- Recap:
  - HMMs with known params $\Theta = \{\pi_0, \Lambda, \Phi\}$ are trees.
  - (Draw model from last time)
  - Belief propagation therefore allows us to compute all singleton marginals $p(z_t \mid \bar{x}_{1:T})$ AKA "beliefs".
  - BP in HMMs $\equiv$ the "forwards-backwards" algo.
  - To learn parameters, we need inference.
  - EM: iterates between updating params, doing inference.

Expectation - Maximization (EM) (1977)

- complete likelihood $P_\theta(x, z)$
- marginal likelihood $P_\theta(x)$ aka "evidence"
- Goal: Type-II MLE

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ P_\theta(x)$$

- Setting up the evidence lower bound (ELBO)

$$\log P_\theta(x) = \log \mathbb{E}_{z \sim Q(z)} \left[\frac{P(x,z)}{Q(z)} \right]$$

trivially true for any density $Q(z)$ s.t.

$Q(z) > 0$ if $P(z \mid x) > 0$.

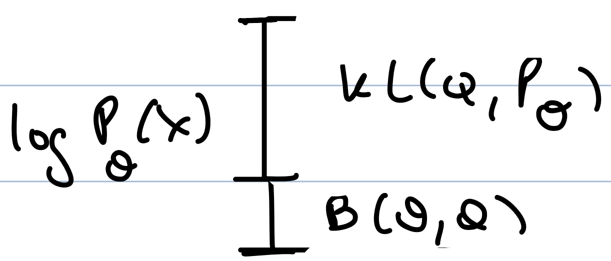
$$\geq \mathbb{E}_Q \left[\log \frac{P_\theta(x,z)}{Q(z)} \right] \triangleq B(Q, \theta)$$

(ELBO)

$$= \mathbb{E}_Q [\log P(x,z)] - \underbrace{\mathbb{E}_Q [Q(z)]}_{= H(Q(z))}$$

$$= \underbrace{\mathbb{E}_Q [\log P(x)]}_{= P(x)} + \underbrace{\mathbb{E}_Q \left[\log \frac{P(z \mid x)}{Q(z)} \right]}_{= -KL(Q(z) \,\|\, P(z \mid x))}$$

Picture:



The EM algorithm

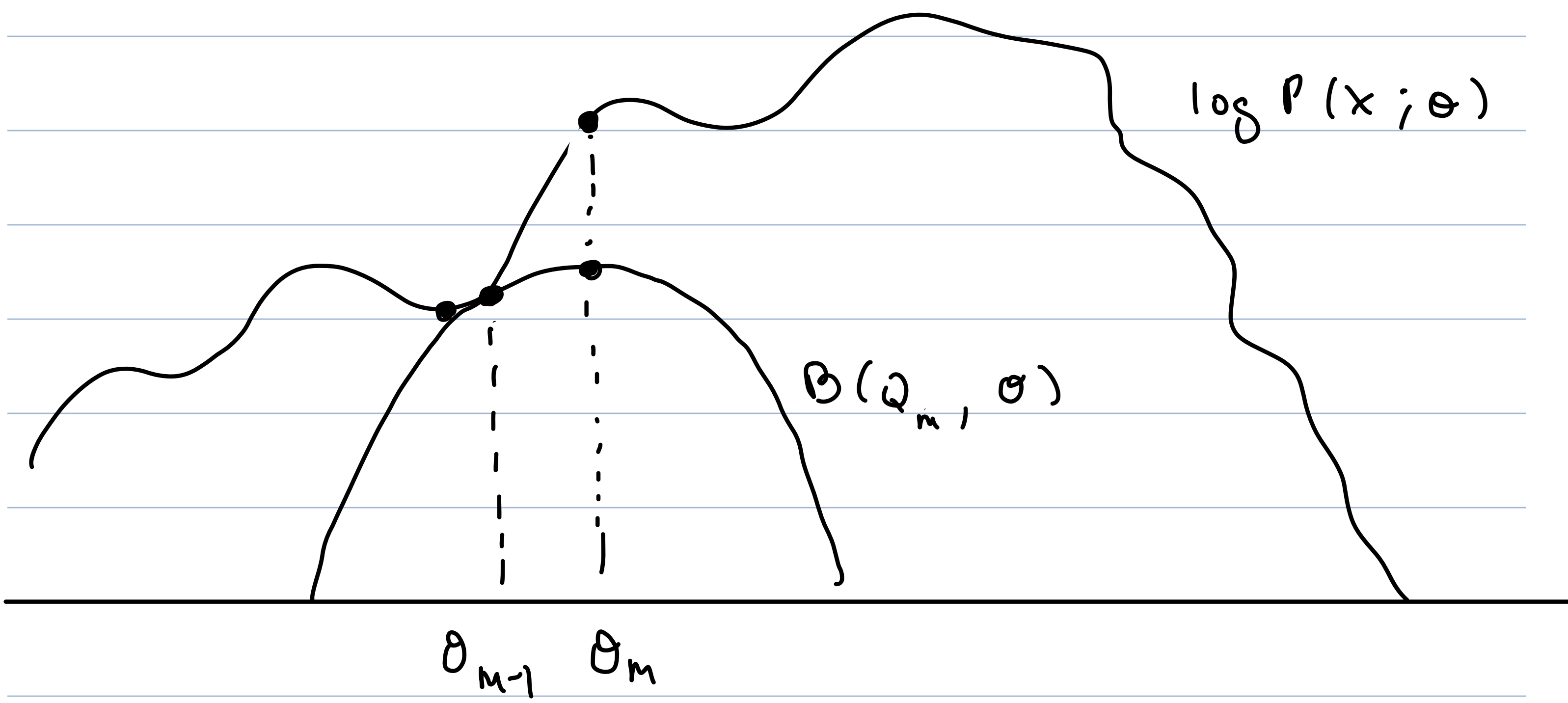
- initialize θ_0

for $m = 1, 2, \ldots$ until convergene in θ_m

- $Q_m = \underset{Q}{\operatorname{argmax}}\, B(\theta_{m-1}, Q)$ "E-step"

- $\theta_m = \underset{\theta}{\operatorname{argmax}}\, B(\theta, Q_m)$ "M-step"

This is a <u>minorize-maximize algorithm</u>



- converges to <u>local</u> mode of $P(x;\theta)$
- random restarts for θ_0 good idea
- ELBO is tight after E-step and always increases

E-step

$$Q_m = \underset{Q}{\arg\max}\ \mathcal{B}(\theta_{m-1}, Q)$$

$$= \underset{Q}{\arg\max}\ -KL(Q \,\|\, P_{\theta_{m-1}}(z|x))$$

$$= P_{\theta_{m-1}}(z|x)$$

M-step

$$\theta_m = \underset{\theta}{\arg\max}\ \mathcal{B}(\theta, Q_m)$$

$$= \underset{\theta}{\arg\max}\ \mathbb{E}_{Q_m}[\log P_\theta(x,z)]$$

$$= \underset{\theta}{\arg\max} \sum_z P_{\theta_{m-1}}(z|x) \log P_\theta(x,z)$$

Notice that the E-step seems redundant...

The entire algorithm could be stated as:

$$\boxed{\begin{array}{l}\text{Until convergence} \\ \quad \theta^{new} = \underset{\theta}{\arg\max}\ \mathbb{E}_{z \sim P(z|x;\theta^{old})}[\log P(x,z;\theta)]\end{array}}$$

... assuming that $\mathbb{E}_{P(z|x)}[\log P(x,z)]$ is tractable.

Is it? Often. But why? $\sum_z P(z|x;\theta^{old}) \log P(x,z;\theta)$

seems just as bad as $\sum_z P(x,z;\theta) = P(x;\theta)$ . . .

Recall last time:

$$P_\theta(x, z) = P_\theta(\bar{x}_{1:T}, z_{1:T}) = \prod_t P_\Lambda(z_t \mid z_{t-1})\, P_\mu(\bar{x}_t \mid z_t)$$

$$= \prod_t \Lambda(z_{t-1}, z_t) \prod_k \mathrm{Pois}(\bar{x}_t ; \mu_k)^{\mathbb{1}(z_t = k)}$$

Think about the M-step for just μ_k...

$$\underset{\mu_k}{\operatorname{argmax}}\ \mathbb{E}\left[\log P(\bar{x}_{1:T}, z_{1:T} ; \Lambda, \mu)\right]$$

$$= \underset{\mu_k}{\operatorname{argmax}}\ \mathbb{E}\left[\log \prod_t \mathrm{Pois}(\bar{x}_t ; \mu_k)^{\mathbb{1}(z_t = k)}\right]$$

$$= \underset{\mu_k}{\operatorname{argmax}} \sum_{z_1} \cdots \sum_{z_T} P(z_1 \cdots z_T \mid \bar{x}_{1:T} ; \mu^{old}, \Lambda^{old}) \log \prod_t \mathrm{Pois}(\cdots)^{\mathbb{1}(\cdots)}$$

$$= \underset{\mu_k}{\operatorname{argmax}} \sum_{z_1} \cdots \sum_{z_T} P(z_1 \ldots z_T \mid \bar{x}_{1:T}, \mu^{old}, \Lambda^{old}) \sum_t \mathbb{1}(z_t = k) \log \mathrm{Pois}(\bar{x}_t ; \mu_k)$$

