Chapter 16

Graphical Models and Exponential Families

This chapter discusses exponential families of distributions with applications to Bayesian networks, focusing on links with convex analysis and specifically with the theory of conjugate duality. This is then applied to updating the probability distribution in a graphical model in the light of new information. One of the key features of exponential families, with mean parameters, is the relative ease with which the entropy and Kullback Leibler distance between two members of the family may be computed.

16.1 Introduction to Exponential Families

The notations are as before. Let \( V = \{X_1, \ldots, X_d\} \) denote the random variables. For \( j = 1, \ldots, d \), \( X_j \) will denote the state space for variable \( X_j \). If \( X_j \) is continuous, then \( X_j \subseteq \mathbb{R} \) (the real numbers). If \( X_j \) is discrete, then \( X_j = \{x_j^{(1)}, \ldots, x_j^{(k_j)}\} \), where \( k_j \) is possibly \( +\infty \). As usual, the notation \( \underline{X} = (X_1, \ldots, X_d) \) denotes the row vector of variates. An instantiation of \( \underline{X} \) will be denoted \( \underline{x} \in X_1 \times \cdots X_d = \mathcal{X} \) (when no subscript is employed, \( \mathcal{X} \) denotes the product space, which is the state space of the row vector \( \underline{X} \)).

An exponential family is a family of probability distributions satisfying certain properties, listed in Definition 16.1 below. For the purposes of Bayesian Networks, the emphasis is on discrete variables and Gaussian variables.

**Definition 16.1** (Exponential Family). An exponential family is a family of probability distributions \( \{\mathbb{P}_\theta: \theta \in \Theta\} \), where \( \Theta \) is a parameter space. These are defined by a probability mass function \( \mathbb{P}_\underline{X}(\underline{x}) \) if \( \underline{X} \) are discrete variables, or a probability density function \( \pi_\underline{X}(\underline{x}) \) for continuous variables, indexed by a parameter set \( \Theta \subseteq \mathbb{R}^p \) (where \( p \) is possibly infinite), where there is a function \( \Phi: \mathcal{X} \to \mathbb{R}^p \), a function \( A: \Theta \to \mathbb{R} \) and a function \( h: \mathcal{X} \to \mathbb{R} \) such that

\[
\mathbb{P}_\underline{X}(\underline{x}|\theta) = \exp\{\langle \theta, \Phi(\underline{x}) \rangle - A(\theta) \} h(\underline{x})
\]

if \( \underline{X} \) is a discrete random vector and
\[ \pi_X(x|\theta) = \exp\{\langle \theta, \Phi(x) \rangle - A(\theta) \} h(x) \]

if \( X \) is a continuous random vector.

It is convenient to use the notation \( \mathcal{I} \) to denote the indexing set for the parameters; \( \theta = (\theta_\alpha)_{\alpha \in \mathcal{I}} \). Then \( \Phi \) denotes a collection of functions \( \Phi = (\phi_\alpha)_{\alpha \in \mathcal{I}} \), where \( \phi_\alpha : \mathcal{X} \to \mathbb{R} \). The inner product notation is defined as
\[
\langle \theta, \Phi(x) \rangle = \sum_{\alpha \in \mathcal{I}} \theta_\alpha \phi_\alpha(x).
\]

The parameters in the vector \( \theta \) are known as the canonical parameters or exponential parameters.

Attention will be restricted to distributions where \(|\mathcal{I}| = p < +\infty\); namely, \( \mathcal{I} \) has a finite number, \( p \), of elements.

Since \( \sum_X \mathbb{P}_X(x|\theta) = 1 \) for discrete variables and \( \int_X \pi_X(x|\theta)dx = 1 \) for continuous variables, it follows that the quantity \( A \), known as the log partition function, is given by the expression
\[
A(\theta) = \log \int_X \exp\{\langle \theta, \Phi(x) \rangle \} h(x)dx,
\]
for continuous variables and
\[
A(\theta) = \log \sum_X \exp\{\langle \theta, \Phi(x) \rangle \} h(x)
\]
for discrete variables. It is assumed that \( h, \theta \) and \( \Phi \) satisfy appropriate conditions so that \( A \) is finite.

Set
\[
P(x; \theta) = \frac{\mathbb{P}_X(x|\theta)}{h(x)}. \tag{16.1}
\]

With the set of functions \( \Phi \) fixed, each parameter vector \( \theta \) indexes a particular probability function \( \mathbb{P}_X(x|\theta) \) belonging to the family. The exponential parameters of interest belong to the parameter space, which is the set
\[
\Theta = \{ \theta \in \mathbb{R}^p | A(\theta) < +\infty \}. \tag{16.2}
\]

It will be seen shortly that \( A \) is a convex function of \( \theta \).

**Definition 16.2** (Regular Families). An exponential family for which the domain \( \Theta \) of Equation (16.2) is an open set is known as a regular family.

Attention will be restricted to regular families.

**Definition 16.3** (Minimal Representation). An exponential family, defined using a collection of functions \( \Phi \) for which there is no linear combination \( \langle \theta, \Phi(x) \rangle = \sum_{\alpha \in \mathcal{I}} a_\alpha \phi_\alpha(x) \) equal to a constant is known as a minimal representation.
16.2. STANDARD EXAMPLES OF EXPONENTIAL FAMILIES

For a minimal representation, there is a unique parameter vector \( \theta \) associated with each distribution.

