8 The Exponential Family

8.1 Introduction

In the previous chapters, we discussed several different representations of complex distributions. These included both representations of global structures (for example, Bayesian networks and Markov networks) and representations of local structures (for example, representations of CPDs and of potentials). In this chapter, we revisit these representations and view them from a different perspective. This view allows us to consider several basic questions and derive generic answers for these questions for a wide variety of representations. As we will see in later chapters, these solutions play a role in both inference and learning for the different representations we consider.

We note, however, that this chapter is somewhat abstract and heavily mathematical. Although the ideas described in this chapter are of central importance to understanding the theoretical foundations of learning and inference, the algorithms themselves can be understood even without the material presented in this chapter. Thus, this chapter can be skipped by readers who are interested primarily in the algorithms themselves.

8.2 Exponential Families

parametric family

Our discussion so far has focused on the representation of a single distribution (using, say, a Bayesian or Markov network). We now consider *families of distributions*. Intuitively, a family is a set of distributions that all share the same parametric form and differ only in choice of particular parameters (for example, the entries in table-CPDs). In general, once we choose the global structure and local structure of the network, we define a family of all distributions that can be attained by different parameters for this specific choice of CPDs.

Example 8.1

Consider the empty graph structure $\mathcal{G}_\emptyset$ over the variables $\mathcal{X} = \{X_1, \ldots, X_n\}$. We can define the family $\mathcal{P}_\emptyset$ to be the set of distributions that are consistent with $\mathcal{G}_\emptyset$. If all the variables in $\mathcal{X}$ are binary, then we can specify a particular distribution in the family by using n parameters, $\boldsymbol{\theta} = \{P(x_i^1) : i = 1, \ldots, n\}$. ■

We will be interested in families that can be written in a particular form.

Definition 8.1

exponential family

Let $\mathcal{X}$ be a set of variables. An exponential family *$\mathcal{P}$ over $\mathcal{X}$ is specified by four components:*

sufficient statistic function

parameter space

legal parameter

natural parameter

- *A* sufficient statistics function τ *from assignments to* $\mathcal{X}$ *to* $\mathcal{R}^K$.
- *A* parameter space *that is a convex set* $\Theta \subseteq \mathcal{R}^M$ *of* legal parameters.
- *A* natural parameter *function* $\mathtt{t}$ *from* $\mathcal{R}^M$ *to* R^K.
- *An auxiliary measure* A *over* $\mathcal{X}$.

Each vector of parameters $\boldsymbol{\theta} \in \Theta$ *specifies a distribution* $P_{\boldsymbol{\theta}}$ *in the family as*

$$P_{\boldsymbol{\theta}}(\xi) = \frac{1}{Z(\boldsymbol{\theta})} A(\xi) \exp\{\langle \mathtt{t}(\boldsymbol{\theta}), \tau(\xi) \rangle\} \tag{8.1}$$

where $\langle \mathtt{t}(\boldsymbol{\theta}), \tau(\xi) \rangle$ *is the inner product of the vectors* $\mathtt{t}(\boldsymbol{\theta})$ *and* $\tau(\xi)$*, and*

$$Z(\boldsymbol{\theta}) = \sum_{\xi} A(\xi) \exp\{\langle \mathtt{t}(\boldsymbol{\theta}), \tau(\xi) \rangle\}$$

partition function

is the partition function *of* $\mathcal{P}$*, which must be finite. The parametric family* $\mathcal{P}$ *is defined as:*

$$\mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}.$$ ■

We see that an exponential family is a concise representation of a class of probability distributions that share a similar functional form. A member of the family is determined by the parameter vector $\boldsymbol{\theta}$ in the set of legal parameters. The sufficient statistic function τ summarizes the aspects of an instance that are relevant for assigning it a probability. The function $\mathtt{t}$ maps the parameters to space of the sufficient statistics.

The measure A assigns additional preferences among instances that do not depend on the parameters. However, in most of the examples we consider here A is a constant, and we will mention it explicitly only when it is not a constant.

Although this definition seems quite abstract, many distributions we already have encountered are exponential families.

Example 8.2

Consider a simple Bernoulli distribution. In this case, the distribution over a binary outcome (such as a coin toss) is controlled by a single parameter θ *that represents the probability of* x^1*. To show that this distribution is in the exponential family, we can set*

$$\tau(X) = \langle \mathbf{1}\{X = x^1\}, \mathbf{1}\{X = x^0\} \rangle, \tag{8.2}$$

a numerical vector representation of the value of X*, and*

$$\mathtt{t}(\theta) = \langle \ln \theta, \ln(1-\theta) \rangle. \tag{8.3}$$

It is easy to see that for $X = x^1$*, we have* $\tau(X) = \langle 1, 0 \rangle$*, and thus*

$$\exp\{\langle \mathtt{t}(\theta), \tau(X) \rangle\} = e^{1 \cdot \ln \theta + 0 \cdot \ln(1-\theta)} = \theta.$$

Similarly, for $X = x^0$*, we get that* $\exp\{\langle \mathtt{t}(\theta), \tau(X) \rangle\} = 1 - \theta$*. We conclude that, by setting* $Z(\theta) = 1$*, this representation is identical to the Bernoulli distribution.* ■

Example 8.3

Consider a Gaussian distribution over a single variable. Recall that

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

Define

$$\tau(x) = \langle x, x^2 \rangle \tag{8.4}$$

$$\mathtt{t}(\mu, \sigma^2) = \langle \frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2} \rangle \tag{8.5}$$

$$Z(\mu, \sigma^2) = \sqrt{2\pi}\sigma \exp\left\{\frac{\mu^2}{2\sigma^2}\right\}. \tag{8.6}$$

We can easily verify that

$$P(x) = \frac{1}{Z(\mu, \sigma^2)} \exp\left\{\langle \mathtt{t}(\theta), \tau(X) \rangle\right\}.$$

■

In fact, most of the parameterized distributions we encounter in probability textbooks can be represented as exponential families. This includes the Poisson distributions, exponential distributions, geometric distributions, Gamma distributions, and many others (see, for example, exercise 8.1).

We can often construct multiple exponential families that encode precisely the same class of distributions. There are, however, desiderata that we want from our representation of a class of distributions as an exponential family. First, we want the parameter space Θ to be "well-behaved," in particular, to be a convex, open subset of $\mathcal{R}^M$. Second, we want the parametric family to be *nonredundant* — to have each choice of parameters represent a unique distribution. More precisely, we want $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$ to imply $P_{\boldsymbol{\theta}} \neq P_{\boldsymbol{\theta}'}$. It is easy check that a family is nonredundant if and only if the function $\mathtt{t}$ is invertible (over the set Θ). Such exponential families are called *invertible*. As we will discuss, these desiderata help us execute certain operations effectively, in particular, finding a distribution Q in some exponential family that is a "good approximation" to some other distribution P.

nonredundant parameterization

invertible exponential family

8.2.1 Linear Exponential Families

A special class of exponential families is made up of families where the function $\mathtt{t}$ is the identity function. This implies that the parameters are the same dimension K as the representation of the data. Such parameters are also called the *natural parameters* for the given sufficient statistic function. The name reflects that these parameters do not need to be modified in the exponential form. When using natural parameters, equation (8.1) simplifies to

natural parameter

$$P_{\boldsymbol{\theta}}(\xi) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left\{\langle \boldsymbol{\theta}, \tau(\xi) \rangle\right\}.$$

Clearly, for any given sufficient statistics function, we can reparameterize the exponential family using the natural parameters. However, as we discussed earlier, we want the space of parameters Θ to satisfy certain desiderata, which may not hold for the space of natural

parameters. In fact, for the case of linear exponential families, we want to strengthen our desiderata, and require that any parameter vector in $\mathcal{R}^K$ defines a distribution in the family. Unfortunately, as stated, this desideratum is not always achievable. To understand why, recall that the definition of a legal parameter space Θ requires that each parameter vector $\boldsymbol{\theta} \in \Theta$ give rise to a legal (normalizable) distribution $P_{\boldsymbol{\theta}}$. These normalization requirements can impose constraints on the space of legal parameters.

