Submitted to
manuscript #

Optimization for Maximizing the Expected Value of Order Statistics

David Bergman
Department of Operations and Information Management, University of Connecticut, 2100 Hillside Rd, Storrs, CT 06268
david.bergman@uconn.edu

Carlos Cardonha
IBM Research, Rua Tutoia 1157, Sao Paulo, SP. Brazil
carloscardonha@br.ibm.com

Jason Imbrogno
Department of Economics and Finance, University of North Alabama, 1 Harrison Plaza, Florence, AL 35632
jimbrogno@una.edu

Leonardo Lozano
Operations, Business Analytics & Information Systems, University of Cincinnati, 2925 Campus Green Drive,
Cincinnati, OH 45221
leolozano@uc.edu

This paper discusses a computational approach to optimization problems with objective criteria consisting of the expectation of order statistics. We propose a branch-and-cut algorithm relying on a series of functional approximations and discretization techniques which allows us to optimally solve problems consisting of two sums of potentially correlated and normally distributed random variables. Through an extensive experimental evaluation on a synthetic benchmark set we show the effectiveness of the algorithm in identifying high-quality solutions that are far superior to a basic heuristic; in particular, the results show that the superiority of the exact approach grows in high-variance settings. In order to evaluate the effectiveness on a real-world data set, we apply the methodology to choosing entries in daily fantasy football, and exhibit how the algorithm can be used to win thousands of dollars on sports-betting platform DraftKings.

Key words: order statistics; optimization; branch and check; fantasy sports

History: This paper was first submitted on X and has been with the authors for 0 years and 0 revisions.

1. Introduction

Decision making under uncertainty has been a major focus in the optimization literature (Dantzig 1955, Charnes and Cooper 1959, Birge and Louveaux 1997, Sahinidis 2004). Coping with uncertain constraints and/or objective function terms results in complex optimization problems that require advanced modeling and solution procedures such as robust optimization (Ben-Tal and Nemirovski 1997, 2002, Bertsimas and Sim 2004, 2006) and scenario-based optimization (Higle and Sen 1991, Birge and Louveaux 1988, Shapiro and Homem-de Mello 1998, Carøe and Tind 1998, Kleywegt et al. 2002, Sherali and Zhu 2006, Gade et al. 2014), among others.

In this paper we explore computational approaches to problems with objective criteria consisting of the expectations of order statistics. Given a set of m random variables $X_1, \dots, X_m$, the k^{th} order statistic $Y_{(k)}$ is the k^{th} lowest value assumed by these variables. We consider a class of stochastic binary optimization problems where we seek to maximize the expectation of $Y_{(k)}$ given that each of the random variables is a sum of random variables whose selection depends on the choice of binary variables, i.e., each component random variable X_i is *defined* by the selection in a vector of decisions variables denoted by x. To the best knowledge of the authors, such an objective function has not been studied in the optimization literature.

Order statistics play a critical role in a variety of settings, and there is a rich literature dedicated to their study. Some applications in which order statistics play a critical role include:

- **Auctions**: In auctions, the prices bidders pay are frequently determined by the first or second higher bidder (Brown and Brown 1986, Kaplan and Zamir 2012).
- **Insurance**: Insurance companies study order statistics (Ramachandran 1982, Dimitrova et al. 2018), for example in determining policies for joint-life insurance.
- **Systems**: It is common practice to create systems that are robust to components failing, for example in wireless communication networks (Yang and Alouini 2011).
- **Risk Management**: Risk management frequently requires the study of order statistics (Koutras and Koutras 2018)

- **Manufacturing**: The lifetime of an item can be modeled through order statistics (Ahmadi et al. 2015)

Due to the practical importance of order statistics and the need from practitioners to understand their underlying distributions, the statistics literature contains myriad papers dedicated to the subject (the interested reader can refer to these books (David and Nagaraja 2004, Arnold et al. 2008, Ahsanullah and Nevzorov 2005, Arnold and Balakrishnan 2012)). However, full distributional knowledge is limited to restrictive cases—in particular, this often requires the *independent and identically distributed* (iid) assumption. For example, if each component random variable is uniformly distributed and they are iid, then the order statistics follow beta distributions. Alternatively, if each component random variable is exponential and they are iid, then the minimum is also exponentially distributed, and the maximum follows the Gumbel distribution. In the presence of correlation between the random variables and/or differences in their distribution, limited distribution knowledge is known, in general. Due to the lack of exact distributional knowledge, some research has been published on bounds for the expected value order statistics for random variables that are correlated. Beyond this, little is known beyond bounds for general distributions (Bertsimas et al. 2006).

Perhaps due to problem complexity, even under the iid assumption, no papers have published exact approaches to problems with objective functions modeled as the optimization of order statistics on the expectation of random variables. In such a problem, the random variables themselves are functions of the decision variables, thereby resulting in a problem where the expectation of the order statistics is determined by the decision vector. Our paper proposes the first approach to solving such a problem exactly.

Due to the generality of such a problem, we focus on the case of two random variables, each of which is the sum of normal random variables that are arbitrarily correlated, with the sum selected through binary decision variables. In general, the exact distribution of order statistics for correlated random variables is complex and unknown, with papers dedicated to exploring bounds even just for their expectations (Ross 2010, D'Eramo et al. 2016). For the normal distribution specifically and for only two random variables, closed-form expressions for the expectation of both order statistics (the maximum and the minimum) are known (Nadarajah and Kotz (2008a)). The expression is nonlinear, in that it requires the evaluation of the p.d.f. and c.d.f. of a nonlinear function of the expectations, variables, and correlation.

When making decisions in order to optimize order statistics, we are therefore presented with a hard combinatorial optimization problem with a highly non-linear objective function, consisting of the square root of a quadratic function, that is the component of integral and rational equations. In order to solve the problem, we explore reformulations which allow us to compute tight relaxations that are enforced through cuts in a branch-and-cut framework.

The algorithm is applied to two problem sets. The first is a large synthetic benchmark set designed to evaluate how the proposed algorithm scales and how the solutions obtained favor against a natural heuristic approach. Our thorough experimental evaluation on this benchmark set indicates that the algorithm scales well and that the solutions obtained are far superior to what the heuristic identifies, especially in scenarios where the variances of the random variables are large.

Decision making under uncertainty is improved by learning parameters from real-world data. Therefore, we do not limit our approach to the domain of synthetic instances, but rather apply it to *daily fantasy sports* (DFS), where participants select entries of players in actual sporting events subject to roster and budget constraints. In order to choose optimal lineups in these DFS contests, one requires expected value estimates for each player in the game associated with a contest and estimates on the covariances of the player scores. There are myriad websites providing projected points for players, but lacking are refined estimates for the correlations between player scores. In this paper, we suggest two different methods for estimating player-by-player covariance of scores, and use all of those parameter estimates to optimize picks in selected contests.

We explore this application in Section 7, showing that our models can achieve selections in sports betting pools with skewed payout structures that could win thousands of dollars. The algorithm using our preferred correlation method would have resulted in $5,000 profit on less than a $10,000 initial investment. The literature on applying advanced optimization to sports betting applications has been the topic of several papers (Brown and Minor 2014, Clair and Letscher 2007, Bergman and Imbrogno 2017, Kaplan and Garstka 2001), and in particular some papers have arose that study daily fantasy sports (Hunter et al. 2016, Urbaczewski and Elmore 2018), but to the best of our knowledge, this is the first paper to study the *showdown* competition format available on the popular DFS betting platform `DraftKings`; showdown contests only include the players in a single NFL game, with entries consisting of 6 players.

We note here that our problem applied to DFS is related to the static stochastic knapsack problem, as items and their sizes are known *a priori* but their rewards are unknown. Literature on the problem is typically focused on scenarios where both rewards and sizes are stochastic (Dean et al. (2008)). In our problem, rewards are actually correlated; Ma (2018) explored an LP relaxation of the problem to derive an approximation algorithm for the problem. Note that the stochastic knapsack problem with correlated rewards and sizes given in binary can be shown to be PSPACE-hard (Dean et al. (2004)). To the best of our knowledge, settings of the stochastic knapsack problem with several knapsacks with correlations between rewards of items assigned to different containers has not been addressed in the literature yet.

In summary, our contributions are as follows:

- A first-ever exact algorithm for finding optimal decisions to problems with criteria specified by order statistics applied to expectations of random variables;
- An introduction of daily fantasy showdown contests to the literature;
- A nearest-neighbor algorithm for estimating covariance of points scored by different NFL players in DFS; and
- A thorough experimental evaluation which highlights the quality of solutions that can be obtained by our methodology in comparison with a basic heuristic, including a detailed report of results obtained on applying the methodology to real-world betting on the 2018 NFL season.

The remainder of the paper is organized as follows. In Section 2 we formally define the class of stochastic optimization problems we study. In Section 3 we describe complexity results. Section 4 provides details of a novel basic branch-and-cut algorithm. Section 5 presents various additional techniques to improve the basic method, thus leading to a significantly improved algorithm. Sections 6 and 7 detail a thorough experimental evaluation, both on synthetic instances and DFS showdown contests. Finally, we conclude in Section 8 with ideas for how this work can be expanded.

2. Problem Description

In this paper we study the problem

$$\max_{x\in\Omega\subseteq\{0,1\}^n} \mathbb{E}\left[Y_{(k)}(x)\right] \tag{P}$$

where $Y_{(k)}(x)$ is the kth order statistic among m random variables $X_1(x),\dots,X_m(x)$, each of which is functionally dependent on decision variables x which are assumed to be binary and restricted to a set Ω. $\mathbb{E}(\cdot)$ is the expected value, and so the objective is to maximize the expectation of the kth order statistic. The kth order statistic is the kth smallest value in a statistical sample among a collection of random variables. And so, for a fixed k, the problem seeks to make choices of x so that we maximize the expectation of the kth smallest of the random variables $X_1(x),\dots,X_m(x)$ over all possible samples. In particular, for $k=m$, we seek to maximize the expectation of the largest of the variables.

This class of optimization problems is challenging to solve in its full generality, and we study a specific case in this paper, where $k=m=2$. Namely, suppose there are p items $\mathcal{I}=\{1,\dots,p\}$, and each item i has a normally distributed *profit* $Z_i\sim\mathcal{N}(\mu_i,\sigma_i^2)$, with mean μ_i and variance σ_i^2. Two sets $\mathcal{S}^1,\mathcal{S}^2\subseteq\mathcal{I}$ are to be selected for two *bins*. The selection of sets must satisfy a collection of constraints $\mathcal{R}$ that can be specific for the individual bins, or based on joint conditions. We assume that the constraints are uniform from bin to bin. The *profit* of a bin $\mathcal{S}$ is a random variable defined as $z(\mathcal{S}):=\sum_{i\in\mathcal{S}} Z_i$. We study the optimization problem of selecting items satisfying $\mathcal{R}$ and maximizing the 2nd order statistic, i.e., the expected value of the bin achieving the maximum total profit; namely,

$$\max_{\mathcal{S}^1,\mathcal{S}^2\subseteq\mathcal{I}}\{\mathbb{E}\left(\max\{z(\mathcal{S}^1),z(\mathcal{S}^2)\}\right):\mathcal{S}^1,\mathcal{S}^2 \text{ satisfy constraints } \mathcal{R}\}.$$

We further assume that the random variables have arbitrary correlation between them. We use Σ to denote the associated covariance matrix, where $\rho_{i,j}$ is the correlation between Z_i and Z_j, so that the i,jth entry of Σ, the covariance of Z_i and Z_j, is $\text{cov}(Z_i,Z_j)=\rho_{i,j}\sigma_i\sigma_j$.

With these assumptions, we can now formulate our problem as follows. For $i=1,2$ and $j=1,\ldots,p$, let $x_{i,j}$ be a binary variable indicating the selection of item j for bin i, and $X(x)=\{X_1(x),X_2(x)\}$ be a set of random variables determined by

$$X_i(x)=\sum_{j=1}^{p} Z_j x_{i,j},\ \forall i\in\{1,2\},$$

that is, each $X_i(x)$ is a sum of random variables, where each binary variable $x_{i,j}$ indicates whether Z_j is selected and should be considered in the sum for set i.

Since each Z_j is normally distributed, $X_i(x)$ is normally distributed. Moreover, for any choice of $x\in\{0,1\}^{2\times p}$, one can calculate the mean and variance of $X_1(x)$ and $X_2(x)$, as well as their covariance (and correlation). In particular:

$$\mathbb{E}\left[X_i(x)\right]=\sum_{j=1}^{p}\mu_j x_{i,j} \qquad \forall i\in\{1,2\} \tag{1}$$

$$\sigma^2\left(X_i(x)\right)=\sum_{j=1}^{p}\sigma_j^2 x_{i,j}+2\sum_{1\le j_1<j_2\le p}\text{cov}(Z_{j_1},Z_{j_2})x_{i,j_1}x_{i,j_2} \qquad \forall i\in\{1,2\} \tag{2}$$

$$\text{cov}\left(X_1(x),X_2(x)\right)=\sum_{j_1=1}^{p}\sum_{j_2=1}^{p}\text{cov}\left(Z_{j_1},Z_{j_2}\right)x_{1,j_1}x_{2,j_2}. \tag{3}$$

We formulate our problem in the space of the x-variables as:

$$\max_{x\in\Omega}\mathbb{E}\left[Y_{(2)}(x)\right]=\max_{x\in\Omega}\mathbb{E}\left[\max\left\{X_1(x),X_2(x)\right\}\right], \tag{4}$$

where Ω is the subset of binary vectors which satisfy the constraints in $\mathcal{R}$. We assume for the remainder of the paper that the constraints in $\mathcal{R}$ are linear in the x-variables.

