# Drafting the Best Baseball Team: An Integer Programming Problem 

Griffin Holt ${ }^{\dagger}$


#### Abstract

We aim to draft a 25-man roster out of all the available players from the 2019 season of Major League Baseball (MLB) that meets the constraints of a typical MLB baseball team and maximizes the total sum of the $b W A R$ scores of all players on the roster. We formulate this goal as an Integer Programming problem and solve it using both a Real-Valued Relaxation approximation method and the Branch-and-Cut method. The Real-Valued Relaxation algorithm returns almost optimal solutions, differing less than 1 point in the optimal objective values returned by the Branch-and-Cut algorithm. Simulations of 162 -game seasons show that the teams yielded from solving the Integer Programming problem achieve recordbreaking winning percentages-. 846 for a team with a large payroll cap of U.S. $\$ 197,683,216, .716$ for a team with a small payroll cap of U.S. $\mathbf{\$ 2 8 , 2 2 9 , 1 0 8 - t h u s ~ d e m o n s t r a t i n g ~ t h e ~ e f f e c t i v e n e s s ~ o f ~}$ both the $b W A R$ score and the model of Integer Programming as a method for making informed drafting decisions.


Index Terms-sabermetrics, integer programming, $b W A R$

## I. InTRODUCTION

The sport of baseball is perfectly suited for the use of statistics to make informed decisions. The discrete nature of the game's events and the large number of these events that occur during every game lend to the ability of statistics to make accurate predictions and inference about long-term trends within the game. Since the early twentieth century, team managers and baseball eccentrics have employed statistics to attempt to quantify player performance, predict longterm team performance, and improve the quality of teams. The field of "sabermetrics"-baseball statistics-became increasingly popular in the 1980 s with the publications of Bill James [1], [2], [3], [4], and again in the 2000's with the publishing of the book Moneyball [5], and once more in the 2010's with the release of the book's movie adaptation [6]. Since Bill James, there have been even more developments in the use of numbers to inform baseball decisions, including the use of Markov chains [7] and machine learning [8].

Our goal in this paper is to draft a 25 -man roster out of all the available players from the 2019 season of Major League Baseball (MLB) that meets the constraints of a typical MLB baseball team and maximizes the total sum of the $b W A R$ scores of all players on the roster. ${ }^{1}$

[^0]There are generally accepted best-practices for how to allocate each spot on the 25 -man roster to different positions [11]: typically, a $25-\mathrm{man}$ roster will consist of:

1) 5 starting pitchers (SP);
2) 7 relief pitchers (RP);
3) 2 catchers (C);
4) 6 infielders (IF); and
5) 5 outfielders (OF).

We also need at least 1 player specialized at each of the infield positions: first baseman (1B), second baseman (2B), shortstop (SS), and third baseman (3B). For this paper, we will ignore the position of designated hitter ( DH ) present in the American League.

It is also advantageous to have a certain number of lefthanded batters and left-handed pitchers on a team. Batters tend to perform better against opposite-handed pitchers (i.e., a left-handed batter performs better, on average, against a right-handed pitcher than a left-handed pitcher) [12]. Thus, every MLB team strives to have at least a few left-handed pitchers and batters on their team.

## A. The bW AR Statistic

Wins Above Replacement ( $W A R$ ) is a baseball statistic designed to explain the number of additional wins a player contributes to his team above the expected number of wins contributed by a replacement-level or average player [13]. A $W A R$ score of 0 means that a player contributes no additional wins above the expected number of wins to be contributed by an average player in the league. $W A R$ scores for batters and pitchers are scaled equally, so as to compare their relative contributions to a team. The statistic is nonstandardized; that is, different sources compute the value of a player's $W A R$ score differently. In this paper, we utilize the computation provided by Baseball-Reference [13]; we refer to this specific definition of the $W A R$ score as the $b W A R$ score.

For batters, the $b W A R$ statistic is computed using six different components:

1) batting runs: runs scored as a direct result of a hit;
2) baserunning runs: runs scored by other means (e.g., base-stealing, tagging up after fly balls, etc.);
3) runs added or lost from hitting a ground ball into a double play;
4) fielding runs: opposite team runs prevented by a player's fielding performance;
5) positional adjustment runs: a measure of runs normalized by a player's position; and
6) replacement-level runs: runs scored above the average player (based on playing time).

