

# Optimizing Baseball Fielder Positioning with Consideration for Adaptable Hitters

William Melville, Samuel Wise, Grant Nielson, Tristan Mott,  
Christopher Archibald, David Grimsman

## 1. Introduction

In baseball, it is the responsibility of a team's defense to prevent runs by getting opposing hitters out. One prominent strategic decision a defense must make is how to position their seven fielders behind the pitcher and catcher. In this paper, we present a novel approach to optimally position fielders individualized for each opposing hitter. We find evidence that our approach effectively increases outs and decreases runs allowed compared to MLB average defenses.

A fundamental requirement of our approach is having batter-specific distributions over the trajectories of batted balls. Section 2 describes how these distributions are estimated from data using a Bayesian hierarchical model of the joint probability distribution over three parameters describing the trajectory of a batted ball: horizontal spray angle, vertical launch angle, and exit speed. Section 3 describes our novel approach to optimizing fielder positioning, given the batter-specific batted ball distributions. Let  $p_s$  be a positioning strategy in the set of admissible positioning strategies  $P_s \subset \mathbb{R}^{14}$ . We consider two choices of objective functions. First, we can maximize the expected outs over a batter's batted ball distribution by identifying a  $p_s^*$  that satisfies

$$p_s^* \in \operatorname{argmax}_{p_s \in P_s} \mathbb{E}_{b \sim d} [f(b, p_s)] \quad (1)$$

where  $f$  is a model that gives the probability of an out given the trajectory  $b$  of a batted ball and a positioning strategy  $p_s$ , and where  $b$  is drawn from  $d$ , one of the batter-specific batted ball distributions described in section 2.

Alternatively, we can choose to minimize the expected runs allowed over a batter's batted ball distribution by identifying a  $p_s^*$  that satisfies

$$p_s^* \in \operatorname{argmin}_{p_s \in P_s} \mathbb{E}_{b \sim d} [(1 - f(b, p_s)) \cdot f_{\text{woba}}(b)] \quad (2)$$

where  $f_{\text{woba}}$  is a model giving the expected weighted on base average (wOBA) of a batted ball under the assumption that the ball is not fielded for an out. wOBA is related to runs by an affine relationship, so minimizing wOBA is equivalent to minimizing runs [29]. The  $f$  and  $f_{\text{woba}}$  models are described in detail in section 3.

Our approach differs from existing positioning strategies that optimize over a discrete positioning space [15, 17, 27] because our admissible set  $P_s$  is continuous. Our approach also differs from methods that choose to optimize spatially [16, 17, 26, 27] because we optimize to either maximize outs or minimize runs.

A common assumption with existing positioning optimization methods (including (1) and (2)) is that a defense can change its positioning and expect the batter to continue to hit according to their historical tendencies. Critics of extreme defensive shifts have raised the concern that hitters can observe the positions of their opponents, and adaptable hitters may change their approach at the plate in response [13]. To address these concerns, in section 4 we model the positioning problem as a zero-sum sequential game between the defense and the batter. The defense first chooses its positioning strategy. The batter observes the selected positioning strategy and chooses a batted ball distribution,  $d$ , from a set of possible batted ball distributions,  $\mathcal{D}$ . The defense and batter have opposing goals: to either maximize/minimize expected outs or minimize/maximize expected runs, depending on the choice of objective.

For the expected outs objective, an equilibrium pair of strategies  $(p_s^*, d^*)$  must satisfy

$$p_s^* \in \operatorname{argmax}_{p_s \in P_s} \min_{d \in \mathcal{D}} \mathbb{E}_{b \sim d} [f(b, p_s)] \quad (3a)$$

$$d^* \in \operatorname{argmin}_{d \in \mathcal{D}} \max_{p_s \in P_s} \mathbb{E}_{b \sim d} [f(b, p_s)] \quad (3b)$$

where, as before,  $b$  is the trajectory of a batted ball drawn from the batted ball distribution  $d$ .

When the objective is expected runs, an equilibrium pair of strategies  $(p_s^*, d^*)$  must satisfy

$$p_s^* \in \operatorname{argmin}_{p_s \in P_s} \max_{d \in \mathcal{D}} \mathbb{E}_{b \sim d} [(1 - f(b, p_s)) \cdot f_{\text{woba}}(b)] \quad (4a)$$

$$d^* \in \operatorname{argmax}_{d \in \mathcal{D}} \min_{p_s \in P_s} \mathbb{E}_{b \sim d} [(1 - f(b, p_s)) \cdot f_{\text{woba}}(b)] \quad (4b)$$

Note the difference between (3) and (4). In (3) the defense is the maximizing player and the batter is the minimizing player because the defense wants to maximize outs and the batter wants to minimize outs. In (4), these roles reverse. The batter becomes the maximizing player and the defense the minimizing player because the batter wants to maximize runs and the defense wants to minimize runs.

Examples of possible batter action sets  $\mathcal{D}$  are provided in section 4. A trivial example is when the set  $\mathcal{D}$  consists only of the batter-specific batted ball distribution estimated using the method described in section 2. When that is the case, (3) is equivalent to (1) and (4) is equivalent to (2). These equilibrium positioning strategies provide principled solutions to the challenge of facing an adaptable batter.

## 2. Learning Batter-Specific Batted Ball Trajectory Distributions

The out model  $f$  is a key component of both objective functions (1) and (2). The likelihood of an out on any batted ball depends on its trajectory relative to the position of the fielders. The trajectory of a batted ball can be approximated from its spray angle, launch angle, and exit speed. In this section we learn batter-specific joint probability distributions of batted ball spray angles, launch angles, and exit speeds from historical data. Note that some research has already been done to model the spray angle distribution [23, 25] as well as the joint distribution of launch angle and exit speed [28]. As our approach requires the joint distribution of all three variables, we create our own model.

Denote  $h$ ,  $v$ , and  $s$  as the horizontal spray angle, vertical launch angle, and exit speed of a batted ball, respectively. Our goal is to estimate the joint distribution  $p(h, v, s)$ . This will be done using MLB ball tracking data from 2018-2022 from Statcast, which was accessed using the pybaseball package [8, 10, 22]. This data provides the exit speed, launch angle, and the hit coordinates of batted balls. Horizontal spray angle  $h$  is defined using the hit coordinates with the formula described in [12]: balls hit down the third baseline have a spray angle of  $-45^\circ$  and balls hit down the first baseline have a spray angle of  $45^\circ$ . A pulled batted ball is defined as one with a negative spray angle for right-handed batters and positive spray angle for left-handed batters. Any other batted ball is called an opposite field (oppo) ball.

To most efficiently estimate the desired joint distribution, we first need to determine if there are any independencies between the variables that can be leveraged. This is done by examining the distributions over one of the variables, given different assignments to another variable. First, figure 1 shows two histograms of batted ball launch angles partitioned by pulled or oppo spray angle. Based on this figure, we make the assumption that launch angle is dependent on spray angle.

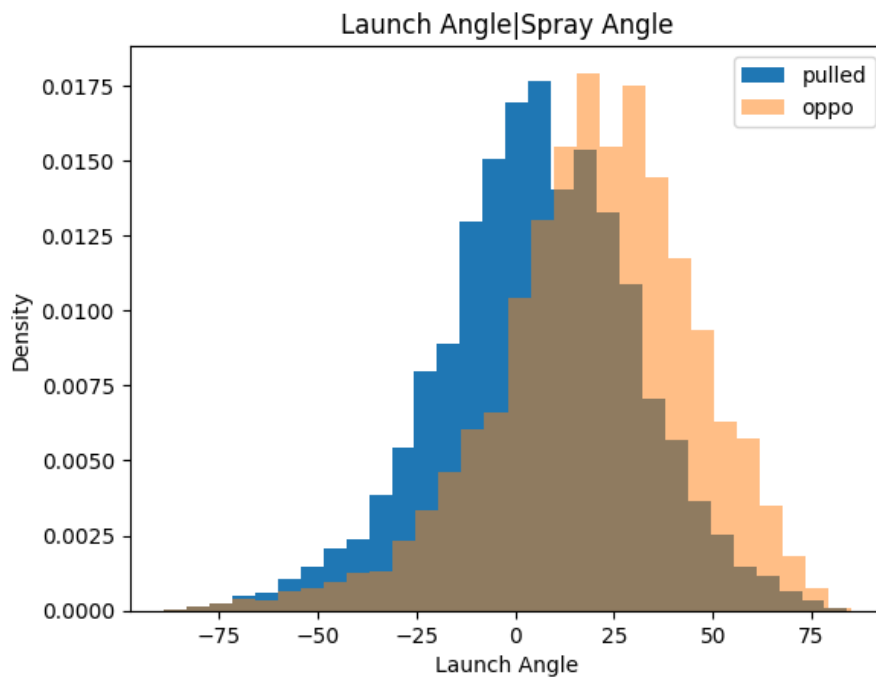


Figure 1: Distributions of batted ball launch angles partitioned by pulled or oppo spray angle, giving evidence that launch angle and spray angle are not independent.

Next, the relationship between exit speed and the two other variables is explored. We define four different batted ball launch angle types: ground balls ( $v \leq 10$ ), line drives ( $10 < v \leq 25$ ), fly balls ( $25 < v \leq 50$ ), and pop-ups ( $50 < v$ ). Figure 2 shows the distribution of exit speeds in these launch angle partitions. Figure 3 shows the distribution of exit speeds in the launch angle and spray angle partitions.

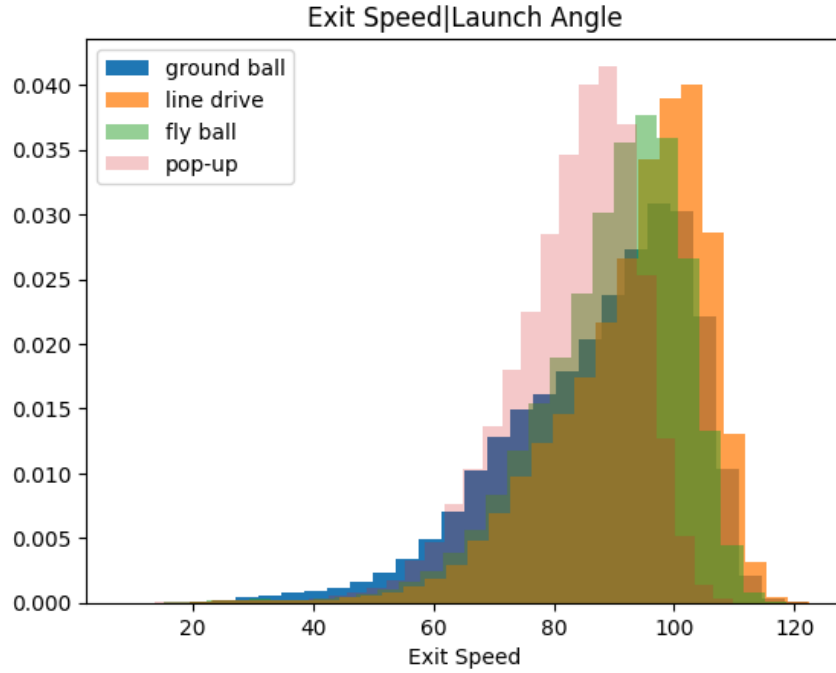


Figure 2: Distribution of exit speeds partitioned by launch angle type. We conclude that exit speed is dependent on launch angle.

Based on these figures, we make the assumption that exit speed is dependent on launch angle and is conditionally independent of spray angle given launch angle. Thus, the joint probability distribution of  $h$ ,  $v$ , and  $s$  can be written as:

$$\begin{aligned}
 p(h, v, s) &= p(h)p(v|h)p(s|v, h) \\
 &= p(h)p(v|h)p(s|v)
 \end{aligned} \tag{5}$$

where the first equality follows from the chain rule and the second equality follows from the assumption that  $s$  is conditionally independent of  $h$  given  $v$ . The following subsections present models for the distributions  $p(h)$ ,  $p(v|h)$ , and  $p(s|v)$ . In each case a Bayesian hierarchical model is used. This provides individual hitter distributions while also facilitating the sharing of information about hitter tendencies throughout the population. Thus, hitters with little ball-in-play data, like rookies, have estimated distributions that borrow heavily from the shared information rather than relying too much on their sparse data.

