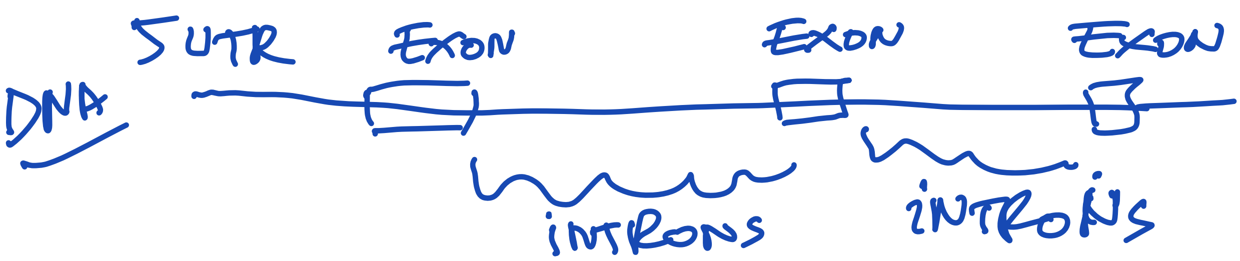
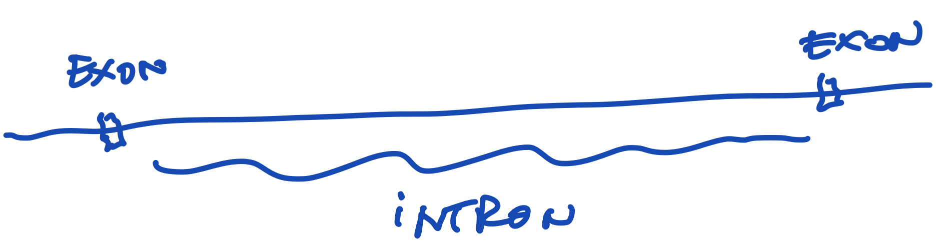
MARKOV CHAINS & HIDDEN MARKOV MODELS

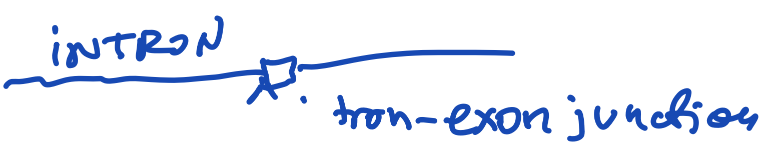
SORIN ISTRAIL

we can find out genes on the genomes

gene







{ in

○ Head
Tail

SOME EXPERIMENTS

can hear you but cannot see her

H, T, T, T, H, T, ...

is the coin fair?

50-50 Head-Tail

coin H 75%
T 25%

Can we infer the bias of the coin?
How do we know?

(1) H, T, H, H, T, T, T, H, T, H, H

(2) H, T, T, T, T, T, H, T, T, T, H

50% = unbiased

	1	2	3	4	5	6	7
(1)	(1,0)	(1,1)	(2,1)	(3,1)	(3,2)	(3,3)	(3,4)

$\uparrow$ H $\uparrow$ T

	1	2	3	4	5	6
(2)	(1,0)	(1,1)	(1,2)	(1,3)	(1,4)	(1,5), ...

The T of the coin 2 is biased

we cannot do this inference when we have a long seq. of tosses.

We can estimate what the bias is: 75-25 or
80-20
55-45

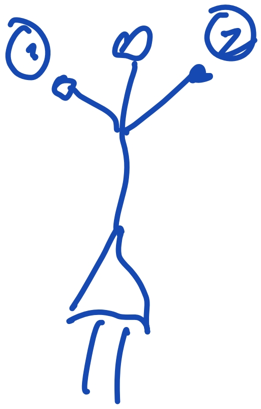
Dishonest Casino Problem

A random sequence:

H, T, H, H, T, T, T, ...

at any point you have to have roughly 50-50
H-T

50,50
53,47



two coins

flip one coin : H or T

use coin #3 — H flips ①, T flips ②

H , T , H , H , T

① ① ② ① ②

① p_H^1 , p_T^1 $\quad p_H^1 + p_T^1 = 1$

② p_H^2 , p_T^2 $\quad p_H^2 + p_T^2 = 1$

③ p_H^3 , p_T^3 $\quad p_H^3 + p_T^3 = 1$

Can we find out : all those 6 numbers $p_H^1, p_T^1, p_H^2, p_T^2, p_H^3, p_T^3$?

HMM we can solve this pb.

