Baseball Pitching Pattern Analyzer Using Double Layer Markov Models

Louisa Kim

Computer Science and Engineering

University of California, Riverside

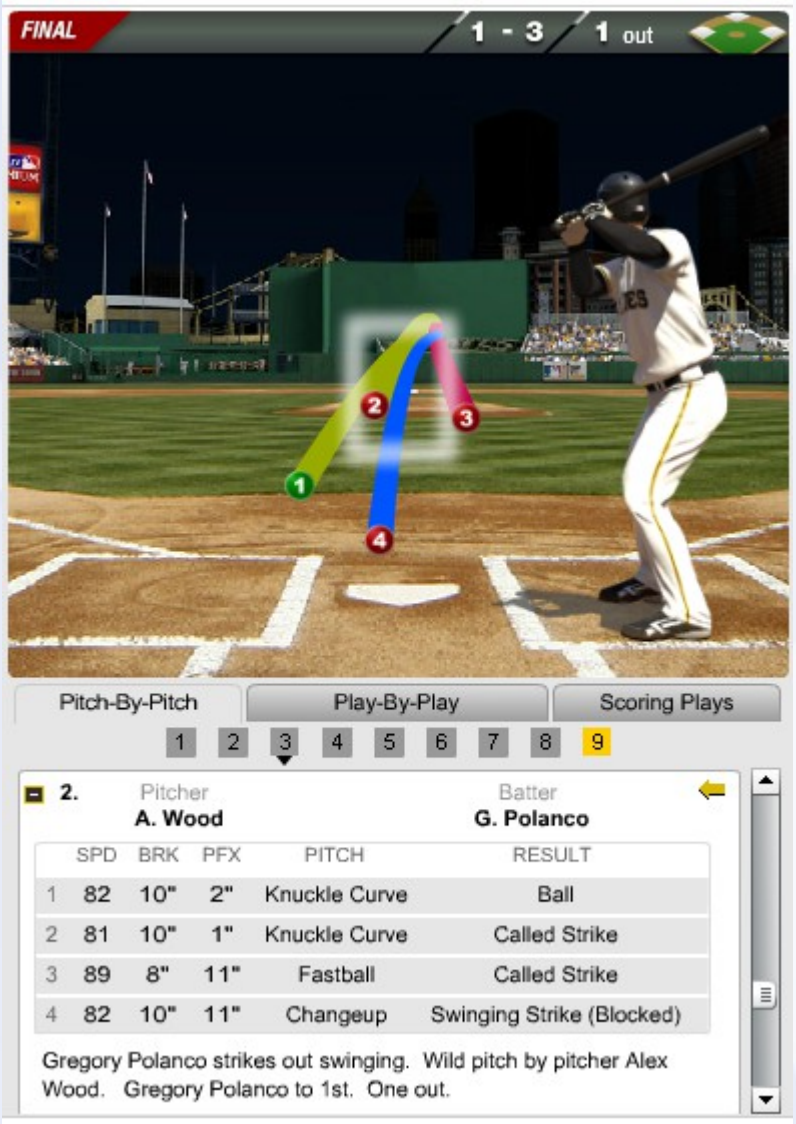
Outline

- Introduction
- Baseball Basics
- Definition as a Probabilistic Model
- Data Processing
- Components of HMM
- Viterbi Algorithm
- Results and Discussion

Introduction

- Raw data: Gameday app of MLB.com
- HMM techniques: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition by Lawrence Rabiner
- Clustering for codebook: k-means
- C++

