

## 13 Expectation-Maximization (EM) and Mixture Models

### Goal

Mixture models for density estimation and the celebrated expectation-maximization algorithm.

### Alert 13.1: Convention

Gray boxes are not required hence can be omitted for unenthusiastic readers.

[This note is likely to be updated again soon.](#)

### Definition 13.2: Density estimation

The central problem of this note is to estimate a **density function** (or more generally a **probability measure**), through a finite training sample. Formally, we are interested in estimating a probability measure  $\chi$  from a (non)parametric family  $\{\chi_\theta\}_{\theta \in \Theta}$ . A typical approach is to minimize some statistical divergence (distance) between a noisy version  $\hat{\chi}$  and  $\chi_\theta$ :

$$\inf_{\theta \in \Theta} D(\hat{\chi} \| \chi_\theta).$$

However, the minimization problem above may not always be easy to solve, and alternative (indirect) strategies have been developed. As we show below, choosing the KL divergence corresponds to the maximum likelihood estimation procedure.

### Definition 13.3: KL and LK

Recall that the **Kullback-Leibler** (KL) divergence between two density functions  $p$  and  $q$  is defined as:

$$\text{KL}(p \| q) = \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x}.$$

Reverse the inputs we obtain the reverse KL divergence:

$$\text{LK}(p \| q) := \text{KL}(q \| p).$$

(In Section 14 we will generalize the above two divergences to a family of  $f$ -divergence.)

### Definition 13.4: Entropy, conditional entropy, cross-entropy, and mutual information

We define the **entropy** of a random vector  $\mathbf{X}$  with pdf  $p$  as:

$$H(\mathbf{X}) := -\mathbb{E} \log p(\mathbf{X}) = - \int p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x},$$

the **conditional entropy** between  $\mathbf{X}$  and  $\mathbf{Z}$  (with pdf  $q$ ) as:

$$H(\mathbf{X} | \mathbf{Z}) := -\mathbb{E} \log p(\mathbf{X} | \mathbf{Z}) = - \int p(\mathbf{x}, \mathbf{z}) \log p(\mathbf{x} | \mathbf{z}) d\mathbf{x} d\mathbf{z},$$

and the **cross-entropy** between  $\mathbf{X}$  and  $\mathbf{Z}$  as:

$$\mathfrak{t}(\mathbf{X}, \mathbf{Z}) := -\mathbb{E} \log q(\mathbf{X}) = - \int p(\mathbf{x}) \log q(\mathbf{x}) d\mathbf{x}.$$

Finally, we define the **mutual information** between  $\mathbf{X}$  and  $\mathbf{Z}$  as:

$$I(\mathbf{X}, \mathbf{Z}) := \text{KL}(p(\mathbf{x}, \mathbf{z}) \| p(\mathbf{x})q(\mathbf{z})) = \int p(\mathbf{x}, \mathbf{z}) \log \frac{p(\mathbf{x}, \mathbf{z})}{p(\mathbf{x})q(\mathbf{z})} d\mathbf{x} d\mathbf{z}$$

### Exercise 13.5: Information theory

Verify the following:

$$H(\mathbf{X}, \mathbf{Z}) = H(\mathbf{Z}) + H(\mathbf{X}|\mathbf{Z})$$

$$\mathfrak{I}(\mathbf{X}, \mathbf{Z}) = H(\mathbf{X}) + \text{KL}(\mathbf{X}|\mathbf{Z}) = H(\mathbf{X}) + \text{LK}(\mathbf{Z}|\mathbf{X})$$

$$I(\mathbf{X}, \mathbf{Z}) = H(\mathbf{X}) - H(\mathbf{X}|\mathbf{Z})$$

$$I(\mathbf{X}, \mathbf{Z}) \geq 0, \text{ with equality iff } \mathbf{X} \text{ independent of } \mathbf{Z}$$

$$\text{KL}(p(\mathbf{x}, \mathbf{z}) \| q(\mathbf{x}, \mathbf{z})) = \text{KL}(p(\mathbf{z}) \| q(\mathbf{z})) + \mathbb{E}[\text{KL}(p(\mathbf{x}|\mathbf{z}) \| q(\mathbf{x}|\mathbf{z}))].$$

All of the above can obviously be iterated to yield formula for more than two random vectors.