**Definition 16.4** (Over-complete). An over-complete representation is a representation that is not minimal; there is a linear combination of the elements of \( \Phi \) which yields a constant.

When the representation is over-complete, there exists an affine subset of parameter vectors \( \theta \), each associated with the same distribution.

Recall the definition of sufficiency, given in Definition 13.17. The following lemma is crucial. Its proof is left as an exercise.

**Lemma 16.5.** Let \( X = (X_1, \ldots, X_d) \) be a random vector with joint probability function

\[
p_X(x|\theta) = \exp\{\langle \theta, \Phi(x) \rangle - A(\theta) \} h(x), \quad x \in \mathcal{X}
\]

then \( \Phi(X) \), which will be denoted \( \Phi \), is a sufficient statistic for \( \theta \). If the representation is minimal, then \( \Phi(X) \) is a minimal sufficient statistic for \( \theta \).

**Proof** Exercise 1 page 346.

### 16.2 Standard Examples of Exponential Families

The purpose of this section is to take some basic distributions, which are well known, and illustrate that they satisfy the definition of an exponential family.

**Bernoulli** Consider the random variable \( X \), taking values 0 or 1, with probability function \( \mathbb{P}_X(1) = p, \mathbb{P}_X(0) = 1 - p \). This may be written as

\[
\mathbb{P}_X(x) = \begin{cases} 
p^x(1-p)^{1-x} & x \in \{0,1\} \\
0 & \text{other } x.
\end{cases}
\]

Then

\[
p_X(x) = \exp\left\{x \log \left( \frac{p}{1-p} \right) + \log(1 - p) \right\} = \exp\{x\theta + \log(1 - p)\} = \exp\{x\theta - \log(1 + e^\theta)\},
\]

where \( \theta = \log \left( \frac{p}{1-p} \right) \).

**Notation** Here, the quantity \( \theta \) denotes the canonical parameter.

In the language of exponential families, \( \mathcal{X} = \{0,1\}, \Phi = \{\phi\} \) where \( \phi(x) = x, h(0) = h(1) = 1, \)
\[ \mathbb{P}_X(0|\theta) = e^{-A(\theta)}, \quad \mathbb{P}_X(1|\theta) = e^{\theta - A(\theta)} \]

In other words

\[ \log \mathbb{P}_X(x|\theta) = \theta x - A(\theta), \]

which gives

\[ 1 = \mathbb{P}_X(0|\theta) + \mathbb{P}_X(1|\theta) = e^{-A(\theta)}(1 + e^{\theta}) \]

so that

\[ A(\theta) = \log(1 + \exp(\theta)). \]

**Gaussian** Recall that the one dimensional Gaussian density is of the form

\[ \pi(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}. \]

This may be expressed in terms of an exponential family as follows: \( \mathcal{X} = \mathbb{R}, h(x) = 1, \Phi = \{\phi_1, \phi_2\} \) where \( \phi_1(x) = x \) and \( \phi_2(x) = -x^2 \).

\[ \log \pi(x|\theta) = \theta_1 x - \theta_2 x^2 - A(\theta) \]

where

\[ 1 = e^{-A(\theta)} \int_{-\infty}^{\infty} e^{\theta_1 x - \theta_2 x^2} dx. \]

The partition function is therefore

\[ A(\theta) = \frac{1}{2} \log \pi - \frac{1}{2} \log \theta_2 + \frac{\theta_2^2}{4\theta_2} \]

and the parameter space is

\[ \Theta = \{(\theta_1, \theta_2) \in \mathbb{R}^2 | \theta_2 > 0\}. \]

Note that in the ‘usual’ notation

\[ \theta_1 = \frac{\mu}{\sigma^2}, \quad \theta_2 = \frac{1}{\sigma^2}. \]

**Exponential** Recall that an Exponential density is of the form

\[ \pi(x|\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases} \]

This is an exponential family, taking \( \mathcal{X} = (0, +\infty), h(x) = dx, \Phi = \phi, \) where \( \phi(x) = -x, \theta = \lambda, \) so that \( e^{-A(\theta)} = \theta, \) yielding \( A(\theta) = -\log \theta, \Theta = (0, +\infty). \)
16.3. GRAPHICAL MODELS AND EXPONENTIAL FAMILIES

**Poisson** Recall that the probability function \( p \) for a Poisson distribution with parameter \( \mu \) is given by
\[
\mathbb{P}(x|\mu) = \frac{\mu^x}{x!} e^{-\mu}, \quad x = 0, 1, 2, \ldots
\]
This is an exponential family with \( h(x) = \frac{1}{x!}, \theta = \log \mu \) so that \( \mathbb{P}(x|\mu) = P(x; \theta) h(x) \), where
\[
P(x; \theta) = e^{\theta x - e^\theta}.
\]
This gives \( A(\theta) = \exp(\theta) \). Since \( \mu \geq 0 \) and \( \theta = \log \mu \), it follows that \( \Theta = \mathbb{R} \).

