Deriving the E-Step and M-Step for Hidden Markov Models

From First Principles to Update Equations

A step-by-step mathematical derivation with intuition, connecting each formula to the project's implementation

Contents

1 The Problem: Why Do We Need EM?

1.1 What We Want

We have observed data $\mathbf{O} = (O_1, \ldots, O_T)$ (daily sector returns) and we want to find the HMM parameters $\lambda = (\mathbf{A}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\pi})$ that **maximise the marginal likelihood**:

$$\hat{\lambda} = \arg\max_{\lambda} \; P(\mathbf{O} \mid \lambda) \tag{1}$$

1.2 Why Direct Maximisation Fails

The marginal likelihood requires summing over all hidden state sequences:

$$P(\mathbf{O} \mid \lambda) = \sum_{\mathbf{Q}} P(\mathbf{O}, \mathbf{Q} \mid \lambda) \tag{2}$$

where the sum is over all K^T possible state sequences $\mathbf{Q} = (q_1, \ldots, q_T)$.
If we try to take the derivative and set it to zero:

$$\frac{\partial}{\partial \lambda} \log \underbrace{\sum_{\mathbf{Q}} P(\mathbf{O}, \mathbf{Q} \mid \lambda)}_{\text{sum of exponentials}} = 0$$

The log of a sum has no closed-form solution. The parameters are entangled inside the summation in a way that prevents direct optimisation.

The Core Difficulty

If we *knew* the hidden states $\mathbf{Q}$, maximisation would be trivial (just count transitions and compute sample means). But $\mathbf{Q}$ is unknown. If we *knew* the parameters λ, we could infer $\mathbf{Q}$. But λ is also unknown. This chicken-and-egg problem is exactly what EM resolves.

2 The EM Framework: The $\mathcal{Q}$-Function

2.1 The Key Idea

Instead of maximising $\log P(\mathbf{O} \mid \lambda)$ directly, EM works with the **complete-data log-likelihood** $\log P(\mathbf{O}, \mathbf{Q} \mid \lambda)$, which is much easier to handle because the log can be pushed inside the product.

Since $\mathbf{Q}$ is unknown, we take its **expectation** under the posterior $P(\mathbf{Q} \mid \mathbf{O}, \lambda^{\text{old}})$ computed using our current best guess λ^{old}.

The $\mathcal{Q}$-Function (Auxiliary Function)

$$\boxed{\mathcal{Q}(\lambda,\, \lambda^{\text{old}}) = \mathbb{E}_{\mathbf{Q}|\mathbf{O},\lambda^{\text{old}}}\Big[\log P(\mathbf{O}, \mathbf{Q} \mid \lambda)\Big] = \sum_{\mathbf{Q}} P(\mathbf{Q} \mid \mathbf{O}, \lambda^{\text{old}}) \log P(\mathbf{O}, \mathbf{Q} \mid \lambda)} \tag{3}$$

- λ^{old}: current parameters (used to compute the expectation)
- λ: new parameters (what we optimise over)

2.2 Why Maximising $\mathcal{Q}$ Works

Theorem 2.1 (EM Monotonicity). *If $\lambda^{new} = \arg\max_\lambda \mathcal{Q}(\lambda, \lambda^{old})$, then:*

$$\log P(\mathbf{O} \mid \lambda^{new}) \geq \log P(\mathbf{O} \mid \lambda^{old})$$

Each EM iteration is guaranteed to increase (or maintain) the marginal log-likelihood.