### Exercise 13.6: Multivariate Gaussian

Compute

- the entropy of the multivariate Gaussian  $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ ;
- the KL divergence between two multivariate Gaussians  $\mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$  and  $\mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2)$ .

### Remark 13.7: MLE = KL minimization

Let us define the empirical “pdf” based on a dataset  $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ :

$$\hat{p}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \delta_{\mathbf{x}_i},$$

where  $\delta_{\mathbf{x}}$  is the “illegal” **delta mass** concentrated at  $\mathbf{x}$ . Then, we claim that

$$\theta_{\text{MLE}} = \underset{\theta \in \Theta}{\text{argmin}} \text{KL}(\hat{p} \| p(\mathbf{x}|\theta)).$$

Indeed, we have

$$\text{KL}(\hat{p} \| p(\mathbf{x}|\theta)) = \int [\log(\hat{p}(\mathbf{x})) - \log p(\mathbf{x}|\theta)] \hat{p}(\mathbf{x}) d\mathbf{x} = C + \frac{1}{n} \sum_{i=1}^n -\log p(\mathbf{x}_i|\theta),$$

where  $C$  is a constant that does not depend on  $\theta$ .

### Exercise 13.8: Why KL is so special

To appreciate the uniqueness of the KL divergence, prove the following:

$\log$  is the only continuous function satisfying  $f(st) = f(s) + f(t)$ .

**Algorithm 13.9: Expectation-Maximization (EM) (Dempster et al. 1977)**

We formulate EM under the density estimation formulation in Definition 13.2, except that we carry out the procedure in a **lifted space**  $\mathbf{X} \times \mathbf{Z}$  where  $\mathbf{Z}$  is the space that some latent random variable  $\mathbf{Z}$  lives in. Importantly, we do not observe the latent variable  $\mathbf{Z}$ : it is “artificially” constructed to aid our job. We fit our model with a **prescribed** family of **joint** distributions

$$\mu_\theta(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{z}) = \zeta_\theta(\mathrm{d}\mathbf{z}) \cdot \mathfrak{D}_\theta(\mathrm{d}\mathbf{x}|\mathbf{z}) = \chi_\theta(\mathrm{d}\mathbf{x}) \cdot \mathfrak{E}_\theta(\mathrm{d}\mathbf{z}|\mathbf{x}), \quad \theta \in \Theta.$$

In EM, we typically specify the joint distribution  $\mu_\theta$  **explicitly**, and in a way that the **posterior distribution**  $\mathfrak{E}_\theta(\mathrm{d}\mathbf{z}|\mathbf{x})$  **can be easily computed**. Similarly, we “lift”  $\chi(\mathrm{d}\mathbf{x})$  (our target of estimation) to the joint distribution

$$\hat{\nu}(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{z}) = \hat{\chi}(\mathrm{d}\mathbf{x}) \cdot \mathcal{E}(\mathrm{d}\mathbf{z}|\mathbf{x}).$$

(We use the hat notation to remind that we do not really have access to the true distribution  $\chi$  but a sample from it, represented by the empirical distribution  $\hat{\chi}$ .) Then, we minimize the discrepancy between the joint distributions  $\hat{\nu}$  and  $\mu_\theta$ , which is an *upper bound* of the discrepancy of the marginals  $\mathrm{KL}(\hat{\chi}||\chi_\theta)$  (Exercise 13.5):

$$\inf_{\theta \in \Theta} \inf_{\mathcal{E}(\mathrm{d}\mathbf{z}|\mathbf{x})} \mathrm{KL}(\hat{\nu}(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{z})||\mu_\theta(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{z})).$$

Note that there is no restriction on  $\mathcal{E}$  (and do not confuse it with  $\mathfrak{E}_\theta$ , which is “prescribed”).

