Chapter 7

Variational Inference

We turn our attention to variational inference, *aka* deterministic approximate inference. Some inference problems have prohibitive complexity because the underlying probability distributions are untractable. A powerful idea is then to approximate such distributions by more tractable ones. This generally involves solving an optimization problem to find the distribution that is best in a prescribed approximation class. The optimization problem can often be solved using variational methods, hence the name of the approach. The origins of variational inference can be traced to statistical physics. In some cases, one can derive bounds on the performance of variational methods.

References on this topic include the review paper by Blei on variational Bayes [1], and the book by Wainwright and Jordan, which emphasizes the deep connections of variational inference to graphical models, exponential families of distribution, and convexity analysis [2]. For a more introductory exposition, see Chapter 10 in Bishop's book [3], or Chapter 28 in Barber's book [4].

We will denote by P the "true" distribution and by Q its approximation. In a typical application of variational inference, an unknown state x is to be estimated given observations y [1]. The posterior distribution $\pi(x|y)$ is to be approximated with some element $q(x)$ of a tractable family, e.g., a parametric family, or a product distribution.

7.1 Naive Mean-Field Methods

The *naive mean field* method approximates a distribution by a product distribution. Namely, the target distribution $p(\boldsymbol{x})$, $\boldsymbol{x} \in \mathcal{X}^n$ is approximated by $q(\boldsymbol{x}) = \prod_{i=1}^n q_i(x_i)$. One may use Kullback-Leibler divergence as an approximation criterion and seek $\{q_i\}_{i=1}^n$ that minimize $D(q\|p)$. One might think that the solution is obtained by simply matching the marginals of p, but this is generally not true. [1]

Whereas KL divergence is convex, the feasible set for q is nonconvex. Fortunately for each i, minimizing KL divergence over q_i alone (with the other coordinates fixed) is a convex program. We derive necessary conditions for optimality of a candidate q using

[1] However for the related problem of minimizing $D(p\|q)$ over q, the solution is obtained by matching the marginals of p. This does not imply the solution is easy to compute. As discussed in the previous chapter, if p is a MRF whose graph contains loops, computing marginals can be a hard problem.

the following variational approach. We use the following notation. Consider the vector $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)^\top$. We denote by $\boldsymbol{x}_{\setminus i}$ the components of $\boldsymbol{x}$ other than component x_i. If the vector is random with product distribution $q(\boldsymbol{x}) = \prod_i q_i(x_i)$, we denote by $q_{\setminus i} = \prod_{j \neq i} q_j$ the resulting product distribution on $\boldsymbol{x}_{\setminus i}$.

We have

$$\begin{aligned} D(q\|p) &= \mathbb{E}_q \left[\ln \frac{\prod_{i=1}^n q_i(X_i)}{p(\boldsymbol{X})} \right] \\ &= \sum_{i=1}^n \mathbb{E}_q[\ln q_i(X_i)] - \mathbb{E}_q[\ln p(\boldsymbol{X})] \\ &= \sum_{i=1}^n \mathbb{E}_q[\ln q_i(X_i)] - \mathbb{E}_{q_i} \underbrace{\mathbb{E}_{q_{\setminus i}}[\ln p(X_i, \boldsymbol{X}_{\setminus i})]}_{=\ln f(X_i)} \\ &= \sum_{i=1}^n \mathbb{E}_q[\ln q_i(X_i)] - \mathbb{E}_{q_i} \ln \tilde{p}(X_i) \\ &= D(q_i\|q_i^*) + c(q_{\setminus i}). \end{aligned} \tag{7.1}$$

where in the third line we have introduced the function

$$f(x_i) = \exp\{\mathbb{E}_{q_{\setminus i}}[\ln p(x_i, \boldsymbol{X}_{\setminus i})]\}$$

which is nonnegative and can be normalized to be a probability distribution

$$q_i^*(x_i) = \frac{1}{Z_i} f(x_i).$$

In the last line of (7.1), $c(q_{\setminus i})$ is a term that does not depend on q_i. Given $q_{\setminus i}$, KL divergence is therefore minimized over q_i by selecting $q_i = q_i^*$.

A reasonable iterative approach to minimize $D(q\|p)$ consists in initializing $\{q_i\}$ and cycling through the update equations

$$q_i(x_i) = \frac{1}{Z_i} \exp\{\mathbb{E}_{q_{\setminus i}}[\ln p(x_i, \boldsymbol{X}_{\setminus i})]\}, \quad 1 \leq i \leq n,\ x_i \in \mathcal{X} \tag{7.2}$$

until convergence. Convergence is guaranteed because this coordinatewise approach is greedy, and the cost function is lower-bounded (by 0).

Finally, note that the optimal q_i may equivalently be written as

$$q_i(x_i) = \frac{1}{Z_i} \exp\{\mathbb{E}_{\setminus i}[\ln p(x_i|\boldsymbol{X}_{\setminus i})]\}, \quad 1 \leq i \leq n,\ x_i \in \mathcal{X} \tag{7.3}$$

(with a different normalization factor Z_i) where $p(x_i|x_{\setminus i})$ is the so-called *complete conditional distribution* of x_i.

7.1.1 Factorized Approximation to Multivariate Gaussian

Consider a n-dimensional Gaussian vector $\boldsymbol{X}$ with mean $\boldsymbol{\mu}$ and precision matrix J. Its pdf is given by

$$p(\boldsymbol{x}) = (2\pi)^{-n/2} |\mathsf{J}|^{1/2} \exp\left\{ -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^\top \mathsf{J} (\boldsymbol{x} - \boldsymbol{\mu}) \right\}.$$

Hence

$$p(x_i, \boldsymbol{X}_{\setminus i}) \propto \exp\left\{-\frac{1}{2}J_{ii}x_i^2 + x_i\left[J_{ii}\mu_i - \sum_{j\setminus i} J_{ij}(X_j - \mu_j)\right] + f(\boldsymbol{X}_{\setminus i})\right\}$$

where $f(\boldsymbol{X}_{\setminus i})$ is a (quadratic) function of $\boldsymbol{X}_{\setminus i}$. Substituting into (7.2), we obtain

$$q_i(x_i) \propto \exp\left\{\mathbb{E}_{q_{\setminus i}}\left(-\frac{1}{2}J_{ii}x_i^2 + x_i\left[J_{ii}\mu_i - \sum_{j\neq i} J_{ij}(X_j - \mu_j)\right] + f(\boldsymbol{X}_{\setminus i})\right)\right\}.$$

We see that the expectation of $f(\boldsymbol{X}_{\setminus i})$ can be absorbed into the proportionality constant and the expectation over X_j only involves q_j. Hence

$$q_i(x_i) \propto \exp\left\{-\frac{1}{2}J_{ii}x_i^2 + x_i\left[J_{ii}\mu_i - \sum_{j\neq i} J_{ij}(\mathbb{E}_{q_j}(X_j) - \mu_j)\right]\right\}$$

which is a Gaussian distribution with precision J_{ii} and mean

$$\mathbb{E}_{q_i}(X_i) = \mu_i - \sum_{j\neq i}\frac{J_{ij}}{J_{ii}}(\mathbb{E}_{q_j}(X_j) - \mu_j), \quad i = 1, 2, \cdots, n.$$

This is an $n \times n$ linear system of equations for the means. By inspection, the solution is simply $\mathbb{E}_{q_i}(X_i) = \mu_i$ for all i. Hence the naive mean-field approximation to the multivariate Gaussian is simply the product of its marginals. [2]

7.1.2 Graphical Models

The system (7.2) does not admit a closed-form solution and exact computation of the expectation is generally intractable. However tractable expressions can be derived for graphical models. Consider the pairwise Markov network

$$p(\boldsymbol{x}) = \frac{1}{Z}\prod_{(i,j)\in\mathcal{E}}\psi_{ij}(x_i, x_j) \tag{7.4}$$

defined over a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Then the expectation of (7.2) can be written as

$$\mathbb{E}_{q_{\setminus i}}[\ln p(x_i, X_{\setminus i})] = \sum_{j\in\mathcal{N}(i)}\mathbb{E}_{q_j}[\ln\psi_{ij}(x_i, X_j)] + \text{cst}$$

and thus

$$q_i(x_i) = \frac{1}{Z_i}\exp\left\{\sum_{j\in\mathcal{N}(i)}\sum_{x_j\in\mathcal{X}} q_j(x_j)\ln\psi_{ij}(x_i, x_j)\right\}. \tag{7.5}$$

[2] This property does not extend to non-Gaussian multivariate distributions.

Here the iterative minimization problem is tractable because the expectation is the sum of a small number of terms. Interestingly this algorithm admits a message-passing interpretation:

$$q_i(x_i) = \frac{1}{Z_i} \prod_{j \in \mathcal{N}(i)} m_{j \to i}(x_i)$$

if we define the messages

$$m_{j \to i}(x_i) = \exp \left\{ \sum_{x_j \in \mathcal{X}} q_j(x_j) \ln \psi_{ij}(x_i, x_j) \right\}.$$

Hence the solution $\{q_i\}_{i=1}^n$ is a stationary point of a belief-propagation algorithm.

Note from the second line of (7.1) and (7.4) that

$$D(q\|p) = -\sum_{i \in \mathcal{V}} H(q_i) - \sum_{(i,j) \in \mathcal{E}} \mathbb{E}_q \left[\ln \psi_{ij}(X_i, X_j) \right] - \ln Z.$$

The last term ($\ln Z$) does not need to be computed, and the progress of the algorithm can be monitored by evaluating the remainder of the right side.

7.1.3 Ising Model

Thie algorithm is now specialized to the 2-D Ising model, a simple and insighful model from statistical physics, where the mean-field method originated. The binary random variables represent electron spins which can have two states, ± 1. The simplest version of this model is one-dimensional and was studied by Ising [5]; the 2-D version which is described below, is much more difficult and was studied by Onsager [6]. Consider a 2-D torus $\mathcal{V}$ with $|\mathcal{V}| = n$ nodes, and $\mathcal{X} = \{\pm 1\}$. Each node is connected to its upper, lower, right, and left neighbors. The distribution is of the form

$$p(\boldsymbol{x}) = \frac{1}{Z} \exp \left\{ \beta \sum_{i \sim j} x_i x_j \right\}, \quad \boldsymbol{x} \in \{\pm 1\}^n \tag{7.6}$$

with $\beta \geq 0$. The parameter β represents the inverse of a temperature. For $\beta = 0$ the distribution is uniform, hence fully factorized. For large positive values of β, configurations $\boldsymbol{x}$ with strong correlations are favored.

Since each X_i is a Bernoulli random variable, q_i can be represented by a single parameter which we choose to be the mean $m_i = q_i(1) - q_i(-1) \in [-1, 1]$. Equivalently,

$$q_i(1) = \frac{1 + m_i}{2}, \quad q_i(-1) = \frac{1 - m_i}{2}.$$

Then (7.5) becomes

$$q_i(x_i) = \frac{1}{Z_i} \exp \left\{ \beta x_i \sum_{j \in \mathcal{N}(i)} m_j \right\}, \quad x_i \in \{\pm 1\}.$$

The normalization constant is given by

$$Z_i = 2\cosh\left(\beta \sum_{j\in\mathcal{N}(i)} m_j\right)$$

hence

$$m_i = q_i(1) - q_i(-1) = \tanh\left(\beta \sum_{j\in\mathcal{N}(i)} m_j\right).$$

Recall the hyperbolic tangent function is antisymmetric and increases from -1 to 1 as its argument goes from $-\infty$ to ∞. Its slope at the origin is 1.

Convergence. We show that the algorithm always converges if a uniform initialization is used, i.e., $m_i^{(0)} = m^{(0)}$ for all $i \in \mathcal{V}$. Then the (simultaneous) update equation for the means is

$$m^{(k+1)} = \tanh(4\beta m^{(k)}), \quad k = 0, 1, 2, \cdots \tag{7.7}$$

where the factor of 4 arises because each vertex has 4 neighbors. Analysis of convergence depends on the value of β, and two cases need to be considered.

Case I: $\beta < \frac{1}{4}$. The mapping of (7.7) is a contraction mapping for $\beta < \frac{1}{4}$, and so the fixed point of this mapping is $\lim_{k\to\infty} m^{(k)} = 0$, for any initialization $m^{(0)}$. Hence the variational approximation is uniform: $q(\boldsymbol{x}) = 2^{-n}$ for all $\boldsymbol{x} \in \{\pm 1\}^{\mathcal{V}}$.

Case II: $\beta > \frac{1}{4}$. In this case, the equation $m = \tanh(4\beta m)$ has three possible solutions 0 and $\pm m^*$ where $m^* > 0$. If the algorithm is initialized with $m^{(0)} = 0$, then subsequent iterations do not change this value. If the algorithm is initialized with $m^{(0)} > 0$, it converges to m^*. Finally, if the algorithm is initialized with $m^{(0)} < 0$, it converges to $-m^*$. In the latter two cases (convergence to either m^* or $-m^*$), the variational approximations q_i are nonuniform.

The case $\beta > \frac{1}{4}$ is related to *percolation theory* in statistical physics. It may be shown that the distribution p favors configurations featuring large homogeneous regions. The correlation between any two nodes is significant, even for large graphs. This behavior is completely different from the case $\beta < \frac{1}{4}$, where the correlation between distant nodes dies out with distance (similarly to a homogeneous, irreducible Markov chain). The case $\beta = \frac{1}{4}$ is known as a *phase transition*.

7.1.4 Structured Mean Field Methods

Since approximating a joint distribution with a fully factored distribution may be quite inaccurate, improvements on the naive mean field method have been sought. Structured mean field methods were introduced by Saul and Jordan [7] and perform the optimization over a conveniently structured classes of distributions. Two examples are given below.

Cluster-Factorized Distributions [9]. Partition the set of nodes $\mathcal{V}$ into k clusters $\mathcal{C}_1, \cdots, \mathcal{C}_k$; for the 2-D Ising model for instance, each cluster could be a 2×2 block of nodes,

hence $k = n/4$. Denote by x^c the restriction of $\boldsymbol{x}$ to cluster c, by $x^{\setminus c}$ the restriction of $\boldsymbol{x}$ to the complement set of nodes, and by q^c the cluster marginal distribution of x^c. A *cluster-factorized distribution* is a distribution of the form $q(\boldsymbol{x}) = \prod_{c=1}^k q^c(x^c)$. We wish to solve the minimization problem $\min_q D(q\|p)$ where q runs over all possible cluster-factorized distributions, and p is the pairwise Markov model of (7.4). Denote by

- $\mathcal{E}_c$ the set of edges with both endpoints in cluster c, and
- $\mathcal{S}_c$ the set of edges (i,j) where node i belongs to cluster c but j does not.

Proceeding as in the naive mean field analysis, we obtain the first-order optimality conditions for $\{q^c\}_{c=1}^k$:

$$\begin{aligned}
\forall c, x^c: \quad q^c(x^c) &\propto \exp\left\{\mathbb{E}_q[\ln p(x^c, X^{\setminus c})\right\} \\
&\propto \exp\left\{\mathbb{E}_q\left[\sum_{(i,j)\in\mathcal{E}_c} \ln\psi_{ij}(x_i,x_j) + \sum_{(i,j)\in\mathcal{S}_c} \ln\psi_{ij}(x_i,X_j)\right]\right\} \\
&= \frac{1}{Z_c}\left(\prod_{(i,j)\in\mathcal{E}_c}\psi_{ij}(x_i,x_j)\right)\left(\prod_{(i,j)\in\mathcal{S}_c}\underbrace{\exp\{\mathbb{E}_{q_j}[\ln\psi_{ij}(x_i,X_j)]\}}_{=m_{j\to i}(x_i)}\right) \quad (7.8)
\end{aligned}$$

In the second line, the edges (i,j) for which neither i nor j belongs to cluster c only contribute to a constant factor. The resulting conditions (7.8) are a generalization of (7.2).

