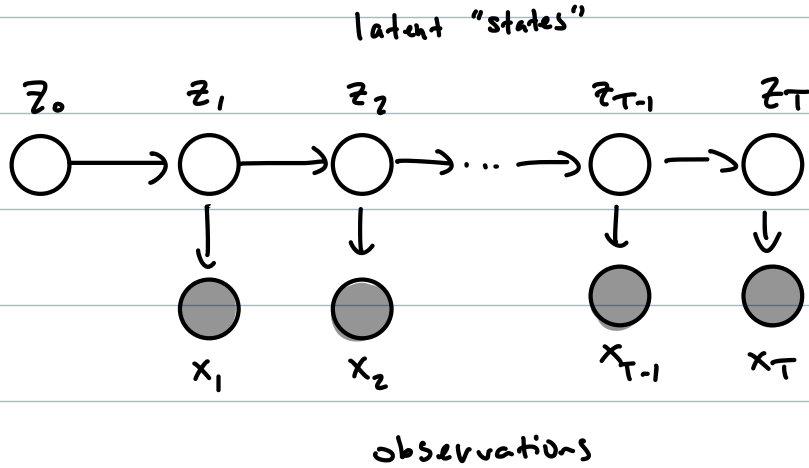
State-space models (SSMs) (very general)



$z_0 \sim p(z_0)$

for $t = 1 \ldots T$:

$\quad z_t \sim p(z_t \mid z_{t-1})$

$\quad x_t \sim p(x_t \mid z_t)$

- $p_\phi(x_t \mid z_t)$ "emission" distribution (likelihood)
  with fixed parameters ϕ
- $p_\Lambda(z_t \mid z_{t-1})$ "transition" distribution
  with fixed parameters Λ
- $p_\pi(z_0)$ "initial" distribution
  with fixed parameters π

Hidden Markov Models (HMMs) are SSMs

- $z_t \in [K] \equiv \{1 \ldots K\}$ discrete states
- $p_\Lambda(z_t = k \mid z_{t-1} = j) = \Lambda(j,k)$
  "transition matrix"
- $\sum_{k=1}^{K} \Lambda(j,k) = 1$ "row-stochastic" matrix
- $p_\pi(z_0 = k) = \pi_0(k)$ "initial dist"
- $x_t \in \mathbb{R}^d$ (no constraints on emission dist)

Examples...

- speech recognition

$x_t \in \mathbb{R}^d$ acoustic signals

$z_t \in$ WORDS word token

$p_\lambda(z_t \mid z_{t-1})$ language model

$p_\phi(x_t \mid z_t)$ acoustic model

- activity recognition

$x_t \in \mathbb{R}^d$ video frame of subject

$z_t \in$ ACTIVITIES activity of subject

e.g. is the Zebrafish "hunting", "sleeping", ... at each $t = 1 \ldots T$

e.g. is the enemy aircraft "attacking", "escorting", etc. at t

- gene finding :

t = locus of nucleotide

$x_t \in \{A, C, G, T\}$

$z_t \in$ REGIONS, e.g. "gene-coding" region

Types of inference in SSMs

filtering: $p(z_t \mid x_{1:t})$ "online"

prediction: $p(z_{t+1} \mid x_{1:t})$

smoothing: $p(z_t \mid x_{1:T})$ "offline"

forecasting: $p(x_{t+1} \mid x_{1:t})$

imputing: $p(x_t \mid x_{1:T \setminus t})$
(or **denoising**, or **anomaly detection**)

