Expectation Propagation Notes

Joaquín Rapela

January 14, 2026

These notes were mostly extracted from Bishop (2016)

Fundamental result

Consider the problem of minimizing $\mathrm{KL}(p||q)$ with respect to $q(\mathbf{z})$, when $p(\mathbf{z})$ is a fixed distribution, and $q(\mathbf{z})$ is a member of the exponential family:

$$q(\mathbf{z}) = h(\mathbf{z})g(\eta)\exp(\eta^\intercal \mathbf{u}(\mathbf{z}))$$

The minimizing distribution $q(\mathbf{z})$ satisfies the moment matching criterion:

$$\mathbb{E}_{q(\mathbf{z})}\left[\mathbf{u}(\mathbf{z})\right] = \mathbb{E}_{p(\mathbf{z})}\left[\mathbf{u}(\mathbf{z})\right]$$

Summary of expectation propagation

We are given the joint probability density function between the data $\mathcal{D}$ and hidden variables and parameters θ:

$$p(\mathcal{D}, \theta) = \prod_i f_i(\theta)$$

For example, if we are given N independent and identically distributed data points $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, then $f_i(\theta) = p(\mathbf{x}_i|\theta)$, for $i \in \{1, \ldots, N\}$, and $f_0(\theta)$ is the prior on the hidden variables and parameters, $f_0(\theta) = p(\theta)$.

We want to estimate

$$p(\theta|\mathcal{D}) = \frac{1}{p(\mathcal{D})}\prod_i f_i(\theta) \quad \text{with} \quad p(\mathcal{D}) = \int \prod_i f_i(\theta) d\theta \tag{1}$$

We assume that the marginalisation in Eq. 1 is intractable, and we will approximate this true posterior with

$$q(\theta) = \frac{1}{Z} \prod_i \tilde{f}_i(\theta)$$

where $\tilde{f}_i(\theta)$ are members of the exponential family.

We could find the factors $\tilde{f}_i(\theta)$ by minimizing

$$\mathrm{KL}\left(\frac{1}{p(\mathcal{D})} \prod_i f_i(\theta) \middle\| \frac{1}{Z} \prod_i \tilde{f}_i(\theta) \right)$$

However, this minimization is intractable, as we need to simultaneously fit all terms of the complex distribution $\frac{1}{p(\mathcal{D})} \prod_i f_i(\theta)$.

An alternative approach would fit separately the factors $f_i(\theta)$ and $\tilde{f}_i(\theta)$. But this approach could overfit some factors, by not taking into consideration the other factors.

Expectation propagation fits the terms $f_i(\theta)$ and $\tilde{f}_i(\theta)$ in the context of previously fitted terms. It starts by initializing all factors $\tilde{f}_i(\theta)$ and then cycles each factor optimizing them one at a time.

References

Bishop, C. M. (2016). *Pattern recognition and machine learning.* Springer-Verlag New York.