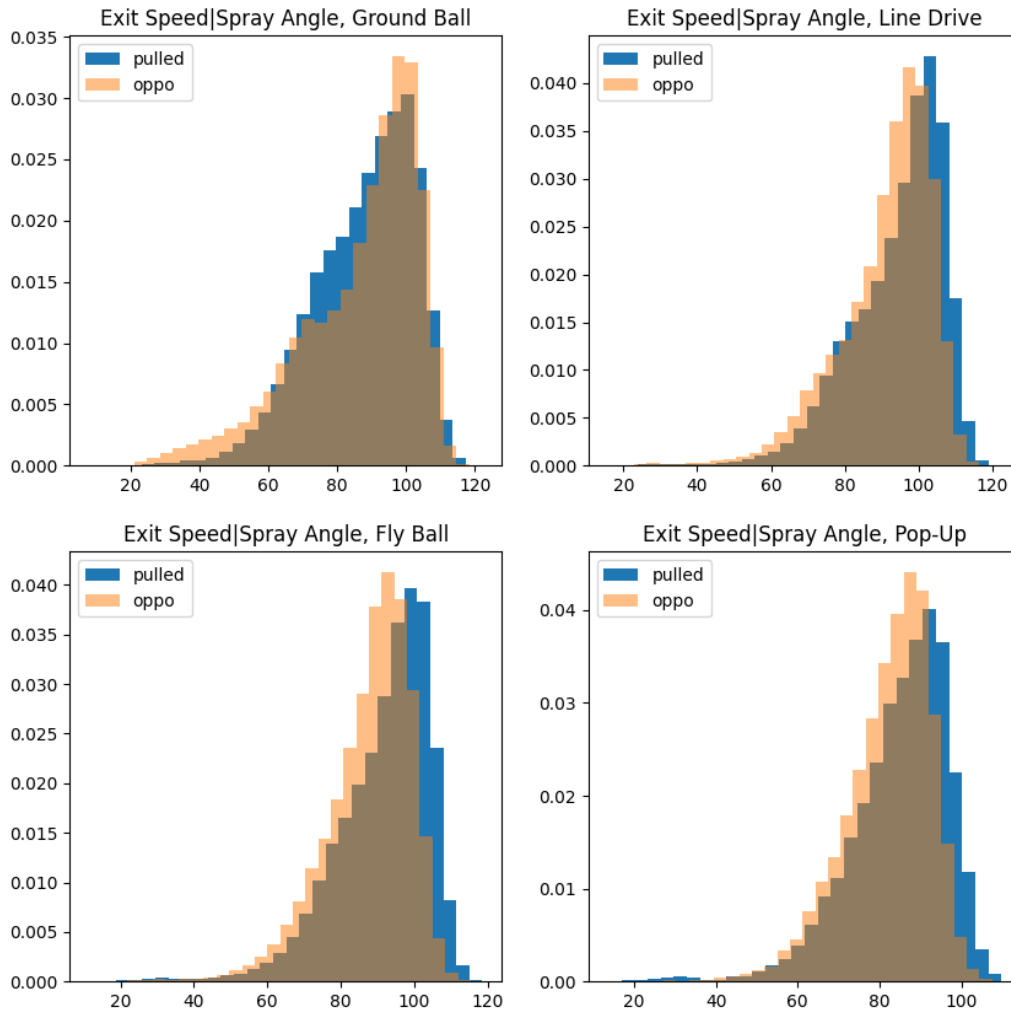


Figure 3: Distribution of exit speeds partitioned by launch angle type and by pulled and oppo spray angles. This suggests that exit speed is conditionally independent of spray angle given launch angle.

## 2.1. Estimating Horizontal Spray Angle Distributions $p(h)$

The data is pre-processed by flipping the sign of the spray angle for all left-handed batters, transforming them into right-handed batters. This gives all pulled balls negative spray angles and all oppo balls positive spray angles. Information can now be shared among all hitters without effects from handedness. Spray angles outside of the interval  $[-55, 55]$  are removed from the data. Although  $-45$  and  $45$  are the angles of the foul lines, a significant number of non-foul balls in the data had spray angles outside of  $[-45, 45]$  but within  $[-55, 55]$ , so bounds of  $[-55, 55]$  were chosen. Lastly, the spray angles are scaled and shifted so that they are between 0 and 1. This allows for the use of the beta distribution, which has support  $[0, 1]$ .

Figure 4 shows the histogram of the shifted and scaled spray angles along with an estimate of a density function to fit these spray angles. This density function is a mixture of two beta distributions with mixing weights  $\hat{\pi} = [0.4, 0.6]$ ,  $\alpha$  parameters  $\hat{\alpha} = [3, 3]$ , and  $\beta$  parameters  $\hat{\beta} = [9, 2]$ . In other words, the density function was defined as  $0.4 \cdot \text{Beta}(3, 9) + 0.6 \cdot \text{Beta}(3, 2)$ .

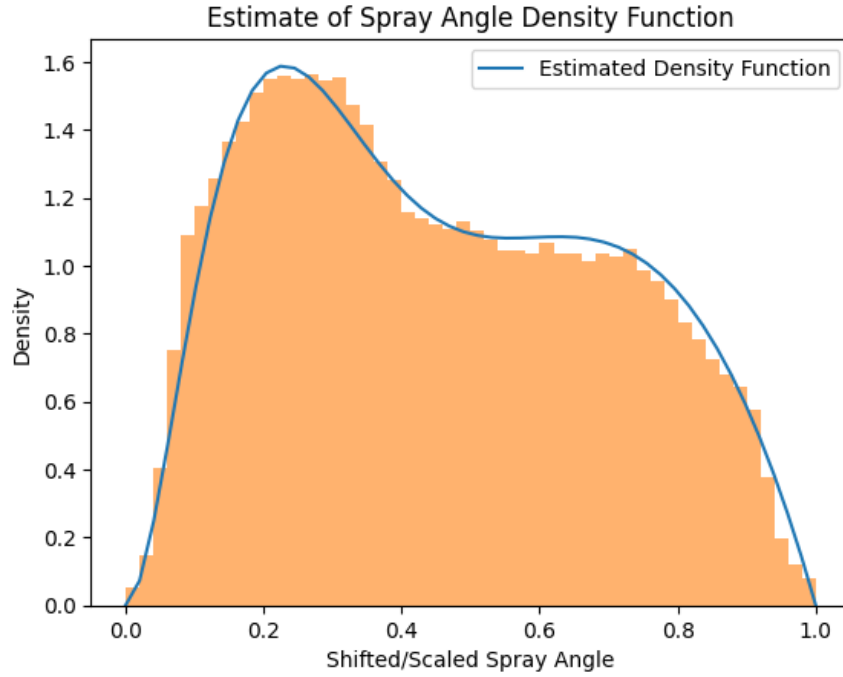


Figure 4: Histogram of shifted and scaled spray angles plotted under the estimated density function.

Given this density estimate, we'd like to sample values of the mixing weights from a distribution with mean  $\hat{\pi}$ , values of  $\alpha$  from a distribution with mean  $\hat{\alpha}$ , and values of  $\beta$  from a distribution with mean  $\hat{\beta}$ .

Our hierarchical model is defined below. Note we use  $\mu$  and  $\sigma$  to signify a distribution's mean and standard deviation. We also abuse notation in the gamma and half-normal distributions. These are distributions of single variables that we have parameterized with vectors of length two. This is meant to signify that we are fitting two of the same kind of distribution with different shapes. We have

$$a_{\pi} \sim \text{Gamma}(\mu = [4,6], \sigma = [10,10]) \quad (6a)$$

$$\mu_{\alpha} \sim \text{Gamma}(\mu = [3,3], \sigma = [10,10]) \quad (6b)$$

$$\mu_{\beta} \sim \text{Gamma}(\mu = [9,2], \sigma = [10,10]) \quad (6c)$$

$$\sigma_{\alpha} \sim \text{Half-Normal}(\sigma = [1,1]) \quad (6d)$$

$$\sigma_{\beta} \sim \text{Half-Normal}(\sigma = [1,1]) \quad (6e)$$

$$\pi_i \sim \text{Dirichlet}(a_{\pi}) \quad (6f)$$

$$\alpha_i \sim \text{Gamma}(\mu = \mu_{\alpha}, \sigma = \sigma_{\alpha}) \quad (6g)$$

$$\beta_i \sim \text{Gamma}(\mu = \mu_{\beta}, \sigma = \sigma_{\beta}) \quad (6h)$$

$$h_i \sim \pi_{i1} \cdot \text{Beta}(\alpha_{i1}, \beta_{i1}) + \pi_{i2} \cdot \text{Beta}(\alpha_{i2}, \beta_{i2}) \quad (6i)$$

where  $\pi_{ij}$  is the  $j^{th}$  value in  $\pi_i$ ,  $\alpha_{ij}$  is the  $j^{th}$  value in  $\alpha_i$ ,  $\beta_{ij}$  is the  $j^{th}$  value in  $\beta_i$ ,  $j \in \{1,2\}$ , and  $i = 1, \dots, N_b$  where  $N_b$  is the number of batters. Note, switch hitters are fitted with two different distributions, one when they bat left-handed and one when they bat right-handed. This model draws  $\pi_i$  from a prior distribution with mean  $\hat{\pi}$ ,  $\alpha_i$  from a prior distribution with mean  $\hat{\alpha}$ , and  $\beta_i$  from a prior distribution with mean  $\hat{\beta}$ , as desired.

We estimate posterior distributions for  $a_\pi, \mu_\alpha, \mu_\beta, \sigma_\alpha, \sigma_\beta, \pi_i, \alpha_i$ , and  $\beta_i$  using automatic differentiation variational inference (ADVI)[1, 20]. We define the batter-specific spray angle distribution,  $p_i(h)$ , using the mean values of the posterior estimates of  $\pi_i, \alpha_i$ , and  $\beta_i$ . If  $\pi_i, \alpha_i$ , and  $\beta_i$  are those means, we have

$$p_i(h) \sim \pi_{i1} \cdot \text{Beta}(\alpha_{i1}, \beta_{i1}) + \pi_{i2} \cdot \text{Beta}(\alpha_{i2}, \beta_{i2})$$

## 2.2. Estimating Vertical Launch Angle Conditional Distributions $p(v|h)$

Since launch angle is dependent on spray angle, we fit two launch angle distributions for each batter, one conditioned on the batter pulling the ball and the other conditioned on the batter going oppo. We estimated priors for the mean and standard deviation of launch angle normal distributions empirically. Figure 5 shows our estimates.

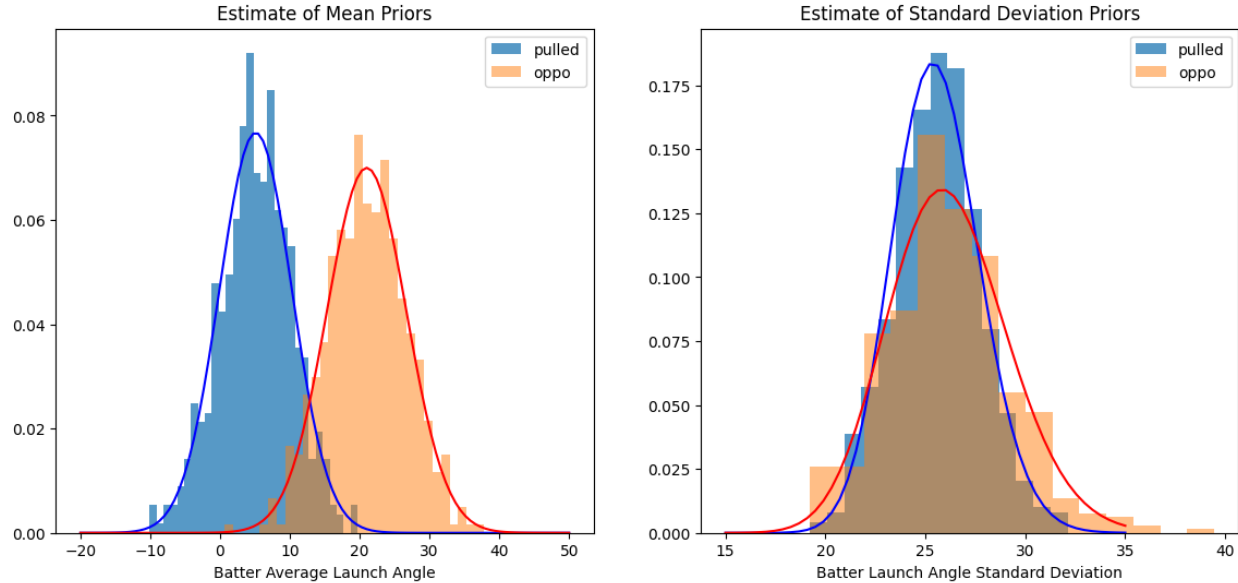


Figure 5: Estimates of the prior distributions of launch angle means and standard deviations partitioned by pulled and oppo spray angles.