evidence: $p(x_{1:T})$

FORWARD pass

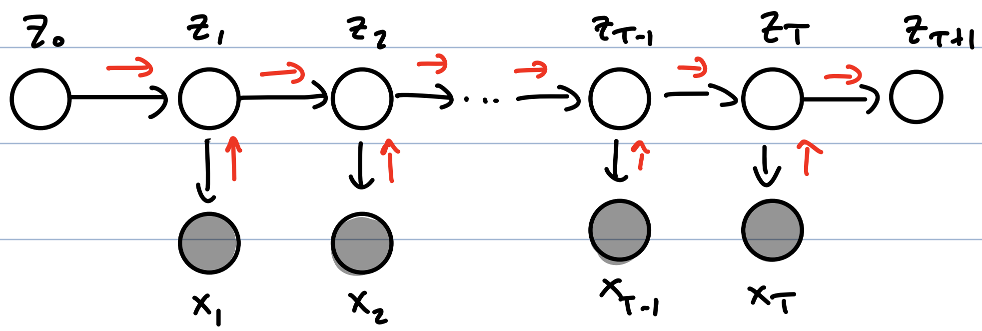
- Goal : $P(x_{T+1} \mid x_{1:T})$ "forecasting"

- Requires: $P(z_{T+1} \mid x_{1:T})$ "predictive"

  target of inference

- HMMs are trees; variable elimination

  - Set z_{T+1} as root of new tree
  - point edges away
  - ORDER by depth-first : $\{x_1, \dots x_T, z_1 \dots z_{T+1}\}$



- Messages from observations $x_1 \dots x_T$

$$m_{x_t \to z_t}(z_t) = \sum_{x_t} \psi(x_t)\, \psi(x_t, z_t)$$

$$= \sum_{x_t} \delta_{\bar{x}_t}(x_t)\, P_\phi(x_t \mid z_t)$$

$$= P_\phi(\bar{x}_t \mid z_t)$$

- Vector notation

$$\ell_t = \begin{bmatrix} P_\phi(\bar{x}_t \mid z_t = 1) \\ \vdots \\ P_\phi(\bar{x}_t \mid z_t = K) \end{bmatrix} \equiv m_{x_t \to z_t}$$

Messages between states

- $\underset{z_0 \to z_1}{m}(z_1) = \sum_{z_0} \varphi(z_0)\, \ell(z_0, z_1)$ "root"

$$= \sum_{z_0} p(z_0)\, p(z_1 \mid z_0) \equiv p(z_1)$$

$$= \sum_{z_0} \pi_0(z_0)\, \Lambda(z_0, z_1) \equiv \pi_0^T \Lambda(:, z_1)$$

- $\underset{z_1 \to z_2}{m}(z_2) = \sum_{z_1} \varphi(z_1)\, \ell(z_1, z_2)\, \underset{z_0 \to z_1}{m}(z_1)\, \underset{x_1 \to z_1}{m}(z_1)$

$$= \sum_{z_1} p(z_2 \mid z_1)\, p(z_1)\, p(\bar{x}_1 \mid z_1) \equiv p(z_2, \bar{x}_1)$$

$$= \sum_{z_1} \Lambda(z_1, z_2)\, \underset{z_0 \to z_1}{m}(z_1)\, \ell_1(z_1)$$

All future messages follow the same recursion:

- Define $\alpha_t(z_t) \triangleq \underset{z_{t-1} \to z_t}{m}(z_t)$ e.g., "$\alpha_2(z_2)$"

- $\alpha_t(z_t) = \sum_{z_{t-1}} \Lambda(z_{t-1}, z_t)\, \alpha_{t-1}(z_{t-1})\, \ell_{t-1}(z_{t-1})$

$$\boxed{\begin{aligned} \alpha_t &= \Lambda^T(\alpha_{t-1} \odot \ell_{t-1}), && t = 2 \dots T+1 \\ \alpha_1 &= \pi^T \Lambda \quad (\text{"base case"}) && t = 1 \end{aligned}}$$

