Markov Chains and Markov Models

Huizhen Yu

janey.yu@cs.helsinki.fi
Dept. Computer Science, Univ. of Helsinki

Probabilistic Models, Spring, 2010

Outline

Outline

Markov Chains

Some facts:

- Named after the Russian mathematician A. A. Markov
- The earliest and "simplest" model for dependent variables
- Has many fascinating aspects and a wide range of applications
  - Markov chains are often used in studying temporal and sequence data, for modeling short-range dependences (e.g., in biological sequence analysis), as well as for analyzing long-term behavior of systems (e.g., in queueing systems).
  - Markov chains are also the basis of sampling-based computation methods called Markov Chain Monte Carlo.

We introduce Markov chains and study a small part of its properties, most of which relate to modeling short-range dependences.

A Motivating Example

Are DNA sequences purely random?

- A sequence segment of bases A, C, G, T from the human preproglucagon gene:

  GTATTAAATCCGTAGTCTGAACTAACTA $\cdots$

- "Words" – such as 'CTGAC' – are suspected to serve some biological function if they seem to occur often in a segment of the sequence.
- So it is of interest to measure the "ofteness" of the occurrences of a "word." A popular way is to model the sequence at the "background."
- Observed frequencies/proportions of pairs of consecutive bases from a sequence of 1571 bases:

1st base \ 2nd base	A	C	G	T
A	0.359	0.143	0.167	0.331
C	0.384	0.156	0.023	0.437
G	0.305	0.199	0.150	0.345
T	0.284	0.182	0.177	0.357
overall	0.328	0.167	0.144	0.360

Outline

Recall Definitions of Independence

Recall: for discrete random variables X, Y, Z with joint distribution P, by definition

- X, Y, Z are mutually independent if

$$P(X = x, Y = y, Z = z) = P(X = x)P(Y = y)P(Z = z), \quad \forall x, y, z;$$

- X, Y are conditionally independent given Z, i.e., $X \perp Y \,|\, Z$, if

$$P(X = x \,|\, Y = y, Z = z) = P(X = x \,|\, Z = z), \quad \forall x, y, z.$$

(Our convention: the equalities hold for all x, y, z such that the quantities involved are well-defined under P.)

Let $X_1, X_2, \ldots$ be an indexed sequence of discrete random variables with joint probability distribution P.

- If $X_1, X_2, \ldots$ are mutually independent, then by definition, for all n,

$$P(X_1, X_2, \ldots, X_n) = P(X_1)P(X_2)\cdots P(X_n),$$

$$P(X_{n+1} \,|\, X_1, \ldots, X_n) = P(X_{n+1}).$$

Definition of a Markov Chain

The sequence $\{X_n\}$ is called a (discrete-time) Markov chain if it satisfies the *Markov property*: for all $n \geq 1$ and $(x_1, \ldots, x_n)$,

$$P(X_{n+1} = x_{n+1} \,|\, X_1 = x_1, \ldots, X_n = x_n) = P(X_{n+1} = x_{n+1} \,|\, X_n = x_n), \qquad (1)$$

i.e., $X_{n+1} \perp X_1, X_2, \ldots, X_{n-1} \mid X_n$.

Recall: the joint PMF of $X_1, X_2, \ldots X_n$ can be expressed as

$$p(x_1, x_2, \ldots x_n) = p(x_1)p(x_2 \,|\, x_1)p(x_3 \,|\, x_1, x_2)\cdots p(x_n \,|\, x_1, x_2, \ldots x_{n-1}).$$

The Markov property $p(x_i \,|\, x_1, \ldots, x_{i-1}) = p(x_i \,|\, x_{i-1}), \forall i$ then implies that $p(x_1, x_2, \ldots, x_n)$ *factorizes* as

$$p(x_1, x_2, \ldots, x_n) = p(x_1)p(x_2 \,|\, x_1)\cdots p(x_n \,|\, x_{n-1}), \quad \forall n. \qquad (2)$$

In turn, this is equivalent to that for all $m > n$,

$$p(x_{n+1}, x_{n+2}, \ldots, x_m \,|\, x_1, x_2, \ldots, x_n) = p(x_{n+1}, x_{n+2}, \ldots, x_m \,|\, x_n).$$

Informally, the "future" is independent of the "past" given the "present."