Nadarajah and Kotz (2008b) provide a closed form (non-analytical) expression for the maximum of two dependent normal random variables. Let $\phi(\cdot)$ and $\Phi(\cdot)$ be the probability density function (p.d.f.) and the cumulative distribution function (c.d.f.), respectively, of standard normal random variables; i.e., for $w \in (-\infty, \infty)$,

$$\begin{aligned} \phi(w) &= \frac{1}{\sqrt{2\pi}} e^{\frac{w^2}{2}} \\ \Phi(w) &= \int_{-\infty}^{w} \phi(u) du. \end{aligned}$$

To handle edge cases in evaluating the c.d.f. at rational expression $w = \frac{a}{b}$ with $b = 0$, we let

$$\Phi(w) = \begin{cases} 0, & a < 0 \\ \frac{1}{2}, & a = 0 \\ 1, & a > 0. \end{cases}$$

We can now write a expression for $\mathbb{E}\left(Y_{(2)}(x)\right)$ as:

$$\begin{aligned} \mathbb{E}\left[Y_{(2)}(x)\right] = \mathbb{E}\left[X_1(x)\right] \cdot \Phi\left(\frac{\mathbb{E}\left[X_1(x)\right] - \mathbb{E}\left[X_2(x)\right]}{\theta(x)}\right) + \\ \mathbb{E}\left[X_2(x)\right] \cdot \Phi\left(\frac{\mathbb{E}\left[X_2(x)\right] - \mathbb{E}\left[X_1(x)\right]}{\theta(x)}\right) + \\ \theta(x) \cdot \phi\left(\frac{\mathbb{E}\left[X_1(x)\right] - \mathbb{E}\left[X_2(x)\right]}{\theta(x)}\right), \end{aligned} \tag{5}$$

where

$$\theta(x) = \sqrt{\sigma^2(X_1(x)) + \sigma^2(X_2(x)) - 2\mathrm{cov}\left(X_1(x), X_2(x)\right)}. \tag{6}$$

As we can see from Equality 5, due to the (possible) dependency among the random variables $X_1(x)$ and $X_2(x)$, the expectation of their maximum is a highly nonlinear function.

The main methodological contribution of the paper is an exact mathematical programming algorithm for tackling this highly complex optimization problem. A motivating example is daily fantasy football, as described in the introduction, where the sets correspond to the entries that a participant can choose from, and the index set $[p]$ of Z_j variables are the players available for each entry. Linear constraints enforce that a viable entry is selected (e.g., based on number of players in each position and budget constraints).

In order to understand the general applicability of the algorithms developed and how they scale with problem size and varied problem characteristics, we investigate the application of the algorithms to synthetic scenarios of the problem where knapsack constraints limit the number of items selected for each bin. We experimentally evaluate cases when the items selected are allowed to be selected for both bins or for only one; i.e, with and without the constraints

$$x_{1,j} + x_{2,j} \leq 1, \quad j = 1, \ldots, p. \tag{7}$$

We note that even though our approach is tailored to the case in which $k = m = 2$, the proposed algorithms and results can be easily extended to the case in which $k = 1$, i.e., the objective seeks to maximize the minimum between $X_1(x)$ and $X_2(x)$.

3. Computational complexity

The following result shows that the variant of problem P we study is NP-hard.

Theorem 1. *P with $k = m = 2$ is NP hard.*

Proof. The result follows from a reduction from the minimum cut problem on a graph with positive and negative weights; the problem is known to be NP-hard in the general case (McCormick et al. (2003)). Let $G = (V, E)$ be an undirected graph with integer weights w_e of arbitrary sign on each edge $e \in E$. A (S,T)-*cut* of G is a 2-partition of V, and its *weight* $w(S,T)$ is the sum of the edges crossing the cut, i.e., $w(S,T) = \sum_{e: e\cap S, e\cap T \neq \emptyset} w_e$. The *minimum weight cut* is a cut of minimum weight. As a decision problem, the problem can be posed as follows: given a constant K, does there exist a (S,T)-cut of G such that $w(S,T) \leq K$.

We create an instance of P as follows. Assuming there are p vertices in G, every vertex $j \in V$ is associated with an item j with a normally distributed profit $Z_j \sim N(0,1)$, for $j = 1, \ldots, p$. The covariance between Z_{j_1} and Z_{j_2} is $\text{cov}(Z_{j_1}, Z_{j_2}) = \frac{w_{\{j_1,j_2\}}}{4M+1}$, where $M = \sum_{e \in E} |w_e|$. As a result, we have that $\sum_{j_1=1}^{p} \sum_{\substack{j_2=1 \\ j_2 \neq j_1}}^{p} \text{cov}(Z_{j_1}, Z_{j_2}) \leq \frac{2M}{4M+1} < \frac{1}{2}$ and $|\sigma_{j_1}^2| \geq \sum_{j_2 \neq j_1} |\text{cov}(Z_{j_1}, Z_{j_2})|$ for all $j_1 =$

$1,\dots,p$. One can then construct a symmetric and diagonally dominant (and, consequently, positive semi-definite) matrix Σ whose columns and rows are indexed by variables Z_j. It follows that Σ is a valid covariance matrix. Finally, let $\Omega = \{0,1\}^{2\times p}$, i.e., the set of feasible solutions is unconstrained.

By construction, $\mathbb{E}[X_1(x)] = \mathbb{E}[X_2(x)] = 0$, so the expression for $\mathbb{E}[Y_{(2)}(x)]$ reduces to:

$$\mathbb{E}[Y_{(2)}(x)] = \theta(x)\frac{1}{\sqrt{2\pi}}.$$

As $p \geq 1$ and all variances are equal to 1, it follows that at optimality $\theta(x) \geq 1$ (this value is achieved if we assign one item to the first set and leave the second set empty), so any x that maximizes $\theta(x)^2$ also maximizes $\theta(x)$. Therefore, the optimization of $\mathbb{E}[Y_{(2)}(x)]$ in our case is equivalent to

$$\max_{x\in\Omega}\ \theta^2(x) = \left(\sigma^2(X_1(x)) + \sigma^2(X_2(x)) - 2\mathrm{cov}(X_1(x) + X_2(x))\right).$$

By expanding the terms of the last expression, we obtain

$$\begin{aligned}\theta(x)^2 = &\sum_{j=1}^{p}\sigma^2(Z_j)x_{1,j} + 2\sum_{j_1=1}^{p-1}\sum_{j_2=j_1+1}^{p}\mathrm{cov}(Z_{j_1},Z_{j_2})x_{1,j_1}x_{1,j_2}\\ &+\sum_{j=1}^{p}\sigma^2(Z_j)x_{2,j} + 2\sum_{j_1=1}^{p-1}\sum_{j_2=j_1+1}^{p}\mathrm{cov}(Z_{j_1},Z_{j_2})x_{2,j_1}x_{2,j_2}\\ &-2\sum_{j_1=1}^{p}\sum_{j_2=1}^{p}\mathrm{cov}(Z_{j_1},Z_{j_2})x_{1,j_1}x_{2,j_2}.\end{aligned}$$

Because all variances are equal to 1, we have

$$\begin{aligned}\theta(x)^2 = &\sum_{j=1}^{p}x_{1,j} + 2\sum_{j_1=1}^{p-1}\sum_{j_2=j_1+1}^{p}\mathrm{cov}(Z_{j_1},Z_{j_2})x_{1,j_1}x_{1,j_2}\\ &+\sum_{j=1}^{p}x_{2,j} + 2\sum_{j_1=1}^{p-1}\sum_{j_2=j_1+1}^{p}\mathrm{cov}(Z_{j_1},Z_{j_2})x_{2,j_1}x_{2,j_2}\\ &-2\sum_{j=1}^{p}x_{1,j}x_{2,j} - 2\sum_{j_1=1}^{p}\sum_{\substack{j_2=1\\ j_2\neq j_1}}^{p}\mathrm{cov}(Z_{j_1},Z_{j_2})x_{1,j_1}x_{2,j_2},\end{aligned}\tag{8}$$

Claim 1 *In any optimal solution for* $\max_{x\in\Omega}\theta(x)^2$, $x_{1,j} + x_{2,j} = 1$ *for* $j = 1,\dots,p$.

Proof First, by reorganizing the terms in 8, we obtain

$$\begin{aligned}\theta(x)^2 = &\sum_{j=1}^{p} x_{1,j} + \sum_{j=1}^{p} x_{2,j} - 2\sum_{j=1}^{p} x_{1,j}x_{2,j}\\ &+2\left\{\sum_{j_1=1}^{p-1}\sum_{j_2=j_1+1}^{p}\operatorname{cov}(Z_{j_1},Z_{j_2})x_{1,j_1}x_{1,j_2} + \sum_{j_1=1}^{p-1}\sum_{j_2=j_1+1}^{p}\operatorname{cov}(Z_{j_1},Z_{j_2})x_{2,j_1}x_{2,j_2}\right.\\ &\qquad\left. - \sum_{j_1=1}^{p}\sum_{\substack{j_2=1\\ j_2\neq j_1}}^{p}\operatorname{cov}(Z_{j_1},Z_{j_2})x_{1,j_1}x_{2,j_2}\right\}.\end{aligned}$$

Let $A(x)$ denote the sum within the brackets ($\{\}$); $A(x)$ belongs to the interval defined by

$$\pm\sum_{j_1=1}^{p}\sum_{\substack{j_2=1\\ j_2\neq j_1}}^{p}\operatorname{cov}(Z_{j_1},Z_{j_2}),$$

because each covariance term appears at most twice with a positive coefficient and at most twice with a negative coefficient. By construction, $\sum_{j_1=1}^{p}\sum_{\substack{j_2=1\\ j_2\neq j_1}}^{p}\operatorname{cov}(Z_{j_1},Z_{j_2}) < \frac{1}{2}$, so we have that $-1 < 2A(x) < 1$. Therefore, any optimal solution to $\max_{x\in\Omega}\ \theta(x)^2$ also optimizes

$$\max_{x\in\Omega}\left\{\sum_{j=1}^{p} x_{1,j} + \sum_{j=1}^{p} x_{2,j} - 2\sum_{j=1}^{p} x_{1,j}x_{2,j}\right\},$$

since the absolute value of each of the coefficients in this expression is greater than or equal to 1. Finally, note that this expression is maximized if and only if $x_{1,j}+x_{2,j}=1$, as desired. □

It follows from Claim 1 that $\sum_{j=1}^{p}(x_{1,j}+x_{2,j}-2x_{1,j}x_{2,j}) = p$, a constant, so we have

$$\begin{aligned}\max_{x\in\Omega}\quad & f(x) = \theta(x)^2 = p + 2\left\{\sum_{j_1=1}^{p-1}\sum_{j_2=j_1+1}^{p}\operatorname{cov}(Z_{j_1},Z_{j_2})x_{1,j_1}x_{1,j_2}\right.\\ &\qquad + \sum_{j_1=1}^{p-1}\sum_{j_2=j_1+1}^{p}\operatorname{cov}(Z_{j_1},Z_{j_2})x_{2,j_1}x_{2,j_2}\\ &\qquad\left. - \sum_{j_1=1}^{p}\sum_{\substack{j_2=1\\ j_2\neq j_1}}^{p}\operatorname{cov}(Z_{j_1},Z_{j_2})x_{1,j_1}x_{2,j_2}\right\} \qquad (9)\\ \text{s.t.}\quad & x_{1,j}+x_{2,j} = 1 \qquad\qquad j=1,\ldots,p\\ & x_{i,j}\in\{0,1\} \qquad\qquad i=1,2,\ j=1,\ldots,p.\end{aligned}$$

Consider the following optimization problem:

$$\begin{aligned}
\min \quad & h(x) = \sum_{j_1=1}^{p} \sum_{\substack{j_2=1\\ j_2\neq j_1}}^{p} \operatorname{cov}(Z_{j_1}, Z_{j_2}) x_{1,j_1} x_{2,j_2} & \\
\text{s.t.} \quad & x_{1,j} + x_{2,j} = 1 & j = 1, \dots, p \\
& x_{i,j} \in \{0,1\} & i = 1,2,\ j = 1, \dots, p.
\end{aligned} \tag{10}$$

Claim 2 *An optimal solution to optimization problem (10) is also optimal to problem (9).*

Proof Both optimization problems have the same set of feasible solutions Ω'. Let x' and x'' be two feasible solutions with $h(x') < h(x'')$. Showing that $f(x') > f(x'')$ establishes the claim. To show this, we first note that for any feasible solution $\tilde{x}$,

$$\begin{aligned}
&\sum_{j_1=1}^{p-1} \sum_{j_2=j_1+1}^{p} \operatorname{cov}(Z_{j_1}, Z_{j_2}) \tilde{x}_{1,j_1} \tilde{x}_{1,j_2} + \sum_{j_1=1}^{p-1} \sum_{j_2=j_1+1}^{p} \operatorname{cov}(Z_{j_1}, Z_{j_2}) \tilde{x}_{2,j_1} \tilde{x}_{2,j_2} \\
&+ \sum_{j_1=1}^{p} \sum_{j_2\neq j_1}^{p} \operatorname{cov}(Z_{j_1}, Z_{j_2}) \tilde{x}_{1,j_1} \tilde{x}_{2,j_2} = \sum_{j_1=1}^{p-1} \sum_{j_2=j_1+1}^{p} \operatorname{cov}(Z_{j_1}, Z_{j_2}).
\end{aligned} \tag{11}$$

This follows because for any two indices $j_1 \neq j_2$, the covariance term $\operatorname{cov}(Z_{j_1}, Z_{j_2})$ is counted in exactly one of the three terms in the left-hand size of equation (11):

1. If $\tilde{x}_{1,j_1} = \tilde{x}_{1,j_2} = 1 \rightarrow \operatorname{cov}(Z_{j_1}, Z_{j_2})$ is counted only in the first term;
2. If $\tilde{x}_{2,j_1} = \tilde{x}_{2,j_2} = 1 \rightarrow \operatorname{cov}(Z_{j_1}, Z_{j_2})$ is counted only in the second term;
3. If $\tilde{x}_{1,j_1} = 1, \tilde{x}_{2,j_2} = 1 \rightarrow \operatorname{cov}(Z_{j_1}, Z_{j_2})$ is counted only in the third term; and
4. If $\tilde{x}_{2,j_1} = 1, \tilde{x}_{1,j_2} = 1 \rightarrow \operatorname{cov}(Z_{j_1}, Z_{j_2})$ is counted only in the third term.