Proof Sketch (Jensen's Inequality)

Write:

$$\begin{aligned}
\log P(\mathbf{O} \mid \lambda) &= \log \sum_{\mathbf{Q}} P(\mathbf{O}, \mathbf{Q} \mid \lambda) \\
&= \log \sum_{\mathbf{Q}} P(\mathbf{Q} \mid \mathbf{O}, \lambda^{\text{old}}) \frac{P(\mathbf{O}, \mathbf{Q} \mid \lambda)}{P(\mathbf{Q} \mid \mathbf{O}, \lambda^{\text{old}})} \\
&\geq \sum_{\mathbf{Q}} P(\mathbf{Q} \mid \mathbf{O}, \lambda^{\text{old}}) \log \frac{P(\mathbf{O}, \mathbf{Q} \mid \lambda)}{P(\mathbf{Q} \mid \mathbf{O}, \lambda^{\text{old}})} \quad \text{(Jensen's inequality: log is concave)} \\
&= \underbrace{\sum_{\mathbf{Q}} P(\mathbf{Q} \mid \mathbf{O}, \lambda^{\text{old}}) \log P(\mathbf{O}, \mathbf{Q} \mid \lambda)}_{\mathcal{Q}(\lambda, \lambda^{\text{old}})} - \underbrace{\sum_{\mathbf{Q}} P(\mathbf{Q} \mid \mathbf{O}, \lambda^{\text{old}}) \log P(\mathbf{Q} \mid \mathbf{O}, \lambda^{\text{old}})}_{\text{entropy; constant w.r.t. } \lambda}
\end{aligned}$$

The second term doesn't depend on λ, so maximising the lower bound over λ reduces to maximising $\mathcal{Q}(\lambda, \lambda^{\text{old}})$. Since the bound is tight at $\lambda = \lambda^{\text{old}}$, any increase in $\mathcal{Q}$ translates to an increase in $\log P(\mathbf{O} \mid \lambda)$.

3 Expanding the Complete-Data Log-Likelihood

3.1 The Joint Probability Factorisation

From the HMM's conditional independence assumptions:

$$P(\mathbf{O}, \mathbf{Q} \mid \lambda) = \pi_{q_1} \cdot b_{q_1}(O_1) \cdot \prod_{t=2}^{T} A_{q_{t-1},q_t} \cdot b_{q_t}(O_t) \tag{4}$$

Taking the logarithm:

Complete-Data Log-Likelihood

$$\log P(\mathbf{O}, \mathbf{Q} \mid \lambda) = \underbrace{\log \pi_{q_1}}_{\text{(I)}} + \underbrace{\sum_{t=2}^{T} \log A_{q_{t-1},q_t}}_{\text{(II)}} + \underbrace{\sum_{t=1}^{T} \log b_{q_t}(O_t)}_{\text{(III)}} \tag{5}$$

Why This is Easier

The log of a product became a **sum** of logs. Each term depends on the hidden state at one or two time steps only. This separability is what makes the M-step tractable. Compare with $\log P(\mathbf{O} \mid \lambda) = \log \sum_{\mathbf{Q}} P(\mathbf{O}, \mathbf{Q} \mid \lambda)$, where the log is *outside* the sum—completely intractable.

3.2 The $\mathcal{Q}$-Function in Terms of γ and ξ

Now we compute $\mathcal{Q}(\lambda, \lambda^{\text{old}}) = \mathbb{E}_{\mathbf{Q}|\mathbf{O},\lambda^{\text{old}}}[\log P(\mathbf{O}, \mathbf{Q} \mid \lambda)]$ by taking the expectation of each term in Eq. (5).

3.2.1 Introduce indicator variables

To handle the expectation formally, define indicator variables:

$$\mathbb{1}[q_t = i] = \begin{cases} 1 & \text{if state at time } t \text{ is } i \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Then:

$$\log \pi_{q_1} = \sum_{i=1}^{K} \mathbb{1}[q_1 = i] \log \pi_i \tag{7}$$

$$\sum_{t=2}^{T} \log A_{q_{t-1},q_t} = \sum_{t=2}^{T} \sum_{i=1}^{K} \sum_{j=1}^{K} \mathbb{1}[q_{t-1} = i, q_t = j] \log A_{ij} \tag{8}$$

$$\sum_{t=1}^{T} \log b_{q_t}(O_t) = \sum_{t=1}^{T} \sum_{k=1}^{K} \mathbb{1}[q_t = k] \log b_k(O_t) \tag{9}$$

