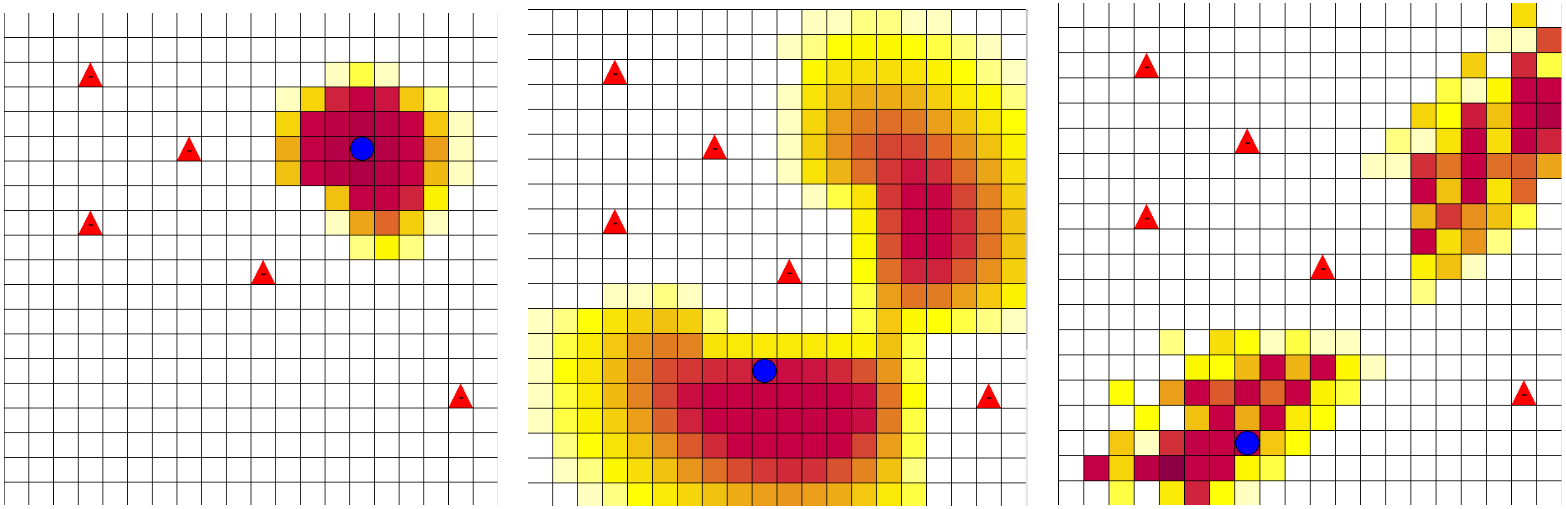
Hidden Markov Models

"…,99,100! Markov, here I come!"



16.410/413 Principles of Autonomy and Decision-Making

Pedro Santana (psantana@mit.edu)

October 7th , 2015.

Based on material by
Brian Williams and Emilio Frazzoli.

Assignments

- Problem set 4
  - Out last Wednesday.
  - Due at midnight **tonight**.
- Problem set 5
  - Out today and **due in a week**.
- Readings
  - Today: “Probabilistic Reasoning Over Time” [AIMA], Ch. 15.

Today's topics

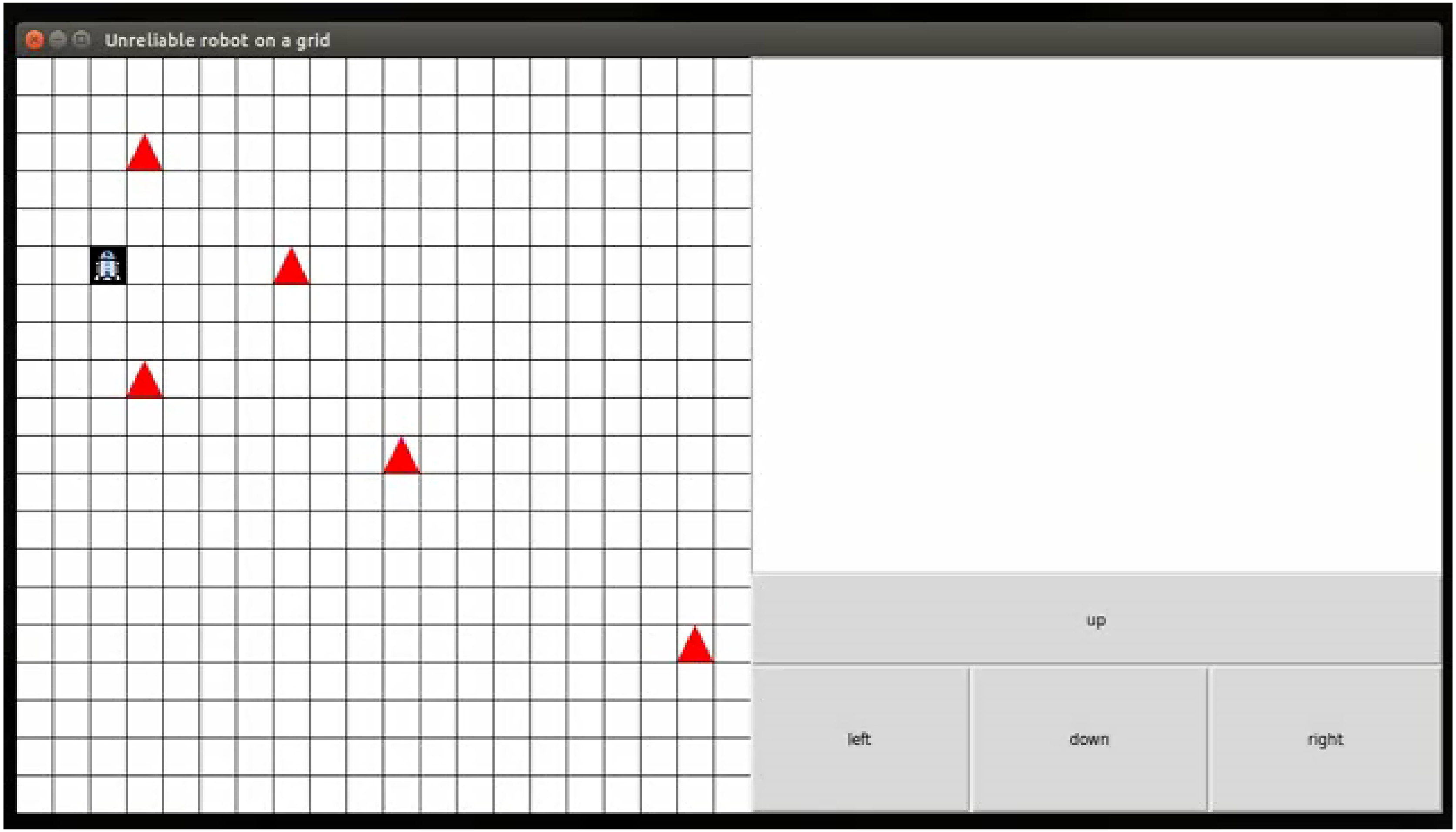
1. Motivation
2. Probability recap
   - Bayes' Rule
   - Marginalization
3. Markov chains
4. Hidden Markov models
5. HMM algorithms
   - Prediction
   - Filtering
   - Smoothing
   - Decoding
   - Learning (Baum-Welch)

Won't be covered today and significantly more involved, but you might want to learn more about it.

