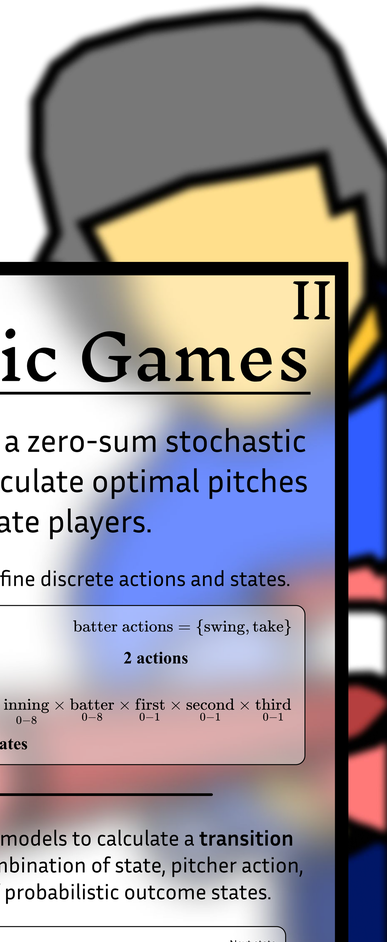
PLAY BALL

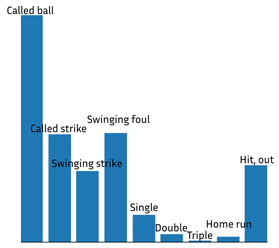
Explorations in Baseball and Game Theory

Start here for the full story! ↓↓↓

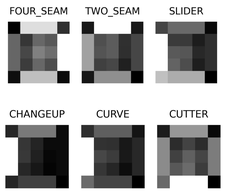


I Data and Models

11 million pitches taken from MLB's Statcast database, from 2008-2024

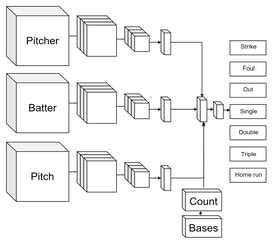


Note how we divide zones and pitch types

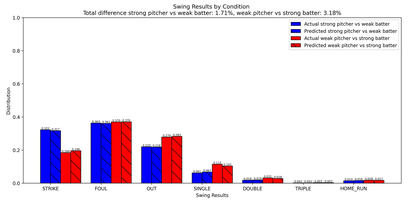


We train 3 models to capture important distributions

Player representations are learned from data, consisting of stats for each zone and pitch type. For batters, we record OBP and relative swinging frequency. Pitchers use normalized speed and relative throwing frequencies.

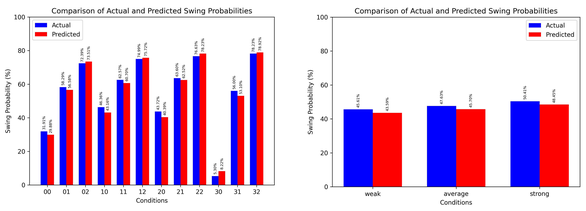


SwingOutcomeDistribution



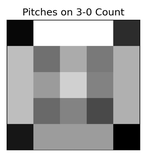
Our representations capture meaningful variations in the data distributions.

BatterPatienceDistribution



PitcherControlDistribution

Pitcher control cannot be directly calculated from real life data. We assume consistent intentions across 3-0 counts and fit Gaussian distributions for each pitch type.



