LECTURE 16 EM algorithm

- Expectation Maximization:
  iterative algorithm for maximizing likelihood

- Given: vectors of "visible" variables v_n
  Hidden: vectors of "hidden" variables h_n
  Where n is example index
- Complete data set: $\underbrace{V = \{v_n\}}_{\text{true data}}$, $H = \{h_n\}$

- Model specifies a joint distribution

$$P(V, H \mid \theta)$$

parameters of model

- Usually i.i.d. data

$$P(V, H \mid \theta) = \prod_n P(v_n, h_n \mid \theta)$$

Then Whole analysis "decomposes"!

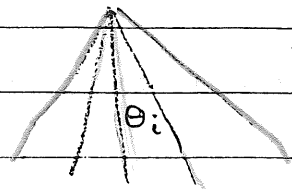
- $$P(V \mid \theta) = \sum_H P(V, H \mid \theta)$$

$$= \sum_H P(V \mid H, \theta)\, P(H \mid \theta)$$

Goal:
- Maximize $\ln P(V \mid \theta)$

Note: Its log of a sum over the hidden variables

Example 1: Mixture of m fixed densities



$P(x|i)$
input distributions

$\theta = \{\theta_1, \theta_2, \ldots, \theta_m)$
Mixture coefficients
Statistitions often use θ for their model / parameter

$$P(x|\theta) = \sum_i \theta_i P(x|i)$$

VISIBLE HIDDEN

PARAMETER

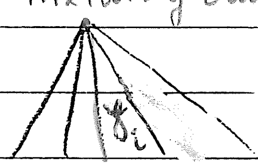
Formally: $\underset{\text{VISIBLE}}{P(x|\theta)} = \sum_i P(x, i | \theta)$

$$= \sum_i P(x|i, \cancel{\theta}) \underbrace{P(i|\theta)}_{\theta_i}$$

fixed densities

Example 2: Mixture of Gaussians

Mixture of Gaussians



$P(x|i,\theta)$

PARAMETERS

$\theta = \{\gamma_i\} \cup \{\mu_i\} \cup \{\Sigma_i\}$ $1 \le i \le k$

mixture coefficients k — means $d \times 1$ — co-variance matrices $d \times d$

$$P(x|\theta) = \sum_i \underbrace{P(i|\theta)} \underbrace{P(x|i,\theta)}$$

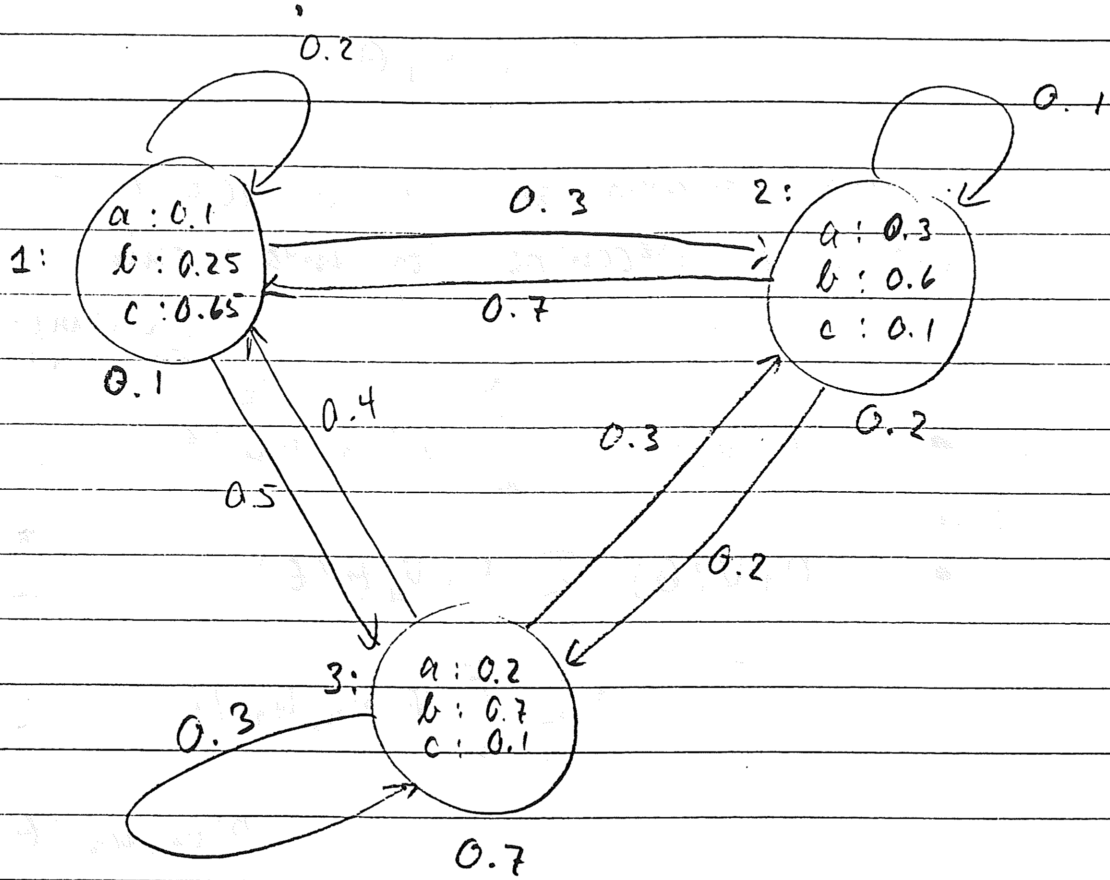
$$= \sum_i \gamma_i \frac{1}{(2\pi)^{d/2} |\Sigma_i|^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^T \Sigma^{-1} (x-\mu_i)}$$