3.2.2 Take the expectation

Under $P(\mathbf{Q} \mid \mathbf{O}, \lambda^{\text{old}})$, the expectations of the indicators are exactly γ and ξ:

This is Where γ and ξ Come From

$$\mathbb{E}\big[\mathbb{1}[q_t = i]\big] = P(q_t = i \mid \mathbf{O}, \lambda^{\text{old}}) = \gamma_t(i) \tag{10}$$

$$\mathbb{E}\big[\mathbb{1}[q_{t-1} = i, q_t = j]\big] = P(q_{t-1} = i, q_t = j \mid \mathbf{O}, \lambda^{\text{old}}) = \xi_t(i,j) \tag{11}$$

This is why the E-step computes γ and ξ: they are exactly the expected sufficient statistics needed to evaluate $\mathcal{Q}$.

Substituting:

$\mathcal{Q}$-Function Expanded

$$\mathcal{Q}(\lambda, \lambda^{\text{old}}) = \underbrace{\sum_{i=1}^{K} \gamma_1(i) \log \pi_i}_{\text{(I): initial state}} + \underbrace{\sum_{t=2}^{T} \sum_{i=1}^{K} \sum_{j=1}^{K} \xi_t(i,j) \log A_{ij}}_{\text{(II): transitions}} + \underbrace{\sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_t(k) \log b_k(O_t)}_{\text{(III): emissions}} \tag{12}$$

The Crucial Observation

The three terms in $\mathcal{Q}$ involve **different parameters**: Term (I) involves only $\boldsymbol{\pi}$, Term (II) involves only $\mathbf{A}$, and Term (III) involves only $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$. We can **maximise each term independently**.

4 Deriving the E-Step: Why γ and ξ Take Those Forms

We've just shown that the E-step *needs* $\gamma_t(i)$ and $\xi_t(i,j)$. Now let us derive their formulas.

4.1 Deriving $\gamma_t(i)$

We want:

$$\gamma_t(i) = P(q_t = i \mid \mathbf{O}, \lambda)$$

Derivation of $\gamma_t(i)$

By Bayes' theorem:

$$\gamma_t(i) = P(q_t = i \mid \mathbf{O}, \lambda) = \frac{P(q_t = i, \mathbf{O} \mid \lambda)}{P(\mathbf{O} \mid \lambda)} \tag{13}$$

Now split $\mathbf{O}$ into past observations $(O_1, \ldots, O_t)$ and future observations $(O_{t+1}, \ldots, O_T)$:

$$P(q_t = i, \mathbf{O} \mid \lambda) = P(O_1, \ldots, O_t, q_t = i, O_{t+1}, \ldots, O_T \mid \lambda) \tag{14}$$

$$= \underbrace{P(O_1, \ldots, O_t, q_t = i \mid \lambda)}_{\alpha_t(i)} \cdot \underbrace{P(O_{t+1}, \ldots, O_T \mid q_t = i, \lambda)}_{\beta_t(i)} \tag{15}$$

The second equality uses the fact that, given $q_t = i$, the future observations are conditionally independent of the past (by the Markov property and output independence). This is the product $\alpha_t(i) \cdot \beta_t(i)$.
The denominator is just the total probability of the data:

$$P(\mathbf{O} \mid \lambda) = \sum_{j=1}^{K} P(q_t = j, \mathbf{O} \mid \lambda) = \sum_{j=1}^{K} \alpha_t(j)\, \beta_t(j) \tag{16}$$

Substituting into Eq. (13):

$$\boxed{\gamma_t(i) = \frac{\alpha_t(i)\, \beta_t(i)}{\sum_{j=1}^{K} \alpha_t(j)\, \beta_t(j)}} \tag{17}$$

Intuition

$\alpha_t(i)$ summarises all evidence from the **past** that we're in state i at time t. $\beta_t(i)$ summarises all evidence from the **future**. Multiplying them combines both directions of evidence. Dividing by the total normalises to a proper probability.

