PGMs as Exponential Families
10708, Fall 2020
Pradeep Ravikumar

1 Introduction: Exponential Families

So far, the representational theory for PGMs, specifically UGMs, mainly required that they factor according to a UG G. But we largely did not further specify what form these factors should take. When the factors are all strictly positive, they can be expressed as the exponential of a general real-valued (i.e. not necessarily positive) function. An important class of such factors are linearly parameterized (inside the exponential), these are then simply a class of exponential family distributions.

Let us first recall the basics of exponential family distributions. Consider a general random vector $X = (X_1, \ldots, X_p)$ taking values in $\mathcal{X} = \otimes_{s=1}^{p} \mathcal{X}_s$. Let $\phi = (\phi_\alpha, \alpha \in \mathcal{I})$ be a collection of functions $\phi_\alpha : \mathcal{X} \mapsto \mathbb{R}$ known as sufficient statistics, where $\mathcal{I}$ is the index set of these sufficient statistics, and suppose $|\mathcal{I}| = d$. Associated with these sufficient statistics, suppose we have a set of canonical exponential family parameters $\theta = (\theta_\alpha, \alpha \in \mathcal{I})$. Then, the exponential family of distributions associated with ϕ is the set of distributions with densities:

$$p_\theta(x) = \exp\{\langle \theta, \phi(x) \rangle - A(\theta)\}$$

with respect to some base measure v, and where

$$A(\theta) = \int_{\mathcal{X}} \exp\{\langle \theta, \phi(x) \rangle\} v(x) dx,$$

is the log normalization constant, also known as the log-partition function. Let

$$\Omega = \{\theta \in \mathbb{R}^d \,|\, A(\theta) < \infty\},$$

denote the domain of the set of exponential family parameters that allow for a finite density.

The family is said to be *regular* if the set Ω is open. In all examples we will consider, this will be the case.

The set of sufficient statistics are said to be minimal or the exponential family to have minimal representation if there is no non-zero vector $a \in \mathbb{R}^d$ s.t.

$$\langle a, \phi(x) \rangle = \text{constant},$$

v-almost everywhere. If it is not minimal, then the exponential family is said to have an overcomplete representation.

Now, suppose that each sufficient statistic only depends on a subset of variables that form cliques in some undirected graph G, so that $\text{scope}(\phi_\alpha) \subseteq C_\alpha \in \mathcal{C}(G)$. Then the corresponding exponential family

$$p_\theta(x) = \exp(-A(\theta)) \prod_\alpha \exp(\theta_\alpha \phi_\alpha(x)),$$

can be seen to factor according to the graph G, and hence is a UGM associated with G. Thus, exponential families can be used as parameteric families of UGMs when we set the sufficient statistics to depend only on cliques of a graph.

2 Examples of PGMs as exponential families

Suppose we have a graph $G = (V, E)$ with nodes V associated with each of the random variables $\{X_i\}_{i \in [p]}$.

Ising Model. Here, $X_s \in \mathcal{X}_s \equiv \{-1, 1\}$, and the set of sufficient statistics are:

$$\phi(x) = (x_s, s \in V; x_s x_t, (s,t) \in E) \in \mathbb{R}^{|V|+|E|}.$$

These then specify the exponential family:

$$p_\theta(x) = \exp\left\{ \sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t - A(\theta) \right\}.$$

Since the domains of the variables are finite, the normalization constant is always finite, and hence $\Omega = \mathbb{R}^d$. This is also a minimal representation.

Such *Ising models* first arose in statistical physics to model orientations of magnets in a field, or presence or absence of molecules in a gas. But they can of course be used in any setting where there are a large number of binary variables: for instance to model voting patterns of US senators. The standard variant above only has nodewise and pairwise sufficient statistics. One can generalize this to include higher order monomials. Allowing for monomials up to order k are called k-spin models in statistical physics. Allowing all possible monomials up to order p allows for a factor involving all variables, and hence allows for representing any distribution over $\mathcal{X}$.

