Markov Chains

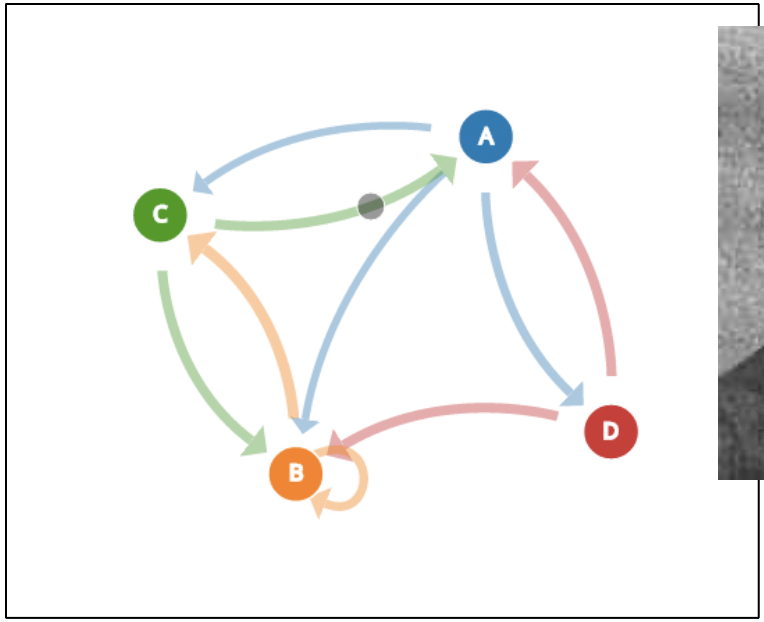
Statistics and data analysis

Ben Galili

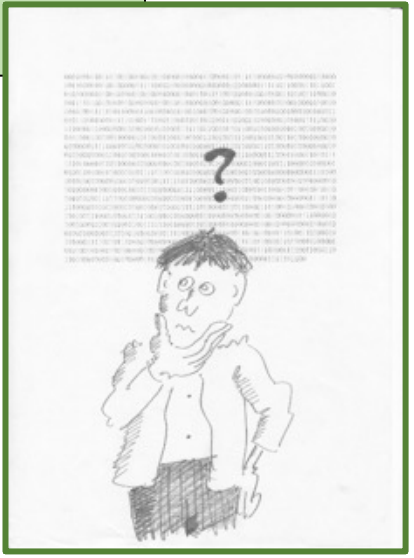
Zohar Yakhini

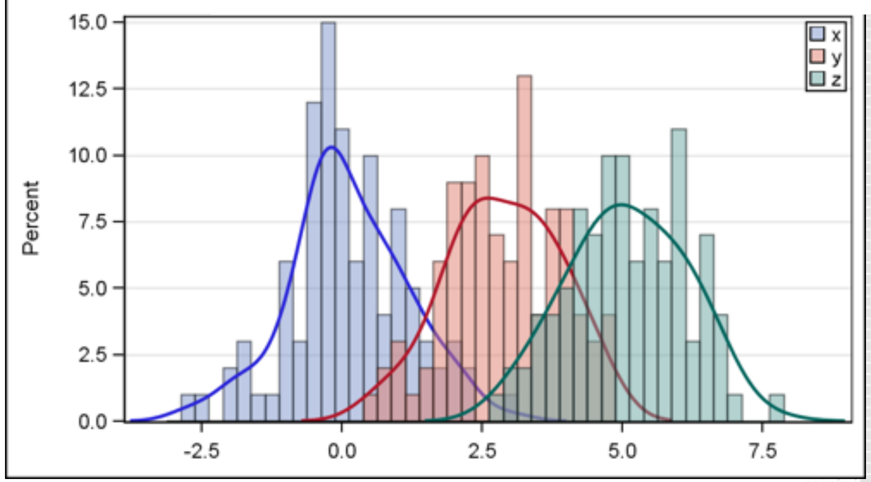
Leon Anavy

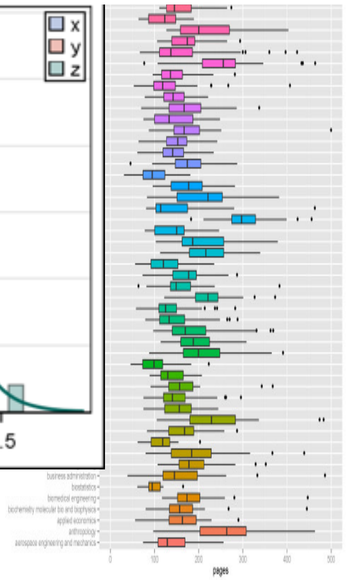
IDC, Herzeliya



00100111010101001010101001001000010
10101000101011111010110100110010 01
11101010100110100101100101101100 10







Stochastic Processes

A (discrete, integer indexed) stochastic process is a sequence of random variables $X_0, X_1, X_2, \dots, X_n, \dots$
with a joint distribution defined over any finite set of variables.

For example, the variables:$(X_8, X_{11}, X_{23}, X_{30}, X_{31}, X_{207})$
have some 6-dimensional probability distribution which is consistent with the (lower-dimensional) probability distributions of all subsets of random variables therein.

Example: independent coin tossing.

IDC HERZLIYA

Stationary Stochastic Processes

A process is called stationary if for every t
the distribution of any finite set of variables $(X_{i1}, X_{i2}, X_{i3}, X_{i4} \;\ldots, X_{ir})$
is the same as the distribution of $(X_{i1+t}, X_{i2+t}, X_{i3+t}, X_{i4+t} \;\ldots, X_{ir+t})$

A process is temporally homogeneous if for every t
the distribution of any finite set of variables $(X_{i1}, X_{i2}, X_{i3}, X_{i4} \;\ldots, X_{ir} | X_{i0})$
is the same as the distribution of $(X_{i1+t}, X_{i2+t}, X_{i3+t}, X_{i4+t} \;\ldots, X_{ir+t} | X_{i0+t})$

Is independent coin tossing w $p = 0.5$ stationary? How about with $p = 0.3$?

IDC HERZLIYA

Markov chains

$$X_0, X_1, X_2, \dots, X_n, \dots$$

s.t. $X_i \in \{Sunny, Cloudy, Rainy\}$

Example sets of outcomes:

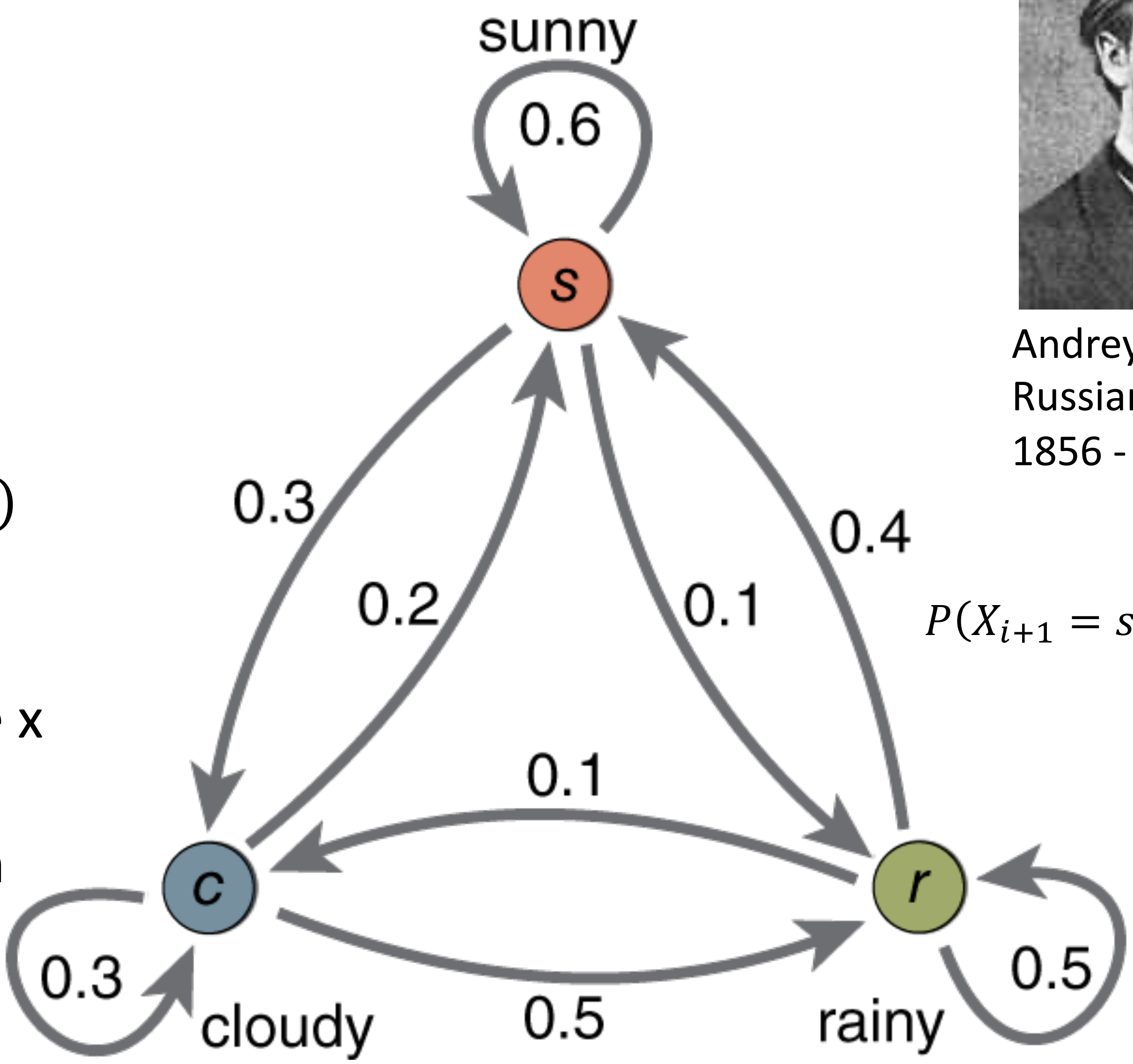
$$(X_8 = s, X_{11} = s, X_{23} = r, X_{30} = r)$$

Transition probability:

The probability of moving from state x to state y

The probability distribution of step n **<u>only depend</u>** on step n-1

$P(X_{i+1} = s | X_i = s) = 0.6$



$P(X_{i+1} = s | X_i = r) = 0.4$

$P(X_{i+1} = r | X_i = c) = 0.5$

Andrey A Markov
Russian Mathematician
1856 - 1922

IDC HERZLIYA

Markov chains

Andrey A Markov
Russian Mathematician
1856 - 1922

A stochastic process $X_0, X_1, X_2, \dots, X_n, \dots$
Taking values in a state space $\{S_1, \dots, S_k, \dots\}$
Is said to be a (finite state space) Markov process if:

- It is **temporally homogeneous**
- It satisfies the **Markovian property**
  That is, for all $t = 0, 1, 2, \dots$ and for every possible set of $t + 1$ states, including S_i and S_j:

$$P(X_{t+1} = j \mid X_0 = k_0, X_1 = k_1, \dots, X_{t-1} = k_{t-1}, X_t = i) = P(X_{t+1} = j \mid X_t = i)$$