Note that:

$$\sum_{z_1} \cdots \sum_{z_T} P(z_1 \cdots z_T \mid \cdots)\, \mathbb{1}(z_t = k)$$

$$= \sum_{z_t} P(z_t \mid \cdots)\, \mathbb{1}(z_t = k)$$

$$= P(z_t = k \mid \cdots) \quad \text{"belief"}$$

$$= \underset{\mu_k}{\operatorname{argmax}} \sum_t \underbrace{P(z_t = k \mid \bar{x}_{1:T}, \mu_k^{old})}_{\stackrel{\circ}{=} q_{tk}} \log \operatorname{Pois}(\bar{x}_t ; \mu_k)$$

So in this case we only need the beliefs q_{tk}.

$$= \underset{\mu_k}{\operatorname{argmax}} \sum_t q_{tk} \left[\bar{x}_t \log \mu_k - \mu_k \right]$$

$$= \underset{\mu_k}{\operatorname{argmax}} \left(\sum_t \bar{x}_t q_{tk} \right) \log \mu_k - \left(\sum_t q_{tk} \right) \mu_k$$

$$\hookrightarrow \quad \mu_k = \frac{\sum_t \bar{x}_t q_{tk}}{\sum_t q_{tk}}$$

This is why it is called the "Expectation" - step.

① Compute all expectations required by M-step:

$$\text{e.g., } q_{tk} = p(z_t = k \mid \bar{x}_{1:T}, \mu^{old}, \Lambda^{old}) \quad \forall\, t, k$$

$$\equiv \mathbb{E}_{z_{1:T} \sim p(z_1, \dots z_T \mid \bar{x}_{1:T}, \mu^{old}, \Lambda^{old})} \left[\mathbb{1}(z_t = k) \right]$$

② Maximize μ, Λ ...

Notice also that the M-step was "nice".

why? Exponential family conditionals.

$$\underset{\mu_k}{\operatorname{argmax}} \sum_t \underbrace{P(z_t = k \mid \bar{x}_{1:T}, \mu_k^{old})}_{\triangleq q_{tk}} \log \operatorname{Pois}(\bar{x}_t; \mu_k)$$

Rewrite in terms of natural parameter $\eta_k = \log \mu_k$.

$$\underset{\eta_k}{\operatorname{argmax}} \sum_t q_{tk} \left(\eta_k^T t(\bar{x}_t) - a(\eta_k) \right)$$

$$\underset{\eta_k}{\operatorname{argmax}} \quad \eta_k^T \left(\sum_t t(\bar{x}_t) q_{tk} \right) - \left(\sum_t q_{tk} \right) a(\eta_k)$$

$$\nabla_{\eta_k} \text{ " } \quad \text{ " } = \sum_t q_{th} t(\bar{x}_t) - \mu_k \left(\sum_t t(\bar{x}_t) \right)$$

(Note for all expfam $\nabla_\eta a(\eta) = \mu$)

$$= 0 \implies \eta_k = \dots \text{ s.t. } \mu_k = \frac{\sum_t q_{tk} t(\bar{x}_t)}{\sum_t t(\bar{x}_t)}.$$

Returning to the question of why

$$\sum_z P(z \mid x ; \theta^{old}) \log P(x, z ; \theta)$$

is often a "nicer" objective than

$$\sum_z P(x, z ; \theta)$$

At a very high level, for the following reason:

$$\mathbb{E} \log \prod \exp(\sum \cdots) = \sum \sum \mathbb{E} [\cdots]$$

we will see this motif again...

e.g. ...

$$f_{tk} \triangleq 1(z_t = k)$$

$$\mathbb{E}\left[\log \prod_t \prod_k \exp(\eta_k^T t(x) - a(\eta_k))^{f_{tk}}\right]$$

$$= \mathbb{E}\left[\log \exp\left(\sum_k \eta_k^T \left(\sum_t f_{tk} t(x_t)\right) - \sum_k \left(\sum_t f_{tk}\right) a(\eta_k)\right)\right]$$

$$\cdots \qquad q_{tk} \triangleq \mathbb{E}[f_{tk}]$$

$$= \sum_k \eta_k^T \left(\sum_t q_{tk} t(x_t)\right) - \sum_k \left(\sum_t q_{tk}\right) a(\eta_k)$$

Models with expfam conditionals and lots of conditional independence tend to lead to "nice" EM.

