Mathematical foundations for ML - UM6P class

Exploring Latent Variable Models

February 2, 2025

Contents

1 Introducing the Expectation Maximization algorithm

1.1 Introducing the context

The Expectation-Maximization (EM) algorithm is an iterative method used for obtaining maximum likelihood estimates of parameters within statistical models. These models are characterized by their reliance on unobserved latent variables or hidden variables. Latent variables, denoted as z, are not directly observed but are inferred through the variables that are observed, denoted as x.

Within the context of the EM algorithm, we operate under the following framework:

- **Assumption:** We consider (x, z) to be random variables, where x represents the observed data, and z represents the hidden or latent variables (for example, unknown cluster centers in a clustering problem). The joint density function of x and z, $p_\theta(x, z)$, is parameterized by θ, indicating the model's parameters.
- **Objective:** The primary goal is to maximize the marginal likelihood of the observed data x with respect to the parameters θ, expressed as:

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

This objective highlights the challenge posed by the presence of latent variables: maximizing the marginal likelihood is not straightforward due to the summation over the latent variable z. The summation introduces complexities, making the problem more challenging than optimizing a likelihood function without latent variables.

Specifically, taking the logarithm of the marginal likelihood does not lead to a simple convex optimization problem. The EM algorithm provides a robust method for addressing this challenge, facilitating the estimation of model parameters in the presence of latent variables.

1.2 The EM algorithm

The Dataset is composed of the pairs $(x_i, z_i)_{1 \leq i \leq n}$ where x_i is the observed data and z_i is the hidden data.

We make the assumption that the $(x_i, z_i)_{1\leq i\leq n}$ are i.i.d.

The aim is to maximize the log likelihood:

$$\log p_\theta(x) = \sum_{i=1}^{n} \log \sum_{z_i} p_\theta\left(x_i, z_i\right)$$

We will use the following properties :

Proposition 1.2.1. *Jensen Inequality:*

1. *if $f : \mathbb{R} \to \mathbb{R}$ is convex and if X is an integrable random variable :*

$$\mathbb{E}_X(f(X)) \geq f\left(\mathbb{E}_X(X)\right)$$

2. *if $f : \mathbb{R} \to \mathbb{R}$ is strictly convex, we have equality in the previous inequality if and only if $X =$ constant a.s.*

The EM algorithm is an iterative method for finding maximum likelihood estimates of parameters in statistical models, where the models depend on unobserved latent variables.

Consider for instance n observations $x_1, \dots, x_n$ and the latent variables associateed with them $z_1, \dots, z_n$.

We assume the pairs (x_i, z_i) to be independent and identically distributed.

For $(x, z) = (x_1, z_1, \dots, x_n, z_n)$, the objective is to maximize:

$$\log(p(x;\theta)) = \sum_{i=1}^{n} \log\left(\sum_{z_i} p(x_i, z_i; \theta)\right)$$

For each $i \in \{1, \dots, n\}$, we introduce a function $z_i \mapsto q(z_i)$ such that $q(z_i) \geq 0$ and $\sum_{z_i} q(z_i) = 1$ in the expression of the likelihood.

By conditioning on a latent variable z_i and using the Jensen inequality, we get a lower bound $\mathcal{L}(q, \theta)$ that depends on both q and θ.

$$
\begin{aligned}
\log(p(x;\theta)) &= \sum_{i=1}^{n} \log\left(\sum_{z_i} p(x_i, z_i; \theta)\right) \\
&= \sum_{i=1}^{n} \log\left(\sum_{z_i} q(z_i)\frac{p(x_i, z_i; \theta)}{q(z_i)}\right) \\
&\geq \sum_{i=1}^{n} \sum_{z_i} q(z_i) \log\left(\frac{p(x_i, z_i; \theta)}{q(z_i)}\right) \\
&= \sum_{i=1}^{n} \underbrace{\mathbb{E}_{q(z_i)}\left[\log\left(\frac{p_\theta(x_i, z_i)}{q(z_i)}\right)\right]}_{\mathcal{L}(q(z_i),\theta)}
\end{aligned}
$$

The EM algorithm can then be summarized as depicted in 1.

Algorithm 1 EM Algorithm

```
Require: Data set $X = \{x_1, \dots, x_n\}$
Ensure: Optimal $\theta$

Initialization: Choose initial parameters $\theta^{(0)}$.
Set iteration counter $i = 0$.
while not converged do
    E-step: Update $q$ to maximize the lower bound with respect to $q$.
        $q_{t+1} \in \arg\max_{q} (\mathcal{L}(q, \theta_t))$
    M-step: Update $\theta$ to maximize the lower bound with respect to $\theta$.
        $\theta_{t+1} \in \arg\max_{\theta} (\mathcal{L}(q_{t+1}, \theta))$
    Check for convergence criterion (e.g., change in $\theta$ below a threshold).
    $i \leftarrow i + 1$
end while
return Optimized parameters $\theta^*$.
```

Exercise:

Show that the gap beween the marginal log-likelihood and the **lower bound** $\sum_{i=1}^{n} \mathcal{L}(q(z_i), \theta)$ is reduced to 0 when $q(z_i) = p_\theta(z_i \mid x_i)\ \forall i \in \{1, \dots, n\}$.

$p_\theta(z_i \mid x_i)$ is called the **posterior distribution**

Solution: Let $d = \log(p_\theta(x)) - \sum_{i=1}^{n} \mathcal{L}(q(z_i), \theta)$.

We have:

$$\begin{aligned}
\mathrm{d} &= \log(p_\theta(x)) - \sum_{i=1}^{n} \mathcal{L}(q(z_i), \theta) \\
&= \sum_{i=1}^{n} \left(\log(p_\theta(x_i)) - \mathcal{L}(q(z_i), \theta)\right) \\
&= \sum_{i=1}^{n} \left(\sum_{z_i} q(z_i) \log(p_\theta(x_i)) - \sum_{z_i} q(z_i) \log\left(\frac{p_\theta(x_i, z_i)}{q(z_i)}\right) \right) \\
&= \sum_{i=1}^{n} \sum_{z_i} q(z_i) \left(\log(p_\theta(x_i)) - \log\left(\frac{p_\theta(x_i, z_i)}{q(z_i)}\right) \right) \\
&= \sum_{i=1}^{n} \sum_{z_i} q(z_i) \log\left(\frac{q(z_i)}{p_\theta(z_i|x_i)}\right) \\
&= \sum_{i=1}^{n} D_{\mathrm{KL}}\left(q(z_i) \,\|\, p_\theta(z_i|x_i)\right)
\end{aligned}$$

Therefore,

$$\begin{aligned}
d = 0 &\iff \sum_{i=1}^{n} \underbrace{D_{\mathrm{KL}}\left(q(z_i) \,\|\, p_\theta(z_i|x_i)\right)}_{\geq 0} \\
&\iff \forall i \in \{1, \ldots, n\} \quad D_{\mathrm{KL}}\left(q(z_i) \,\|\, p_\theta(z_i|x_i)\right) \\
&\iff \forall i \in \{1, \ldots, n\} \quad q(z_i) = p_\theta(z_i|x_i)
\end{aligned}$$

Therefore, maximizing the lower bound $\log(p_\theta(x))$ with respect to q consists in taking the posterior distributions $\forall i \in \{1, \ldots, n\} \quad q(z_i) = p_\theta(z_i|x_i)$.

Let's recall the expression of the lower bound:

$$\mathcal{L}(q,\theta) = \sum_{i=1}^{n} \left(\sum_{z_i} q(z_i) \log p_\theta(x_i, z_i) - \sum_{z_i} q(z_i) \log q(z_i) \right)$$

Since $\sum_{z_i} q(z_i) \log q(z_i)$ doesn't depend on θ, maximizing the lower bound with respect to θ is equivalent to maximizing w.r.t θ the expected value of the complete log likelihood function $\log\left(p_{\theta_t}(x,z)\right)$.

The final recipe is given in algorithm 5. It consists in the following steps:

1. Compute the probability of Z given X : $p_{\theta_t}(z \mid x)$ (Corresponding to $q_{t+1} = \arg\max_q \mathcal{L}\left(q, \theta_t\right)$)
2. Write the complete loglikelihood $l_c = \log\left(p_{\theta_t}(x,z)\right)$.
3. **E-Step**: Calculate the expected value of the complete log likelihood function, with respect to the conditional distribution of Z given X under the current estimate of the parameter $\theta_t : \mathbb{E}_{Z|X}\left(l_c\right)$.
4. **E-Step**: Find θ_{t+1} by maximizing $\mathcal{L}\left(q_{t+1}, \theta\right)$ with respect to θ.