**Beta** Recall that the probability density function for a Beta distribution is given by
\[
\pi(x|\alpha, \beta) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} & x \in [0,1] \\ 0 & \text{other } x. \end{cases}
\]
This is an exponential family, with \( X = (0,1), h \equiv 1, \alpha - 1 = \theta_1, \beta - 1 = \theta_2, \Phi = \{\phi_1, \phi_2\} \) where \( \phi_1(x) = \log x, \phi_2(x) = \log(1-x) \). Then
\[
\log \pi(x|\theta) = \theta_1 \log x + \theta_2 \log(1-x) - A(\theta),
\]
where the partition function \( A \) is given by
\[
A(\theta) = \log \Gamma(\theta_1 + 1) + \log \Gamma(\theta_2 + 1) - \log \Gamma(\theta_1 + \theta_2 + 2)
\]
and the parameter space is \( \Theta = (-1, \infty)^2 \).

16.3 Graphical Models and Exponential Families

The scalar examples described in section 16.2 serve as building blocks for the construction of exponential families, which have an underlying graphical structure.

**Example 16.6** (Sigmoid Belief Network Model).

The *sigmoid belief network model*, described below, was introduced by R. Neal (1992) [93]. It is an exponential family, with an underlying graphical structure.

Consider a directed acyclic graph \( G = (V, D) \), where \( V = \{X_1, \ldots, X_d\} \) is the set of variables, along which the probability distribution of \( X = (X_1, \ldots, X_d) \) may be factorised. Suppose that for each \( X_j \in V, j = 1, \ldots, d \), the random variable \( X_j \) takes values 0 or 1, each with probability 1/2. For any two components \( X_a \) and \( X_i \) of the random vector \( X \), component \( X_a \) has a direct causal effect on \( X_i \) only if \( \langle X_a, X_i \rangle \in D \).

The notation will be simplified in the following way: \( V \) and \( D \) will be used to denote the sets of nodes (variables) and directed edges respectively; the same notation will also be used to denote the indexing
sets of nodes and directed edges. In other words the notations

\[ V = \{1, \ldots, d\} \quad \text{and} \quad D = \{(s,t) \mid (X_s, X_t) \in D\} \]

will also be used. The meaning will be clear from the context. The probability distribution over the possible configurations is modelled by an exponential family with probability function \( \mathbb{P}_X(\theta) \) of the form

\[ \mathbb{P}_X(x | \theta) = \exp \left\{ \sum_{s=1}^d \theta_s x_s + \sum_{(s,t) \in D} \theta_{(s,t)} x_s x_t - A(\theta) \right\}. \]

The notation \( \text{Pa}_i \) denotes the parent set of node \( X_i \) and \( \pi_i(x) \) denotes the instantiation of \( \text{Pa}_i \) corresponding to the instantiation \( \{X = x\} \), this may be rewritten as

\[ \mathbb{P}_X(x | \theta) = \prod_{i=1}^d \mathbb{P}_{X_i \mid \text{Pa}_i}(x_i | \pi_i(x), \theta), \]

where (clearly)

\[ \mathbb{P}_{X_i \mid \text{Pa}_i}(x_i | \pi_i(x), \theta) = \frac{\exp \left\{ x_i \left( \theta_i + \sum_{j \in \pi_i(x)} \theta_{(ij)} x_j \right) \right\}}{1 + \exp \left\{ \theta_i + \sum_{j \in \pi_i(x)} \theta_{(ij)} x_j \right\}}, \]

where the notation \( x_j \in \pi_i(x) \) is clear. The index set is \( I = V \cup D \). The domain \( \Theta = \mathbb{R}^n \), where \( n = |I| \).

Since the sum that defines \( A(\theta) \) is finite for all \( \theta \in \mathbb{R}^n \), it follows that the family is regular. It is minimal, since there is no linear combination of the functions equal to a constant.

This model may be generalised. For example, one may consider higher order interactions. To include coupling of triples \( (X_s, X_t, X_u) \), one would add a monomial \( x_s x_t x_u \) with corresponding exponential parameter \( \theta_{(s,t,u)} \). More generally, the set \( C \) of indices of interacting variables may be considered, giving

\[ \mathbb{P}_X(x | \theta) = \exp \left\{ \sum_{C \in C} \theta_{(C)} \prod_{s \in C} X_s - A(\theta) \right\}. \]

\( \square \)

**Example 16.7** (Noisy 'or' as an Exponential Family).

The QMR - DT (Quick Medical Reference - Decision Theoretic) database is a large scale probabilistic data base that is intended to be used as a diagnostic aid in the domain of internal medicine. It is a bipartite graphical model; that is, a graphical model where the nodes may be of one of two types. The upper layer of nodes (the parents) represent diseases and the lower layer of nodes represent symptoms. There are approximately 600 disease nodes and 4000 symptom nodes in the database.

An evidence, or finding will be a set of observed symptoms, denoted by a vector of length 4000, each entry being a 1 or 0 depending upon whether or not the symptom is present or absent. This will be denoted \( \underline{d} \), which is an instantiation of the random vector \( \underline{F} \). The vector \( \underline{d} \) will be used to represents
the diseases; this is considered as an instantiation of the random vector $D$. Let $d_j$ denote component $j$ of vector $d$ and let $f_j$ denote component $j$ of vector $f$. Then, if the occurrence of various diseases are taken to be independent of each other, the following factorisation holds:

$$\prod_{j} \Pr(f_j|d_j) = \prod_{j} \Pr(f_j|d_j) \prod_{j} \Pr(d_j).$$

This may be represented by noisy ‘or’ model. Let $q_0$ denote the probability that symptom $i$ is present in the absence of any disease and $q_{ij}$ the probability that disease $j$ induces symptom $i$, then the probability that symptom $i$ is absent, given a vector of diseases $d$ is

$$p_{f_i|d}(0|d) = (1 - q_0) \prod_j (1 - q_{ij})^{d_j}.$$ 

The noisy or may then be rewritten in an exponential form:

$$p_{f_i|d}(0|d) = \exp \left\{ - \sum_j \theta_{ij} d_j - \theta_i \right\},$$

where $\theta_{ij} = \log (1 - q_{ij})$ are the transformed parameters.

