An Introduction to Probabilistic Graphical Models

Michael I. Jordan
University of California, Berkeley

June 30, 2003

Chapter 19

Features, maximum entropy, and duality

Chapter 16 and Chapter 17 have presented aspects of the general theory of graphical models. As we have seen, this theory is based on associating potential functions with the cliques in the graph, focusing in particular on the maximal cliques of the graph. By ranging over all possible potentials on the maximal cliques of a graph, we obtain all of the probability distributions that respect the Markov properties of the graph. Moreover, by forming junction trees that include all of the maximal cliques of a triangulated graph, we obtain a general inference algorithm that fully exploits these Markov properties.

While the focus on maximal cliques is necessary for the theory to take its simplest and most general form, in practical applications of graphical models one does not necessarily want to work with families of probability distributions that range over arbitrary potentials on maximal cliques. Large, fully-parameterized cliques are problematic both for computational reasons (inference is exponential in the clique sizes) and for statistical reasons (the estimation of large numbers of parameters requires large amounts of data). One generally wants to work with reduced parameterizations that range over proper subsets of the set of all possible potential functions on maximal cliques.

Consider, for example, the applied problem of building a graphical that assigns high probability to strings that respect the orthographic rules of English and low probability to strings that do not. Let us use a very simple model of strings of letters in which, for a given string length, each position in the string is represented by a multinomial random variable that takes on one of 26 values. Thus, for strings of length five we have five nodes and a sample space of size 26^5. One fact about English is that strings ending in `ing` have relatively high probability, and we would like to incorporate this fact into our model. Assigning high probability to `ing`, above and beyond the probabilities that we assign to singletons `i`, `n` and `g`, and the pairs `in` and `ng`, requires that we include a third-order term in our model—a potential function on three nodes. A table on three nodes has 26^3 entries, however, and such a large parameter count gives us pause. It would be unfortunate if we were required to assign values to all of these entries just so that we can assign a value to a particular cell, the `ing` cell. Clearly the way out of this dilemma is to consider reduced parameterizations of the clique potential. In particular, we might construct a parameterized "feature" that varies on the ing cell and is uniform on all other cells.

Another example of a reduced parameterization arises when we build a potential function on a maximal clique from potential functions on non-maximal cliques. For example, we may want to parameterize the clique potential on three nodes, $\psi_{123}(x_1, x_2, x_3)$, in terms of pairwise potentials $\psi_{12}(x_1, x_2)$, $\psi_{13}(x_1, x_3)$, and $\psi_{23}(x_2, x_3)$. We have:

$$\psi_{123}(x_1, x_2, x_3) = \psi_{12}(x_1, x_2)\psi_{13}(x_1, x_3)\psi_{23}(x_2, x_3). \tag{19.1}$$

Letting the pairwise potentials range over all possible values, the product $\psi_{123}(x_1, x_2, x_3)$ ranges over a proper subset of the set of possible potentials on three nodes.

In this chapter we discuss parameterizations of graphical models, focusing on representational issues. We have already touched on some of this material before, in particular in Chapter 8 and Chapter 9. In the current chapter we present a broader and more systematic treatment of this material.

We begin by discussing "features," and their relationship to exponential family representations. We then step back and provide a general motivation for using exponential family representations, introducing the variational principle of *maximum entropy*. Finally, we explore the relationships between maximum entropy and maximum likelihood, showing that they are dual optimization methods, and providing a geometrical interpretation of their duality.

19.1 Features

The goal is to find useful parameterizations of clique potentials, and to be able to estimate parameters from data.

One way to conceptualize this problem is to try to represent potential functions in terms of *features*—elementary functions on subsets of nodes. We put the features together to construct clique potentials. By using a relatively small number of features we obtain a reduced parameterization with relatively few independent degrees of freedom.

For example, to represent the high probability associated with the substring ing, we might use a binary feature $f_{\texttt{ing}}(x)$ that is equal to one if the three nodes in question are in the states i, n and g, respectively, and is equal to zero otherwise. Thus we have an indicator function that picks out a particular cell among the 26^3 cells that form the domain of a potential function on three nodes. We associate with this function, and thus with this cell, a parameter $\theta_{\texttt{ing}}$ that represents the numerical strength of the feature. In particular, we will be working with exponential representations in this chapter, so let us define a potential which is equal to one in all cells except the ing cell where it is equal to $e^{\theta_{\texttt{ing}}}$. We achieve this by defining a base potential $h(x)$ that is equal to one in all 26^3 of the cells, and multiplying this base potential by the factor $e^{\theta_{\texttt{ing}} f_{\texttt{ing}}(x)}$.