Finally, because $\tilde{x}_{1,j_1} + \tilde{x}_{2,j_1} = 1$ and $\tilde{x}_{1,j_2} + \tilde{x}_{2,j_2} = 1$, it follows that the list above is exhaustive and contains all possible assignments of j_1 and j_2. Therefore,

$$\begin{gathered}
h(x') < h(x'') \implies \\
-2h(x') > -2h(x'') \implies \\
\sum_{j_1=1}^{p-1} \sum_{j_2=j+1+1}^{p} \operatorname{cov}(Z_{j_1}, Z_{j_2}) - 2h(x') > \sum_{j_1=1}^{p-1} \sum_{j_2=j+1+1}^{p} \operatorname{cov}(Z_{j_1}, Z_{j_2}) - 2h(x'') \implies \\
f(x') > f(x''),
\end{gathered}$$

as desired. $\square$

We conclude that an optimal solution to (10) is also optimal to the original problem

$$\max_{x\in\Omega}\{\mathbb{E}[Y_{(2)}(x)]\}.$$

There is a one-to-one mapping between solutions to (10) and (S,T)-cuts of G. Namely, for a feasible solution x' we associate it with the $(S(x'),T(x'))$-cut defined by $x'_{1,j}=1\leftrightarrow j\in S(x')$ and $x'_{2,j}=1\leftrightarrow j\in T(x')$. Additionally, $w(S(x'),T(x'))\leq K$ if and only if $h(x')\leq \frac{K}{4M+1}$. This follows because

$$\begin{aligned} w(S(x'),T(x')) &= \sum_{j_1\in S(x')}\sum_{j_2\in T(x;)} w_{\{j_1,j_2\}} \\ &= \sum_{j_1\in S(x')}\sum_{j_2\in T(x')} (4M+1)\operatorname{cov}(Z_{j_1},Z_{j_2}) \\ &= (4M+1)\sum_{j_1=1}^{p}\sum_{j+2=1,j_2\neq j_1}^{p} \operatorname{cov}(Z_{j_1},Z_{j_2}) \\ &= (4M+1)\,h(x'). \end{aligned}$$

Therefore, if one can solve (10), one can also solve the family of instances of (P) presented in our construction, thus giving an algorithm that decides exactly the minimum cut problem with positive and negative edge weights. □

4. A Cutting-Plane Algorithm

Expression (5) is highly nonlinear, thus making the application of direct formulations combined with off-the-shelf solvers unlikely to be satisfactory for (P) and its extensions. In this work, we present an exact cutting-plane algorithm to solve problem (P) when $k=m=2$. The main component of our approach is a mixed-integer linear program (MIP), which we refer to as the *relaxed master problem* (RMP), that provides upper bounds on the optimal objective value and is iteratively updated through the inclusion of cuts. Lower bounds are obtained through the evaluation of $\mathbb{E}\left(Y_{(2)}(x)\right)$ on the feasible solutions obtained from the optimization of the RMP. Section 4.1 describes our proposed cutting-plane algorithm, and Section 4.2 presents a basic formulation of the RMP, based on a standard McCormick linearization technique.

4.1. Algorithm Description

Our approach to solve the problem is presented in Algorithm 1. A key component of our algorithm is the construction of a linear upper-bounding function for $\mathbb{E}\left(Y_{(2)}(x)\right)$ defined over the set Ω of feasible solutions. Namely, we wish to work with a function $g(x)$ such that

$$g(x) \geq \mathbb{E}\left[Y_{(2)}(x)\right] \quad \forall x \in \Omega.$$

Given $g(x)$, the relaxed master problem can be stated as

$$\bar{z} = \max_{x \in \Omega} g(x), \tag{RMP}$$

where $\Omega(\mathcal{C})$ represents the restriction of Ω to solutions that also satisfy a set of constraints $\mathcal{C}$. Note that $\bar{z}$ provides an upper bound on the optimal objective value of problem (P).

Algorithm 1 solves problem RMP iteratively, adding *no-good constraints* (or simply *cuts*) to a set $\mathcal{C}$, which are incorporated to RMP in order to prune solutions previously explored (Balas and Jeroslow 1972). RMP($\mathcal{C}$) denotes RMP enriched by the inclusion of (all) cuts in $\mathcal{C}$, so that a solution x for RMP($\mathcal{C}$) belongs to Ω and satisfies all constraints in $\mathcal{C}$.

Our cutting-plane algorithm keeps a lower bound *LB* and an upper bound *UB* for expression (5), which are non-decreasing and non-increasing over time, respectively. Both values are computed (and iteratively updated) based on each incumbent solution $\hat{x}$ of RMP; upper bounds are given by $g(\hat{x})$ and lower bounds follow from the exact evaluation of $\mathbb{E}\left[Y_{(2)}(\hat{x})\right]$. Algorithm 1 terminates if *LB* becomes equal to *UB* or if RMP($\mathcal{C}$) becomes infeasible. Because Ω is finite set, each cut incorporated to $\mathcal{C}$ removes at least one solution from Ω, and $|\mathcal{C}|$ increases in each iteration, it follows that Algorithm 1 terminates with an optimal solution in a finite number of iterations.

Algorithm 1 A Cutting-Plane Algorithm

```
1: Set $LB = -\infty$, $UB = \infty$, $\mathcal{C} = \emptyset$, and incumbent solution $\bar{x} = 0$
2: Solve RMP($\mathcal{C}$). If the problem is infeasible, then go to Step 5. Otherwise, let $UB$ be the optimal objective function value obtained for this problem, and record the optimal solution $\hat{x}$ found.
3: Compute $\mathbb{E}\left[Y_{(2)}(\hat{x})\right]$ using Equation (5). If $\mathbb{E}\left[Y_{(2)}(\hat{x})\right] > LB$, set $LB = \mathbb{E}\left[Y_{(2)}(\hat{x})\right]$ and update incumbent $\bar{x} = \hat{x}$.
4: If $LB = UB$, go to Step 5. Otherwise, add a no-good constraint to $\mathcal{C}$ which is violated solely by $\hat{x}$ among all solutions in $\Omega$. Return to Step 2.
5: If $LB = -\infty$, then terminate and conclude that the original problem is infeasible. Otherwise, terminate with an optimal solution given by $\bar{x}$.
```

4.2. A Standard Approach to Obtain Upper Bounds

A challenging task involved on the optimization of (P) is the evaluation of $\theta(x)$, a nonlinear expression that appears in all terms of expression (5), including the denominators of the c.d.f. and the p.d.f. of the normal distribution. In the following proposition, we show how to derive an upper-bounding function for (5) which avoids technical issues involved in the evaluation of $\theta(x)$.

In order to simplify notation, we assume w.l.o.g. throughout the manuscript that $\mathbb{E}[X_1(x)] \geq \mathbb{E}[X_2(x)]$; we also use $\delta(x) = \mathbb{E}[X_1(x)] - \mathbb{E}[X_2(x)]$ to simplify a recurring expression in the text.

Proposition 1. *For every $x \in \Omega$,*

$$\mathbb{E}\left[Y_{(2)}(x)\right] \leq \mathbb{E}[X_1(x)] + \frac{1}{\sqrt{2\pi}}\left(1 + \theta(x)^2\right). \tag{12}$$

Proof From the symmetry around 0 of the c.d.f. of the standard normal distribution, we have

$$\Phi(a) + \Phi(-a) = P[X \leq a] + P[X \leq -a] = P[X \leq a] + 1 - P[X \geq -a] = 1$$

for any $a \in \mathbb{R}$. Therefore, it follows that

$$\Phi\left(\frac{\delta(x)}{\theta(x)}\right) + \Phi\left(\frac{-\delta(x)}{\theta(x)}\right) = 1.$$

As Φ is a c.d.f. and, therefore, non-negative, and as $\mathbb{E}\left[X_1(x)\right] \geq \mathbb{E}\left[X_2(x)\right]$ by assumption, we have

$$\mathbb{E}\left[X_1(x)\right] \geq \Phi\left(\frac{\delta(x)}{\theta(x)}\right)\mathbb{E}\left[X_1(x)\right] + \Phi\left(\frac{-\delta(x)}{\theta(x)}\right)\mathbb{E}\left[X_2(x)\right], \tag{13}$$

because the multipliers form a convex combination, so we have an upper bounding expression for the first two terms of $\mathbb{E}\left[Y_{(2)}(\hat{x})\right]$.

In order to find an upper bound for $\theta(x)\cdot\phi\left(\frac{\delta(x)}{\theta(x)}\right)$, first note that $\phi\left(\frac{\delta(x)}{\theta(x)}\right) \leq \phi(0) = \frac{1}{\sqrt{2\pi}}$, because the p.d.f. of the standard normal random variable is maximized at its mean 0. Additionally, because $\theta(x) \geq 0$, $\theta(x) \leq 1+\theta(x)^2$ for every x, so we have

$$\frac{1}{\sqrt{2\pi}}\left(1+\theta(x)^2\right) \geq \theta(x)\cdot\phi\left(\frac{\delta(x)}{\theta(x)}\right). \tag{14}$$

Finally, Inequality (12) follows from the addition of Inequality (13) with Inequality (14):

$$\begin{aligned}\mathbb{E}\left[Y_{(2)}(x)\right] &= \mathbb{E}[X_1(x)]\cdot\Phi\left(\frac{\delta(x)}{\theta(x)}\right) + \mathbb{E}[X_2(x)]\cdot\Phi\left(\frac{-\delta(x)}{\theta(x)}\right) + \theta(x)\cdot\phi\left(\frac{\delta(x)}{\theta(x)}\right)\\ &\leq \mathbb{E}[X_1(x)] + \frac{1}{\sqrt{2\pi}}\left(1+\theta(x)^2\right). \quad \square\end{aligned}$$

By using the right-hand side expression of Inequality (12) as the objective function of RMP, one needs to evaluate $\theta(x)^2$ rather than $\theta(x)$, thus avoiding the square root operator. $\theta(x)^2$ is given by

$$\theta(x)^2 = \sigma^2(X_1(x)) + \sigma^2(X_2(x)) - 2\mathrm{cov}\left(X_1(x), X_2(x)\right). \tag{15}$$

A direct formulation of this expression, based solely on variables x, would have quadratic terms. However, this issue can be directly (and exactly) addressed via a McCormick linearization technique (McCormick 1976), in which linear constraints and binary variables are incorporated to the formulation in order to replace quadratic terms.

We now present a mixed-integer linear programming formulation of RMP, based on the results of Proposition 1 and on the McCormick linearization technique.

$$\text{RMP:} \quad \max \quad u_1 + \frac{1}{\sqrt{2\pi}}(1+s) \tag{16a}$$

$$\text{s.t.} \qquad u_1 = \sum_{j=1}^{p}\mu_j x_{1,j};\;\; u_2 = \sum_{j=1}^{p}\mu_j x_{2,j};\;\; u_1 \geq u_2 \tag{16b}$$

$$s = \sum_{i=1}^{2} \left(\sum_{j=1}^{p} \sigma_j^2 x_{i,j} + 2 \sum_{1 \leq j_1 < j_2 \leq p} \operatorname{cov}(Z_{j_1}, Z_{j_2}) v_{i,j_1,j_2} \right) - 2 \sum_{j_1=1}^{p} \sum_{j_2=1}^{p} \operatorname{cov}(Z_{j_1}, Z_{j_2}) r_{j_1,j_2} \tag{16c}$$

$$v_{i,j_1,j_2} \leq x_{i,j_1};\ v_{i,j_1,j_2} \leq x_{i,j_2} \qquad \forall j_1, j_2 \in \{1, \ldots, p\},\ i \in \{1,2\} \tag{16d}$$

$$v_{i,j_1,j_2} \geq x_{i,j_1} + x_{i,j_2} - 1 \qquad \forall j_1, j_2 \in \{1, \ldots, p\},\ i \in \{1,2\} \tag{16e}$$

$$r_{j_1,j_2} \leq x_{1,j_1};\ r_{j_1,j_2} \leq x_{2,j_2} \qquad \forall j_1, j_2 \in \{1, \ldots, p\} \tag{16f}$$

$$r_{j_1,j_2} \geq x_{1,j_1} + x_{2,j_2} - 1 \qquad \forall j_1, j_2 \in \{1, \ldots, p\} \tag{16g}$$

$$v_{i,j_1,j_2} \in \{0,1\} \qquad \forall j_1, j_2 \in \{1, \ldots, p\},\ i \in \{1,2\} \tag{16h}$$

$$r_{j_1,j_2} \in \{0,1\} \qquad \forall j_1, j_2 \in \{1, \ldots, p\} \tag{16i}$$

$$x \in \Omega. \tag{16j}$$

Variables u_1 and u_2 denote $\mathbb{E}[X_1(x)]$ and $\mathbb{E}[X_2(x)]$, respectively, and s represents $\theta(x)^2$. Binary variable v_{i,j_1,j_2} takes a value of 1 iff $x_{i,j_1} = x_{i,j_2} = 1$, i.e., items j_1, j_2 are selected for set i. Similarly, r_{j_1,j_2} is a binary variable that equals 1 iff $x_{1,j_1} = x_{2,j_2} = 1$, i.e., items j_1, j_2 are selected for sets 1 and 2, respectively. The objective function (16a) maximizes $g(x)$ as defined in Proposition 1. Constraints (16b) define the u-variables according to equation (1) and impose the symmetry breaking condition $u_1 \geq u_2$, so that $u_1 = \max(\mathbb{E}[X_1(x)], \mathbb{E}[X_2(x)])$. Constraint (16c) imposes $s = \theta(x)^2$ as described by equation (15), where $\sigma^2(X_1(x))$, $\sigma^2(X_2(x))$, and $\operatorname{cov}(X_1(x), X_2(x))$ are computed according to equations (2) and (3), respectively. Constraints (16d)–(16g) are the McCormick linearization constraints. Constraints (16h)–(16i) require v-variables and r-variables to be binary. Constraint (16j) requires the x-variables to be feasible to the original problem. We later refer to Ψ as the space defined by constraints (16b)–(16j).

5. Strengthened Formulation

We describe next SRMP, a strengthened version of the restricted master problem formulation presented previously. SRMP employs a more accurate bounding function for $\mathbb{E}\left(Y_{(2)}(x)\right)$, based on a joint discretization of $\theta(x)^2$ and $\delta(x)$. The estimated values are coupled via supervalid inequalities, thus further enhancing the computational performance of SRMP.