Example 8.4

Consider again the Gaussian distribution. Suppose we define a new parameter space using the definition of t. *That is let* $\boldsymbol{\eta} = \mathtt{t}(\mu, \sigma^2) = \langle \frac{2\mu}{2\sigma^2}, -\frac{1}{2\sigma^2} \rangle$ *be the natural parameters that corresponds to* $\boldsymbol{\theta} = \langle \mu, \sigma^2 \rangle$. *Clearly, we can now write*

$$P_{\boldsymbol{\eta}}(x) \propto \exp\{\langle \eta, \tau(x) \rangle\}.$$

However, not every choice of $\boldsymbol{\eta}$ *would lead to a legal distribution. For the distribution to be normalized, we need to be able to compute*

$$\begin{aligned} Z(\boldsymbol{\eta}) &= \int \exp\{\langle \boldsymbol{\eta}, \tau(x) \rangle\}\, dx \\ &= \int_{-\infty}^{\infty} \exp\left\{\eta_1 x + \eta_2 x^2\right\} dx. \end{aligned}$$

If $\eta_2 \geq 0$ *this integral is undefined, since the function grows when* x *approaches* ∞ *and* $-\infty$. *When* $\eta_2 < 0$, *the integral has a finite value. Fortunately, if we consider* $\boldsymbol{\eta} = \mathtt{t}(\mu, \sigma^2)$ *of equation (8.5), we see that the second component is always negative (since* $\sigma^2 > 0$*). In fact, we can see that the* image *of the original parameter space,* $\langle \mu, \sigma^2 \rangle \in \mathcal{R} \times \mathcal{R}^+$, *through the function* $\mathtt{t}(\mu, \sigma^2)$, *is the space* $\mathcal{R} \times \mathcal{R}^-$. *We can verify that, for every* $\boldsymbol{\eta}$ *in that space, the normalization constant is well defined.* ■

natural parameter space

More generally, when we consider natural parameters for a sufficient statistics function τ, we define the set of allowable natural parameters, the *natural parameter space*, to be the set of natural parameters that can be normalized

$$\Theta = \left\{ \boldsymbol{\theta} \in \mathcal{R}^K : \int \exp\{\langle \boldsymbol{\theta}, \tau(\xi) \rangle\}\, d\xi < \infty \right\}.$$

In the case of distributions over finite discrete spaces, all parameter choices lead to normalizable distributions, and so $\Theta = \mathcal{R}^K$. In other examples, such as the Gaussian distribution, the natural parameter space can be more constrained. An exponential family over the natural parameter space, and for which the natural parameter space is open and convex, is called a *linear exponential family*.

linear exponential family

The use of linear exponential families significantly simplifies the definition of a family. To specify such a family, we need to define only the function τ; all other parts of the definition are implicit based on this function. This gives us a tool to describe distributions in a concise manner. As we will see, linear exponential families have several additional attractive properties.

Where do find linear exponential families? The two examples we presented earlier were not phrased as linear exponential families. However, as we saw in example 8.4, we may be able to provide an alternative parameterization of a nonlinear exponential family as a linear exponential family. This example may give rise to the impression that any family can be reparameterized in a trivial manner. However, there are more subtle situations.

Example 8.5

Consider the Bernoulli distribution. Again, we might reparameterize θ by $\mathtt{t}(\theta)$. However, the image of the function $\mathtt{t}$ of example 8.2 is the curve $\langle \ln\theta, \ln(1-\theta)\rangle$. This curve is not a convex set, and it is clearly a subspace of the natural parameter space.

Alternatively, we might consider using the entire natural parameter space $\mathcal{R}^2$, corresponding to the sufficient statistic function $\tau(X) = \langle \mathbf{1}\{X = x^1\}, \mathbf{1}\{X = x^0\}\rangle$ of equation (8.2). This gives rise to the parametric form:

$$P_{\boldsymbol{\theta}}(x) \propto \exp\{\langle \boldsymbol{\theta}, \tau(x)\rangle\} = \exp\left\{\theta_1 \mathbf{1}\{X = x^1\} + \theta_2 \mathbf{1}\{X = x^0\}\right\}.$$

Because the probability space is finite, this form does define a distribution for every choice of $\langle \theta_1, \theta_2\rangle$. However, it is not difficult to verify that this family is redundant: for every constant c, the parameters $\langle \theta_1 + c, \theta_2 + c\rangle$ define the same distribution as $\langle \theta_1, \theta_2\rangle$.

Thus, a two-dimensional space is overparameterized for this distribution; conversely, the one-dimensional subspace defined by the natural parameter function is not well behaved. The solution is to use an alternative representation of a one-dimensional space. Since we have a redundancy, we may as well clamp θ_2 to be 0. This results in the following representation of the Bernoulli distribution:

$$\begin{aligned} \tau(x) &= \mathbf{1}\{x = x^1\} \\ \mathtt{t}(\theta) &= \ln\frac{\theta}{1-\theta}. \end{aligned}$$

We see that

$$\begin{aligned} \exp\left\{\langle \mathtt{t}(\theta), \tau(x^1)\rangle\right\} &= \frac{\theta}{1-\theta} \\ \exp\left\{\langle \mathtt{t}(\theta), \tau(x^0)\rangle\right\} &= 1. \end{aligned}$$

Thus,

$$Z(\theta) = 1 + \frac{\theta}{1-\theta} = \frac{1}{1-\theta}.$$

Using these, we can verify that

$$P_\theta(x^1) = (1-\theta)\frac{\theta}{1-\theta} = \theta.$$

We conclude that this exponential representation captures the Bernoulli distribution. Notice now that, in the new representation, the image of $\mathtt{t}$ is the whole real line $\mathcal{R}$. Thus, we can define a linear exponential family with this sufficient statistic function. ■

Example 8.6

Now, consider a multinomial variable X with k values $x^1, \ldots, x^k$. The situation here is similar to the one we had with the Bernoulli distribution. If we use the simplest exponential representation, we find that the legal natural parameters are on a curved manifold of $\mathcal{R}^k$. Thus, instead we define the sufficient statistic as a function from values of x to $\mathcal{R}^{k-1}$:

$$\tau(x) = \langle \mathbf{1}\{x = x^2\}, \ldots, \mathbf{1}\{x = x^k\}\rangle.$$

Using a similar argument as with the Bernoulli distribution, we see that if we define

$$\mathtt{t}(\boldsymbol{\theta}) = \langle \ln \frac{\theta_2}{\theta_1}, \ldots, \ln \frac{\theta_k}{\theta_1} \rangle,$$

then we reconstruct the original multinomial distribution. It is also easy to check that the image of $\mathtt{t}$ *is* $\mathcal{R}^{k-1}$*. Thus, by reparameterizing, we get a linear exponential family.* ■

All these examples define linear exponential families. An immediate question is whether there exist families that are not linear. As we will see, there are such cases. However, the examples we present require additional machinery.

8.3 Factored Exponential Families

The two examples of exponential families so far were of univariate distributions. Clearly, we can extend the notion to multivariate distributions as well. In fact, we have already seen one such example. Recall that, in definition 4.15, we defined log-linear models as distributions of the form:

$$P(X_1, \ldots, X_n) \propto \exp\left\{ \sum_{i=1}^{k} \theta_i \cdot f_i(\boldsymbol{D}_i) \right\}$$

where each feature f_i is a function whose scope is $\boldsymbol{D}_i$. Such a distribution is clearly a linear exponential family where the sufficient statistics are the vector of features

$$\tau(\xi) = \langle f_1(\boldsymbol{d}_1), \ldots, f_k(\boldsymbol{d}_k) \rangle.$$

As we have shown, by choosing the appropriate features, we can devise a log-linear model to represent a given discrete Markov network structure. This suffices to show that discrete Markov networks are linear exponential families.

8.3.1 Product Distributions

What about other distributions with product forms? Initially the issues seem deceptively easy. A product form of terms corresponds to a simple composition of exponential families

Definition 8.2 exponential factor family

An (unnormalized) exponential factor family Φ *is defined by* τ, $\mathtt{t}$, A, *and* Θ *(as in the exponential family). A factor in this family is*

$$\phi_{\boldsymbol{\theta}}(\xi) = A(\xi) \exp\left\{ \langle \mathtt{t}(\boldsymbol{\theta}), \tau(\xi) \rangle \right\}.$$

■

Definition 8.3 family composition

Let $\Phi_1, \ldots, \Phi_k$ *be exponential factor families, where each* Φ_i *is specified by* τ_i, $\mathtt{t}_i$, A_i, *and* Θ_i. *The* composition *of* $\Phi_1, \ldots, \Phi_k$ *is the family* $\Phi_1 \times \Phi_2 \times \cdots \times \Phi_k$ *parameterized by* $\boldsymbol{\theta} = \boldsymbol{\theta}_1 \circ \boldsymbol{\theta}_2 \circ \cdots \circ \boldsymbol{\theta}_k \in \Theta_1 \times \Theta_2 \times \cdots \times \Theta_k$*, defined as*

$$P_{\boldsymbol{\theta}}(\xi) \propto \prod_i \phi_{\boldsymbol{\theta}_i}(\xi) = \left(\prod_i A_i(\xi) \right) \exp\left\{ \sum_i \langle \mathtt{t}_i(\boldsymbol{\theta}_i), \tau_i(\xi) \rangle \right\}$$

where $\phi_{\boldsymbol{\theta}_i}$ *is a factor in the* i*'th factor family.* ■

It is clear from this definition that the composition of exponential factors is an exponential family with $\tau(\xi) = \tau_1(\xi) \circ \tau_2(\xi) \circ \cdots \circ \tau_k(\xi)$ and natural parameters $\mathtt{t}(\boldsymbol{\theta}) = \mathtt{t}_1(\boldsymbol{\theta}_1) \circ \mathtt{t}_2(\boldsymbol{\theta}_2) \circ \cdots \circ \mathtt{t}_k(\boldsymbol{\theta}_k)$.

This simple observation suffices to show that if we have exponential representation for potentials in a Markov network (not necessarily simple potentials), then their product is also an exponential family. Moreover, it follows that the product of linear exponential factor families is a linear exponential family.

8.3.2 Bayesian Networks

Taking the same line of reasoning, we can also show that, if we have a set of CPDs from an exponential family, then their product is also in the exponential family. Thus, we can conclude that a Bayesian network with exponential CPDs defines an exponential family. To show this, we first note that many of the CPDs we saw in previous chapters can be represented as exponential factors.

