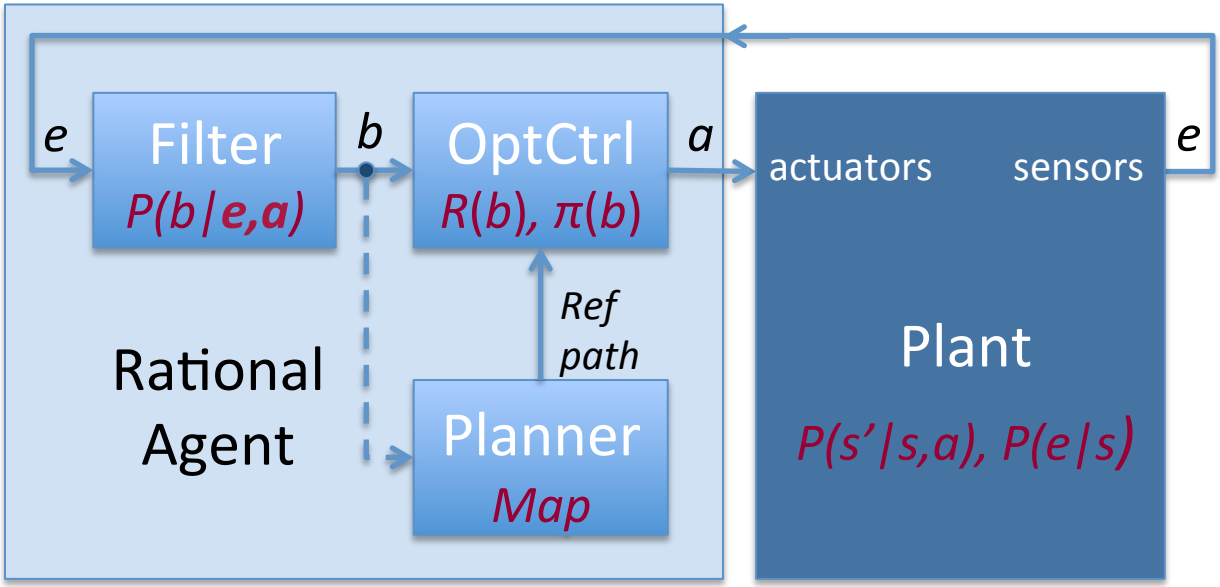
Learning Probabilistic Models

Chapter 20



Temporal-Models Problems

Inference from HMM and observations :

- Filtering $P(X_t \mid \mathbf{e}_{0:t})$: Current-state estimation
- Prediction $P(X_{t+k} \mid \mathbf{e}_{0:t})$: Future-state estimation
- Smoothing $P(X_k \mid \mathbf{e}_{0:t})$: Past-state estimation
- MLE $\text{argmax}_{\mathbf{x}_{0:t}} P(\mathbf{x}_{0:t} \mid \mathbf{e}_{0:t})$: Most-likely explanation

Learning best HMM from observations:

- EM $P(X_0)$, $P(X' \mid X)$, $P(E \mid X)$: Expectation Maximisation

Prediction

Filtering without the addition of new evidence :

$$P(X_{t+k} \mid \mathbf{e}_{0:t}) = \sum_{x_{t+k-1}} P(X_{t+k} \mid x_{t+k-1})\, P(x_{t+k-1} \mid \mathbf{e}_{0:t})$$

Predicting further and further into the future :

- Converges to the stationary distribution of the HMM
- Doomed to failure for more than a few steps ahead

Mixing time:

- Roughly the time it takes to reach the fixed point
- The more uncertainty the shorter the mixing time

Smoothing

Compute past-state distribution given evidence:

$$P(X_k \mid \mathbf{e}_{0:t}) = P(X_k \mid \mathbf{e}_{0:k}, \mathbf{e}_{k+1:t})$$

$$= \alpha P(X_k \mid \mathbf{e}_{0:k})\, P(\mathbf{e}_{k+1:t} \mid X_k, \mathbf{e}_{0:k}) \quad \text{By ext. Bayes rule}$$

$$= \alpha\, P(X_k, \mathbf{e}_{0:k})\, P(\mathbf{e}_{k+1:t}, X_k) \quad \text{By cond. independence}$$

$$P(X_k \mid \mathbf{e}_{0:t}) = \alpha\, \mathbf{f}_{0:k} \times \mathbf{b}_{k+1:t} \quad \text{forward} \times \text{backward algorithms}$$

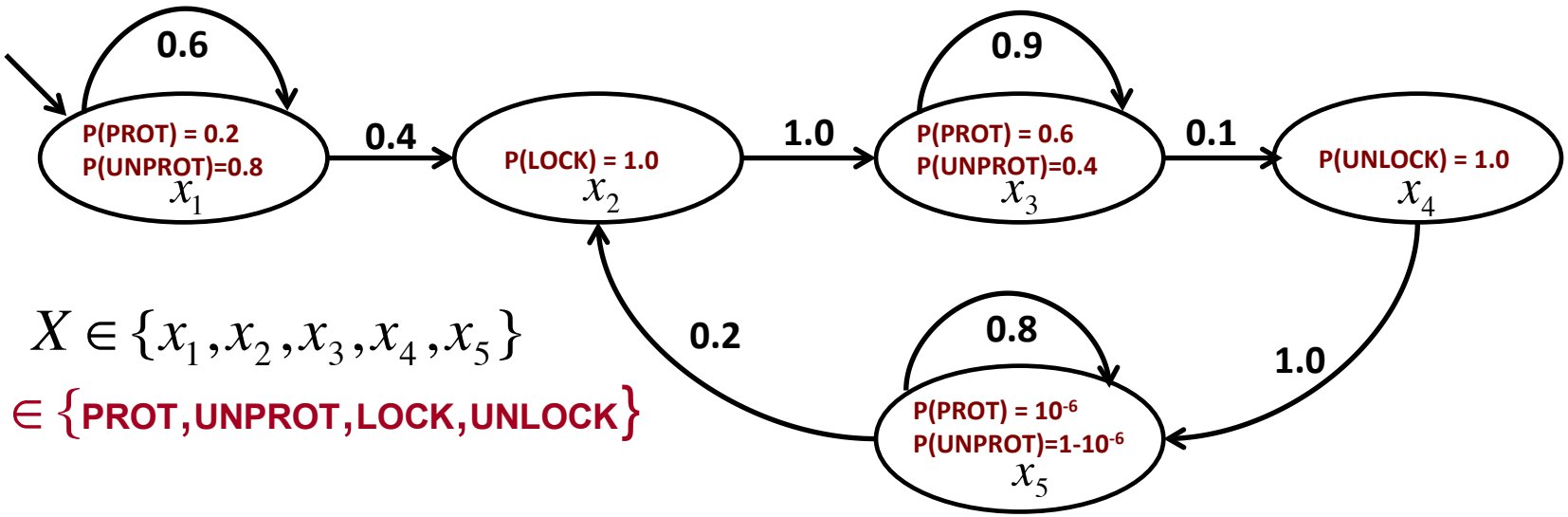
where $\times$ represents pointwise multiplication of vectors.

