Expectation Maximization

STATS 305C: Applied Statistics

Scott Linderman

April 20, 2022

Recall our Bayesian Mixture Model

1. Sample the proportions from a Dirichlet prior:

$$\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha}) \tag{1}$$

2. Sample the parameters for each component:

$$\boldsymbol{\theta}_k \overset{\text{iid}}{\sim} p(\boldsymbol{\theta} \mid \boldsymbol{\phi}, \nu) \qquad \text{for } k = 1, \ldots, K \tag{2}$$

3. Sample the assignment of each data point:

$$z_n \overset{\text{iid}}{\sim} \boldsymbol{\pi} \qquad \text{for } n = 1, \ldots, N \tag{3}$$

4. Sample data points given their assignments:

$$\boldsymbol{x}_n \sim p(\boldsymbol{x} \mid \boldsymbol{\theta}_{z_n}) \qquad \text{for } n = 1, \ldots, N \tag{4}$$

Joint distribution

- This generative model corresponds to the following factorization of the joint distribution,

$$p(\pi, \{\boldsymbol{\theta}_k\}_{k=1}^K, \{(z_n, \boldsymbol{x}_n)\}_{n=1}^N \mid \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) = p(\pi \mid \boldsymbol{\alpha}) \prod_{k=1}^K p(\boldsymbol{\theta}_k \mid \boldsymbol{\phi}, \nu) \prod_{n=1}^N p(z_n \mid \pi)\, p(\boldsymbol{x}_n \mid \boldsymbol{z}_n, \{\boldsymbol{\theta}_k\}_{k=1}^K) \tag{5}$$

- Equivalently,

$$p(\pi, \{\boldsymbol{\theta}_k\}_{k=1}^K, \{(z_n, \boldsymbol{x}_n)\}_{n=1}^N \mid \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) = p(\pi \mid \boldsymbol{\alpha}) \prod_{k=1}^K p(\boldsymbol{\theta}_k \mid \boldsymbol{\phi}, \nu) \prod_{n=1}^N \prod_{k=1}^K \left[\Pr(z_n = k \mid \pi)\, p(\boldsymbol{x}_n \mid \boldsymbol{\theta}_k)\right]^{\mathbb{I}[z_n = k]} \tag{6}$$

- Substituting in the assumed forms

$$p(\pi, \{\boldsymbol{\theta}_k\}_{k=1}^K, \{(z_n, \boldsymbol{x}_n)\}_{n=1}^N \mid \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) = \mathrm{Dir}(\pi \mid \boldsymbol{\alpha}) \prod_{k=1}^K p(\boldsymbol{\theta}_k \mid \boldsymbol{\phi}, \nu) \prod_{n=1}^N \prod_{k=1}^K \left[\pi_k\, p(\boldsymbol{x}_n \mid \boldsymbol{\theta}_k)\right]^{\mathbb{I}[z_n = k]} \tag{7}$$

Exponential family mixture models

What about $p(\boldsymbol{x} \mid \boldsymbol{\theta}_k)$ and $p(\boldsymbol{\theta}_k \mid \boldsymbol{\phi}, \nu)$?

Let's assume an **exponential family** likelihood,

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}_k) = h(\boldsymbol{x}_n) \exp\left\{\langle t(\boldsymbol{x}_n), \boldsymbol{\theta}_k\rangle - A(\boldsymbol{\theta}_k)\right\}. \tag{8}$$

Then assume a **conjugate prior**,

$$p(\boldsymbol{\theta}_k \mid \boldsymbol{\phi}, \nu) \propto \exp\left\{\langle \boldsymbol{\phi}, \boldsymbol{\theta}_k\rangle - \nu A(\boldsymbol{\theta}_k)\right\}. \tag{9}$$

The hyperparmeters $\boldsymbol{\phi}$ are **pseudo-observations** of the sufficient statistics (like statistics from fake data points) and ν is a **pseudo-count** (like the number of fake data points).

Note that the product of prior and likelihood remains in the same family as the prior. That's why we call it conjugate.