Example 8.7

We start by examining a simple table-CPD $P(X \mid \boldsymbol{U})$. Similar to the case of Bernoulli distribution, we can define the sufficient statistics to be indicators for different entries in $P(X \mid \boldsymbol{U})$. Thus, we set

$$\tau_{P(X|\boldsymbol{U})}(\mathcal{X}) = \langle \mathbb{I}\{X = x, \boldsymbol{U} = \boldsymbol{u}\} : x \in Val(X), \boldsymbol{u} \in Val(\boldsymbol{U}) \rangle.$$

We set the natural parameters to be the corresponding parameters

$$\mathtt{t}_{P(X|\boldsymbol{U})}(\boldsymbol{\theta}) = \langle \ln P(x \mid \boldsymbol{u}) : x \in Val(X), \boldsymbol{u} \in Val(\boldsymbol{U}) \rangle.$$

It is easy to verify that

$$P(x \mid \boldsymbol{u}) = \exp\left\{ \langle \mathtt{t}_{P(X|\boldsymbol{U})}(\boldsymbol{\theta}), \tau_{P(X|\boldsymbol{U})}(x, \boldsymbol{u}) \rangle \right\},$$

since exactly one entry of $\tau_{P(X|\boldsymbol{U})}(x, \boldsymbol{u})$ is 1 and the rest are 0. Note that this representation is not a linear exponential factor. ■

Clearly, we can use the same representation to capture any CPD for discrete variables. In some cases, however, we can be more efficient. In tree-CPDs, for example, we can have a feature set for each leaf in tree, since all parent assignment that reach the leaf lead to the same parameter over the children.

What happens with continuous CPDs? In this case, not every CPD can be represented by an exponential factor. However, some cases can.

Example 8.8

Consider a linear Gaussian CPD for $P(X \mid \boldsymbol{U})$ where

$$X = \beta_0 + \beta_1 u_1 + \cdots + \beta_k u_k + \epsilon,$$

where ϵ is a Gaussian random variable with mean 0 and variance σ^2, representing the noise in the system. Stated differently, the conditional density function of X is

$$P(x \mid \boldsymbol{u}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2}(x - (\beta_0 + \beta_1 u_1 + \cdots + \beta_k u_k))^2 \right\}.$$

By expanding the squared term, we find that the sufficient statistics are the first and second moments of all the variables

$$\tau_{P(X|U)}(\mathcal{X}) = \langle 1, x, u_1, \ldots, u_k, x^2, xu_1, \ldots, xu_k, u_1^2, u_1u_2, \ldots, u_k^2 \rangle,$$

and the natural parameters are the coefficients of each of these terms. ■

As the product of exponential factors is an exponential family, we conclude that a Bayesian network that is the product of CPDs that have exponential form defines an exponential family.

However, there is one subtlety that arises in the case of Bayesian networks that does not arise for a general product form. When we defined the product of a set of exponential factors in definition 8.3, we ignored the partition functions of the individual factors, allowing the partition function of the overall distribution to ensure global normalization.

However, in both of our examples of exponential factors for CPDs, we were careful to construct a normalized conditional distribution. This allows us to use the chain rule to compose these factors into a joint distribution without the requirement of a partition function. This requirement turns out to be critical: We cannot construct a Bayesian network from a product of unnormalized exponential factors.

Example 8.9

Consider the network structure $A \rightarrow B$, with binary variables. Now, suppose we want to represent the CPD $P(B \mid A)$ using a more concise representation than the one of example 8.7. As suggested by example 8.5, we might consider defining

$$\tau(A, B) = \langle \mathbf{1}\{A = a^1\}, \mathbf{1}\{B = b^1, A = a^1\}, \mathbf{1}\{B = b^1, A = a^0\} \rangle.$$

That is, for each conditional distribution, we have an indicator only for one of the two relevant cases. The representation of example 8.5 suggests that we should define

$$\mathsf{t}(\boldsymbol{\theta}) = \left\langle \ln \frac{\theta_{a^1}}{\theta_{a^0}}, \ln \frac{\theta_{b^1|a^1}}{\theta_{b^0|a^1}}, \ln \frac{\theta_{b^1|a^0}}{\theta_{b^0|a^0}} \right\rangle.$$

Does this construction give us the desired distribution? Under this construction, we would have

$$P_{\boldsymbol{\theta}}(a^1, b^1) = \frac{1}{Z(\boldsymbol{\theta})} \frac{\theta_{a^1}\theta_{b^1|a^1}}{\theta_{a^0}\theta_{b^0|a^1}}.$$

Thus, if this representation was faithful for the intended interpretation of the parameter values, we would have $Z(\boldsymbol{\theta}) = \frac{1}{\theta_{a^0}\theta_{b^0|a^1}}$. On the other hand,

$$P_{\boldsymbol{\theta}}(a^0, b^0) = \frac{1}{Z(\boldsymbol{\theta})},$$

which requires that $Z(\boldsymbol{\theta}) = \frac{1}{\theta_{a^0}\theta_{b^0|a^0}}$ in order to be faithful to the desired distribution. Because these two constants are, in general, not equal, we conclude that this representation cannot be faithful to the original Bayesian network. ■

The failure in this example is that the global normalization constant cannot play the role of a local normalization constant within each conditional distribution. This implies that to have an exponential representation of a Bayesian network, we need to ensure that each CPD is locally

normalized. For every exponential CPD this is easy to do. We simply increase the dimension of τ by adding another dimension that has a constant value, say 1. Then the matching element of $\mathtt{t}(\boldsymbol{\theta})$ can be the logarithm of the partition function. This is essentially what we did in example 8.8.

We still might wonder whether a Bayesian network defines a linear exponential family.

Example 8.10

Consider the network structure $A \rightarrow C \leftarrow B$, with binary variables. Assuming a representation that captures general CPDs, our sufficient statistics need to include features that distinguish between the following four assignments:

$$\begin{aligned}\xi_1 &= \langle a^1, b^1, c^1 \rangle \\ \xi_2 &= \langle a^1, b^0, c^1 \rangle \\ \xi_3 &= \langle a^0, b^1, c^1 \rangle \\ \xi_4 &= \langle a^0, b^0, c^1 \rangle\end{aligned}$$

More precisely, we need to be able to modify the CPD $P(C \mid A, B)$ to change the probability of one of these assignments without modifying the probability of the other three. This implies that $\tau(\xi_1), \ldots, \tau(\xi_4)$ must be linearly independent: otherwise, we could not change the probability of one assignment without changing the others. Because our model is a linear function of the sufficient statistics, we can choose any set of orthogonal basis vectors that we want; in particular, we can assume without loss of generality that the first four coordinates of the sufficient statistics are $\tau_i(\xi) = \mathbf{1}\{\xi = \xi_i\}$, and that any additional coordinates of the sufficient statistics are not linearly dependent on these four. Moreover, since the model is over a finite set of events, any choice of parameters can be normalized. Thus, the space of natural parameters is $\mathcal{R}^K$, where K is dimension of the sufficient statistics vector. The linear family over such features is essentially a Markov network over the clique $\{A, B, C\}$. Thus, the parameterization of this family includes cases where A and B are not independent, violating the independence properties of the Bayesian network. ■

Thus, this simple Bayesian network cannot be represented by a linear family. **More broadly, although a Bayesian network with suitable CPDs defines an exponential family, this family is not generally a linear one. In particular, any network that contains immoralities does not induce a linear exponential family.**

8.4 Entropy and Relative Entropy

We now explore some of the consequences of representation of models in factored form and of their exponential family representation. These both suggest some implications of these representations and will be useful in developments in subsequent chapters.

8.4.1 Entropy

We start with the notion of entropy. Recall that the entropy of a distribution is a measure of the amount of "stochasticity" or "noise" in the distribution. A low entropy implies that most of the distribution mass is on a few instances, while a larger entropy suggests a more uniform distribution. Another interpretation we discussed in appendix A.1 is the number of bits needed, on average, to encode instances in the distribution.

In various tasks we need to compute the entropy of given distributions. As we will see, we also encounter situations where we want to choose a distribution that maximizes the entropy subject to some constraints. A characterization of entropy will allow us to perform both tasks more efficiently.

8.4.1.1 Entropy of an Exponential Model

We now consider the task of computing the entropy for distributions in an an exponential family defined by τ and $\mathtt{t}$.

Theorem 8.1 *Let $P_{\boldsymbol{\theta}}$ be a distribution in an exponential family defined by the functions τ and $\mathtt{t}$. Then*

$$\boldsymbol{H}_{P_{\boldsymbol{\theta}}}(\mathcal{X}) = \ln Z(\boldsymbol{\theta}) - \langle \boldsymbol{E}_{P_{\boldsymbol{\theta}}}[\tau(\mathcal{X})], \mathtt{t}(\boldsymbol{\theta}) \rangle. \tag{8.7}$$

While this formulation seems fairly abstract, it does provide some insight. The entropy decomposes as a difference of two terms. The first is the partition function $Z(\boldsymbol{\theta})$. The second depends only on the *expected value* of the sufficient statistics $\tau(\mathcal{X})$. Thus, instead of considering each assignment to $\mathcal{X}$, we need to know only the expectations of the statistics under $P_{\boldsymbol{\theta}}$. As we will see, this is a recurring theme in our discussion of exponential families.