One needs to compute the backward message

Hidden Markov Model

T_H	x_1	x_2	x_3	x_4	x_5
x_1	0.6	0.4			
x_2			1		
x_3			0.9	0.1	$P(x_5 \mid x_3)$
x_4					1
x_5		0.2			0.8

O_H	PROT	UNPROT	LOCK	UNLOCK
x_1	0.2	0.8		
x_2			1	
x_3	0.6	0.4		$P(\text{UNLOCK} \mid x_3)$
x_4				1
x_5	10^{-6}	$1-10^{-6}$		



$X \in \{x_1, x_2, x_3, x_4, x_5\}$

$Y \in \{\text{PROT}, \text{UNPROT}, \text{LOCK}, \text{UNLOCK}\}$

Forward Algorithm

$$P(X_k \mid \mathbf{e}_{0:k}) = \alpha\, P(e_k \mid X_k) \sum_{x_{k-1}} P(X_k \mid x_{k-1})\, P(x_{k-1} \mid \mathbf{e}_{0:k-1})$$

For HMM $H = (\boldsymbol{\pi}, \mathbf{T}, \mathbf{O})$

function Forward ($\mathbf{e}_{0:k}$, H) returns $P(X_k \mid \mathbf{e}_{0:k})$

local $\mathbf{f} = \pi \times \mathbf{O}[e_0]^t$ // forward row vector

for $(i = 1 : k : +1)$ $\mathbf{f} = (\mathbf{f}\ \mathbf{T}) \times \mathbf{O}[e_i]^t$ // Forward

return Normalize($\mathbf{f}$)

Backward Algorithm

Analogous to the the forward algorithm but from t:

$$P(\mathbf{e}_{k+1:t} \mid X_k) = \sum_{x_{k+1}} P(\mathbf{e}_{k+1:t} \mid X_k, x_{k+1})\, P(x_{k+1} \mid X_k)$$

$$= \sum_{x_{k+1}} P(\mathbf{e}_{k+1:t} \mid x_{k+1})\, P(x_{k+1} \mid X_k)$$

$$= \sum_{x_{k+1}} P(\mathbf{e}_{k+1}, \mathbf{e}_{k+2:t} \mid x_{k+1})\, P(x_{k+1} \mid X_k)$$

$$P(\mathbf{e}_{k+1:t} \mid X_k) = \sum_{x_{k+1}} P(\mathbf{e}_{k+1} \mid x_{k+1})\, P(\mathbf{e}_{k+2:t} \mid x_{k+1})\, P(x_{k+1} \mid X_k)$$

$$\mathbf{b}_{k+1:t} = \text{Backward}(\mathbf{b}_{k+2:t}, \mathbf{e}_{k+2:t}), \quad \mathbf{b}_{t+1:t} = \mathbf{1}$$

Complexity of smoothing $P(X_k \mid e_{1:t})$ is $O(t)$

Backward Algorithm

$$P(\mathbf{e}_{k+1:t} \mid X_k) = \sum_{x_{k+1}} P(x_{k+1} \mid X_k)\, P(\mathbf{e}_{k+1} \mid x_{k+1})\, P(\mathbf{e}_{k+2:t} \mid x_{k+1})$$

For HMM $H = (\boldsymbol{\pi}, \mathbf{T}, \mathbf{O})$

function Backward ($\mathbf{e}_{k+1:t}$, H) returns $P(\mathbf{e}_{k+1:t} \mid X_k)$

local $\mathbf{b} = \mathbf{1}$ // backward row vector

for ($i = t : k+1 : -1$) $\mathbf{b} = (\mathbf{b} \times \mathbf{O}[e_i]^t)\, \mathbf{T}^t$ // Backward

return $\mathbf{b}$

Smoothing Algorithm (Fwd-Bwd)

$$P(X_k \mid \mathbf{e}_{1:t}) = \alpha\, P(X_k \mid \mathbf{e}_{1:k})\, P(\mathbf{e}_{k+1:t} \mid X_k)$$

For HMM $H = (\boldsymbol{\pi}, \mathbf{T}, \mathbf{O})$

function Smoothing ($\mathbf{e}_{1:t}$, H, $k \le t$) returns $P(X_k \mid \mathbf{e}_{1:t})$

local $\mathbf{f} = \boldsymbol{\pi} \times \mathbf{O}[e_0]^t$, $\mathbf{b} = \mathbf{1}$ // fwd & bwd row vectors

for $(i = 1 : k : +1)$ $\mathbf{f} = (\mathbf{f}\ \mathbf{T}) \times \mathbf{O}[e_i]^t$ // forward

for $(i = t : k+1 : -1)$ $\mathbf{b} = (\mathbf{b} \times \mathbf{O}[e_i]^t)\ \mathbf{T}^t$ // backward

return Normalise($\mathbf{f} \times \mathbf{b}$)

Baum-Welch Algorithm

Given

N : the number of states

Υ : The set of observations

ε : The error margin

$\mathbf{e}_{0,T}$: An observation sequence

Return HMM (π, **T**, **O**)

maximizes $\mathcal{L}(H)$ and $\mathcal{L}(H)$ changed less than ε

$$\mathcal{L}(H) = \log P(\mathbf{e}_{0:T} \mid H)$$

Baum-Welch Algorithm

function Learn-HMM ($\mathbf{e}_{0:t}$, N, Υ, ε) returns HMM (π, **T**, **O**)

local $H^* = (\boldsymbol{\pi}, \mathbf{T}, \mathbf{O})$ random // initialize H* randomly

repeat // main fixpoint loop

$H = H^*$;

(* E-Step *)

$\mathbf{f}_{i,0} = \boldsymbol{\pi}_i \mathbf{O}_{i,e_0}$; $\mathbf{f}_{i,t} = \mathbf{O}_{i,e_t} \sum_{j=1}^{N} \mathbf{f}_{j,t-1}\, \mathbf{T}_{j,i}$; $\forall i = 1{:}N,\ t = 1{:}T$ // fwd

