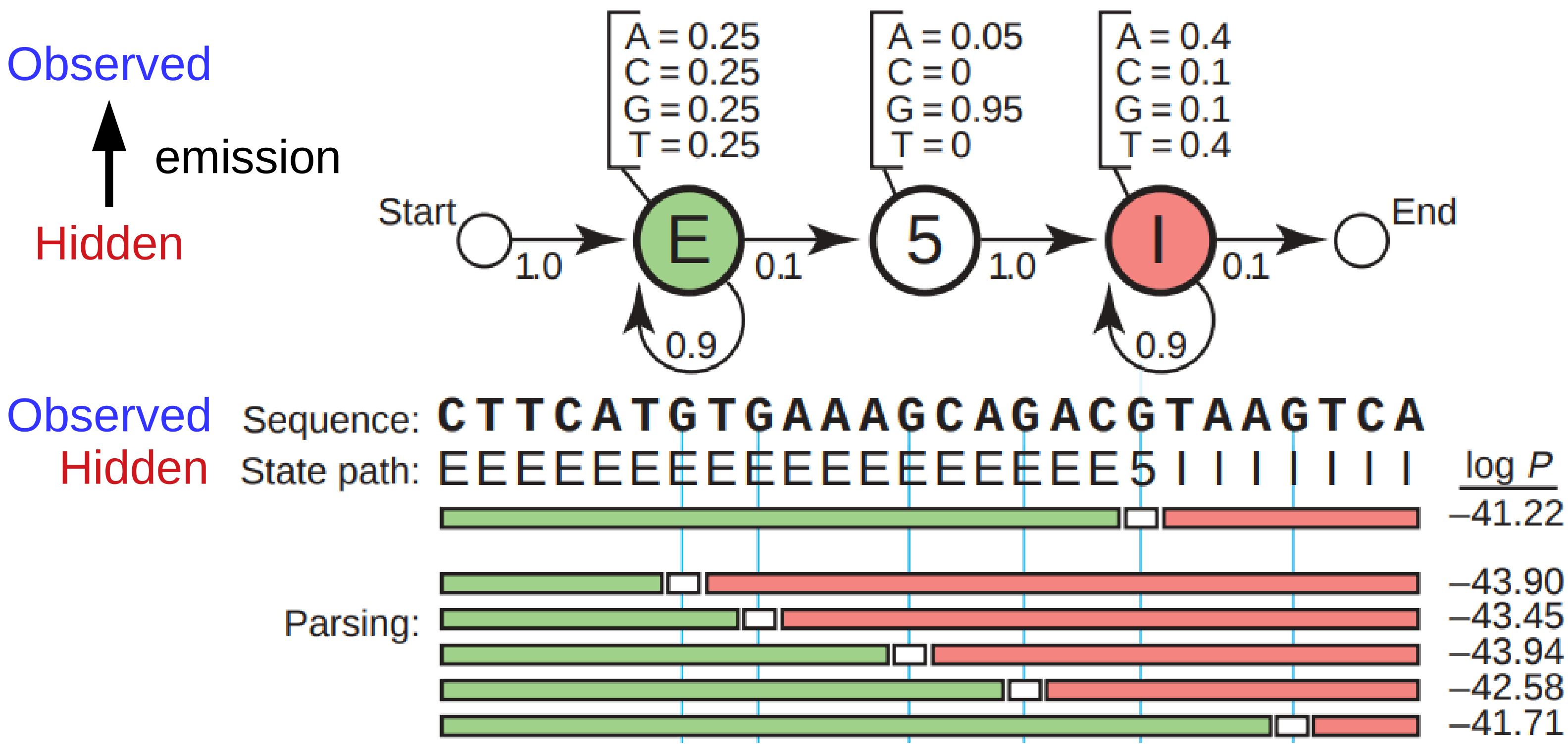
Today's objectives

- Hidden markov models overview
- Markov review (CpG sites)
- Hidden markov models
- Viterbi algorithm

Hidden Markov Models

- Hidden Markov Model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (i.e. hidden) states.
- Hidden Markov models (HMMs) are a formal foundation for making probabilistic models of linear sequence 'labeling' problems
- HMMs are the Legos of computational sequence analysis.

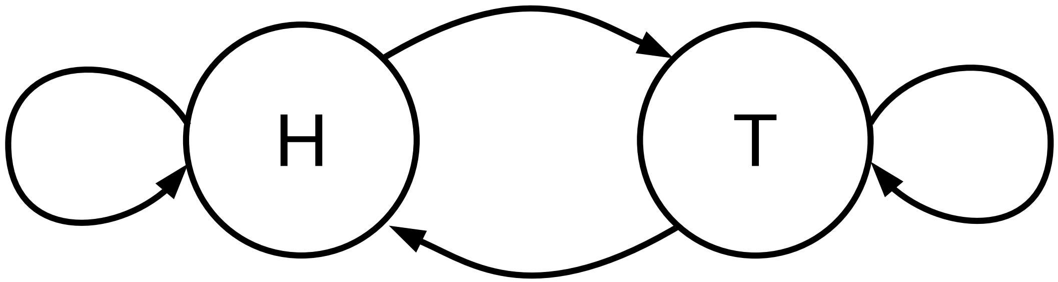
Example: intron/exon labels



Applications
What are the labels?

1) Gene finding
2) Finding conserved sequences
3) Annotation of regulatory sequences
4) Annotation of chromatin states
5) Copy number variation
6) Search for weak similarity (profile alignment HMMs)

Coin toss models



Fair

	H	T
H	0.5	0.5
T	0.5	0.5

A zeroth-order Markov chain
$P(X_n \mid X_{n-1}) = P(X_n)$
Essentially a Binomial Random Variable

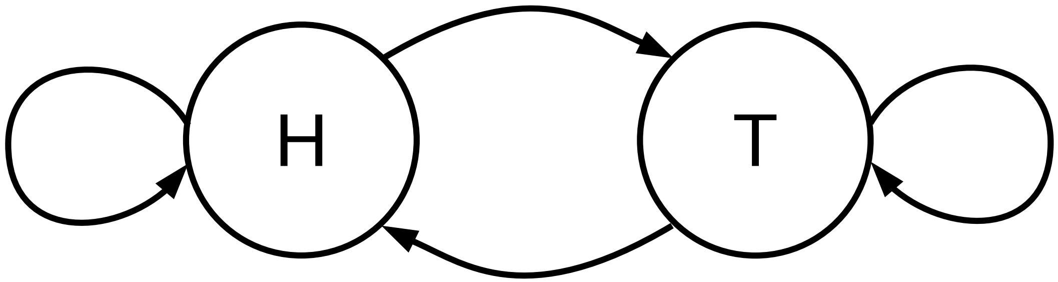
THTHHHTHTHTHTHHTHTHTHTHHTHTHT

Loaded

	H	T
H	0.9	0.1
T	0.9	0.1

THHHHHTHHHHHHTHHHHHHHHHTHHH

Coin toss models



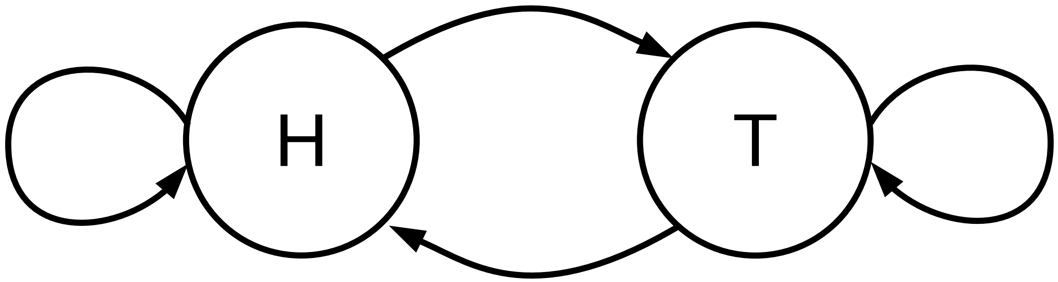
	H	T
H	0.9	0.1
T	0.1	0.9

A weird coin: 1st order Markov chain

TTTTTTTTHHHHHHHHHHTTTTTTTT

A 1st order Markov chain
$P(X_n \mid X_{n-1}) \mathrel{!}= P(X_n)$
An mth order Markov chain
$P(X_n \mid X_{n-1} \ldots X_{n-m}) \mathrel{!}= P(X_n)$
The n state depends on the past m states.

Hidden Coin toss models



Fair

	H	T
H	0.5	0.5
T	0.5	0.5

Loaded

	H	T
H	0.9	0.1
T	0.9	0.1

What if there are two coins (fair and loaded), but you don't know which one is used and they switch?