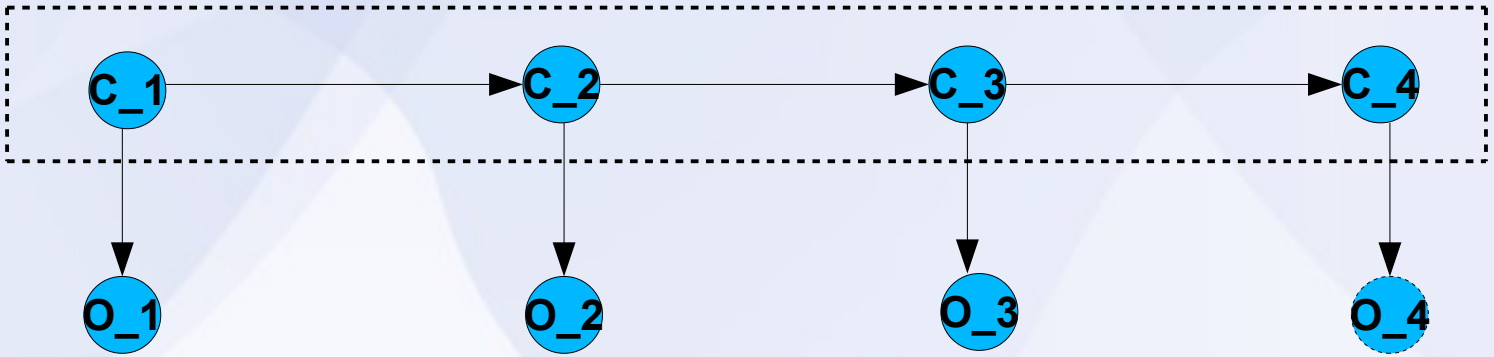
Baseball Basics



Definition as a Probabilistic Model

Let :
O_i *be the information about* i_{th} *pitch*
X_i *be the result of the* i_{th} *pitch* (*ball* or *strike*)
r *be the result of the at−bat* (*out* or *not−out*)
T *be the length of the at−bat* $(1,2,\ldots,6)$

Our goal is to *build a model* to *maximize* $P(O_1, X_1, \ldots, O_T, X_T | T, r)$.
If we let $C_n(X_1, \ldots, X_n)$ *be the count after n pitches with results* $X_1, \ldots, X_n$,
then we assume that $O_i \perp O_j | C_i(X_1, \ldots, X_i)$ *for* $j \neq i$ and *that* $X_{i+1} \perp O_j | C_i$ *for* $j < i$.



Data Processing

Input Folder	Dates	Size	# of Files	# of Data Points (Pitches)
1 Day	6/30/2015	3.9 MB	150	4529
5 Days	6/26/2015 ~ 6/30/2015	16 MB	650	19828
10 Days	6/21/2015 ~ 6/30/2015	30.7 MB	1229	38448
30 Days	6/1/2015 ~ 6/30/2015	92.4 MB	3752	116442

Data Processing (cont.)

{"top_inning":"Y","s":"0","b":"0","reason":"","ind":"F","status":"Final","o":"3","inning":"9","inning_state":"","note":""},"home_loss":"25","home_games_back":"-","home_code":"nya","away_sport_code":"mlb","home_win":"32","time_hm_lg":"1:05","away_name_abbrev":"LAA","league":"AA","time_zone_aw_lg":"-4","away_games_back":"5.5","home_file_code":"nyy","game_data_directory":"/components/game/mlb/year_2015/month_06/day_07/gid_2015_06_07_anamlb_nyamlb_1","time_zone":"ET","away_league_id":"103","home_team_id":"147","day":"SUN","time_aw_lg":"1:05","away_team_city":"LA

http://mlb.mlb.com/gdcross/components/game/mlb/year_2015/month_05/day_01/gid_2015_05_01_detmlb_kcamlb_1/inning/inning_1.xml?live

<atbat num="1" b="1" s="2" o="0" start_tfs="230838" start_tfs_zulu="2015-06-30T23:08:38Z" batter="570256" stand="L" b_height="6-5" pitcher="434378" p_throws="R" des="Gregory Polanco singles on a fly ball to left fielder Yoenis Cespedes. " des_es="Gregory Polanco pega sencillo con elevado a jardinero izquierdo Yoenis Cespedes. " event_num="8" event="Single" event_es="Sencillo" play_guid="300f6b47-d2dc-4830-96dd-9f9476b14829" home_team_runs="0" away_team_runs="0"><pitch des="Foul" des_es="Foul" id="3" type="S" tfs="230906" tfs_zulu="2015-06-30T23:09:06Z" x="165.07" y="157.51" event_num="3" sv_id="150630_191004" play_guid="df958229-8b84-4fe0-8160-45ddffb684f1" start_speed="91.8" end_speed="84.8" sz_top="3.91" sz_bot="1.81" pfx_x="-7.59" pfx_z="10.43" px="-1.261" pz="3.01" x0="-2.095" y0="50.0" z0="6.624" vx0="4.834" vy0="-134.326" vz0="-7.154" ax="-13.92" ay="28.163" az="-12.987" break_y="23.8" break_angle="41.8" break_length="4.3" pitch_type="FF" type_confidence=".904" zone="11" nasty="65" spin_dir="215.959" spin_rate="2560.204" cc="" mt=""/>

Data Processing (cont.)

