Expectation–Maximization - Derivations & Proofs

Paul F. Roysdon, Ph.D.

Contents

1 Mathematical Derivations & Proofs

1.1 Introduction

Expectation–Maximization (EM) is a general-purpose algorithm for maximum-likelihood (ML) or maximum a posteriori (MAP) estimation in models with *latent* (unobserved) variables. The key difficulty in such models is the intractability of the marginal likelihood due to summation/integration over latent variables. EM alternates between (i) computing expectations of the complete-data log-likelihood under the posterior of the latents given current parameters (*E-step*) and (ii) maximizing this expected complete-data log-likelihood w.r.t. parameters (*M-step*). We derive EM from first principles by constructing a tight variational lower bound via Jensen's inequality and prove the monotone ascent property. We then instantiate the algorithm for Gaussian mixture models, deriving all parameter updates with explicit dimensions.

1.2 Data and Notation

Let the observed dataset be

$$\mathcal{D} = \{(\mathbf{x}_i)\}_{i=1}^{n}, \qquad \mathbf{x}_i \in \mathbb{R}^d.$$

Let $\mathbf{z}_i$ denote the latent variable associated with $\mathbf{x}_i$; its support may be discrete or continuous. Let $\boldsymbol{\theta}$ denote the model parameters (a vector collecting all free parameters). The joint model factorizes as

$$p(\mathbf{x}, \mathbf{z} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x}_i, \mathbf{z}_i \mid \boldsymbol{\theta}),$$

and the marginal (incomplete-data) likelihood is

$$\mathcal{L}(\boldsymbol{\theta}) \;=\; p(\mathcal{D} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \underbrace{\sum_{\mathbf{z}_i} p(\mathbf{x}_i, \mathbf{z}_i \mid \boldsymbol{\theta})}_{\text{sum/integral over } \mathbf{z}_i} .$$

We work with the log-likelihood

$$\ell(\boldsymbol{\theta}) \;=\; \log p(\mathcal{D} \mid \boldsymbol{\theta}) \;=\; \sum_{i=1}^{n} \log \sum_{\mathbf{z}_i} p(\mathbf{x}_i, \mathbf{z}_i \mid \boldsymbol{\theta}).$$

Dimensions and objects.

- $\mathbf{x}_i \in \mathbb{R}^d$: observed vector; n samples; d features.
- $\mathbf{z}_i$: latent variable (discrete indicator or continuous vector).
- $\boldsymbol{\theta}$: parameter vector (dimension depends on the model).
- $p(\mathbf{x}_i, \mathbf{z}_i \mid \boldsymbol{\theta})$: complete-data model.
- $p(\mathbf{z}_i \mid \mathbf{x}_i, \boldsymbol{\theta})$: posterior over latents (E-step distribution).

1.3 Model Formulation and the Difficulty

The ML estimator solves

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} \in \arg\max_{\boldsymbol{\theta}} \; \ell(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \; \sum_{i=1}^{n} \log \sum_{\mathbf{z}_i} p(\mathbf{x}_i, \mathbf{z}_i \mid \boldsymbol{\theta}).$$

The log of a sum couples the parameters across latent configurations, obstructing direct optimization. EM breaks this log-sum via a tight lower bound.

1.4 Variational Lower Bound via Jensen

Fix any set of distributions $\{q_i(\mathbf{z}_i)\}$ with $q_i(\mathbf{z}_i) \geq 0$ and $\sum_{\mathbf{z}_i} q_i(\mathbf{z}_i) = 1$. For each i,

$$\log \sum_{\mathbf{z}_i} p(\mathbf{x}_i, \mathbf{z}_i \mid \boldsymbol{\theta}) = \log \sum_{\mathbf{z}_i} q_i(\mathbf{z}_i) \, \frac{p(\mathbf{x}_i, \mathbf{z}_i \mid \boldsymbol{\theta})}{q_i(\mathbf{z}_i)} \;\geq\; \sum_{\mathbf{z}_i} q_i(\mathbf{z}_i) \, \log \frac{p(\mathbf{x}_i, \mathbf{z}_i \mid \boldsymbol{\theta})}{q_i(\mathbf{z}_i)},$$