Markov chains

A sequence of random variables $X_0, X_1, X_2, \dots, X_n, \dots$
A finite state space $\{S_1, \dots, S_k\}$
An initial distribution vector π_0 : $P(X_0 = S_i) = \pi_0(i)$
A transition square matrix T : $T_{i,j} = P\left(X_{n+1} = S_j | X_n = S_i\right)$

$$\pi_{n+1}(j) = P\left(X_{n+1} = S_j\right)$$

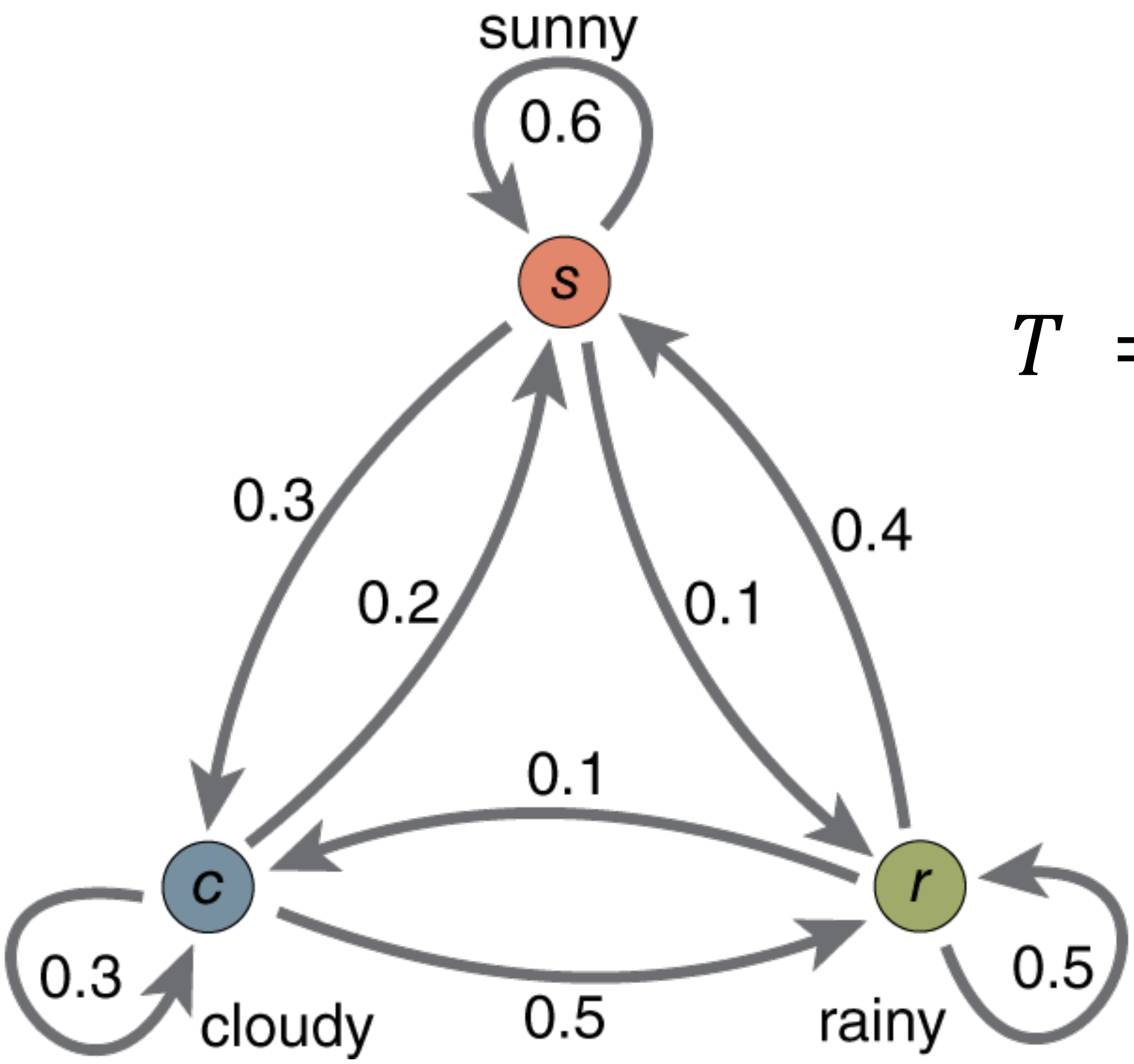
$$= \sum_{i=1}^{k} P(X_n = S_i) P\left(X_{n+1} = S_j | X_n = S_i\right)$$

$$= \sum_{i=1}^{k} \pi_n(i) T_{i,j}$$

$$\begin{bmatrix} T(1,1) & T(1,2) & \cdots & & T(1,k) \\ T(2,1) & T(2,2) & & & T(2,k) \\ \vdots & & \ddots & & \vdots \\ & & \cdots & & \\ T(k,1) & T(k,2) & & T(k,k-1) & T(k,k) \end{bmatrix}$$