Terminologies

- X_n: the *state* of the chain at time n
- $S = \{\text{possible } x_n, \forall n\}$: the *state space* (we assume S is finite)
- $P(X_{n+1} = j \,|\, X_n = i), i, j \in S$: the *transition probabilities*
- The chain is said to be *homogeneous*, if for all n and $i, j \in S$,

$$P(X_{n+1} = j \,|\, X_n = i) = p_{ij}$$

  independently of n; and *inhomogeneous*, otherwise.
- For a homogeneous chain, the $|S| \times |S|$ matrix

$$\begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mm} \end{bmatrix} \quad \text{where } m = |S|,$$

  is called the *transition probability matrix* of the Markov chain, (or *transition matrix* for short).

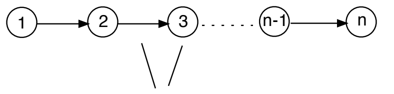
Graphical Model of a Markov Chain

The joint PMF of $X_1, X_2, \dots X_n$ factorizes as

$$p(x_1, x_2, \dots, x_n) = p(x_1)p(x_2 \,|\, x_1) \cdots p(x_n \,|\, x_{n-1}).$$

A pictorial representation of this factorization form:

- A *directed* graph $G = (V, E)$
- Vertex set: $V = \{1, \dots, n\}$
- Edge set: $E = \{(i-1, i), 1 \le i < n\}$
- i is associated with X_i, $\forall i \in V$



Each vertex i with its incoming edge represents a term in the factorized expression of the PMF p, $p(x_i \,|\, x_{i-1})$.

This is a graphical model for Markov chains with length n.

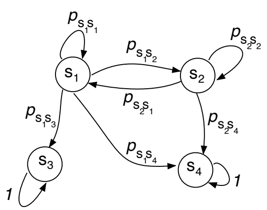
Parameters of a Markov Chain: Transition Probabilities

Transition probabilities determine entirely the behavior of the chain:

$$\begin{aligned} p(x_1, x_2, \dots x_n) &= p(x_1)p(x_2|x_1) \cdots p(x_n|x_{n-1}) \\ &= p(x_1) \prod_{j=1}^{n-1} p_{x_j x_{j+1}} \qquad \text{(for a homoneneous chain)} \end{aligned}$$

For a homogeneous chain, P is determined by $\{p_{ij}, i, j \in S\}$ and the *initial distribution* of X_1. (Thus the number of free parameters is $|S|^2$.)

Transition probability graph: another pictorial representation of a homogeneous Markov chain; it shows the "structure" of the chain in the state space:



- Nodes represent possible states and arcs possible transitions. (Not to be confused with the graphical model.)
- Using this graph one may classify states as being *recurrent*, *absorbing*, or *transient* – notions important for understanding long term behavior of the chain.

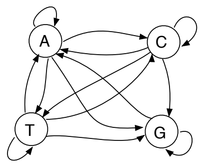
Outline

Simple Examples of DNA Sequence Modeling

A Markov chain model for the DNA sequence shown earlier:

- State space $S = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$
- Transition probabilities (taken to be the observed frequencies)

	A	C	G	T
A	0.359	0.143	0.167	0.331
C	0.384	0.156	0.023	0.437
G	0.305	0.199	0.150	0.345
T	0.284	0.182	0.177	0.357



- The probability of 'CTGAC' given the first base being 'C':

$$P(\mathtt{CTGAC} \mid X_1 = \mathtt{C}) = p_{\mathtt{CT}} \cdot p_{\mathtt{TG}} \cdot p_{\mathtt{GA}} \cdot p_{\mathtt{AC}}$$
$$= 0.437 \times 0.177 \times 0.305 \times 0.143 \approx 0.00337.$$

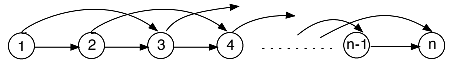
(The probabilities soon become too small to handle as the sequence length grows. In practice we work with $\ln P$ and the log transition probabilities instead: $\ln P(\mathtt{CTGAC} \mid X_1 = \mathtt{C}) = \ln p_{\mathtt{CT}} + \ln p_{\mathtt{TG}} + \ln p_{\mathtt{GA}} + \ln p_{\mathtt{AC}}$.)

