Ch I Markov models on chains and acyclic graphs

1. Markov models on chains

1A Definitions & basic properties

- Sequence \( S = (S_1, S_n) \) of \( K \)-valued random variables \( S_i \in K \)
- \( K \) is a finite set, its elements are called states
- \( p(S) = p(S_1, S_n) \) is a joint probability distribution on \( K^n \)

W.l.o.g. we can write

\[
p(S_1, S_n) = p(S_1 | S_2, \ldots, S_{n-1}) p(S_n | S_1, \ldots, S_{n-1})
= \ldots
= p(S_1 | S_2, S_{n-1}) p(S_{n-1} | S_1, \ldots, S_{n-2}) \ldots p(S_2 | S_1) p(S_1)
\]

**Definition 1a**: A p.d. on \( K^n \) is a Markov chain if

\[
p(S) = \prod_{i=2}^{n} p(S_i | S_{i-1})
\]

holds \( \forall S \in K^n \)

**Definition 1b**: A p.d. on \( K^n \) is a Markov chain if

\[
p(S) = \prod_{i=2}^{n} g_i(S_{i-1}, S_i)
\]

holds \( \forall S \in K^n \) where \( g_i : K^2 \to \mathbb{R}_+ \) are some functions

**Equivalence**

a) \( \implies \) b) trivial

b) \( \implies \) a) recursively apply the following step

\[
p(S_{n-1}, S_n) = \left\{ \sum_{S_{n-2}} \prod_{l=2}^{n-1} g_l(S_{i-1}, S_i) \right\} g_n(S_{n-1}, S_n)
\]

\[
\implies g_n(S_{n-1}, S_n) = p(S_n | S_{n-1}) f_{n-1}(S_{n-1}) \text{ with same } f_{n-1}
\]
Therefore, we have

\[ p(s_1, s_n) = \left[ \prod_{i=2}^{n-1} \frac{b_i (s_{i-1}, s_i)}{\hat{b}_{n-1}(s_{n-1})} \right] \frac{b_{n-1}(s_{n-1}) \cdot p(s_n | s_{n-1})}{p(s_{n-1}, s_n)} \]

Another useful formula

\[ p(s_1, s_n) = \frac{p(s_1, s_2) \cdot p(s_2, s_3) \cdots p(s_{n-1}, s_n)}{p(s_2) \cdot p(s_3) \cdots p(s_{n-1})} \]

**Example 1 (Ehrenfest model)**

Consider \( N \) particles in two containers. At each discrete time \( t = 1, 2, \ldots \), independently from the past, a particle is selected at random and moved to the other container. Let \( s_t \) denote the number of particles in the first container at time \( t \). Then we have

\[ p(s_t = k | s_{t-1} = l) = \begin{cases} \frac{N-l}{N} & \text{if } k = l+1 \\ \frac{l}{N} & \text{if } k = l-1 \\ 0 & \text{otherwise} \end{cases} \]

**Q:** How do \( p(s_t = k), k=1, \ldots, N \) behave as \( t \to \infty \)?

**Example 2 (Random walk on a graph)**

Consider a random walk on an undirected graph \( V, E \)

- \( V \) = set of vertices,
- \( E \) = set of edges
- \( p(s_t) \) = some p.d. for the start vertex
- \( p(s_t = i | s_{t-1} = j) = \begin{cases} W_{ij} & \text{if } \{i, j\} \in E \\ 0 & \text{otherwise} \end{cases} \)

where the \( W_{ij} \) fulfill \( \sum_{i \in V(j)} W_{ij} = 1 \) for all \( j \in V \).
18. Homogeneous Markov chains, stationary distributions

Definition 2 A Markov chain is homogeneous if its conditional probabilities \( P(s_i | s_{i-1} = k') \) do not depend on the position \( i \), i.e.
\[
P(s_i = k | s_{i-1} = k') = q(k, k') \quad \forall i = 2, \ldots, n.
\]

We know that
\[
P(s_i = k) = \sum_{k' \in K} P(s_i = k | s_{i-1} = k') P(s_{i-1} = k')
\]

Consider \( P(s_i = k), k \in K \) as components of a vector \( \pi_i \in \mathbb{R}^K \)
and \( P(s_i = k | s_{i-1} = k') \), \( k, k' \in K \) as elements of a \( K \times K \) matrix \( P \). Then the previous eq. reads
\[
\pi_i = P \pi_{i-1}
\]
and more general, we have \( \pi_i = \pi_0 = P^{i-1} \pi_0 \).