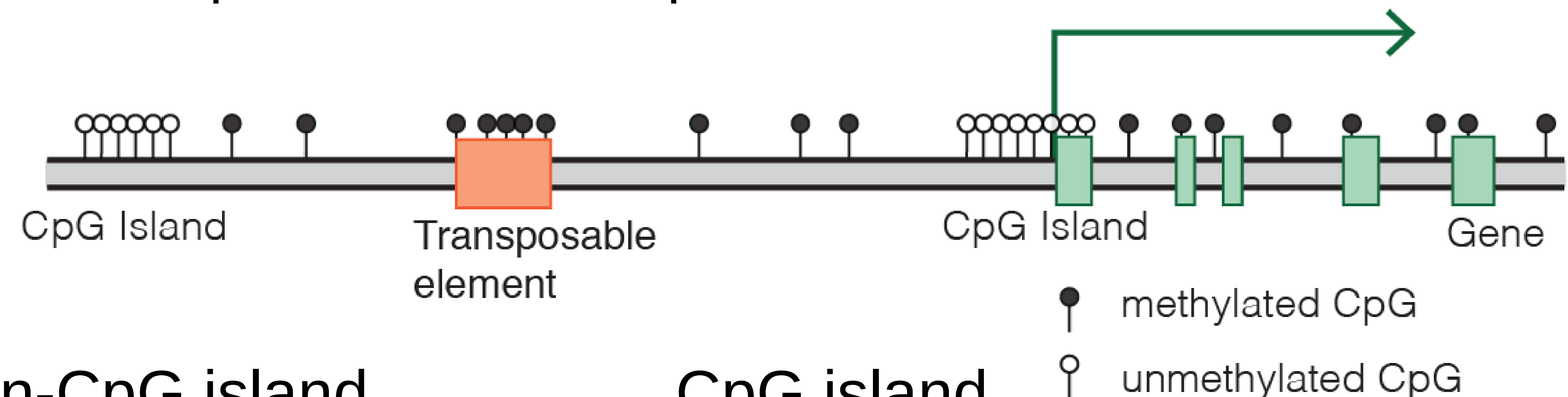
FFFFFFFFLLLLLLLLLLFFFFFFF
THTHHHTHHHHHTHHHHHTHTHTTH

A first order Markov chain would only model the average $P(X_n|X_{n-1})$

A Hidden Markov Model can account for fair/loaded coin this is not observed

A Markov Chain for genomic sequences

CpG island = 200 bp with GC content of >50%



Non-CpG island

	A	G	C	T
A				
G				
C		low		
T				

CpG island

	A	G	C	T
A				
G				
C				
T				

Rows = from
Columns = to

Red = high prob
Orange = medium
Yellow = low

CpG = high G/C
nonCpG = low CG

Simulations, probabilities & building a Markov Model

	A	G	C	T
A	0.19	0.27	0.40	0.14
G	0.17	0.33	0.36	0.14
C	0.19	0.36	0.25	0.20
T	0.10	0.34	0.38	0.19

Probabilities

x = ATCG

$P(x) = P(x_4|x_3)P(x_3|x_2)P(x_2|x_1)P(x_1)$

P(x) = P(G|C)P(C|T)P(T|A)P(A)

P(x) = 0.36 * 0.38 * 0.14 * 0.16

P(A) approx = mean P(A|X)=0.16

```
a = random(1)
if a < 0.19
    pick A
elseif a < 0.46
    pick G
elseif a < 0.86
    pick C
else
    pick T
```

Simulating

P(C | A) = 0.40

Building

P(C | A) = # times AC occurs / # times AX occurs

Which model is more likely?

CpG island

	A	G	C	T
A	0.19	0.27	0.40	0.14
G	0.17	0.33	0.36	0.14
C	0.19	0.36	0.25	0.20
T	0.10	0.34	0.38	0.19

x = ATCG

P(x) = 0.36*0.38*0.14*0.16

P(x) = 0.00306

4.7 times more likely

Non-CpG island

	A	G	C	T
A	0.34	0.23	0.18	0.25
G	0.30	0.25	0.20	0.25
C	0.38	0.04	0.26	0.33
T	0.22	0.26	0.21	0.31

x = ATCG

P(x) = 0.04*0.21*0.25*0.31

P(x) = 0.000651

Log10(CpG/non-CpG) = Log10(CpG) – Log10(non-CpG)
= 0.673 (log10 likelihood ratio)

P(Cpg/non-Cpg) = 10^0.673 = 4.7 (likelihood ratio)

P(non-Cpg/Cpg) = 10^-0.673 = 0.21 (likelihood ratio)

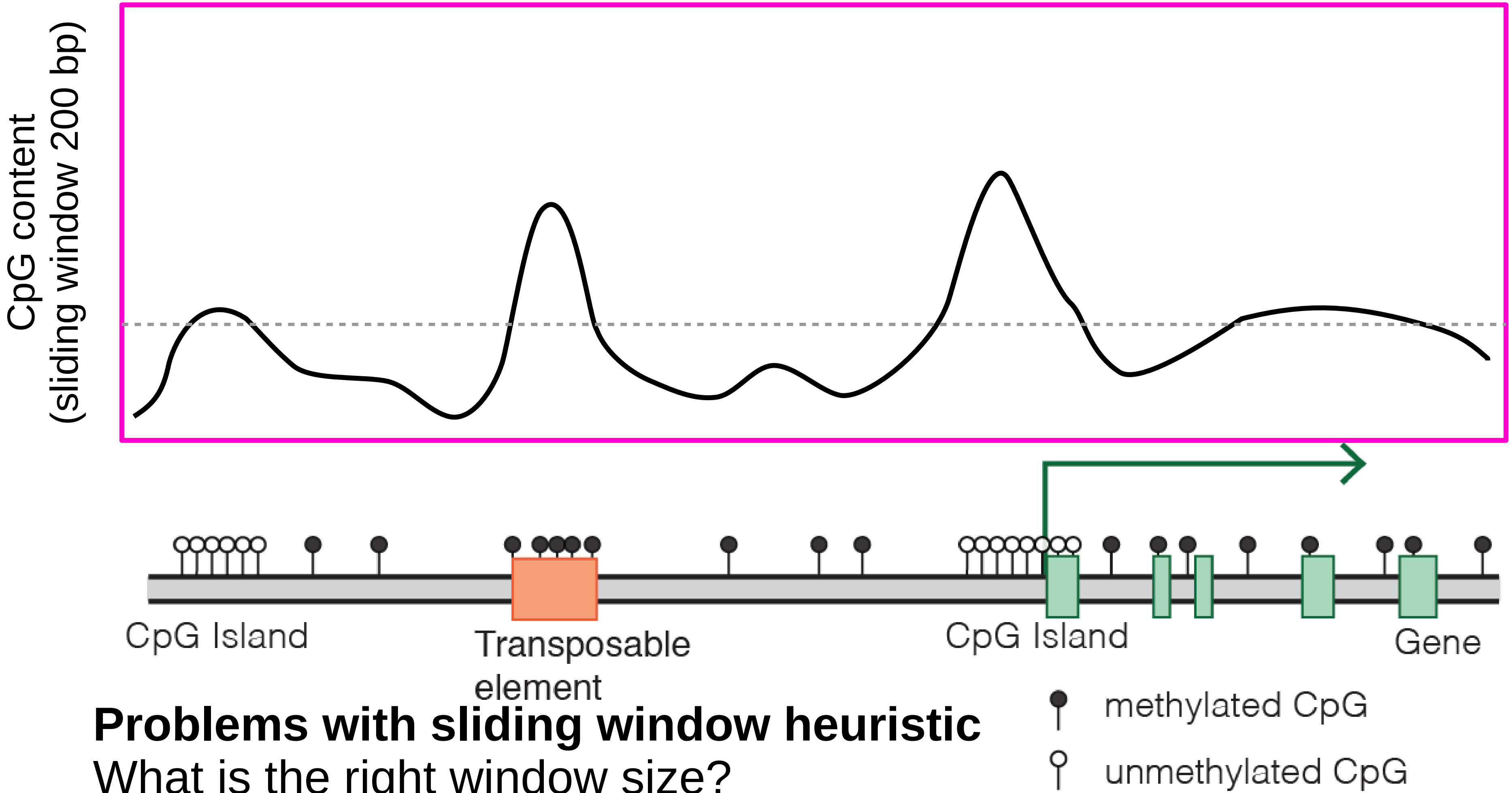
Overflow

Overflow occurs when an arithmetic operation attempts to create a numeric value that is outside of the range that can be represented with a given number of bits – either larger than the maximum or lower than the minimum representable value.

x = ATCGCGATCGATCGCAGTACGTCGATCG
P(x) = P(G|C)*P(C|T) .... P(A)
= 0.00000000000000000000000000000001
10 + P(x) = 10 (if not enough bits)

Solution is to work in log space (e.g. ln, log10, log2)
Log(P(x)) = log(P(G|C)) + log(P(G|C)) ... + log(P(A))
= -31.192812731927

Finding CpG islands



Problems with sliding window heuristic

What is the right window size?