### 16.4 Properties of the log Partition Function

Firstly, some basic properties of the log partition function $A(\theta)$ are discussed, which are then developed using convex analysis, discussed in [3]. Let $E_\theta[.]$ denote expectation with respect to $p(\cdot|\theta)$ for discrete variables, or $\pi(\cdot|\theta)$ for continuous variables. Of particular importance is the idea that the vector $\mu$, where $\mu_i := E_\theta[\phi_i(X)]$ provides an alternative parametrisation of the exponential family. Here expectation is defined as

$$E_\theta[f(X)] = \int f(x) \pi_X(x|\theta) dx$$

if $X$ is a continuous random vector and

$$E_\theta[f(X)] = \sum_{x \in X} f(x) \pi_X(x|\theta)$$

if $X$ is a discrete random vector. Recall that, for discrete variables,

$$A(\theta) = \log \sum_{x \in X} e^{\langle \theta, \phi(x) \rangle} h(x). \quad (16.3)$$

Provided expectations and variances exist, it follows that

$$\frac{\partial}{\partial \theta_\alpha} A(\theta) = \sum_{x \in X} e^{\langle \theta, \phi(x) \rangle} h(x) = E_\theta[\phi_\alpha(X)]. \quad (16.4)$$

Taking second derivatives yields
\[
\frac{\partial}{\partial \theta_{0,0}} A(\theta) = \mathbb{E}_\theta [\phi_{\theta_0}(X) \phi_{\theta_1}(X)] - \mathbb{E}_\theta [\phi_{\theta_0}(X) \mathbb{E}_\theta [\phi_{\theta_1}(X)]] = \text{Cov}_\theta (\phi_{\theta_0}(X), \phi_{\theta_1}(X)).
\]
It is and easy to show, and a standard fact, that any covariance matrix is non negative definite. It now follows that, on \( \Theta \), \( A \) is a convex function.

**Mapping to Mean Parameters** Given a vector of functions \( \Phi \), set \( F(\theta) = \mathbb{E}_\theta [\Phi(X)] \) and let \( \mathcal{M} = F(\Theta) \). For an arbitrary exponential family defined by

\[
\mathbb{P}_X(x|\theta) = \exp \{ \langle \theta, \Phi(x) \rangle - A(\theta) \} h(x),
\]

a mapping \( \Lambda : \Theta \to \mathcal{M} \) may be defined as follows:

\[
\Lambda(\theta) := \mathbb{E}_\theta [\Phi(X)].
\]

To each \( \theta \in \Theta \), the mapping \( \Lambda \) associates a vector of mean parameters \( \mu = \Lambda(\theta) \) belonging to the set \( \mathcal{M} \). Note that, by Equation (16.4),

\[
\Lambda(\theta) = \nabla A(\theta).
\]

The mapping \( \Lambda \) is one to one, and hence invertible on its image, when the representation is minimal. The image of \( \Theta \) is the interior of \( \mathcal{M} \).

**Example 16.8 (Bernoulli Trial).**

Consider a Bernoulli random variable \( X \) with state space \( \{0,1\} \). That is, \( p_X(0) = 1 - p \) and \( p_X(1) = p \). Now consider an Overcomplete exponential representation

\[
\mathbb{P}_X(x|\theta) = \exp \{ \theta_0 (1-x) + \theta_1 x - A(\theta_0, \theta_1) \}
\]

so that

\[
A(\theta_0, \theta_1) = \log \left( e^{\theta_0} + e^{\theta_1} \right).
\]

Here \( \Theta = \mathbb{R}^2 \). \( \phi_0(x) = 1 - x \) and \( \phi_1(x) = x \).

\[
\frac{\partial}{\partial \theta_0} A(\theta) = e^{\theta_0 - A(\theta_0, \theta_1)} = 1 - p = \mu_0
\]

\[
\frac{\partial}{\partial \theta_1} A(\theta) = e^{\theta_1 - A(\theta_0, \theta_1)} = p = \mu_1.
\]

The set \( \mathcal{M} \) of mean parameters is the simplex \( \{ (\mu_0, \mu_1) \in \mathbb{R}_+ \times \mathbb{R}_+ | \mu_0 + \mu_1 = 1 \} \). For any fixed \( \mu = (\mu_0, \mu_1) \) where \( \mu_0 \geq 0, \mu_1 \geq 0, \mu_0 + \mu_1 = 1 \), the inverse image is,

\[
\Lambda^{-1}(\mu) = \left\{ (\theta_0, \theta_1) \in \mathbb{R}^2 \left| \frac{e^{\theta_0}}{e^{\theta_0} + e^{\theta_1}} = \mu_0 \right. \right\}
\]
which may be rewritten as
\[ \Lambda^{-1}(\mu) = \left\{ (\theta_0, \theta_1) \in \mathbb{R}^2 \mid \theta_1 - \theta_0 = \log \frac{\mu_1}{\mu_0} \right\}. \]

In an over-parametrised, or over-complete representation, there is no longer a bijection between \( \Theta \) and \( \Lambda(\Theta) \). Instead, there is a bijection between elements of \( \Lambda(\Theta) \) and an affine subset of \( \Theta \). A pair \((\overline{\theta}, \mu)\) is said to be dualy coupled if \( \overline{\mu} = \Lambda(\overline{\theta}) \), and hence \( \theta \in \Lambda^{-1}(\mu) \).