Example 8.11 *We now apply this result to a Gaussian distribution $X \sim N(\mu, \sigma^2)$, as formulated in the exponential family in example 8.3. Plugging into equation (8.7) the definitions of τ, $\mathtt{t}$, and Z from equation (8.4), equation (8.5), and equation (8.6), respectively, we get*

$$\begin{aligned}
\boldsymbol{H}_P(X) &= \frac{1}{2}\ln 2\pi\sigma^2 + \frac{\mu^2}{2\sigma^2} - \frac{2\mu}{2\sigma^2}\boldsymbol{E}_P[X] + \frac{1}{2\sigma^2}\boldsymbol{E}_P\left[X^2\right] \\
&= \frac{1}{2}\ln 2\pi\sigma^2 + \frac{\mu^2}{2\sigma^2} - \frac{2\mu}{2\sigma^2}\mu + \frac{1}{2\sigma^2}(\sigma^2 + \mu^2) \\
&= \frac{1}{2}\ln 2\pi e\sigma^2
\end{aligned}$$

where we used the fact that $\boldsymbol{E}_P[X] = \mu$ and $\boldsymbol{E}_P\left[X^2\right] = \mu^2 + \sigma^2$. ■

We can apply the formulation of theorem 8.1 directly to write the entropy of a Markov network.

Proposition 8.1 *If $P(\mathcal{X}) = \frac{1}{Z}\prod_k \phi_k(\boldsymbol{D}_k)$ is a Markov network, then*

$$\boldsymbol{H}_P(\mathcal{X}) = \ln Z + \sum_k \boldsymbol{E}_P[-\ln \phi_k(\boldsymbol{D}_k)].$$

Example 8.12 *Consider a simple Markov network with two potentials $\beta_1(A, B)$ and $\beta_2(B, C)$, so that*

		$\beta_1(A,B)$
a^0	b^0	2
a^0	b^1	1
a^1	b^0	1
a^1	b^1	5

		$\beta_2(B,C)$
b^0	c^0	6
b^0	c^1	1
b^1	c^0	1
b^1	c^1	0.5

Simple calculations show that $Z = 30$, and the marginal distributions are

A	B	$P(A,B)$
a^0	b^0	0.47
a^0	b^1	0.05
a^1	b^0	0.23
a^1	b^1	0.25

B	C	$P(B,C)$
b^0	c^0	0.6
b^0	c^1	0.1
b^1	c^0	0.2
b^1	c^1	0.1

Using proposition 8.1, we can calculate the entropy:

$$\begin{aligned}
\boldsymbol{H}_P(A,B,C) &= \ln Z + \boldsymbol{E}_P[-\ln \beta_1(A,B)] + \boldsymbol{E}_P[-\ln \beta_2(B,C)] \\
&= \ln Z \\
&\quad -P(a^0,b^0)\ln\beta_1(a^0,b^0) - P(a^0,b^1)\ln\beta_1(a^0,b^1) \\
&\quad -P(a^1,b^0)\ln\beta_1(a^1,b^0) - P(a^1,b^1)\ln\beta_1(a^1,b^1) \\
&\quad -P(b^0,c^0)\ln\beta_2(b^0,c^0) - P(b^0,c^1)\ln\beta_2(b^0,c^1) \\
&\quad -P(b^1,c^0)\ln\beta_2(b^1,c^0) - P(b^1,c^1)\ln\beta_2(b^1,c^1) \\
&= 3.4012 \\
&\quad -0.47 * 0.69 - 0.05 * 0 - 0.23 * 0 - 0.25 * 1.60 \\
&\quad -0.6 * 1.79 - 0.1 * 0 - 0.2 * 0 - 0.1 * -0.69 \\
&= 1.670.
\end{aligned}$$

■

In this example, the number of terms we evaluated is the same as what we would have considered using the original formulation of the entropy where we sum over all possible joint assignments. However, if we consider more complex networks, the number of joint assignments is exponentially large while the number of potentials is typically reasonable, and each one involves the joint assignments to only a few variables.

Note, however, that to use the formulation of proposition 8.1 we need to perform a global computation to find the value of the partition function Z as well as the marginal distribution over the scope of each potential $\boldsymbol{D}_k$. As we will see in later chapters, in some network structures, these computations can be done efficiently.

Terms such as $\boldsymbol{E}_P[-\ln \beta_k(\boldsymbol{D}_k)]$ resemble the entropy of $\boldsymbol{D}_k$. However, since the marginal over $\boldsymbol{D}_k$ is usually not identical to the potential β_k, such terms are not entropy terms. In some sense we can think of $\ln Z$ as a correction for this discrepancy. For example, if we multiply all the entries of β_k by a constant c, the corresponding term $\boldsymbol{E}_P[-\ln \beta_k(\boldsymbol{D}_k)]$ will decrease by $\ln c$. However, at the same time $\ln Z$ will increase by the same constant, since it is canceled out in the normalization.

8.4.1.2 Entropy of Bayesian Networks

We now consider the entropy of a Bayesian network. Although we can address this computation using our general result in theorem 8.1, it turns out that the formulation for Bayesian networks is simpler. Intuitively, as we saw, we can represent Bayesian networks as an exponential family where the partition function is 1. This removes the global term from the entropy.

Theorem 8.2 *If $P(\mathcal{X}) = \prod_i P(X_i \mid \mathrm{Pa}_i^{\mathcal{G}})$ is a distribution consistent with a Bayesian network $\mathcal{G}$, then*

$$\boldsymbol{H}_P(\mathcal{X}) = \sum_i \boldsymbol{H}_P(X_i \mid \mathrm{Pa}_i^{\mathcal{G}})$$

PROOF

$$\begin{aligned}
\boldsymbol{H}_P(\mathcal{X}) &= \boldsymbol{E}_P[-\ln P(\mathcal{X})] \\
&= \boldsymbol{E}_P\left[-\sum_i \ln P(X_i \mid \mathrm{Pa}_i^{\mathcal{G}})\right] \\
&= \sum_i \boldsymbol{E}_P\left[-\ln P(X_i \mid \mathrm{Pa}_i^{\mathcal{G}})\right] \\
&= \sum_i \boldsymbol{H}_P(X_i \mid \mathrm{Pa}_i^{\mathcal{G}}),
\end{aligned}$$

where the first and last steps invoke the definitions of entropy and conditional entropy. ■

We see that the entropy of a Bayesian network decomposes as a sum of conditional entropies of the individual conditional distributions. This representation suggests that the entropy of a Bayesian network can be directly "read off" from the CPDs. This impression is misleading. Recall that the conditional entropy term $\boldsymbol{H}_P(X_i \mid \mathrm{Pa}_i^{\mathcal{G}})$ can be written as a weighted average of simpler entropies of conditional distributions

$$\boldsymbol{H}_P(X_i \mid \mathrm{Pa}_i^{\mathcal{G}}) = \sum_{\mathrm{pa}_i^{\mathcal{G}}} P(\mathrm{pa}_i^{\mathcal{G}}) \boldsymbol{H}_P(X_i \mid \mathrm{pa}_i^{\mathcal{G}}).$$

While each of the simpler entropy terms in the summation can be computed based on the CPD entries alone, the weighting term $P(\mathrm{pa}_i^{\mathcal{G}})$ is a marginal over $\mathrm{pa}_i^{\mathcal{G}}$ of the joint distribution, and depends on other CPDs upstream of X_i. Thus, computing the entropy of the network requires that we answer probability queries over the network.

However, based on local considerations alone, we can analyze the amount of entropy introduced by each CPD, and thereby provide bounds on the overall entropy:

Proposition 8.2 *If $P(\mathcal{X}) = \prod_i P(X_i \mid \mathrm{Pa}_i^{\mathcal{G}})$ is a distribution consistent with a Bayesian network $\mathcal{G}$, then*

$$\sum_i \min_{\mathrm{pa}_i^{\mathcal{G}}} \boldsymbol{H}_P(X_i \mid \mathrm{pa}_i^{\mathcal{G}}) \leq \boldsymbol{H}_P(\mathcal{X}) \leq \sum_i \max_{\mathrm{pa}_i^{\mathcal{G}}} \boldsymbol{H}_P(X_i \mid \mathrm{pa}_i^{\mathcal{G}}).$$

Thus, if all the CPDs in a Bayesian network are almost deterministic (low conditional entropy given each parent configuration), then the overall entropy of the network is small. Conversely, if all the CPDs are highly stochastic (high conditional entropy) then the overall entropy of the network is high.

8.4.2 Relative Entropy

A related notion is the relative entropy between models. This measure of distance plays an important role in many of the developments of later chapters.

If we consider the relative entropy between an arbitrary distribution Q and a distribution P_θ within an exponential family, we see that the form of P_θ can be exploited to simplify the form of the relative entropy.

Theorem 8.3 *Consider a distribution Q and a distribution $P_{\boldsymbol{\theta}}$ in an exponential family defined by τ and $\mathtt{t}$. Then*

$$\boldsymbol{D}(Q \| P_{\boldsymbol{\theta}}) = -\boldsymbol{H}_Q(\mathcal{X}) - \langle \boldsymbol{E}_Q[\tau(\mathcal{X})], \mathtt{t}(\boldsymbol{\theta}) \rangle + \ln Z(\boldsymbol{\theta}).$$

The proof is left as an exercise (exercise 8.2).

We see that the quantities of interest are again the expected sufficient statistics and the partition function. Unlike the entropy, in this case we compute the expectation of the sufficient statistics according to Q.

If both distributions are in the same exponential family, then we can further simplify the form of the relative entropy.