The computations involving these six different components are quite complex, and explaining them in the paper would detract from this paper's purpose. For a more detailed explanation of the computation, see [14].

The $b W A R$ statistic for pitchers is computed using only two components:

1) Runs Allowed (RA), both earned and unearned; and
2) Innings Pitched (IP).

The computations involving these two different components are also complex, and explaining them in the paper would detract from this paper's purpose. For a more detailed explanation of the computation, see [15].

## II. Problem Formulation

Let $i=1, \ldots, 801$ be the 801 different baseball players in the MLB a) that have more than 10 game appearances (GAs) in a single position and b) whose 2019 yearly salaries are publicly available. Let $x_{i} \in\{0,1\}$ represent whether player $i$ is included on our baseball team (with 1 meaning player $i$ is included on the team, and 0 means the player is not on the team). Let $b W A R_{i}$ be the $b W A R$ score of player $i$.

Let $P_{S}$ be the set of starting pitchers out of the 801 possible baseball players. Similarly, let $P_{R}$ be the set of relief pitchers, $C$ be the set of catchers, $I$ be the set of infielders, $I_{1 B}$ be the set of first basemen, $I_{2 B}$ be the set of second basemen, $I_{S S}$ be the set of shortstops, $I_{3 B}$ be the set of third basemen, and $O$ be the set of outfielders.

Let $B_{l}$ be the set of left-handed batters and let $P_{l}$ be the set of left-handed pitchers. For this paper, we constrain the problem so that we have at least 3 left-handed batters and 4 left-handed pitchers. As stated earlier in the paper, it is advantageous to have a certain number of left-handed batters and left-handed pitchers on a team.

Let $s_{i}$ be the 2019 salary (in U.S.\$) of player $i$. To explore the effect of salary constraints on the draft, we also introduce five different payroll caps (in U.S.\$), to be introduced separately:

1) $S_{\text {None }}=\infty$ (no payroll cap; that is, unlimited funds);
2) $S_{\mathrm{Max}}=197,683,216$ (the maximum team payroll in 2019, belonging to the Chicago Cubs);
3) $S_{\text {Mean }}=100,487,138$ (the mean team payroll in 2019);
4) $S_{\text {Median }}=85,304,101$ (the median team payroll in 2019); and
5) $S_{\text {Min }}=28,229,108$ (the minimum team payroll in 2019, belonging to the Pittsburgh Pirates).

Thus, we actually aim to solve five different Integer Programming (IP) problems, formulated as

$$
\begin{align*}
& \max _{x} \sum_{i=1}^{801} b W A R_{i} x_{i}  \tag{1}\\
& \text { subject to: } \sum_{i=1}^{801} x_{i}=25  \tag{2}\\
& \sum_{i \in P_{S}} x_{i}=5  \tag{3}\\
& \sum_{i \in P_{R}} x_{i}=7  \tag{4}\\
& \sum_{i \in C} x_{i}=2  \tag{5}\\
& \sum_{i \in I} x_{i}=6  \tag{6}\\
& 1 \leq \sum_{i \in I_{1 B}} x_{i} \leq 3  \tag{7}\\
& 1 \leq \sum_{i \in I_{2 B}} x_{i} \leq 3  \tag{8}\\
& 1 \leq \sum_{i \in I_{S S}} x_{i} \leq 3  \tag{9}\\
& 1 \leq \sum_{i \in I_{3 B}} x_{i} \leq 3  \tag{10}\\
& 3 \leq \sum_{i \in B_{l}} x_{i} \leq 25  \tag{11}\\
& 4 \leq \sum_{i \in P_{l}} x_{i} \leq 12  \tag{12}\\
& \sum_{i=1}^{801} s_{i} x_{i} \leq S_{j}  \tag{13}\\
& x_{i} \in\{0,1\}, i=1, \ldots, 801 \tag{14}
\end{align*}
$$

for $j=$ None, Max, Mean, Median, Min.
We will refer to the value of the objective function $\sum_{i=1}^{801} b W A R_{i} x_{i}$ for a specific team as the "utility" of that team.