- Too small, we break real islands up
- Too large, we miss islands

What is the right cutoff?

- Changing the cutoff changes the results

Markov Model

CpG island

	A	G	C	T
A	0.19	0.27	0.40	0.14
G	0.17	0.33	0.36	0.14
C	0.19	0.36	0.25	0.20
T	0.10	0.34	0.38	0.19

Non-CpG island

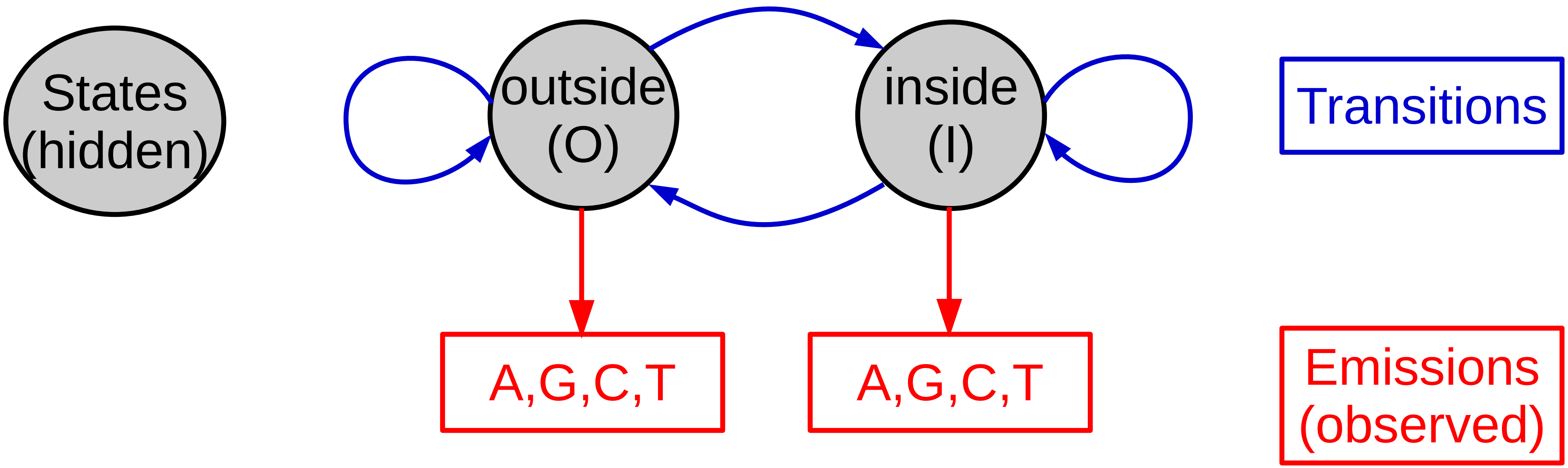
	A	G	C	T
A	0.34	0.23	0.18	0.25
G	0.30	0.25	0.20	0.25
C	0.38	0.04	0.26	0.33
T	0.22	0.26	0.21	0.31

Cutoff problem is solved: calculate which model is more likely
But.. Still have the window size/boundary problem

A MM with two states: inside a CpG island or outside a CpG island
Fixed window size = 3 bp (or 200 bp), slide the window and evaluate

```
ATACGATCAGTACTGTACGATATCGCGTACTCGGCGCTAGCGCTAG
P(ATA|CpG) vs P(ATA|non-CpG)
P(TAC|CpG) vs P(TAC|non-CpG)
etc
```

Hidden Markov Model

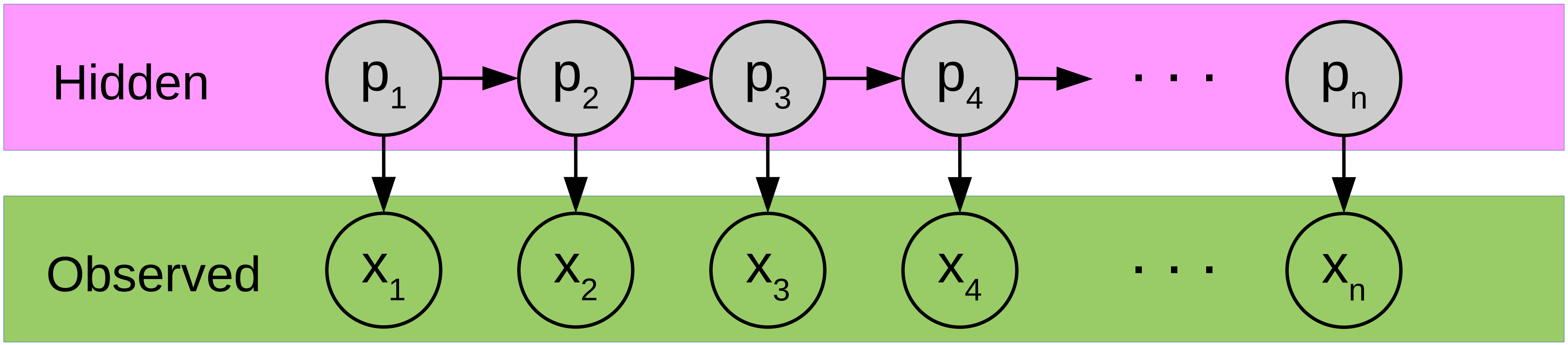


An HMM with two states: inside a CpG island or outside a CpG island

ATACGATCAGTACTGTACGATATCGCGTACTCGGCGCTAGCGCTAG
OOOOOOOOOOOOOOOOOOOOOOOIIIIIIIIIIIIIIIIIIIOOOO
or
OOOOOOOOOOOOOOOOOIIIIIIIIIIIIIIIIIIIIIIIIIOOOO

HMM solution: all possible boundaries and are considered, so a CpG island can be any size, there is no 'window'.
Size of island depends on Transition probabilities

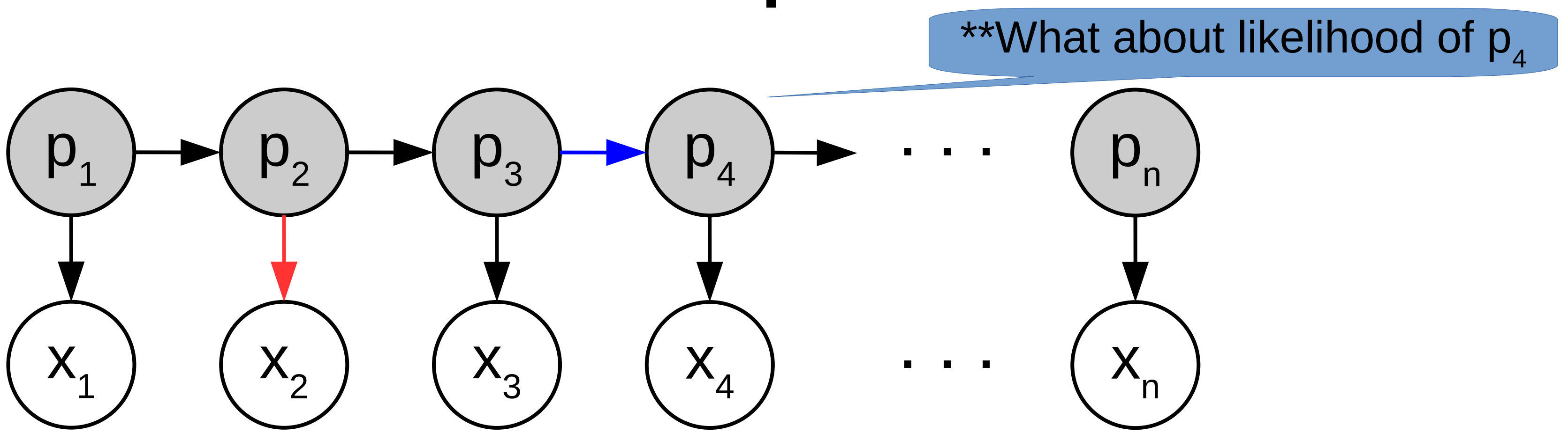
Trellis diagram



$p = \{ p_1, p_2, ..., p_n \}$ is a sequence of states (AKA a path). Each p_i takes a value from set Q. We do not observe p.

