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 multiple sufficient conditions for polynomial time solvability - one using extended formulations, another exploiting the interaction of combinatorial structure and optimal transport. 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 et al. (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:

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

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:

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

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 et al. (2004)), the set of distributions with confidence sets and mean values residing on an affine
manifold (see Wiesemann et al. (2014)), the set of distributions that lie in a ball around a reference probability distribution where the distance is defined using the $\phi$-divergence measure (see Ben-Tal (2013)) or the Wasserstein distance measure (see Esfahani and Kuhn (2017), Gao and Kleywegt (2016), Blanchet et al. (2017)). 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 $h(s, \tilde{\xi})$ 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 et al. (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 $\mu_i$ 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 $\mu_i$. 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:

\[
\min_{s \in S} \sup_{\theta \in \Gamma} \mathbb{E}_{\tilde{c} \sim \theta}[Z(s, \tilde{c})],
\] (1)

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

\[
Z(s, \tilde{c}) = \min \sum_{i=1}^{n+1} w_i \quad \text{s.t. } w_1 = 0,
\] (2)

\[
w_{i+1} \geq 0, \quad \forall i = 1, \ldots, n,
\]
\[
w_{i+1} \geq w_i + \tilde{c}_i - s_i, \quad \forall i = 1, \ldots, n.
\]

Applying linear programming duality, this is equivalent to:

\[
Z(s, \tilde{c}) = \max \sum_{i=1}^{n} (\tilde{c}_i - s_i) x_i \quad \text{s.t. } x_i - x_{i-1} \geq -1, \forall i = 2, \ldots, n,
\]
\[
x_n \leq 1,
\]
\[
x_i \geq 0, \quad \forall i = 1, \ldots, n.
\] (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 et al. (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 et al. (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.

Following the form in (3), and noting that the $\tilde{c}_i - s_i$ terms are all distributed according to $\tilde{c}_i$, except for a fixed translation, this paper will focus on Distributionally Robust optimization problems in which the cost is of the form:

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

(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. While we have omitted the dependency on the argument $s$, it turns out that study of this form will suffice for problems like the Appointment Scheduling problem, and we address the details of how this is done in section 5. So the main theoretical focus of this paper is on finding a tight upper bound on the expected value of (4), the value of a linear optimization problem in which the objective coefficients are chosen randomly from a distribution in the set $\Gamma(\mu_1, \ldots, \mu_n)$ as follows:

$$Z^* = \sup_{\theta \in \Gamma(\mu_1, \ldots, \mu_n)} E_{\tilde{c} \sim \theta} [Z(\tilde{c})],$$

$$= \sup_{\theta \in \Gamma(\mu_1, \ldots, \mu_n)} E_{\tilde{c} \sim \theta} [\tilde{c}^T x^*(\tilde{c})],$$

(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 $\text{conv}(\mathcal{X})$, is a polytope with $Z(c) = \max \{ \sum_{i=1}^{n} c_i x_i : x \in \text{conv}(\mathcal{X}) \}$. Our interest in this problem is primarily in cases where $\mathcal{X}$ is given implicitly by the extreme points to an H-polytope $\{ x : Ax \leq b \}$. Following the terminology in Natarajan et al. (2009), we refer to the model in (5) as the Marginal Distribution Model (MDM).
1.1. Our Contributions

We add to the previous literature in the following manner. In Section 2, whereas previous papers have focused on the case where the points in $\mathcal{X}$ are explicitly given, we focus on the case when $\mathcal{X}$ is implicitly given as the extreme points of an H-polytope (i.e., the input provides the constraints of the polytope in the form $Ax \leq b$) and show that computing $Z^*$ given this form of input is generally NP-hard.

In Section 3, we use results from the field of Optimal Transport to identify problem (5) as more generally part of a primal-dual pairing of optimization problems. In doing so, we generalize the primal formulation result of Natarajan et al. (2009) to arbitrary marginal distributions. As well, the accompanying dual problem we derive is a convex mathematical program that with respect to the existing literature either captures, generalizes, or provides a simpler program than those derived in the studies of Meilijson and Nadas (1979), Haneveld (1986), and Weiss (1986) that focused on the special case of $\mathcal{X}$ as a collection of 0/1 vectors.

In Section 4, we identify two sufficiency conditions for when $Z^*$ is computable in polynomial-time. Included in the study is the case where $\mathcal{X} \subseteq \{0,1\}^n$ (studied in Haneveld (1986) but assumed that set $\mathcal{X}$ was explicitly given) as well as the more general case when $\mathcal{X}$ is the set of extreme points to a polytope given in the form of halfspace representation (H-polytope form) that has not been studied in the previous literature. The first sufficiency condition is motivated from examination of the dual form derived in Section 3, and it is a statement regarding the existence of an extended formulation for $\text{conv}(\mathcal{X})$ (see Theorem 3). The second sufficiency condition on the input $\mathcal{X}$ is derived from the connection between monotone couplings and supermodularity. The results of this study extend the current body of tractability results of Meilijson and Nadas (1979) and Bertsimas et al. (2004) for 0/1 polytopes and the result of Mak et al. (2015) for the appointment scheduling problem to a class of integer polytopes.

Finally, we discuss the implications of the results on the solvability of some distributionally robust optimization problems in Section 5 and provide computational experiments in Section 6. For example, the results involving the computation of $Z^*$ under H-polytope input are used for the Appointment Scheduling and Ranking applications and simulations.
(Detailed) Literature Review

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 $X$ in this example is defined as the set of directed paths (i.e., their 0/1 incidence vectors) from the start to the end node in a directed acyclic graph representation of the PERT network. The key contribution of Meilijson and Nadas (1979) was to propose a convex optimization formulation to compute the tight upper bound on the expected completion time that is valid for the class of joint distributions with the specified marginals. In a follow-up paper, Nadas (1979) proposed a numerical solution to solve this problem efficiently by applying a network flow algorithm for project cost curves. A lower bound on the probability that a given path in this PERT network is critical in the worst-case distribution is obtained from the Lagrange multiplier associated with the constraint determined by the path. Haneveld (1986) also derived similar analytical results and more with a convex duality approach. And Weiss (1986) furthered this kind of study to obtain bounds for other combinatorial optimization problems such as the shortest path, maximum flow and network reliability problem by using the theory of clutters and blocking clutters. It should be noted that in all these instances, $X$ is given explicitly as a finite collection of 0/1 vectors, a special case of the MDM problem (5). As well, these studies did not venture far beyond an analytical perspective to study computation.

The study of the bound in (5) for general discrete optimization was initiated in Natarajan et al. (2009). The feasible region considered in their work was $X = \{x \in \mathbb{Z}_+^n : Ax \leq b, x_i \in [\alpha_i, \beta_i], \forall i \in [n]\}$. Under the assumption that the linear optimization problem (the integrand) almost surely admits a unique solution, which complies with an absolute continuity assumption of the marginal distributions, Natarajan et al. (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 in general is not discussed in their work.

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 et al. (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 follow-up paper, Bertsimas et al. (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 et al. (2004). The advantage of the primal formulation is that it provides the “persistence” of the binary variables which is the marginal distribution of the optimal solution. Other more recent works using this approach of incorporating marginal moments include Van Parys et al. (2016), but for stochastic cost functions of the form given by $Z(\tilde{c}) := l(\sum_{i=1}^{n} \tilde{c}_i)$, for some convex loss function $l$.

In the special case of the appointment scheduling problem with marginal moment information, Mak et al. (2015) considered an alternative binary reformulation to that proposed in Natarajan et al. (2009). Interestingly in this special case, the distributionally robust appointment scheduling problem was shown to be solvable in polynomial time. Natarajan et al. (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 et al. (2009) also showed that the persistence 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 \( \min_{s \in S} \sup_{\theta \in \Theta} \mathbb{E}_{\tilde{\xi} \sim \theta} \left[ h(s, \tilde{\xi}) \right] \) 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 \( \xi \). Towards studying this further, they defined the “price of correlation”, a measure of the cost of model misspecification, by comparing the performance of operating under the assumption of independence as opposed to operating under the assumption of worst-case coupling. Using ideas from cost-sharing analysis in cooperative game theory, they identified sufficient conditions under which this price is bounded and provided applications to stochastic combinatorial problems. In contrast, in this paper we restrict our attention to support functions over polytopes.

Lastly, the field of optimal transport theory and the related Monge-Kantorovich problem provides several fundamental tools including a duality theory. This has been applied in a wide range of areas from physics to engineering and to 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 \((X, \mu)\) and \((Y, \nu)\) 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 may not always exist. For instance, consider the case of \( \mu \) having total mass at a single point while \( \nu \) 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.

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 an H-polytope denoted as $\{x : Ax \leq b\}$, namely $\mathcal{X} = \text{Extr}(\{x : Ax \leq b\})$. 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 is solvable in polynomial time as a linear program, we show that the distributionally robust bound is NP-hard to compute.

**Theorem 1.** Computing $Z^*$ in MDM for the class of linear optimization problems given discrete marginal distributions and a H-polytope is NP-hard.

**Proof:** We refer the reader to the Appendix section EC.1 for a proof that involves a reduction from 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. $\square$

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 expected value of a linear program over the Fréchet class of distributions is also NP-Hard. In the next section, we present a primal-dual formulation, which later will be used to identify conditions to recognize tractable subclasses of instances.
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 establish and recall some basic facts, notations, and related measure-theoretic concepts that will be useful for discussion.

For any function $f : \mathbb{R}^n \to \mathbb{R}$, let $f^* : \mathbb{R}^n \to \mathbb{R}$ denote the Legendre Fenchel conjugate, defined by

$$f^*(c) := \sup_{x \in \mathbb{R}^n} \{c^T x - f(x)\} \quad \forall c \in \mathbb{R}^n.$$ 

Let $f^{**} := (f^*)^*$, which is equivalent to $f$ in the case of convex $f$, and let $f^{***} := (f^{**})^* = f^*$ (Rockafellar (1997)). Notice that these definitions are constructed for functions defined on a connected space, as in $\mathbb{R}^n$. However, throughout this text we will often encounter functions defined on a finite subset $\mathcal{X} \subset \mathbb{R}$. For example, consider a function $\psi : \mathcal{X} \to \mathbb{R}$. So that we may be able to discuss notions of “convexity” as well as the Legendre Fenchel conjugate, we consider extending $\psi$ to all of $\mathbb{R}$ in a manner that preserves the value of $\max_{x \in \mathcal{X}} \{cx - \psi(x)\}$. For any $\xi \in \mathbb{R}$, let

$$\psi(\xi) = \begin{cases} 
\psi(x), & \xi \in \mathcal{X}, \\
\lambda \psi(x_\lambda) + (1 - \lambda) \psi(x), & \xi = \lambda x_\lambda + (1 - \lambda)x, \text{ for some } x \in \mathcal{X}, x_\lambda := \sup \{\alpha \in \mathcal{X} : \alpha < x\}, \lambda \in (0, 1), \\
+\infty, & \xi \notin \text{ conv}(\mathcal{X}).
\end{cases}$$

(6)

Henceforth, for the sake of convenience, we will use the symbol $\psi$ when in fact referring to the extension of $\psi$ that is defined in the manner above. And we will call $\psi$ convex when its extension in the manner above results in a convex function. One advantage to this convention will be that it affords us the shorthand

$$\max_x \{\bar{c}x - \psi(x)\} = \psi^*(\bar{c}) \quad \forall \bar{c} \in \mathbb{R}.$$ 

(7)
Next, for any convex function \( f : \mathbb{R}^n \to (-\infty, +\infty] \), we denote the set of subgradients to \( f \) at \( x \) by 
\[
\partial f(x) := \{ x^* \in \mathbb{R}^n : f(y) \geq f(x) + \langle x^*, y - x \rangle \quad \forall y \in \mathbb{R}^n \}
\]

Finally, if \( f \) is finite at the point \( x \), then let the right derivative be denoted by 
\[
 f'_+(x) := \lim_{\lambda \downarrow 0} \frac{f(x + \lambda y) - f(x)}{\lambda}.
\]

For the following discussion, when \( S \) is a topological space, the Borel sigma algebra will be the sigma algebra generated by the family of open sets, and we let the set of all probability measures defined on this Borel sigma algebra (called Borel probability measures) be denoted by \( \mathcal{P}(S) \). Whenever the space is a finite set \( \mathcal{X} \), we will assume the discrete topology, in which all subsets of \( \mathcal{X} \) are considered open. If \( S \) is a topological space with sigma algebra \( \mathcal{F} \), \( f \) is a measurable function from a probability space \((S, \mathcal{F}, \mu)\) into the reals, we will refer to \( f \# \mu := \mu \circ f^{-1} \) as the “measure induced by \( f \)” on the reals.

**Definition 1.** Let \( \gamma \) be a Borel probability measure on \( \mathbb{R}^n \). Then for any \( i \), define the “\( i \)-th projection” \( \Pi_i \gamma \) of the measure \( \gamma \) by:
\[
\Pi_i \gamma (A) := \gamma([R \times \ldots \times R \times A \times R \times \ldots \times R]) \quad \text{for all Borel measurable subsets } A \text{ of } \mathbb{R}.
\]

Given a collection of Borel probability measures \( \{\mu_i\}_n \), each defined on \( \mathbb{R} \), let:
\[
\Gamma(\mu_1, \ldots, \mu_n) := \{ \gamma \in \mathcal{P}(\mathbb{R}^n) : \Pi_i \gamma = \mu_i, \forall i \in [n] \}.
\]

Observe that this set is nonempty because the product measure \( \mu_1 \otimes \mu_2 \otimes \ldots \otimes \mu_n \) is trivially a member.

Unlike Natarajan et al. (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 \( Z(\hat{c}) := \)
max \{c^T x : x \in \mathcal{X} \}, the optimal solution multifunction (see Shapiro et al. (2014)) \( x^{OPT} : \mathbb{R}^n \rightarrow \mathbb{R}^n \) 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 \(A \subset \mathbb{R}^n\), \(\{\tilde{c} : x^{OPT}(\tilde{c}) \cap A \neq \emptyset\}\) is measurable.

Hence, there exists a measurable selection of \(x^{OPT}\), which we will write as \(x^* : \mathbb{R}^n \rightarrow \mathbb{R}^n\), where \(x^*(\tilde{c}) \in x^{OPT}(\tilde{c})\). The measurable mapping \(x^*(\cdot)\) then returns a possible optimal solution to any given vector \(\tilde{c}\).

### 3.2. A Lagrangian Formulation

In this section, we derive a Lagrangian formulation of MDM. The following core duality result from optimal transport theory for the Monge-Kantorovich problem will prove useful for the analysis.

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

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

where \(\Phi(h)\) is the set of pairs \((u, v)\) of Borel functions \(u : X \rightarrow (-\infty, \infty]\) and \(v : Y \rightarrow (-\infty, \infty]\) such that \(u \in L^1(\mu)\) and \(v \in L^1(\nu)\) 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 2 (Monotone Coupling Lemma, Theorem 3.1.2 in Rachev and Rüschendorf (1998)).** Let \(h : \mathbb{R}^2 \rightarrow \mathbb{R}\) be a right-continuous supermodular function. Then,

\[
\sup_{\gamma \in \Gamma(\mu, \nu)} E_{(X, Y) \sim \gamma} [h(x, y)] = \int_0^1 h(F^{-1}_\mu(t), F^{-1}_\nu(t)) dt,
\]
\[
\inf_{\gamma \in \Gamma(\mu, \nu)} E_{(x,y) \sim \gamma}[h(x,y)] = \int_0^1 h(F^{-1}_\mu(t), F^{-1}_\nu(1-t)) dt
\]

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

From a probabilistic perspective, the lemma says the monotone coupling on the probability space \(((0,1), \mathcal{B}, \lambda)\) - with \( \lambda \) denoting Lebesgue measure - given by \((X_1(t), \ldots X_n(t)) := (F^{-1}_1(t), \ldots, F^{-1}_n(t))\) yields the largest expectation.