Single Called S 88.49 184.7 93.0 FF 14

Single In X 120.13 164.48 91.3 FF 5

/ 2

Strikeout Called S 82.08 181.24 91.5 FF 14

Strikeout Ball B 73.62 180.14 91.5 FF 14

Strikeout Called S 92.64 177.52 78.1 SL 9

Strikeout Ball B 64.4 223.17 78.9 SL 14

Strikeout Foul S 115.28 166.93 92.3 Sl 5

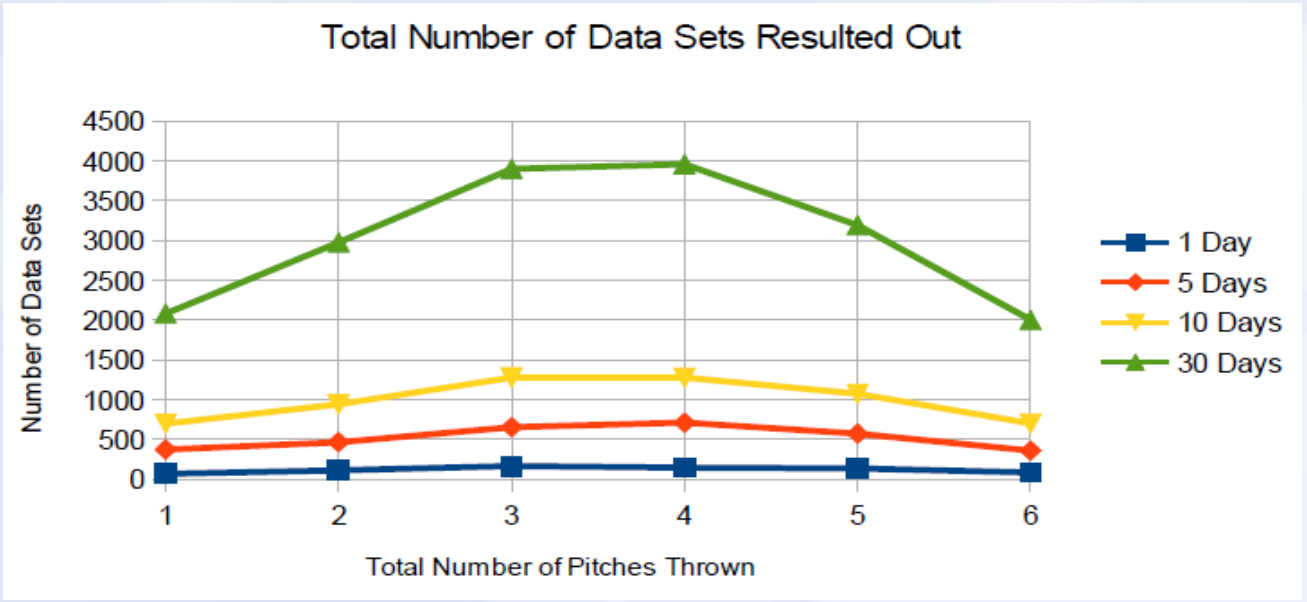
Strikeout Called S 108.8 154.73 92.7 FF 2

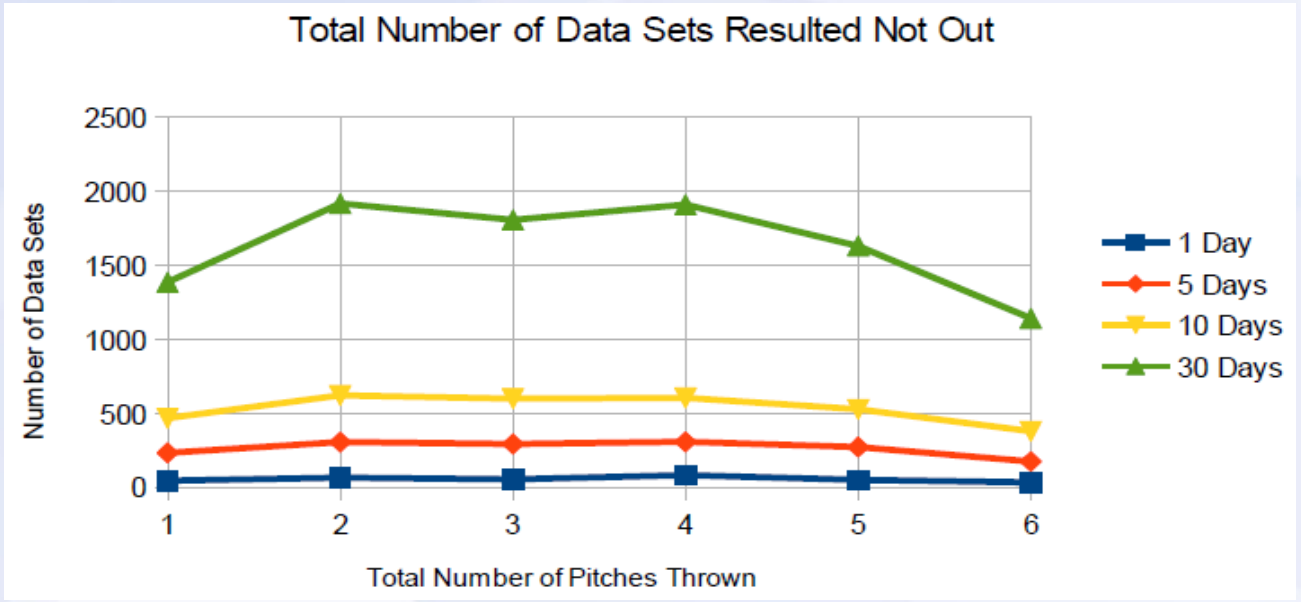
/ 6

Data Processing (cont.)