Simple Examples of DNA Sequence Modeling

Second-order Markov Chain:

- Joint PMF and graphical model

$$p(x_1, x_2, \ldots x_n) = p(x_1, x_2) p(x_3 | x_1, x_2) \cdots p(x_n | x_{n-1}, x_{n-2})$$

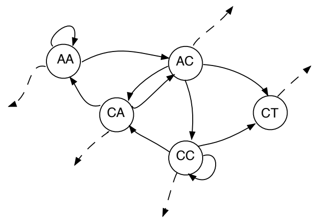


Correspondingly, $\{Y_n\}$ is a Markov chain where

$$Y_n = (Y_{n,1}, Y_{n,2}) = (X_{n-1}, X_n), \quad P(Y_{n+1,1} = Y_{n,2} \mid Y_n) = 1.$$

Second-order model for the DNA sequence example:

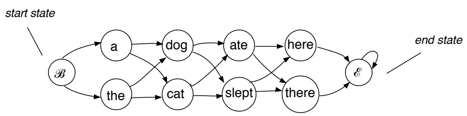
- State space
  $S = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\} \times \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$
- Size of transition probability matrix: 16×16
  The number of parameters grows exponentially with the order.
- Transition probability graph (shown partially on the right)



Higher order Markov chains are analogously defined.

Simple Examples of Language Modeling

A trivial example: the transition probability graph of a Markov chain that can generate a sentence "*a cat slept here*" with start and end:



Sequences of English generated by two Markov models:

- Third-order letter model:
  THE GENERATED JOB PROVIDUAL BETTER TRAND THE DISPLAYED CODE, ABOVERY UPONDULTS WELL THE CODERST IN ...
- First-order word model:
  THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHODS FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD ...

Applications in the biology and language modeling contexts include

- analyzing potentially important "words"
- sequence classification
- coding/compression

Outline

Two Extreme Cases

Two extreme cases

- If $X_1, X_2, \ldots$ are mutually independent, $\{X_n\}$ is certainly a Markov chain (of zeroth-order).
- If $\{X_n\}$ is an arbitrary random sequence, then $\{Y_n\}$, define by $Y_n = (X_1, X_2, \ldots X_n)$, is a Markov chain.
  Because $Y_{n+1} = (Y_n, X_{n+1})$ and in Y_n the entire "past" is "kept."

Notes: Consider a Markov chain $\{X_n\}$.

- At the current state $x_n = s$, the past is not necessarily "forgotten." It may even happen that for some $m < n$, there is only one possible path $(s_m, s_{m+1}, \ldots s_{n-1})$ for $X_m, X_{m+1}, \ldots, X_{n-1}$ such that $X_n = s$.
- But each time the chain visits s, its behavior starting from s is probabilistically the same regardless of the paths it took to reach s in the past.

Geometrically Distributed Duration of Stay

The geometric distribution with parameter $q \in [0, 1]$ has the PMF

$$p(m) = (1-q)^{m-1} q, \quad m = 1, 2, \ldots$$

For independent trials each with success probability q, this is the distribution of the number M of trials until (and including) the first success. The mean and variance of M are

$$E[M] = \frac{1}{q}, \quad var(M) = \frac{1-q}{q^2}.$$

Consider a homogeneous Markov chain. Let s be some state with *self-transition* probability $p_{ss} > 0$.

Let L_s be the duration of stay at s after entering s at some arbitrary time $n+1$:

$$L_s = m \ \text{ if } \ X_n \neq s, \ X_{n+1} = \cdots = X_{n+m} = s, \ X_{n+m+1} \neq s,$$

Then, for $m \geq 1$,

$$P(L_s = m \mid X_n \neq s, X_{n+1} = s) = P(L_s = m \mid X_{n+1} = s) = p_{ss}^{m-1}(1 - p_{ss}).$$

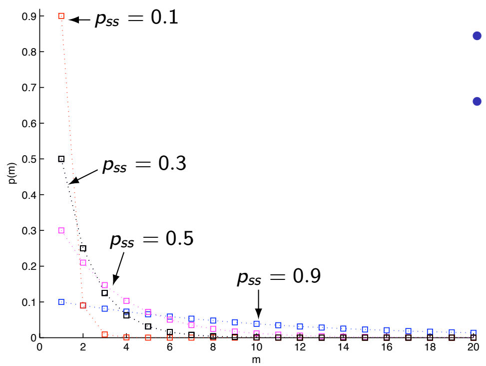
So the duration has the geometric distribution with mean $\frac{1}{1-p_{ss}}$.

Geometrically Distributed Duration of Stay

Geometric distribution has a memoryless property:

$$P\big(\text{time of the first success} = r + m \mid \text{ failed the first } r \text{ trials}\big) = p(m).$$

Illustration of geometric distributions with $q = 1 - p_{ss}$:



- The shape of such distributions may not match that found in data.
- When the duration distribution reflects an important aspect of the nature of data, a general approach is to model the duration distribution explicitly by including as part of the state variable the time already spent at s after entering s.