Fair $\qquad\qquad$ biased

MARKOV CHAINS

N = # of states

$S = \{s_1, \dots, s_N\}$

A = the state transition probability matrix

$$A \quad \begin{array}{c|ccc|} & s_1 & \cdots s_j \cdots & s_N \\ \hline s_1 & & \vdots & \\ s_i & \cdots & a(i,j) & \\ s_N & & & \\ \hline \end{array}$$

$a(i,j)$ = prob. of transitioning from state s_i to state s_j

---

Providence

s_1 = Sunny
s_2 = Rainy

day by day

A Markov Model

$N = 2$

$S = \{s_1, s_2\}$

	s_1	s_2
s_1	0.8	0.2
s_2	0.9	0.1

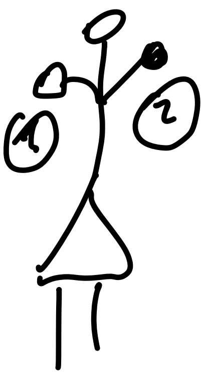
States: are observable

later: <u>Hidden</u> states

"decode": reveal the hidden states

<u>Hidden Markov Models</u> (HMM)

Back to the example



coin #3

T, H, H, T, T, H, T

1 1 2 1 2 2 2

HIDDEN

$M=2$ alphabet $V=\{T, H\}$

sequence of <u>symbols</u> or

sequence of <u>observations</u>

DEF (HMM)

(N, M, A, B, π)

1) N = number of states

$$S = \{S_1, \dots, S_N\}$$

2) M = number of symbols observed per state

$$V = \{v_1, v_2, \dots, v_M\}$$

3) A = the transition probability matrix

$$A = [a_{ij}] \quad \begin{matrix} 1 \leq i \leq N \\ 1 \leq j \leq N \end{matrix}$$

$$a_{ij} = \text{Prob}[q_{t+1} = s_j \mid q_t = s_i]$$

$t = 1, 2, 3, \ldots$ — time units

4) <u>states emit symbols from</u> $V = \{v_1, \ldots, v_M\}$

$$B = [b_j(k)]$$

$$1 \leq k \leq M$$

$$1 \leq j \leq N$$

$$b_j(k) = \text{Prob}[\text{observing symbol } v_k \text{ at time } t \mid q_t = s_j]$$

5) The initial state distribution

$\Pi = \{\pi_i\} \qquad 1 \le i \le N$

$$\pi_i = \mathrm{Prob}\{q_1 = S_i\}$$

A, B, π are probabilistic matrices and vectors

$A: \sum_{j=1}^{N} a_{ij} = 1, \quad 0 \le a_{ij} \le 1$

$\Pi: \sum_{i=1}^{N} \pi_i = 1, \quad 0 \le \pi_i \le 1$ constraints as probabilistic distributions

$B: \sum_{k=1}^{M} b_j(k) = 1$

$0 \le b_j(k) \le 1$

what is time unit?

Markov Chains

$X_1, X_2, X_3 \ldots$

$t = 1, 2, 3, \ldots$

$t=1 \quad q_1$

$t=2 \quad q_2$

$\vdots$

---

3 FUNDAMENTAL COMPUTATIONAL PROBLEMS FOR HMMs

INPUT: an HMM $\lambda = (A, B, \pi)$

λ

A is $N \times N$

B is $N \times M$

π is N

~~and~~ an observation sequence

$O = o_1 o_2 \cdots o_T$

INPUT: $\lambda = (A, B, \pi)$,
O

OUTPUT:

PROBLEM 1: The EVALUATION PB

INPUT: $\lambda = (A, B, \pi)$
$O = o_1 o_2 \cdots o_T$, $o_i \in V = \{v_1 \cdots v_M\}$
$1 \leq i \leq T$

OUTPUT: $P(O|\lambda)$ =
Probability of observing the observation sequence O in the model λ.

PROBLEM 2: $\boxed{\text{The DECODING PB}}$

INPUT: $\lambda = (A, B, \pi)$,
O

OUTPUT: A sequence of states $Q = q_1 q_2 \cdots q_T$

which optimally "explains" the observation sequence".

PROBLEM 3 : The LEARNING PB

INPUT : O

OUTPUT: to find values of the parameters of
$\lambda = (A, B, \pi)$
such that $P(O \mid \lambda) = \text{MAX}$

SOLUTION TO PB 1

Problem 1 is the Evaluation PB, i.e. given seq $\mathcal{O}$ and model λ, compute the probability of observing $\mathcal{O}$ in the model λ.

