Expectation Maximization

Yutao Chen

⧖ Oct 06 2025
⧗ Oct 11 2025

Contents

Expectation maximization (EM) (Dempster et al., 1977) is designed for *maximum likelihood* estimation of parameters in probabilistic models with *missing data* or *hidden variables*.

Let $\{\boldsymbol{x}_n\}$ denote the set of observed data, and $\{\boldsymbol{z}_n\}$ the set of hidden data. We want to maximize the likelihood w.r.t. the observed data:

$$\begin{aligned} &\arg\max_{\boldsymbol{\theta}} \sum_{\boldsymbol{x}_n} \log p(\boldsymbol{x}_n|\boldsymbol{\theta}) \\ &= \arg\max_{\boldsymbol{\theta}} \sum_{\boldsymbol{x}_n} \log\left(\int p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}) \,\mathrm{d}\boldsymbol{z}_n\right), \end{aligned}$$

where $p(\boldsymbol{x}|\boldsymbol{\theta})$ is known as the *incomplete-data* likelihood, and $p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})$ is known as the *complete-data* likelihood.

Evidence Lower Bound

Unfortunately, this maximization is generally intractable, because of the $\log \int p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}) \,\mathrm{d}\boldsymbol{z}$ term.

We can bypass the intractability by transforming $\log p(\boldsymbol{x}|\boldsymbol{\theta})$ as follows:

$$\begin{aligned} \log p(\boldsymbol{x}|\boldsymbol{\theta}) &= \mathbb{E}_{q(\boldsymbol{z})}[\log p(\boldsymbol{x}|\boldsymbol{\theta})] \\ &= \mathbb{E}_{q(\boldsymbol{z})}[\log(p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})/p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta}))] \\ &= \underbrace{\mathbb{E}_{q(\boldsymbol{z})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})}{q(\boldsymbol{z})}\right]}_{\mathcal{F}(q(\boldsymbol{z}),\boldsymbol{\theta})} + \mathbb{D}_{\mathrm{KL}}(q(\boldsymbol{z}) \parallel p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta})), \end{aligned}$$

where $\mathcal{F}(q(\boldsymbol{z}), \boldsymbol{\theta})$ is known as the *evidence lower bound* (ELBO). We have

$$\mathcal{F}(q(\boldsymbol{z}), \boldsymbol{\theta}) \leq \log p(\boldsymbol{x}|\boldsymbol{\theta})$$

for any $q(\boldsymbol{z})$ and $\boldsymbol{\theta}$, with equality holding iff $q(\boldsymbol{z}) = p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta})$.

The EM algorithm then maximizes $\log p(\boldsymbol{x}|\boldsymbol{\theta})$ by instead maximizing the lower bound $\mathcal{F}(q(\boldsymbol{z}), \boldsymbol{\theta})$ iteratively. For each iteration t, we perform *coordinate ascent* on $\mathcal{F}(q(\boldsymbol{z}), \boldsymbol{\theta})$ alternating between $q(\boldsymbol{z})$ and $\boldsymbol{\theta}$.

- In the **E-step**, we maximize $\mathcal{F}(q(\boldsymbol{z}), \boldsymbol{\theta})$ with $\boldsymbol{\theta} = \boldsymbol{\theta}_t$ fixed:

$$q_t(\boldsymbol{z}) = \arg\max_{q(\boldsymbol{z})} \mathcal{F}(q(\boldsymbol{z}), \boldsymbol{\theta}_t) = p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta}_t).$$

- In the **M-step**, we maximize $\mathcal{F}(q(\boldsymbol{z}), \boldsymbol{\theta})$ with $q(\boldsymbol{z}) = q_t(\boldsymbol{z})$ fixed:

$$\begin{aligned}\boldsymbol{\theta}_{t+1} &= \arg\max_{\boldsymbol{\theta}} \mathcal{F}(q_t(\boldsymbol{z}), \boldsymbol{\theta}) \\ &= \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{q_t(\boldsymbol{z})}[\log p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})].\end{aligned}$$