5.1. Framework

SRMP relies on the following proposition, which is valid for any functions $u_\theta(x)$, $l_\theta(x)$, and $u_\delta(x)$ such that $0 \le l_\theta(x) \le \theta(x) \le u_\theta(x)$ and $u_\delta(x) \ge \delta(x)$:

Proposition 2. *For every $x \in \Omega$,*

$$\mathbb{E}\left[X_1(x)\right] \cdot \Phi\left(\frac{u_\delta(x)}{l_\theta(x)}\right) + \mathbb{E}\left[X_2(x)\right]\left(1 - \Phi\left(\frac{u_\delta(x)}{l_\theta(x)}\right)\right) + \frac{1}{\sqrt{2\pi}} u_\theta(x) \ge \mathbb{E}\left[Y_{(2)}(x)\right]. \tag{17}$$

Proof This follows because

$$\begin{aligned}\mathbb{E}[Y_{(2)}(x)] &\le \mathbb{E}\left[X_1(x)\right] \cdot \Phi\left(\frac{\delta(x)}{\theta(x)}\right) + \mathbb{E}\left[X_2(x)\right] \cdot \Phi\left(\frac{-\delta(x)}{\theta(x)}\right) + \frac{1}{\sqrt{2\pi}}\theta(x) \\ &\le \mathbb{E}\left[X_1(x)\right] \cdot \Phi\left(\frac{\delta(x)}{\theta(x)}\right) + \mathbb{E}\left[X_2(x)\right]\left(1 - \Phi\left(\frac{\delta(x)}{\theta(x)}\right)\right) + \frac{1}{\sqrt{2\pi}} u_\theta(x) \\ &\le \mathbb{E}\left[X_1(x)\right] \cdot \Phi\left(\frac{u_\delta(x)}{l_\theta(x)}\right) + \mathbb{E}\left[X_2(x)\right]\left(1 - \Phi\left(\frac{u_\delta(x)}{l_\theta(x)}\right)\right) + \frac{1}{\sqrt{2\pi}} u_\theta(x).\end{aligned}$$

The first inequality follows because the p.d.f. of a standard normal is bounded by $\frac{1}{\sqrt{2\pi}}$. The second inequality follows because $u_\theta(x) \ge \theta(x)$ and because a standard normal is symmetric around 0. With respect to the last inequality, first note that, for any constants a, b with $a \ge b$, $a\lambda_1 + b(1-\lambda_1) \le a\lambda_2 + b(1-\lambda_2)$, for any values $0 \le \lambda_1 \le \lambda_2 \le 1$; intuitively, this holds because the largest term gains a larger weight on a convex combination when λ_1 is replaced for λ_2. This implies the second inequality, since the c.d.f. is non-decreasing on its domain and $\frac{\delta(x)}{\theta(x)} \le \frac{u_\delta(x)}{l_\theta(x)}$. □

Suppose we are given d continuous (and not necessarily disjoint) intervals $\left\{[\theta_q^2, \theta_{q+1}^2]\right\}_{q=1}^d$ for which $\theta(x)^2 \in [\theta_1^2, \theta_{d+1}^2]$ and l continuous (and also not necessarily disjoint) intervals $\{[\delta_k, \delta_{k+1}]\}_{k=1}^l$ for which $\delta(x) \in [\delta_1, \delta_{l+1}]$, for any $x \in \Omega$. Furthermore, suppose that for $q = 1, \ldots, d$, $\overline{\theta}_q$ and $\underline{\theta}_q$ are upper and lower bounds, respectively, of $\theta(x)$ for $\theta(x)^2 \in [\theta_q^2, \theta_{q+1}^2]$. Using these intervals, we can construct the following SRMP, where binary variables w_q, $q = 1, \ldots, d$, and y_k, $k = 1, \ldots, l$, indicate which interval $\theta(x)^2$ and $\delta(x)$ belong to, respectively:

$$\max \quad u' + \frac{1}{\sqrt{2\pi}} s' \tag{18}$$

$$\text{s.t.} \quad \sum_{q=1}^{d} w_q = 1 \tag{19}$$

$$\theta_q^2 w_q \le s \le \theta_{q+1}^2 + \theta_{d+1}^2(1-w_q) \qquad q=1,\dots,d \tag{20}$$

$$s' = \sum_{q=1}^{d} \overline{\theta}_q w_q \tag{21}$$

$$\sum_{k=1}^{l} y_k = 1 \tag{22}$$

$$\delta_k y_k \le u_1 - u_2 \le \delta_{k+1} + \delta_{l+1}(1-y_k) \qquad k=1,\dots,l \tag{23}$$

$$u' \le u_1 \cdot \Phi\left(\frac{\delta_{k+1}}{\underline{\theta}_q}\right) + u_2\left(1-\Phi\left(\frac{\delta_{k+1}}{\underline{\theta}_q}\right)\right) + M\left(2-w_q-y_k\right) \quad q=1,\dots,d,\ k=1,\dots,l \tag{24}$$

$$(x,v,r,u_1,u_2,s) \in \Psi \tag{25}$$

$$w_q \in \{0,1\},\ y_k \in \{0,1\} \quad q=1,\dots,d,\ k=1,\dots,l. \tag{26}$$

Constraints (19) and (22) ensure that one interval is chosen for $\theta(x)^2$ and $\delta(x)$, respectively. Constraint (25) ensures that s equates to $\theta(x)^2$, and so Constraint (20) relates the w_q variables with s. Constraints (21) then sets s' equal to the upper bound of $\theta(x)$ for the interval that $\theta(x)^2$ belongs to; note that s' is unrestricted in the model. Constraints (23) select the right interval for $\delta(x)$, noting that u_1 and u_2 are defined by the constraints in Ψ. Constraints (24) are only active for the selected intervals and enforce u' to be bounded by a linear combination of u_1 and u_2, defined by the evaluation of the c.d.f. at appropriately chosen constants associated with the intervals that $\theta(x)^2$ and $\delta(x)$ lie in; M is a sufficiently large value. Constraints (26) defines the domains of the variables appropriately. Finally, note that the objective function is also directly affected by the discretization, as its terms are pre-computed for the given set of discretization intervals.

We describe next the strategies to define the discretization intervals and compute the corresponding upper and lower bounds on $\theta(x)$ and $\delta(x)$.

5.2. Discretization of $\theta(x)^2$

A trivial upper bound for $\theta(x)^2$ is given by $4\max_{x\in\Omega}\sigma^2(X_1(x))$. This is because, for any $x \in \Omega$,

$$\sigma^2(X_1(x)) \le \max_{x\in\Omega}\sigma^2(X_1(x));$$

$$\sigma^2(X_2(x)) \le \max_{x\in\Omega}\sigma^2(X_1(x)); \text{ and}$$

$$\operatorname{cov}(X_1(x),X_2(x)) = \rho(X_1(x),X_2(x))\sigma(X_1(x))\sigma(X_2(x)) \quad \ge (-1)\max_{x\in\Omega}\sigma^2(X_1(x))$$

Note that the simple upper bound of $\theta_d^2 = 4\max_{x\in\Omega}\sigma^2(X_1(x))$ can be tightened by solving $\max\{s \mid (x,v,r,u_1,u_2,s)\in\Psi\}$; from Theorem 1, though, it follows that this problem is NP-hard. Additionally, our computational evaluations show that the resulting MIP is indeed challenging, so our strategy consists of solving the MIP for a limited amount of time in order to obtain a relatively refined upper bound $\hat{\theta}^2$, which is used in our discretization strategy.

We define d intervals for $\theta(x)^2$ as follows: $\theta_1^2 = 0$, $\theta_2^2 = 1$, and $\theta_q^2 = \theta_{q-1}^2 + \frac{\hat{\theta}^2-1}{d-1}$ for $q = 3,\dots,d+1$. Note that for $x\in\Omega$, $0 \le \theta(x)^2 \le \theta_{d+1}^2 = \hat{\theta}^2$ and $\theta_q^2 \le \theta(x)^2 \le \theta_{q+1}^2$ for some $q = 1,...,d+1$, so this discretization is valid. Given these intervals, upper bounds for $\theta(x)$ are given by

$$\overline{\theta}_q = \begin{cases} 1, & q = 1 \\ \\ \sqrt{\theta_{q+1}^2}, & q = 2,\dots,d, \end{cases}$$

whereas lower bounds are given by

$$\underline{\theta}_q = \begin{cases} 0, & q = 1 \\ \\ \sqrt{\theta_q^2}, & q = 2,\dots,d. \end{cases}$$

5.3. Discretization of $\delta(x)$

Our algorithm relies on an upper bound $\hat{\delta}$ for $\delta(x) = \mathbb{E}[X_1(x)] - \mathbb{E}[X_2(x)]$ for $x\in\Omega$. A trivial bound is given by $\hat{\delta} = 2\sum_{j=1}^p |\mu_j|$; similarly to $\theta(x)^2$, though, tighter bounds may be obtained from feasible upper bounds for the problem $\max\{u_1 - u_2 \mid (x,v,r,u_1,u_2,s)\in\Psi\}$.

Equipped with $\hat{\delta}$, the discretization for $\delta(x)$ is simpler than that for $\theta(x)^2$. Namely, we generate l discretization intervals defined by $\delta_k = \frac{k-1}{l}\hat{\delta}$ for $k = 1,\dots l+1$. The intervals span the possible values for $\delta(x)$, and so this is a valid discretization. Finally, upper bounds for $\delta(x)$ are obtained directly from the upper limits of the intervals.

5.4. Strengthening Inequalities for the RMP

We show next how estimates on $\delta(x)$ and $\theta(x)$ can be coupled via supervalid inequalities (SVIs), which potentially eliminate integer solutions without removing all the optimal ones (Israeli and Wood 2002). We propose SVIs of the form

$$\delta(x)\in[\delta_k,\delta_{k+1}] \implies \theta(x) \ge \theta_{min}^k \quad \forall k = 1,\dots,l, \tag{27}$$

where θ^k_{min} establishes a lower bound on $\theta(x)$ assuming that $\delta(x)$ belongs to the interval $[\delta_k, \delta_{k+1}]$. We leverage current knowledge about the problem by considering the satisfaction of a global lower bound z^{LB} for (5) when computing values for θ^k_{min}. We obtain lower bound z^{LB} using a simple heuristic procedure described in the computational experiments section.

Propositions 3 and 4 establish the results needed to obtain valid values for θ^k_{min}. Proposition 3 shows how to obtain a lower bound on $\theta(x)$ assuming that $\delta(x)$ is equal to some constant δ. Proposition 4 shows how to extend this result to the case in which $\delta(x)$ belongs to a given interval. Both propositions use value $\bar{\mu} = \max_{x\in\Omega} \mathbb{E}[X_1(x)]$, which typically can be quickly computed.

Proposition 3. *For a given value δ and lower bound z^{LB}, let $\Omega\left(\delta, z^{LB}\right)$ be the set of solutions x such that $\mathbb{E}\left[Y_{(2)}(x)\right] \geq z^{LB}$ and $\delta(x) = \delta$. A lower bound $\theta_{min}(\delta)$ for $\min_{x\in\Omega\left(\delta, z^{LB}\right)} \theta(x)$ can be obtained through the solution of the following optimization problem:*

$$LBP(\delta): \quad \theta_{min}(\delta, z^{LB}) = \min_{\theta\geq 0}\left\{\theta \mid \bar{u} + \theta\cdot\phi\left(\frac{\delta}{\theta}\right) \geq z^{LB}\right\}. \tag{28}$$

Proof For any $x \in \Omega_\delta$, we have that

$$\mathbb{E}\left[X_1(x)\right]\cdot\Phi\left(\frac{\delta}{\theta(x)}\right) + \mathbb{E}\left[X_2(x)\right]\cdot\Phi\left(\frac{-\delta}{\theta(x)}\right) + \theta(x)\cdot\phi\left(\frac{\delta}{\theta(x)}\right) \geq z^{LB}.$$

As $\bar{u} \geq \max_{x\in\Omega} \mathbb{E}[X_1(x)]$, we have

$$\mathbb{E}\left[X_1(x)\right]\cdot\Phi\left(\frac{\delta}{\theta(x)}\right) + \mathbb{E}\left[X_2(x)\right]\cdot\Phi\left(\frac{-\delta}{\theta(x)}\right) \leq \bar{u}.$$

As a result,

$$\bar{u} + \theta(x)\cdot\phi\left(\frac{\delta}{\theta(x)}\right) \geq \mathbb{E}\left[X_1(x)\right]\cdot\Phi\left(\frac{\delta}{\theta(x)}\right) + \mathbb{E}\left[X_2(x)\right]\cdot\Phi\left(\frac{-\delta}{\theta(x)}\right) \geq z^{LB}$$

for every $x \in \Omega\left(\delta, z^{LB}\right)$. Therefore, the value of θ obtained from the optimization $LBP(\delta)$ is at least as small as the value attained by any $x \in \Omega\left(\delta, z^{LB}\right)$, so the result follows. □

Proposition 4. *Given two values δ_1 and δ_2 such that $\delta_1 \leq \delta_2$ and a fixed value z^{LB}, $\theta_{min}(\delta_1, z^{LB}) \leq \theta_{min}(\delta_2, z^{LB})$.*

Proof For $\theta \geq 0$ and fixed δ, both θ and $\phi\left(\frac{\delta}{\theta}\right)$ are continuous and non-decreasing functions of θ, so we have that $\theta \cdot \phi\left(\frac{\delta}{\theta}\right)$ is also a continuous and non-decreasing function of θ. Therefore, at optimality we have $\theta_{min}\left(\delta_1, z^{LB}\right) \cdot \phi\left(\frac{\delta_1}{\theta_{min}\left(\delta_1, z^{LB}\right)}\right) = \theta_{min}(\delta_2, z^{LB}) \cdot \phi\left(\frac{\delta_2, z^{LB}}{\theta_{min}\left(\delta_2, z^{LB}\right)}\right) = z^{LB} - \bar{u}$. Assume by contradiction that $\theta_{min}(\delta_2, z^{LB}) < \theta_{min}\left(\delta_1, , z^{LB}\right)$; in this case, $\phi\left(\frac{\delta_2, z^{LB}}{\theta_{min}\left(\delta_2, , z^{LB}\right)}\right) < \phi\left(\frac{\delta_1, z^{LB}}{\theta_{min}\left(\delta_1, , z^{LB}\right)}\right)$ and, consequently we have that $\theta_{min}(\delta_2, z^{LB}) \cdot \phi\left(\frac{\delta_2}{\theta_{min}\left(\delta_2, z^{LB}\right)}\right) < \theta_{min}\left(\delta_1, z^{LB}\right) \cdot \phi\left(\frac{\delta_1, z^{LB}}{\theta_{min}\left(\delta_1, z^{LB}\right)}\right)$, thus contradicting the optimality of $\theta_{min}\left(\delta_1, z^{LB}\right)$. $\square$

Corollary 1. *For every $x \in \Omega$ such that $\mathbb{E}\left(Y_{(k)}(x)\right) \geq z^{LB}$, if $\delta(x) \in [a, b]$, $\theta_{min}\left(a, z^{LB}\right) \leq \theta(x)$.*

From Corollary 1, we obtain the following expression for the lower bounds of θ^k_{min} given δ_k:

$$\theta^k_{min} = \theta_{min}\left(\delta_k, z^{LB}\right) \quad \forall k = 1, \ldots, l. \tag{29}$$

From this, we can add the following inequality to SRMP:

$$s \geq \left(\theta^k_{min}\right)^2 y_k, \qquad \forall k = 1, \ldots, l.$$

This inequality provides a significant tightening of the bound by relating the possible choices of k for $y_k = 1$ to the best known solution. Note that, for this, one needs to find a lower bound on $\mathbb{E}\left[Y_{(2)}(x)\right]$ and solve l problems $LBP(\delta)$ at the initialization of the algorithm.