Theorem 8.4 *Consider two distribution $P_{\boldsymbol{\theta}_1}$ and $P_{\boldsymbol{\theta}_2}$ within the same exponential family. Then*

$$\boldsymbol{D}(P_{\boldsymbol{\theta}_1} \| P_{\boldsymbol{\theta}_2}) = \langle \boldsymbol{E}_{P_{\boldsymbol{\theta}_1}}[\tau(\mathcal{X})], \mathtt{t}(\boldsymbol{\theta}_1) - \mathtt{t}(\boldsymbol{\theta}_2) \rangle - \ln \frac{Z(\boldsymbol{\theta}_1)}{Z(\boldsymbol{\theta}_2)}$$

Proof Combine theorem 8.3 with theorem 8.1. ■

When we consider Bayesian networks, we can use the fact that the partition function is constant to simplify the terms in both results.

Theorem 8.5 *If P is a distribution consistent with a Bayesian network $\mathcal{G}$, then*

$$\boldsymbol{D}(Q \| P) = -\boldsymbol{H}_Q(\mathcal{X}) - \sum_i \sum_{\mathrm{pa}_i^{\mathcal{G}}} Q(\mathrm{pa}_i^{\mathcal{G}}) \boldsymbol{E}_{Q(X_i \mid \mathrm{pa}_i^{\mathcal{G}})} \left[\ln P(X_i \mid \mathrm{pa}_i^{\mathcal{G}}) \right];$$

If Q is also consistent with $\mathcal{G}$, then

$$\boldsymbol{D}(Q \| P) = \sum_i \sum_{\mathrm{pa}_i^{\mathcal{G}}} Q(\mathrm{pa}_i^{\mathcal{G}}) \boldsymbol{D}(Q(X_i \mid \mathrm{pa}_i^{\mathcal{G}}) \| P(X_i \mid \mathrm{pa}_i^{\mathcal{G}})).$$

The second result shows that, analogously to the form of the entropy of Bayesian networks, we can write the relative entropy between two distributions consistent with $\mathcal{G}$ as a weighted sum of the relative entropies between the conditional distributions. These conditional relative entropies can be evaluated directly using the CPDs of the two networks. The weighting of these relative entropies depends on the the joint distribution Q.

8.5 Projections

As we discuss in appendix A.1.3, we can view the relative entropy as a notion of distance between two distributions. We can therefore use it as the basis for an important operation — the *projection* operation — which we will utilize extensively in subsequent chapters. Similar to the geometric concept of projecting a point onto a hyperplane, we consider the problem of finding the distribution, within a given exponential family, that is closest to a given distribution

projection

in terms of relative entropy. For example, we want to perform such a projection when we approximate a complex distribution with one with a simple structure. As we will see, this is a crucial strategy for approximate inference in networks where exact inference is infeasible. In such an approximation we would like to find the best (that is, closest) approximation within a family in which we can perform inference. Moreover, the problem of *learning* a graphical model can also be posed as a projection problem of the empirical distribution observed in the data onto a desired family.

Suppose we have a distribution P and we want to approximate it with another distribution Q in a class of distributions $\mathcal{Q}$ (for example, an exponential family). For example, we might want to approximate P with a product of marginal distributions. Because the notion of relative entropy is not symmetric, we can use it to define two types of approximations.

Definition 8.4

Let P be a distribution and let $\mathcal{Q}$ be a convex set of distributions.

I-projection

- *The* I-projection *(information projection) of P onto $\mathcal{Q}$ is the distribution*

$$Q^I = \arg\min_{Q \in \mathcal{Q}} \boldsymbol{D}(Q \| P).$$

M-projection

- *The* M-projection *(moment projection) of P onto $\mathcal{Q}$ is the distribution*

$$Q^M = \arg\min_{Q \in \mathcal{Q}} \boldsymbol{D}(P \| Q).$$

■

8.5.1 Comparison

We can think of both Q^I and Q^M as the projection of P into the set $\mathcal{Q}$ in the sense that it is the distribution closest to P. Moreover, if $P \in \mathcal{Q}$, then in both definitions the projection would be P. However, because the relative entropy is not symmetric, these two projections are, in general, different. To understand the differences between these two projections, let us consider a few examples.

Example 8.13

Suppose we have a non-Gaussian distribution P over the reals. We can consider the M-projection and the I-projection on the family of Gaussian distributions. As a concrete example, consider the distribution P of figure 8.1. As we can see, the two projections are different Gaussian distributions. (The M-projection was found using the analytic form that we will discuss, and the I-projection by gradient ascent in the (μ, σ^2) space.) Although the means of the two projected distributions are relatively close, the M-projection has larger variance than the I-projection. ■

We can better understand these differences if we examine the objective function optimized by each projection. Recall that the M-projection Q^M minimizes

$$\boldsymbol{D}(P \| Q) = -\boldsymbol{H}_P(X) + \boldsymbol{E}_P[-\ln Q(X)].$$

We see that, in general, we want Q^M to have high density in regions that are probable according to P, since a small $-\ln Q(X)$ in these regions will lead to a smaller second term. At the same time, there is a high penalty for assigning low density to regions where $P(X)$ is nonnegligible.

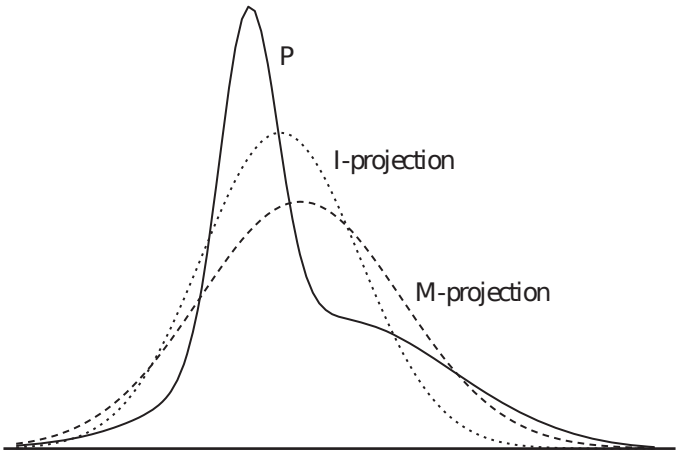

Figure 8.1 Example of M- and I-projections into the family of Gaussian distributions

As a consequence, although the M-projection attempts to match the main mass of P, its high variance is a compromise to ensure that it assigns reasonably high density to all regions that are in the support of P.

On the other hand, the I-projection minimizes

$$\boldsymbol{D}(Q\|P) = -\boldsymbol{H}_Q(X) + \boldsymbol{E}_Q[-\ln P(X)].$$

Thus, the first term incurs a penalty for low entropy, which in the case of a Gaussian Q translates to a penalty on small variance. The second term, $\boldsymbol{E}_Q[-\ln P(X)]$, encodes a preference for assigning higher density to regions where $P(X)$ is large and very low density to regions where $P(X)$ is small. Without the first term, we can minimize the second by putting all of the mass of Q on the most probable point according to P. The compromise between the two terms results in the distribution we see in figure 8.1.

A similar phenomenon occurs in discrete distributions.

Example 8.14

Now consider the projection of a distribution $P(A, B)$ onto the family of factored distributions $Q(A, B) = Q(A)Q(B)$. Suppose $P(A, B)$ is the following distribution:

$$\begin{aligned} P(a^0, b^0) &= 0.45 \\ P(a^0, b^1) &= 0.05 \\ P(a^1, b^0) &= 0.05 \\ P(a^1, b^1) &= 0.45. \end{aligned}$$

That is, the distribution P puts almost all of the mass on the event $A = B$. This distribution is a particularly difficult one to approximate using a factored distribution, since in P the two variables A and B are highly correlated, a dependency that cannot be captured using a fully factored Q.

Again, it is instructive to compare the M-projection and the I-projection of this distribution (see figure 8.2). It follows from example A.7 (appendix A.5.3) that the M-projection of this distribution is

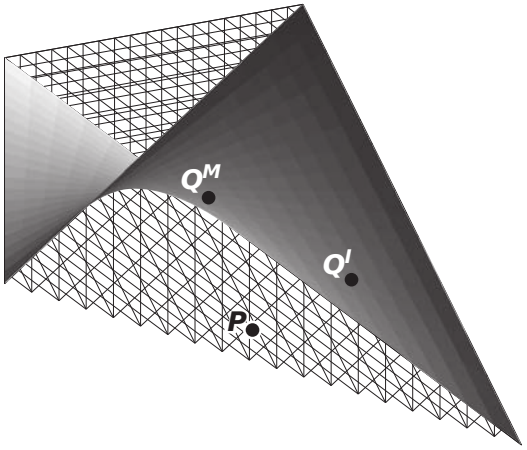

Figure 8.2 Example of M- and I-projections of a two variable discrete distribution where $P(a^0 = b^0) = P(a^1 = b^1) = 0.45$ and $P(a^0 = b^1) = P(a^0 = b^1) = 0.05$ onto factorized distribution. Each axis denotes the probability of an instance: $P(a^1, b^1)$, $P(a^1, b^0)$, and $P(a^0, b^1)$. The wire surfaces mark the region of legal distributions. The solid surface shows the distributions where A and independent of B. The points show P and its two projections.

the uniform distribution:

$$\begin{aligned} Q^M(a^0, b^0) &= 0.5 * 0.5 = 0.25 \\ Q^M(a^0, b^1) &= 0.5 * 0.5 = 0.25 \\ Q^M(a^1, b^0) &= 0.5 * 0.5 = 0.25 \\ Q^M(a^1, b^1) &= 0.5 * 0.5 = 0.25. \end{aligned}$$

In contrast, the I-projection focuses on one of the two "modes" of the distribution, either when both A and B are true or when both are false. Since the distribution is symmetric about these modes, there are two I-projections. One of them is

$$\begin{aligned} Q^I(a^0, b^0) &= 0.25 * 0.25 = 0.0625 \\ Q^I(a^0, b^1) &= 0.25 * 0.75 = 0.1875 \\ Q^I(a^1, b^0) &= 0.75 * 0.25 = 0.1875 \\ Q^I(a^1, b^1) &= 0.75 * 0.75 = 0.5625. \end{aligned}$$

The second I-projection is symmetric around the opposite mode a^0, b^0. ■

As in example 8.13, we can understand these differences by considering the underlying mathematics. **The M-projection attempts to give all assignments reasonably high probability, whereas the I-projection attempts to focus on high-probability assignments in P while maintaining a reasonable entropy.** In this case, this behavior results in a uniform distribution for the M-projection, whereas the I-projection places most of the probability mass on one of the two assignments where P has high probability.

