally convex and whose extreme points lie in $\mathcal{X}_i$ (recollect that a function is polyhedral convex if and only if the epigraph $epi(f)$ is polyhedral). To see why this is true, we observe that for any i, we can extend ψ_i to all of $\mathbb{R}$ in a manner that preserves its integral w.r.t. any measure $\nu \in \mathcal{P}(\mathcal{X})$. Letting $x_{i-} := \sup\{\alpha \in \mathcal{X}_i : \alpha < x_i\}$, define the extension via:

$$\psi_i(\xi_i) = \begin{cases} \psi_i(\xi_i), & \xi_i \in \mathcal{X}_i, \\ \lambda\psi_i(x_{i-}) + (1-\lambda)\psi_i(x_i), & \xi_i = \lambda x_{i-} + (1-\lambda)x_i, \text{ for some } x_i \in \mathcal{X}_i, \lambda \in (0,1), \\ +\infty, & \xi_i \notin conv(\mathcal{X}_i). \end{cases}$$

Henceforth it will be assumed without loss of generality that ψ_i is defined on all of $\mathbb{R}$, taking finite values only over the interval $conv(\mathcal{X}_i)$, Next, recall that for any $\tilde{c}_i$,

$$max_{x_i}\{\tilde{c}_i x_i - \psi_i(x_i)\} = \psi_i^*(\tilde{c}_i) = \psi_i^{***}(\tilde{c}_i)$$

where * denotes the Legendre-Fenchel transformation. Since additionally it always holds that $\psi_i^{**} \leq \psi_i$, we see that for all $\nu \in \mathcal{P}(\mathcal{X})$,

$$\sum_i \int \psi_i^{**}(x_i)\ \Pi_i\nu(dx_i) + \int \psi_i^{***}(\tilde{c}_i)\ \mu_i(d\tilde{c}_i) \leq \sum_i \int \psi_i(x_i)\ \Pi_i\nu(dx_i) + \int \psi_i^*(\tilde{c}_i)\ \mu_i(d\tilde{c}_i).$$

In other words, under the goal of minimizing the Lagrangian $L(\{\psi_i\}_i, \nu)$ over ψ_i, with fixed $\nu \in \mathcal{P}(\mathcal{X})$, any proposed solution ψ_i can be modifed to ψ_i^{**}, the largest convex minorant of ψ_i, to yield a solution that does no worse. Since in our case $\psi_i^{**} = \psi_i$ is equivalent to ψ_i being a univariate polyhedral convex function with extreme points in $\mathcal{X}_i$, we can restrict the search, as claimed. △

It turns out that in a sense all the distributional information of $\Pi_i\nu$ is encoded in the subdifferential mapping $\partial\psi_i(\cdot)$. Indeed, the graphs of the subdifferential mappings $\partial\psi_i(\cdot)$ are complete nondecreasing curves in $\mathbb{R}^2$ (see Rockafellar (1997)) and vertical line segments at each $x_i \in \mathcal{X}_i$, not unlike the graph of a cumulative distribution function for a discrete distribution. In the following corollary, we see it is the measure μ_i that translates the nondecreasing curve into a distribution.

Corollary 1 **("KKT/Saddle-point" conditions)**. *Following Remark 1, we restrict attention to $\{\psi_i\}_i$ that are convex. Hence, when the given marginals $\mu_1, \ldots, \mu_n$ are absolutely continuous, $(\bar{\nu}, \{\bar{\psi}_i\}_i)$ forms an optimal primal-dual pair(aka saddle point) if and only if the following conditions are satisfied:*

1. $\Pi_i\bar{\nu}(x_i) = \mu_i(\partial\bar{\psi}_i(x_i)), \quad \forall i, x_i \in \mathcal{X}_i,$
2. $\int_{\mathcal{X}} \sum_i \bar{\psi}_i(x_i) d\nu \leq \int_{\mathcal{X}} \sum_i \bar{\psi}_i(x_i) d\bar{\nu}, \ \ \forall \nu \in \mathcal{P}(\mathcal{X}).$

Furthermore, 1. yields:

$$\Pi_i\bar{\nu}(-\infty, x_i] = \mu_i(-\infty, (\bar{\psi}_i)'_+(x_i)], \ \forall x_i \in \mathcal{X}_i, \ \forall i \in [n].$$

Proof: The KKT conditions in this case are equivalent to the following:

1. $\frac{\partial L}{\partial \psi_i}(\bar{\psi},\bar{\nu}) = 0, \quad \forall i \in [n]$,
2. $\int \frac{\partial L}{\partial \nu}(\bar{\psi},\bar{\nu})\ d(\nu - \bar{\nu}) \leq 0, \quad \forall \nu \in \mathcal{P}(\mathcal{X})$.

Regarding condition 1, by the absolute continuity hypothesis,

$$\int \bar{\psi}_i d\Pi_i \bar{\nu} + \int \bar{\psi}_i{}^* d\mu_i = \int \bar{\psi}_i d\Pi_i \bar{\nu} + \sum_{x_i \in \mathcal{X}_i} \int_{\tilde{c}_i : x_i = \arg\max_{\xi_i \in \mathcal{X}_i} \tilde{c}_i \xi_i - \psi_i(\xi_i)} \tilde{c}_i x_i - \bar{\psi}_i(x_i) d\mu_i, \quad \forall i,$$

so that upon differentiating with respect to $\psi_i(x_i)$ and setting it equal to zero, we arrive at:

$$\Pi_i \bar{\nu}(x_i) + \int_{\tilde{c}_i \in \partial \bar{\psi}_i(x_i)} -1 d\mu_i = 0.$$

Similarly, for condition 2,

$$\frac{\partial L(\bar{\psi},\bar{\nu})}{\partial \nu(x)} = \sum_i \psi_i(x_i),$$

as desired.

Remark 2 (On the Absolute Continuity Assumption). The absolute continuity assumption yields the following fact: for any i, the set of $\tilde{c}_i$ such that more than one member of $\mathcal{X}_i$ solves the expression $\max_{\xi_i \in \mathcal{X}_i} \tilde{c}_i \xi_i - \bar{\psi}_i(\xi_i)$ has μ_i probability measure zero. Equivalently, $\tilde{c}_i$ lies in precisely one and only one of the set of subgradients from the collection of $\{\partial \bar{\psi}_i(x_i)\}_{x_i \in \mathcal{X}_i}$, with μ_i probability measure 1. In the case of an arbitrary probability measure μ_i, however, there may exist $x_i \in \mathcal{X}_i$ s.t. the singleton $\partial \bar{\psi}_i(x_i) \cap \partial \bar{\psi}_i(x_{i-})$ has positive μ_i measure. In Theorem 3 in the next subsection, the possible existence of such atoms will in general preclude the existence of Monge transport maps, i.e., measurable solution mappings $x^*(\cdot)$. However, as seen in the proof, a Kantorovich transport map exists, if we perform careful splitting. △

Remark 3 (Extent of Uniqueness to the dual and primal variables). As we have seen, the dual optimal ψ_i are unique up to an additive constant - they are identified by their subdifferential mappings. In the next section, we will find the optimal primal variables exhibit a similar uniqueness quality, whereby all optimal primal solutions are identified by specific marginals.

3.3. The MDM Primal Problem

We now generalize Theorem 1 in Natarajan, Song and Teo (2009) to arbitrary marginal distributions. In order to state the theorem, we introduce the following notation. Let $x_i - e_i$, where $x_i \in \mathcal{X}_i$, 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.

THEOREM 3 **(Maximization Form)**. *Let $\mathcal{X}$ be a finite set. Let $\tilde{c}_1, \dots, \tilde{c}_n$ be real-valued random variables with $\mu_1, \dots, \mu_n$ denoting the probability measures they induce on the real line. Then,*

$$\max_{\theta \in \Gamma(\mu_1,\dots,\mu_n)} E_{\tilde{c} \sim \theta} Z(\tilde{c}) = \max_{\nu \in \mathcal{P}(\mathcal{X})} \sum_i \max_{\gamma_i \in \Gamma(\mu_i, \Pi_i \nu)} E_{(\tilde{c}_i, x_i) \sim \gamma_i} \tilde{c}_i x_i \tag{Primal}$$

$$= \max_{\nu \in \mathcal{P}(\mathcal{X})} \sum_i \sum_{x_i \in \mathcal{X}_i} x_i \int_{\Pi_i \nu((-\infty, x_i - e_i])}^{\Pi_i \nu((-\infty, x_i])} F_{\mu_i}^{-1}(t) dt. \tag{8}$$

If ν_{rel} solves the concave maximization problem (8), there exists a maximizing joint distribution $\theta^* \in \Gamma(\mu_1, \dots, \mu_n)$. Further, if $\mu_1, \dots, \mu_n$ are absolutely continuous w.r.t. the Lebesgue measure, there exists some suitably defined measurable selection $x^* : \mathbb{R}^n \to \mathcal{X}$ of x^{OPT} s.t. $x^*(c) = (x_1^*(c_1), \dots, x_n^*(c_n))$ and the "persistence values" $P_{\tilde{c} \sim \theta^*}(x_i^*(\tilde{c}) = x_i)$ are given by:

$$P_{\tilde{c} \sim \theta^*}(x_i^*(\tilde{c}_i) = x_i) = \Pi_i \nu_{rel}(x_i), \ \forall x_i \in \mathcal{X}_i, \ \forall i \in [n],$$

i.e., the x_i^* are Monge transport maps:

$$x_i^* \# \mu_i = \Pi_i \nu_{rel}, \ \forall i \in [n].$$

Proof: We refer the reader to the Appendix section EC.2.2.

□

REMARK 4 (COMPARISON TO NATARAJAN, SONG AND TEO (2009)). Theorem 3 is a generalization of Theorem 1 in Natarajan, Song and Teo (2009) in the following two respects. First, Theorem 3 is derived without any assumptions on the given marginals such as absolute continuity. So, for example, discrete marginals are permissible. However, the absolute continuity assumption does afford us the explicit characterization of a transport map that is at the same time an optimal solution mapping. Second, the constraint set $\mathcal{X}$ might be an arbitrary set of points instead of just integers as in Natarajan, Song and Teo (2009). △

In the next corollary, we show the uniqueness of the marginal optimal solutions under the assumption of absolutely continuous marginals.

Corollary 2 (**Uniqueness of ν**). *Assume $\mu_1,\dots,\mu_n$ are absolutely continuous w.r.t. the Lebesgue measure. If ν and τ both solve (Primal), then $\Pi_i\nu=\Pi_i\tau$ for all i.*

Proof: This immediately follows from Corollary 1. Below is an alternative argument.

Let ψ_i be an optimal dual solution. By Theorem 3, let x_i^* and y_i^* be such that $x_i^*\#\mu_i=\Pi_i\nu$ and $y_i^*\#\mu_i=\Pi_i\tau$. By Corollary 1,

$$x_i^*(\tilde c_i)=x_i \iff F_{\mu_i}^{-1}(\Pi_i\nu((-\infty,x_{i-}]))\le\tilde c_i\le F_{\mu_i}^{-1}(\Pi_i\nu((-\infty,x_i])),$$
$$\iff \tilde c_i\in\partial\psi_i(x_i),$$

and similarly for y_i^*. Hence,

$$\tilde c_i x_i^*(\tilde c_i)-\psi_i(x_i^*(\tilde c_i))=\tilde c_i y_i^*(\tilde c_i)-\psi_i(y_i^*(\tilde c_i)),\ \forall\tilde c_i,$$

which means that upon differentiating with respect to $\tilde c_i$, we get:

$$x_i^*(\tilde c_i)+\tilde c_i(x_i^*)'_+(\tilde c_i)-(\psi_i)'_+(x_i^*(\tilde c_i))\cdot(x_i^*)'_+(\tilde c_i)=y_i^*(\tilde c_i)+\tilde c_i(y_i^*)'_+(\tilde c_i)-(\psi_i)'_+(y_i^*(\tilde c_i))\cdot(y_i^*)'_+(\tilde c_i).$$

This implies that:

$$x_i^*(\tilde c_i)=y_i^*(\tilde c_i),\ \mu_i-a.s.,$$

because the derivatives of x_i^* and y_i^* are both zero except at places that altogether constitute a set of μ_i-zero measure. Therefore, $\Pi_i\nu=x_i^*\#\mu_i=y_i^*\#\mu_i=\Pi_i\tau$, as desired.

□

Remark 5. This corollary can be seen as a justification of the use of MDM as a choice model in Natarajan, Song and Teo (2009). Indeed, in Natarajan, Song and Teo (2009), $\mathcal{X}:=\{x:\sum_i x_i=1,\ \ x\in\{0,1\}\}$, so for any $\nu\in\mathcal{P}(\mathcal{X})$, $\Pi_i\nu$ yields the probability that option i is included in a choice. So this corollary shows that under the assumption of absolutely continuous marginals, all primal optimal probability measures over $\mathcal{X}$ yield the same choice probabilities. △

4. The MDM Dual Problem: Sufficient Conditions for Tractability

In this section, we exploit the dual formulation of MDM to identify conditions under which the problem is tractable for the class of linear and discrete optimization problems. Note that from Proposition 1, the bound on the expected value for the class of linear optimization problems is NP-hard to compute though the deterministic problem is efficiently solvable. Thus, our next best goal is to find specific instances under which the bound is computable in polynomial time. We make the following assumption throughout this section to aid in the analysis.

ASSUMPTION 1. *Given the marginals $\mu_1, \dots, \mu_n$ as input, we assume that the expected value of univariate piecewise linear convex functions of the form $\int \max_{x_i \in \mathcal{X}_i} \tilde{c}_i x_i - \psi_i(x_i) d\mu_i$ are efficiently computable for all $i \in [n]$. Furthermore, the subgradients of the functions are efficiently computable.*

The dual formulation of MDM is:

$$\max_{\theta \in \Gamma(\mu_1, \dots, \mu_n)} E_{\tilde{c} \sim \theta}[Z(\tilde{c})] = \min_{\{\psi_i : \mathcal{X}_i \to \mathbb{R}\}_{i=1}^n} \max_{\nu \in \mathcal{P}(\mathcal{X})} \sum_i \left(\int \psi_i d\Pi_i \nu + \int \psi_i^*(\tilde{c}_i) d\mu_i \right). \tag{9}$$

Observe that this assumption is mild, includes the case when all marginals have finite support, which is the case in our numerical experiments section 6.

Given the assumptions, the term $\int \psi_i^*(\tilde{c}_i) d\mu_i$ is can be computed tractably since ψ_i^* is a polyhedral convex function and the integral is a univariate expectation. The key to analyzing the computational complexity is the complexity of the inner maximization problem over probability measures of $\mathcal{X}$:

$$\max_{\nu \in \mathcal{P}(\mathcal{X})} \sum_{x \in \mathcal{X}} \left[\sum_{i=1}^n \psi_i(x_i) \right] \nu(x) = \max_{\nu \in \mathcal{P}(\mathcal{X})} \sum_{i=1}^n \sum_{x_i \in \mathcal{X}_i} \psi_i(x_i) \cdot \Pi_i \nu(x_i), \tag{*}$$

which depends on the characterization of $\mathcal{P}(\mathcal{X})$. In this section, we identify sufficient conditions under which this is tractable.

In deriving the sufficient conditions, we begin by considering the input set $\mathcal{X}$ given as the set of vertices to a polytope (V-polytope) before going on to study, in increasing generality, when it is given as the extreme points to a half-space representation of a polytope (H-polytope). We will then comment on other specific cases and discuss the sharpness of the derived sufficiency conditions.

4.1. $\mathcal{X}$ as V-polytope

We start with the simple case when $\mathcal{X}$ is explicitly provided as a V-polytope. In this case, if the size of the set $|\mathcal{X}|$ is polynomially-bounded, the left-hand side of (*) is an efficiently-sized linear optimization problem for a given set of values $\{\psi_i(x_i)\}$:

$$\max_{\nu\in\mathcal{P}(\mathcal{X})}\sum_{x\in\mathcal{X}}\left[\sum_{i=1}^{n}\psi_i(x_i)\right]\nu(x) = \max_{x\in\mathcal{X}}\sum_{i=1}^{n}\psi_i(x_i),$$
$$= \min\left\{y : y \geq \sum_{i=1}^{n}\psi_i(x_i),\ \forall x\in\mathcal{X}\right\}.$$

Substituting into the dual form in equation (9), implies that the bound Z^* is efficiently solvable as a convex minimization problem. Such a result is also discussed in Theorem 3 in Meilijson (1991) and in special csaes such as $Z(\tilde{c}) = \max_i \tilde{c}_i$ by Lai and Robbins (1976). However, in general, this set can be very large which will be our focus of interest in the remaining part of this section.

4.2. $\mathcal{X}$ as extreme points to a 0-1 H-polytope

Let $P\subset\mathbb{R}^n$ be a polytope, whose set of extreme points are given by $\mathcal{X}$. Naturally, any element $y\in P$ can be written as some convex combination of finitely many members from $\mathcal{X}$, i.e., $y=\sum_i\lambda_i x^i = E_{\tilde{y}}[\tilde{y}]$, with $x^i\in\mathcal{X}$, $\sum_i\lambda_i = 1$, $\lambda_i\geq 0$, and the expectation is taken with respect to the discrete probability distribution dictated by the λ_i. This representation of y is unique if and only if the set of vectors in $\mathcal{X}$ is affinely independent, in which case $P\,(=conv(\mathcal{X}))$ and $\mathcal{P}(\mathcal{X})$ are in bijection in such a setting. If, however, $\mathcal{X}$ is not a set of affine independent vectors, then there is no longer a bijection. In this case, there exists a surjective function $f:\mathcal{P}(\mathcal{X})\to P$ that is not injective. For any $y\in P$ let us write $[y]=f^{-1}(y)\subset\mathcal{P}(\mathcal{X})$ for the set of probability measures that correspond to the same $y\in P$.