6. Numerical Evaluation

In this section we describe results of an extensive evaluation on synthetic instances. The models and algorithms are implemented in CPLEX version 12.7.1 (ILOG 2018) through the Java API using the best bound exploration strategy in the solver. We utilize `Python` 3.6 to generate the random instances that will be available to the interested reader. All experiments are conducted on an Intel(R) Xeon(R) CPU E5-2680 v2 at 2.80GHz. Each execution is restricted to a single thread, a time limit of 60 minutes, and a memory limit of 10 GB.

6.1. Generation of Synthetic Instances

We generate synthetic instances to measure runtime for understanding the scalability of the algorithm and the quality of the solution obtained by our exact approach. For the latter, we are interested in understanding how the solutions we find compare with those produced by a simple heuristic (which we describe below) and which instance characteristics lead to significant improvement in objective function over the highest single entry solution.

The selection of items for instances of this family is restricted by knapsack constraints on the bins. The instances composing this dataset were randomly generated, with each one being associated with a configuration (p, w, d, α). Items have weights and each bin has a capacity on the sum of the weights of the packed elements; both the capacity of the bins and the item weights are constructed in a way that the number of items that can be assigned to each bin is w in expectation. Binary parameter d indicates if items may belong to both bins or if they can be selected at most once. Finally, α indicates an integer *variance multiplier*. The benchmark consists of 5 randomly generated instances per configuration. The set of configurations consists of all combinations of p in $\{15, 20, 25, 30\}$, w in $\{2, 3, 4, 5\}$, d in $\{0, 1\}$, and α in $\{50, 100, 150, 200, 250\}$.

Given the parameters of the configuration, we generate each instance as follows. First, item weights are integer values drawn independently and uniformly at random from the interval $[1, 19]$, and the knapsack capacity per bin is set to $10w$; note that items have weight 10 in expectation, so the expected number of items that can fit in each bin is w. Profit means μ_i are continuous values drawn independently and uniformly at random from the interval $[15, 25]$. In order to generate the covariance matrix, we use `Scikit-learn` functions to generate a random positive semidefinite matrix and to find the nearest covariance (matrix Pedregosa et al. (2011)); the resulting covariance matrix is multiplied by α.

6.2. Runtime Analysis and Effect of Strengthened Formulation

We evaluate the impact of the strengthened formulation presented in Section 5 over the basic version of the algorithm presented in Section 4.2, which we refer to as $\mathbf{A}$ and $\mathbf{A}^0$, respectively. The metrics we use for comparison are runtime and ending solution quality. The analysis provides significant evidence that the machinery developed in Section 5 is required for scalable implementation of the branch-and-cut algorithm since $\mathbf{A}$ improves substantially over $\mathbf{A}^0$ on both metrics.

Figure 1 **Performance profile comparing runtimes of the different problem classes tested.**

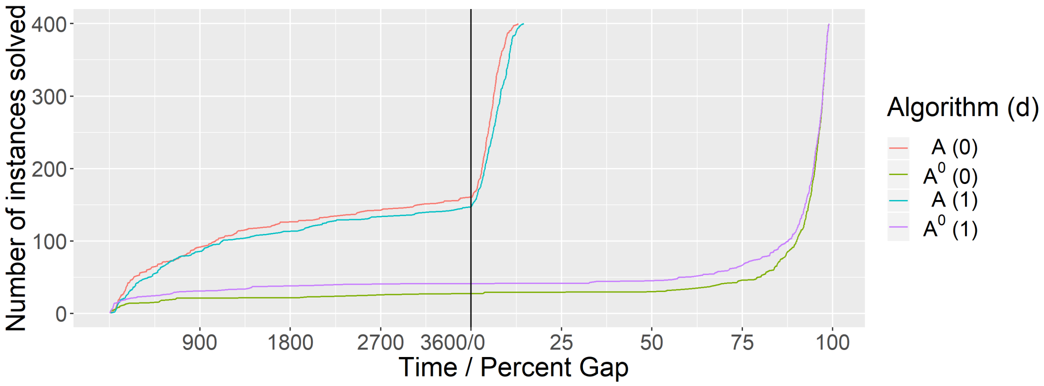
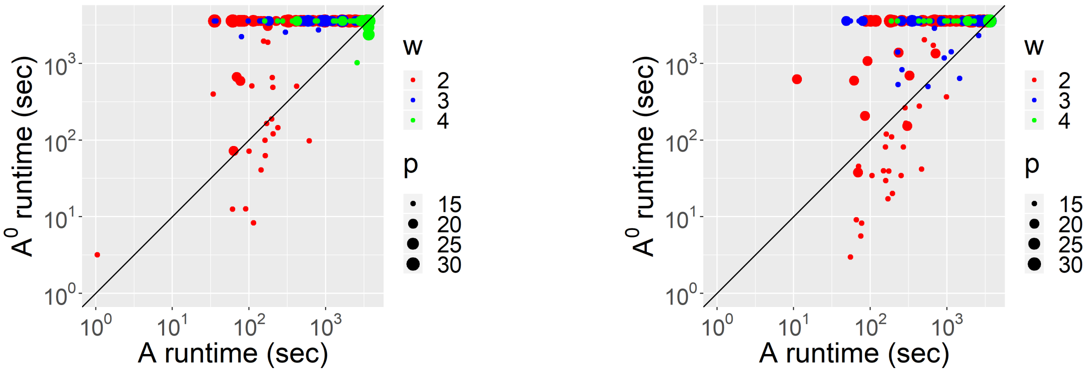


Figure 1 is a performance profile depicting runtimes and ending optimality gaps (measured as percent difference of ending lower bound from upper bound) of **A** and $\mathbf{A}^0$, both allowing $(d = 0)$ and disallowing $(d = 1)$ items from being selected for both bins. Each line corresponds to an algorithm and a value of d. In the left portion of the plot, the height of the line is the number of instances solved by the corresponding time on the x-axis; in the right portion, the height of the line is the number of instances with optimality gap less than or equal to the corresponding value on the x-axis. In particular, the lines are all non-decreasing, and the higher the line the better the performance.

The figure readily exhibits substantial performance enhancement in **A** over $\mathbf{A}^0$, regardless of d. For $d = 0$, **A** solves 160 of the 400 instances versus only 27 for $\mathbf{A}^0$; for $d = 1$, **A** solves 147 and $\mathbf{A}^0$ solves 41. The ending gaps are also substantially tighter for **A**; averaged for all instances and over all runs with $d = 0$ and $d = 1$, the optimality gap is 4% for **A** and 84% for $\mathbf{A}^0$. Figure 1 therefore provides clear evidence of the efficiency of **A**.

It is also interesting to see which problem characteristics result in harder instances. The scatter plot in Figure 2 depicts the runtime of **A** and $\mathbf{A}^0$ for $d = 0$ (left) and $d = 1$ (right). Each point has coordinates given by the runtime of **A** and $\mathbf{A}^0$, with dots being sized proportionally to the number of items, m, and colored by the knapsack size, w. Note that we only depict those instances with $w \in \{2, 3, 4\}$, omitting those with $w = 5$ for clarity. The plot shows that, as w and m increase, it becomes critical to utilize **A**. $\mathbf{A}^0$ generally only solves instances with $w = 2$ and the overhead of its use only contributes to longer runtimes for the easiest instances.

Figure 2 **Scatter plots comparing runtimes for A and A^0**



(a) Instances with $d = 0$.

(b) Instances with $d = 1$.

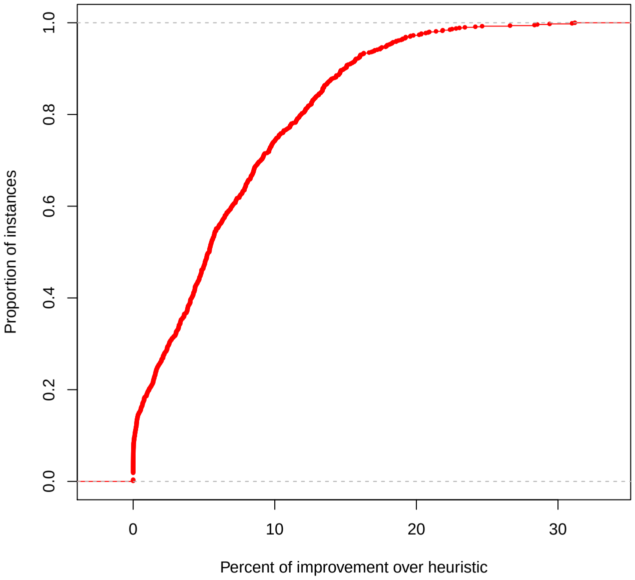
6.3. Comparison with a Heuristic

We discuss now the improvements brought by the strengthened formulation over a natural heuristic approach that consists of selecting bins $\mathcal{S}_1$ and $\mathcal{S}_2$ to be the highest and second-highest expected profit bins. Note that the problem of maximizing the expected profit for a single bin reduces to a traditional deterministic knapsack problem. Our heuristic first solves this knapsack problem to determine $\mathcal{S}_1$, i.e., the highest expected profit bin. To find $\mathcal{S}_2$, we add a no-good cut that prevents $\mathcal{S}_1$ to be selected again resulting in the second-highest expected profit bin. Note that if $d = 0$, then the no-good cut ensures that $\mathcal{S}_1$ and $\mathcal{S}_2$ differ in at least one item, as items are allowed to belong to both bins. If $d = 1$, then the no-good cut ensures that none of the items in $\mathcal{S}_1$ are selected in $\mathcal{S}_2$.

The running time of the heuristic is negligible (i.e., less than 1 second), so we restrict the comparison to the quality of the solutions; namely, we investigate the relative gains of the strengthened formulation over the heuristic. The comparison is always between the solution obtained by the heuristic and the best feasible solution found by the algorithm before the time or the memory limit was exceeded.

Figure 3 shows the results in an aggregate way, displaying the proportion of instances for which the percent gap is a given constant or less. In particular, a point (x, y) on the line reports the proportion of instances y for which the percent improvement over the heuristic is at most x. Improvements over the heuristic were approximately 6.7%, going up to over 30% in some scenarios.

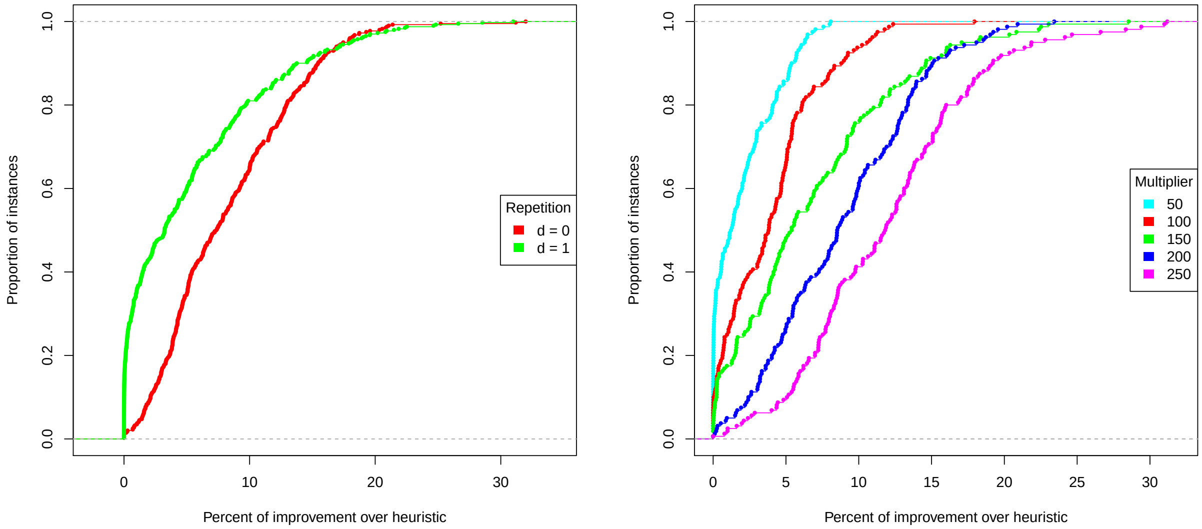
Figure 3 **Proportion of instances with percent improvement over heuristic.**



The plots composing Figure 4 present a more refined analysis, where the results are decomposed according to some of the parameters defining the underlying instances. Figure 4(b) shows that the variance multiplier has a significant impact on solution quality. Namely, whereas the average improvement is slightly below 2% for $\alpha = 50$, the value grows to almost 12% for $\alpha = 250$. Our results also show that performance improvements are positively correlated with repetition of items, as we can see in Figure 4(a). Namely, if items can be assigned to both bins, the average improvement is approximately 5.2%; conversely, the value goes up to 8.1% if repetitions are forbidden. Finally, the plots associated with the number of items and the knapsack constraint were similar to Figure 3, i.e., these parameters did not have a significant impact on performance improvements.