- Each message computes:

$$\alpha_{t+1}(z_{t+1}) = P(z_{t+1}, \bar{x}_{1:t})$$

- e.g. $\alpha_3(z_3) = \sum_{z_2} \Lambda(z_2, z_3)\, \alpha_2(z_2)\, l_2(z_2)$

$$= \sum_{z_2} P(z_3 \mid z_2)\, P(z_2, \bar{x}_1)\, P(\bar{x}_2 \mid z_2)$$

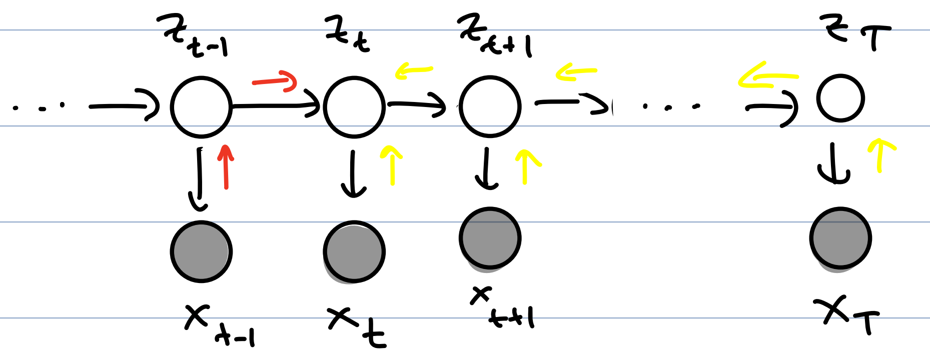
$$= P(z_3, \bar{x}_{1:2})$$

- last message $\alpha_{T+1}(z_{T+1}) = P(z_{T+1}, \bar{x}_{1:T})$
- $P(z_{T+1} = k \mid \bar{x}_{1:T}) = \dfrac{\alpha_{T+1}(k)}{\sum_j \alpha_{T+1}(j)}$
- $P(\bar{x}_{1:T}) = \sum_{j=1}^{K} \alpha_{T+1}(j)$
- What about $P(z_t \mid \bar{x}_{1:T})$?
- Can we compute it for $t < T+1$ using only the forward messages?
- No. Why not? z_t needs "backward" info from $\bar{x}_{t+1:T}$.

BACKWARD pass :

Goal : $P(z_t \mid \bar{x}_{1:T})$

- Set z_t as root of new tree
- point edges away
- ORDER by depth-first : $\{x_1 \dots x_T, z_1 \dots z_{T+1}\}$



- $$m_{z_T \to z_{T-1}}(z_{T-1}) = \sum_{z_T} \varphi(z_T)\,\varphi(z_{T-1}, z_T)\, m_{x_T \to z_T}(z_T)$$

$$= \sum_{z_T} P(z_T \mid z_{T-1})\, P(\bar{x}_T \mid z_T) \equiv P(\bar{x}_T \mid z_{T-1})$$

- $$\beta_t(z_t) \hat{=} m_{z_{t-1} \to z_t}(z_t) \quad , \quad \text{e.g. } \beta_{T-1}(z_{T-1})$$

- $$\beta_{T-2}(z_{T-2}) = \sum_{z_{T-1}} \varphi(z_{T-1})\,\varphi(z_{T-2}, z_{T-1})\, m_{x_{T-1} \to z_{T-1}}(z_{T-1})\, \beta_{T-1}(z_{T-1})$$

$$= \sum_{z_{T-1}} P(z_{T-1} \mid z_{T-2})\, P(\bar{x}_{T-1} \mid z_{T-1})\, \beta_{T-1}(z_{T-1})$$

$$= P(\bar{x}_{T-1:T} \mid z_{T-2})$$

- $\beta_t(z_t) = \sum_{z_{t+1}} \Lambda(z_t, z_{t+1})\, \ell_{t+1}(z_{t+1})\, \beta_{t+1}(z_{t+1})$

$$= P(\bar{x}_{t+1:T} \mid z_t)$$

$$\boxed{\begin{aligned} &\beta_T = \underline{1}_K \quad \text{"base case"} \quad t = T \\ &\beta_t = \Lambda(\ell_{t+1} \odot \beta_{t+1}) \quad \text{for } t = T-1 \ldots 0 \end{aligned}}$$

Putting it all together...