POINT $\bar{x} \in \mathbb{R}^d$ VISIBLE

$i \in \{1, 2, \ldots, k\}$ HIDDEN

↑ # OF GAUSSIANS

Example 3: Hidden Markov Models



$$P(\underbrace{abca}_{\text{observed sequence}}, \underbrace{1231}_{\text{hidden state sequence}} \mid \Theta) = \begin{array}{cccccccc} & a & & b & & c & & a \\ 0.1 & 0.1 & 0.3 & 0.6 & 0.2 & 0.1 & 0.4 & 0.1 \\ & 1 & & 2 & & 3 & & 1 \end{array}$$

Θ parameters

PARAMETERS: $\Theta = \{\Theta_i\}_{1 \le i \le m}$ initial state, output, transition probabilities

$$P(x, s \mid \Theta) = \prod_{i=1}^{m} \Theta_i^{\,n_i(x,s)},$$

where n_i is # of times Θ_i occurs in (x, s)

$$P(x \mid \Theta) = \sum_{s} P(x, s \mid \Theta)$$

sum over all hidden paths

Loss: $-\ln P(V|\Theta)$

$$= -\ln \sum_H P(V,H|\Theta)$$

Hard to minimize because log of sum!

I.i.d. case: Decomposition into a sum

$V = \{v_n\}, \quad H = \{h_n\}$

SEQUENCES OF FIXED LENGTH N | ONE (v_n, h_n) PER EXAMPLE

ASSUMPTION: $P(V,H|\Theta) = \prod_{n=1}^{N} p(v_n, h_n|\Theta)$

THIS IMPLIES:

- $$P(V|\Theta) = \sum_H P(V,H|\Theta)$$
  $$= \sum_H \prod_n P(v_n, h_n|\Theta)$$
  $$= \sum_{h_1,h_2,\ldots,h_N} \prod_n P(v_n, h_n|\Theta)$$
  $$= \prod_n \sum_{h_n} P(v_n, h_n|\Theta)$$
  $$= \prod_n P(v_n|\Theta)$$
- $$P(H|V,\Theta) = \frac{P(H,V|\Theta)}{P(V|\Theta)} = \frac{\prod_n P(h_n, v_n|\Theta)}{\prod_n P(v_n|\Theta)} = \prod_n P(h_n|v_n,\Theta)$$
- $$-\ln P(V|\Theta) = -\sum_n \ln \underbrace{P(v_n|\Theta)}_{\text{STILL SUM: } \sum_{h_n} P(v_n,h_n|\Theta)}$$

Product becomes sum

- Loss is essentially a relative entropy

$$\sum_n \frac{1}{N} \ln \frac{\frac{1}{N}}{P(v_n|\Theta)} = -\frac{1}{N} \sum_n \ln(v_n|\Theta) - \ln N$$

I.I.D. MAKES RELATIVE ENTROPIES DECOMPOSE AS WELL

$$\sum_H P(H|V,\theta) \ln \frac{p(H|V,\theta)}{p(H|V,\tilde{\theta})}$$

$$= \sum_{\{h_n\}} \prod_n p(h_n|v_n,\theta) \ln \frac{\prod_n p(h_n|v_n,\theta)}{\prod_n p(h_n|v_n,\tilde{\theta})}$$

$$= \sum_n \sum_{h_n} p(h_n|v_n,\theta) \ln \frac{p(h_n|v_n,\theta)}{p(h_n|v_n,\tilde{\theta})}$$

SIMPLEST CASE

$$\sum_{x,y} p(x)p(y) \ln \frac{p(x)p(y)}{q(x)q(y)}$$

$$= \sum_{x,y} p(x)p(y) \left(\ln \frac{p(x)}{q(x)} + \ln \frac{p(y)}{q(y)} \right)$$

$$= \sum_x p(x) \ln \frac{p(x)}{q(x)} + \sum_y p(y) \ln \frac{p(y)}{q(y)}$$

UPSHOT:

ALL BECOMES SUMS OVER EXAMPLES

$$\underbrace{\sum_H P(H|V,\theta) \ln \frac{P(H|V,\theta)}{P(H|V,\tilde{\theta})}}_{\text{Divergence that motivates EM}} + \eta \underbrace{\left(-\ln P(V|\tilde{\theta})\right)}_{\text{Too hard to minimize !}} \qquad (*)$$

$$= \sum_H P(H|V,\theta) \ln \frac{P(H,V|\theta)/P(V|\theta)}{P(H,V|\tilde{\theta})/P(V|\tilde{\theta})} - \eta \ln P(V|\tilde{\theta})$$

$$= \sum_H P(H|V,\theta) \ln \frac{P(H,V|\theta)}{P(H,V|\tilde{\theta})} + \cancel{\ln P(V|\tilde{\theta})} - \ln P(V|\theta) - \eta \cancel{\ln P(V|\tilde{\theta})}$$

$\eta = 1$

$$= \underbrace{\sum_H P(H|V,\theta) \ln \frac{P(H,V|\theta)}{P(H,V|\tilde{\theta})}}_{\text{easy to minimize}} \underbrace{- \ln P(V|\theta)}_{\text{constant}}$$

Estimation Step:

Compute posterior $P(H|V,\theta)$

Maximization Step:

$$\tilde{\theta} := \arg\max_{\tilde{\theta}} \sum_H P(H|V,\theta) \ln P(H,V|\tilde{\theta})$$

IID CASE:

$$P(V,H|\theta) = \prod_{n=1}^{N} P(v_n, h_n|\theta)$$

(*) becomes: $-\sum_n \sum_{h_n} P(h_n|v_n,\theta) \ln P(h_n, v_n|\tilde{\theta})$ + constant

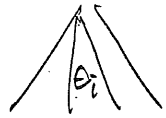
E-step:

Compute posteriors $P(h_n|v_n,\theta)$

M-step:

$$\tilde{\theta} = \arg\max_{\tilde{\theta}} \sum_n \sum_{h_n} P(h_n|v_n,\theta) \ln P(h_n, v_n|\tilde{\theta})$$

Example 1:



$P(x|i)$

Parameters $\theta = \{\theta_1, \dots, \theta_m\}$
visable x
hidden i
fixed densities : $P(x|i,\theta) = P(x|i)$

$$P(x|\theta) = \sum_i P(x, i|\theta)$$

$$= \sum_i \underbrace{P(i|\theta)} \underbrace{P(x|i,\theta)}$$

$$= \sum_i \theta_i \quad P(x|i)$$

E-step:

$$P(i|x_n,\theta) = \frac{P(i, x_n|\theta)}{P(x_n|\theta)} = \frac{\theta_i P(x_n|i)}{\sum_j \theta_j P(x_n|j)}$$

EXPECTED USAGE OF DISTRIBUTION i

M-step:

Maximize

$$\sum_n \sum_{x_n} P(i|x_n,\theta) \ln P(x_n, i|\tilde{\theta}) + \lambda\left(\sum \tilde{\theta}_i - 1\right)$$

$$= \sum_n \sum_{x_n} P(i|x_n,\theta) \ln \tilde{\theta}_i \underbrace{p(x_n|i)}_{\text{constant}} + \lambda\left(\sum \tilde{\theta}_i - 1\right)$$

$$\frac{\partial}{\partial \tilde{\theta}_i} = \sum_n \frac{P(i|x_n,\theta)}{\tilde{\theta}_i} + \lambda = 0$$

$$\Longleftrightarrow \quad \sum_n P(i|x_n,\theta) + \lambda \tilde{\theta}_i = 0 \qquad (*)$$

$$\sum_i \sum_n P(i|x_n,\theta) + \lambda \sum_i \tilde{\theta}_i = 0$$

$$\lambda = -N$$

From $(*)$: $\tilde{\theta}_i = \frac{1}{N} \sum_n P(i|x_n,\theta)$

Average Posterior

$N=1$ Bayes Rule

Example 2: Mixture of Gaussians

$P(x|i,\theta)$ not independend of θ

$$P(x|\theta) = \sum_i \underbrace{P(i|\theta)}\,\underbrace{P(x|i,\theta)}$$

$$= \sum_i \gamma_i \,(2\pi)^{-d/2} \, |\Sigma_i|^{-1/2} \, e^{-\frac{1}{2}(x-\mu_i)^T \Sigma^{-1} (x-\mu_i)}$$

E step:

$$P(i|x_n,\theta) = \frac{P(i,x_n|\theta)}{P(x_n|\theta)} = \frac{\gamma_i \, P(x|i,\theta)}{\sum_j \gamma_j \, P(x|j,\theta)}$$

M step:

Maximize

$$\sum_n \sum_i P(i|x_n,\theta) \ln P(i,x_n|\tilde{\theta})$$

$$= \sum_n \sum_i P(i|x_n,\theta) \ln \underbrace{P(i|\tilde{\theta})}_{\tilde{\gamma}_i} P(x_n|i,\tilde{\theta}) \quad (*)$$