We can also define additional features, corresponding to other cells. For example, ate and ion also appear with have high probability at the end of English words. Each such substring corresponds to a different cell, which we can represent with indicator features $f_{\texttt{ate}}(x)$ and $f_{\texttt{ion}}(x)$. Multiplying the potential by the factors $e^{\theta_{\texttt{ate}} f_{\texttt{ate}}(x)}$ and $e^{\theta_{\texttt{ion}} f_{\texttt{ion}}(x)}$ parameterizes the ate and ion cells, while leaving the potential unchanged in the other cells. The resulting potential, $h(x)e^{\theta_{\texttt{ing}} f_{\texttt{ing}}(x)+\theta_{\texttt{ate}} f_{\texttt{ate}}(x)+\theta_{\texttt{ion}} f_{\texttt{ion}}(x)}$, has independently varying parameters in three cells, and is equal to one in all other cells. In the limiting case, if we utilize one binary feature for each cell in the table, we obtain a full parameterization of the potential function in which all cells in the table have an independently adjustable parameter. See Figure 19.1 for a simple example of this construction.

There is no reason, however, for us to be required to use fully parameterized potentials; we may wish to allow only a few of the cells to vary. Moreover, we are not required to utilize features that are indicators of single cells—features that pick out subsets of cells may be of interest. Also, we need not be restricted to binary-valued features or to discrete variables. In particular, for continuous-valued random variables, the notion of "cell" loses its meaning, and it is natural to consider "features" as arbitrary functions on subsets of nodes.

Generalized linear models (GLIMs) can be thought of in terms of features. By taking a linear combination of the parents of a given node, we are considering a subset of all possible probability distributions (all configurations of parents that yield the same value of the linear combination must have the same probability). Here the setting is a directed graph, and the clique of interest is a clique in the moral graph (it is the subset of nodes consisting of a node and its parents). The linear combination used in the GLIM parameterization is a feature on this clique.

Let us formalize these ideas. In the general case, we parameterize a clique C as follows. Define a set of features $f_i(x_{C_i})$, where $C_i \subseteq C$ is the subset of variables that feature f_i references. The subsets C_i are unconstrained, overlapping subsets; in fact, the same subset can appear multiple times. (Indeed, this occurs in Figure 19.1 and in any example in which we represent the cells in a potential table using binary features). Associated with each feature there is a scalar parameter θ_i. Given the features and the parameters we define the clique potential $\psi_C(x_C)$ as follows:

$$\psi_C(x_C) \triangleq \prod_{i \in \mathcal{I}_C} \exp\{\theta_i f_i(x_{C_i})\} \tag{19.2}$$

$$= \exp\left\{\sum_{i \in \mathcal{I}_C} \theta_i f_i(x_{C_i})\right\}, \tag{19.3}$$

where $\mathcal{I}_C$ is a set of indices for the features associated with clique C. If the functions f_i are linearly independent, we obtain a representation of the clique potential that has $|\mathcal{I}_C|$ degrees of freedom.

Taking the product over clique potentials and normalizing, we obtain the usual joint probability distribution associated with the graph:

$$p(x \,|\, \theta) = \frac{1}{Z(\theta)} \prod_{C \in \mathcal{C}} \psi_C(x_C) \tag{19.4}$$

$$= \frac{1}{Z(\theta)} \prod_{C \in \mathcal{C}} \exp\left\{\sum_{i \in \mathcal{I}_C} \theta_i f_i(x_{C_i})\right\} \tag{19.5}$$

$$= \frac{1}{Z(\theta)} \exp\left\{\sum_{C \in \mathcal{C}} \sum_{i \in \mathcal{I}_C} \theta_i f_i(x_{C_i})\right\}. \tag{19.6}$$

This representation makes explicit the association of features to maximal cliques. In many cases, however, we may not need to make this association explicit, and we can simplify our representation