$x = \{ x_1, x_2, ..., x_n \}$ is a sequence of emissions. Each x_i takes a value from set $\sum$. We do observe x.

Conditional independence



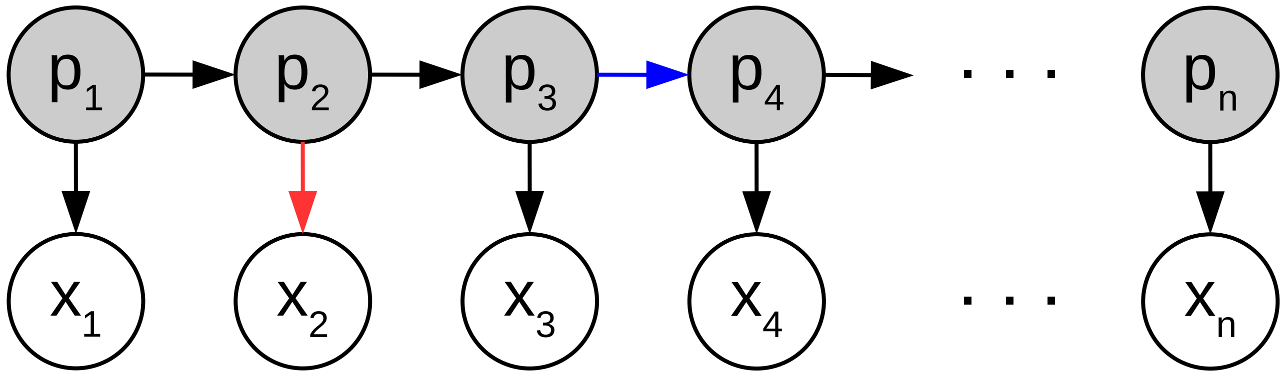
Like Markov chains, edges capture conditional independence:

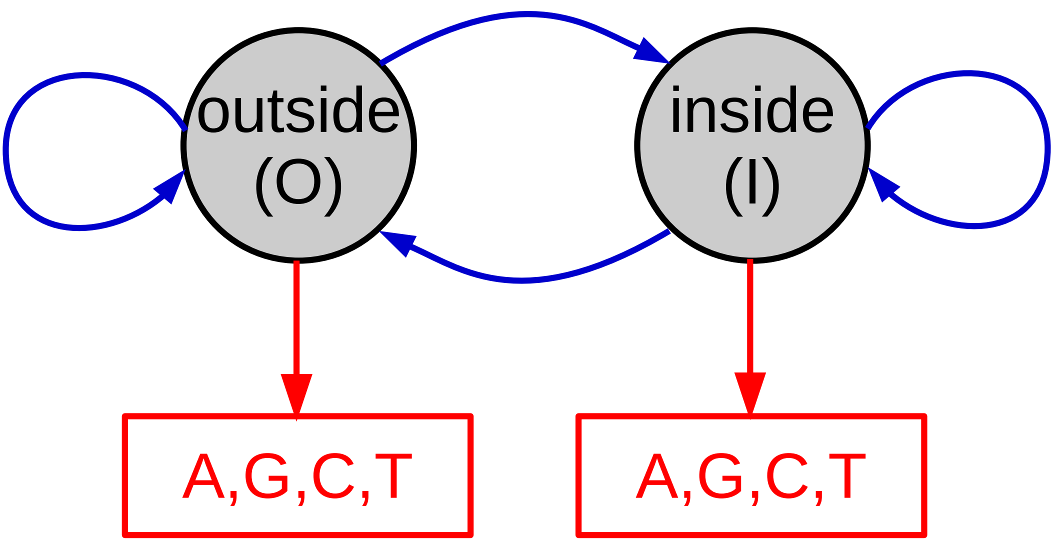
- x_2 is conditionally independent of everything else given p_2
- p_4 is conditionally independent of everything else given p_3

Probability of being in a particular state at step i is known once we know what state we were in at step i-1. Probability of seeing a particular emission at step i is known once we know what state we were in at step i.

** However, the likelihood of a state (p_4) depends on data ($x_{1\text{-}n}$)

HMMs have two matrices: transition and emission





Transitions

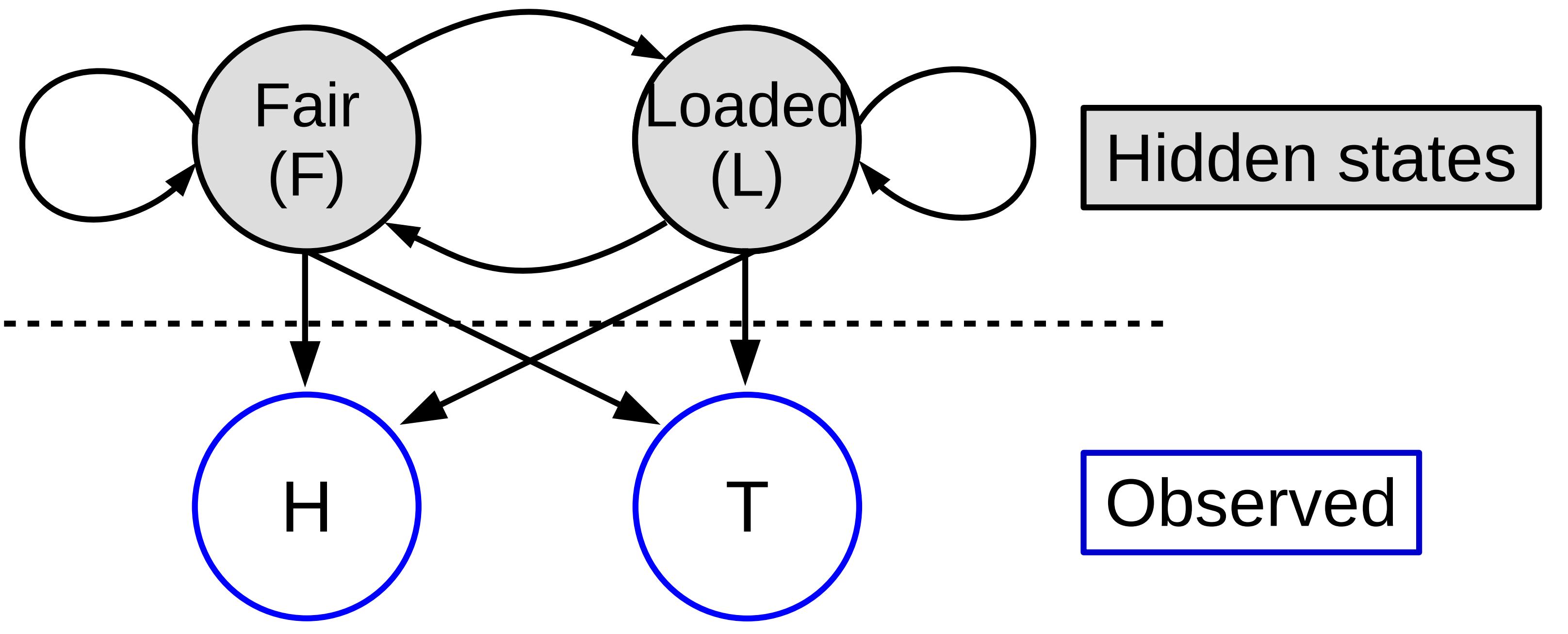
	O	I
O		
I		

Emissions (observed)

	A	G	C	T
O				
I				

Occasionally dishonest casino

Dealer repeatedly flips a coin.
Sometimes the coin is fair, with
    P(heads) = 0.5,
Sometimes it's loaded, with
    P(heads) = 0.8.
Dealer occasionally switches coins, invisibly to you.