Given the one-dimensional marginal distributions \(\mu_1, \ldots, \mu_n\), we define the convex-concave Lagrangian function \(L:\left(\mathbb{R}^{X_1} \times \cdots \times \mathbb{R}^{X_n}\right) \times \mathcal{P}(\mathcal{X})\) by:

\[
L\left(\{\psi_i\}_{i=1}^n, \nu\right) := \sum_i \left( \int \psi_i d\Pi_{i,\nu} + \int \max x_i \{c_i x_i - \psi_i(x_i)\} d\mu_i \right).
\]

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 \(n\) Borel probability measures \(\{\mu_i\}_{i=1}^n\) over \(\mathbb{R}\). And let \(\mathcal{X} \subset \mathbb{R}^n\) be an arbitrary finite point set. Then

\[
\max_{\theta \in \Gamma(\mu_1, \ldots, \mu_n)} E_{\hat{\theta}} \tilde{Z}(\tilde{c}) = \max_{\nu \in \mathcal{P}(\mathcal{X})} \sum_i \max_{\gamma_i \in \Gamma(\mu_i, \Pi_i, \nu)} E_{(\tilde{c}_i, x_i) \sim \gamma_i} \tilde{c}_i x_i.
\]

If, in addition, \(\mathbb{E}_{\hat{\theta}, \mu_i} [\tilde{c}_i] \in \mathbb{R}\) for all \(i\), then we find the additional equalities,

\[
= \max_{\nu \in \mathcal{P}(\mathcal{X})} \inf_{\{\psi_i : \mathcal{X} \to \mathbb{R}\}_{i=1}^n} L\left(\{\psi_i\}_{i=1}^n, \nu\right) \quad \text{(Primal)}
\]

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

**Proof Sketch.** We defer a full proof of Theorem 2 to Appendix EC.2.1, but here provide a sketch of the high-level ideas which can also be used as a roadmap while reading the proof in the appendix.

**Establishing equality (†).** We separately show first \(\text{LHS} \leq \text{RHS}\) and then \(\text{LHS} \geq \text{RHS}\).

- **LHS \leq RHS:**

  Let \(\theta^* \in \Gamma(\mu_1, \ldots, \mu_n)\) be optimal for the optimization in the expression of LHS. Based on the discussion in Section 3.1, there exists a solution mapping, call it \(x^*\), i.e., \(x^*(c) = Z(c)\) for any \(c \in \mathbb{R}^n\). Consequently, \(x^*\) yields a probability law \(\nu \in \mathcal{P}(\mathcal{X})\), as well as the collection of measures \(\gamma_i := ((\text{Proj})_i, x_i^*)\#\theta \in \Gamma(\mu_i, \Pi_i, \nu)\) - all measures that are feasible to the optimization problem comprising RHS. What’s more, LHS is equal to the objective function of RHS evaluated with these feasible measures.
\bullet\ LHS \geq \text{RHS}:

Let \( \nu^* \in \mathcal{P}(\mathcal{X}) \) and \( \gamma^*_1 \in \Gamma(\mu_1, \Pi_1 \nu^*), \ldots, \gamma^*_n \in \Gamma(\mu_n, \Pi_n \nu^*) \) be optimal solutions to the max-max problem on the RHS. These elements together suffice to compose a random vector \( \tilde{c} \) distributed according to a valid coupling. The procedure to compose \( \tilde{c} \) is summarized as follows. Given the random vector \( x \sim \nu^* \), the couplings \( \gamma^*_1, \ldots, \gamma^*_n \) yield distributions we can understand as \( \tilde{c}_i|x_i \), a collection of conditional distributions which, upon arbitrary coupling, yield the joint conditional distribution \( \tilde{c}|x \), and hence \( \tilde{c} \). Indeed, this \( \tilde{c} \) constructed is not only a valid coupling but also yields an expectation \( \mathbb{E}_{\tilde{c}}[Z(\tilde{c})] \) equal to the value of RHS.

This implicit construction of \( \tilde{c} \) conveniently completes the proof; in contrast, Natarajan et al. (2009) relies on an explicit construction in its proof for the less general setting of absolutely continuous marginals. In Remark EC.1, we also extend this approach, outlining an explicit construction for the case of general marginals, which can serve as an alternative proof.

**Establishing equality (Primal).** For each coordinate \( i \), the inner maximization problem over \( \gamma_i \) is an optimal transport problem, and after applying Lemma 1 (assuming \( \mathbb{E}_{\tilde{c}_i \sim \mu_i}[\tilde{c}_i] \in \mathbb{R} \)), we arrive at the form

\[
\max_{\gamma_i \in \Gamma(\mu_i, \Pi_i \nu)} \mathbb{E}_{(\tilde{c}_i, x_i) \sim \gamma_i} \tilde{c}_i x_i = \min_{\psi_i : \mathcal{X}_i \to \mathbb{R}} \int \psi_i d\Pi_i \nu + \int \max_{x_i \in \mathcal{X}_i} \{\tilde{c}_i x_i - \psi_i(x_i)\} d\mu_i,
\]

after further massaging.

**Establishing equality (Dual).** To establish the exchange of the sup and inf, we show that a saddle-point exists for the convex-concave Lagrangian function \( L(\cdot, \cdot) \).

**Definition 3.1** A measure \( \bar{\nu} \in \mathcal{P}(\mathcal{X}) \) will be called a primal optimal solution if it solves the maximization problem in (Primal), or equivalently, (†). A collection of functions \( (\psi_1, \ldots, \psi_n) \) such that \( \psi_i : \mathcal{X}_i \to \mathbb{R} \) for every \( i \) will be called a dual optimal solution if it solves the minimization problem in (Dual).

**Remark 1 (Two Player Zero-Sum Game and Nash Equilibria).** We note that the above max-min and min-max exchange in Theorem 2 naturally characterizes the existence of a mixed
Nash Equilibrium to a game that we describe here. Player 1 decides on a mixed strategy that selects a random action from $X$. Player 2 decides on a collection of real-valued vectors. Player 1’s utility is given by the Lagrangian function $L$; consequently, Player 2’s utility is $-L$. Observe that Theorem 2 only comments on mixed rather than pure nash equilibria. This is in fact because it is not true in general that a pure equilibrium exists for this game; in other words, all equilibria may exclude having Player 1 select a pure strategy, or deterministic action, from $X$. Consider the following example.

Let $X = \{(1,0,0),(0,1,0),(0,0,1)\}$ and $\mu_1, \mu_2, \mu_3 = \frac{1}{3} \cdot 1 + \frac{2}{3} \cdot 1_0$. Then

$$\max_{x \in X} \inf_{(\psi_i:X_i \to \mathbb{R})_{i=1}^3} L(\{\psi_i\}_i, x) = 0 \neq 1 = \inf_{(\psi_i:X_i \to \mathbb{R})_{i=1}^3} \max_{x \in X} L(\{\psi_i\}_i, x).$$

We next show that all of the identities in Theorem 2 continue to hold even if the functions $\psi_i$ are restricted to be convex. This will allow for a discussion of the primal-dual relationship that exists between any primal optimal solution $\bar{\nu}$ and any dual optimal solution $(\psi_1, \ldots, \psi_n)$, which we discuss in Corollary 1; further, it will yield a uniqueness property to be discussed in Section 3.3.

**Remark 2.** All infima in the theorem can be taken equivalently over functions $\psi_i: \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ that are univariate polyhedrally convex and whose extreme points lie in $X_i$ (recollect that a function is polyhedral convex if and only if the epigraph $\text{epi}(f)$ is polyhedral). Recall that for any $\bar{c}_i$,

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

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

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

In other words, under the goal of minimizing the Lagrangian $L(\{\psi_i\}_i, \nu)$ over $\psi_i$, with fixed $\nu \in \mathcal{P}(X)$, any proposed solution $\psi_i$ can be modified to $\psi_i^{**}$, the largest convex minorant of $\psi_i$, to yield a solution that does no worse. Since in our case $\psi_i^{**} = \psi_i$ is equivalent to $\psi_i$ being a univariate polyhedral convex function with extreme points in $X_i$, we can restrict the search, as claimed. $\triangle$
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)) with vertical line segments at each $x_i \in \mathcal{X}_i$, not unlike the graph of any cumulative distribution function for a discrete distribution. In the following corollary, we see it is the measures $\mu_i$ that translate the distributional information in the nondecreasing curves.

**Corollary 1 ("Saddle-point" conditions).** Following Remark 2, we restrict attention to $\{\psi_i\}_i$ that are convex. Then, $(\bar{\nu}, \{\bar{\psi}_i\}_i)$ is a pair of primal and dual optimal solutions (aka saddle point) if and only if the following conditions are satisfied:

1. $\Pi_i\bar{\nu}(x_i) = \mu_i(\partial\bar{\psi}_i(x_i) \setminus \cup x_i' x_i, \partial\bar{\psi}_i(x_i'))$, $\mu_i(\partial\bar{\psi}_i(x_i))$ for all $x_i \in \mathcal{X}_i$.
2. $\int_{\mathcal{X}} \sum_i \bar{\psi}_i(x_i) d\nu \leq \int_{\mathcal{X}} \sum_i \bar{\psi}_i(x_i) d\tilde{\nu}$, for all $\nu \in \mathcal{P}(\mathcal{X})$.

Furthermore, if $\mu_i$ is absolutely continuous, the condition in 1. yields:

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

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

**Proof:** Observe that $L(\{\bar{\psi}_i\}_i) \leq L(\{\psi_i\}_i)$ if and only if

$$\sum_{x_i \in \mathcal{X}_i} \psi_i(x_i) \cdot \Pi_i\bar{\nu}(x_i) + \int_{x_i \in \mathcal{X}_i} \max\{\bar{c}_i - \psi_i(x_i)\} d\mu_i$$

is minimized in $\psi_i$ at $\bar{\psi}_i$. For notation’s sake, let $F_{\bar{c}_i}(\psi_i) := \max_{x_i \in \mathcal{X}_i} \{\bar{c}_i - \psi_i(x_i)\}$, and $f(\psi_i) := \int \max_{x_i \in \mathcal{X}_i} \{\bar{c}_i - \psi_i(x_i)\} d\mu_i$. Then

$$\partial F_{\bar{c}_i}(\psi_i) = \text{conv}(-1_{x_i} : \bar{c}_i \in \partial\psi_i(x_i))$$

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

$$\partial f(\psi_i) = \int \partial F_{\bar{c}_i}(\psi_i) d\mu_i = [-1, 0] \cdot \mu_i(\partial\psi_i(x_i) \setminus \cup x_i' x_i, \partial\psi_i(x_i')) - \mu_i(\partial\bar{\psi}_i(x_i) \setminus \cup x_i' x_i, \partial\bar{\psi}_i(x_i'))$$

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

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

as desired for condition 1. Condition 2 is equivalent to $L(\{\bar{\psi}_i\}_i, \bar{\nu}) \geq L(\{\psi_i\}_i, \nu)$. □
3.3. The MDM Primal Problem

We now turn attention to (Primal). Previously, Natarajan et al. (2009) studied a variant of (Primal), derived under the assumption of absolutely continuous marginals. So we begin the study of the (Primal) problem by noting an immediate consequence of Theorem 2, a generalization of Theorem 1 of Natarajan et al. (2009). In order to state the theorem, we introduce the following notation.

Let $x_i - e_i$, where $x_i \in X_i$, denote the largest element in $X_i$ that is less than but not equal to $x_i$; we let $x_i - e_i = -\infty$, if no such largest element exists.

**Corollary 2.** Let $X$ be a finite set. Let $\tilde{c}_1, \ldots, \tilde{c}_n$ be real-valued random variables with $\mu_1, \ldots, \mu_n$ denoting the Borel probability measures they induce on the real line. Then,

$$\max_{\theta \in \Gamma(\mu_1, \ldots, \mu_n)} E_{\tilde{c} \sim \theta} Z(\tilde{c}) = \max_{\nu \in \mathcal{P}(X)} \sum_i \sum_{x_i \in X_i} x_i \int_{\Pi_i \nu((-\infty, x_i])} F_{\mu_i}^{-1}(t) dt. \quad (8)$$

In particular, for any $\nu_*$ solving the concave maximization problem (8), we may construct a maximizing joint distribution $\theta^* \in \Gamma(\mu_1, \ldots, \mu_n)$.

**Proof:** Equality (†) was established in Theorem 2. The final inequality holds due to Lemma 2.

Now let us consider the set of primal optimal solutions, and we ask about uniqueness matters. At a glance, we observe that there exist both cases of uniqueness as well as cases of multiple primal solutions - see Remark EC.2 for examples of both these types of behavior. In other words, in the most general case, the set of primal optimal solutions does not really exhibit any conclusive type of uniqueness behavior, in terms of uniqueness.

So let us restrict to the case when the measures $\mu_1, \ldots, \mu_n$ correspond to continuous distributions. As previously noted, Natarajan et al. (2009) conducted their MDM study under this assumption. But the following result indicates that this assumption in fact yields a structural property to the set of primal optimal solutions in the way of uniqueness.

**Corollary 3 (Marginal Uniqueness in Primal Optimality).** Suppose that the Borel probability measures $\mu_1, \ldots, \mu_n$ are all absolutely continuous. If $\tilde{\nu}$ and $\tilde{\tau}$ are both primal optimal solutions, then $\Pi_i \tilde{\nu} = \Pi_i \tilde{\tau}$ for all $i$. 


Proof: Let \( \{\tilde{\psi}_i\}_i \) be a dual optimal solution. Then, with \((\tilde{\nu}, \{\tilde{\psi}_i\}_i)\) and \((\tilde{\tau}, \{\tilde{\psi}_i\}_i)\) both being pairs of primal-dual optimal solutions, Corollary 1 yields the desired result. \(\square\)

This corollary along with the following proposition illustrate that any primal optimal distribution can be viewed as the action of an optimal-solution mapping \(x^*\) on an extremal coupling. For a concrete example, consider discrete choice modeling in Natarajan et al. (2009), in which \(X := \{x : \sum_i x_i = 1, \ x_i \in \{0, 1\}\}\) and \(\tilde{c}_i \sim \mu_i\) represents the random utility an agent has for item \(i\), and \(Z(\tilde{c}) := \max_{x \in X} \tilde{c}^T x\) is the utility an agent achieves upon choosing the utility-maximizing item. The content of the following result indicates that under the assumption of absolutely continuous marginals, for any primal optimal solution \(\tilde{\nu}\), the action of a randomly drawn agent (equivalently, a random draw of utilities from the extremal coupling) making a utility-maximizing choice \(x^*(\tilde{c})\) will simulate \(\tilde{\nu}\).

As \(\Pi_i \tilde{\nu}\) yields the probability that option \(i\) is included in a choice, Corollary (3) indicates that under the assumption of absolutely continuous marginals, all primal optimal probability measures over \(X\) yield the same choice probabilities.

**Proposition 1.** Suppose that the Borel probability measures \(\mu_1, \ldots, \mu_n\) are all absolutely continuous. Then there exists a measurable function \(x^* : \mathbb{R}^n \to X\) that takes the form \(x^*(c) = (x^*_1(c_1), \ldots, x^*_n(c_n))\) and satisfies \(x^*(c) \in \arg \max_{x \in X} c^T x\) for all \(c \in \times_{i=1}^n \text{supp}(\mu_i)\), such that for any primal optimal solution \(\tilde{\nu}\),

\[
P_{\tilde{c} - \theta^*}(x^*_i(\tilde{c}) = x_i) = P_{\tilde{c}_i - \mu_i}(x^*_i(\tilde{c}_i) = x_i) = \Pi_i \tilde{\nu}(x_i), \ \forall x_i \in X_i, \ \forall i \in [n],
\]

where \(\theta^* \in \Gamma(\mu_1, \ldots, \mu_n)\) is any optimal solution to the MDM problem (5).