Subsequences of a Markov Chain

Let $\{n_k\}$ be a sequence of integers with $1 \leq n_1 \leq n_2 \leq \ldots$
Let $Y_k = X_{n_k}, k \geq 1$. Is $\{Y_k\}$ a Markov chain?

We check the form of the joint PMF $p(y_1, \ldots, y_{k+1})$. We have

$$p(y_1, \ldots, y_{k+1}) = \sum_{\substack{i < n_{k+1} \\ i \notin \{n_1, \ldots, n_{k+1}\}}} \sum_{x_i} p(x_1, x_2, \ldots, x_{n_{k+1}}),$$

and by the Markov property of $\{X_n\}$,

$$p(x_1, x_2, \ldots, x_{n_{k+1}}) = p(x_1, x_2, \ldots, x_{n_k}) \cdot p(x_{n_k+1}, x_{n_k+2}, \ldots x_{n_{k+1}} \mid x_{n_k}).$$

So $p(y_1, \ldots, y_{k+1})$ equals

$$\Bigg(\sum_{\substack{i < n_k \\ i \notin \{n_1, \ldots, n_k\}}} \sum_{x_i} p(x_1, x_2, \ldots, x_{n_k}) \Bigg) \cdot \Bigg(\sum_{n_k < i < n_{k+1}} \sum_{x_i} p(x_{n_k+1}, x_{n_k+2}, \ldots x_{n_{k+1}} \mid x_{n_k}) \Bigg),$$

$$= p(x_{n_1}, x_{n_2}, \ldots, x_{n_k}) \cdot p(x_{n_{k+1}} \mid x_{n_k}) = p(y_1, y_2, \ldots, y_k) \cdot p(y_{k+1} \mid y_k).$$

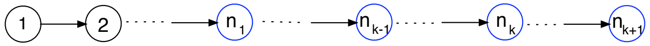
This shows

$$Y_{k+1} \perp Y_1, Y_2, \ldots, Y_{k-1} \mid Y_k,$$

so $\{Y_k\}$ is a Markov chain.

Subsequences of a Markov Chain

The independence relation $X_{n_{k+1}} \perp X_{n_1}, X_{n_2}, \ldots, X_{n_{k-1}} \mid X_{n_k}$ can actually be "read off" from the graphical model:



X_{n_k} "separates" $X_{n_{k+1}}$ from $X_{n_1}, X_{n_2}, \ldots, X_{n_{k-1}}$ in the graph. (The exact meaning of this will be explained together with more general results in the future.)

A special choice of the index sequence $\{n_k\}$ is

$$n_k = (k-1) \cdot m + 1, \quad \text{for some fixed integer } m > 1.$$

The corresponding Markov chain is $X_1, X_{m+1}, X_{2m+1}, \ldots$.

The transition probabilities $p_{ij}^{(m)}$ of this chain (homogeneous case) are the *m-step transition probabilities* of the original chain:

$$p_{ij}^{(m)} = P(X_{m+1} = j \mid X_1 = i), \quad \forall i, j \in S.$$

We examine next the relation between $p_{ij}^{(m)}$ and p_{ij}, $i, j \in S$.

m-step Transition Probabilities of a Markov chain

The *m*-step transition probabilities $p_{ij}^{(m)}$ satisfy the recursive formula

$\forall\, 1 \leq r \leq m$:

$$p_{ij}^{(m)} = P(X_{m+1} = j \mid X_1 = i) = \sum_{\ell \in S} P(X_{r+1} = \ell \mid X_1 = i)\, P(X_{m+1} = j \mid X_{r+1} = \ell)$$
$$= \sum_{\ell \in S} p_{i\ell}^{(r)} p_{\ell j}^{(m-r)}, \qquad \forall i, j \in S. \tag{3}$$

This is called the *Chapman-Kolmogorov equation*. (Once it was conjectured to be an equivalent definition for Markov chains, but this turned out to be false.)