by Jensen's inequality (concavity of log). Summing over i, we obtain the evidence lower bound (ELBO)

$$\ell(\boldsymbol{\theta}) \geq \mathcal{F}(\{q_i\}, \boldsymbol{\theta}) \;\triangleq\; \sum_{i=1}^{n} \sum_{\mathbf{z}_i} q_i(\mathbf{z}_i) \, \log p(\mathbf{x}_i, \mathbf{z}_i \mid \boldsymbol{\theta}) \;+\; \sum_{i=1}^{n} H(q_i), \tag{1}$$

where $H(q_i) = -\sum_{\mathbf{z}_i} q_i(\mathbf{z}_i) \log q_i(\mathbf{z}_i)$ is the entropy. The gap between $\ell(\boldsymbol{\theta})$ and the bound is a KL divergence:

$$\ell(\boldsymbol{\theta}) - \mathcal{F}(\{q_i\}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \mathrm{KL}\Big(q_i(\mathbf{z}_i) \,\Big\|\, p(\mathbf{z}_i \mid \mathbf{x}_i, \boldsymbol{\theta}) \Big) \;\geq\; 0, \tag{2}$$

with equality iff $q_i(\mathbf{z}_i) = p(\mathbf{z}_i \mid \mathbf{x}_i, \boldsymbol{\theta})$ for all i.

1.5 The EM Algorithm as Coordinate Ascent on the ELBO

EM alternates maximization of $\mathcal{F}$ in Eqn. (1) w.r.t. q and $\boldsymbol{\theta}$.

E-step (tighten the bound). Given current parameters $\boldsymbol{\theta}^{\text{old}}$, set

$$q_i^\star(\mathbf{z}_i) \;\leftarrow\; p(\mathbf{z}_i \mid \mathbf{x}_i, \boldsymbol{\theta}^{\text{old}}), \qquad i = 1, \dots, n.$$

By Eqn. (2), this choice maximizes $\mathcal{F}$ over $\{q_i\}$ and makes the bound tight:

$$\mathcal{F}(\{q_i^\star\}, \boldsymbol{\theta}^{\text{old}}) = \ell(\boldsymbol{\theta}^{\text{old}}).$$

M-step (raise the bound). With $q_i^\star$ fixed, maximize $\mathcal{F}$ over $\boldsymbol{\theta}$. The entropy terms $H(q_i^\star)$ do not depend on $\boldsymbol{\theta}$, so the M-step reduces to

$$\boldsymbol{\theta}^{\text{new}} \in \arg\max_{\boldsymbol{\theta}} \; Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{\text{old}}), \qquad Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{\text{old}}) \;\triangleq\; \sum_{i=1}^{n} \mathbb{E}_{\mathbf{z}_i \sim p(\cdot \mid \mathbf{x}_i, \boldsymbol{\theta}^{\text{old}})} \big[\log p(\mathbf{x}_i, \mathbf{z}_i \mid \boldsymbol{\theta}) \big]. \tag{3}$$

This is the *expected complete-data log-likelihood.*

1.6 Monotone Ascent

Proof. Let $\{q_i^\star\}$ be the E-step distributions at $\boldsymbol{\theta}^{\text{old}}$.

$$\ell(\boldsymbol{\theta}^{\text{new}}) \;\geq\; \mathcal{F}(\{q_i^\star\}, \boldsymbol{\theta}^{\text{new}}) \;\geq\; \mathcal{F}(\{q_i^\star\}, \boldsymbol{\theta}^{\text{old}}) \;=\; \ell(\boldsymbol{\theta}^{\text{old}}).$$

The first inequality is the ELBO property; the second holds because the M-step Eqn. (3) maximizes $\mathcal{F}$ over $\boldsymbol{\theta}$; the equality uses tightness at $\boldsymbol{\theta}^{\text{old}}$. Hence EM produces a non-decreasing sequence $\ell(\boldsymbol{\theta}^{(t)})$ and converges to a stationary point (local maximum or saddle) of the likelihood under mild regularity. ■

1.7 Optimal MAP-EM

With a prior $p(\boldsymbol{\theta})$, replace $\log p(\mathbf{x}_i, \mathbf{z}_i \mid \boldsymbol{\theta})$ by $\log p(\mathbf{x}_i, \mathbf{z}_i \mid \boldsymbol{\theta}) + \frac{1}{n} \log p(\boldsymbol{\theta})$ in Q (or equivalently add $\log p(\boldsymbol{\theta})$ to the M-step objective). The same derivation applies.