4.2 Deriving $\xi_t(i,j)$

We want:

$$\xi_t(i,j) = P(q_t = i, q_{t+1} = j \mid \mathbf{O}, \lambda)$$

Derivation of $\xi_t(i,j)$

Again, by Bayes' theorem:

$$\xi_t(i,j) = \frac{P(q_t = i, q_{t+1} = j, \mathbf{O} \mid \lambda)}{P(\mathbf{O} \mid \lambda)} \tag{18}$$

Expand the numerator by splitting $\mathbf{O}$ at time t:

$$\begin{aligned} &P(q_t = i, q_{t+1} = j, \mathbf{O} \mid \lambda) \\ &= P(O_1, \ldots, O_t, q_t = i \mid \lambda) \cdot P(q_{t+1} = j \mid q_t = i, \lambda) \cdot P(O_{t+1} \mid q_{t+1} = j, \lambda) \\ &\quad \cdot P(O_{t+2}, \ldots, O_T \mid q_{t+1} = j, \lambda) \end{aligned} \tag{19}$$

Identifying each factor:

$$= \underbrace{\alpha_t(i)}_{\text{past evidence}} \cdot \underbrace{A_{ij}}_{\text{transition}} \cdot \underbrace{b_j(O_{t+1})}_{\text{emission at } t+1} \cdot \underbrace{\beta_{t+1}(j)}_{\text{future evidence}} \tag{20}$$

The denominator is $P(\mathbf{O} \mid \lambda) = \sum_{i'} \sum_{j'} \alpha_t(i') A_{i'j'} b_{j'}(O_{t+1}) \beta_{t+1}(j')$.
Therefore:

$$\boxed{\xi_t(i,j) = \frac{\alpha_t(i)\ A_{ij}\ b_j(O_{t+1})\ \beta_{t+1}(j)}{\sum_{i'=1}^{K} \sum_{j'=1}^{K} \alpha_t(i')\ A_{i'j'}\ b_{j'}(O_{t+1})\ \beta_{t+1}(j')}} \tag{21}$$

Intuition for ξ

To compute the probability of the specific transition $i \to j$ at time t, we need four things:

1. $\alpha_t(i)$: evidence from days $1, \ldots, t$ that we're in state i
2. A_{ij}: the probability of transitioning from i to j
3. $b_j(O_{t+1})$: how well state j explains tomorrow's observation
4. $\beta_{t+1}(j)$: evidence from days $t+2, \ldots, T$ that we were in state j at $t+1$

Multiply all four, then normalise.

4.3 Consistency Check: γ from ξ

Notice that $\gamma_t(i)$ is the **marginal** of $\xi_t(i,j)$:

$$\gamma_t(i) = \sum_{j=1}^{K} \xi_t(i,j) \tag{22}$$

This makes sense: the probability of being in state i at time t equals the sum of probabilities of being in state i and transitioning to *any* state j.

5 Deriving the M-Step: The Update Equations

We now maximise $\mathcal{Q}(\lambda, \lambda^{\text{old}})$ from Eq. (12) with respect to each group of parameters. Since the three terms are separable, we handle each independently.

5.1 Deriving $\hat{A}_{ij}$: The Transition Matrix Update

We maximise Term (II) of $\mathcal{Q}$:

$$\max_{\mathbf{A}} \sum_{t=2}^{T} \sum_{i=1}^{K} \sum_{j=1}^{K} \xi_t(i,j) \log A_{ij} \quad \text{subject to} \quad \sum_{j=1}^{K} A_{ij} = 1 \;\; \forall i$$

Derivation via Lagrange Multipliers

The constraint is that each row of $\mathbf{A}$ sums to 1. We introduce a Lagrange multiplier λ_i for each row i:

$$\mathcal{L} = \sum_{t=2}^{T} \sum_{i=1}^{K} \sum_{j=1}^{K} \xi_t(i,j) \log A_{ij} - \sum_{i=1}^{K} \lambda_i \left(\sum_{j=1}^{K} A_{ij} - 1 \right) \tag{23}$$