8.5.2 M-Projections

Can we say more about the form of these projections? We start by considering M-projections onto a simple family of distributions.

Proposition 8.3 *Let P be a distribution over $X_1, \ldots, X_n$, and let $\mathcal{Q}$ be the family of distributions consistent with $\mathcal{G}_\emptyset$, the empty graph. Then*

$$Q^M = \arg \min_{Q \models \mathcal{G}_\emptyset} \boldsymbol{D}(P \| Q)$$

is the distribution:

$$Q^M(X_1, \ldots, X_n) = P(X_1)P(X_2) \cdots P(X_n).$$

PROOF Consider a distribution $Q \models \mathcal{G}_\emptyset$. Since Q factorizes, we can rewrite $\boldsymbol{D}(P \| Q)$:

$$\begin{aligned}
\boldsymbol{D}(P \| Q) &= \boldsymbol{E}_P[\ln P(X_1, \ldots, X_n) - \ln Q(X_1, \ldots, X_n)] \\
&= \boldsymbol{E}_P[\ln P(X_1, \ldots, X_n)] - \sum_i \boldsymbol{E}_P[\ln Q(X_i)] \\
&= \boldsymbol{E}_P\left[\ln \frac{P(X_1, \ldots, X_n)}{P(X_1) \cdots P(X_n)}\right] + \sum_i \boldsymbol{E}_P\left[\ln \frac{P(X_i)}{Q(X_i)}\right] \\
&= \boldsymbol{D}(P \| Q^M) + \sum_i \boldsymbol{D}(P(X_i) \| Q(X_i)) \\
&\geq \boldsymbol{D}(P \| Q^M).
\end{aligned}$$

The last step relies on the nonnegativity of the relative entropy. We conclude that $\boldsymbol{D}(P \| Q) \geq \boldsymbol{D}(P \| Q^M)$ with equality only if $Q(X_i) = P(X_i)$ for all i. That is, only when $Q = Q^M$. ∎

Hence, the M-projection of P onto factored distribution is simply the product of marginals of P.

This theorem is an instance of a much more general result. To understand the generalization, we observe that the family $\mathcal{Q}$ of fully factored distributions is characterized by a vector of sufficient statistics that simply counts, for each variable X_i, the number of occurrences of each of its values. The marginal distributions over the X_i's are simply the expectations, relative to P, of these sufficient statistics. We see that, by selecting Q to match these expectations, we obtain the M-projection.

As we now show, this is not an accident. The characterization of a distribution P that is relevant to computing its M-projection into $\mathcal{Q}$ is precisely the expectation, relative to P, of the sufficient statistic function of $\mathcal{Q}$.

Theorem 8.6

Let P be a distribution over $\mathcal{X}$, and let $\mathcal{Q}$ be an exponential family defined by the functions $\tau(\xi)$ and $\mathtt{t}(\boldsymbol{\theta})$. If there is a set of parameters $\boldsymbol{\theta}$ such that $\boldsymbol{E}_{Q_{\boldsymbol{\theta}}}[\tau(\mathcal{X})] = \boldsymbol{E}_P[\tau(\mathcal{X})]$, then the M-projection of P is $Q_{\boldsymbol{\theta}}$.

PROOF Suppose that $\boldsymbol{E}_P[\tau(\mathcal{X})] = \boldsymbol{E}_{Q_{\boldsymbol{\theta}}}[\tau(\mathcal{X})]$, and let $\boldsymbol{\theta}'$ be some set of parameters. Then,

$$\begin{aligned}
\boldsymbol{D}(P\|Q_{\boldsymbol{\theta}'}) - \boldsymbol{D}(P\|Q_{\boldsymbol{\theta}}) &= -\boldsymbol{H}_P(\mathcal{X}) - \langle \boldsymbol{E}_P[\tau(\mathcal{X})], \mathtt{t}(\boldsymbol{\theta}')\rangle + \ln Z(\boldsymbol{\theta}') \\
&\quad + \boldsymbol{H}_P(\mathcal{X}) + \langle \boldsymbol{E}_P[\tau(\mathcal{X})], \mathtt{t}(\boldsymbol{\theta})\rangle - \ln Z(\boldsymbol{\theta}) \\
&= \langle \boldsymbol{E}_P[\tau(\mathcal{X})], \mathtt{t}(\boldsymbol{\theta}) - \mathtt{t}(\boldsymbol{\theta}')\rangle - \ln \frac{Z(\boldsymbol{\theta})}{Z(\boldsymbol{\theta}')} \\
&= \langle \boldsymbol{E}_{Q_{\boldsymbol{\theta}}}[\tau(\mathcal{X})], \mathtt{t}(\boldsymbol{\theta}) - \mathtt{t}(\boldsymbol{\theta}')\rangle - \ln \frac{Z(\boldsymbol{\theta})}{Z(\boldsymbol{\theta}')} \\
&= \boldsymbol{D}(Q_{\boldsymbol{\theta}}\|Q_{\boldsymbol{\theta}'}) \geq 0.
\end{aligned}$$

We conclude that the M-projection of P is $Q_{\boldsymbol{\theta}}$. ■

This theorem suggests that we can consider both the distribution P and the distributions in $\mathcal{Q}$ in terms of the expectations of $\tau(\mathcal{X})$. Thus, instead of describing a distribution in the family by the set of parameters, we can describe it in terms of the *expected sufficient statistics*. To formalize this intuition, we need some additional notation. We define a mapping from legal parameters in Θ to vectors of sufficient statistics

expected sufficient statistics

$$\mathit{ess}(\boldsymbol{\theta}) = \boldsymbol{E}_{Q_{\boldsymbol{\theta}}}[\tau(\mathcal{X})].$$

Theorem 8.6 shows that if $\boldsymbol{E}_P[\tau(\mathcal{X})]$ is in the image of *ess*, then the M-projection of P is the distribution $Q_{\boldsymbol{\theta}}$ that matches the expected sufficient statistics of P. In other words,

$$\boldsymbol{E}_{Q^M}[\tau(\mathcal{X})] = \boldsymbol{E}_P[\tau(\mathcal{X})].$$

moment matching

This result explains why M-projection is also referred to as *moment matching*. In many exponential families the sufficient statistics are moments (mean, variance, and so forth) of the distribution. In such cases, the M-projection of P is the distribution in the family that matches these moments in P.

We illustrate these concepts in figure 8.3. As we can see, the mapping $\mathit{ess}(\boldsymbol{\theta})$ directly relates parameters to expected sufficient statistics. By comparing the expected sufficient statistics of P to these of distributions in $\mathcal{Q}$, we can find the M-projection.

Moreover, using theorem 8.6, we obtain a general characterization of the M-projection function M-project(s), which maps a vector of expected sufficient statistics to a parameter vector:

Corollary 8.1

Let $\boldsymbol{s}$ be a vector. If $\boldsymbol{s} \in \mathit{image}(\mathit{ess})$ and ess is invertible, then

$$\text{M-project}(\boldsymbol{s}) = \mathit{ess}^{-1}(\boldsymbol{s}).$$

That is, the parameters of the M-projection of P are simply the inverse of the *ess* mapping, applied to the expected sufficient statistic vector of P. This result allows us to describe the M-projection operation in terms of a specific function. This result assumes, of course, that $\boldsymbol{E}_P[\tau]$ is in the image of *ess* and that *ess* is invertible. In many examples that we consider, the image of *ess* includes all possible vectors of expected sufficient statistics we might encounter. Moreover, if the parameterization is nonredundant, then *ess* is invertible.