The 3 messages sent to z_t:

1. $\alpha_t(z_t) = P(z_t, \bar{x}_{1:t-1})$
2. $\beta_t(z_t) = P(\bar{x}_{t+1:T} \mid z_t)$
3. $\ell_t(z_t) = P(\bar{x}_t \mid z_t)$

$$\alpha_t(z_t)\, \beta_t(z_t)\, \ell_t(z_t) = P(z_t, \bar{x}_{1:T})$$

$$P(Z_t = k \mid \bar{x}_{1:T}) = \frac{\alpha_t(k)\, \beta_t(k)\, \ell_t(k)}{\sum_j \alpha_t(j)\, \beta_t(j)\, \ell_t(j)}$$

- The "Forwards - Backwards" algo (1986):

$$\alpha_t = \dots \quad \text{for } t = 1 \dots T+1$$

$$\beta_t = \dots \quad \text{for } t = T \dots 0$$

- Invented in 1980s; special case of belief prop.
- Estimates all singleton posterior marginals exactly

$$\underbrace{P(z_t \mid \bar{x}_{1:T})}_{\text{"beliefs"}} \quad t = 1 \dots T$$

- $O(K^2)$

What about $P(z_1 \dots z_T \mid \bar{x}_{1:T})$?

- Many applications motivate inference of sequences
- Instantiating entire joint distribution is $O(K^T)$
- Working with individual sequences is tractable:

① Evaluating probability of a sequence

$$P(z_1 \dots z_n \mid \bar{x}_{1:T}) = \frac{\prod_{t=1}^{T} p(z_t \mid z_{t-1})\, p(\bar{x}_t \mid z_t)}{p(\bar{x}_{1:T})}$$

② sampling a sequence:

$$z_1 \dots z_n \sim p(z_1 \dots z_n \mid \bar{x}_{1:T})$$

③ maximizing

$$z^*_{1:T} = \underset{z_{1:T}}{\operatorname{argmax}}\; p(z_{1:T} \mid \bar{x}_{1:T})$$

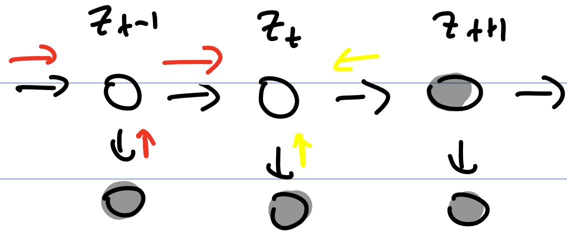
Forward-filtering backward sampling (FFBS)

① Forward pass $\alpha_t = \cdots$

② $z_T \sim p(z_T \mid \bar{x}_{1:T})$

③ $z_t \sim p(z_t \mid \bar{z}_{t+1}, \bar{x}_{1:T}) \quad t = T-1 \ldots 0$

$$p(z_t \mid \bar{z}_{t+1}, \bar{x}_{1:T}) \propto \alpha_t(z_t)\, \ell_t(z_t)\, p(\bar{z}_{t+1} \mid z_t)$$



MAP

$$\max_z p(z \mid \bar{x}) = \max_{z_1} p(z_1 \mid \bar{x}) \cdots \max_{z_T} p(z_T \mid z_{T-1}, \bar{x})$$

$$\tilde{\beta}_T(z_T) = 1$$

$$\tilde{\beta}_t(z_t) = \max_{z_{t+1}} p(z_{t+1} \mid z_t)\, p(\bar{x}_{t+1} \mid z_{t+1})\, \tilde{\beta}_{t+1}(z_{t+1}) \quad t = T-1 \ldots 0$$