**Proof:** We refer the reader to the Appendix section EC.2.2. \(\square\)

We now make several remarks on this result in the case of absolutely continuous marginals. First, the quantity \(P_{\tilde{c} - \theta}(x^*_i(\tilde{c}) = x_i)\), for some distribution \(\theta\), is significant for applications (e.g. discrete choice) and is referred to as a persistence value in Natarajan et al. (2009). In (9), we see that the persistence value under any worst-case coupling \(\theta^*\) in fact is entirely determined by the marginals alone, for the fact that the mapping behavior of \(x^*\) is separable. Second, the fact that each \(x_i\)
can be identified from $\tilde{c}_i$ separately across $i = 1, \ldots, n$ is possible because $\mu_1, \ldots, \mu_n$ are absolutely continuous. Indeed, this separability may not exist otherwise-for example, consider the discrete measures $\mu_1, \mu_2$ which both have mass 3/4 on 1 and mass 1/4 on -1. If $\mathcal{X} = \{(-1, 2), (1, -2)\}$, then the optimal value of $x_1$ must depend on $\tilde{c}_2$—even if $\tilde{c}_1 = 1$, whether we can choose the point $x = (1, -2)$ depends on the value of $\tilde{c}_2$.

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

Recall that from Proposition 1, the bound on the expected value is NP-hard to compute. Thus, the interest is to identify a class of inputs under which the bound is computable efficiently. Recall that the input is two-fold: a collection of Borel (marginal) probability measures $\mu_1, \ldots, \mu_n$ on $\mathbb{R}$, and a finite point set $\mathcal{X} \subseteq \mathbb{R}^n$. We note that $\mathcal{X}$ will always be understood as the extreme points to a given polytope, regardless of V or H representation. In the case of V-polytope input, $\mathcal{X}$ is given explicitly. In the case of H-polytope input, $\mathcal{X}$ is given implicitly. And we note that while the previous literature has tackled the case of V-polytope input, the work of this section tackles H-polytope input in addition.

We discuss two kinds of sufficiency conditions on the input under which the problem of computing $Z^*$ is tractable (we make this notion precise momentarily): 1.) The first one is derived from the dual formulation of MDM and involves extended formulations; 2.) The second one highlights the role that supermodularity may play.

Throughout this section, we will consider the following assumption on the given marginals $\mu_1, \ldots, \mu_n$, which admits the special case of finite-supported measures.

**Assumption 1.** Given the marginals $\mu_1, \ldots, \mu_n$ as input, we assume that the expected value of univariate piecewise linear convex functions of the form $\int \max_{x_i \in \mathcal{X}_i} \tilde{c}_i x_i - \psi_i(x_i) d\mu_i$ - where $\psi_i : \mathcal{X}_i \rightarrow \mathbb{R}$- are efficiently computable for all $i \in [n]$. Furthermore, the subgradients of the functions are efficiently computable.
Observe that Assumption 1 is satisfied in the case when all marginals have finite support, which is the case in the computational experiments of section 6. But more generally, in this assumption we are assuming oracle-access to subgradients.

**Definition 2.** “Efficiently computable” and “tractable” in this section will mean, in the case of discrete marginals, existence of poly-time algorithms for exact computation of $Z^*$, and in the general case, existence of convergent algorithms with oracle complexity that is polynomial in the target approximation error and input.

### 4.1. Sufficiency by MDM Dual

Now, recall the dual formulation of MDM from Theorem 2:

$$\max_{\theta \in \Gamma(\mu_1, \ldots, \mu_n)} E_{\tilde{c} \sim \theta}[Z(\tilde{c})] = \min_{\psi_i : X_i \to \mathbb{R}} \max_{\nu \in \mathcal{P}(X)} \sum_{i=1}^{n} \int \psi_i^*(\tilde{c}_i) d\mu_i + \int \left[ \sum_{i=1}^{n} \psi_i(x_i) \right] d\nu(x) \tag{10}$$

Given that the terms $\int \psi_i^*(\tilde{c}_i) d\mu_i$ are already mentioned in Assumption 1, the quantity that remains in question is the inner maximization problem over probability measures on $X$:

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

Notice that for any $\{\psi_i(x_i)\}_{x_i \in X_i}$, the computation of this quantity depends on the finite point set $X$. In this section, we identify sufficient conditions on $X$ under which $(*)$ admits a tractable reformulation, to be incorporated into an efficient scheme to tackle the entire optimization problem (10) at once.

Naturally, we will be interested in the polytope $\text{conv}(X)$ that $X$ presents. Interestingly, observe that the above suggests that it in fact suffices to characterize the collection of marginals $\{\{\Pi_i \nu\}_{i=1}^{n} : \nu \in \mathcal{P}(X)\}$. Indeed, this insight is at the center of the forthcoming Theorem 3. In any case, for a discussion on complexity, we need to be clear on how $X$ is given as input. With regards to $X$, we will consider two kinds of input: 1.) a list of points, or 2.) a list of linear inequalities to a
polytope. In the first kind of input, the list is precisely \( \mathcal{X} \). In the second, and the main point of interest, \( \mathcal{X} \) is the set of extreme points to the polytope given by linear inequalities. We will refer to these types of input representations as V-polytope (explicit list of vertices) and H-polytope (list of linear inequalities) respectively. All polytopes can be represented with a list of vertices or hyperplanes (facets), so the distinction merely highlights how the polytope is presented as input.

We begin with the (trivial) case of 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.

4.1.1. \( \mathcal{X} \) as V-polytope We start with the simple case when \( \mathcal{X} \) is explicitly provided as a V-polytope. Perhaps as expected, MDM with V-polytope input is similarly trivial as the case of linear programming with V-polytope input. More precisely, the left-hand side of (*) is an efficiently-sized (in \(|\mathcal{X}|\)) 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 (10), implies that the bound \( Z^* \) is efficiently solvable as a convex minimization problem (assuming the marginals satisfy Assumption 1). Such a result is also discussed in Theorem 3 in Meilijson (1991) and in special cases 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.

4.1.2. \( \mathcal{X} \) as extreme points to a 0/1 H-polytope Now let \( \mathcal{X} \) be given implicitly as the set of extreme points to a given 0/1 H-polytope \( P \subset \mathbb{R}^n \) - note that we can immediately identify the projections \( \mathcal{X}_i = \{0,1\} \). Naturally, any element \( y \in P \) can be written as some convex combination of finitely many members from \( \mathcal{X} \), i.e., \( y = \sum \lambda_i x^i \), with \( x^i \in \mathcal{X} \), \( \sum \lambda_i = 1 \), \( \lambda_i \geq 0 \). Writing \( \nu(x^i) := \lambda_i \), we arrive at a corresponding member \( \nu \in \mathcal{P}(<\mathcal{X}>); \) furthermore, we also find that all the
marginal measures $\Pi_i: \mathcal{X} \to [0,1]$ satisfy $\Pi_i(1) = y_i$, for all $i = 1, \ldots, n$. So when the set $\mathcal{X}$ is affinely independent, $\mathcal{P}(\mathcal{X})$ and $P$ are in bijection. And this immediately implies we can perform the optimization (*) efficiently with the rewritten form:

$$\max_{\nu \in \mathcal{P}(\mathcal{X})} \sum_{i=1}^{n} \sum_{x_i \in \mathcal{X}_i} \psi_i(x_i) \cdot \Pi_i(\nu) = \max_{y \in P} \sum_{i=1}^{n} \psi_i(1) \cdot y_i + \psi_i(0) \cdot (1-y_i)$$

But even if $\mathcal{X}$ is not a set of affinely independent vectors, the above still holds, and this is precisely where $P$ being a 0/1 H-polytope comes into play. To see this, for any $y \in P$ let us write $[y] \subset \mathcal{P}(\mathcal{X})$ for the set of probability measures that correspond to the same $y \in P$. And observe that given any two probability measures $\nu_1, \nu_2 \in [y]$, that $y_i = \Pi_i \nu_1(\{1\}) = \Pi_i \nu_2(\{1\}) \in [0,1]$. In other words, $\{\{\Pi,\nu\}_{i=1}^{n} : \nu \in \mathcal{P}(\mathcal{X})\}$ and $P$ are in bijection. And this is sufficient for the above rewriting to hold, so that the bound $Z^*$ is efficiently solvable as a convex minimization problem.

More precisely, based on our discussion and the fact that we can without loss of generality assume that $\psi_i(0) = 0$,

$$\max_{\theta \in \Gamma(\mu_1, \ldots, \mu_n)} E_{\theta \to Z}[Z(\bar{c})] = \min_{\{\psi_i(0,1) \to R\}_{i=1}^{n}} \max_{\nu \in \mathcal{P}(\mathcal{X})} \sum_{i} \left( \int \psi_i d\Pi_i \nu + \int \psi_i^*(\bar{c}_i) d\mu_i \right)$$

$$= \min_{\{\psi_i(0,1) \to R\}_{i=1}^{n}} \max_{y \in P} \sum_{i} \psi_i(1) \cdot y_i + \psi_i(0) \cdot (1-y_i) + \left( \sum_{i} \int \max_{x_i=0,1} (\bar{c}_i \cdot x_i - \psi_i(x_i)) d\mu_i \right)$$

$$= \min_{\{w_i\}_{i=1}^{n}} \max_{y \in P} \sum_{i} w_i \cdot y_i + \left( \sum_{i} \int \max(\bar{c}_i - w_i, 0) d\mu_i \right)$$

$$= \min_{\{w_i\}_{i=1}^{n}} Z(w) + \left( \sum_{i} \int \max(\bar{c}_i - w_i, 0) d\mu_i \right)$$

In summary, optimization over $\{\{\Pi,\nu\}_{i=1}^{n} : \nu \in \mathcal{P}(\mathcal{X})\}$ being equivalent to the efficient optimization over $P$ (in the 0/1 H-polytope case) admits tractability. This MDM tractability result was developed for PERT networks in Meilijson and Nadas (1979) and Haneveld (1986), shortest path, maximum flow and network reliability problems in Weiss (1986). And analogous results were provided for combinatorial optimization problems in the settings of Bertsimas et al. (2010, 2004, 2006) in which only the marginal moments are known.
4.1.3. \( \mathcal{X} \) 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, and we do not make the assumption that the extreme points are binary vectors. As discussed in Proposition 1, \( Z^* \) is NP-hard to compute. But with the previous section revealing that compact 0/1 H-polytopes present tractability, we seek to exploit this in the following sufficient condition. The idea, simply put, is that if \( \mathcal{X} \) is the image of a compact, 0/1 H-Polytope under a linear mapping of particular form, then (*) admits a tractable reformulation. In this case, we may use the MDM dual form as part of an efficient procedure to compute \( Z^* \).

**Theorem 3 (Sufficiency by Extended Formulation).** Suppose a class of polytopes in \( \mathbb{R}^n \) is such that for any member \( Q \) - with \( \mathcal{X} = \text{Ext}(Q) \) - we can efficiently identify the sets \( \{ \mathcal{X}_i \}_{i=1}^n \) as well as a finite set \( B \) with subsets \( \{ B_i(\xi) \}_{\xi \in \mathcal{X}_i} \subseteq B \) for every \( i = 1, \ldots, n \), such that \( Q \) has a compact extended formulation:

\[
Q = \prod_{x} \left\{ (x, y) : y \in P \subseteq \mathbb{R}^B, \ x_i = \sum_{\xi \in \mathcal{X}_i} \xi_i \sum_{j \in B_i(\xi)} y_j \ \forall i = 1, \ldots, n \right\}, \quad \text{(Ext. Form)}
\]

where \( P \subseteq [0,1]^B \) is a 0/1 polytope of the form:

\[
P \subseteq \left\{ y \in [0,1]^B : \sum_{\xi \in \mathcal{X}_i} \sum_{j \in B_i(\xi)} y_j = 1, \ \forall i \in [n] \right\},
\]

then \( Z^* \) is efficiently computable over this class, for any collection of marginals \( \mu_1, \ldots, \mu_n \) that satisfy Assumption 1.

**Proof:** Assume there exists such a compact extended formulation. This means there exists a linear transformation \( A \) such that \( A(P) \supseteq \mathcal{X} \) - in fact, \( A(\text{Ext}(P)) \supseteq \mathcal{X} \). To start, observe that for any \( x \in \mathcal{X} \),

\[
x_i = \xi_i \iff \sum_{j \in B_i(\xi)} y_j = 1 \quad \forall y^x \in A^{-1}(x) \cap \text{Ext}(P).
\]

We next show a type of equivalence between \( \mathcal{P}(\mathcal{X}) \) and \( P \) with the following two claims:

1. For any \( \nu \in \mathcal{P}(\mathcal{X}) \), there exists a \( y \in P \) such that \( \Pi_i \nu(\cdot) = \sum_{j \in B_i(\cdot)} y_j \)
2. For any \( y \in P \), there exists a \( \nu \in \mathcal{P}(\mathcal{X}) \) such that \( \Pi_i \nu(\cdot) = \sum_{j \in B_i(\cdot)} y_j \)
To see 1.), let \( \nu \in \mathcal{P}(\mathcal{X}) \). And for each \( x \in \text{supp}(\nu) \), select a \( y^x \in A^{-1}(x) \cap \text{Ext}(P) \). Next, we construct a \( y \in P \) by letting \( y = \sum_{x \in \text{supp}(\nu)} y^x \cdot \nu(x) \). Then for arbitrary \( \xi_i \in \mathcal{X}_i \), by (15) we find that

\[
\Pi_i \nu(\xi_i) = \sum_{x \in \text{supp}(\nu): x_i = \xi_i} \nu(x) = \sum_{x \in \text{supp}(\nu)} \nu(x) \cdot 1_{x_i = \xi_i} = \sum_{x \in \text{supp}(\nu)} \nu(x) \cdot \left( \sum_{j \in B_i(\xi_i)} y_j^\ast \right) = \sum_{j \in B_i(\xi_i)} y_j
\]

To see 2.), let \( y \in P \). Let us enumerate the set of extreme points to \( P \), writing \( \text{Ext}(P) = \{y_1^1, \ldots, y_{\text{Ext}(P)}^1\} \). With this, \( y = \sum_k \lambda_k y^k \), for some \( \lambda_1, \ldots, \lambda_{\text{Ext}(P)} \geq 0 \) and that sum to 1. Next, we construct a \( \nu \in \mathcal{P}(\mathcal{X}) \) by letting \( \nu(x) = \sum_{k: A^{-1}(x) = y^k} \lambda_k \). Then for arbitrary \( \xi_i \in \mathcal{X}_i \), by (15) we find that

\[
\Pi_i \nu(\xi_i) = \sum_{x \in \text{supp}(\nu): x_i = \xi_i} \nu(x) = \sum_{x \in \text{supp}(\nu)} \nu(x) \cdot \left( \sum_{k: A^{-1}(x) = y^k} \lambda_k \cdot \sum_{j \in B_i(\xi_i)} y_j^k \right) = \sum_k \lambda_k \sum_{j \in B_i(\xi_i)} y_j^k = \sum_{j \in B_i(\xi_i)} y_j
\]

In summary, using this equivalence, problem (*) becomes a linear program that is efficiently solvable:

\[
\max_{\nu \in \mathcal{P}(\mathcal{X})} \sum_{i=1}^n \sum_{x_i \in \mathcal{X}_i} \psi_i(x_i) \cdot \Pi_i \nu(x_i) = \max_{y \in P} \sum_{i=1}^n \sum_{x_i \in \mathcal{X}_i} \psi_i(x_i) \cdot \sum_{j \in B_i(x_i)} y_j
\]

Applying the dual of this linear program into formulation (10), we see that the resulting mathematical program is a convex minimization problem (linear program in case of marginals with finite support), so we have arrived at an efficient procedure to compute \( Z^\ast \). \( \square \)

We now make the following observations and remarks. First, it in fact suffices to identify, for each \( i \), a superset \( \mathcal{X}_i' \supset \mathcal{X}_i \), in the sufficient condition of Theorem 3. This fact can prove useful for the class of polytopes we discuss next in the forthcoming Theorem 4, in which such approximations to the projections are readily obtainable. For more details on this, we refer the reader to the Appendix section EC.3. Regarding the notation in Theorem 3’s extended formulation, the reader may interpret \( B \) as a family of “events,” rich enough to describe \( \mathcal{X} \) (a la \( A(\text{Ext}(P)) \supset \mathcal{X} \)), and yet efficiently sized, with \( \{B_i(\xi_i)\}_{\xi_i \in \mathcal{X}_i} \) also efficiently describing \( \mathcal{X}_i \) (a la \( \sum_{\xi_i \in \mathcal{X}_i} \sum_{j \in B_i(\xi_i)} y_j = 1 \)). As it turns out, we may use Theorem 3 to identify the already-considered V-polytope and 0/1 H-polytope cases as tractable. For the sake of unifying our understanding or perspective on these cases, we now formalize the discussions above before returning to the topic at hand- studying the case of general H-polytopes and identifying a tractable subclass.
Corollary 4 (V-Polytopes). Let \( \tilde{c}_1, \ldots, \tilde{c}_n \) be real-valued random variables with \( \mu_1, \ldots, \mu_n \) denoting the probability measures they induce on the real line, and satisfying Assumption 1. For these measures and the class of V-polytopes, \( Z^* \) is efficiently computable.