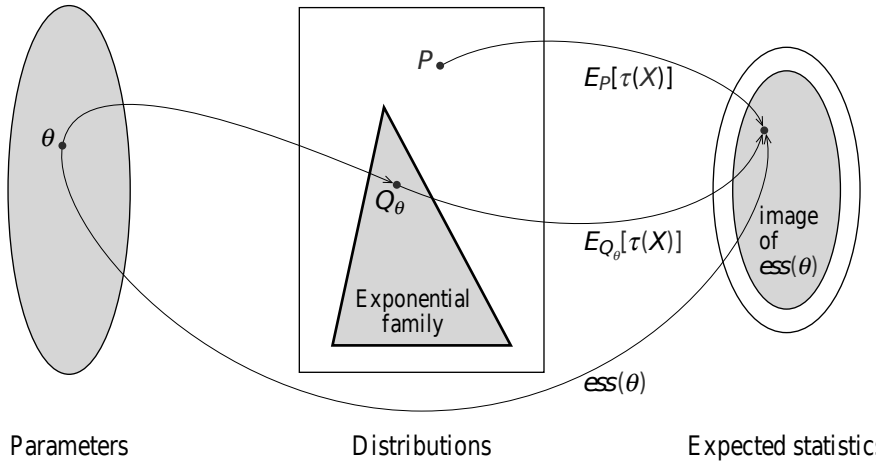

Figure 8.3 Illustration of the relations between parameters, distributions and expected sufficient statistics. Each parameter corresponds to a distribution, which in turn corresponds to a value of the expected statistics. The function *ess* maps parameters directly to expected statistics. If the expected statistics of P and $Q_{\boldsymbol{\theta}}$ match, then $Q_{\boldsymbol{\theta}}$ is the M-projection of P.

Example 8.15

Consider the exponential family of Gaussian distributions. Recall that the sufficient statistics function for this family is $\tau(x) = \langle x, x^2 \rangle$. Given parameters $\boldsymbol{\theta} = \langle \mu, \sigma^2 \rangle$, the expected value of τ is

$$ess(\langle \mu, \sigma^2 \rangle) = \boldsymbol{E}_{Q_{\langle \mu, \sigma^2 \rangle}}[\tau(X)] = \langle\ \mu, \sigma^2 + \mu^2 \rangle.$$

It is not difficult to show that, for any distribution P, $\boldsymbol{E}_P[\tau(X)]$ must be in the image of this function (see exercise 8.4). Thus, for any choice of P, we can apply theorem 8.6.

Finally, we can easily invert this function:

$$\text{M-project}(\langle s_1, s_2 \rangle) = ess^{-1}(\langle s_1, s_2 \rangle) = \langle s_1, s_2 - s_1^2 \rangle.$$

Recall that $s_1 = \boldsymbol{E}_P[X]$ and $s_2 = \boldsymbol{E}_P[X^2]$. Thus, the estimated parameters are the mean and variance of X according to P, as we would expect. ■

This example shows that the "naive" choice of Gaussian distribution, obtained by matching the mean and variance of a variable X, provides the best Gaussian approximation (in the M-projection sense) to a non-Gaussian distribution over X. We have also provided a solution to the M-projection problem in the case of a factored product of multinomials, in proposition 8.3, which can be viewed as a special case of theorem 8.6. In a more general application of this result, we show in section 11.4.4 a general result on the form of the M-projection for a linear exponential family over discrete state space, including the class of Markov networks.

The analysis for other families of distributions can be subtler.

Example 8.16

We now consider a more complex example of M-projection onto a chain network. Suppose we have a distribution P over variables $X_1, \ldots, X_n$, and want to project it onto the family of distributions Q of the distributions that are consistent with the network structure $X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_n$.

What are the sufficient statistics for this network? Based on our previous discussion, we see that each conditional distribution $Q(X_{i+1} \mid X_i)$ requires a statistic of the form

$$\tau_{x_i, x_{i+1}}(\xi) = \mathbb{1}\{X_i = x_i, X_{i+1} = x_{i+1}\} \ \forall \langle x_i, x_{i+1} \rangle \in Val(X_i) \times Val(X_{i+1}).$$

These statistics are sufficient but are redundant. To see this, note that the "marginal statistics" must agree. That is,

$$\sum_{x_i} \tau_{x_i, x_{i+1}}(\xi) = \sum_{x_{i+2}} \tau_{x_{i+1}, x_{i+2}}(\xi) \ \forall x_{i+1} \in Val(X_{i+1}). \tag{8.8}$$

Although this representation is redundant, we can still apply the mechanisms discussed earlier and consider the function ess that maps parameters of such a network to the sufficient statistics. The expectation of an indicator function is the marginal probability of that event, so that

$$\mathbf{E}_{Q_{\boldsymbol{\theta}}}\left[\tau_{x_i, x_{i+1}}(\mathcal{X})\right] = Q_{\boldsymbol{\theta}}(x_i, x_{i+1}).$$

Thus, the function ess simply maps from $\boldsymbol{\theta}$ to the pairwise marginals of consecutive variables in $Q_{\boldsymbol{\theta}}$. Because these are pairwise marginals of an actual distribution, it follows that these sufficient statistics satisfy the consistency constraints of equation (8.8).

How do we invert this function? Given the statistics from P, we want to find a distribution Q that matches them. We start building Q along the structure of the chain. We choose $Q(X_1)$ and $Q(X_2 \mid X_1)$ so that $Q(x_1, x_2) = \mathbf{E}_P[\tau_{x_1, x_2}(\mathcal{X})] = P(x_1, x_2)$. In fact, there is a unique choice that satisfies this equality, where $Q(X_1, X_2) = P(X_1, X_2)$. This choice implies that the marginal distribution $Q(X_2)$ matches the marginal distribution $P(X_2)$. Now, consider our choice of $Q(X_3 \mid X_2)$. We need to ensure that

$$Q(x_3, x_2) = \mathbf{E}_P[\tau_{x_2, x_3}(\mathcal{X})] = P(x_2, x_3).$$

We note that, because $Q(x_3, x_2) = Q(x_3 \mid x_2)Q(x_2) = Q(x_3 \mid x_2)P(x_2)$, we can achieve this equality by setting $Q(x_3 \mid x_2) = P(x_3 \mid x_2)$. Moreover, this implies that $Q(x_3) = P(x_3)$. We can continue this construction recursively to set

$$Q(x_{i+1} \mid x_i) = P(x_{i+1} \mid x_i).$$

Using the preceding argument, we can show that this choice will match the sufficient statistics of P. This suffices to show that this Q is the M-projection of P.

Note that, although this choice of Q coincides with P on pairwise marginals of consecutive variables, it does not necessarily agree with P on other marginals. As an extreme example, consider a distribution P where X_1 and X_3 are identical and both are independent of X_2. If we project this distribution onto a distribution Q with the structure $X_1 \rightarrow X_2 \rightarrow X_3$, then P and Q will not necessarily agree on the joint marginals of X_1, X_3. In Q this distribution will be

$$Q(x_1, x_3) = \sum_{x_2} Q(x_1, x_2)Q(x_3 \mid x_2).$$

Since $Q(x_1, x_2) = P(x_1, x_2) = P(x_1)P(x_2)$ and $Q(x_3 \mid x_2) = P(x_3 \mid x_2) = P(x_3)$, we conclude that $Q(x_1, x_3) = P(x_1)P(x_3)$, losing the equality between X_1 and X_3 in P. ■

This analysis used a redundant parameterization; exercise 8.6 shows how we can reparameterize a directed chain within the linear exponential family and thereby obtain an alternative perspective on the M-projection operation.

So far, all of our examples have had the characteristic that the vector of expected sufficient statistics for a distribution P is always in the image of *ess*; thus, our task has only been to invert *ess*. Unfortunately, there are examples where not every vector of expected sufficient statistics can also be derived from a distribution in our exponential family.

Example 8.17

Consider again the family $\mathcal{Q}$ from example 8.10, of distributions parameterized using network structure $A \rightarrow C \leftarrow B$, with binary variables A, B, C. We can show that the sufficient statistics for this distribution are indicators for all the joint assignments to A, B, and C except one. That is,

$$\begin{aligned}\tau(A,B,C) \quad = \langle \quad & \mathbf{1}\{A = a^1, B = b^1, C = c^1\}, \\ & \mathbf{1}\{A = a^0, B = b^1, C = c^1\}, \\ & \mathbf{1}\{A = a^1, B = b^0, C = c^1\}, \\ & \mathbf{1}\{A = a^1, B = b^1, C = c^0\}, \\ & \mathbf{1}\{A = a^1, B = b^0, C = c^0\}, \\ & \mathbf{1}\{A = a^0, B = b^1, C = c^0\}, \\ & \mathbf{1}\{A = a^0, B = b^0, C = c^1\}\rangle.\end{aligned}$$

If we look at the expected value of these statistics given some member of the family, we have that, since A and B are independent in $Q_{\boldsymbol{\theta}}$, $Q_{\boldsymbol{\theta}}(a^1, b^1) = Q_{\boldsymbol{\theta}}(a^1)Q_{\boldsymbol{\theta}}(b^1)$. Thus, the expected statistics should satisfy

$$\begin{aligned}&\mathbf{E}_{Q_{\boldsymbol{\theta}}}\left[\mathbf{1}\{A = a^1, B = b^1, C = c^1\}\right] + \mathbf{E}_{Q_{\boldsymbol{\theta}}}\left[\mathbf{1}\{A = a^1, B = b^1, C = c^0\}\right] = \\ &\quad \left(\mathbf{E}_{Q_{\boldsymbol{\theta}}}\left[\mathbf{1}\{A = a^1, B = b^1, C = c^1\}\right] + \mathbf{E}_{Q_{\boldsymbol{\theta}}}\left[\mathbf{1}\{A = a^1, B = b^1, C = c^0\}\right]\right. \\ &\qquad \left. + \mathbf{E}_{Q_{\boldsymbol{\theta}}}\left[\mathbf{1}\{A = a^1, B = b^0, C = c^1\}\right] + \mathbf{E}_{Q_{\boldsymbol{\theta}}}\left[\mathbf{1}\{A = a^1, B = b^0, C = c^0\}\right]\right) \\ &\quad \left(\mathbf{E}_{Q_{\boldsymbol{\theta}}}\left[\mathbf{1}\{A = a^1, B = b^1, C = c^1\}\right] + \mathbf{E}_{Q_{\boldsymbol{\theta}}}\left[\mathbf{1}\{A = a^1, B = b^1, C = c^0\}\right]\right. \\ &\qquad \left. + \mathbf{E}_{Q_{\boldsymbol{\theta}}}\left[\mathbf{1}\{A = a^0, B = b^1, C = c^1\}\right] + \mathbf{E}_{Q_{\boldsymbol{\theta}}}\left[\mathbf{1}\{A = a^0, B = b^1, C = c^0\}\right]\right).\end{aligned}$$

This constraint is not typically satisfied by the expected statistics from a general distribution P we might consider projecting. Thus, in this case, there are expected statistics vectors that do not fall within the image of ess. ■

In such cases, and in Bayesian networks in general, the projection procedure is more complex than inverting the *ess* function. Nevertheless, we can show that the projection operation still has an analytic solution.