The transition rule:

$$\pi_{n+1} = \pi_n \cdot T$$



$$T = \begin{array}{c|ccc} & s & c & r \\ \hline s & 0.6 & 0.3 & 0.1 \\ c & 0.2 & 0.3 & 0.5 \\ r & 0.4 & 0.1 & 0.5 \end{array}$$

Independent coin tossing

What is the transition matrix?

Fair coin (p=0.5)

$$\begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$$

Biased coin (p=0.7)

$$\begin{pmatrix} 0.3 & 0.7 \\ 0.3 & 0.7 \end{pmatrix}$$

What is the initial probability distribution?

Fair coin (p=0.5)

$$\begin{pmatrix} 0.5 & 0.5 \end{pmatrix}$$

Biased coin (p=0.7)

$$\begin{pmatrix} 0.3 & 0.7 \end{pmatrix}$$

Transitions further into the future

		Weather on next day		
		Dry	Wet	Total
Weather	Dry	57	12	69
on one day	Wet	12	8	20

$$T = \begin{array}{c|cc} & 0 & 1 \\ \hline 0 & 0.826 & 0.174 \\ 1 & 0.600 & 0.400 \end{array} = \begin{bmatrix} T(0,0) & T(0,1) \\ T(1,0) & T(1,1) \end{bmatrix}$$

Two days into the future:

$$P(X_2 = 0 |\, X_0 = 1) = ?$$

Transitions further into the future

$$T = \begin{bmatrix} T(0,0) & T(0,1) \\ T(1,0) & T(1,1) \end{bmatrix}$$

	0	1
0	0.826	0.174
1	0.600	0.400

$$P(X_2 = 0 | X_0 = 1) = ?$$

$$\pi_1 = \pi_0 \cdot T = (0,1) \cdot T = (\mathrm{T}(1,0), \mathrm{T}(1,1))$$

$$P(X_2 = 0 | X_0 = 1) = \pi_2(0) = \pi_1 \cdot \begin{bmatrix} T(0,0) \\ T(1,0) \end{bmatrix} = T(1,0)T(0,0) + T(1,1)T(1,0)$$

And therefore, we get :

$$0.6 \cdot 0.826 + 0.4 \cdot 0.6 = 0.7356$$

IDC HERZLIYA

Transitions further into the future: $T^{(r)} = T^r$

What if we want to talk about day 7?
Note that the above calculation actually means that:

$$\pi_2 = \pi_1 \cdot T = (\pi_0 \cdot T) \cdot T = \pi_0 \cdot T^2$$

and we can continue to get

$$\pi_r = \pi_0 \cdot T^r$$

which we can summarize as: $T^{(r)} = T^r$

IDC HERZLIYA

The stationary distribution

A probability distribution, σ , over the state space $\{S_1, \dots, S_k\}$ that satisfies:

$$\sigma \cdot T = \sigma$$

$$\pi_0 = \sigma \rightarrow \pi_1 = \pi_0 \cdot T = \sigma \cdot T = \sigma = \pi_0 \rightarrow \pi_n = \sigma \; \forall n$$