## III. Methodology

## A. Data Acquisition and Preparation

All of the data used to solve this problem comes from three sources:

1) The Lahman Database [16]: for a player's game appearances, handedness, and position;
2) Baseball-Reference [17]: for a player's salary and $b W A R$ score; and
3) Spotrac [18]: for each MLB team's 2019 payroll.

For players that appeared on different teams during the 2019 season (i.e., they were traded mid-season), we calculated their season-long $b W A R$ score as a weighted average of their separate $b W A R$ scores for each team, with the weight depending on how many game appearances the player had on that team.

|  |  | Real-Valued Relaxation |  |  | Branch-and-Cut |  |  |
| :---: | :---: | ---: | ---: | ---: | ---: | ---: | ---: |
| Team | Payroll Cap | Utility | Payroll | Cost/bWAR |  | Utility | Payroll | Cost/bWAR

TABLE I
Optimal Utilities, total Payrolls, and Cost (U.S.\$) per 1 unit of $b W A R$ For the solutions to the five IP problems

We marked a player as able to play a certain position $-i \in$ $C, I_{1 B}, I_{2 B}, I_{S S}, I_{3 B}, O$, respectively-if the player appeared in more than 10 games at that position; note that these sets are not necessarily disjoint.

We classified a pitcher as a starting pitcher $i \in P_{S}$ if he started more games than he finished; otherwise, we classified the pitcher as a relief pitcher $i \in P_{R}$.

## B. Solving the Integer Programming Problems

Generally, Integer Programming problems belong to the set of NP-complete problems. There exist no known polynomial-time algorithms for solving such problems. Thus, large NP-complete problems are often intractable and cannot be solved by a computer within any reasonable amount of time. There are a number of known approximation methods for solving Integer Programming problems to arrive at nearoptimal solutions in polynomial time [19]. There are also a number of known intelligent search methods for finding the optimal solution to an Integer Programming problem [19], [20]; although such methods do not run in polynomial time, they may run quickly for Integer Programming problems with a small number of decision variables and constraints.

We used two methods, one approximation method and one intelligent search method, to solve the five Integer Programming problems described in Section II.

First, we aimed to find near-optimal solutions to the problems using Real-Valued Relaxation [21]. We relaxed the integer constraint $x_{i} \in\{0,1\}$ so that the decision variables can be real-valued:

$$
\begin{equation*}
0 \leq x_{i} \leq 1, i=1, \ldots, 801 \tag{16}
\end{equation*}
$$

Then, the problems are reduced to simple standard Linear Programming (LP) problems. We solved these relaxed problems using the LP solver provided by SciPy [22]. For a single relaxed problem, the LP solver returns values of the decision variables $x_{i}, i=1, \ldots, 801$. For our specific problems, the majority of these variables were already 0 - or 1 -valued; a few, however, were fractional and remained between 0 and 1 . To arrive at a near-optimal team, we selected the combination of these fraction-valued decision variables that result in a feasible and maximal solution (with respect to the other combinations).

Second, we aimed to find exactly optimal solutions to the Integer Programming problems using the Branch-andCut method [23]. To accomplish this, we used the Gurobi

| Team | Utility | Payroll | Cost/bWAR |
| :---: | ---: | :---: | :---: |
| Astros | 60.57 | $\$ 155,036,881$ | $\$ 2,559,736.15$ |
| Dodgers | 41.54 | $\$ 157,394,622$ | $\$ 3,789,199.96$ |
| Yankees | 41.58 | $\$ 171,935,047$ | $\$ 4,135,042.02$ |

TABLE II
Utility, Payroll, and Cost (U.S.\$) per 1 unit of $b W A R$ For the top three teams in the MLB 2019 SEASON

Optimizer [24], which uses the Branch-and-Cut method to solve Integer Programming problems.