The plot on the left shows our estimates of the prior distributions of mean launch angles. The blue histogram shows the means of pulled launch angles for all batters with at least 50 pulled balls, and the orange histogram shows the same thing but for oppo launch angles. We fit normal distributions to these histograms using maximum likelihood estimation. These normal distributions had means 5.1 and 21.0 and standard deviations 5.2 and 5.7 respectively.

The plot on the right shows our estimates of the prior distributions of standard deviations. Like before, the blue histogram is the standard deviations of pulled launch angles for all batters with 50 pulled balls, and the orange is the same but for oppo launch angles. We fit gamma distributions to these histograms using maximum likelihood estimation. These gamma distributions had shape ( $\alpha$ ) parameters of 137.9 and 76.8 and rate ( $\beta$ ) parameters of 5.4 and 2.9 respectively.

We then defined our hierarchical model such that each batter's mean launch angles we're drawn from a prior of  $\mathcal{N}([5.1, 21], [5.2, 5.7])$ , and each batter's launch angle standard deviations were drawn from a prior of  $\text{Gamma}(\alpha = [137.9, 76.8], \beta = [5.4, 2.9])$ . Therefore, our launch angle model is

$$\mu_i \sim \mathcal{N}([5.1, 21], [5.2, 5.7]) \quad (7a)$$

$$\sigma_i \sim \text{Gamma}(\alpha = [137.9, 76.8], \beta = [5.4, 2.9]) \quad (7b)$$

$$v_i | h_i \sim \begin{cases} \mathcal{N}(\mu_{i1}, \sigma_{i1}) & h_i < 0 \\ \mathcal{N}(\mu_{i2}, \sigma_{i2}) & h_i \geq 0 \end{cases} \quad (7c)$$

where  $\mu_{ij}$  is the  $j^{th}$  value in  $\mu_i$ ,  $\sigma_{ij}$  is the  $j^{th}$  value in  $\sigma_i$ ,  $j \in \{1, 2\}$ , and  $i = 1, \dots, N_b$  where  $N_b$  is the number of batters. We once again abuse notation in the gamma and normal distributions. The vectors of length two that we used to parameterize them are meant to signify that we are fitting two separate distributions of one variable rather than a single distribution of two variables. As is apparent in (7c), the first batter-specific normal distribution applies when the batter pulls the ball ( $h < 0$ ) and the second applies when the batter goes oppo ( $h \geq 0$ ).

We sampled from the posteriors of  $\mu_i$  and  $\sigma_i$  for all  $i$  using the NUTS algorithm [2, 18]. We then defined the batter-specific launch angle distributions using the mean of the samples of  $\mu_i$  and  $\sigma_i$ .

### 2.3. Estimating Exit Speed Conditional Distributions $p(s|v)$

It is clear in figures 2 and 3 that exit speeds are left skewed. We therefore started by transforming exit speeds from a left skewed distribution into a right skewed one by setting  $z = s_{\max} - s$ . This would allow us to fit a gamma distribution to the transformed variable since gamma distributions have support  $(0, \infty)$  and can fit right skewed data.

Since exit speed is conditioned on launch angle, we fit four exit speed distributions for each batter, one for each launch angle type. We estimated priors for the shape ( $\alpha$ ) and rate ( $\beta$ ) of these batter-specific gamma distributions empirically. For each of the launch angle types, and for each batter with at least 20 of that type of batted ball, we used maximum likelihood estimation to fit a gamma distribution to the batter's transformed exit speeds for that hit type. This gave us a histogram of shapes and scales ( $\frac{1}{\beta}$ ) for each hit type. Figure 6 shows those histograms.

Since all shapes and scales have to be positive, and since some of these histograms appear right skewed, we defined gamma distributions for each of these histograms using their sample means and standard deviations. Thus in the exit speed hierarchical model defined below, we have the prior gamma distributions of  $\alpha$  and  $\frac{1}{\beta}$  parameterized using the mean  $\mu$  and standard deviation  $\sigma$ . Then the transformed exit speed gamma distributions are parameterized using the batter-specific  $\alpha_i$  and  $\frac{1}{\beta_i}$  drawn from the prior distributions. The hierarchical model follows



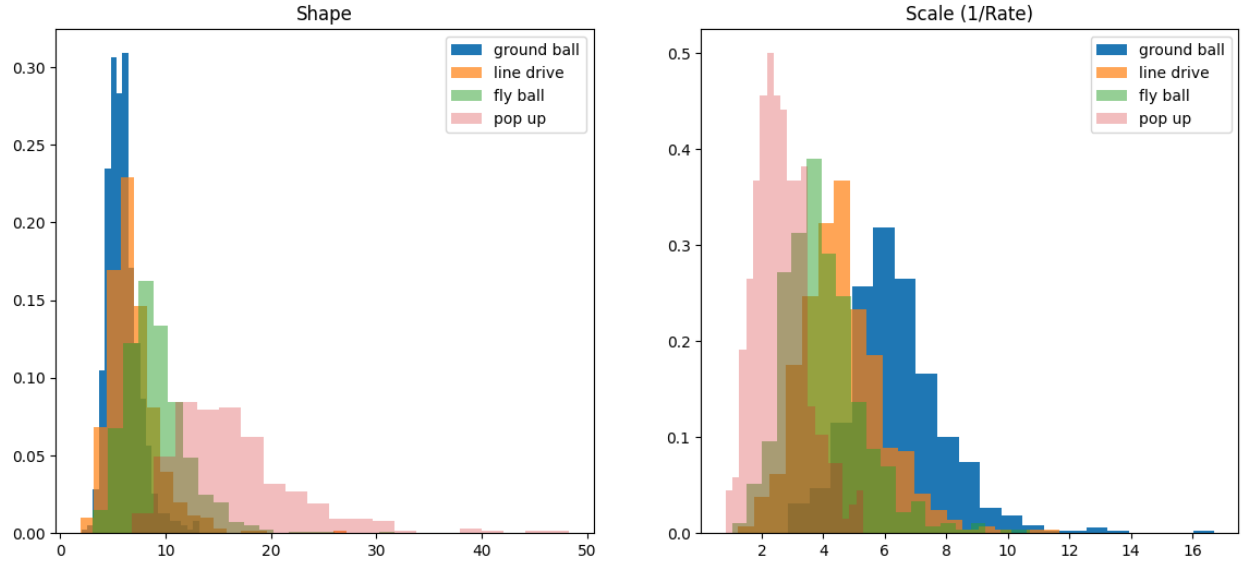


Figure 6: Histograms of gamma distribution shapes and scales found by doing maximum likelihood estimation on each batter's exit speed data.

$$\alpha_i \sim \text{Gamma}(\mu = [6.0, 6.9, 9.1, 16.7], \sigma = [1.5, 2.4, 3.2, 6.1]) \quad (8a)$$

$$1/\beta_i \sim \text{Gamma}(\mu = [6.4, 4.6, 4.0, 2.6], \sigma = [1.6, 1.4, 1.3, 0.8]) \quad (8b)$$

$$z_i | v_i \sim \begin{cases} \text{Gamma}(\alpha_{i1}, \beta_{i1}) & v_i \leq 10 \\ \text{Gamma}(\alpha_{i2}, \beta_{i2}) & 10 < v_i \leq 25 \\ \text{Gamma}(\alpha_{i3}, \beta_{i3}) & 25 < v_i \leq 50 \\ \text{Gamma}(\alpha_{i4}, \beta_{i4}) & 50 < v_i \end{cases} \quad (8c)$$

where  $\alpha_{ij}$  is the  $j^{th}$  value in  $\alpha_i$ ,  $\beta_{ij}$  is the  $j^{th}$  value in  $\beta_i$ ,  $j \in \{1, 2, 3, 4\}$ , and  $i = 1, \dots, N_b$  where  $N_b$  is the number of batters. We once again abuse notation in the gamma distributions. The vectors of length four that we used to parameterize them signify that we are fitting four separate gamma distributions. As is apparent in (8c), the first batter-specific gamma distribution applies when the batter hits a ground ball, the second when he hits a line drive, the third when he hits a fly ball, and the fourth when he hits a pop-up.

Just as we did in the previous subsection, we sample from the posteriors of  $\alpha_i$  and  $\beta_i$  for all  $i$  using the NUTS algorithm. We then define the batter-specific transformed exit speed distributions using the mean of the samples of  $\alpha_i$  and  $\beta_i$ . To conclude this section, we show an example of our posterior estimates of  $p(h)$ ,  $p(v|h)$ , and  $p(s|v)$  for a specific hitter, Nathaniel Lowe.

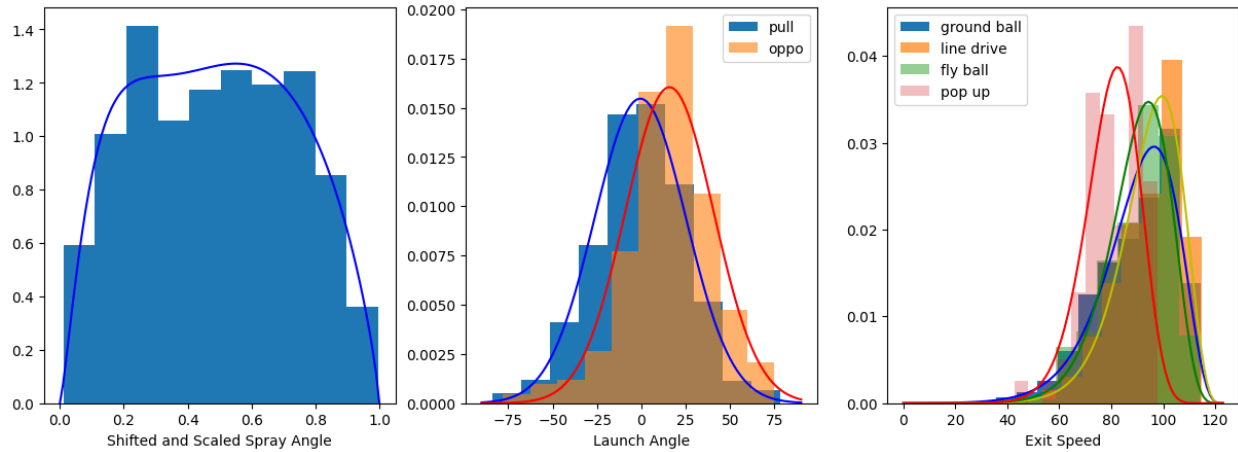


Figure 7: Posterior estimates of  $p(h)$ ,  $p(v|h)$ , and  $p(s|v)$  for Nathaniel Lowe. Note the spray angles are reversed since Lowe is a left-handed batter.

### 3. Finding Optimal Positions

In this section, we describe how we use equations (1) and (2) to optimize fielder positioning. We start by learning the out model  $f$  and the expected wOBA model  $f_{\text{woba}}$ . Then we describe how we use mini-batch stochastic gradient descent/ascent to optimize our objective functions. We conclude this section by testing our optimal solutions on 2023 batted balls.

Since ground ball outs require a throw to first base, whereas non-ground ball outs just require catching the ball in the air, we build separate out models for ground balls and non-ground balls before combining them to create  $f$ . In the next two subsections we describe the ground ball and non-ground ball out models.