$\mathbf{b}_{i,T} = \mathbf{1}$; $\mathbf{b}_{i,t} = \sum_{j=1}^{N} \mathbf{b}_{j,t+1}\, \mathbf{T}^t_{j,i}\, \mathbf{O}_{j,e_{t+1}}$; $\forall i = 1{:}N,\ t = 0{:}T-1$ // bwd

$\mathbf{s}_{i,t} = \mathbf{f}_{i,t}\, \mathbf{b}_{i,t} / \sum_{k=1}^{N} \mathbf{f}_{k,t}\, \mathbf{b}_{k,t}$; $\forall i = 1{:}N,\ t = 0{:}T$ // fwd-bwd

$\mathbf{A}_{i,j,t} = (\mathbf{f}_{i,t}\, \mathbf{T}_{i,j})(\mathbf{b}_{j,t+1}\, \mathbf{O}_{j,e_{t+1}}) / \sum_{k=1}^{N} \mathbf{f}_{k,t}\, \mathbf{b}_{k,t}$; $\forall i = 1{:}N,\ t = 0{:}T-1$ // trans

Baum-Welch Algorithm

(* M-Step *)

$\boldsymbol{\pi}^*_i = \mathbf{s}_{0,i};\ \forall i = 1{:}N$ // initial distribution

$\mathbf{T}^*_{i,j} = \sum_{t=0}^{T-1} \mathbf{A}_{i,j,t} \,/\, \sum_{t=0}^{T-1} \mathbf{s}_{i,t} \quad \forall i,j = 1{:}N$ // Transition **T** *

$\mathbf{O}^*_{i,e} = \sum_{t=0}^{T} 1_{e_t=e} \mathbf{s}_{i,t} \,/\, \sum_{t=0}^{T} \mathbf{s}_{i,t} \quad \forall i = 1{:}N,\ e \in \Upsilon$ // Output **O** *

until ($\mathcal{L}(H^*) - \mathcal{L}(H) \leq \varepsilon$) // main fixpoint loop

return (H*) // learned HMM

Baum-Welch Algorithm

function Learn-HMM ($\mathbf{e}_{0:t}$, N, Υ, ε) returns HMM (π, **T**, **O**)

local $H^* = (\boldsymbol{\pi}, \mathbf{T}, \mathbf{O})$ random // initialize H* randomly

repeat // main fixpoint loop

$H = H^*$;

(* E-Step *)

$\mathbf{f}_0 = \boldsymbol{\pi} \times \mathbf{O}^t_{e_0}$; $\mathbf{f}_t = (\mathbf{f}_{t-1}\,\mathbf{T}) \times \mathbf{O}^t_{e_t}$; $\forall t = 1:T$ // $P(X_t \mid \mathbf{e}_{1:t})$

$\mathbf{b}_T = \mathbf{1}$; $\mathbf{b}_t = (\mathbf{b}_{t+1} \times \mathbf{O}_{e_{t+1}})\,\mathbf{T}^t$; $\forall t = 1:T-1$ // $P(\mathbf{e}_{t+1:T} \mid X_t)$

$\mathbf{s}_t = \mathbf{f}_t \times \mathbf{b}_t / \mathbf{f}_t . \mathbf{b}_t$; $\forall t = 0:T$ // $P(X_t \mid \mathbf{e}_{1:T})$

$\mathbf{A}_t = (\mathbf{f}_t\,\mathbf{T}) \times (\mathbf{b}_{t+1} \times \mathbf{O}^t_{e_{t+1}})) / \mathbf{f}_t . \mathbf{b}_t$; $\forall t = 0:T-1$ // $P(X_{t+1} \mid X_t, e_{1:T})$

Baum-Welch Algorithm

(* M-Step *)

$\boldsymbol{\pi}^*_i = \mathbf{s}_{0,i}; \quad \forall i = 1:M$ // initial-state distribution

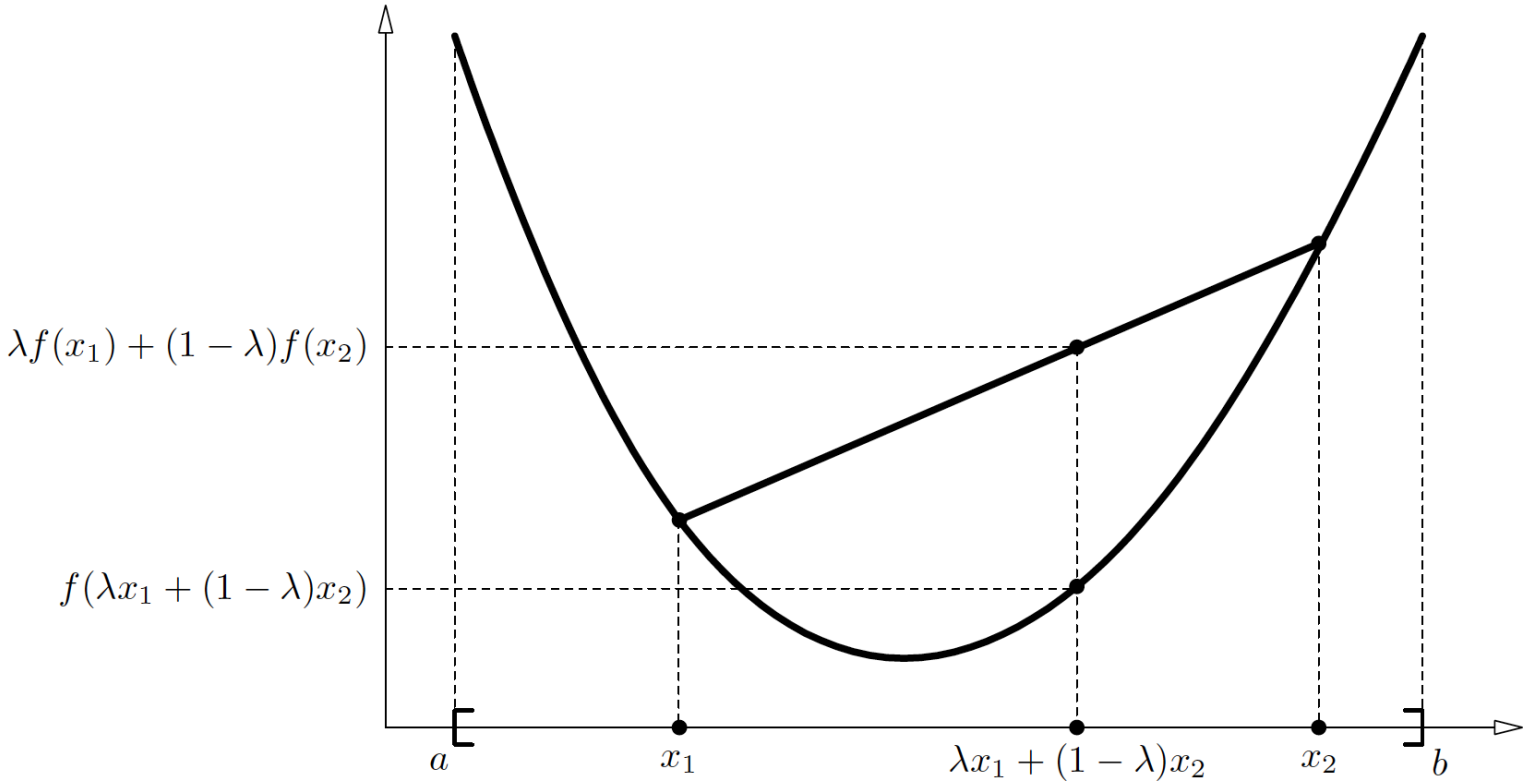
$\mathbf{T}^*_{i,j} = \sum_{t=0}^{T-1} \mathbf{A}_{i,j,t} \,/\, \sum_{t=0}^{T-1} \mathbf{s}_{i,t} \quad \forall i,j = 1:N$ // $E(P(X_{t+1} \mid X_t, \mathbf{e}_{1:T}))$