Proof: To see this, we exploit the fact that every V-polytope is the projection of a simplex. In other words, to each V-polytope is an H-polytope that satisfies the sufficiency condition described in Theorem 3. More precisely, let \( G \) be the \( n \times m \) matrix formed by appending the elements of \( \mathcal{X} \) together as columns, i.e., \( G = (x^1 \ x^2 \ \ldots \ x^m) \). Then,

\[
\text{conv}(\mathcal{X}) = \Pi_x \left( \{(x, y) : y \in \Delta_{m-1}, x = Gy\} \right)
\]

where

\[
\Delta_{m-1} := \{ y \in \mathbb{R}^m : y \geq 0, \sum_i y_i = 1 \} \subseteq \left\{ y \in [0,1]^B : \sum_{\xi \in \mathcal{X}_i, \ j \in B(\xi)} y_j = 1, \ \forall i \in [n] \ \forall i, \xi_i \in \mathcal{X}_i \right\}
\]

where for each \( \xi_i \in \mathcal{X}_i, B_i(\xi_i) := \{ j \in B : x^j_i = \xi_i \}, B := \{1, \ldots, m\} \). \( \square \)

Similarly, we can also provide the sufficient compact extended formulations to 0/1 H-polytopes.

Corollary 5 (0/1 H-Polytopes). Let \( \tilde{c}_1, \ldots, \tilde{c}_n \) be real-valued random variables with \( \mu_1, \ldots, \mu_n \) denoting the probability measures they induce on the real line, and satisfying Assumption 1. For these measures and the class of 0/1 H-polytopes, \( Z^* \) is efficiently computable.

Proof: Let \( Q \subset \mathbb{R}^n \) be a 0/1 H-polytope. Then set \( B = [n] \cup [n'] \), where \( [n] = \{1, \ldots, n\} \) and \( [n'] = \{1', \ldots, n'\} \). Set \( B_i(1) = \{i\}, B_i(0) = \{i'\} \). Set \( P := \{(y, y') \in [0,1][n] \times [0,1][n'] : y \in Q, y'_i = 1 - y_i \} \). With these quantities defined in this manner for arbitrary \( Q \), we see that the sufficiency conditions of Theorem 3 are satisfied. \( \square \)

Indeed, there is an analogous result identified for the case when only the marginal moments are known (see Bertsimas et al. (2006)). In the next section, we discuss application problems that satisfy this condition for general linear and discrete optimization problems.
4.1.4. \( L^1 \) Convex Polytopes  In this section, we examine a special class of integral polytopes that are not necessarily 0/1, called \( L^1 \) convex polytopes and show that this class satisfies the conditions of Theorem 3. This class of polytopes can arise in dual formulations of network flow problems, as well as other settings. For example, in the Appointment Scheduling problem introduced in Section 1, the polytope in (3) belongs to this class, and remains so even when the overtime cost is different from the waiting time cost, i.e., the overtime cost is some integer \( \gamma > 1 \).

We begin by presenting the input parameters that specify the \( H \)-polytope representation of a member of this class. Consider a directed graph \( (V = \{1, \ldots, n\}, A) \), an integer-valued distance function \( d : V \times V \to \mathbb{Z} \cup \{+\infty\} \), and collections of bounds \( \{l_i\}_{i=1}^n \subseteq \mathbb{Z} \cup \{-\infty\} \), \( \{u_i\}_{i=1}^n \subseteq \mathbb{Z} \cup \{+\infty\} \). Let

\[
X = \{x \in \mathbb{Z}^n : x_i - x_j \leq d_{ij} \; \forall (i,j) \in A, \; l_i \leq x_i \leq u_i \; \forall i \in \{1, \ldots, n\}\}
\]

A set of this form is known as an \( L^1 \) convex set. It is nonempty if and only if \( d \) satisfies the triangle inequality: \( d_{ik} \leq d_{ij} + d_{jk} \), for all \( i, j, k \in V \). Its convex hull \( \text{conv}(X) \) is an integral polyhedron (e.g. see Murota (2003)) that when bounded we'll identify as an \( L^1 \) convex polytope.

When assuming \( \text{conv}(X) \) is bounded, observe that for any \( i \),

\[
l_i := \max_{k \neq i} \{l_k - d_{ki}\} \geq -\infty,
\]

\[
\bar{u}_i := \min_{j \neq i} \{u_j + d_{ij}\} < +\infty,
\]

so without loss of generality, let us assume that \( \{l_i\}_{i=1}^n \subseteq \mathbb{Z} \), \( \{u_i\}_{i=1}^n \subseteq \mathbb{Z} \). In summary,

**Definition 4.1 (\( L^1 \) convex polytope)** Let \( A \subseteq [n] \times [n] \), \( \{l_i\}_{i=1}^n \subseteq \mathbb{Z} \), \( \{u_i\}_{i=1}^n \subseteq \mathbb{Z} \), \( \{d_{ij}\}_{(i,j) \in A} \subseteq \mathbb{Z} \).

Then \( P = \{x \in \mathbb{R}^n : x_i - x_j \leq d_{ij} \; \forall (i,j) \in A, \; l_i \leq x_i \leq u_i \; \forall i \in \{1, \ldots, n\}\} \) is an \( L^1 \) convex polytope.

**Theorem 4.** Let \( \bar{c}_1, \ldots, \bar{c}_n \) be real-valued random variables with \( \mu_1, \ldots, \mu_n \) denoting the probability measures they induce on the real line, and satisfying Assumption 1. For these measures and the class of \( L^1 \) convex polytopes, \( Z^* \) is efficiently computable.
Proof: Let $G = (V, \mathcal{A})$, $d : V \times V \to \mathbb{Z} \cup \{+\infty\}$ (satisfying the triangle inequality) and $\{l_i\}_{i=1}^n \subset \mathbb{Z} \cup \{-\infty\}$, $\{u_i\}_{i=1}^n \subset \mathbb{Z} \cup \{+\infty\}$ be given as in the above for the characterization of $X$. Define $u := \max_i u_i$ and $l := \min_i l_i$. We claim that

$$x_i - x_j \leq d_{ij} \quad \forall (i, j) \in \mathcal{A} \iff \left[ x_i \in \{t, \ldots, u\} \implies x_j \not\in \{l, \ldots, t - d_{ij} - 1\} \right] \quad \forall (i, j) \in \mathcal{A}, \forall t \in [l, u] \cap \mathbb{Z}.$$ 

The “only if” direction is clear. The “if” direction is also clear, as given an $(i, j) \in \mathcal{A}$, $x_i \in \{x_i, \ldots, u\} \implies x_j \not\in \{l, \ldots, x_i - d_{ij} - 1\}$, by hypothesis. Hence, $x_j \geq x_i - d_{ij}$, as desired.

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

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

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

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

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

What remains is to show that $P$ is an integral polytope. Please refer to Appendix section EC.3.1 for an involved argument of this point. □

4.2. Sufficiency by Supermodularity

We now discuss a second sufficiency condition on $X$, based on a result from the study of Optimal Transport. It will identify a different subclass of tractable instances. Following this, we will compare and contrast with the first sufficiency condition of Theorem 3. In this section, submodular functions will be useful, so we briefly review here.

Let $\vee$ and $\wedge$ be binary operators defined on $\mathbb{R}^n \times \mathbb{R}^n$ by:

$$(x_1, \ldots, x_n) \vee (y_1, \ldots, y_n) = (x_1 \vee y_1, \ldots, x_n \vee y_n) \quad \forall x, y \in \mathbb{R}^n$$
\[(x_1, \ldots, x_n) \land (y_1, \ldots, y_n) = (x_1 \land y_1, \ldots, x_n \land y_n) \quad \forall x, y \in \mathbb{R}^n\]

We will refer to \(\mathbb{R}^n\) as a lattice with join and meet operations in \(\lor\) and \(\land\) respectively. Further, if a subset \(S \subseteq \mathbb{R}^n\) is closed under the join and meet operations, then it is a sublattice.

**Definition 4.2 (Submodular Function)** Let \(f : \mathbb{R}^n \to \mathbb{R}\). We call \(f\) submodular if

\[f(x \lor y) + f(x \land y) \leq f(x) + f(y) \quad \forall x, y \in \mathbb{R}^n.\]

If \(-f\) is submodular, we say \(f\) is supermodular.

As already stated, the sufficiency condition of this section is based on a standard result from Optimal Transport, in fact a generalization of the two-marginals (\(n = 2\)) result in Lemma 2 to a multi-marginal version (\(n \geq 3\)). Proof sketches and discussions abound in the Optimal Transport literature, and we provide a proof for sake of completeness in the Appendix.

**Lemma 3.** Let \(X_i\) be a totally ordered set for all \(i\), and consider \(X = X_1 \times \ldots \times X_n\). Let \(c : X \to \mathbb{R}\) be a continuous, submodular function. If \(\{\mu_i\}_{i=1}^n\) is any collection of marginal measures, where \(\mu_i \in \mathcal{P}(X_i)\) for all \(i\), then

\[
\inf_{\gamma : \Pi_i \gamma = \mu_i (\forall i)} \int_{X = X_1 \times \ldots \times X_n} c(x) \ d\gamma(x) = \int_0^1 c(F_{\mu_1}^{-1}(t), \ldots, F_{\mu_n}^{-1}(t)) \ dt,
\]

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

**Proof:** The technical proof is relegated to the Appendix section EC.3.2. □

The significance of this result for the MDM problem is that, assuming the marginals \(\mu_1, \ldots, \mu_n\) are all finite-supported, all the \(X\) that make \(Z(c) := \max_{x \in X} c^\top x\) supermodular form a subclass of instances for which the computation of \(Z^*\) is tractable. And this follows because the evaluation of the expectation of \(Z\) under the monotone coupling would be polynomial in the input.

A natural follow-up question is: what kinds of \(X\) yield supermodular \(Z\)? It turns out that in fact there is a class that can be quite conveniently characterized.

**Corollary 6.** Let \(\mu_1, \ldots, \mu_n\) all be finite-supported marginals. Consider the class of V-polytopes in which the vertex set \(X\) is a sublattice. Then for this class of inputs, \(Z^*\) is efficiently computable.
Proof: Define the function \( \delta_{\mathcal{X}}(x) := \begin{cases} 0 & x \in \mathcal{X} \\ +\infty & x \notin \mathcal{X} \end{cases} \). It follows that

\[
\delta_{\mathcal{X}}(x \lor y) + \delta_{\mathcal{X}}(x \land y) \leq \delta_{\mathcal{X}}(x) + \delta_{\mathcal{X}}(y) \quad \forall x, y \in \mathcal{X}
\]

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

**Corollary 7.** Let \( \mu_1, \ldots, \mu_n \) all be finite-supported marginals. For the class of H-polytopes that are sublattices, i.e., closed under the lattice operations \( \lor \) and \( \land \), \( Z^* \) is efficiently computable.

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

We now make a few remarks. The distinction between these two kinds of input is made because in fact a finite set \( \mathcal{X} \) being a lattice does not imply that \( \text{conv}(\mathcal{X}) \) is a lattice - unless \( n \leq 2 \), as shown in an example in Queyranne and Tardella (2006)- and vice versa. Checking the sublattice structure of \( \mathcal{X} \) in the case of V-polytope input can be clearly done efficiently. But for the case of H-polytope input, we can in fact also recognize if it is a sublattice in polynomial-time. Veinott Jr (1989) demonstrates a method via linear programming. Further, Queyranne and Tardella (2006) prove that all closed, polyhedral sublattices can be written as the solution set of a finite system of bimonotone linear inequalities (inequalities of the form \( a_i x_i + a_j x_j \leq c \), for some \( i, j \) with the product \( a_i a_j \) nonpositive), which could make recognition in the case of H-polytopes possible by inspection.

The next result indicates that when it comes to the case of two dimensions, the supermodularity of \( Z \) is equivalent to \( \text{conv}(\mathcal{X}) \) being a sublattice.

**Lemma 4.** Let \( n \leq 2 \), \( \mathcal{X} \) a finite subset of \( \mathbb{R}^n \), then the polytope \( \text{conv}(\mathcal{X}) \) is a sublattice if and only if \( Z \) is supermodular.
Proof: The “only if” direction was established previously for all \( n \). For \( n = 2 \), if \( Z \) is supermodular, then its conjugate \( \delta_{\text{conv}(\mathcal{X})} \) is submodular (see Murota (2003)), as desired. The case of \( n = 1 \) is trivial. \( \Box \)

To conclude this section, we now ask how this second sufficiency condition relates to the first one of Theorem 3, derived by way of considering the Dual Problem. How are they similar, and how are they different? The next result states that the two conditions both capture the class of \( L^b \) convex polytopes.

**Proposition 2.** Let \( \mathcal{X} \) be a finite point set such that \( \text{conv}(\mathcal{X}) \) is an \( L^b \) convex polytope. Then \( Z(c) := \max_{x \in \mathcal{X}} c^T x \) is supermodular. Consequently, \( Z^* \) is efficiently computable for the class of \( L^b \) convex polytopes when the marginals \( \mu_1, \ldots, \mu_n \) are finite-supported.

Proof: By hypothesis, \( \delta_{\text{conv}(\mathcal{X})} \) is polyhedral \( L^b \) convex. With \( Z \) being the conjugate of \( \delta_{\text{conv}(\mathcal{X})} \), it follows that \( Z \) is polyhedral \( M^b \) convex (see Fujishige (2005)), hence supermodular. \( \Box \)

As for how the sufficiency conditions are different, we emphasize that the class of \( \mathcal{X} \) that satisfy the sufficiency condition of Theorem 3 do indeed include \( \mathcal{X} \) that DO NOT result in supermodular \( Z \). We will see this illustrated in one of the three examples of the next section.

**5. Polynomial Time Solvable Instances**

In this section, we illustrate the application of the sufficiency conditions to identify three kinds of 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 et al. (2015). We show that this extends to the case where the entire marginal distributions are known. The second example is a ranking problem with random utilities with applications to allocating resources in a scheduling context. The third example deals with finding bounds for project scheduling problems with irregular, random starting time costs.