## IV. Results and Discussion

The results of our Real-Valued Relaxation method and the Branch-and-Cut method for all five Integer Programming problems are displayed in Table I. This table displays the maximized (or nearly maximized) Utility, total Payroll, and cost in U.S.\$ per 1 unit of $b W A R$ for the teams returned by the Real-Valued Relaxation algorithm and the Branch-and-Cut algorithm for the five Integer Programming problems, each characterized by a different payroll cap: $S_{\text {None }}, S_{\text {Max }}, S_{\text {Mean }}, S_{\text {Median }}, S_{\text {Min }}$. Note that the Utility score for the Real-Valued Relaxation problem represents the Utility of the team once the integer constraint was re-enforced on the problem; it is not the Utility returned by the relaxed LP problem.
In Table II, we provide for comparison the Utility, Payroll, and Cost for 1 unit of $b W A R$ for the three teams that had the highest win percentage in the MLB 2019 Season: the Houston Astros, the Los Angeles Dodgers, and the New York Yankees.

The optimal solutions to the IP problem characterized by the payroll constraints $S_{\text {None }}$ and $S_{\mathrm{Max}}$ are the same; that is, the optimal team with no payroll constraint is the same optimal team for the maximum payroll constraint. However, the other three IP problems each had unique solutions due to the lowered payroll constraints.

The Real-Valued Relaxation algorithm returned the actual optimal team (returned by Branch-and-Cut) for the two problems characterized by the payroll constraints $S_{\text {None }}$ and $S_{\text {Max }}$. Also, although the Real-Valued Relaxation algorithm returned less-than-optimal solutions for the three problems characterized by the mean, median, and minimum payroll

| Name | MLB Team | Position | Bats | Pitches | Salary | bWAR |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Mitch Garver | MIN | C | R | - | \$575,000 | 4.14 |
| J.T. Realmuto | PHI | C | R | - | \$5,900,000 | 4.56 |
| Pete Alonso | NYM | 1B | R | - | \$555,000 | 5.45 |
| D.J. LeMahieu | NYY | 2B | R | - | \$12,000,000 | 5.60 |
| Marcus Semien | OAK | SS | R | - | \$5,900,000 | 8.36 |
| Trevor Story | COL | SS | R | - | \$5,000,000 | 6.93 |
| Alex Bregman | HOU | 3B | R | - | \$640,500 | 8.96 |
| Matt Chapman | OAK | 3B | R | - | \$580,000 | 7.71 |
| Mookie Betts | BOS | OF | R | - | \$20,000,000 | 7.26 |
| Aaron Judge | NYY | OF | R | - | \$684,300 | 5.61 |
| George Springer | HOU | OF | R | - | \$12,000,000 | 6.45 |
| Mike Trout | LAA | OF | R | - | \$36,833,333 | 7.89 |
| Christian Yelich | MIL | OF | L | - | \$9,750,000 | 6.95 |
| Gerrit Cole | HOU | SP | R | R | \$13,500,000 | 6.62 |
| Jacob deGrom | NYM | SP | L | R | \$9,000,000 | 8.42 |
| Lance Lynn | TEX | SP | L | R | \$9,333,333 | 7.51 |
| Mike Minor | TEX | SP | R | L | \$9,833,333 | 7.79 |
| Justin Verlander | HOU | SP | R | R | \$28,000,000 | 7.35 |
| Josh Hader | MIL | RP | L | L | \$687,600 | 2.66 |
| Liam Hendriks | OAK | RP | R | R | \$2,150,000 | 3.49 |
| Felipe Vazquez | PIT | RP | L | L | \$4,500,000 | 2.96 |
| Hansel Robles | LAA | RP | R | R | \$1,400,000 | 2.62 |
| Taylor Rogers | MIN | RP | L | L | \$1,525,000 | 2.45 |
| Brandon Workman | BOS | RP | R | R | \$1,150,000 | 3.23 |
| Kirby Yates | SD | RP | L | R | \$3,062,000 | 2.87 |