The EM algorithm proceeds with alternating minimization:

- (E-step) Fix  $\theta_t$ , we solve  $\mathcal{E}_{t+1}$  by (recall Exercise 13.5)

$$\inf_{\mathcal{E}} \mathrm{KL}(\hat{\nu}(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{z})||\mu_{\theta_t}(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{z})) = \mathrm{KL}(\hat{\chi}||\chi_{\theta_t}) + \mathbb{E}_{\hat{\chi}} \mathrm{KL}(\mathcal{E}||\mathfrak{E}_{\theta_t}),$$

which leads to the “closed-form” solution:

$$\mathcal{E}_{t+1} = \mathfrak{E}_{\theta_t}.$$

- (M-step) Fix  $\mathcal{E}_{t+1}$ , we solve  $\theta_{t+1}$  by

$$\inf_{\theta \in \Theta} \mathrm{KL}(\hat{\nu}_{t+1}(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{z})||\mu_\theta(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{z})) = \underbrace{\mathrm{KL}(\hat{\chi}||\chi_\theta)}_{\text{likelihood}} + \underbrace{\mathbb{E}_{\hat{\chi}} \mathrm{KL}(\mathcal{E}_{t+1}||\mathfrak{E}_\theta)}_{\text{regularizer}}.$$

For the generalized EM algorithm, we need only decrease the above (joint) KL divergence if finding a (local) minima is expensive. It may be counter-intuitive that **minimizing the sum of two terms above can be easier than minimizing the first likelihood term only!**

Obviously, the EM algorithm monotonically decreases our (joint) KL divergence  $\mathrm{KL}(\hat{\nu}, \mu_\theta)$ . Moreover, thanks to construction, the EM algorithm also ascends the likelihood:

$$\mathrm{KL}(\hat{\chi}||\chi_{\theta_{t+1}}) \leq \mathrm{KL}(\hat{\nu}_{t+1}||\mu_{\theta_{t+1}}) \leq \mathrm{KL}(\hat{\nu}_{t+1}||\mu_{\theta_t}) = \mathrm{KL}(\hat{\chi}||\chi_{\theta_t}).$$

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). “Maximum Likelihood from Incomplete Data via the EM Algorithm”. *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38.

**Definition 13.10: Exponential family distribution**

The exponential family distributions have the following density form:

$$p(\mathbf{x}) = h(\mathbf{x}) \exp(\langle \boldsymbol{\eta}, T(\mathbf{x}) \rangle - A(\boldsymbol{\eta})),$$

where  $T(\mathbf{x})$  is the sufficient statistics,  $\boldsymbol{\eta}$  is the natural parameter,  $A$  is the log-partition function, and  $h$  represents the base measure. Since  $\mathbf{p}$  integrates to 1, we have

$$A(\boldsymbol{\eta}) = \log \int \exp(\langle \boldsymbol{\eta}, T(\mathbf{x}) \rangle) \cdot h(\mathbf{x}) d\mathbf{x}$$

We can verify that  $A$  is a convex function.

### Example 13.11: Gaussian distribution in exponential family

Recall that the multivariate Gaussian density is:

$$\begin{aligned} \mathbf{p}(\mathbf{x}) &= (2\pi)^{-d/2} [\det(\Sigma)]^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) \\ &= \exp\left(\left\langle \begin{bmatrix} \mathbf{x} \\ -\frac{1}{2}\mathbf{x}\mathbf{x}^\top \end{bmatrix}, \begin{bmatrix} \Sigma^{-1}\boldsymbol{\mu} \\ \Sigma^{-1} \end{bmatrix} \right\rangle - \frac{1}{2}(\boldsymbol{\mu}^\top \Sigma^{-1}\boldsymbol{\mu} + d \log(2\pi) + \log \det \Sigma)\right). \end{aligned}$$

Thus, we identify

$$\begin{aligned} T(\mathbf{x}) &= (\mathbf{x}, -\frac{1}{2}\mathbf{x}\mathbf{x}^\top) \\ \boldsymbol{\eta} &= (\Sigma^{-1}\boldsymbol{\mu}, \Sigma^{-1}) =: (\boldsymbol{\xi}, S) \\ A(\boldsymbol{\mu}, \Sigma) &= \frac{1}{2}(\boldsymbol{\mu}^\top \Sigma^{-1}\boldsymbol{\mu} + d \log(2\pi) + \log \det \Sigma) \\ A(\boldsymbol{\eta}) &= A(\boldsymbol{\xi}, S) = \frac{1}{2}(\boldsymbol{\xi}^\top S^{-1}\boldsymbol{\xi} + d \log(2\pi) - \log \det S). \end{aligned}$$