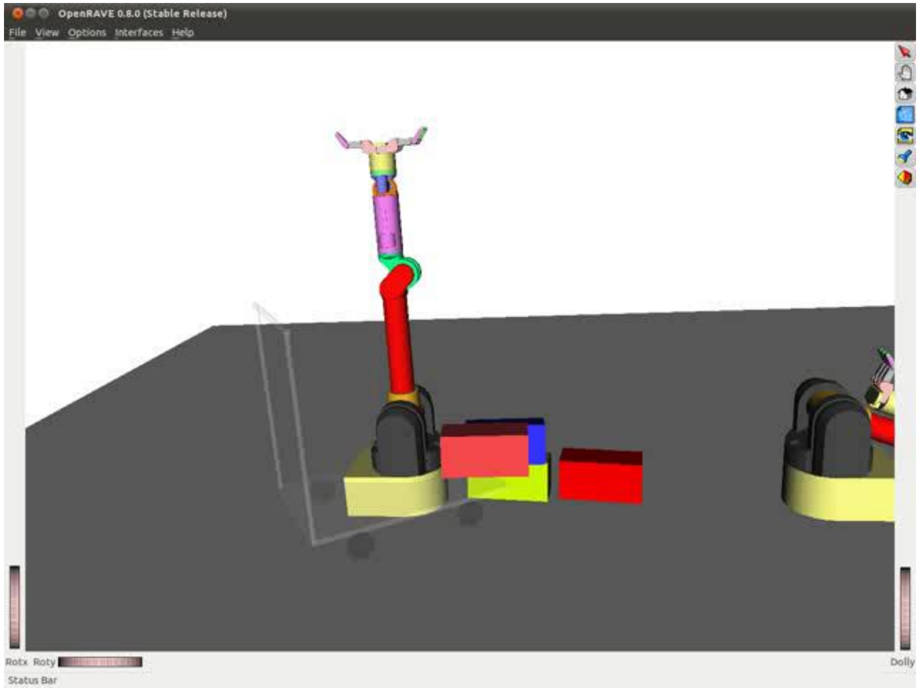
1. Motivation

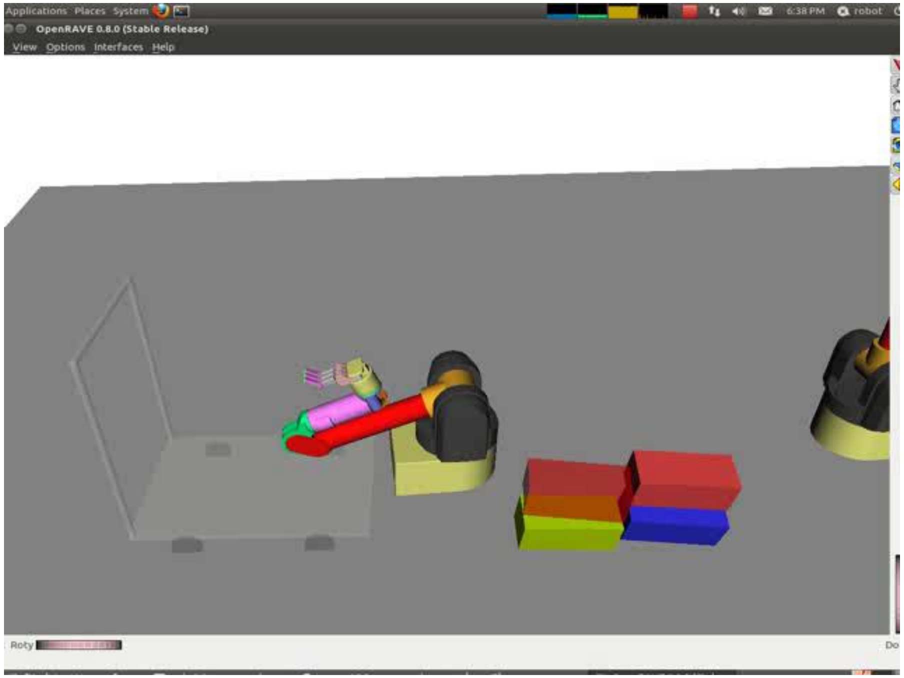
Why are we learning this?

Robot navigation

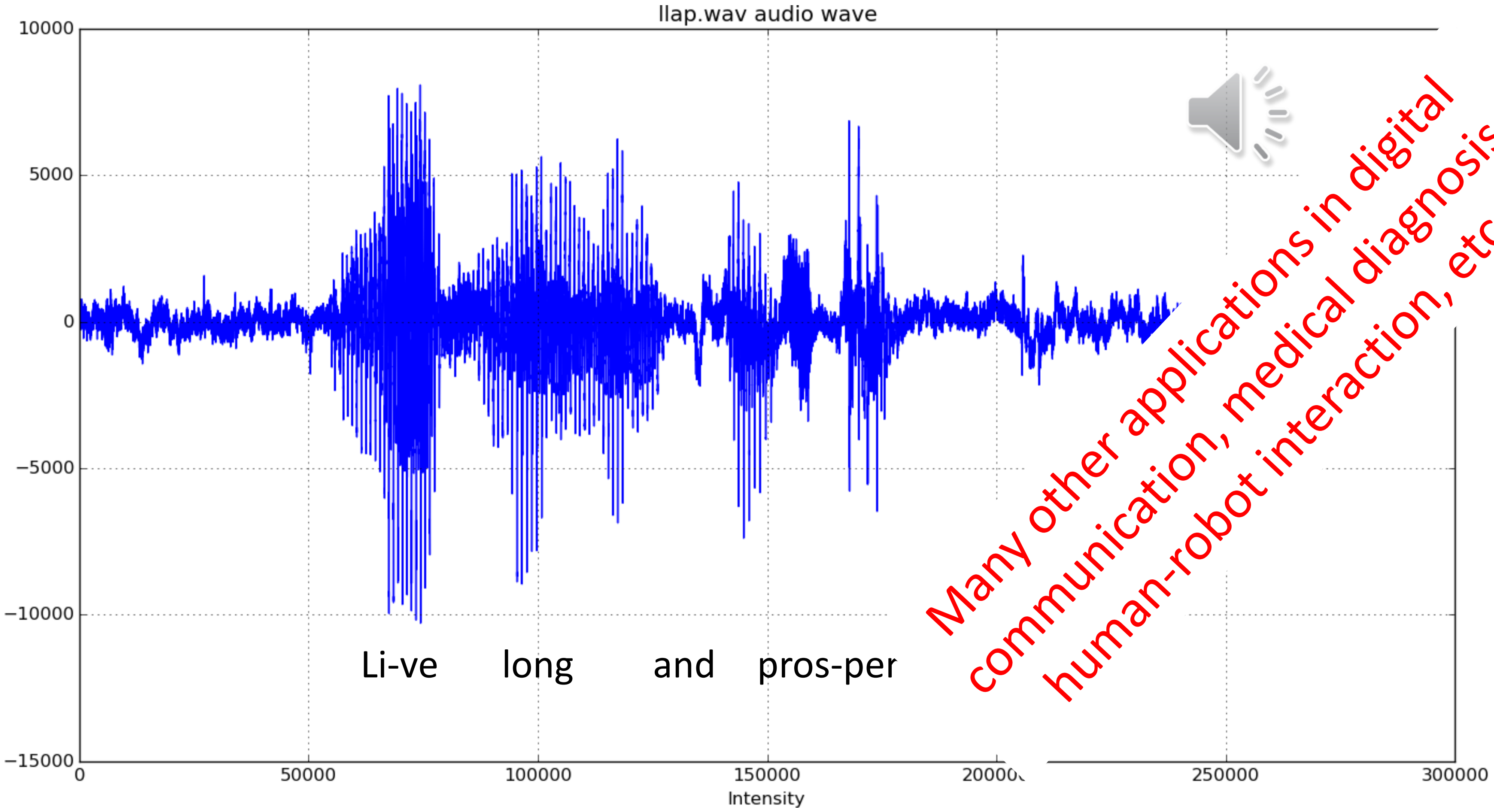


Robust sensor fusion (visual tracking)





Natural language processing (NLP)

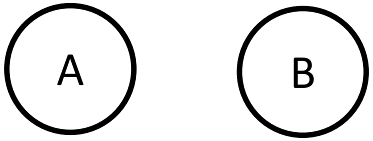


Many other applications in digital communication, medical diagnosis, human-robot interaction, etc.!