Now, let us consider the special case when P is in fact specified as a 0-1 H-polytope, i.e., $\mathcal{X}$ consists of integral 0-1 vectors. The set of extreme points may not be affine independent. But, if we have two probability measures $\nu_1,\nu_2\in[y]$, then $y_i=\Pi_i\nu_1(\{1\})=\Pi_i\nu_2(\{1\})\in[0,1]$. In other words, $\{\{\Pi_i\nu\}_{i=1}^n : \nu\in\mathcal{P}(\mathcal{X})\}$ and P are in bijection. In this case, the right-hand side of (*) becomes a linear program over the 0-1 H-polytope, implying that the bound Z^* is efficiently solvable as a convex minimization problem. This tractability result was developed for PERT networks in Meilijson and Nadas (1979), shortest path, maximum flow and network reliability problems in Weiss (1986) and for general combinatorial optimization problems in Bertsimas et al. (2010), Bertsimas, Natarajan and Teo (2004, 2006).

4.3. $\mathcal{X}$ as extreme points to a general H-polytope

In this section we consider the case where the set $\mathcal{X}$ is defined implicitly as the extreme points to a H-polytope. However we do not make the assumption that the extreme points are binary vectors. As discussed in Proposition 1, Z^* is NP-hard to compute in this case. We next identify conditions under which this bound is computationally tractable.

Theorem 4. *Under assumption 1, if there exists a compact extended formulation of conv($\mathcal{X}$) of the form:*

$$conv(\mathcal{X}) = \Pi_x\left(\left\{(x,y): y \in P \subset \mathbb{R}^B,\ x_i = \sum_{\bar{x}_i \in \mathcal{X}_i} \bar{x}_i \sum_{j=1}^{n_{\bar{x}_i}} y_{F_j^{\bar{x}_i}} \ for\ \{F_j^{\bar{x}_i}\} \in B,\ \forall j = 1,\dots,n_{\bar{x}_i},\ \forall i \in [n]\right\}\right), \tag{Ext. Form}$$

where $n_{\bar{x}_i}$ is a finite integer for each i, $\bar{x}_i \in \mathcal{X}_i$ and $P \subset \mathbb{R}^B$ is a 0-1 polytope of the form:

$$P \subseteq \left\{ y \in [0,1]^B : \sum_{\bar{x}_i \in \mathcal{X}_i} \sum_{j=1}^{n_{\bar{x}_i}} y_{F_j^{\bar{x}_i}} = 1,\ \forall i \in [n] \right\},$$

then Z^* is efficiently computable.

Proof: Assume there exists such a compact extended formulation. Then based on the previous subsection, with P being a 0-1 polytope, $\{\{\Pi_i \nu\}_{i=1}^n : \nu \in \mathcal{P}(\mathcal{X})\}$ and P are in bijection, with $\Pi_i \nu(\bar{x}_i) = \sum_{j=1}^{n_{\bar{x}_i}} y_{F_j^{\bar{x}_i}}$. In summary, problem (*) becomes a linear program that is efficiently solvable:

$$\max_{\nu \in \mathcal{P}(\mathcal{X})} \sum_{i=1}^n \sum_{x_i \in \mathcal{X}_i} \psi_i(x_i) \cdot \Pi_i \nu(x_i) = \max_{y \in P} \sum_{i=1}^n \sum_{x_i \in \mathcal{X}_i} \psi_i(x_i) \cdot \sum_{j=1}^{n_{x_i}} y_{F_j^{x_i}}\ .$$

Applying the dual of this linear program into formulation (9), we see that Z^* is efficiently computable.

□

Regarding the notation in Theorem 4's extended formulation, the reader may interpret B as a family of "events". With the proper specification of B, Theorem 4 can explain multiple types of problems. For instance, we can bring the V-polytope representation case into this framework.

COROLLARY 3. *Let $\tilde{c}_1,\dots,\tilde{c}_n$ be real-valued random variables with $\mu_1,\dots,\mu_n$ denoting the probability measures they induce on the real line. If $\mathcal{X}$ is given in V-polytope representation with compact size, then $conv(\mathcal{X})$ admits a compact extended formulation of the kind in Theorem 4, and consequently Z^* is efficiently computable.*

Proof: To see this, we exploit the fact that every V-polytope is the projection of a simplex. Let A be the $n\times m$ matrix given by appending the elements of $\mathcal{X}$ together as columns, i.e., $A = (x^1\ x^2\ \dots\ x^m)$. Then,

$$conv(\mathcal{X}) = \Pi_x\left(\{(x,y) : y\in\Delta_{m-1}, x = Ay\}\right),$$

where

$$\Delta_{m-1} := \{y\in\mathbb{R}^m : y\geq 0, \sum_i y_i = 1\}$$

where for each $\bar{x}_i\in\mathcal{X}_i$ for some i, we can identify the set of indices $J_{\bar{x}_i} := \{j : x_i^j = \bar{x}_i\}$ as $\{F_j^{\bar{x}_i}\}_{j=1}^{n_{\bar{x}_i}}$.

□

Similarly, we can also provide the sufficient compact extended formulations to 0, 1 H-polytope and $L^\natural$ Convex Polytopes.

4.3.1. 0, 1 H-Polytopes

COROLLARY 4 **(0,1 H-Polytopes)**. *Let $\tilde{c}_1,\dots,\tilde{c}_n$ be real-valued random variables with $\mu_1,\dots,\mu_n$ denoting the probability measures they induce on the real line. If $\mathcal{X}$ is the set of extreme points to a compact-sized 0-1 H-polytope, then $conv(\mathcal{X})$ admits a compact extended formulation of the kind in Theorem 4, and consequently Z^* is efficiently computable.*

Indeed, there is an analogous result identified for the case when only the marginal moments are known (see Bertsimas, Natarajan and Teo (2006)). In the next section, we discuss examples of problems that satisfy this condition for general linear and discrete optimization problems.

4.3.2. $L^\natural$ **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

$$\mathcal{X} = \{x \in \mathbb{Z}^n : x_i - x_j \le d_{ij} \;\; \forall (i,j) \in \mathcal{A},\; l_i \le x_i \le u_i \;\; \forall i \in \{1,\dots,n\}\}$$

A set of this form is known as an $L^\natural$-convex set. Its convex hull $conv(\mathcal{X})$ is an integral polyhedron that when bounded we'll identify as an $L^\natural$ -convex polytope.

When assuming $conv(\mathcal{X})$ is bounded, observe that for any j,

$$\underline{l}_j := \min_{i:(i,j)\in\mathcal{A}} \{l_i - d_{ij}\} > -\infty$$

$$\bar{u}_j := \min_{j:(i,j)\in\mathcal{A}} \{u_j - d_{ij}\} < +\infty,$$

so without loss of generality, let us assume that $\{l_i\}_{i=1}^n \subset \mathbb{Z}$, $\{u_i\}_{i=1}^n \subset \mathbb{Z}$.

Theorem 5. *Let $\mathcal{X}$ be a bounded $L^\natural$- convex set. Then $conv(\mathcal{X})$ admits an integral extended form as in Theorem 4. In particular, under Assumption 1, the MDM problem with $\mathcal{X}$ as input can be solved efficiently.*

Proof: Define $u := \max_i u_i$ and $l := \min_i l_i$. We claim that

$$x_i - x_j \le d_{ij} \;\; \forall (i,j) \in \mathcal{A} \iff [x_i \in \{t,\dots,u\} \implies x_j \notin \{0,\dots,t - d_{ij} - 1\}] \;\; \forall (i,j) \in \mathcal{A}.$$

The "only if" direction is clear. The "if" direction is also clear, as given an $(i,j) \in \mathcal{A}$, $x_i \in \{x_i,\dots,u\} \implies x_j \notin \{0,\dots,x_i - d_{ij} - 1\}$, by hypothesis. Hence, $x_j \ge x_i - d_{ij}$, as desired. Thus, the following 0-1 polytope provides an extended form

$$P := \{y \in \mathbb{R}_+^{n\times[l,u]_\mathbb{Z}} : \sum_{x_i\in[l_i,u_i]_\mathbb{Z}} y_{i,x_i} = 1 \;\; \forall i,\; \sum_{s=t}^{u} y_{i,s} + \sum_{s=0}^{t-d_{ij}-1} y_{js} \le 1 \;\; \forall (i,j) \in \mathcal{A}, t = l, l+1,\dots,u\}$$

□

4.4. Interpreting the Compact Extended Formulation

In this section, we attempt to further shed some light on how the compact extended formulation of Theorem 4 captures the probability and measure-theoretic features of $\mathcal{P}(\mathcal{X})$. Consider a pair $(\mathcal{X}',\mathcal{F})$, where $\mathcal{X}'$ is a finite point set, and $\mathcal{F}$ is a set of "events." Suppose we find a surjective function $f:\mathcal{X}'\to\mathcal{X}$, written as $f(\xi)=(f_1(\xi),\dots,f_n(\xi))$, such that each f_i is $\mathcal{F}$-measurable, meaning $f_i^{-1}(x_i)\in\mathcal{F}$ for all i and $x_i\in\mathcal{X}_i$. Then, f can be thought of as analogous to a discrete random vector taking values in $\mathcal{X}$. The probability law of the random vector f is determined by a probability measure ν' on the set system $\mathcal{F}\subset 2^{\mathcal{X}'}$. Since f is surjective, we have:

$$(\forall \nu\in\mathcal{P}(\mathcal{X}))\,(\exists\nu'\in\mathcal{P}(\mathcal{X}'))\quad \nu=f\#\nu'.$$

Next, as $\mathcal{F}$ represents a $\sigma-$algebra, which is a semiring, there exists a collection B of pairwise disjoint sets such that all $f_i^{-1}(x_i)\in\mathcal{F}$ admit finite expansions with respect to the sets in B (see Lemma 2, Page 33 of Kolmogorov and Fomin (1970)). If there exists a 0-1 polytope $P\subset\mathbb{R}^B$, where $B\subset\mathcal{F}$, and $Extr(P)$ are in bijection with $\mathcal{X}'$, then $\{\{\nu'(F)\}_{F\in B}:\nu'\in\mathcal{P}(\mathcal{X}')\}$ is in bijection with P. If this polytope is compact-sized, then (*) admits a tractable optimization form:

$$\max_{\nu\in\mathcal{P}(\mathcal{X})}\sum_{i=1}^{n}\sum_{x_i\in\mathcal{X}_i}\psi_i(x_i)\cdot\Pi_i\nu(x_i)=\max_{\nu'\in\mathcal{P}(\mathcal{X}')}\sum_{i=1}^{n}\sum_{x_i\in\mathcal{X}_i}\psi_i(x_i)\cdot\nu'\left(f_i^{-1}(x_i)\right)=\max_{y\in P}\sum_{i=1}^{n}\sum_{x_i\in\mathcal{X}_i}\psi_i(x_i)\cdot\sum_{j=1}^{n_{x_i}}y_{F_j^{x_i}}.$$

We formalize this discussion in the following result.

THEOREM 6. *Let measures $\mu_1,\dots,\mu_n$ be given and Assumption 1 holds. Let $\mathcal{X}$ be given as the set of extreme points to a V- or H-polytope. If the following conditions hold:*

1. *The projection sets $\mathcal{X}_i$ are polynomial-sized.*
2. *There exists a set $\mathcal{X}'$, a set system $\mathcal{F}\subseteq 2^{\mathcal{X}'}$, and a surjective function $f:\mathcal{X}'\to\mathcal{X}$ given by $f(\xi)=(f_1(\xi),f_2(\xi),\dots,f_n(\xi))^\top$. Further, the sets $f_i^{-1}(x_i)$ are all members in $\mathcal{F}$ and each set can be efficiently identified/enumerated.*
3. *There exists an integral 0-1 polytope $P\subseteq\mathbb{R}^B$ $(B\subseteq\mathcal{F})$ s.t.*
   (a) *B is a finite system of pairwise disjoint sets*

(b) *There exists a bijective mapping between Ext(P) and $\mathcal{X}'$.*

(c) *All $f_i^{-1}(x_i)$ have a finite expansion with respect to the sets in B, i.e., for each $f_i^{-1}(x_i)$, there exists n_{x_i} such that there exists the representation $f_i^{-1}(x_i) = \cup_{j=1}^{n_{x_i}} F_j^{x_i}$, where $F_1^{x_i}, \ldots, F_{n_{x_i}}^{x_i} \in B$ are pairwise disjoint.*

(d) *P has compact description in variables and constraints*

then MDM is solvable in time polynomial in the input size.

Clearly, the existence of the compact extended formulation models the action of a surjective random vector, with abstract probability space $\mathcal{X}'$. The following result establishes that the reverse is in fact true as well, showing the two conditions are equivalent.

Corollary 5. *Let $\mathcal{X}$ be the set of extreme points to a polytope. If there exists a set $\mathcal{X}'$, a surjective function $f:\mathcal{X}'\to\mathcal{X}$, and an integral 0-1 polytope P that satisfy the sufficiency conditions in Theorem 6, then the polytope $conv(\mathcal{X})$ admits a compact extended formulation.*

Proof: Let $x \in conv(\mathcal{X})$. Then there exists a $\nu \in \mathcal{P}(\mathcal{X})$ s.t. $x = E_{\tilde{x}\sim\nu}[\tilde{x}]$. By the surjectivity of f, there exists a $\nu' \in \mathcal{P}(\mathcal{X}')$ s.t. $f\#\nu' = \nu$. To this ν' corresponds a $y \in P$, for which $\nu'(F) = y_F, \forall F \in B$; in particular, the marginals $\Pi_i\nu$ can be obtained via $\Pi_i\nu(x_i) = \nu' \circ f_i^{-1}(x_i) = \sum_{j=1}^{n_{x_i}} y_{F_j^{x_i}}, \forall i, \forall x_i \in \mathcal{X}_i$. Thus, for any i, $x_i = E_{\tilde{x}_i\sim\Pi_i\nu}[\tilde{x}_i] = \sum_{\bar{x}_i\in\mathcal{X}_i} \bar{x}_i \sum_{j=1}^{n_{\bar{x}_i}} y_{F_j^{\bar{x}_i}}$. In other words, we have:

$$conv(\mathcal{X}) \subset \Pi_x\left(\left\{(x,y) : y \in P, x_i = \sum_{\bar{x}_i\in\mathcal{X}_i} \bar{x}_i \sum_{j=1}^{n_{\bar{x}_i}} y_{F_j^{\bar{x}_i}} \;\; \forall i \in [n]\right\}\right).$$

For the reverse inclusion, let $y \in P$. Then by the sufficiency conditions, there exists a corresponding collection $\{\nu'(F)\}_{F\in B}$, with $\nu' \in \mathcal{P}(\mathcal{X}')$. Defining $\nu := f\#\nu'$, we find that:

$$E_{\tilde{x}\sim\nu\in\mathcal{P}(\mathcal{X})}\begin{bmatrix}\tilde{x}_1\\ \tilde{x}_2\\ \tilde{x}_3\\ \vdots\\ \tilde{x}_n\end{bmatrix} = E_{\xi\sim\nu'}\begin{bmatrix}f_1(\xi)\\ f_2(\xi)\\ f_3(\xi)\\ \vdots\\ f_n(\xi)\end{bmatrix} = \begin{bmatrix}\sum_{\bar{x}_1\in\mathcal{X}_1}\bar{x}_1\sum_{j=1}^{n_{\bar{x}_1}} y_{F_j^{\bar{x}_1}}\\ \sum_{\bar{x}_2\in\mathcal{X}_2}\bar{x}_2\sum_{j=1}^{n_{\bar{x}_2}} y_{F_j^{\bar{x}_2}}\\ \sum_{\bar{x}_3\in\mathcal{X}_3}\bar{x}_3\sum_{j=1}^{n_{\bar{x}_3}} y_{F_j^{\bar{x}_3}}\\ \vdots\\ \sum_{\bar{x}_n\in\mathcal{X}_n}\bar{x}_n\sum_{j=1}^{n_{\bar{x}_n}} y_{F_j^{\bar{x}_n}}\end{bmatrix}.$$

Hence, defining $x := E_{\tilde{x}\sim\nu}[\tilde{x}]$, we see that (x,y) belongs to the extended formulation, with $x \in conv(\mathcal{X})$, as desired.

□

4.5. Sharpness of the Sufficiency Condition

A natural question is whether the sufficiency conditions are sharp. In other words, to what extent are they also necessary? In this section, we show that the conditions are close to being sharp by focusing on a class of polytopes - parallelotopes. A parallelotope is a polytope of the form $c + \sum_{i=1}^{m}[-x^i, x^i]$, where $x^1, \ldots, x^m$ are linearly independent vectors. Let us consider here the parallelotopes of the form $Q = \sum_{i=1}^{m}[-x^i, x^i]$, in which the vectors take values $x^i \in \{-1,0,1\}^n$. Setting $\mathcal{X} := Extr(Q)$, we can immediately see from the representation:

$$conv(\mathcal{X}) = Z = \Pi_x \left\{ (x, \epsilon) : x = \sum_i \epsilon_i x^i, -1 \le \epsilon_i \le 1, \forall i \right\},$$

when Q has 2^m extreme points in general.