Tag	Event	Call	Pitch-Type	Zone	x	y	Speed	Count
61	11	0	14	13	137	221	86	10
62	11	8	5	12	84	165	89	11
63	11	0	14	13	204	223	85	21
64	11	8	5	13	144	208	89	22
65	11	4	5	8	122	191	90	0
71	15	1	8	5	125	179	83	1
72	15	1	1	11	149	158	67	2
73	15	0	5	11	195	67	83	12
74	15	0	1	11	121	153	67	22
75	15	4	8	12	83	157	84	0

Data Set Statistics





Components of HMM

$$\lambda = (A, B, \pi)$$

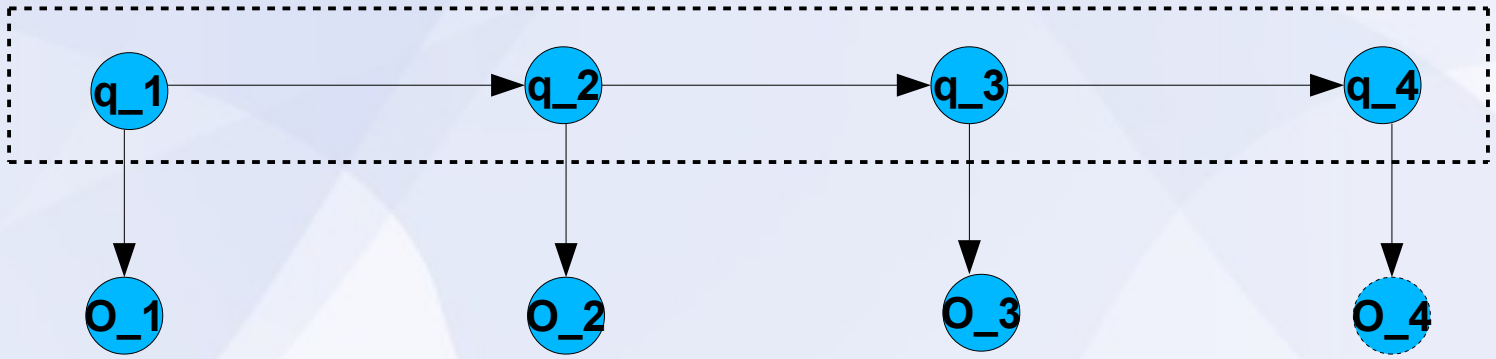
$$A = \{a_{ij}\} \text{ where } a_{ij} = P[q_{t+1} = S_j \mid q_t = S_i],\ 1 \le i, j \le N$$

$$B = \{b_j(k)\} \text{ where } b_j(k) = P[v_k \text{ at } t \mid q_t = S_j],\ 1 \le j \le N \text{ and } 1 \le k \le M$$

$$\pi = \{\pi_i\} \text{ where } \pi_i = P[q_1 = S_i],\ 1 \le i \le N$$

Left-Right Model

$$a_{ij} = 0 \text{ for } j < i \text{ and } \sum_{j=1}^{N} a_{ij} = 1 \text{ for } 1 \le i \le N$$



Computed Pi, A, B for T=5

Pi

1	0.563959
10	0.436041

B

Observation Probability Distribution with size 6736

t	Count	Prob	CodebookInd
1	1	0.00111982	1
1	1	0.00111982	4
1	1	0.00055991	5
1	1	0.00223964	6
1	1	0.00167973	7
1	1	0.00223964	8
1	1	0.00111982	10
1	1	0.00167973	11

A

	1	2	10	11	12	20	21	22	30	31	0
1	0	0.456954	0	0.543046	0	0	0	0	0	0	0
2	0	0.243836	0	0	0.709589	0	0	0	0	0	0.0465753
10	0	0	0	0.633833	0	0.366167	0	0	0	0	0
11	0	0	0	0	0.578526	0	0.421474	0	0	0	0
12	0	0	0	0	0.226933	0	0	0.477556	0	0	0.295511
20	0	0	0	0	0	0	0.748538	0	0.251462	0	0
21	0	0	0	0	0	0	0	0.777494	0	0.222506	0
22	0	0	0	0	0	0	0	0	0	0	1
30	0	0	0	0	0	0	0	0	0	1	0
31	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	1

B

- Create codebook using k-means clustering
- Vector quantization of observation vectors using codebook
- Compute observation probabilities B on pg. 12 using vector quantization and counting

