Hidden Markov Models

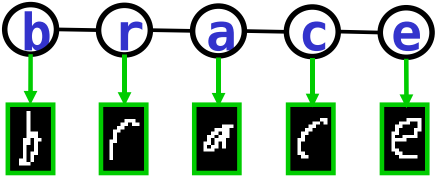
Guillaume Obozinski

Swiss Data Science Center

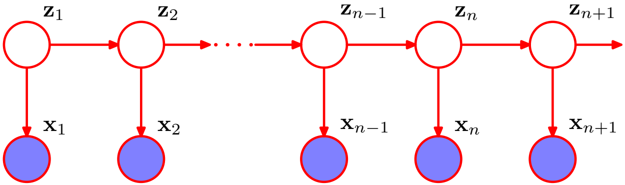
African Masters of Machine Intelligence, 2018-2019, AIMS, Kigali

Hidden Markov Model (HMM)

- voice recognition
- natural langage processing
- handwritten character recognition
- modelling biological sequence (protein, DNA)



Hidden Markov Model (HMM)

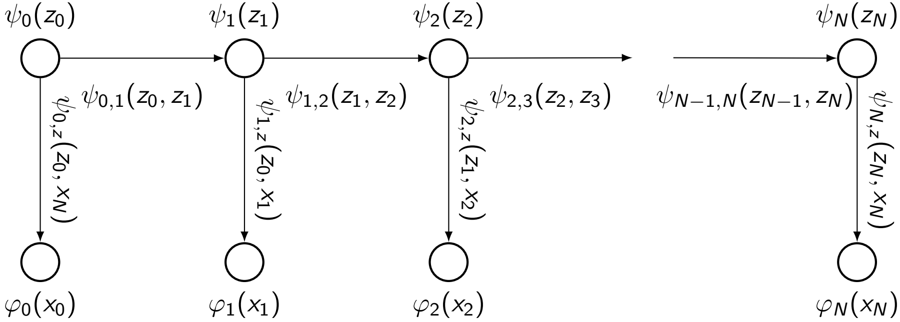


$$p(\mathbf{x}_1, \dots, \mathbf{x}_N, \boldsymbol{z}_1, \dots, \boldsymbol{z}_N) = p(\boldsymbol{z}_1) \prod_{i=2}^{N} p(\boldsymbol{z}_i | \boldsymbol{z}_{i-1}) \prod_{i=1} p(\mathbf{x}_i | \boldsymbol{z}_i)$$

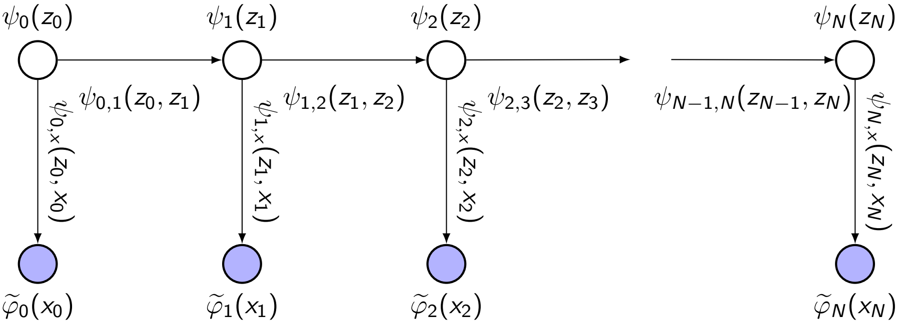
Homogeneous Markov chains

- $\boldsymbol{z}_i \in \{0,1\}^K$ state indicator variable $(1, \dots, K)$
- *homogeneous* Markov chain: $\forall n,\ p(\boldsymbol{z}_i | \boldsymbol{z}_{i-1}) = p(\boldsymbol{z}_2 | \boldsymbol{z}_1)$
- emitted symbol $\mathbf{x}_i$ $(\{0,1\}^K)$ / observation $(\mathbb{R}^d)$