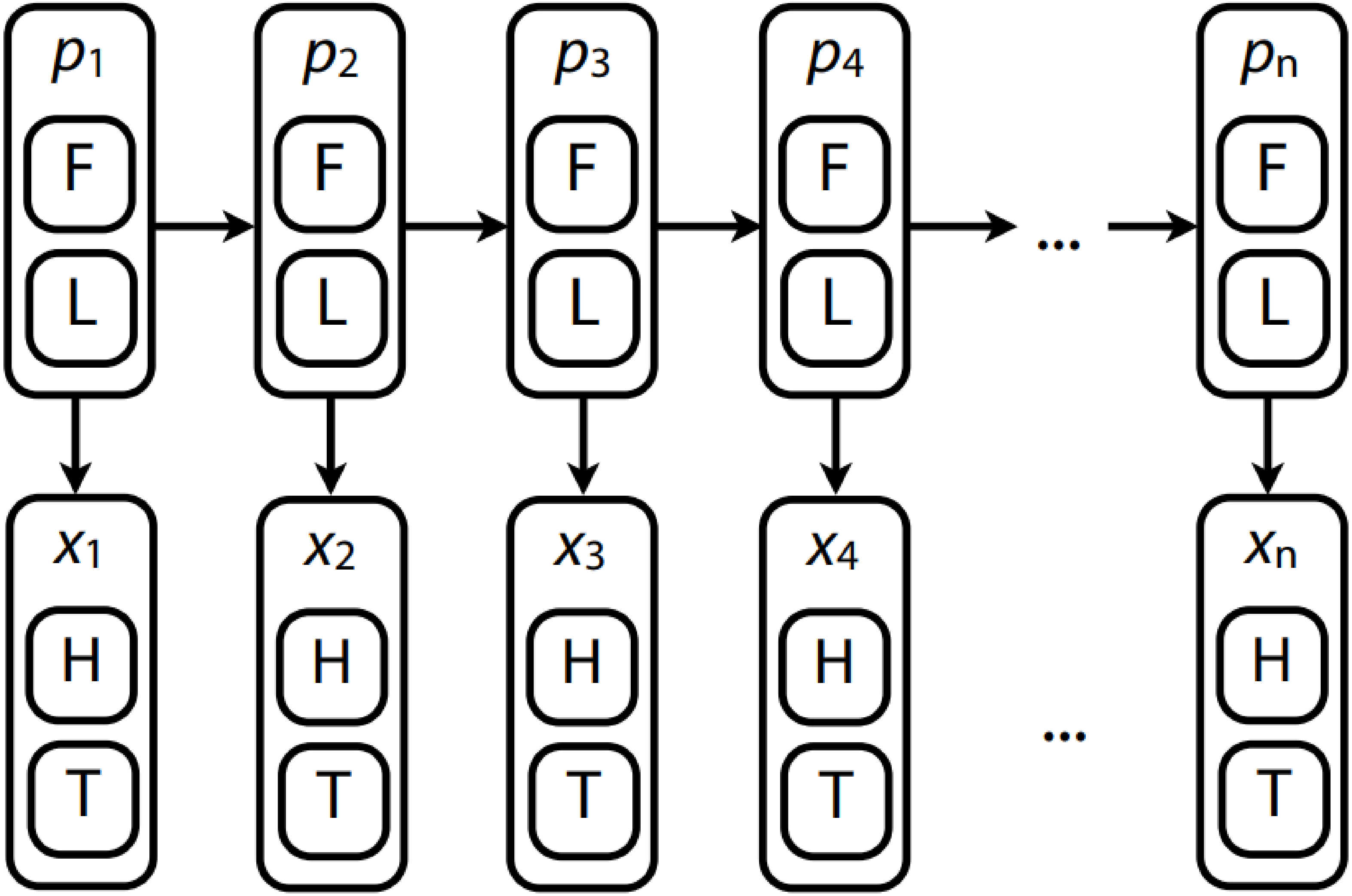
Casino Trellis

States encode which coin is used

F = fair
L = loaded

Emissions encode flip outcomes

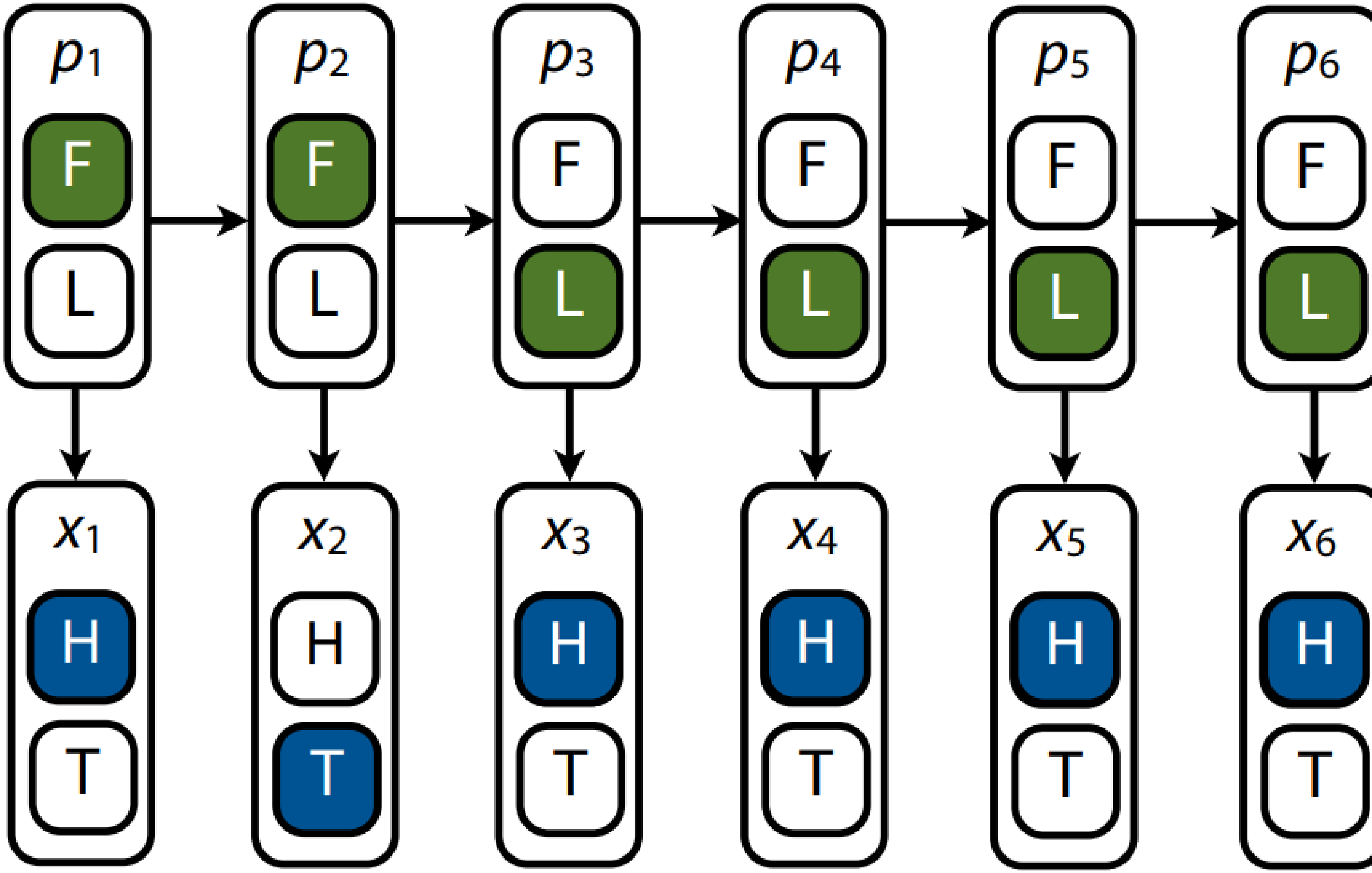
H = heads
T = tails



Casino example with 6 flips

$p = \text{FFLLLL}$

$x = \text{HTHHHH}$



Forward and Backward Algorithm

What is the joint probability of p and x?

$P(p_1,\ldots, p_n, x_1,\ldots., x_n)$

What is the most likely path? (decoding = viterbi algorithm)

$p^* = \text{argmax } P(p_1,\ldots, p_n \mid x_1,\ldots., x_n)$

What is the probability p is in state t and emitting $x_1 \ldots x_i$

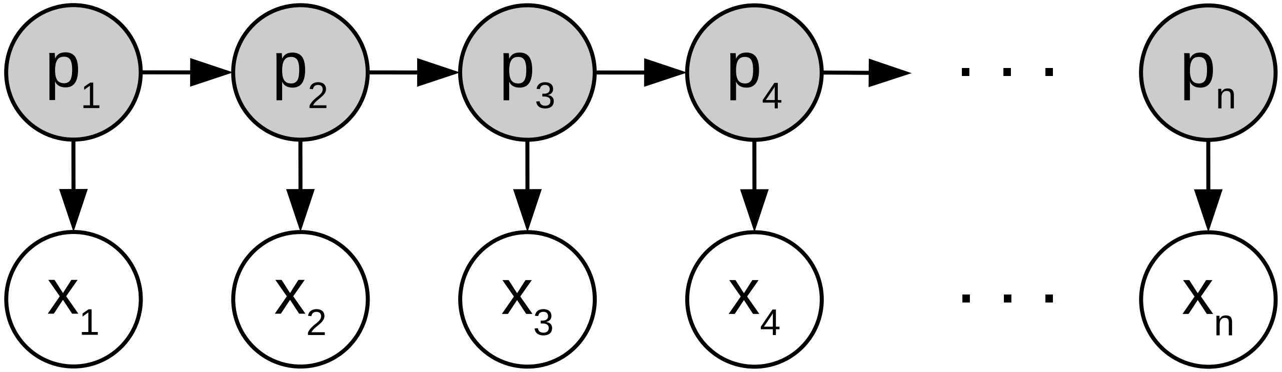
$P(p_i = t, x_1,\ldots., x_i)$ – forward algorithm

What is the probability of emitting $x_{i+1} \ldots x_n$ given $p_i = t$?