Take the derivative with respect to A_{ij} and set to zero:

$$\frac{\partial \mathcal{L}}{\partial A_{ij}} = \frac{\sum_{t=2}^{T} \xi_t(i,j)}{A_{ij}} - \lambda_i = 0 \tag{24}$$

Solving for A_{ij}:

$$A_{ij} = \frac{\sum_{t=2}^{T} \xi_t(i,j)}{\lambda_i} \tag{25}$$

Now determine λ_i using the constraint $\sum_j A_{ij} = 1$:

$$\sum_{j=1}^{K} A_{ij} = \frac{\sum_{j=1}^{K} \sum_{t=2}^{T} \xi_t(i,j)}{\lambda_i} = 1 \qquad \Rightarrow \qquad \lambda_i = \sum_{j=1}^{K} \sum_{t=2}^{T} \xi_t(i,j) = \sum_{t=2}^{T} \gamma_t(i) \tag{26}$$

where we used $\sum_j \xi_t(i,j) = \gamma_t(i)$. Substituting back into Eq. (25):

$$\boxed{\hat{A}_{ij} = \frac{\sum_{t=2}^{T} \xi_t(i,j)}{\sum_{t=2}^{T} \gamma_t(i)} = \frac{\text{expected number of transitions } i \to j}{\text{expected number of times in state } i}} \tag{27}$$

Intuition: Relative Frequency

If the soft counts say: "state Bull appeared 100 times (expected), and 90 of those transitioned to Bull, 7 to Sideways, 3 to Bear," then:

$$\hat{A}_{\text{Bull,Bull}} = \frac{90}{100} = 0.90, \quad \hat{A}_{\text{Bull,Side}} = \frac{7}{100} = 0.07, \quad \hat{A}_{\text{Bull,Bear}} = \frac{3}{100} = 0.03$$

This is the same formula as the maximum likelihood estimate for a multinomial, but with **soft** counts from γ and ξ instead of hard counts.

5.2 Deriving $\hat{\pi}$: The Initial Distribution Update

The derivation is nearly identical. Maximise Term (I):

$$\max_{\boldsymbol{\pi}} \sum_{i=1}^{K} \gamma_1(i) \log \pi_i \quad \text{subject to} \quad \sum_{i=1}^{K} \pi_i = 1$$

Derivation

Lagrangian: $\mathcal{L} = \sum_i \gamma_1(i) \log \pi_i - \mu \left(\sum_i \pi_i - 1 \right)$
Taking derivative and setting to zero:

$$\frac{\partial \mathcal{L}}{\partial \pi_i} = \frac{\gamma_1(i)}{\pi_i} - \mu = 0 \quad \Rightarrow \quad \pi_i = \frac{\gamma_1(i)}{\mu}$$

Using $\sum_i \pi_i = 1$: $\mu = \sum_i \gamma_1(i) = 1$, so:

$$\boxed{\hat{\pi}_i = \gamma_1(i)} \tag{28}$$

This is intuitive: the probability of starting in state i is simply the posterior probability of being in state i at $t = 1$.

5.3 Deriving $\hat{\mu}_k$: The Emission Mean Update

Now we maximise Term (III). For Gaussian emissions, $b_k(O_t) = \mathcal{N}(\mathbf{r}_t; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, so:

$$\log b_k(\mathbf{r}_t) = -\frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\mathbf{r}_t - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{r}_t - \boldsymbol{\mu}_k) \tag{29}$$

Term (III) of $\mathcal{Q}$, collecting only terms involving $\boldsymbol{\mu}_k$ for a specific regime k:

$$\mathcal{Q}_k^{(\mu)} = -\frac{1}{2}\sum_{t=1}^{T}\gamma_t(k)(\mathbf{r}_t - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{r}_t - \boldsymbol{\mu}_k) + \text{const} \tag{30}$$