Example: Gaussian mixture model

Assume the conditional distribution of $\boldsymbol{x}_n$ is a Gaussian with mean $\boldsymbol{\theta}_k \in \mathbb{R}^D$ and identity covariance,

$$
\begin{aligned}
p(\boldsymbol{x}_n \mid \boldsymbol{\theta}_k) &= \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\theta}_k, \boldsymbol{I}) && (10)\\
&= (2\pi)^{-D/2} \exp\left\{-\tfrac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\theta}_k)^\top (\boldsymbol{x}_n - \boldsymbol{\theta}_k)\right\} && (11)\\
&= (2\pi)^{-D/2} \exp\left\{-\tfrac{1}{2}\boldsymbol{x}_n^\top \boldsymbol{x}_n + \boldsymbol{x}_n^\top \boldsymbol{\theta}_k - \tfrac{1}{2}\boldsymbol{\theta}_k^\top \boldsymbol{\theta}_k\right\}, && (12)
\end{aligned}
$$

which is an exponential family distribution with base measure $h(\boldsymbol{x}_n) = (2\pi)^{-D/2} e^{-\frac{1}{2}\boldsymbol{x}_n^\top \boldsymbol{x}_n}$, sufficient statistics $t(\boldsymbol{x}_n) = \boldsymbol{x}_n$, and log normalizer $A(\boldsymbol{\theta}_k) = \underline{\tfrac{1}{2}\boldsymbol{\theta}_k^\top \boldsymbol{\theta}_k}$.

$\nabla A(\theta) = \theta$

The conjugate prior is a Gaussian prior on the mean,

$$
p(\boldsymbol{\theta}_k \mid \boldsymbol{\phi}, \nu) = \mathcal{N}(\nu^{-1}\boldsymbol{\phi}, \nu^{-1}\boldsymbol{I}) \propto \exp\left\{\boldsymbol{\phi}^\top \boldsymbol{\theta}_k - \tfrac{\nu}{2}\boldsymbol{\theta}_k^\top \boldsymbol{\theta}_k\right\} = \exp\left\{\boldsymbol{\phi}^\top \boldsymbol{\theta}_k - \nu A(\boldsymbol{\theta}_k)\right\}. \quad (13)
$$

Note that $\boldsymbol{\phi}$ sets the location and ν sets the precision (i.e. inverse variance).

EM in the Gaussian mixture model

K-Means made **hard assignments** of data points to clusters in each iteration. What if we used **soft assignments** instead?

Instead of assigning $z_n^\star$ to the closest cluster, we compute *responsibilities* for each cluster:

1. For each data point n and component k, set the *responsibility* to,

$$\omega_{nk} = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\theta}_k, \boldsymbol{I})}{\sum_{j=1}^K \pi_j \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\theta}_j, \boldsymbol{I})}. \tag{14}$$

2. For each component k, set the new mean to

$$\boldsymbol{\theta}_k^\star = \frac{1}{N_k} \sum_{n=1}^K \omega_{nk} \boldsymbol{x}_n, \tag{15}$$

   where $N_k = \sum_{n=1}^N \omega_{nk}$.

This is called the **expectation maximization (EM)** algorithm.

What is EM doing?

Rather than maximizing the **joint probability**, EM is maximizing the **marginal probability**,

$$\log p(\boldsymbol{X}, \boldsymbol{\theta}) = \log p(\boldsymbol{\theta}) + \log \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta}) \tag{16}$$

$$= \log p(\boldsymbol{\theta}) + \log \prod_{n=1}^{N} \sum_{z_n} p(\boldsymbol{x}_n, z_n \mid \boldsymbol{\theta}) \tag{17}$$

$$= \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} \log \sum_{z_n} p(\boldsymbol{x}_n, z_n \mid \boldsymbol{\theta}) \tag{18}$$

For discrete mixtures (with small enough K) we can evaluate the log marginal probability (with what complexity?).

We can usually evaluate its gradient too, so we could just do gradient ascent to find $\boldsymbol{\theta}^*$.

However, EM typically obtains faster convergence rates.