Observation with size 15955

t	Count	Pitch-Type	Zone	Speed	x	y
1	1	0	1	82	136	152
1	1	0	1	83	142	148
1	1	0	1	83	142	157
1	1	0	1	85	137	157
1	1	0	2	80	118	156
1	1	0	2	82	125	158
1	1	0	2	83	121	164
1	1	0	2	87	117	148
1	1	0	2	89	110	152

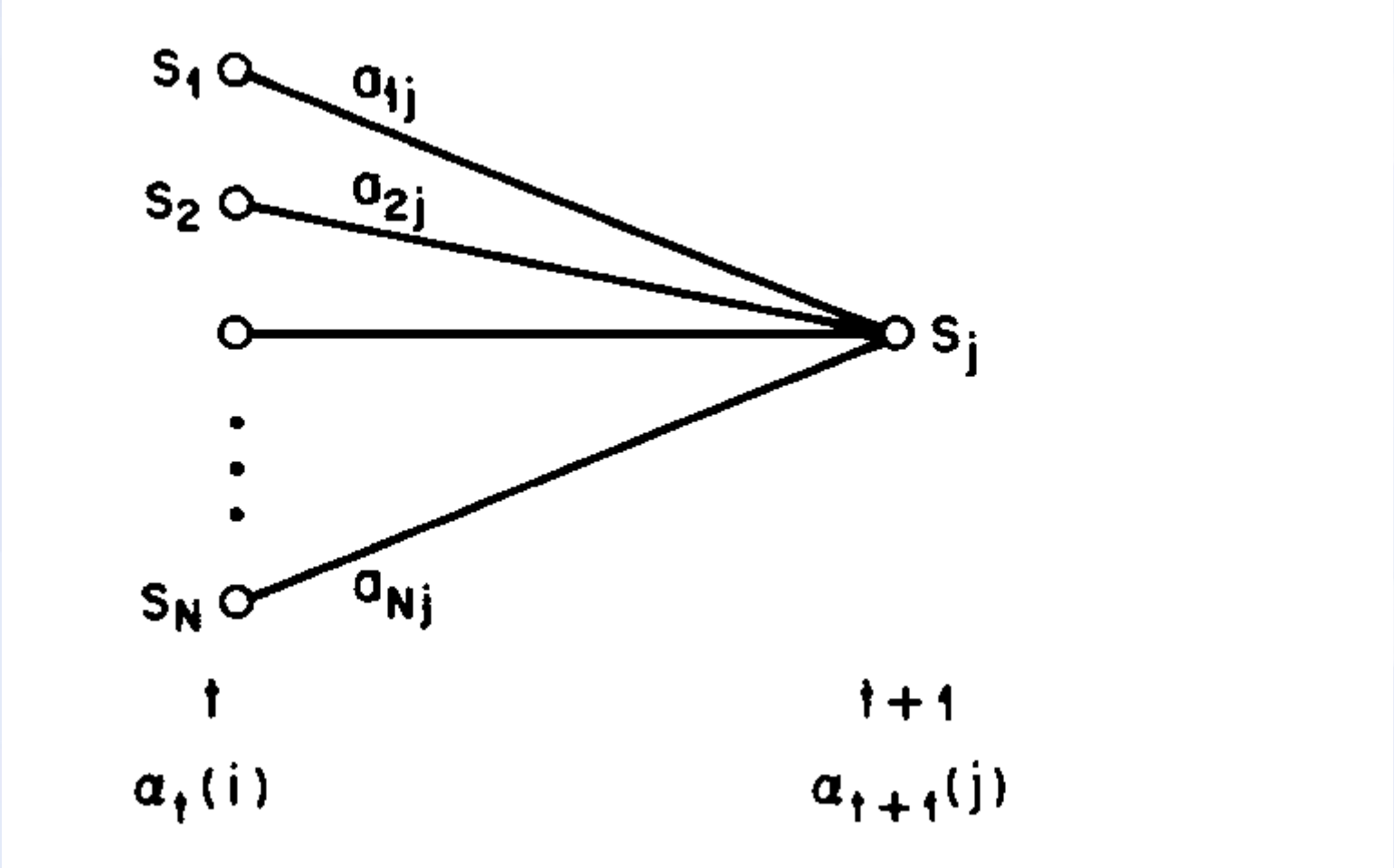
Codebook with size 1041

0	0	10	81	132	136
1	0	1	81	131	149
2	0	10	81	111	134
3	0	10	82	93	141
4	0	3	82	120	152
5	0	2	81	129	161
6	0	3	81	136	170
7	0	2	83	138	159
8	0	4	82	124	167
9	0	4	77	114	169

Vector Quantization with size 15955

t	Count	CodebookInd
1	1	1
1	1	1
1	1	4
1	1	4
1	1	5
1	1	6
1	1	6
1	1	6
1	1	6

Viterbi Algorithm



Viterbi Algorithm (cont.)

