A Detailed and Intuitive Tutorial on the Baum–Welch Algorithm

1 Introduction

Hidden Markov Models (HMMs) are probabilistic models for sequential data where the system evolves through a sequence of *hidden states* that cannot be directly observed. Instead, each hidden state probabilistically generates an *observable output*.

The Baum–Welch algorithm is an Expectation–Maximization (EM) algorithm used to learn HMM parameters when only the observation sequences are available.

2 Hidden Markov Model Definition

An HMM is defined by the parameter set

$$\lambda = (A, B, \pi).$$

2.1 Hidden States

Let

$$Q = \{1, 2, \dots, N\}$$

be the set of hidden states. At any time, the system is in exactly one hidden state, but this state is not directly observable.

2.2 Observations

Let

$$O = \{O_1, O_2, \dots, O_M\}$$

be the set of observable symbols. Observations are the only data we see.

2.3 Model Parameters

- **Initial probabilities**

  $$\pi_i = P(q_1 = i)$$

  Meaning: belief that the system starts in state i.

- **Transition probabilities**

  $$a_{ij} = P(q_{t+1} = j \mid q_t = i)$$

  Meaning: how the system evolves internally.

- **Emission probabilities**

  $$b_i(o) = P(O_t = o \mid q_t = i)$$

  Meaning: how hidden states generate visible observations.

3 Observation Sequence and Indexing

Let

$$O = (O_1, O_2, \dots, O_T)$$

be an observation sequence of length T.

- t: position in the sequence
- T: total number of observations

A longer sequence provides more evidence and reduces statistical uncertainty.

4 Forward Probability (Prefix Evidence)

4.1 Definition

$$\alpha_t(i) = P(O_1, O_2, \dots, O_t, q_t = i \mid \lambda)$$

Semantic meaning: The probability that the first t observations have occurred and the system is currently in state i.

4.2 Recurrence Relations

4.2.1 Initialization

$$\alpha_1(i) = \pi_i \, b_i(O_1)$$

This combines initial belief with compatibility of the first observation.

4.2.2 Recursion

$$\alpha_{t+1}(j) = \left(\sum_{i=1}^{N} \alpha_t(i) a_{ij} \right) b_j(O_{t+1})$$

Data dependency:

- depends on the entire past through α_t,
- directly influenced by the current observation.

5 Backward Probability (Suffix Evidence)

5.1 Definition

$$\beta_t(i) = P(O_{t+1}, O_{t+2}, \ldots, O_T \mid q_t = i, \lambda)$$

Semantic meaning: If the system were in state i at time t, how well could it explain all future observations?

5.2 Recurrence Relations

5.2.1 Initialization

$$\beta_T(i) = 1$$

5.2.2 Recursion

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)$$

Data dependency: Backward probability summarizes the entire future into a single value.

6 Probability of the Observation Sequence

$$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

This measures how well the model explains the data.

7 Auxiliary Variables (Soft Credit Assignment)

7.1 State Responsibility

$$\gamma_t(i) = P(q_t = i \mid O, \lambda) = \frac{\alpha_t(i)\beta_t(i)}{P(O \mid \lambda)}$$

Meaning: Fraction of belief that the system was in state i at time t.

7.2 Transition Responsibility

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{P(O \mid \lambda)}$$

Meaning: Fraction of probability flow through transition $i \to j$.

8 Parameter Update Equations

8.1 Initial Probabilities

$$\pi_i^{new} = \gamma_1(i)$$

8.2 Transition Probabilities

$$a_{ij}^{new} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

8.3 Emission Probabilities

$$b_i(o)^{new} = \frac{\sum_{t:O_t=o} \gamma_t(i)}{\sum_{t=1}^{T} \gamma_t(i)}$$

Key idea: Emission probabilities are normalized expected counts.

9 Numerical Example Setup

9.1 States and Observations

States: Rainy (R), Sunny (S)
Observations: Walk (W), Shop (H)

9.2 Initial Parameters

$$\pi = (0.6, 0.4)$$

$$A = \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}$$

$$B = \begin{array}{c|cc} & W & H \\ \hline R & 0.1 & 0.9 \\ S & 0.6 & 0.4 \end{array}$$

10 Short Observation Sequence

$$O^{(short)} = (W, H)$$

10.1 Forward and Backward Computation

$$\alpha_1(R) = 0.06, \quad \alpha_1(S) = 0.24$$

$$\alpha_2(R) = 0.1242, \quad \alpha_2(S) = 0.0648$$

$$P(O) = 0.189$$

$$\beta_2(R) = \beta_2(S) = 1$$

$$\beta_1(R) = 0.75, \quad \beta_1(S) = 0.60$$

10.2 State Responsibilities

$$\gamma_1(R) = 0.238, \quad \gamma_1(S) = 0.762$$

$$\gamma_2(R) = 0.657, \quad \gamma_2(S) = 0.343$$

10.3 Emission Update

$$b_R(W) = \frac{0.238}{0.895} = 0.266 \quad b_R(H) = 0.734$$

$$b_S(W) = 0.689 \quad b_S(H) = 0.311$$

Observation: Estimates change sharply due to limited data.

11 Long Observation Sequence

$$O^{(long)} = (W, H, H, W, H)$$

11.1 State Responsibilities (Iteration 1)

Position	Observation	$\gamma(R)$	$\gamma(S)$
1	W	0.23	0.77
2	H	0.66	0.34
3	H	0.71	0.29
4	W	0.31	0.69
5	H	0.74	0.26

11.2 Emission Update

Rainy total responsibility:

$$2.65$$

Rainy emitting Walk:

$$b_R(W) = \frac{0.54}{2.65} = 0.204$$

Sunny emitting Walk:

$$b_S(W) = \frac{1.46}{2.35} = 0.621$$

11.3 Second Iteration Refinement

Using updated emissions, responsibilities stabilize. Example refinement:

$$b_R(W) : 0.204 \to 0.187$$

Changes become smaller, indicating convergence.

12 Why Longer Sequences Are Better

- More observations = more expected counts
- Variance reduces through averaging
- Emission probabilities converge smoothly
- Transitions are reinforced repeatedly

13 Final Intuition

Baum–Welch learns by repeatedly:

- propagating belief through recurrences,
- assigning soft credit using auxiliary variables,
- normalizing expected counts into probabilities.

Key takeaway:

> Observations do not update probabilities directly; they reshape belief flow, and belief flow updates the parameters.