We wish to compute:

$$\boxed{\Pr(\mathcal{O} \mid \lambda)}$$

$\mathcal{O} = o_1 \cdots o_T$

Remember symbols are emitted by states

$V = \{a, b, c\}$

Ex.

$s_{10}\ s_1\ s_2\ s_5\ s_{10}\ s_1$

$\mathcal{O} = a\,a\,b\,c\,c\,c$

$s_1\ s_3\ s_{10}\ s_1\ s_2\ s_{25}$

For any sequence of states of length T we can emit a sequence of symbols of length T.

This example $O = aabccc$
can be emitted in any sequence
of states of length 6

We have N states

How many sequences of states of
length 6 are there?

N^6

in general N^T for a
sequence of length T.

N^T is exponential

If $N = 50$, $T = 100$

N^T = more than atoms in
universe

$T = 1{,}000{,}000$.

OUR

AMAIZINGLY : Alg $O(N^2 T)$

consider one fixed state sequence

$$Q = q_1, \ldots, q_T \quad , \quad q_1 = \text{initial state}$$

The prob of observing $O = o_1 \ldots o_T$ in the state sequence Q

$$P(O|Q) = P(o_1|q_1)\,P(o_2|q_2) \cdots P(o_T|q_T)$$

<u>statistical independence</u>:
a state s_i emits a symbol completely independent of any other state emission of symbols

$$P(O|Q) = b_{q_1}(o_1)\, b_{q_2}(o_2) \cdots b_{q_T}(o_T)$$

---

what is Prob $P(Q)$?

$$P(Q) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \cdots a_{q_{T-1} q_T}$$

We computed $P(Q)$ and $P(O|Q)$

We want now $P(O, Q)$ joint prob.

$$P(O, Q) = P(O|Q) \cdot P(Q)$$

by Prob. theorem

$\underline{\underline{P(O|\lambda)}}$ = the objective we want to compute

$$P(O|\lambda) = P(O) =$$

$$P(O) = \sum_{\substack{\text{all seq of} \\ \text{states } Q}} P(O|Q) \cdot P(Q)$$

$$= \sum_{q_1, q_2 \cdots q_T} b_{q_1}(o_1)\, b_{q_2}(o_2) \cdots b_{q_T}(o_T) \cdot \pi_{q_1}\, a_{q_1 q_2}\, a_{q_2 q_3} \cdots a_{q_{T-1} q_T}$$

$$\boxed{P(O) = \sum_{q_1, q_2, \cdots q_T} \pi_{q_1}\, b_{q_1}(o_1)\, a_{q_1 q_2}\, b_{q_2}(o_2) \cdots a_{q_{T-1} q_T}\, b_{q_T}(o_T)}$$

$O(N^T)$ alg: impractical.

The FORWARD ALGORITHM

SOL TO PB 1 (Dynamic Programming)

Def a variable called forward variable

$$\alpha_t(i) = P[\underbrace{o_1 o_2 \cdots o_t}_{\text{prefix}}, q_t = S_i]$$

= The probability of the partial observation sequence $o_1 o_2 \cdots o_t$ until time t and being in state S_i at time t.

The FORWARD ALGORITHM

1) INITIALIZATION

$$\alpha_1(i) = \pi_i \, b_i(o_1), \quad 1 \le i \le N$$

2) RECURRENCE

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i)\, a_{ij} \right] b_j(o_{t+1})$$

$$1 \le t \le T-1$$
$$1 \le j \le N$$

3) TERMINATION

$$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

$\Theta(N^2T)$ practical
fast
exact

Solution PB 2 The Viterbi Algorithm.

variable δ

The new

$$\delta_t(i) = \max_{q_1, q_2, \ldots q_{t-1}} P[q_1, q_2 \ldots q_{t-1}, q_t = S_i, O_1 O_2 \ldots O_t]$$

~~est~~ probability) which
≈ the <u>best</u> score (highest prob.)
along the state sequence
accounts for the prefix $O_1 \ldots$
of the obs. sequence, and which
ends in state $q_t = S_i$

TRACE-BACK = to find the seq of states that "best explain" the obs seq O

$\boxed{P(O, \underline{\underline{Q}})}$ = MAX most probable path $\underline{\underline{Q}}$

RABINER TUTORIAL ON HMMs

SOL TO PB 1. More details

The RECURRENCE:

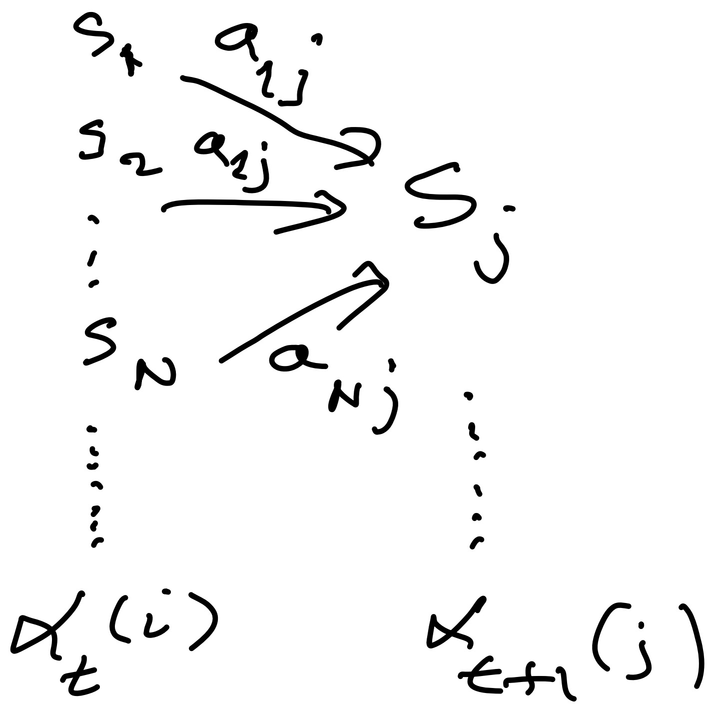
RECURRENCE (step 2)

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i)\, a_{ij}\right] b_j(o_{t+1})$$

$1 \le t \le T-1$

$1 \le j \le N$

step 2



- S_j can be reached at time $t+1$ from N possible states

s_i, $1 \leq i \leq N$, at time t.

- Since $\alpha_t(i)$ is the prob. of the joint event "$\sigma_1 \sigma_2 \ldots \sigma_t$" is observed and "at time t in state s_i"

- Then $\alpha_t(i) \cdot a_{ij}$ = is the prob. of the joint event "$\sigma_1 \sigma_2 \cdots \sigma_t$" observed and state s_j is reached at time $t+1$ via s_i at time t

- Summing up all N states s_i gives the joint prob.

of "$o_1 o_2 \cdots o_t$" observed and ~~state~~ S_j at time $t+1$.

- The concluding $\alpha_{t+1}(j)$ accounts also for the symbol o_{t+1} to be emitted

so we need to multiply by its prob $b_j(o_{t+1})$

- This recurrence is performed for all states $1 \leq j \leq N$ and for all $t = 1, 2, \ldots, T-1$

STEP 3

3 TERMINATION

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

$P(O/\lambda)$ is computed via

$$\alpha_T(i) = P\{o_1, o_2, \dots, o_T, q_T = S_i\}$$

$$\sum_{i=1}^{N} \alpha_T(i) \text{ gives you } P(O).$$

Time complexity is $O(N^2 T)$

SOLUTION TO PB 2

THE VITERBI ALGORITHM

a new variable

$$\delta_t(i) = \max_{q_1 q_2 \cdots q_{t-1}} P\left[q_1, \ldots, q_{t-1}, q_t = S_i, o_1 o_2 \cdots o_t\right]$$

$O = o_1 o_2 \cdots o_T$

$\delta_t(i)$ = the best score (highest probability) along the single path of states, at time t, which accounts for the first t observations $o_1 o_2 \dots o_t$ and ends in state S_i

The Recurrence:

$$\delta_{t+1}(j) = \left[\max_i \delta_t(i)\, a_{ij}\right] b_j(o_{t+1})$$

we will need to backtrack
we will use a new variable for backtracking

THE VITERBI ALGORITHM

1 INITIALIZATION

$$\delta_1(i) = \pi_i b_i(o_1)$$

$$1 \leq i \leq N$$

$$\psi_1(i) = 0 \qquad 1 \leq i \leq N$$

2 RECURRENCE

$$\delta_t(j) = \max_{1 \leq i \leq N} [\delta_{t-1}(i) a_{ij}] b_j(o_t)$$

$$1 \leq j \leq N$$

$$2 \leq t \leq T$$

in PB1

Σ

$$\psi_t(j) = \underset{1 \le i \le N}{\mathrm{Argmax}} \left[\delta_{t-1}(i) a_{ij} \right]$$

$2 \le t \le T$
$1 \le j \le N$

③ TERMINATION

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

$$q_T^* = \underset{1 \le i \le N}{\mathrm{Argmax}} \left[\delta_T[i] \right]$$

q^* = state index

④ BACKTRACKING

$$q_t^* = \psi_{t+1}(q_{t+1}^*)$$

$t = T-1, T-2, \dots, 2, 1$

The matrix Ψ is for backpointers.