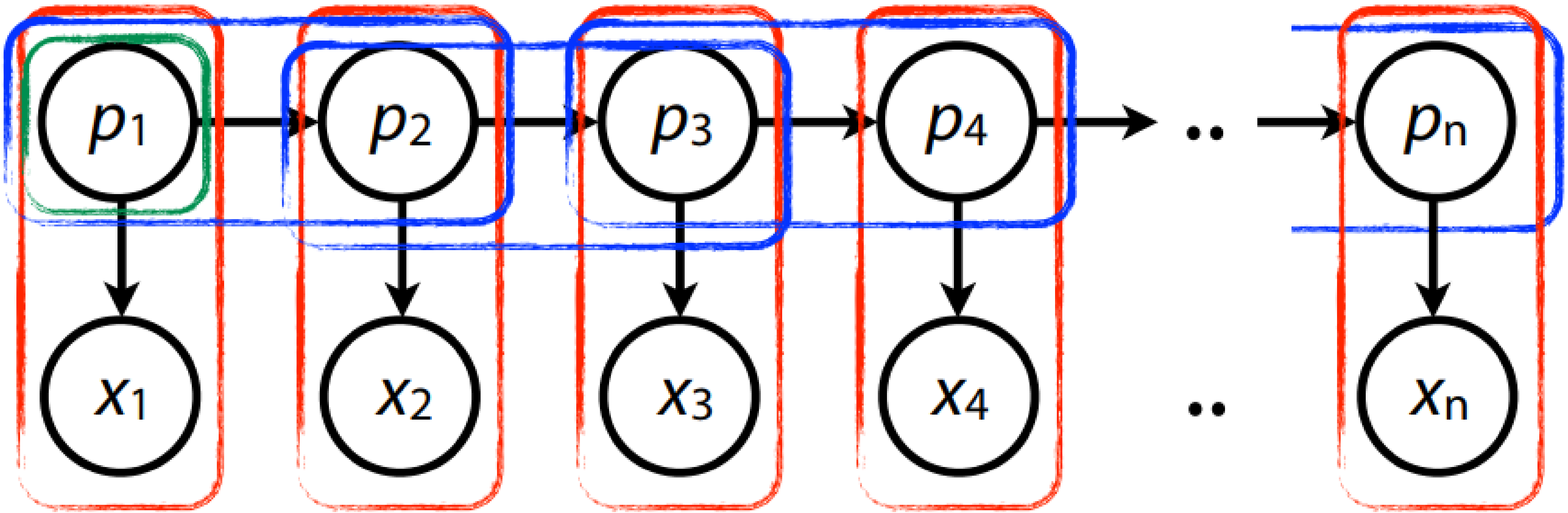
$P(x_{i+1}\ldots x_n \mid p_i = t)$ – backward algorithm

What is the conditional probability of hidden state p at site i

$P(p_i \mid x_1,\ldots x_n)$ -- forward and backward algorithm



Likelihood under an HMM



$$P(p_1, p_2, \ldots, p_n, x_1, x_2, \ldots, x_n) = \prod_{k=1}^{n} P(x_k \mid p_k) \prod_{k=2}^{n} P(p_k \mid p_{k-1})\, P(p_1)$$

$|Q| \times |\Sigma|$ emission matrix E encodes $P(x_i \mid p_i)$s $E[p_i, x_i] = P(x_i \mid p_i)$

$|Q| \times |Q|$ transition matrix A encodes $P(p_i \mid p_{i-1})$s $A[p_{i-1}, p_i] = P(p_i \mid p_{i-1})$

$|Q|$ array I encodes initial probabilities of each state $I[p_i] = P(p_1)$

What is the joint probability of p and x?
If $P(p_1 = F) = 0.5$,
$P(p,x) = 0.5^9 \times 0.8^3 \times 0.6^8 \times 0.4^2 = 0.0000026874$

A	F	L
F	0.6	0.4
L	0.4	0.6

E	H	T
F	0.5	0.5
L	0.8	0.2

p	F	F	F	L	L	L	F	F	F	F	F
x	T	H	T	H	H	H	T	H	T	T	H
$P(x_i \mid p_i)$	0.5	0.5	0.5	0.8	0.8	0.8	0.5	0.5	0.5	0.5	0.5
$P(p_i \mid p_{i-1})$	-	0.6	0.6	0.4	0.6	0.6	0.4	0.6	0.6	0.6	0.6

Viterbi Algorithm

What is the most likely path (p^*) given the emissions?

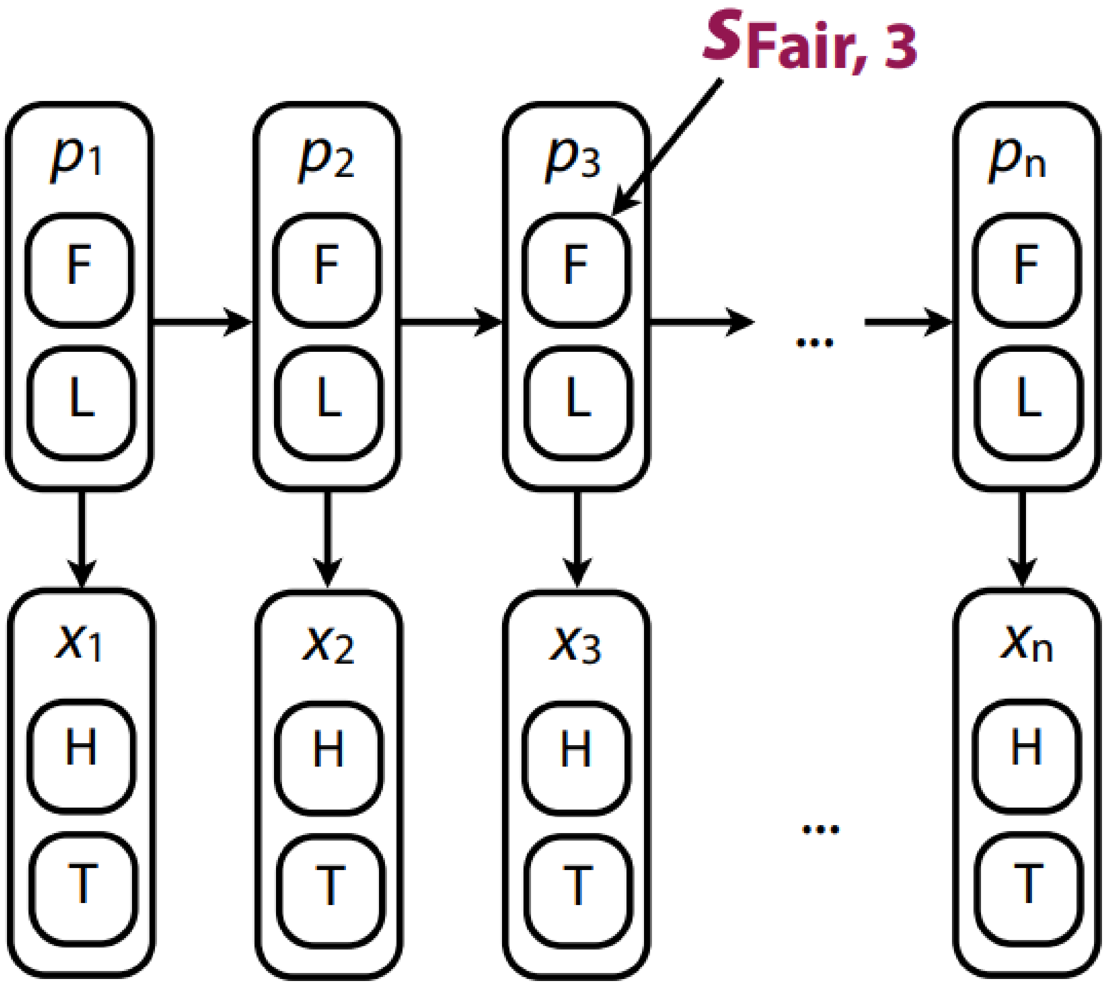
$$p^* = \underset{p}{\operatorname{argmax}} P(p \mid x)$$

Bottom-up dynamic programming

$s_{k,i}$ = score of the most likely path up to step i with $p_i = k$

Start at step 1, calculate successively longer $s_{k,i}$'s

Keep track of $s_{k,i}$ for backtrace to find the most likely path



Given transition matrix *A* and emission matrix *E* (right), what is the most probable path *p* for the following *x*?