This iterative process guarantees monotonic improvement of $\log p(\boldsymbol{x}|\boldsymbol{\theta})$ until convergence to some *local* maxima, because for each iteration t

$$\log p(\boldsymbol{x}|\boldsymbol{\theta}_t) = \underbrace{\mathcal{F}(q_t(\boldsymbol{z}), \boldsymbol{\theta}_t) \leq}_{\text{E-step}} \underbrace{\mathcal{F}(q_t(\boldsymbol{z}), \boldsymbol{\theta}_{t+1}) \leq}_{\text{M-step}} \log p(\boldsymbol{x}|\boldsymbol{\theta}_{t+1}).$$

> The EM algorithm can also be applied to *maximum a posteriori* with a prior distribution $p(\boldsymbol{\theta})$ over the parameters. This simply amounts to a modified lower bound objective $\tilde{\mathcal{F}}$:
>
> $$\tilde{\mathcal{F}}(q(\boldsymbol{z}), \boldsymbol{\theta}) = \mathcal{F}(q(\boldsymbol{z}), \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \leq \log p(x|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

Extensions and Connections

Variational EM

One of the basic assumption we have made in EM is that we can easily evaluate $q_t(\boldsymbol{z}) = p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta}_t)$ in the E-step.

However, evaluating the posterior $p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta}_t)$ itself could be intractable, especially if $\boldsymbol{z}$ is a continuous r.v. We can instead use *variational inference* (VI) to pick q_t such that

$$q_t(\boldsymbol{z}) = \arg\max_{q \in \mathcal{Q}} \mathbb{D}_{\mathrm{KL}}(q(\boldsymbol{z}) \parallel p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta})),$$

where Q is the variational family. Intuitively, we pick a distribution $q_t(\boldsymbol{z}) \in \mathcal{Q}$ that can best approximate the exact posterior $p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta})$.

This approach, unfortunately, does not guarantee monotonic improvement of $\log p(\boldsymbol{x}|\boldsymbol{\theta})$ due to approximation errors. Only when the variational family Q is sufficiently versatile such that $p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta}) \in \mathcal{Q}$ can we (in theory) recover the behaviors of regular EM.

Stochastic Gradient EM

Another basic assumption we have made in EM is that we can compute $\boldsymbol{\theta}_{t+1} = \arg\max_{\boldsymbol{\theta}} \mathcal{F}(q_t(\boldsymbol{z}), \boldsymbol{\theta})$ in the M-step.

For many practical problems, however, such maximization is not easy. Fortunately, note that in the M-step, as long as we can find some $\boldsymbol{\theta}_{t+1}$ that guarantees

$$\mathcal{F}(q_t(\boldsymbol{z}), \boldsymbol{\theta}_t) \leq \mathcal{F}(q_t(\boldsymbol{z}), \boldsymbol{\theta}_{t+1}),$$

the monotonic improvement of $\log p(\boldsymbol{x}|\boldsymbol{\theta})$ (and hence convergence) still holds. Therefore, we can find $\boldsymbol{\theta}_{t+1}$ by taking one or a few gradient ascent steps following $\nabla_{\boldsymbol{\theta}} \mathcal{F}$:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \eta \nabla_{\boldsymbol{\theta}} \mathcal{F}(q_t(\boldsymbol{z}), \boldsymbol{\theta}_t).$$

> The varational auto-encoders (VAEs) (Kingma & Welling, 2013) can be interpreted as an instance of variational stochastic gradient EM.

However, EM becomes less appealling when there is no close form for the M-step, as one might just as well directly optimize $\log p(\boldsymbol{x}|\boldsymbol{\theta})$ using gradient-based methods. Particularly, one can show that

$$\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{x}|\boldsymbol{\theta}_t) = \nabla_{\boldsymbol{\theta}} \mathcal{F}(q_t(\boldsymbol{z}), \boldsymbol{\theta}_t).$$

References

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

Kingma, D. P., & Welling, M. (2013,). Auto-Encoding Variational Bayes. *International Conference on Learning Representations.* https://openreview.net/forum?id=33X9fd2-9FyZd