1.8 Example: Gaussian Mixture Model

Model. For K components with mixing weights π_k (nonnegative, $\sum_k \pi_k = 1$), means $\boldsymbol{\mu}_k \in \mathbb{R}^d$, covariances $\boldsymbol{\Sigma}_k \in \mathbb{R}^{d \times d}$ (symmetric positive definite), define latent indicators $\mathbf{z}_i \in \{1, \dots, K\}$. The complete-data model is

$$p(\mathbf{x}_i, \mathbf{z}_i = k \mid \boldsymbol{\theta}) = \pi_k \, \mathcal{N}\big(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\big), \quad \boldsymbol{\theta} = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K.$$

The marginal is the mixture $p(\mathbf{x}_i \mid \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k \, \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

E-step (responsibilities). Posterior component probabilities for each sample:

$$\gamma_{ik} \;\triangleq\; p(\mathbf{z}_i = k \mid \mathbf{x}_i, \boldsymbol{\theta}^{\text{old}}) \;=\; \frac{\pi_k^{\text{old}} \, \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k^{\text{old}}, \boldsymbol{\Sigma}_k^{\text{old}})}{\sum_{t=1}^K \pi_t^{\text{old}} \, \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_t^{\text{old}}, \boldsymbol{\Sigma}_t^{\text{old}})}, \qquad \sum_{k=1}^K \gamma_{ik} = 1. \tag{4}$$

M-step (weighted MLEs). Let effective counts $N_k = \sum_{i=1}^{n} \gamma_{ik}$.

$$\pi_k^{\text{new}} = \frac{N_k}{n}, \tag{5}$$

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k} \sum_{i=1}^{n} \gamma_{ik}\, \mathbf{x}_i, \tag{6}$$

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k} \sum_{i=1}^{n} \gamma_{ik}\, (\mathbf{x}_i - \boldsymbol{\mu}_k^{\text{new}})(\mathbf{x}_i - \boldsymbol{\mu}_k^{\text{new}})^\top. \tag{7}$$

Derivation (complete-data log-likelihood). Introduce one-hot indicators $z_{ik} = \mathbf{1}\{\mathbf{z}_i = k\}$. Then

$$\log p(\mathcal{D}, \{\mathbf{z}_i\} \mid \boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \Big(\log \pi_k - \tfrac{1}{2} \log |2\pi \boldsymbol{\Sigma}_k| - \tfrac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \Big).$$

Taking expectation over the posterior $p(\mathbf{z}_i \mid \mathbf{x}_i, \boldsymbol{\theta}^{\text{old}})$ replaces z_{ik} by γ_{ik}, yielding

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{\text{old}}) = \sum_{k=1}^{K} \left[\Big(\textstyle\sum_i \gamma_{ik} \Big) \log \pi_k - \tfrac{1}{2} \textstyle\sum_i \gamma_{ik} \Big(\log |2\pi \boldsymbol{\Sigma}_k| + (\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \Big) \right].$$

Maximizing w.r.t. π_k under $\sum_k \pi_k = 1$ via a Lagrange multiplier gives Eqn. (5). Setting $\partial Q / \partial \boldsymbol{\mu}_k = \mathbf{0}$ gives Eqn. (6). Setting $\partial Q / \partial \boldsymbol{\Sigma}_k = \mathbf{0}$ (using matrix calculus identities for Gaussian log-likelihoods) yields Eqn. (7). All updates are weighted MLEs with weights γ_{ik}.

Dimensions.

- $\gamma_{ik} \in [0, 1]$, $\sum_k \gamma_{ik} = 1$; $N_k \in [0, n]$; $\sum_k N_k = n$.
- $\boldsymbol{\mu}_k \in \mathbb{R}^d$; $\boldsymbol{\Sigma}_k \in \mathbb{R}^{d \times d}$ (SPD).
- $\pi_k \in [0, 1]$, $\sum_k \pi_k = 1$.