Derivation of $\hat{\mu}_k$

Take the derivative with respect to $\boldsymbol{\mu}_k$ and set to zero.
Using the matrix calculus identity $\frac{\partial}{\partial \boldsymbol{\mu}}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{M}(\mathbf{x}-\boldsymbol{\mu}) = -2\mathbf{M}(\mathbf{x}-\boldsymbol{\mu})$:

$$\frac{\partial \mathcal{Q}_k^{(\mu)}}{\partial \boldsymbol{\mu}_k} = -\frac{1}{2}\sum_{t=1}^{T}\gamma_t(k)\cdot(-2)\boldsymbol{\Sigma}_k^{-1}(\mathbf{r}_t - \boldsymbol{\mu}_k) = \mathbf{0} \tag{31}$$

$$= \sum_{t=1}^{T}\gamma_t(k)\,\boldsymbol{\Sigma}_k^{-1}(\mathbf{r}_t - \boldsymbol{\mu}_k) = \mathbf{0} \tag{32}$$

Since $\boldsymbol{\Sigma}_k^{-1}$ is non-singular, we can multiply both sides by $\boldsymbol{\Sigma}_k$:

$$\sum_{t=1}^{T}\gamma_t(k)(\mathbf{r}_t - \boldsymbol{\mu}_k) = \mathbf{0} \tag{33}$$

$$\sum_{t=1}^{T}\gamma_t(k)\,\mathbf{r}_t = \boldsymbol{\mu}_k\sum_{t=1}^{T}\gamma_t(k) \tag{34}$$

Solving for $\boldsymbol{\mu}_k$:

$$\boxed{\hat{\boldsymbol{\mu}}_k = \frac{\sum_{t=1}^{T}\gamma_t(k)\,\mathbf{r}_t}{\sum_{t=1}^{T}\gamma_t(k)}} \tag{35}$$

Intuition: Weighted Average

$\hat{\boldsymbol{\mu}}_k$ is the weighted average of all observations, where the weight of observation $\mathbf{r}_t$ is $\gamma_t(k)$—the probability that $\mathbf{r}_t$ was generated by regime k.

Concrete example: Suppose we have 3 days of returns and $K = 2$ regimes:

t	$\mathbf{r}_t$	γ_t(Bull)	γ_t(Bear)
1	$+0.5\%$	0.9	0.1
2	$+0.3\%$	0.8	0.2
3	-0.4%	0.1	0.9

$$\hat{\mu}_{\text{Bull}} = \frac{0.9\times 0.5 + 0.8\times 0.3 + 0.1\times(-0.4)}{0.9+0.8+0.1} = \frac{0.45+0.24-0.04}{1.8} = \frac{0.65}{1.8} = +0.361\%$$

$$\hat{\mu}_{\text{Bear}} = \frac{0.1\times 0.5 + 0.2\times 0.3 + 0.9\times(-0.4)}{0.1+0.2+0.9} = \frac{0.05+0.06-0.36}{1.2} = \frac{-0.25}{1.2} = -0.208\%$$

Day 3 (-0.4%) barely affects $\hat{\mu}_{\text{Bull}}$ (weight 0.1) but dominates $\hat{\mu}_{\text{Bear}}$ (weight 0.9). The soft assignments naturally route each observation to the regime that likely generated it.

5.4 Deriving $\hat{\Sigma}_k$: The Emission Covariance Update

Collecting terms from $\mathcal{Q}$ that involve $\boldsymbol{\Sigma}_k$:

$$\mathcal{Q}_k^{(\Sigma)} = -\frac{1}{2}\sum_{t=1}^{T}\gamma_t(k)\left[\log|\boldsymbol{\Sigma}_k| + (\mathbf{r}_t - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{r}_t - \boldsymbol{\mu}_k)\right] + \text{const} \tag{36}$$