In light of Lemma 3 of the previous section, when we can recognize the supermodularity of \( Z \), we are justified in replacing \( \max_{\theta \in \Gamma(\mu_1, \ldots, \mu_n)} E_{\tilde{c} \sim \theta}[Z(\tilde{c})] \) with the expectation \( E[Z(\tilde{c})] \) under
the monotone coupling. And while this coupling will indeed present itself in the examples to follow (see Appointment Scheduling), we would like to emphasize that the monotone coupling is certainly not always the extremal coupling. For a very basic illustration, consider letting \( X = \{(-1,0,0), (0,-1,0), (0,0,-1)\} \) and \( \mu_1, \mu_2, \mu_3 = 1/3 \mathbb{1}_1 + 2/3 \mathbb{1}_0 \). Then

\[
\max_{\tilde{c}} \mathbb{E}[Z(\tilde{c})] = - \min_{\tilde{c}} \mathbb{E}\left[ \min_{x \in \mathcal{X}} \tilde{c}^T x \right],
\]

and the optimal coupling is found in: \( (\tilde{c}_1, \tilde{c}_2, \tilde{c}_3) = \frac{1}{3} (1,0,0) \), with expected value of 0. By contrast, the monotone coupling yields an expected value of \(-1/3\), indicating its suboptimality.

We now present the three examples, illustrating how the use of the sufficiency conditions of the previous section. To start, we finally address the problem we motivated the paper to begin with.

### 5.1. Appointment Scheduling

Recall that in this setting, we have \( n \) patients who arrive in a fixed order \( \{1, 2, \ldots, n\} \) who need to be scheduled in a given time interval \([0,T]\). We assume that for any patient \( i \), the distribution \( \mu_i \) 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 \( S = \{ s \in \mathbb{R}^n : \sum_{i=1}^n s_i \leq T, s_i \geq 0 \ \forall i \in [n] \} \), 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 weighted (positive integer weights \( d_i \)) collection of wait times of the \( n \) patients seen by one doctor, plus any overtime weighted by an integer \( \gamma \):

\[
\min_{s \in S} \sup_{\theta \in \Gamma(\mu_1, \ldots, \mu_n)} E_{\tilde{c} \sim \theta} [Z(s, \tilde{c})],
\]
where:

\[
Z(s, \tilde{c}) = \min \sum_{i=1}^{n} d_i w_i + \gamma w_{n+1}
\]

\[
\text{s.t. } w_1 = 0,
\]

\[
w_{i+1} \geq 0, \quad \forall i = 1, \ldots, n,
\]

\[
w_{i+1} \geq w_i + \tilde{c}_i - s_i, \quad \forall i = 1, \ldots, n,
\]

or equivalently,

\[
Z(s, \tilde{c}) = \max \sum_{i=1}^{n} (\tilde{c}_i - s_i) x_i
\]

\[
\text{s.t. } x_i - x_{i-1} \geq -d_i, \quad \forall i = 2, \ldots, n,
\]

\[
x_n \leq \gamma,
\]

\[
x_i \geq 0, \quad \forall i = 1, \ldots, n.
\]

5.1.1. Exploiting Supermodularity in MDM Appointment Scheduling

At first reading, the reader’s intuition may suggest 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 of supermodularity.

**Proposition 3.** For any \(s\), \(Z(s, \tilde{c})\) is supermodular in the \(\tilde{c}\) variables. Hence, the monotone coupling provides a worst-case coupling.

**Proof:** Let \(s\) be given. By Proposition 2, since \(Z(s, \cdot)\) is equivalent to the support function of an \(L^1\) convex polytope, it is necessarily supermodular. The final conclusion follows from Lemma 3.

In the case of finite-supported marginals, the problem then reduces to solving the LP given by

\[
\min_{s \in \mathcal{S}} E_{\tilde{c} \sim \theta_{\text{mon}}} [Z(s, \tilde{c})],\]

where \(\theta_{\text{mon}}\) denotes the monotone coupling of the marginals. As for the more general case of marginals that satisfy Assumption 1, we may address the problem with the dual formulation, an approach we discuss now.
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

$$\max_{\theta \in \Theta(\mu_1, \ldots, \mu_n)} E_{\theta}[Z(s, c)] = \max_{\theta \in \Theta(\mu_1-s_1, \ldots, \mu_n-s_n)} E_{\theta}[Z(0, c')]$$

$$= \min_{\{\nu_i: X \to \mathbb{R}\}_{i=1}^n} \max_{\nu \in \mathcal{P}(\mathcal{X})} \left( \int \psi_i d\Pi, \nu + \int \psi_i^*(\tilde{c}_i - s_i) d\mu_i \right),$$

where $\mu_i - s_i$ denotes the measure on $\mathbb{R}$ induced by the random variable $\tilde{c}_i - s_i$, when $\tilde{c}_i \sim \mu_i$.

Define the set $\mathcal{X} := \text{Extr} \left( \{ x \in \mathbb{R}_+: x_i - x_{i-1} \geq -d_i \ \forall i = 2, \ldots, n, \ x_n \leq \gamma \} \right)$. As $\mathcal{X}$ turns out to be a bounded $L^3$ convex set, Theorem 4 provides an integral extended form representation of $\text{conv}(\mathcal{X})$ as in Theorem 3.

However, for the sake of demonstrating an alternative representation here, let us consider all the weights to be 1, i.e., $d_i = 1$ for all $i$, and $\gamma = 1$, as in the introduction. Indeed, this $\mathcal{X}$ has in previous works been well-characterized (see Zangwill (1966), Zangwill (1969)) which we exploit as in Mak et al. (2015) to construct a representation of $\mathcal{P}(\mathcal{X})$ in the manner of Theorem 3. More precisely, consider the set

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

It can be shown that $\mathcal{X}$ and $\mathcal{X}'$ are in bijection, where members $x$ and $\xi$ of the respective sets satisfy the following:

$$x_i = u - i \iff i \in [l, u] \in \xi$$

Bringing this into the framework of Theorem 3, we let $B := \{(k, j) \in [n+1] \times [n+1] : k \leq j \}$, $B_i(\xi) := \{(k, i + \xi_i) : k \leq i \}$, and $P := \{ y \in [0, 1]^2 : \sum_{(k,j) \in i} y_{kj} = 1, \ \forall i \in [n+1] \}$, then with $P$ being a 0/1 polytope,

$$\text{conv}(\mathcal{X}) = \{ x : \exists y \in P \} \ x_i = \sum_{\xi_i = 0}^{n+1-i} \xi_i \cdot \sum_{(k, i + \xi_i) \in B_i(\xi_i)} y_{k, i + \xi_i} \}.$$ 

Hence, (*) takes the linear programming form:

$$\max_{\nu \in \mathcal{P}(\mathcal{X})} \sum_{x \in \mathcal{X}} \left[ \sum_{i=1}^n \psi_i(x_i) \right] \nu(x) = \max_{\nu \in \mathcal{P}(\mathcal{X})} \sum_{x \in \mathcal{X}} \psi_i(x_i) \cdot \Pi_i \nu(x_i) = \max_{\nu \in \mathcal{P}} \sum_{i=1}^n \sum_{x \in \mathcal{X}_i} \psi_i(x_i) \left( \sum_{(k, i + x_i) \leq i} y_{k, i + x_i} \right),$$
and we obtain the following tractable formulation for the distributionally robust appointment scheduling problem:

$$\min_{s \in S} \max_{\theta \in \Theta(\mu_1, \ldots, \mu_n)} E_\theta [Z(s, \hat{c})] = $$

$$\min_{s \in S} \min_{(d_{ij})_{i,j \leq n+1}} \max_{y \in P} \sum_{i=1}^{n} \sum_{j \in S \cup \{n\}} d_{ij} \left( \sum_{k \in S \cup \{i\}} y_{k,j} \right) + \sum_{i=1}^{n} E_{\hat{c}_i \sim \nu_i} \left[ \max_{j \in S \cup \{n\}} \left( (\hat{c}_i - s_i)(j - i) - d_{ij} \right) \right]$$  \hspace{1cm} (16)

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 Assumption 1.

5.2. Ranking with Applications to a Scheduling Problem

Consider \( \mathcal{X}_\text{perm} := \{ x : [n] \mapsto [n] \} \), the set of all permutations on \([n] := \{1, 2, \ldots, n\}\). Equivalently, this set \( \mathcal{X}_\text{perm} \) can be viewed as the set of extreme points to the permutahedron \( P_{\text{perm}} \) which is defined by the set of inequalities:

$$\sum_{i \in S} x_i \geq \frac{|S|(|S| + 1)}{2}, \ \forall S \subseteq [n], S \neq \phi,$$

$$\sum_{i=1}^{n} x_i = \frac{n(n+1)}{2}.$$

While this is not an efficiently-sized H-polytope representation, the 0/1 Birkhoff polytope shows that we can still efficiently solve the right-hand side of (*). Indeed, to see this, observe that if \( x \) is a permutation, then we can define an associated 0/1 permutation matrix \( Y \in \{0, 1\}^{n \times n} \) via \( Y_{i,j} = 1 \iff x(j) = i \), i.e., \( x = (1, \ldots, n) \cdot Y \). Let \( \mathcal{X}' \) be the set of such 0/1 permutation matrices, which is clearly in bijection with \( \mathcal{X}_\text{perm} \). Next, Birkhoff’s theorem tells us that the Birkhoff polytope \( \{ Y \in [0,1]^{n \times n} : \sum_i Y_{ij} = 1 \ \forall j, \ \sum_j Y_{ij} = 1 \ \forall i \} \) is in fact \( \text{conv}(\mathcal{X}') \); in other words, \( \text{conv}(\mathcal{X}') \) can be formulated as a 0-1 polytope in \( n^2 \) variables \( \{ y_{ij} \}_{i,j \in [n]} \). Thus, we set \( B = [n] \times [n], B_{ij}(x_i) = \{(i, x_i)\} \) for all \( i, x_i \in [n] \), and \( P \) as the Birkhoff polytope. This implies that the right-hand side of (*) can be formulated as a set of primal and dual linear programs:

$$\max_{\{y_{ij}\}_{i,j=1}^n} \sum_{i=1}^{n} \sum_{j=1}^{n} \psi_i(j) y_{ij} = \min_{\{\alpha_j\}_{j=1}^n, \{\beta_i\}_{i=1}^n} \sum_{j=1}^{n} \alpha_j + \sum_{i=1}^{n} \beta_i$$

s.t. \( \sum_j y_{ij} = 1, \ \forall j \),

s.t. \( \alpha_j + \beta_i \geq \psi_i(j), \ \forall i, j, \)

\( \sum_j y_{ij} = 1, \ \forall i \),

\( y_{ij} \geq 0, \ \forall i, j. \)
Using the dual formulation in (10), we have a min-min problem that is tractable.

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:

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

(17)

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

$$Z(t, \tilde{c}) := \max \sum_{i=1}^{n} (\tilde{c}_i - t_i)^+ x_i$$

subject to:

$$x \in \mathcal{X}_{\text{perm}},$$

(18)

where $t$ is the reduction in the individual job times which is assumed to lie in a polyhedral set $\mathcal{T}$.

Combining the discussion above with Theorem 2, problem (17) can be formulated as:

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

(19)

which by applying linear programming duality to the inner maximization problem reduces to the tractable (under Assumption 1) optimization problem:

$$\min \sum_{j=1}^{n} \alpha_j + \sum_{i=1}^{n} \beta_i + \sum_{i=1}^{n} E_{\tilde{c}_i \sim \mu_i} \left[ \max_{j \in [n]} \{(\tilde{c}_i - t_i)^+ j - d_{ij}\} \right]$$

subject to:

$$t \in \mathcal{T},$$

$$\alpha_j + \beta_i \geq d_{ij}, \quad \forall i, j.$$

(20)

To conclude this example, we now highlight some structure to this problem. In particular, whereas in the previous example, supermodularity was present. Here, we find the reverse in submodularity.
**Proposition 4.** For any $t \in \mathbb{R}^n$, 
\[
\max_{x \in \text{perm} \sum_{i=1}^{n} (\tilde{c}_i - t_i) x_i
\]

is monotone, submodular in the $\tilde{c}$ variables. Consequently, $Z(t, \tilde{c})$ is monotone, submodular in the $\tilde{c}$ variables.

**Proof:** We refer the reader to the Appendix section EC.4.1. □

The significance of Proposition 4 is that, though we may not conclude tractability (as with the supermodular case), we may still conclude a kind of approximation property. Indeed, this is the subject of the study conducted in Agrawal et al. (2012), which has something to say about a practice in stochastic optimization in which stochastic quantities are assumed to be independently distributed. Namely, with the established difficulty of computing the worst-case expected value of the cost in general, how much “excess cost” would we incur by assuming independence? Agrawal et al. (2012) answers this with a bound on the “Price of Correlations”, or POC, ratio.

**Corollary 8.** (Agrawal et al. (2012) Theorem 2) If $Z(t, \cdot)$ is integrable with respect to the independent coupling of the $\tilde{c}_i$ variables, then the price of correlation

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

where $t^R$ denote the robust solution and $t^I$ denote the solution under independent assumption.

### 5.3. Bounds for Project Scheduling with Random, Irregular Starting Time Costs

In this section, we discuss an extension of the polynomial time complexity results in Möhring et al. (2001) for project scheduling problems with irregular starting time costs to the case where randomness is incorporated in the cost function. Consider a set of jobs denoted by $N = \{1, \ldots, n\}$ with a fixed time horizon $\mathcal{T} = \{0, 1, \ldots, T\}$ 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 integer processing time of each job $j$ is denoted by $p_j$. Then this can capture an ordinary precedence constraint that job $j$ is started only after job $i$ by incorporating the constraint $S_j \geq S_i + p_i$. The precedence among the jobs is denoted by the digraph $G = (N, A)$ where $A = \{(i, j) | d_{ij} > -\infty\}$. We assume that there is no directed cycle of positive length in the graph, to prevent conflicts in scheduling. Assume that $\Gamma(\mu_1, \ldots, \mu_n)$ is the set of distributions for the random terms $\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_n$. The optimization problem is to solve:

\[
\inf_{\theta \in \Gamma(\mu_1, \ldots, \mu_n)} E \left[ \min_{S \in S_n} \sum_{j=1}^{n} \tilde{c}_j(S_j) \right], \tag{21}
\]

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

**Remark 3.** Consider

\[
\begin{align*}
\text{minimize} & \quad \frac{\epsilon_1}{\sqrt{S_1}} + \epsilon_2 S_2^{1/3} \\
\text{subject to} & \quad 1 \leq S_1 \leq 3 \\
& \quad 1 \leq S_2 \leq 2 \\
& \quad S_1 \leq S_2,
\end{align*}
\]

where the deterministic cost functions are defined by

\[
c^0_1(S_1) := \begin{cases} 
\frac{1}{\sqrt{S_1}} & S_1 \in [1, 3] \\
+\infty & \text{o.w.}
\end{cases}
\]

and

\[
c^0_2(S_2) := \begin{cases} 
S_2^{1/3} & S_2 \in [1, 2] \\
+\infty & \text{o.w.}
\end{cases}
\]

and we let the marginals be given by $\epsilon_1 = \epsilon_2 = \begin{cases} 
1 & \text{w.p.1/2} \\
0 & \text{w.p.1/2}
\end{cases}$. Then it is easy to verify that the monotone coupling yields an expected value of $\approx 0.980$, while the anti-monotone yields a smaller expected value at $\approx 0.853$. $\Delta$
With the above remark suggesting that there is no standard analytical characterization of the optimal coupling, we proceed with using the dual formulation of Theorem 2 and the extended formulation technique we illustrated in Section 4.

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

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

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

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

The linear programming relaxation of this integer program provides integral $x$ solutions (see the proof of Theorem 4). In other words, $\text{conv}(X)$ is a 0-1 polytope; hence, Theorem 3 implies that the robust bound in (21) in MDM can be computed efficiently.