#### 3.1 Ground Ball Outs

To predict the probability of an out on any batted ball, we need the starting position of the relevant fielders, which is not publicly available data. However, Baseball Savant has a tool that estimates the average starting position of fielders in certain situations [3]. Thus, for each play in the 2023 ball tracking data we assumed that the fielders were standing in these average positions. Limiting our work to only the 2023 data ensures that there are no infield shifts (due to the recent MLB shift ban rule [4]) and reduces variance in the average positions. Batted balls with runners on base were removed, since baserunners may move fielders out of their typical positions (for instance, when a first baseman has to hold a runner on first). Finally, we used different average starting positions for left-handed batters and right-handed batters and when the infield was in a “standard” alignment or in a “shaded” alignment.

A model of ground ball out probabilities is learned using logistic regression. Denote  $a_d$  as the minimum absolute difference between the spray angle of the batted ball and the spray angle of the infielders' starting positions. Let  $b_t$  be the “ball time to fielder” which is simply the depth from home plate of the closest infielder divided by the exit speed of the ground ball. We can think of  $b_t$  as the time it would take the ball to reach the fielder's depth assuming no deceleration. Then the ground ball out probability model is given by

$$f_{gb}(b, p_s) = p(o|b = gb) = \sigma(-2.78 - 0.15a_d + 5.87b_t - 1.32b_t^2) \quad (9)$$

where  $b$  is the batted ball,  $p_s$  is the positioning strategy,  $o$  is the event that an out occurs,  $gb$  is the event of a ground ball, and  $\sigma$  is the sigmoid function,  $\sigma(x) = \frac{1}{1+e^{-x}}$ . Note, we intentionally use the notation  $f_{gb}(b, p_s)$  to match the notation for our out function  $f(b, p_s)$  in equations (1) and (2) for consistency. This choice is justified because  $a_d$  and  $b_t$  depend on both the fielder positions and the batted ball characteristics, so  $f_{gb}(b, p_s)$  is a function of  $b$  and  $p_s$ .

Figure 8 shows a calibration plot [9] for our ground ball out model on a holdout validation data set. It shows that  $f_{gb}$  makes well-calibrated probability predictions, suggesting that maximizing the outputs of  $f_{gb}$  will maximize the true probability of ground ball outs.

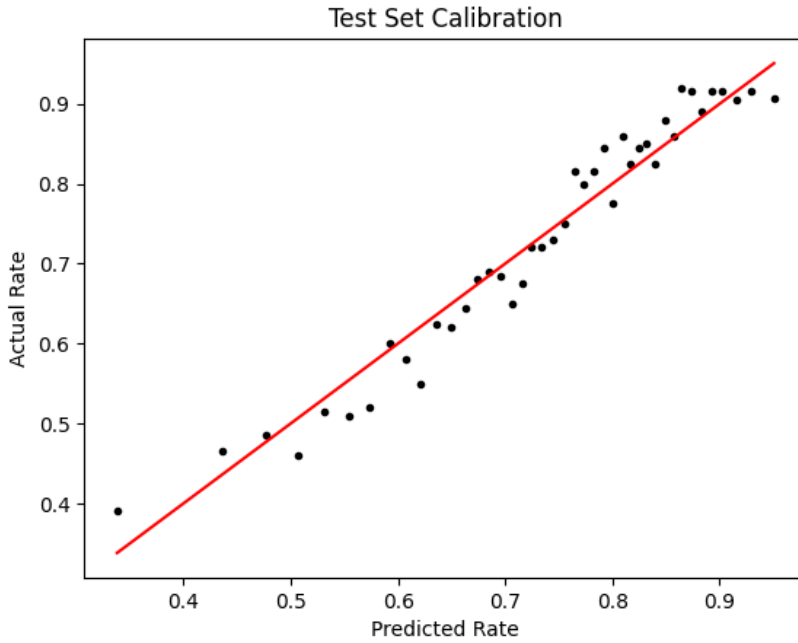


Figure 8: Ground ball out model calibration on a holdout data set. This suggests that  $f_{gb}$  makes well-calibrated out probability predictions.

Figure 9 shows the effects that minimum absolute angle difference and ball time to fielder have on the log odds of an out. Note, the log odds of an out are given by the input to  $\sigma$  in equation (9). Since  $\sigma$  is a monotone increasing function, then we know that increasing log odds will increase the probability of an out. The ball time to fielder plot shows that for low times, which are caused by infielders playing too shallow or by hard hit ground balls, we can improve log odds and therefore probabilities by increasing the ball time to fielder, which can be done by having the fielder play deeper. For large times, caused by infielders playing too deep or weakly hit ground balls, we can improve log odds by decreasing the ball time to fielder, which can be done by having the fielder play shallower. The minimum absolute angle difference plot shows that decreasing the angle difference will increase the log odds/probability of an out.

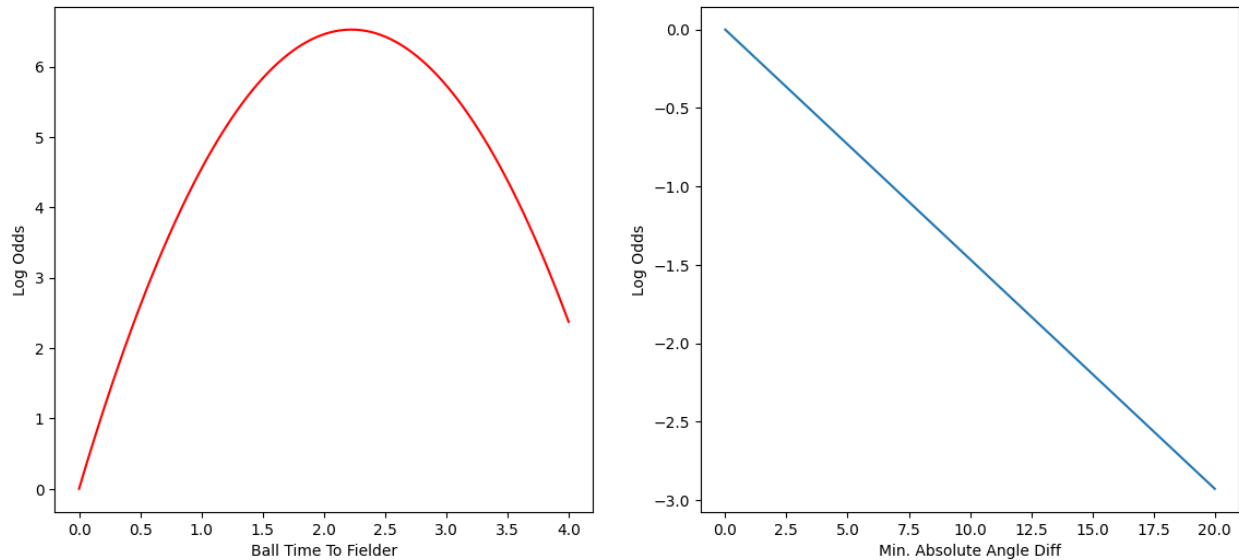


Figure 9: The effects of minimum absolute angle difference and ball time to fielder on the log odds of an out in  $f_{gb}$ .

Data that shows the exact starting position of infielders (such data is available to MLB teams) would facilitate more accurate ground ball out models than what we show in this work. However, one can use publicly available data to make well-calibrated predictions on a holdout data set, as shown in figure 8. Additionally, the effects plots in figure 9 suggest that  $f_{gb}$  incentivizes moving infielders closer to home plate when they need to field a weakly hit ball with large time to fielder and moving them further from home plate when they need to field a hard-hit ball with low time to fielder. The function  $f_{gb}$  also incentivizes decreasing the angle between the ground ball and the fielder. These incentives are appropriate for optimizing positioning to field ground balls.

### 3.2. Non-Ground Ball Outs

As in the previous section, we define the starting positions for fielders in the ball tracking data by assuming the fielders were positioned in the average starting position. We again only use 2023 data to avoid any extreme shifts and remove all plays with baserunners. We also use different averages depending on batter handedness and whether the infield was shaded. In this section, however, the data is further filtered in one additional way: that the outfield fielding alignment is defined as “standard.” This ensures any extreme outfield alignments (like the two-man outfield that the Royals deployed several times in 2023, see figure 10) are removed.

Since the likelihood of an out on a non-ground ball depends on its hang time and landing spot, one needs to estimate hang times and landing coordinates given spray angle, launch angle, and exit speed. We use the baseball trajectory calculator created by physicist Alan Nathan [19] to estimate the hang time and landing coordinates of all batted balls hit in Tropicana Field in 2019. Tropicana Field was chosen because it is a dome; its atmospheric conditions are consistent. We then fit XGBoost models [5] to predict the hang time and distance of a batted ball given the batted ball's spray angle, launch angle, and exit speed as well as the batter's handedness. The models' predictions on a holdout validation set shown in figure 11 suggest these are accurate enough models to use in our analysis.

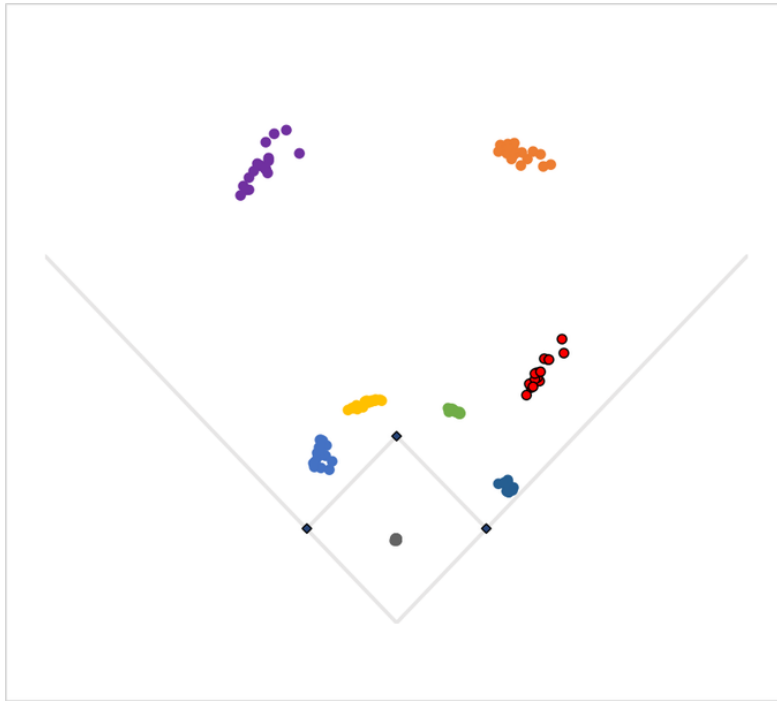


Figure 10: “the 2-OF configuration that the Royals are using” [31]

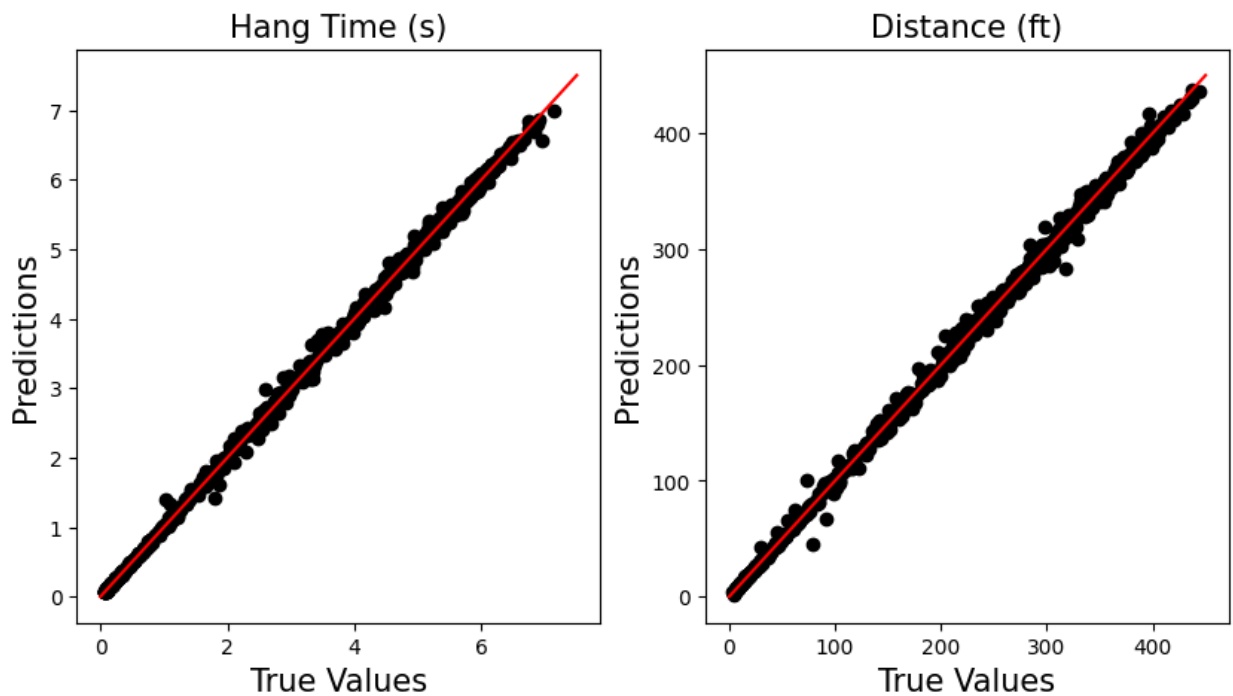


Figure 11: XGBoost hang time and landing distance model predictions on a validation set.