### Example 13.12: Bernoulli and Multinoulli in exponential family

The Bernoulli distribution is given as:

$$\begin{aligned} \mathbf{p}(z) &= \pi^z (1 - \pi)^{1-z} = \exp(z \log \pi + (1 - z) \log(1 - \pi)) \\ &= \exp\left(z \log \frac{\pi}{1-\pi} + \log(1 - \pi)\right). \end{aligned}$$

Thus, we may identify

$$\begin{aligned} T(z) &= z \\ \eta &= \log \frac{\pi}{1-\pi} \\ A(\eta) &= \log(1 + \exp(\eta)). \end{aligned}$$

We can also consider the multinoulli distribution:

$$\begin{aligned} \mathbf{p}(\mathbf{z}) &= \prod_{k=1}^c \pi_k^{z_k} = \exp(\langle \mathbf{z}, \log \boldsymbol{\pi} \rangle) \\ &= \exp\left(\tilde{\mathbf{z}}^\top \log \frac{\tilde{\boldsymbol{\pi}}}{1 - \langle \mathbf{1}, \tilde{\boldsymbol{\pi}} \rangle} + \log(1 - \langle \mathbf{1}, \tilde{\boldsymbol{\pi}} \rangle)\right), \end{aligned}$$

where recall that  $\mathbf{z} \in \{0, 1\}^c$  is one-hot (i.e.  $\mathbf{1}^\top \mathbf{z} = 1$ ), and we use the tilde notation to denote the subvector with the last entry removed. Thus, we may identify

$$\begin{aligned} T(\tilde{\mathbf{z}}) &= \tilde{\mathbf{z}} \\ \tilde{\boldsymbol{\eta}} &= \log \frac{\tilde{\boldsymbol{\pi}}}{1 - \langle \mathbf{1}, \tilde{\boldsymbol{\pi}} \rangle} \\ A(\tilde{\boldsymbol{\eta}}) &= \log(1 + \langle \mathbf{1}, \exp(\tilde{\boldsymbol{\eta}}) \rangle). \end{aligned}$$

(Here, we use the tilde quantities to remove one redundancy since  $\mathbf{1}^\top \mathbf{z} = \mathbf{1}^\top \boldsymbol{\pi} = 1$ ).

**Exercise 13.13: Mean parameter and moments**

Prove that for the exponential family distribution,

$$\begin{aligned}\nabla A(\boldsymbol{\eta}) &= \mathbb{E}[T(\mathbf{X})] \\ \nabla^2 A(\boldsymbol{\eta}) &= \mathbb{E}[T(\mathbf{X}) \cdot T(\mathbf{X})^\top] - \mathbb{E}[T(\mathbf{X})] \cdot \mathbb{E}[T(\mathbf{X})]^\top = \text{Cov}(T(\mathbf{X})),\end{aligned}$$

where the last equality confirms again that the log-partition function  $A$  is convex (since the covariance matrix is positive semidefinite).

**Exercise 13.14: Marginal, conditional and product of exponential family**

Let  $\mathbf{p}(\mathbf{x}, \mathbf{z})$  be a joint distribution from the exponential family. Prove the following:

- The marginal  $\mathbf{p}(\mathbf{x})$  need **not** be from the exponential family.
- The conditional  $\mathbf{p}(\mathbf{z}|\mathbf{x})$  is again from the exponential family.
- The product of two exponential family distributions is again in exponential family.

**Exercise 13.15: Exponential family approximation under KL**

Let  $\mathbf{p}(\mathbf{x})$  be an arbitrary distribution and  $\mathbf{q}_\eta(\mathbf{x})$  from the exponential family with sufficient statistics  $T$ , natural parameter  $\boldsymbol{\eta}$  and log-partition function  $A$ . Then,

$$\boldsymbol{\eta}^* := \underset{\boldsymbol{\eta}}{\operatorname{argmin}} \text{KL}(\mathbf{p} \parallel \mathbf{q}_\eta)$$

is given by moment-matching:

$$\mathbb{E}_{\mathbf{p}} T(\mathbf{X}) = \mathbb{E}_{\mathbf{q}_\eta} T(\mathbf{X}) = \nabla A(\boldsymbol{\eta}), \quad \text{i.e.,} \quad \boldsymbol{\eta} = \nabla A^{-1}(\mathbb{E}_{\mathbf{p}} T(\mathbf{X})).$$