We now turn to the alternative expression in (*). All x^i belong to $\{-1,0,1\}^n$ implies that $\mathcal{X}_i \subseteq \{-m, -m+1, \ldots, m\}$. Furthermore, if we introduce the set $\mathcal{X}' := \{0,1\}^n$ and the function $f : \mathcal{X}' \to \mathcal{X}$, where $f(\xi) = 2[x^1 \ldots x^m]\xi - (x^1 + x^2 + \ldots + x^m)$, then we see that there exists a bijection between the two sets $\mathcal{X}'$ and $\mathcal{X}$. Furthermore, $\mathcal{X}'$ is a 0/1 polytope of compact size, and sufficiency condition 3 is satisfied. However, while the size of the sets $\mathcal{X}_i$ are polynomially bounded, explicitly characterizing the sets $f_i^{-1}(x_i)$ will require a combinatorial search in general. Thus, this perhaps natural attempt at a combination of $\mathcal{X}'$ and f fail to satisfy the sufficiency conditions. Interestingly, where the set $\mathcal{X}$ is defined as the extreme points to a parallelotope leads to a hard problem.

Corollary 6. *Computing the tight bound Z^* for linear optimization problems over a parallelotope of the form $Q = \sum_{i=1}^{m}[-x^i, x^i]$, in which all $x^i \in \{-1,0,1\}^n$ is NP-hard.*

Proof: Theorem 12 in Bodlaender et al. (1990) shows that the 1-norm maximization over parallelotopes of the form $\sum_{i=1}^{m}[-x^i, x^i]$, in which all $x^i \in \{-1,0,1\}^n$ is NP-hard. Applying the same reduction argument as in Proposition 2 leads to the hardness result. □

5. Polynomial Time Solvable Instances

In this section, we illustrate the application of the dual formulation to identify four polynomial time solvable instances of MDM. The first example is the distributionally robust appointment scheduling which was introduced in Section 1. The polynomial time solvability of this problem under marginal moment information was first shown in Mak, Rong and Zhang (2015). We show that this extends to the case where the entire marginal distributions are known. The second example is the Max Flow problem with random arc capacities. The third example is a ranking problem with random utilities with applications to allocating resources in a scheduling context. The fourth example deals with finding bounds for project scheduling problems with irregular random starting time costs.

5.1. Appointment Scheduling

Recall that in this setting, we have n patients who arrive in a fixed order $\{1,2,\dots,n\}$ who 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 amount of service times scheduled for each patient i, denoted by s_i. Patient 1 arrives at the normalized time 0 while we instruct patient 2 to arrive at time s_1, patient 3 to arrive at time s_1+s_2, and so on. Let $\mathcal{S}=\{s\in\mathbb{R}^n:\sum_{i=1}^n s_i\leq T,\ s_i\geq 0\ \forall i\in[n]\}$, where we want to schedule all patients before time T.

We are interested in the problem of minimizing the worst-case (over all distributions consistent with the marginals) expected total wait time of the n patients to be seen by one doctor and any overtime:

$$\min_{s\in\mathcal{S}}\ \sup_{\theta\in\Gamma(\mu_1,\dots,\mu_n)} E_{\tilde{c}\sim\theta}\left[Z(s,\tilde{c})\right],$$

where:

$$\begin{aligned}
Z(s,\tilde{c}) = \max\ & \sum_{i=1}^n(\tilde{c}_i-s_i)x_i\\
\text{s.t. }\ & x_i-x_{i-1}\geq -1,\ \forall i=2,\dots,n,\\
& x_n\leq 1,\\
& x_i\geq 0,\qquad\qquad \forall i=1\dots,n.
\end{aligned}$$

5.1.1. Exploiting Supermodularity in MDM Appointment Scheduling It is perhaps immediate to the reader that in some sense the "worst" possible realization of the patient processing times $(\tilde{c}_1, \ldots, \tilde{c}_n)$ occurs when all these values assume their largest (or nearly largest, in case of no upper-bounded marginal support) possible value simultaneously. In transferring this intuition to imagining a worst-case coupling of patient processing times, we could propose the monotone coupling of the $\tilde{c}_1, \ldots, \tilde{c}_n$ random variables. Indeed, this turns out to be correct, and this is so because supermodularity, yet again, is part of the picture.

Proposition 2. *For any s, $Z(s, \tilde{c})$ is supermodular in the $\tilde{c}$ variables.*

Proof: We refer the reader to the Appendix section EC.3.1. □

The following result, the technical proof of which is relegated to the Appendix section EC.3.2, establishes the aforementioned intuition that for all s,

$$\sup_{\gamma \in \Gamma(\mu_1, \ldots, \mu_n)} E_{\tilde{c} \sim \gamma}[Z(s, \tilde{c})]$$

is solved by the monotone coupling. Whereas this result is known for the two-marginals (n = 2) case, as in Lemma 1, a multi-marginal $(n \geq 3)$ generalization is now established here.

Lemma 2. *Let $\mathcal{X}_i$ be a totally ordered set for all i, and consider $\mathcal{X} = \mathcal{X}_1 \ldots \mathcal{X}_n$ Let $c : \mathcal{X} \to \mathbb{R}$ be a continuous, submodular function. If $\{\mu_i\}_{i=1\ldots n}$ is any collection of marginal measures, where $\mu_i \in \mathcal{P}(\mathcal{X}_i)$ for all i,*

$$\inf_{\gamma:\ \Pi_i \gamma = \mu_i\ (\forall i)} \int_{\mathcal{X} := \mathcal{X}_1 \times \ldots \times \mathcal{X}_n} c(x)\ d\gamma(x) = \int_0^1 c(F_{\mu_1}^{-1}(t), \ldots, F_{\mu_n}^{-1}(t))\, dt,$$

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

5.1.2. Exploiting Theorem 2 in MDM Appointment Scheduling To properly account for the effect of the scheduling decision variables s_i in the inner MDM problem, observe that

$$\begin{aligned}
\max_{\theta \in \Gamma(\mu_1, \ldots, \mu_n)} E_{\tilde{c} \sim \theta}[Z(s, \tilde{c})] &= \max_{\theta' \in \Gamma(\mu_1 - s_1, \ldots, \mu_n - s_n)} E_{\tilde{c}' \sim \theta'}[Z(0, \tilde{c}')] \\
&= \min_{\{\psi_i : \mathcal{X}_i \to \mathbb{R}\}_{i=1}^n} \max_{\nu \in \mathcal{P}(\mathcal{X})} \sum_i \left(\int \psi_i d\Pi_i \nu + \int \psi_i^*(\tilde{c}_i') d(\mu_i - s_i) \right) \\
&= \min_{\{\psi_i : \mathcal{X}_i \to \mathbb{R}\}_{i=1}^n} \max_{\nu \in \mathcal{P}(\mathcal{X})} \sum_i \left(\int \psi_i d\Pi_i \nu + \int \psi_i^*(\tilde{c}_i - s_i) d\mu_i \right),
\end{aligned}$$

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 $\mathcal{X} := Extr\left(\{x \in \mathbb{R}^n_+ : x_i - x_{i-1} \geq -1 \ \forall i = 2, \ldots, n, \ x_n \leq 1\}\right)$. As $\mathcal{X}$ turns out to be a bounded $L^\natural$ convex set, Theorem 5 provides an integral extended form representation of $conv(\mathcal{X})$ as in Theorem 4.

However, we elect to use a different representation here. Indeed, $\mathcal{X}$ has in previous works been well-characterized (see Zangwill (1966), Zangwill (1969)) which we exploit to construct a representation of $\mathcal{P}(\mathcal{X})$. More precisely, if we let:

$$\mathcal{X}' := \{\xi = \{[1, a_1], [a_1+1, a_2], \ldots, [a_{k-1}+1, a_k = n+1]\} : 0 = a_0 < a_1 < \ldots < a_k = n+1, \ k \in [n+1], \{a_i\}_{i=1}^k \subset \mathbb{Z}_+\},$$

and define the mapping $f : \mathcal{X}' \to \mathcal{X}$ with component functions $f_i : \mathcal{X}' \to \mathcal{X}_i = \{0, 1, \ldots, n+1-i\}$:

$$f_i(\xi) := \sum_{x_i=0}^{n+1-i} x_i \cdot \mathbb{1}_{\{\xi \ni [k, i+x_i] : \exists k \leq i\}}, \quad \forall i = 1, \ldots, n,$$

then we can obtain a 0-1 polytope P for an extended formulation of the form outlined in Theorem 4 and 6 with $P = \{y \in \mathbb{R}^{(n+1)\times(n+1)} : \sum_{(k,j):k \leq i \leq j} y_{kj} = 1, \ \forall i = 1, \ldots, n, \ 0 \leq y_{ij} \leq 1, \forall (k,j) \in [n+1] \times [n+1]\}$, where we have set B $= \mathcal{F} := \cup_{k \leq j} \{\xi : \xi \ni [k,j]\}$.

Hence, (*) takes the linear programming form:

$$\max_{\nu \in \mathcal{P}(\mathcal{X})} \sum_{x \in \mathcal{X}} \left[\sum_{i=1}^n \psi_i(x_i)\right] \nu(x) = \max_{\nu \in \mathcal{P}(\mathcal{X})} \sum_{i=1}^n \sum_{x_i \in \mathcal{X}_i} \psi_i(x_i) \cdot \Pi_i \nu(x_i) = \max_{y \in P} \sum_{i=1}^n \sum_{x_i \in \mathcal{X}_i} \psi_i(x_i) \left(\sum_{(k, i+x_i):k \leq i} y_{k,i+x_i}\right),$$

and we obtain the following tractable formulation for the distributionally robust appointment scheduling problem:

$$\min_{s \in \mathcal{S}} \max_{\theta \in \Gamma(\mu_1, \ldots, \mu_n)} E_\theta\left[Z(s, \tilde{c})\right] =$$

$$\min_{s \in \mathcal{S}} \min_{\{d_{i,j}\}_{i \leq j \leq n+1}} \max_{y \in P} \sum_{i=1}^n \sum_{j:i \leq j \leq n+1} d_{i,j} \left(\sum_{k:1 \leq k \leq i} y_{kj}\right) + \sum_{i=1}^n E_{\tilde{c}_i \sim \mu_i}\left[\max_{j:i \leq j \leq n+1} \{(\tilde{c}_i - s_i)(j-i) - d_{i,j}\}\right] \qquad (10)$$

Note that the inner maximization can be dualized in a straightforward manner since P is a 0-1 polytope. This leads to a tractable instance under the assumption that the univariate expectations are easy to handle.

5.2. Max Flow with Random Arc Capacities

In the max flow problem (cf. Ahuja, Magnanti and Orlin (1993)), we are given a graph $G = (N, A)$. N is the node set consisting of at least two distinct elements—the *source* s and the *sink* t. A is the arc set consisting of ordered pairs of the form (i, j), which indicates that flow can be directed from i to j, for distinct nodes $i, j \in N$. For each arc $(i, j) \in A$, we let $\tilde{c}_{ij} \geq 0$ denote the random *capacity* of that arc (an upper bound on the flow that can be directed from node i to node j); we use $\tilde{c}$ to refer to the set $\{\tilde{c}_{ij} : (i, j) \in A\}$. We are given the marginal distribution μ_{ij} of $\tilde{c}_{ij}$ for each arc (i, j), and would like to evaluate the expected value of the maximum flow that can be directed from s to t under the worst-case correlation between $\{\mu_{ij} : (i, j) \in A\}$.

For a realization of the arc capacities $\tilde{c}$, the value of the max flow $Z(\tilde{c})$ is given by the optimal objective value of the following LP.

$$\max v \tag{MaxFlow-1}$$

$$\text{s.t.} \sum_{j:(i,j)\in A} x_{ij} - \sum_{j:(j,i)\in A} x_{ji} = \begin{cases} v, & i = s \\ 0, & i \in N \setminus \{s, t\} \\ -v, & i = t \end{cases} \quad \forall i \in N$$

$$0 \leq x_{ij} \leq \tilde{c}_{ij}, \ (i, j) \in A$$

We would like to evaluate $\inf_{\theta \in \Gamma(\{\mu_{ij}:(i,j)\in A\})} \mathbb{E}_{\tilde{c}\sim\theta}[Z(\tilde{c})]$. Towards this goal, observe that $Z(\tilde{c})$ can be written as a discrete minimization problem (the *min cut* problem) over the set $\mathcal{C}$ of incidence vectors to cuts in the graph. Futher, $conv(\mathcal{C})$ is a 0-1, integral polyhedron. The following corollary is then obtained after applying our main result Theorem 2 to this inf-min problem.

Corollary 7. *The expected value of the max flow under the worst-case correlation, defined as* $\inf_{\theta \in \Gamma(\{\mu_{ij}:(i,j)\in A\})} \mathbb{E}_{\tilde{c}\sim\theta}[Z(\tilde{c})]$, *is equal to the optimal objective value of the following problem.*

$$\max \left(v - \sum_{(i,j)\in A} \int_{\tilde{c}_{ij}} \max\{w_{ij} - \tilde{c}_{ij}, 0\} d\mu_{ij} \right) \tag{MaxFlow-2}$$

$$
s.t. \sum_{j:(i,j)\in A} x_{ij} - \sum_{j:(j,i)\in A} x_{ji} = \begin{cases} v, & i = s \\ 0, & i \in N \setminus \{s,t\} \\ -v, & i = t \end{cases} \qquad \forall i \in N
$$

$$
0 \le x_{ij} \le w_{ij}, \;\; (i,j) \in A
$$

Problem (MaxFlow-2) resembles the original LP (MaxFlow-1), except now the arc capacities, instead of being fixed realizations $\tilde{c}_{ij}$, are decision variables w_{ij}. For each arc (i,j) there is a "penalty" term $\int_{\tilde{c}_{ij}} \max\{w_{ij} - \tilde{c}_{ij}, 0\} d\mu_{ij}$ in the objective function, dependent on the marginal distribution μ_{ij}, which dissuades w_{ij} from being as large as possible.

In general, the worst-case joint distribution θ between the random arc capacities will be high-dimensional and intractable to compute. Nonetheless, Corollary 7 shows that we can still evaluate the worst-case expected value of the max flow. Indeed, (MaxFlow-2) is tractable since the penalty terms are convex—in fact, if each μ_{ij} is given as a discrete distribution, then (MaxFlow-2) can be reformulated as a polynomial-sized LP. Below, we directly prove Corollary 7 in this case of discrete distributions, to provide a more intuitive explanation of our main result which is *free from measure-theoretic technicalities.*

Proof of Corollary 7 under Discrete Distributions: First, we reference the first statement in our main result Theorem 2 to establish that

$$
\inf_{\theta\in\Gamma(\{\mu_{ij}:(i,j)\in A\})} \mathbb{E}_{\tilde{c}\sim\theta}[Z(\tilde{c})] = \min_{\nu\in\mathcal{P}(\mathcal{C})} \sum_{(i,j)\in A} \min_{\gamma_{ij}\in\Gamma(\mu_{ij},\Pi_{ij}\nu)} E[\tilde{c}_{ij} x_{ij}]. \tag{11}
$$

This will be the only time we reference Theorem 2 in the present proof, and furthermore we provide an intuitive explanation of equation (11). Consider a joint distribution θ^* which minimizes the LHS of (11). θ^* induces a distribution over the feasible region $\mathcal{C}$; namely, the distribution of the random vector $x^*(\tilde{c})$, where $x^*(\cdot)$ is an appropriately defined mapping taking any realization of $\tilde{c}$ to an optimal solution of the problem $Z(\tilde{c})$. Using this distribution induced by $x^*(\cdot)$ as the ν in the RHS of (11), one can see that the RHS is a lower bound on the LHS. To show the converse, i.e. the

LHS is a lower bound on the RHS, one can similarly take any distribution ν^* over $\mathcal{C}$ minimizing the RHS and construct a corresponding joint distribution θ which achieves the same value on the LHS.

Now, let us consider an arbitrary (i,j)-th summand term. For notational convenience, define $c_0 := 0$ and write $supp(\mu_{ij}) := \{c_1, \ldots, c_{m_{ij}}\}$, with $c_1 < \ldots < c_{m_{ij}}$, where $m_{ij} := |supp(\mu_{ij})|$, and observe that

$$\min_{\gamma_{ij} \in \Gamma(\mu_{ij}, \Pi_{ij}\nu)} E\left[\tilde{c}_{ij} x_{ij}\right] = \min_{\gamma_{ij} \in \Gamma(\mu_{ij}, \Pi_{ij}\nu)} \sum_{k=1}^{m_{ij}} c_k \cdot \gamma_{ij}\left((c_k, 1)\right).$$

In the RHS, the smallest value that $\sum_{k=1}^{m_{ij}} \gamma_{ij}\left((c_k, 1)\right) = \Pi_{ij}\nu(1)$ can take is

$$\sum_{k=1}^{k^*-1} c_k \cdot \mu_{ij}(c_k) + c_{k^*} \cdot (\Pi_{ij}\nu(1) - F_{\mu_{ij}}(c_{k^*-1})),$$

where $k^* := min\{k \in [m_{ij}] : F(c_k) \geq \Pi_{ij}\nu(1)\}$. This is because the γ_{ij} which minimizes $\sum_{k=1}^{m_{ij}} \gamma_{ij}\left((c_k, 1)\right) = \Pi_{ij}\nu(1)$ will "couple" the smallest elements of $supp(\mu_{ij})$ with the event $x_{ij} = 1$ until the total mass reaches $\Pi_{ij}\nu(1)$.