2. Probability recap

> “Probability is common sense reduced to calculation.”
>
> —— Pierre-Simon Laplace

Bayes’ rule



A, B: random variables

$$\overbrace{\Pr(A,B)}^{\text{Joint}} = \overbrace{\Pr(\mathrm{A}|\mathrm{B})}^{\text{Conditional}}\overbrace{\Pr(\mathrm{B})}^{\text{Marginal}}$$

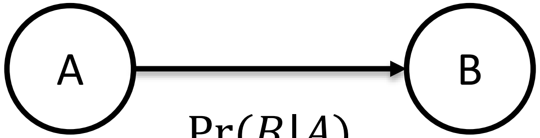
$$\Pr(A,B) = \Pr(\mathrm{B}|\mathrm{A})\Pr(\mathrm{A})$$

$$\Pr(\mathrm{A}|\mathrm{B})\Pr(\mathrm{B}) = \Pr(\mathrm{B}|\mathrm{A})\Pr(\mathrm{A})$$

Bayes’ rule!

$$\Pr(\mathrm{A}|\mathrm{B}) = \frac{\Pr(\mathrm{B}|\mathrm{A})\Pr(\mathrm{A})}{\Pr(\mathrm{B})} \propto \Pr(\mathrm{B}|\mathrm{A})\Pr(\mathrm{A})$$

Marginalization & graphical models



$\Pr(B|A)$

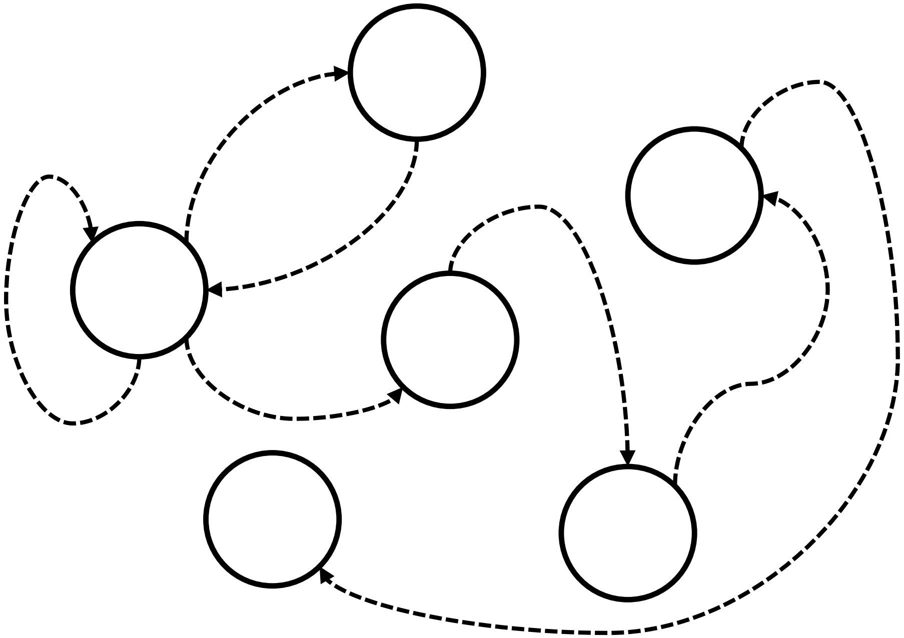
A "causes" B

$$\underbrace{\Pr(B)}_{\text{Distribution of the "effect" B}} = \sum_a \overbrace{\Pr(A=a,B)}^{\text{Marginalizes A out}} = \sum_a \underbrace{\Pr(\mathrm{B}|\mathrm{A}=a)}_{\text{Conditioning on "cause" makes the computation easier.}} \overbrace{\Pr(\mathrm{A}=\mathrm{a})}^{\text{Prior on "cause"}}$$

Our goal for today

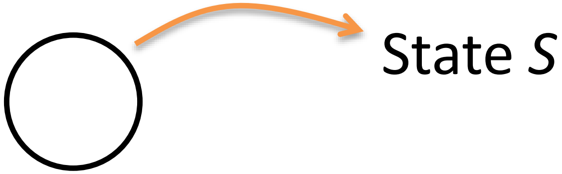
How can we ***estimate*** *the* ***hidden state*** *of a system from* ***noisy sensor observations****?*

3. Markov chains



Andrey Markov

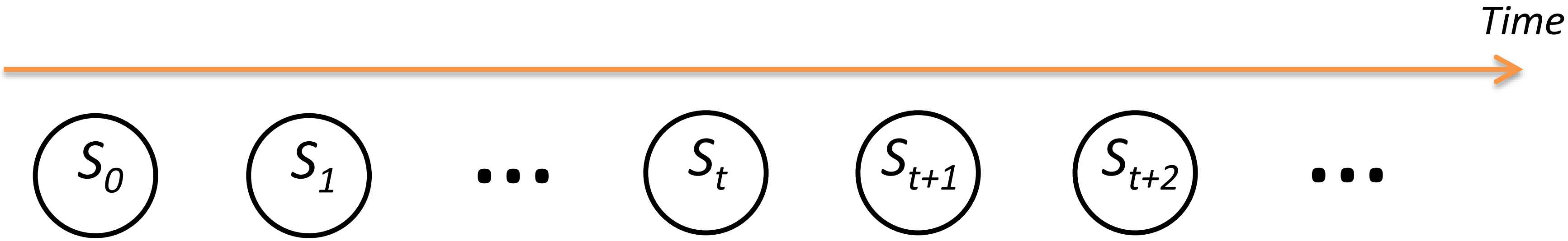
State transitions over time



S_t: state at time t (random variable)

$S_t=s$: particular value of S_t (not random)

$s \in \mathbb{S}$, $\mathbb{S}$ is the *state space.*



State transitions over time

$$\Pr(S_0, S_1, \ldots, S_t, S_{t+1}) = \Pr(S_{0:t+1})$$

$$\Pr(S_{0:t+1}) = \Pr(S_0)\Pr(S_1|S_0)\Pr(S_2|S_{0:1})\Pr(S_3|S_{0:2})\Pr(S_4|S_{0:3})\ldots$$

$\Pr(S_t|S_{0:t-1})$ → "Past influences present" models

→ Models grow exponentially with time!

The Markov assumption

$$\Pr(S_t|S_{0:t-1}) = \boxed{\Pr(S_t|S_{t-1})}$$

Constant size! ☺

"Path" to S_t isn't relevant, given knowledge of S_{t-1}.

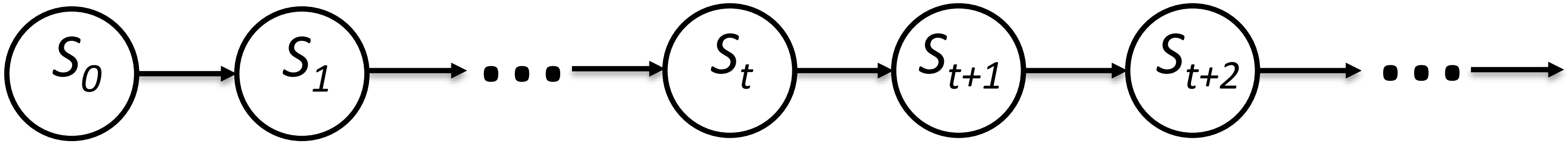
Definition: Markov chain

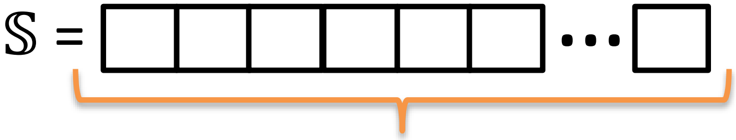
If a sequence of random variable $S_0, S_1, \ldots, S_{t+1}$ is such that

$$\Pr(S_{0:t+1}) = \Pr(S_0)\Pr(S_1|S_0)\Pr(S_2|S_1)\ldots = \Pr(S_0)\prod_{i=1}^{t+1}\Pr(S_i|S_{i-1}),$$

we say that $S_0, S_1, \ldots, S_{t+1}$ form a **Markov chain**.

Markov chains





Discrete set with *d* values.