### 16.5 Fenchel Legendre Conjugate

The *Fenchel Legendre conjugate* of the log partition function \( A \) is defined as follows:

\[ A^*(\mu) := \sup_{\overline{\theta} \in \overline{\Theta}} \{ \langle \mu, \overline{\theta} \rangle - A(\overline{\theta}) \}. \]  
(16.5)

The choice of notation is deliberately suggestive; the variables in the Fenchel Legendre dual turn out to have interpretation as the mean parameters. Recall the definition of \( P \) given by Equation (16.1); namely, if \( \mathbb{P}_X (x|\theta) \) is the probability function (or density function), then

\[ P(x; \theta) = \frac{\mathbb{P}_X (x|\theta)}{h(x)}. \]

**Definition 16.9** (Boltzmann - Shannon Entropy). *The Boltzmann - Shannon entropy* of \( \mathbb{P}_X (x|\theta) \) with respect to \( h \) is defined as

\[ H(\mathbb{P}_X (x|\theta)) = -\mathbb{E}_x [\log P(x; \theta)]. \]

The following is the main result of the chapter.

**Theorem 16.10.** For any \( \mu \in \mathcal{M} \), let \( \overline{\theta}(\mu) \in \Lambda^{-1}(\mu) \). Then

\[ A^*(\mu) = -H(\mathbb{P}_X (x|\overline{\theta}(\mu))). \]

In terms of this dual, for \( \overline{\theta} \in \Theta \), the log partition satisfies be expressed:

\[ A(\overline{\theta}) = \sup_{\mu \in \mathcal{M}} \{ \langle \overline{\theta}, \mu \rangle - A^*(\mu) \}. \]

(16.6)

**Proof of Theorem 16.10** From the definition \( \overline{\mu} = \mathbb{E}_\theta [\Phi(X)] \), it follows that

\[ -H(\mathbb{P}_X (x|\overline{\theta})) = \mathbb{E}_\theta [\log P(X; \overline{\theta})] = \mathbb{E}_\theta [(\overline{\theta}, \Phi(X))] - A(\overline{\theta}) = \langle \overline{\theta}, \overline{\mu} \rangle - A(\overline{\theta}). \]  
(16.7)

Consider the function

\[ F(\overline{\mu}, \overline{\theta}) = \langle \overline{\mu}, \overline{\theta} \rangle - A(\overline{\theta}). \]
Let $\theta(\mu)$ denote a value of $\theta$ that maximises $F(\mu, \theta)$ if such a value exists in $\Theta$. The result follows directly by using the definition given by Equation (16.5) together with Equation (16.7). Otherwise, let $\theta^{(n)}(\mu)$ denote a sequence such that $\lim_{n \to +\infty} F(\mu, \theta^{(n)}(\mu)) = A^*(\mu)$. The first statement of the theorem follows directly.

For the second part, choose $\theta \in \Theta$ and choose $\mu(\theta) = \nabla_{\theta} A(\theta)$. By definition of $\mathcal{M}$, note that $\mu(\theta) \in \mathcal{M}$. Since $A$ is convex, it follows that $\mu(\theta)$ maximises $\langle \theta, \mu \rangle - A(\theta)$, so that

$$A(\theta) = \langle \mu(\theta), \theta \rangle - A^*(\mu(\theta)).$$

But, from the definition of $A^*(\mu)$, it follows that for all $\mu \in \mathcal{M}$,

$$A(\theta) \geq \langle \mu, \theta \rangle - A^*(\mu).$$

From this,

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \mu, \theta \rangle - A^*(\mu) \}$$

and Theorem 16.10 is established. $\square$

The conjugate dual pair $(A, A^*)$ for the families of exponential variables given before are now given.

**Bernoulli** Recall that $A(\theta) = \log(1 + \exp(\theta))$ for $\theta \in \mathbb{R}$. It follows that

$$A^*(\mu) = \sup_{\theta \in \mathbb{R}} \{ \theta \mu - \log(1 + e^\theta) \}$$

The supremum is attained for $\theta(\mu)$ satisfying

$$\frac{\mu = e^{\theta(\mu)}}{1 + e^{\theta(\mu)}}.$$ 

It follows that

$$e^{\theta(\mu)} = \frac{\mu}{1 - \mu}$$

and

$$\theta(\mu) = \log \mu - \log(1 - \mu)$$

so that

$$A^*(\mu) = \mu \log \mu - \mu \log(1 - \mu) - \log(1 + \frac{\mu}{1 - \mu}),$$

which gives

$$A^*(\mu) = \mu \log \mu + (1 - \mu) \log(1 - \mu).$$
16.5. **Fenchel Legendre Conjugate**

**Gaussian**  Recall that \( \Theta = \{ (\theta_1, \theta_2) | \theta_2 > 0 \} \) and

\[
A(\theta) = \frac{1}{2} \log \pi - \frac{1}{2} \log \theta_2 + \frac{\theta_1^2}{4\theta_2}.
\]

\[
A^*(\mu) = \sup_{\theta \in \Theta} (\theta_1 \mu_1 + \theta_2 \mu_2 - \frac{1}{2} \log \pi + \frac{1}{2} \ln \theta_2 - \frac{\theta_1^2}{4\theta_2}).
\]