Next, we derive the following:

$$\begin{aligned}
&\sum_{k=1}^{k^*-1} c_k \cdot \mu_{ij}(c_k) + c_{k^*} \cdot (\Pi_{ij}\nu(1) - F_{\mu_{ij}}(c_{k^*-1})) \\
&= \sum_{k=1}^{k^*-1} \sum_{l=0}^{k-1} (c_{l+1} - c_l) \mu_{ij}(c_k) + \sum_{k=0}^{k^*-1} (c_{k+1} - c_k) \cdot (\Pi_{ij}\nu(1) - F_{\mu_{ij}}(c_{k^*-1})) \\
&= \sum_{l=0}^{k^*-2} (c_{l+1} - c_l) \sum_{k=l+1}^{k^*-1} \mu_{ij}(c_k) + \sum_{k=0}^{k^*-1} (c_{k+1} - c_k) \cdot (\Pi_{ij}\nu(1) - F_{\mu_{ij}}(c_{k^*-1})) \\
&= \sum_{l=0}^{k^*-2} (c_{l+1} - c_l) \left(F_{\mu_{ij}}(c_{k^*-1}) - F_{\mu_{ij}}(c_l)\right) + \sum_{k=0}^{k^*-1} (c_{k+1} - c_k) \cdot (\Pi_{ij}\nu(1) - F_{\mu_{ij}}(c_{k^*-1})) \\
&= \sum_{k=0}^{k^*-1} (c_{k+1} - c_k) \cdot (\Pi_{ij}\nu(1) - F_{\mu_{ij}}(c_k)).
\end{aligned}$$

The final expression is equal to

$$\begin{aligned}
\underset{b_1^{ij}, \ldots, b_{m_{ij}}^{ij}}{\text{maximize}} \quad & \sum_{k=1}^{m_{ij}} b_k^{ij} (\Pi_{ij}\nu(1) - F_{\mu_{ij}}(c_{k-1})) \\
\text{subject to} \quad & 0 \leq b_k^{ij} \leq c_k - c_{k-1},\ k = 1, \ldots, m_{ij}
\end{aligned}$$

because it is optimal to set $b_k^{ij} = c_k - c_{k-1}$ for all $k = 1, \ldots, k^* - 1$ and $b_k^{ij} = 0$ otherwise.

Let $\mathcal{B}^{ij}$ denote the feasible set to this optimization problem. We can subsequently derive:

$$
\begin{aligned}
&\max_{b^{ij}\in\mathcal{B}}\left[\sum_{k=1}^{n} b_k^{ij}\right]\Pi_{ij}\nu(1)-\sum_{k=1}^{n} b_k^{ij}F_{\mu_{ij}}(c_{k-1})\\
&=\max_{b^{ij}\in\mathcal{B}^{ij}}\left[\sum_{k=1}^{n} b_k^{ij}\right]\Pi_{ij}\nu(1)-\sum_{l=1}^{n-1}\mu_{ij}(c_l)\left(\sum_{k=1}^{n} b_k^{ij}-\sum_{k=1}^{l} b_k^{ij}\right)\\
&=\max_{b^{ij}\in\mathcal{B}}\left[\sum_{k=1}^{n} b_k^{ij}\right]\Pi_{ij}\nu(1)-\sum_{l=1}^{n-1}\mu_{ij}(c_l)\max\left(\sum_{k=1}^{n} b_k^{ij}-\sum_{k=1}^{l} b_k^{ij},0\right)\\
&=\max_{w_{ij}\in supp(\mu_{ij})} w_{ij}\cdot\Pi_{ij}\nu(1)-\int\max(w_{ij}-\tilde{c}_{ij},0)d\mu_{ij}(\tilde{c}_{ij})
\end{aligned}
$$

where the last equality holds because there always exists an optimal b^{ij} vector that satisfies $\sum_{k=1}^{n} b_k^{ij}\in supp(\mu_{ij})$.

In summary,

$$
\begin{aligned}
\inf_{\theta\in\Gamma(\{\mu_{ij}:(i,j)\in A\})}\mathbb{E}_{\tilde{c}\sim\theta}[Z(\tilde{c})] &= min_{\nu\in\mathcal{P}(\mathcal{C})}\sum_{(i,j)\in A}\min_{\gamma_{ij}\in\Gamma(\mu_{ij},\Pi_{ij}\nu)} E[\tilde{c}_{ij}x_{ij}]\\
&= min_{\nu\in\mathcal{P}(\mathcal{C})}\max_{\{w_{ij}\in supp(\mu_{ij})\}_{(i,j)\in A}}\sum_{(i,j)\in A} w_{ij}\cdot\Pi_{ij}\nu(1)-\int\max(w_{ij}-c,0)d\mu_{ij}(\tilde{c}_{ij})\\
&= \max_{\{w_{ij}\in supp(\mu_{ij})\}_{(i,j)\in A}} min_{\nu\in\mathcal{P}(\mathcal{C})}\sum_{(i,j)\in A} w_{ij}\cdot\Pi_{ij}\nu(1)-\int\max(w_{ij}-c,0)d\mu_{ij}(\tilde{c}_{ij}),
\end{aligned}
$$

where the exchange of max and min is justified by von Neumann's minimax theorem. All that remains is to see that

$min_{\nu\in\mathcal{P}(\mathcal{C})}\sum_{(i,j)\in A} w_{ij}\cdot\Pi_{ij}\nu(1)$

$=$

$$
\begin{aligned}
\max\quad & v\\
\text{s.t.}\quad & \sum_{j:(i,j)\in A} x_{ij}-\sum_{j:(j,i)\in A} x_{ji}=\begin{cases} v, & i=s\\ 0, & i\in N\setminus\{s,t\}\\ -v, & i=t\end{cases}\qquad \forall i\in N\\
& 0\le x_{ij}\le w_{ij},\ (i,j)\in A
\end{aligned}
$$

and then the proof is complete. □

Interestingly, the *robust max-flow* problem, while no longer a max-flow problem, remains a min-cost flow problem.

COROLLARY 8. *The expected value of the max flow under the worst-case correlation, $\inf_{\theta\in\Gamma(\{\mu_{ij}:(i,j)\in A\})}\mathbb{E}_{\tilde{c}\sim\theta}[Z(\tilde{c})]$, when each probability measure μ_{ij} has finite support, can be reformulated as a min-cost flow problem.*

Proof:

$$\begin{aligned}\inf_{\theta\in\Gamma(\{\mu_{ij}:(i,j)\in A\})}\mathbb{E}_{\tilde{c}\sim\theta}[Z(\tilde{c})] &= \min_{\nu\in\mathcal{P}(\mathcal{C})}\sum_{(i,j)\in A}\min_{\gamma_{ij}\in\Gamma(\mu_{ij},\Pi_{ij}\nu)} E\left[\tilde{c}_{ij}x_{ij}\right]\\ &= \min_{\nu\in\mathcal{P}(\mathcal{C})}\sum_{(i,j)\in A}\max_{b^{ij}\in\mathcal{B}^{ij}}\sum_{k=1}^{m_{ij}} b_k^{ij}(\Pi_{ij}\nu(1)-F_{\mu_{ij}}(c_{k-1}))\\ &= \max_{b:b^{ij}\in\mathcal{B}^{ij}\,\forall(i,j)\in A}\ \min_{\nu\in\mathcal{P}(\mathcal{C})}\sum_{(i,j)\in A}\sum_{k=1}^{m_{ij}} b_k^{ij}\Pi_{ij}\nu(1)-\sum_{(i,j)\in A}\sum_{k=1}^{m_{ij}} b_k^{ij}F_{\mu_{ij}}(c_{k-1}),\end{aligned}$$

where

$$\min_{\nu\in\mathcal{P}(\mathcal{C})}\sum_{(i,j)\in A}\sum_{k=1}^{m_{ij}} b_k^{ij}\Pi_{ij}\nu(1)$$

has as its linear programming dual

$$\begin{aligned}\max\quad & v\\ \text{s.t.}\quad & \sum_{j:(i,j)\in A}x_{ij}-\sum_{j:(j,i)\in A}x_{ji}=\begin{cases}v, & i=s\\ 0, & i\in N\setminus\{s,t\}\\ -v, & i=t\end{cases}\qquad \forall i\in N\\ & 0\le x_{ij}\le\sum_{k=1}^{m_{ij}} b_k,\ (i,j)\in A\end{aligned}$$

Thus, after replacing each occurence of x_{ij} with $\sum_{k=1}^{m_{ij}} b_k$, we obtain a max-cost flow problem in the flow variables b_k^{ij}. □

Furthermore, while (MaxFlow-2) merely evaluates the max flow under the worst-case joint distribution, it can easily be modified to optimize for robust pre-positioning decisions against this distribution. For example, suppose that at each arc (i,j), we can deterministically increase the capacity on that arc by $s_{ij} \ge 0$ units, at a cost of P_{ij} per unit increase. We would like to find the minimum total cost required to guarantee that the final flow delivered is at least some value F. This problem can be formulated as follows.

$$\begin{aligned}
&\min \sum_{(i,j)\in A} P_{ij} s_{ij} \\
&\text{s.t. } v - \sum_{(i,j)\in A} \int_{\tilde{c}_{ij}} \max\{w_{ij} - \tilde{c}_{ij} - s_{ij}, 0\} d\mu_{ij} \ge F \\
&\sum_{j:(i,j)\in A} x_{ij} - \sum_{j:(j,i)\in A} x_{ji} = \begin{cases} v, & i = s \\ 0, & i \in N \setminus \{s,t\} \\ -v, & i = t \end{cases} \quad \forall i \in N \\
&0 \le x_{ij} \le w_{ij}, \ (i,j) \in A \\
&0 \le s_{ij}, \ (i,j) \in A
\end{aligned}$$

The first constraint guarantees that for the set of s_{ij}-variables chosen, it is indeed possible to choose v and w_{ij} to ensure that the final flow is at least F. This constraint is also tractable since the penalty terms subtracted are convex.

5.3. Ranking with Applications to a Scheduling Problem

Consider a set of n random utilities, denoted by $\tilde{c}_1, \ldots, \tilde{c}_n$, which for example could represent the skills of n players. The ranking problem with random utilities is expressed as:

$$\begin{aligned}
Z_{perm}(c) = \max &\sum_{i=1}^{n} \tilde{c}_i x_i \\
\text{s.t. } &x \in \mathcal{X}_{perm},
\end{aligned} \tag{12}$$

where $\mathcal{X}_{perm}$ is the set of all permutations on $[n]$, i.e., $x : [n] \to [n] \quad \forall x \in \mathcal{X}_{perm}$. Unfortunately, $|\mathcal{X}_{perm}| = n!$ is superpolynomial in n. Equivalently this set $\mathcal{X}_{perm}$ can be viewed as the set of extreme points to the permutahedron P_{perm} which is defined by the sets of inequalities:

$$\begin{aligned}
\sum_{i\in S} x_i &\ge \frac{|S|(|S|+1)}{2}, \ \forall S \subset [n], S \ne \phi, \\
\sum_{i=1}^{n} x_i &= \frac{n(n+1)}{2}.
\end{aligned}$$

While this is not a compact representation, an application of Birkhoff's theorem implies that we can efficiently solve the right-hand side of (*). Indeed, to see this, begin by observing that if $\mathcal{X}'$ is the set of 0/1 permutation matrices, then $f:\mathcal{X}'\to\mathcal{X}$ defined by $f(Y):=Y\cdot(1,2,\dots,n)^T$ is an invertible linear mapping between the sets. Next, Birkhoff's theorem tells us that the Birkhoff polytope $B_n = conv(\mathcal{X}')$; in other words, $conv(\mathcal{X}')$ can be formulated as a 0-1 polytope in n^2 variables $\{y_{ij}\}_{i,j\in[n]}$. Next, with $f_i^{-1}(x_i) = \cup\{Y : Y\in\mathcal{X}', Y_{i,x_i}=1\}$, we have $\Pi_i\nu(x_i) = y_{i,x_i}$. This implies that right-hand side of (*) can be formulated as a set of primal and dual linear programs:

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

Using the dual formulation in (9), we have a min-min problem that is tractable. This follows from an extended formulation of the permutahedron, in which the additional variables belong to a 0-1 polytope: $\{(x,y): x_i = \sum_j j\cdot y_{ij}, \forall i, \sum_i y_{ij}=1, \sum_j y_{ij}=1, y\ge 0\}$.

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:

$$\min_{t\in\mathcal{T}} \sup_{\theta\in\Gamma(\mu_1,\dots,\mu_n)} E_{\tilde{c}\sim\theta}\left[Z(t,\tilde{c})\right]. \tag{13}$$

Here $Z(t,\tilde{c})$ is the optimal value to the linear optimization problem:

$$\begin{aligned} Z(t,\tilde{c}) &:= \max \sum_{i=1}^{n}(\tilde{c}_i - t_i)x_i \\ &\text{s.t. } x \in \mathcal{X}_{perm}, \end{aligned} \tag{14}$$

where t is the reduction in the individual job times which is assumed to lie in a set $\mathcal{T}$. A direct application of the result indicates that this problem can be formulated as:

$$\min_{t\in\mathcal{T},d_{ij}} \max_{\sum_i y_{ij}=1,\sum_j y_{ij}=1,y\geq 0} \sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}y_{ij} + \sum_{i=1}^{n} E_{\tilde{c}_i\sim\mu_i}[\max_{j\in[n]}\{(\tilde{c}_i - t_i)\cdot j - d_{ij}\}], \tag{15}$$

which by applying linear programming duality to the inner maximization problem reduces to the tractable optimization problem:

$$\begin{aligned} \min \ & \sum_{j=1}^{n}\alpha_j + \sum_{i=1}^{n}\beta_i + \sum_{i=1}^{n} E_{\tilde{c}_i\sim\mu_i}[\max_{j\in[n]}\{(\tilde{c}_i - t_i)\cdot j - d_{ij}\}] \\ \text{s.t. } & t \in \mathcal{T}, \\ & \alpha_j + \beta_i \geq d_{ij}, \quad \forall i,j. \end{aligned} \tag{16}$$

PROPOSITION 3. *For any t, $Z(t,\tilde{c})$ is monotone, submodular in the $\tilde{c}$ variables.*

Proof: We refer the reader to the Appendix section EC.3.3. □

COROLLARY 9. *(Agrawal et al. (2012) Corollary 3) If $Z(t,\cdot)\geq 0$ $\forall t\in\mathcal{T}$, i.e., when $t\leq\tilde{c}$ $\forall(t,\tilde{c})\in \mathcal{T}\times supp(\tilde{c})$, and is integrable with respect to the independent coupling of the $\tilde{c}_i$ variables, then*

$$POC := \frac{\sup_{\theta\in\Gamma(\mu_1,\ldots,\mu_n)} Z(\mathbf{t}^I,\tilde{\mathbf{c}})}{\sup_{\theta\in\Gamma(\mu_1,\ldots,\mu_n)} Z(\mathbf{t}^R,\tilde{\mathbf{c}})} \leq e/(e-1)$$

where $\mathbf{t}^R$ denote the robust solution and $\mathbf{t}^I$ denote the solution under independent assumption.

5.4. 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 is denoted by the constraint $S_j \geq S_i + d_{ij}$ where $d_{ij} \in (-\infty, \infty)$ is an integer number imposing a time lag between the jobs. Assume the 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 directed 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:

$$\inf_{\theta \in \Gamma(\mu_1, \ldots, \mu_n)} E\left[\min_{S \in \mathcal{S}} \sum_{j=1}^{n} \tilde{c}_j(S_j)\right], \tag{17}$$

where $\mathcal{S}$ is the set of feasible starting time vectors.

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" = $conv\{x^F : F \text{ is a feasible schedule}\} = conv(\mathcal{X})$. 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:

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

The linear programming relaxation of this integer program provides integral solutions since the constraint matrix is totally unimodular (see Möhring et al. (2001)). In other words, $conv(\mathcal{X})$ is a 0-1 polytope; hence, Theorem 4 implies that the robust bound in (17) in MDM can be computed efficiently.

To illustrate the tractable computational form:

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

where

$$\begin{aligned}
Q := \{\psi,\lambda,\gamma:\ &\lambda_j+\sum_{(i,t):i\in N^+(j),t=0,\ldots s}\gamma_{ji}^t+\sum_{(i,t):i\in N^-(j),t=(s-d_{ij}+1)\vee 0,\ldots,T}\gamma_{ij}^t\ge\psi_{js},\forall j\in\{1,\ldots,n\},\forall s\in\{0,\ldots,T\},\\
&\gamma_{ij}^t\ge 0,\forall(i,j)\in E,t\in\{0,\ldots,T\}\},
\end{aligned}$$

and for each j, the integrand term

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

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

has an equivalent dual minimization form:

$$\begin{aligned}
\underset{\{\alpha_{j'}\}_{j'=1}^{n},\{\beta_{ij'}^{t}\}_{(i,j')\in E,t=0:T}}{\text{minimize}} \quad & \sum_{j'=1}^{n}\alpha_{j'}+\sum_{(i,j',t)\in E,t\in\{0,\dots,T\}}\beta_{ij'}^{t} \\
\text{s.t.} \quad & \alpha_{j'}+\sum_{(i,j',t):(j',i)\in E,t\in\{0,\dots,s\}}\beta_{j'i}^{t}+\sum_{(i,j',t):(i,j')\in E,t\in\{(s-d_{ij'}+1)\vee 0,\dots,T\}}\beta_{ij'}^{t}\geq 0,\ \forall j'\neq j, s\in\{0,\dots,T\} \\
& \alpha_{j}+\sum_{(i,t):(j,i)\in E,t\in\{0,\dots,s\}}\beta_{ji}^{t}+\sum_{(i,t):(i,j)\in E,t\in\{(s-d_{ij'}+1)\vee 0,\dots,T\}}\beta_{ij}^{t}\geq -\tilde{\epsilon}_j c_j^0(s)-\psi_{js},\ \forall s\in\{0,\dots,T\} \\
& \beta_{ij'}^{t}\geq 0\ \forall (i,j')\in E, t\in\{0,\dots,T\}
\end{aligned}$$

6. Computational Experiments

6.1. Appointment Scheduling

Mak, Rong and Zhang (2015) has studied the appointment scheduling problem considered in this paper with first two marginal moments. They derived a second order cone to model the distributionally robust version of the problem. Although tractable, the distributionally robust model with only first two marginal moments is often challenged by the conservativeness of the solution. Incorporating more information of uncertainties helps to mitigate this effect at the cost of sacrificing tractability. In Section 5.1, we have shown with marginal distribution information, we can still get a tractable model using our dual formula. We are interested in understanding to what extent incorporating marginal distribution information manages to mitigate the conservativeness of the marginal moment based model. To see it, we investigate the performance of the robust solutions from MDM and MMM model respectively when the distributions of the service duration are known and independent.