3 Classes of parameters: a) Mixture coefficients γ_i

b) means μ_i

c) covariance matrices Σ_i

Maximize above for one class at a time while keeping the other classes fixed.

a) μ_i and Σ_i fixed

Thus $P(x_n | i, \tilde{\theta}) = P(x_n | i, \theta)$

(*) becomes $\sum_n \sum_i P(i | x_n, \theta) \ln \tilde{\gamma}_i$ + constant

Minimize as before with Lagrangien :

$$\tilde{\gamma}_i = \frac{1}{N} \sum_n P(i | x_n, \theta)$$

b) and c) : γ_i fixed

Thus $P(i | \tilde{\theta}) = P(i | \theta) = \gamma_i$

(*) becomes $\sum_n \sum_i P(i | x_n, \theta) \ln P(x_n | i, \tilde{\theta})$ + const.

$$\sum_n \sum_i P(i | x_n, \theta) \ln \left((2\pi)^{-d/2} |\tilde{\Sigma}_i|^{-\frac{1}{2}} e^{-\frac{1}{2}(x_n - \tilde{\mu}_i)^T \tilde{\Sigma}_i^{-1} (x_n - \tilde{\mu}_i)} \right) + \text{const.}$$

$$\sum_n \sum_i P(i | x_n, \theta) \left(-\frac{1}{2} \ln |\tilde{\Sigma}_i| - \frac{1}{2} (x_n - \tilde{\mu}_i)^T \tilde{\Sigma}_i^{-1} (x_n - \tilde{\mu}_i) \right) + \text{const}$$

$\tilde{\Sigma}_i = \Sigma_i$ | $\tilde{\mu}_i = \mu_i$

$$\sum_n \sum_i P(i | x_n, \theta) \left(-\frac{1}{2} (x_n - \tilde{\mu}_i)^T \Sigma_i^{-1} (x_n - \tilde{\mu}_i) \right) + \text{const}$$

b)

$$\sum_n \sum_i P(i | x_n, \theta) \left(-\frac{1}{2} \ln |\tilde{\Sigma}_i| - \frac{1}{2} (x_n - \mu_i)^T \tilde{\Sigma}_i^{-1} (x_n - \mu_i) \right) + \text{const}$$

c)

b) $$\frac{\partial}{\partial \tilde{\mu}_i} = \sum_n P(i|x_n,\theta)\, \Sigma_i^{-1} (x_n - \tilde{\mu}_i) = 0$$

$$\iff \sum_n P(i|x_n,\theta)\,(x_n - \tilde{\mu}_i) = 0$$

$$\sum_n P(i|x_n,\theta)\, x_n = \sum_n P(i|x_n,\theta)\, \tilde{\mu}_i$$

$$\tilde{\mu}_i = \frac{\sum_n P(i|x_n,\theta)\, x_n}{\sum_n P(i|x_n,\theta)}$$

c) $$\frac{\partial}{\partial \tilde{\Sigma}_i} = \sum_n P(i|x_n,\theta)\left(-\frac{1}{2}\tilde{\Sigma}_i^{-1} + \frac{1}{2}\tilde{\Sigma}_i^{-1}(x_n-\mu_i)(x_n-\mu_i)^T\tilde{\Sigma}_i^{-1}\right) = 0$$

$$\iff \sum_n P(i|x_n,\theta)\left(-\frac{1}{2}\tilde{\Sigma}_i + \frac{1}{2}(x_n-\mu_i)(x_n-\mu_i)^T\right) = 0$$

$$\tilde{\Sigma}_i = \frac{\sum_n P(i|x_n,\theta)(x_n-\mu_i)(x_n-\mu_i)^T}{\sum_n P(i|x_n,\theta)}$$

High-level intuition:
In EM parameters are often updated to their usage when generating the data.

Intuition why EM works:

A) Minimizing $-\ln P(V|\tilde{\theta}) = -\ln \sum_H P(V,H|\tilde{\theta})$

is hard because "ln" of a sum.

Minimizing $-\sum_H P(H|V,\theta) \ln P(H,V|\tilde{\theta})$

"easy" when $P(H,V|\tilde{\theta})$ is product!

EM avoids minimizing "ln" of a sum directly.

By adding distance, minimization simplified.

B) $-\ln P(V|\tilde{\theta}) = -\ln \sum_H P(V,H|\tilde{\theta})$ has many symmetries.

For example by renaming any local minima

for a mixture of m distinct Gaussians becomes

$m!$ local minima.

In $-\sum_H P(H|V,\theta) \ln P(H,V|\tilde{\theta})$ not as many

symmetries. Hidden variables are "coupled"

with visable variables.