What is EM doing? II

Idea: Obtain a lower bound on the marginal probability,

$$\log p(\boldsymbol{X}, \boldsymbol{\theta}) = \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} \log \sum_{z_n} p(\boldsymbol{x}_n, z_n \mid \boldsymbol{\theta}) \tag{19}$$

$$= \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} \log \sum_{z_n} q(z_n) \frac{p(\boldsymbol{x}_n, z_n \mid \boldsymbol{\theta})}{q(z_n)} \tag{20}$$

$$= \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} \log \mathbb{E}_{q(z_n)} \left[\frac{p(\boldsymbol{x}_n, z_n \mid \boldsymbol{\theta})}{q(z_n)} \right] \tag{21}$$

where $q(z_n)$ is any distribution on $z_n \in \{1, \ldots, K\}$ such that $p(\boldsymbol{x}_n, z_n \mid \boldsymbol{\theta})$ is **absolutely continuous** w.r.t. $q(z_n)$.

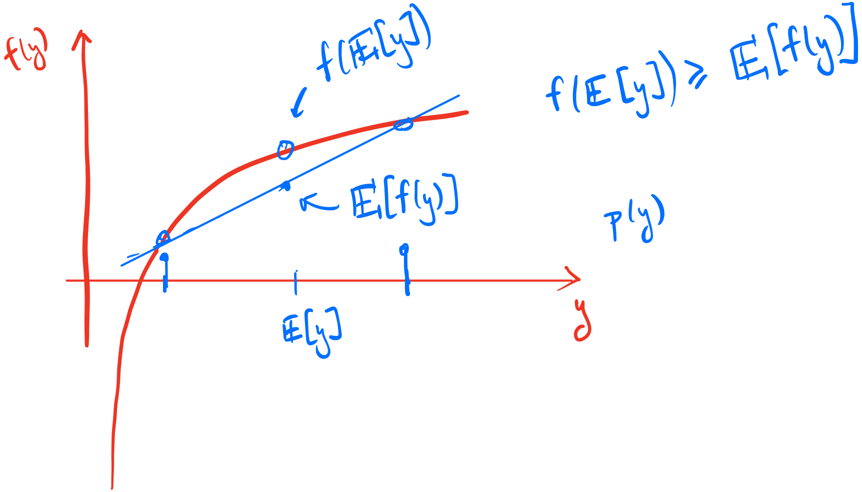
Jensen's Inequality

Jensen's inequality states that,

$$f(\mathbb{E}_{p(y)}[y]) \geq \mathbb{E}_{p(y)}[f(y)] \tag{22}$$

if f is a **concave function**, with equality iff f is linear.

[Picture]



What is EM doing? III

Applied to the log marginal probability, Jensen's inequality yields,

$$\log p(\boldsymbol{X}, \boldsymbol{\theta}) = \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} \log \mathbb{E}_{q_n(z_n)} \left[\frac{p(\boldsymbol{x}_n, z_n \mid \boldsymbol{\theta})}{q_n(z_n)} \right] \tag{23}$$

$$\geq \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} \mathbb{E}_{q_n(z_n)} [\log p(\boldsymbol{x}_n, z_n \mid \boldsymbol{\theta}) - \log q_n(z_n)] \tag{24}$$

$$\triangleq \mathscr{L}[\boldsymbol{\theta}, \boldsymbol{q}] \tag{25}$$

ELBO

where $\boldsymbol{q} = (q_1, \ldots, q_N)$ is a tuple of densities.

This is called the **evidence lower bound**, or **ELBO** for short.

It is a function of $\boldsymbol{\theta}$ and a **functional** of $\boldsymbol{q}$, since each q_n is a probability density function.

We can think of **EM as coordinate ascent on the ELBO**.

M-step: Maximizing the ELBO wrt θ (Gaussian case)

Suppose we fix $\boldsymbol{q}$. Since each z_n is a discrete latent variable, q_n must be a probability mass function. Let it be denoted by,

$$q_n(z_n) = [q_n(z_n = 1), \ldots, q_n(z_n = K)]^\top = [\omega_{n1}, \ldots, \omega_{nK}]^\top. \tag{26}$$

(These will be the **responsibilities** from before.)