TABLE III
THE OPTIMAL TEAM FOR THE MAXIMUM PAYROLL CAP IP PROBLEM; AKA "TEAM MAX CAP"

| Name | MLB Team | Position | Bats | Pitches | Salary | bWAR |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Willson Contreras | CHC | C | R | - | \$684,000 | 3.18 |
| Mitch Garver | MIN | C | R | - | \$575,000 | 4.14 |
| Pete Alonso | NYM | 1B | R | - | \$555,000 | 5.45 |
| Max Muncy | LAD | 2B | L | - | \$575,000 | 5.28 |
| Marcus Semien | OAK | SS | R | - | \$5,900,000 | 8.36 |
| Trevor Story | COL | SS | R | - | \$5,000,000 | 6.93 |
| Alex Bregman | HOU | 3B | R | - | \$640,500 | 8.96 |
| Matt Chapman | OAK | 3B | R | - | \$580,000 | 7.71 |
| Ronald Acuna Jr. | ATL | OF | R | - | \$560,000 | 5.08 |
| Aaron Judge | NYY | OF | R | - | \$684,300 | 5.61 |
| Austin Meadows | TB | OF | L | - | \$557,400 | 3.94 |
| Victor Robles | WSH | OF | R | - | \$557,800 | 4.36 |
| Juan Soto | WSH | OF | L | - | \$578,300 | 5.02 |
| Shane Bieber | CLE | SP | R | R | \$559,600 | 4.66 |
| Luis Castillo | CIN | SP | R | R | \$557,500 | 4.53 |
| Jack Flaherty | STL | SP | R | R | \$562,100 | 5.57 |
| Lucas Giolito | CWS | SP | R | R | \$573,000 | 5.81 |
| John Means | BAL | SP | L | L | \$555,500 | 4.87 |
| Josh Hader | MIL | RP | L | L | \$687,600 | 2.66 |
| Liam Hendriks | OAK | RP | R | R | \$2,150,000 | 3.49 |
| Francisco Liriano | PIT | RP | L | L | \$100,000 | 1.19 |
| Seth Lugo | NYM | RP | R | R | \$591,875 | 2.45 |
| Hansel Robles | LAA | RP | R | R | \$1,400,000 | 2.62 |
| Taylor Rogers | MIN | RP | L | L | \$1,525,000 | 2.45 |
| Brandon Workman | BOS | RP | R | R | \$1,150,000 | 3.23 |

TABLE IV
The optimal team for the minimum payroll cap ip problem; aka "Team Min Cap"

| Opponent | Team Max Cap |  |  |  |  |  |  |  | Team Min Cap |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | W | L | Pct | R | RA | H | HA | E | W | L | Pct | R | RA | H | HA | E |
| Astros | 20 | 10 | . 667 | 154 | 119 | 223 | 205 | 14 | 19 | 11 | . 633 | 155 | 128 | 228 | 247 | 11 |
| Dodgers | 20 | 10 | . 667 | 137 | 93 | 244 | 169 | 7 | 20 | 10 | . 667 | 156 | 129 | 264 | 255 | 9 |
| Yankees | 24 | 6 | 0.8 | 289 | 121 | 360 | 207 | 6 | 18 | 12 | 0.6 | 246 | 184 | 324 | 294 | 7 |

TABLE V
Results of Simulated 30-Game Series against Top Three MLB Teams in 2019
caps, the near-optimal teams each had a utility within 1 $b W A R$ of the optimal utility.

Also notable is the fact that the optimal utility 117.55 of the team with even the lowest payroll cap $S_{\text {Min }}$ had a higher utility score than all three of the top teams in the MLB 2019 season.

## A. A Closer Look at Two Optimal Teams

Let us refer to the optimal team found by the Branch-and-Cut algorithm for the IP problem characterized by a payroll cap of $S_{\text {Max }}$ as "Team Max Cap". Similarly, let us refer to the optimal team found by Branch-and-Cut for the IP problem characterized by a payroll cap of $S_{\text {Min }}$ as "Team Min Cap". The chosen players of these two teams, along with information about each player, are displayed in Tables III and IV, respectively.