Example 3: HMM's $\quad P(x,s|\theta) = \prod_{i=1}^{I} \theta_i^{n_i(x,s)}$

Maximize $\sum_n \sum_{s_n} P(s_n|x_n,\theta) \ln P(x_n,s_n|\tilde{\theta})$

$$= \sum_n \sum_{s_n} P(s_n|x_n,\theta) \sum_i n_i(x_n,s_n) \ln \tilde{\theta}_i$$

$$= \sum_n \sum_i \ln \tilde{\theta}_i \underbrace{\sum_{s_n} P(s_n|x_n,\theta)\, n_i(x_n,s_n)}_{\hat{n}_i(x_n|\theta)}$$

Expected "usage" of param. θ_i

Computable by dynamic program.

$[i] ::= \{ j : \theta_i$ and θ_j in same "class"$\}$

All parameters of a class must sum to one.

- two classes associated with a state.
- one class associated with the initial state probabilities

Maximize

$$\sum_n \sum_i (\ln \tilde{\theta}_i)\, \hat{n}_i(x_n|\theta) + \sum_{[i]} \lambda_{[i]} \left(\sum_{j \in [i]} \theta_j - 1 \right)$$

not right

$$\frac{\partial}{\partial \tilde{\theta}_i} = \sum_n \frac{\hat{n}_i(x_n|\theta)}{\tilde{\theta}_i} + \lambda_{[i]} = 0$$

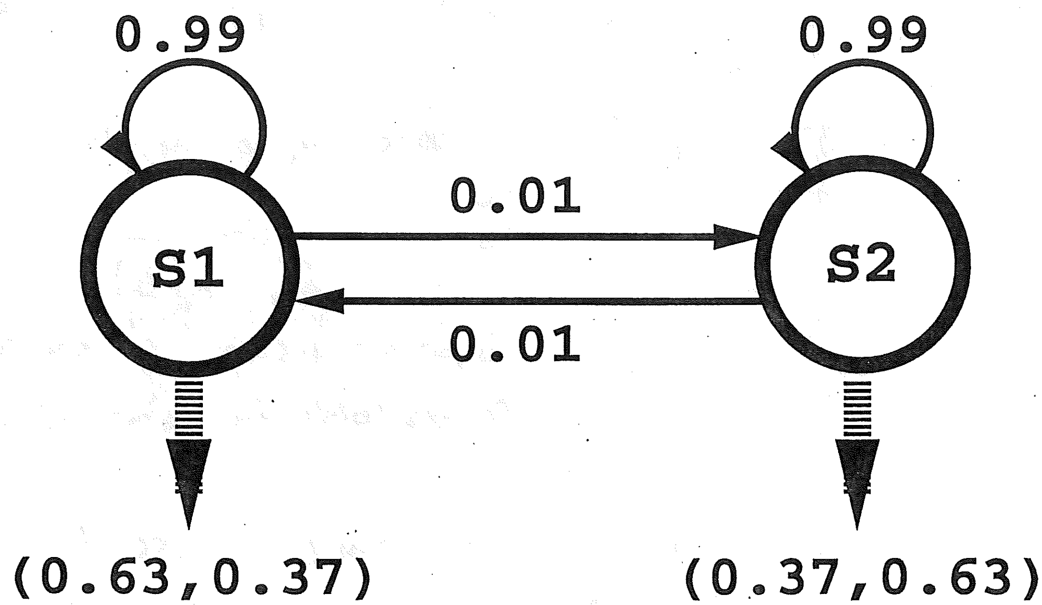
$$\tilde{\theta}_i = \sum_n \frac{\hat{n}_i(x_n|\theta)}{-\lambda_{[i]}}$$