Our experimental setup is based on Mak, Rong and Zhang (2015). Namely, we consider the case of $n = 5$ jobs, and assume that the job durations follow three types of probability distributions: normal, gamma, and log-normal. Under each type of probability assumption, we generate our random problem instances as follows:

- Randomly generate 100 instances each by sampling:
  - —Mean $m_i \sim U[30,60]$ with parameter $\epsilon \sim U[0,0,3]$,

—Standard deviation $\sigma_i = m_i \cdot \epsilon$,

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

For each instance, using the generated m_i and σ_i, we can solve a second order conic model proposed in Mak, Rong and Zhang (2015) to get the robust solution under the marginal moment based model, denoted as a MMM solution. We can also solve (10) with the marginal distribution specified as the given type of probability distribution (normal/ gamma/ log-normal) with the generated mean m_i and standard deviation σ_i. We use the sample average approximation method with $N = 5000$ samples to compute univariate expectation in (10). The expected performance of the two robust solutions under the independent probability distribution can be evaluated by the sample mean in simulation.

In Figures 1-3, we compare the expected costs of the MDM solution and the MMM solution under the assumption that the true distribution is the independent distribution with the specified marginals. The figures plot the empirical cumulative distribution function of the total wait times for the 100 randomly generated problem instances. As the figures illustrate, MDM can present a sizeable reduction of the expected waiting time over MMM. The mean of the relative reduction for the three cases are 0.2462 (Normal), 0.2420 (Log-Normal), 0.2001 (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. From the other perspective, it also indicates that when the marginal distribution is available, the usage of only marginal moments will incur some loss in getting an efficient schedule.

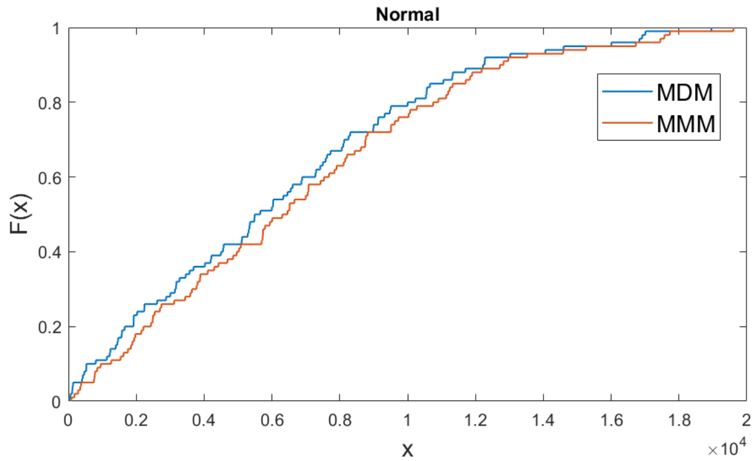

Figure 1 Performance Comparison Between MDM Solution and MMM Solution (Normal)

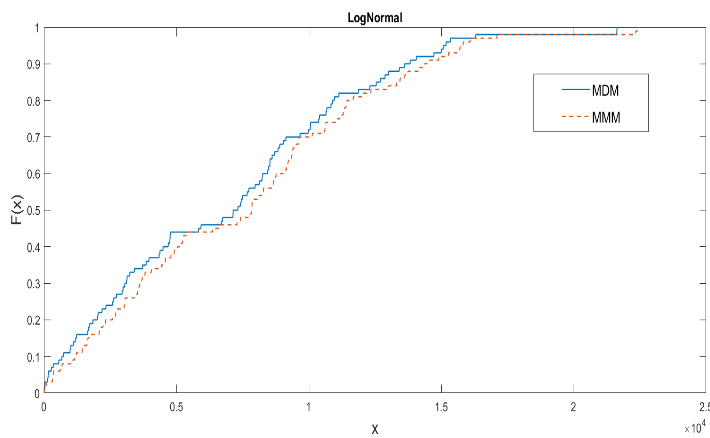

Figure 2 Performance Comparison Between MDM Solution and MMM Solution (Log-Normal)

6.2. Ranking with Scheduling

We have shown the ranking problem (14) is submodular in $\mathbf{c}$. Hence the price of correlation is low for the ranking problem according to Corollary 9. Indeed, we can test it using a small numeric example. Consider the resource allocation application in Section 5.3. Suppose there are $n = 5$ jobs and assume the job duration is discrete distributed. Suppose the cardinality of the support set for each job is the same and denote it as N. We take $N = 5$ in this example. We generate 100 problem instances and each instance is generated as follows:

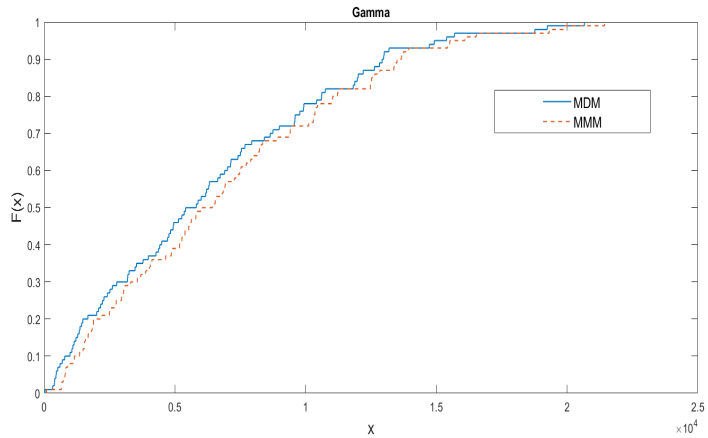

Figure 3 Performance Comparison Between MDM Solution and MMM Solution (Gamma)

.

- We randomly generate an upper bound of the time reduction u_i for each job i.
- The support is generated by multiplying u_i by N random numbers between $(1,2)$
- Randomly generate the probability of each realization in the support set.
- Calculate the first two marginal moments based on the probability generated.

We can calculate the price of correlation (POC) for each problem instance and plot the empirical distribution of the POC in Figure 6.2. From the figure, we observe the price of correlation for this allocation problem is close to 1, which is below the upper bound derived.

Given the low price of correlation, one may argue that we can use independent coupling to approximate the worst case distribution without involving much risk. However, to solve the ranking problem under independent coupling exactly is not necessarily an easy task even when the marginals are discrete distributed. Indeed, consider the example above. To solve problem (13) exactly, we must enumerate the full distribution which involves N^n scenarios, a large linear program. In contrast,our dual formula of the robust model (16) only requires a linear program of size $O(Nn)$, which can be efficiently solved.

Similar to the appointment scheduling problem, we also have a tractable second order conic program for the robust model with first two marginal moments (see Appendix EC.4 for the formula

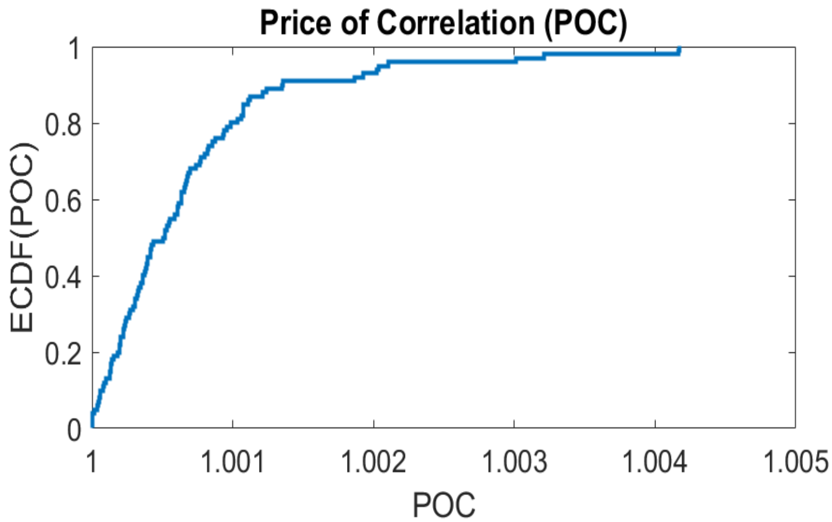

Figure 4 Empirical cumulative distribution of POC for 100 problem instances

.

of the second order conic program). Although also tractable, we claim the MMM solution does not perform as well as the MDM solution. To see this, we plot the empirical cumulative distribution function of the difference between the performance of the MDM solution and optimal cost under independent coupling versus the difference between the performance of the MMM solution and optimal cost under independent coupling. From Figure 5, we can see the MDM solution generates a lower performance difference than does the MMM solution. In summary, for the ranking problem, our dual formula can achieve a less conservative robust solution and preserve tractability.

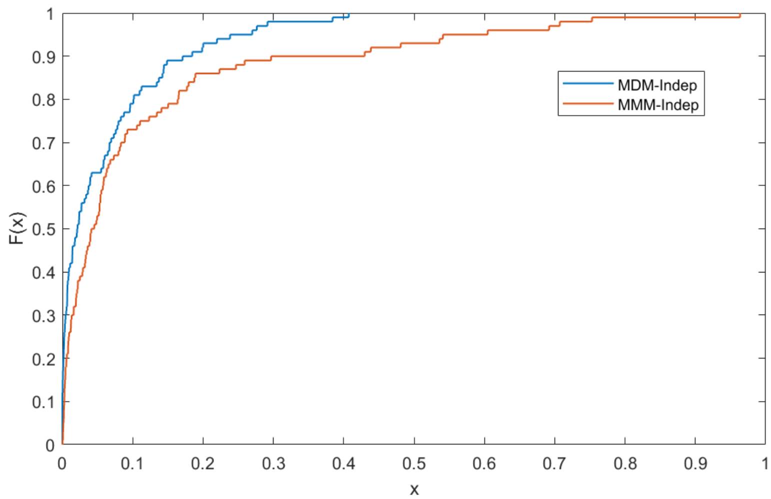

Figure 5 Performance comparison between MDM solution and MMM solution under independent coupling

.

Acknowledgments

We would like to thank Professors James B. Orlin and Chung-Piaw Teo for several interesting discussions on this topic.

References

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

R. K. Ahuja, T. L. Magnanti and J. B. Orlin. (1993). *Network flows: theory, algorithms, and applications.* Prentice Hall.

F. Bach. (2018). *Submodular Functions: from Discrete to Continous Domains.* Mathematical Programming.

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

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

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

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

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

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

J. Blanchet, Y. Kang, F. Zhang and K. Murthy. (2017). *Data-driven optimal transport cost selection for distributionally robust optimization.*

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

G. Carlier and I. Ekeland. (2010). *Matching for teams.* Economic Theory, 42(2), 397-418.

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

I. Ekeland and R. Temam. (1999). *Convex analysis and variational problems.* SIAM Classics in Applied Mathematics, Philadelphia.

P. M. Esfahani and D. Kuhn. (2017). *Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations.* Mathematical Programming, 1-52.

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

R. Gao and A. J. Kleywegt. (2016). *Distributionally robust stochastic optimization with Wasserstein distance.*

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

A. N. Kolmogorov and S. V. Fomin. (1970). *Introductory real analysis.* Translated and edited by R. A. Silverman. Englewood Cliffs, N.J: Prentice-Hall.

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

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

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

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

I. Meilijson (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.

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

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

R. H. Möhring, A. S. Schulz, F. Stork and M. Uetz. (2001). *On project scheduling with irregular starting time costs.* Operations Research Letters, 28, 149154.

Murota. (2003). *Discrete Convex Analysis.* SIAM, Philadelphia.

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

K. Natarajan, D. Shi and K-C. Toh. (2017). *Bounds for random binary quadratic programs.* To appear in SIAM Journal on Optimization.

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

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

B. Pass. (2012). *On the local structure of optimal measures in the multi-marginal optimal transportation problem.* Calculus of Variations, 43: 529-536.

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

R. T. Rockafellar. (1997). *Convex analysis.* Princeton University Press.

H. L. Royden and P. M. Fitzpatrick. (2010). *Real analysis.* Fourth Edition, Prentice Hall.

F. Santambrogio. (2015). *Optimal transport for applied mathematicians.* Birkhäuser Basel.

H. Scarf. (1958). *A min-max solution of an inventory problem*, Studies in The Mathematical Theory of Inventory and Production. (K. Arrow, S. Karlin and H. Scarf, Eds.) 201-209, Stanford University Press, California.

A. Shapiro, D. Dentcheva and A. Ruszczyński. (2014). *Lectures on stochastic programming: Modeling and theory.* Second edition. MOS-SIAM Series on Optimization.

D. Simchi-Levi, X. Chen, J. Bramel. (2014). *The Logic of Logistics.* Third Edition, Springer.

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

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

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

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

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

W. I. Zangwill. (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.

Appendix(Proofs of Statements)

EC.1. MDM Primal-Dual: An Economic Interpretation of Theorem 2

In this section, we relate the problem to a multi-marginal optimal transport problem that has been studied in the mathematical economics literature, building on which we provide an economic interpretation of the dual problem formulation. Carlier and Ekeland (2010) studied the problem of equilibrium matching where a list of tasks Z and a population comprising of n groups of equal size is given. For each group j, X_j denotes the set of various skill levels of the members in group j. Each task in Z must be done by a team, comprising of a single member from each group. If a member of group j chooses to be part of a team that performs task $z \in Z$, then they receive a payment $\psi_j(z)$. The goal is to find an equilibrium, which specifies a distribution ν over the tasks Z, as well as a payment/transfer system $\{\psi_j : X_j \to \mathbb{R}\}_j$ that leads to the entire system of agents distributing themselves into distinct teams, each assigned to a task in Z. The authors showed that this equilibrium can be obtained from the solution to the following pair of optimization problems.

Theorem EC.1 **(Matching Equilibria for Teams from Ekeland and Temam (1999))**. *Let $Z, X_1, \ldots, X_n$ be compact metric spaces, and let $\mu_i \in P(X_i)$, for all i = 1, ..., n. Let $c_i \in C^0(X_i \times Z, \mathbb{R})$, for all i = 1, ..., n. Then*

$$\inf_{\nu \in P(\mathcal{Z})} \sum_{i=1}^{n} \inf_{\gamma_i \in \Gamma(\mu_i, \nu)} \int_{X_i \times Z} c_i(x_i, z) d\gamma_i = \begin{array}{rl} \sup\limits_{\psi_i \in C^0(Z, \mathbb{R})} & \sum\limits_{i=1}^{n} \int_{X_i} \psi_i^{c_i} d\mu_i \\ s.t. & \sum\limits_{i=1}^{n} \psi_i = 0, \end{array}$$

where $\psi_i^{c_i}(x_i) := \inf_{z \in Z}\{c_i(x_i, z) - \psi_i(z)\}$.

The left hand side of this formulation has similarities to the primal formulation in Theorem 3. The only difference is that all the summands involve couplings with ν, i.e., $\gamma_i \in \Gamma(\mu_i, \nu)$ in place of our problem with $\gamma_i \in \Gamma(\mu_i, \Pi_i \nu)$. This difference becomes accounted for in a more pronounced way on the right hand side, as their dual problem now consists of a "balancing" constraint. This leads us to provide a related interpretation for the dual problem in MDM, which we now describe.

Let there be a customer population segmented into groups indexed by $1,\ldots,n$. A customer from group i has type represented by $\tilde{c}_i$, distributed according to a given measure μ_i. Each group i has demand for a different type of house, for example a mansion, condominium, lake house, etc. whose set of qualities such as size, level of luxury, etc. is denoted by $\mathcal{X}_i$. Denote by $x_i \in \mathcal{X}_i$ a particular quality of a type i house. A builder/contractor produces the houses that the groups $1,\ldots,n$ have demand for. However, resource constraints restrict it to producing its lineup of houses in specific quality bundles, e.g., $x = (x_1,\ldots,x_n) \in \mathcal{X}$. Denote by $\mathcal{X}$ the finite set of quality bundles the builder/contractor can produce. The customers have quasilinear utility functions. An item of quality x_i obtained at price $\psi_i(x_i)$, paid to the builder, provides utility $\tilde{c}_i x_i - \psi_i(x_i)$ to a group i customer of type $\tilde{c}_i$. We assume the customers are utility maximizing. Given a system of prices $\{\psi_i(x_i)\}_{x_i\in\mathcal{X}_i}$, a group i customer of type $\tilde{c}_i$ obtains, through compartative shopping, the "indirect utility" $\max_{x_i\in\mathcal{X}_i} \tilde{c}_i x_i - \psi_i(x_i)$.

The dual minimization form:

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

is then the search for a payment system $\{\psi_i : \mathcal{X}_i \to \mathbb{R}\}_{i=1}^n$ that leads to an equilibrium (strong duality) for all the parties (customers and builder) acting in their own self-interest, i.e., the customers doing comparative shopping and the company supplying a distribution that would maximize its own revenue. The primal formulation:

$$\max_{\theta\in\Gamma(\mu_1,\ldots,\mu_n)} E_{\tilde{c}\sim\theta}[\max_{x\in\mathcal{X}} \sum_i \tilde{c}_i x_i], \tag{Primal}$$

is the problem of finding the best joint arrival rate of customers that, assuming the contractor always provides the best quality-bundle, results in a social utility that is highest in expectation. From Corollary 1,

$$\Pi_i\nu(-\infty, x_i] = \mu_i(-\infty, (\psi_i)'_+(x_i)],$$

indicates that the steeper the increase in "price" for the qualities higher than x_i, the more provision for qualities less than or equal to x_i by the optimal supply distribution ν. Lastly,

$$\int_{\mathcal{X}} \sum_i \psi_i(x_i) d\nu \le \int_{\mathcal{X}} \sum_i \psi_i(x_i) d\bar{\nu}, \quad \forall \nu \in \mathcal{P}(\mathcal{X}),$$

indicates that given a pricing scheme, the optimal supply distribution should provide the contractor with largest payoff paid by the consumers. In summary, the primal and dual optimization problems in Section 3 offer two different ways of organizing the market - one centralized, another decentralized that yield the same results - an equilibrium.