Note that the distribution above could be re-written as:

$$p_\theta(x) = \exp(-A(\theta)) \prod_{s \in V} \exp(\theta_s x_s) \prod_{(s,t) \in E} \exp(\theta_{st} x_s x_t).$$

It can be seen to factor over nodes (vacuous clique) and edges (smallest non-vacuous cliques) of the graph, and consequently factors according the graph G i.e. is an undirected graphical model distribution associated with G.

Discrete PGM. Here, suppose $|\mathcal{X}_s| = n_s$, and the set sufficient statistics are:

$$\phi(x) = \{\mathbb{I}[x_s = j]\}_{s\in V, j\in\mathcal{X}_s} \cup \{\mathbb{I}[(x_s, x_t) = (j,k)]\}_{(s,t)\in E,\, j\in\mathcal{X}_s, k\in\mathcal{X}_t}.$$

This can be seen to be an overcomplete representation since the indicator functions for any node, or for any pair of nodes, all sum to one:

$$\sum_j \mathbb{I}[X_s = j] = 1$$
$$\sum_{j,k} \mathbb{I}[(X_s, X_t) = (j,k)] = 1.$$

These specify the exponential family:

$$p_\theta(x) = \exp\left\{ \sum_{s\in V, j\in\mathcal{X}_s} \theta_{s;j}\mathbb{I}[X_s = j] + \sum_{(s,t)\in E,\, s\in\mathcal{X}_s, t\in\mathcal{X}_t} \theta_{st;jk}\mathbb{I}[(x_s, x_t) = (j,k)] - A(\theta) \right\}.$$

It can be seen that allowing for higher order factors will allow for more and more possible distributions over the set of discrete variables.

Here again, it can be seen that the distribution above factors according to the graph G.

Gaussian graphical model. Here, $X_s \in \mathcal{X}_s \equiv \mathbb{R}$, and the set of sufficient statistics are:

$$\phi(x) = (x_s, x_s^2, s\in V;\, x_s x_t, (s,t)\in E),$$

which specifies the exponential family:

$$p_\theta(x) = \exp\left\{ \langle\theta, x\rangle + \frac{1}{2}\langle\!\langle \Theta, xx^T\rangle\!\rangle - A(\theta,\Theta) \right\},$$

where

$$\langle\theta, x\rangle = \sum_{i=1}^p \theta_i x_i$$
$$\langle\!\langle \Theta, xx^T\rangle\!\rangle = \sum_{i=1}^p\sum_{j=1}^p \Theta_{ij}x_i x_j.$$

We thus have two sets of parameters: $\theta \in \mathbb{R}^p$ associated the linear statistics x, and $\Theta \in \mathbb{R}^{p\times p}$ associated with the quadratic statistics xx^T. As we will see in a later section, it can be shown that the parameter Θ is actually the inverse of the covariance matrix of X. It can be seen that the sparsity pattern of Θ corresponds to the edges of the UGM. The domain of the set of parameters is given as:

$$\Omega = \left\{(\theta,\Theta) \in \mathbb{R}^p \times \mathbb{R}^{p\times p} \,|\, \Theta \prec 0, \Theta = \Theta^T\right\}.$$

3 Mean Parameterization

Let P be any distribution over X, not necessarily an exponential family. The mean parameter μ_α associated with the sufficient statistic ϕ_α is given as:

$$\mu_\alpha = \mathbb{E}_{X\sim p}[\phi_\alpha(X)] = \int \phi_\alpha(x)p(x)v(x)dx.$$

The set of these mean parameters $\{\mu_\alpha, \alpha \in \mathcal{I}\}$ then defines the set:

$$\mathcal{M} = \{\mu \in \mathbb{R}^d \,|\, \exists p \text{ s.t. } \mathbb{E}_p[\phi_\alpha(X)] = \mu_\alpha,\ \forall \alpha \in \mathcal{I}\}.$$