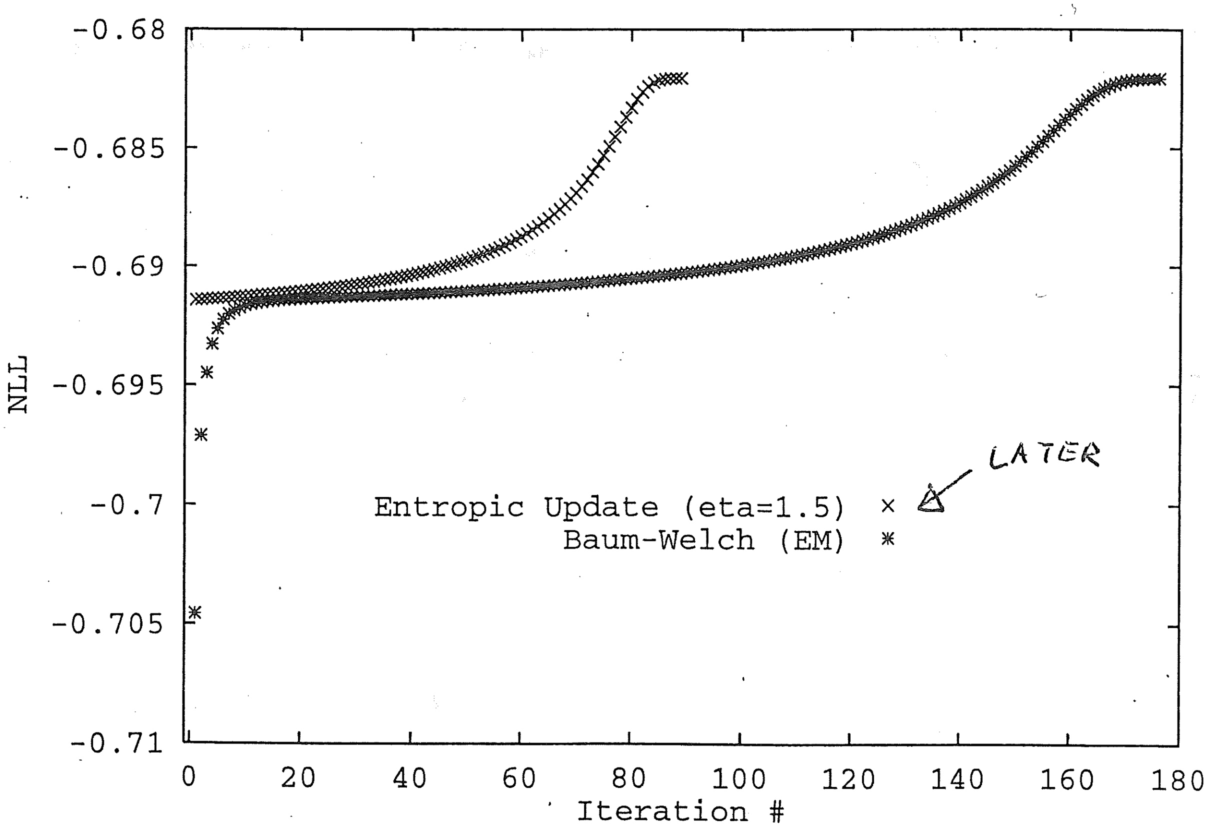
Enforcing constraint: $\sum_{j \in [i]} \tilde{\theta}_j = 1$

$$\tilde{\theta}_i = \frac{\sum_n \hat{n}_i(x_n|\theta)}{\sum_{j \in [i]} \sum_n \hat{n}_j(x_n|\theta)}$$

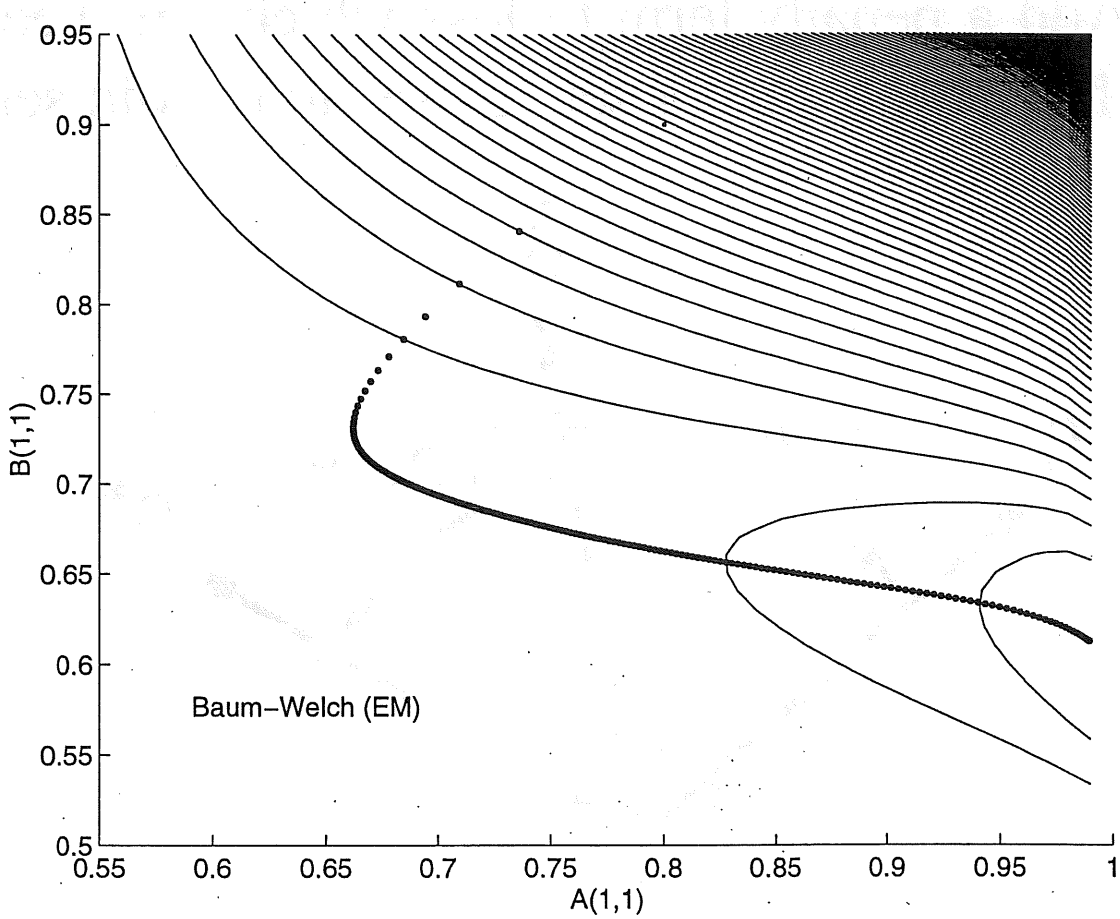
EM is too slow!