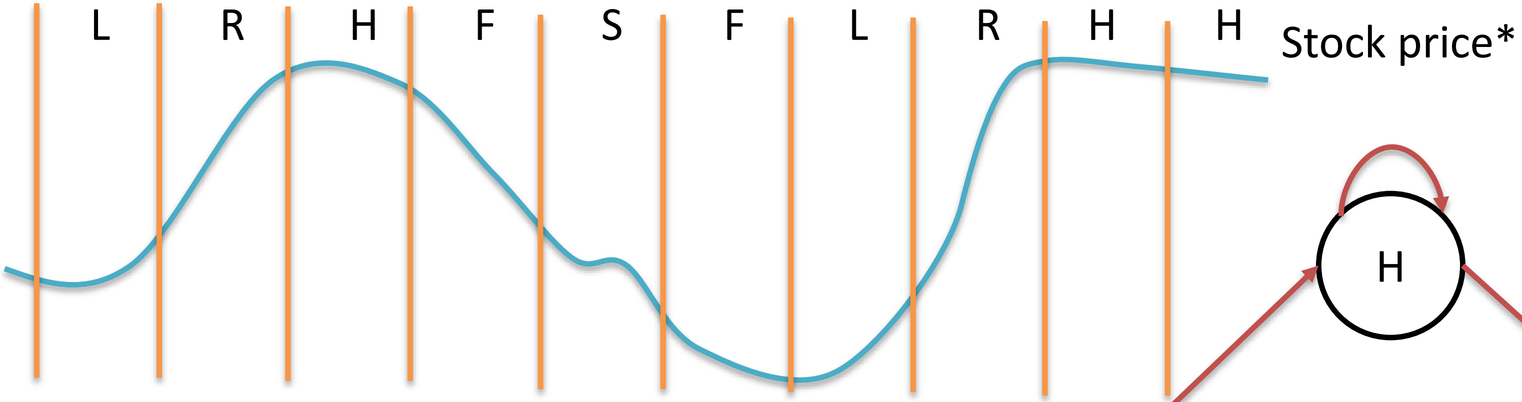
→ $\Pr(S_t|S_{t-1})$: $d \times d$ matrix T^t

$$T^t_{i,j} = \Pr(S_t = i|S_{t-1} = j)$$

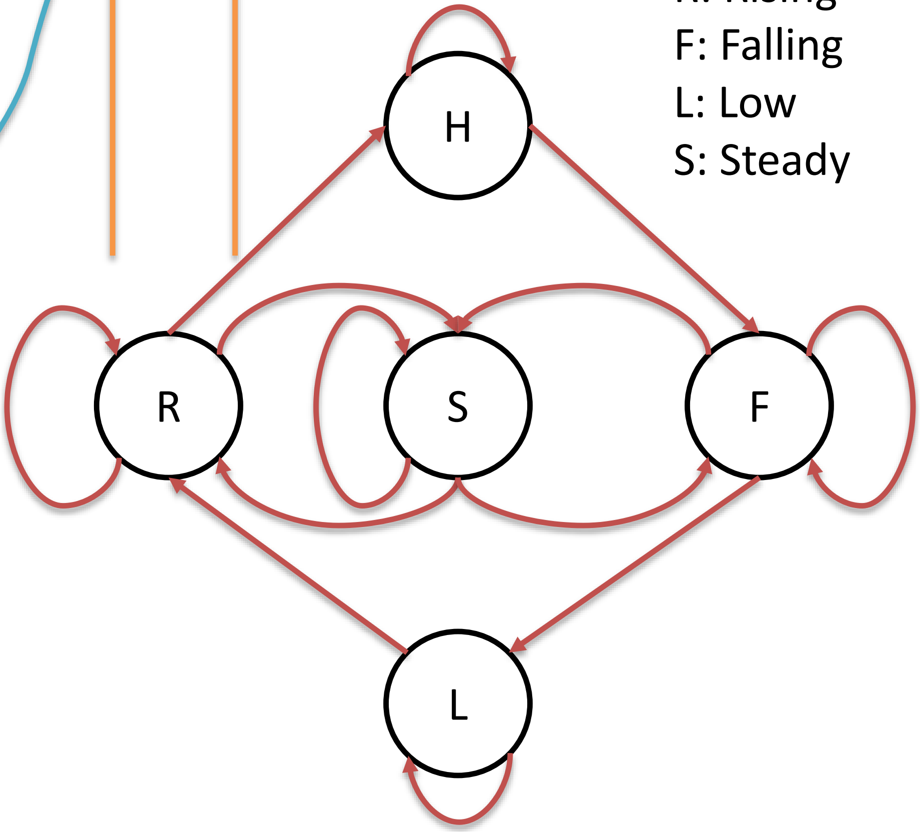
If T^t does not depend on t → Markov chain is ***stationary***.

$$T_{i,j} = \Pr(S_t = i|S_{t-1} = j)\,, \forall t$$

(Very) Simple Wall Street



H: High
R: Rising
F: Falling
L: Low
S: Steady

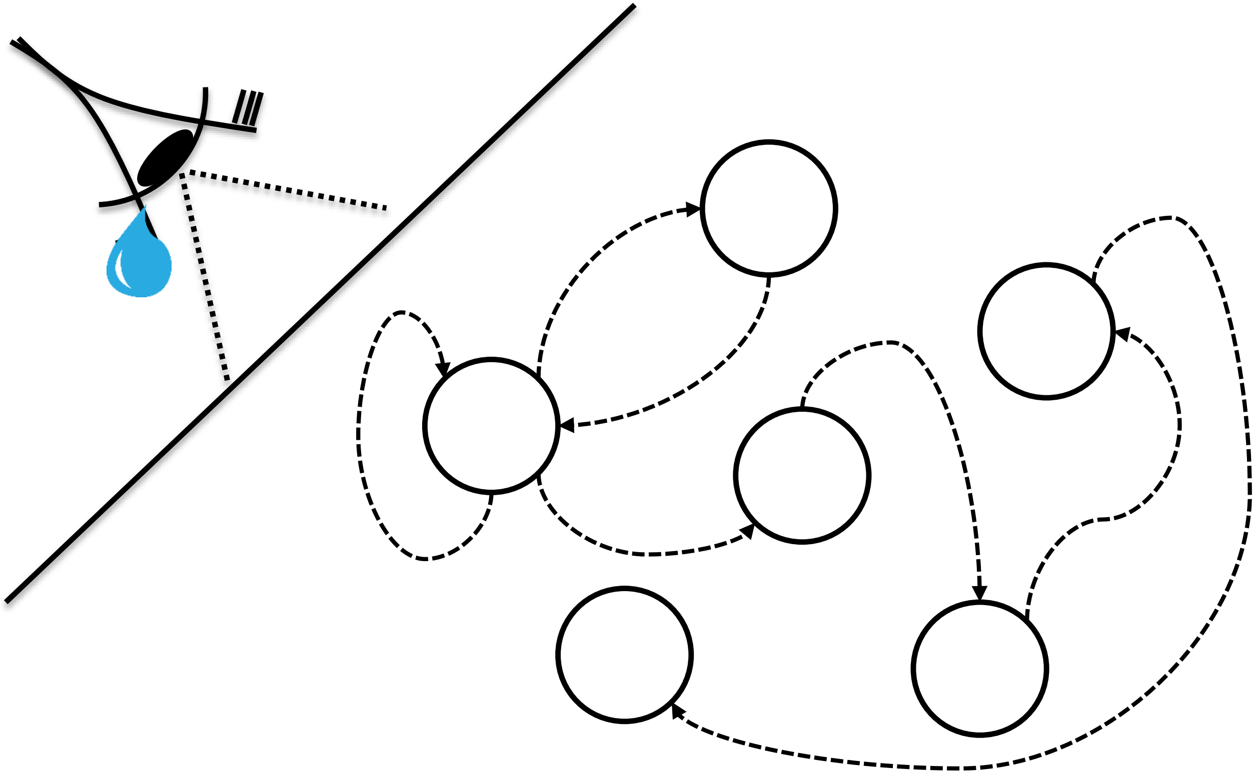


T

	H_{k-1}	R_{k-1}	F_{k-1}	L_{k-1}	S_{k-1}
H_k	0.1	0.05		0.2	
R_k		0.5		0.8	0.25
F_k	0.9		0.6		0.25
L_k			0.1		
S_k		0.45	0.3		0.5