While it may be the case that these models are less predictive in a more extreme atmosphere like Coors Field in Denver, the process used to fit the models could easily be applied to any other

ballpark or atmospheric conditions. For the remainder of this paper we assume the batters and defenses are playing in Tropicana Field.

The model of non-ground ball outs is learned using logistic regression. Denote  $\delta_t$  to be the difference between a batted ball's hang time and the time that it takes the closest fielder to reach the landing spot of the batted ball if we assume they are moving at 27 ft/s (MLB average sprint speed). In other words,  $\delta_t = \text{hang time} - \frac{\text{distance to ball}}{27}$ , where hang time is measured in seconds and distance in feet. Intuitively, a positive value of  $\delta_t$  suggests that the fielder will reach the landing spot before the ball hits the ground, so it will likely be caught for an out. A negative value of  $\delta_t$  means the ball will hit the ground before the fielder can reach it, so it will likely be a hit. Thus, our non-ground ball out model is given by

$$f_{\neg\text{gb}}(b, p_s) = p(o|b = \neg\text{gb}) = \sigma(-2.17 + 1.5\delta_t) \quad (10)$$

where once again  $\sigma$  is the sigmoid function,  $b$  is the batted ball,  $p_s$  is the positioning strategy,  $o$  is the event that an out occurs, and  $\text{gb}$  is the event of a groundball, so  $\neg\text{gb}$  is the event of a non-ground ball.

Figure 12 shows a calibration plot for the non-ground ball out model on a holdout validation data set. It shows that  $f_{\neg\text{gb}}$  makes well-calibrated probability predictions, suggesting that maximizing the outputs of this model will maximize non-ground ball outs.

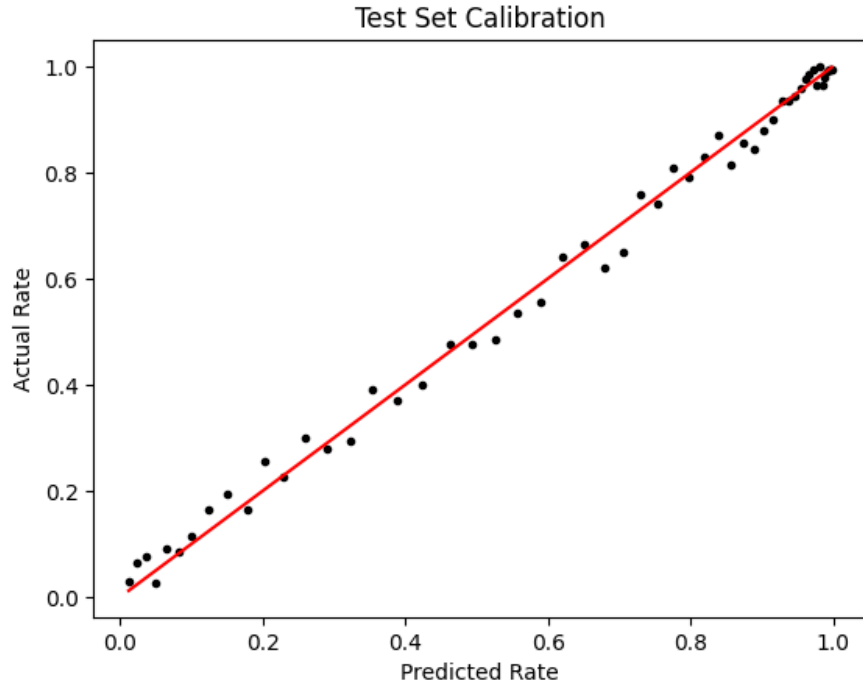


Figure 12: Non-ground ball out model calibration on a holdout data set. This suggests that  $f_{\neg\text{gb}}$  makes well-calibrated out probability predictions.

As with the ground ball out model, precise starting position data would likely be able to make a more accurate non-ground ball out model. However, our model on publicly available data again

makes well-calibrated predictions. Additionally,  $f_{\neg gb}$  incentivizes fielders to increase the difference between a batted ball's hang time and the time it takes them to get to the ball's landing spot. The only way to do this is by decreasing the time to the landing spot by moving closer to it. Thus, this model will place fielders into positions that will help them catch more batted balls in the air.

### 3.3. Final Preparations for Optimization

We can now define the overall out model  $f$  used in our optimization objective functions (1) and (2). Applying the law of total probability, we have

$$\begin{aligned} f(b, p_s) &= p(o|b, p_s) = f_{gb}(b, p_s)p(b = gb) + f_{\neg gb}(b, p_s)p(b = \neg gb) \\ &= f_{gb}(b, p_s)\mathbb{1}_{gb}(b) + f_{\neg gb}(b, p_s)(1 - \mathbb{1}_{gb}(b)) \end{aligned} \quad (11)$$

where  $\mathbb{1}_{gb}(b)$  is the indicator function,

$$\mathbb{1}_{gb}(b) = \begin{cases} 1 & \text{if } b = gb \\ 0 & \text{if } b = \neg gb \end{cases}$$

The last equality in (11) results because we know whether  $b$  is a ground ball or not, so  $p(b = gb) \in \{1, 0\}$  and  $p(b = \neg gb) = 1 - p(b = gb)$ .

The last piece of formulating (2) is to define  $f_{woba}$ . Using just the plays in the batted ball tracking data that resulted in hits, we created a machine learning hit type classification model that takes as input the spray angle, launch angle, exit speed, and batter handedness and outputs the probability of a single (1B), double (2B), triple (3B), and home run (HR) conditioned on the batted ball not being an out. If the wOBA weights of a single, double, triple, and home run are  $w_1, w_2, w_3$ , and  $w_4$  respectively, then we calculate the expected wOBA of a batted ball given that it is not an out as

$$f_{woba}(b) = w_1p(1B|b) + w_2p(2B|b) + w_3p(3B|b) + w_4p(HR|b) \quad (12)$$

where  $p(1B|b), p(2B|b), p(3B|b)$ , and  $p(HR|b)$  are defined using the hit type classification model. We used the wOBA weights from 2023:  $w_1 = 0.883, w_2 = 1.244, w_3 = 1.569$ , and  $w_4 = 2.004$  [6].

Note that the wOBA weight of an out is always 0, so we could calculate the expected wOBA of a batted ball without the condition that it is not an out as

$$p(o|b) \cdot 0 + (1 - p(o|b)) \cdot f_{woba}(b) = (1 - p(o|b)) \cdot f_{woba}(b),$$

where  $p(o|b)$  is the probability of an out on batted ball  $b$ . Substituting  $f(b, p_s)$  in for  $p(o|b)$  reveals how we derived the minimize expected wOBA objective function (2).

### 3.4. Optimization with Gradient Descent

In practice, we approximate the optimal solutions to (1) and (2). Let  $\mathbf{b} = [b_1, b_2, \dots, b_N]$  be a sample of batted balls from the  $i^{th}$  batter's batted ball distribution,  $p_i(h, v, s)$  (5). Then our approximately optimal solution to (1) for batter  $i$  is given by maximizing the average out probability over this large sample of batted balls,

$$p_s^* \in \operatorname{argmax}_{p_s \in P_s} \mathbb{E}_{\mathbf{b}} [f(\mathbf{b}, p_s)] = \operatorname{argmax}_{p_s \in P_s} \frac{1}{N} \sum_{j=1}^N f(b_j, p_s). \quad (13)$$

Likewise, our approximately optimal solution to (2) for batter  $i$  is given by minimizing the average expected wOBA over the large sample of batted balls,

$$p_s^* \in \operatorname{argmin}_{p_s \in P_s} \mathbb{E}_{\mathbf{b}} [(1 - f(\mathbf{b}, p_s)) \cdot f_{\text{woba}}(\mathbf{b})] = \operatorname{argmin}_{p_s \in P_s} \frac{1}{N} \sum_{j=1}^N (1 - f(b_j, p_s)) \cdot f_{\text{woba}}(b_j). \quad (14)$$

We use mini-batch stochastic gradient descent with momentum to find these approximately optimal solutions [7].

When we draw batted ball samples, we use Tropicana Field's outfield dimensions, acquired from [14], to identify batted balls that would land foul or over the Trop's outfield fence. Those balls get removed from the sample. Figure 13 shows an example of the landing coordinates and hang times of batted ball samples that stayed in fair territory in the Trop.

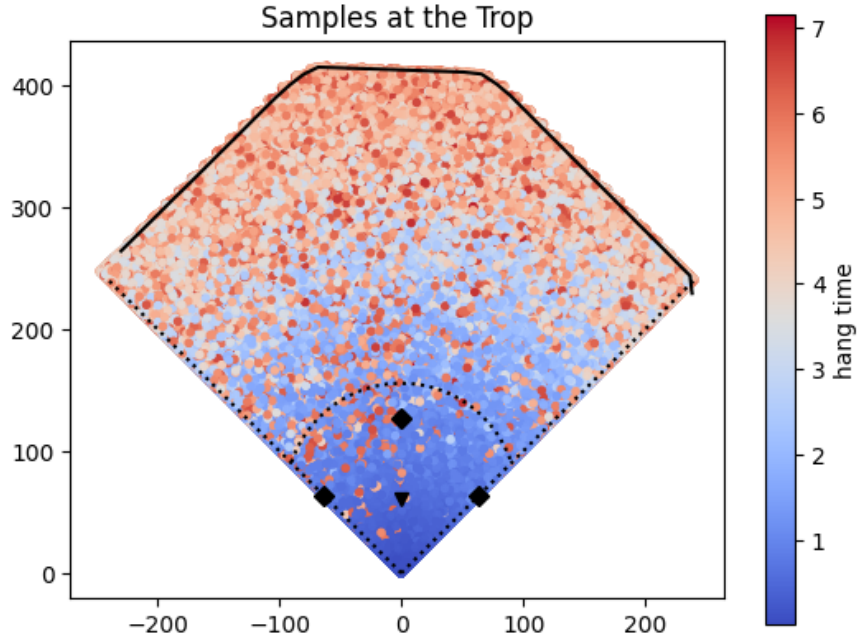


Figure 13: A set of batted ball samples that stayed in fair territory at Tropicana Field.

Recall that  $P_s$  is the set of admissible positioning strategies. Prior to 2023,  $P_s$  was any positioning strategy that had the fielders standing in fair territory with the first baseman close enough to first base to cover it on any ground balls in the infield. When the new shift ban rule was added in 2023,  $P_s$  was constrained further, requiring four infielders in the dirt with two on either side of second base [4]. Stochastic gradient descent is an unconstrained optimization method, so we use projected stochastic gradient descent to meet the constraints on  $P_s$  [11]. After each step of gradient descent the algorithm checks that three conditions are met:



1. The first baseman must be within 40 feet of first base. If not, his position is projected back into  $P_s$  by maintaining the relative angle with first base while decreasing the distance back down to 40 feet.
2. Every infielder must be in the infield dirt, i.e., they must be within 95 feet of the pitching rubber [32]. If not, their position is projected back into  $P_s$  by maintaining the relative angle with the rubber while decreasing the distance back down to 95 feet.
3. There must be two infielders on either side of second base. If the third baseman or shortstop have spray angles greater than  $0^\circ$ , they are projected back into  $P_s$  by maintaining the depth from home plate while decreasing the spray angle back down to  $0^\circ$ . Likewise, if the second baseman has a spray angle less than  $0^\circ$ , his position is projected back into  $P_s$  by maintaining the depth from home plate while increasing the spray angle back up to  $0^\circ$ .