Algorithm 2 EM algorithm

```
Require: Observations $x_1, \dots, x_n$
Ensure: Optimal $\theta$

Initialize $\theta^{(0)}$
while not converged do
    E-step: $q(z) = p(z|x; \theta^{(i-1)})$
    M-step: $\theta^{(i)} = \arg\max_\theta \mathbb{E}_q[\log p(x, z; \theta)]$
    $i \leftarrow i + 1$
end while
```

Remarks:

- It is an ascent algorithm, indeed it goes up in term of likelihood (compare to before where we were descending along the distortion).
- The sequence of log-likelihoods converges.
- It does not converge to a global maximum but rather to a local maximum because we are dealing here with a non-convex problem. An illustration is given in Figure 1

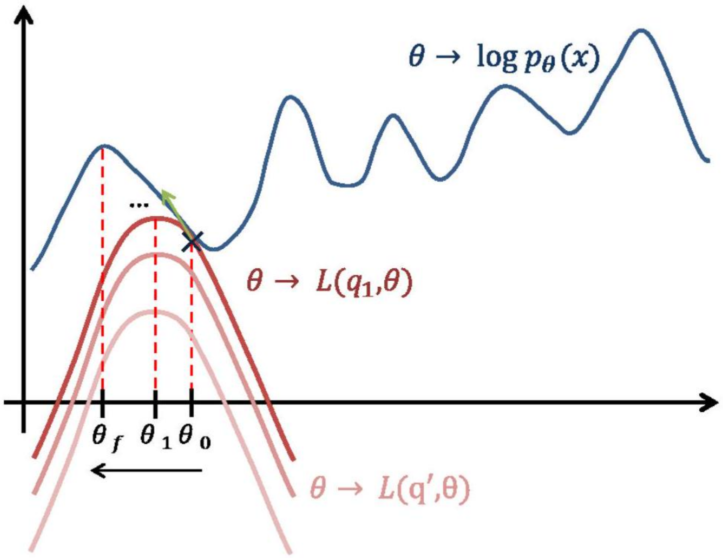
$\theta \rightarrow \log p_\theta(x)$
...
$\theta \rightarrow L(q_1,\theta)$
θ_f
θ_1
θ_0
$\theta \rightarrow L(q',\theta)$

Figure 1: An illustration of the EM algorithm that converges to a local minimum.

- As it was already the case for K-means, we reiterate the result in order to be more confident. Then we keep the one with the highest likelihood.
- Because EM gives a local maximum, it is clever to choose a θ_0 relatively close to the final solution. For Gaussian mixtures, it is quite usual to initiate EM by a K-means.

1.3 Estimating the parameters of a Gaussian Mixture Model using the EM algorithm

Gaussian Mixture Models (GMMs) are a probabilistic model for representing normally distributed subpopulations within an overall population. Unlike single Gaussian models, which assume that all observations are drawn from a single distribution, GMMs consider a mixture of several Gaussian distributions, each with its own mean and variance, thus providing a more flexible approach to modeling data distributions. This flexibility makes GMMs particularly useful for modeling complex data sets with hidden or latent variables—where observations may originate from one of several unknown subpopulations.

Let's present a simple example to illustrate what we just said. The probability density represented on Figure 2 is akin to an average of two Gaussians. Thus, it is natural to use a mixture model and to introduce an hidden variable z, following a Bernoulli distribution defining which Gaussian the point is sampled from.

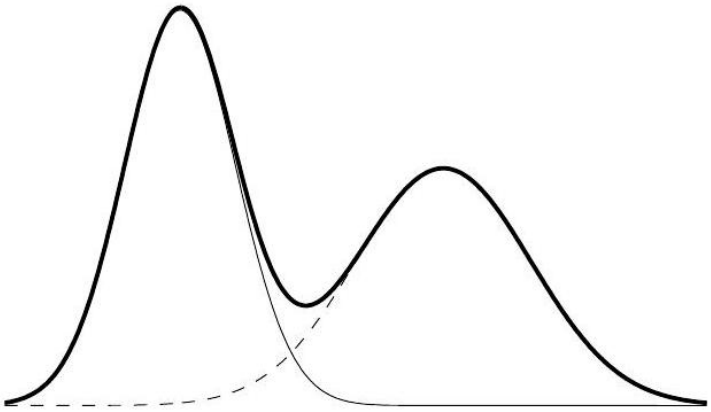

Figure 2: Average of two probability distributions of two Gaussian for which it is natural to introduce a mixture model

Thus we have : $z \in \{1, 2\}$ and $x \mid z = i \sim \mathcal{N}(\mu_i, \Sigma_i)$. The density $p(x)$ is a convex combination of normal density:

$$p(x) = p(x, z = 1) + p(x, z = 2) = p(x \mid z = 1)p(z = 1) + p(x \mid z = 2)p(z = 2)$$

It is a mixture model. It represents a simple way to model complicated phenomena.

Exercise:

Suppose we have observations n observations $x_1, \dots x_n$ in $\mathbb{R}^p$.

We make the assumption of the existence of latent variables $z_1, \dots, z_n$ from a multinomial distribution with K possible outcomes.

i.e:

$\forall i \in \{1, \dots, n\} \quad x_i \in \mathbb{R}^p, z_i \sim \mathcal{M}(1, \pi_1, \dots, \pi_K)$ and $(x_i \mid z_i = j) \sim \mathcal{N}(\mu_j, \Sigma_j)$.

Here we have $\theta = (\pi, \mu, \Sigma)$.

Use the EM algorithm to estimate θ.

Solution:

1. **Calculation of the posterior distributions** $p_\theta(z_i \mid x_i)$:

   We write $p_\theta\left(x_i\right)$:

$$
\begin{aligned}
p_\theta\left(x_i\right) &= \sum_{z_i} p_\theta\left(x_i, z_i\right) = \sum_{z_i} p_\theta\left(x_i \mid z_i\right) p_\theta\left(z_i\right) \\
&= \sum_{j=1}^{K} p_\theta\left(x_i \mid z_i = j\right) p_\theta\left(z_i = j\right)
\end{aligned}
$$

   Then we use the Bayes formula to estimate $p_\theta(z \mid x)$:

$$
\begin{aligned}
p_\theta\left(z_i = j \mid x_i\right) &= \frac{p_\theta\left(x_i \mid z_i = j\right) p_\theta\left(z_i = j\right)}{p_\theta\left(x_i\right)} \\
&= \frac{\pi_j \mathcal{N}\left(x_i \mid \mu_j, \Sigma_j\right)}{\sum_{j'} \pi_{j'} \mathcal{N}\left(x_i \mid \mu_j', \Sigma_j'\right)} \\
&= \tau_i^j(\theta).
\end{aligned}
$$

   We recall that $\mathcal{N}\left(x_i \mid \mu, \Sigma\right) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$.

   Suppose that we are at the t-th iteration of the algorithm.

2. **Complete likelihood**

   Let's write the complete likelihood of the problem.

$$
\begin{aligned}
l_{c,t} = \log p_{\theta_t}(x,z) &= \sum_{i=1}^{n} \log p_{\theta_t}\left(x_i, z_i\right) \\
&= \sum_{i=1}^{n} \log\left(p_{\theta_t}\left(z_i\right) p_{\theta_t}\left(x_i \mid z_i\right)\right) \\
&= \sum_{i=1}^{n} \log\left(p_{\theta_t}\left(z_i\right)\right) + \log\left(p_{\theta_t}\left(x_i \mid z_i\right)\right) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{K} z_i^j \log\left(\pi_{j,t}\right) \\
&\quad + \sum_{i=1}^{n} \sum_{j=1}^{K} z_i^j \log\left(\mathcal{N}\left(x_i \mid \mu_{j,t}, \Sigma_{j,t}\right)\right)
\end{aligned}
$$

where $z_i^j \in \{0,1\}$ with $z_i^j = 1$ if $z_i = j$ and 0 otherwise.

3. **E-Step** In the E-step, we compute the expectation of the complete log-likelihood with respect to the conditional distribution of the latent variables Z given the observed data X. This involves replacing the indicator variables z_i^j with their expected values:

$$
\mathbb{E}_{Z|X}\left(z_i^j\right) = p_{\theta_t}(z = j | x_i) = \tau_i^j(\theta_t),
$$

where τ_i^j represents the posterior probability that observation x_i belongs to component j, given the current parameter estimates. By substituting z_i^j with τ_i^j, we obtain the expected complete log-likelihood:

$$
\mathbb{E}_{Z|X}(l_{c,t}) = \sum_{i=1}^{n} \sum_{j=1}^{K} \tau_i^j \log(\pi_{j,t}) + \sum_{i=1}^{n} \sum_{j=1}^{K} \tau_i^j \log\left(\mathcal{N}(x_i | \mu_{j,t}, \Sigma_{j,t})\right).
$$

4. **M-Step**

For the M-step, we this need to maximize:

$$
\sum_{i=1}^{n} \sum_{j=1}^{K} \tau_i^j \log\left(\pi_{j,t}\right) + \sum_{i=1}^{n} \sum_{j=1}^{K} \tau_i^j \left[\log\left(\frac{1}{(2\pi)^{\frac{p}{2}}}\right) + \log\left(\frac{1}{|\Sigma_{j,t}|^{\frac{1}{2}}}\right) \right.
$$
$$
\left. -\frac{1}{2}\left(x_i - \mu_{j,t}\right)^T \Sigma_{j,t}^{-1}\left(x_i - \mu_{j,t}\right)\right]
$$

We want to maximize the previous equation with respect to $\theta_t = (\pi_t, \mu_t, \Sigma_t)$

As the sum is separated into two terms independent along the variables we can first maximize with respect to π_t :

$$\max_{\pi} \sum_{j=1}^{K} \sum_{i=1}^{n} \tau_i^j \log \pi_j \quad \Rightarrow \quad \pi_{j,t+1} = \frac{\sum\limits_{i=1}^{n} \tau_i^j}{\sum_{i=1}^{n} \sum\limits_{j'=1}^{K} \tau_i^{j'}} = \frac{1}{n} \sum_{i=1}^{n} \tau_i^j$$

We can now maximize with respect to μ_t and Σ_t. By computing the gradient along the $\mu_{j,t}$ and along the $\Sigma_{j,t}$, we obtain :

$$\mu_{j,t+1} = \frac{\sum_i \tau_i^j x_i}{\sum_i \tau_i^j}$$

$$\Sigma_{j,t+1} = \frac{\sum_i \tau_i^j \left(x_i - \mu_{j,t+1}\right) \left(x_i - \mu_{j,t+1}\right)^T}{\sum_i \tau_i^j}$$

The M-step in the EM algorithm corresponds to the estimation of means step in K-means. Note that the value of τ_i^j in the expressions above are taken for the parameter values of the previous iterate, i.e., $\tau_i^j = \tau_i^j\left(\theta_t\right)$.