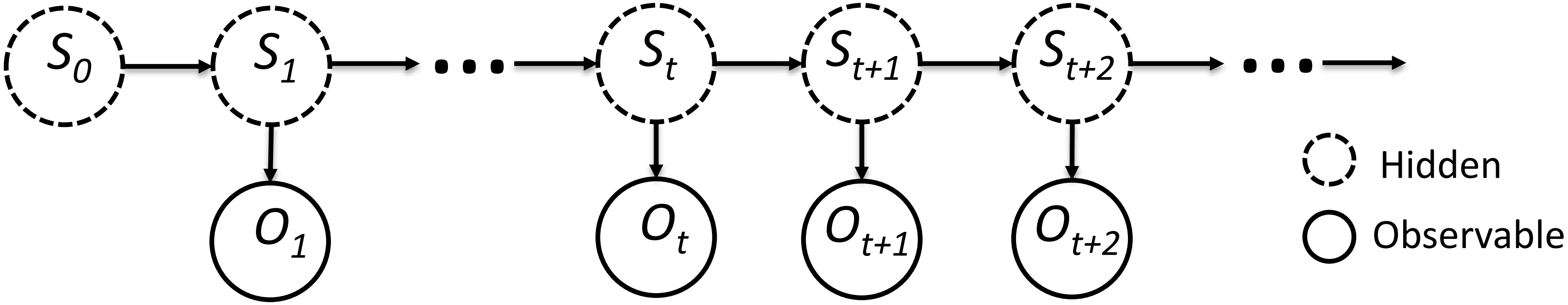
*Pedagogical example. In no circumstance shall the author be responsible for financial losses due to decisions based on this model.

4. Hidden Markov models (HMMs)



Andrey Markov

Observing hidden Markov chains



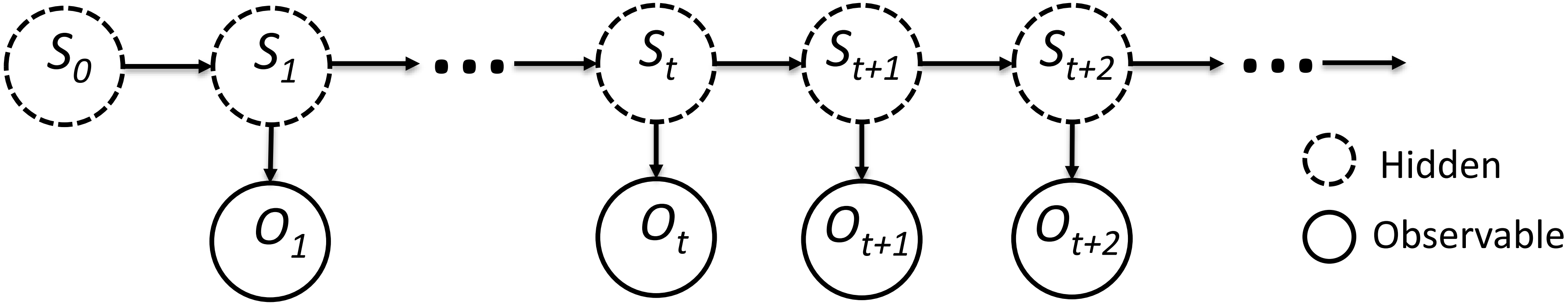
Definition: Hidden Markov Model (HMM)

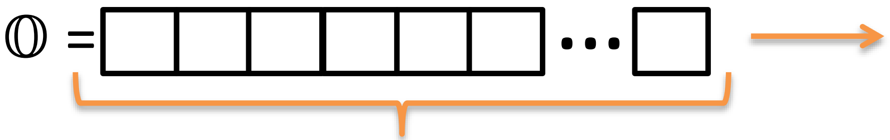
A sequence of random variables $O_1, O_2, \ldots, O_t, \ldots$, is an HMM if the distribution of O_t is completely defined by the current (hidden) state S_t according to

$$\Pr(O_t | S_t),$$

where S_t is part of an underlying Markov chain.

Hidden Markov models



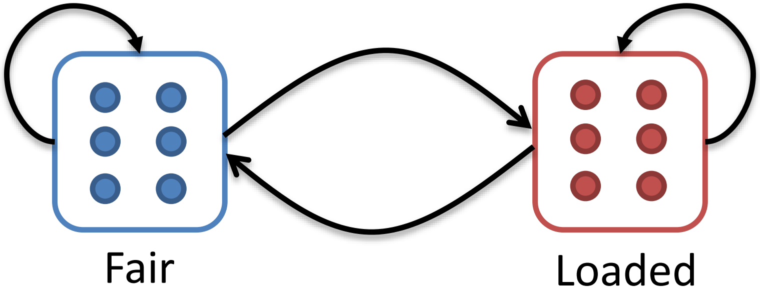


Discrete set with *m* values.

$\Pr(O_t|S_t)$: d × m matrix M

$$M_{i,j} = \Pr(O_t = j|S_t = i)$$

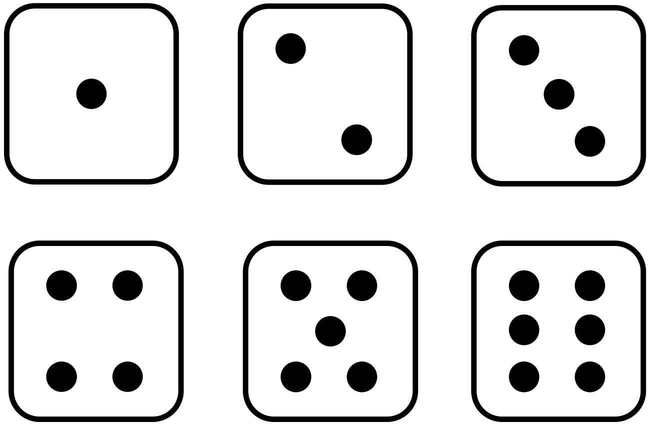
The dishonest casino



T	F_{k-1}	L_{k-1}
F_k	0.95	0.05
L_k	0.05	0.95

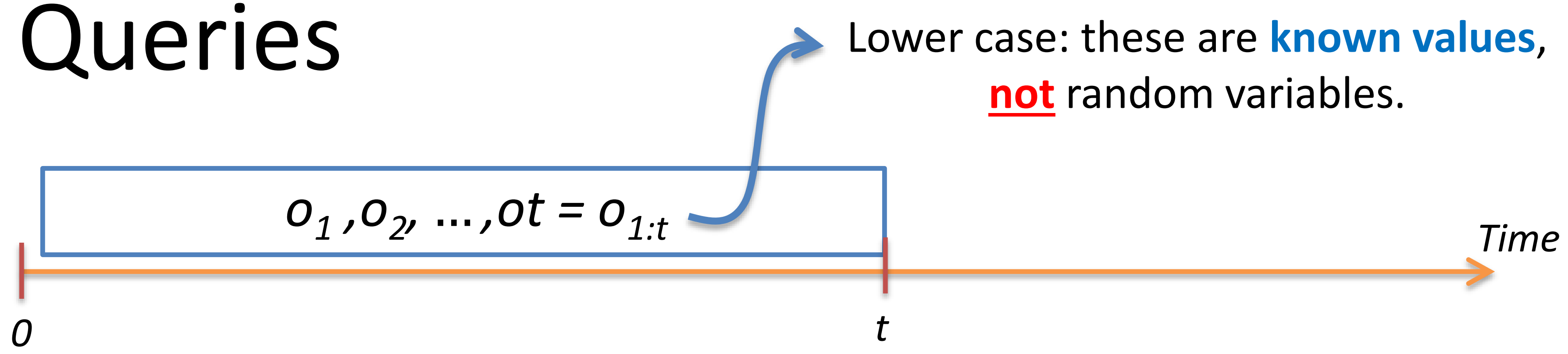
Hidden states

Observations



M	1	2	3	4	5	6
F_k	1/6	1/6	1/6	1/6	1/6	1/6
L_k	1/10	1/10	1/10	1/10	1/10	1/2

Queries



Lower case: these are **known values**, **not** random variables.

“Given the available history of observations, what’s the belief about the current hidden state?”

