Expectation-Maximization Algorithm

Statistics · Module 9 — Computational Statistics

Node 9.5

Position in Knowledge Graph

System: Statistics

Structural Domain: Computational Statistics

Node: 9.5

The EM algorithm is the canonical method for maximum likelihood estimation in latent variable models, where the observed-data log-likelihood involves an intractable integral over latent variables. It converts this intractable optimization into a sequence of tractable complete-data optimizations by alternating between computing expected sufficient statistics (E-step) and maximizing the complete-data likelihood (M-step). This node derives the ELBO decomposition that justifies EM, proves monotonicity of the likelihood sequence, analyzes convergence rate, and establishes the connection to variational inference.

Formal Definition

Let Y denote observed data and Z denote latent (unobserved) variables. The complete-data log-likelihood is $\ell_c(\theta; Y, Z) = \log p(Y, Z \mid \theta)$. The observed-data log-likelihood is

$$\ell(\theta; Y) = \log p(Y \mid \theta) = \log \int p(Y, Z \mid \theta)\, dZ.$$

The **Q-function** at iteration t is

$$Q(\theta \mid \theta^{(t)}) = \mathbb{E}_{Z \mid Y, \theta^{(t)}}[\log p(Y, Z \mid \theta)] = \int \log p(Y, Z \mid \theta)\, p(Z \mid Y, \theta^{(t)})\, dZ.$$

The **EM algorithm** iterates:

E-step: Compute $Q(\theta \mid \theta^{(t)})$ (or equivalently, the sufficient statistics of $p(Z \mid Y, \theta^{(t)})$).

M-step: Set $\theta^{(t+1)} = \arg\max_\theta Q(\theta \mid \theta^{(t)})$.

Intuition

The observed-data likelihood marginalizes over the latent variables, creating a complex surface with potential local maxima. The EM algorithm avoids directly climbing this surface. Instead, it constructs a simpler surrogate function (the Q-function) that touches the log-likelihood at the current parameter value and lies below it everywhere else. Maximizing this surrogate is guaranteed to increase (or maintain) the log-likelihood. The E-step computes what the latent variables "should" look like given the current parameters; the M-step finds the best parameters given those imputed latent variables.

Structural Role

EM is the bridge between latent variable models and computational tractability. It depends on numerical optimization (Node 9.1) for the M-step, on Bayesian posterior computation (Node 6.1) for the E-step, and on the MLE framework (Node 3.2) for the overall objective. It enables variational inference (Node 9.6) by providing the ELBO framework that VI generalizes. EM is the canonical algorithm for Gaussian mixtures, hidden Markov models, factor analysis, and probabilistic PCA.

Mathematical Development

The ELBO Decomposition

For any distribution $q(Z)$ over the latent variables, we can decompose the observed log-likelihood:

$$\ell(\theta; Y) = \log p(Y \mid \theta) = \log \int p(Y, Z \mid \theta)\, dZ.$$

Introduce $q(Z)$ and apply Jensen's inequality:

$$\ell(\theta; Y) = \log \int \frac{p(Y, Z \mid \theta)}{q(Z)} q(Z)\, dZ \geq \int q(Z) \log \frac{p(Y, Z \mid \theta)}{q(Z)}\, dZ.$$

The right-hand side is the **Evidence Lower BOund (ELBO)**:

$$\text{ELBO}(q, \theta) = \mathbb{E}_q[\log p(Y, Z \mid \theta)] - \mathbb{E}_q[\log q(Z)].$$

The gap between $\ell(\theta; Y)$ and the ELBO is exactly the KL divergence:

$$\ell(\theta; Y) = \text{ELBO}(q, \theta) + \text{KL}(q(Z) \| p(Z \mid Y, \theta)).$$

Proof. Write

$$\text{ELBO}(q, \theta) = \mathbb{E}_q[\log p(Y, Z \mid \theta)] - \mathbb{E}_q[\log q(Z)] = \mathbb{E}_q\left[\log \frac{p(Y, Z \mid \theta)}{q(Z)}\right].$$

Since $p(Y, Z \mid \theta) = p(Z \mid Y, \theta) p(Y \mid \theta)$:

$$\text{ELBO}(q, \theta) = \mathbb{E}_q[\log p(Z \mid Y, \theta)] + \log p(Y \mid \theta) - \mathbb{E}_q[\log q(Z)].$$

Rearranging:

$$\log p(Y \mid \theta) = \text{ELBO}(q, \theta) + \mathbb{E}_q\left[\log \frac{q(Z)}{p(Z \mid Y, \theta)}\right] = \text{ELBO}(q, \theta) + \text{KL}(q \| p(Z \mid Y, \theta)). \quad \square$$

EM as Coordinate Ascent on the ELBO

The EM algorithm maximizes the ELBO by alternating:

E-step (optimize over q at fixed $\theta^{(t)}$): The ELBO is maximized when the KL gap is zero, i.e., $q^{(t)}(Z) = p(Z \mid Y, \theta^{(t)})$. At this q, the ELBO equals $\ell(\theta^{(t)}; Y)$.