On average, a player on Team Max Cap has played in the MLB for 5.2 years by 2019, whereas a player on Team Min Cap has only played for 3.2 years. Team Min Cap is mostly comprised of players that have played between 0 3 years and thus, despite however good they may be at the game, they have restricted salaries as beginner players. Also, 18 out of 25 players on Team Max Cap made the MLB AllStar Team in 2019; 14 out of 25 players on Team Min Cap made the MLB All-Star Team. Finally, both teams would have increased in worth since 2019 (according to the most recent contracts of the players). Although Team Min Cap would have cost only U.S. $\$ 27.9$ million in 2019, the same team would cost U.S. $\$ 146.1$ million in 2021: this is due to the fact that many of these players completed their rookie contracts by 2021 and have since been picked up on larger contracts because of their talent. Team Max Cap, costing U.S. $\$ 194.6$ million in 2019, would increase to U.S. $\$ 361.8$ million in 2021-a payroll way beyond any that has ever existed in Major League Baseball.

## B. Simulations

To somewhat verify the quality of Team Max Cap and Team Min Cap-produced by the solutions to their respective IP problems-we decided to run game simulations using the MLB SimMatchup software created by WhatIfSports [25]. The software takes a 25 -man roster as an input and can simulate a game between the input roster and any historical baseball team as far back as 1885 . We ran four sets of simulations for each team:

1) a 30-game series ( 15 home, 15 away) against the 2019 Houston Astros;
2) a 30-game series ( 15 home, 15 away) against the 2019 Los Angeles Dodgers;
3) a 30-game series ( 15 home, 15 away) against the 2019 New York Yankees; and
4) a full 162-game season using the MLB teams from 2019.

Statistical summaries of the results of the three sets of 30-game series for Team Max Cap and Team Min Cap are presented in Table V. Notice that both Team Max Cap and Team Min Cap had winning records against all three teams.

To simulate the 2019 162-game season for one of our teams (Team Max Cap or Team Min Cap), we removed the San Francisco Giants from the set of 30 team in the MLB and replaced them with our team. We gave our team the same schedule as the San Francisco Giants had in the 2019 season, and then simulated each of the 162 games in this schedule using the MLB SimMatchup software.

A statistical summary of the results of these two simulated seasons and the actual 2019 seasons for the top three teams that year in the MLB are presented in Table VI. The results of each game individually are presented graphically in Figure 1. Both Team Max Cap and Team Min Cap finished with the best winning percentages of the season (.846 and .716 both exceed the highest actual winning percentage of that year: .660, belonging to the Houston Astros). In addition, Team Max Cap beat and Team Min Cap tied the record for most MLB wins in a single season: 116 wins, held by both the 2001 Seattle Mariners and the 1906 Chicago Cubs [26]. Also remarkable are the high number of runs ( R ), low number of runs allowed (RA), high number of hits (H), low number of hits allowed (HA), and low number of errors (E) achieved by both teams, especially in comparison to the same statistics for the top three teams in 2019.

## V. Conclusions and Future Work

Overall, it should be clear that modeling a baseball draft of the best baseball team as an Integer Programming problem and using $b W A R$ as a utility score were successful endeavors. When our optimally drafted teams were pitted against real teams in a simulation, our teams outperformed the real teams drastically. The team produced by solving the maximum payroll cap problem-"Team Max Cap"-scored over 300 runs more than the Houston Astros during its simulated 2019 season. This is especially notable as this was the year that the Houston Astros won the World Series, and, notoriously, it was later revealed that they had been cheating the entire 2019 season by stealing pitch signs [27]; stealing

| Team | W | L | Pct | R | RA | H | HA | E |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Max Cap | 137 | 25 | .846 | 1234 | 518 | 1626 | 1009 | 37 |
| Min Cap | 116 | 46 | .716 | 1032 | 641 | 1424 | 1316 | 49 |
| Astros | 107 | 55 | .660 | 920 | 640 | 1538 | 1205 | 71 |
| Dodgers | 106 | 56 | .654 | 886 | 613 | 1414 | 1201 | 106 |
| Yankees | 103 | 59 | .636 | 943 | 739 | 1493 | 1374 | 102 |

TABLE VI
Results of Simulated and Actual 2019 162-Game Seasons

(a) Team Max Cap

(b) Team Min Cap

(c) Houston Astros

(d) Los Angeles Dodgers

(e) New York Yankees

Fig. 1. The game results of each of the 162 games in the simulated 2019 seasons for Team Max Cap and Team Min Cap, and the actual 2019 seasons for the Houston Astros, the Los Angeles Dodgers, and the New York Yankees. Each bar represents a different game. Green bars represent wins; red bars represent losses. The magnitude of the bar is the run difference between the opposing teams (e.g., a large green bar represents a win by a large margin).
pitch signs inflates the number of hits and runs that a team achieves as it allows the batter to know the type of pitch that is about to be thrown. Also notable is the fact that even the optimal minimum payroll team outperformed the Houston Astros.