An iterative solution can be derived by initializing the cluster distributions $\{q^c\}_{c=1}^k$ and iteratively updating them by application of (7.8). Note that these updates require knowledge of the node marginals q_j, which are obtained by marginalizing $q^{c(j)}$ where $c(j)$ is the cluster to which j belongs. This marginalization is easy if the cluster is small or if the junction-tree algorithm can be conveniently used. The paper [9] shows the quality of structured mean field when applied to a 2-D Ising model on a 8×8 grid, and clusters of size either 2×2 or 4×4. As expected, both outperform naive mean-field for large β (in which case p is a strongly correlated distribution).

Directed Acyclic Graphs. A more elaborate idea is to let q be a distribution over a Bayesian network with node set $\mathcal{V}$ and edge set $\mathcal{E}'$. For notational simplicity, we consider only Markov chains:

$$q(\boldsymbol{x}) = q_1(x_1)\prod_{i=2}^n r_i(x_i|x_{i-1})$$

where q_1 is the distribution of the initial state, and r_i are the transition probabilities. We again assume p is the pairwise Markov model of (7.4) with graph $(\mathcal{V},\mathcal{E})$; note that $\mathcal{E}'$ is a strict subset of $\mathcal{E}$. Then

$$\begin{aligned}
D(q\|p) &= \mathbb{E}_q\left[\ln q_1(X_1) + \sum_{i=2}^n \ln r_i(X_i|X_{i-1}) - \sum_{(i,j)\in\mathcal{E}} \ln\psi_{ij}(X_i,X_j)\right] + \ln Z \\
&= -H(q_1) - \sum_{i=2}^n H(r_i|q_{i-1}) - \sum_{(i,j)\in\mathcal{E}} \mathbb{E}_q[\ln\psi_{ij}(X_i,X_j)] + \ln Z
\end{aligned}$$

where

$$H(r_i|q_{i-1}) = -\mathbb{E}[\ln r_i(X_i|X_{i-1})] = -\sum_{x_i,x_{i-1}} q_{i-1}(x_{i-1}) r_i(x_i|x_{i-1}) \ln r_i(x_i|x_{i-1}), \quad i \geq 2$$

is the *conditional entropy* of X_i given X_{i-1}. The marginal distributions $\{q_i\}_{i=2}^n$ are obtained recursively as

$$q_i(x_i) = \sum_{x_{i-1}} q_{i-1}(x_{i-1}) r_i(x_i|x_{i-1}) \quad : \; i \geq 2, \tag{7.9}$$

and similarly the 2-D marginal distributions $\{q_{ij}\}_{1 \leq i < j \leq n}$ are obtained recursively as

$$q_{ij}(x_i, x_j) = \begin{cases} q_i(x_i) r_{i+1}(x_{i+1}|x_i) & : \; j = i+1 \geq 2 \\ \sum_{x_{j-1}} q_{i,j-1}(x_i, x_{j-1}) r_j(x_j|x_{j-1}) & : \; j > i+1 \geq 2. \end{cases} \tag{7.10}$$

We will need the following partial derivatives of these marginals:

$$\begin{aligned}
\frac{\partial q_i(x_i)}{\partial q_1(x_1)} &= \begin{cases} r_2(x_2|x_1) & : \; i = 2 \\ \sum_{x_{i-1}} \frac{\partial q_{i-1}(x_{i-1})}{\partial q_1(x_1)} r_i(x_i|x_{i-1}) & : \; i \geq 3 \end{cases} \\
\frac{\partial q_i(x_i)}{\partial r_j(x_j|x_{j-1})} &= \begin{cases} q_{i-1}(x_{i-1}) & : \; i = j \geq 2 \\ \sum_{x_{i-1}} \frac{\partial q_{i-1}(x_{i-1})}{\partial r_j(x_j|x_{j-1})} r_i(x_i|x_{i-1}) & : \; i \geq j+1 \geq 3. \end{cases} \\
\frac{\partial q_{ij}(x_i, x_j)}{\partial q_1(x_1)} &= \begin{cases} r_2(x_2|x_1) & : \; j = i+1 = 2 \\ \frac{\partial q_i(x_i)}{\partial q_1(x_1)} r_{i+1}(x_{i+1}|x_i) & : \; j = i+1 > 2 \\ \sum_{x_{j-1}} \frac{\partial q_{i,j-1}(x_i, x_{j-1})}{\partial q_1(x_1)} r_j(x_j|x_{j-1}) & : \; j > i+1 \geq 2. \end{cases} \\
\frac{\partial q_{ij}(x_i, x_j)}{\partial r_k(x_k|x_{k-1})} &= \begin{cases} q_i(x_i) & : \; j = i+1 = k \\ \frac{\partial q_i(x_i)}{\partial r_k(x_k|x_{k-1})} r_{i+1}(x_{i+1}|x_i) & : \; j = i+1 > k \\ \sum_{x_{j-1}} \frac{\partial q_{i,j-1}(x_i, x_{j-1})}{\partial r_k(x_k|x_{k-1})} r_j(x_j|x_{j-1}) & : \; j > i+1 \geq k. \end{cases}
\end{aligned} \tag{7.11}$$

Define the Lagrangian

$$\mathcal{L}(q, \lambda) \triangleq D(q\|p) + \lambda_1 \left(\sum_{x_1} q_1(x_1) - 1 \right) + \sum_{i=2}^{n} \sum_{x_{i-1}} \lambda_{i,x_{i-1}} \left(\sum_{x_i} r_i(x_i|x_{i-1}) - 1 \right)$$

The necessary first-order conditions for optimality of q are

$$\begin{aligned}
0 &= \frac{\partial \mathcal{L}(q, \lambda)}{\partial q_1(x_1)} = \frac{\partial D(q\|p)}{\partial q_1(x_1)} + \lambda_1 \quad \forall x_1 \\
0 &= \frac{\partial \mathcal{L}(q, \lambda)}{\partial r_i(x_i|x_{i-1})} = \frac{\partial D(q\|p)}{\partial r_i(x_i|x_{i-1})} + \lambda_{i,x_{i-1}} \quad \forall i \geq 2, x_i, x_{i-1}.
\end{aligned}$$

To compute the partial derivatives of $D(q\|p)$, we express the partial derivatives of its

individual terms in terms of (7.11):

$$\begin{aligned}
\frac{\partial H(q_1)}{\partial q_1(x_1)} &= -1 - \ln q_1(x_1) \\
\frac{\partial H(r_i|q_{i-1})}{\partial q_1(x_1)} &= \begin{cases} -\sum_{x_2} r_2(x_2|x_1) \ln r_2(x_2|x_1) & : \ i = 2 \\ -\sum_{x_i, x_{i-1}} \frac{\partial q_{i-1}(x_{i-1})}{\partial q_1(x_1)} r_i(x_i|x_{i-1}) \ln r_i(x_i|x_{i-1}) & : \ i \geq 3 \end{cases} \\
\frac{\partial H(r_i|q_{i-1})}{\partial r_j(x_j|x_{j-1})} &= \begin{cases} -q_{i-1}(x_{i-1})[1 + \ln r_i(x_i|x_{i-1})] & : \ i = j \geq 2 \\ -\sum_{x_i} \frac{\partial q_{i-1}(x_{i-1})}{\partial r_j(x_j|x_{j-1})} r_i(x_i|x_{i-1}) \ln r_i(x_i|x_{i-1}) & : \ i = j+1 \geq 3 \\ -\sum_{x_i, x_{i-1}} \frac{\partial q_{i-1}(x_{i-1})}{\partial r_j(x_j|x_{j-1})} r_i(x_i|x_{i-1}) \ln r_i(x_i|x_{i-1}) & : \ i > j+1 \geq 3. \end{cases} \\
\frac{\partial \mathbb{E}_q[\ln \psi_{ij}(X_i, X_j)]}{\partial q_1(x_1)} &= \sum_{x_i, x_j} \frac{\partial q_{ij}(x_i, x_j)}{\partial q_1(x_1)} \ln \psi_{ij}(x_i, x_j) : \quad j \geq i+1 \geq 2 \\
\frac{\partial \mathbb{E}_q[\ln \psi_{ij}(X_i, X_j)]}{\partial r_k(x_k|x_{k-1})} &= \sum_{x_i, x_j} \frac{\partial q_{ij}(x_i, x_j)}{\partial r_k(x_k|x_{k-1})} \ln \psi_{ij}(x_i, x_j) : \quad j \geq i+1 \geq k \geq 1.
\end{aligned}$$

We obtain

$$\begin{aligned}
\frac{\partial D(q\|p)}{\partial q_1(x_1)} &= -1 - \ln q_1(x_1) + f_1(x_1; \{r_i\}) \\
\frac{\partial D(q\|p)}{\partial r_i(x_i|x_{i-1})} &= -q_{i-1}(x_{i-1})[1 + \ln r_i(x_i|x_{i-1})] + f_i(x_i, x_{i-1}; \{r_j\}_{j \neq i}, q_1).
\end{aligned}$$

where the function f_1 does not depend on q_1, and the function f_i does not depend on r_i, as can be verified by inspection of the partial derivatives above. Hence the first-order optimality conditions take the form

$$q_1(x_1) = \frac{1}{Z_1} \exp\{f_1(x_1; \{r_i\})\} \tag{7.12}$$

$$r_i(x_i|x_{i-1}) = \frac{1}{Z_{i,x_{i-1}}} \exp\left\{ \frac{f_i(x_i, x_{i-1}; \{r_j\}_{j \neq i}, q_1)}{q_{i-1}(x_{i-1})} \right\}, \quad 2 \leq i \leq n \tag{7.13}$$

where the right side of (7.12) does not involve q_1, and the right side of (7.13) does not involve r_i. Hence the fixed point can be sought using an iterative procedure in which q_1 and $\{r_i\}_{i=2}^n$ are initialized, then recursively updated using (7.12) and (7.13).

7.2 Variational Bayesian Inference

Variational inference has been successfully used to approximate complicated posterior distributions that arise in Bayesian inference problems. The general framework is introduced in this section and applied to estimation of mixtures of Gaussians in the next section. Denote by Y the observations and by X the variables to be estimated. in the Bayesian setup, these variables follow a known prior distribution $\pi(x)$ and the observations follow the conditional distribution $p(y|x)$. The posterior distribution $\pi(x|y)$ plays a crucial role in Bayesian inference but is sometimes untractable. Hence the idea of constructing a varation approximation $q(x)$, solving the problem

$$\min_{q \in \mathcal{Q}} D(q\|\pi(.|y)) = \mathbb{E}_q \left[\ln \frac{q(X)}{\pi(X|Y=y)} \right] \tag{7.14}$$

where $\mathcal{Q}$ is a suitable class of variational approximations.

ELBO. Write the identity

$$\ln p(y) = \ln \frac{p(x,y)}{\pi(x|y)}$$

and taking the expectation over X following any distribution q, we have

$$\ln p(y) = \underbrace{\mathbb{E}_q[\ln p(X,y)] - \mathbb{E}_q[\ln q(X)]}_{=\mathrm{ELBO}(q)} + D(q\|\pi(.|y)).$$

Since the left side of this equation is independent of q, the variational problem (7.14) is equivalent to minimization of the function

$$\mathrm{ELBO}(q) = \mathbb{E}_q[\ln p(X,y)] - \mathbb{E}_q[\ln q(X)]. \tag{7.15}$$

Variational Bayesian inference has been particularly useful in problems where the data are a sequence $\{Y_i\}_{i=1}^n$. Associated with each data point Y_i is a *local latent variable* J_i. Let $(\boldsymbol{Y}, \boldsymbol{J}) = \{(Y_i, J_i)\}_{i=1}^n$ where the pairs (Y_i, J_i) are assumed to be drawn iid from a distribution $p(y, j|\theta)$, and θ is a *global parameter* with prior distribution $\pi(\theta)$. Both the local latent variables and the global parameter are to be inferred. Thus $X = (\boldsymbol{J}, \theta)$ are the variables to be inferred, and a variational approximation

$$q(X) = q_1(\boldsymbol{J})q_2(\theta)$$

wil be constructed to approximate the posterior distribution $\pi(X|\boldsymbol{Y} = \boldsymbol{y})$.

For this problem, the ELBO (7.15) takes the form

$$\mathrm{ELBO}(q_1, q_2) = \mathbb{E}_q[\ln p(\theta, \boldsymbol{J}, \boldsymbol{Y})] - \mathbb{E}_{q_1}[\ln q_1(\boldsymbol{J})] - \mathbb{E}_{q_2}[\ln q_2(\theta)]. \tag{7.16}$$

The variational approximations will be obtained as

$$\begin{aligned} q_1(\boldsymbol{j}) &\propto \exp\{\mathbb{E}_{q_2} \ln p(\boldsymbol{y}, \boldsymbol{j}|\theta)\}, \\ q_2(\theta) &\propto \pi(\theta) \exp\{\mathbb{E}_{q_1} \ln p(\boldsymbol{y}, \boldsymbol{J}|\theta)\}. \end{aligned}$$

Regarding the selection of $\pi(\theta)$, the conjugate prior distribution to the distribution p_θ will be used as this considerably simplifies the derivations and results in distributions with parameters that admit closed-form expressions. The resulting algorithm is called *Coordinate Ascent Variational Inference* (CAVI).

7.3 Variational Mixture of Gaussians

This section studies variational Bayesian mixtures of Gaussians. We start with univariate distributions in Secs 7.3.1 and 7.3.2. The data $\boldsymbol{Y} = \{Y_i\}_{i=1}^n$ are drawn iid from a pdf p_θ which is the mixture of m univariate Gaussians with respective probabilities ω_j, means μ_j, and precisions $\tau_j = 1/\sigma_j^2$, for $1 \le j \le m$:

$$p_\theta(y) = \sum_{j=1}^m \omega_j \, \phi(y; \mu_j, 1/\tau_j), \quad y \in \mathbb{R} \tag{7.17}$$

where

$$\phi(y;\mu,\sigma^2) \triangleq \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$$

denotes the Gaussian pdf with mean μ and variance σ^2. The problem is extended to mixtures of multivariate Gaussians in Sec. 7.3.3.

7.3.1 Unknown Means

To simplify the exposition, we initially assume that the mixture probabilities $\{\omega_j\}$ and the varances $\{\sigma_j^2\}$ are given but do not know the means $\{\mu_j\}$.

To generate Y with the prescribed mixture distribution, we may draw a random label that takes value j with probability ω_j, and then generate $Y \sim \mathcal{N}(\mu_j, \sigma_j^2)$. An EM algorithm was derived to find the ML estimator of $\theta = \{\mu_j\}_{j=1}^m$ in Sec. 4.3, viewing the labels $\{j_i\}_{i=1}^n$ as hidden variables distributed iid with probability mass function π.

Here we view θ as a random variable with a known prior distribution π (a convenient prior will be identified soon). Write $\boldsymbol{j} = \{j_i\}_{i=1}^n$, then we seek to approximate the posterior distribution $\pi(\boldsymbol{j}, \theta|\boldsymbol{y})$ with the factorized distribution

$$q(\boldsymbol{j}, \theta) = q_1(\boldsymbol{j}) q_2(\theta).$$

Note that we do not assume the the labels $\{J_i\}_{i=1}^n$ are independent under q_1, or that the means $\{\mu_j\}_{j=1}^m$ are independent under q_2. The conditional distribution of $(\boldsymbol{y}, \boldsymbol{j})$ given θ is

$$p(\boldsymbol{y}, \boldsymbol{j}|\theta) = \prod_{i=1}^n \omega_{j_i} \sqrt{\tau_{j_i}/2\pi} \exp\left\{-\frac{\tau_{j_i}}{2}(y_i - \mu_{j_i})^2\right\}.$$

It will be convenient to encode each label j_i as a one-hot vector: $e_{ij} = \mathbb{1}\{j_i = j\}$. Note that $\sum_{j=1}^m e_{ij} = 1$. Then we can conveniently write the conditional distribution in product form as

$$p(\boldsymbol{y}, \boldsymbol{j}|\theta) = \prod_{i=1}^n \prod_{j=1}^m \left[\omega_j \sqrt{\tau_j/2\pi} \exp\left\{-\frac{\tau_j}{2}(y_i - \mu_j)^2\right\}\right]^{e_{ij}}. \tag{7.18}$$

The joint distribution $p(\boldsymbol{y}, \boldsymbol{j}, \theta)$ is the product of the conditional distribution above with the prior $\pi(\theta)$. To facilitate evaluation of the expressions, we choose a Gaussian prior

$$\pi(\theta) = \prod_{j=1}^m \phi(\mu_j; \tilde{\mu}_j, 1/\tilde{\tau}_j) = \prod_{j=1}^m \sqrt{\tilde{\tau}_j/2\pi} \exp\left\{-\frac{\tilde{\tau}_j}{2}(\mu_j - \tilde{\mu}_j)^2\right\} \tag{7.19}$$

as $\mathcal{N}(\tilde{\mu}_j, 1/\tilde{\tau}_j)$ is a conjugate prior for the mean of a Gaussian variable.