EC.2. Section 3 Proofs

EC.2.1. Proof of Theorem 2

Theorem 2 (Lagrange Form) *Let $Z(c) := \max_{x \in \mathcal{X}} c^T x$. Let there be given n probability measures $\{\mu_i\}_{i=1}^n$ over $\mathbb{R}$, each with finite variance. And let $\mathcal{X} \subset \mathbb{R}^n$ be an arbitrary finite point set. Then*

$$\max_{\theta \in \Gamma(\mu_1, \ldots, \mu_n)} E_{\tilde{c} \sim \theta} Z(\tilde{c}) = \max_{\nu \in \mathcal{P}(\mathcal{X})} \sum_i \max_{\gamma_i \in \Gamma(\mu_i, \Pi_i \nu)} E_{(\tilde{c}_i, x_i) \sim \gamma_i} \tilde{c}_i x_i = \max_{\nu \in \mathcal{P}(\mathcal{X})} \inf_{\{\psi_i : \mathcal{X}_i \to \mathbb{R}\}_{i=1}^n} L(\{\psi_i\}_i, \nu) \quad \text{(Primal)}$$

$$= \min_{\{\psi_i : \mathcal{X}_i \to \mathbb{R}\}_{i=1}^n} \sup_{\nu \in \mathcal{P}(\mathcal{X})} L(\{\psi_i\}_i, \nu) \quad \text{(Dual)}$$

For an economic interpretation of this max-min, min-max duality, the reader is referred to Appendix section EC.1.

Proof: We establish the equalities in order, justifying the "max" notations along the way. Regarding the first equality, on the right hand side, the inner max is attained from Theorem 1. To see that the outer max is attained, note that $\mathcal{P}(\mathcal{X})$ is weak-* compact. Next, observe that the mapping $\nu \mapsto \sum_i \max_{\gamma_i \in \Gamma(\mu_i, \Pi_i \nu)} E_{(\tilde{c}_i, x_i) \sim \gamma_i} [\tilde{c}_i x_i]$ is concave and weak-* upper semi-continuous for all i. This follows because if for arbitrary i, we let:

$$F(\psi_i) := -\int \psi_i^* d\mu_i, \quad \forall \psi_i \in C^0(\mathcal{X}_i, \mathbb{R}),$$

then by the (inf) Fenchel-conjugacy operation and Theorem 1, we have:

$$F^*(\nu_i) := \begin{cases} \max_{\gamma_i \in \Gamma(\mu_i, \nu_i)} E_{(\tilde{c}_i, x_i) \sim \gamma_i} \tilde{c}_i x_i, & \nu_i \in \mathcal{P}(\mathcal{X}_i), \\ -\infty, & otherwise, \end{cases}$$

revealing that the mapping $\Pi_i\nu \mapsto \max_{\gamma_i \in \Gamma(\mu_i,\Pi_i\nu)} E_{(\tilde{c}_i,x_i)\sim\gamma_i}[\tilde{c}_i x_i]$ is indeed concave and weak-* upper semi-continuous. As a result, $\nu \mapsto \max_{\gamma_i \in \Gamma(\mu_i,\Pi_i\nu)} E_{(\tilde{c}_i,x_i)\sim\gamma_i}[\tilde{c}_i x_i]$ is concave and weak-* upper semi-continuous for all i, as desired. As for the left hand side "max", a constructive argument for this can be found in the coming Theorem 3.

Next, we establish the first equality. Let $\nu \in \mathcal{P}(\mathcal{X}), \gamma_i \in \Gamma(\mu_i, \Pi_i\nu)$. Then, writing $\eta = \gamma_1 \otimes \ldots \otimes \gamma_n$, define $\theta := \Pi_{\mathbb{R}^n}\eta \in \Gamma(\mu_1,\ldots,\mu_n)$. This constructed coupling in θ results in the following inequality:

$$\begin{aligned}\int_{\mathbb{R}^n} \max_{x\in\mathcal{X}} \sum_i \tilde{c}_i x_i d\theta &= \int_{\mathbb{R}^n\times\mathcal{X}} \max_{x\in\mathcal{X}} \sum_i \tilde{c}_i x_i d\eta,\\ &\geq \int_{\mathbb{R}^n\times\mathcal{X}} \sum_i \tilde{c}_i x_i d\eta(\tilde{c},x),\\ &= \int_{\mathbb{R}\times\mathcal{X}_1} \ldots \int_{\mathbb{R}\times\mathcal{X}_n} \sum_i \tilde{c}_i x_i d\gamma_n \ldots d\gamma_1,\\ &= \sum_i \int_{\mathbb{R}\times\mathcal{X}_i} \tilde{c}_i x_i d\gamma_i,\end{aligned}$$

which in summary gives the inequality:

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

Next, let $\theta \in \Gamma(\mu_1,\ldots,\mu_n)$. Then, taking any measurable selection x^* s.t. $x^*(c) \in \arg\max_{x\in\mathcal{X}} \sum_i c_i x_i$ $\forall c$, define $\nu := x^* \# \theta$ and $\gamma_i := ((Proj)_i, x_i^*)\#\theta \in \Gamma(\mu_i, \Pi_i\nu)$ to find:

$$\begin{aligned}E_{\tilde{c}\sim\theta}[\max_{x\in\mathcal{X}} \sum_{i=1}^n \tilde{c}_i x_i] &= \int_\Omega \sum_i \tilde{c}_i x_i^*(\tilde{c}) d\theta,\\ &= \sum_i \int_{\Omega_i\times\mathcal{X}_i} \tilde{c}_i x_i d\gamma_i,\\ &\leq \max_{\nu\in\mathcal{P}(\mathcal{X})} \quad \sum_i \max_{\gamma_i\in\Gamma(\mu_i,\Pi_i\nu)} E_{(\tilde{c}_i,x_i)\sim\gamma_i}[\tilde{c}_i x_i],\end{aligned}$$

where $Proj_i(\tilde{c}) := \tilde{c}_i$. This shows the first equality.

To establish the remaining equalities, let us note that for an arbitrary i, since $\tilde{c}_i \cdot x_i \leq \tilde{c}_i^2/2 + x_i^2/2$, for all $\tilde{c}_i$ and x_i, and since all marginal probability measures μ_i are assumed to have finite variance, applying Theorem 1 yields:

$$\begin{aligned}\max_{\gamma_i\in\Gamma(\mu_i,\Pi_i\nu)} E_{(\tilde{c},x)\sim\gamma_i}[\tilde{c}_i x_i] &= \inf_{(\psi_i,\phi_i):\phi_i(\tilde{c}_i)+\psi_i(x_i)\geq\tilde{c}_i x_i} \int \phi_i d\mu_i + \int \psi_i d\Pi_i\nu,\\ &= \inf_{(\psi_i):\mathcal{X}_i\to\mathbb{R}} \int \psi_i d\Pi_i\nu + \int max_{x_i}\{\tilde{c}_i x_i - \psi_i(x_i)\} d\mu_i.\end{aligned}$$

Hence,

$$\begin{aligned}
\max_{\theta\in\Gamma(\mu_1,\dots,\mu_n)} E_{\tilde{c}\sim\theta}[Z(\tilde{c})] &= \max_{\nu\in\mathcal{P}(\mathcal{X})} \sum_i \max_{\gamma_i\in\Gamma(\mu_i,\Pi_i\nu)} E_{(\tilde{c}_i,x_i)\sim\gamma_i}[\tilde{c}_i x_i], \\
&= \max_{\nu\in\mathcal{P}(\mathcal{X})} \sum_i \inf_{(\psi_i):\mathcal{X}_i\to\mathbb{R}} \int \psi_i d\Pi_i\nu + \int max_{x_i}\{\tilde{c}_i x_i - \psi_i(x_i)\} d\mu_i, \quad \text{(Theorem 1)} \\
&= \max_{\nu\in\mathcal{P}(\mathcal{X})} \inf_{\{\psi_i:\mathcal{X}_i\to\mathbb{R}\}_{i=1}^n} \sum_i \int \psi_i d\Pi_i\nu + \int max_{x_i}\{\tilde{c}_i x_i - \psi_i(x_i)\} d\mu_i, \\
&= \inf_{\{\psi_i:\mathcal{X}_i\to\mathbb{R}\}_{i=1}^n} \sup_{\nu\in\mathcal{P}(\mathcal{X})} \sum_i \int \psi_i d\Pi_i\nu + \int max_{x_i}\{\tilde{c}_i x_i - \psi_i(x_i)\} d\mu_i,
\end{aligned}$$

where the last equality is established using a "partial saddle point" result (see Proposition 2.3-Remark 2.3 in Ekeland and Temam (1999)). This result is valid because the following conditions are satisfied:

- $max_{x_i}\{\tilde{c}_i x_i - \psi_i(x_i)\}$ is convex in ψ_i,
- $max_{x_i}\{\tilde{c}_i x_i - \psi_i(x_i)\}$ is continuous (hence lower semi-continuous) in ψ_i (wrt $\|\cdot\|_\infty$),
- $\int \psi_i d\Pi_i\nu$ is linear in ν (hence concave),
- $\int \psi_i d\Pi_i\nu$ is continuous in ν and hence upper semicontinuous (w.r.t. norm of total variation),
- $\mathbb{R}^{\mathcal{X}_i}$ is reflexive for any i,
- $M(\mathcal{X})$, the space of signed radon measures (with norm of total variation) is reflexive, and the subset $\mathbb{P}(\mathcal{X})$ is bounded.

To conclude the proof, we show that the infimum in the dual problem is attained. This is true since $L(\cdot,\cdot)$ 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 (Slater's Condition / dual "strong consistency"), so that the infimum is indeed attained (see Corollary 30.5.2 in Rockafellar (1997)). □

EC.2.2. Proof of Theorem 3

THEOREM 3 *Let $\mathcal{X}$ be a finite set. Let $\tilde{c}_1,\dots,\tilde{c}_n$ be real-valued random variables with $\mu_1,\dots,\mu_n$ denoting the probability measures they induce on the real line. Then,*

$$\max_{\theta\in\Gamma(\mu_1,\dots,\mu_n)} E_{\tilde{c}\sim\theta} Z(\tilde{c}) = \max_{\nu\in\mathcal{P}(\mathcal{X})} \sum_i \max_{\gamma_i\in\Gamma(\mu_i,\Pi_i\nu)} E_{(\tilde{c}_i,x_i)\sim\gamma_i}\tilde{c}_i x_i \quad \text{(Primal)}$$

$$= \max_{\nu \in \mathcal{P}(\mathcal{X})} \quad \sum_i \sum_{x_i \in \mathcal{X}_i} x_i \int_{\Pi_i \nu((-\infty, x_i - e_i])}^{\Pi_i \nu((-\infty, x_i])} F_{\mu_i}^{-1}(t) dt. \tag{EC.1}$$

If ν_{rel} solves the concave maximization problem (8), there exists a maximizing joint distribution $\theta^* \in \Gamma(\mu_1, \ldots, \mu_n)$. Further, if $\mu_1, \ldots, \mu_n$ are absolutely continuous w.r.t. the Lebesgue measure, there exists some suitably defined measurable selection $x^* : \mathbb{R}^n \to \mathcal{X}$ of x^{OPT} s.t. $x^*(c) = (x_1^*(c_1), \ldots, x_n^*(c_n))$ and the "persistence values" $P_{\tilde{c} \sim \theta^*}(x_i^*(\tilde{c}) = x_i)$ are given by:

$$P_{\tilde{c} \sim \theta^*}(x_i^*(\tilde{c}_i) = x_i) = \Pi_i \nu_{rel}(x_i), \ \forall x_i \in \mathcal{X}_i, \ \forall i \in [n],$$

i.e., the x_i^* are Monge transport maps:

$$x_i^* \# \mu_i = \Pi_i \nu_{rel}, \ \forall i \in [n].$$

Proof: We begin by upper bounding $\max_{\theta \in \Gamma(\mu_1, \ldots, \mu_n)} E_{\tilde{c} \sim \theta} Z(\tilde{c})$ as follows:

$$\max_{\theta \in \Gamma(\mu_1, \ldots, \mu_n)} E_{\tilde{c} \sim \theta} Z(\tilde{c}) \le \max_{\nu \in \mathcal{P}(\mathcal{X})} \max_{\gamma \in \Gamma(\mu_1, \ldots, \mu_n, \Pi_1 \nu, \ldots \Pi_n \nu)} E_{(\tilde{c}, x) \sim \gamma} \tilde{c}^T x.$$

The maximization problem is indeed an upper bound. To see this, let a $\theta \in \Gamma(\mu_1, \ldots, \mu_n)$ be given. Then, given a measurable selection x^*, let us define a candidate γ and ν by:

$$\gamma = \theta \circ (id \times x^*)^{-1} \text{ and } \nu(B) = \gamma[\mathbb{R}^n \times B], \ \forall B \in \mathcal{B},$$

where $\circ$ is the composition operator and id is the identity mapping. It is easy to verify that $\gamma \in \Gamma(\mu_1, \ldots, \mu_n, \Pi_1 \nu, \ldots \Pi_n \nu)$ and $\nu \in \mathcal{P}(\mathcal{X})$. Now, observe:

$$\begin{aligned}
E_{\tilde{c} \sim \theta}[\tilde{c}^T x^*(\tilde{c})] &= \int_{\mathbb{R}^n} \tilde{c}^T x^*(\tilde{c}) \ d\theta(\tilde{c}), \\
&= \sum_{x' \in \mathcal{X}} \int_{\{\tilde{c} : x^*(\tilde{c}) = x'\}} \tilde{c}^T x' \ d\theta(\tilde{c}), \\
&= \sum_{x' \in \mathcal{X}} \int_{\{(\tilde{c}, x^*(\tilde{c})) : \tilde{c} \in \mathbb{R}^n, x^*(\tilde{c}) = x'\}} \tilde{c}^T x' \ \gamma(d\tilde{c} dx), \\
&= \int_{\{(\tilde{c}, x^*(\tilde{c})) : \tilde{c} \in \mathbb{R}^n\}} \tilde{c}^T x \ \gamma(d\tilde{c} dx), \\
&= \int_{\mathbb{R}^n \times \mathcal{X}} \tilde{c}^T x \ \gamma(d\tilde{c} dx), \\
&= E_{(\tilde{c}, x) \sim \gamma} \left[\tilde{c}^T x\right],
\end{aligned}$$

which shows that given any feasible θ to the left hand side maximization problem, there exists a feasible solution to the right hand size optimization problem with equal value. The intuition is that given a distribution for the random vector $\tilde{c}$, we can construct a distribution over $\mathbb{R}^n \times \mathcal{X}$ that is concentrated on the graph of x^* in a fashion that yields the same integral.