Because they belong to the class of NP-Complete problems, large Integer Programming problems can take an immense amount of time to solve. However, in our specific case, our Integer Programming problems only had 801 decision variables and 19 constraints. As a result, the Gurobi Integer Program Solver was able to arrive at a solution for all five of our problems quickly, each in less than a second. Thus, for drafting problems of this kind, and even those slightly larger, Integer Programming represents a quick way to model and solve these problems.

It is also worth noting how effective the $b W A R$ statistic was in serving as a utility score. There is much debate surrounding the truthfulness of a $b W A R$ statistic in describing a player's contribution to his team [28]. However, it does appear that $b W A R$ was effective enough in producing a team composed of players from 2019 that outperformed all other teams from that same year.

## A. Future Work

One of the noticeable flaws in the way that we have modeled these Integer Programming problems is the unrealistic composition of the teams produced by solving them. Although its players possess a wide range of experience, the optimal maximum payroll team ("Team Max Cap") represents a rare combination of elite players. Each of these players has contracts of differing lengths and for all contracts to line up for the players to join the same team at the same time would require near impossible timing. In addition, the fact that no team has ever come close to achieving the number of wins achieved by this team seems to suggest the near impossibility of the existence of such a team. The cost of such a team is also so heavy that very few MLB franchises would even be able to afford it.

The optimal minimum payroll team ("Team Min Cap"), although more reasonable in price, is possibly even more improbable than the optimal maximum payroll team. The team is almost entirely composed of rookie or beginning players that have been in the MLB for 3 years or less. Unless a MLB franchise were to fire all of its existing players and start completely from scratch, the concentration of so many rookies on a single team could not happen. In addition, because of the structure of the MLB Draft, it would be unlikely for so many high quality rookies-recall that 14 of the players on Team Min Cap made the All-Star team in 2019-to be on the same team.

However, the use of Integer Programming as a drafting technique could be adapted to produce more realistic teams. Additional constraints might be added to the problem in order to enforce a greater range of experience among the players. It is also possible that franchises could use Integer Programming to inform their trading decisions during the off-season. Every year, player contracts expire and team
managers have to make decisions about whom to bring onto the team. Such a dilemma could be framed as an Integer Programming problem similar to our own; however, instead of seeking to fill an entire 25 -man roster, the problem would only seek to fill the available positions out of the pool of all currently contract-less players.

I would propose that future researchers in this area experiment with Integer Programming problems for multiple teams in Major League Baseball for the purpose of instructing trade and drafting decisions. It would also be useful to consider the employment of multi-agent systems theory and game theory to approach the dilemma of how all 30 MLB franchises might be attempting to optimize their teams by competitively selecting from the same pool of candidates. In either case, the use of Integer Programming and the use of $b W A R$ both seem to be beneficial in the problem of drafting an optimized baseball team.