Now we derive the desired variational factorization. First

$$\begin{aligned} q_1(\boldsymbol{j}) &\propto \exp\{\mathbb{E}_{q_2} \ln p(\boldsymbol{y}, \boldsymbol{j}|\theta)\} \\ &\propto \exp\left\{\mathbb{E}_{q_2}\left(\sum_{j=1}^{m}\sum_{i=1}^{n} e_{ij}\left[\ln(\omega_j\sqrt{\tau_j}) - \frac{\tau_j}{2}(y_i-\mu_j)^2\right]\right)\right\} \\ &= \exp\left\{\sum_{j=1}^{m}\sum_{i=1}^{n} e_{ij}\underbrace{\left[\ln(\omega_j\sqrt{\tau_j}) - \frac{\tau_j}{2}\mathbb{E}_{q_2}(y_i-\mu_j)^2\right]}_{=\ln\rho_{ij}}\right\} \\ &= \prod_{j=1}^{m}\prod_{i=1}^{n}\rho_{ij}^{e_{ij}} \end{aligned}$$

where we have introduced

$$\rho_{ij} = \omega_j\sqrt{\tau_j}\exp\left\{-\frac{\tau_j}{2}\mathbb{E}_{q_2}(y_i-\mu_j)^2\right\}. \tag{7.20}$$

Observe that ρ_{ij} are positive quantities, and rescale them as follows:

$$r_{ij} \triangleq \frac{\rho_{ij}}{\sum_{j=1}^{m}\rho_{ij}} \tag{7.21}$$

so that $\sum_{j=1}^{m} r_{ij} = 1$ for each i. Using the identity $\sum_{j=1}^{m} e_{ij} = 1$ for each i, we obtain the exact expression

$$q_1(\boldsymbol{j}) = \prod_{j=1}^{m}\prod_{i=1}^{n} r_{ij}^{e_{ij}}. \tag{7.22}$$

This is a product (over i) of categorical distributions with respective probabilities $\{r_{ij}\}_{j=1}^{m}$, hence $\{J_i\}_{i=1}^{n}$ are independent under q_1 with respective probability distributions

$$r_{ij} = \mathbb{E}_{q_1}(E_{ij}), \quad 1 \le j \le m.$$

These quantities are called responsibilities and are akin to posterior probabilities for the labels given observation y_i. Observe the striking similarity of r_{ij} with the posterior probabilities $\pi_\theta(j|y_i)$ derived in the EM chapter (no prior distribution was assumed on the means μ_j there.) Finally, observe from (7.20) and (7.21) that only the expectation of $(y_i-\mu_j)^2$ and not the full distribution q_2 was needed to derive (7.22).

Comparing $q_1(\boldsymbol{j})$ with the prior probability for $\boldsymbol{j}$,

$$p(\boldsymbol{j}) = \prod_{j=1}^{m}\prod_{i=1}^{n}\omega_j^{e_{ij}}$$

we see that the prior probabilities ω_j on the labels have been replaced with the responsibilities r_{ij} to obtain the variational probability distribution q_1.

Next we derive the variational distribution for θ. We have

$$\begin{aligned} q_2(\theta) &= \exp\{\mathbb{E}_{q_1} \ln p(\boldsymbol{y}, \boldsymbol{J}, \theta)\} \\ &= \pi(\theta) \exp\{\mathbb{E}_{q_1} \ln p(\boldsymbol{y}, \boldsymbol{J}|\theta)\} \\ &\propto \pi(\theta) \exp\left\{ \mathbb{E}_{q_1} \left(\sum_{j=1}^{m} \sum_{i=1}^{n} E_{ij} \left[\ln(\omega_j \sqrt{\tau_j}) - \frac{\tau_j}{2}(y_i - \mu_j)^2 \right] \right) \right\} \\ &= \pi(\theta) \exp\left\{ \sum_{j=1}^{m} \sum_{i=1}^{n} \underbrace{\mathbb{E}_{q_1}(E_{ij})}_{=r_{ij}} \left[\ln(\omega_j \sqrt{\tau_j}) - \frac{\tau_j}{2}(y_i - \mu_j)^2 \right] \right\} \qquad (7.23) \\ &\propto \pi(\theta) \exp\left\{ \sum_{j=1}^{m} -\frac{\tau_j}{2} \sum_{i=1}^{n} r_{ij}(y_i - \mu_j)^2 \right\} \\ &= \prod_{j=1}^{m} \exp\left\{ -\frac{\tilde{\tau}_j}{2}(\mu_j - \tilde{\mu}_j)^2 - \frac{\tau_j}{2} \sum_{i=1}^{n} r_{ij}(y_i - \mu_j)^2 \right\} \end{aligned}$$

where in the last line we have used the expression (7.19) for the conjugate prior. Now define

$$\overline{\mu}_j = \frac{\tilde{\tau}_j \tilde{\mu}_j + \tau_j \sum_{i=1}^{n} r_{ij} y_i}{\tilde{\tau}_j + \tau_j \sum_{i=1}^{n} r_{ij}} \qquad (7.24)$$

$$\overline{\tau}_j = \tilde{\tau}_j + \tau_j \sum_{i=1}^{n} r_{ij}, \quad 1 \le j \le m. \qquad (7.25)$$

Completing the square in the exponent and normalizing, we obtain the Gaussian variational density for the means:

$$q_2(\theta) = \prod_{j=1}^{m} \phi(\mu_j; \overline{\mu}_j, 1/\overline{\tau}_j)$$

and conclude that

$$\mathbb{E}_{q_2}[(y_i - \mu_j)^2] = (y_i - \overline{\mu}_j)^2 + 1/\overline{\tau}_j. \qquad (7.26)$$

Moreover note from (7.25) that the precision $\overline{\tau}_j = 1/\mathrm{Var}_{q_2}(\mu_j)$ is typically large for labels j that are strongly associated with many data points (so $\sum_{i=1}^{n} r_{ij}$ is large).

The responsibilities r_{ij} are expressed in terms of $\overline{\mu}_j$ and $\overline{\tau}_j$ via (7.20), (7.21), and (7.26), while $\overline{\mu}_j$ and $\overline{\tau}_j$ are expressed in terms of r_{ij} via (7.24). The algorithm iterates between these expressions to obtain the final parameters of the variational approximation.

At each iteration we can evaluate the ELBO

$$\begin{aligned}
\text{ELBO}(q_1, q_2) &= \mathbb{E}_q[\ln p(\theta, \boldsymbol{J}, \boldsymbol{y})] - \mathbb{E}_{q_1}[\ln q_1(\boldsymbol{J})] - \mathbb{E}_{q_2}[\ln q_2(\theta)] \\
&= \mathbb{E}_{q_2}[\ln \pi(\theta)] + \sum_{i=1}^{n} \mathbb{E}_q[\ln p(y_i, J_i|\theta)] - \sum_{i=1}^{n} \mathbb{E}_{q_1}[\ln q_1(J_i)] - \mathbb{E}_{q_2}[\ln q_2(\theta)] \\
&= -D(q_2\|\pi) - \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{E}_{q_1}(E_{ij}) \ln r_{ij} \\
&\quad + \sum_{i=1}^{n} \mathbb{E}_{q_1}(E_{ij}) \left[\ln(\omega_j \sqrt{\tau_j/2\pi}) - \frac{\tau_j}{2}\mathbb{E}_{q_2}(y_i - \mu_j)^2\right] \\
&= -\sum_{j=1}^{m} D(\mathcal{N}(\overline{\mu}_j, 1/\overline{\tau}_j)\|\mathcal{N}(\tilde{\mu}_j, 1/\tau_j)) - \sum_{i=1}^{n}\sum_{j=1}^{m} r_{ij} \ln r_{ij} \\
&\quad + \sum_{i=1}^{n} r_{ij} \left[\ln(\omega_j \sqrt{\tau_j/2\pi}) - \frac{\tau_j}{2}((y_i - \overline{\mu}_j)^2 + 1/\overline{\tau}_j)\right] \\
&= -\frac{1}{2}\sum_{j=1}^{m} \left[\ln \frac{\overline{\tau}_j}{\tau_j} - 1 + \frac{\tau_j}{\overline{\tau}_j} + \tau_j(\overline{\mu}_j - \tilde{\mu}_j)^2\right] - \sum_{i=1}^{n}\sum_{j=1}^{m} r_{ij} \ln r_{ij} \\
&\quad + \sum_{i=1}^{n} r_{ij} \left[\ln(\omega_j \sqrt{\tau_j/2\pi}) - \frac{\tau_j}{2}((y_i - \overline{\mu}_j)^2 + 1/\overline{\tau}_j)\right]. \qquad (7.27)
\end{aligned}$$

In summary, the variational algorithm proceeds as follows:

1. Select the prior means $\tilde{\mu}_j$ and precisions $\tilde{\tau}_j$ and initialize the responsibilities r_{ij} (e.g., uniform).
2. Evaluate the variational means $\overline{\mu}_j$ and precisions $\overline{\tau}_j$ using (7.24) and (7.25).
3. Update the responsibilities r_{ij} using (7.20), (7.21), and (7.26).
4. Evaluate ELBO in (7.27). If the improvement over the previous iteration exceeds a threshold, return to Step 2. Else END.

7.3.2 Unknown Means, Variances, and Mixture Probabilities

Now we assume that the Gaussian means, precisions, and mixture probabilities in (7.17) are unknown, and view them as a vector θ of parameters with prior π to be specified. Again this will be a conjugate prior. The distribution of the mixture probability vector $\boldsymbol{\omega}$ is Dirichlet with parameter-vector $\boldsymbol{\alpha}_0 = (\alpha_{0,1}, \cdots, \alpha_{0,m})$:

$$\pi(\boldsymbol{\omega}) = \frac{\Gamma(\sum_{j=1}^{m} \alpha_{0,j})}{\prod_{j=1}^{m} \Gamma(\alpha_{0,j})} \prod_{j=1}^{m} \omega_j^{\alpha_{0,j}-1} = \text{Dir}(\boldsymbol{\omega}; \boldsymbol{\alpha}_0). \qquad (7.28)$$

where $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, dt$ is the Gamma function. Choosing $\alpha_{0,j} \equiv 1$ results in a uniform prior on $\boldsymbol{\omega}$. Choosing $\alpha_{0,j} < 1$ encourages $\boldsymbol{\omega}$ to be a sparse vector (the most likely probability vectors $\boldsymbol{\omega}$ are the ones with a single component, i.e., degenerate distributions).

The precision parameters τ_j are mutually independent and follow a $\Gamma(a,b)$ distribution $(a,b>0)$:

$$\pi(\tau_j) = \frac{b^a}{\Gamma(a)}\tau_j^{a-1}\, e^{-b\tau_j}, \quad j = 1,2,\cdots,k.$$

The parameters μ_j are Gaussian with respective means $\tilde{\mu}_j$ and precisions $c\tau_j$ (with parameter $c>0$):

$$\pi(\mu_j|\tau_j) = \phi(\mu_j;\tilde{\mu}_j, 1/c\tau_j), \quad j = 1,2,\cdots,k.$$

Hence (μ_j,τ_j) follow a so-called Gaussian-Gamma distribution with parameters $(\tilde{\mu}_j, c, a, b)$ and pdf given by

$$\pi(\mu_j,\tau_j) = \frac{b^a\sqrt{c}}{\Gamma(a)\sqrt{2\pi}}\,\tau_j^{a-1/2}\exp\left\{-\frac{1}{2}c\tau_j(\mu_j-\tilde{\mu}_j)^2 - b\tau_j)\right\}. \tag{7.29}$$

The m pairs $(\mu_j,\tau_j)_{j=1}^m$ and the random probability vector $\boldsymbol{\omega}$ are mutually independent. Hence

$$\pi(\theta) = \prod_{j=1}^m \omega_j^{\alpha_0-1}\tau_j^{a-1/2}\exp\left\{-\frac{1}{2}c\tau_j(\mu_j-\tilde{\mu}_j)^2 - b\tau_j)\right\}. \tag{7.30}$$

The derivation of the variational approximation $q_1(\boldsymbol{j})$ is similar to that in the previous section. We have

$$\begin{aligned} q_1(\boldsymbol{j}) &= \exp\{\mathbb{E}_{q_2}\ln p(\boldsymbol{y},\boldsymbol{j}|\theta)\} \\ &\propto \exp\left\{\mathbb{E}_{q_2}\left(\sum_{j=1}^m\sum_{i=1}^n e_{ij}\left[\ln\omega_j + \ln\sqrt{\tau_j} - \frac{\tau_j}{2}(y_i-\mu_j)^2\right]\right)\right\} \\ &= \exp\left\{\sum_{j=1}^m\sum_{i=1}^n e_{ij}\left(\mathbb{E}_{q_2}[\ln\omega_j] + \frac{1}{2}\mathbb{E}_{q_2}[\ln\tau_j] - \frac{1}{2}\mathbb{E}_{q_2}[\tau_j(y_i-\mu_j)^2]\right)\right\}. \end{aligned} \tag{7.31}$$

Observe that only three expectations and not the full distribution q_2 were needed to derive (7.31). The variational distribution q_1 is still given by (7.22) and (7.21), but here ρ_{ij} takes the form

$$\rho_{ij} = \exp\left\{\mathbb{E}_{q_2}[\ln\omega_j] + \frac{1}{2}\mathbb{E}_{q_2}[\ln\tau_j] - \frac{1}{2}\mathbb{E}_{q_2}[\tau_j(y_i-\mu_j)^2]\right\}. \tag{7.32}$$

The derivation of the variational approximation $q_2(\theta)$ is obtained as follows from (7.23) and (7.30):

$$\begin{aligned} q_2(\theta) &= \exp\{\mathbb{E}_{q_1}\ln p(\boldsymbol{y},\boldsymbol{J},\theta)\} \\ &\propto \pi(\theta)\exp\left\{\sum_{j=1}^m\sum_{i=1}^n r_{ij}\left[\ln\omega_j + \ln\sqrt{\tau_j} - \frac{\tau_j}{2}(y_i-\mu_j)^2\right]\right\} \\ &= \prod_{j=1}^m \omega_j^{\sum_i r_{ij}+\alpha_0-1}\tau_j^{a+\sum_i r_{ij}/2-1/2}\exp\left\{-\frac{\tau_j}{2}\sum_i r_{ij}(y_i-\mu_j)^2 - \frac{1}{2}c\tau_j(\mu_j-\tilde{\mu}_j)^2 - b\tau_j)\right\} \\ &= \prod_{j=1}^m \omega_j^{\alpha_j-1}\tau_j^{a_j-1/2}\exp\left\{-\frac{1}{2}c_j\tau_j(\mu_j-\overline{\mu}_j)^2 - b_j\tau_j)\right\} \end{aligned}$$

where

$$\begin{aligned}
\alpha_j &= \alpha_0 + \sum_i r_{ij} \\
a_j &= a + \frac{1}{2}\sum_i r_{ij} \\
c_j &= c + \sum_{i=1}^{n} r_{ij} \\
b_j &= b + \frac{1}{2}\sum_{i=1}^{n} r_{ij} y_i^2 + \frac{1}{2} c\tilde{\mu}_j^2 \\
\overline{\mu}_j &= \frac{c\tilde{\mu}_j + \sum_{i=1}^{n} r_{ij} y_i}{c + \sum_{i=1}^{n} r_{ij}}.
\end{aligned} \tag{7.33}$$

Hence $q_2(\theta)$ is in the same functional form as the prior $\pi(\theta)$, with the parameters (c, a, b) replaced with (c_j, a_j, b_j) for the Gaussian-Gamma distribution on (μ_j, τ_j), and the parameter-vector $\boldsymbol{\alpha}_0$ replaced with $\boldsymbol{\alpha} = \{\alpha_j\}_{j=1}^m$ for the Dirichlet prior on $\boldsymbol{\omega}$.

