Markov models on sequences
Inference algorithms for Markov models
Parameter learning for Markov models
1. Structured hidden states

Models discussed so far: mainly classifiers predicting a categorical (class) variable $y \in \mathcal{Y}$

Often in applications: the hidden state $y$ is a structured variable.

Here: the hidden state $y$ is given by a sequence of categorical variables.

**Application examples:**

- text recognition (printed, handwritten, “in the wild”),
- speech recognition (single word recognition, continuous speech recognition, translation),
- robot self localisation.

**Markov Models and Hidden Markov Models on chains:**
a class of generative probabilistic models for sequences of features and sequences of categorical variables.
2. Markov Models

Let \( s = (s_1, s_2, \ldots, s_n) \) denote a sequence of length \( n \) with elements from a finite set \( K \).

Any joint probability distribution on \( K^n \) can be written as

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

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

\[
p(s) = p(s_1) p(s_2 | s_1) p(s_3 | s_2) \cdot \ldots \cdot p(s_n | s_{n-1}) = p(s_1) \prod_{i=2}^{n} p(s_i | s_{i-1})
\]

holds for any \( s = (s_1, s_2, \ldots, s_n) \).
2. Markov Models

**Example 1** (Random walk on a graph).

- Let \((V, E)\) be a directed graph. A random walk in \((V, E)\) is described by a sequence \(s = (s_1, \ldots, s_t, \ldots)\) of visited nodes, i.e. \(s_t \in V\).

- The walker starts in node \(i \in V\) with probability \(p(s_1 = i)\).

- The edges of the graph are weighted by \(w : E \rightarrow \mathbb{R}_+\), s.t.

\[
\sum_{j : (i, j) \in E} w_{ij} = 1 \quad \forall i \in V
\]

- In the current position \(s_t = i\), the walker randomly chooses an outgoing edge with probability given by the weights and moves along this edge, i.e.

\[
p(s_{t+1} = j \mid s_t = i) = \begin{cases} 
w_{ij} & \text{if } (i, j) \in E \\
0 & \text{otherwise} \end{cases}
\]
3. Algorithms: Computing the most probable sequence

How to compute the most probable sequence \( s^* \in \arg\max_{s \in K^n} \left[ p(s_1) \prod_{i=2}^{n} p(s_i | s_{i-1}) \right] \)?

Take the logarithm of \( p(s) \): \( s^* \in \arg\max_{s \in K^n} \left[ g_1(s_1) + \sum_{i=2}^{n} g_i(s_{i-1}, s_i) \right] \)

and apply dynamic programming: Set \( \phi_1(s_1) \equiv g_1(s_1) \) and compute

\[
\phi_i(s_i) = \max_{s_{i-1} \in K} \left[ \phi_{i-1}(s_{i-1}) + g_i(s_{i-1}, s_i) \right] \quad \forall s_i \in K.
\]

Finally, find \( s^*_n \in \arg\max_{s_n \in K} \phi_n(s_n) \) and back-track the solution. This corresponds to searching the best path in the graph.
3. Algorithms: Computing marginal probabilities

How to compute marginal probabilities for the sequence element $s_j$ in position $j$

$$p(s_j) = \sum_{s_1 \in K} \cdots \sum_{s_j \in K} \cdots \sum_{s_n \in K} p(s_1) \prod_{i=2}^{n} p(s_i | s_{i-1})$$

Summation over the trailing variables is easily done because:

$$\sum_{s_n \in K} p(s_1) \cdots p(s_{n-1} | s_{n-2}) p(s_n | s_{n-1}) = p(s_1) \cdots p(s_{n-1} | s_{n-2})$$

The summation over the leading variables is done dynamically: Begin with $p(s_1)$ and compute

$$p(s_i) = \sum_{s_{i-1} \in K} p(s_i | s_{i-1}) p(s_{i-1}) \quad \forall s_i \in K$$
3. Algorithms: Computing marginal probabilities

This computation is equivalent to a matrix vector multiplication: Consider the values
\[ p(s_i = k \mid s_{i-1} = k') \]
as elements of a matrix \( P_{kk'}(i) \) and the values of \( p(s_i = k') \) as elements
of a vector \( \pi_i \). Then the computation above reads as
\[ \pi_i = P(i)\pi_{i-1}. \]

Remark 1.