Now, recall our basic model, $\boldsymbol{x}_n \sim \mathcal{N}(\boldsymbol{\theta}_{z_n}, \boldsymbol{I})$, and assume a prior $\boldsymbol{\theta}_k \sim \mathcal{N}(\boldsymbol{\phi}, \nu^{-1}\boldsymbol{I})$, Then,

$$\begin{aligned}
\mathcal{L}[\boldsymbol{\theta}, \boldsymbol{q}] &= \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} \mathbb{E}_{q_n(z_n)}[\log p(\boldsymbol{x}_n, z_n \mid \boldsymbol{\theta})] + c && (27)\\
&= \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} \sum_{k=1}^{K} \omega_{nk} \log p(\boldsymbol{x}_n, z_n = k \mid \boldsymbol{\theta}) + c && (28)\\
&= \sum_{k=1}^{K} \left[\boldsymbol{\phi}^\top \boldsymbol{\theta}_k - \tfrac{\nu}{2} \boldsymbol{\theta}_k^\top \boldsymbol{\theta}_k\right] + \sum_{n=1}^{N} \sum_{k=1}^{K} \omega_{nk} \left[\boldsymbol{x}_n^\top \boldsymbol{\theta}_k - \tfrac{1}{2} \boldsymbol{\theta}_k^\top \boldsymbol{\theta}_k\right] + c && (29)
\end{aligned}$$

M-step: Maximizing the ELBO wrt θ (Gaussian case) II

Zooming in on just $\boldsymbol{\theta}_k$,

$$\mathscr{L}[\boldsymbol{\theta}, \boldsymbol{q}] = \boldsymbol{\phi}_{N,k}^\top \boldsymbol{\theta}_k - \tfrac{1}{2} \nu_{N,k} \boldsymbol{\theta}_k^\top \boldsymbol{\theta}_k \tag{30}$$

where

$$\boldsymbol{\phi}_{N,k} = \boldsymbol{\phi} + \sum_{n=1}^{N} \omega_{nk} \boldsymbol{x}_n \qquad \nu_{N,k} = \nu + \sum_{n=1}^{N} \omega_{nk} \tag{31}$$

Taking derivatives and setting to zero yields,

$$\boldsymbol{\theta}_k^\star = \frac{\boldsymbol{\phi}_{N,k}}{\nu_{N,k}} = \frac{\boldsymbol{\phi} + \sum_{n=1}^{N} \omega_{nk} \boldsymbol{x}_n}{\nu + \sum_{n=1}^{N} \omega_{nk}}. \tag{32}$$

In the improper uniform prior limit where $\boldsymbol{\phi} \to 0$ and $\nu \to 0$, we recover the EM updates shown on slide 6.

E-step: Maximizing the ELBO wrt q (Gaussian case)

As a function of q_n, for discrete Gaussian mixtures with identity covariance,

$$\mathscr{L}[\boldsymbol{\theta}, \boldsymbol{q}] = \mathbb{E}_{q_n(z_n)}\left[\log p(\boldsymbol{x}_n, z_n \mid \boldsymbol{\theta}) - \log q_n(z_n)\right] + c \tag{33}$$

$$= \sum_{k=1}^{K} \omega_{nk}\left[\log \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\theta}_k, \boldsymbol{I}) + \log \pi_k - \log \omega_{nk}\right] + c \tag{34}$$

where $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_K]^\top$ is the vector of cluster probabilities.