This is maximised when

\[
\begin{align*}
\mu_1 - \frac{\theta_1 (\mu)}{2 \theta^2 (\mu)} &= 0, \\
\mu_2 + \frac{1}{2 \theta^2 (\mu)} + \frac{\theta_2 (\mu)}{4 \theta^3 (\mu)} &= 0,
\end{align*}
\]

which gives

\[
\begin{align*}
\theta_2 (\mu_1, \mu_2) &= -\frac{1}{2(\mu_1^2 + \mu_2)} \\
\theta_1 (\mu) &= -\frac{\mu_1}{\mu_1^2 + \mu_2}
\end{align*}
\]

and

\[
A^*(\mu_1, \mu_2) = -\frac{1}{2} - \frac{1}{2} \log \pi - \frac{1}{2} \log (-2(\mu_1^2 + \mu_2)).
\]

Note that

\[
\mathcal{M} = \{ (\mu_1, \mu_2) | \mu_1^2 + \mu_2 < 0 \}.
\]

**Exponential Distribution**  Recall that \( \Theta = (0, +\infty) \) and that \( A(\theta) = -\log(\theta) \). By a straightforward computation,

\[
A^*(\mu) = -1 - \log(-\mu)
\]

and

\[
\mathcal{M} = (-\infty, 0).
\]

**Poisson Distribution**  Recall that \( \Theta = \mathbb{R} \) and that \( A(\theta) = \exp(\theta) \). It is a straightforward computation to see that

\[
A^*(\mu) = \mu \log \mu - \mu
\]

and that

\[
\mathcal{M} = (0, +\infty).
\]
16.6 Kullback-Leibler Divergence

Recall Definition 4.9, the Kullback-Leibler distance between two probability distributions \( p \in [0,1]^M \) and \( q \in [0,1]^M \)

\[
D_{KL}(q||p) = \sum_{j=1}^{M} q_j \ln \frac{q_j}{p_j}.
\]

This may be written as

\[
D_{KL}(q||p) = \mathbb{E}_q \left[ \ln \frac{q(X)}{p(X)} \right],
\]

where \( X \) is a random vector with state space \( X = (x_1, \ldots, x_M) \) and \( \mathbb{E}_q \) is expectation with respect to the measure such that \( q_j = \mathbb{P}(X = x_j) \). The definition of Kullback-Leibler may be extended to continuous distributions using Equation (16.8), where \( q \) and \( p \) denote the respective density functions. In this case, Equation (16.8) is taken as

\[
D_{KL}(q||p) = \int_{\mathbb{R}^M} q(x) \log \frac{q(x)}{p(x)} dx.
\]

When \( q \) and \( p \) are members of the same exponential family, the Kullback-Leibler distance may be computed in terms of the parameters. The key result, for expressing the distance in terms of the partition function, is the Fenchel’s inequality given in Equation (16.9), which can be seen directly from the definition of \( A^*(\mu) \).

\[
A(\theta) + A^*(\mu) \geq \langle \mu, \theta \rangle,
\]

with equality if and only if \( \mu = \Lambda(\theta) \) and \( \theta = \Lambda^{-1}(\mu) \). That is, for \( \mu = \Lambda(\theta) \) and \( \theta = \Lambda^{-1}(\mu) \),

\[
A(\theta) + A^*(\mu) = \langle \mu, \theta \rangle.
\]

Consider an exponential family of distributions, and consider two exponential parameter vectors, \( \theta_1 \in \Theta \) and \( \theta_2 \in \Theta \). When distributions are from the same exponential family, the notation \( D(\theta_1||\theta_2) \) is used to denote \( D_{KL}(p(\theta_1)||p(\theta_2)) \). Set \( \mu_\lambda = \Lambda(\theta_\lambda) \). Using the parameter to denote the distribution with respect to which the expectation is taken, note that

\[
D(\theta_1||\theta_2) = \mathbb{E}_{\mu_\lambda} \left[ \log \frac{p(X|\theta_1)}{p(X|\theta_2)} \right] = A(\theta_2) - A(\theta_1) - \langle \mu_1, \theta_2 - \theta_1 \rangle.
\]

The representation of the Kullback-Leibler divergence given in Equation (16.11) is known as the primal form of the KL divergence.

Taking \( \mu_1 = \Lambda(\theta_1) \) and applying Equation (16.10), the Kullback-Leibler distance may also be written

\[
D(\theta_1||\theta_2) = \tilde{D}(\mu_1||\mu_2) = A(\theta_2) + A^*(\mu_1) - \langle \mu_1, \theta_2 \rangle.
\]
16.7. MEAN FIELD THEORY

The representation given in Equation (16.12) is known as the mixed form of the KL divergence. Recall the definition of $A^*$ given by

$$A^*(\mu) := \sup_{\theta \in \Theta} \{ \langle \mu, \theta \rangle - A(\theta) \}$$

and recall Equation (16.6) from Theorem 16.10,

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \}.$$  

Equation (16.6) may be rewritten as

$$\inf_{\mu \in \mathcal{M}} \{ A(\theta) + A^*(\mu) - \langle \theta, \mu \rangle \} = 0.$$  