Possible forms for Σ_j

- isotropic: $\Sigma_j = \sigma_j^2 \mathrm{Id}, 1$ parameter, the cluster is a sphere.
- diagonal: Σ_j is a diagonal matrix, p parameters, the cluster is an ellipse oriented along the axis.
- general: Σ_j, $\frac{p(p+1)}{2}$ parameters, the cluster is an ellipse.

2 Hidden Markov Models

2.1 Introduction

We denote $(\tilde{X}_1, \ldots, \tilde{X}_T)$ a sequence of vectors in $\mathbb{R}^d$.

The objective of this section is to model the dynamics of the sequence $(\tilde{X}_1, \ldots, \tilde{X}_T)$ using a Hidden Markov Model (HMM) with m possible hidden states.

Section 2.2 gives a brief description of the parameterization of the HMM graphical model. Section 2.3 deals with the inference problem and introduces the filtering and smoothing probabilities. In section 2.5, we present the learning process using the Expectation Maximization algorithm. Finally, in section 2.6, we predict the distribution of the next hidden states given a sequence of observations.

2.2 The parameterization of the graphical model

The hidden state at time t is denoted by H_t and the observation at time t by $\tilde{X}_t$. Let us assume there are M possible hidden states and that the observations are continuous in $\mathbb{R}^d$. Let us also suppose we have T continuous observations $(\tilde{X}_t)_{1\leq T} \in \mathbb{R}^{T\times d}$

Figure 3 shows the difference between the graphical representation associated with the vanilla Gaussian mixture model introduced in section 1.3 and the graphical model associated with the HMM.

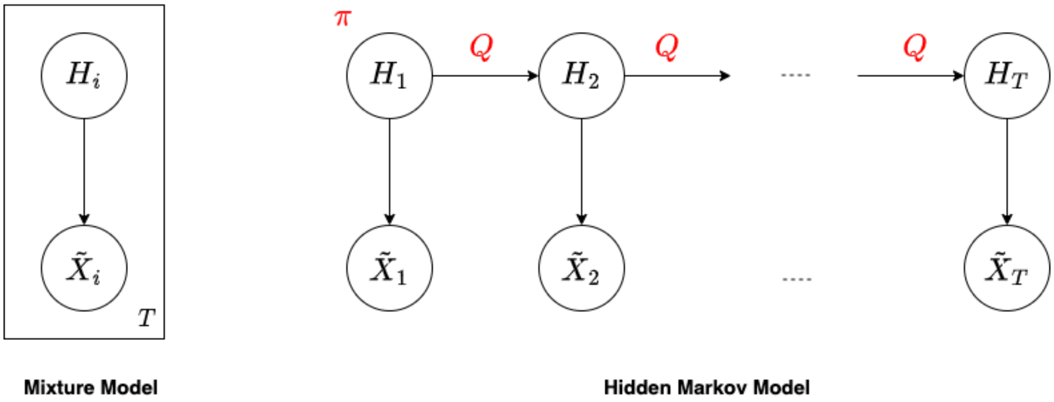

Figure 3: Comparing the graphical representations of the mixture model and the HMM

A Gaussian Mixture model would be parameterized by a vector $\pi = (\pi_1, \ldots, \pi_M) \in \mathbb{R}^M$ such that $\sum\limits_{m=1}^{M} \pi_m = 1$ and $(\mu_m, \Sigma_m) \in \mathbb{R}^d \times \mathbb{R}^{d\times d}$ for each $m \in \{1, \ldots, M\}$, such that:

$$H_i \sim \mathcal{M}(1, \pi_1, \dots, \pi_M)$$
$$\forall t \in \{1, \dots, T\} \; \tilde{X}_t | H_t = m \sim \mathcal{N}_d(\mu_m, \Sigma_m)$$

In the HMM graphical model, each vertical slice represents a time step. Applying d-separation to the the graphical model, we can retrieve the well known result that the future is independent of the past given the present.[1].

To parameterize the Hidden Markov Model, we need to assign local conditional probabilities to each of the nodes. We represent the state at time t as a multinomial random variable H_t with **M** hidden states.

The first state node has no parents, thus we endow this node with an unconditional distribution π (called the **intital state distribution**), such that:

$$\forall i \in [\![1, M]\!] \quad \pi_i = p(H_0 = i)$$

Each successive state node has the previous state node in the chain as its parent, thus we need a $M \times M$ matrix to specify its local conditional probability.

We define a **state transition matrix** Q, where the (i, j)th entry Q_{ij} of Q is defined to be the transition probability $p(H_{t+1} = j | H_t = i)$. We assume a *homogeneous* HMM, so the transition probability is independent of t.

Each of the output node has a single state node as a parent, thus we require a probability distribution $p(\tilde{X}_t | h_t)$ called the **emission distribution**. We assume the emission distribution to be Gaussian and independent of t.

Therefore, the parameterization can be summarized as follows:

$$\begin{aligned} \forall m \in \{1, \dots, M\} \quad & \pi_m = p(H_1 = m) \\ \forall h, h' \in \{1, \dots, M\} \quad & Q_{h,h'} = p(H_{t+1} = h' | H_t = h) \\ \forall t \in \{1, \dots, T\} \; \forall h \in \{1, \dots, M\} \quad & \tilde{X}_t | H_t = h \sim \mathcal{N}_d(\mu_h, \Sigma_h) \end{aligned}$$

We denote $\mu = (\mu_m)_{m \in [\![1,M]\!]} \in \mathbb{R}^{M \times d}$ and $\Sigma = (\Sigma_m)_{m \in [\![1,M]\!]} \in \mathbb{R}^{M \times d \times d}$ the parameters of the emission distributions.

The parameters associated with an HMM are: $\theta = (\pi, Q, \mu, \Sigma)$.

[1] By present, we mean conditioning on the state note H_t, not the output node $\tilde{X}_t$

2.3 The Inference problem

2.3.1 The joint probability

For a particular sequence $(\mathbf{h}, \tilde{\mathbf{x}}) = (h_1, \ldots, h_T, \tilde{x}_1, \ldots, \tilde{x}_t)$, we obtain the following joint probability:

$$p_\theta(\mathbf{h}, \tilde{\mathbf{x}}) = p(h_1) \prod_{t=1}^{T-1} p(h_{t+1}|h_t) \prod_{t=1}^{T} p(\tilde{x}_t|h_t) \tag{1}$$

To introduce Q and π into this equation and adopt a notation in which state variables can be used as indices, we use the one hot encoding notation: The one hot vector $\tilde{h}_t \in \{0,1\}^M$ is defined as follows:

$$\forall t \in [\![1, T]\!] \ \forall i \in [\![1, M]\!] \quad \tilde{h}_t^i = 1 \iff h_t = m \tag{2}$$

We can then define $Q_{h_t,h_{t+1}}$ and π_{h_0} as follows:

$$Q_{h_t,h_{t+1}} := \prod_{i,j=1}^{M} [Q_{ij}]^{\tilde{h}_t^i \tilde{h}_{t+1}^j} \quad \text{and} \quad \pi_{h_0} := \prod_{i=1}^{M} [\pi_i]^{\tilde{h}_0^i} \tag{3}$$

Similarly, we define μ_{h_t} and Σ_{h_t} as follows:

$$\mu_{h_t} := \left(\prod_{i=1}^{M} [\mathrm{diag}(\mu_i)]^{\tilde{h}_t^i} \right) \mathbf{1}_d \quad \text{and} \quad \Sigma_{h_t} := \prod_{i=1}^{M} [\Sigma_i]^{\tilde{h}_t^i}$$

Plugging the definitions into the joint probability, we obtain the parameterized distribution:

$$p_\theta(\mathbf{h}, \tilde{\mathbf{x}}) = \pi_{h_1} \prod_{t=1}^{T-1} Q_{h_t,h_{t+1}} \prod_{t=1}^{T} \mathcal{N}(\tilde{x}_t; \mu_{h_t}, \Sigma_{h_t}) \tag{4}$$

Hence,

$$p_\theta(\tilde{\mathbf{x}}) = \sum_{h_1} \sum_{h_2} \cdots \sum_{h_T} \pi_{h_1} \prod_{t=1}^{T-1} Q_{h_t,h_{t+1}} \prod_{t=1}^{T} \mathcal{N}(\tilde{x}_t; \mu_{h_t}, \Sigma_{h_t}) \tag{5}$$

2.3.2 Smoothing - Filtering probabilities

We introduce the following probabilities:

- **Filtering probabilities:** $\xi \in \mathbb{R}^{T\times M}$

$$\xi(t,h) := p(H_t = h|\tilde{X}_1,\dots,\tilde{X}_t) \quad \forall (t,h) \in [\![1,T]\!] \times [\![1,M]\!]$$

- **Smoothing probabilities:** $\psi \in \mathbb{R}^{T\times M}$ and $\phi \in \mathbb{R}^{T-1\times M\times M}$

$$\begin{aligned}\psi(t,h) &:= p(H_t = h|\tilde{X}_1,\dots,\tilde{X}_T) \quad \forall (t,h) \in [\![1,T]\!] \times [\![1,M]\!]\\ \phi(t,h,h') &:= p(H_t = h, H_{t+1} = h'|\tilde{X}_1,\dots,\tilde{X}_T)\\ &\qquad \forall (t,h,h') \in [\![1,T-1]\!] \times [\![1,M]\!] \times [\![1,M]\!]\end{aligned}$$

2.4 The Forward Backward Algorithm for calculating filtering and smoothing probabilities

2.4.1 Filtering probabilities: Forward Algorithm

In order to compute the filtering probabilities $\xi \in \mathbb{R}^{T\times M}$, we will take advantage of the conditional independencies in the graphical model, breaking the problem into pieces as follows:

1. **Expressing the Filtering probabilities in term of alpha variables:**

   First, let's introduce $\alpha \in \mathbb{R}^{T\times M}$:

$$\forall (t,h) \in [\![1,T]\!] \times [\![1,M]\!] \quad \alpha(t,h) := p(\tilde{X}_1,\dots,\tilde{X}_t, H_t = h) \tag{6}$$

   The filtering probabilities can be expressed using the α variables, which are easy to compute (recursively).