We also have two constraints: $\omega_{nk} \geq 0$ and $\sum_k \omega_{nk} = 1$. Let's ignore the non-negative constraint for now (it will automatically be satisfied anyway) and write the Lagrangian with the simplex constraint,

$$\mathscr{J}(\boldsymbol{\omega}_n, \lambda) = \sum_{k=1}^{K} \omega_{nk}\left[\log \mathcal{N}(\boldsymbol{x}_n \mid \theta_k, \boldsymbol{I}) + \log \pi_k - \log \omega_{nk}\right] - \lambda\left(1 - \sum_{k=1}^{K} \omega_{nk}\right) \tag{35}$$

E-step: Maximizing the ELBO wrt q (Gaussian case) II

Taking the partial derivative wrt ω_{nk} and setting to zero yields,

$$\frac{\partial}{\partial \omega_{nk}} \mathscr{J}(\boldsymbol{\omega}_n, \lambda) = \log \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\theta}_k, \boldsymbol{I}) + \log \pi_k - \log \omega_{nk} - 1 + \lambda = 0 \tag{36}$$

$$\Rightarrow \log \omega^\star_{nk} = \log \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\theta}_k, \boldsymbol{I}) + \log \pi_k + \lambda - 1 \tag{37}$$

$$\Rightarrow \omega^\star_{nk} \propto \pi_k \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\theta}_k, \boldsymbol{I}) \tag{38}$$

Enforcing the simplex constraint yields,

$$\omega^\star_{nk} = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\theta}_k, \boldsymbol{I})}{\sum_{j=1}^K \pi_j \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\theta}_j, \boldsymbol{I})}, \tag{39}$$

just like on slide 6.

Note that

$$\omega^\star_{nk} \propto p(z_n = k)\, p(\boldsymbol{x}_n \mid z_n = k, \boldsymbol{\theta}) = p(z_n = k \mid \boldsymbol{x}_n, \boldsymbol{\theta}) \tag{40}$$

The ELBO is tight after the E-step

Equivalently, q_n equals the posterior, $p(z_n \mid \boldsymbol{x}_n, \boldsymbol{\theta})$. At that point, the ELBO simplifies to,

$$\mathscr{L}[\boldsymbol{\theta}, \boldsymbol{q}] = \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} \mathbb{E}_{q_n(z_n)}\left[\log p(\boldsymbol{x}_n, z_n \mid \boldsymbol{\theta}) - \log q_n(z_n)\right] \tag{41}$$

$$= \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} \mathbb{E}_{p(z_n \mid \boldsymbol{x}_n, \boldsymbol{\theta})}\left[\log p(\boldsymbol{x}_n, z_n \mid \boldsymbol{\theta}) - \log p(z_n \mid \boldsymbol{x}_n, \boldsymbol{\theta})\right] \tag{42}$$

$$= \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} \mathbb{E}_{p(z_n \mid \boldsymbol{x}_n, \boldsymbol{\theta})}\left[\log p(\boldsymbol{x}_n \mid \boldsymbol{\theta})\right] \tag{43}$$

$$p(z \mid x, \theta) = \frac{p(z, x \mid \theta)}{p(x \mid \theta)}$$

$$= \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} \log p(\boldsymbol{x}_n \mid \boldsymbol{\theta}) \tag{44}$$

$$= \log p(\boldsymbol{X}, \boldsymbol{\theta}) \tag{45}$$

In other words, **after the E step, the bound is tight**!