**Exercise 13.16: EM for exponential family**

Prove that the M-step of EM simplifies to the following, if we assume the joint distribution  $\mu_\eta$  is from the exponential family with natural parameter  $\boldsymbol{\eta}$ , sufficient statistics  $T$  and log-partition function  $A$ :

$$\boldsymbol{\eta}_{t+1} = \nabla A^{-1}(\mathbb{E}_{\hat{\nu}_{t+1}}(T(\mathbf{X}))).$$

**Definition 13.17: Mixture Distribution**

We define the joint distribution over a discrete latent random variable  $Z \in \{1, \dots, c\}$  and an observed random variable  $\mathbf{X}$ :

$$\mathbf{p}(\mathbf{x}, \mathbf{z}) = \prod_{k=1}^c [\pi_k \cdot \mathbf{p}_k(\mathbf{x}; \theta_k)]^{z_k},$$

where we represent  $\mathbf{z}$  using one-hot encoding. We easily obtain the marginal and conditional:

$$\mathbf{p}(\mathbf{x}) = \sum_{k=1}^c \pi_k \cdot \mathbf{p}_k(\mathbf{x}; \theta_k) \tag{13.1}$$

$$\begin{aligned}
p(\mathbf{z} = \mathbf{e}_k) &= \pi_k \\
p(\mathbf{x}|\mathbf{z} = \mathbf{e}_k) &= p_k(\mathbf{x}; \theta_k) \\
p(\mathbf{z} = \mathbf{e}_k|\mathbf{x}) &= \frac{\pi_k \cdot p_k(\mathbf{x}; \theta_k)}{\sum_{j=1}^c \pi_j \cdot p_j(\mathbf{x}; \theta_j)}.
\end{aligned}$$

The marginal distribution  $p(\mathbf{x})$  can be interpreted as follows: There are  $c$  component densities  $p_k$ . We choose a component  $p_k$  with probability  $\pi_k$  and then we sample  $\mathbf{x}$  from the resulting component density. However, in reality we do not know which component density an observation  $\mathbf{x}$  is sampled from, i.e., the discrete random variable  $\mathbf{Z}$  is not observed (missing).

Let  $p_k(\mathbf{x}; \theta_k)$  be multivariate Gaussian (with  $\theta_k$  denoting its mean and covariance) we get the popular Gaussian mixture model (GMM).

#### Algorithm 13.18: Mixture density estimation – MLE

Replacing the parameterization  $\chi_\theta$  with the mixture model in (13.1) we get a direct method for estimating the density function  $\chi$  based on a sample:

$$\min_{\pi \in \Delta, \theta \in \Theta} \text{KL}(\hat{\chi} \| p), \quad p(\mathbf{x}) = \sum_{k=1}^c \pi_k \cdot p_k(\mathbf{x}; \theta_k)$$

where  $\Delta$  denotes the simplex constraint (i.e.,  $\pi \geq \mathbf{0}$  and  $\mathbf{1}^\top \pi = 1$ ). The number of components  $c$  is a hyperparameter that needs to be determined *a priori*. We may apply (projected) gradient descent to solve  $\pi$  and  $\theta$ . However, it is easy to verify that the objective function is nonconvex hence convergence to a reasonable solution may not be guaranteed.

We record the gradient here for later comparison. We use  $p_W(\mathbf{x}, \mathbf{z})$  for the joint density whose marginalization over the latent  $\mathbf{z}$  gives  $p(\mathbf{x})$ . For mixtures, the parameter  $W$  includes both  $\pi$  and  $\theta$ .

$$\frac{\partial}{\partial W} = -\mathbb{E}_{\hat{p}_W(\mathbf{z}, \mathbf{x})} \frac{\partial \log p_W(\mathbf{x}, \mathbf{z})}{\partial W}, \quad \text{where} \quad \hat{p}_W(\mathbf{z}, \mathbf{x}) := \hat{\chi}(\mathbf{d}\mathbf{x}) \cdot p_W(\mathbf{z}|\mathbf{x}).$$