$\mathbf{O}^*_{i,e} = \sum_{t=0}^{T} 1_{e_t=e} \mathbf{s}_{i,t} \,/\, \sum_{t=0}^{T} \mathbf{s}_{i,t} \quad \forall i = 1:N,\ e \in \Upsilon$ // Output $\mathbf{O}^*$

until ($\mathcal{L}(H^*) - \mathcal{L}(H) \leq \varepsilon$) // main fixpoint loop

return (H*) // learned HMM

Convex Functions



A function f: [a,b] $\rightarrow \mathbb{R}$ is said to be (strictly) convex if:

$\forall x_1, x_2 \in [a,b], \lambda \in [0,1].\ f(\lambda x_1 + (1-\lambda) x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$

Convex Functions

Jensen's inequality: For any convex function

$$\forall x_1,..,x_n \in [a,b], \lambda_1,...,\lambda_n \in [0,1], \sum_{i=1}^{n} \lambda_i = 1.\ f(\sum_{i=1}^{n} \lambda_i x_i) \leq \sum_{i=1}^{n} \lambda_i f(x_i)$$

Proof: By induction.

$n = 1$: $f(x) \leq f(x)$

$n = 2$: By definition of convex function

Assume it holds for n: Prove it holds for n+1

$$f(\sum_{i=1}^{n+1} \lambda_i x_i) = f(\lambda_{n+1} x_{n+1} + \sum_{i=1}^{n} \lambda_i x_i) = f(\lambda_{n+1} x_{n+1} + (1-\lambda_{n+1}) \sum_{i=1}^{n} \frac{\lambda_i}{1-\lambda_{n+1}} x_i)$$

$$\leq \lambda_{n+1} f(x_{n+1}) + (1-\lambda_{n+1}) f(\sum_{i=1}^{n} \frac{\lambda_i}{1-\lambda_{n+1}} x_i)$$

$$\leq \lambda_{n+1} f(x_{n+1}) + (1-\lambda_{n+1}) \sum_{i=1}^{n} \frac{\lambda_i}{1-\lambda_{n+1}} f(x_i) = \sum_{i=1}^{n+1} \lambda_i f(x_i)$$

Convex Functions

Definition: The function f is said to be concave if $-f$ is convex.

Proposition: If f is twice differentiable and $\ddot{f}(x) \geq 0$ then f is convex.

Proposition: $f(x) = -\ln(x)$ is strictly convex on $(0,\infty)$.

Jensen's inequality corollary:

$$\forall x_1,..,x_n \in [a,b], \lambda_1,...,\lambda_1 \in [0,1], \sum_{i=1}^{n} \lambda_i = 1. \ \ln\left(\sum_{i=1}^{n} \lambda_i x_i\right) \geq \sum_{i=1}^{n} \lambda_i \ln(x_i)$$

Conditional Expectation

Expectation: Let RV X have domain $\mathcal{X}$. Then:

$$E[X] = \sum_{x \in \mathcal{X}} x\, P(X = x) \qquad E[X] = \int_{\mathcal{X}} x\, f(x)\, dx$$

Jensen's Corollary : For X RV and f concave $f(E[X]) \geq E[f(X)]$

Conditional Expectation: Let RV X have domain $\mathcal{X}$. Then:

$$E[X \mid Y = y] = \sum_{x \in \mathcal{X}} x\, P(X = x \mid Y = y) = \sum_{x \in \mathcal{X}} x \frac{P(X = x, Y = y)}{P(Y = y)}$$

$$E[X \mid Y = y] = \int_{\mathcal{X}} x\, f_X(x \mid Y = y)\, dx = \int_{\mathcal{X}} x \frac{f_{X,Y}(x,y)}{f_Y(y)}\, dx$$

Maximum Likelihood Estimation

Let E be a RV with pdf $P(E \mid \theta)$ with unkown parameter $\theta \in \Theta$.