Having identified q_2, we are now able to derive the three expectations that appear in (7.31) and are taken with respect to the Dirichlet, Gamma, and Gaussian-Gamma distributions, respectively:

$$\begin{aligned}
\mathbb{E}_{q_2}[\ln \omega_j] &= \psi(\alpha_j) - \psi\left(\sum_j \alpha_j\right) \\
\mathbb{E}_{q_2}[\ln \tau_j] &= \psi(a_j) - \ln b_j \\
\mathbb{E}_{q_2}[\tau_j (y_i - \mu_j)^2] &= \frac{1}{c_j} + \frac{a_j}{b_j}(y_i - \overline{\mu}_j)^2
\end{aligned} \tag{7.34}$$

where $\psi(a) = \frac{d \ln \Gamma(a)}{da}$ is the digamma function.

The responsibilities r_{ij} are expressed in terms of these three expectations via (7.32) and (7.21), and the three expectations are expressed in terms of r_{ij} via (7.34) and (7.33). The algorithm iterates between these expressions to obtain the final parameters of the variational approximation.

7.3.3 Mixture of Multivariate Gaussians

We now extend the analysis of Sec. 7.3.2 to the multivariate case. Here

$$p_\theta(y) = \sum_{j=1}^{m} \omega_j \, \phi(y; \mu_j, \Lambda_j), \quad y \in \mathbb{R}^d \tag{7.35}$$

where μ_j are mean vectors, Λ_j are $d \times d$ precision matrices, and

$$\phi(y; \mu, \Lambda) = (2\pi)^{-d/2} |\Lambda|^{1/2} \exp\left\{-\frac{1}{2}(y-\mu)^\top \Lambda (y-\mu)\right\}.$$

The parameter of the distribution (7.35) is $\theta = \left(\boldsymbol{\omega}, \{(\mu_j, \Lambda_j)\}_{j=1}^m\right)$.

Given observations $\{y_i\}_{i=1}^n$ iid P_θ, we again denote by j_i the label associated with datapoint y_i and write $\boldsymbol{j} = \{j_i\}_{i=1}^n$. Similarly to (7.18), the conditional distribution of $(\boldsymbol{y}, \boldsymbol{j})$ given θ is

$$p(\boldsymbol{y}, \boldsymbol{j}|\theta) = \prod_{i=1}^n \prod_{j=1}^m \left[\omega_j (2\pi)^{-d/2} |\Lambda_j|^{1/2} \exp\left\{-\frac{1}{2}(y_i - \mu_j)^\top \Lambda_j (y_i - \mu_j)\right\}\right]^{e_{ij}}. \tag{7.36}$$

We assume the same Dirichlet prior (7.28) as in the scalar case. The Gamma prior on the precisions in the scalar case ($d = 1$) is extended to the multivariate case using the *Wishart distribution*

$$W(\Lambda; \mathsf{B}, \nu) = c(\mathsf{B}, \nu) |\Lambda|^{(\nu-d-1)/2} \exp\left\{-\frac{1}{2}\mathrm{Tr}(\mathsf{B}^{-1}\Lambda)\right\} \tag{7.37}$$

where Λ is a precision matrix, B is the (symmetric and positive-definite) scale matrix, $\nu > d - 1$ is the number of degrees of freedom, and the normalization constant

$$c(\mathsf{B}, \nu) = |\mathsf{B}|^{-\nu/2} \Bigg/ \left(2^{\nu d/2} \pi^{d(d-1)/4} \prod_{i=1}^d \Gamma\left(\frac{\nu + 1 - i}{2}\right)\right).$$

The Wishart distribution is the conjugate prior for the precision matrix of a Gaussian vector. The Gaussian-Wishart distribution $\mathrm{GW}(\tilde{\mu}, c, \mathsf{B}, \nu)$ (with parameters $(\tilde{\mu}, c, \mathsf{B}, \nu)$) on (μ, Λ) is the product $\phi(\mu; \tilde{\mu}, (c\Lambda)^{-1}) W(\Lambda; \mathsf{B}, \nu)$ and is the conjugate prior for the mean μ and precision matrix Λ of a Gaussian vector. In the scalar case, this distribution reduces to the Gaussian-Gamma prior of (7.29). We assume that the pairs (μ_j, Λ_j) follow a product Gaussian-Wishart distribution $\mathrm{GW}(\tilde{\mu}_j; \Lambda, c, \mathsf{B}, \nu)$ and are independent of $\boldsymbol{\omega}$. Hence the prior on θ takes the product form

$$\begin{aligned}
\pi(\theta) &= \mathrm{Dir}(\boldsymbol{\omega}; \boldsymbol{\alpha}_0) \prod_{j=1}^m \phi(\mu_j; \tilde{\mu}_j, (c\Lambda_j)^{-1}) W(\Lambda_j; \mathsf{B}, \nu) \\
&\propto \prod_{j=1}^m \omega_j^{\alpha_0 - 1} |\Lambda_j|^{(\nu-d)/2} \exp\left\{-\frac{1}{2}(\mu_j - \tilde{\mu}_j)^\top c\Lambda_j (\mu_j - \tilde{\mu}_j) - \frac{1}{2}\mathrm{Tr}(\mathsf{B}^{-1}\Lambda_j)\right\}
\end{aligned} \tag{7.38}$$

The derivation of the variational approximation $q_1(\boldsymbol{j})$ is similar to that in the scalar case. We obtain

$$\begin{aligned}
q_1(\boldsymbol{j}) &= \exp\{\mathbb{E}_{q_2} \ln p(\boldsymbol{y}, \boldsymbol{j}|\theta)\} \\
&\propto \exp\left\{\mathbb{E}_{q_2}\left(\sum_{j=1}^m \sum_{i=1}^n e_{ij}\left[\ln \omega_j + \frac{1}{2}\ln|\Lambda_j| - \frac{1}{2}(y_i - \mu_j)^\top \Lambda_j (y_i - \mu_j)\right]\right)\right\} \\
&= \exp\left\{\sum_{j=1}^m \sum_{i=1}^n e_{ij}\left(\mathbb{E}_{q_2}[\ln \omega_j] + \frac{1}{2}\mathbb{E}_{q_2}[\ln|\Lambda_j|] - \frac{1}{2}\mathbb{E}_{q_2}[(y_i - \mu_j)^\top \Lambda_j (y_i - \mu_j)]\right)\right\}.
\end{aligned} \tag{7.39}$$

The variational distribution q_1 is still given by (7.22) and (7.21), but here ρ_{ij} takes the form

$$\rho_{ij} = \exp\left\{\mathbb{E}_{q_2}[\ln \omega_j] + \frac{1}{2}\mathbb{E}_{q_2}[\ln|\Lambda_j|] - \frac{1}{2}\mathbb{E}_{q_2}[(y_i - \mu_j)^\top \Lambda_j (y_i - \mu_j)]\right\}. \tag{7.40}$$

Next we derive the variational distribution for θ. Similarly to (7.23), we have

$$\begin{aligned} q_2(\theta) &= \exp\{\mathbb{E}_{q_1} \ln p(\boldsymbol{y}, \boldsymbol{J}, \theta)\} \\ &= \pi(\theta) \exp\left\{ \sum_{j=1}^{m} \sum_{i=1}^{n} r_{ij} \left[\ln \omega_j + \frac{1}{2} \ln |\Lambda_j| - \frac{1}{2}(y_i - \mu_j)^\top \Lambda_j (y_i - \mu_j) \right] \right\} \\ &\propto \prod_{j=1}^{m} \omega_j^{\sum_{i=1}^n r_{ij} + \alpha_0 - 1} |\Lambda_j|^{(\sum_{i=1}^n r_{ij} + \nu - d - 1)/2} \exp\left\{ -\frac{1}{2}(\mu_j - \tilde{\mu}_j)^\top c \Lambda_j (\mu_j - \tilde{\mu}_j) \right. \\ &\quad \left. -\frac{1}{2} \sum_{i=1}^{n} r_{ij} (y_i - \mu_j)^\top \Lambda_j (y_i - \mu_j) - \frac{1}{2} \mathrm{Tr}(\mathsf{B}^{-1} \Lambda_j) \right\}. \end{aligned} \tag{7.41}$$

Now define the following parameters:

$$\begin{aligned} \alpha_j &= \alpha_0 + \sum_{i=1}^{n} r_{ij} \\ c_j &= c + \sum_{i=1}^{n} r_{ij} \\ \nu_j &= \nu + \sum_{i=1}^{n} r_{ij} \\ \mathsf{B}_j^{-1} &= \mathsf{B}^{-1} + \sum_{i=1}^{n} r_{ij} y_i y_i^\top + c \tilde{\mu}_j \tilde{\mu}_j^\top \\ \overline{\mu}_j &= \frac{c\tilde{\mu}_j + \sum_{i=1}^n r_{ij} y_i}{c + \sum_{i=1}^n r_{ij}}, \quad 1 \le j \le m. \end{aligned} \tag{7.42}$$

The quadratic form in the argument of the exponential of (7.41) may be written as

$$-\frac{1}{2}\mu_j^\top \underbrace{(c + \sum_{i=1}^{n} r_{ij})}_{=c_j} \Lambda_j \mu_j + \mu_j^\top \Lambda_j \underbrace{(c\tilde{\mu}_j + \sum_{i=1}^{n} r_{ij} y_i)}_{=c_j \overline{\mu}_j} - \frac{1}{2} \underbrace{(\tilde{\mu}_j^\top c \Lambda_j \tilde{\mu}_j + \sum_{i=1}^{n} r_{ij} y_i^\top \Lambda_j y_i + \mathrm{Tr}(\mathsf{B}^{-1}\Lambda_j))}_{=\mathrm{Tr}(\mathsf{B}_j^{-1}\Lambda_j)}$$

$$= -\frac{1}{2}(\mu_j - \overline{\mu}_j)^\top c_j \Lambda_j (\mu_j - \overline{\mu}_j) - \frac{1}{2}\mathrm{Tr}(\mathsf{B}_j^{-1}\Lambda_j)$$

Hence

$$\begin{aligned} q_2(\theta) &\propto \prod_{j=1}^{m} \omega_j^{\alpha_j - 1} |\Lambda_j|^{(\nu_j - d)/2} \exp\left\{ -\frac{1}{2}(\mu_j - \overline{\mu}_j)^\top c_j \Lambda_j (\mu_j - \overline{\mu}_j) - \frac{1}{2}\mathrm{Tr}(\mathsf{B}_j^{-1}\Lambda_j) \right\} \\ &= \mathrm{Dir}(\boldsymbol{\omega}; \boldsymbol{\alpha}) \prod_{j=1}^{m} \phi(\mu_j; \overline{\mu}_j, (c_j \Lambda_j)^{-1}) W(\Lambda_j; \mathsf{B}_j, \nu_j) \end{aligned}$$

has the same functional form as the prior $\pi(\theta)$, with parameters now given by (7.42). Having identified q_2, we are able to evaluate the three expectations that feature in (7.40)

using known formulas for the Dirichlet and Gaussian-Wishart distributions:

$$\begin{aligned}
\mathbb{E}_{q_2}[\ln \omega_j] &= \psi(\alpha_j) - \psi\left(\sum_{j=1}^{m} \alpha_j\right) \\
\mathbb{E}_{q_2}[\ln |\Lambda_j|] &= \sum_{i=1}^{d} \psi\left(\frac{\nu_j + 1 - i}{2}\right) + d \ln 2 + \ln |\mathsf{B}_j| \\
\mathbb{E}_{q_2}[(y_i - \mu_j)^\top \Lambda_j (y_i - \mu_j)] &= d/c_j + \nu_j (y_i - \overline{\mu}_j)^\top \mathsf{B}_j (y_i - \overline{\mu}_j). \qquad (7.43)
\end{aligned}$$

The ELBO is given by

$$\begin{aligned}
&\mathrm{ELBO}(q_1, q_2) \\
&= \mathbb{E}_q[\ln p(\theta, \boldsymbol{J}, \boldsymbol{y})] - \mathbb{E}_{q_1}[\ln q_1(\boldsymbol{J})] - \mathbb{E}_{q_2}[\ln q_2(\theta)] \\
&= \mathbb{E}_{q_2}[\ln \pi(\theta)] + \sum_{i=1}^{n} \mathbb{E}_q[\ln p(y_i, J_i|\theta)] - \sum_{i=1}^{n} \mathbb{E}_{q_1}[\ln q_1(J_i)] - \mathbb{E}_{q_2}[\ln q_2(\theta)] \\
&= -D(q_2 \| \pi) - \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{E}_{q_1}(E_{ij}) \ln r_{ij} \\
&\quad + \sum_{j=1}^{m} \sum_{i=1}^{n} \mathbb{E}_{q_1}(E_{ij}) \mathbb{E}_{q_2}\left[\ln \omega_j + \frac{1}{2} \ln |\Lambda_j| - \frac{1}{2}(y_i - \mu_j)^\top \Lambda_j (y_i - \mu_j)\right] \\
&= -D(\mathrm{Dir}(\cdot; \boldsymbol{\alpha}) \| \mathrm{Dir}(\cdot; \boldsymbol{\alpha}_0)) - \sum_{j=1}^{m} D(\mathrm{GW}(\overline{\mu}_j, c_j, \mathsf{B}_j, \nu_j) \| \mathrm{GW}(\tilde{\mu}_j, c, \mathsf{B}, \nu)) \\
&\quad - \sum_{i=1}^{n} \sum_{j=1}^{m} r_{ij} \ln r_{ij} + \sum_{i=1}^{n} r_{ij} \left[\mathbb{E}_{q_2}(\ln \omega_j) + \frac{1}{2} \mathbb{E}_{q_2}(\ln |\Lambda_j|) - \frac{1}{2} \mathbb{E}_{q_2}\{(y_i - \mu_j)^\top \Lambda_j (y_i - \mu_j)\}\right]. \qquad (7.44)
\end{aligned}$$

The three expectations in the last line are given in (7.43), and closed-formed expressions are available for the Kullback-Leibler divergences as well.

In summary, the variational algorithm proceeds as follows:

1. Select the parameters of the prior and initialize the responsibilities r_{ij} (e.g., uniform).
2. Evaluate the parameters of (7.42), then the three expectations of (7.43).
3. Update the responsibilities r_{ij} using (7.40) and (7.21).
4. Evaluate ELBO. If the improvement over the previous iteration exceeds a threshold, return to Step 2. Else END.

7.4 Exponential Families

Much insight into variational problems is gained by considering a broad and flexible class of probability distributions called exponential families.

Definition. A d-dimensional exponential family in canonical form is defined by

$$p_\theta(x) = \frac{h(x)}{Z(\theta)} \exp\left\{\sum_{k=1}^{d} \theta_k T_k(x)\right\}, \quad x \in \mathcal{X}$$

where $\theta = (\theta_1, \cdots, \theta_d)^\top$ is the parameter of the family, the functions $T_k(\cdot)$ are called sufficient statistics, and the *partition function* $Z(\theta)$ is the normalization constant ensuring that the density p_θ integrates to 1. It is often convenient to absorb $h(x)$ into the base measure ν, in which case

$$Z(\theta) = \int_{\mathcal{X}} \exp\{\theta^\top T(x)\}\, d\nu(x)$$

where $T(x)$ is the vector-valued function with components $T_1(x), \cdots, T_d(x)$. Now defining the *log partition function* (aka *cumulant function*) $A(\theta) = \ln Z(\theta)$, we write

$$p_\theta(x) = \exp\left\{\theta^\top T(x) - A(\theta)\right\}, \quad x \in \mathcal{X}. \tag{7.45}$$

The *natural parameter set*

$$\Theta \triangleq \left\{\theta \ : \ \int_{\mathcal{X}} \exp\{\theta^\top T(x)\}\, d\nu(x) < \infty\right\}$$

is convex. If Θ is an open set, the family is said to be *regular*. If there exists no vector $a \in \mathbb{R}^d$ such that $a^\top T(x)$ is constant a.e. ν, the above representation of the family is said to be *minimal*. Otherwise the representation is said to be *overcomplete*. [3]

Example 1. The 2-D Ising model of (7.6) (with $\beta \in \mathbb{R}$) is a one-dimensional exponential family with parameter $\beta \in \mathbb{R}$ and sufficient statistic $T(x) = \sum_{i \sim j} x_i x_j$.