Gibbs model for the HMM

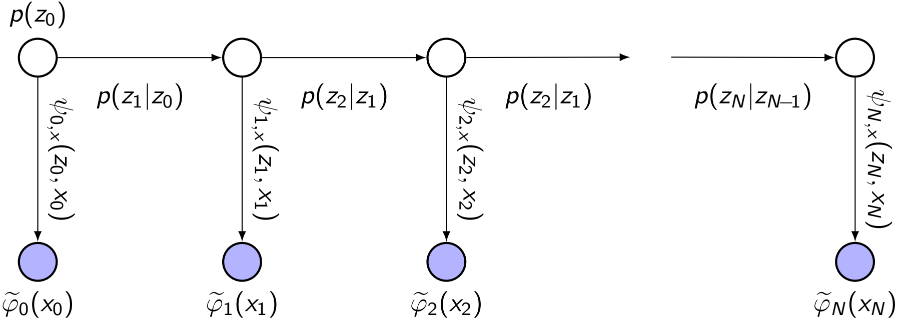


Gibbs model for $p(z \mid \mathbf{x})$



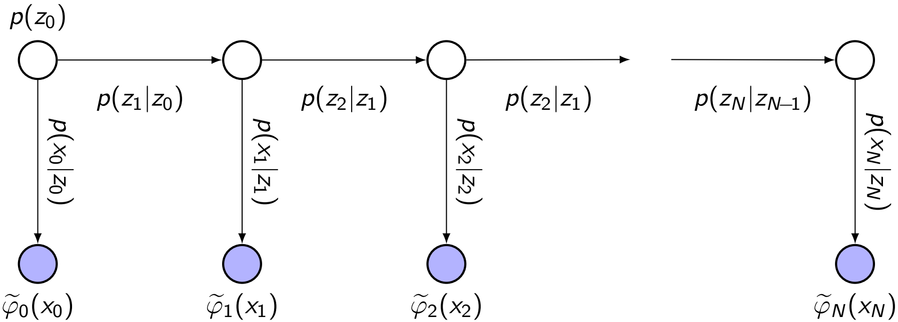
- with $\widetilde{\varphi}_i(x_i) = \varphi_i(x_i)\,\delta(x_i, x_i^{\mathrm{obs}})$

Gibbs model for $p(z \mid \mathbf{x})$ with explicit potentials



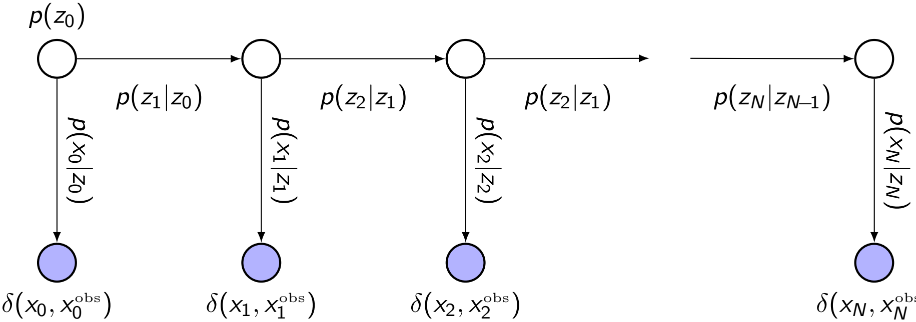
- with $\widetilde{\varphi}_i(x_i) = \varphi_i(x_i)\,\delta(x_i, x_i^{\mathrm{obs}})$

Gibbs model for $p(z \mid \mathbf{x})$ with explicit potentials



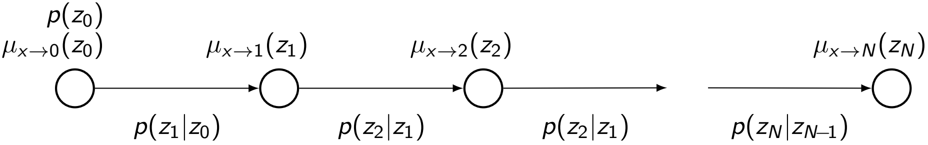
- with $\widetilde{\varphi}_i(x_i) = \varphi_i(x_i)\,\delta(x_i, x_i^{\mathrm{obs}})$

Gibbs model for $p(z \mid \mathbf{x})$ with explicit potentials



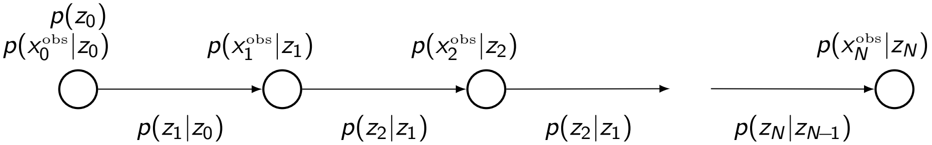
- Marginalize out all the variables x_i
- $\mu_{x \to i}(z_i) = \int_{x_i} \psi_{i,x}(z_i, x_i)\, \widetilde{\varphi}_i(x_i)\, dx_i = \int_{x_i} p(x_i|z_i)\, \delta(x_1, x_1^{\mathrm{obs}})\, dx_i$