Backtracking reconstruct the "alignment" between O and Q_{opt}

$O = o_1 o_2 \cdots, o_T$

$Q_{opt} = q_1 q_2 \cdots q_T$

o_i

$\uparrow$

q_i

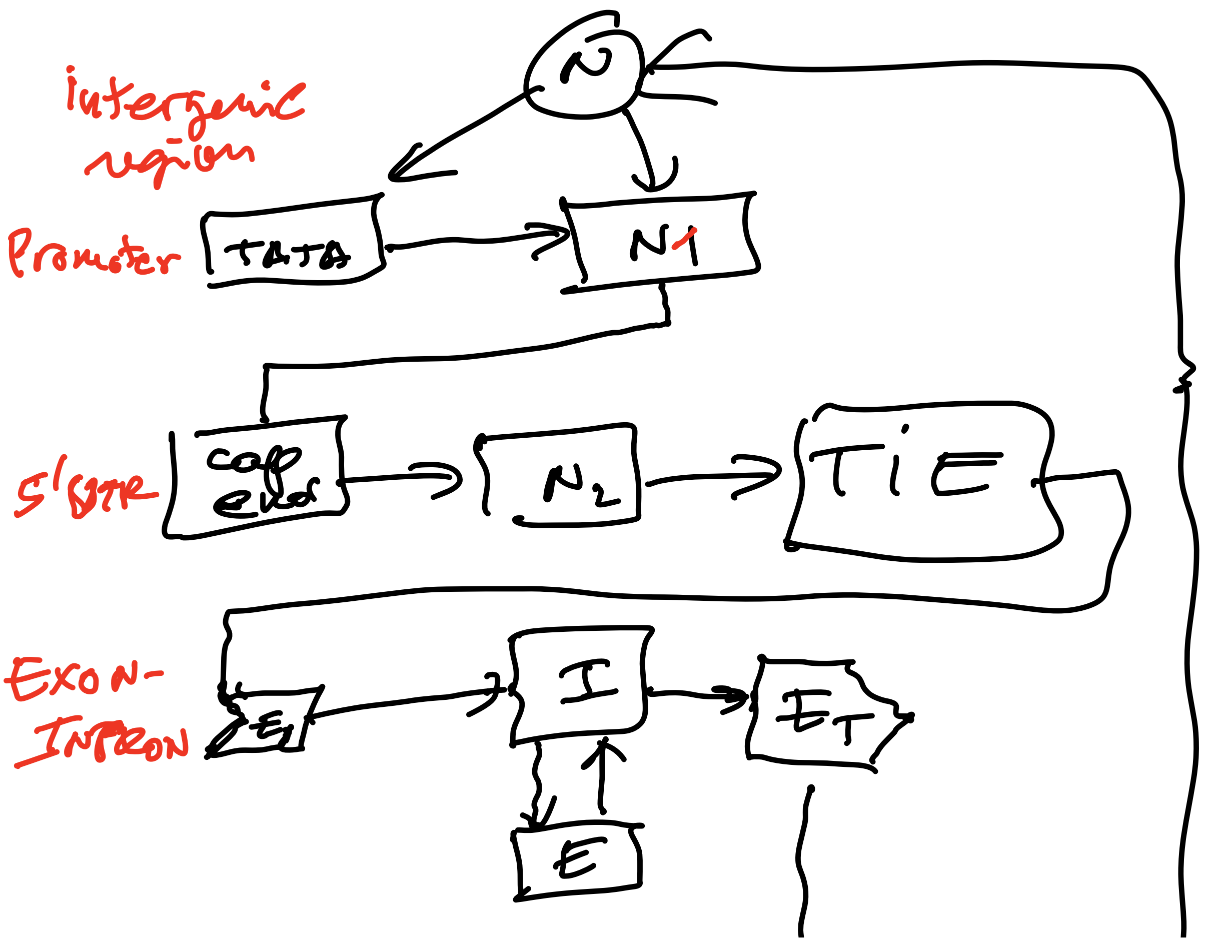
Reconstruction through backtracking goes from right to left

(like in the seq alignment backtracking)

$Q_{opt} = q_1 \cdots q_T$ is the "decoding"

revealed the hidden states

Gene Finding



Post translational region