Example 2. Generalized Ising model with $\mathcal{X} = \{\pm 1\}^{\mathcal{V}}$ and

$$p_\theta(x) = \frac{1}{Z(\theta)} \exp\left\{\sum_{i \in \mathcal{V}} \theta_i x_i + \sum_{i \sim j} \theta_{ij} x_i x_j\right\}.$$

This is an exponential family with $T_i(x) = x_i$ for $i \in \mathcal{V}$ and $T_{ij}(x) = x_i x_j$ for $i \sim j$. Since $Z(\theta)$ is given by a finite sum, the natural parameter set is $\Theta = \mathbb{R}^d$ where $d = |\mathcal{V}| + |\mathcal{E}| = 3n$.

Example 3. k-spin model with $\mathcal{X} = \{\pm 1\}^{\mathcal{V}}$, maximal clique size 3, and $p_\theta(x) = \frac{1}{Z(\theta)} \exp\{\sum_{i \in \mathcal{V}} \theta_i x_i + \sum_{\mathcal{C}=(i,j,k)} \theta_{ijk} x_i x_j x_k\}$ where the second sum in the exponent runs over all maximal cliques.

Example 4. n-variate Gaussian family with zero mean and inverse covariance matrix $\theta \succ 0$. Then $p_\theta(x) = \frac{1}{Z(\theta)} \exp\{-\frac{1}{2} x^\top \theta x\}$ with $Z(\theta) = (2\pi)^{n/2} |\theta|^{-1/2}$. This is a $\frac{1}{2} n(n+1)$ dimensional regular exponential family with sufficient statistics $T_{ij}(x) = -x_i x_j$

[3] Overcompleteness is often undesirable from a statistical perspective (θ is identifiable only up to an affine subset of Θ) but will be useful in our study of graphical models.

for $1 \le i < j \le n$, and $T_{ii}(x) = -\frac{1}{2}x_i^2$. This representation of the family is minimal.

Example 5. n-variate Gaussian family with potential vector h and inverse covariance matrix $\mathsf{J} \succ 0$. Then

$$p_\theta(x) = \underbrace{(2\pi)^{-n/2}|\mathsf{J}|^{1/2}\, e^{-\frac{1}{2}h^\top \mathsf{J} h}}_{=1/Z(\mathsf{J},h)} \exp\left\{-\frac{1}{2}x^\top \mathsf{J} x + x^\top h\right\}.$$

If we let $\theta = (\mathsf{J}, h)$, this is a $\frac{1}{2}n(n+3)$ dimensional exponential family in canonical form with the quadratic statistics of Example 4, augmented with the n linear statistics $T_i(x) = x_i$ for $1 \le i \le n$.

Example 6. Poisson random variable with parameter λ. Then $p_\lambda(x) = \frac{\lambda^x}{x!}e^{-\lambda}$ for $x = 0, 1, 2, \cdots$. This may also be written as a 1-dimensional exponential family in canonical form, with parameter $\theta = \ln\lambda \in \mathbb{R}$, statistic $T(x) = x$, and

$$p_\theta(x) = \frac{\exp\{-e^\theta\}}{x!}\, e^{\theta x}, \quad x = 0, 1, 2, \cdots.$$

KL Divergence. The Kullback-Leibler divergence between two distributions p_θ and $p_{\theta'}$ in the same exponential family (7.45) takes the form

$$\begin{aligned} D(p_\theta \| p_{\theta'}) = \mathbb{E}_\theta\left[\ln\frac{p_\theta(X)}{p_{\theta'}(X)}\right] &= \mathbb{E}_\theta\left[A(\theta') - A(\theta) + (\theta-\theta')^\top T(X)\right] \\ &= A(\theta') - A(\theta) + (\theta-\theta')^\top \mathbb{E}_\theta[T(X)], \quad \theta, \theta' \in \Theta. \end{aligned}$$

Cumulant-generating function. The cgf for p_θ takes the form

$$\begin{aligned} \kappa(u) &= \ln \mathbb{E}_\theta[e^{u^\top T(X)}] \\ &= \ln \int_{\mathcal{X}} e^{(\theta+u)^\top T(x) - A(\theta)} d\nu(x) \\ &= A(\theta+u) - A(\theta), \quad u \in \mathbb{R}^d. \end{aligned}$$

7.4.1 Mean Parameter

It follows from the definition of $A(\theta) = \ln\int_{\mathcal{X}} \exp\{\theta^\top T(x)\}\, d\nu(x)$ and its relation to the cgf that the gradient and the Hessian of $A(\theta)$ are given by

$$\nabla A(\theta) = \nabla\kappa(0) = \mathbb{E}_\theta[T(X)], \tag{7.46}$$

$$\nabla^2 A(\theta) = \nabla^2\kappa(0) = \mathrm{Cov}_\theta[T(X)]. \tag{7.47}$$

The expected value (7.46) of the sufficient statistics vector $T(X)$, is called the *mean parameter* of the distribution and is often denoted by μ. It is often convenient to parameterize the distribution using μ instead of θ. The gradient mapping $\nabla A(\cdot)$ and its inverse will play a fundamental role in the analysis.

Definition. The set of *realizable mean parameters* $\mathcal{N}$ is the set of μ that are the expected value of $T(X)$ under *some* distribution p (not necessarily in the exponential family). Thus

$$\mathcal{N} \triangleq \{\mu \in \mathbb{R}^d \ : \ \exists p \ : \ \mathbb{E}_p[T(X)] = \mu\} \tag{7.48}$$

which is a convex set. (Proof as an exercise).

Example 7. If $T(x) = xx^\top \in \mathbb{R}^{n\times n}$ then μ is a correlation matrix, and so $\mathcal{N}$ is the set of all $n \times n$ symmetric nonnegative definite matrices.

Example 8. For discrete random variables, $\mathcal{N}$ is the convex hull of the points $\{T(x)\}$ and is therefore a convex polytope in $\mathbb{R}^d$. Consider for instance a modification of the generalized Ising model of Example 2, with $\mathcal{X} = \{0, 1\}$ instead of $\{\pm 1\}$. Then we have $\mu_i = \mathbb{E}_p[X_i] = P\{X_i = 1\}$ for all $i \in \mathcal{V}$ and $\mu_{ij} = \mathbb{E}_p[X_i X_j] = P\{(X_i, X_j) = (1, 1)\}$ for all $i \sim j$.

Theorem [2]. The gradient mapping $\nabla A : \Theta \to \text{int}(\mathcal{N})$ is one-to-one if and only if the exponential representation is minimal.

An immediate consequence of this important theorem is that for each $\mu \in \text{int}(\mathcal{N})$, there exists an exponential distribution with parameter $\theta \in \Theta$ such that $\mathbb{E}_\theta[T(X)] = \mu$. In Examples 4 and 7 respectively, θ is an inverse covariance matrix, $\nabla A(\theta) = -\frac{1}{2}\theta^{-1}$, and $-\mu$ is a correlation matrix. Thus $\Theta = \text{int}(\mathcal{N})$ is the set of all $n \times n$ positive definite matrices.

7.4.2 ML Estimation

Consider n iid samples $X^{(i)}$, $1 \le i \le n$ drawn from the exponential distribution p_θ. The ML estimator of θ given these n samples is obtained by solving

$$\begin{aligned} \max_\theta \frac{1}{n}\sum_{i=1}^n \ln p_\theta(X^{(i)}) &= \max_\theta \frac{1}{n}\sum_{i=1}^n \left[\theta^\top T(X^{(i)}) - A(\theta)\right] \\ &= \max_\theta [\theta^\top \hat{\mu} - A(\theta)] \end{aligned} \tag{7.49}$$

where

$$\hat{\mu} \triangleq \frac{1}{n}\sum_{i=1}^n T(X^{(i)})$$

is the sample mean estimator of the mean parameter μ. Since $A(\theta)$ is a convex function, the maximization problem of (7.49) is concave. Assuming the maximizer is an interior point of Θ, it can be found by setting the gradient of the loglikelihood function to zero, hence $\hat{\theta}$ satisfies the nonlinear equation

$$\nabla A(\hat{\theta}) = \hat{\mu}.$$

The gradient mapping could be hard to invert, however. For instance, for the Ising model of Example 1 we easily obtain

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^n \sum_{j\sim k} X_j^{(i)} X_k^{(i)}$$

but inverting the gradient mapping is a hard problem. Such is generally the case if p is a distribution over a Markov network with cycles.

7.4.3 Maximum Entropy

Consider a random variable X over a finite set $\mathcal{X}$. Its probability distribution p is unknown, however we are given the expected value $\mu_k = \mathbb{E}_p[T_k(X)]$ of d statistics $T_k(X)$, $1 \le k \le d$. A classical problem, which originates from statistical physics, is to find p that maximizes the entropy $H(p) = -\sum_x p(x) \ln p(x)$ subject to the d constraints above. Assuming the feasible set is nonvoid, the resulting distribution is called the *maximum-entropy* (or *maxent*) distribution. [4]

Since $H(p)$ is concave, the constraints are linear in p, and the probability simplex is a convex set, the maxent problem is concave. Its solution is obtained by introducing d Lagrange multipliers λ_k, $1 \le k \le d$ associated with the mean constraints, and a Lagrange multiplier λ_{d+1} associated with the constraint $\sum_x p(x) = 1$. Ignoring momentarily the nonnegativity constraints, we maximize the Lagrangian

$$\mathcal{L}(p, \lambda) \triangleq -\sum_{x \in \mathcal{X}} p(x) \ln p(x) + \sum_{k=1}^{d} \lambda_k \left(\sum_{x \in \mathcal{X}} p(x) T_k(x) - \mu_k \right) + \lambda_{d+1} \left(\sum_{x \in \mathcal{X}} p(x) - 1 \right)$$

over p, subject to the $d+1$ equality constraints

$$\begin{aligned} \sum_{x \in \mathcal{X}} p(x) T_k(x) &= \mu_k, \quad 1 \le k \le d, \\ \sum_{x \in \mathcal{X}} p(x) &= 1. \end{aligned} \tag{7.50}$$

The first-order optimality conditions are given by

$$0 = \frac{\partial \mathcal{L}(p, \lambda)}{\partial p(x)} = -\ln p(x) - 1 + \sum_{k=1}^{d} \lambda_k T_k(x) + \lambda_{d+1}, \quad \forall x \in \mathcal{X}$$

hence

$$p(x) = \frac{1}{Z} \exp\left\{ \sum_{k=1}^{d} \lambda_k T_k(x) \right\} \tag{7.51}$$

where $Z = \exp\{1 - \lambda_{d+1}\}$ is the normalization factor, and the nonnegativity constraints on p are automatically satisfied. The Lagrange multipliers are chosen so as to satisfy the constraints of (7.50). We see that the maxent solution (7.51) belongs to a d-dimensional exponential family of the form (7.45); the maximum entropy is given by

$$H(p) = -\mathbb{E}_p[\ln p(X)] = -\mathbb{E}_p[\theta^\top T(X) - A(\theta)] = -[\theta^\top \mu - A(\theta)].$$

Example 9. Let $X = (X_1, X_2) \in \{0, 1\}^2$ and consider maximizing entropy subject to the constraint $\mathbb{E}[X_1 X_2] = \mu$ where $\mu \in (0, 1)$. We obtain $p(x) = \frac{1}{Z} \exp\{\lambda x_1 x_2\}$. Since $\sum_x p(x) = 1$, the normalization constant is obtained as $Z = e^\lambda + 3$. We obtain λ from the constraint

$$\mu = \mathbb{E}_P[X_1 X_2] = \frac{e^\lambda}{e^\lambda + 3} \quad \Rightarrow \quad \lambda = \ln \frac{3\mu}{1 - \mu}.$$

[4] In the absence of constraint, the maxent solution is u, the uniform distribution over $\mathcal{X}$. Indeed the following inequality holds for any distribution p over $\mathcal{X}$: $0 \le D(p\|u) = -H(p) - \sum_x p(x) u(x) = -H(p) + \log |\mathcal{X}|$, and the lower bound is achieved by $p = u$.

The maxent solution takes the form

$$p(x) = \begin{cases} \mu & : (x_1, x_2) = (1,1) \\ \frac{1-\mu}{3} & : \text{ else} \end{cases}$$

and has entropy is $H(p) = -\mu \ln \mu - (1-\mu) \ln \frac{1-\mu}{3}$.

A similar version of the maxent problem exists for continuous random variables. The entropy function is replaced with the *differential entropy functional* $h(p) \triangleq -\int p \ln p$, and the maxent solution again takes an exponential form.

Example 10. Consider $X \in \mathbb{R}$ and the constraint $\mathbb{E}_p[X^2] = \sigma^2$. Then the maxent distribution is $p = \mathcal{N}(0, \sigma^2)$, and $h(p) = \frac{1}{2} \ln(2\pi e \sigma^2)$.

7.4.4 Conjugate Dual

Assume the exponential family is regular (Θ is an open set) and consider the *Fenchel transform*, aka *conjugate dual function* [11]

$$A^*(\mu) \triangleq \sup_{\theta \in \Theta} [\theta^\top \mu - A(\theta)], \quad \mu \in \mathcal{N}. \tag{7.52}$$

By duality we have

$$A(\theta) = \sup_{\mu \in \mathcal{N}} [\theta^\top \mu - A^*(\mu)], \quad \theta \in \Theta. \tag{7.53}$$

If the supremum in (7.53) is achieved at a point $\mu \in \text{int}(\mathcal{N})$, the gradient of the objective function at that point must be zero. Hence $\theta = \nabla A^*(\mu)$, which is dual to the relation $\mu = \nabla A(\theta)$ established earlier.

For $\mu \in \text{int}(\mathcal{N})$, the supremum in (7.52) is achieved at $\theta^*(\mu)$ satisfying the zero-gradient condition $\mu = \nabla A(\theta^*(\mu))$. Note that the maximum is not unique if the representation of the exponential family is overcomplete. By the property (7.46), we have $\mu = \mathbb{E}_{\theta^*(\mu)}[T(X)]$. Then

$$\begin{aligned} A^*(\mu) &= \theta^*(\mu)^\top \mu - A(\theta^*(\mu)) \\ &= \theta^*(\mu)^\top \mathbb{E}_{\theta^*(\mu)}[T(X)] - A(\theta^*(\mu)) \\ &= \mathbb{E}_{\theta^*(\mu)}[\ln p_{\theta^*(\mu)}(X)] \\ &= -H(p_{\theta^*(\mu)}). \end{aligned} \tag{7.54}$$

Hence $A^*(\mu)$ is the negative entropy of the maxent distribution.