We also allow for the option to relax the 2023 shift ban constraints, in which case we only check that the first baseman was close to first base. Since the optimizer never moves fielders into foul ground there is no need to enforce the fielders in fair territory.

When comparing the optimal positioning strategies from maximizing outs (13) and from minimizing wOBA (14) on 2023 batted balls, the wOBA optimized positioning outperforms the out optimized positioning in terms of expected number of outs as well as expected wOBA. The wOBA optimized results are presented in the next subsection.

### 3.5. Optimization Results

Figure 14 shows our recommended positions with the 2023 constraints at Tropicana Field for 2023 AL and NL MVPs Shohei Ohtani and Ronald Acuña Jr. The gray X's are the initial positions in the optimizer.

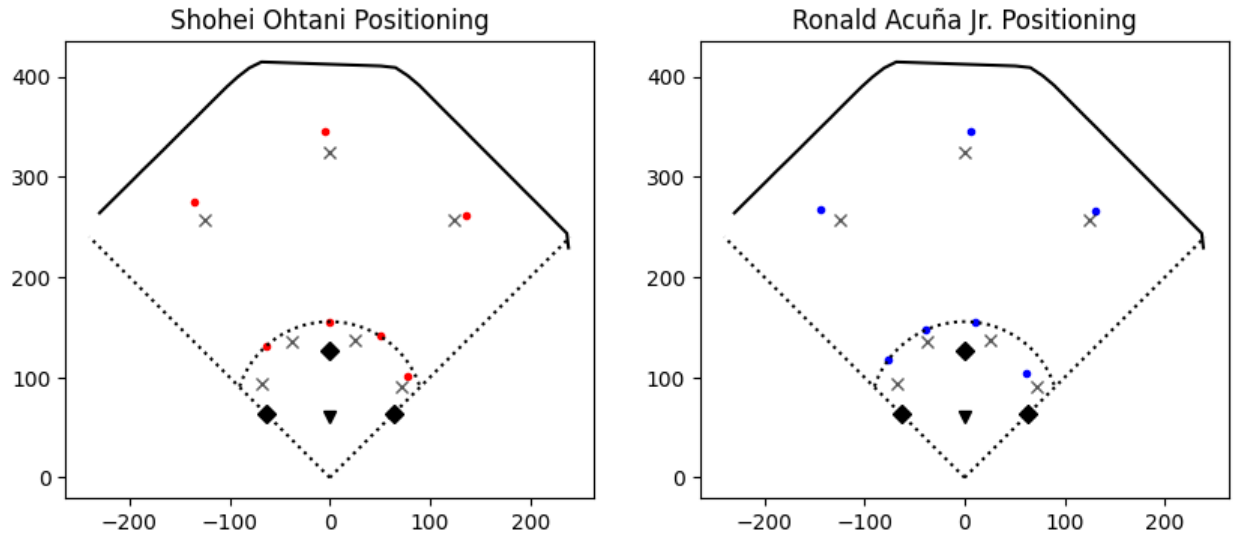


Figure 14: 2023 wOBA optimized positioning recommendations against MVPs Shohei Ohtani and Ronald Acuña Jr. The gray X's are the starting points of the fielders in the optimizer.

To test the effectiveness of our recommended positioning, we use  $f$  and  $f_{woba}$  to calculate the expected outs and expected wOBA on all batted balls in 2023 using our recommended positioning, and then we compared that to the expected outs and expected wOBA using the MLB average starting positions. We removed all home runs in 2023 as well as any ball that would have gotten out

of Tropicana Field since our recommended positions were for the Trop. We also removed plays with non-standard outfield positioning, like we did when we trained  $f_{\neg_{gb}}$ . This helped ensure that our MLB average starting positioning assumption was never too far off. We ultimately had 62,552 plays left for testing.

Using our positioning, the expected batting average in 2023 was 0.305, and the expected wOBA was 0.308. The average starting positions gave an expected batting average of 0.320 and an expected wOBA of 0.321. Thus, over the 62,552 plays our positioning is expected to prevent about 938 more hits (31 per team). Converting wOBA to runs above average [29], our positioning saves an expected 675 more runs (22.5 per team). However, there is a notable discrepancy between the actual batting average on these plays (0.293) and the expected batting average when we assume average starting positions (0.319). Thus, even though we see improvement over the average positioning with our method, there are inaccuracies resulting from the outs model trained using average starting positions.

As a final check on our method, we built another expected out model using the proprietary MLB data, which includes starting positions, courtesy of the Texas Rangers. While we are not permitted to share the specifics of the models that were learned, we can state that these models gave an expected batting average of 0.292 when using the average starting positions. This is much closer to the actual batting average of 0.293. The optimal positions determined by these new models had an expected batting average of 0.281. This is an improvement over the actual batting average of 0.293 by about 750 outs (25 per team).

## 4. Strategizing Against Adaptable Hitters

We conclude this paper by addressing the concern in [13] that hitters will observe the positions of the opposing defenders and change their strategy at the plate in response. Recall that  $\mathcal{D}$  is a set of possible choices of batted ball distributions, and in our game model of defensive positioning (3)-(4), the batter gets to choose the distribution  $d$  in  $\mathcal{D}$  that his batted balls will be sampled from. For example, hitters are occasionally asked to sacrifice bunt in certain game situations to advance a baserunner, so  $\mathcal{D}$  could be reasonably defined as {swing away, bunt}. A particularly skilled hitter may have some ability to go more oppo when he observes an extreme pull-side shift from the defense, so he would have  $\mathcal{D} = \{\text{pull, oppo}\}$ . An equilibrium pair of strategies is a positioning strategy  $p_s^* \in P_s$  and a batter strategy  $d^* \in \mathcal{D}$  satisfying (3) in the case where our objective function is expected outs and satisfying (4) in the case where our objective function is expected wOBA.

Just as in section 3.4, we approximate these equilibria: a large number of samples are drawn from each of the possible batter distributions in  $\mathcal{D}$ . Then the approximate equilibria when the objective function is outs in (3) are  $(p_s^*, d^*)$  satisfying

$$p_s^* \in \operatorname{argmax}_{p_s \in P_s} \min_{d \in \mathcal{D}} \mathbb{E}_{\mathbf{b} \sim d} [f(\mathbf{b}, p_s)] = \operatorname{argmax}_{p_s \in P_s} \min_{d \in \mathcal{D}} \frac{1}{N} \sum_{j=1}^N f(b_j, p_s) \quad (15a)$$

$$d^* \in \operatorname{argmin}_{d \in \mathcal{D}} \max_{p_s \in P_s} \mathbb{E}_{\mathbf{b} \sim d} [f(\mathbf{b}, p_s)] = \operatorname{argmin}_{d \in \mathcal{D}} \max_{p_s \in P_s} \frac{1}{N} \sum_{j=1}^N f(b_j, p_s) \quad (15b)$$

where  $\mathbf{b}$  is the sample of size  $N$  of batted balls from  $d$ , and  $b_j$  is the  $j^{th}$  batted ball in the sample. Likewise, the approximate equilibria when the objective function is wOBA in (4) are  $(p_s^*, d^*)$  satisfying

$$p_s^* \in \operatorname{argmin}_{p_s \in P_s} \max_{d \in \mathcal{D}} \mathbb{E}_{\mathbf{b} \sim d} [(1 - f(\mathbf{b}, p_s)) \cdot f_{\text{woba}}(\mathbf{b})] = \operatorname{argmin}_{p_s \in P_s} \max_{d \in \mathcal{D}} \frac{1}{N} \sum_{j=1}^N (1 - f(b_j, p_s)) \cdot f_{\text{woba}}(b_j) \quad (16a)$$

$$d^* \in \operatorname{argmax}_{d \in \mathcal{D}} \min_{p_s \in P_s} \mathbb{E}_{\mathbf{b} \sim d} [(1 - f(\mathbf{b}, p_s)) \cdot f_{\text{woba}}(\mathbf{b})] = \operatorname{argmax}_{d \in \mathcal{D}} \min_{p_s \in P_s} \frac{1}{N} \sum_{j=1}^N (1 - f(b_j, p_s)) \cdot f_{\text{woba}}(b_j) \quad (16b)$$

To find these approximate equilibria, we once again use stochastic gradient descent. Since the batter can observe the positions of the fielders, we assume that for any choice of  $p_s$  he will use the  $d \in \mathcal{D}$  that either minimizes outs or maximizes wOBA against  $p_s$ . Thus, in (15) we calculate the gradient of the minimum expected outs over  $\mathcal{D}$  with respect to  $p_s$ , whereas before we just calculated the gradient of expected outs with respect to  $p_s$  in (13). Likewise, in (16) we calculate the gradient of the maximum expected wOBA over  $\mathcal{D}$  with respect to  $p_s$ .

This approach can be generalized to any finite number of batter actions  $d \in \mathcal{D}$ . The examples above had  $|\mathcal{D}| = 2$ , but one could apply the same methodology to some omnipotent batter who can choose the precise spray angle at which they hit the ball. Their action space would have 90 possible actions, one for each spray angle. It is likely that the positioning strategy against such a batter would be the least exploitable strategy in  $P_s$ . It would have to fill in as many gaps as possible. Nevertheless, in the following two subsections we find equilibria for two possible batter action spaces,  $\mathcal{D} = \{\text{pull, oppo}\}$  and  $\mathcal{D} = \{\text{bunt, swing away}\}$ .

#### 4.1. Shift-Beaters

We suspect that there are some hitters who make an effort to go more oppo when they observe the defense shifting them to the pull side. To identify such hitters, we took the ball in play data from 2021-2022, and we refit the spray angle distributions using the method in section 2.1. However, this time we fit two different  $p(h)$  distributions for each hitter, one using only data where the infield was shifted against them, and the other using only data where the infield was not shifted against them. We compared the means of the two distributions to identify hitters that went more oppo against a shift. Of the hitters with at least 100 shifted and non-shifted batted balls in the training data, the biggest “shift-beaters” were Josh Bell, Josh Rojas, Chas McCormick, Adam Frazier, and Tyler O'Neill. Note that Josh Bell is a switch hitter, so to clarify, we found him to be a shift-beater when he bats left-handed. Figure 15 shows our posterior estimates for Josh Bell and Chas McCormick, along with a histogram of the spray angles used to learn these distributions. Recall that we flip the sign of spray angles for left-handed batters, so Bell's actual spray angles would be flipped the other way. For the remainder of this section we will focus our attention on Chas McCormick.