Given outcome e of E

Maximize the likelihood function $\mathcal{L}(\theta) = P(e \mid \theta)$ wrt. θ over Θ

$$\hat{\theta} = \text{argmax}_{\theta \in \Theta} \; \mathcal{L}(\theta)$$

Theorem: $\text{argmax}_{\theta \in \Theta} \; \mathcal{L}(\theta) = \text{argmax}_{\theta \in \Theta} \; \log \mathcal{L}(\theta)$

Notation: $L(\theta) = \log \mathcal{L}(\theta)$

Problem: ML estimation has generally no closed-form solution

Problem: $L(\theta)$ is generally hard to maximize

Expectation Maximization

A class of optimizers specifically taylored to ML problems.
Works iteratively by maximizing successive local approx of L

Each iteration has two steps:

- E-step: Performs the approximation
- M-step: Performs the maximization

Key underlying idea of EM:

- Introduce hidden RV X: With $P(X \mid \theta)$ easier to maximize
- Any RV X such that: $\theta \rightarrow X \rightarrow E$ is a Markov Chain
- Hence: $P(X,E \mid \theta) = P(X \mid \theta)\, P(E \mid X,\theta) = P(X \mid \theta)\, P(E \mid X)$

Conceptually:

- X is a complete data space: If fully observed, easy to estimate θ

EM as Jensen's Ineq Consequence

EM formulation stems from a simple variational argument

Assumption: Pdf does not vanish to zero

$$L(\theta) - L(\theta^{(n)}) = \log \frac{P(e \mid \theta)}{P(e \mid \theta^{(n)})}$$ log property

$$= \log \int_x \frac{P(x,e \mid \theta)}{P(e \mid \theta^{(n)})} dx$$ sum up over z

$$= \log \int_x \frac{P(x,e \mid \theta)}{P(x,e \mid \theta^{(n)})} P(x \mid e,\theta^{(n)})\ dx$$ Bayes' rule

$$\geq \underbrace{\int_x \log \frac{P(x,e \mid \theta)}{P(x,e \mid \theta^{(n)})} P(x \mid e,\theta^{(n)})\ dx}_{Q(\theta,\theta^{(n)})}$$ Jensen's inequality

Auxiliary Function Q(θ,θ')

$$L(\theta) - L(\theta^{(n)}) \geq \underbrace{\int_x \log \frac{P(x,e \mid \theta)}{P(x,e \mid \theta^{(n)})} P(x \mid e,\theta^{(n)})\ dx}_{Q(\theta,\theta^{(n)})}$$

The likelihood variation from $\theta^{(n)}$ to θ: Always greater than $Q(\theta,\theta^{(n)})$

If we do not change parameter $\theta^{(n)}$ then: $Q(\theta^{(n)},\theta^{(n)}) = 0$

We are guaranteed to increase likelyhood if: $Q(\theta,\theta^{(n)}) > 0$

Iterating such a process: Defines an EM algorithm

Convergence theorem: Can be proven under mild conditions

Only trick behind EM: Exploit concavity of logarithm function!

EM as Expectation-Maximization

Define: $Q(\theta|\theta^{(n)}) = \int_x \ln P(x,e\,|\,\theta)\; P(x\,|\,e,\theta^{(n)})\; dx$

Then: $Q(\theta,\theta^{(n)}) = \int_x \ln \frac{P(x,e\,|\,\theta)}{P(x,e\,|\,\theta^{(n)})} P(x\,|\,e,\theta^{(n)})\; dx \;= Q(\theta|\theta^{(n)}) - Q(\theta^{(n)}|\theta^{(n)})$

For fixed $\theta^{(n)}$: Maximizing $Q(\theta,\theta^{(n)})$ is equivalent to maximizing $Q(\theta|\theta^{(n)})$

Residual: $L(\theta) = Q(\theta|\theta^{(n)}) + R(\theta|\theta^{(n)})$

- EM-phylosophy: Replace $\text{argmax}_{\theta\in\Theta} L(\theta)$ with $\text{argmax}_{\theta\in\Theta} Q(\theta|\theta^{(n)})$
- Can ignore $R(\theta|\theta^{(n)})$: Because $R(\theta|\theta^{(n)}) \geq R(\theta^{(n)}|\theta^{(n)})$

Given a current parameter estimate $\theta^{(n)}$:

- E-step: Form $Q(\theta|\theta^{(n)})$ which involves computing $P(x|e,\theta^{(n)})$
- M-step: Update $\theta^{(n)}$ by maximization, $\theta^{(n+1)} = \arg\max_{\theta\in\Theta} Q(\theta|\theta^{(n)})$

Generalization of M-step: replace mazimization with increase

EM Probabilistic Interpretation

By definition: $Q(\theta|\theta^{(n)}) = \int_x \ln P(x,e\,|\,\theta)\, P(x\,|\,e,\theta^{(n)})\,dx = E[\,\ln P(X,e\,|\,\theta)\,|\,e,\theta^{(n)}]$

$Q(\theta,\theta^{(n)})$ is an estimate: Of the complete-data log-likelihood built upon:

- The knowledge: Of the incomplete data
- Under the assumption: That the true parameter θ is known

Interpreting the residual:

$$R(\theta|\theta^{(n)}) = \log P(x,e\,|\,\theta) - \int_x \log P(x,e\,|\,\theta)\; P(x\,|\,e,\theta^{(n)})\; dx$$

$$= \int_x \log \frac{P(x,e\,|\,\theta)}{P(x,e\,|\,\theta^{(n)})} P(x\,|\,e,\theta^{(n)})\; dx \qquad \int_x P(x\,|\,e,\theta^{(n)})\; dx = 1$$

$$R(\theta|\theta^{(n)}) - R(\theta^{(n)}|\theta^{(n)}) = \underbrace{\int_x \log \frac{q_{\theta^{(n)}}(x)}{q_\theta(x)} q_{\theta^{(n)}}(x)\; dx}_{D(q_{\theta^{(n)}} \| q_\theta)\ \text{Kullback-Leibler distance}} \quad \text{with } q_\theta(x) = P(x,e\,|\,\theta)$$

EM Probabilistic Interpretation

Kulback-Leibler distance: Tool to asses the deviation between pdfs

- Not a genuine distance: It is not symmetric
- It is always positive
- It vanishes iff the pdfs are equal

What does it mean in our case?

- Perfect approximation: One in which $P(x \mid e,\theta)$ is independent of θ
- In other words we would like: $\theta \rightarrow E \rightarrow X$
- But we have already assumed: $\theta \rightarrow X \rightarrow E$
- Can we pemute X and E?: Not in general

Validity of $Q(\theta|\theta^{(n)})$ as local MLE:

- Controlled by amount of info that e and θ have about x

EM as a Fixed-Point

Quite clearly:

$$\theta^{(n+1)} = \Phi(\theta^{(n)}) \quad \text{with} \quad \Phi(\theta^{(n)}) = \arg\max_{\theta \in \Theta} Q(\theta, \theta^{(n)})$$

Assume that the sequence θ_n converges to $\hat{\theta}$:

- Hoepfully the maximum likelihood estimate
- But possibly some other local maximum or saddle point

Assuming that Φ is continuous:

- $\hat{\theta}$ must be a fix point of Φ: $\hat{\theta} = \Phi(\hat{\theta})$
- Can approximate Φ around $\hat{\theta}$ with a Taylor series

$$\theta^{(n+1)} \simeq \Phi(\hat{\theta}) + (\hat{\theta} - \theta^{(n)}) \frac{\partial \Phi(\theta)}{\partial \theta}(\hat{\theta})$$

EM for HMM

Given observation $\mathbf{e}_{0:T} = e_0,...,e_T$ Learn HMM $\theta^* = (\pi, T, O)$

Such that $\theta^* = \arg\max_{\theta \in \Theta} \log P(\mathbf{e}_{0:T} \mid \theta)$

- Key idea: Introduce hidden RVs $\mathbf{X}_{0:T} = X_0,...,X_T$, iid
- Estimate: $\arg\max_{\theta \in \Theta} Q(\theta \mid \theta^{(n)})$

The expectation-maximization algorithm has two steps:

- E-step: $Q(\theta \mid \theta^{(n)}) = \sum_{\mathbf{x}_{0:T} \in \mathbf{X}_{0:T}} \log(P(\mathbf{x}_{0:T}, \mathbf{e}_{0:T} \mid \theta))\, P(\mathbf{x}_{0:T} \mid \mathbf{e}_{0:T}, \theta^{(n)})$
- M-step: $\theta^{(n+1)} = \arg\max_{\theta \in \Theta} Q(\theta \mid \theta^{(n)})$

$P(\mathbf{x}_{0:T}, \mathbf{e}_{0:T} \mid \theta^{(n)}) = P(\mathbf{x}_{0:T} \mid \mathbf{e}_{0:T}, \theta^{(n)})\, P(\mathbf{e}_{0:T} \mid \theta^{(n)})$ and $P(\mathbf{e}_{0:T} \mid \theta^{(n)})$ not affected by θ choice:

- E-step: $\hat{Q}(\theta \mid \theta^{(n)}) = \sum_{\mathbf{x}_{0:T} \in \mathbf{X}_{0:T}} \log(P(\mathbf{x}_{0:T}, \mathbf{e}_{0:T} \mid \theta))\, P(\mathbf{x}_{0:T}, \mathbf{e}_{0:T} \mid \theta^{(n)})$
- M-step: $\theta^{(n+1)} = \arg\max_{\theta \in \Theta} \hat{Q}(\theta \mid \theta^{(n)})$

E-Step for HMM

$$\hat{Q}(\theta|\theta^{(n)}) = \sum_{\mathbf{x}_{0:T} \in \mathbf{X}_{0:T}} P(\mathbf{x}_{0:T}, \mathbf{e}_{0:T} \mid \theta^{(n)}) \log P(\mathbf{x}_{0:T}, \mathbf{e}_{0:T} \mid \theta)$$

$$P(\mathbf{x}_{0:T}, \mathbf{e}_{0:T} \mid \theta) = (\pi_{x_0} \times O_{x_0, e_0}) \prod_{t=0}^{T-1} (T_{x_t, x_{t+1}} \times O_{x_{t+1}, e_{t+1}})$$

$$\log P(\mathbf{x}_{0:T}, \mathbf{e}_{0:T} \mid \theta) = \log \pi_{x_0} + \sum_{t=0}^{T-1} \log T_{x_t, x_{t+1}} + \sum_{t=0}^{T} \log O_{x_t, e_t}$$

$$\begin{aligned}
\hat{Q}(\theta|\theta^{(n)}) = & \sum_{\mathbf{x}_{0:T} \in \mathbf{X}_{0:T}} P(\mathbf{x}_{0:T}, \mathbf{e}_{0:T} \mid \theta^{(n)}) \log \pi_{x_0} \\
& + \sum_{\mathbf{x}_{0:T} \in \mathbf{X}_{0:T}} P(\mathbf{x}_{0:T}, \mathbf{e}_{0:T} \mid \theta^{(n)}) \sum_{t=0}^{T-1} \log T_{x_t, x_{t+1}} \\
& + \sum_{\mathbf{x}_{0:T} \in \mathbf{X}_{0:T}} P(\mathbf{x}_{0:T}, \mathbf{e}_{0:T} \mid \theta^{(n)}) \sum_{t=1}^{T} \log O_{x_t, e_t}
\end{aligned}$$

Marginalization Trick for π

$$\sum_{\mathbf{x}_{0:T} \in \mathbf{X}_{0:T}} P(\mathbf{x}_{0:T}, \mathbf{e}_{0:T} \mid \theta^{(n)}) \log \pi_{x_0}$$
$$=$$
$$\sum_{x_0 \in X, x_1 \in X \ldots, x_T \in X} P(x_0, x_1 \ldots, x_T, \mathbf{e}_{0:T} \mid \theta^{(n)}) \log \pi_{x_0}$$
$$=$$
$$\sum_{x_0 \in X, x_2 \in X \ldots, x_T \in X} P(x_0, x_2 \ldots, x_T, \mathbf{e}_{0:T} \mid \theta^{(n)}) \log \pi_{x_0}$$
$$=$$
$$\sum_{x_0 \in X, x_T \in X} P(x_0, x_T, \mathbf{e}_{0:T} \mid \theta^{(n)}) \log \pi_{x_0}$$
$$=$$
$$\sum_{x_0 \in X} P(x_0, \mathbf{e}_{0:T} \mid \theta^{(n)}) \log \pi_{x_0}$$
$$=$$
$$\sum_{m=1}^{M} P(X_0 = m, \mathbf{e}_{0:T} \mid \theta^{(n)}) \log \pi_m$$

Marginalization Trick for T_{ij}

$$\sum_{\mathbf{x}_{0:T} \in \mathbf{X}_{0:T}} P(\mathbf{x}_{0:T}, \mathbf{e}_{0:T} \mid \theta^{(n)}) \sum_{t=0}^{T-1} \log T_{x_t, x_{t+1}}$$

$$= \sum_{\mathbf{x}_{0:T} \in \mathbf{X}_{0:T}} \sum_{t=0}^{T-1} P(\mathbf{x}_{0:T}, \mathbf{e}_{0:T} \mid \theta^{(n)}) \log T_{x_t, x_{t+1}}$$

$$= \sum_{t=0}^{T-1} \sum_{\mathbf{x}_{0:T} \in \mathbf{X}_{0:T}} P(\mathbf{x}_{0:T}, \mathbf{e}_{0:T} \mid \theta^{(n)}) \log T_{x_t, x_{t+1}}$$