Example 11. Let $\mathcal{X} = \{0, 1\}$ and consider the family of Bernoulli distributions with $p(1) \in (0, 1)$. Let $T(x) = x$, then this family can be represented as a 1-D regular exponential family $p_\theta(x) = e^{\theta x - A(\theta)}$ with natural parameter set $\Theta = \mathbb{R}$ and $A(\theta) = \ln(1 + e^\theta)$. The gradient mapping is given by $\mu = A'(\theta) = e^\theta / (1 + e^\theta) = p_\theta(1) \in (0, 1)$ and the inverse gradient mapping by $\theta^*(\mu) = \ln \frac{\mu}{1-\mu}$ for $\mu \in (0, 1)$. We have

$$A^*(\mu) = -H(p_{\theta^*(\mu)}) = \mu \ln \mu + (1-\mu) \ln(1-\mu), \quad \forall \mu \in (0, 1).$$

7.4.5 Variational Inequalities

Assume the "true" distribution is an element of an exponential family in canonical form: $p_\theta(x) = \exp\{\theta^\top T(x) - A(\theta)\}$. We will approximate p_θ using a tractable distribution q. The approximation criterion is the KL divergence

$$D(q\|p_\theta) = \mathbb{E}_q\left[\ln \frac{q(X)}{p_\theta(X)}\right] = A(\theta) - \theta^\top \mathbb{E}_q[T(X)] - H(q). \tag{7.55}$$

Our goal is to minimize this divergence over an interesting class $\mathcal{Q}$ of distributions that does not include p_θ. Consider a convex subset $\mathcal{N}_\mathcal{F} \subset \mathcal{N}$ of the set of realizable mean parameters, and the associated class of distributions

$$\mathcal{Q}_\mathcal{F} \triangleq \{q \ : \ \mu = \mathbb{E}_q[T(X)] \quad \text{for some } \mu \in \mathcal{N}_\mathcal{F}\}.$$

As shown in Sec. 7.4.3, given sufficient statistics $T(\cdot)$ and a realizable mean vector $\mu \in \mathcal{N}_\mathcal{F}$, the maxent distribution in class $\mathcal{Q}_\mathcal{F}$ admits an exponential form $p_{\theta^*(\mu)}$. Hence $\max_{q\in\mathcal{Q}_\mathcal{F}} H(q) = H(p_{\theta^*(\mu)}) = -A^*(\mu)$, where the last equality follows from (7.54). From (7.55) we obtain

$$\begin{aligned} \min_{q\in\mathcal{Q}_\mathcal{F}} D(q\|p_\theta) &= A(\theta) - \max_{q\in\mathcal{Q}_\mathcal{F}} [\theta^\top \mathbb{E}_q[T(X)] + H(q)] \\ &= A(\theta) - \max_{\mu\in\mathcal{N}_\mathcal{F}} [\theta^\top \mu - A^*(\mu)]. \end{aligned} \tag{7.56}$$

Therefore our variational problem (7.56) can be expressed as

$$\max_{\mu\in\mathcal{N}_\mathcal{F}} [\theta^\top \mu - A^*(\mu)] \tag{7.57}$$

which takes the same form as (7.53), except that the maximization is over the restricted set $\mathcal{N}_\mathcal{F}$. The idea behind this approximation strategy is that the maximization problem (7.57) should be tractable, as discussed in the next section.

7.5 Exponential distributions and graphical models

We now consider distributions on sequences $\boldsymbol{x} \in \mathcal{X}^\mathcal{V}$ associated with a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{X}$ is a finite set. These distributions will be pairwise Markov graphs of the form

$$p(x) = \frac{1}{Z} \prod_{i\in\mathcal{V}} \psi_i(x_i) \prod_{(i,j)\in\mathcal{E}} \psi_{ij}(x_i, x_j). \tag{7.58}$$

It is then useful to work with binary-valued sufficient statistics, namely

$$\begin{aligned} \forall i \in \mathcal{V}, \forall x \in \mathcal{X}: \quad T_{ix}(\boldsymbol{x}) &= \mathbb{1}\{x_i = x\} \\ \forall (i,j) \in \mathcal{E}, \forall x, x' \in \mathcal{X}: \quad T_{ijxx'}(\boldsymbol{x}) &= \mathbb{1}\{x_i = x, x_j = x'\} \end{aligned} \tag{7.59}$$

in the case of a pairwise Markov network. Any probability distribution (7.58) may be represented by the following (overcomplete) exponential family:

$$p_\theta(\boldsymbol{x}) = \exp\left\{\sum_{i\in\mathcal{V}} \sum_{x\in\mathcal{X}} \theta_i(x)\mathbb{1}\{x_i = x\} + \sum_{(i,j)\in\mathcal{E}} \sum_{x,x'\in\mathcal{X}} \theta_{ij}(x,x')\mathbb{1}\{x_i = x, x_j = x'\} - A(\theta)\right\}.$$

where $\theta_i(x) = \ln \psi_i(x)$ and $\theta_{ij}(x, x') = -\ln \psi_{ij}(x, x')$. The dimension of this family is $d = |\mathcal{V}|\,|\mathcal{X}| + |\mathcal{E}|\,|\mathcal{X}|^2$.

7.5.1 Marginal polytope

The mean parameters associated with the distribution p_θ are the 1-dimensional marginals for the vertices,

$$\forall i \in \mathcal{V}, \forall x \in \mathcal{X}: \quad \mu_i(x) = \mathbb{E}_\theta[T_{ix}(\boldsymbol{X})] = P_\theta\{X_i = x\}$$

and the pairwise marginals associated with the edge set $\mathcal{E}$,

$$\forall (i,j) \in \mathcal{E}, \forall x, x' \in \mathcal{X}: \quad \mu_{ij}(x,x') = \mathbb{E}_\theta[T_{ijxx'}(\boldsymbol{X})] = P_\theta\{X_i = x, X_j = x'\}.$$

The set $\mathcal{N}$ of realizable mean parameters is then called the *marginal polytope* and denoted by $\mathcal{N}(\mathcal{G})$. It follows from the discussion below (7.48) that $\mathcal{N}(\mathcal{G})$ is the convex hull of the points $T(\boldsymbol{x})$, $\boldsymbol{x} \in \mathcal{X}^\mathcal{V}$. The number of such points grows exponentially with the size of the graph. While $\mathcal{N}(\mathcal{G})$ can in principle be described by a system of linear equations, the size of this linear system is unfortunately exponential in the size of the graph.

7.5.2 Locally Consistent Marginals

Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, consider the set of marginal distributions τ_i on individual nodes $i \in \mathcal{V}$ and pairwise marginals τ_{ij} on edges $(i,j) \in \mathcal{E}$ that are locally consistent in the sense that

$$\begin{aligned}
\sum_{x_i \in \mathcal{X}} \tau_i(x_i) &= 1, \quad \forall i \in \mathcal{V} \\
\sum_{x_j \in \mathcal{X}} \tau_{ij}(x_i, x_j) &= \tau_i(x_i), \quad \forall (i,j) \in \mathcal{E}, x_i \in \mathcal{X} \qquad (7.60) \\
\sum_{x_i \in \mathcal{X}} \tau_{ij}(x_i, x_j) &= \tau_j(x_j), \quad \forall (i,j) \in \mathcal{E}, x_j \in \mathcal{X} \\
\tau_{ij}(x_i, x_j) &\geq 0, \quad \forall (i,j) \in \mathcal{E}, x_i, x_j \in \mathcal{X}. \qquad (7.61)
\end{aligned}$$

Definition. The *local marginal polytope* $\mathcal{L}(\mathcal{G})$ is the set of $\tau = (\{\tau_i\}_{i\in\mathcal{V}}, \{\tau_{ij}\}_{(i,j)\in\mathcal{E}})$ that satisfy the above consistency conditions.

This is a fairly simple polytope defined by $|\mathcal{V}| + (2|\mathcal{X}| + |\mathcal{X}|^2)|\mathcal{E}|$ linear constraints. Clearly the marginal polytope $\mathcal{N}(\mathcal{G})$ is a subset of $\mathcal{L}(\mathcal{G})$, but is the converse true?

Proposition. If $\mathcal{G}$ is a forest then $\mathcal{N}(\mathcal{G}) = \mathcal{L}(\mathcal{G})$. Any probability distribution on $\mathcal{G}$ can be expressed as follows in terms of its 1-D and pairwise marginals:

$$p(\boldsymbol{x}) = \left(\prod_{i\in\mathcal{V}} \mu_i(x_i)\right) \left(\prod_{(i,j)\in\mathcal{E}} \frac{\mu_{ij}(x_i, x_j)}{\mu_i(x_i)\mu_j(x_j)}\right). \qquad (7.62)$$

Proof. We first prove the claim for a tree. Any tree can be generated starting from a single node and adding one edge at a time. The claim (7.62) can be proven by induction. It clearly holds for a graph consisting of two nodes and a single edge (i,j), since $p(x) = \mu_{ij}(x_i, x_j)$ in this case. If a new node k and a new edge (j,k) are added to an existing tree

$(\mathcal{V}', \mathcal{E}')$ where $j \in \mathcal{V}'$, we obtain a new tree $(\mathcal{V}, \mathcal{E})$ with $\mathcal{V} = \mathcal{V}' \cup \{k\}$ and $\mathcal{E} = \mathcal{E}' \cup \{(j, k\}$. If

$$p(\boldsymbol{x}_{\mathcal{V}'}) = \left(\prod_{i \in \mathcal{V}'} \mu_i(x_i) \right) \left(\prod_{(i,j) \in \mathcal{E}'} \frac{\mu_{ij}(x_i, x_j)}{\mu_i(x_i)\mu_j(x_j)} \right)$$

then

$$p(\boldsymbol{x}) = p(\boldsymbol{x}_{\mathcal{V}'}, x_k) = p(\boldsymbol{x}_{\mathcal{V}'}) p(x_k | x_j) = p(\boldsymbol{x}_{\mathcal{V}'}) \frac{\mu_{jk}(x_j, x_k)}{\mu_j(x_j)}$$

satisfies (7.62). Next, if the forest contains more than one tree, the distribution $p(\boldsymbol{x})$ factors over the trees, and (7.62) still holds.

Finally, to any $\mu \in \mathcal{L}(\mathcal{G})$ we can associate a global distribution p using (7.62). The marginals and pairwise marginals of p are given by μ, hence $\mu \in \mathcal{N}(\mathcal{G})$. This proves the first part of the claim. □

However for a graph that is not a tree, $\mathcal{N}(\mathcal{G})$ is in general a strict subset of $\mathcal{L}(\mathcal{G})$. Consider for instance the 3-cycle with node set $\mathcal{V} = \{1, 2, 3\}$ and edge set $\mathcal{E} = \{(1, 2), (2, 3), (3, 1)\}$. Let $\mathcal{X} = \{0, 1\}$ and consider τ_1, τ_2, τ_3 that are uniform over $\mathcal{X}$, and

$$\tau_{12} = \tau_{23} = \begin{bmatrix} 0.5 - \epsilon & \epsilon \\ \epsilon & 0.5 - \epsilon \end{bmatrix}, \quad \tau_{31} = \begin{bmatrix} \epsilon & 0.5 - \epsilon \\ 0.5 - \epsilon & \epsilon \end{bmatrix}$$

for some $\epsilon \in (0, 0.5)$. By inspection, $\tau \in \mathcal{L}(\mathcal{G})$. However, for ϵ small enough, the definitions of $\tau_{12}, \tau_{23}, \tau_{31}$ imply respectively that $X_1 = X_2$, $X_2 = X_3$, and $X_3 \neq X_1$ with high probability. These conditions are incompatible, hence $\tau \notin \mathcal{N}(\mathcal{G})$.

In this example, the edge set is small, and it is relatively easy to determine that $\tau \notin \mathcal{N}(\mathcal{G})$. For a large graph, this would generally not be computationally feasible. Since τ may not be the marginals of any joint distribution on $\mathcal{G}$, τ are often referred to as *pseudomarginals*.

7.5.3 Entropy on Tree Graphs

Any distribution p defined on a tree graph is of the form (7.62). Hence its entropy is given by

$$\begin{aligned} H(p) &= \mathbb{E}_p[-\ln p(\boldsymbol{X})] \\ &= \sum_{i \in \mathcal{V}} \mathbb{E}_p[-\ln \mu_i(X_i)] - \sum_{(i,j) \in \mathcal{E}} \mathbb{E}_p \left[\ln \frac{\mu_{ij}(X_i, X_j)}{\mu_i(X_i)\mu_j(X_j)} \right] \\ &= \sum_{i \in \mathcal{V}} H(\mu_i) - \sum_{(i,j) \in \mathcal{E}} I(\mu_{ij}) \end{aligned} \tag{7.63}$$

where

$$\begin{aligned} I(\mu_{ij}) &\triangleq \mathbb{E}_{\mu_{ij}} \left[\ln \frac{\mu_{ij}(X_i, X_j)}{\mu_i(X_i)\mu_j(X_j)} \right] \\ &= D(\mu_{ij} \| \mu_i \mu_j) \\ &= H(\mu_i) + H(\mu_j) - H(\mu_{ij}) \end{aligned} \tag{7.64}$$

is the *mutual information* associated with the pairwise marginal μ_{ij}. Since this is a Kullback-Leibler divergence, it is nonnegative. The mutual information is zero if X_i and X_j are independent random variables and is upper-bounded by both $H(\mu_i)$ and $H(\mu_j)$ (value achieved if X_j is a function of X_i, or vice-versa).

The entropy of p is easily computed but is not concave in μ. Equivalently, the set of distributions p on a tree graph is generally nonconvex. (Consider a length-3 chain for instance.)

7.5.4 Gaussian Graphs

If $\mathcal{X} = \mathbb{R}$ and the joint distribution p on the graph is Gaussian, then the 1-D and pairwise marginal distributions are also Gaussian, and the sufficient statistics are linear and quadratic functions:

$$\begin{aligned} \forall i \in \mathcal{V}: \quad T_i(\boldsymbol{x}) &= x_i \\ T_{ii}(\boldsymbol{x}) &= -\frac{1}{2}x_i^2 \\ \forall (i,j) \in \mathcal{E}: \quad T_{ij}(\boldsymbol{x}) &= -\frac{1}{2}x_i x_j. \end{aligned} \tag{7.65}$$

The mean parameters are $\mu_i = \mathbb{E}[X_i]$, $\mu_{ii} = -\frac{1}{2}\mathbb{E}[X_i^2]$ for all $i \in \mathcal{V}$, and $\mu_{ij} = -\frac{1}{2}\mathbb{E}[X_i X_j]$ for all $(i,j) \in \mathcal{E}$. Any Gaussian distribution on $\mathcal{G}$ may be represented by the following (overcomplete) exponential family:

$$p_\theta(\boldsymbol{x}) = \exp\left\{ \sum_{i\in\mathcal{V}} \theta_i x_i - \frac{1}{2}\sum_{i\in\mathcal{V}} \theta_{ii} x_i^2 - \frac{1}{2}\sum_{(i,j)\in\mathcal{E}} \theta_{ij} x_i x_j - A(\theta) \right\}.$$

The dimension of this family is $d = 2|\mathcal{V}| + |\mathcal{E}|$.