It may happen that there exists a p.d. \( \pi^* \) on \( K \) s.t. \( P \pi^* = \pi^* \).
We call it a stationary p.d. of \( P \).

Definition 3 A homogeneous Markov chain is irreducible if for each pair of states \( k, k' \) there is an \( m > 0 \) s.t.
\[
P^m_{kk'} > 0.
\]
I.e., there is a non-zero probability to reach state \( k \) starting from state \( k' \) (after \( m \) transitions).

A somewhat stronger condition ensures the existence & uniqueness of a stationary distribution and convergence to it.

Theorem 1 (w/o proof) If for some \( m > 0 \) all elements of the matrix \( P^m \) are strictly positive, then the Markov chain has a unique stationary distribution \( \pi^* \), which is a fixpoint
\[
P^m \pi \rightarrow \pi^* \quad \forall \pi
\]
Moreover
\[ \mathcal{P}^n = \mathcal{P}^* \otimes \mathcal{P} + \mathcal{E}(h), \]
where \( \mathcal{E} = (\epsilon_{hi}) \) and \( \mathcal{E}_{hi} = O(h^n) \) with some \( 0 < h < 1 \).

**Definition 4** A Markov chain satisfies the detailed balance condition if it has a stationary distribution \( \pi \in \mathbb{R}_+^k \) s.t.
\[ \pi(s_i) \pi(s_j) = \pi(s_j) \pi(s_i) \]
This means the reverse Markov chain has the same transition probability matrix as the forward chain.

**E. Hidden Markov models on chains**

Common models in pattern recognition
\( X = (x_1, \ldots, x_n) \) sequence of features (observable)
\( S = (s_1, \ldots, s_n) \) sequence of states (hidden)

**Hidden Markov model (HMM):** a p.d. on pairs \((x, s)\) s.t.

a) \[ p(x, s) = \prod_{i=1}^n p(x_i | s_i) \cdot p(s_n) \cdot \prod_{i=2}^n p(s_i | s_{i-1}) \]

\[ p(x | s) \quad p(s) \text{ - Markov model} \]

b) or slightly more general
\[ p(x, s) = p(s_0) \prod_{i=1}^n p(x_i | s_i) \cdot p(s_i | s_{i-1}) \]

\[ p(x_1, x_2, \ldots, x_n) \]

**Remark** This describes a stochastic regular language.
Conditional HMM

As before, \( X = (x_1, \ldots, x_n) \) - sequence of features and \( S = (s_1, \ldots, s_n) \) - sequence of hidden states

Discriminative model \( \Rightarrow \) we model only \( p(s \mid x) \)

\[
p(s \mid x) = \frac{1}{Z(x)} \frac{1}{n} \prod_{i=2}^{n} g_i(s_{i-1}, s_i, x),
\]

where \( Z(x) \) is a normalisation constant

Such models allow to model a direct dependence of \( s_i \) on a larger context window of features

Hierarchical variational autoencoders & diffusion models

Generative latent variable models (deep learning)

\( Z_0, \ldots, Z_n \) - latent variables (vectors, tensors) \( X \) - image

The model is specified by

\( p(Z_0) \) - simple distribution (uniform, standard Gaussian, etc.)

\( p_i(Z_{i-1} \mid Z_{i-1}) \) - parametrised conditional distributions

\( p(X \mid Z_n) \) - conditional distribution on images

If \( Z_k \in \mathbb{R}^k \), i.e. \( Z_k \) is a binary valued vector \( \Rightarrow \)

\[
\log p(Z_0 \mid Z_{k-1}) = \langle Z_k, f(Z_{k-1}, \theta) \rangle - C(Z_{k-1}),
\]

where \( f(Z_{k-1}, \theta) \) is modelled by a (deep) network.

\( C(Z_{k-1}) \) is the log-partition function (normalising constant)