Do we only ever need singleton beliefs? No...

$$B(\theta, Q) \propto_{\Lambda} \mathbb{E}_Q[\log P(x, z; \theta)]$$

$$\propto_{\Lambda} \mathbb{E}_Q\left[\log \prod_{t=1}^{T} \prod_{k=1}^{K} \prod_{j=1}^{K} P(z_t = k \mid z_{t-1} = j)^{\mathbb{1}(z_t = k)\mathbb{1}(z_{t-1} = j)}\right]$$

$$= \sum_t \sum_k \sum_j \mathbb{E}_Q[\mathbb{1}(z_t = k, z_{t-1} = j)] \log \Lambda(j,k)$$

$$= \sum_k \sum_j \left[\sum_t Q(z_t = k, z_{t-1} = j)\right] \log \Lambda(j,k)$$

lagrange multiplier to enforce that Λ_j sums to 1

$$\frac{\partial}{\partial \Lambda(j,k)}\left[B(\theta, Q) + \eta_0\left(\sum_k \Lambda(j,k) - 1\right)\right]$$

$$= \frac{\sum_t Q(z_t = k, z_{t-1} = j)}{\Lambda(j,k)} - \eta_0$$

$$0 = \cdots$$

$$\Lambda(j,k) = \frac{\sum_t Q(z_t = k, z_{t-1} = j)}{\eta_0}$$

Setting $\eta_0 = \sum_t \sum_{k'} Q(z_t = k', z_{t-1} = j) = \sum_t Q(z_{t-1} = j)$

satisfies the constraint that η_0 enforces.

$$\Lambda(j,k) = \frac{\sum_t Q(z_t = k, z_{t-1} = j)}{\sum_t Q(z_{t-1} = j)}$$

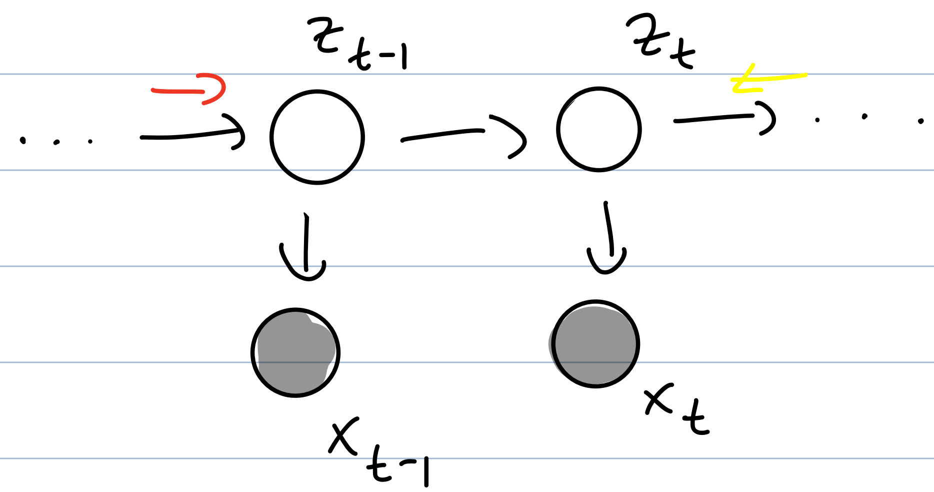
Summary: To maximize the ELBO with respect to Λ, we need the pairwise marginals $Q(z_t, z_{t-1})$.

Pairwise marginals in HMMs

The optimal $Q^*(z_t, z_{t-1}) = p(z_t, z_{t-1} \mid \bar{x}_{1:T})$

→ How do we compute the pairwise marginals?

Think of (z_{t-1}, z_t) as a "super node"...



Then $p(z_{t-1}, z_t \mid \bar{x}_{1:T}) \propto p(z_{t-1}, z_t, \bar{x}_{1:T})$

$$\propto \alpha_{t-1}(z_{t-1})\, \beta_t(z_t)\, \ell_{t-1}(z_{t-1})\, \ell(z_t)\, \Lambda(z_{t-1}, z_t)$$

$$= p(z_{t-1}, z_t, \bar{x}_{1:T})$$

Note that this is still $O(K^2)$.

More generally for N-way marginal $O(K^N)$.