The expression (7.63) holds for Gaussian graphs, if (discrete) entropy is replaced with differential entropy: $h(p) \triangleq -\int p \ln p \, d\nu$. For a scalar Gaussian random variable $\mathcal{N}(m, \sigma^2)$ we have $h(p) = \frac{1}{2}\ln(2\pi e \sigma^2)$, independently of m. For a n-variate Gaussian distribution $\mathcal{N}(m, \mathsf{C})$ we have $h(p) = \frac{n}{2}\ln(2\pi e |\mathsf{C}|^{1/n})$. The mutual information between two Gaussian random variables with normalized correlation coefficient $\rho \in (-1, 1)$ is given by $I(p) = -\frac{1}{2}\ln(1-\rho^2)$. Analogously to (7.63), the differential entropy for a Gaussian graph with node variances σ_i^2 and normalized correlation coefficients $\rho_{ij} \in (-1,1)$ for adjacent nodes is given by

$$h(p) = \sum_{i\in\mathcal{V}} h(\mu_i) - \sum_{(i,j)\in\mathcal{E}} I(\mu_{ij}) = \frac{1}{2}\sum_{i\in\mathcal{V}} \ln(2\pi\sigma_i^2) + \frac{1}{2}\sum_{(i,j)\in\mathcal{E}} \ln(1-\rho_{ij}^2). \tag{7.66}$$

7.5.5 Naive Mean Field

We now revisit the naive mean field method of Sec. 7.1 using the exponential distribution formalism. Consider a fully factorized approximation $q(\boldsymbol{x}) = \prod_{i\in\mathcal{V}} q_i(x_i)$ to p_θ. Hence $\mathcal{F} = (\mathcal{V}, \emptyset)$. For any $\mu \in \mathcal{N}(\mathcal{F})$ and $(i,j) \in \mathcal{E}$ we have $\mu_{ij} = \mu_i \mu_j$. Restating (7.56), we have

$$\min_{q\in\mathcal{Q}_\mathcal{F}} D(q\|p_\theta) = A(\theta) - \max_{\mu\in\mathcal{N}_\mathcal{F}} [\theta^\top \mu - A^*(\mu)]$$

where the negative entropy $A^*(\mu) = -\sum_{i\in\mathcal{V}} H(\mu_i)$ for $\mu \in \mathcal{N}_\mathcal{F}$ and

$$\begin{aligned}&\max_{\mu\in\mathcal{N}_\mathcal{F}} [\theta^\top \mu - A^*(\mu)] \\ &= \max_{\{\mu_i\}_{i\in\mathcal{V}}} \left\{ \sum_{i\in\mathcal{V}} \sum_{x\in\mathcal{X}} \theta_i(x)\mu_i(x) + \sum_{(i,j)\in\mathcal{E}} \sum_{x,x'\in\mathcal{X}} \theta_{ij}(x,x')\mu_i(x)\mu_j(x') + \sum_{i\in\mathcal{V}} H(\mu_i) \right\} \quad (7.67)\end{aligned}$$

The cost function is concave in each μ_i individually (but not concave in μ). The first-order necessary conditions for a maximizer are

$$\forall i, x_i: \quad 0 = \frac{\partial[\theta^\top \mu - A^*(\mu)]}{\partial \mu_i(x_i)} = \theta_i(x_i) + \sum_{j\in\mathcal{N}(i)} \sum_{x_j} \theta_{ij}(x_i,x_j)\mu_j(x_j) - 1 - \ln \mu_i(x_i)$$

hence

$$\mu_i(x_i) = \frac{1}{Z} \exp\left\{ \theta_i(x_i) + \sum_{j\in\mathcal{N}(i)} \sum_{x_j} \theta_{ij}(x_i,x_j)\mu_j(x_j) \right\}. \quad (7.68)$$

A coordinatewise maximization algorithm similar to that of Sec. 7.1.2 converges to a local maximum of the cost function.

Gaussian distribution. Refer to Example 5 earlier. Let $\theta = (h, \mathsf{J})$ and p_θ be a n-variate Gaussian distribution with and inverse covariance matrix $\mathsf{J} \succ 0$ and potential vector h. The sufficient statistics are $T_i(\boldsymbol{x}) = x_i$ and $T_{ij}(\boldsymbol{x}) = -x_i x_j$ for $1 \le i, j \le n$. The distribution

$$p_\theta(\boldsymbol{x}) = \exp\left\{ \sum_{i=1}^n h_i x_i - \frac{1}{2} \sum_{i,j=1}^n \mathsf{J}_{ij} x_i x_j - A(\theta) \right\}$$

(where $A(\theta) = -\frac{1}{2}\ln|\mathsf{J}| + \frac{1}{2} h^\top \mathsf{J} h + \frac{n}{2}\ln(2\pi)$) is to be approximated by a Gaussian distribution with mean $m \in \mathbb{R}^n$ and *diagonal* covariance matrix $\mathsf{J} - mm^\top$. Let $\mu = (m, -\mathsf{J})$. Then

$$\mathcal{N}_\mathcal{F} = \{(m, -\mathsf{J}) : \ \mathsf{J} - mm^\top = \mathrm{diag}(\mathsf{J} - mm^\top) \succ 0\}.$$

For $\mu = (m, -\mathsf{J}) \in \mathcal{N}_\mathcal{F}$ we have $\mathsf{J}_{ij} = m_i m_j$ for $i \ne j$. The differential entropy of a $\mathcal{N}(m, \mathsf{C})$ distribution is given by $h(p) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathsf{C}|$ (independently of m) hence

$$A^*(\mu) = -h(p_{\theta^*(\mu)}) = \frac{n}{2}\ln(2\pi) + \frac{1}{2}\sum_{i=1}^n \ln(\mathsf{J}_{ii} - m_i^2), \quad \forall \mu \in \mathcal{N}_\mathcal{F}$$

and

$$\max_{\mu\in\mathcal{N}_\mathcal{F}} [\theta^\top \mu - A^*(\mu)] = \max_{(m,-\mathsf{J})\in\mathcal{N}_\mathcal{F}} \left\{ \sum_{i=1}^n h_i m_i - \frac{1}{2}\sum_{i,j=1}^n \mathsf{J}_{ij}\mathsf{J}_{ij} + \frac{1}{2}\sum_{i=1}^n \ln(\mathsf{J}_{ii} - m_i^2) + \frac{n}{2}\ln(2\pi) \right\}.$$

Setting the derivative with respect to each J_{ii} to zero, we obtain

$$0 = -\mathsf{J}_{ii} + \frac{1}{\mathsf{J}_{ii} - m_i^2}, \quad 1 \le i \le n. \quad (7.69)$$

Setting the derivative with respect to each m_i to zero, we obtain

$$0 = h_i - \sum_{j \neq i} \mathsf{J}_{ij} m_j - \frac{m_i}{\mathsf{J}_{ii} - m_i^2}, \quad 1 \leq i \leq n. \tag{7.70}$$

It is easily verified that the solution to this system is given by $m = \mathsf{J}^{-1}h$ (match the mean of X) and $\mathsf{J}_{ii} = m_i^2 + 1/\mathsf{J}_{ii}$ (recall $1/\mathsf{J}_{ii}$ is the conditional variance of X_i given the other random variables). For large n, the following Gauss-Seidel type method can be used to iterate to the stationary point:

$$\hat{m}_i^{(k+1)} = \frac{1}{\mathsf{J}_{ii}} \left(h_i - \sum_{j \neq i} \mathsf{J}_{ij} \hat{m}_j^{(k)} \right) \tag{7.71}$$

with $\hat{\mathsf{J}}_{ii}^{(k)} = \frac{1}{\mathsf{J}_{ii}} + (\hat{m}_i^{(k)})^2$.

7.5.6 Structured Mean Field

The naive mean field approach can be extended to tractable subgraphs, typically acyclic graphs, for which an exact expression is available for the negative entropy $A^*(\mu)$ [2, Sec. 5.5] [10]. Again let p be an exponential family associated with a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Consider a subgraph $\mathcal{F} = (\mathcal{V}, \mathcal{E}')$ where $\mathcal{E}'$ is a strict subset of $\mathcal{E}$.

The vector of mean parameters associated with a distribution on $\mathcal{G}$ will be denoted by $\mu = (\mu_\beta)_{\beta \in \mathcal{I}(\mathcal{G})}$ where $\mathcal{I}(\mathcal{G}) = (\mathcal{V} \times \mathcal{X}) \cup (\mathcal{E} \times \mathcal{X}^2)$ is the index set and has cardinality $d = |\mathcal{I}(\mathcal{G})| = |\mathcal{V}||\mathcal{X}| + |\mathcal{E}||\mathcal{X}|^2$. Similarly, we define a restricted index set $\mathcal{I}(\mathcal{F}) = (\mathcal{V} \times \mathcal{X}) \cup (\mathcal{E}' \times \mathcal{X}^2) \subset \mathcal{I}(\mathcal{G})$ associated with the nodes and edges of $\mathcal{F}$; its cardinality $d' = |\mathcal{I}(\mathcal{F})| = |\mathcal{V}||\mathcal{X}| + |\mathcal{E}'||\mathcal{X}|^2$ is lower than d. The restriction of the vector μ to $\mathcal{F}$ will be denoted by $\mu' = (\mu_\alpha)_{\alpha \in \mathcal{I}(\mathcal{F})}$ and has dimension d'.

To solve our variational problem (7.57), we can express the components of μ in $\mathcal{I}(\mathcal{G}) \setminus \mathcal{I}(\mathcal{F})$ as a function of those in $\mathcal{I}(\mathcal{F})$. Specifically, we define the *embedding* $\Gamma : \mathbb{R}^{d'} \to \mathbb{R}^{d-d'}$ whose components are

$$\Gamma_\beta(\mu') = \mathbb{E}_{\mu'}[T_\beta(X)], \quad \beta \in \mathcal{I}(\mathcal{G}) \setminus \mathcal{I}(\mathcal{F}).$$

Then we may write the variational problem (7.57) as

$$\max_{\mu \in \mathcal{N}_\mathcal{F}} [\theta^\top \mu - A^*(\mu)] = \max_{\mu' \in \mathcal{N}'_\mathcal{F}} \left[\sum_{\alpha \in \mathcal{I}(\mathcal{F})} \theta_\alpha \mu'_\alpha + \sum_{\beta \in \mathcal{I}(\mathcal{G}) \setminus \mathcal{I}(\mathcal{F})} \theta_\beta \Gamma_\beta(\mu') - A^*(\mu') \right].$$

Denote the cost function in brackets above by $f(\mu')$, and define the vectors $\theta' = (\theta_\alpha)_{\alpha \in \mathcal{I}(\mathcal{F})}$ and $\theta'' = (\theta_\beta)_{\beta \in \mathcal{I}(\mathcal{G}) \setminus \mathcal{I}(\mathcal{F})}$. Also define the $d' \times (d - d')$ *embedding Jacobian* matrix $\mathsf{J}(\mu')$ with components

$$\mathsf{J}_{\alpha\beta}(\mu') = \frac{\partial \Gamma_\beta(\mu')}{\partial \mu'_\alpha}, \quad \alpha \in \mathcal{I}(\mathcal{F}),\ \beta \in \mathcal{I}(\mathcal{G}) \setminus \mathcal{I}(\mathcal{F}).$$

The first-order necessary conditions for optimality are

$$0 = \nabla f(\mu') = \theta' + \mathsf{J}(\mu')\theta'' - \nabla A^*(\mu')$$

which can also be written as

$$\nabla A^*(\mu') = \theta' + \mathsf{J}(\mu')\theta''. \tag{7.72}$$

Since $\nabla A(\nabla A^*(\mu')) = \mu'$, we obtain

$$\mu' = \nabla A(\theta' + \mathsf{J}(\mu')\theta'').$$

The paper [10] shows that this nonlinear system can be solved relatively easily for an interesting class of tractable subgraphs.

Example. If $\mathcal{E}' = \emptyset$, we verify that we recover the naive mean field equations of the previous section. For $\alpha = (i, x_i)$ and $\beta = ((j,k), x_j, x_k)$ where $i \in \mathcal{V}$, $(j,k) \in \mathcal{E}$ and $x_i, x_j, x_k \in \mathcal{X}$, we have

$$\Gamma_\beta(\mu') = \mathbb{E}_{\mu'}[T_\beta(X)] = \Pr\{X_j = x_j, X_k = x_k\} = \mu_{jk}(x_j, x_k) = \mu_j(x_j)\mu_k(x_k)$$

hence

$$\mathsf{J}_{\alpha\beta}(\mu') = \frac{\partial \Gamma_\beta(\mu')}{\partial \mu_i(x_i)} = \begin{cases} \mu_j(x_j) & : \ (i,x_i) = (k,x_k) \\ \mu_k(x_k) & : \ (i,x_i) = (j,x_j) \\ 0 & : \ \text{else.} \end{cases}$$

Since $\partial A^*(\mu')/\partial \mu_\alpha = -1 - \ln \mu_i(x_i)$, the fixed-point equation (7.72) becomes

$$-1 - \ln \mu_i(x_i) = \theta_i(x_i) + \sum_{j \in \mathcal{N}(i)} \sum_{x_j} \theta_{ij}(x_i, x_j)\mu_j(x_j), \quad \forall i, x_i$$

hence (7.68) follows.

7.5.7 Bethe Entropy Approximation

For a distribution p that is not defined on a tree graph, $H(p)$ does not admit a simple expression, and cannot be expressed simply in terms of 1-D marginals and pairwise marginals. (Verify on a 3-cycle). However if these marginals are known, one could use (7.63) as an approximation to $H(p)$. This approximation is known as the *Bethe approximation*, and the functional

$$H_{\text{Bethe}}(\tau) \triangleq \sum_{i \in \mathcal{V}} H(\tau_i) - \sum_{(i,j) \in \mathcal{E}} I(\tau_{ij}), \quad \tau \in \mathcal{L}(\mathcal{G}) \tag{7.73}$$

is known as the *Bethe entropy*. This "entropy" is well defined for all pseudomarginals $\tau \in \mathcal{L}(\mathcal{G})$.

The *Bethe variational problem* is defined as

$$A_{\text{Bethe}}(\theta) \triangleq \max_{\tau \in \mathcal{L}(\mathcal{G})} [\theta^\top \tau + H_{\text{Bethe}}(\tau)] \tag{7.74}$$

and is relatively tractable owing to the simple nature of $\mathcal{L}(\mathcal{G})$ and the availability of a closed-form expression for $H_{\text{Bethe}}(\tau)$. Compare with the expression

$$A(\theta) = \sup_{\mu \in \mathcal{N}(\mathcal{G})} [\theta^\top \mu + H(p_{\theta(\mu)})]$$

that follows from (7.54) (7.53) and is unfortunately intractable because of the complex nature of $\mathcal{N}(\mathcal{G})$ and the lack of an explicit form for $H(p_\mu)$. For a general graph, $\mathcal{N}(\mathcal{G}) \subset \mathcal{L}(\mathcal{G})$ and Bethe entropy is an approximation to entropy; $A_{\text{Bethe}}(\theta)$ is not a bound on $A(\theta)$, only an approximation (see example below). For a tree graph however, $\mathcal{N}(\mathcal{G}) = \mathcal{L}(\mathcal{G})$ and $A_{\text{Bethe}}(\theta) = A(\theta)$.

Inexactness of Bethe approximation. Consider a fully connected graph with four nodes, $\mathcal{V} = \{1, 2, 3, 4\}$, uniform 1-D marginals μ_i, $i \in \mathcal{V}$, and pairwise marginals

$$\mu_{ij} = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix} \quad \forall i, j \in \mathcal{V}.$$

We have $\mu \in \mathcal{N}(\mathcal{G})$; indeed the distribution p that places probability $\frac{1}{2}$ on the sequences $(0, 0, 0, 0)$ and $(1, 1, 1, 1)$ satisfies the marginal constraints above. We have $H(\mu_i) = \ln 2$ for all $i \in \mathcal{V}$ and $I(\mu_{ij}) = \ln 2$ for all $i, j \in \mathcal{V}$. Since there are 6 edges, we obtain

$$H_{\text{Bethe}}(\mu) = 4 \ln 2 - 6 \ln 2 = -2 \ln 2 < 0$$

which shows that the Bethe entropy does not satisfy the same properties as an entropy (it can be negative). The actual entropy $H(p) = \ln 2 > 0$.