- A Markov model is called *homogeneous* if the transition probabilities
  \[ p(s_i = k \mid s_{i-1} = k') \]
do not depend on the position \( i \) in the sequence. In this case the
  formula \( \pi_i = P_{i-1}^{i-1}\pi_1 \) holds for the computation of the marginal probabilities.

- Notice that the preferred direction (from first to last) in the Def. 1 of a Markov model
  is only apparent. By computing the marginal probabilities \( p(s_i) \) and by using
  \[ p(s_i \mid s_{i-1})p(s_{i-1}) = p(s_{i-1}, s_i) = p(s_{i-1} \mid s_i)p(s_i), \]
  we can rewrite the model in reverse order.
3. Algorithms: Learning a Markov model

Suppose we are given i.i.d. training data $T^m = \{ s^j \in K^n | j = 1, \ldots, m \}$ and want to estimate the parameters of the Markov model by the maximum likelihood estimate. This is very easy:

- Denote by $\alpha(s_{i-1} = \ell, s_i = k)$ the number of sequences in $T^m$ for which $s_{i-1} = \ell$ and $s_i = k$.

- The estimates for the conditional probabilities are then given by

$$p(s_i = k | s_{i-1} = \ell) = \frac{\alpha(s_{i-1} = \ell, s_i = k)}{\sum_k \alpha(s_{i-1} = \ell, s_i = k)}.$$

**Proof (idea):**

Consider all terms in the log-likelihood that depend on the transition probability from $(i-1) \rightarrow i$ and rewrite them using transition counts $\alpha(s_{i-1} = \ell, s_i = k)$

$$\frac{1}{m} \sum_{s \in T^m} \log p(s_i | s_{i-1}) = \frac{1}{m} \sum_{k, \ell \in K} \alpha(s_{i-1} = \ell, s_i = k) \log p(s_i = k | s_{i-1} = \ell)$$

Maximise this w.r.t. $p(s_i | s_{i-1})$ under the constraint $\sum_{s_i \in K} p(s_i | s_{i-1}) = 1$. 
3. Algorithms: Learning a Markov model

Markov models are **exponential families**. For simplicity we show this for the family of homogeneous Markov models on sequences $s = (s_1, s_2, \ldots, s_n)$ of length $n$ under the additional assumption that $p(s_1) = \frac{1}{K}$.

We have

$$p(s) = \frac{1}{K} \prod_{i=2}^{n} p(s_i | s_{i-1})$$

- **sufficient statistic:** $\Phi(s)$ is a $K \times K$ matrix with entries $\Phi_{kl}(s)$ counting the number of transitions from state $l$ to state $k$ in the sequence $s$.

- **natural parameter:** $H$ is a $K \times K$ matrix with entries $H_{kl} = \log p(s_i = k | s_{i-1} = l)$

We can write the probability of sequences as

$$p(s; H) = \exp [\langle \Phi(s), H \rangle - \log(K)]$$

**Remark 2.** This can be generalised for models with non-uniform $p(s_1)$ and also for general (i.e. non-homogeneous) Markov models.
4. Return times and limiting distributions

- A homogeneous Markov model is *irreducible* if each state \( l \) can be reached starting from any state \( k \) with non-zero probability (after some number of transitions).

- A state \( k \) has *return time* \( \tau \) if it can be reached with non-zero probability after \( \tau \) transitions when starting from itself.

- A state \( k \in K \) is *a-periodic* if the greatest common divisor of its return times is 1.

**Theorem 1.** Let \( P \) be the transition probability matrix of an irreducible homogeneous Markov model with a-periodic states. Then there exists a unique marginal probability vector \( \pi^* \) s.t. \( P\pi^* = \pi^* \). Moreover, it is a limiting distribution, i.e.

\[
\lim_{t \to \infty} P^t \pi = \pi^*
\]

for arbitrary starting distributions \( \pi \).

**Q:** What conditions on the graph in Example 1 ensure that this theorem applies for the random walk considered there?