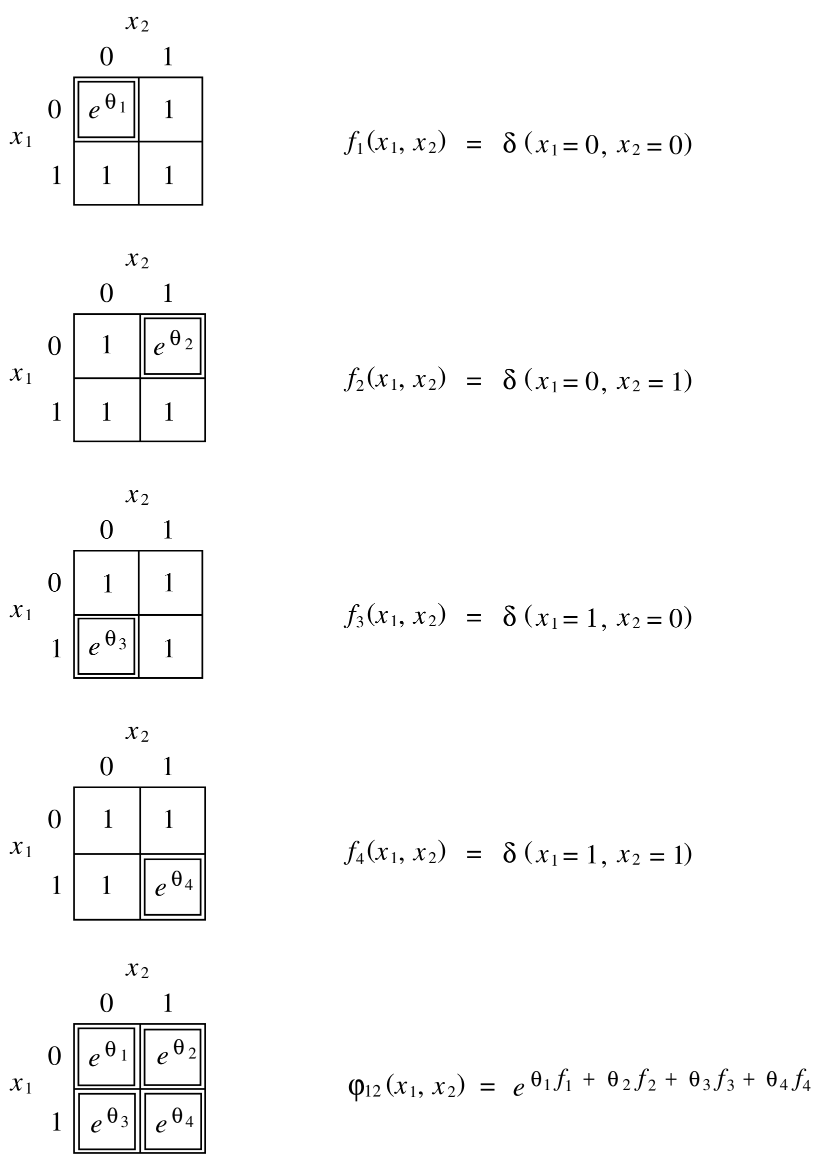

Figure 19.1: The representation of a clique potential using binary "indicator" features. The first four diagrams show potentials with a single degree of freedom corresponding to a single cell. The potential in each case is defined as $e^{\theta_i f_i}$. The final diagram shows a fully parameterized clique potential formed by the product of the four one-degree-of-freedom potentials.

by simply summing over all features associated with a given graph, irrespective of their association with maximal cliques. Thus, consider a set of features $\{f_i(x_{C_i})\}$, for i in an index set $\mathcal{I}$, where C_i is the set of nodes referred to by feature f_i. We have:

$$p(x \mid \theta) = \frac{1}{Z(\theta)} \exp\left\{ \sum_{i \in \mathcal{I}} \theta_i f_i(x_{C_i}) \right\}. \tag{19.7}$$

We see that the featural representation is nothing but the exponential family representation from Chapter 8. The "features" are the sufficient statistics in the exponential family representation. Thus, in working with "features," as we do in this chapter, we are back on familiar terrain.

19.1.1 Graph first or features first?

Thus far we have assumed that a graph has been specified and our task is that of finding a parameterization for that graph. This reduces to finding a parameterization for the maximal cliques in the graph, which, as we have shown, can be achieved using a featural representation.

