today: EM for HMM info theory

\[ \beta\text{-recursion (smoothing)} \]

\[ p(z_t, x_{1:T}) = \frac{1}{\beta_t(z_t)} \sum_{z_{t+1}} \beta_{t+1}(z_{t+1}) p(z_{t+1} | z_t) p(x_{t+1} | z_{t+1}) \]

\[ m_{Z_{t+1} \rightarrow Z_t}(z_t) = \frac{\beta_t(z_t)}{\beta_t(z_t)} \]

\[ \beta_t(z_t) = \frac{1}{\beta_t(z_t)} \sum_{z_{t+1}} p(z_{t+1} | z_t) p(x_{t+1} | z_{t+1}) \beta_{t+1}(z_{t+1}) \]

\[ \beta_{t+1}(z_{t+1}) = \frac{1}{\beta_{t+1}(z_{t+1})} \sum_{z_t} p(z_t | z_{t+1}) p(x_{t+1} | z_t) \beta_t(z_t) \]

\[ \text{initialization:} \quad \beta_1(z_1) = 1 \]

\[ \text{edge marginal} \]

\[ p(z_T, x_{1:T}) = \frac{1}{\beta_T(z_T)} \sum_{z_{T-1}} \beta_{T-1}(z_{T-1}) p(z_{T-1} | z_T) p(x_{T-1} | z_{T-1}) \]

\[ p(z_T, x_{1:T}) = \alpha_T(z_T) \beta_T(z_T) \]

\[ \text{numerical stability trick:} \]

\[ \text{issue:} \quad \alpha_t \text{ and } \beta_t \text{ can easily go to } 1e-100 \]

\[ \text{two possibilities:} \]

\[ \begin{align*}
  & a) \ (\text{general}) \text{ store } \log(\alpha_t) \text{ instead} \\
  & \log \left( \frac{\alpha_t}{a_t} \right) = \log \left( \frac{\alpha_t}{a_t} \right) \\
  & (a_t, \gamma_t) 
\end{align*} \]
b) normalize the message

- $\alpha$-recursion: use $\hat{\alpha}_t(z_t) \triangleq p(z_t | x_{1:t})$

before $\alpha_t = \alpha_t \odot A \hat{\alpha}_{t-1}$

\[
\alpha_t = \alpha_t \odot A \hat{\alpha}_{t-1}
\]

you can show that $c_t = \frac{1}{Z_t} (\alpha_t \odot A \hat{\alpha}_{t-1})(z_t) = p(z_t | x_{1:t})$

\[
p(x_{1:t} | z_t) = \frac{1}{Z_t} p(x_t | x_{1:t-1}) c_t
\]

- $\beta$-recursion:

define $\hat{\beta}_t(z_t) = \frac{p(x_{t+1} | z_t)}{p(x_{t+1} | x_{1:t})} \frac{T}{c_{t+1}}$

\[
\hat{\beta}_t(z_t) = \frac{p(x_{t+1} | z_t)}{p(x_{t+1} | x_{1:t})} \frac{T}{c_{t+1}}
\]

note: $\hat{\beta}_k(z_k) \neq 1$

exercises: derive $\hat{\beta}$-recursion

15h50

**ML for HMM**

- suppose $p(x_t | z_t = k) = f(x_t | \theta_k)$

\[
p(x_t | z_t = k) = f(x_t | \theta_k)
\]

- $p(z_{t+1} = i | z_t = j) = \pi_{ij}$

\[
p(z_{t+1} = i | z_t = j) = \pi_{ij}
\]