   Exercise:

Show that:

$$\forall t \in [\![1, T]\!] \quad \xi_t = \frac{\alpha_t}{\mathbf{1}_M^T \alpha_t}$$

Solution:

Indeed, for all $(t, h) \in [\![1, T]\!] \times [\![1, M]\!]$:

$$\begin{aligned}
\xi(t, h) &:= p(H_t = h | \tilde{X}_1, \ldots, \tilde{X}_t) \\
&= \frac{p(H_t = h, \tilde{X}_1, \ldots, \tilde{X}_t)}{p(\tilde{X}_1, \ldots, \tilde{X}_t)} \\
&= \frac{p(H_t = h, \tilde{X}_1, \ldots, \tilde{X}_t)}{\sum\limits_{h'=1}^{M} p(H_t = h', \tilde{X}_1, \ldots, \tilde{X}_t)} \\
&= \frac{\alpha(t, h)}{\sum\limits_{h'=1}^{M} \alpha(t, h')}
\end{aligned}$$

So,

$$\boxed{\forall (t, h) \in [\![1, T]\!] \times [\![1, M]\!] \quad \xi(t, h) = \frac{\alpha(t, h)}{\sum\limits_{h'=1}^{M} \alpha(t, h')}} \tag{7}$$

Notation:

$\forall t \in [\![1, T]\!] \quad \alpha_t = (\alpha(t, 1), \ldots, \alpha(t, M))^T$ and $\xi_t = (\xi(t, 1), \ldots, \xi(t, M))^T$

The equation 7 can then be written as follows:

$$\boxed{\forall t \in [\![1, T]\!] \quad \xi_t = \frac{\alpha_t}{\mathbf{1}_M^T \alpha_t}} \tag{8}$$

2. **Calculating the alpha variables recursively:**

   Let's introduce the emission tensor $\Gamma \in [\![1, T]\!] \times [\![1, M]\!] \times [\![1, M]\!]$ such that:

$$\forall t \in [\![1, T]\!] : \Gamma(t) = \begin{pmatrix} p(\tilde{X}_t | H_t = 1) & & \\ & \ddots & \\ & & p(\tilde{X}_t | H_t = M) \end{pmatrix}$$

Exercise:

Using the marginalization on the previous hidden state and exploiting the graphical model independencies, show that:

$$\alpha_1 = \Gamma(1)\pi, \quad \forall t \geq 2 \ \alpha_t = \Gamma(t) Q^T \alpha_{t-1}$$

Solution:

For all $(t, h) \in [\![1, T]\!] \times [\![1, M]\!]$:

$$\begin{aligned}
\alpha(t,h) &= p(\tilde{X}_1, \dots, \tilde{X}_t, H_t = h) \\
&= \sum_{h'=1}^{M} p(\tilde{X}_1, \dots, \tilde{X}_t, H_t = h, H_{t-1} = h') \\
&= \sum_{h'=1}^{M} p(\tilde{X}_t | \tilde{X}_1, \dots, \tilde{X}_{t-1}, H_t = h, H_{t-1} = h') p(\tilde{X}_1, \dots, \tilde{X}_{t-1}, H_t = h, H_{t-1} = h') \\
&= \sum_{h'=1}^{M} p(\tilde{X}_t | H_t = h) p(H_t = h | \tilde{X}_1, \dots, \tilde{X}_{t-1}, H_{t-1} = h') p(\tilde{X}_1, \dots, \tilde{X}_{t-1}, H_{t-1} = h') \\
&= \sum_{h'=1}^{M} p(\tilde{X}_t | H_t = h) \underbrace{p(H_t = h | H_{t-1} = h')}_{Q_{h'h}} \underbrace{p(\tilde{X}_1, \dots, \tilde{X}_{t-1}, H_{t-1} = h')}_{\alpha(t-1,h')}
\end{aligned}$$

So,

$$\forall (t,h) \in [\![1, T]\!] \times [\![1, M]\!] \quad \alpha(t,h) = \sum_{h'=1}^{M} p(\tilde{X}_t | H_t = h) Q_{h'h} \alpha(t-1, h') \tag{9}$$

The equation 9 can be written as follows:

$$\boxed{\alpha_1 = \Gamma(1)\pi, \quad \forall t \geq 2 \ \alpha_t = \Gamma(t)Q^T\alpha_{t-1}} \tag{10}$$

3. **Handling numerical issues:**

From the above sections, we conclude that, given an observed sequence, the filtering probabilities can be obtained by:

- Calculating the alpha variables recursively, as shown in the equations 10.
- Deducing the filtering probabilities, as shown in equation 8.

However, attention must be paid to numerical issues during the implementation. Since the recursion in 10 implies several multiplications of small numbers, the numbers underflow quite rapidly.

To avoid underflow, it suffices to normalize. Indeed, normalizing the alpha variables (6), which means dividing $p(\tilde{X}_1, \ldots, \tilde{X}_t, H_t = h)$ by $p(\tilde{X}_1, \ldots, \tilde{X}_t)$ yields conditionals $p(H_t = h|\tilde{X}_1, \ldots, \tilde{X}_t)$. These conditionals, which represent the filtered estimates of the states, scaled in a stable manner. In sum, one should compute directly the filtering probabilities instead of computing recursively the alpha variables and then deducing the filtering estimates.

Therefore, let's introduce $(\tilde{\xi}_t)_{1\leq t\leq T}$ as follows:

$$\boxed{\forall t \in [\![1, T]\!] \quad \tilde{\xi}_t := \Gamma(t)Q^T\xi_{t-1}} \tag{11}$$

And $(c_t)_{1\leq t\leq T}$ defined as follows:

$$\boxed{\forall t \in [\![1, T]\!] \quad c_t := \mathbf{1}^T\tilde{\xi}_t} \tag{12}$$

Exercise:

(a) Show that:

$$\forall t \in [\![1, T]\!] \quad c_t = \frac{\mathbf{1}_M^T\alpha_t}{\mathbf{1}_M^T\alpha_{t-1}}$$

(b) Deduce that:

$$\forall t \in [\![1, T]\!] \quad \xi_t = \frac{\tilde{\xi}_t}{c_t}$$

(c) Deduce an algorithm to calculate the filtering probabilities recursively

We have then:

$$\begin{aligned} c_t &:= \mathbf{1}^T \tilde{\xi}_t \\ &= \mathbf{1}_M^T \Gamma(t) Q^T \xi_{t-1} \\ &= \mathbf{1}_M^T \frac{\overbrace{\Gamma(t) Q^T \alpha_{t-1}}^{\alpha_t}}{\mathbf{1}_M^T \alpha_{t-1}} \quad \text{(from 8 and 10)} \\ &= \frac{\mathbf{1}_M^T \alpha_t}{\mathbf{1}_M^T \alpha_{t-1}} \end{aligned}$$

So,

$$\forall t \in [\![1, T]\!] \quad c_t = \frac{\mathbf{1}_M^T \alpha_t}{\mathbf{1}_M^T \alpha_{t-1}} \tag{13}$$

And we have:

$$\begin{aligned} \xi_t &= \frac{1}{\mathbf{1}_M^T \alpha_t} \alpha_t \quad \text{(from 8)} \\ &= \frac{1}{\mathbf{1}_M^T \alpha_t} \Gamma(t) Q^T \alpha_{t-1} \quad \text{(from 10)} \\ &= \Gamma(t) Q^T \underbrace{\frac{\alpha_{t-1}}{\mathbf{1}_M^T \alpha_{t-1}}}_{\xi_{t-1}} \frac{\mathbf{1}_M^T \alpha_{t-1}}{\mathbf{1}_M^T \alpha_t} \\ &= \underbrace{\Gamma(t) Q^T \xi_{t-1}}_{:=\tilde{\xi}_t} \frac{\mathbf{1}_M^T \alpha_{t-1}}{\mathbf{1}_M^T \alpha_t} \\ &= \frac{\tilde{\xi}_t}{c_t} \quad \text{(from 13)} \end{aligned}$$

So,

$$\boxed{\forall t \in [\![1, T]\!] \quad \xi_t = \frac{\tilde{\xi}_t}{c_t}} \tag{14}$$