It follows that $\inf_{\mu \in \mathcal{M}} \tilde{D}(\mu|\theta) = 0.$

Finally, taking $\mu_2 = \Lambda(\theta_2)$ and applying Equation (16.10) once again to Equation (16.12) yields the so-called dual form of the KL divergence;

$$\tilde{D}(\mu_1 | \mu_2) = D(\theta_1 | \theta_2) = A^*(\mu_1) - A^*(\mu_2) - \langle \theta_2, \mu_1 - \mu_2 \rangle.$$  \hspace{1cm} (16.13)

16.7 Mean Field Theory

In this section, probability distributions of the form

$$P_X(x|\theta) = \exp \left\{ \sum_{\alpha} \theta_\alpha \phi_\alpha(x) - A(\theta) \right\} h(x)$$

are considered. Mean field theory techniques are discussed and it is shown how they may be used to obtain estimates of the log partition function $A(\theta)$. This is equivalent to the problem of finding an appropriate normalising constant to make a function into a probability density, a problem that often arises when updating using Bayes rule.

Mean Field Theory is based on the variational principle of Equation (16.6). The two fundamental difficulties associated with the variational problem are the nature of the constraint set $\mathcal{M}$ and the lack of an explicit form for the dual function $A^*$. Mean field theory entails limiting the optimization to a subset of distributions for which $A^*$ is relatively easy to characterise.

More specifically, the discussion of this chapter is restricted to the case where the functions $\phi_\alpha$ are either linear or quadratic. The problem therefore reduces to considering a graph $G = (V, U)$, where the node set $V$ denotes the variables and the edge set $U$ denotes a direct association between the variables. For this discussion, the edges in $U$ are assumed to be undirected. As usual, $V$ and $U$ denote the node (variable) and undirected edge sets; the same notation is used for the indexing sets. That is, with
minor abuse of notation (clear from the context), \( V \) and \( U \) are also used to mean: \( V = \{1, \ldots, d\} \) and \( U = \{(s,t)\vert (X_s, X_t) \in E\} \). Specifically, the probability distributions under consideration are of the form

\[
\mathbb{P}_X(x\mid \theta) = \exp \left\{ \sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{(s,t)} x_s x_t - A(\theta) \right\}.
\]

Let \( H \) denote a sub-graph of \( \mathcal{G} \) over which it is feasible to perform exact calculations. In an exponential formulation, the set of all distributions that respect the structure of \( H \) can be represented by a linear subspace of the exponential parameters. Let \( \mathcal{I}(H) \) denote the subset of indices associated with cliques in \( H \). Then the set of exponential parameters corresponding to distributions structured according to \( H \) is given by

\[
\mathcal{E}(H) := \{ \theta \in \Theta \mid \theta_\alpha = 0, \; \alpha \in \mathcal{I}(H) \}.
\]

The simplest example is to consider the completely disconnected graph \( H = (V, \emptyset) \). Then

\[
\mathcal{E}(H) = \{ \theta \in \Theta \mid \theta(s,t) = 0, \; (s,t) \in E \}.
\]

The associated distributions are of the product form

\[
\mathbb{P}_X(X \mid \theta) = \prod_{s \in V} \mathbb{P}_X(x_s \mid \theta_s).
\]

**Optimisation and Lower Bounds** Let \( \mathbb{P}_X(x\mid \theta) \) denote the target distribution that is to be approximated. The basis of mean field approximation is the following: any valid mean parameter specifies a lower bound on the log partition function, established using Jensen’s inequality.

**Proposition 16.11** (Mean Field Lower Bound).

\[
A(\theta) \geq \sup_{\mu \in \mathcal{M}} \left\{ (\theta, \mu) - A^*(\mu) \right\}
\]

**Proof** The proof is given for discrete variables; the proof for continuous variables is exactly the same, replacing the sum with an integral.

\[
A(\theta) = \log \sum_{x \in \mathcal{X}} \exp\{ (\theta, \Phi(x)) \} \\
= \log \sum_{x \in \mathcal{X}} \mathbb{P}_X(x\mid \theta) \exp\{ (\theta, \Phi(X)) - \log \mathbb{P}_X(x\mid \theta) \} \\
= \log \mathbb{E}_\theta [\exp\{ (\theta, \Phi(X)) - \log \mathbb{P}_X(X\mid \theta) \}]
\]

\[
\overset{(a)}{\geq} \quad (\theta, \mathbb{E}_\theta [\Phi(X)]) - \mathbb{E}_\theta [\log \mathbb{P}_X(X\mid \theta)] \\
= \quad (\theta, \mu) - A^*(\mu).
\]

The inequality \((a)\) follows from Jensen’s inequality; the last line follows from Theorem 16.10. \qed
There are difficulties in computing the lower bound in cases where there is not an explicit form for \( A^*(\mu) \). The mean field approach circumvents this difficulty by restricting to

\[
\mathcal{M}(G; H) := \{ \mu \in \mathbb{R}^d \mid \mu = \mathbb{E}_{\theta} [\Phi(X)] , \theta \in \mathcal{E}(H) \}.
\]

Note that \( \mathcal{M}(G; H) \subset \mathcal{M} \), hence

\[
A(\theta) \geq \sup_{\mu \in \mathcal{M}} \{ (\theta, \mu) - A^*(\mu) \} \geq \sup_{\mu \in \mathcal{M}(G; H)} \{ (\theta, \mu) - A^*(\mu) \}.
\]

This lower bound is the best that can be obtained by restricting to \( H \).