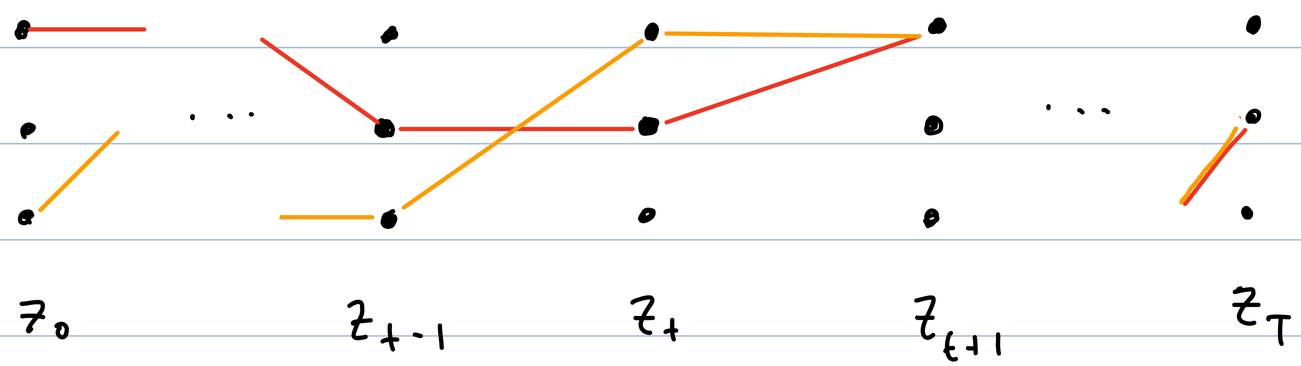
$$\tilde{\beta}^*_t(z_t) = \arg\max_{z_{t+1}} \quad \text{"} \quad \text{"} \quad t = T-1 \ldots 0$$

$$z_0^* = \arg\max_{z_0} \pi_0(z_0)\, \tilde{\beta}_0(z_0)$$

$$z_t^* = \tilde{\beta}^*_{t-1}(z^*_{t-1}) \quad \text{for } t = 1 \ldots T$$

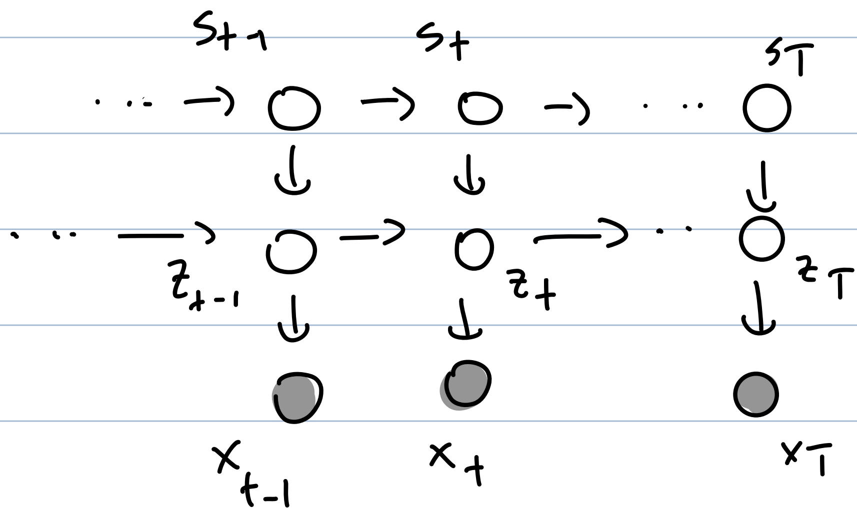
This is the Viterbi algorithm (1967) (special case of max-prod).

Often visualized with a lattice:



"Switching" HMMs

- $P(z_t = k \mid z_{t-1} = j, s_t = s) = A^s(j,k)$
- $P(s_t = s \mid s_{t-1} = r) = \Delta(r,s)$