What is or is not tractable in HMMs

- Evaluating the joint at given $z_1 \dots z_T$:

$$P(Z_1 = z_1 \dots Z_T = z_t \mid \bar{x}_{1:T}, \Theta) \qquad O(K^2 T) \text{ to run BP.}$$

- Evaluating the evidence

$$P(\bar{x}_{1:T} ; \Theta) \qquad O(K^2 T)$$

- Evaluating a gradient with backprop

$$\nabla_\Theta P(\bar{x}_{1:T} ; \Theta) \qquad O(K^2 T)$$

- Storing the joint:

$$P(z_1 \dots z_T \mid \bar{x}_{1:T}, \Theta) \quad \forall\, z_1 \dots z_T \qquad O(K^T) \text{ values}$$

- Storing the N-marginals:

$$O(K^N) \text{ values}$$ depends on N

Reasoning about the joint $P(z_1 \dots z_T \mid \dots)$ is often important.

e.g. speech recognition. Tractable options:

- MAP: $\underset{z_{1:T}}{\operatorname{argmax}}\ P(z_1 \dots z_T \mid \bar{x}_{1:T}; \theta)$

  "Viterbi algorithm"

- Sample: $z_1 \dots z_T \sim P(z_1, \dots z_T \mid \bar{x}_{1:T}, \Theta)$

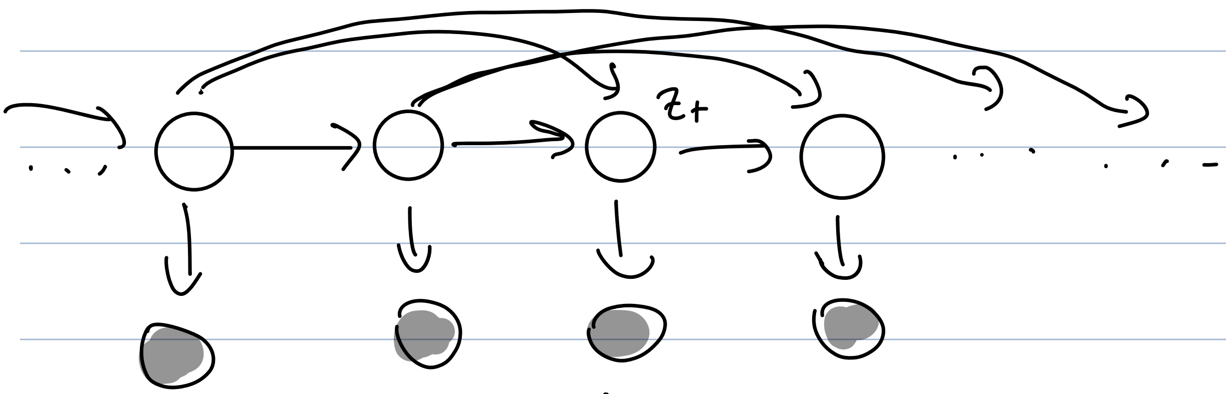
  "forward filtering backward sampling" (FFBS)

Return to last lecture to cover:

Viterbi, FFBS, extensions of HMMs

Monte Carlo (MC) - EM

Say we have an m-order HMM



- $z_{t+1} \sim p(z_{t+1} | z_t \dots z_{t-m})$
- Λ is a $(km \times k)$ matrix
  (or a $\underbrace{k \times \cdots \times k}_{m+1}$ tensor)
- Exact EM would require all
  $(m+1)$-marginals $p(z_{t+1} \dots z_{t-m} | \bar{x}_{1:T})$
- Large m → intractable
- Instead, replace the E-step with:
  - $z_1^s \dots z_T^s \sim p(z_1 \dots z_T | \bar{x}_{1:T}, \Theta^{old})$ $\quad s = 1 \dots S$
  - $Q(z_{1:T}) = \frac{1}{S} \sum_{s=1}^{S} \mathbb{1}(z_{1:T} = z^s_{1:T})$

    "atomic measure", each unique $z^s_{1:T}$ is an "atom".
- $\Theta^{new} = \underset{\Theta}{\arg\max} \frac{1}{S} \sum_s \log p(\bar{x}_{1:T}, z^s_{1:T}; \Theta)$
- $S = 1$ is called "Stochastic EM".

Variational EM

- Constrain $Q(z_{1:T})$ to be from a tractable family $\mathcal{F}$

- e.g. $Q(z_{1:T}) = \prod_t Q(z_t)$ "factorized"

- Variational E-step:

$$Q^{new} = \underset{Q \in \mathcal{F}}{\operatorname{argmin}} \; KL\left(Q(z) \,\|\, P(z \mid x; \Theta)\right)$$