On the other hand, one might approach a modeling problem with a given set of features and Eq. (19.7) in mind, quite apart from any explicit graphical representation. Indeed, in the literature on exponential family models (also known as *loglinear models*), this is quite commonly done.

Given a set of features and an exponential family model, one can of course always construct a graphical representation. In particular, given features $f_i(x_{C_i})$, construct a graph by linking any pair of nodes that appear together in a subset C_i for some feature i. Let $C \in \mathcal{C}$ range over the maximal cliques of the resulting graph. Each subset C_i is contained in (at least) one such maximal clique by construction; pick one such clique C and associate C_i to that C. Define the potential $\psi_C(x_C)$ to be the product of the factors $\exp\{\theta_i f_i(x_{C_i})\}$, for all C_i associated with C. Given that each feature appears only once, the usual definition of the joint probability for an undirected graph:

$$p(x \mid \theta) = \frac{1}{Z(\theta)} \prod_{C \in \mathcal{C}} \psi_C(x_C) \tag{19.8}$$

is equivalent to Eq. (19.7).

This reduction to a graphical model is useful for a number of purposes, including defining junction-tree-based exact inference algorithms for an exponential family model. Yet it also involves a loss of information. Let us return to our orthography example. The presence of an `ing` feature implies a dependence between the three corresponding nodes in the model. Our graph-construction procedure would accordingly add links between these three nodes. We obtain a clique on three nodes, and we have no way of distinguishing this model from an alternative model in which all 26^3 configurations in this clique are independently parameterized—the featural representation has disappeared.

Indeed, for short strings of letters we are very likely to want to use a few features that span all of the letters in the string. Any such features lead to all nodes being connected, a rather uninformative graphical model.

[Something on the alternative, "factor graph" representation of features.]

19.2 Maximum entropy and maximum likelihood

In Chapter 8 we introduced the exponential family of distributions, motivating the family as a natural generalization of the Bernoulli, Gaussian, and Poisson distributions. In this section we provide a stronger justification for referring to this as a "natural generalization." We show that the exponential family can be motivated as the expression of a variational principle—the principle of *maximum entropy*.

The maximum entropy formalism also provides us with new insights into the problem of fitting a model to data. From the maximum entropy point of view this problem is treated as a constrained optimization problem, in which we place constraints on expectations of a set of "features," but otherwise make no specific parametric assumptions regarding the form of the model. It is a *consequence* of the variational principle that we obtain a specific parametric form, in particular the exponential family form. The parameters in the exponential family representation are Lagrange multipliers that enforce the constraints.

Although the insights afforded by the maximum entropy point of view are interesting and important, it is also important to understand that the resulting procedure is the same as the procedure that we obtain from the maximum likelihood point of view. Maximum entropy and maximum likelihood are *dual problems*, yielding the same optimum (there is no "duality gap"). Maximum entropy simply provides an alternative, dual point of view on the problem of maximum likelihood estimation in the exponential family.

In this section, we develop these ideas in detail, beginning by recalling the maximum likelihood problem and its solution, then stating and solving the maximum entropy problem, and finally proceeding to the identification of the maximum likelihood and maximum entropy as dual problems. We work with discrete variables throughout for simplicity, but all of our results also go through for continuous variables.

19.2.1 Maximum likelihood

In Chapter 8 we presented the general solution to the maximum likelihood problem for exponential family models. The reader can turn back to that chapter to retrieve the result, but it is just as easy to rederive it. From Eq. (19.7) we have the following log likelihood:

$$
\begin{aligned}
l(\theta; \mathcal{D}) &= \sum_x m(x) \log p(x \,|\, \theta) && (19.9) \\
&= \sum_x m(x) \left(\sum_i \theta_i f_i(x) - \log Z(\theta) \right) && (19.10) \\
&= \sum_x m(x) \sum_i \theta_i f_i(x) - N \log Z(\theta), && (19.11)
\end{aligned}
$$

where $m(x)$ is the number of occurences of configuration x in the data set $\mathcal{D}$, and where we have lightened our notation somewhat by suppressing the dependence of the sufficient statistics on the subsets C_i.