- $p(z_1 = i) = \pi_i$

\[
p(z_1 = i) = \pi_i
\]

want to estimate $\hat{\theta}$, $\hat{A}$, $\hat{\pi}$ by ML from data $x = (x^{(i)})_{i=1}^N$

\[
x = (x^{(i)})_{i=1}^N
\]

\[
x^{(i)} = x^{(i)}_{1:T}
\]

\[
\text{use EM at 5th lecture}
\]

\[
\begin{align*}
\text{E step:} & \quad Q_{st} (z) = p(z | x, \Theta^{(s)}) \\
\text{M step:} & \quad \Theta^{(s+1)} = \arg \max \quad \mathbb{E}_{q_{st}} [ \log p(x | z, \Theta) ]
\end{align*}
\]

complete log-likelihood

\[
\log \mathcal{L}(x, z | \Theta) = \sum_{i=1}^N \left\{ \log \alpha(z^{(i)}_{1:T}) + \sum_{t=1}^{T_i} \log p(x_t | z_t^{(i)}) + \sum_{t=1}^{T_i} \log p(z_{t+1}^{(i)} | z_t^{(i)}) \right\}
\]
Information theory

**K-L divergence** for discrete dist. $p, q$

$$K-L(p \| q) = \sum_{x \in \Omega} p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_p [\log \frac{p(x)}{q(x)}]$$

- $\lim_{x \to 0^+} x \log x = 0$ (for $0 < x < 1$) but $q(x) > 0$
- $\lim_{x \to 0^-} x \log x = 0$
- $\lim_{x \to \infty} \frac{-p(x) \log q(x)}{q(x)} = +\infty$

**Example**

`\begin{align*}
\hat{\omega}_k &= \frac{\sum_{\beta=2}^{T_k} \gamma_1^{(i)} \delta_{e}\delta_{u,m}^{(i)}}{\sum_{\beta=2}^{T_k} \gamma_1^{(i)} \delta_{e}\delta_{u,m}^{(i)}} \\
\hat{\omega}_k &= \frac{\sum_{\beta=2}^{T_k} \gamma_1^{(i)} \delta_{e}\delta_{u,m}^{(i)}}{\sum_{\beta=2}^{T_k} \gamma_1^{(i)} \delta_{e}\delta_{u,m}^{(i)}}
\end{align*}`
\[
\text{is support of } p \subseteq \text{ support of } q \\
\Rightarrow \text{KL}(p \| q) = \text{two}
\]

**Motivation from density estimation**

Recall statistical decision theory

(Statistical) loss \( L(p, \hat{q}) \)

Standard (MLE) loss is log-loss

\[
L(p, \hat{q}) = \mathbb{E}_{x \sim p} [-\log \hat{q}(x)]
\]

If \( \hat{q} = p \), then get

\[
\frac{1}{\text{entropy of } p} \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = H(p)
\]

Excess loss for action \( a = \hat{q} \)

\[
L(p, \hat{q}) - \min_{q} L(p, q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{\hat{q}(x)}{q(x)} = \text{KL}(p \| \hat{q})
\]

**Coding theory**

Use length of code \( -\log p(x) \)

Expected length of code:

\[
\mathbb{E} \left[ -\log p(x) \right]
\]

**KL divergence**

Interpreted as excess length of code in terms of length of code.

*Use dist. \( q \) to design code vs. optimal dist. \( p \).

**Example:**

- Entropy of a Bernoulli

\[
-H(p) = \log(p) + \log(1-p)
\]

- Entropy for a uniform dist. over \( k \) states

\[
\frac{1}{k} \sum_{x=1}^{k} \log \frac{1}{k} = \log k
\]

**Properties of KL**

- \( \text{KL}(p \| q) \geq 0 \)

  (by Jensen’s inequality)

  \( \mathbb{E} f_X \leq f(\mathbb{E} X) \)

  when \( f \) is convex

- \( \text{KL} \) is strictly convex in each of its arguments: \( \text{KL}(p \| q; r) \)
\* not symmetric: \( KL(p\|q) \neq KL(q\|p) \) in general

\[
KL(p\|q) = 0 \Longleftrightarrow \exists x \text{ s.t. } p(x) = q(x)
\]

symmetrized version:

\[
\frac{1}{2} (KL(p\|q) + KL(q\|p))
\]