$$= \sum_{t=0}^{T-1} \sum_{x_0 \in X, \ldots, x_T \in X} P(x_0, \ldots, x_T, \mathbf{e}_{0:T} \mid \theta^{(n)}) \log T_{x_t, x_{t+1}}$$

$$= \sum_{t=0}^{T-1} \sum_{x_t \in X, x_{t+1} \in X} P(x_t, x_{t+1}, \mathbf{e}_{0:T} \mid \theta^{(n)}) \log T_{x_t, x_{t+1}}$$

$$= \sum_{t=0}^{T-1} \sum_{m=1}^{M} \sum_{n=1}^{M} P(x_t = m, x_{t+1} = n, \mathbf{e}_{0:T} \mid \theta^{(n)}) \log T_{m,n}$$

$$= \sum_{m=1}^{M} \sum_{n=1}^{M} \sum_{t=0}^{T-1} P(x_t = m, x_{t+1} = n, \mathbf{e}_{0:T} \mid \theta^{(n)}) \log T_{m,n}$$

Marginalization Trick for O_{ij}

$$\sum_{\mathbf{x}_{0:T} \in \mathbf{X}_{0:T}} P(\mathbf{x}_{0:T}, \mathbf{e}_{0:T} \mid \theta^{(n)}) \sum_{t=0}^{T} \log O_{x_t, e_t}$$

$$= \sum_{\mathbf{x}_{0:T} \in \mathbf{X}_{0:T}} \sum_{t=0}^{T-1} P(\mathbf{x}_{0:T}, \mathbf{e}_{0:T} \mid \theta^{(n)}) \log O_{x_t, e_t}$$

$$= \sum_{t=0}^{T-1} \sum_{\mathbf{x}_{0:T} \in \mathbf{X}_{0:T}} P(\mathbf{x}_{0:T}, \mathbf{e}_{0:T} \mid \theta^{(n)}) \log O_{x_t, e_t}$$

$$= \sum_{t=0}^{T-1} \sum_{x_0 \in X, \ldots, x_T \in X} P(x_0, \ldots, x_T, \mathbf{e}_{0:T} \mid \theta^{(n)}) \log O_{x_t, e_t}$$

$$= \sum_{t=0}^{T-1} \sum_{x_t \in X} P(x_t, \mathbf{e}_{0:T} \mid \theta^{(n)}) \log O_{x_t, e_t}$$

$$= \sum_{t=0}^{T-1} \sum_{m=1}^{M} \sum_{n=1}^{N} P(x_t = m, \mathbf{e}_{0:T} \mid \theta^{(n)}) \, 1(e_t = n) \log O_{m,n}$$

$$= \sum_{m=1}^{M} \sum_{n=1}^{M} \sum_{t=0}^{T-1} P(x_t = m, \mathbf{e}_{0:T} \mid \theta^{(n)}) \, 1(e_t = n) \log O_{m,n}$$

M-Step for HMM

Maximize $Q(\theta|\theta^{(n)})$ subject to the constraints:

$$\sum_{i=1}^{n} \pi_i = 1, \quad \forall i \in \{1,...,M\}. \sum_{j=1}^{M} T_{ij} = 1, \quad \forall i \in \{1,...,M\}. \sum_{i=1}^{N} O_{ij} = 1,$$

Use the Lagrange Multipliers method:

$$LM(\theta|\theta^{(n)}) = \hat{Q}(\theta|\theta^{(n)}) - \lambda_{\pi}(\sum_{i=1}^{M} \pi_i - 1) - \sum_{i=1}^{M} \lambda_{T_i}(\sum_{j=1}^{M} T_{ij} - 1) - \sum_{i=1}^{M} \lambda_{O_i}(\sum_{j=1}^{N} O_{ij} - 1)$$

Basic idea:

- Compute: $\partial LM(\theta|\theta^{(n)})/\partial \pi_i = 0$ and $\partial LM(\theta|\theta^{(n)})/\partial \lambda_{\pi} = 0$. Solve.
- Compute: $\partial LM(\theta|\theta^{(n)})/\partial T_{ij} = 0$ and $\partial LM(\theta|\theta^{(n)})/\partial \lambda_{T_i} = 0$. Solve.
- Compute: $\partial LM(\theta|\theta^{(n)})/\partial O_{ij} = 0$ and $\partial LM(\theta|\theta^{(n)})/\partial \lambda_{O_i} = 0$. Solve.

Maximize for π

$$\frac{\partial LM(\theta|\theta^{(n)})}{\partial \pi_i} = \frac{\partial \sum_{\mathbf{x}_{0:T} \in \mathbf{X}_{0:T}} P(\mathbf{x}_{0:T}, \mathbf{e}_{0:T} \mid \theta^{(n)}) \log \pi_{x_0}}{\partial \pi_i} - \lambda_\pi = 0$$

$$= \frac{\partial \sum_{m=1}^{M} P(X_0 = m, \mathbf{e}_{0:T} \mid \theta^{(n)}) \log \pi_m}{\partial \pi_i} - \lambda_\pi = 0$$

$$= \frac{P(X_0 = i, \mathbf{e}_{0:T} \mid \theta^{(n)})}{\pi_i} - \lambda_\pi = 0$$

$$\frac{\partial LM(\theta|\theta^{(n)})}{\partial \lambda_i} = -(\sum_{m=1}^{M} \pi_m - 1) = 0$$

Maximize for π

$$\pi_i = P(X_0 = i, \mathbf{e}_{0:T} \mid \theta^{(n)}) / \lambda_\pi$$

$$\sum_{m=1}^{M} P(X_0 = m, \mathbf{e}_{0:T} \mid \theta^{(n)})) / \lambda_\pi - 1 = 0$$

Hence:

$$\lambda_\pi = \sum_{m=1}^{M} P(X_0 = m, \mathbf{e}_{0:T} \mid \theta^{(n)}) = 1$$

$$\pi_i = P(X_0 = i, \mathbf{e}_{0:T} \mid \theta^{(n)}) = \mathbf{s}_{0,i} \quad \forall i = 1:M$$