IDC HERZLIYA

The stationary distribution

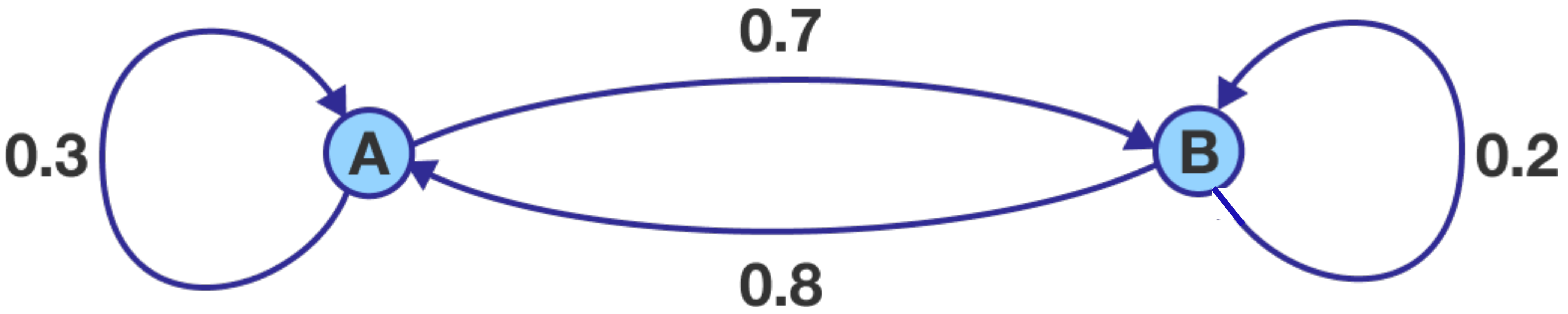
$$\sigma \cdot T = \sigma$$

- 1 is an eigenvalue of T
- σ is a left eigenvector of T, with eigenvalue 1

$$\begin{pmatrix} \qquad & \qquad \end{pmatrix}\begin{pmatrix} \quad \end{pmatrix} = 1\begin{pmatrix} \quad \end{pmatrix}$$

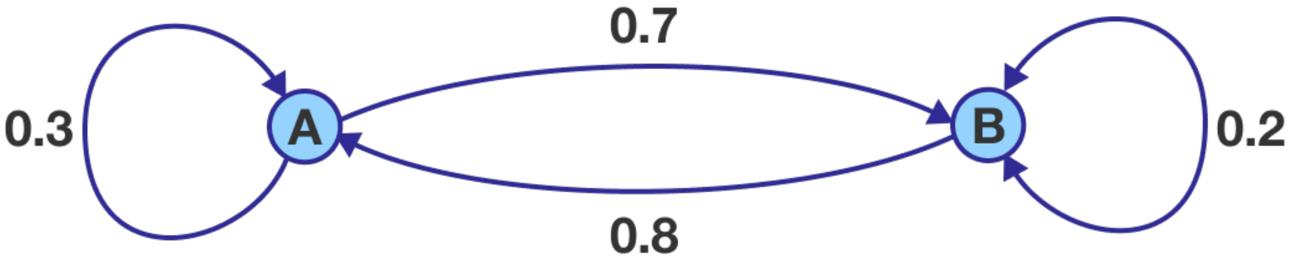
$$\sum_{j=1}^{k} T_{i,j} v_j = v_j \rightarrow v = \vec{\mathbf{1}}$$

What is the stationary distribution here?



$$T = \begin{bmatrix} 0.3 & 0.7 \\ 0.8 & 0.2 \end{bmatrix}$$

What is the stationary distribution here?



$$T = \begin{bmatrix} 0.3 & 0.7 \\ 0.8 & 0.2 \end{bmatrix}$$

$$(x \quad y)\mathrm{T} = (x \quad y)\begin{pmatrix} 0.3 & 0.7 \\ 0.8 & 0.2 \end{pmatrix} = (x \quad y)$$

$$0.3\mathrm{x} + 0.8\mathrm{y} = \mathrm{x} \rightarrow \mathrm{x} = \frac{0.8\mathrm{y}}{0.7}$$

$$0.7\mathrm{x} + 0.2\mathrm{y} = \mathrm{y}$$

$$x + y = 1 \rightarrow \frac{0.8y}{0.7} + y = 1 \rightarrow 1.5y = 0.7 \rightarrow y = \frac{7}{15}, x = \frac{8}{15}$$

Higher order Markov

$$P(X_{t+1} = j \mid X_0 = k_0, X_1 = k_1, \dots, X_{t-1} = k_{t-1}, X_t = i) = P(X_{t+1} = j \mid X_t = i)$$

Need more terms in the condition.
Will depend not only on the most
recent time but on r recent times.