Derivation of $\hat{\Sigma}_k$

We use two matrix calculus identities. Let $\mathbf{S} = \boldsymbol{\Sigma}_k^{-1}$ (the precision matrix):

1. $\frac{\partial}{\partial \mathbf{S}}\log|\mathbf{S}^{-1}| = -\mathbf{S}^{-1}$, equivalently $\frac{\partial}{\partial \boldsymbol{\Sigma}_k}\log|\boldsymbol{\Sigma}_k| = \boldsymbol{\Sigma}_k^{-1}$
2. $\frac{\partial}{\partial \boldsymbol{\Sigma}_k^{-1}}\mathbf{x}^T\boldsymbol{\Sigma}_k^{-1}\mathbf{x} = \mathbf{x}\mathbf{x}^T$

It is cleaner to differentiate with respect to $\mathbf{S} = \boldsymbol{\Sigma}_k^{-1}$. Rewriting $\mathcal{Q}_k^{(\Sigma)}$:

$$\mathcal{Q}_k^{(\Sigma)} = -\frac{1}{2}\sum_{t=1}^{T}\gamma_t(k)\Big[-\log|\mathbf{S}| + (\mathbf{r}_t - \boldsymbol{\mu}_k)^T\mathbf{S}\,(\mathbf{r}_t - \boldsymbol{\mu}_k)\Big] \tag{37}$$

Taking the derivative with respect to $\mathbf{S}$ and setting to zero:

$$\frac{\partial \mathcal{Q}_k^{(\Sigma)}}{\partial \mathbf{S}} = -\frac{1}{2}\sum_{t=1}^{T}\gamma_t(k)\Big[-\mathbf{S}^{-1} + (\mathbf{r}_t - \boldsymbol{\mu}_k)(\mathbf{r}_t - \boldsymbol{\mu}_k)^T\Big] = \mathbf{0} \tag{38}$$

$$\sum_{t=1}^{T}\gamma_t(k)\,\mathbf{S}^{-1} = \sum_{t=1}^{T}\gamma_t(k)(\mathbf{r}_t - \boldsymbol{\mu}_k)(\mathbf{r}_t - \boldsymbol{\mu}_k)^T \tag{39}$$

$$\boldsymbol{\Sigma}_k \cdot \sum_{t=1}^{T}\gamma_t(k) = \sum_{t=1}^{T}\gamma_t(k)(\mathbf{r}_t - \boldsymbol{\mu}_k)(\mathbf{r}_t - \boldsymbol{\mu}_k)^T \tag{40}$$

Solving:

$$\boxed{\hat{\boldsymbol{\Sigma}}_k = \frac{\sum_{t=1}^{T}\gamma_t(k)\,(\mathbf{r}_t - \hat{\boldsymbol{\mu}}_k)(\mathbf{r}_t - \hat{\boldsymbol{\mu}}_k)^T}{\sum_{t=1}^{T}\gamma_t(k)}} \tag{41}$$

Intuition: Weighted Sample Covariance

This is the standard sample covariance formula, but with soft weights $\gamma_t(k)$. Only observations that "belong to" regime k (high $\gamma_t(k)$) contribute significantly to that regime's covariance estimate.

6 Connection to the Implementation

All of the above is executed internally by `hmmlearn` when `model.fit(X)` is called (`models.py`, line 174). The project wraps this with multiple initialisations:

```
# models.py, lines 161-183
for i in range(n_init):
    model = hmm.GaussianHMM(
        n_components=self.n_regimes,      # K = 3
        covariance_type=self.covariance_type,  # 'full'
        n_iter=self.n_iter,               # max 100 EM iterations
        tol=self.tol,                     # convergence threshold 1e-4
    )
    model.fit(X)                          # <-- EM runs here!
    #   Internally:
    #     for iter in range(n_iter):
    #       alpha, beta = forward_backward(X)  --> E-step
    #       gamma, xi = compute_posteriors(alpha, beta)
    #       update A, mu, Sigma using gamma, xi  --> M-step
    #       if improvement < tol: break

    self._regularize_covariance(model)    # Sigma_k += lambda * I
    score = model.score(X)                # final log P(O|lambda)
```