To illustrate the tractable computational form:

$$\begin{align*}
\inf_{\theta \in \Gamma(\mu_1, \ldots, \mu_n)} E \left[ \min_{S \in S} \sum_{j=1}^n \tilde{c}_j(S_j) \right] \\
= -\sup_{\theta \in \Gamma(\mu_1, \ldots, \mu_n)} E \left[ -\min_{S \in S} \sum_{j=1}^n \tilde{c}_j(S_j) \right] \\
= -\sup_{\theta \in \Gamma(\mu_1, \ldots, \mu_n)} E \left[ \max_{S \in S} \sum_{j=1}^n -\tilde{c}_j(S_j) \right] \\
= -\min_{\psi_{jt}} \max_{x \in \text{conv}(X)} \sum_{j=1}^n \sum_{t=0}^T \psi_{jt} x_{jt} + \sum_{j=1}^n E_{\tilde{\epsilon}_j - \mu_j} \left[ \max_{x \in X_j} \left\{ -\tilde{c}_j - \sum_{t=0}^T c^0_j(t) \tilde{x}_{jt} - \sum_{t=0}^T \psi_{jt} \tilde{x}_{jt} \right\} \right] \\
= -\min_{(\psi_{jt}, \gamma_{ij}, \tilde{\gamma}_{ij}) \in Q} \sum_{j=1}^n \lambda_j + \sum_{(i, j) \in E} \gamma_{ij} + \sum_{j=1}^n E_{\tilde{\epsilon}_j - \mu_j} \left[ \max_{x \in X_j} \left\{ -\tilde{c}_j - \sum_{t=0}^T c^0_j(t) \tilde{x}_{jt} - \sum_{t=0}^T \psi_{jt} \tilde{x}_{jt} \right\} \right].
\end{align*}$$
where

\[ Q = \{ \psi, \lambda, \gamma : \lambda_j + \sum_{(i,t),(j,s) \in A,t=0,\ldots,s} \gamma_{js}^t + \sum_{(i,t),(i,j) \in A,t \in [(s-d_{ij}+1)v0,T]} \gamma_{ij}^t \geq \psi_{js}, \forall j \in \{1,\ldots,n\}, \forall s \in \{0,\ldots,T\}, \gamma_{ij}^t \geq 0, \forall (i,j) \in A, t \in \{0,\ldots,T\} \}, \]

and for each j, the integrand term

\[
\max_{\hat{x}_{jt} \in X_j} \{ -\tilde{\epsilon}_j - \sum_{t=0}^T c_j^0(t) \hat{x}_{jt} - \sum_{t=0}^T \psi_{jt} \hat{x}_{jt} \} = \\
\max \sum_{t=0}^T (-\tilde{\epsilon}_j c_j^0(t) - \psi_{jt}) \hat{x}_{jt} \\
s.t. \sum_{t=0}^T \hat{x}_{jt} = 1, \forall j = 1,\ldots,n, \\
\sum_{s=0}^{t-d_{ij}-1} \hat{x}_{is} + \sum_{s=0}^T \hat{x}_{js} \leq 1, \forall (i,j) \in A, t = 0,\ldots,T, \\
\hat{x}_{jt} \geq 0, \forall j = 1,\ldots,n, t = 0,\ldots,T.
\]

has an equivalent dual minimization form:

\[
\begin{aligned}
\minimize_{\{\alpha_{j'},\alpha_j\}} \sum_{j'=1}^n \alpha_j' + \sum_{(i,j',t),(i,j) \in A,t \in [0,\ldots,T]} \beta_{ij'}^t \\
s.t. \alpha_j' + \sum_{(i,j,t),(j',i) \in A,t \in [0,\ldots,s]} \beta_{j'j}^t + \sum_{(i,j,t),(i,j') \in A,t \in [(s-d_{ij}+1)v0,T]} \beta_{ij'}^t \geq 0, \forall j' \neq j, s \in \{0,\ldots,T\} \\
\alpha_j + \sum_{(i,t),(j,i) \in A,t \in [0,\ldots,s]} \beta_{ij}^t + \sum_{(i,t),(i,j) \in A,t \in [(s-d_{ij}+1)v0,T]} \beta_{ij'}^t \geq -\tilde{\epsilon}_j c_j^0(s) - \psi_{js}, \forall s \in \{0,\ldots,T\} \\
\beta_{ij}^t \geq 0 \forall (i,j') \in A, t \in \{0,\ldots,T\} \\
\end{aligned}
\]

6. Computational Experiments

6.1. Appointment Scheduling

Mak et al. (2015) have studied the appointment scheduling problem considered in this paper with first two marginal moments. They used the second order cone to model the distributionally robust version of the problem. Although tractable, the distributionally robust model in which the uncertainty set is comprised of all joint distributions in which the marginals satisfy given first two marginal moments - hereafter referred to as the Marginal Moment Model (MMM model) - is often challenged for the conservativeness of its solution. Generally, incorporating more information of
uncertainties helps to mitigate this effect at the cost of sacrificing tractability. In Section 5.1, though, we have shown that incorporating marginal distribution information can yield a tractable model. We are interested in understanding to what extent incorporating marginal distribution information manages to mitigate the conservativeness of the marginal moment based model. Towards this, we investigate the performance of the robust solutions from MDM and MMM model respectively when the distributions of the service durations are known and independent.

Our experimental setup is based on Mak et al. (2015). Namely, we consider the case of \( n = 5 \) jobs, and assume that the 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 \( \sigma_i \), we can solve a second order conic model proposed in Mak et al. (2015) to get the robust solution under the marginal moment based model, denoted as an MMM solution. We can also solve (16) 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 \( \sigma_i \). We use the sample average approximation method with \( N = 5000 \) samples to compute univariate expectation in (16). 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
Figure 1  Performance Comparison Between MDM Solution and MMM Solution (Normal)

the three cases are 0.2462 (Normal), 0.2420 (Log-Normal), 0.2001 (Gamma). The results show that the usage of only marginal moments will incur some loss in getting an efficient schedule.
Figure 2  Performance Comparison Between MDM Solution and MMM Solution (Log-Normal)

Figure 3  Performance Comparison Between MDM Solution and MMM Solution (Gamma)
6.2. Ranking with Scheduling

We have shown the ranking problem (2) is submodular in \( \tilde{c} \). Hence the price of correlation is low for the ranking problem according to Corollary 8. Indeed, we can test it using a small numeric example. Consider the resource allocation application in Section 5.2. 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:

- 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 4. 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 (17) 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 (20) 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.5 for the formula 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.

7. Conclusion

We studied the problem of computing the robust bound on the expected optimal value of linear and discrete optimization problems, in which the stochastic uncertainty is specified only up to its marginal distributions. We generalized and unified primal-dual formulations for this problem. And while we showed the robust bound is in fact NP-hard to compute for linear optimization over an $H$-polytope, we also used the dual formulation to find a sufficient condition on the input polytope for polynomial time solvability. In the process, we showed that the class of $L^k$ convex polytopes form a tractable subclass. Finally, we concluded by examining some application problems in scheduling.

Future directions of research may examine how to incorporate additional kinds of constraints on the joint in addition to the marginal constraints. Another direction is to consider when we can not only efficiently compute the bound but also efficiently construct a worst-case distribution. Other directions could consider costs apart from support functions, like costs arising from a multi-stage process.
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


Appendix (Proofs of Statements)

EC.1. Section 2 Proofs

**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:

$$\max \sum_{S \subseteq [n]} p_S \max \left\{ \sum_{i \in S} x_i - \sum_{i \in S^c} x_i : Ax \leq b \right\} \quad \text{subject to} \quad \sum_{S \subseteq [n]} p_S = p_i, \quad \forall i \in [n],$$

$$\sum_{S \subseteq [n]} p_S = 1,$$

$$p_S \geq 0, \quad \forall S \subseteq [n],$$

(EC.1)

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:

$$\min y_0 + \sum_{i \in [n]} p_i y_i \quad \text{subject to} \quad y_0 + \sum_{i \in S} y_i \geq \max \left\{ \sum_{i \in S} x_i - \sum_{i \in S^c} x_i : Ax \leq b \right\}, \forall S \subseteq [n].$$

(EC.2)

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:

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

(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

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

which is equivalent to

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

\[= \max_{x : Ax \leq b} \|x\|_1.\]

The right hand side corresponds to 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. \( \square \)

**EC.2. Section 3 Proofs**

**EC.2.1. Proof of Theorem 2**

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

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

(†)

**Proof:** We establish the equalities in order, justifying the “max” notations along the way.

**Equality (†):**

**From “sup” to “max”:** Regarding the first equality, the left hand side max will follow shortly, when an optimal solution \( \theta^* \) will be constructed. On the right hand side, each of the inner max problems indexed by \( i \) are attained because of Theorem 1. To see that the outer max is attained, note that \( \mathcal{P}(X) \) is compact, as \( X \) is a finite set. Next, observe that the mapping \( \nu \mapsto \)
\[ \sum_{i} \max_{\gamma_i \in \Gamma(\mu_i, \Pi_i \nu)} E(c_i, x_i) - \gamma_i [\tilde{c}_i x_i] \text{ is concave and upper semi-continuous for all } i. \] This follows because if for arbitrary \( i \), we let:

\[ F(\psi_i) := - \int \psi_i^* \, d\mu_i, \quad \forall \psi_i : X_i \to \mathbb{R}, \]

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

\[ F^*(\nu_i) := \begin{cases} 
\max_{\gamma_i \in \Gamma(\mu_i, \nu_i)} E(c_i, x_i) - \gamma_i [\tilde{c}_i x_i], & \nu_i \in \mathcal{P}(X_i), \\
-\infty, & \text{otherwise,}
\end{cases} \]

revealing that the mapping \( \Pi_i \nu \mapsto \max_{\gamma_i \in \Gamma(\mu_i, \Pi_i \nu)} E(c_i, x_i) - \gamma_i [\tilde{c}_i x_i] \) is indeed concave and upper semi-continuous. As a result, \( \nu \mapsto \max_{\gamma_i \in \Gamma(\mu_i, \Pi_i \nu)} E(c_i, x_i) - \gamma_i [\tilde{c}_i x_i] \) is concave and upper semi-continuous for all \( i \), as desired.

**Establishing “\( = \)”**: We establish the first equality now. Let \( \theta \in \Gamma(\mu_1, \ldots, \mu_n) \). Then, taking any measurable selection \( x^* \) s.t. \( x^*(c) \in \arg\max_{x \in X} \sum_i c_i x_i \) \( \forall \, c \), define \( \nu := x^* \# \theta \) and \( \gamma_i := ((\text{Proj})_i x_i^*) \# \theta \in \Gamma(\mu_i, \Pi_i \nu) \) to find:

\[ E_{\theta} \left[ \max_{x \in X} \sum_{i=1}^n \tilde{c}_i x_i \right] = \int_{\mathbb{R}^n} \sum_{i} \tilde{c}_i x_i^* (\tilde{c}) \, d\theta, \]

\[ = \sum_{i} \int_{\mathbb{R} \times X_i} \tilde{c}_i x_i \, d\gamma_i, \]

\[ \leq \max_{\nu \in \mathcal{P}(X)} \sum_{i} \max_{\gamma_i \in \Gamma(\mu_i, \Pi_i \nu)} E(c_i, x_i) - \gamma_i [\tilde{c}_i x_i], \]

where \( \text{Proj}(\tilde{c}) := \tilde{c}_i \). Hence, \( \max_{\theta \in \Gamma(\mu_1, \ldots, \mu_n)} E_{\theta} \left[ \max_{x \in X} \sum_{i=1}^n \tilde{c}_i x_i \right] \leq \max_{\nu \in \mathcal{P}(X)} \sum_{i} \max_{\gamma_i \in \Gamma(\mu_i, \Pi_i \nu)} E(c_i, x_i) - \gamma_i [\tilde{c}_i x_i] \).

What remains is to show tightness by showing the reverse inequality. Let \( \tilde{\nu} \in \mathcal{P}(X) \), be optimal to the right hand side, with \( \gamma_i \in \Gamma(\mu_i, \Pi_i \tilde{\nu}) \) solving the i-th transport problem in the right hand side. Towards this end, we make use of the following consequence of the Disintegration theorem Ambrosio et al. (2008).

**Lemma EC.1.** Let \( X \) be a compact subset of \( \mathbb{R} \). Let \( \gamma \) be a Borel probability measure defined on the product set \( \mathbb{R} \times X \), and with \( \Pi_X : \mathbb{R} \times X \to X \) being the projection function, suppose \( \nu = \Pi_X \# \gamma \).

Then there exists a family of probability measures on \( \mathbb{R} \), indexed by members \( x \) in \( X \), denoted by \( \{\mu^x\}_{x \in X} \), such that

\[ \gamma(B \times E) = \int_E \int_B d\mu^x d\nu = \int_E \mu^x(B) \, d\nu \]
We will refer to this “disintegration” of the measure $\gamma$ with $\gamma = \mu^x \otimes \nu$ as shorthand.

By Lemma EC.1, there exists a collection of measures $\{\mu_i^x\}_{x \in \mathcal{X}_i}$, where $\mu_i^x$ is a distribution over $\mathbb{R}$ for each $x_i \in \mathcal{X}_i$, so that $\gamma_i = \mu_i^x \otimes \Pi_i \nu$. Then,

$$
\max_{\nu \in \mathcal{P}(\mathcal{X})} \sum_i \max_{\gamma_i \in \Gamma(\mu_i, \Pi_i, \nu)} E(\tilde{c}_i, x_i) - \gamma_i \tilde{c}_i x_i = \sum_i \int \tilde{c}_i x_i \, d\gamma_i = \sum_i \int \tilde{c}_i x_i \, (\mu_i^x \otimes \Pi_i \nu) \\
= \sum_i \int \left[ \int \tilde{c}_i x_i \, \mu_i^x (d\tilde{c}_i) \right] \Pi_i \nu (dx_i) \\
= \int \left[ \sum_i \int \tilde{c}_i x_i \, \mu_i^x (d\tilde{c}_i) \right] \nu (dx) \\
= \int \left[ \int \left( \sum_i \tilde{c}_i x_i \right) d(\mu_1^x \otimes \ldots \otimes \mu_n^x) \right] \nu (dx).
$$

In order to relate this last integral to $\max_{\theta \in \Gamma(\mu_1, \ldots, \mu_n)} E_{\tilde{c} \otimes \theta} \left[ \max_{x \in \mathcal{X}} \sum_{i=1}^n \tilde{c}_i x_i \right]$, let us rewrite it as an integral with respect to a measure $\eta$, defined in the following way.

$$
\eta(E \times \{x\}) := (\theta^*)^x (E) \cdot \nu(x) \quad \text{for any Borel measurable set } E, \ x \in \mathcal{X},
$$

with $(\theta^*)^x := (\mu_1^x \otimes \ldots \otimes \mu_n^x)$. In other words,

$$
\int \left[ \int \left( \sum_i \tilde{c}_i x_i \right) d(\theta^*)^x \right] \nu (dx) = \int \sum_i \tilde{c}_i x_i \, d\eta(c, x).
$$

And notice when we project the measure $\eta$ onto $\mathbb{R}^n$ to obtain a measure, call it $\theta^* \in \mathcal{P}(\mathbb{R}^n)$,

$$
\theta^*(E) := \int (\theta^*)^x (E) \, d\nu(x) \quad \text{for any Borel measurable set } E,
$$

then for arbitrary $i$ and Borel set $B_i \subseteq \mathbb{R}$,

$$
\theta^*(\mathbb{R}^{i-1} \times B_i \times \mathbb{R}^{n-i}) = \int (\mu_1^x \otimes \ldots \otimes \mu_n^x)(\mathbb{R}^{i-1} \times B_i \times \mathbb{R}^{n-i}) \, d\nu(x) \\
= \int \mu_i^x (B_i) \, d\nu(x) = \int \mu_i^x (B_i) \, d\Pi_i \nu(x_i) \\
= \gamma_i (B_i \times \mathcal{X}_i) = \mu_i (B_i),
$$

which means $\theta^* \in \Gamma(\mu_1, \ldots, \mu_n)$. Therefore,

$$
\max_{\nu \in \mathcal{P}(\mathcal{X})} \sum_i \max_{\gamma_i \in \Gamma(\mu_i, \Pi_i, \nu)} E(\tilde{c}_i, x_i) - \gamma_i \tilde{c}_i x_i = \int \left[ \int \left( \sum_i \tilde{c}_i x_i \right) d(\mu_1^x \otimes \ldots \otimes \mu_n^x) \right] \nu (dx) \leq \max_{\theta \in \Gamma(\mu_1, \ldots, \mu_n)} E_{\tilde{c} \otimes \theta} \left[ \max_{x \in \mathcal{X}} \sum_{i=1}^n \tilde{c}_i x_i \right],
$$

as desired. Further, this establishes $\theta^*$ as an optimal solution.
For a more explicit construction of $\theta^*$ that could possibly provide more intuition, we refer the reader to Remark EC.1. In probabilistic terms, $(\theta^*)^x$ provides a joint conditional probability distribution $(\tilde{c}_1, \ldots, \tilde{c}_n)|_x$, while $\tilde{c}_i|_{x_i} \sim \mu_i^{c_i}$. Further, under this joint conditional distribution, for all $i$, the $\tilde{c}_i$ are independent, conditioned on $x = (x_1, \ldots, x_n) \in \mathcal{X}$; in fact, for each $i$, $\tilde{c}_i$ is independent of all $x_j$ for $j \neq i$, conditioned on $x_i$.