Recalling that the derivatives of $\log Z(\theta)$ with respect to θ_i are the expectations of the corresponding sufficient statistics, we obtain:

$$\frac{\partial l}{\partial \theta_i} = \sum_x m(x) f_i(x) - N \frac{\partial}{\partial \theta_i} \log Z(\theta) \tag{19.12}$$

$$= \sum_x m(x) f_i(x) - N \sum_x p(x \mid \theta) f_i(x). \tag{19.13}$$

This yields the following characterization of maximum likelihood estimates:

$$\sum_x p(x \mid \theta) f_i(x) = \sum_x \tilde{p}(x) f_i(x), \tag{19.14}$$

where $\tilde{p}(x) \triangleq m(x)/N$ is the empirical distribution. We see that the marginals of the sufficient statistics (the "features") must be equal to the empirical marginals. This provides an implicit set of equations for the parameters θ_i.

As an aside, recall that in Chapter 9 we discussed the problem of finding maximum likelihood estimates in the setting of general undirected graphical models. The characterization of maximum likelihood estimates in that setting was the following—for each maximal clique in the graph, the marginals of the random variables in the clique must be equal to the empirical marginals. We now see that this result is, unsurprisingly, a consequence of the more general result in Eq. (19.14). Marginal probabilities can be viewed as expectations of indicator variables. If we use features that are indicator variables that pick out the cells in the maximal cliques, as in Figure 19.1, then Eq. (19.14) implies that the clique marginals must equal the empirical marginals.

19.2.2 Maximum entropy

Let us now consider the maximum entropy formulation. We begin with a random vector X and with a set of "features" $f_i(x)$, where $i \in \mathcal{I}$ for an index set $\mathcal{I}$. Rather than assuming a specific functional form for the distribution of X, we instead specify a probability distribution for X indirectly, by specifying constraints that such a distribution must satisfy. In particular, we require that the expectations of the features f_i be equal to particular given constants α_i:

$$\sum_x p(x) f_i(x) = \alpha_i, \tag{19.15}$$

The unknowns here are the probabilities $p(x)$. Note that the constraints are linear in the unknowns.

We ask that a distribution satisfy all of these constraints. In general, of course, it is not clear that such a distribution exists, but for now we skirt that problem by assuming the existence of at least one such distribution. In essence, we assume that our constraints are consistent. We return to the problem of consistency later in this section.

There may be many distributions that satisfy the constraints, and thus we make an additional requirement—among those distributions that satisfy the constraints we choose the distribution that has maximum entropy. Actually, we will generalize somewhat and ask for a distribution that has minimum Kullback-Leibler divergence with respect to a given reference distribution $h(x)$. (If $h(x)$

is the uniform distribution we obtain the maximum entropy problem). We thus have the following constrained optimization problem:

$$\begin{aligned} \min \quad & D(p \parallel h) \triangleq \sum_x p(x) \log \frac{p(x)}{h(x)} && (19.16) \\ \text{subject to} \quad & \sum_x p(x) f_i(x) = \alpha_i && (19.17) \\ & \sum_x p(x) = 1, && (19.18) \end{aligned}$$

where the final constraint embodies our desire to obtain a probability distribution.

To solve this problem, we form the Lagrangian:

$$\mathcal{L} = \sum_x p(x) \log \frac{p(x)}{h(x)} - \sum_i \theta_i \left(\sum_x p(x) f_i(x) - \alpha_i \right) - \mu \left(\sum_x p(x) - 1 \right), \tag{19.19}$$

and take the derivative of $\mathcal{L}$ with respect to the unknowns $p(x)$:

$$\frac{\partial \mathcal{L}}{\partial p(x)} = 1 + \log p(x) - \log h(x) - \sum_i \theta_i f_i(x) - \mu. \tag{19.20}$$

Setting to zero and rearranging yields:

$$p(x) = e^{\mu - 1} h(x) \exp\left\{ \sum_i \theta_i f_i(x) \right\}. \tag{19.21}$$

Summing over x must yield one, and thus:

$$e^{-(\mu-1)} = \sum_x h(x) \exp\left\{ \sum_i \theta_i f_i(x) \right\} \tag{19.22}$$

is the normalization factor. Letting $Z(\theta) \triangleq e^{-(\mu-1)}$ denote this normalization factor, we obtain:

$$p(x \mid \theta) = \frac{1}{Z(\theta)} h(x) \exp\left\{ \sum_i \theta_i f_i(x) \right\}. \tag{19.23}$$

This is of course the exponential family distribution. We have derived the exponential family distribution as an expression of the maximum entropy principle. The features $f_i(x)$ are the sufficient statistics, and the Lagrange multipliers θ_i are the canonical parameters.