Let \( \mu^{(n)} \) denote a sequence such that for each \( n \), \( \mu^{(n)} \in \mathcal{M}(G, H) \), such that \( \mu^{(n)} \xrightarrow{n \to \infty} \mu \) and such that

\[
\langle (\theta, \mu^{(n)}) - A^*(\mu^{(n)}) \rangle \xrightarrow{n \to \infty} \sup_{\mu \in \mathcal{M}(G; H)} \{ (\theta, \mu) - A^*(\mu) \}.
\]

Note that \( \mu \in \mathcal{M}(G; H) \). Since \( \theta \in \Theta \), it follows that \( \mu \in \mathcal{M} \). The distribution associated with \( \mu \) minimises the Kullback-Leibler divergence between the approximating distribution and the target distribution, subject to the constraint that \( \mu \in \mathcal{M}(G; H) \). Recall the mixed form of the Kullback-Leibler divergence; namely, Equation (16.12).

\[
\hat{D}(\mu; \theta) = A(\theta) - A^*(\mu) - \langle \theta, \mu \rangle.
\]

**Naive Mean Field Updates** In the naive mean field approach, a fully factorised distribution is chosen. This is equivalent to the approximation obtained by taking an empty edge set to approximate the original distribution. The naive mean field updates are a set of recursions for finding a stationary point of the resulting optimisation problem.

**Example 16.12** (Sigmoid Network Model).

Let \( X = (X_1, \ldots, X_d) \) be a random vector with state space \( \mathcal{X} = \{0, 1\}^d \) (\( d \) binary variables). Suppose that the distribution may be factorised along an undirected graph \( G = (V, E) \). The probability function is given by

\[
\mathbb{P}_X(x|\theta) = \exp \left\{ \sum_{j=1}^{n} \theta_j x_j + \sum_{\langle i,j \rangle \in E} \theta_{i,j} x_i x_j - A(\theta) \right\}.
\]

The naive mean field approach involves considering the graph with no edges. In this restricted class,

\[
p_X(x|\theta) = \exp \left\{ \sum_{j=1}^{n} \theta_j x_j - A(\theta^{(H)}) \right\},
\]

where \( \theta^{(H)} \) is the collection of parameters \( \theta^{(H)}_s = \theta_s, s = 1, \ldots, d \) and \( \theta^{(H)}(s, t) = 0 \). Note that

\[
\mu_s = \mathbb{E}_{\theta} [\phi_s(X)] = \mathbb{E}_{\theta} [X_s]
\]
and

\[ \mu_{(s,t)} = \mathbb{E}_q[\phi_{s,t}(X)] = \mathbb{E}_q[X_sX_t]. \]

When \( \theta \in H \), it follows that \((X_s)_{s=1}^d\) are independent, so that

\[ \mu_{(s,t)} = \mathbb{E}_q[X_sX_t] = \mu_s\mu_t. \]

The optimisation is therefore restricted to the set of parameters

\[ \mathcal{M}(G; H) = \{ (\mu_s)_{s=1}^d, (\mu_{(s,t)})_{(s,t)\in\{1,\ldots,d\}^2} : 0 \leq \mu_s \leq 1, \mu_{(s,t)} = \mu_s\mu_t \}. \]

With the restriction to product form distributions, \((X_s)_{s=1}^d\) are independent Bernoulli variables and hence

\[ A^*_H(\mu) = \sum_{s=1}^d (\mu_s \log \mu_s + (1 - \mu_s) \log(1 - \mu_s)). \]

Set

\[ F(\mu; \theta) = \sum_{s=1}^d \theta_s \mu_s + \sum_{(s,t)\in U'} \theta_{(s,t)} \mu_s \mu_t - \sum_{s=1}^d (\mu_s \log \mu_s + (1 - \mu_s) \log(1 - \mu_s)), \]

then the lower bound is given by

\[ A(\theta) \geq \sup_{(\mu_s)_{s=1}^d \in [0,1]^d} F(\mu; \theta). \]

Note that, for each \( \mu_s \), the function \( F \) is strictly convex. It is easy to see that the maximum is attained when, for all \( 1 \leq s \leq t \), \((\mu_s)_{s=1}^d\) satisfies

\[ \theta_s + \sum_{t \neq s, t \in U} \theta_{(s,t)} \mu_t - \log \frac{\mu_s}{1 - \mu_s} = 0, \]

or

\[ \log \frac{\mu_s}{1 - \mu_s} = \theta_s + \sum_{t \in N(s)} \theta_{(s,t)} \mu_t. \]

Note that if

\[ \log \frac{y}{1-y} = x, \]

then

\[ y = \sigma(x), \]

where
16.7. MEAN FIELD THEORY

\[ \sigma(x) = \frac{1}{1 + e^{-x}}. \]

The algorithm then proceeds by setting

\[ \mu^{(j+1)}_s = \sigma \left( \theta_s + \sum_{l \in \mathcal{N}(s)} \theta_{(s,l)} \mu^{(j)}_l \right). \]

As discussed in [69] (page 222), the lower bound thus computed seems to provide a good approximation to the true value.

Notes

The material for Chapter 16 is taken mostly from Wainwright and Jordan [129]. It is developed further in [69]. Possible improvements to the lower bound are proposed by Humphreys and Titterington in [63]. The book by Barndorff-Nielsen [3] is the standard treatise of exponential families and the required convex analysis.