The function $f(c,x) = c^T x$ is a supermodular function in (c,x). Noting that:

$$E_{(\tilde{c},x)\sim\gamma} \tilde{c}^T x = \sum_i E_{(\tilde{c},x)}[\tilde{c}_i x_i],$$

we can apply Lemma 1 to each of these separable components. More precisely, the upper bound we derived above can be rewritten as:

$$\max_{\nu \in \mathcal{P}(\mathcal{X})} \max_{\gamma \in \Gamma(\mu_1,\ldots,\mu_n,\Pi_1\nu,\ldots\Pi_n\nu)} \sum_i E_{(\tilde{c}_i,x_i)\sim\gamma}[\tilde{c}_i x_i]. \tag{EC.2}$$

Upon relaxing the inner maximization we get the upper bound:

$$\max_{\nu \in \mathcal{P}(\mathcal{X})} \sum_i \max_{\gamma_i \in \Gamma(\mu_i,\Pi_i\nu)} E_{(\tilde{c}_i,x_i)\sim\gamma_i} \tilde{c}_i x_i \tag{Primal}$$

Now, a direct application of Lemma 1 to each of the inner i-subproblems in (Primal) yields the last optimization problem in the theorem statement:

$$\begin{aligned}\max_{\nu \in \mathcal{P}(\mathcal{X})} \sum_i \int_0^1 F_{\mu_i}^{-1}(t) F_{\Pi_i\nu}^{-1}(t)dt &= \max_{\nu \in \mathcal{P}(\mathcal{X})} \sum_i \int_0^1 F_{\mu_i}^{-1}(t) \cdot \sum_{x_i \in \mathcal{X}_i} x_i \cdot \mathbb{1}_{[\Pi_i\nu((-\infty,x_i-e_i])<t\le\Pi_i\nu((-\infty,x_i])}(t)dt,\\ &= \max_{\nu \in \mathcal{P}(\mathcal{X})} \sum_i \sum_{x_i \in \mathcal{X}_i} x_i \int_{\Pi_i\nu((-\infty,x_i-e_i])}^{\Pi_i\nu((-\infty,x_i])} F_{\mu_i}^{-1}(t)dt. \end{aligned} \tag{8}$$

Hence,

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

We now need to show that the inequalities are in fact tight. We construct a distribution $\theta^* \in \Gamma(\mu_1,\ldots,\mu_n)$ for the problem $\max_{\theta \in \Gamma(\mu_1,\ldots,\mu_n)} E_{\tilde{c}\sim\theta} Z(\tilde{c})$, and show that under this distribution, the value of the objective equals the upper bound on the right side of the inequality presented by (8). Letting $\nu_{rel} \in \mathcal{P}(\mathcal{X})$ be an optimal solution to the right hand side of the inequality, we will construct a measure θ^* of the following (disintegrated) form:

$$\theta^* = \sum_{x \in \mathcal{X}} (\mu_1^{x_1} \otimes \ldots \otimes \mu_n^{x_n}) \cdot \nu_{rel}(x),$$

where it is implicit that $(\theta^*)^x = (\mu_1^{x_1} \otimes \ldots \otimes \mu_n^{x_n})$. In other words, with $\tilde{c} \sim \theta^*$, the $\tilde{c}_i$ are conditionally independent given x. In fact, this form also reveals that for any i, $\tilde{c}_i$ is conditionally independent of all x_j for $j \neq i$. With θ^* of this form, what remains is to define the measures $\mu_i^{x_i}$ for all i and $x_i \in \mathcal{X}_i$. Define for all i and $x_i \in \mathcal{X}_i$,

$$I_{i,x_i} := \{c_i \in \mathbb{R} : L_{i,x_i} \leq c_i \leq R_{i,x_i}\},$$

where $L_{i,x_i} := F_{\mu_i}^{-1}(\Pi_i \nu_{rel}((-\infty, x_i - e_i])$ and $R_{i,x_i} := F_{\mu_i}^{-1}(\Pi_i \nu_{rel}((-\infty, x_i])$. Further, to capture the (possible) discontinuous jumps in F_{μ_i} at L_{i,x_i} and R_{i,x_i}, define the following nonnegative quantities:

$$J_{i,x_i}^L := \begin{cases} F_{\mu_i}(L_{i,x_i}) - \Pi_i \nu_{rel}(-\infty, x_i - e_i], & F_{\mu_i}(L_{i,x_i}) \leq \Pi_i \nu_{rel}(-\infty, x_i], \\ 0, & \text{otherwise}, \end{cases}$$

$$J_{i,x_i}^R := \Pi_i \nu_{rel}(-\infty, x_i] - \lim_{c_i \uparrow R_{i,x_i}} .F_{\mu_i}(c_i)$$

Define the measure as follows:

1. Define the desired measure $\mu_i^{x_i}$, for $x_i : \Pi_i \nu_{rel}(x_i) > 0$, by:

$$\mu_i^{x_i}(B) := \frac{J_{i,x_i}^L \cdot \mathbb{1}_{L_{i,x_i} \in B} + \mu_i(B \cap (L_{i,x_i}, R_{i,x_i})) + J_{i,x_i}^R \cdot \mathbb{1}_{R_{i,x_i} \in B}}{\Pi_i \nu_{rel}(x_i)} \leq 1, \quad \forall B \in \mathcal{B}$$

One can verify that this is indeed a probability measure, with $I_{i,\alpha_i^{(k)}}$ as support, as:

$$\mu_i^{x_i}(I_{i,x_i}) = \frac{J_{i,x_i}^L + J_{i,x_i}^R + \mu_i((L_{i,x_i}, R_{i,x_i}))}{\Pi_i \nu_{rel}(x_i)} = \frac{\Pi_i \nu_{rel}(\infty, x_i] - \Pi_i \nu_{rel}(-\infty, x_i - e_i]}{\Pi_i \nu_{rel}(x_i)} = 1.$$

2. We can define $\mu_i^{x_i}$, for $x_i : \Pi_i \nu_{rel}(x_i) = 0$ arbitrarily.

One can verify that this definition satisfies the marginal constraint:

$$\Pi_i \theta^* = \Pi_i \left[\sum_{x \in \mathcal{X}} (\mu_1^{x_1} \otimes \ldots \otimes \mu_n^{x_n}) \cdot \nu_{rel}(x) \right] = \sum_{x \in \mathcal{X}} \mu_i^{x_i} \cdot \nu_{rel}(x) = \mu_i, \quad \forall i,$$

so that θ^* indeed is an admissible joint distribution, i.e., $\theta^* \in \Gamma(\mu_1, \ldots, \mu_n)$. The intuition comes from Lemma 1, which tells us that an optimal coupling of the random variables $\tilde{c}_i \sim \mu_i$ and $\tilde{x}_i \sim \Pi_i \nu_{rel}$ can be obtained if we "match up" quantiles. Thus, the above simply proposes that we concentrate

$\tilde{c}_i$ on the interval I_{i,x_i}, given knowledge that the random variable $\tilde{x} \sim \nu_{rel}$ has a realization x with i-th component equal to x_i.

We check tightness where all expectation operators are taken with respect to the constructed distribution θ^*:

$$
\begin{aligned}
E_{\tilde{c}}[Z(\tilde{c})] &\geq \sum_{x \in \mathcal{X}} \nu_{rel}(x) \cdot E[\sum_i \tilde{c}_i x_i \| \tilde{x} = x], \\
&= \sum_{x \in \mathcal{X}} \sum_i \nu_{rel}(x) \cdot x_i \cdot E[\tilde{c}_i \| \tilde{x} = x], \\
&= \sum_{x \in \mathcal{X}} \sum_i \nu_{rel}(x) \cdot x_i \cdot E[\tilde{c}_i \| \tilde{x}_i = x_i], \\
&= \sum_{x \in \mathcal{X}} \sum_i \nu_{rel}(x) \cdot x_i \cdot \frac{E[\tilde{c}_i \cdot \mathbb{1}_{\tilde{x}_i = x_i}]}{\Pi_i \nu_{rel}(x_i)}, \\
&= \sum_{x \in \mathcal{X}} \sum_i \nu_{rel}(x) \cdot x_i \cdot \frac{\int \tilde{c}_i \cdot \mathbb{1}_{\tilde{x}_i = x_i} \cdot (\mu_i^{x_i} \otimes \Pi_i \nu_{rel})(d\tilde{c}_i d\tilde{x}_i)}{\Pi_i \nu_{rel}(x_i)}, \\
&= \sum_i \sum_{x_i \in \mathcal{X}_i} \sum_{x' \in \mathcal{X} : x'_i = x_i} \nu_{rel}(x') \cdot x_i \cdot \frac{\int c_i \cdot (\mu_i^{x_i})(dc_i) \Pi_i \nu_{rel}(x_i)}{\Pi_i \nu_{rel}(x_i)}, \\
&= \sum_i \sum_{x_i \in \mathcal{X}_i} \sum_{x' \in \mathcal{X} : x'_i = x_i} \nu_{rel}(x') \cdot x_i \cdot \frac{\int_{\Pi_i \nu_{rel}((-\infty, x_i - e_i])}^{\Pi_i \nu_{rel}((-\infty, x_i])} F_{\mu_i}^{-1}(t) dt}{\Pi_i \nu_{rel}(x_i)}, \\
&= \sum_i \sum_{x_i \in \mathcal{X}_i} x_i \int_{\Pi_i \nu_{rel}((-\infty, x_i - e_i])}^{\Pi_i \nu_{rel}((-\infty, x_i])} F_{\mu_i}^{-1}(t) dt \sum_{x' : x'_i = x_i} \frac{\nu_{rel}(x')}{\Pi_i \nu_{rel}(x_i)}, \\
&= \sum_i \sum_{x_i \in \mathcal{X}_i} x_i \int_{\Pi_i \nu_{rel}((-\infty, x_i - e_i])}^{\Pi_i \nu_{rel}((-\infty, x_i])} F_{\mu_i}^{-1}(t) dt \cdot \frac{\Pi_i \nu_{rel}(x_i)}{\Pi_i \nu_{rel}(x_i)}, \\
&= \sum_i \sum_{x_i \in \mathcal{X}_i} x_i \int_{\Pi_i \nu_{rel}((-\infty, x_i - e_i])}^{\Pi_i \nu_{rel}((-\infty, x_i])} F_{\mu_i}^{-1}(t) dt,
\end{aligned}
$$

where the inequality at the top is justified by:

$$
\begin{aligned}
& Z(\tilde{c}) = max\{\tilde{c}^T x : x \in \mathcal{X}\} \ \geq \ \tilde{c}^T x, \ \forall x \in \mathcal{X}, \\
\Longrightarrow\ & E[Z(\tilde{c}) \| \tilde{x} = x] \geq E[\sum_i \tilde{c}_i x_i \| \tilde{x} = x], \ \forall x \in \mathcal{X}, \\
\Longrightarrow\ & E_{\tilde{c} \sim \theta^*}[Z(\tilde{c})] \geq E[E[\sum_i \tilde{c}_i x_i \| \tilde{x} = x]] = \sum_{x \in \mathcal{X}} \nu_{rel}(x) \cdot E[\sum_i \tilde{c}_i x_i \| \tilde{x} = x].
\end{aligned}
$$

We have hence shown that the upper bound is indeed tight.

□

One final point for the case when $\mu_1, \ldots, \mu_n$ represent continuous distributions; since now we know the inequality should in fact be equality:

$$
\begin{aligned}
& E[Z(\tilde{c})] = EE[\tilde{c}^T x \| \tilde{x} = x] \\
\Longrightarrow\ & EE[Z(\tilde{c}) \| \tilde{x} = x] = EE[\tilde{c}^T x \| \tilde{x} = x] \\
\Longrightarrow\ & E[Z(\tilde{c}) \| \tilde{x} = x] = E[\tilde{c}^T x \| \tilde{x} = x] \ \ \forall x \in \mathcal{X} : \nu_{rel}(x) > 0 \\
\Longrightarrow\ & Z(\tilde{c})|_{\tilde{x}=x} = \tilde{c}^T x|_{\tilde{x}=x} \ \ a.e., \forall x \in \ : \nu_{rel}(x) > 0
\end{aligned}
$$

This last line tells us that $x \in argmax_{x \in \mathcal{X}} c^T x$ for all $c \in supp(\tilde{c}|_{\tilde{x}=x})$. So in order to create a measurable mapping $x^*(\cdot) : \mathbb{R}^n \to \mathcal{X}$ that maps every $c \in \mathbb{R}^n$ to a maximizing solution, we could define one via:

$$\text{For all i,} \quad x_i^*(c_i) := x_i, \ \text{for } c_i \in I_{i, x_i}.$$

Then this yields:

$$x^*(c) := x, \ \forall c \in supp(\tilde{c}|_{\tilde{x}=x}) = \{\hat{c} : \hat{c}_i \in I_{i,x_i} \ \ \forall i\} = I_{1,x_1} \times \ldots \times I_{n,x_n},$$

so that it is guaranteed that $x^*(c) \in argmax_{x \in \mathcal{X}} c^T x$ whenever $c \in supp(\tilde{c}|_{\tilde{x}=x})$ for some x. When $c \notin supp(\tilde{c}|_{\tilde{x}=x})$ for any x, $x^*(\cdot)$ will be defined arbitrarily (say a constant, to guarantee measurability).

Remark EC.1. The above may appear to be not well-defined since there may exist c that lie in the intersection $I_{1,x_1} \times \ldots \times I_{n,x_n} \cap I_{1,x_1'} \times \ldots \times I_{n,x_n'}$ for a pair $x, x' \in \mathcal{X}$. But such cases are negligible in the case of atomless $\mu_1, \ldots, \mu_n$, because any such c must lie on the boundary of both hyperrectangles, which means the inconsistencies present themselves in regions that have Lebesgue zero-measure.

Thus,

$$
\begin{aligned}
& x^*(\tilde{c}) \|_{\tilde{x}=x} = x \ \ (\forall x \in \mathcal{X} : \nu_{rel}(x) > 0) \\
\Longrightarrow\ & P_{\tilde{c} \sim \theta^*}(x_i^*(\tilde{c}) = x_i) = \sum_{x' : \nu_{rel}(x') > 0} P(x_i^*(\tilde{c}) = x_i \| \tilde{x} = x)\, \nu_{rel}(x') \\
& = \sum_{x' : \nu_{rel}(x') > 0} \mathbb{1}_{x_i' = x_i}\, \nu_{rel}(x') = \Pi_i \nu_{rel}(x_i)
\end{aligned}
$$

In summary, we can read off the persistence values for an appropritely defined "solution" mapping straight off the solved-for ν_{rel}.

REMARK EC.2. The reason this last discussion was limited to continuous distributions was because of the following example. Consider:

- n = 1
- $\mu_1 = \mathbb{1}_0$
- $\mathcal{X} = \{0,1\}$

Then any distribution over $\mathcal{X}$ can solve (8). Let us consider one in particular in ν_{rel} defined by $\nu_{rel}(0) = \nu_{rel}(1) = 1/2$. Then it is clear that no solution mapping x^* exists s.t.

$$P_{\tilde{c}\sim\theta^*}(x^*(\tilde{c}) = 0) = \nu_{rel}(0),$$

and

$$P_{\tilde{c}\sim\theta^*}(x^*(\tilde{c}) = 1) = \nu_{rel}(1)$$

Indeed, the previous remark's justification that the boundary of the hyperrectangles having zero-measure no longer holds here. △

EC.3. Section 5 Proofs

EC.3.1. Proof of Proposition 2

PROPOSITION 2 *For any s, $Z(s,\tilde{c})$ is supermodular in the $\tilde{c}$ variables.*

Proof: We will fix an arbitrary scheduling variable s and omit it from our notation during this proof. By Proposition 2.2.2 of Simchi-Levi, Chen, and Bramel (2014), submodularity on $\mathbb{R}^n$ 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 $\hat{c}_{ij} = (c_1,\ldots,c_{i-1},c_{i+1}\ldots,c_{j-1},c_{j+1},\ldots,c_n) \in \mathbb{R}^{n-2}$.

$$Z_{\hat{c}_{ij}}(c_i,c_j') - Z_{\hat{c}_{ij}}(c_i,c_j) \le Z_{\hat{c}_{ij}}(c_i',c_j') - Z_{\hat{c}_{ij}}(c_i',c_j) \enspace .$$

For our analysis, let us define the notion of a "late-chain". Given $(c_1,\ldots,c_n)$, we will call a finite sequence of consecutive integers $\{k,\ldots,l\}$ a late-chain in $(c_1,\ldots,c_n)$ if every patient $j \in \{k,\ldots,l\}$

experiences positive wait time. Observe that $\{1,\ldots,n\}$ can be partitioned into maximal late-chains as well as consecutive sequences of integers wherein all members have 0 wait time.

Now, let us analyze the LHS= $Z_{\hat{c}_{ij}}(c_i,c_j') - Z_{\hat{c}_{ij}}(c_i,c_j)$. Starting with $(c_1,\ldots,c_n)$, let us consider what happens when c_j is increased to c_j'. The LHS is zero if and only if j is not in a late chain in $(c_1,\ldots,c_n)$ and j is not in a late chain in $(c_1,\ldots,c_{j-1},c_j',c_{j+1},\ldots,c_n)$. As it is clear that Z is montonic in the c variables for any s, we thus need only concern ourselves with the case that $c_j < c_j'$ and that the change is sufficiently large for j to join a late-chain if it is not already part of one. Such a change would mean the LHS scales with (1 + # of patients following j that are in the maximal late-chain in $(c_1,\ldots,c_{j-1},c_j',c_{j+1},\ldots,c_n)$ with j).

Now when considering the RHS, in which c_i increases to c_i', observe that any maximal late chain in $(c_1,\ldots,c_{i-1},c_i,c_{i+1},\ldots,c_n)$ is contained in some maximal late chain in $(c_1,\ldots,c_{i-1},c_i',c_{i+1},\ldots,c_n)$. Using the same analysis as we did for the LHS, we see that since the maximal late chains are no smaller, the RHS can be no smaller as well, as desired.

□

EC.3.2. Proof of Lemma 2

LEMMA 2 *Let $\mathcal{X}_i$ be a totally ordered set for all i, and consider $\mathcal{X} = \mathcal{X}_1 \ldots \mathcal{X}_n$ Let $c : \mathcal{X} \to \mathbb{R}$ be a continuous, submodular function. If $\{\mu_i\}_{i=1\ldots n}$ is any collection of marginal measures, where $\mu_i \in \mathcal{P}(\mathcal{X}_i)$ for all i,*

$$\inf_{\gamma:\ \Pi_i\gamma=\mu_i\ (\forall i)} \int_{\mathcal{X}:=\mathcal{X}_1\times\ldots\times\mathcal{X}_n} c(x)\ d\gamma(x) = \int_0^1 c(F_{\mu_1}^{-1}(t),\ldots,F_{\mu_n}^{-1}(t))\, dt,$$

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

Proof: Given the probability space, $([0,1],\mathcal{B},\lambda)$, where λ is the lebesgue measure restricted to [0,1], let us define the monotone coupling by $\gamma_{mon} := (F_{\mu_1}^{-1},\ldots,F_{\mu_n}^{-1})\#\lambda \in \mathcal{P}(\mathbb{R}^n)$, where

$$F_{\mu_i}^{-1}(t) := \inf\{y : F_{\mu_i}(y) \ge t\}, \quad \forall t \in [0,1]$$

Notice then that by right-continuity, $F_{\mu_i}\left(F_{\mu_i}^{-1}(t)\right) \ge t$. Further, we find $F_{\mu_i}^{-1}(t) \le x_i \iff t \le F_{\mu_i}(x_i)$.

The plan is first to show that any measure $\gamma \in \Gamma(\mu_1, \dots, \mu_n)$ whose support is totally-ordered, i.e.,

$$(\forall y, y' \in supp(\gamma)) \quad y \le y' \text{ OR } y \ge y', \tag{EC.3}$$

must be equal to γ_{mon}. To arrive at this conclusion, it suffices to show that any such γ agrees with γ_{mon} on the semiring of "downward-closed rectangles" $\mathcal{S}$, since, both being σ-finite, would mean they agree on $\sigma(\mathcal{S})$, the family of Borel sets of $\mathbb{R}^n$.