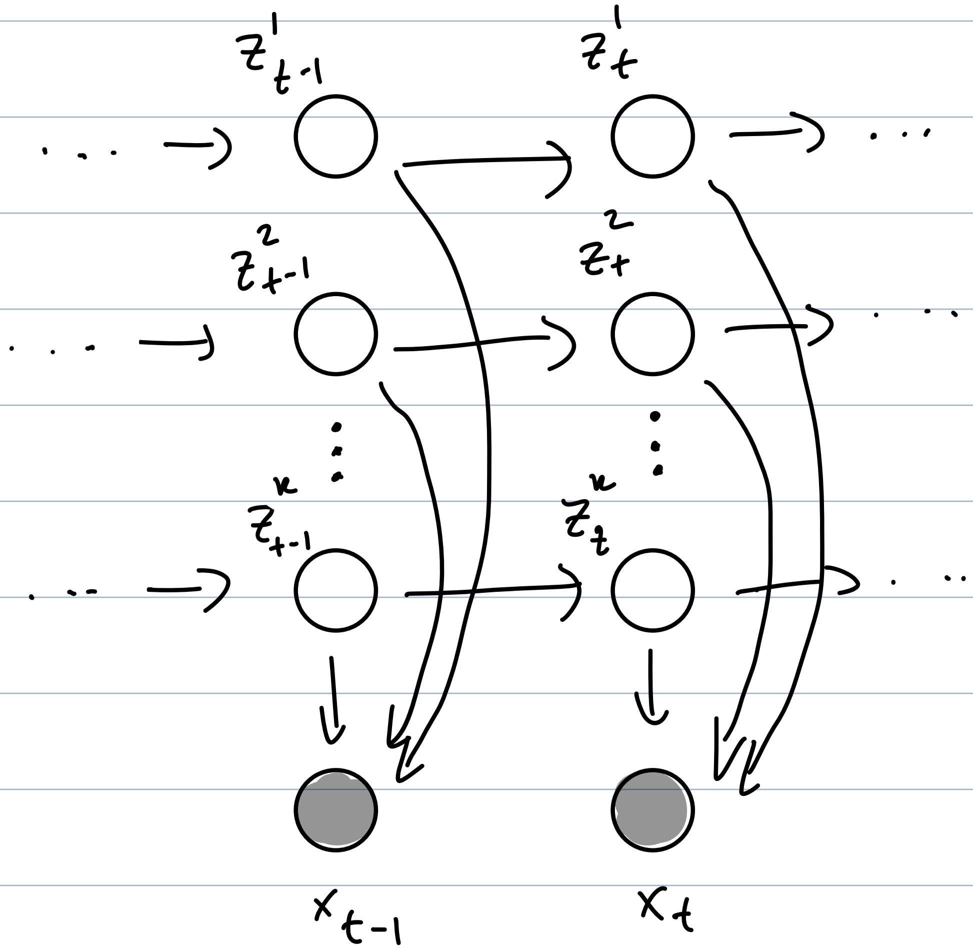
No longer a tree. How do we do inference?

$$\ldots \sum_{s_T} P(s_T \mid s_{T-1}) \sum_{z_T} P(z_T \mid z_{T-1}, s_T)\, P(\bar{x}_T \mid z_T)$$

$$= \ldots \sum_{s_T, z_T} P(s_T, z_T \mid s_{T-1}, z_{T-1})\, P(\bar{x}_T \mid s_T, z_T)$$

Redefine (s_t, z_t) as a "super node"

"Factorial" HMMs



$$\vec{z}_t \equiv \left[z^1_t \cdots z^k_t \right]^T, \quad z^k_t \in [p]$$

"distributed representation"

- e.g. to represent 30 bits of info from the history $\bar{x}_{1:T}$, a vanilla HMM would require $z_{T+1} \in [2^{30}]$, whereas a factorial HMM could do this with $p = 2$, $k = 30$.

- Connection to general state-space models

Learning in HMMs

- We have only thus far considered <u>inference</u> in graphical models (e.g. $p(x_F \mid \bar{x}_E)$) assuming that the parameters are <u>known</u>

- Parameters in an HMM:

$$\Theta = \{ \Lambda, \pi_0, \Phi \}$$

"transition" "initial" "emission"

- Type-II MLE would be:

"evidence"

$$\hat{\Theta} \leftarrow \underset{\Theta}{\arg\max}\ p(\bar{x}_{1:T}; \Theta)$$

- For a given Θ, we can compute the evidence using just the forward pass

- One option is gradient ascent ...