Where do the values α_i come from? The variational principle itself does not specify any particular source of the α_i, and indeed there are applications of maximum entropy in optimization theory in which the α_i are simply boundary conditions that come from prior knowledge. In our application, however—a statistical application—we need to make the α_i depend in some way on the data. Noting that Eq. (19.15) sets the expectations of the features equal to α_i, it is natural to choose the α_i to be equal to the empirical expectations of the features. This is the "method of moments," an estimation method that is motivated by the law of large numbers. Given this assumption we have:

$$\alpha_i = \sum_x \tilde{p}(x) f_i(x), \tag{19.24}$$

and thus our optimization problem becomes:

$$\begin{aligned} \min \quad & D(p \parallel h) \triangleq \sum_x p(x) \log \frac{p(x)}{h(x)} && (19.25) \\ \text{subject to} \quad & \sum_x p(x) f_i(x) = \sum_x \tilde{p}(x) f_i(x) && (19.26) \\ & \sum_x p(x) = 1; && (19.27) \end{aligned}$$

a convex minimization problem under linear constraints.

With this form of the constraints we have also solved our consistency problem. There clearly is at least one distribution $p(x)$ that is consistent with the constraints—the empirical distribution $\tilde{p}(x)$. Thus there must exist a solution to the maximum entropy problem.

Moreover, that solution must be unique. As we ask the reader to show in Exercise **??**, the KL divergence is strictly convex in the variables $p(x)$. The problem of minimizing a strictly convex function with respect to linear constraints has at most a single solution.

19.2.3 Duality

The results in the previous two sections suggest that the maximum likelihood problem and the maximum entropy problem are very closely related. In one case (maximum likelihood) we assume the exponential family distribution and show that the model expectations must equal the empirical expectations. In the other case (maximum entropy) we assume that the model expectations must equal the empirical expectations and show that we must use an exponential family distribution.

The relationship turns out to be one of *duality*—the maximum likelihood problem and the maximum entropy problem are Lagrangian duals. We prove this fact in this section and also provide a geometric interpretation of the result.

A dual function is obtained by solving for the primal variables $p(x)$ and substituting the solution back into the Lagrangian $\mathcal{L}$. (See Appendix XXX for a review of Lagrangian duality). This yields a function $g(\theta)$ that we must maximize.

Returning to Eq. (19.19), we plug Eq. (19.23) into the Lagrangian. We drop the term involving μ; given that Eq. (19.23) is explicitly normalized this term must be zero. We have:

$$\begin{aligned} g(\theta) &= \sum_x p(x \mid \theta) \log \frac{p(x \mid \theta)}{h(x)} - \sum_i \theta_i \left(\sum_x p(x \mid \theta)(f_i(x) - \sum_x \tilde{p}(x) f_i(x) \right) \\ &= \sum_x p(x \mid \theta) \left(\sum_i \theta_i f_i(x) - \log Z(\theta) \right) - \sum_i \theta_i \sum_x p(x \mid \theta) f_i(x) + \sum_i \theta_i \sum_x \tilde{p}(x) f_i(x) \end{aligned}$$

$$= \sum_x \tilde{p}(x) \sum_i \theta_i f_i(x) - \log Z(\theta). \tag{19.28}$$

Up to a scaling factor N, this expression is simply the log likelihood (cf. Eq. 19.11). Thus we see that maximum likelihood is dual to maximum entropy.

For convex problems it is a fact of optimization theory that there is no duality gap between the primal and dual problems (see Appendix XXX. Actually there is a technicality here, which is discussed in the Appendix—we require a regularity condition such as the Slater condition). That is, the primal problem and the dual problem must yield the same value of the objective. Indeed, we have already seen, in Eq. (19.14), that maximum likelihood sets the parameters θ_i such that the expectations of the features are equal to the empirical expectations. This is the same constraint that defines the maximum entropy problem, and both problems therefore pick out the same exponential family distribution.

Geometric interpretation

We now show that the duality result can be given an appealing geometric interpretation.