Reduced Gibbs model for $p(z|\mathbf{x}^{\mathrm{obs}})$



- With $\mu_{x\to i}(z_i) = p(x_i^{\mathrm{obs}}|z_i)$

Sum-product for the HMM



Messages for the sum-product algorithm

$$\mu_{i-1\to i}(z_i) = \sum_{z_{i-1}} p(z_i|z_{i-1}) p(x_{i-1}^{\mathrm{obs}}|z_{i-1}) \mu_{i-2\to i-1}(z_{i-1})$$

$$\mu_{i+1\to i}(z_i) = \sum_{z_{i+1}} p(z_{i+1}|z_i) p(x_{i+1}^{\mathrm{obs}}|z_{i+1}) \mu_{i+2\to i+1}(z_{i+1})$$

Rewriting the sum-product as the α and β recursions

Messages for the sum-product algorithm

$$\mu_{i-1\to i}(z_i) := \sum_{z_{i-1}} p(z_i|z_{i-1}) p(x_{i-1}^{\mathrm{obs}}|z_{i-1}) \mu_{i-2\to i-1}(z_{i-1})$$

$$\mu_{i+1\to i}(z_i) = \sum_{z_{i+1}} p(z_{i+1}|z_i) p(x_{i+1}^{\mathrm{obs}}|z_{i+1}) \mu_{i+2\to i+1}(z_{i+1})$$

Messages for the sum-product algorithm

$$\alpha_i(z_i) = \mu_{i-1\to i}(z_i) p(x_i^{\mathrm{obs}}|z_i) \quad \text{and} \quad \beta_i(z_i) = \mu_{i+1\to i}(z_i)$$

$$\begin{aligned}
\alpha_i(z_i) &= p(x_i^{\mathrm{obs}}|z_i) \sum_{z_{i-1}} p(z_i|z_{i-1}) \alpha_{i-1}(z_{i-1}) \\
\beta_i(z_i) &= \sum_{z_{i+1}} p(z_{i+1}|z_i) p(x_{i+1}^{\mathrm{obs}}|z_{i+1}) \beta_{i+1}(z_{i+1})
\end{aligned}$$

Properties of α and β messages

$$
\begin{aligned}
\alpha_i(z_i) &= p(x_i^{\mathrm{obs}}|z_i) \sum_{z_{i-1}} p(z_i|z_{i-1}) \alpha_{i-1}(z_{i-1}) \\
\beta_i(z_i) &= \sum_{z_{i+1}} p(z_{i+1}|z_i) p(x_{i+1}^{\mathrm{obs}}|z_{i+1}) \beta_{i+1}(z_{i+1})
\end{aligned}
$$

$$\alpha_i(z_i) = p(x_1, \ldots, x_i, z_i) \qquad \beta_i(z_i) = p(x_{i+1}, \ldots, \mathbf{x}_N|z_i)$$

Finally one gets the marginals:

$$p(z_i|\mathbf{x}) = \frac{\alpha_i(z_i)\beta_i(z_i)}{p(\mathbf{x})}$$

and

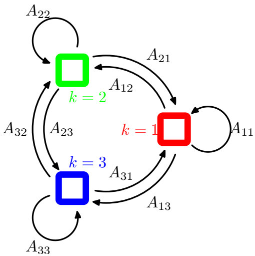
$$p(z_{i-1}, z_i|\mathbf{x}) = \frac{\alpha_{i-1}(x_{i-1}) p(x_i|z_i) p(z_i|z_{i-1}) \beta_i(x_i)}{p(\mathbf{x})}$$

Hidden Markov Model (HMM)

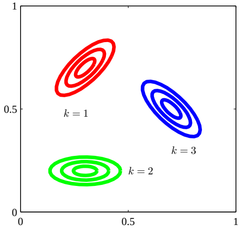
Parameterization

distribution of the initial state	$p(\boldsymbol{z}_1; \pi) = \prod_{\ell=1}^{K} \pi_k^{z_{1\ell}}$
transition matrix	$p(\boldsymbol{z}_i \vert \boldsymbol{z}_{i-1}; A) = \prod_{k=1}^{K} \prod_{\ell=1}^{K} A_{k\ell}^{z_{i-1,j}\, z_{i\ell}}$
emission probabilities	$p(\mathbf{x}_i \vert \boldsymbol{z}_i; \phi)$ e.g. Gaussian Mixture