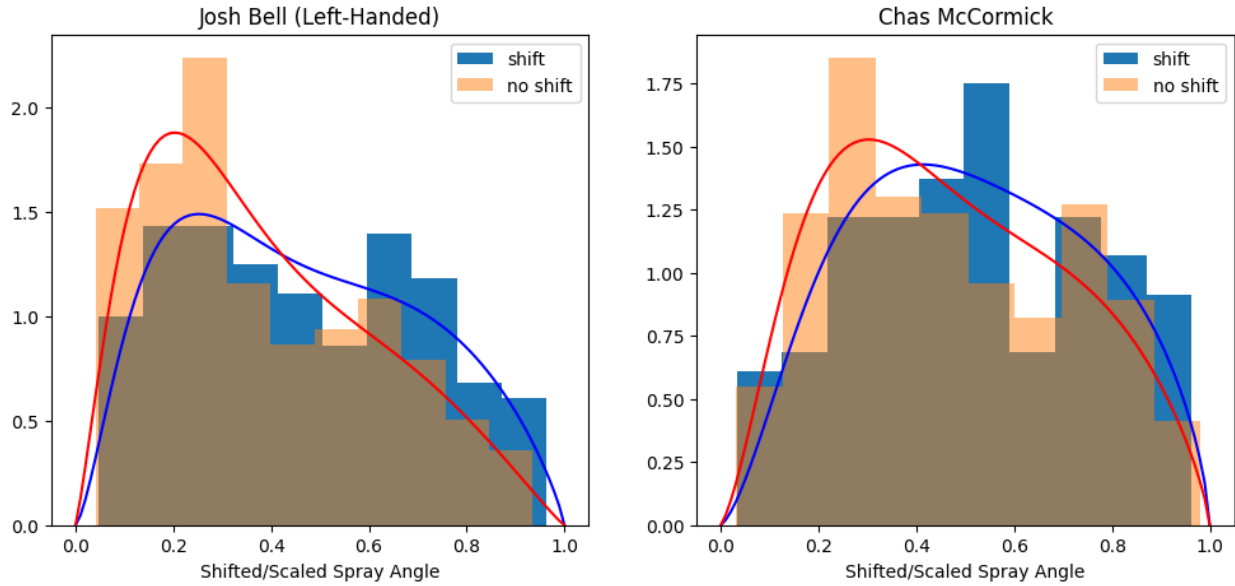


Figure 15: Posterior estimates of Josh Bell and Chas McCormick's shift and non-shift spray angle distributions

Let  $d_{\text{oppo}}$  be McCormick's batted ball distribution against the shift and  $d_{\text{pull}}$  his batted ball distribution against a non-shift. Then McCormick's action set is  $\mathcal{D} = \{d_{\text{oppo}}, d_{\text{pull}}\}$ . Applying (15), the approximate equilibrium positioning strategy when maximizing outs against Chas McCormick is given by

$$p_s^* \in \operatorname{argmax}_{p_s \in P_s} \min_{d \in \{d_{\text{oppo}}, d_{\text{pull}}\}} \frac{1}{N} \sum_{j=1}^N f(b_j, p_s) \quad (17)$$

The optimal positioning strategy is calculated using (13) when  $d = d_{\text{oppo}}$  and  $d = d_{\text{pull}}$  in order to compare those positioning strategies with the equilibrium.

Figure 16 shows the equilibrium strategy (red dots), the strategy against  $d_{\text{pull}}$  (blue triangles), and the strategy against  $d_{\text{oppo}}$  (orange triangles). Again the gray X's are the starting positions of the fielders in the optimizer. The optimizer initially suggested moving the infielders to an unrealistic depth, so we added a new constraint that the infielders had to be within 100 feet of the pitching rubber.

Using our outs model  $f$ , we found the expected batting average when McCormick uses the pull strategy against the blue defense is 0.274 and when he uses the oppo strategy against the blue defense it is 0.271. When he uses the pull strategy against the orange defense he has an expected batting average of 0.275, and when he uses the oppo strategy against the orange defense he has an expected batting average of 0.269. It is not surprising that the orange defense does better against the oppo strategy than the blue defense, nor is it surprising that the blue defense does better against the pull strategy than the orange defense. However, it is interesting that whether the defense uses the blue or orange strategy, McCormick is better off using his pull strategy. That suggests that the pull strategy dominates the oppo strategy, and it explains why the equilibrium

positioning matches the blue positioning. If McCormick's pull strategy dominates his oppo strategy, then when we solve for the equilibrium we assume that regardless of positioning McCormick chooses the pull strategy. Thus, each step of gradient descent in (17) uses the samples from  $d_{\text{pull}}$ , so we end up in the same place as we do when we solve (13) with  $d = d_{\text{pull}}$ . Not surprisingly, the equilibrium strategy for McCormick is  $d^* = d_{\text{pull}}$ . The expected batting average in the equilibrium is 0.274, which matches the expected batting average we calculated when McCormick uses his pull strategy against the blue defense, as expected.

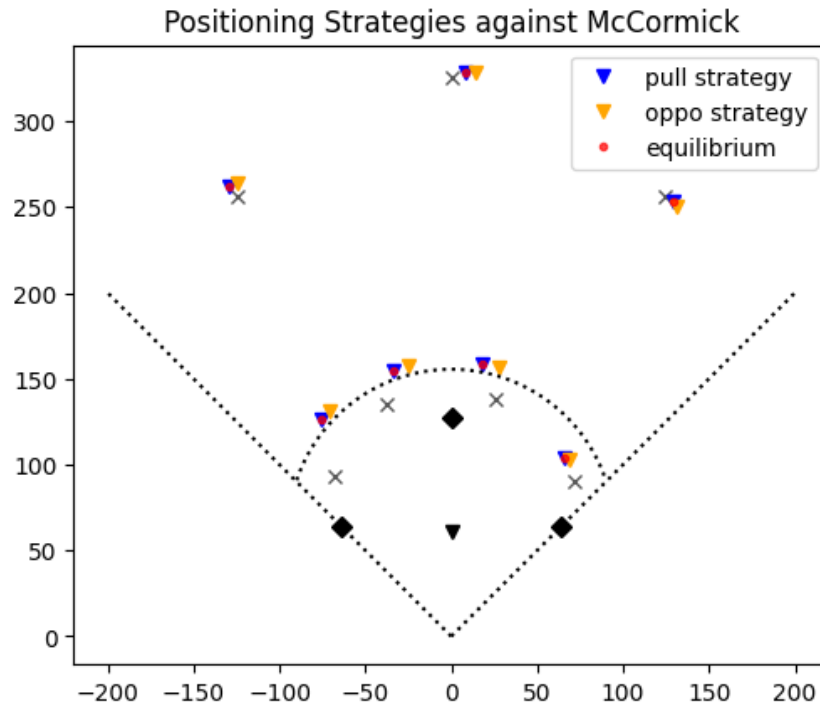


Figure 16: The optimal positioning strategies and equilibrium positioning strategies against “shift-beater” Chas McCormick. McCormick’s pull strategy is dominant, so the equilibrium defense is the same as the defense against the pull strategy.

It is not true in general that one of the hitter's strategies will dominate the other. In the next section we provide an example where that is not the case.

## 4.2. Bunting or Swinging Away

Unlike Chas McCormick, many hitters may struggle to go more oppo against the shift. However, we believe that most hitters, perhaps with some practice, should be able to bunt to beat an extreme fielding alignment. Left-handed batters in particular could benefit from bunting down the third baseline against a shift since these alignments often leave the area by third base wide open. In this section, we determine how a defense should position when a very pull-heavy left-handed batter begins to show a willingness to bunt down the third baseline.

We started by estimating a bunt batted ball distribution,  $d_{\text{bunt}}$ . Our assumption is that we will have a left-handed batter who is trying to bunt down the third baseline, so we took sacrifice bunts from 2019-2023 that were hit by left-handed batters with spray angles less than  $-22.5^\circ$ . We shifted and

scaled the spray angles of the bunts so that they were between 0 and 1, and then we used maximum likelihood estimation to fit a beta distribution to the data. The resulting distribution was  $p_{\text{bunt}}(h) \sim \text{Beta}(0.945, 1.142)$ . Similarly, we fit a normal distribution to the exit speeds of these bunts using maximum likelihood estimation,  $p_{\text{bunt}}(s) \sim \mathcal{N}(34, 7)$ . Finally, we shifted launch angles so that they were between 0 and 180, and we used maximum likelihood estimation to fit a gamma distribution to these shifted launch angles,  $p_{\text{bunt}}(v) \sim \text{Gamma}(\alpha = 4, \beta = 0.07)$ . Since we defined these distributions independently, their joint distribution is just the product, so we have  $d_{\text{bunt}} = p_{\text{bunt}}(h)p_{\text{bunt}}(v)p_{\text{bunt}}(s)$ . Figure 17 shows an example of a sample of batted balls drawn from  $d_{\text{bunt}}$ .

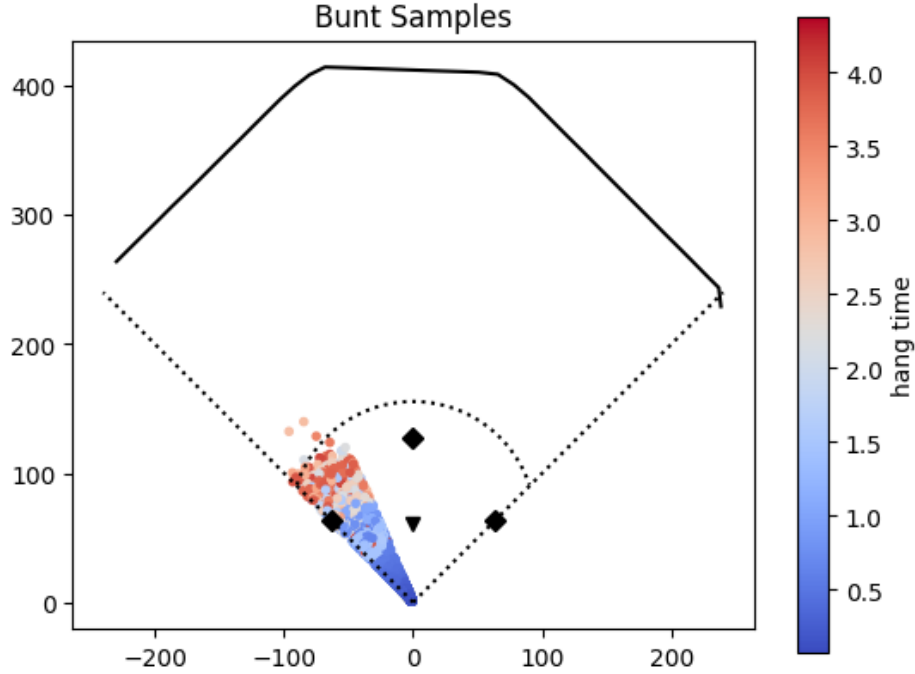


Figure 17: A sample of shift-beating bunts from left-handed hitters.

Consider switch-hitter Carlos Santana when he bats left-handed. Our batter action set  $\mathcal{D}$  is therefore given by  $\{d_{\text{bunt}}, d_{\text{swing}}\}$  where  $d_{\text{swing}}$  is Santana's left-handed batted ball distribution from section 2. Applying (15) to this specific example gives the optimization problem we need to solve to find the equilibrium positioning strategy

$$p_s^* \in \underset{p_s \in P_s}{\operatorname{argmax}} \min_{d \in \{d_{\text{bunt}}, d_{\text{swing}}\}} \frac{1}{N} \sum_{j=1}^N f(b_j, p_s) \quad (18)$$

Figure 18 shows our positioning recommendations against Santana. The blue triangles show the positioning under the assumption that Santana will swing away, and the orange triangles show the positioning under the assumption that he will bunt. It is notable that most of the positions in the equilibrium match the positions of the blue defense. Essentially the only difference between the equilibrium and the blue defense is that we move the third baseman a little shallower and closer to the third baseline to defend against the threat of a bunt.

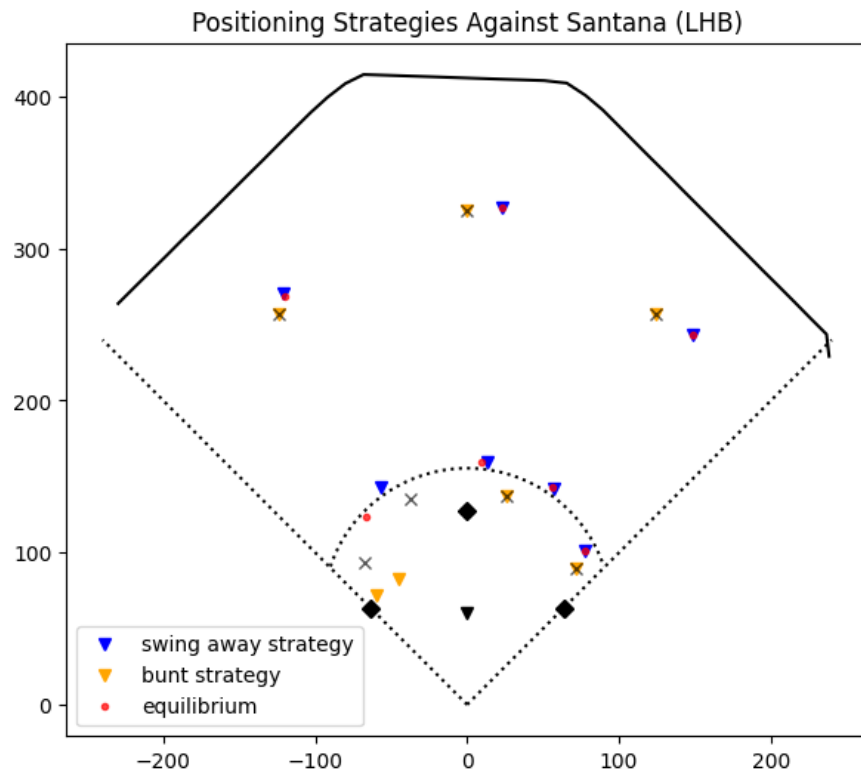


Figure 18: The optimal positioning strategies and equilibrium strategy against Carlos Santana. The threat of a well-placed bunt forces the third baseman to move from the shifted defense given by the blue triangle to a shallower depth in the equilibrium strategy.