III Batter Selection

BREAKING NEWS

Researchers Find New Cardinals Batting Order with .04 ERA Expected Improvement*

Existing	Ours
1. Masyn Winn (R) SS	1. Brendan Donovan (L) LF
2. Alec Burleson (L) RF	2. Willson Contreras (R) C
3. Willson Contreras (R) C	3. Nolan Gorman (L) 2B
4. Paul Goldschmidt (R) 1B	4. Paul Goldschmidt (R) 1B
5. Brendan Donovan (L) LF	5. Nolan Arenado (R) 3B
6. Nolan Arenado (R) 3B	6. Alec Burleson (L) RF
7. Lars Nootbaar (L) CF	7. Matt Carpenter (L) DH
8. Matt Carpenter (L) DH	8. Lars Nootbaar (L) CF
9. Nolan Gorman (L) 2B	9. Masyn Winn (R) SS

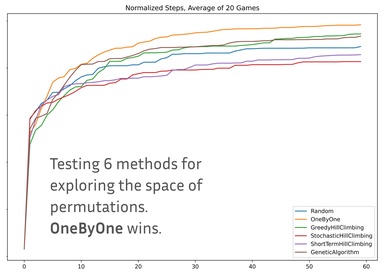
Two factors complicate finding an optimal batting order with our setup. **1.** Testing a single game takes 8 minutes and **2.** There are 9! = 362880 choices for a batter lineup!

We improve on the former by simplifying the game even further; since reducing the number of states linearly improves runtime. We also allow fouls to end at-bats, which speeds up value iteration by 3x. Of course, this is all costs accuracy.

$$\text{game states} = \underset{0-2}{\text{balls}} \times \underset{0-1}{\text{strikes}} \times \underset{0-1}{\text{outs}} \times \underset{0-2}{\text{inning}} \times \underset{0-8}{\text{batter}} \times \underset{0-1}{\text{first}} \times \underset{0-1}{\text{second}} \times \underset{0-1}{\text{third}}$$

1296 states

Instead of searching through every permutation, we devise an algorithm to find the best permutation with a limited budget of searches. We gave our "OneByOne" strategy a budget of 120 searches, and it came up with the batting order listed above.



*Ultimately, simulating batting orders in this manner requires many simplifications in a game known for its subtleties. It's hard to say how an improvement in our model translates to real baseball strategy. More research is needed to make any concrete statements.

II Stochastic Games

Modeling baseball as a zero-sum stochastic game allows us to calculate optimal pitches and evaluate players.

This technique requires we define discrete actions and states.

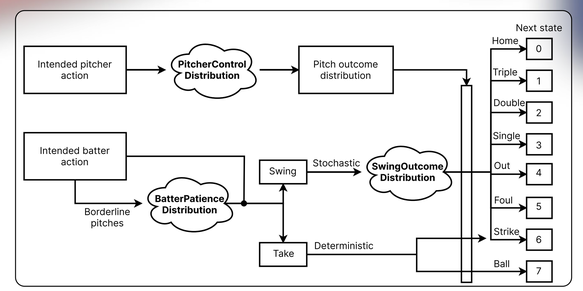
$$\text{pitcher actions} = \underset{0-5}{\text{types}} \times \underset{0-16}{\text{zones}} \qquad \text{batter actions} = \{\text{swing}, \text{take}\}$$

102 actions **2 actions**

$$\text{game states} = \underset{0-3}{\text{balls}} \times \underset{0-2}{\text{strikes}} \times \underset{0-2}{\text{outs}} \times \underset{0-8}{\text{inning}} \times \underset{0-8}{\text{batter}} \times \underset{0-1}{\text{first}} \times \underset{0-1}{\text{second}} \times \underset{0-1}{\text{third}}$$

23328 states

We use our previously trained models to calculate a **transition distribution**, mapping each combination of state, pitcher action, and batter action to a set of probabilistic outcome states.



The resulting data-structure looks like a 4d array, essentially an adjacency list for each pitcher-batter action pair. With simple preprocessing, the procedure can be entirely vectorized.

With the transition distribution in place, we can use value iteration to find a mixed strategy equilibrium. In practice, it takes optimizing ~200,000 linear programs for the values to converge.

The result is an expected run average for every state, along with probabilistic recommendations for each pitcher action.

Binyamin Friedman
Yevgeniy Vorobeychik

2024