#### Algorithm 13.19: Mixture density estimation – EM

Let us now apply the EM Algorithm 13.9 to the mixture density estimation problem. As mentioned before, we minimize the upper bound:

$$\min_{\pi \in \Delta, \theta \in \Theta} \min_{\mathcal{E}} \text{KL}(\hat{\nu}(\mathbf{x}, \mathbf{z}) \| p(\mathbf{x}, \mathbf{z})), \quad \hat{\nu}(\mathbf{x}, \mathbf{z}) = \hat{\chi}(\mathbf{x}) \mathcal{E}(\mathbf{z}|\mathbf{x}), \quad p(\mathbf{x}, \mathbf{z}) = \prod_{k=1}^c [\pi_k \cdot p_k(\mathbf{x}; \theta_k)]^{z_k},$$

with the following two steps alternated until convergence:

- E-step: Fix  $\pi^{(t)}$  and  $\theta^{(t)}$ , we solve

$$\mathcal{E}_{t+1} = p^{(t)}(\mathbf{z} = \mathbf{e}_k|\mathbf{x}) = \frac{\pi_k^{(t)} \cdot p_k(\mathbf{x}; \theta_k^{(t)})}{\sum_{j=1}^c \pi_j^{(t)} \cdot p_j(\mathbf{x}; \theta_j^{(t)})} =: r_k^{(t+1)}(\mathbf{x}). \quad (13.2)$$

- M-step: Fix  $\mathcal{E}_{t+1}$  hence  $\hat{\nu}_{t+1}$ , we solve

$$\begin{aligned}
\min_{\pi \in \Delta} \min_{\theta \in \Theta} \text{KL}(\hat{\nu}_{t+1}(\mathbf{x}, \mathbf{z}) \| p(\mathbf{x}, \mathbf{z})) &\equiv \max_{\pi \in \Delta} \max_{\theta \in \Theta} \mathbb{E}_{\hat{\chi}} \mathbb{E}_{\mathcal{E}_{t+1}} [\langle \mathbf{z}, \log \pi \rangle + \langle \mathbf{z}, \log p(\mathbf{X}; \theta) \rangle] \\
&= \max_{\pi \in \Delta} \max_{\theta \in \Theta} \mathbb{E}_{\hat{\chi}} \left[ \left\langle \mathbf{r}^{(t+1)}(\mathbf{X}), \log \pi \right\rangle + \left\langle \mathbf{r}^{(t+1)}(\mathbf{X}), \log p(\mathbf{X}; \theta) \right\rangle \right].
\end{aligned}$$

It is clear that the optimal

$$\boldsymbol{\pi}^{(t+1)} = \mathbb{E}_{\hat{\chi}} \mathbf{r}^{(t+1)}(\mathbf{X}) = \sum_{i=1}^n \hat{\chi}(\mathbf{x}_i) \cdot \mathbf{r}^{(t+1)}(\mathbf{x}_i). \quad (13.3)$$

(For us the empirical training distribution  $\hat{\chi} \equiv \frac{1}{n}$ , although we prefer to keep everything abstract and general.)

The  $\theta_k$ 's can be solved independently:

$$\max_{\theta_k \in \Theta_k} \mathbb{E}_{\hat{\chi}} \left[ r_k^{(t+1)}(\mathbf{X}) \cdot \log \mathbf{p}_k(\mathbf{X}; \theta_k) \right] \equiv \min_{\theta_k \in \Theta_k} \text{KL}(\hat{\chi}_k^{(t+1)} \parallel \mathbf{p}_k(\cdot; \theta_k)), \quad (13.4)$$

where we define  $\hat{\chi}_k^{(t+1)} \propto \hat{\chi} \cdot r_k^{(t+1)}$ . (This is similar to Adaboost, where we reweigh the training examples!)

