Today: EM for HMM; info theory

\[ p(Z_t, \bar{X}_{1:T}) = \frac{\alpha_t(Z_t) \cdot m_{2t+1:t+1}(Z_t)}{\beta_t(Z_t)} \]

\[ m_{2t+1:t+1}(Z_t) = \sum_{Z_{t+1}} p(Z_{t+1}|Z_t) p(\bar{X}_{t+1}|Z_{t+1}) m_{2t+1:t}(Z_{t+1}) \]

\[ \beta_t(Z_t) = \sum_{Z_{t+1}} p(Z_{t+1}|Z_t) p(\bar{X}_{t+1}|Z_{t+1}) \beta_{t+1}(Z_{t+1}) \]

\[ \beta_t(Z_t) = p(\bar{X}_{1:T} \mid Z_t) \]

\[ \text{initialization: } \beta_0(Z_0) = 1 \quad \forall Z_0 \]

**Numerical stability trick:**
**numerical stability trick:**

**issue:** $\alpha_t \leq \beta_t$ can easily go to $1 \times 10^{-1000}$

**two possibilities:**

a) **(general) use** $\log(\alpha_t)$ instead

$$
\log (\xi a_t) = \log (\tilde{\alpha}_t (\xi a_t)) \quad (a_t > 0)
$$

call $\tilde{\alpha}_t \triangleq \max_{\alpha_t} = \log \tilde{\alpha} + \log (\xi a_t)

$$
i_{\max} \triangleq \arg \max a_t \quad \log ((\xi \exp(\log(\xi) - \log(\tilde{\alpha})))

$$
\log(\tilde{\alpha}_t) = \log(a_{i_{\max}}) = \max_i \log(a_i)

$$
\log (\xi a_t) = \log(\tilde{\alpha}_t) + \log(1 + \xi \exp(\log(a_t) - \log(\tilde{\alpha}_t)))
$$

b) **normalize the message**

- **N-reduction:** use $\hat{z}_t(z_t) \triangleq p(z_t | x_{1:t})$

  before $\alpha_t = \alpha_{t-1} A \alpha_{t-1}$

  $\hat{z}_t = \frac{\prod_{i=1}^t (\xi z_i^{\hat{a}_{i-1}(z_{i-1})})}{\sum_{z_t} (\alpha_{t-1} A \alpha_{t-1}) (z_t)} \triangleq c_t$

  you can show that $c_t = \frac{\prod_{i=1}^t (\xi z_i^{\hat{a}_{i-1}(z_{i-1})})}{\sum_{z_t} (\alpha_{t-1} A \alpha_{t-1}) (z_t)} = p(x_{1:T} | x_{1:t-1})$

  $\prod_{t=1}^T p(z_t | x_{1:t-1}) = \prod_{t=1}^T c_t$

- **B-reduction:**

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

  $\frac{\prod_{t=1}^T c_t}{\sum_{z_t} (\alpha_{t-1} A \alpha_{t-1}) (z_t)}$ note: $\frac{\prod_{t=1}^T c_t}{\sum_{z_t} (\alpha_{t-1} A \alpha_{t-1}) (z_t)} \neq 1$

  **converse:** define $\hat{\beta}$-reduction

---

**ML for HMM**

(for some parametric model for $x_{1:T}$)
Suppose \( p(x_t | z_t = k) = f(x_t | \pi_k) \) for some parametric model for \( z_t \).

\[ p(z_{t+1} = k | z_t = k) = \Theta_k^k \]

\[ p(z_t = k) = \pi_k \]

\[ \Theta = (\pi, A, \pi) \]

Want to estimate \( \hat{\Theta} \) by ML from data \( x = (x(i))^N_{i=1} \)

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

Sure EM at \( s \)th iteration

**E step:** \( q_{s+1}(z) = p(z | x, \Theta_s) \)

**M step:** \( \Theta_{s+1} = \text{argmax} \ E_{q_{s+1}(z)} \left[ \log p(x, z | \Theta) \right] \)

Complete log-likelihood

\[ \log p(x, z | \Theta) = \sum_{i=1}^{N} \left[ \log p(z^{(i)} | \pi) + \sum_{t=1}^{T_t} \log p(z_t^{(i)} | z_{t-1}^{(i)}) + \sum_{t=2}^{T_t} \log p(z_t^{(i)} | z_{t-1}^{(i)}, x_{1:T_t}) \right] \]

**E step:** \( q_{s+1}(z) = p(z | x, \Theta_s) \)

**M step:** \( \Theta_{s+1} = \text{argmax} \ E_{q_{s+1}(z)} \left[ \log p(x, z | \Theta) \right] \)

Smoothing auxillary variable

\[ q_{s+1}(z_{t:k}) = q_{s+1}(z_{t:k} | z_t^{(i)}, x_{1:T_t}) = \left[ z_{t:k} | z_t^{(i)}, x_{1:T_t} \right] \]

Smoothing edge marginal prob in HMM

\[ A_{km}^{(s)} = \sum_{i=1}^{N} \gamma_{t+1, k}^{(i)} \]

\[ \hat{\Theta} \]

Max. with respect to \( \Theta \):
Information theory

KL divergence: for discrete dist. $p$ vs $q$

$$KL(p || q) = \mathbb{E}_p \left[ \log \frac{p(x)}{q(x)} \right] = \mathbb{E}_p \left[ \log p(x) - \log q(x) \right]$$

- $\log 0 = 0$
- $\lim_{x \to 0^+} x \log x = 0$
- $\lim_{x \to \infty} x \log x = \infty$
- $\lim_{x \to 0^-} x \log x = -\infty$
- $\log x$ is log-convex

Motivation from density estimation
Recall statistical decision theory: the goal is to estimate the distribution $p(x)$.

Standard (MLE) loss is log-convex: $L(p_0, \hat{p}) = \mathbb{E}_x \left[ -\log \hat{p}(x) \right]$.

If we choose $\hat{p} = p_0$, then get

$$\mathbb{E}_x \left[ -p_0(x) \log p_0(x) \right] \equiv H(p_0)$$

Cross-entropy
excess loss for action \( a = q \)

\[
L(p, q) - \min_{q'} L(p, q') = \sum_{x \in \Omega} -q(x) \log \frac{q(x)}{p(x)} = KL(p || q)
\]

**Coding Theory:**

- Use length of code \( \sum -\log p(x) \)
- \( \log_e \rightarrow \text{"nats"} \)

Expected length of code: \( \sum -\log p(x) \) (entropy measured in nats)

**KL Divergence:** interpreted as excess length cost (in terms of length of code) to use dist \( q \) to design code vs optimal dist (max p)

**Example:**

- **Entropy of a Bernoulli**
  \[
  H(p) = -p \log p - (1-p) \log (1-p)
  \]

- **Entropy of a uniform dist. on \( k \) states**
  \[
  \sum_{x=1}^{k} -\frac{1}{k} \log \left( \frac{1}{k} \right) = \log k
  \]

  (max entropy dist. over \( k \) states)

**Properties of KL**

- \( KL(p || q) \geq 0 \) & to show this, use Jensen's inequal., \( f(\mathbb{E} x) \leq \mathbb{E} f(x) \) when \( f \) is convex

- **KL is strictly concave in each of its arguments**
  \( \Delta_{KL} : \mathbb{R}^k \rightarrow \mathbb{R} \)
  strictly convex

- **Not symmetric**
  \( KL(p || q) \neq KL(q || p) \) in general
  \( KL(p || q) = 0 \iff p = q \)
in general: \( \frac{1}{2}(KL_{pq} + KL_{qp}) \)

symmetrized version