M-step (optimize over θ at fixed $q^{(t)}$): Maximize $\text{ELBO}(q^{(t)}, \theta) = \mathbb{E}_{q^{(t)}}[\log p(Y, Z \mid \theta)] + H(q^{(t)})$. Since $H(q^{(t)})$ does not depend on θ, this reduces to maximizing $Q(\theta \mid \theta^{(t)}) = \mathbb{E}_{p(Z \mid Y, \theta^{(t)})}[\log p(Y, Z \mid \theta)]$.

Proof of Monotone Likelihood Ascent

Theorem. $\ell(\theta^{(t+1)}; Y) \geq \ell(\theta^{(t)}; Y)$ for every EM iteration.

Proof. After the E-step, $q^{(t)} = p(Z \mid Y, \theta^{(t)})$, so

$$\ell(\theta^{(t)}; Y) = \text{ELBO}(q^{(t)}, \theta^{(t)}).$$

After the M-step, $\theta^{(t+1)}$ maximizes the ELBO over θ, so

$$\text{ELBO}(q^{(t)}, \theta^{(t+1)}) \geq \text{ELBO}(q^{(t)}, \theta^{(t)}) = \ell(\theta^{(t)}; Y).$$

Since the ELBO is always a lower bound on the log-likelihood:

$$\ell(\theta^{(t+1)}; Y) \geq \text{ELBO}(q^{(t)}, \theta^{(t+1)}) \geq \ell(\theta^{(t)}; Y). \quad \square$$

Convergence Rate

EM converges **linearly** near a local maximum. The rate matrix is

$$M = I_c(\theta^*)^{-1} I_m(\theta^*),$$

where $I_c(\theta^*)$ is the complete-data Fisher information and $I_m(\theta^*) = I_c(\theta^*) - I_{\text{obs}}(\theta^*)$ is the missing information. The convergence rate is the largest eigenvalue of M:

$$\rho = \lambda_{\max}(M) = 1 - \frac{\lambda_{\min}(I_{\text{obs}}(\theta^*))}{\lambda_{\max}(I_c(\theta^*))}.$$

High missingness ($\rho \to 1$) implies slow convergence. When the fraction of missing information is large, the E-step provides little new information and the parameter updates are small.

Convergence to Stationary Points

Under regularity conditions (compactness of Θ, continuity of Q, identifiability), the EM sequence $\{\theta^{(t)}\}$ converges to a stationary point of $\ell(\theta; Y)$. The monotone ascent property, combined with boundedness of ℓ, guarantees that $\ell(\theta^{(t)})$ converges; additional regularity ensures that the parameter sequence converges and the limit satisfies the score equation $\nabla \ell(\theta^*; Y) = 0$. Convergence to a local rather than global maximum is possible; multiple initializations are standard practice.

Worked Examples

Example 1: Gaussian Mixture Model

Let $Y_i \sim \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \sigma_k^2)$ with latent indicators $Z_i \in \{1, \ldots, K\}$.

E-step. Compute responsibilities:

$$r_{ik}^{(t)} = \frac{\pi_k^{(t)} \phi(Y_i; \mu_k^{(t)}, \sigma_k^{2(t)})}{\sum_{j=1}^{K} \pi_j^{(t)} \phi(Y_i; \mu_j^{(t)}, \sigma_j^{2(t)})}.$$

M-step. Update:

$$n_k = \sum_{i=1}^{n} r_{ik}^{(t)}, \quad \mu_k^{(t+1)} = \frac{\sum_i r_{ik}^{(t)} Y_i}{n_k}, \quad \sigma_k^{2(t+1)} = \frac{\sum_i r_{ik}^{(t)} (Y_i - \mu_k^{(t+1)})^2}{n_k}, \quad \pi_k^{(t+1)} = \frac{n_k}{n}.$$

For a mixture of $K = 2$ components with $n = 200$ observations, initializing with k-means, EM typically converges in 20–50 iterations. The log-likelihood increases monotonically, with the largest increases in the first few iterations.