Finally, the recursion consists in the following steps:

- Initialization of c_1 and ξ_1

- We move from t to $t+1$ as follows:

$$\xi_t \xrightarrow{\text{eq11}} \tilde{\xi}_{t+1} \xrightarrow{\text{eq12}} c_{t+1}$$

$$(c_{t+1} \text{ and } \tilde{\xi}_{t+1}) \xrightarrow{\text{eq14}} \xi_{t+1}$$

4. **The Forward Algorithm:**

   From the previous section, we conclude the following algorithm, called **Forward Algorithm**:

Algorithm 3 Forward Algorithm

Require: Observations $\tilde{X}_1 \dots \tilde{X}_T$
Ensure: $(\xi_t)_{1 \leq t \leq T}$ (The filtering probabilties)

```
$c_1 \leftarrow \mathbf{1}^T \Gamma(1) \pi$
$\xi_1 \leftarrow \Gamma(1) \pi / c_1$
for $t \leftarrow 2, \dots, T$ do
    $\tilde{\xi}_t \leftarrow \Gamma(t) Q^T \xi_{t-1}$
    $c_t \leftarrow \mathbf{1}^T \tilde{\xi}_t$
    $\xi_t \leftarrow \tilde{\xi}_t / c_t$
end for
```

2.4.2 Smoothing probabilities: Forward Backward Algorithm

1. **Expressing the smoothing probabilities in term of alpha and beta variables:**

   First, let's introduce $\beta \in \mathbb{R}^{T \times M}$:

$$\forall (t,h) \in [\![1,T]\!] \times [\![1,M]\!] \quad \beta(t,h) := p(\tilde{X}_{t+1}, \dots, \tilde{X}_T | H_t = h) \tag{15}$$

   We introduce the following notations:

$$\forall t \in [\![1,T]\!] \quad \beta_t = (\beta(t,1), \dots, \beta(t,M))^T \quad \phi_t = [\phi(t,h,h')]_{h,h'} \quad \psi_t = [\psi(t,h)]_h$$

   The smoothing probabilities can be expressed using the alpha and beta variables, which we can calculate (recursively).

   Exercise:

   Show that:

$$\boxed{\forall t \in [\![1, T]\!] \quad \psi_t = \frac{\alpha_t \circ \beta_t}{\mathbf{1}_M^T \alpha_t}}$$

And

$$\boxed{\forall t \in [\![1, T]\!] \quad \phi_t = \frac{\operatorname{diag}(\alpha_t) Q \Gamma(t+1) \operatorname{diag}(\beta_{t+1})}{\mathbf{1}_M^T \alpha_T}}$$

Solution:

Indeed, for all $(t, h) \in [\![1, T]\!] \times [\![1, M]\!]$:

$$\begin{aligned}
\psi(t,h) &:= p(H_t = h | \tilde{X}_1, \ldots, \tilde{X}_T) \\
&= \frac{p(H_t = h, \tilde{X}_1, \ldots, \tilde{X}_T)}{p(\tilde{X}_1, \ldots, \tilde{X}_T)} \\
&= \frac{p(H_t = h, \tilde{X}_1, \ldots, \tilde{X}_t) p(\tilde{X}_{t+1}, \ldots, \tilde{X}_T | H_t = h, \tilde{X}_1, \ldots, \tilde{X}_t)}{\sum_{h'=1}^M p(H_t = h', \tilde{X}_1, \ldots, \tilde{X}_T)} \\
&= \frac{p(H_t = h, \tilde{X}_1, \ldots, \tilde{X}_t) p(\tilde{X}_{t+1}, \ldots, \tilde{X}_T | H_t = h)}{\sum_{h'=1}^M p(H_t = h', \tilde{X}_1, \ldots, \tilde{X}_T)} \\
&= \frac{\alpha(t,h)\beta(t,h)}{\sum\limits_{h'=1}^{M} \alpha(t,h')}
\end{aligned}$$

So,

$$\forall (t, h) \in [\![1, T]\!] \times [\![1, M]\!] \quad \psi(t,h) = \frac{\alpha(t,h)\beta(t,h)}{\sum\limits_{h'=1}^{M} \alpha(t,h')} \tag{16}$$

i.e,

$$\boxed{\forall t \in [\![1, T]\!] \quad \psi_t = \frac{\alpha_t \circ \beta_t}{\mathbf{1}_M^T \alpha_t}}$$

We also have for all $(t, h, h') \in [\![1, T]\!] \times [\![1, M]\!] \times [\![1, M]\!]$:

$$\begin{aligned}
\phi(t,h,h') &:= p(H_t = h, H_{t+1} = h' | \tilde{X}_1, \dots, \tilde{X}_T) \\
&= \frac{p(\tilde{X}_1, \dots, \tilde{X}_T | H_t = h, H_{t+1} = h') p(H_t = h, H_{t+1} = h')}{p(\tilde{X}_1, \dots, \tilde{X}_T)} \\
&= \frac{\begin{array}{r} p(\tilde{X}_1, \dots, \tilde{X}_t | \tilde{X}_{t+1}, \dots, \tilde{X}_T, H_t = h, H_{t+1} = h') p(\tilde{X}_{t+1}, \dots, \tilde{X}_T | H_t = h, H_{t+1} = h') \\ p(H_t = h, H_{t+1} = h') \end{array}}{p(\tilde{X}_1, \dots, \tilde{X}_T)} \\
&= \frac{\overbrace{p(\tilde{X}_1, \dots, \tilde{X}_t | H_t = h)} p(\tilde{X}_{t+1}, \dots, \tilde{X}_T | H_{t+1} = h') p(H_{t+1} = h' | H_t = h) \overbrace{p(H_t = h)}}{p(\tilde{X}_1, \dots, \tilde{X}_T)} \\
&= \frac{\alpha(t,h) p(\tilde{X}_{t+1}, \dots, \tilde{X}_T | H_{t+1} = h') p(H_{t+1} = h' | H_t = h)}{p(\tilde{X}_1, \dots, \tilde{X}_T)} \\
&= \frac{\alpha(t,h) p(\tilde{X}_{t+1} | \tilde{X}_{t+2}, \dots, \tilde{X}_T, H_{t+1} = h') p(\tilde{X}_{t+2}, \dots, \tilde{X}_T | H_{t+1} = h') p(H_{t+1} = h' | H_t = h)}{p(\tilde{X}_1, \dots, \tilde{X}_T)} \\
&= \frac{\alpha(t,h) p(\tilde{X}_{t+1} | H_{t+1} = h') \overbrace{p(\tilde{X}_{t+2}, \dots, \tilde{X}_T | H_{t+1} = h')}^{:=\beta(t+1,h')} \overbrace{p(H_{t+1} = h' | H_t = h)}^{Q_{hh'}}}{\sum_{h''=1}^{M} p(\tilde{X}_1, \dots, \tilde{X}_T, H_T = h'')} \\
&= \frac{\alpha(t,h) p(\tilde{X}_{t+1} | H_{t+1} = h') \beta(t+1,h') Q_{hh'}}{\sum_{h''=1}^{M} \alpha(T,h'')}
\end{aligned}$$

So,

$$\forall (t,h,h') \in [\![1,T]\!] \times [\![1,M]\!] \times [\![1,M]\!] \quad \phi(t,h,h') = \frac{\alpha(t,h) p(\tilde{X}_{t+1} | H_{t+1} = h') \beta(t+1,h') Q_{hh'}}{\sum_{h''=1}^{M} \alpha(T,h'')} \tag{17}$$

So, The equation 17 can then be written as:

$$\boxed{\forall t \in [\![1,T]\!] \quad \phi_t = \frac{\text{diag}(\alpha_t) Q \Gamma(t+1) \text{diag}(\beta_{t+1})}{\mathbf{1}_M^T \alpha_T}} \tag{18}$$

2. **Calculating the beta variables recursively:**

Exercise:

Using the marginalization on the next hidden state and exploiting the graphical model independencies, show that

$$\beta_T = \mathbf{1}_M, \quad \forall 0 \leq t \leq T-1 \ \beta_t = Q\Gamma(t+1)\beta_{t+1}$$

Solution:

For all $(t, h) \in [\![1, T]\!] \times [\![1, M]\!]$:

$$\begin{aligned}
\beta(t,h) &= p(\tilde{X}_{t+1}, \ldots, \tilde{X}_T | H_t = h) \\
&= \sum_{h'=1}^{M} p(\tilde{X}_{t+1}, \ldots, \tilde{X}_T, H_{t+1} = h' | H_t = h) \\
&= \sum_{h'=1}^{M} p(\tilde{X}_{t+1}, \ldots, \tilde{X}_T | H_{t+1} = h', H_t = h) p(H_{t+1} = h' | H_t = h) \\
&= \sum_{h'=1}^{M} p(\tilde{X}_{t+1}, \ldots, \tilde{X}_T | H_{t+1} = h') p(H_{t+1} = h' | H_t = h) \\
&= \sum_{h'=1}^{M} p(\tilde{X}_{t+1} | \tilde{X}_{t+2}, \ldots, \tilde{X}_T, H_{t+1} = h') p(\tilde{X}_{t+2}, \ldots, \tilde{X}_T | H_{t+1} = h') p(H_{t+1} = h' | H_t = h) \\
&= \sum_{h'=1}^{M} p(\tilde{X}_{t+1} | H_{t+1} = h') p(\tilde{X}_{t+2}, \ldots, \tilde{X}_T | H_{t+1} = h') p(H_{t+1} = h' | H_t = h) \\
&= \sum_{h'=1}^{M} p(\tilde{X}_{t+1} | H_{t+1} = h') \beta(t+1, h') Q_{hh'}
\end{aligned}$$

So,

$$\forall (t,h) \in [\![1, T]\!] \times [\![1, M]\!] \quad \beta(t,h) = \sum_{h'=1}^{M} p(\tilde{X}_{t+1} | H_{t+1} = h') \beta(t+1, h') Q_{hh'} \tag{19}$$

The equation 19 can be written as follows:

$$\boxed{\beta_T = \mathbf{1}_M, \quad \forall 0 \leq t \leq T-1 \ \beta_t = Q\Gamma(t+1)\beta_{t+1}} \tag{20}$$

2.4.3 Wrap up: Calculating the filtering and the smoothing probabilities using the Forward Backward Algorithm

In order to compute the filtering and the smoothing probabilities and the likelihood efficiently, we use the Forward Backward algorithm 4.