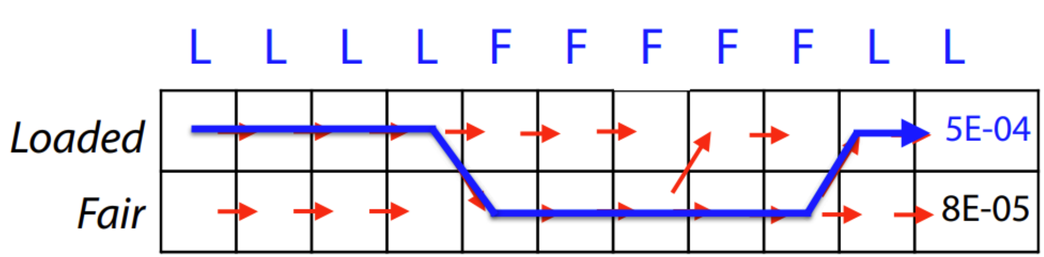
Initial probabilities of F/L are 0.5

A	F	L
F	0.6	0.4
L	0.4	0.6

E	H	T
F	0.5	0.5
L	0.8	0.2

p	?	?	?	?	?	?	?	?	?	?	?
x	T	H	T	H	H	H	T	H	T	T	H
$s_{\text{Fair}, i}$	0.25	?	?	?	?	?	?	?	?	?	?
$s_{\text{Loaded}, i}$	0.1	?	?	?	?	?	?	?	?	?	?

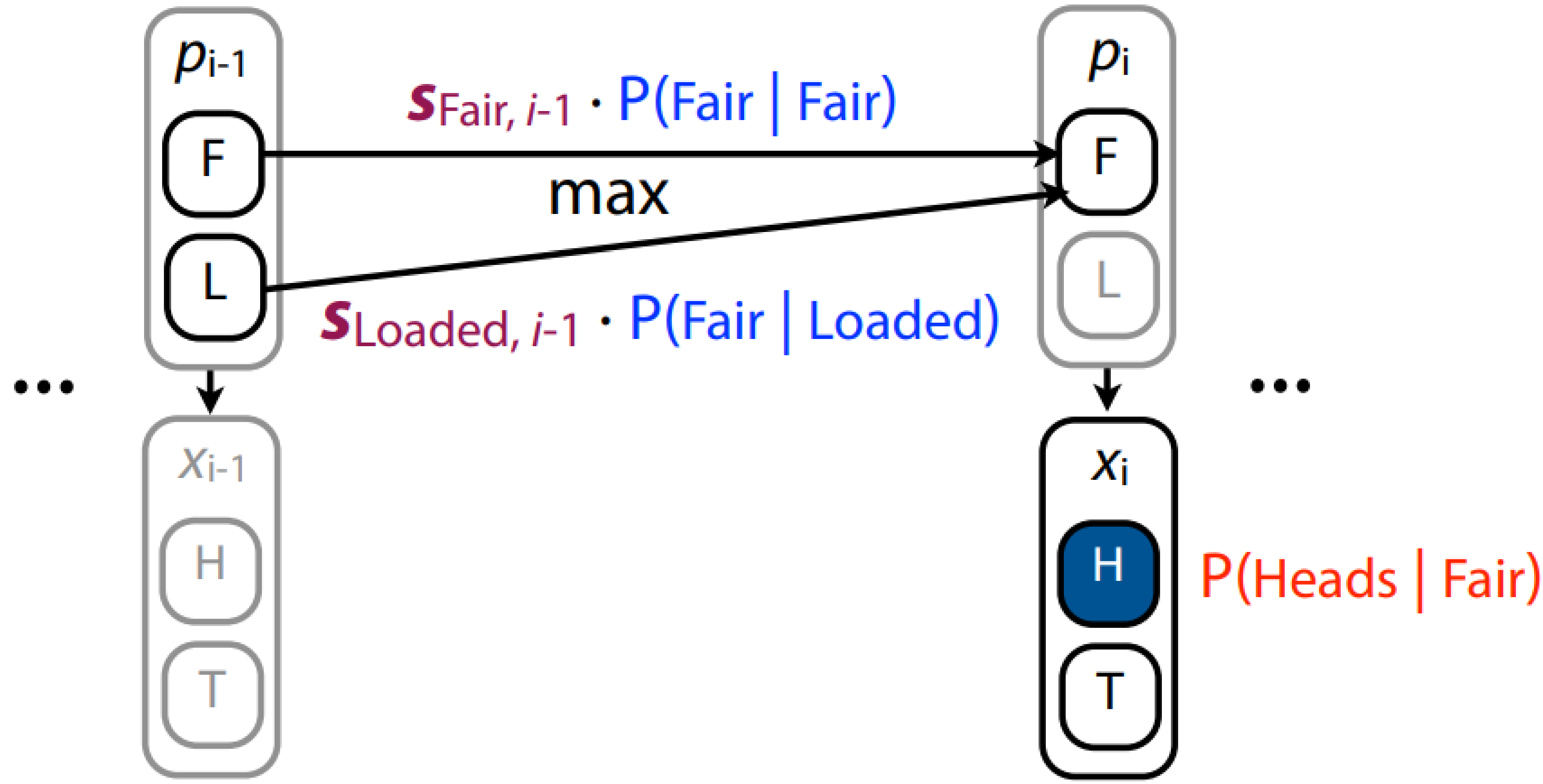
Viterbi fills in all the question marks



$$s_{\text{Fair},i} = P(\text{Heads} \mid \text{Fair}) \cdot \max_{k \in \{\text{Fair, Loaded}\}} \{ s_{k,i-1} \cdot P(\text{Fair} \mid k) \}$$

Emission prob

Transition prob



$S_{F,i} = E \times \max\{S_{k,i-1} \times A\}$

$S_{L,i} = E \times \max\{S_{k,i-1} \times A\}$

assume $p_1(F) = 0.5$

A	F	L
F	0.6	0.4
L	0.4	0.6

E	H	T
F	0.5	0.5
L	0.8	0.2

x	T	H	T	H
$S_{F,i}$	$E(T\|F) = .5$ $A(F) = .5$ $S_{F,1}=.25$			
$S_{L,i}$	$E(T\|L) = .2$ $A(L) = .5$ $S_{L,1}=.1$			

$S_{F,2} = E \times \max\{S_{k,1} \times A\}$

$E(H|F) = 0.5$

$S_{F,1} \times A(F|F) = .25 \times .6 = .15$ (max)

$S_{L,1} \times A(F|L) = .1 \times .4 = .04$

$S_{F,2} = 0.075$

$S_{L,2} = E \times \max\{S_{k,1} \times A\}$

$E(H|L) = 0.8$

$S_{F,1} \times A(L|F) = .25 \times .4 = .1$ (max)

$S_{L,1} \times A(L|L) = .1 \times .6 = .06$

$S_{L,2} = 0.08$

$S_{F,i}$ = E x max{$S_{k,i-1}$ x A}

$S_{L,i}$ = E x max{$S_{k,i-1}$ x A}

assume $p_1(F) = 0.5$

A	F	L
F	0.6	0.4
L	0.4	0.6

E	H	T
F	0.5	0.5
L	0.8	0.2

x	T	H	T	H
$S_{F,i}$	E(T|F) = .5 A(F) = .5 $S_{F,1}$=.25	E(H|F)=.5 $\underline{S_{F,1}\ A(F\|F)}$ $S_{L,1}$ A(F|L) $S_{F,2}$=.075		
$S_{L,i}$	E(T|L) = .2 A(L) = .5 $S_{L,1}$=.1	E(H|L) = .8 $\underline{S_{F,1}\ A(L\|F)}$ $S_{L,1}$ A(L|L) $S_{L,2}$=.08		

$S_{F,3}$ = E x max{$S_{k,2}$ x A}

E(T|F) = 0.5

$S_{F,2}$ x A(F|F) = .075 x .6 =.045 (max)

$S_{L,2}$ x A(F|L) = .08 x .4 = .032

$S_{F,3}$ = 0.0225

$S_{L,3}$ = E x max{$S_{k,2}$ x A}

E(T|L) = 0.2

$S_{F,2}$ x A(L|F) = .075 x .4 = .03

$S_{L,2}$ x A(L|L) = .08 x .6 = .048 (max)