1) Initialization:

$$\delta_1(i) = \pi_i b_i(O_1), \qquad 1 \le i \le N \tag{32a}$$

$$\psi_1(i) = 0. \tag{32b}$$

2) Recursion:

$$\delta_t(j) = \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}] b_j(O_t), \qquad 2 \le t \le T, \quad 1 \le j \le N \tag{33a}$$

$$\psi_t(j) = \operatorname*{argmax}_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}], \qquad 2 \le t \le T, \quad 1 \le j \le N. \tag{33b}$$

3) Termination: (34a)

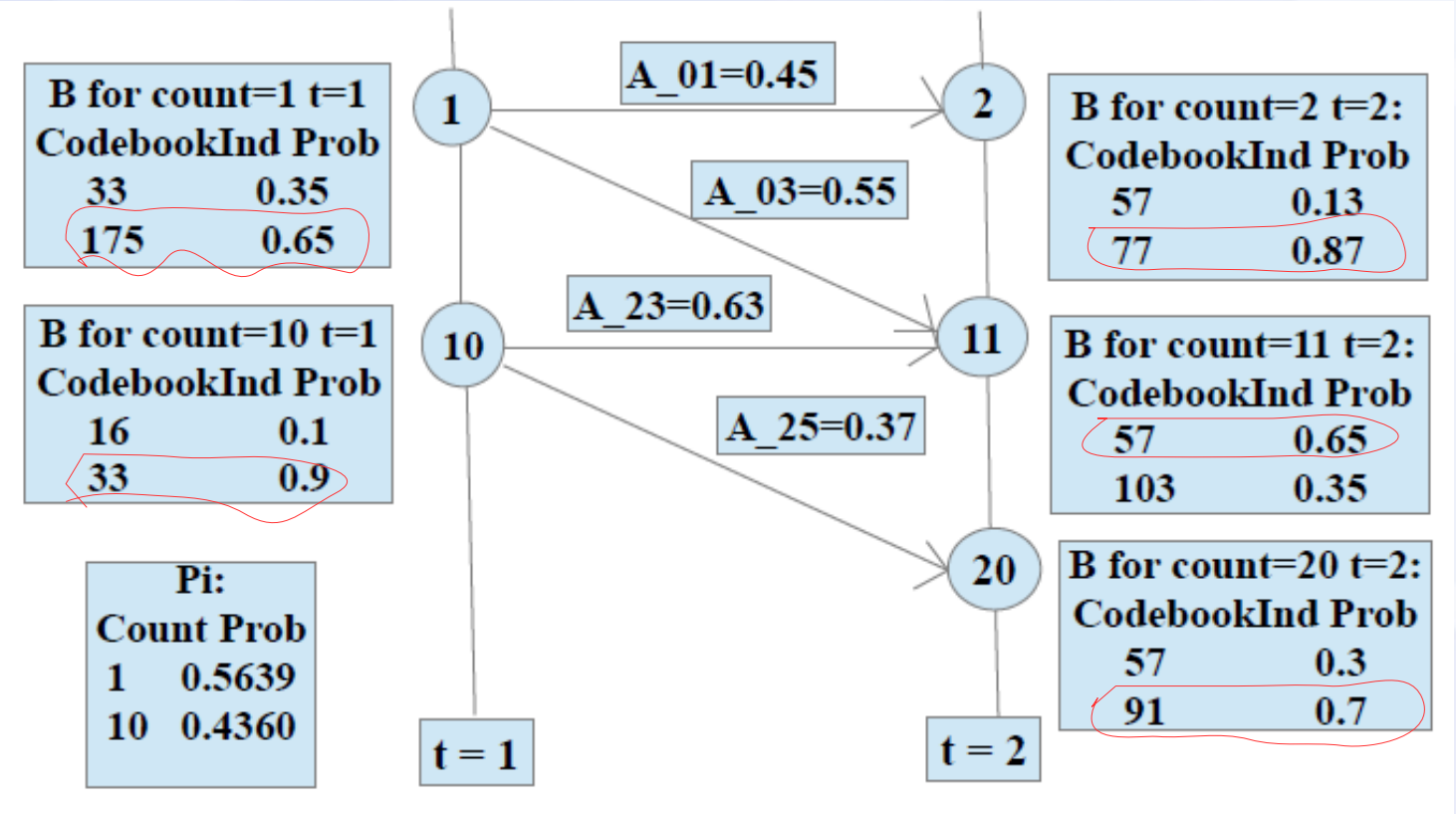
$$P^* = \max_{1 \le i \le N} [\delta_T(i)]$$