Ising Model mean parameters. Recall that the Ising model had sufficient statistics:

$$\phi(x) = (x_s, s \in V; x_s x_t, (s,t) \in E) \in \mathbb{R}^{|V|+|E|}.$$

Suppose we consider the case where the variables take values in $\{0, 1\}$. The associated mean parameters are:

$$\begin{aligned}\mu_s &= \mathbb{E}_p[X_s] = \mathbb{P}[X_s = 1],\ \forall s \in V \\ \mu_{st} &= \mathbb{E}_p[X_s X_t] = \mathbb{P}[(X_s, X_t) = (1,1)],\ \forall (s,t) \in E.\end{aligned}$$

The set of mean parameters $\mathcal{M}$ thus consist of all node-wise and pairwise marginals that can be realized by some distribution over $\{0,1\}^p$. This is also known as the cut polytope, or the correlation polytope.

Discrete Exponential Family mean parameters. The Ising model is a special instance of the more general class of discrete GMs where the domain $\mathcal{X}_s$ of each variable has finite cardinality. The associated space $\mathcal{X}$ of the entire random vector is also finite. Then, given any set of sufficient statistics $\phi : \mathcal{X} \mapsto \mathbb{R}^d$, the set of mean parameters

$$\begin{aligned}\mathcal{M} &= \{\mu \in \mathbb{R}^d \,|\, \mu = \sum_{x\in\mathcal{X}} \phi(x)p(x),\ \text{for some dist. } p \text{ over } \mathcal{X}\} \\ &= \text{conv}\{\phi(x), x \in \mathcal{X}\},\end{aligned}$$

is a convex polytope. By the Minkowski-Weyl theorem, it thus follows that an alternative representation of the set of mean parameters is an intersection of some set $\mathcal{J}$ of halfspaces:

$$\mathcal{M} = \{\mu \in \mathbb{R}^d \,|\, \langle a_j, \mu, \geq \rangle b_j, \forall j \in \mathcal{J}\}.$$

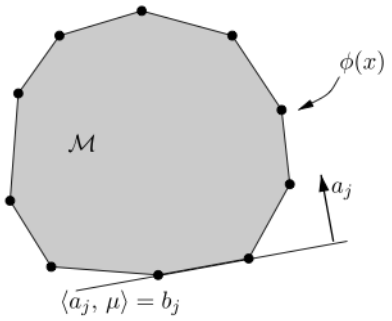

Figure 1: Mean parameter convex polytope for discrete exponential families

Overcomplete Discrete GM mean parameters. With the overcomplete sufficient statistics for a discrete GM, the mean parameters are:

$$\mu_{s;j} = \mathbb{E}_p[\mathbb{I}[X_s = j]] = \mathbb{P}[X_s = j], \ \forall s \in V, j \in \mathcal{X}_s$$
$$\mu_{st;jk} = \mathbb{E}_p[\mathbb{I}[(X_s, X_t) = (j, k)]] = \mathbb{P}[X_s = j, X_t = k], \ \forall (s,t) \in E, j \in \mathcal{X}_s, k \in \mathcal{X}_t,$$

so that the mean parameters correspond to node-wise and pairwise marginals of the discrete random vector. From above, we know that $\mathcal{M}$ is a convex polytope; in this overcomplete case, it is often referred to as the marginal polytope (since it comprise nodewise and pairwise marginals). This will play an important role when we study approximate inference algorithms.

4 Forward Mapping from Canonical to Mean Parameters

We first note the following proposition.

Proposition 1 *The log-partition function $A(\theta)$ is convex, and strictly so if the representation is minimal. Moreover its gradient specifies the mean parameter:*

$$\nabla A(\theta) = \mathbb{E}_{p_\theta}[\phi(X)].$$

It can thus be seen that the gradient mapping ∇A maps a parameter $\theta \in \Omega$ to a mean parameter $\mu \in \mathcal{M}$. It is thus referred to as the "forward" mapping in exponential family UGMs. We can provide some additional properties for this forward mapping.