After EM finishes, the learned parameters are stored as:

```
# models.py, lines 321-340
model.transmat_    # A_hat: the transition matrix
model.means_       # mu_hat: emission means (K x D array)
model.covars_      # Sigma_hat: emission covariances (K x D x D)
```

Then the covariance regularisation (`models.py`, lines 104-120) applies the stabilising correction:

```
# After EM, prevent singular covariance:
for k in range(self.n_regimes):
    n_features = model.covars_[k].shape[0]
    model.covars_[k] += self.regularization * np.eye(n_features)
    # This is: Sigma_k <-- Sigma_k + lambda * I
```

7 Summary: The Complete Derivation Map

The Logical Flow of the Entire Derivation

1. We want $\hat{\lambda} = \arg\max_\lambda \log P(\mathbf{O} \mid \lambda)$ but can't optimise directly
2. Jensen's inequality gives us a tractable lower bound: the $\mathcal{Q}$-function
3. Writing out $\mathcal{Q}$ requires the complete-data log-likelihood $\log P(\mathbf{O}, \mathbf{Q} \mid \lambda)$
4. Taking expectation over $\mathbf{Q}$ introduces two sufficient statistics:
   - $\gamma_t(i) = \mathbb{E}[\mathbb{1}(q_t = i)]$ — derived via Bayes' rule from $\alpha_t(i)$ and $\beta_t(i)$
   - $\xi_t(i,j) = \mathbb{E}[\mathbb{1}(q_t = i, q_{t+1} = j)]$ — derived similarly
5. **This is the E-step**: compute γ and ξ using forward-backward with current λ^{old}
6. $\mathcal{Q}$ separates into three independent terms (initial, transitions, emissions)
7. Maximising each via Lagrange multipliers / calculus gives:
   - $\hat{A}_{ij}$: ratio of expected transitions $i \to j$ to expected time in state i
   - $\hat{\boldsymbol{\mu}}_k$: γ-weighted average of observations
   - $\hat{\boldsymbol{\Sigma}}_k$: γ-weighted sample covariance
8. **This is the M-step**: update parameters using the formulas above
9. Iterate E-step $\to$ M-step until convergence

7.1 All Update Equations at a Glance

Complete Set of EM Update Equations for Gaussian HMM

$$\hat{\pi}_i = \gamma_1(i) \qquad \text{(Initial)}$$

$$\hat{A}_{ij} = \frac{\sum_{t=2}^{T} \xi_t(i,j)}{\sum_{t=2}^{T} \gamma_t(i)} \qquad \text{(Transitions)}$$

$$\hat{\boldsymbol{\mu}}_k = \frac{\sum_{t=1}^{T} \gamma_t(k)\, \mathbf{r}_t}{\sum_{t=1}^{T} \gamma_t(k)} \qquad \text{(Means)}$$

$$\hat{\boldsymbol{\Sigma}}_k = \frac{\sum_{t=1}^{T} \gamma_t(k)\, (\mathbf{r}_t - \hat{\boldsymbol{\mu}}_k)(\mathbf{r}_t - \hat{\boldsymbol{\mu}}_k)^T}{\sum_{t=1}^{T} \gamma_t(k)} \qquad \text{(Covariances)}$$

where:

$$\gamma_t(i) = \frac{\alpha_t(i)\, \beta_t(i)}{\sum_j \alpha_t(j)\, \beta_t(j)} \qquad \text{(E-step 1)}$$

$$\xi_t(i,j) = \frac{\alpha_t(i)\, A_{ij}\, b_j(O_{t+1})\, \beta_{t+1}(j)}{\sum_{i'} \sum_{j'} \alpha_t(i')\, A_{i'j'}\, b_{j'}(O_{t+1})\, \beta_{t+1}(j')} \qquad \text{(E-step 2)}$$