Let us define two subsets of the simplex of probability distributions $\{p(x)\}$. The first will be the subset $\mathcal{E}$ of all exponential family distributions based on a given set of features $\{f_i(x)\}$ and a given base measure $h(x)$:

$$\mathcal{E} = \left\{ p(x) \;:\; p(x) = \frac{1}{Z(\theta)} h(x) \exp\left\{ \sum_i \theta_i f_i(x) \right\} \right\}. \tag{19.29}$$

Note that any two distributions in this family are related by a multiplicative factor. That is, given a distribution $p(x \mid \theta_M)$ and a distribution $p(x \mid \theta)$, we have:

$$p(x \mid \theta) = \frac{Z(\theta_M)}{Z(\theta)} \exp\left\{ \sum_i \Delta\theta_i f_i(x) \right\} p(x \mid \theta_M), \tag{19.30}$$

where $\Delta\theta \triangleq \theta - \theta_M$.[1]

The second subset that we define is the set $\mathcal{M}$ of all distributions that meet the moment constraints:

$$\mathcal{M} = \left\{ p(x) \;:\; \sum_x p(x) f_i(x) = \sum_x \tilde{p}(x) f_i(x) \right\}. \tag{19.31}$$

Note that in general the distributions in $\mathcal{M}$ are not exponential family distributions. Indeed, our results in the previous section have shown that there is a single distribution that is in $\mathcal{M}$ and is also in $\mathcal{E}$. We denote this distribution, the maximum entropy or maximum likelihood distribution, as $p(x \mid \theta_M)$, or p_M for short.

We thus have the geometry suggested in Figure 19.2; subsets $\mathcal{M}$ and $\mathcal{E}$ that meet at the single point p_M.

[1] Taking logarithms, it is interesting to note that this can be viewed as an affine transformation on a log scale, where f_i are the "coordinates."

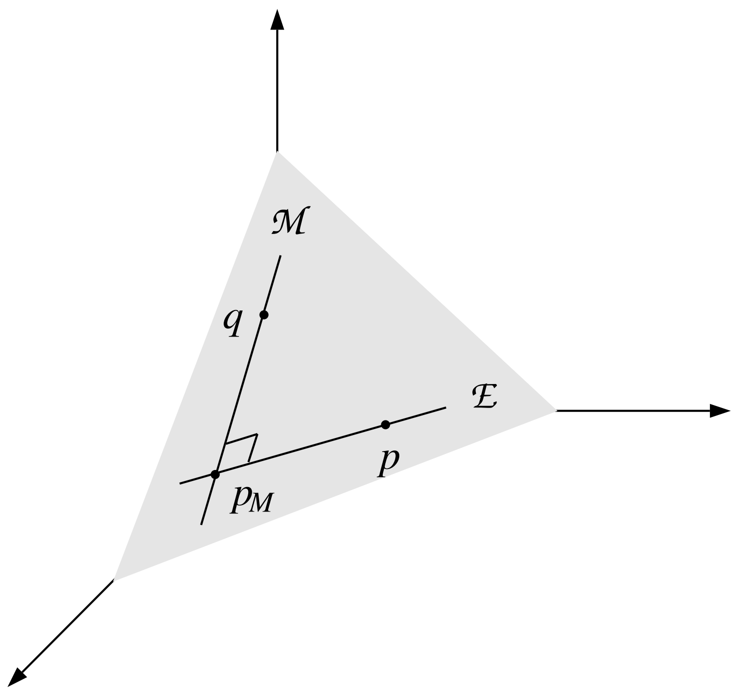

Figure 19.2: The set $\mathcal{E}$ of exponential family distributions and the set $\mathcal{M}$ of distributions that meet the moment constraints lie in a probability simplex and intersect at the point p_M, the maximum likelihood or maximum entropy point. A generalized Pythagorean theorem for the KL divergence holds among the points p, q, and p_M.

Moreover, there is an interesting "orthogonality" relationship between these subsets. Let p be any other point in $\mathcal{E}$ (an exponential family distribution with parameters θ), and let q be any other point in $\mathcal{M}$. We now prove the following fact:

$$D(q \parallel p) = D(q \parallel p_M) + D(p_M \parallel p), \tag{19.32}$$

which can be viewed as a generalized "Pythagorean theorem."