**Equalities (Primal) and (Dual):** To establish the remaining equalities, let us note that for an arbitrary $i$, since $\tilde{c}_i \cdot x_i \leq \tilde{c}_i \cdot \max_{x_i \in \mathcal{X}_i} x_i$, for all $\tilde{c}_i$ and $x_i$, and since all marginal probability measures $\mu_i$ are assumed to have finite mean, applying Theorem 1 yields:

$$
\max_{\gamma_i \in \Gamma(\mu_i, \Pi_i, \nu)} E_{(\tilde{c}, \gamma)}[\tilde{c}_i, x_i] = \inf_{(\psi_{i, X_i} \rightarrow \mathbb{R})} \int \psi_i d\Pi_i \nu + \int \max_{x_i} \{\tilde{c}_i x_i - \psi_i(x_i)\} d\mu_i.
$$

Hence,

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

$$
= \max_{\nu \in \mathcal{P}(\mathcal{X})} \sum_i \inf_{(\psi_{i, X_i} \rightarrow \mathbb{R})} \int \psi_i d\Pi_i \nu + \int \max_{x_i} \{\tilde{c}_i x_i - \psi_i(x_i)\} d\mu_i, \quad (\text{Theorem 1})
$$

$$
= \max_{\nu \in \mathcal{P}(\mathcal{X})} \inf_{(\psi_{i, X_i} \rightarrow \mathbb{R})} \sup_{i = 1}^n \sum_i \int \psi_i d\Pi_i \nu + \int \max_{x_i} \{\tilde{c}_i x_i - \psi_i(x_i)\} d\mu_i,
$$

where the last equality is established using a “partial saddle point” result (see Proposition 2.3-Remark 2.3 in Ekeland and Temam (1999)). To conclude the proof, we must argue that the infimum in the dual problem is attained, and this follows because $L(\cdot, \cdot)$ is “upper closed concave-convex” and hence is necessarily the Lagrangian of a convex program (see Theorem 36.5 in Rockafellar (1997)), whose dual program has a feasible solution in the relative interior (Slater’s Condition / dual “strong consistency”), so that the infimum is indeed attained (see Corollary 30.5.2 in Rockafellar (1997)). $\Box$

**Remark EC.1 (Explicit Construction).** Following the outline of the proof of Theorem 2, we provide here for the reader a more explicit construction of the optimal distribution $\theta^* \in \Gamma(\mu_1, \ldots, \mu_n)$
for the problem $\max_{\theta \in \Gamma(\mu_1, \ldots, \mu_n)} E_{\Xi, \theta}Z(\tilde{c})$. Letting $\tilde{\nu} \in \mathcal{P}(\mathcal{X})$ be an optimal solution to the right hand side of the inequality, we will construct a measure $\theta^*$ of the following (disintegrated) form:

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

where $(\theta^*)^x = (\mu_1^{x_1} \otimes \ldots \otimes \mu_n^{x_n})$. With $\theta^*$ 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 $L_{i,x_i} = F_{\mu_i}^{-1}(\Pi, \tilde{\nu}((\infty, x_i - e_i])$ and $R_{i,x_i} = F_{\mu_i}^{-1}(\Pi, \tilde{\nu}((\infty, x_i])$. Further, to capture the (possible) discontinuous jumps in $F_{\mu_i}$ at $L_{i,x_i}$ and $R_{i,x_i}$, define the following nonnegative quantities:

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

$$J_{i,x_i}^{R} = \Pi, \tilde{\nu}(-\infty, x_i] - \lim_{c_i \uparrow R_{i,x_i}} F_{\mu_i}(c_i).$$

Define the measure as follows:

1. If $L_{i,x_i} = R_{i,x_i}$, then define the desired measure $\mu_i^{x_i}$ for $x_i$ s.t. $\Pi, \tilde{\nu}(x_i) > 0$, by

$$\mu_i^{x_i} := \delta_{L_{i,x_i}}$$

2. Else when $L_{i,x_i} < R_{i,x_i}$, define the desired measure $\mu_i^{x_i}$, for $x_i$ s.t. $\Pi, \tilde{\nu}(x_i) > 0$, by:

$$\mu_i^{x_i}(B) := \frac{J_{i,x_i}^{L} \cdot 1_{E_{i,x_i} \in B} + \mu_i(B \cap (L_{i,x_i}, R_{i,x_i})) + J_{i,x_i}^{R} \cdot 1_{R_{i,x_i} \in B}}{\Pi, \tilde{\nu}(x_i)} \leq 1, \quad \forall B \in \mathcal{B}$$

One can verify that this is indeed a probability measure, with $I_{i,x_i}$ as support, as:

$$\mu_i^{x_i}(I_{i,x_i}) = \frac{J_{i,x_i}^{L} + J_{i,x_i}^{R} + \mu_i((L_{i,x_i}, R_{i,x_i}))}{\Pi, \tilde{\nu}(x_i)} = \frac{\Pi, \tilde{\nu}(\infty, x_i] - \Pi, \tilde{\nu}(\infty, x_i - e_i]}{\Pi, \tilde{\nu}(x_i)} = 1.$$
One can verify that this definition satisfies the marginal constraint:
\[ \Pi_i \theta^* = \Pi_i \left[ \sum_{x \in \mathcal{X}} \left( \mu_1^{x_1} \otimes \ldots \otimes \mu_n^{x_n} \right) \cdot \tilde{\nu}(x) \right] = \sum_{x \in \mathcal{X}} \mu_i^{x_i} \cdot \tilde{\nu}(x) = \mu_i, \quad \forall i, \]
so that \( \theta^* \) indeed is an admissible joint distribution, i.e., \( \theta^* \in \Gamma(\mu_1, \ldots, \mu_n) \). The intuition comes from Lemma 2, which tells us that an optimal coupling of the random variables \( \hat{c}_i \sim \mu_i \) and \( \hat{x}_i \sim \Pi_i \tilde{\nu} \) can be obtained if we “match up” quantiles. Thus, the above simply proposes that we concentrate \( \hat{c}_i \) on the interval \( I_i, x_i \), given knowledge that the random variable \( \hat{x} \sim \tilde{\nu} \) has a realization \( x \) with \( i \)-th component equal to \( x_i \). \( \Delta \)

**EC.2.2. Proof of Proposition 1**

**Proposition 1.** Suppose that the Borel probability measures \( \mu_1, \ldots, \mu_n \) are all absolutely continuous. Then there exists a measurable function \( x^* : \mathbb{R}^n \rightarrow \mathcal{X} \) that takes the form \( x^*(c) = (x_1^*(c_1), \ldots, x_n^*(c_n)) \) and satisfies \( x^*(c) \in \arg\max_{x \in \mathcal{X}} c^T x \) for all \( c \in \mathcal{X}_{i=1}^n \supp(\mu_i) \), such that for any primal optimal solution \( \tilde{\nu} \),
\[
P_{c \sim \theta^*}(x_i^*(\hat{c}) = x_i) = P_{c \sim \mu_i}(x_i^*(\hat{c}) = x_i) = \Pi_i \tilde{\nu}(x_i), \quad \forall x_i \in \mathcal{X}_i, \quad \forall i \in [n], \tag{EC.4}
\]
where \( \theta^* \in \Gamma(\mu_1, \ldots, \mu_n) \) is any optimal solution to the MDM problem (5).

**Proof:** Suppose \( \mu_1, \ldots, \mu_n \) are absolutely continuous, \( \tilde{\nu} \) is a primal optimal solution. Let \( \theta^* \) be constructed as in the proof of Theorem 2. By tightness of (EC.3), we find
\[
\int \left[ \sum_i \hat{c}_i x_i \right] (\theta^*)^x(d\hat{c}) = \int Z(\hat{c}) (\theta^*)^x(d\hat{c}) \quad \forall x : \tilde{\nu}(x) > 0.
\]
This tells us that for \( x \) that has positive \( \nu \) probability mass, \( x \in \arg\max_{x \in \mathcal{X}} c^T x \), \( (\theta^*)^x \) - almost surely. Consequently, this motivates the definition of a measurable mapping \( x^*(\cdot) : \mathbb{R}^n \rightarrow \mathcal{X} \) that maps every \( c \in \mathcal{X}_{i=1}^n \supp(\mu_i) \) to a maximizing solution:

For all \( i \), \( x_i^*(c_i) := \sum_{x_i \in \Pi_i \supp(\tilde{\nu})} x_i \cdot 1_{c_i \in \supp(\mu_i^{x_i})} \),

with \( x_i^*(c_i) \) defined to be an arbitrary constant when \( c_i \notin \cup_{x_i \in \Pi_i \supp(\tilde{\nu})} \supp(\mu_i^{x_i}) = \supp(\mu_i) \)- to guarantee measurability.
Regarding $c_i$ that could possibly lie in more than one support, by Lemma 2, for any $i$, the supports $\{\text{supp}(\mu_i^{x_i})\}_{x_i \in \mathcal{X}_i}$ are intervals, with pairwise disjoint interiors that cover the support of $\mu_i$. Hence, the points of intersection to these intervals form a set of $\mu_i$- zero measure. With the well-definedness and measurability issues aside, it can be verified that, given any $x \in \text{supp}(\bar{\nu})$, $x^*(c) = x \in \text{argmax}_{x \in \mathcal{X}} c^T \bar{x}$, $(\theta^*)_x$ - almost surely; further, $\forall x_i \in \mathcal{X}_i, ~ \forall i \in [n]$

$$P_{\bar{\nu},\theta^*}(x_i^*(\bar{c}_i) = x_i) = \frac{\sum (\theta^*)_x (x^*_i(\bar{c}_i) = x_i) \cdot \bar{\nu}(\bar{x})}{\sum_{\bar{x} : \bar{x}_i = x_i} (\theta^*)_x (x^*_i(\bar{c}_i) = x_i)} = \frac{\sum \bar{\nu}(\bar{x}) = \Pi_i \bar{\nu}(x_i)}{\text{supp}(\mu_i)}.$$  

Finally, we show that with this definition of $x^*$, equation (EC.4) holds not just for the specific primal optimal solution but for all primal optimal solutions. Begin by observing that $x^*$ has the separable form $x^*(c) = (x^*_1(c_1), \ldots, x^*_n(c_n))$ for all $c \in \mathbb{R}^n$. Further, observe that, for all $i$, the action of $x_i^*$ over $\bigcup_{x_i \in \text{supp}(\mu_i)} \text{supp}(\mu_i^{x_i})$ is determined by the corresponding marginal $\bar{\nu}$ to the primal optimal solution $\bar{\nu}$. Since Corollary 3 reveals that these marginals are unique, for any other possible primal optimal solution $\tau_\ast$, the resulting collection of $x_i^*$ functions would have the exact same action over $\text{supp}(\mu_i)$. □

**Remark EC.2 (On Primal Optimal Solutions and Persistency).** Consider now when $\mu_1, \ldots, \mu_n$ are not absolutely continuous but in fact finite-supported measures. For a first example, consider $n = 1$, $\mu_1 = \mathbb{1}_0$, and $\mathcal{X} = \{0, 1\}$. Then any distribution over $\mathcal{X}$ can solve (8). Let us consider one in particular in $\bar{\nu}$ defined by $\bar{\nu}(0) = \bar{\nu}(1) = 1/2$. Then it is clear that no solution mapping (i.e., a measurable selection of $x^{OPT}$) $x^*$ exists such that

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

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

This might lead one to ask, in the case of $\mu_1, \ldots, \mu_n$ that are not absolutely continuous, will there always exist a primal optimal $\bar{\nu}$ such that a measurable selection of $x^{OPT}$, call it $x^*$, of form $x^*(c) = (x^*_1(c_1), \ldots, x^*_n(c_n))$ can be defined to yield the “persistence value” statements? The following example indicates that the answer is no. Consider $n = 2$, $\mathcal{X} = \{(0,1),(1,0)\},$
Then the unique optimal $\tilde{\nu}$ to (8) is $\tilde{\nu} = 1/2 \cdot 1_{(0,1)} + 1/2 \cdot 1_{(1,0)}$. By inspection, or alternatively by

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

Remark EC.1, we find that the unique $\theta^*$ is given by: $$(\tilde{c}_1, \tilde{c}_2) = \begin{cases} 
(1,2.5) & \text{w.p. } 1/4 \\
(2,2.5) & \text{w.p. } 1/4 \\
(3,1.5) & \text{w.p. } 1/5 \\
(3,2.5) & \text{w.p. } 3/10
\end{cases}$$

Any solution mapping $x^*$ maps all $(c_1, c_2) : c_1 < c_2$ to $(0,1)$ and all $(c_1, c_2) : c_1 > c_2$ to $(1,0)$, with arbitrary decision for the case of $c_1 = c_2$. So an example solution mapping $x^*$ would involve:

$$
x^*(c_1, c_2) = \begin{cases} 
(0,1) & \text{if } (c_1, c_2) \in \{(1,2.5), (2,2.5)\} \\
(1,0) & \text{if } (c_1, c_2) \in \{(3,1.5), (3,2.5)\}
\end{cases}
$$

so that $x^* \# \theta^* = \tilde{\nu}$. But, $x^*$ is not of the form $x^* = (x_1^*(c_1), \ldots, x_n^*(c_n))$.

$\Delta$

**EC.3. Section 4 Proofs**

**EC.3.1. Proof of Theorem 4**

**Theorem 4** Let $\tilde{c}_1, \ldots, \tilde{c}_n$ be real-valued random variables with $\mu_1, \ldots, \mu_n$ denoting the probability measures they induce on the real line, and satisfying Assumption 1. For these measures and the class of $L^1$ convex polytopes, $Z^*$ is efficiently computable.