Following through with this plan, we observe that we immediately have the action of γ_{mon} on any downward-closed rectangle. More precisely, for any $x \in \mathbb{R}^n$,

$$\gamma_{mon}(\{y : y \le x\}) = \lambda\left(\{t : F_{\mu_i}^{-1}(t) \le x_i \ \forall i\}\right) = \lambda(\{t : t \le F_{\mu_i}(x_i) \ \forall i\}) = \min_i F_{\mu_i}(x_i).$$

For comparison, we now investigate the action of an arbitrary $\gamma \in \Gamma(\mu_1, \dots, \mu_n)$, whose support satisfies the totally-ordered condition in (EC.3), on a downward-closed rectangle $\{y : y \le x\}$, for some arbitrary $x \in \mathbb{R}^n$. Towards this, let us introduce the following definitions:

- $x^{\downarrow} := \{y : y \le x\}$
- $x^{\uparrow} := \{y : y > x\}$
- Given any partition $P_x^{\le} \cup P_x^{>} = \{1, \dots, n\}$, where $P_x^{\le} \ne \varnothing$ and $P_x^{>} \ne \varnothing$, write

$$P_x := \{y : y_i \le x_i \ \forall i \in P_x^{\le}, \ y_i > x_i \ \forall i \in P_x^{>}\}.$$

And let $\mathcal{P}_x$ denote the family of all such sets P_x.

- Observe that $\mathbb{R}^n = x^{\downarrow} \cup x^{\uparrow} \cup \dot{\bigcup}_{P_x \in \mathcal{P}_x} P_x$. But because of (EC.3), $\dot{\bigcup}_{P_x \in \mathcal{P}_x} P_x \not\subset supp(\gamma)$. Instead, there exists a subcollection $\mathcal{P}_x^{comp} \subset \mathcal{P}_x$ that satisfies

$$\left(\forall y, y' \in \dot{\bigcup_{P_x \in \mathcal{P}_x^{comp}}} P_x\right) \quad y \le y' \text{ OR } y \ge y'$$

and

$$\gamma(\dot{\bigcup_{P_x \in \mathcal{P}_x}} P_x) = \gamma(\dot{\bigcup_{P_x \in \mathcal{P}_x^{comp}}} P_x). \tag{EC.4}$$

Let us assume that this quantity is greater than zero, otherwise $\gamma(x^{\downarrow}) = \gamma_{mon}(x^{\downarrow})$ trivially. Define $A := \{i : i \in P_x^{\le}, \text{ some } P_x \in \mathcal{P}_x^{comp}\}$.

With the definitions set, let us first observe that for any $i \notin A$,

$$\gamma(\{y : y_i \leq x_i\}) + \gamma(\{y : y_i > x_i\}) = 1$$
$$\Longrightarrow \gamma(\{y : y_i \leq x_i\}) + \gamma(x^{\uparrow}) + \gamma(\dot{\bigcup}_{P_x \in \mathcal{P}_x^{comp}} P_x) = 1$$

This tells us that

$$\gamma(\{y : y_i \leq x_i\}) = \gamma(\{y : y_j \leq x_j\}) \quad \forall i, j \notin A \tag{EC.5}$$

Further, we observe that

$$\gamma(x^{\downarrow}) = \gamma(\{y : y_i \leq x_i, i \notin A\}) \tag{EC.6}$$

because by definition of A, any y that satisfies $y_i \leq x_i$ for all $i \notin A$ and $y_i > x_i$ for some $i \in A$, cannot be in $\dot{\bigcup}_{P_x \in \mathcal{P}_x^{comp}} P_x$, and hence by (EC.4), it does not belong in the support of γ. By the same reasoning,

$$supp(\gamma) \cap \{y : y_j \leq x_j, \forall j \notin A\} \subseteq \{y : y_i \leq x_i\} \cap supp(\gamma) \subseteq x^{\downarrow} \cap supp(\gamma) \quad \forall i \notin A$$

implies that

$$\gamma(x^{\downarrow}) = \gamma(\{y : y_i \leq x_i, i \notin A\}) \leq \gamma(\{y : y_i \leq x_i\}) \leq \gamma(x^{\downarrow}) \quad \forall i \notin A,$$

so conclude

$$\gamma(\{y : y_i \leq x_i\}) = \gamma(x^{\downarrow}), \quad \forall i \notin A \tag{EC.7}$$

Finally, let $i \notin A$ and $j \in A$ be arbitrary. Then

$$\gamma(\{y : y_i \leq x_i\}) = \gamma(\{y : y_j \leq x_j\} \cap x^{\downarrow}) < \gamma(\{y : y_j \leq x_j\} \cap x^{\downarrow}) + \gamma(\{y : y_j \leq x_j\} \cap \dot{\bigcup}_{P_x \in \mathcal{P}_x} P_x) = \gamma(\{y : y_j \leq x_j\}) \tag{EC.8}$$

Altogether, (EC.7), (EC.8) and (EC.5) indicate that $\gamma(x^{\downarrow}) = \min_i F_{\mu_i}(x_i)$, as desired.

What remains is to show that any optimal coupling γ^* must satisfy the totally-ordered condition of (EC.3). For this, we cite the following lemma:

Lemma EC.1 **(Lemma 2.2 from Pass (2012)).** *Let $c : \mathcal{X}_1 \times \ldots \times \mathcal{X}_n \to \mathbb{R}$ be a continuous cost function, with the $\mathcal{X}_i$ being smooth manifolds with given corresponding Borel probability measures μ_i. Then with regards to the optimization problem,*

$$\min_{\gamma \in \Gamma(\mu_1, \ldots, \mu_n)} \int c \; d\gamma,$$

if the optimal value is $< \infty$, then for any optimal solution γ^*, supp(γ^*) necessarily satisfies the following: Given any $y = (y_1, \ldots, y_n)$ and $y' = (y'_1, \ldots, y'_n)$ in supp(γ^*), then for any partition of $\{1, \ldots, n\}$, written $p = (p_+, p_-)$,

$$c(y) + c(y') \le c(z^p) + c(z'^p),$$

where $z_i^p = y_i$, $z_i'^p = y'_i$ if $i \in p_+$, and $z_i^p = y'_i$, $z_i'^p = y_i$ if $i \in p_-$.

By this lemma,

$$c(y) + c(y') \le c(y \vee y') + c(y \wedge y') \;\; \forall y, y' \in supp(\gamma^*),$$

which yields a contradiction to the strict submodularity of c if there exists a pair y and y' in the support of γ^* that cannot be related by the canonical component-wise partial-order $\le$. Hence, we conclude that any optimal coupling for a continuous, strictly submodular cost function must be equal to γ_{mon}.

What remains is a basic limiting argument. Indeed, by the monotone convergence theorem, letting m tend to infinity,

$$\int c + \frac{1}{m} \sum_{i \ne j} (x_i - x_j)^2 \; d\gamma_{mon} \le \inf_\gamma \int c + \frac{1}{m} \sum_{i \ne j} (x_i - x_j)^2 \; d\gamma$$
$$\Longrightarrow \int c \; d\gamma_{mon} \le \inf_\gamma \int c \; d\gamma,$$

as desired. □

EC.3.3. Proof of Proposition 3

Proposition 3 *For any t, $Z(t, \tilde{c})$ is monotone, submodular in the $\tilde{c}$ variables.*

Proof: For simplicity we absorb $\mathbf{t}$ into $\mathbf{c}$, considering $c_i - t_i$ as a whole in the following argument, and with a bit abuse of notation, we use c_i to denote $c_i - t_i$. Regarding monotonicity, this is clear. By Proposition 2.2.2 of Simchi-Levi, Chen, and Bramel (2014), submodularity on $\mathbb{R}^n$ 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 $i, j \in \{1, 2, \dots, n\}$ such that $i < j$, $c_i, c_i', c_j, c_j' \in \mathbb{R}$ with $c_i \leq c_i'$ and $c_j \leq c_j'$, and for any $\hat{c}_{ij} = (c_1, \dots, c_{i-1}, c_{i+1} \dots, c_{j-1}, c_{j+1}, \dots, c_n) \in \mathbb{R}^{n-2}$.

$$Z_{\hat{c}_{ij}}(c_i, c_j') - Z_{\hat{c}_{ij}}(c_i, c_j) \geq Z_{\hat{c}_{ij}}(c_i', c_j') - Z_{\hat{c}_{ij}}(c_i', c_j) \tag{EC.9}$$

In the lines of argument to follow, we will denote by

- x:= The position of the number c_j' among the numbers $(c_1, \dots, c_l, c_{l+1}, \dots, c_j', \dots, c_n)$ when enumerating them from smallest to largest.

- y:= The position of the number c_j among the numbers $(c_1, \dots, c_l, c_{l+1}, \dots, c_j, \dots, c_n)$ when enumerating them from smallest to largest.

Case 1 $(c_j < c_i \leq c_j', c_j' < c_i')$ *In this case, the LHS of (EC.9)*

$$Z_{\hat{c}_{ij}}(c_i, c_j') - Z_{\hat{c}_{ij}}(c_i, c_j) = c_j' \cdot x - c_j \cdot y - \sum_{k: c_j \leq c_k \leq c_j'} c_k$$

while the RHS of (EC.9)

$$Z_{\hat{c}_{ij}}(c_i', c_j') - Z_{\hat{c}_{ij}}(c_i', c_j) = c_j' \cdot (x-1) - c_j \cdot y - \sum_{k: c_j \leq c_k \leq c_j', k \neq i} c_k$$

so that $LHS - RHS = c_j' - c_i \geq 0$

Case 2 $(c_j < c_i \leq c_j', c_j \leq c_i' \leq c_j')$ *In this case, the LHS*

$$Z_{\hat{c}_{ij}}(c_i, c_j') - Z_{\hat{c}_{ij}}(c_i, c_j) = c_j' \cdot x - c_j \cdot y - \sum_{k: c_j \leq c_k \leq c_j'} c_k$$

while the RHS

$$Z_{\hat{c}_{ij}}(c_i', c_j') - Z_{\hat{c}_{ij}}(c_i', c_j) = c_j' \cdot x - c_j \cdot y - \sum_{k: c_j \leq c_k \leq c_j'} c_k$$

so that $LHS - RHS = 0$

Case 3 $(c_i > c_j', c_i' > c_j')$ *In this case, the LHS*

$$Z_{\hat{c}_{ij}}(c_i, c_j') - Z_{\hat{c}_{ij}}(c_i, c_j) = c_j' \cdot x - c_j \cdot y - \sum_{k: c_j \le c_k \le c_j'} c_k$$

while the RHS

$$Z_{\hat{c}_{ij}}(c_i, c_j') - Z_{\hat{c}_{ij}}(c_i, c_j) = c_j' \cdot x - c_j \cdot y - \sum_{k: c_j \le c_k \le c_j'} c_k$$

so that $LHS - RHS = 0$

Case 4 $(c_i < c_j, c_i' > c_j' \ge c_j)$ *I n this case, the LHS*

$$Z_{\hat{c}_{ij}}(c_i, c_j') - Z_{\hat{c}_{ij}}(c_i, c_j) = c_j' \cdot x - c_j \cdot y - \sum_{k: c_j \le c_k \le c_j'} c_k$$

while the RHS

$$Z_{\hat{c}_{ij}}(c_i, c_j') - Z_{\hat{c}_{ij}}(c_i, c_j) = c_j' \cdot (x-1) - c_j \cdot (y-1) - \sum_{k: c_j \le c_k \le c_j'} c_k$$

so that $LHS - RHS = c_j' - c_j \ge 0$

Case 5 $(c_i < c_j, c_j \le c_i' \le c_j')$ *n this case, the LHS*

$$Z_{\hat{c}_{ij}}(c_i, c_j') - Z_{\hat{c}_{ij}}(c_i, c_j) = c_j' \cdot x - c_j \cdot y - \sum_{k: c_j \le c_k \le c_j'} c_k$$

while the RHS

$$Z_{\hat{c}_{ij}}(c_i, c_j') - Z_{\hat{c}_{ij}}(c_i, c_j) = c_j' \cdot x - c_j \cdot (y-1) - \sum_{k: c_j \le c_k \le c_j'} c_k - c_i'$$

so that $LHS - RHS = -c_j + c_i' \ge 0$

□

EC.4. MMM model for the Ranking with Scheduling problem

Define

$$u_{ij} = \mathbb{P}(y_{ij}(\tilde{\mathbf{c}}) = 1)$$

$$w_{ij} = \mathbb{E}[\tilde{c}_i \mid y_{ij}(\tilde{\mathbf{c}}) = 1]\mathbb{P}(y_{ij}(\tilde{\mathbf{c}}) = 1)$$

$$z_{ij} = \mathbb{E}[\tilde{c}^2 \mid y_{ij}(\tilde{\mathbf{c}}) = 1]\mathbb{P}(y_{ij}(\tilde{\mathbf{c}}) = 1)$$

Consider the following model

$$\begin{aligned} \max\ & \sum_{i=1}^{n}\sum_{j=1}^{n} jw_{ij} - t_i j u_{ij} \\ s.t.\ & \sum_{j=1}^{n} u_{ij} = 1, \forall i = 1,\dots,n \\ & \sum_{j=1}^{n} w_{ij} = \mu_i, \forall i = 1,\dots,n \\ & \sum_{j=1}^{n} z_{ij} = \mu_i^2 + \sigma_i^2, \forall i = 1,\dots,n \\ & \sum_{i=1}^{n} u_{ij} = 1, \forall j = 1,\dots,n \\ & \begin{pmatrix} u_{ij} & w_{ij} \\ w_{ij} & z_{ij} \end{pmatrix} \succeq 0, \forall i,j = 1,\dots,n \end{aligned} \tag{EC.10}$$

Denote the dual variable in (EC.10) as α_i, β_i, ξ_i, for $i = 1,\dots,n$ and $\epsilon_j, \forall j = 1,\dots,n$. Consider the dual of (EC.10).

$$\begin{aligned} \min\ & \sum_{i=1}^{n} \left(\alpha_i + \beta_i\mu_i + \xi_i(\mu_i^2 + \sigma_i^2)\right) + \sum_{j=1}^{n} \epsilon_j \\ s.t.\ & \begin{pmatrix} \alpha_i + \epsilon_j + t_i j & \frac{1}{2}(\beta_i - j) \\ \frac{1}{2}(\beta_i - j) & \xi_i \end{pmatrix} \succeq 0, \forall i,j = 1,\dots,n \end{aligned} \tag{EC.11}$$

When solving the first stage problem, we can incorporate the constraints on $\mathbf{t}$ in (EC.11). Notice $\begin{pmatrix} \alpha_i + \epsilon_j + t_i j & \frac{1}{2}(\beta_i - j) \\ \frac{1}{2}(\beta_i - j) & \xi_i \end{pmatrix} \succeq 0$ can be equivalently written as a socp constraint as

$$\sqrt{(\beta_i - j)^2 + (\alpha_i + \epsilon_j + t_i j - \xi_i)^2} \le \alpha_i + \epsilon_j + t_i j + \xi_i$$

References

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

R. K. Ahuja, T. L. Magnanti and J. B. Orlin. (1993). *Network flows: theory, algorithms, and applications.* Prentice Hall.

F. Bach. (2018). *Submodular Functions: from Discrete to Continous Domains.* Mathematical Programming.

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

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

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

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

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

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

J. Blanchet, Y. Kang, F. Zhang and K. Murthy. (2017). *Data-driven optimal transport cost selection for distributionally robust optimization.*

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

G. Carlier and I. Ekeland. (2010). *Matching for teams.* Economic Theory, 42(2), 397-418.

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

I. Ekeland and R. Temam. (1999). *Convex analysis and variational problems.* SIAM Classics in Applied Mathematics, Philadelphia.

P. M. Esfahani and D. Kuhn. (2017). *Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations.* Mathematical Programming, 1-52.

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

R. Gao and A. J. Kleywegt. (2016). *Distributionally robust stochastic optimization with Wasserstein distance.*

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

A. N. Kolmogorov and S. V. Fomin. (1970). *Introductory real analysis*. Translated and edited by R. A. Silverman. Englewood Cliffs, N.J: Prentice-Hall.

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

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

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

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

I. Meilijson (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.

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

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

R. H. Möhring, A. S. Schulz, F. Stork and M. Uetz. (2001). *On project scheduling with irregular starting time costs*. Operations Research Letters, 28, 149154.

Murota. (2003). *Discrete Convex Analysis*. SIAM, Philadelphia.

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

K. Natarajan, D. Shi and K-C. Toh. (2017). *Bounds for random binary quadratic programs*. To appear in SIAM Journal on Optimization.

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

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

B. Pass. (2012). *On the local structure of optimal measures in the multi-marginal optimal transportation problem.* Calculus of Variations, 43: 529-536.

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

R. T. Rockafellar. (1997). *Convex analysis.* Princeton University Press.

H. L. Royden and P. M. Fitzpatrick. (2010). *Real analysis.* Fourth Edition, Prentice Hall.

F. Santambrogio. (2015). *Optimal transport for applied mathematicians.* Birkhäuser Basel.

H. Scarf. (1958). *A min-max solution of an inventory problem*, Studies in The Mathematical Theory of Inventory and Production. (K. Arrow, S. Karlin and H. Scarf, Eds.) 201-209, Stanford University Press, California.

A. Shapiro, D. Dentcheva and A. Ruszczyński. (2014). *Lectures on stochastic programming: Modeling and theory.* Second edition. MOS-SIAM Series on Optimization.

D. Simchi-Levi, X. Chen, J. Bramel. (2014). *The Logic of Logistics.* Third Edition, Springer.

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

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

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

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

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

W. I. Zangwill. (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.