Algorithm 4 Forward Backward Algorithm

Require: Observations $\tilde{X}_1 \dots \tilde{X}_T$
Ensure: $(\xi_t)_{1 \leq t \leq T}$ and $(\psi_t)_{1 \leq t \leq T}$ (The filtering and smoothing probabilties)

```
$c_1 \leftarrow 1^T\Gamma(1)\pi$
$\xi_1 \leftarrow \Gamma(1)\pi/c_1$
for $t \leftarrow 2, \dots, T$ do
    $\tilde{\xi}_t \leftarrow \Gamma(t)Q^T\xi_{t-1}$
    $c_t \leftarrow 1^T\tilde{\xi}_t$
    $\xi_t \leftarrow \tilde{\xi}_t/c_t$
end for
$\tilde{\beta}_T \leftarrow 1/c_T$
for $t \leftarrow 1, \dots, T-1$ do
    $\tilde{\beta}_{T-t} \leftarrow Q\Gamma(T-t+1)\tilde{\beta}_{T-t+1}$
    $\psi_{T-t} \leftarrow \text{diag}(\xi_{T-t})Q\Gamma(T-t+1)\text{diag}(\tilde{\beta}_{T-t+1})$
    $\phi_{T-t} \leftarrow \psi_{T-t}1_M$
end for
```

2.5 Learning the parameters of the HMM using the Expectation Maximization (EM) Algorithm

In order to learn the parameters of the HMM, we use the Expectation Maximization algorithm.

1. **Introducing the EM algorithm**

   The EM algorithm is an iterative method for finding maximum likelihood estimates of parameters in statistical models, where the models depend on unobserved latent variables.

   Consider for instance N observations $a_1, \dots, a_N$ and the latent variables associateed with them $z_1, \dots, z_N$.

We assume the pairs (a_i, z_i) to be independent and identically distributed. For $(a, z) = (a_1, z_1, \dots, a_N, z_N)$, the objective is to maximize:

$$\log(p(a;\theta)) = \sum_{i=1}^{N} \log\left(\sum_{z_i} p(a_i, z_i; \theta)\right)$$

By conditioning on a latent variable z and using the Jensen inequality, we get:

$$\begin{aligned}
\log(p(a;\theta)) &= \log\left(\sum_{z} p(a, z; \theta)\right) \\
&= \log\left(\sum_{z} q(z)\frac{p(a, z; \theta)}{q(z)}\right) \\
&\geq \sum_{z} q(z) \log\left(\frac{p(a, z; \theta)}{q(z)}\right) \\
&= \underbrace{\mathbb{E}_q[\log(p(a, z; \theta))] + f(q)}_{\mathcal{L}(q,\theta)}
\end{aligned}$$

With equality iff $q(z) = p_\theta(z|a)$

The EM algorithm can then be summarized as depicted in 5

Algorithm 5 EM algorithm

```
Require: Observations $a_1 \dots a_N$
Ensure: Optimal $\theta$

Initialize $\theta$
$\xi_1 \leftarrow \Gamma(1)\pi/c_1$
while (Not converged) do
    E-step: $q(z) = p(z|a;\theta^{(i-1)})$
    M-step: $\theta^{(i)} = \arg\max_\theta \mathbb{E}_q[\log(p(a, z; \theta)]$
end while
```

2. **Learning the parameters of the HMM using the EM algorithm**

   (a) **The E-step**

   At the iteration i, the complete loglikelihood $\log(p_{\theta_i}(\mathbf{h}, \tilde{\mathbf{x}}))$ is expressed as follows:

$$\log(p_{\theta_i}(\mathbf{h},\tilde{\mathbf{x}})) = \log\left(\pi^{(i)}_{h_1}\prod_{t=1}^{T-1} Q^{(i)}_{h_t,h_{t+1}}\prod_{t=1}^{T}\mathcal{N}(\tilde{x}_t;\mu^{(i)}_{h_t},\Sigma^{(i)}_{h_t})\right) \tag{21}$$

$$= \log(\pi^{(i)}_{h_1}) + \sum_{t=1}^{T-1}\log(Q^{(i)}_{h_t,h_{t+1}}) + \sum_{t=1}^{T-1}\log(\mathcal{N}(\tilde{x}_t;\mu^{(i)}_{h_t},\Sigma^{(i)}_{h_t})) \tag{22}$$

The E-step consists in computing : $\mathbb{E}_{\mathbf{H}|\tilde{\mathbf{x}}}[\log(p_{\theta_i}(\mathbf{h},\tilde{\mathbf{x}}))]$.
We have:

$$\mathbb{E}_{\mathbf{H}|\tilde{\mathbf{x}}}[\log(\pi^{(i)}_{h_1})] = \sum_{h=1}^{M}\log(\pi^{(i)}_{h})p(H_1 = h|\tilde{\mathbf{x}}) \tag{23}$$

$$\mathbb{E}_{\mathbf{H}|\tilde{\mathbf{x}}}[\log(Q^{(i)}_{h_t,h_{t+1}})] = \sum_{h=1}^{M}\sum_{h'=1}^{M}\log(Q^{(i)}_{hh'})p(H_t = h, H_{t+1} = h'|\tilde{\mathbf{x}}) \tag{24}$$

$$\mathbb{E}_{\mathbf{H}|\tilde{\mathbf{x}}}[\log(\mathcal{N}(\tilde{X}_t;\mu^{(i)}_{h_t},\Sigma^{(i)}_{h_t}))] = \sum_{h=1}^{M}\log(\mathcal{N}(\tilde{X}_t;\mu^{(i)}_{h},\Sigma^{(i)}_{h}))p(h_t = h|\tilde{\mathbf{x}}) \tag{25}$$

By summing the three parts, we obtain the following expression:

$$\begin{aligned}\mathbb{E}_{\mathbf{H}|\tilde{\mathbf{x}}}[\log(p_{\theta_i}(\mathbf{h},\tilde{\mathbf{x}}))] =& \sum_{h=1}^{M}\log(\pi^{(i)}_{h})p(H_1 = h|\tilde{\mathbf{x}}) + \sum_{t=1}^{T-1}\sum_{h=1}^{M}\sum_{h'=1}^{M}\log(Q^{(i)}_{hh'})p(H_t = h, H_{t+1} = h'|\tilde{\mathbf{x}})\\ &+ \sum_{t=1}^{T-1}\sum_{h=1}^{M}\log(\mathcal{N}(\tilde{X}_t;\mu^{(i)}_{h},\Sigma^{(i)}_{h}))p(H_t = h|\tilde{\mathbf{x}})\end{aligned} \tag{26}$$

Introducing the smoothing probabilities, we obtain:

$$\boxed{\begin{aligned}\mathbb{E}_{\mathbf{H}|\tilde{\mathbf{x}}}[\log(p_{\theta_i}(\mathbf{h},\tilde{\mathbf{x}}))] = \quad & \sum_{h=1}^{M}\log(\pi^{(i)}_{h})\psi(1,h) + \sum_{t=1}^{T-1}\sum_{h=1}^{M}\sum_{h'=1}^{M}\log(Q^{(i)}_{hh'})\phi(t,h,h')\\ &+ \sum_{t=1}^{T-1}\sum_{h=1}^{M}\log(\mathcal{N}(\tilde{X}_t;\mu^{(i)}_{h},\Sigma^{(i)}_{h}))\psi(t,h)\end{aligned}} \tag{27}$$

(b) **The M-step**

The objective is to maximize $\mathbb{E}_{\mathbf{H}|\tilde{\mathbf{x}}}[\log(p_{\theta}(\mathbf{h},\tilde{\mathbf{x}}))]$ with respect to θ:

$$\theta_{i+1} = \arg\max_{\theta}\mathbb{E}_{\mathbf{H}|\tilde{\mathbf{x}}}[\log(p_{\theta}(\mathbf{h},\tilde{\mathbf{x}}))] \tag{28}$$

Exercise:

Show that we obtain the following update equations, expressed with the smoothing probabilities.
$\forall (h,h') \in [\![1,M]\!] \times [\![1,M]\!]$

$$\pi_h^{(i+1)} = \psi(1,h) \tag{29}$$

$$Q_{h,h'}^{(i+1)} = \frac{\sum\limits_{t=1}^{T} \phi(t,h,h')}{\sum\limits_{t=1}^{T} \psi(t,h)} \tag{30}$$

$$\mu_h^{(i+1)} = \frac{\sum\limits_{t=1}^{T} \psi(t,h)\tilde{X}_t}{\sum\limits_{t=1}^{T} \psi(t,h)} \tag{31}$$

$$\Sigma_h^{(i+1)} = \frac{\sum\limits_{t=1}^{T} \psi(t,h)(\tilde{X}_t - \mu_h^{(i)})(\tilde{X}_t - \mu_h^{(i)})^T}{\sum\limits_{t=1}^{T} \psi(t,h)} \tag{32}$$

Solution: See Appendix A

2.6 Predicting the distribution of the hidden state over the next period of time

Once the model is trained using the EM algorithm, we would like to compute the probability of being in each hidden state over the next period of time, as shown in figure 4.