Maximize for T_{ij}

$$\frac{\partial LM(\theta|\theta^{(n)})}{\partial T_{ij}} = \frac{\partial \sum_{\mathbf{x}_{0:T} \in \mathbf{X}_{0:T}} \sum_{t=0}^{T-1} \log T_{x_t, x_{t+1}} P(\mathbf{x}_{0:T}, \mathbf{e}_{0:T} \mid \theta^{(n)})}{\partial T_{ij}} - \lambda_{T_i} = 0$$

$$= \frac{\partial \sum_{m=1}^{M} \sum_{n=1}^{M} \sum_{t=0}^{T-1} \log T_{m,n} \; P(x_t = m, x_{t+1} = n, \mathbf{e}_{0:T} \mid \theta^{(n)})}{\partial T_{ij}} - \lambda_{T_i} = 0$$

$$= \frac{\sum_{t=0}^{T-1} P(x_t = i, x_{t+1} = j, \mathbf{e}_{0:T} \mid \theta^{(n)})}{T_{ij}} - \lambda_{T_i} = 0$$

$$\frac{\partial LM(\theta|\theta^{(n)})}{\partial \lambda_{T_i}} = -(\sum_{j=1}^{M} T_{ij} - 1) = 0$$

Maximize for T_{ij}

$$T_{ij} = \frac{\sum_{t=0}^{T-1} P(x_t = i, x_{t+1} = j, \mathbf{e}_{0:T} \mid \theta^{(n)})}{\lambda_{T_i}} \qquad \sum_{j=1}^{M} T_{ij} = 1$$

$$\lambda_{T_i} = \sum_{j=1}^{M} \sum_{t=0}^{T-1} P(x_t = i, x_{t+1} = j, \mathbf{e}_{0:T} \mid \theta^{(n)})$$

Hence:

$$T_{ij} = \frac{\sum_{t=0}^{T-1} P(x_{t-1} = i, x_t = j, \mathbf{e}_{0:T} \mid \theta^{(n)})}{\sum_{j=1}^{M} \sum_{t=0}^{T-1} P(x_t = i, x_{t+1} = j, \mathbf{e}_{0:T} \mid \theta^{(n)})}$$

$$= \frac{\sum_{t=0}^{T-1} P(x_t = i, x_{t+1} = j, \mathbf{e}_{0:T} \mid \theta^{(n)})}{\sum_{t=0}^{T-1} P(x_t = i, \mathbf{e}_{0:T} \mid \theta^{(n)})} = \frac{\sum_{t=0}^{T-1} \mathbf{A}_{i,j,t}}{\sum_{t=0}^{T-1} \mathbf{s}_{i,t}} \quad \forall i, j = 1:N$$

Maximize for O_{ij}

$$\frac{\partial LM(\theta|\theta^{(n)})}{\partial O_{ij}} = \frac{\partial \sum_{\mathbf{x}_{0:T} \in \mathbf{x}_{1:T}} \sum_{t=0}^{T} P(\mathbf{x}_{1:T}, \mathbf{e}_{1:T} \mid \theta^{(n)}) \log O_{x_t, e_t}}{\partial O_{ij}} - \lambda_{\pi} = 0$$

$$= \frac{\partial \sum_{m=1}^{M} \sum_{n=1}^{N} \sum_{t=0}^{T} P(x_t = m, \mathbf{e}_{1:T} \mid \theta^{(n)})\ 1(e_t = n)\ \log O_{m,n}}{\partial O_{ij}} - \lambda_{T_i} = 0$$

$$= \frac{\sum_{t=0}^{T} P(x_t = i, \mathbf{e}_{1:T} \mid \theta^{(n)})\ 1(e_t = j)}{O_{ij}} - \lambda_{O_i} = 0$$

$$\frac{\partial LM(\theta|\theta^{(n)})}{\partial \lambda_{O_i}} = -(\sum_{j=1}^{N} O_{ij} - 1) = 0$$

Maximize for O_{ij}

$$O_{ij} = \frac{\sum_{t=0}^{T} P(x_t = i, \mathbf{e}_{1:T} \mid \theta^{(n)})\, 1(e_t = j)}{\lambda_{O_i}} \qquad \sum_{j=1}^{N} O_{ij} = 1$$

$$\lambda_{O_i} = \sum_{j=1}^{N} \sum_{t=0}^{T} P(x_t = i, \mathbf{e}_{1:T} \mid \theta^{(n)})\, 1(e_t = j)$$

Hence:

$$O_{ij} = \frac{\sum_{t=0}^{T} P(x_t = i, \mathbf{e}_{1:T} \mid \theta^{(n)})\, 1(e_t = j)}{\sum_{j=1}^{N} \sum_{t=0}^{T} P(x_t = i, \mathbf{e}_{1:T} \mid \theta^{(n)})\, 1(e_t = j)}$$

$$= \frac{\sum_{t=0}^{T} P(x_t = i, \mathbf{e}_{0:T} \mid \theta^{(n)})\, 1(e_t = j)}{\sum_{t=0}^{T} P(x_t = i, \mathbf{e}_{0:T} \mid \theta^{(n)})} = \frac{\sum_{t=0}^{T} \mathbf{s}_{i,t} 1_{e_t = j}}{\sum_{t=0}^{T} \mathbf{s}_{i,t}} \quad \forall i = 1{:}M,\ j = 1{:}N$$

M-Step in Summary

$$\theta^{(n+1)} = \arg\max_{\theta \in \Theta} Q(\theta|\theta^{(n)}) \quad \text{where} \quad \theta^{(n)} = (\pi^{(n)}, T^{(n)}, O^{(n)})$$

$$\pi_i^{(n+1)} = P(x_0 = i, \mathbf{e}_{0:T} \mid \theta^{(n)}) = \mathbf{s}_{0,i} \quad \forall i = 1:M$$

$$T_{ij}^{(n+1)} = \frac{\sum_{t=0}^{T-1} P(x_t = i, x_{t+1} = j, \mathbf{e}_{0:T} \mid \theta^{(n)})}{\sum_{t=0}^{T-1} P(x_t = i, \mathbf{e}_{0:T} \mid \theta^{(n)})} = \frac{\sum_{t=0}^{T-1} \mathbf{A}_{i,j,t}}{\sum_{t=0}^{T-1} \mathbf{s}_{i,t}} \quad \forall i, j = 1:M$$

$$O_{ij}^{(n+1)} = \frac{\sum_{t=0}^{T} P(x_t = j, \mathbf{e}_{0:T} \mid \theta^{(n)}) \, 1(e_t = j)}{\sum_{t=0}^{T} P(x_t = i, \mathbf{e}_{0:T} \mid \theta^{(n)})} = \frac{\sum_{t=0}^{T} 1_{e_t=j} \mathbf{s}_{i,t}}{\sum_{t=0}^{T} \mathbf{s}_{i,t}} \quad \forall i = 1:M, \; j = 1:N$$