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

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

The “only if” direction is clear. The “if” direction is also clear, as given an $(i, j) \in A$, $x_i \in \{x_i, \ldots, u\} \Rightarrow x_j \notin \{l, \ldots, x_i-d_{ij}-1\}$, by hypothesis. Hence, $x_j \geq x_i - d_{ij}$, as desired.
Using indicator functions, we can view the equivalence just as well in the form:

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

Additionally, since \( x_i = \sum_{x=1}^{u} \bar{x}_i \cdot \bar{x}_i \), we have found an extended form of the kind in Theorem 3, with

\[ P := \{ y \in \mathbb{R}_+^{n \times [l, u]} : \sum_{x_i \in [l_1, u_1]} y_{i,x_i} = 1 \forall i, \sum_{s=1}^{l-1} y_{i,s} + \sum_{t=1}^{u-1} y_{j,s} \leq 1 \forall (i, j) \in \mathcal{A}, t = l, l + 1, \ldots, u \} \]

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

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

Next, we will assign an orientation to the edges to show that this graph has a transitive orientation. To do this, for \( (i, j) \in \mathcal{A} \), let the edge between nodes \( x_{i,t} \) and \( x_{j,t'} \) (where \( t' \leq t - d_{ij} - 1 \)) be drawn from \( x_{i,t} \) towards \( x_{j,t'} \). Let the edges among the clique \( \{x_{i,s}\}_{s \in [l, u] \cap \mathbb{Z}} \) be drawn from \( x_{i,s} \) towards \( x_{i,t} \) if and only if \( s \geq t \). Now, we verify the transitivity. Observe that by the triangle inequality of \( d \), for \( (i, j), (j, k) \in \mathcal{A} \),

\[ x_{i,t} \rightarrow x_{j,t'} \land [x_{j,t'} \rightarrow x_{k,t''}] \implies [t' \leq t - d_{ij} - 1] \land [t'' \leq t' - d_{jk} - 1] \implies t'' \leq t - d_{ik} - 2 < t - d_{ik} - 1 \implies x_{i,t} \rightarrow x_{k,t''}, \]

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

\[ x_{i,t} \rightarrow x_{i,t'} \land x_{i,t'} \rightarrow x_{j,t''} \implies t' \leq t \land t'' \leq t' - d_{ij} - 1 \implies t'' \leq t - d_{ij} - 1 \implies x_{i,t} \rightarrow x_{j,t''}. \]
This suffices to conclude that the orientation exhibits transitivity in this directed graph, making $G$ a comparable graph. Then, by Mirsky’s theorem, $\bar{G}$ is a perfect graph. By an established result, perfectness yields that the 0/1 stable set polytope for $\bar{G}$ is characterized by the following collection of facet-defining inequalities: $\{ z \in \mathbb{R}^\bar{V} : \sum_{v \in K} z_v \leq 1 \}$, where $K \subset V$ is some maximal clique in $G$.

Observe that for arbitrary $(i,j) \in A$ and $t \in \{1, \ldots, u\}$, the collection of nodes $\{x_{i,s}\}_{s=t}^{u} \cup \{x_{j,s}\}_{s=t}^{u-d_{ij}-1}$ forms precisely a maximal clique - further, it is clear that all maximal cliques of $\bar{G}$ are of this form.

Hence, we can view $P$ as a face of the stable set polytope of $\bar{G}$, concluding it is integral (as the stable set polytope is integral), as desired.

**Remark EC.3 (Inexact Projections $X_i$).** In the dual formulation of Theorem 2, for any $i$ the potential function $\psi_i$ is defined on $X_i$. In practice, we may not be able to identify $X_i$ precisely.

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

$$\inf_{\{\psi_i : X_i \rightarrow \mathbb{R}\}_{i=1}^{n}} \sup_{\nu \in \mathcal{P}(\mathcal{X})} \sum_{i} \int \psi_i d\Pi_i \nu + \int \max_{x_i \in \mathcal{X}_i} \{ \tilde{c}_i x_i - \psi_i(x_i) \} d\mu_i$$

This follows because the potential functions $\psi_i$ in both formulations above may be assumed to be convex as a result of Remark 2; further, if $x'_i \notin \mathcal{X}_i$, then $\Pi_i \nu(x'_i) = 0$ for all $\nu \in \mathcal{P}(\mathcal{X})$. More precisely, extending the definition of $\psi_i$ from $\mathcal{X}_i$ to $\mathcal{X}'_i$ - in a manner that preserves convexity - can only result in increasing $\psi_i^*$ (if any change at all) pointwise. And with $\psi_i(x'_i) \cdot \Pi_i \nu(x'_i) = 0$, for any $x'_i \in \mathcal{X}'_i \setminus \mathcal{X}_i$, we see that no such extension could help with minimization.

\[\triangle\]

**EC.3.2. Proof of Lemma 3**

**Lemma 3** Let $\mathcal{X}_i$ be a totally ordered set for all $i$, and consider $\mathcal{X} = \mathcal{X}_1 \ldots \mathcal{X}_n$. Let $c : \mathcal{X} \rightarrow \mathbb{R}$ be a continuous, submodular function. If $\{\mu_i\}_{i=1}^{n}$ is any collection of marginal measures, where $\mu_i \in \mathcal{P}(\mathcal{X}_i)$ for all $i$,

$$\inf_{\gamma : \Pi \gamma = \mu_i (\forall i)} \int_{\mathcal{X} := \mathcal{X}_1 \times \ldots \times \mathcal{X}_n} c(x) \, d\gamma(x) = \int_{0}^{1} c\left(F_{\mu_1}^{-1}(t), \ldots, F_{\mu_n}^{-1}(t)\right) \, dt,$$
where $F_{\mu_i}(x_i) := \mu_i(y_i \in (-\infty, x_i])$, and $F_{\mu_i}^{-1}(t_i) := \inf\{x_i : F_{\mu_i}(x_i) \geq t_i\}$.

**Proof:** Given the probability space, $([0,1], \mathcal{B}, \lambda)$, where $\lambda$ is the Lebesgue measure restricted to $[0,1]$, let us define the monotone coupling by

$$
\gamma_{\text{mon}} := (F^{-1}_{\mu_1}, \ldots, F^{-1}_{\mu_n}) \# \lambda \in \mathcal{P}(\mathbb{R}^n),
$$

where $F^{-1}_{\mu_i}(t_i) := \inf\{y_i : F_{\mu_i}(y_i) \geq t_i\}, \forall t_i \in [0,1]$.

Notice then that by right-continuity, $F_{\mu_i}(F^{-1}_{\mu_i}(t_i)) \geq t_i$. Further, we find $F^{-1}_{\mu_i}(t_i) \leq x_i \iff t_i \leq F_{\mu_i}(x_i)$.

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

$$
(\forall y, y' \in \text{supp}(\gamma)) \quad y \leq y' \text{ OR } y \geq y',
$$

(EC.5) must be equal to $\gamma_{\text{mon}}$. To arrive at this conclusion, it suffices to show that any such $\gamma$ agrees with $\gamma_{\text{mon}}$ on the semiring of “downward-closed rectangles” $\mathcal{S}$, since, both being $\sigma$-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 $\gamma_{\text{mon}}$ on any downward-closed rectangle. More precisely, for any $x \in \mathbb{R}^n$,

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

For comparison, we now investigate the action of an arbitrary $\gamma \in \Gamma(\mu_1, \ldots, \mu_n)$, whose support satisfies the totally-ordered condition in (EC.5), on a downward-closed rectangle $\{y : y \leq x\}$, for some arbitrary $x \in \mathbb{R}^n$. Towards this, let us introduce the following definitions:

- $x^\downarrow := \{y : y \leq x\}$
- $x^\uparrow := \{y : y > x\}$

- Given any partition $P_x^\leq \cup P_x^\geq = \{1, \ldots, n\}$, where $P_x^\leq \neq \emptyset$ and $P_x^\geq \neq \emptyset$, write

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

And let $\mathcal{P}_x$ denote the family of all such sets $P_x$. 
Observe that $\mathbb{R}^n = \mathbb{x}^\uparrow \cup \mathbb{x}^\downarrow \cup \bigcup_{P_x \in \mathcal{P}_x} P_x$ (\(\cup\) represents disjoint union). But because of (EC.5), \(\bigcup_{P_x \in \mathcal{P}_x} P_x \notin \text{supp}(\gamma)\). Instead, there exists a subcollection \(\mathcal{P}_x^{\text{comp}} \subset \mathcal{P}_x\) that satisfies

\[
\left( \forall y, y' \in \bigcup_{P_x \in \mathcal{P}_x^{\text{comp}}} P_x \right) y \preceq y' \text{ OR } y \succeq y',
\]

and

\[
\gamma\left( \bigcup_{P_x \in \mathcal{P}_x^{\text{comp}}} P_x \right) = \gamma\left( \bigcup_{P_x \in \mathcal{P}_x^{\text{comp}}} P_x \right).
\]  

(EC.6)

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

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

\[
\gamma(\{y : y_i \leq x_i\}) + \gamma(\{y : y_i > x_i\}) = 1
\]

\[
\Longrightarrow \gamma(\{y : y_i \leq x_i\}) + \gamma(\mathbb{x}^\downarrow) + \gamma(\bigcup_{P_x \in \mathcal{P}_x^{\text{comp}}} P_x) = 1
\]

This tells us that

\[
\gamma(\{y : y_i \leq x_i\}) = \gamma(\{y : y_j \leq x_j\}) \quad \forall i, j \notin A
\]  

(EC.7)

Further, we observe that

\[
\gamma(\mathbb{x}^\downarrow) = \gamma(\{y : y_i \leq x_i, i \notin A\})
\]  

(EC.8)

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 $\bigcup_{P_x \in \mathcal{P}_x^{\text{comp}}} P_x$, and hence by (EC.6), it does not belong in the support of $\gamma$. By the same reasoning,

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

implies that

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

so conclude

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

(EC.9)
Finally, let $i \not\in A$ and $j \in A$ be arbitrary. Then
\[ \gamma(\{y: y_i \leq x_i\}) = \gamma(\{y: y_j \leq x_j\} \cap x_i^+) + \gamma(\{y: y_j \leq x_j\} \cap \bigcup_{P_x \in P_x} P_x) = \gamma(\{y: y_j \leq x_j\}) \]

(EC.10)

Altogether, (EC.9), (EC.10) and (EC.7) indicate that $\gamma(x_i) = \min_i F_{\mu_i}(x_i)$, as desired.

What remains is to show that any optimal coupling $\gamma^*$ must satisfy the totally-ordered condition of (EC.5). For this, we cite the following lemma:

**Lemma EC.2 (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 $\mu_i$. 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 $\gamma^*$, supp$(\gamma^*)$ necessarily satisfies the following: Given any $y = (y_1, \ldots, y_n)$ and $y' = (y'_1, \ldots, y'_n)$ in supp$(\gamma^*)$, then for any partition of $\{1, \ldots, n\}$, written $p = (p_+, p_-)$,
\[ c(y) + c(y') \leq c(z^p) + c(z'^p), \]
where $z_i^p = y_i$, $z_i' = y_i'$ if $i \in p_+$, and $z_i^p = y_i'$, $z_i'^p = y_i$ if $i \in p_-$. By this lemma,
\[ c(y) + c(y') \leq c(y \vee y') + c(y \wedge y') \forall y, y' \in \text{supp}(\gamma^*), \]
which yields a contradiction to the strict submodularity of $c$ if there exists a pair $y$ and $y'$ in the support of $\gamma^*$ that cannot be related by the canonical component-wise partial-order $\leq$. Hence, we conclude that any optimal coupling for a continuous, strictly submodular cost function must be equal to $\gamma_{\text{mon}}$.

What remains is a basic limiting argument. Indeed, by the monotone convergence theorem, letting $m$ tend to infinity,
\[ \int c + \frac{1}{m} \sum_{i \neq j} (x_i - x_j)^2 \, d\gamma_{\text{mon}} \leq \inf_\gamma \int c + \frac{1}{m} \sum_{i \neq j} (x_i - x_j)^2 \, d\gamma \]
\[ \implies \int c \, d\gamma_{\text{mon}} \leq \inf_\gamma \int c \, d\gamma, \]
as desired. \( \square \)
EC.4. Section 5 Proofs

EC.4.1. Proof of Proposition 4

**Proposition 4** For any $t \in \mathbb{R}^n$,

$$\max_{x \in \mathcal{X}_{\text{perm}}} \sum_{i=1}^{n} (\tilde{c}_i - t_i) \cdot x_i$$

is monotone, submodular in the $\tilde{c}$ variables. Consequently, $Z(t, \tilde{c})$ is monotone, submodular in the $\tilde{c}$ variables.

**Proof:** For the sake of notational simplicity we absorb $t$ into $c$, 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 $c_i - t_i$. Let $F : \mathbb{R}^n \to \mathbb{R}$ be defined by

$$F(c) := \max_{y \in \mathbb{R}^n} \sum_{i} c_i \sum_{j=1}^{n} j \cdot y_{ij}$$

subject to

$$\sum_{j} y_{ij} = 1, \quad i = 1, \ldots, n.$$

$$\sum_{i} y_{ij} = 1, \quad j = 1, \ldots, n$$

$$y \geq 0.$$

Then $F$ is submodular in $c$. To see this, consider a directed network constructed as follows. There are two special nodes $s$ and $t$, and two sets of nodes $S = \{s_1, \ldots, s_n\}$ and $T = \{t_1, \ldots, t_n\}$. Let the arc set be $\mathcal{A} := \{(s, s_i) : i = 1, \ldots, n) \cup \{(s_i, t_j) : \forall i \in [n], \forall j \in [n]\} \cup \{(t_j, t) : j = 1, \ldots, n\} \cup \{(t, t)\}$. Let all the arcs, except the arc $(t, s)$, have upper capacity set to 1. The arc $(t, s)$ has upper capacity set to $n$. Given the vector $c$, let any arc of the form $(s_i, t_j)$ have weight $c_i \cdot j$. All other arcs have zero weight. Hence, we can understand $F(c)$ as the value of the max-weight circulation in the network just described. Because all the arcs of the form $(s_i, t_j)$ are parallel, the value of the max-weight circulation is submodular in the weights $\{c_i \cdot j\}_{i,j}$—hence, $F$ is submodular in $c$.

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

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

as $[x \vee y]^\star = x^\star \vee y^\star$, $[x \wedge y]^\star = x^\star \wedge y^\star$. □
EC.5. MMM model for the Ranking with Scheduling problem

Define

\[ u_{ij} = \mathbb{P}(y_{ij}(\tilde{c}) = 1) \]
\[ w_{ij} = \mathbb{E}[\tilde{c}_i | y_{ij}(\tilde{c}) = 1] \mathbb{P}(y_{ij}(\tilde{c}) = 1) \]
\[ z_{ij} = \mathbb{E}[\tilde{c}_i^2 | y_{ij}(\tilde{c}) = 1] \mathbb{P}(y_{ij}(\tilde{c}) = 1) \]

Consider the following model

\[
\begin{align*}
\max & \sum_{i=1}^{n} \sum_{j=1}^{n} jw_{ij} - t_{ij}u_{ij} \\
\text{s.t.} & \sum_{j=1}^{n} u_{ij} = 1, \forall i = 1, \ldots, n \\
& \sum_{j=1}^{n} w_{ij} = \mu_i, \forall i = 1, \ldots, n \\
& \sum_{j=1}^{n} z_{ij} = \mu_i^2 + \sigma_i^2, \forall i = 1, \ldots, n \\
& \sum_{i=1}^{n} u_{ij} = 1, \forall j = 1, \ldots, n \\
\end{align*}
\]

(EC.11)

Denote the dual variable in (EC.11) as \( \alpha_i, \beta_i, \xi_i \), for \( i = 1, \ldots, n \) and \( \epsilon_j, \forall j = 1, \ldots, n \). Consider the dual of (EC.11).

\[
\begin{align*}
\min & \sum_{i=1}^{n} \left( \alpha_i + \beta_i \mu_i + \xi_i (\mu_i^2 + \sigma_i^2) \right) + \sum_{j=1}^{n} \epsilon_j \\
\text{s.t.} & \begin{pmatrix} \alpha_i + \epsilon_j + t_{ij} \frac{1}{2} (\beta_i - j) \\ \frac{1}{2} (\beta_i - j) \end{pmatrix} \geq 0, \forall i, j = 1, \ldots, n \\
\end{align*}
\]

(EC.12)

When solving the first stage problem, we can incorporate the constraints on \( t \) in (EC.12). Notice \( \begin{pmatrix} \alpha_i + \epsilon_j + t_{ij} \frac{1}{2} (\beta_i - j) \\ \frac{1}{2} (\beta_i - j) \end{pmatrix} \geq 0 \) can be equivalently written as a socp constraint as

\[
\sqrt{(\beta_i - j)^2 + (\alpha_i + \epsilon_j + t_{ij} - \xi_i)^2} \leq \alpha_i + \epsilon_j + t_{ij} + \xi_i
\]