Theorem 8.7

Let P be a distribution over $X_1, \ldots, X_n$, and let $\mathcal{G}$ be a Bayesian network structure. Then the M-projection Q^M is:

$$Q^M(X_1, \ldots, X_n) = \prod_i P(X_i \mid \mathrm{Pa}_{X_i}^{\mathcal{G}}).$$

Because the mapping *ess* for Bayesian networks is not invertible, the proof of this result (see exercise 8.5) does not build on theorem 8.6 but rather directly on theorem 8.5. This result turns out to be central to our derivation of Bayesian network learning in chapter 17.

8.5.3 I-Projections

What about I-projections? Recall that

$$\boldsymbol{D}(Q\|P) = -\boldsymbol{H}_Q(\mathcal{X}) - \boldsymbol{E}_Q[\ln P(\mathcal{X})].$$

If Q is in some exponential family, we can use the derivation of theorem 8.1 to simplify the entropy term. However, the exponential form of Q does not provide insights into the second term. When dealing with the I-projection of a general distribution P, we are left without further simplifications. However, if the distribution P has some structure, we might be able to simplify $\boldsymbol{E}_Q[\ln P(\mathcal{X})]$ into simpler terms, although the projection problem is still a nontrivial one. We discuss this problem in much more detail in chapter 11.

8.6 Summary

In this chapter, we presented some of the basic technical concepts that underlie many of the techniques we explore in depth later in the book. We defined the formalism of exponential families, which provides the fundamental basis for considering families of related distributions. We also defined the subclass of linear exponential families, which are significantly simpler and yet cover a large fraction of the distributions that arise in practice.

We discussed how the types of distributions described so far in this book fit into this framework, showing that Gaussians, linear Gaussians, and multinomials are all in the linear exponential family. Any class of distributions representable by parameterizing a Markov network of some fixed structure is also in the linear exponential family. By contrast, the class of distributions representable by a Bayesian network of some fixed structure is in the exponential family, but is not in the linear exponential family when the network structure includes an immorality.

We showed how we can use the formulation of an exponential family to facilitate computations such as the entropy of a distribution or the relative entropy between two distributions. The latter computation formed the basis for analyzing a basic operation on distributions: that of projecting a general distribution P into some exponential family $\mathcal{Q}$, that is, finding the distribution within $\mathcal{Q}$ that is closest to P. Because the notion of relative entropy is not symmetric, this concept gave rise to two different definitions: I-projection, where we minimize $\boldsymbol{D}(Q\|P)$, and M-projection, where we minimize $\boldsymbol{D}(P\|Q)$. We analyzed the differences between these two definitions and showed that solving the M-projection problem can be viewed in a particularly elegant way, constructing a distribution Q that matches the expected sufficient statistics (or moments) of P.

As we discuss later in the book, both the I-projection and M-projection turn out to play an important role in graphical models. The M-projection is the formal foundation for addressing the learning problem: there, our goal is to find a distribution in a particular class (for example, a Bayesian network or Markov network of a given structure) that is closest (in the M-projection sense) to the *empirical distribution* observed in a data set from which we wish to learn (see equation (16.4)). The I-projection operation is used when we wish to take a given graphical model P and answer probability queries; when P is too complex to allow queries to be answered

efficiently, one strategy is to construct a simpler distribution Q, which is a good approximation to P (in the I-projection sense).

8.7 Relevant Literature

The concept of exponential families plays a central role in formal statistic theory. Much of the theory is covered by classic textbooks such as Barndorff-Nielsen (1978). See also Lauritzen (1996). Geiger and Meek (1998) discuss the representation of graphical models as exponential families and show that a Bayesian network usually does not define a linear exponential family.

The notion of I-projections was introduced by Csiszàr (1975), who developed the "information geometry" of such projections and their connection to different estimation procedures. In his terminology, M-projections are called "reverse I-projections." The notion of M-projection is closely related to parameter learning, which we revisit in chapter 17 and chapter 20.

8.8 Exercises

Exercise 8.1⋆

Poisson distribution

A variable X with $Val(X) = 0, 1, 2, \ldots$ is *Poisson-distributed* with parameter $\theta > 0$ if

$$P(X = k) = \frac{1}{k!} \exp -\theta\theta^k.$$

This distribution has the property that $\boldsymbol{E}_P[X] = \theta$.

a. Show how to represent the Poisson distribution as a linear exponential family. (Note that unlike most of our running examples, you need to use the auxiliary measure A in the definition.)

b. Use results developed in this chapter to find the entropy of a Poisson distribution and the relative entropy between two Poisson distributions.

c. What is the function *ess* associated with this family? Is it invertible?

Exercise 8.2

Prove theorem 8.3.

Exercise 8.3

In this exercise, we will provide a characterization of when two distributions P_1 and P_2 will have the same M-projection.

a. Let P_1 and P_2 be two distribution over $\mathcal{X}$, and let $\mathcal{Q}$ be an exponential family defined by the functions $\tau(\xi)$ and $\mathtt{t}(\boldsymbol{\theta})$. If $\boldsymbol{E}_{P_1}[\tau(\mathcal{X})] = \boldsymbol{E}_{P_2}[\tau(\mathcal{X})]$, then the M-projection of P_1 and P_2 onto $\mathcal{Q}$ is identical.

b. Now, show that if the function $ess(\boldsymbol{\theta})$ is invertible, then we can prove the converse, showing that the M-projection of P_1 and P_2 is identical only if $\boldsymbol{E}_{P_1}[\tau(\mathcal{X})] = \boldsymbol{E}_{P_2}[\tau(\mathcal{X})]$. Conclude that this is the case for linear exponential families.

Exercise 8.4

Consider the function *ess* for Gaussian variables as described in example 8.15.

a. What is the image of *ess*?

b. Consider terms of the form $\boldsymbol{E}_P[\tau(X)]$ for the Gaussian sufficient statistics from that example. Show that for any distribution P, the expected sufficient statistics is in the image of *ess*.

Exercise 8.5⋆

Prove theorem 8.7. (Hint: Use theorem 8.5.)

Exercise 8.6⋆

Let $X_1, \ldots, X_n$ be binary random variables. Suppose we are given a family $\mathcal{Q}$ of chain distributions of the form $Q(X_1, \ldots, X_n) = Q(X_1)Q(X_2 \mid X_1) \cdots Q(X_n \mid X_{n-1})$. We now show how to reformulate this family as a linear exponential family.

a. Show that the following vector of statistics is sufficient and nonredundant for distributions in the family:

$$\tau(X_1, \ldots, X_n) = \begin{pmatrix} \mathbf{1}\{X_1 = x_1^1\}, \\ \ldots \\ \mathbf{1}\{X_n = x_n^1\}, \\ \mathbf{1}\{X_1 = x_1^1, X_2 = x_2^1\}, \\ \ldots \\ \mathbf{1}\{X_{n-1} = x_{n-1}^1, X_n = x_n^1\} \end{pmatrix}.$$

b. Show that you can reconstruct the distributions $Q(X_1)$ and $Q(X_{i+1} \mid X_i)$ from the the expectation $E_Q[\tau(X_1, \ldots, X_n)]$. This shows that given the expected sufficient statistics you can reconstruct Q.

c. Suppose you know Q. Show how to reparameterize it as a linear exponential model

$$Q(X_1, \ldots, X_n) = \frac{1}{Z} \exp\left\{ \sum_i \theta_i \mathbf{1}\{X_i = x_i^1\} + \sum_i \theta_{i,i+1} \mathbf{1}\{X_i = x_i^1, X_{i+1} = x_{i+1}^1\} \right\}. \quad (8.9)$$

Note that, because the statistics are sufficient, we know that there are some parameters for which we get equality; the question is to determine their values. Specifically, show that if we choose:

$$\theta_i = \ln \frac{Q(x_1^0, \ldots, x_{i-1}^0, x_i^1, x_{i+1}^0, \ldots, x_n^0)}{Q(x_1^0, \ldots, x_n^0)}$$

and

$$\theta_{i,i+1} = \ln \frac{Q(x_1^0, \ldots, x_{i-1}^0, x_i^1, x_{i+1}^1 x_{i+2}^0, \ldots, x_n^0)}{Q(x_1^0, \ldots, x_n^0)} - \theta_i - \theta_{i+1}$$

then we get equality in equation (8.9) for all assignments to $X_1, \ldots, X_n$.