IDC HERZLIYA

Higher order Markov

$$P(X_{t+1} = j \ | X_0 = k_0, X_1 = k_1, \dots, X_{t-r+1} = k_{t-r+1}, \dots, X_{t-1} = k_{t-1}, X_t = k_t) =$$

$$P(X_{t+1} = j \ | X_{t-r+1} = k_{t-r+1}, \dots, X_{t-1} = k_{t-1}, X_t = k_t)$$

r most recent events

IDC HERZLIYA

Markov chains application: text prediction

- Use the previous N-1 words in a sequence to predict the next word
- In auto completion and in auto translation these are called N-Grams
- Language Model (LM)
  - unigrams, bigrams, trigrams,…

Using N-Grams

- For N-gram models $P(w_n|w_1^{n-1}) \approx P(w_n|w_{n-N+1}^{n-1})$
- Bi-grams: N = 2. A Markov assumption.
- By the Chain Rule we can compute probabilities of sentences:

$P(w_1, w_2, w_3, \dots, w_n) =$

$$= P(w_n|w_1, w_2, \dots, w_{n-1})P(w_{n-1}|w_1, w_2, \dots, w_{n-2}) \dots P(w_2|w_1)P(w_1)$$

What would this be under a Markov assumption?

$$= P(w_n|w_{n-1})P(w_{n-1}|w_{n-2}) \dots P(w_2|w_1)P(w_1)$$

IDC HERZLIYA

Bigram Counts - example

- Out of 9222 sentences

	i	want	to	eat	chinese	food	lunch	spend
i	5	827	0	9	0	0	0	2
want	2	0	608	1	6	6	5	1
to	2	0	4	686	2	0	6	211
eat	0	0	2	0	16	2	42	0
chinese	1	0	0	0	0	82	1	0
food	15	0	15	0	1	4	0	0
lunch	2	0	0	0	0	1	0	0
spend	1	0	1	0	0	0	0	0

i	want	to	eat	chinese	food	lunch	spend
2533	927	2417	746	158	1093	341	278

IDC HERZLIYA

Using the Bigram

$$P(\text{"}i\ want\ chinese\ food\text{"}) =$$

$$P(i) * P(want|i) * P(chinese \mid want) \ * P(\ food \mid chinese) =$$

$$\frac{\#(i)}{\#(all\ sentences)} * \frac{\#(\text{"}i\ want\text{"})}{\#(i)} * \frac{\#(\text{"}want\ chinese\text{"})}{\#(want)} * \frac{\#(\text{"}chinese\ food\text{"})}{\#(chinese)} =$$

$$\left(\frac{2533}{9222}\right) * \left(\frac{827}{2533}\right) * \left(\frac{6}{927}\right) * \left(\frac{82}{158}\right) = 0.0003$$

IDC HERZLIYA

Useful for …

- Speech recognition:
  “I ate a cherry” is a more likely sentence than “Eye eight a Jerry”
- Machine translation
  Pr[high acceptance rate] > Pr[tall acceptance rate]
  Pr[finite list] >? Pr[final list]
- Context sensitive spelling correction
  “Their are problems wit this sentence.”
- Sentence completion “Please turn off your …” , “I want to eat …”

CLT for Markov chains

In the homework …

$Cov(X_i, X_{i+t})$

- How can we compute it?
- Assume stationarity for simplicity
- First, note that $Cov(X_i, X_{i+t}) = Cov(X_0, X_t)$
- How would we compute $Cov(X_0, X_1)$?
- Use the definition and compute from the joint distribution, over k^2 elements
- Now use the fact that $T^{(t)} = T^t$

IDC HERZLIYA

Summary

- Brief intro to Markov Chains
- The transition matrix and its powers
- The stationary distribution
- N-grams
- Covariance and the CLT