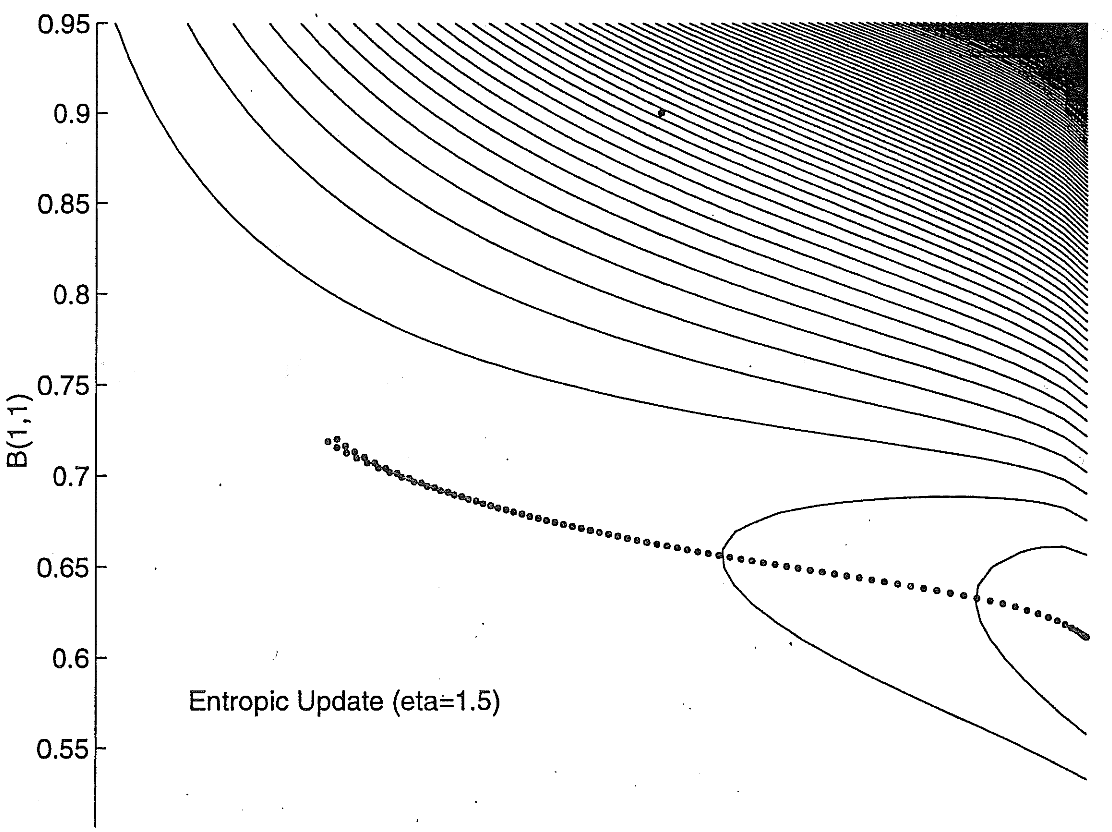
Synthetic Data 1: 2 State HMM





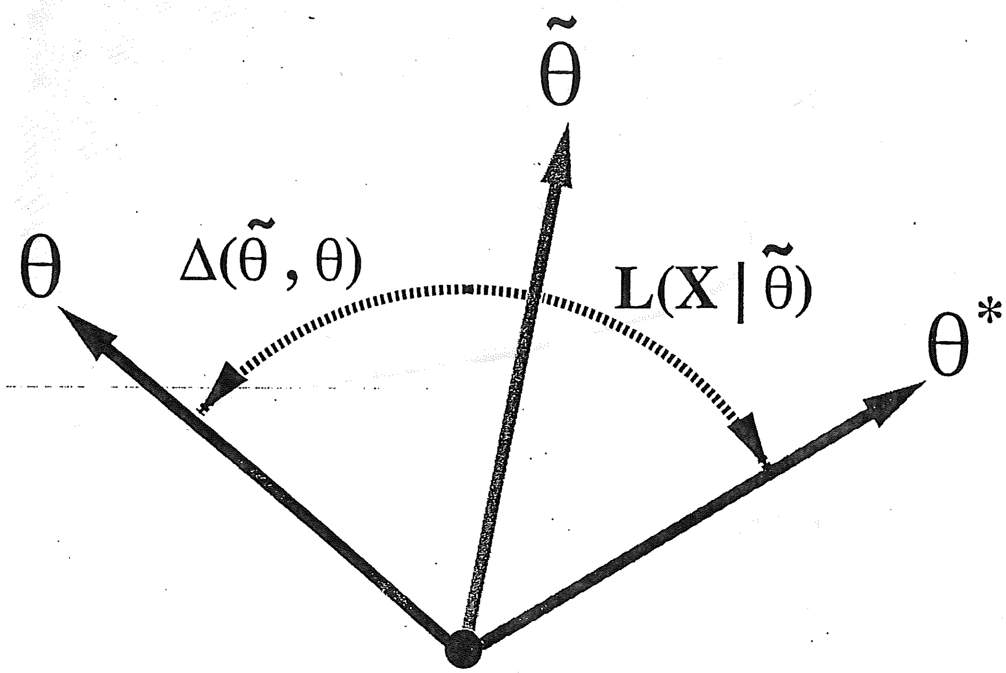
Synthetic Data 1 (cont.)





Framework for Parameter Update

- Add a penalty term to loss which will keep the new parameter set "close" to the old set.



- Solve:

$$\Theta^{t+1} = \arg\min_{\widetilde{\Theta}} U^t(\widetilde{\Theta}) \qquad (*)$$
$$U^t(\widetilde{\Theta}) = \Delta(\widetilde{\Theta}, \Theta^t) + \eta \operatorname{loss}(S|\widetilde{\Theta})$$

η is a non-negative trade-off parameter that becomes the learning rate of the algorithm.

Any update of the form (*) called

Implicit Update

Minimal Properties of the Divergence

(i) $\Delta(\Theta, \Theta) = 0$

(ii) $\Delta(\widetilde{\Theta}, \Theta) > 0$ whenever $\widetilde{\Theta} \neq \Theta$.

Key Lemma

If $U^t(\widetilde{\Theta}) < U^t(\Theta^t)$
then $\text{loss}(S|\widetilde{\Theta}) < \text{loss}(S|\Theta^t)$.

Proof:

$$\begin{aligned} U^t(\widetilde{\Theta}) &= \Delta(\widetilde{\Theta}, \Theta^t) + \eta \, \text{loss}(S|\widetilde{\Theta}) \\ < \; U^t(\Theta^t) &= \Delta(\Theta^t, \Theta^t) + \eta \, \text{loss}(S|\Theta^t) \overset{(i)}{=} \eta \, \text{loss}(S|\Theta^t) \end{aligned}$$

This is equivalent to

$$\begin{aligned} \text{loss}(S|\widetilde{\Theta}) &= \text{loss}(S|\Theta^t) - \frac{\Delta(\widetilde{\Theta}, \Theta^t)}{\eta} \\ &\overset{(ii)}{<} \text{loss}(S|\Theta^t) \end{aligned}$$