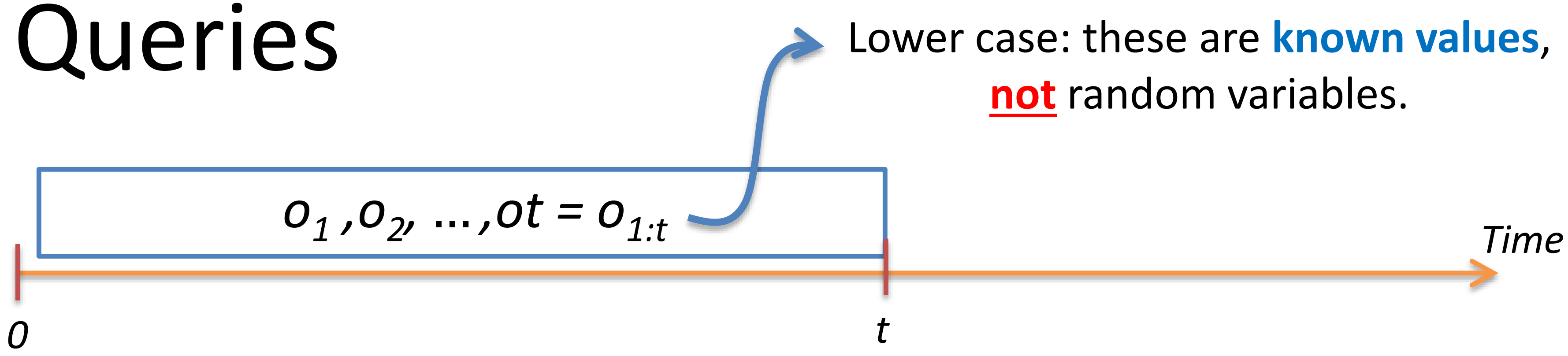
$\Pr(S_t|o_{1:t})$ → Filtering

“Given the available history of observations, what’s the belief about a past hidden state?”

$\Pr(S_k|o_{1:t}), k < t$ → Smoothing

Queries

Lower case: these are **known values**, **not** random variables.



“Given the available history of observations, what’s the belief about a future hidden state?”

$$\Pr(S_k | o_{1:t}), k > t \longrightarrow \text{Prediction}$$

“Given the available history of observations, what’s the most likely **sequence** of hidden states?”

$$\mathrm{s}^*_{0:\mathrm{t}} = \arg\max_{s_{0:t}} \Pr(S_{0:t} = s_{0:t} | o_{1:t}) \longrightarrow \text{Decoding}$$

5. HMM algorithms

Where we'll learn how to compute answers to the previously seen HMM queries.

Notation

Random variable!

$$\underbrace{\Pr(\overset{\nearrow}{S_t}|\cdot)}_{\text{Vector of } d \text{ probability values.}} \longrightarrow \text{Probability distribution of } S_t$$

$$\underbrace{\Pr(S_t = s|\cdot) = \Pr(\mathrm{s}_t|\cdot)}_{\text{Probability} \in [0,1]} \longrightarrow \text{Probability of observing } S_t = s \text{ according to } \Pr(S_t|\cdot)$$

Filtering (*forward*)

“Given the available history of observations, what’s the belief about the current hidden state?”

$$\Pr(S_t|o_{1:t}) = \hat{p}_t$$

$$
\begin{aligned}
\Pr(S_t|o_{1:t}) &= \Pr(S_t|o_t, o_{1:t-1}) \\
&\propto \Pr(o_t|S_t, o_{1:t-1})\Pr(S_t|o_{1:t-1}) && \textit{Bayes} \\
&= \Pr(o_t|\overset{\checkmark}{S_t})\Pr(S_t|o_{1:t-1}) && \textit{Obs. model}
\end{aligned}
$$

$$
\begin{aligned}
\Pr(S_t|o_{1:t-1}) &= \sum_{i=1}^{d} \Pr(S_t|S_{t-1} = i, o_{1:t-1}) \Pr(S_{t-1} = i|o_{1:t-1}) && \textit{Marg.} \\
&= \sum_{i=1}^{d} \Pr(S_t|\overset{\checkmark}{S_{t-1}} = i) \underbrace{\Pr(S_{t-1} = i|o_{1:t-1})}_{\textit{Recursion!}} && \textit{Trans. model}
\end{aligned}
$$

Filtering

"Given the available history of observations, what's the belief about the current hidden state?"

$$\Pr(S_t | o_{1:t}) = \hat{p}_t$$

1. One-step prediction:

$$\Pr(S_t | o_{1:t-1}) = \overline{p_t} = \sum_{i=1}^{d} \Pr(S_t | S_{t-1} = i) \Pr(S_{t-1} = i | o_{1:t-1}) = T\hat{p}_{t-1}$$

2. Measurement update:

$$\widehat{p_t}[i] = \eta \Pr(o_t | S_t = i) \overline{p_t}[i]$$

3. Normalize belief (to get rid of η):

$$\widehat{p_t}[i] \leftarrow \frac{\widehat{p_t}[i]}{\eta}, \eta = \sum_{j=1}^{d} \widehat{p_t}[j]$$

Prediction

“Given the available history of observations, what’s the belief about a future hidden state?”

$$\Pr(S_k | o_{1:t}), k > t$$

$$\Pr(S_{t+1} | o_{1:t}) = T\hat{p}_t$$

Previous slide.

$$\Pr(S_{t+2} | o_{1:t}) = \sum_{i=1}^{d} \Pr(S_{t+2} | S_{t+1} = i) \Pr(S_{t+1} = i | o_{1:t}) = T^2 \widehat{p}_t$$

$$\vdots$$

$$\Pr(S_k | o_{1:t}) = T^{k-t} \hat{p}_t$$

Smoothing (*forward-backward)*

"Given the available history of observations, what's the belief about a past hidden state?" $\Pr(S_k|o_{1:t}), k < t$

$$\Pr(S_k|o_{1:t}) = \Pr(S_k|o_{1:k}, o_{k+1:t})$$

$$\propto \Pr(\mathrm{o}_{\mathrm{k+1:t}}|\mathrm{S}_{\mathrm{k}}, \mathrm{o}_{1:\mathrm{k}})\Pr(S_k|o_{1:k})$$ *Bayes*

$$= \Pr(\mathrm{o}_{\mathrm{k+1:t}}|\mathrm{S}_{\mathrm{k}})\underbrace{\Pr(S_k|o_{1:k})}_{\textit{Filtering!}}\checkmark$$ *Obs. model*

$$\Pr(\mathrm{o}_{\mathrm{k+1:t}}|\mathrm{S}_{\mathrm{k}}) = \sum_{i=1}^{d} \Pr(\mathrm{o}_{\mathrm{k+1:t}}|\mathrm{S}_{\mathrm{k+1}} = \mathrm{i}, \mathrm{S}_{\mathrm{k}})\Pr(S_{k+1} = i|S_k)$$ *Marg.*

$$= \sum_{i=1}^{d} \Pr(o_{k+2:t}, \mathrm{o}_{\mathrm{k+1}}|S_{k+1} = i)\Pr(S_{k+1} = i|S_k)$$ *Obs. model*

$$= \sum_{i=1}^{d} \underbrace{\Pr(o_{k+2:t}|S_{k+1} = i)}_{\textit{Recursion!}}\Pr(\mathrm{o}_{\mathrm{k+1}}|\overset{\checkmark}{\mathrm{S}_{\mathrm{k+1}}} = \mathrm{i})\Pr(S_{k+1} = \overset{\checkmark}{i}|S_k)$$

Smoothing

"Given the available history of observations, what's the belief about a past hidden state?"

$$\Pr(S_k|o_{1:t}) = \tilde{p}_{k,t}, k < t$$

1. Perform filtering from *0* to *k* (*forward*):

$$\Pr(S_k|o_{1:k}) = \hat{p}_k$$

2. Compute the backward recursion from *t* to *k*:

$$\Pr(\mathrm{o}_{\mathrm{k+1:t}}|\mathrm{S}_\mathrm{k}) = b_{k,t}, \quad b_{t,t} = \mathbf{1}$$

$$b_{m-1,t}[i] = \sum_{j=1}^{d} b_{m,t}[j] \Pr(\mathrm{o}_\mathrm{m}|\mathrm{S}_\mathrm{m} = \mathrm{j})\Pr(S_m = j|S_{m-1} = i) \ , k+1 \leq m \leq t$$