Iterative solution for Bethe Variational Problem. The maximization problem of (7.74) can be solved using a Lagrangian approach, introducing Lagrange multipliers for the linear constraints of (7.61) that define $\mathcal{L}(\mathcal{G})$. The Lagrangian is given by

$$\begin{aligned} \mathcal{L}(\tau, \lambda) &= \theta^\top \tau + H_{\text{Bethe}}(\tau) - \sum_{i \in \mathcal{V}} \lambda_i \left(\sum_{x_i} \tau_i(x_i) - 1 \right) \\ &\quad - \sum_{(i,j) \in \mathcal{E}} \sum_{x_i} \lambda_{ji}(x_i) \left(\sum_{x_j} \tau_{ij}(x_i, x_j) - \tau_i(x_i) \right) \\ &\quad - \sum_{(i,j) \in \mathcal{E}} \sum_{x_j} \lambda_{ij}(x_j) \left(\sum_{x_i} \tau_{ij}(x_i, x_j) - \tau_j(x_j) \right). \end{aligned}$$

Setting to zero the partial derivatives, we obtain

$$\begin{aligned} 0 = \frac{\partial \mathcal{L}(\tau, \lambda)}{\partial \tau_i(x_i)} &= \theta_i(x_i) + \frac{\partial H_{\text{Bethe}}(\tau)}{\partial \tau_i(x_i)} - \lambda_i + \sum_{j \in \mathcal{N}(i)} \lambda_{ji}(x_i) \\ 0 = \frac{\partial \mathcal{L}(\tau, \lambda)}{\partial \tau_{ij}(x_i, x_j)} &= \theta_{ij}(x_i, x_j) + \frac{\partial H_{\text{Bethe}}(\tau)}{\partial \tau_{ij}(x_i, x_j)} - \lambda_{ji}(x_i) - \lambda_{ij}(x_j) \end{aligned} \tag{7.75}$$

for all $(i,j) \in \mathcal{E}$ and $x_i, x_j \in \mathcal{X}$. We have

$$\begin{aligned}
\frac{\partial H(\tau_i)}{\partial \tau_i(x_i)} &= -\ln \tau_i(x_i) - 1 \\
\frac{\partial I(\tau_{ij})}{\partial \tau_i(x_i)} &= -\sum_{x_j} \tau_{ij}(x_i, x_j) \frac{1}{\tau_i(x_i)} = -1 \\
\frac{\partial I(\tau_{ij})}{\partial \tau_{ij}(x_i, x_j)} &= \ln \frac{\tau_{ij}(x_i, x_j)}{\tau_i(x_i)\tau_j(x_j)} + 1.
\end{aligned}$$

The partial derivatives of the Bethe entropy are then obtained from (7.73) as

$$\begin{aligned}
\frac{\partial H_{\text{Bethe}}(\tau)}{\partial \tau_i(x_i)} &= -\ln \tau_i(x_i) - 1 \\
\frac{\partial H_{\text{Bethe}}(\tau)}{\partial \tau_{ij}(x_i, x_j)} &= -\ln \frac{\tau_{ij}(x_i, x_j)}{\tau_i(x_i)\tau_j(x_j)}.
\end{aligned}$$

Substituting back into (7.75) we obtain

$$0 = \theta_i(x_i) - \ln \tau_i(x_i) - \lambda_i + \sum_{j \in \mathcal{N}(i)} \lambda_{ji}(x_i) \tag{7.76}$$

$$0 = \theta_{ij}(x_i, x_j) - \ln \frac{\tau_{ij}(x_i, x_j)}{\tau_i(x_i)\tau_j(x_j)} - \lambda_{ji}(x_i) - \lambda_{ij}(x_j). \tag{7.77}$$

Let

$$\psi_i(x_i) = \exp\{\theta_i(x_i)\}, \quad \psi_{ij}(x_i, x_j) = \exp\{\theta_{ij}(x_i, x_j)\}$$

and define the message from node i to node j as

$$m_{i \to j}(x_j) = \exp\{\lambda_{ij}(x_j)\}, \quad x_j \in \mathcal{X}.$$

Then from (7.76) we have

$$\tau_i(x_i) = \frac{1}{Z_i} \psi_i(x_i) \prod_{j \in \mathcal{N}(i)} m_{j \to i}(x_i) \tag{7.78}$$

and from (7.77) and (7.78) we obtain

$$\tau_{ij}(x_i, x_j) = \frac{1}{Z_{ij}} \psi_{ij}(x_i, x_j) \psi_i(x_i) \psi_j(x_j) \left(\prod_{k \in \mathcal{N}(i) \setminus \{j\}} m_{k \to i}(x_i) \right) \left(\prod_{k \in \mathcal{N}(j) \setminus \{i\}} m_{k \to j}(x_j) \right).$$

The marginal compatibility condition (7.60) implies

$$\begin{aligned}
&\frac{1}{Z_i} \psi_i(x_i) \prod_{k \in \mathcal{N}(i)} m_{k \to i}(x_i) \\
&= \left(\prod_{k \in \mathcal{N}(i) \setminus \{j\}} m_{k \to i}(x_i) \right) \psi_i(x_i) \frac{1}{Z_{ij}} \sum_{x_j} \psi_{ij}(x_i, x_j) \psi_j(x_j) \left(\prod_{k \in \mathcal{N}(j) \setminus \{i\}} m_{k \to j}(x_j) \right)
\end{aligned}$$

for all $(i,j) \in \mathcal{E}$ and $x_i \in \mathcal{X}$. Simplifying, we obtain

$$m_{j\to i}(x_i) = \frac{1}{Z_i} \sum_{x_j} \psi_{ij}(x_i, x_j)\psi_j(x_j) \left(\prod_{k \in \mathcal{N}(j)\setminus\{i\}} m_{k\to j}(x_j) \right). \tag{7.79}$$

Equations (7.79) (7.78) are the same as the belief propagation equations for a pairwise Markov network with joint distribution $p(\boldsymbol{x}) = \frac{1}{Z}\prod_{(i,j)\in\mathcal{E}} \psi_i(x_i)\psi_{ij}(x_i, x_j)$. Hence the stationary points of the belief propagation algorithm satisfy the first-order optimality conditions for the Bethe variational problem.

7.5.8 General Markov Networks

We have focused on pairwise Markov networks for simplicity of the exposition. The concepts described above can be applied to more general graphs. For instance, one may apply the junction tree algorithm and represent the network with a tree of cliques. If the maximal clique size is small (e.g., 3, such as the k-spin model from statistical physics presented earlier), the distribution factorizes as follows:

$$p(\boldsymbol{x}) = \frac{\prod_{\mathcal{C}\in\mathscr{C}} \mu_{\mathcal{C}}(x_{\mathcal{C}})}{\prod_{\mathcal{S}\in\mathscr{F}} \mu_{\mathcal{S}}(x_{\mathcal{S}})^{d(\mathcal{S})-1}}$$

where $\mathcal{C}$ denotes a clique, $\mathcal{S}$ a separator set, and $d(\mathcal{S})$ is the degree of $\mathcal{S}$, namely, the number of maximal cliques to which $\mathcal{S}$ is adjacent. For instance, in a 4-node network with edges $(1,2),(1,3),(2,3),(2,4),(3,4)$, we have two maximal cliques $(1,2,3)$ and $(2,3,4)$ and one separator set $(2,3)$. One can define a set of realizable mean parameters $\mathcal{N}(\mathcal{G})$ for $(\{\mu_{\mathcal{C}}\}_{\mathcal{C}\in\mathscr{C}}, \{\mu_{\mathcal{S}}\}_{\mathcal{S}\in\mathscr{F}})$, as well as a set $\mathcal{L}(\mathcal{G})$ of locally consistent pseudomarginals $(\{\tau_{\mathcal{C}}\}_{\mathcal{C}\in\mathscr{C}}, \{\tau_{\mathcal{S}}\}_{\mathcal{S}\in\mathscr{F}})$ such that

$$\sum_{x_{\mathcal{S}}\in\mathcal{X}^{\mathcal{S}}} \tau_{\mathcal{S}}(x_{\mathcal{S}}) = 1, \quad \forall \mathcal{S} \in \mathscr{F}$$

and

$$\sum_{x'_{\mathcal{C}}\in\mathcal{X}^{\mathcal{C}}: x'_{\mathcal{S}} = x_{\mathcal{S}}} \tau_{\mathcal{C}}(x'_{\mathcal{C}}) = \tau_{\mathcal{S}}(x_{\mathcal{S}}), \quad \forall \mathcal{C} \in \mathscr{C}, \mathcal{S} \subset \mathcal{C},\, x_{\mathcal{S}} \in \mathcal{X}^{\mathcal{S}}.$$

This approach is unwieldy if the maximal clique size is large. In this case it is more convenient to represent the network with a factor graph [2, 12].

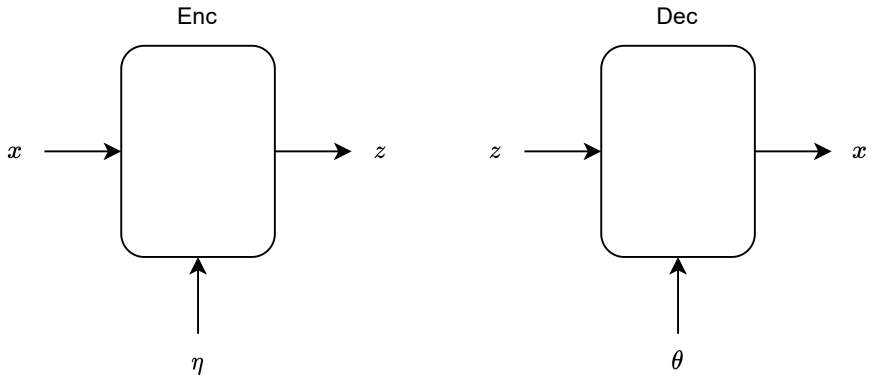
7.6 Variational Autoencoders

The previous section applied variational inference to problems where the true distribution p has a known analytical form and is approximated by a distribution q from a suitably designed exponential family. This scenario requires iterative evaluation of closed-form expressions for certain expectations.

This section describes variational autoencoders (VAEs), a more general scenario where p is completely unknown but a training set drawn iid from p is available [13]. The model

assumes that the sample x (e.g., an image) can be accurately represented by a lower-dimensional "feature vector" z learned using an *encoder network*, in the sense that a mapping applied to z (the *decoding network*) produces $\hat{x}$ that is close to x in some sense. Applications include generative AI, where realistic pictures of faces, art, etc. can be produced by passing a random sample Z through the decoding network.

Specifically, a tractable joint distribution $r_\theta(x, z) = \pi_\theta(z)p_\theta(x|z)$ is assumed. However the posterior distribution $\pi_\theta(z|x)$ is intractable, so apparently it is difficult to obtain features. The idea is to approximate this posterior with a tractable $q_\eta(z|x)$. When the method is applied to neural networks, the parameters η and θ describe the encoding and decoding networks, respectively.



As an example, $q_\eta(z|x)$ could be a Gaussian distribution with mean vector $\mu(x)$ and diagonal covariance matrix $\Lambda(x)$ both depending generally nonlinearly on x, and $p_\theta(x|z)$ could be a Gaussian distribution with mean $f_\theta(z)$ and covariance matrix equal to ϵ times the identity matrix. The prior distribution $\pi_\theta(z)$ could be a standard normal distribution.

The variational approximation maximizes an evidence lower bound (ELBO) on the log-likelihood function $\ln p_\theta(x)$. The following identity holds for any conditional distribution $q(z|x)$:

$$\begin{aligned} \ln p_\theta(x) &= \ln\left(\frac{p_\theta(x|z)\pi_\theta(z)}{\pi_\theta(z|x)}\,\frac{q(z|x)}{q(z|x)}\right) \\ &= \ln p_\theta(x|z) - \ln\frac{q(z|x)}{\pi_\theta(z)} + \ln\frac{q(z|x)}{\pi_\theta(z|x)}. \end{aligned}$$

Taking the expectation with respect to $q(\cdot|x)$, we obtain

$$\ln p_\theta(x) = \underbrace{\mathbb{E}_{q(\cdot|x)}[\ln p_\theta(x|Z)] - D(q(\cdot|x)\|\pi_\theta)}_{=\mathcal{L}(q,\theta,x)} + D(q(\cdot|x)\|\pi_\theta(\cdot|x)). \tag{7.80}$$

The difference between the first two terms in the right side is the ELBO:

$$\mathcal{L}(q, \theta, x) = \mathbb{E}_{q(\cdot|x)}[\ln p_\theta(x|Z)] - D(q(\cdot|x)\|\pi_\theta) \tag{7.81}$$

and is to be maximized over θ and over a tractable class of distributions $\{q_\eta\}$.

Evaluation and maximization of the ELBO turns out to be tractable. From (7.80), maximizing the ELBO over q is equivalent to minimizing the discrepancy $D(q(\cdot|x)\|\pi_\theta(\cdot|x))$

between $q(\cdot|x)$ and the (intractable) posterior $\pi_\theta(\cdot|x)$. Maximizing the ELBO over θ is then approximately equivalent to maximizing the log-likelihood function $\ln p_\theta(x)$.

The ELBO in (7.81) is the difference between two terms. The first term is viewed as a reconstruction loss: for the conditionally Gaussian model mentioned above, we have $\ln p_\theta(x|z)] = \frac{1}{\epsilon}\|x - f_\theta(z)\|^2 + \text{cst}$. The second term penalizes misfit between $q(\cdot|x)$ and the prior π_θ on the latent variables.

In practice a sample-based approximation could be used to evaluate both expectations in (7.81):

$$\mathcal{L}(q, \theta, x) \approx \widetilde{\mathcal{L}}(q, \theta, x) \triangleq \frac{1}{L} \sum_{\ell=1}^{L} \left[\ln p_\theta(x|Z^\ell) - \ln \frac{q(Z^\ell|x)}{\pi(Z^\ell)} \right] \tag{7.82}$$

where $\{Z^\ell\}_{\ell=1}^L$ are i.i.d. samples of the conditional distribution $q(\cdot|x)$. In fact, the KL divergence term admits a closed-form expression. For $\pi_\theta = \mathcal{N}(0, \mathsf{I})$ and $q(\cdot|x) = \mathcal{N}(\mu(x), \Lambda(x))$, we have

$$D(q(\cdot|x)\|\pi_\theta) = \frac{1}{2} \sum_i [-1 - \ln \sigma_i^2(x) + \sigma_i^2(x) + \mu_i^2(x)]$$

where $\{\sigma_i^2(x)\}$ are the diagonal entries of $\Lambda(x)$.

Given a parametric model $(r_\theta,\ q_\eta)$, we would like to learn parameters θ and η that maximize the function

$$\mathcal{L}(\theta, \eta) = \mathbb{E}_p[\mathcal{L}(q_\eta, \theta, X)]$$

where P is the true distribution of X. We do not know P but have access to a training data set $\{X^t\}_{t\in\mathcal{T}}$ drawn i.i.d. P. Then $\mathcal{L}(\theta, \eta)$ is replaced by

$$\widehat{\mathcal{L}}(\theta, \eta) = \frac{1}{|\mathcal{T}|} \sum_{t\in\mathcal{T}} \widetilde{\mathcal{L}}(q_\eta, \theta, X^t)]. \tag{7.83}$$

The following example from [13] shows random faces generated from only a 2-dimensional space of features.

Bibliography

[1] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational Bayes: A Review for Statisticians," *Journal of the American Statistical Association*, Vol. 112, No. 518, pp. 859—877, 2017.

[2] M. Wainwright and M. Jordan, *Graphical models, exponential families, and variational inference*, NOW publisher, 2008. Available online.

[3] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.

[4] D. Barber, *Bayesian Reasoning and Machine Learning*, Cambridge University Press, UK, 2012. Available online: http://web4.cs.ucl.ac.uk/staff/D.Barber/textbook/090310.pdf

[5] E. Ising, "Beitrag zur Theorie des Ferromagnetismus," *Zeitschrift für Physik*, Vol. 31, No. 1, pp. 253—258, 1925.

[6] L. Onsager, "Crystal statistics. I. A two-dimensional model with an order-disorder transition," *Physical Review*, Series II, Vol. 65, Nos. 3,4, pp. 117—149, 1944.

[7] L. K. Saul and M. I. Jordan,"Exploiting tractable substructures in intractable networks," in *Advances in Neural Information Processing Systems*, MIT Press, pp. 486—492, 1996.

[8] D. Barber and W. Wiegerinck, "Tractable Variational Structures for Approximating Graphical Models," *Proc. NIPS*, 1998.

[9] E. P. Xing, M. I. Jordan, and S. Russell, "A Generalized Mean Field Algorithm for Variational Inference in Exponential Families," *Proc. 19th Conf. on Uncertainty in Artificial Intelligence*, 2003.

[10] A. Bouchard-Côté and M. I. Jordan, "Optimization of Mean Field Objectives," *Proc. 25th Conf. on Uncertainty in Artificial Intelligence*, 2009.

[11] R. T. Rockafellar, *Convex Analysis*, 1972.

[12] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms," *IEEE Trans. Information Theory*, Vol. 51, No. 7, pp. 2282—2312.

[13] D. Kingsma and M. Welling, "Autoencoding Variational Bayes," *Proc. Int. Conf. on Learning Representations (ICLR)*, Banff, Canada, Apr. 2014.