EM as a minorize-maximize (MM) algorithm

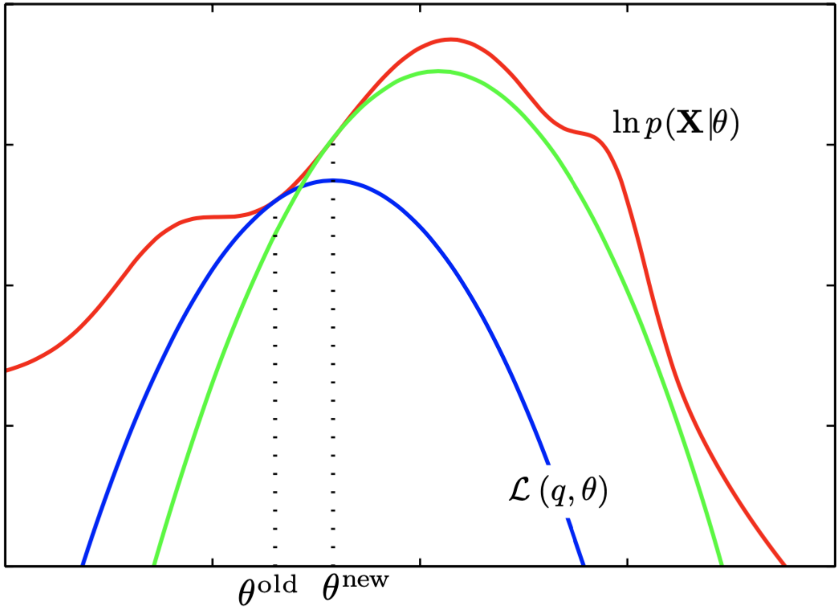


Figure: Bishop, Figure 9.14: EM alternates between constructing a lower bound (minorizing) and finding new parameters that maximize it.

M-step: Maximizing the ELBO wrt θ (generic exp. fam.)

Now let's consider the general Bayesian mixture with exponential family likelihoods and conjugate priors. As a function of $\boldsymbol{\theta}$,

$$\mathscr{L}[\boldsymbol{\theta}, \boldsymbol{q}] = \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} \mathbb{E}_{q_n(z_n)}[\log p(\boldsymbol{x}_n, z_n \mid \boldsymbol{\theta})] + \mathrm{c} \tag{46}$$

$$= \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} \sum_{k=1}^{K} \omega_{nk} \log p(\boldsymbol{x}_n, z_n = k \mid \boldsymbol{\theta}) + \mathrm{c} \tag{47}$$

$$= \sum_{k=1}^{K} \left[\boldsymbol{\phi}^\top \boldsymbol{\theta}_k - \nu A(\boldsymbol{\theta}_k) \right] + \sum_{n=1}^{N} \sum_{k=1}^{K} \omega_{nk} \left[t(\boldsymbol{x}_n)^\top \boldsymbol{\theta}_k - A(\boldsymbol{\theta}_k) \right] + \mathrm{c} \tag{48}$$

M-step: Maximizing the ELBO wrt θ (generic exp. fam.) II

Zooming in on just $\boldsymbol{\theta}_k$,

$$\mathscr{L}[\boldsymbol{\theta}, \boldsymbol{q}] = \boldsymbol{\phi}_{N,k}^\top \boldsymbol{\theta}_k - \nu_{N,k} A(\boldsymbol{\theta}_k) \tag{49}$$

$\nabla_\theta \mathcal{L}[\theta, q] = \phi_{Nk} - \nu_{Nk} \nabla A(\theta) \rightarrow \theta^k = [\nabla A]^{-1}\left(\frac{\phi_{Nk}}{\nu}\right)$

where

$$\boldsymbol{\phi}_{N,k} = \boldsymbol{\phi} + \sum_{n=1}^{N} \underbrace{\omega_{nk} t(\boldsymbol{x}_n)} \qquad \nu_{N,k} = \nu + \sum_{n=1}^{N} \omega_{nk} \tag{50}$$

Taking derivatives and setting to zero yields, $\mathbb{E}_{q_n}[\mathbb{I}[z_n = k]\, t(x_n)] \longrightarrow$ "E step"

$$\boldsymbol{\theta}_k^* = [\nabla A]^{-1} \left(\frac{\boldsymbol{\phi}_{N,k}}{\nu_{N,k}} \right) \tag{51}$$

M-step: Maximizing the ELBO wrt θ (generic exp. fam.) III

What is the gradient of the log normalizer? We have,

$$\nabla A(\boldsymbol{\theta}_k) = \nabla_{\boldsymbol{\theta}_k} \log \int h(\boldsymbol{x}) \exp\left\{\langle t(\boldsymbol{x}), \boldsymbol{\theta}_k \rangle\right\} \mathrm{d}\boldsymbol{x} \tag{52}$$

$$= \frac{\int h(\boldsymbol{x}) \exp\left\{\langle t(\boldsymbol{x}), \boldsymbol{\theta}_k \rangle\right\} t(\boldsymbol{x}) \; \mathrm{d}\boldsymbol{x}}{\int h(\boldsymbol{x}) \exp\left\{\langle t(\boldsymbol{x}_n), \boldsymbol{\theta}_k \rangle\right\} \mathrm{d}\boldsymbol{x}} \tag{53}$$