1.9 Other Canonical Instances

Mixture of Bernoulli vectors. For $\mathbf{x}_i \in \{0, 1\}^d$, component parameter $\boldsymbol{\eta}_k \in (0, 1)^d$ (independent Bernoulli per coordinate): E-step as in Eqn. (4); M-step gives $\eta_{kj}^{\text{new}} = \frac{1}{N_k} \sum_i \gamma_{ik} x_{ij}$.

Mixture of Poissons (counts). For $x_i \in \mathbb{N}$, component rate $\lambda_k > 0$: $\lambda_k^{\text{new}} = \frac{1}{N_k} \sum_i \gamma_{ik} x_i$.

All are weighted MLEs under the component family.

1.10 Generalized EM and Convergence

In some models the M-step maximization may be hard. Generalized EM (GEM) requires only that the M-step *increases* $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{\text{old}})$ (not necessarily to its global maximum). The same monotone ascent proof applies, since $\mathcal{F}$ increases each iteration.

1.11 Algorithm (Expectation–Maximization)

1. **Input:** data $\{\mathbf{x}_i\}_{i=1}^n$, latent-variable model $p(\mathbf{x}_i, \mathbf{z}_i \mid \boldsymbol{\theta})$.
2. **Initialize** $\boldsymbol{\theta}^{(0)}$ (e.g., via k-means for GMM).
3. For $t = 1, 2, \ldots$ until convergence:

(a) **E-step:** For each i, set $q_i^{(t)}(\mathbf{z}_i) = p(\mathbf{z}_i \mid \mathbf{x}_i, \boldsymbol{\theta}^{(t-1)})$.

(b) **M-step:** Set $\boldsymbol{\theta}^{(t)} \in \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \mathbb{E}_{q_i^{(t)}}[\log p(\mathbf{x}_i, \mathbf{z}_i \mid \boldsymbol{\theta})]$.

4. **Output:** parameter estimate $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(T)}$; optional posterior summaries $q_i^{(T)}$.

1.12 Connections and Interpretations

- **Variational view:** EM is coordinate ascent on the ELBO Eqn. (1) in the restricted family where the variational posterior q_i is *exact* (E-step) and parameters are optimized exactly (M-step).
- **KL gap:** The EM increase equals the ELBO increase; the KL gap Eqn. (2) is zero after the E-step and generally positive after the M-step (since the posterior changes).
- **Missing data:** EM applies whenever maximizing the complete-data log-likelihood is easier than the incomplete-data one.

1.13 Summary of Variables and Their Dimensions

- $\mathbf{x}_i \in \mathbb{R}^d$: observed sample (column vector, $d \times 1$).
- $\mathbf{z}_i$: latent variable (discrete category or continuous vector).
- $\boldsymbol{\theta}$: parameter vector (model specific).
- $q_i(\mathbf{z}_i)$: E-step distribution (equals $p(\mathbf{z}_i \mid \mathbf{x}_i, \boldsymbol{\theta}^{\text{old}})$ in EM).
- $\mathcal{F}(\{q_i\}, \boldsymbol{\theta})$: ELBO; $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{\text{old}})$: expected complete-data log-likelihood.
- (GMM) $\pi_k \in [0,1]$, $\sum_k \pi_k = 1$; $\boldsymbol{\mu}_k \in \mathbb{R}^d$; $\boldsymbol{\Sigma}_k \in \mathbb{R}^{d \times d}$; responsibilities $\gamma_{ik} \in [0,1]$ with $\sum_k \gamma_{ik} = 1$.

1.14 Summary

From first principles, EM arises by lower-bounding the intractable log-likelihood with a variational ELBO that becomes tight when q_i equals the true posterior of the latent variables. Alternating exact maximization in q (E-step) and $\boldsymbol{\theta}$ (M-step) yields monotone ascent of the likelihood and closed-form weighted-MLE updates in many exponential-family latent-variable models (e.g., mixtures). The same derivation extends to MAP-EM and to GEM when exact M-steps are impractical.