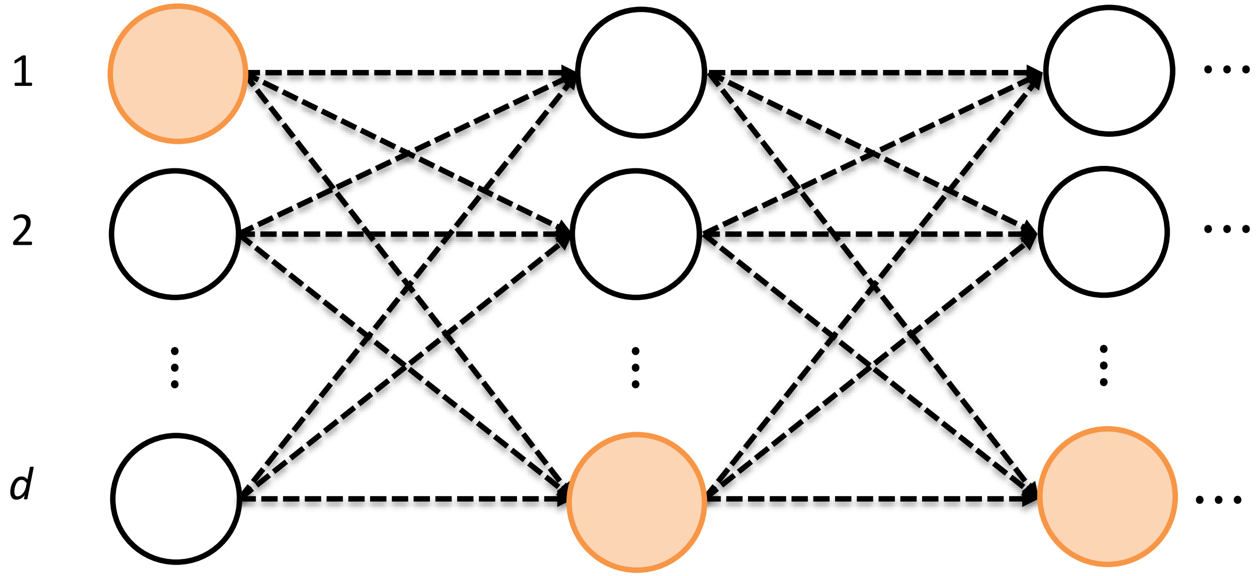
3. Combine the two results and normalize:

$$\tilde{p}_{k,t}[i] = b_{k,t}[i]\hat{p}_k[i], \qquad \tilde{p}_{k,t}[i] \leftarrow \frac{\tilde{p}_{k,t}[i]}{\eta}, \eta = \sum_{j=1}^{d} \tilde{p}_{k,t}[j]$$

Decoding

"Given the available history of observations, what's the most likely **sequence** of hidden states so far?"

$$s^*_{0:t} = \arg\max_{s_{0:t}} \Pr(S_{0:t} = s_{0:t} | o_{1:t})$$



Decoding (simple algorithm)

"Given the available history of observations, what's the most likely **sequence** of hidden states so far?"

$$\mathrm{s}^*_{0:\mathrm{t}} = \arg\max_{s_{0:t}} \Pr(S_{0:t} = s_{0:t} | o_{1:t})$$

$$\Pr(s_{0:t} | o_{1:t}) \propto \Pr(o_{1:t} | s_{0:t}) \Pr(s_{0:t})$$

Bayes

$$= \Pr(s_0) \prod_{i=1}^{t} \Pr(s_i | s_{i-1}) \Pr(o_i | s_i)$$

HMM model

Decoding (simple algorithm)

"Given the available history of observations, what's the most likely **sequence** of hidden states so far?"

$$\mathrm{s}^*_{0:\mathrm{t}} = \arg\max_{s_{0:t}} \Pr(S_{0:t} = s_{0:t} | o_{1:t})$$

Compute all possible state trajectories from *0* to *t*.

$$\mathbb{Tr}_{0:\mathrm{t}} = \{s_{0:t} | s_i \in \mathbb{S}, \mathrm{i}=0, \dots, t\} \longrightarrow$$

How big is $\mathbb{Tr}_{0:\mathrm{t}}$?

$$d^{t+1}$$

$$\downarrow$$

Choose the most likely trajectory according to

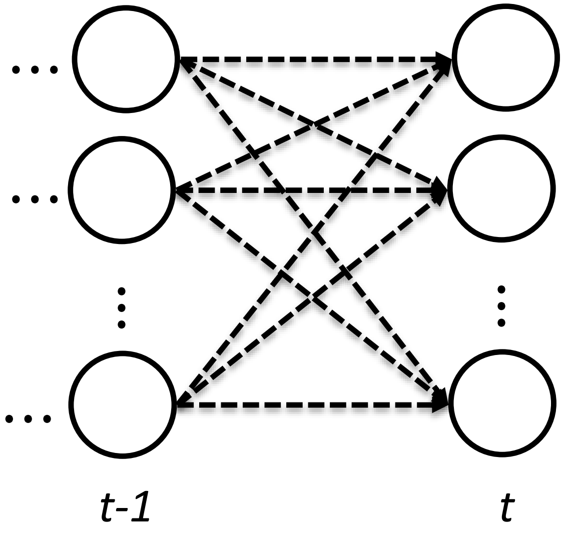
$$\mathrm{s}^*_{0:\mathrm{t}} = \arg\max_{s_{0:t} \in \mathbb{Tr}_{0:\mathrm{t}}} \Pr(s_0) \prod_{i=1}^{t} \Pr(s_i | s_{i-1}) \Pr(o_i | s_i)$$

Can we do better?

Decoding (the *Viterbi* algorithm)

"Given the available history of observations, what's the most likely **sequence** of hidden states so far?"

$$\mathrm{s}^*_{0:\mathrm{t}} = \arg\max_{s_{0:t}} \Pr(S_{0:t} = s_{0:t} | o_{1:t})$$



$$\Pr(S_{0:t}|o_{1:t}) = \Pr(S_{0:t}|o_{1:t-1}, o_t)$$

$$\propto \Pr(\mathrm{o_t}|\mathrm{S}_t)\Pr(S_t, S_{0:t-1}|o_{1:t-1})$$

$$= \Pr(\mathrm{o_t}|\overset{\checkmark}{\mathrm{S}_t})\Pr(\mathrm{S_t}|\overset{\checkmark}{\mathrm{S}_{\mathrm{t}-1}})\underbrace{\Pr(S_{0:t-1}|o_{1:t-1})}_{\textit{Recursion!}}$$

"From all paths arriving at $s_{t\text{-}1}$, record only the most likely one."

$$\max_{s_{0:t}} \Pr(s_{0:t}|o_{1:t}) = \max_{s_t, s_{t-1}} \Pr(o_t|s_t)\Pr(s_t|s_{t-1}) \overbrace{\max_{s_{0:t-1}} \Pr(s_{0:t-1}|o_{1:t-1})}$$

Decoding (the *Viterbi* algorithm)

"Given the available history of observations, what's the most likely **sequence** of hidden states so far?"

$\delta_k[s]$: most likely path ending in $s_t = s$

$\delta_0[s]$: (s), $l_0[s]$: $\Pr(S_0 = s)$

$l_k[s]$: likelihood of $\delta_k[s]$ (unnormalized probability)

1. Expand paths in δ_k according to the transition model

$$\mathrm{pred}_{k+1}[s] = \arg\max_{s'} \Pr(s_{k+1} = s | s_k = s')\, l_k[s'], s = 1, \ldots, d$$

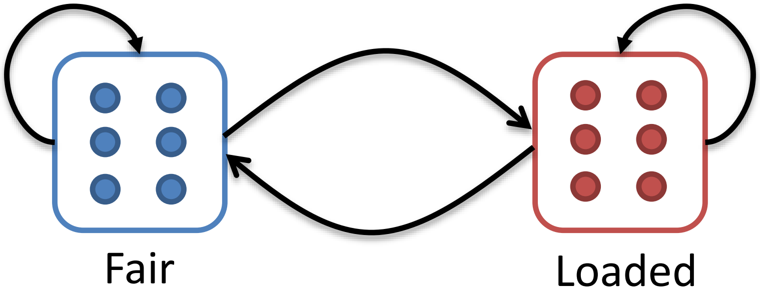
$$\delta_{k+1}[s] = \delta_k\left[\mathrm{pred}_{k+1}[s]\right].\mathrm{append}(s)$$