$$q_T^* = \operatorname*{argmax}_{1 \le i \le N} [\delta_T(i)]. \tag{34b}$$

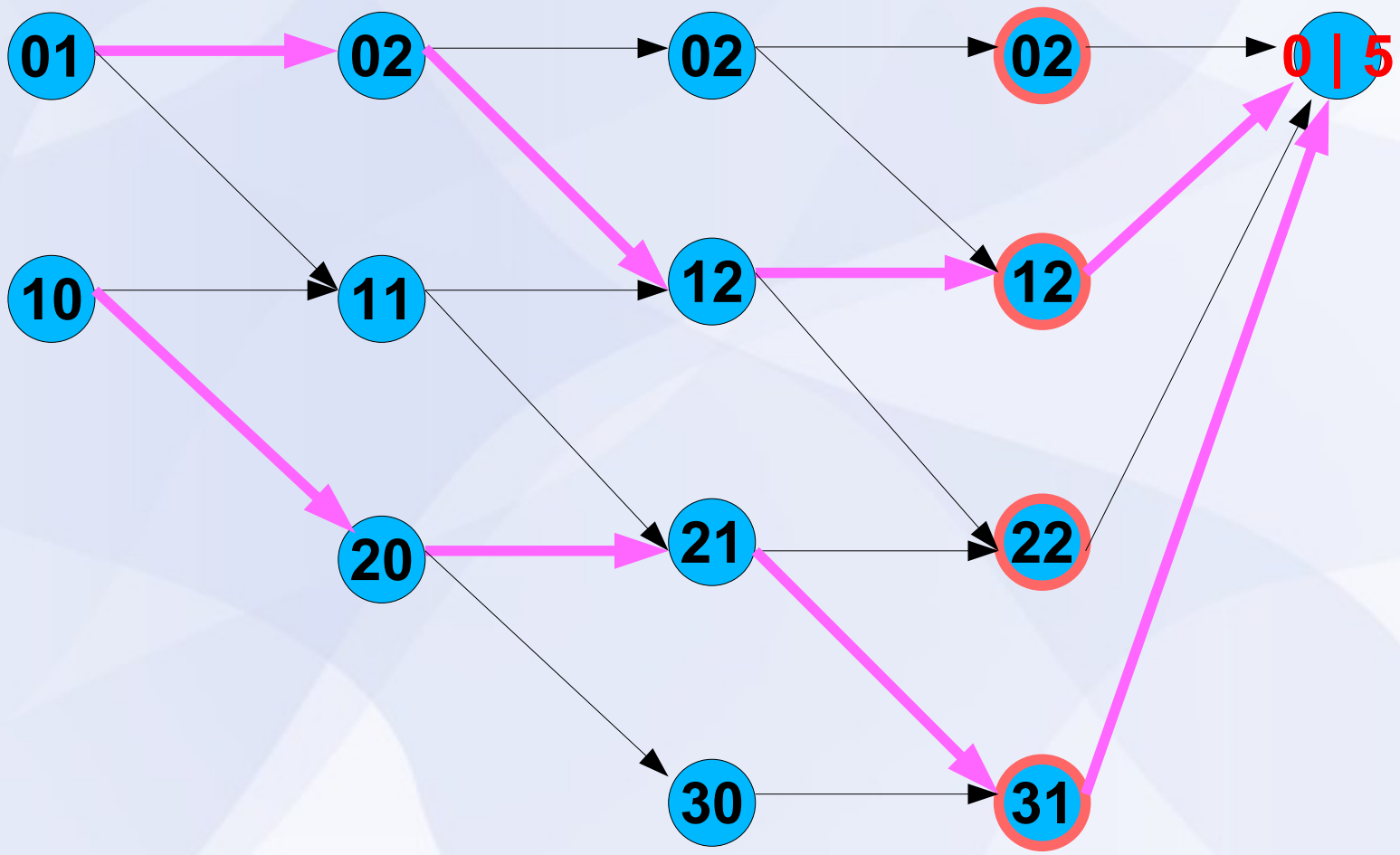
4) Path (state sequence) backtracking:

$$q_t^* = \psi_{t+1}(q_{t+1}^*), \qquad t = T-1, T-2, \cdots, 1. \tag{35}$$

Modified Viterbi



Non-zero Trans. Lattice for T=5



Results

Top 2 out of top 30 output for 5 days, 10 days, 30 days data [count, pitch-type, zone, speed]