In Figure 5, we present the differences in the expected profit of the bins selected by the two different approaches. Let $\mathcal{S}_1^h$ and $\mathcal{S}_2^h$ be the sets selected by the heuristic for bin 1 and 2, respectively, and $\mathcal{S}_1^e$ and $\mathcal{S}_2^e$ be the bins selected by the exact approach. Recall that by construction $z(\mathcal{S}_1^h) \geq z(\mathcal{S}_2^h)$ and $z(\mathcal{S}_1^e) \geq z(\mathcal{S}_2^e)$. Each point in the plot represents one instance, with its x and y coordinate indicating $z(\mathcal{S}_1^h) - z(\mathcal{S}_1^e)$, and $z(\mathcal{S}_2^h) - z(\mathcal{S}_2^e)$, respectively. Note that since the heuristic selects the two highest expected profit bins, when $d = 0$ we have that $z(\mathcal{S}_1^h) \geq z(\mathcal{S}_1^e)$ and $z(\mathcal{S}_2^h) \geq z(\mathcal{S}_2^e)$. When

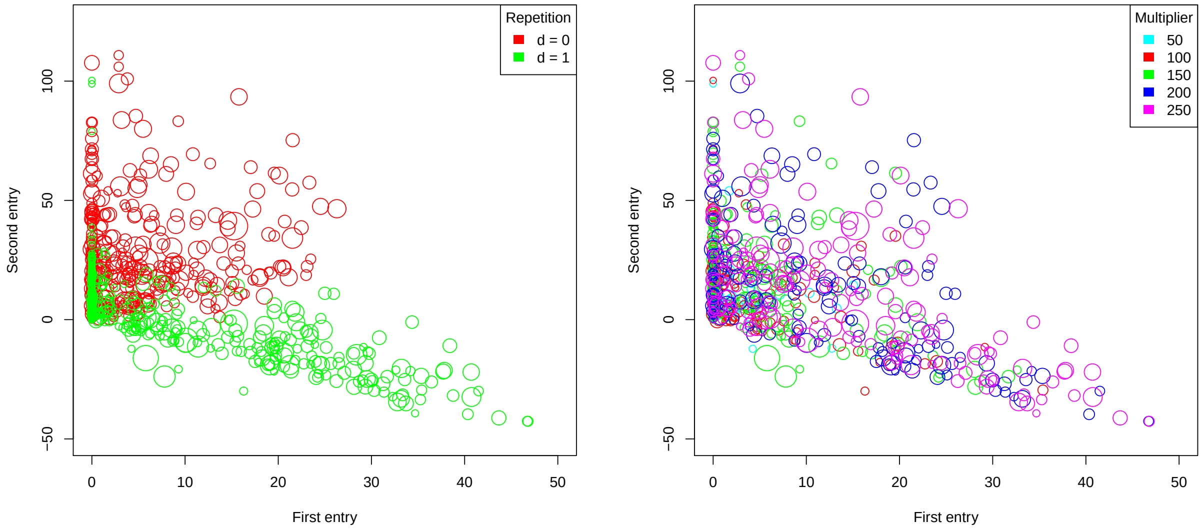
Figure 4 **Proportion of instances with percent improvement over heuristics refined by instance characteristics.**



(a) Results defined by item repetition.

(b) Results defined by α.

Figure 5 **Differences on expected profits of selected bins.**



(a) Instances decomposed by item repetition.

(b) Instances decomposed by α.

$d = 1$ we still have that $z(\mathcal{S}_1^h) \geq z(\mathcal{S}_1^e)$ but cannot establish the same for $\mathcal{S}_2^h$ and $\mathcal{S}_2^e$ since items can only be selected in at most one bin. The relative sizes of the points depict the percent improvement in the solution quality of the exact approach over the heuristic.

Similarly to previous plots, we investigate the distribution of these values based on the possibility to repeat items (Figure 5(a)) and on the variance multiplier (Figure 5(b)). Lack of repetition is correlated with larger difference on the second bin, whereas repetitions may increase the difference on the first. Also, larger variance multipliers lead to larger differences on the expected profit of both bins. These results suggest the existence of correlation at some level between differences in the expected profits of selected items and solution quality. Our results show that these values are indeed slightly correlated: differences in profit for the first bin selected have a correlation of 0.35 with solution quality improvement and differences in the second bin have a correlation of 0.18.

Perhaps the most interesting takeaway from Figure 5 is how different the optimal entries are from the heuristic, exhibiting how important using an exact approach can be for this problem. It is often the case that $z(S_1^e) = z(S_1^h)$, but when this happens, $z(S_2^e)$ is typically much smaller than $z(S_2^h)$. It is also often the case that both of the entries found by the exact approach have significantly lower individual expectations than both entries identified by the heuristic, but the expectation of their maximum is much higher. For example, for one of the large instances composing this dataset ($m = 30$, $w = 5$, $\alpha = 250$, and $d = 0$) both entries selected by the heuristic were considerably larger than the respective entries generated by our algorithm ($z(S_1^h) - z(S_1^e) = 14.59$ and $z(S_2^h) - z(S_2^e) = 41.60$); nevertheless, the improvement brought by the exact approach was over 17%, thus reinforcing the superiority over a greedy heuristic strategy.

7. Application to Daily Fantasy Football

Fantasy sports are games in which participants serve as “general managers” of imaginary “rosters” made up of real players from a professional sport. The players’ real performance during game play translates into fantasy points, and fantasy sports participants compete against each other by accumulating those points. The companies running fantasy sports contests set a price for all eligible players and allow participants to select a roster while remaining under a specified spending limit, or salary cap; eligible players are those participating in games covered by the specific daily fantasy contest.

In daily fantasy football (DFF), fantasy contests in which participants compete are arranged based on the starting times of each week's slate of National Football League (NFL) games, and only players in those games are eligible for inclusion on a fantasy roster. Further, not all players in the NFL games are eligible for fantasy rosters because the fantasy scoring system only rewards points for specific tasks. Namely, only the offensive "skill position" players (wide receivers, running backs, and tight ends), quarterbacks, and kickers are eligible as single players; the other players are grouped and selected collectively as "team" defenses. Generally the individual players receive fantasy points for gaining yardage and scoring points in the actual game (via touchdowns, extra points, two-point conversions, and field goals) and the team defense earns fantasy points by preventing the opposing team from scoring or by having its members scoring game points itself.

`DraftKings` is one of the two major DFF providers. Three different types of football contests are offered: showdowns, classics, and tiers. Showdown contests only include the players in a single NFL game; entries consist of 6 players, regardless of position, with one designated as captain who costs 1.5 times the normal salary and earns 1.5 times the normal fantasy points. Classic contests include players in at least two and up to any number of NFL games and entries consist of 9 players: 1 quarterback, 2 running backs, 3 wide receivers, 1 tight end, 1 team defense, and 1 "flex" position that can be filled by a running back, wide receiver, or tight end. Tier contests feature no player salaries; rather, participants select a lineup by choosing a single player, regardless of position, from each of 6 pre-set "tiers" of players. Finally, across all contests, eligible lineups must include players from at least two different NFL teams. We focus on showdown contests in this paper.

It is helpful to reiterate the terminology used in the preceding paragraphs. Specifically, *player* refers to a professional athlete who is eligible for selection on a fantasy roster, *team* refers to the actual team the athlete plays for on the field, *game* refers to the actual athletic competition, *contest* refers to a competition on the `DraftKings` platform that costs money to enter and pays out monetary prizes to highest scoring entries, *entry* or *roster* refers to a given selection of players that forms a full fantasy roster and is submitted for a given contest, *entry fee* is the cost paid to submit one entry in a given contest, *payout* is the amount rewarded to high-scoring entries that differs across contests, and *participant* refers to the person who selects at least one entry. Each player may appear no more than once on a given entry, and different contests allow different number of entries per participant (some as high as 150). We focus here on smaller, high-entry-fee contests that limit each participant to no more than 2 or 3 entries.

The last important point to understand about DFF contests in the payout structure. It is heavily skewed toward the top performers in each contest. Generally the top 1/4 of entries place in the money. The very top entry receives approximately 1/4 of all entry fees, and the payouts fall quickly after that, although exact amounts can depend on the contest. The last set of payouts is usually 1.5 times the entry fee.

7.1. Problem definition

We model the two-entry selection problem in a showdown contest as a special case of Problem P. Let p' be the number of players, and $p = 2p'$. The first p' players represent standard versions of the players ("flex" in the lingo used by `DraftKings`), whereas the next p represent their "captain" versions, in a common order. Players are therefore represented by items in P, with selection costs representing salaries and associated stochastic profit representing expected fantasy points; let $\{\mathcal{I}_1, \mathcal{I}_2\}$ be the partition of $\mathcal{I}$ containing the players from the first and second team, respectively.

For $i \in \{1,2\}$ and $j \in [p]$, binary variable $x_{i,j}$ indicates the selection of player j for roster i; that is, rosters correspond to the bins. The following constraints apply to each roster:

- Exactly 5 flex players and 1 captain must be selected:

$$\sum_{j=1}^{p'} x_{i,j} = 5, \sum_{j=p'+1}^{p} x_{i,j} = 1, \quad i = 1,2.$$

- The same player cannot be selected both for a flex and the captain positions:

$$x_{i,j} + x_{i,j+p'} \leq 1, \quad i = 1,2, \ j = 1,\ldots,p'.$$

- At least one player from each team appears on each roster:

$$\sum_{j\in\mathcal{I}^i} x_{i,j} \geq 1, \quad i=1,2.$$

Finally, all decisions are binary, so $x_{i,j}\in\{0,1\}$ completes the constraint set.

There are additional constraints that we add, which implement standard "stacking" rules. These rules help reduce the search space by eliminating certain combinations of players that generally perform do or do not perform well with one another. Specifically, we implement the following rules for any entries selected:

1. if a QB is selected as captain, the defense from either team is not selected; and
2. if a WR or TE is selected as captain, always pick a QB on his team.

7.2. Parameter Estimation

There are several parameters that need to be estimated—in particular, $\forall j\in[p]$, parameters μ_j and σ_j defining the normal distribution for the points scored by player j; and, $\forall j_1, j_2\in[p]$, the covariance ρ_{j_1,j_2} of the performance of players j_1 and j_2. We discuss each in turn.

7.2.1. Expected Value Estimation Due to the growth of the industry of fantasy sports, estimated DFS points for a player is the topic of many non-academic articles and websites (e.g., rotogrinders.com). Tremendous resources are put into calculating reliable and accurate estimates for how a player will perform. Although there is potentially room to improve upon published estimates, for this paper we use estimated player data from fantasydata.com, which is updated frequently for upcoming games and contains historical NFL game data since 2014 of projected fantasy points for players, actual fantasy points for players, and player salaries on the `DraftKings` platform, all for a monthly fee. The projected points estimates used for this paper are the final fantasydata.com estimates set just before game time (changes may occur due to weather, injury updates, or a myriad of other reasons). This data therefore provides μ_j, for $j=1,\ldots p'$, and, because $\mathbb{E}(1.5X)=1.5\mathbb{E}(X)$, we have $\mu_j = 1.5\mu_{j-p'}$, for $j\geq p'+1$.

7.2.2. Variance Estimation We used the data from fantasydata.com in order to learn variances for player scores using two different methods. The first method uses all the projected points data for games played from 2014-2017 as the training set. First, for each position, we partition the data into quintiles. Then, for the 2018 data, we slot each player j into his respective position-specific quintile of projected points, and use the variance of the actual points scored by the players in that quintile in 2014-2017 as the estimate for the variance of that player's actual score. For example, if a running back is *projected* to score 20.5 fantasy points in a game, and that falls in the top quintile for running backs, we set his actual score variance as the computed variance of the *actual* scores of all top-quintile running backs from 2014-2017.

The second method employs a k-nearest-neighbor-like algorithm; again, we use 2014-2017 as our training set. For a player j, his variance is estimated as the variance of the actual scores of the 50 players that share a common position with player j in the data for 2014-2017 and whose expected value is as close as possible (in terms of squared difference) to player j. For example, if a running back is *projected* to score 20.5 fantasy points, we select the 50 running backs in the training set with *projected* fantasy score as close as possible to 20.5, and use the *actual* fantasy scores of those 50 players to calculate the variance of that player's fantasy score.

7.2.3. Correlation Estimation Due to the nature of the sport itself and the way fantasy scoring works, players on the same (or opposing) teams often have correlated scoring. For example, if a quarterback throws a touchdown to a wide receiver, both players receive a number of fantasy points (and the opposing defense would possibly lose some). But since that quarterback had to have completed the pass to someone, fantasy points for individual players do not exist in a vacuum. Rather, we would expect that players at certain positions would have significantly correlated scores with teammates, and even opponents, at other positions. Therefore, given the heavily skewed payoff structure of most `DraftKings` showdown contests, when trying to maximize the expected value of the maximum score of the entries, participants must take into account these correlations. The daily fantasy sports betting community is well aware of this strategy of choosing teammates (or opponents) whose scores should correlate by design; it is called "stacking." It should also be noted that opponents may see their actual scores correlate *even if the players are not on the field at the same time* due to the nature of the game of football. For example, the actual scores of the quarterbacks on opposing teams (who are *never* on the field at the same time) are positively correlated because higher scoring NFL games tend to generate more fantasy points for the quarterbacks. As one quarterback scores fantasy points, the other team tends to throw more to try and catch up, and this tends to raise the opponent quarterback's fantasy scores as a result.

We estimate the correlation of players j_1 and j_2 in similar ways to our single player variance estimates. Since we are only reporting results for showdown contests, all available player pairs are either on the same or opposing teams.