When Santana bunts against the orange defense, the expected batting average is 0.062. When he swings away, the expected batting average is 0.354. When he bunts against the blue defense, his expected batting average is 0.453, and when he swings away his expected batting average is 0.268. In the equilibrium, Santana's optimal strategy is to swing away, which gives an expected batting average of 0.272. Clearly it is extremely advantageous for Santana to bunt against the shift. A defense that shifts against Santana gifts him the incredible opportunity to be the greatest hitter for average of all time, and all he has to do is bunt the ball where the third baseman normally stands. It is truly mystifying to us that so few hitters bunt when an infield plays in an extreme alignment such as our blue defense. Additionally, by being willing to bunt when that is advantageous, Santana forces the defense to move to the equilibrium positioning, and in the equilibrium he has an expected batting average of 0.272, which is four points better than his expected average when he swings away against the blue defense. In other words, a willingness to bunt forces the defense into a positioning strategy where Santana can swing away and expect more hits than he would get if he always swung away and never bunted. This conclusion echoes the statement in Tango, Lichtman, and Dolphin's *the Book* [30], which reads, "The batting team must sometimes attempt a [bunt] to keep the defense from playing all the way back." As great admirers of Tango, Lichtman, and Dolphin, we are thrilled to have provided further evidence to support their claim.

Admittedly we are not convinced that bunting would not still be the most effective strategy against the equilibrium positioning, even though Santana's equilibrium strategy is to swing away. We hypothesize that the third baseman is still too deep to effectively field a bunt, even in the



equilibrium, another shortcoming of the publicly available data used to train our out models. The models did not appropriately learn the relationship between depth and out likelihood. Thus, we once again make use of the models learned from the proprietary data shared with us by the Texas Rangers.

Consider the batter Joey Gallo, who is another pull-heavy left-handed batter, and who is often at the center of shift-ban discussions as a hitter who seems likely to benefit from the shift-ban rule [24, 21]. The set  $d_{\text{bunt}}$  is the same, but we switch  $d_{\text{swing}}$  from Santana's batted ball distribution to Gallo's. Then we solve for the equilibrium in (18) using the Rangers out model.

Figure 19 shows the positioning recommendations for Gallo. Note, the Rangers requested that we only display the location of the third baseman, but the main position of interest is the third baseman anyways since he is primarily responsible for the batted balls in  $d_{\text{bunt}}$ . This time the depth of the third baseman in the equilibrium looks more reasonable for fielding bunts than the depth in Santana's equilibrium.

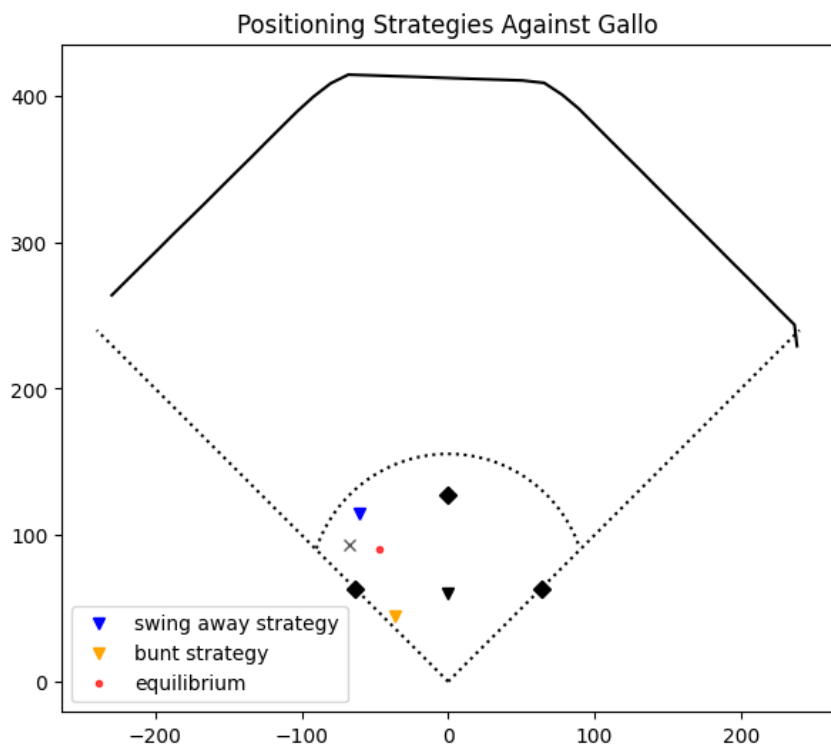


Figure 19: The optimal and equilibrium positioning strategies of just the third baseman against Joey Gallo. The threat of a well-placed bunt forces the third baseman to move to a shallower depth.

The expected batting average when Gallo bunts against the orange defense is 0.036. When he swings away against the orange defense, the expected average is 0.290. When he bunts against the blue defense, the expected average is 0.574, and when he swings away against the blue defense the expected average is 0.223. In the equilibrium, Gallo's optimal strategy is to swing away, which gives an expected batting average of 0.226. These results are very similar to the results we had for Santana. Like Santana, Gallo has the potential to be the greatest hitter ever if he is willing to bunt against a shifted defense. Also like Santana, if Gallo is willing to bunt when that is advantageous, he forces the fielders to adjust to the equilibrium positioning. From this positioning, he has a better



expected batting average when he swings away than he does when he swings away against the original blue defense. Just by being willing to bunt, Gallo would eventually face defenses that will give up more hits against him when he swings away.

## 5. Conclusion

In this paper, we estimated joint probability distributions of batted ball spray angle, launch angle, and exit speed as well as models to estimate the likelihood of an out and expected wOBA given a batted ball trajectory and fielder positioning. We used those models in a novel approach to optimize fielder positioning. We determined that our approach decreases expected hits allowed and expected runs allowed relative to MLB average fielder positioning strategies. We also developed a zero-sum game model to position against adaptable hitters who change their batted ball tendencies in response to the defense's positioning. Notably, we discovered that some left-handed hitters can significantly improve their batting average by bunting against a shifted defense. This leads us to wonder if the 2023 shift-ban was necessary at all. If the goal of the ban was to increase batting averages across the league, a better way to achieve that goal may have been to encourage shifting while also encouraging pull-heavy hitters to practice and start implementing some opposite field bunts.

## References

- [1] <https://www.pymc.io/projects/docs/en/latest/api/generated/pymc.ADVI.html>.
- [2] [https://www.pymc.io/projects/docs/en/stable/api/generated/pymc.sampling.jax.sample\\_numpyro\\_nuts.html](https://www.pymc.io/projects/docs/en/stable/api/generated/pymc.sampling.jax.sample_numpyro_nuts.html).
- [3] <https://baseballsavant.mlb.com/visuals/fielder-positioning>.
- [4] <https://www.mlb.com/glossary/rules/defensive-shift-limits>.
- [5] <https://xgboost.readthedocs.io/en/stable/>
- [6] <https://www.fangraphs.com/guts.aspx?type=cn>.
- [7] <https://pytorch.org/docs/stable/generated/torch.optim.SGD.html>.
- [8] Baseball savant. <https://baseballsavant.mlb.com/>.
- [9] Probability calibration. <https://scikit-learn.org/stable/modules/calibration.html>.
- [10] Statcast. <https://www.mlb.com/glossary/statcast>.
- [11] Charu C. Aggarwal. Constrained Optimization and Duality, pages 255-297. Springer International Publishing, Cham, 2020.
- [12] Jim Albert. Chance of hit as function of launch angle, exit velocity and spray angle. <https://baseballwithr.wordpress.com/2018/01/15/chance-of-hit-as-function-of-launch-angle-exit-velocity-and-spray-angle/>, Jan 2018.
- [13] Russell A Carleton. Baseball therapy: The pretty good case that the shift doesn't work. <https://www.baseballprospectus.com/news/article/29085/baseball-therapy-the-pretty-good-case-that-the-shift-doesnt-work/>, May 2016.
- [14] danmorse314. Danmorse314/dinger-machine: An app to test whether any batted ball would've been a homer in another park. <https://github.com/danmorse314/dinger-machine>.
- [15] Todd Easton and Kyle Becker. Optimizing baseball defensive alignments through integer programming and simulation. International Journal of Modelling and Simulation, 37(2):82-87, 2017.

- [16] Jeffrey Gerlica, Izaiah LaDuke, Garrett O' Shea, Pierce Pluemer, and Johnathon Dulin. Quantifying the outfield shift using k-means clustering. 2021.
- [17] John M Harris, Elizabeth L Bouzarth, Benjamin C Grannan, Andrew J Hartley, Kevin R Hutson, and Ella M Morton. Swing shift: A mathematical approach to defensive positioning in baseball. <https://www.sloansportsconference.com/research-papers/swing-shift-a-mathematical-approach-to-defensive-positioning-in-baseball>.
- [18] Matthew D Hoffman and Andrew Gelman. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15:1593-1623, Apr 2014.
- [19] David Kagan and Alan M. Nathan. Statcast and the baseball trajectory calculator. *The Physics Teacher*, 55(3):134-136, 2017.
- [20] Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M Blei. Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18:1-45, Jan 2017.
- [21] Abhishek Kumar. "Joey Gallo is about to have a career" - fans troll the yankees slugger after mlb and mlbpa mutually agree to ban shift, Mar 2022.
- [22] James LeDoux. Introducing pybaseball: An open source package for baseball data analysis. <https://jamesrledoux.com/projects/open-source/introducing-pybaseball/>, Jul 2017.
- [23] Myles Lewis and Reid Bailey. Batted ball spray charts: a system to determine infield shifting. In *2015 Systems and Information Engineering Design Symposium*, pages 206-211, 2015.
- [24] Shanna McCarriston. Yankees' Joey Gallo says mlb should ban the shift: "at some point, you have to fix the game a little bit" . <https://www.cbssports.com/mlb/news/yankees-joey-gallo-says-mlb-should-ban-the-shift-at-some-point-you-have-to-fix-the-game-a-Feb-2022>.
- [25] Michael W Model. Hitting around the shift: Evaluating batted-ball trends across major league baseball. 2020.
- [26] Anthony Montes. Optimizing outfield positioning: Creating an area-based alignment using outfielder ability and hitter tendencies. *Spring 2021 Baseball Research Journal*, 2021.
- [27] Alan T. Murray, Antonio Ortiz, and Seonga Cho. Enhancing strategic defensive positioning and performance in the outfield. *Journal of Geographical Systems*, 24(2):223-240, 2022.
- [28] Scott Powers. Toward a probability distribution over batted-ball trajectories. <https://tth.fangraphs.com/toward-a-probability-distribution-over-batted-ball-trajectories/>, Aug 2016.
- [29] Piper Slowinski. woba. <https://library.fangraphs.com/offense/woba/>, Feb 2010.
- [30] Tom M. Tango, Mitchel G. Lichtman, and Andrew E. Dolphin. *The book: playing the percentages in baseball*. Potomac Books, 2007.
- [31] Tangotiger. <https://twitter.com/tangotiger/status/1645068710804111360>, Apr 2023.
- [32] Blake Williams. Mlb to begin enforcing length of infield dirt at stadiums. <https://angelsnation.com/mlb-to-begin-enforcing-length-of-infield-dirt-at-stadiums/2022/10/23/>, Oct 2022.