Example 2: Censored Exponential Data

Let $Y_i \sim \text{Exp}(\lambda)$ with right-censoring at c. For censored observations, $Y_i = c$ and the true value $Z_i > c$ is unobserved. The complete-data log-likelihood is $\ell_c(\lambda) = n \log \lambda - \lambda \sum Z_i$.

E-step. For uncensored observations, $\mathbb{E}[Z_i \mid Y_i, \lambda^{(t)}] = Y_i$. For censored observations ($Y_i = c$),

$$\mathbb{E}[Z_i \mid Z_i > c, \lambda^{(t)}] = c + 1/\lambda^{(t)}.$$

M-step. $\lambda^{(t+1)} = n / \sum_i \tilde{Z}_i^{(t)}$, where $\tilde{Z}_i^{(t)}$ uses the imputed expectations.

With $n = 50$, $\lambda = 2$, and censoring at $c = 0.5$ (about 63% censored), EM converges in 10–15 iterations. The high censoring fraction corresponds to large missing information and slower convergence, consistent with the rate formula.

Cross-Links

- **Linear Algebra:** The M-step for Gaussian mixtures involves weighted least squares computations; the convergence rate matrix M is analyzed through its eigenvalue decomposition.
- **AI:** The EM structure underlies training of latent variable models including Gaussian mixture models for clustering, hidden Markov models for sequence labeling, and probabilistic PCA for dimensionality reduction. At the population level, VAE training optimizes the same ELBO that EM maximizes.
- **Economics:** EM is used in econometrics for models with missing data (e.g., sample selection models), regime-switching models (Hamilton filter), and finite mixture models for unobserved heterogeneity.
- **Quantum:** Quantum state tomography with missing measurement settings uses EM to estimate the density matrix; the latent variables correspond to the unobserved measurement outcomes.

Structural Compression

Invariant: Each EM iteration increases or maintains the observed log-likelihood; the algorithm converges to a stationary point of $\ell(\theta; Y)$.

Mechanism: The ELBO decomposition $\ell(\theta; Y) = \text{ELBO}(q, \theta) + \text{KL}(q \| p(Z \mid Y, \theta))$ shows that EM performs coordinate ascent: the E-step sets $q = p(Z \mid Y, \theta^{(t)})$ to close the KL gap, and the M-step maximizes the

ELBO over θ. Since the ELBO is a lower bound that touches ℓ at the current iterate, each M-step guarantees ascent.

Cross-System Echo: EM is a special case of the majorize-maximize (MM) principle: construct a tight lower bound on the objective and maximize the bound. The same principle appears in variational inference (Node 9.6), where the bound is looser because the E-step uses an approximate rather than exact posterior, and in convex optimization, where proximal operators play the role of the M-step. The ELBO is the central object connecting EM, variational inference, and variational autoencoders across statistics and machine learning.

Exercises

1. Derive the ELBO decomposition $\ell(\theta; Y) = \text{ELBO}(q, \theta) + \text{KL}(q \| p(Z \mid Y, \theta))$ from first principles, without using Jensen's inequality, by direct algebraic manipulation.
2. For the two-component Gaussian mixture, write the complete E-step and M-step updates including the variance parameters σ_k^2. Verify that the M-step for π_k involves a Lagrange multiplier enforcing $\sum_k \pi_k = 1$.
3. Prove that the EM convergence rate for the normal mean with known variance and a fraction f of missing observations is $\rho = f$. Interpret this in terms of the missing information principle.
4. Show that Gibbs sampling (Node 9.3) applied to the joint posterior $p(\theta, Z \mid Y)$ produces an ergodic chain, while EM applied to the same model produces a deterministic sequence converging to a point estimate. Explain the structural difference.
5. Implement EM for a $K = 3$ Gaussian mixture on simulated data. Plot the log-likelihood as a function of iteration and verify monotonicity. Run from 5 different initializations and compare the local maxima found.
6. For censored exponential data, derive the E-step formula $\mathbb{E}[Z_i \mid Z_i > c, \lambda] = c + 1/\lambda$ from the memoryless property of the exponential distribution.
7. Argue, structurally, why the ELBO decomposition is the fundamental identity connecting EM, variational inference, and the marginal likelihood: in each case, the decomposition $\log p(Y) = \text{ELBO} + \text{KL}$ determines which quantity is being optimized (the ELBO), what the gap measures (approximation quality), and how the algorithm trades exactness for tractability.

Statistics · Module 9 — Computational Statistics · Node 9.5
Concepts: maximum-likelihood-estimation, convergence, fixed-point, likelihood
Prerequisites: 9.1, 6.1, 3.2
Enables: 9.6

hbar.university — own.your.cognition