For any implicit update s.t. $\theta^{t+1} \neq \theta^t$,
$\text{loss}(S|\theta^{t+1}) < \text{loss}(S|\theta^t)$

What's good about EM:

- Implicit update. Thus negative log likelihood decreases in each iteration
- simple and elegand

Bad: Converges too slowly

$$\sum_H P(H|\underbrace{V}_{data},\theta) \ln \frac{P(H|V,\theta)}{P(H|V,\tilde{\theta})} - \eta \ln P(\underbrace{V}_{data}|\tilde{\theta})$$

Simplification $\Big\downarrow$ for $\eta = 1$

$$-\sum_H P(H|V,\theta) \ln P(H,V|\tilde{\theta}) + \text{const.}$$

Want $\eta > 1$. In that case simplification does not work!!!

Idea1: Use $\eta > 1$ but approximate $-\ln(V|\tilde{\theta})$ by 1. order Taylor. Does not seem to work.

Idea2: Use $\eta > 1$, different distance, and 1. order Taylor.

$$\sum_{V'}\sum_H P(H,V'|\tilde{\theta}) \ln \frac{P(H,V'|\tilde{\theta})}{P(H,V'|\theta)}$$

$$- \eta \left(\ln P(V|\theta) + (\tilde{\theta}-\theta) \frac{\partial \ln P(\overset{data}{V},\theta)}{\partial \theta}\right)$$

Explicit!

Distance we use:

- different direction of entropy
- V' integrated over domain
- aviods "ln" of sum in different way.

In all three examples our method converges faster.

$$\left.\frac{\partial -\ln P(V\mid\tilde{\theta}_\eta)}{\partial \eta}\right|_{\eta=0} < 0 \quad \text{unless at extremum}$$

Provided η is close enough to 0, then loss decreases.

We don't know why our method is so good.