In matrix form, Eq. (3) can be expressed as

$$\widehat{P}^{(m)} = \widehat{P}^m = \widehat{P}^r \cdot \widehat{P}^{m-r}, \quad 1 \leq r \leq m.$$

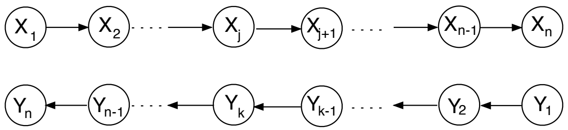
where $\widehat{P}$ denotes the transition probability matrix of $\{X_n\}$, and $\widehat{P}^{(m)}$ denotes the *m-step transition probability matrix* with $\widehat{P}_{ij}^{(m)} = p_{ij}^{(m)}, i, j \in S$.

Reversing the Chain

Let $X_1, X_2, \ldots, X_n$ be a Markov chain.

Let $Y_k = X_{n+1-k}, 1 \leq k \leq n$. Is $\{Y_k\}$ a Markov chain?

j, k: j+k = n + 1



The joint PMF $p(x_j, x_{j+1}, \ldots, x_n)$ can be expressed as

$$\begin{aligned} p(x_j, x_{j+1}, \ldots, x_n) &= p(x_j, x_{j+1}) \cdot p(x_{j+2}, \ldots, x_n \mid x_{j+1}) \\ &= p(x_j \mid x_{j+1}) \cdot p(x_{j+1}) \cdot p(x_{j+2}, \ldots, x_n \mid x_{j+1}) \\ &= p(x_j \mid x_{j+1}) \cdot p(x_{j+1}, x_{j+2}, \ldots, x_n), \end{aligned}$$

so

$$p(x_j \mid x_{j+1}, \ldots, x_n) = p(x_j \mid x_{j+1}).$$

This shows the reversed sequence $\{Y_k\}$ is a Markov chain.

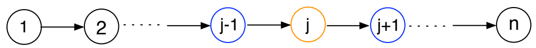
Another Conditional Independence Relation

Let $X_1, X_2, \ldots, X_n$ be a Markov chain.

Denote $X_{-j} = (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_n)$, all variables but X_j.
Then, for all $j \leq n$,

$$P(X_j = x_j \mid X_{-j} = x_{-j}) = P(X_j = x_j \mid X_{j-1} = x_{j-1}, X_{j+1} = x_{j+1}). \tag{4}$$

Visualize the positions of the variables in the graph:



Derivation of Eq. (4): the joint PMF $p(x_j, x_{-j})$ can be written as

$$p(x_j, x_{-j}) = p(x_1, \ldots, x_{j-1}) \cdot p(x_j \mid x_{j-1}) p(x_{j+1} \mid x_j) \cdot p(x_{j+2}, \ldots, x_n \mid x_{j+1}),$$

so it has the form

$$p(x_j, x_{-j}) = h(x_1, \ldots, x_{j-1}) \cdot g(x_j, x_{j-1}, x_{j+1}) \cdot \omega(x_{j+1}, \ldots, x_n)$$

for some functions h, g, ω. This shows (one of the exercises) that

$$X_j \perp X_1, \ldots, X_{j-2}, X_{j+2}, \ldots, X_n \mid X_{j-1}, X_{j+1}.$$

Outline

Hidden Markov Models (HMM)

Hidden Markov Models refer loosely to a broad class of models in which

- a subset of the random variables, denoted by $X = \{X_1, X_2, \ldots\}$, is modeled as a Markov chain, and their values are not observed in practice;
- the rest of the random variables, denoted by Y, are observable in practice, and $P(Y \mid X)$ usually has a simple form.

A common case: $Y = \{Y_1, Y_2, \ldots\}$, and conditional on X, Y_is are independent with

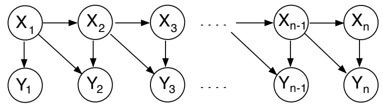
$$P(Y_i \mid X) = P(Y_i \mid X_{i-1}, X_i).$$

(Treat X_0 as a dummy variable.) The joint PMF of (X, Y) then factorizes as

$$p(x, y) = \prod_i p(x_i \mid x_{i-1}) p(y_i \mid x_{i-1}, x_i).$$

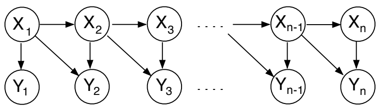
$p(y_i \mid x_{i-1}, x_i)$: often called *observation (or emission) probabilities*

A graphical model indicating the factorization form of p:



The sequence of (X_i, Y_i) jointly is a Markov chain.

HMM and Data-Generating Processes



In some applications, the above structure of HMM corresponds intuitively to the data-generating process: for instance,

- Speech recognition:
  X – speech, Y – acoustic signals
- Tracking a moving target:
  X – positions/velocities of the target, Y – signals detected
- Robot navigation:
  X – positions of a robot, Y – observations of landmarks

In other applications, the model can have nothing to do with the underlying data-generating process, and is introduced purely for questions at hand. Examples include sequence alignment, sequence labeling applications.