T = 3 out	1 FF 3 93 2 FS 10 83 0 FF 12 93 1 KC 9 79 2 SL 5 83 0 KC 9 79	1 FF 9 92 2 FF 5 93 0 CU 9 78 1 CU 3 78 2 SI 11 90 0 KC 9 80	10 FF 13 92 20 FT 13 90 0 SL 6 85 1 FF 3 93 2 CU 8 77 0 FF 4 92
T = 3 not out	10 FT 12 91 20 FF 12 91 5 FT 13 91 10 CU 12 76 20 CH 13 82 5 FF 4 92	10 FC 12 88 20 FF 12 91 5 FF 5 91 10 FF 12 92 20 FF 11 92 5 FF 8 91	10 FF 12 91 20 FF 12 91 5 FF 11 91 10 FF 13 92 20 FF 13 92 5 FT 11 91
T = 4 out	10 FF 11 92 20 CH 13 82 21 FT 14 90 0 KC 13 80 10 CU 12 76 20 FT 13 90 30 FF 12 91 0 KC 9 79	10 FF 12 92 20 FF 13 92 21 CH 13 83 0 CU 12 77 10 FF 11 91 20 SL 14 85 21 CH 12 83 0 CU 13 78	10 FF 12 91 20 FF 11 92 21 FF 11 92 0 FF 11 92 10 FF 13 91 20 FF 13 92 21 SL 13 84 0 CU 12 76
T = 4 not out	10 IN 11 73 20 IN 11 73 30 IN 11 73 5 IN 11 73 10 IN 11 69 20 IN 11 69 30 IN 11 69 5 IN 11 69	10 IN 11 74 20 IN 11 74 30 IN 11 68 5 IN 11 68 10 IN 11 68 20 IN 11 68 30 IN 11 74 5 IN 11 74	10 IN 11 72 20 IN 11 72 30 IN 11 72 5 IN 11 72 10 IN 11 73 20 IN 11 73 30 IN 11 73 5 IN 11 73

Results (cont.)

Other results from top 30 output resulted out for 5 days, 10 days, 30 days data

5 Days

Count	Index	Pitch-Type	Zone	Speed	x	y
1	51	CU	8	77	119	187
2	414	SL	11	83	146	184
0	131	FF	3	93	129	168

Count	Index	Pitch-Type	Zone	Speed	x	y
1	355	SI	8	89	98	185
2	70	CU	13	77	98	204
0	278	FT	4	91	138	172

Count	Index	Pitch-Type	Zone	Speed	x	y
10	63	CU	13	76	153	206
20	425	SL	13	83	47	183
30	89	FC	11	87	147	198
0	225	FF	7	92	140	190

Count	Index	Pitch-Type	Zone	Speed	x	y
1	332	KC	6	79	132	171
2	109	FF	1	91	130	154
12	416	SL	13	83	154	196
0	380	SL	2	84	121	158

10 Days

Count	Index	Pitch-Type	Zone	Speed	x	y
1	313	FF	12	90	154	176
2	182	FF	3	94	133	166
0	558	SL	11	83	145	189

Count	Index	Pitch-Type	Zone	Speed	x	y
1	115	FC	3	88	125	164
2	8	CH	5	82	123	176
0	536	SL	8	86	112	189

Count	Index	Pitch-Type	Zone	Speed	x	y
10	427	FT	13	91	166	197
20	20	CH	11	82	151	152
21	357	FF	14	93	83	185
0	99	CU	13	78	91	201

Count	Index	Pitch-Type	Zone	Speed	x	y
10	429	FT	12	90	173	181
20	348	FF	14	91	82	207
30	44	CH	13	83	119	202
0	217	FF	4	94	137	175

30 Days

Count	Index	Pitch-Type	Zone	Speed	x	y
1	830	SI	5	91	114	171
2	789	KC	8	79	114	187
0	1023	SL	13	84	103	218

Count	Index	Pitch-Type	Zone	Speed	x	y
1	382	FF	5	90	116	175
2	338	FF	3	94	118	158
0	213	FC	11	88	151	159

Count	Index	Pitch-Type	Zone	Speed	x	y
10	593	FF	13	92	110	208
20	76	CH	13	83	144	210
21	535	FF	13	90	161	199
0	928	SL	7	86	121	187

Count	Index	Pitch-Type	Zone	Speed	x	y
10	484	FF	11	93	83	138
20	891	SI	14	90	78	191
21	76	CH	13	83	144	210
0	369	FF	7	92	124	191

Discussion

	5 Days Data	10 Days Data	30 Days Data
Execution Time	21.14 seconds	50.59 seconds	230.19 seconds

- Printed top 30 results for Total Number of Pitches Thrown = 1, 2, …, 6
- Printing top 100 results is also possible to broaden selection pool
- There are results seem unreasonable, we discard them.
- Printing results for a particular player (i.e. batter=570256 or pitcher=434378 on pg. 7) is also possible by adding a few lines of code.
- Printing results for a particular type of play event (i.e. single, double, strikeout, etc.) is also possible by adding a few lines of code.

Appendix

CH	CU	EP	FA	FC	FF	FO	FS
changeup	curveball	eephus	fastball	cutter	4-seam FA	pitch out	sinking FA
FT	IN	KC	KN	PO	SI	SL	UN
2-seam FA	intent ball	knuckle-curve	knuckleball	pitch out	sinker	slider	unidentified