Proposition 2 *The forward mapping ∇A satisfies the following properties.*

- *∇A is one-to-one iff the exponential family representation is minimal.*

- *It is onto the interior int($\mathcal{M}$), so that for all $\mu \in$ int($\mathcal{M}$), there exists a unique exponential family parameter $\theta(\mu) \in \Omega$ such that $\mathbb{E}_{p_\theta(\mu)}[\phi(X)] = \mu$ i.e. $(\nabla A)^{-1}$ is well-defined on int($\mathcal{M}$).*

Note that the set of exponential family distributions is only a subset of the set all possible distributions, and the set of mean parameters $\mathcal{M}$ can be achieved by taking expectation of the sufficient statistics with respect to any possible distribution. So for any μ, there could be many distributions p that give rise to those mean parameters. But the proposition above notes that there is a unique exponential family distribution to do so.

Importance in context of PGMs. Note that the mean parameters in the case of pairwise PGMs exactly corresponds to the node-wise and pairwise marginals (and more generally, to clique-wise marginals). This is precisely the objective in graphical model inference.

5 Conjugate Duality

Given the log-normalization function A, its conjugate dual, denoted by A^* is given by:

$$A^*(\mu) = \sup_{\theta \in \Omega}(\langle \theta, \mu \rangle - A(\theta)). \tag{1}$$

Here $\mu \in \mathbb{R}^d$ is a set of dual parameters, of the same dimension as θ.

Theorem 3 *For any $\mu \in$ int($\mathcal{M}$), the unique optimal solution to the conjugate dual problem in* (1) *is given by:*

$$\theta(\mu) = (\nabla A)^{-1}(\mu),$$

and the conjugate dual function A^* takes the form:

$$A^*(\mu) = -H(p_{\theta(\mu)}) = \mathbb{E}_{p_\theta(\mu)}[\log p_{\theta(\mu)}(X)].$$

Thus, the conjugate dual for a mean parameter in the interior is simply the negative entropy of the associated exponential family distribution. Note that we can take the conjugate dual of the conjugate dual in turn. Since the function A is convex, this yields back A again.

Theorem 4 *The log-partition function A has the variational representation:*

$$A(\theta) = \sup_{\mu \in \mathcal{M}}(\langle \theta, \mu \rangle - A^*(\mu)). \tag{2}$$

For all $\theta \in \Omega$, the optimum above is attained at $\mu = \nabla A(\theta) \in$ int($\mathcal{M}$).

It can be shown that $\nabla A^* = (\nabla A)^{-1}$ when restricted to $\text{int}(\mathcal{M})$. Thus, the forward mapping ∇A maps an exponential family parameter $\theta \in \Omega$ to an interior mean parameter $\mu \in \text{int}(\mathcal{M})$. While the backward mapping $\nabla A^* = (\nabla A)^{-1}$ maps an interior mean parameter $\mu \in \text{int}(\mathcal{M})$ to an exponential family parameter $\theta \in \Omega$.

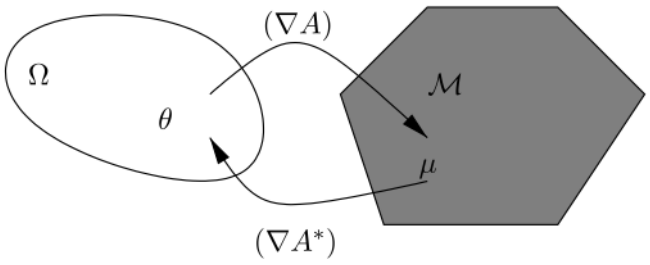

Figure 2: Forward and Backward Mappings for exponential families

Importance in context of PGMs. As we will see later on in the class, the backward mapping will allow us to *learn* PGMs from data, which can be summarized via empirical expectations $\widehat{\mu}$ of the sufficient statistics. The backward mapping $\widehat{\theta} = \nabla A^*(\widehat{\mu})$ will then provide us the unique exponential family parameters corresponding to these empirical sufficient statistics.

The caveat of course is that both the forward and backward mappings are in general intractable for general exponential family PGMs. We would thus need to approximate these somehow. This is where the variational characterization of both the log-partition function as well as its conjugate dual will come in handy, as we will see in the sequel.