$\exp\{A(\theta_k)\}$

$$= \int \underbrace{h(\boldsymbol{x}) \exp\left\{\langle t(\boldsymbol{x}), \boldsymbol{\theta}_k \rangle - A(\boldsymbol{\theta}_k)\right\}}_{p(x \mid \theta_k)} t(\boldsymbol{x}) \; \mathrm{d}\boldsymbol{x} \tag{54}$$

$$= \mathbb{E}_{p(\boldsymbol{x} \mid \boldsymbol{\theta}_k)}[t(\boldsymbol{x})] \tag{55}$$

Gradients of the log normalizer yield expected sufficient statistics!

M-step: Maximizing the ELBO wrt θ (generic exp. fam.) IV

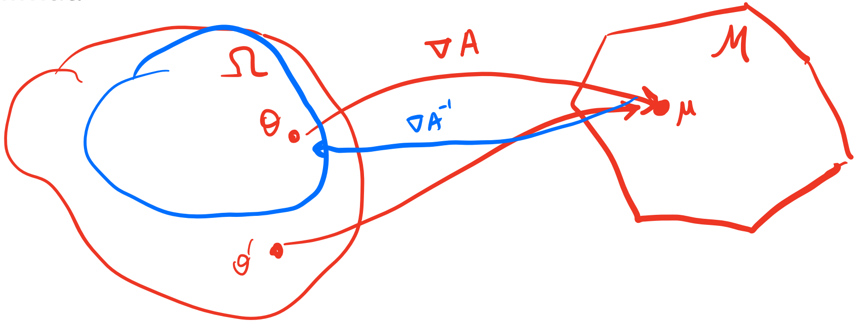
The gradient ∇A is a map from the set of valid natural parameters Ω (those for which the log normalizer is finite) to the set of realizable mean parameters $\mathscr{M}$, ← marginal polytope

$$\mathscr{M} = \{\boldsymbol{\mu} \in \mathbb{R}^D : \exists p \text{ s.t. } \mathbb{E}_p[t(\boldsymbol{x})] = \boldsymbol{\mu}\} \tag{56}$$

An exponential family is **minimal** if its sufficient statistics are linearly independent.

Fact: The gradient mapping $\nabla A : \Omega \to \mathscr{M}$ is one-to-one (and hence invertible) if and only if the exponential family is minimal.

<Picture>



M-step: Maximizing the ELBO wrt θ (generic exp. fam.) V

Thus, the generic M-step in eq. 51 amounts to finding the natural parameters $\boldsymbol{\theta}_k^*$ that yield the expected sufficient statistics $\boldsymbol{\phi}_{N,k}/\nu_{N,k}$ by inverting the gradient mapping.

Note: There is a longer and much more technical story about exponential families, maximum likelihood, convex analysis, and conjugate duals that you can read about in [Wainwright et al., 2008, Ch. 3] if you are interested.

E-step: Maximizing the ELBO wrt q (generic exp. fam.)

In our first pass, we assumed q_n was a finite pmf. More generally, q_n will be a probability density function, and optimizing over functions usually requires the **calculus of variations**. (Ugh!)

However, note that we can write the ELBO in a slightly different form,

$$\mathcal{L}[\boldsymbol{\theta}, \boldsymbol{q}] = \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} \mathbb{E}_{q_n(z_n)} [\log p(\boldsymbol{x}_n, z_n \mid \boldsymbol{\theta}) - \log q_n(z_n)] \tag{57}$$

$$= \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} \mathbb{E}_{q_n(z_n)} [\log p(z_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}) + \log p(\boldsymbol{x}_n \mid \boldsymbol{\theta}) - \log q_n(z_n)] \tag{58}$$

$$= \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} [\log p(\boldsymbol{x}_n \mid \boldsymbol{\theta}) - D_{\mathrm{KL}} (q_n(z_n) \,\|\, p(z_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}))] \tag{59}$$

$$= \log p(\boldsymbol{X}, \boldsymbol{\theta}) - \sum_{n=1}^{N} D_{\mathrm{KL}} (q_n(z_n) \,\|\, p(z_n \mid \boldsymbol{x}_n, \boldsymbol{\theta})) \tag{60}$$

where $D_{\mathrm{KL}}(\cdot \,\|\, \cdot)$ denote the **Kullback-Leibler divergence**.