In the first method, we slot all teammate pairs of players j_1 and j_2 into which of the 25 possible quintile-quintile buckets the pair falls. For example, if a quarterback projects for 30 points (first quintile) and his wide receiver projects for 15 points (second quintile), we find all 1st quintile-2nd quintile quarterback-wide receiver pairs in the training set, and use the correlations of those actual player scores as our parameter estimate for ρ_{j_1,j_2}. The same technique is applied to estimate correlations of pairs of players on opposing teams.

In the second method, we estimate ρ_{j_1,j_2} (for players on the same team) by finding the 50 pairs of players and games in the training set for which the players are teammates and played the same positions as j_1 and j_2 and have expected value as close as possible to that of j_1 and j_2, and use the sample correlation of their actual game scores to estimate their correlation. For example, if a quarterback and wide receiver pair are on the same team and are expected to score 30 and 15 points, respectively, we find the 50 instances in the training set of quarterback and wide receiver teammates with the sum of squared differences from 30 and 15 in expected values as low as possible. We calculate the corresponding correlation of actual scores on this subset of players, and use that as the parameter estimate. We do the same for players on opposing teams. One can then set the correlation to 0 if the significance of the `Person's correlation test` is above some threshold β.

Using either estimation technique, there are instances in which the correlation identifies values that are incompatible, because of the way the estimation is done and the actual scores achieved by certain player pairs. Formally, the estimated covariance matrix Σ for a multivariate normal distribution must be positive semi-definite (PSD). Correlation estimates, multiplied by the corresponding standard deviations, are found only to provide an estimate for Σ, and the estimation procedure described above often yields correlations for which Σ is not PSD. There are several ways to correct for this. One is to use packages available in common statistical software to adjust the covariances, like the function `cov_nearest` in `Python`'s `statsmodels` module. Another is to dampen the correlations by a fixed constant (i.e., set $\rho_{j_1,j_2} := \frac{\rho_{j_1,j_2}}{\alpha}$, with $\alpha > 0$, for all j_1,j_2). Yet another way is to apply the PSD condition and, whenever a solution is found for which $\theta < 0$, redefine $\theta := 0$. Throughout the season of betting several procedures were tested.

7.3. Results

In Tables 1 through 3, we report results that would have been obtained on 16 showdown contests from the 2018 season by our exact algorithm using `statsmodels` to fix the covariance matrices and the different variance / correlation estimation procedures defined above; these contests were selected because we actually bet on them in `DraftKings` and therefore have their results. The actual entries we selected and used for betting vary slightly from those reported here due to evolving techniques for estimating parameters over the course of the season—we report a uniform technique in the paper to ensure reproducibility. Our actual results turned an even larger $6,900 profit.

The content of the first seven columns of Tables 1 through 3 is clear from their titles: the first column indicates the game and the date when it took place; the second indicates the total number of entries; the third indicates the price paid by participants per entry; the fourth and fifth indicates the results by showing the winnings and the profit obtained by the algorithm, respectively; columns six and seven indicate the best score and the minimum score for entries in the money; columns eight and nine contain the expected values of the first and the second entry, respectively; and, finally, columns ten and eleven represent the actual scores of the selected entries.

Because 2-entry showdown contests are not very common on the DraftKings platform, we included a number of 3-entry contests on this list, though we only ever enter 2 entries for those contests. As baseline, we used entries generated by a simple heuristic that would select a first entry with maximum expected value and a second entry that differs from the first by at least one player and maximizes the expected value.

Table 1 shows results obtained using the "tiered" method (hereafter, C1), Table 2 shows results obtained using the nearest-neighbor method (hereafter, C2, using $\beta = 0.25$ and cov_nearest for correcting Σ, i.e., setting to 0 any correlation whose p-value is above 0.25), and Table 3 shows the results for the heuristic. Each of the tables include the NFL teams involved and date of the game, total number of entries in the contest, total entry fee for our 2 entries, total winnings for both of our entries, profit for the game, score of the top finishing entry, minimum payout score (i.e., the cutoff score for winning any money back), expected value of our first entry, expected value of our second entry, actual value of our first entry, and actual value of our second entry. The actual value appears in bold if it would have finished "in the money" in that particular contest. The first seven columns are repeated in all three tables. Finally, the total entry fees, winnings, and profit for each method are computed in the last row of each table. Note however that the money won shown in Tables 1 through 3 is an approximation in the sense that if our 2 entries from the algorithms in this paper were *actually* included in the contests, two other entries that were in the contest necessarily would have been excluded since all contests fill up to capacity. If, for example, the two "eliminated" entries were top scoring ones, the payoffs for the remaining entries, and ours, would necessarily be the same or larger. There is no way to account for this game of elimination in a reasonable way, so we report results as if our entries were simply added to the contest without changing any other details about other entries, payout structure, etc.

There are two important initial takeaways in Tables 1 through 3. First, the contests differ in cost and entries considerably, though the most common version of this contest features 100 total entries at a cost of $444 per entry. Second, the scores needed to win, or even place in the money, differ tremendously by game. Namely, higher-scoring NFL games result in more fantasy points for the involved players; for instance, in the game `Vikings vs. Rams on 9-27`, which finished with a score of 38-31, the *minimum payout score* would have been the *winner score* in every single other contest shown. Because of the vast discrepancy in contest results, it is difficult to say *a priori* what score is going to be necessary to win or place in a given contest.

The comparison of overall results across the three methods reveals the power of our algorithm. The heuristic ends up losing nearly 50% of the money wagered; our optimal algorithm with C1 loses approximately 10%, which is the cut that `DraftKings` takes for running the contests; finally, our optimal algorithm with C2 shows a positive return of over 50%. There were two games whose contests featured different profits for the three methods shown: `Redskins vs. Saints on 10-8` and `Broncos vs. Chiefs on 10-1`. There were two games whose contests featured positive profit for all three methods shown: `Giants vs. Falcons on 10-22` and `Falcons vs. Saints on 11-22`; for them, both entries selected by the heuristic and C2 methods won money. Although both entries finishing “in the money” might seem at first glance to be a great outcome, we actually view it in the context of C2 as a missed opportunity. In general, small differentials between the actual scores of our entries represents a failure of the covariance matrix to manifest itself in the game played on the field, and it usually results in us losing money on the contest. Because of the skewed payout structure of the `DraftKings` contests, we would prefer our entries to completely play off of each other. One entry finishing in first place and the other in last place would be a more profitable outcome than both finishing above the minimum payout score but below the top few spots. For more detail, look at the contest for the `Giants vs. Falcons on 10-22` game in Table 2. There, our entries would have placed 5th and 19th (the top 23 entries were “in the money”). We would gladly have traded down from the 19th spot out of the money to move up just a few more points with our 5th place finishing entry. Just 6.5 more actual points for entry 1 would have earned a $3,000 payoff, and 9.8 more would have earned first place and a $10,000 payoff.

In nearly all contests, across all three specifications, the winner score is significantly higher than the expected value scores (the only major exception being the contest featuring the `Falcons vs.`

`Saints on 11-22`). The players on the winning entry, as a whole, significantly outperform their projections in nearly every contest. Something akin to the heuristic method is what we expect many participants in these contests are using as their DFF strategy. Although it unsurprisingly outperforms our single entries in both C1 and C2 in *expected value*, its actual results leave a lot to be desired. As we have observed in our experiments with synthetic instances, in general, the maximum expected value entry does not necessarily belong to an optimal configuration of entries, and it also will not necessarily result in a payout, so bettors need to exploit joint decision-making across multiple entries in order to elevate their actual scores above the threshold needed to win money.

Table 4 further illustrates the power of our algorithm for P. We again compare to the heuristic by evaluating expression 5 using the solution that the heuristic obtains together with the covariance matrices produced by methods C1 and C2. Game is the same as before, EV represents the single highest expected value entry, AV represents the single highest actual score on our entries, and MEVM represents the maximum of the expected value of the maximum for the two entries, i.e., the evaluation of the objective function at the solution obtained. Notice that while the heuristic dominates P in EV, it falls behind in MEVM in every single game considered, for both covariance considerations. The last row of the table considers the average values above it. On average, the heuristic solution evaluated using C1 is 1.64 points higher in EV than P with C1, but P ends up 3.82 points higher than the heuristic in MEVM. Likewise, the heuristic solution evaluated using C2 is 1.19 points higher in EV than P with C2, but P ends up 6.68 points higher than the heuristic in MEVM. This shows the power of our algorithm, with the consummate benefit being the ability to win money in DFF for multiple entry showdown contests. Finally, the MEVM for P with C2 is 6.21 points higher than the MEVM for P with C1, which explains why it was able to earn such higher profits. This is despite the fact that the average AV for P with C1 actually outperformed the average AV for P with C2 by 1.05 points. In other words, P with C2 better achieves the goal of producing a really high score, at least some of the time, and often enough to win significant money.

Table 1 Results from real-world betting scenarios using P and covariance 1

Game	Entrants	Entry Fee	Winnings	Profit	Winner Score	Min. Pay Score	EV E1	EV E2	AV E1	AV E2
Jets vs. Browns on 9-20	100	$666	$0	-$666	96.51	67.96	77.06	76.05	66.71	43.60
Seahawks vs. Cowboys on 9-23	1,189	$8	$0	-$8	109.78	76.78	79.63	77.97	75.22	61.57
Vikings vs. Rams on 9-27	83	$800	$800	$0	205.28	175.13	96.69	93.77	118.88	**185.43**
Ravens vs. Steelers on 9-30	294	$40	$0	-$40	111.01	96.73	91.02	90.69	66.56	87.68
Broncos vs. Chiefs on 10-1	69	$800	$5,000	$4,200	110.36	90.69	109.85	104.34	72.75	**124.30**
Colts vs. Patriots on 10-4	100	$888	$700	-$188	148.09	131.85	93.59	90.83	112.11	**134.50**
Redskins vs. Saints on 10-8	100	$888	$0	-$888	134.97	99.67	109.51	93.16	78.95	78.50
Giants vs. Falcons on 10-22	100	$888	$1,200	$312	143.56	120.70	100.39	97.97	**130.12**	98.36
Raiders vs.49ers on 11-1	100	$888	$0	-$888	95.81	78.61	76.99	76.29	76.12	37.61
Panthers vs. Steelers on 11-8	100	$888	$700	-$188	149.93	118.70	101.64	101.60	**118.70**	105.09
Giants vs. 49ers on 11-12	100	$888	$0	-$888	120.60	95.18	91.00	90.61	81.52	81.37
Packers vs. Seahawks on 11-15	100	$888	$0	-$888	126.53	114.43	91.12	88.11	107.43	96.68
Steelers vs. Jaguars on 11-18	70	$66	$0	-$66	115.86	96.49	92.96	88.87	94.45	81.31
Falcons vs. Saints on 11-22	151	$150	$300	$150	122.22	106.52	129.62	121.39	**114.12**	90.47
Redskins vs. Eagles on 12-3	100	$888	$0	-$888	111.36	84.29	87.68	83.68	81.94	76.59
Jaguars vs. Titans on 12-6	882	$40	$0	-$40	134.60	74.83	78.48	74.82	44.73	61.07
TOTALS		$9,674	$8,700	-$974						

Table 2 Results from real-world betting scenarios using P and covariance 2

Game	Entrants	Entry Fee	Winnings	Profit	Winner Score	Min. Pay Score	EV E1	EV E2	AV E1	AV E2
Jets vs. Browns on 9-20	100	$666	$0	-$666	96.51	67.96	77.97	76.55	64.07	42.02
Seahawks vs. Cowboys on 9-23	1,189	$8	$0	-$8	109.78	76.78	79.63	77.97	75.22	61.57
Vikings vs. Rams on 9-27	83	$800	$0	-$800	205.28	175.13	98.50	95.90	159.28	150.13
Ravens vs. Steelers on 9-30	294	$40	$0	-$40	111.01	96.73	93.40	89.16	84.66	86.17
Broncos vs. Chiefs on 10-1	69	$800	$0	-$800	110.36	90.69	106.88	106.46	79.69	71.50
Colts vs. Patriots on 10-4	100	$888	$700	-$188	148.09	131.85	93.13	92.75	114.19	**132.34**
Redskins vs. Saints on 10-8	100	$888	$10,000	$9,112	134.97	99.67	104.74	102.74	65.55	**140.80**
Giants vs. Falcons on 10-22	100	$888	$2,700	$1,812	143.56	120.70	103.03	101.08	**133.76**	**123.26**
Raiders vs.49ers on 11-1	100	$888	$1,200	$312	95.81	78.61	77.54	77.35	**86.72**	43.61
Panthers vs. Steelers on 11-8	100	$888	$0	-$888	149.93	118.70	103.03	99.76	107.00	115.49
Giants vs. 49ers on 11-12	100	$888	$0	-$888	120.60	95.18	91.00	90.61	81.52	81.37
Packers vs. Seahawks on 11-15	100	$888	$0	-$888	126.53	114.43	91.12	88.11	107.43	96.68
Steelers vs. Jaguars on 11-18	70	$66	$0	-$66	115.86	96.49	92.96	88.87	94.45	81.31
Falcons vs. Saints on 11-22	151	$150	$450	$300	122.22	106.52	130.90	129.62	**110.94**	**114.12**
Redskins vs. Eagles on 12-3	100	$888	$0	-$888	111.36	84.29	88.60	83.49	82.74	64.14
Jaguars vs. Titans on 12-6	882	$40	$0	-$40	134.60	74.83	78.30	77.25	54.93	47.83
TOTALS		$9,674	$15,050	$5,376						