- Another: <u>EM</u> (next class)

HMM with Poisson observations

$$P(x_t \mid z_t = k) = \text{Pois}(\mu_k)$$

Goal: $\hat{\mu} \leftarrow \underset{\mu}{\text{argmax}}\; P(x_{1:T} ; \mu)$

This is easier WRT the complete data likelihood:

$$P(x_{1:T}, z_{1:T} ; \mu) \propto_\mu \prod_t \text{Pois}(x_t ; \mu_{z_t})$$

Define $f_{tk} \doteq \mathbb{1}(z_t = k)$

$$\rightarrow \quad \propto_\mu \prod_t \prod_k \text{Pois}(x_t ; \mu_k)^{f_{tk}}$$

$$\frac{\partial}{\partial \mu_k} \log \prod_t \prod_k \text{Pois}(x_t ; \mu_k)^{f_{tk}}$$

$$\propto_{\mu_k} \frac{\partial}{\partial \mu_k} \sum_t \sum_k f_{tk} \left[x_t \log \mu_k - \mu_k \right]$$

$$\propto_{\mu_k} \sum_t f_{tk} x_t \frac{1}{\mu_k} - \sum_t f_{tk}$$

$$= \frac{\sum_t f_{tk} x_t}{\mu_k} - N_{tk} = 0 \rightarrow \hat{\mu}_k = \frac{1}{N_{tk}} \sum_t f_{tk} x_t$$

Not surprising: $\hat{\mu}_k = \hat{\mathbb{E}}[x_t \mid z_t = k]$

$P(x, z; \mu)$ is easily optimizeable.

We don't observe $z_{1:T}$. <u>Instead</u>:

$$\log p_\theta(x) = \log \sum_z p_\theta(x, z)$$

$$= \log \sum_z \frac{Q(z)}{Q(z)} p_\theta(x, z)$$

any surrogate dist.

$$= \log \mathbb{E}_Q \left[\frac{p_\theta(x, z)}{Q(z)} \right]$$

$$\geq \mathbb{E}_Q \left[\log \frac{p_\theta(x, z)}{Q(z)} \right] \qquad \text{(Jensen)}$$

$$\triangleq B(Q, \Theta) \qquad \text{(evidence lower bound)}$$

"ELBO"

$$B(Q, \Theta) \propto_{\mu_k} \mathbb{E}_Q \left[\log p(x_{1:T} \mid z_{1:T}; \mu) \right]$$

$$\propto_{\mu_k} \mathbb{E}_Q \left[\sum_t \sum_k \delta_{tk} \left[x_t \log \mu_k - \mu_k \right] \right]$$

$$\propto_{\mu_k} \sum_t \sum_k \mathbb{E}_Q [\delta_{tk}] \left[x_t \log \mu_k - \mu_k \right]$$

$$\frac{\partial}{\partial \mu_k} \cdots = 0 \longrightarrow \hat{\mu}_k = \frac{\sum_t \mathbb{E}_Q[\delta_{tk}] x_t}{\sum_t \mathbb{E}_Q[\delta_{tk}]}$$

So: if we have $Q(z)$, we compute the beliefs

$$\mathbb{E}_Q[y_{tk}] = \mathbb{E}_Q[\mathbb{1}(z_t = k)] = Q(z_t = k),$$

and then maximize ELBO WRT μ_k.

Where do we get $Q(z)$?

$$\hat{Q}(z) \leftarrow \underset{Q(z)}{\operatorname{argmax}}\ B(Q, \Theta)$$

$$B(Q, \Theta) = \mathbb{E}_Q\left[\log \frac{p(x, z)}{Q(z)}\right]$$

$$\underset{Q}{\propto} \qquad\qquad\qquad\qquad - \log p(x)$$

$$= -\mathrm{KL}\left(Q(z) \,\|\, p(z \mid x) \right)$$

Therefore:

$$\underset{Q(z)}{\operatorname{argmax}}\ B(Q, \Theta) = p_\theta(z \mid x)$$

(the exact posterior)

Can we set $Q(z_{1:T}) = p(z_{1:T} \mid x_{1:T})$?
(tractably)

No. But, do we need the whole distribution?

No! Only the beliefs, which we can compute with forwards-backwards.