To prove this theorem, we compute the difference between the left-hand side and the right-hand sides of Eq. (19.32):

$$\begin{aligned}
&D(q \parallel p) - D(q \parallel p_M) - D(p_M \parallel p) \\
&= \sum_x q(x) \log \frac{q(x)}{p(x)} - \sum_x q(x) \log \frac{q(x)}{p_M(x)} - \sum_x p_M(x) \log \frac{p_M(x)}{p(x)} && (19.33)\\
&= \sum_x p_M(x) \log \frac{p(x)}{p_M(x)} - \sum_x q(x) \log \frac{p(x)}{p_M(x)} && (19.34)\\
&= \sum_x p_M(x) \left(\sum_i \Delta\theta_i f_i(x) - \log \frac{Z(\theta_M)}{Z(\theta)} \right) - \sum_x q(x) \left(\sum_i \Delta\theta_i f_i(x) - \log \frac{Z(\theta_M)}{Z(\theta)} \right) && (19.35)\\
&= \sum_i \Delta\theta_i \left(\sum_x p_M(x) f_i(x) - \sum_x q(x) f_i(x) \right) && (19.36)\\
&= 0, && (19.37)
\end{aligned}$$

by the definition of $\mathcal{M}$.

We have derived a Pythagorean theorem by making use of our duality results, but we can also turn the argument around. Let us assume Eq. (19.32), and assume the existence of a point p_M in $\mathcal{E} \cap \mathcal{M}$. Write the maximum entropy problem as follows:

$$\begin{aligned}
&\min && D(q \parallel h) && (19.38)\\
&\text{subject to} && q \in \mathcal{M}. && (19.39)
\end{aligned}$$

From Eq. (19.32) we have:

$$D(q \parallel h) = D(q \parallel p_M) + D(p_M \parallel h). \tag{19.40}$$

The second term is independent of q and can be ignored. Given that the KL divergence—and $D(q \parallel p_M)$ in particular—is greater than or equal to zero, we see immediately that p_M minimizes $D(q \parallel h)$ for $q \in \mathcal{M}$.

Moreover, we can write the maximum likelihood problem succinctly as follows:

$$\begin{aligned}
&\min && D(\tilde{p} \parallel p) && (19.41)\\
&\text{subject to} && p \in \mathcal{E}. && (19.42)
\end{aligned}$$

From Eq. (19.32) we have:

$$D(\tilde{p} \parallel p) = D(\tilde{p} \parallel p_M) + D(p_M \parallel p). \tag{19.43}$$

The first term is independent of p and can be ignored. The second term is minimized by choosing p equal to p_M, which is therefore the maximum likelihood distribution.

Finally, let us introduce a bit of terminology to summarize our results. Given a subset of probability distributions $\mathcal{M}$ defined by moment constraints, and given a fixed reference distribution h, let us refer to the distribution $q \in \mathcal{M}$ that minimizes $D(q \parallel h)$ as the *I-projection* of h on $\mathcal{M}$. Thus the Pythagorean theorem holds for an arbitrary point in $\mathcal{M}$, an arbitrary point h and the I-projection of h on $\mathcal{M}$.

19.3 Summary

In this chapter we have focused on the parameterization of graphical models, emphasizing discrete, undirected graphical models. We have seen how to represent potential functions in terms of collections of "features." In particular, we have shown how indicator features can be used to pick out the cells in potential tables and associate independently-varying parameters with each of these cells. In general, however, we are not restricted to indicator features—we can define a potential in terms of arbitrary collections of features.

The relationship between exponential family models and graphical models is a very close one. If we use an exponential representation for the contribution of each individual feature, then the product of potential functions leads to an exponential family representation for the joint distribution associated with the graphical model. Alternatively, we can also represent an arbitrary exponential family model as a graphical model by connecting nodes that appear together as arguments to the features.

In the final section of the chapter, we showed that the exponential family distribution can be viewed as the expression of a variational principle, namely that of maximum entropy. We also showed that the maximum entropy principle leads to a perspective on parameter estimation that is dual to maximum likelihood.

In the following chapter, we continue our exploration of exponential family parameterizations of graphical models. In particular, we will discuss the problem of parameter estimation in general exponential families, presenting algorithms that exploit graphical structure, and exploring some of the relationships between these algorithms and the inference algorithms discussed in earlier chapters.

19.4 Historical remarks and bibliography