Table 3 **Results from real-world betting scenarios using heuristic**

Game	Entrants	Entry Fee	Winnings	Profit	Winner Score	Min. Pay Score	EV E1	EV E2	AV E1	AV E2
Jets vs. Browns on 9-20	100	$666	$0	-$666	96.51	67.96	78.79	78.50	63.82	60.81
Seahawks vs. Cowboys on 9-23	1,189	$8	$0	-$8	109.78	76.78	80.93	80.91	64.20	66.70
Vikings vs. Rams on 9-27	83	$800	$0	-$800	205.28	175.13	98.50	97.91	159.28	147.68
Ravens vs. Steelers on 9-30	294	$40	$0	-$40	111.01	96.73	94.07	93.40	73.80	84.66
Broncos vs. Chiefs on 10-1	69	$800	$1,200	$400	110.36	90.69	109.91	109.85	**99.60**	72.75
Colts vs. Patriots on 10-4	100	$888	$0	-$888	148.09	131.85	94.25	93.59	119.44	112.11
Redskins vs. Saints on 10-8	100	$888	$1,000	$112	134.97	99.67	109.51	108.88	78.95	**114.20**
Giants vs. Falcons on 10-22	100	$888	$2,700	$1,812	143.56	120.70	103.03	103.03	**133.76**	**123.96**
Raiders vs.49ers on 11-1	100	$888	$0	-$888	95.81	78.61	80.21	80.21	71.49	76.79
Panthers vs. Steelers on 11-8	100	$888	$0	-$888	149.93	118.70	103.93	103.59	118.50	99.00
Giants vs. 49ers on 11-12	100	$888	$0	-$888	120.60	95.18	93.45	93.30	80.72	88.72
Packers vs. Seahawks on 11-15	100	$888	$0	-$888	126.53	114.43	92.54	92.26	105.43	105.08
Steelers vs. Jaguars on 11-18	70	$66	$0	-$66	115.86	96.49	95.26	94.92	80.25	82.67
Falcons vs. Saints on 11-22	151	$150	$400	$250	122.22	106.52	130.90	130.02	**110.94**	**113.06**
Redskins vs. Eagles on 12-3	100	$888	$0	-$888	111.36	84.29	88.70	88.60	78.94	82.74
Jaguars vs. Titans on 12-6	882	$40	$0	-$40	134.60	74.83	79.53	79.38	49.17	36.73
TOTALS		$9,674	$5,300	-$4,374						

Table 4 Comparing P and Heuristic

	P with C1			Heuristic with C1			P with C2			Heuristic with C2		
Game	EV	MEVM	AV	EV	MEVM	AV	EV	MEVM	AV	EV	MEVM	AV
Jets vs. Browns on 9-20	77.06	88.54	66.71	78.79	83.72	63.82	77.97	93.30	64.07	78.79	63.82	88.35
Seahawks vs. Cowboys on 9-23	79.63	89.80	75.22	80.93	85.32	66.70	79.63	96.81	75.22	80.93	88.69	66.70
Vikings vs. Rams on 9-27	96.69	109.85	185.43	98.50	105.28	159.28	98.50	115.53	159.28	98.50	105.93	159.28
Ravens vs. Steelers on 9-30	91.02	105.52	87.68	94.07	102.71	84.66	93.40	110.12	86.17	94.07	109.82	84.66
Broncos vs. Chiefs on 10-1	109.85	124.61	124.30	109.91	121.98	99.60	106.88	133.03	79.69	109.91	125.27	99.60
Colts vs. Patriots on 10-4	93.59	103.82	134.50	94.25	101.40	119.44	93.13	107.85	132.34	94.25	107.29	119.44
Redskins vs. Saints on 10-8	109.51	123.83	78.95	109.51	118.88	114.20	104.74	131.43	140.80	109.51	119.91	114.20
Giants vs. Falcons on 10-22	100.39	113.03	130.12	103.03	107.72	133.76	103.03	119.65	133.76	103.03	107.43	133.76
Raiders vs.49ers on 11-1	76.99	88.41	76.12	80.21	84.66	76.79	77.54	94.74	86.72	80.21	85.07	76.79
Panthers vs. Steelers on 11-8	101.64	114.10	118.70	103.93	109.89	118.50	103.03	118.79	115.49	103.93	108.71	118.50
Giants vs. 49ers on 11-12	91.00	102.21	81.52	93.45	97.01	88.72	91.00	109.97	81.52	93.45	96.02	88.72
Packers vs. Seahawks on 11-15	91.12	103.71	107.43	92.54	97.93	105.43	91.12	110.23	107.43	92.54	102.43	105.43
Steelers vs. Jaguars on 11-18	92.96	105.04	94.45	95.26	101.52	82.67	92.96	110.96	94.45	95.26	109.91	82.67
Falcons vs. Saints on 11-22	129.62	137.32	114.12	130.90	134.73	113.06	130.90	146.04	114.12	130.90	145.32	113.06
Redskins vs. Eagles on 12-3	87.68	97.19	81.94	88.70	94.53	82.74	88.60	102.29	82.74	88.70	94.51	82.74
Jaguars vs. Titans on 12-6	78.48	88.04	61.07	79.53	86.61	49.17	78.30	93.68	54.93	79.53	92.93	49.17
AVERAGES	94.20	105.94	101.14	95.84	102.12	97.41	94.65	112.15	100.09	95.84	105.47	97.41

8. Conclusion

In this paper, we introduced a novel class of optimization problems, which can be framed as a stochastic packing problem where each bin is associated with a random variable whose value is defined according to the selected items and the goal is to optimize order statistics on the expected values of these random variables. This work focuses on studying the maximum of two random variables, defined as the sum of other random and potentially correlated variables that follow the normal distribution.

The paper presents results of theoretical and algorithmic relevance for the problem. Namely, we show that the maximization of order statistics with two normally distributed variables is NP-hard even in its simplest setting, where the assignment of items to sets is unconstrained. We also introduce the first exact algorithm to optimize problems whose objective function are specified by order statistics; the proposed solutions is a cutting-plane algorithm that leverages discretization techniques and approximation results tailored for the problem.

The usefulness of the approach is obvious in settings where obtaining one extremely high outcome is more valuable than two reasonably high outcomes. In particular, we apply the approach to the increasingly popular betting contests known as *daily fantasy sports*, where the payoff structure is heavily skewed toward top finishers and all participants are allowed to purchase multiple entries. We show that our algorithm improves on results delivered by a reasonable heuristic that does not explicitly consider the full spectrum of the optimization problem and instead focuses on treating the entries separately. Combining our approach with a specific player-pair covariance structure learned from prior data for selecting the rosters of the two entries leads to significant positive returns for bettors.

Real-world settings of the problem, such as daily fantasy sports, allow for scenarios where three or more sets can be selected. The resulting problems are even challenging from a mathematical perspective, and investigating them is an exciting possibility for an extension in future work. Also, studying discrete optimization in settings where each individual random variable follows alternative distributions than the normal is of interest for future work.

References

Ahmadi MV, Doostparast M, Ahmadi J (2015) Statistical inference for the lifetime performance index based on generalised order statistics from exponential distribution. *Int. J. Systems Science* 46(6):1094–1107, URL http://dx.doi.org/10.1080/00207721.2013.809611.

Ahsanullah M, Nevzorov V (2005) *Order Statistics: Examples and Exercises* (Nova Science Publishers), ISBN 9781594540714, URL https://books.google.com/books?id=UAVv4k3lZasC.

Arnold B, Balakrishnan N (2012) *Relations, Bounds and Approximations for Order Statistics*. Lecture Notes in Statistics (Springer New York), ISBN 9781461236443, URL https://books.google.com/books?id=Tb7kBwAAQBAJ.

Arnold B, Balakrishnan N, Nagaraja H (2008) *A First Course in Order Statistics*. Classics in Applied Mathematics (Society for Industrial and Applied Mathematics), ISBN 9780898716481, URL https://books.google.com/books?id=gUD-S8USlDwC.

Balas E, Jeroslow R (1972) Canonical cuts on the unit hypercube. *SIAM Journal on Applied Mathematics* 23(1):61–69.

Ben-Tal A, Nemirovski A (1997) Stable truss topology design via semidefinite programming. *SIAM Journal on Optimization* 7:991–1016.

Ben-Tal A, Nemirovski A (2002) Robust optimization–methodology and applications. *Mathematical Programming* 92(3):453–480.

Bergman D, Imbrogno J (2017) Surviving a national football league survivor pool. *Operations Research* 65(5):1343–1354.

Bertsimas D, Natarajan K, Teo CP (2006) Tight bounds on expected order statistics. *Probab. Eng. Inf. Sci.* 20(4):667–686, ISSN 0269-9648, URL http://dx.doi.org/10.1017/S0269964806060414.

Bertsimas D, Sim M (2004) The price of robustness. *Operations Research* 52(1):35–53.

Bertsimas D, Sim M (2006) Tractable approximations to robust conic optimization problems. *Mathematical Programming* 107(1–2):5–36.

Birge JR, Louveaux F (1997) *Introduction to Stochastic Programming* (Springer-Verlag).

Birge JR, Louveaux FV (1988) A multicut algorithm for two-stage stochastic linear programs. *European Journal of Operational Research* 34(3):384–392.

Brown J, Minor DB (2014) Selecting the best? spillover and shadows in elimination tournaments. *Management Science* 60(12):3087–3102.

Brown KC, Brown DJ (1986) Using order statistics to estimate real estate bid distributions. *Management science* 32(3):289–297.

Carøe CC, Tind J (1998) L-shaped decomposition of two-stage stochastic programs with integer recourse. *Mathematical Programming* 83(1-3):451–464.

Charnes A, Cooper WW (1959) Chance-constrained programming. *Management science* 6(1):73–79.

Clair B, Letscher D (2007) Optimal strategies for sports betting pools. *Operations Research* 55(6):1163–1177.

Dantzig GB (1955) Linear programming under uncertainty. *Management Science* 1(3-4):197–206.

David H, Nagaraja H (2004) *Order Statistics*. Wiley Series in Probability and Statistics (Wiley), ISBN 9780471654018, URL https://books.google.com/books?id=bdhzFXg6xFkC.

Dean BC, Goemans MX, Vondrák J (2008) Approximating the stochastic knapsack problem: The benefit of adaptivity. *Mathematics of Operations Research* 33(4):945–964.

Dean BC, Goemans MX, Vondrdk J (2004) Approximating the stochastic knapsack problem: The benefit of adaptivity. *45th Annual IEEE Symposium on Foundations of Computer Science*, 208–217 (IEEE).

D'Eramo C, Nuara A, Restelli M (2016) Estimating the maximum expected value through gaussian approximation. *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, 1032–1040, ICML'16 (JMLR.org), URL http://dl.acm.org/citation.cfm?id=3045390.3045500.

Dimitrova DS, Kaishev VK, Ignatov ZG (2018) Ruin and deficit under claim arrivals with the order statistics property. *Methodology and Computing in Applied Probability* URL http://dx.doi.org/10.1007/s11009-018-9669-5, © The Author(s) 2018. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Gade D, Küçükyavuz S, Sen S (2014) Decomposition algorithms with parametric gomory cuts for two-stage stochastic integer programs. *Mathematical Programming* 144(1-2):39–64.

Higle JL, Sen S (1991) Stochastic decomposition: An algorithm for two-stage linear programs with recourse. *Mathematics of operations research* 16(3):650–669.

Hunter DS, Vielma JP, Zaman T (2016) Picking winners in daily fantasy sports using integer programming.

ILOG I (2018) Cplex optimization studio. URL http://www.cplex.com.

Israeli E, Wood RK (2002) Shortest-path network interdiction. *Networks* 40(2):97–111.

Kaplan EH, Garstka SJ (2001) March madness and the office pool. *Management Science* 47(3):369–382.

Kaplan TR, Zamir S (2012) Asymmetric first-price auctions with uniform distributions: analytic solutions to the general case. *Economic Theory* 50(2):269–302, ISSN 1432-0479, URL http://dx.doi.org/10.1007/s00199-010-0563-9.

Kleywegt AJ, Shapiro A, Homem-de Mello T (2002) The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization* 12(2):479–502.

Koutras VM, Koutras MV (2018) Exact distribution of random order statistics and applications in risk management. *Methodology and Computing in Applied Probability* ISSN 1573-7713, URL http://dx.doi.org/10.1007/s11009-018-9662-z.

Ma W (2018) Improvements and generalizations of stochastic knapsack and markovian bandits approximation algorithms. *Mathematics of Operations Research* 43(3):789–812.

McCormick GP (1976) Computability of global solutions to factorable nonconvex programs: Part i — convex underestimating problems. *Mathematical Programming* 10(1):147–175, ISSN 1436-4646, URL http://dx.doi.org/10.1007/BF01580665.

McCormick ST, Rao MR, Rinaldi G (2003) Easy and difficult objective functions for max cut. *Mathematical programming* 94(2-3):459–466.

Nadarajah S, Kotz S (2008a) Exact distribution of the max/min of two gaussian random variables. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on* 16:210 – 212, URL http://dx.doi.org/10.1109/TVLSI.2007.912191.

Nadarajah S, Kotz S (2008b) Exact distribution of the max/min of two gaussian random variables. *IEEE Transactions on very large scale integration (VLSI) systems* 16(2):210–212.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Ramachandran G (1982) Properties of extreme order statistics and their application to fire protection and insurance problems. *Fire Safety Journal* 5(1):59 – 76, ISSN 0379-7112, URL http://dx.doi.org/https://doi.org/10.1016/0379-7112(82)90007-8.

Ross AM (2010) Computing bounds on the expected maximum of correlated normal variables. *Methodology and Computing in Applied Probability* 12(1):111–138, ISSN 1573-7713, URL http://dx.doi.org/10.1007/s11009-008-9097-z.

Sahinidis NV (2004) Optimization under uncertainty: state-of-the-art and opportunities. *Computers & Chemical Engineering* 28(6-7):971–983.

Shapiro A, Homem-de Mello T (1998) A simulation-based approach to two-stage stochastic programming with recourse. *Mathematical Programming* 81(3):301–325.

Sherali HD, Zhu X (2006) On solving discrete two-stage stochastic programs having mixed-integer first-and second-stage variables. *Mathematical programming* 108(2-3):597–616.

Urbaczewski A, Elmore R (2018) Big data, efficient markets, and the end of daily fantasy sports as we know it? *Big Data* 6(4):239–247, URL http://dx.doi.org/10.1089/big.2018.0057, pMID: 30457878.

Yang HC, Alouini MS (2011) *Distributions of order statistics*, 4071 (Cambridge University Press), URL http://dx.doi.org/10.1017/CBO9781139043328.005.