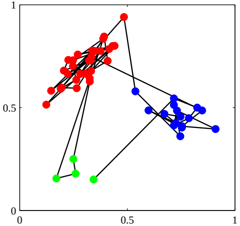
Interpretation



Transistions of $\boldsymbol{z}_i$



$p(\mathbf{x}_i \vert \boldsymbol{z}_i)$



Path of $\mathbf{x}_i$

Maximum likelihood for HMMs

Application of the EM algorithm

$$\gamma(\boldsymbol{z}_i) = p(\boldsymbol{z}_i|\boldsymbol{X}, \boldsymbol{\theta}^t) \qquad \xi(\boldsymbol{z}_{i-1}, \boldsymbol{z}_i) = p(\boldsymbol{z}_{i-1}, \boldsymbol{z}_i|\boldsymbol{X}, \boldsymbol{\theta}^t)$$

Expected value of the log-likelihood:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^t) = \sum_{\ell=1}^{K} \gamma(z_{1\ell}) \log \pi_k + \sum_{i=2}^{N} \sum_{k=1}^{K} \sum_{\ell=1}^{K} \xi(z_{i-1,k}, z_{i\ell}) \log A_{k\ell} + \sum_{i=1}^{N} \sum_{\ell=1}^{K} \gamma(z_{i\ell}) \log p(x_i|\phi_\ell)$$

Maximizing with respect to the parameters $\{\pi, A\}$, one gets

$$\boxed{\pi_k^{t+1} = \frac{\gamma(z_{1k})}{\sum_{k'=1}^{K} \gamma(z_{1k'})}} \qquad \boxed{A_{k\ell}^{t+1} = \frac{\sum_{i=2}^{N} \xi(z_{i-1,k}, z_{i\ell})}{\sum_{\ell'=1}^{K} \sum_{i=2}^{N} \xi(z_{i-1,k}, z_{n\ell'})}}$$

If the emissions are Gaussian then we also have:

$$\boldsymbol{\mu}_k^{t+1} = \frac{\sum_{i=1}^{N} \gamma(z_{ik}) \mathbf{x}_i}{\sum_{i=1}^{N} \gamma(z_{ik})} \qquad \boldsymbol{\Sigma}_k^{t+1} = \frac{\sum_{i=1}^{N} \gamma(z_{ik}) (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top}{\sum_{i=1}^{N} \gamma(z_{ik})}$$

Maximum likelihood for HMMs

Baum-Welch algorithm

The Baum-Welch algorithm is a special instance of the sum-product algorithm. It is also known under the name *forward-backward*.

One propagates the messages

- forward $\alpha(\boldsymbol{z}_i) = p(\mathbf{x}_i|\boldsymbol{z}_i) \sum_{\boldsymbol{z}_{i-1}} \alpha(\boldsymbol{z}_{i-1}) p(\boldsymbol{z}_i|\boldsymbol{z}_{i-1})$
- backward $\beta(\boldsymbol{z}_i) = \sum_{\boldsymbol{z}_{i+1}} \beta(\boldsymbol{z}_{i+1}) p(\mathbf{x}_{i+1}|\boldsymbol{z}_{i+1}) p(\boldsymbol{z}_{i+1}|\boldsymbol{z}_i)$

that satisfy the following properties:

$$\alpha(\boldsymbol{z}_i) = p(\mathbf{x}_1, \ldots, \mathbf{x}_i, \boldsymbol{z}_i) \qquad \beta(\boldsymbol{z}_i) = p(\mathbf{x}_{i+1}, \ldots, \mathbf{x}_N|\boldsymbol{z}_i)$$

Finally one gets the marginals:

$$\gamma(\boldsymbol{z}_n) = p(\boldsymbol{z}_i|\boldsymbol{X}, \boldsymbol{\theta}^t) = \frac{\alpha(\boldsymbol{z}_i)\beta(\boldsymbol{z}_i)}{p(\boldsymbol{X}|\boldsymbol{\theta}^t)}$$

et

$$\xi(\boldsymbol{z}_{i-1}, \boldsymbol{z}_i) = \frac{\alpha(\mathbf{x}_{i-1}) p(\mathbf{x}_i|\boldsymbol{z}_i) p(\boldsymbol{z}_i|\boldsymbol{z}_{i-1}) \beta(\mathbf{x}_i)}{p(\boldsymbol{X}|\boldsymbol{\theta}^t)}$$