Kullback-Leibler (KL) divergence

The KL divergence is defined as,

$$D_{\mathrm{KL}}\left(q(z) \,\|\, p(z)\right) = \int q(z) \log \frac{q(z)}{p(z)} \,\mathrm{d}z. \tag{61}$$

It gives a notion of how similar two distributions are, but it is **not a metric!** (It is not symmetric, e.g.)
Still, it has some intuitive properties:

- It is non-negative, $D_{\mathrm{KL}}\left(q(z) \,\|\, p(z)\right) \geq 0$.
- It equals zero iff the distributions are the same, $D_{\mathrm{KL}}\left(q(z) \,\|\, p(z)\right) = 0 \iff q(z) = p(z)$ almost everywhere.

E-step: Maximizing the ELBO wrt q (generic exp. fam.) II

Maximizing the ELBO wrt q_n amounts to minimizing the KL divergence to the posterior $p(z_n \mid \boldsymbol{x}_n, \boldsymbol{\theta})$,

$$\mathscr{L}[\boldsymbol{\theta}, \boldsymbol{q}] = \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} [\log p(\boldsymbol{x}_n \mid \boldsymbol{\theta}) - D_{\mathrm{KL}}(q_n(z_n) \,\|\, p(z_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}))] \tag{62}$$

$$= -D_{\mathrm{KL}}(q_n(z_n) \,\|\, p(z_n \mid \boldsymbol{x}_n, \boldsymbol{\theta})) + c \tag{63}$$

As we said, the KL is minimized when $q_n(z_n) = p(z_n \mid \boldsymbol{x}_n, \boldsymbol{\theta})$, so the optimal update is,

$$q_n^\star(z_n) = p(z_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}), \tag{64}$$

just like we found on slide 14.

Problem: $\theta^* = \underset{\theta}{\arg\max} \log p(x, \theta)$

References I

Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

$$NIW(\mu, \Sigma) = N(\mu \mid \mu_0, \kappa_0^{-1}\Sigma)\, IW(\Sigma \mid \nu_0, \Sigma_0)$$

$$= |\kappa_0^{-1}\Sigma|^{-1/2} \exp\left\{-\frac{\kappa_0}{2}(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0)\right\} |\Sigma|^{-\nu_0/2} \exp\left\{-\frac{1}{2}Tr(\Sigma^{-1}\Sigma_0)\right\}$$

$$\log NIW(\mu, \Sigma) = -\frac{1}{2}\log|\Sigma| - \frac{\kappa_0}{2}\mu^T\Sigma^{-1}\mu + \kappa_0\mu_0^T\Sigma^{-1}\mu - \frac{\kappa_0}{2}\mu_0^T\Sigma^{-1}\mu_0 - \frac{\nu_0}{2}\log|\Sigma| - \frac{1}{2}Tr(\Sigma^{-1}\Sigma_0)$$

$$= -\frac{(\nu_0+1)}{2}\log|\Sigma| - \frac{\kappa_0}{2}\mu^T\Sigma^{-1}\mu + \kappa_0\mu_0^T\Sigma^{-1}\mu - \frac{1}{2}Tr\left((\Sigma_0 + \kappa_0\mu_0\mu_0^T)\Sigma^{-1}\right)$$

$$= \left\langle -\frac{(\nu_0+1)}{2}, \log|\Sigma| \right\rangle + \left\langle -\frac{\kappa_0}{2}, \mu^T\Sigma^{-1}\mu \right\rangle + \left\langle \kappa_0\mu_0, \Sigma^{-1}\mu \right\rangle + \left\langle -\frac{1}{2}(\Sigma_0 + \kappa_0\mu_0\mu_0^T), \Sigma^{-1} \right\rangle$$