HMM for Parts-of-Speech Tagging

Parts-of-speech tags for a sentence:

pron	*v*	*adv*	*final punct.*
`I`	`drove`	`home`	.

- Possible tags:
  `I`: *n, pron*
  `drove`: *v, n*
  `home`: *n, adj, adv, v*

Represent a sentence as a sequence of words $W = (W_1, W_2, \ldots, W_n)$. Let the associated tags be $T = (T_1, T_2, \ldots, T_n)$.

A Markov chain model for $(W_i, T_i), 1 \le i \le n$ used in practice is:

$$p(w, t) = \prod_{i=1}^{n} p(w_i \mid t_i) p(t_i \mid t_{i-1}, t_{i-2}).$$

Question: Why W is treated as “generated” by T and not the other way around?

Another Artificial Example of Sequence Labeling

Suppose two different types of regions, type '+' and type '-', can occur in the same DNA sequence, and each type has its own characteristic pattern of transitions among the bases A, C, G, T. Given a sequence un-annotated with region types, we want to find where changes between the two types may occur and whether they occur sharply.

An HMM for this problem:

- Let $Y = \{Y_i\}$ correspond to a DNA sequence.
- Introduce variables Z_i with $z_i \in \{+, -\}$ to indicate which type is in force at position i.
- Let $X_i = (Y_i, Z_i)$, and model $X = \{X_i\}$ as a Markov chain on $S = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\} \times \{+, -\}$.

Then, for an un-annotated sequence y, we solve

$$\arg\max_x P(X = x \,|\, Y = y), \quad \text{equivalently}, \quad \arg\max_z P(Z = z \,|\, Y = y).$$

Any z^* in the latter set gives one of the most probable configurations of boundaries between type '+' and type '-' regions for the sequence y.

Inference and Learning with HMM

Quantities of interest in applications usually include

- $P(Y = y)$, and $P(X = x \,|\, Y = y)$ for a single x
- $\arg\max_x P(X = x \,|\, Y = y)$, the most probable path given $Y = y$
- $P(X_i \,|\, Y = y)$, the marginal distribution of X_i given $Y = y$; and $\arg\max_{x_i} P(X_i = x_i \,|\, Y = y)$

Efficient inference algorithms are available – utilizing the structure/factorization form of p, computation can be streamlined.

To specify an HMM, we need to specify its topology (space of X_i, relation between $\{X_i\}$ and $\{Y_i\}$), and then its parameters.

Parameters of HMM in the earlier examples:
transition probabilities and observation probabilities

Parameter estimation:

- Complete data case: sequences of states $\{x_i\}$ are given for estimation
  This case reduces to the case of Markov chains.
- Incomplete data case: sequences of states $\{x_i\}$ are unknown
  In this case estimation is based on the observed sequences of $\{y_i\}$, and is typically done with the so-called expectation-maximization (EM) algorithm.

We will study these topics in some future lectures.

HMM vs. High Order Markov Models

In the application contexts shown earlier, the main interest is on the hidden/latent variables $\{X_i\}$.

Consider now the case where our interest is solely on $\{Y_i\}$ (for instance, to predict Y_{n+1} given $Y_1, \ldots, Y_n$). We compare two choices of models for $\{Y_i\}$:

- a Markov model for $\{Y_i\}$, possibly of high order;
- an HMM for $\{Y_i\}$, in which we introduce auxiliary, latent random variables $\{X_i\}$.

In an HMM for $\{Y_i\}$,

$$p(y_1, y_2, \ldots, y_n) = \sum_{x_1, \ldots, x_n} \prod_{i=1}^{n} p(x_i \,|\, x_{i-1})\, p(y_i \,|\, x_{i-1}, x_i),$$

so generally, $Y_1, \ldots, Y_n$ are fully dependent as modeled.

In a Markov or high order Markov model, certain conditional independence among $\{Y_i\}$ is assumed.

This shows with HMM we approximate the true distribution of $\{Y_i\}$ by relatively simple distributions of another kind than those in a Markov or high order Markov model.

Further Readings

On Markov chains:

1. A. C. Davison. *Statistical Models*, Cambridge Univ. Press, 2003. Chap. 6.1.

(You may skip materials on pp. 229-232 about classification of states if not interested.)