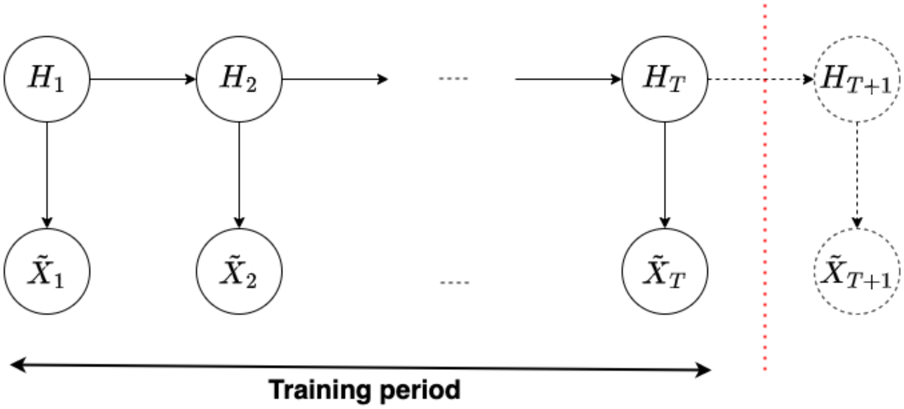

Figure 4: Predicting the turbulence state over the next period of time

The prediction over the next period can be calculated as follows:

$$\begin{aligned}
\forall h \in [\![1, M]\!] \quad & p(H_{T+1} = h | \tilde{X}_1 = \tilde{x}_1, \dots, \tilde{X}_T = \tilde{x}_T) \\
&= \sum_{h'=1}^{M} p(H_{T+1} = h, H_T = h' | \tilde{X}_1 = \tilde{x}_1, \dots, \tilde{X}_T = \tilde{x}_T) \\
&= \sum_{h'=1}^{M} \underbrace{p(H_{T+1} = h | H_T = h')}_{=Q_{h'h}} \underbrace{p(H_T = h' | \tilde{X}_1 = \tilde{x}_1, \dots, \tilde{X}_T = \tilde{x}_T)}_{\xi(T,h')}
\end{aligned}$$

The filtering probabilities $\xi(t, h')$ are calculated using the Forward Backward algorithm as explained before.

A The M-step of the HMM model

1. **Introducing the problem**

   We want to maximize : $\mathbb{E}_{\mathbf{H}|\tilde{\mathbf{x}}}[\log(p_\theta(\mathbf{h}, \tilde{\mathbf{x}}))]$ with respect to $\theta = (\pi, Q, \mu, \Sigma)$

   We have:

$$\begin{aligned}\mathbb{E}_{\mathbf{H}|\tilde{\mathbf{x}}}[\log(p_\theta(\mathbf{h}, \tilde{\mathbf{x}}))] = &\sum_{h=1}^{M} \log(\pi_h)\psi(1,h) + \sum_{t=1}^{T-1}\sum_{h=1}^{M}\sum_{h'=1}^{M} \log(Q_{hh'})\phi(t,h,h') \\ &+ \sum_{t=1}^{T-1}\sum_{h=1}^{M} \log(\mathcal{N}(\tilde{X}_t; \mu_h, \Sigma_h))\psi(t,h)\end{aligned} \tag{33}$$

   So, the optimization problem can be decomposed in three sub-problems.

$$\pi^{(i+1)} = \arg\min_{\pi} - \sum_{h=1}^{M} \log(\pi_h)\psi(1,h) \quad \text{such that} \quad \pi \perp \nVdash_{\mathbb{M}} \tag{34}$$

$$Q^{(i+1)} = \arg\min_{Q} - \sum_{t=1}^{T-1}\sum_{h=1}^{M}\sum_{h'=1}^{M} \log(Q_{hh'})\phi(t,h,h') \quad \text{such that} \quad \sum_{h'=1}^{M} Q_{hh'} = 1 \tag{35}$$

$$(\mu^{(i+1)}, \Sigma^{(i+1)}) = \arg\min_{(\mu,\Sigma)} - \sum_{t=1}^{T-1}\sum_{h=1}^{M} \log(\mathcal{N}(\tilde{X}_t; \mu_h, \Sigma_h))\psi(t,h) \tag{36}$$

2. **Update of the initial state**

   We use the Langrangian to solve the first optimization problem 34:

$$\mathcal{L}(\pi, \lambda) = - \sum_{h=1}^{M} \log(\pi_h)\psi(1,h) + \lambda\left(\sum_{h=1}^{M} \pi_h - 1\right) \tag{37}$$

   We have:

$$\forall h \in [\![1, M]\!] \quad 0 = \frac{\partial \mathcal{L}}{\partial \pi_h}(\pi^{(i+1)}, \lambda) = -\frac{\psi(1,h)}{\pi_h^{(i+1)}} + \lambda \tag{38}$$

   Thus,

$$\forall h \in [\![1, M]\!] \quad \pi_h^{(i+1)} = \frac{\psi(1,h)}{\lambda} \tag{39}$$

   And from:

$$1 = \sum_{h=1}^{M} \pi_h^{(i+1)} = \sum_{h=1}^{M} \frac{\psi(1,h)}{\lambda} = \frac{1}{\lambda} \underbrace{\sum_{h=1}^{M} \psi(1,h)}_{=1} \tag{40}$$

We get:

$$\lambda = 1 \tag{41}$$

Therefore,

$$\boxed{\forall h \in [\![1, M]\!] \quad \pi_h^{(i+1)} = \psi(1,h)} \tag{42}$$

3. **Update of the transition matrix**

Again, we use the Langrangian to solve the second optimization problem 35:

$$\mathcal{L}(Q, \lambda_1, \dots, \lambda_M) = - \sum_{t=1}^{T-1} \sum_{h=1}^{M} \sum_{h'=1}^{M} \log(Q_{hh'})\phi(t,h,h') + \sum_{h=1}^{M} \lambda_h \left(\sum_{h'=1}^{M} Q_{hh'} - 1 \right) \tag{43}$$

We have:

$$\forall (h,h') \in [\![1, M]\!] \times [\![1, M]\!] \quad 0 = \frac{\partial \mathcal{L}}{\partial Q_{hh'}}(Q^{(i+1)}, \lambda) = - \sum_{t=1}^{T-1} \frac{\phi(t,h,h')}{Q_{hh'}^{(i+1)}} + \lambda_h \tag{44}$$

Thus,

$$\forall (h,h') \in [\![1, M]\!] \times [\![1, M]\!] \quad Q_{hh'}^{(i+1)} = \frac{\sum_{t=1}^{T-1} \phi(t,h,h')}{\lambda_h} \tag{45}$$

And from:

$$\forall h \in [\![1, M]\!] \quad 1 = \sum_{h'=1}^{M} Q_{hh'}^{(i+1)} = \sum_{h'=1}^{M} \left(\frac{\sum_{t=1}^{T-1} \phi(t,h,h')}{\lambda_h} \right) = \frac{1}{\lambda_h} \sum_{t=1}^{T-1} \underbrace{\sum_{h'=1}^{M} \phi(t,h,h')}_{=\psi(t,h)} \tag{46}$$

We get:

$$\forall h \in [\![1, M]\!] \quad \lambda_h = \sum_{t=1}^{T-1} \psi(t,h) \tag{47}$$

Therefore,

$$\boxed{\forall (h,h') \in [\![1, M]\!] \times [\![1, M]\!] \quad Q_{hh'}^{(i+1)} = \frac{\sum\limits_{t=1}^{T-1} \phi(t,h,h')}{\sum\limits_{t=1}^{T-1} \psi(t,h)}} \tag{48}$$

4. **Update of the emission distribution** We have:

$$(\mu^{(i+1)}, \Sigma^{(i+1)}) = \arg\min_{(\mu,\Sigma)} \underbrace{-\sum_{t=1}^{T-1} \sum_{h=1}^{M} \log(\mathcal{N}(\tilde{X}_t; \mu_h, \Sigma_h))\psi(t,h)}_{J(\mu,\Sigma)} \tag{49}$$

(a) <u>**Update μ**</u>:

Let's fix $h \in [\![1, M]\!]$ and $t \in [\![1, T]\!]$
We have:

$$\log(\mathcal{N}(\tilde{X}_t; \mu_h, \Sigma_h)) = \frac{D}{2}\log(2\pi) - \frac{1}{2}\log(\det(\Sigma_h)) - \frac{1}{2}(\tilde{X}_t - \mu_h)^T \Sigma_h^{-1} (\tilde{X}_t - \mu_h) \tag{50}$$

We define

$$\xi : \mu_h \overset{f}{\mapsto} \mu_h - \tilde{X}_t \overset{g}{\mapsto} (\tilde{X}_t - \mu_h)^T \Sigma_h^{-1} (\tilde{X}_t - \mu_h) \tag{51}$$