$S_{L,3}$ = 0.0096

$S_{F,i} = E \times \max\{S_{k,i-1} \times A\}$

$S_{L,i} = E \times \max\{S_{k,i-1} \times A\}$

assume $p_1(F) = 0.5$

A	F	L
F	0.6	0.4
L	0.4	0.6

E	H	T
F	0.5	0.5
L	0.8	0.2

x	T	H	T	H
$S_{F,i}$	$E(T\mid F) = .5$ $A(F) = .5$ $S_{F,1}=.25$	$E(H\mid F)=.5$ $\underline{S_{F,1}\ A(F\mid F)}$ $S_{L,1}A(F\mid L)$ $S_{F,2}=.075$	$E(T\mid F)=.5$ $\underline{S_{F,2}\ A(F\mid F)}$ $S_{L,2}A(F\mid L)$ $S_{F,3}=.0225$	
$S_{L,i}$	$E(T\mid L) = .2$ $A(L) = .5$ $S_{L,1}=.1$	$E(H\mid L) = .8$ $\underline{S_{F,1}\ A(L\mid F)}$ $S_{L,1}A(L\mid L)$ $S_{L,2}=.08$	$E(T\mid L) = .2$ $S_{F,2}\ A(L\mid F)$ $\underline{S_{L,2}A(L\mid L)}$ $S_{L,3}= .0096$	

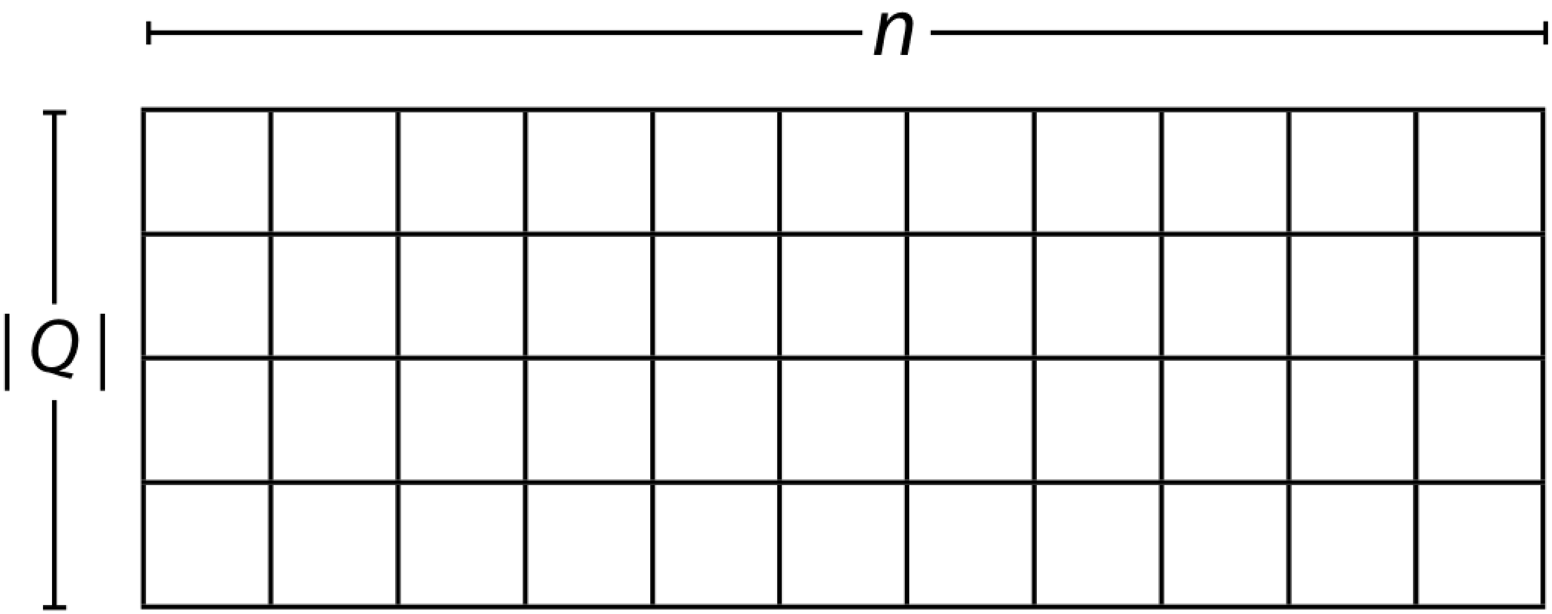
Backtrace

1. Pick state in last step with highest score
2. Backtrace for most likely path according to which state k "won" the max

x	T	H	T	H
$S_{F,i}$	$E(T\|F) = .5$ $A(F) = .5$ $S_{F,1}=.25$	$E(H\|F)=.5$ $\underline{S_{F,1}\,A(F\|F)}$ $S_{L,1}A(F\|L)$ $S_{F,2}=.075$	$E(T\|F)=.5$ $\underline{S_{F,2}\,A(F\|F)}$ $S_{L,2}A(F\|L)$ $\boxed{S_{F,3}=.0225}$	
$S_{L,i}$	$E(T\|L) = .2$ $A(L) = .5$ $S_{L,1}=.1$	$E(H\|L) = .8$ $\underline{S_{F,1}\,A(L\|F)}$ $S_{L,1}A(L\|L)$ $S_{L,2}=.08$	$E(T\|L) = .2$ $S_{F,2}\,A(L\|F)$ $\underline{S_{L,2}A(L\|L)}$ $S_{L,3}= .0096$	

Complexity

How much work did we do, given Q is the set of states and n is the length of the sequence?



$\# \, s_{k,\,i}$ values to calculate $= n \cdot |Q|$, each involves max over $|Q|$ products

$O(n \cdot |Q|^2)$

Matrix A has $|Q|^2$ elements, E has $|Q||\Sigma|$ elements, I has $|Q|$ elements

Exercises

1) Give four examples of application (uses) of HMMs in computational biology.

2) What is the probability of x = AATTCG under the CpG island Markov chain and under the non-CpG island Markov chain (described in the slides)?

CpG island

	A	G	C	T
A	0.19	0.27	0.40	0.14
G	0.17	0.33	0.36	0.14
C	0.19	0.36	0.25	0.20
T	0.10	0.34	0.38	0.19

x = AATTCG

Non-CpG island

	A	G	C	T
A	0.34	0.23	0.18	0.25
G	0.30	0.25	0.20	0.25
C	0.38	0.04	0.26	0.33
T	0.22	0.26	0.21	0.31

x = AATTCG

3) How do you avoid overflow – errors caused by operations on really small numbers?

4) What are two disadvantages of using a sliding window with a cutoff to identify CpG islands?

5) In HMMs, the labels (states) are hidden/observed, and the emissions are hidden/observed?

6) What is the probability of AACG with hidden states OOII under the following HMM:

A	I	O
I	0.8	0.2
O	0.2	0.8

E	A	G	C	T
I	0.1	0.4	0.4	0.1
O	0.25	0.25	0.25	0.25

```
AACG
OOII
```

7) In the diagram below, we observed x1-x4 but not p1-p4:

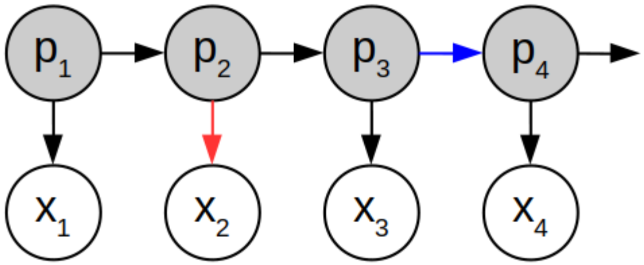
a) does P(p3) depend on p2?

b) does P(p3) depend on x3?

c) does P(p3) depend on x2?

d) does P(p3) depend on x4?

e) does P(p3|p2) depend on x2?



8) Fill in the last column using viterbi and A and E from prior slides.

9) Whats the most likely path?

A	F	L
F	0.6	0.4
L	0.4	0.6

E	H	T
F	0.5	0.5
L	0.8	0.2

x	T	H	T	H
$S_{F,i}$	E(T|F) A(F) $S_{F,1}$=.25	E(H|F)=.5 $S_{F,1}$ A(F|F) $S_{L,1}$ A(F|L) $S_{F,2}$=.075	E(T|F)=.5 $S_{F,2}$ A(F|F) $S_{L,2}$ A(F|L) $S_{F,3}$=.0225	$S_{F,4}$ =
$S_{L,i}$	E(T|L) A(L) $S_{L,1}$=.1	E(H|L) = .8 $S_{F,1}$ A(L|F) $S_{L,1}$ A(L|L) $S_{L,2}$=.08	E(T|L) = .2 $S_{F,2}$ A(L|F) $S_{L,2}$ A(L|L) $S_{L,3}$= .0096	$S_{L,4}$ =