2. Update likelihood:

$$l_{k+1}[s] = \Pr(o_{k+1} | s_{k+1} = s) \Pr(s_{k+1} = s | s_k = \mathrm{pred}_{k+1}[s])\, l_k\left[\mathrm{pred}_{k+1}[s]\right]$$

3. When *k*=*t*, choose $\delta_t[s]$ with the highest $l_t[s]$.

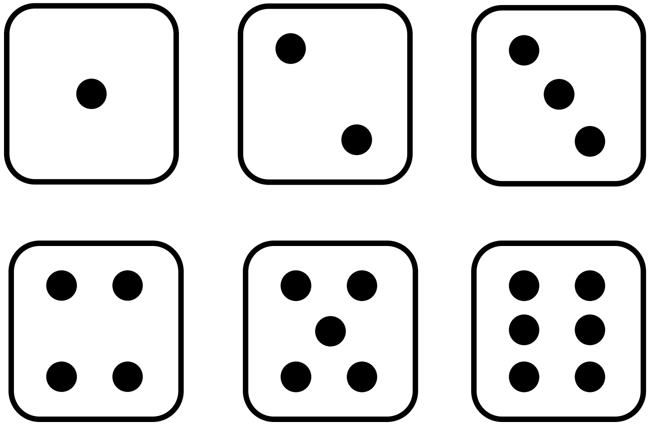
Dishonest casino example



T	F_{k-1}	L_{k-1}
F_k	0.95	0.05
L_k	0.05	0.95

Hidden states

Observations



M	1	2	3	4	5	6
F_k	1/6	1/6	1/6	1/6	1/6	1/6
L_k	1/10	1/10	1/10	1/10	1/10	1/2

Dishonest casino example

$$\Pr(S_0) = \begin{bmatrix}0.8\\0.2\end{bmatrix}\begin{array}{l}\rightarrow \text{Fair}\\ \rightarrow \text{Loaded}\end{array}$$

Observations = 1,2,4,6,6,6,3,6

T

	F_{k-1}	L_{k-1}
F_k	0.95	0.05
L_k	0.05	0.95

M

	1	2	3	4	5	6
F_k	1/6	1/6	1/6	1/6	1/6	1/6
L_k	1/10	1/10	1/10	1/10	1/10	1/2

	Filtering		Smoothing	
	Fair	Loaded	Fair	Loaded
t=0	**0.8000**	0.2000	**0.7382**	0.2618
t=1	**0.8480**	0.1520	**0.6940**	0.3060
t=2	**0.8789**	0.1211	**0.6116**	0.3884
t=3	**0.8981**	0.1019	0.4679	**0.5321**
t=4	**0.6688**	0.3312	0.2229	**0.7771**
t=5	0.3843	**0.6157**	0.1444	**0.8556**
t=6	0.1793	**0.8207**	0.1265	**0.8735**
t=7	0.3088	**0.6912**	0.1449	**0.8551**
t=8	0.1399	**0.8601**	0.1399	**0.8601**

Coincidence?

Dishonest casino example

$$\Pr(S_0) = \begin{bmatrix}0.8\\0.2\end{bmatrix} \begin{array}{l}\rightarrow \text{Fair}\\ \rightarrow \text{Loaded}\end{array}$$

Observations = 1,2,4,6,6,6,3,6

Filtering (MAP): ['Fair ', 'Fair ', 'Fair ', 'Fair ', 'Fair ', 'Loaded', 'Loaded', 'Loaded', 'Loaded']

Smoothing (MAP): ['Fair ', 'Fair ', 'Fair ', 'Loaded ', 'Loaded ', 'Loaded', 'Loaded', 'Loaded', 'Loaded']

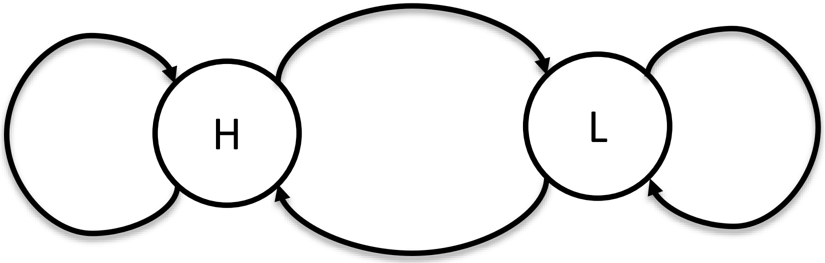
Decoding:

t=0: ['Fair']
t=1: ['Fair', 'Fair']
t=2: ['Fair', 'Fair', 'Fair']
t=3: ['Fair', 'Fair', 'Fair', 'Fair']
t=4: ['Fair', 'Fair', 'Fair', 'Fair', 'Fair']
t=5: ['Fair', 'Fair', 'Fair', 'Fair', 'Fair', 'Fair']
t=6: ['Loaded', 'Loaded', 'Loaded', 'Loaded', 'Loaded', 'Loaded', 'Loaded']
t=7: ['Fair', 'Fair', 'Fair', 'Fair', 'Fair', 'Fair', 'Fair', 'Fair']
t=8: ['Loaded', 'Loaded', 'Loaded', 'Loaded', 'Loaded', 'Loaded', 'Loaded', 'Loaded', 'Loaded']

Borodovsky & Ekisheva (2006), pp 80-81

DNA A G T C A T ••• G

H: High genetic content (coding DNA)
L: Low genetic content (non-coding DNA)



T	H_{k-1}	L_{k-1}
H_k	0.5	0.4
L_k	0.5	0.6

M	A	C	G	T
H_k	0.2	0.3	0.3	0.2
L_k	0.3	0.2	0.2	0.3

Borodovsky & Ekisheva (2006), pp 80-81

$$\Pr(S_0) = \begin{bmatrix}0.5\\0.5\end{bmatrix}\begin{matrix}\rightarrow \text{High}\\ \rightarrow \text{Low}\end{matrix}$$

Observations = G,G,C,A,C,T,G,A,A

T	H_{k-1}	L_{k-1}
H_k	0.5	0.4
L_k	0.5	0.6

M	A	C	G	T
H_k	0.2	0.3	0.3	0.2
L_k	0.3	0.2	0.2	0.3

	Filtering		Smoothing	
	H	L	H	L
t=0	**0.5000**	**0.5000**	**0.5113**	0.4887
t=1	**0.5510**	0.4490	**0.5620**	0.4380
t=2	**0.5561**	0.4439	**0.5653**	0.4347
t=3	**0.5566**	0.4434	**0.5478**	0.4522
t=4	0.3582	**0.6418**	0.3668	**0.6332**
t=5	**0.5368**	0.4632	**0.5278**	0.4722
t=6	0.3563	**0.6437**	0.3648	**0.6352**
t=7	**0.5366**	0.4634	**0.5259**	0.4741
t=8	0.3563	**0.6437**	0.3474	**0.6526**
t=9	0.3398	**0.6602**	0.3398	**0.6602**

Borodovsky & Ekisheva (2006), pp 80-81

$$\Pr(S_0) = \begin{bmatrix}0.5\\0.5\end{bmatrix}$$

Observations = G,G,C,A,C,T,G,A,A

Filtering (MAP): ['H/L', 'H', 'H', 'H', 'L', 'H', 'L', 'H', 'L', 'L']

Smoothing (MAP): ['H', 'H', 'H', 'H', 'L', 'H', 'L', 'H', 'L', 'L']

Decoding:

t=0: ['H']
t=1: ['H', 'H']
t=2: ['H', 'H', 'H']
t=3: ['H', 'H', 'H', 'H']
t=4: ['H', 'H', 'H', 'H', 'L']
t=5: ['H', 'H', 'H', 'H', 'L', 'L']
t=6: ['H', 'H', 'H', 'H', 'L', 'L', 'L']
t=7: ['H', 'H', 'H', 'H', 'L', 'L', 'L', 'H']
t=8: ['H', 'H', 'H', 'H', 'L', 'L', 'L', 'L', 'L']
t=9: ['H', 'H', 'H', 'H', 'L', 'L', 'L', 'L', 'L', 'L']