## References

[1] B. James, "Baseball abstract, self-published," Lawrence, KS, 1981.
[2] ——, "The 1984 baseball abstract," 1984.
[3] __, The Bill James Baseball Abstract 1987. Ballantine Books, 1987.
[4] _-, The Bill James Historical Baseball Abstract. Random House Incorporated, 1988.
[5] M. Lewis, Moneyball: The Art of Winning an Unfair Game, ser. Norton paperback. W.W. Norton, 2003. [Online]. Available: https://books.google.com/books?id=RWOX_2eYPcAC
[6] D. M. Luca, R. Horovitz, B. Pitt, S. Zaillian, A. Sorkin, and S. Chervin, Moneyball. Sony Pictures Entertainment, 2011.
[7] B. Bukiet, E. R. Harold, and J. L. Palacios, "A markov chain approach to baseball," Operations Research, vol. 45, no. 1, pp. 14-23, 1997.
[8] K. Koseler and M. Stephan, "Machine learning applications in baseball: A systematic literature review," Applied Artificial Intelligence, vol. 31, no. 9-10, pp. 745-763, 2017.
[9] D. Adler, "Team rosters can stay at 28 through season," Aug 2020. [Online]. Available: https://www.mlb.com/news/mlb-teams-28-man-rosters-through-2020-season
[10] "26-man roster: Glossary," 2022. [Online]. Available: https://www.mlb.com/glossary/transactions/26-man-roster
[11] I. Baseball Almanac, "Major league baseball roster history," 2022. [Online]. Available: https://www.baseballalmanac.com/articles/baseball_rosters.shtml
[12] D. Adler, "Why being a lefty matters in baseball", Jun 2019. [Online]. Available: https://www.mlb.com/news/why-left-handed-pitching-matters-in-baseball
[13] "War explained." [Online]. Available: https://www.baseballreference.com/about/war_explained.shtml
[14] "Position player war calculations and details." [Online]. Available: https://www.baseballreference.com/about/war_explained_position.shtml
[15] "Pitcher war calculations and details." [Online]. Available: https://www.baseball-reference.com/about/war_explained_pitch.shtml
[16] S. Lahman, "The lahman baseball database 1871-2021," Mar 2022. [Online]. Available: https://www.seanlahman.com/baseballarchive/statistics/
[17] 2022. [Online]. Available: https://www.baseball-reference.com/
[18] "Mlb 2019 payroll tracker." [Online]. Available: https://www.spotrac.com/mlb/payroll/2019/
[19] S. Dasgupta, C. Papadimitriou, and U. Vazirani, Algorithms. McGraw-Hill Higher Education, 2006. [Online]. Available: https://books.google.com/books?id=DJSUCgAAQBAJ
[20] R. Vanderbei, Linear Programming: Foundations and Extensions, ser. International Series in Operations Research \& Management Science. Springer US, 2013. [Online]. Available: https://books.google.com/books?id=HjznBwAAQBAJ
[21] S. Agmon, "The relaxation method for linear inequalities," Canadian Journal of Mathematics, vol. 6, p. 382-392, 1954.
[22] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," Nature Methods, vol. 17, pp. 261-272, 2020.
[23] M. Padberg and G. Rinaldi, "A branch-and-cut algorithm for the resolution of large-scale symmetric traveling salesman problems," SIAM Review, vol. 33, no. 1, pp. 60-100, 1991. [Online]. Available: https://doi.org/10.1137/1033004
[24] Gurobi Optimization, LLC, "Gurobi Optimizer Reference Manual," 2022. [Online]. Available: https://www.gurobi.com
[25] WhatIfSports, "Mlb simmatchup," Dec 2021. [Online]. Available: https://www.whatifsports.com/mlb/
[26] D.-H. P. Chad Thornburg, "Looking at baseball's winningest seasons," Dec 2021. [Online]. Available: https://www.mlb.com/news/most-mlb-wins-in-a-season-c289159676
[27] J. Robert D. Manfred, "Statement of the commissioner," Jan 2020. [Online]. Available: https://img.mlbstatic.com/mlbimages/image/upload/mlb/cglrhmlrwwbkacty2717.pdf
[28] R. McDaniel, "What is war good for?" Aug 2012. [Online]. Available: https://tht.fangraphs.com/what-is-war-good-for/


[^0]:    ${ }^{\dagger}$ C S 412: Linear Programming / Convex Optimization, Department of Computer Science, Brigham Young University, Provo, Utah 84602 griffinbholt@gmail.com
    ${ }^{1}$ Note that 2019 is the last season with an active roster size of 25 players. The MLB expanded the active roster size temporarily to 28 in 2020 [9] and then permanently to 26 in 2021 [10]. We use a roster size of 25 players in order for our results to be compatible with baseball game simulation software which has yet to be updated to accommodate the 26-man roster. This is also why we use player data from 2019 instead of a later season.