Obviously,

$$\nabla f(x) = x \tag{52}$$

We have

$$\begin{aligned} \forall (x,\epsilon) \in \mathbb{R}^D \times \mathbb{R}^D \quad g(x+\epsilon) - g(x) &= (x+\epsilon)^T \Sigma_h^{-1} (x+\epsilon) - x^T \Sigma_h^{-1} x \\ &= \epsilon^T \Sigma_h^{-1} x + x^T \Sigma_h^{-1} \epsilon + \epsilon^T \Sigma_h^{-1} \epsilon \\ &= \epsilon^T (\Sigma_h^{-1} x + (\Sigma_h^{-1})^T x) + o(||\epsilon||) \\ &= \epsilon^T (2\Sigma_h^{-1} x) + o(||\epsilon||) \\ &= \underbrace{\langle 2\Sigma_h^{-1} x), \epsilon \rangle}_{dg_x(\epsilon)} + o(||\epsilon||) \\ &= \langle \nabla g(x), \epsilon \rangle \end{aligned}$$

So,

$$\nabla g(x) = 2\Sigma_h^{-1}x \tag{53}$$

Therefore,

$$\begin{aligned}
\forall x \in \mathbb{R}^D \quad \langle \nabla \xi(\mu_h), x\rangle &= d(\xi)_{\mu_h}(x) \\
&= d(g \circ f)_{\mu_h}(x) \\
&= d(g)_{\mu_h - \tilde{X}_t}(d(f)_{\mu_h}(x)) \\
&= d(g)_{\mu_h - \tilde{X}_t}(\langle \nabla f(\mu_h), x\rangle) \\
&= d(g)_{\mu_h - \tilde{X}_t}(x) \\
&= \langle \nabla g(\mu_h - \tilde{X}_t), x\rangle \\
&= \langle 2\Sigma_h^{-1}(\mu_h - \tilde{X}_t), x\rangle
\end{aligned}$$

So,

$$\nabla \xi(\mu_h) = 2\Sigma_h^{-1}(\mu_h - \tilde{X}_t) \tag{54}$$

Thus,

$$0 = \nabla_{\mu_h} J(\mu_h^{(i+1)}, \Sigma) = -\sum_{t=1}^{T} \psi(t,h)\Sigma_h^{-1}(\mu_h - \tilde{X}_t) \tag{55}$$

Therefore,

$$\boxed{\forall h \in [\![1, M]\!] \quad \mu_h^{(i+1)} = \frac{\sum\limits_{t=1}^{T} \psi(t,h)\tilde{X}_t}{\sum\limits_{t=1}^{T} \psi(t,h)}} \tag{56}$$

(b) **Update Σ**:

Let's denote $\Omega = (\Omega_h)_{h\in[\![1,M]\!]} = (\Sigma_h^{-1})_{h\in[\![1,M]\!]}$
So J becomes a function of μ and Ω:

$$J(\mu, \Omega) = -\sum_{t=1}^{T-1}\sum_{h=1}^{M} \log(\mathcal{N}(\tilde{X}_t; \mu_h, \Omega_h))\psi(t,h) \tag{57}$$

We also have for $h \in [\![1, M]\!]$ and $t \in [\![1, T]\!]$:

$$\log(\mathcal{N}(\tilde{X}_t; \mu_h, \Omega_h)) = \frac{D}{2}\log(2\pi) - \frac{1}{2}\log(\det(\Omega_h)) - \frac{1}{2}(\tilde{X}_t - \mu_h)^T\Omega_h(\tilde{X}_t - \mu_h) \tag{58}$$

Thus, for $h \in [\![1, M]\!]$:

$$\nabla_{\Omega_h} J(\mu, \Omega) = \nabla_{\Omega_h} \left(- \sum_{t=1}^{T-1} \log(\mathcal{N}(\tilde{X}_t; \mu_h, \Omega_h))\psi(t, h) \right) \tag{59}$$

$$= - \sum_{t=1}^{T-1} \psi(t, h) \left(-\frac{1}{2} \nabla_{\Omega_h} \underbrace{\log(\det(\Omega_h)}_{:=u(\Omega_h)} + \frac{1}{2} \nabla_{\Omega_h} \underbrace{(\tilde{X}_t - \mu_h)^T \Omega_h (\tilde{X}_t - \mu_h)}_{:=v(\Omega_h)} \right) \tag{60}$$

Let's calculate $\nabla_{\Omega_h} u(\Omega_h)$

$$\begin{aligned}
\forall H \in \mathbb{R}^{D \times D} \quad u(\Omega_h + H) - u(\Omega_h) &= \log(\det(\Omega_h + H)) - \log(\det(\Omega_h)) \\
&= \log(\det(\Omega_h^{\frac{1}{2}} (I_D + \Omega_h^{\frac{-1}{2}} H \Omega_h^{\frac{-1}{2}}) \Omega_h^{\frac{1}{2}})) - \log(\det(\Omega_h)) \\
&= \log(\det(\Omega_h)) + \log(\det(I_D + \Omega_h^{\frac{-1}{2}} H \Omega_h^{\frac{-1}{2}})) - \log(\det(\Omega_h)) \\
&= \log(\det(I_D + \Omega_h^{\frac{-1}{2}} H \Omega_h^{\frac{-1}{2}}))
\end{aligned} \tag{61}$$

$$\tag{62}$$

We can decompose $\Omega_h^{\frac{-1}{2}} H \Omega_h^{\frac{-1}{2}}$ as follows:

$$\Omega_h^{\frac{-1}{2}} H \Omega_h^{\frac{-1}{2}} = U \begin{pmatrix} \omega_1 & & \\ & \ddots & \\ & & \omega_D \end{pmatrix} U^T$$

The equation 61 becomes:

$$\forall H \in \mathbb{R}^{D \times D} \quad u(\Omega_h + H) - u(\Omega_h) = \sum_{i=1}^{D} \log(1 + \omega_i) \tag{63}$$

$$= \sum_{i=1}^{D} \omega_i + o(||H||) \tag{64}$$

$$= \mathrm{tr}(\Omega_h^{\frac{-1}{2}} H \Omega_h^{\frac{-1}{2}}) + o(||H||) \tag{65}$$

$$= \underbrace{\mathrm{tr}(\Omega_h^{-1} H)}_{du_{\Omega_h}(H)} + o(||H||) \tag{66}$$

$$= \langle \nabla_{\Omega_h} u(\Omega_h), H \rangle + o(||H||) \tag{67}$$

$$\tag{68}$$

Therefore,

$$\nabla_{\Omega_h} u(\Omega_h) = \Omega_h^{-1} = \Sigma_h \tag{69}$$

Let's calculate $\nabla_{\Omega_h} v(\Omega_h)$. We have:

$$\forall H \in \mathbb{R}^{D\times D} \quad v(\Omega_h + H) - v(\Omega_h) = (\tilde{X}_t - \mu_h)^T(\Omega_h + H)(\tilde{X}_t - \mu_h) - (\tilde{X}_t - \mu_h)^T\Omega_h(\tilde{X}_t - \mu_h) \tag{70}$$

$$= \mathrm{tr}((\tilde{X}_t - \mu_h)^T(\Omega_h + H)(\tilde{X}_t - \mu_h)) - \mathrm{tr}((\tilde{X}_t - \mu_h)^T\Omega_h(\tilde{X}_t - \mu \tag{71}$$

$$= \mathrm{tr}(H(\tilde{X}_t - \mu_h)^T(\tilde{X}_t - \mu_h)) \tag{72}$$

$$= \mathrm{tr}((\tilde{X}_t - \mu_h)^T(\tilde{X}_t - \mu_h)H) \tag{73}$$

$$= \underbrace{\mathrm{tr}(((\tilde{X}_t - \mu_h)(\tilde{X}_t - \mu_h)^T)^T H)}_{dv_{\Omega_h}(H)} \tag{74}$$

$$= \langle \nabla_{\Omega_h} v(\Omega_h), H\rangle \tag{75}$$

$$\tag{76}$$

Therefore,

$$\nabla_{\Omega_h} v(\Omega_h) = (\tilde{X}_t - \mu_h)(\tilde{X}_t - \mu_h)^T \tag{77}$$

From the equation 60, 69 and 77, we conclude:

$$\nabla_{\Omega_h} J(\mu, \Omega) = -\sum_{t=1}^{T-1} \psi(t,h)\left(-\frac{1}{2}\Sigma_h + \frac{1}{2}(\tilde{X}_t - \mu_h)(\tilde{X}_t - \mu_h)^T\right) \tag{78}$$

We set the gradient with respect to Σ to zero, we obtain,

$$0 = \nabla_{\mu_h} J(\mu, \Sigma_h^{(i+1)}) = -\sum_{t=1}^{T-1} \psi(t,h)\left(-\frac{1}{2}\Sigma_h^{(i+1)} + \frac{1}{2}(\tilde{X}_t - \mu_h)(\tilde{X}_t - \mu_h)^T\right) \tag{79}$$

Finally,

$$\boxed{\forall h \in [\![1, M]\!] \quad \Sigma_h^{(i+1)} = \frac{\sum_{t=1}^{T} \psi(t,h)(\tilde{X}_t - \mu_h)(\tilde{X}_t - \mu_h)^T}{\sum_{t=1}^{T} \psi(t,h)}} \tag{80}$$