If we choose the component density  $\mathbf{p}_k(\mathbf{x}; \theta_k)$  from the exponential family with sufficient statistics  $T_k$  and log-partition function  $A_k$ , then from Exercise 13.15 we know (13.4) can be solved in closed-form:

$$\theta_k^{(t+1)} = \nabla A_k^{-1} \left[ \mathbb{E}_{\hat{\chi}_k^{(t+1)}} T_k(\mathbf{X}) \right] \quad (13.5)$$

#### Alert 13.20: implicit EM vs. explicit ML

We now make an important connection between EM and ML. We follow the notation in Algorithm 13.18. For the joint density  $\mathbf{p}_W(\mathbf{x}, \mathbf{z})$ , EM solves

$$W_{t+1} = \underset{W}{\operatorname{argmin}} \text{KL}(\hat{\mathbf{p}}_{t+1} \parallel \mathbf{p}_W), \quad \text{where} \quad \hat{\mathbf{p}}_{t+1}(\mathbf{x}, \mathbf{z}) := \hat{\mathbf{p}}_{W_t}(\mathbf{x}, \mathbf{z}) = \hat{\chi}(\mathbf{d}\mathbf{x}) \cdot \mathbf{p}_{W_t}(\mathbf{z}|\mathbf{x}).$$

In particular, at a minimizer  $W_{t+1}$  the gradient vanishes:

$$-\mathbb{E}_{\hat{\mathbf{p}}_{t+1}} \frac{\partial \log \mathbf{p}_W(\mathbf{x}, \mathbf{z})}{\partial W} = \mathbf{0}.$$

In other words, EM solves the above nonlinear equation (in  $W$ ) to get  $W_{t+1}$  while ML with gradient descent simply performs one fixed-point iteration.

#### Example 13.21: Gaussian Mixture Model (GMM) – EM

Using the results in multivariate Gaussian Example 13.11 we derive:

$$\begin{aligned} \nabla A(\boldsymbol{\eta}) &= \begin{bmatrix} S^{-1} \boldsymbol{\xi} \\ -\frac{1}{2}(S^{-1} \boldsymbol{\xi} \boldsymbol{\xi}^\top S^{-1} + S^{-1}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu} \\ -\frac{1}{2}(\boldsymbol{\mu} \boldsymbol{\mu}^\top + \Sigma) \end{bmatrix}, \\ \mathbb{E}_{\hat{\chi}^{(t+1)}} T(\mathbf{X}) &= \begin{bmatrix} \mathbb{E}_{\hat{\chi}^{(t+1)}} \mathbf{X} \\ -\frac{1}{2} \mathbb{E}_{\hat{\chi}^{(t+1)}} \mathbf{X} \mathbf{X}^\top \end{bmatrix} \propto \sum_{i=1}^n (\hat{\chi} \cdot r^{(t+1)})(\mathbf{x}_i) \begin{bmatrix} \mathbf{x}_i \\ -\frac{1}{2} \mathbf{x}_i \mathbf{x}_i^\top \end{bmatrix}, \end{aligned}$$

where we have omitted the component subscript  $k$ . Thus, from (13.5) we obtain:

$$\boldsymbol{\mu}_k^{(t+1)} = \mathbb{E}_{\hat{\chi}_k^{(t+1)}} \mathbf{X} = \sum_{i=1}^n \frac{(\hat{\chi} \cdot r_k^{(t+1)})(\mathbf{x}_i)}{\sum_{\ell=1}^n (\hat{\chi} \cdot r_k^{(t+1)})(\mathbf{x}_\ell)} \cdot \mathbf{x}_i \quad (13.6)$$

$$\Sigma_k^{(t+1)} = \mathbb{E}_{\hat{\chi}_k^{(t+1)}} \mathbf{X} \mathbf{X}^\top - \boldsymbol{\mu}_k^{(t+1)} \boldsymbol{\mu}_k^{(t+1)\top} = \sum_{i=1}^n \frac{(\hat{\chi} \cdot r_k^{(t+1)})(\mathbf{x}_i)}{\sum_{\ell=1}^n (\hat{\chi} \cdot r_k^{(t+1)})(\mathbf{x}_\ell)} \cdot (\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)})(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)})^\top, \quad (13.7)$$

where we remind that the empirical training distribution  $\hat{\chi} \equiv \frac{1}{n}$ .

The updates on “responsibility”  $\mathbf{r}$  in (13.2), mixing distribution  $\boldsymbol{\pi}$  in (13.3), on the means  $\boldsymbol{\mu}_k$  in (13.6), and on the covariance matrices  $S_k$  in (13.7), consist of the main steps for estimating a GMM using EM.