IFT 6269: Probabilistic Graphical Models **Fall 2021**

Lecture 19 — November 12

Lecturer: Simon Lacoste-Julien Scribe: Jarrid Rector-Brooks, Mohammad-Hadi Sotoudeh

Disclaimer: These notes have only been lightly proofread.

19.1 Exponential Families

Despite often looking very different in regards to their forms, many of the probability distributions we've encountered thus far in class are actually families of distributions, and more specifically they are families we call an **Exponential Family**. We may find it more natural at first to use the more familiar definitions of the distributions. However, we will see that the form of the exponential family indeed provides us an simple and natural way to manipulate the distributions. Often, it will be easier for us to work within the framework of the exponential family than with the specific definition of the distribution itself.

19.1.1 Definitions

We'll begin with some preliminaries, and then move to more central definitions. Assume we have some distribution defined on a space $\mathcal{X}$.

Definition 19.1 *A **statistic** is any function ϕ defined on a random variable X.*

Definition 19.2 *A statistic $T(X)$ is a **sufficient statistic** for a parametric distribution p_θ defined on $\mathcal{X}$ if θ is conditionally independent of X given $T(X)$, i.e.,*

$$\theta \perp\!\!\!\perp X \mid T(X)$$

Note that the above definition of a sufficient statistic assumes a Bayesian point of view where we view θ itself is a random variable. Effectively, this definition says that there's no additional information in the observed data X for us to infer anything about θ than what is in $T(X)$. One may view sufficient statistics from a frequentist point of view as well, and their equivalence is shown in the Fisher-Neyman factorization theorem.[1] We will use sufficient statistics to help us define exponential families.

Definition 19.3 *Let X be a random variable on $\mathcal{X}$. An **exponential family** is a parametric family of distributions of the form*

$$p(x;\theta)d\mu(x) = h(x)\exp\left\{b(\theta)^T T(x) - A(\theta)\right\} d\mu(x) \tag{19.1}$$

where

[1] https://en.wikipedia.org/wiki/Sufficient_statistic#Fisher-Neyman_factorization_theorem

1. *$h(x)d\mu(x)$ is the **reference measure**. It allows us to ensure that any observation x is in the correct space. In particular, $d\mu(x)$ is termed the **base measure** and can be the Lebesgue measure in the case of a continuous R.V., or a counting measure for a discrete R.V.. $h(x)$ is the **reference density**. It has no constraints besides being non-negative*[2].

2. *$T(x)$ is the **sufficient statistic**, as defined above.*

3. *$\eta = b(\theta)$ is a parameter vector, termed the **canonical parameter** of the family*

4. *θ is the **parameter** of the family. However, we will usually end up working with η, not θ (as discussed below, many families have that $\eta = \theta$ and so the θ disappears altogether).*

5. *$A(\theta)$ is the **log-partition** function that ensures the density is normalized to 1. This is also often called the **log-normalization** or **cumulant generating** function. This is a very important function and it will make our lives much easier. E.g., successive gradients of this function yield cumulants of the family (this is why $A(\theta)$ is sometimes called the cumulant generating function).*

It is useful to note here that the exponential family does not refer to only one family of distributions. Instead, there are many different exponential families. Each family may be defined by choosing a form for the reference density $h(x)$ and the sufficient statistics $T(x)$, as well as the image of X — Ω_X. Then the members of this exponential family will be indexed by the valid values of the parameter θ.

Remark 19.1.1 *If $\Omega_{\mathcal{X}}$ is discrete, then $p(x;\eta)$ is a probability mass function. For continuous $\Omega_{\mathcal{X}}$, $p(x;\eta)$ is a probability density function.*

We continue now with more definitions.

Definition 19.4 *A **canonical exponential family** is an exponential family where $\theta = b(\theta) = \eta$. This allows us to write the density as*

$$p(x;\eta) = h(x)\exp\left\{\eta^T T(x) - A(\eta)\right\} d\mu(x) \tag{19.2}$$

Definition 19.5 *The **domain** of an exponential family is the set of canonical parameters Ω for which the log-partition function is finite, i.e.,*

$$\Omega = \{\eta \in \mathbb{R}^p : A(\eta) < \infty\}$$

[2] If curious, the following post on mathexchange may be of use in understanding the reference measure more fully: https://math.stackexchange.com/questions/1489330/effect-of-the-measure-on-an-exponential-family

Definition 19.6 *A **minimal** exponential family is an exponential family whose sufficient statistic vector $T(x)$ is linearly independent.*

From here on out we'll assume that we are working with a canonical exponential family and make use of the canonical parameter η. Now, given what we've discussed so far we can now take a closer look at the form of our log-partition function $A(\eta)$.

Proposition 19.7 *We will show that $A(\eta)$ is indeed the log-partition[3] function, i.e., that*

$$A(\eta) = \log \int_{\mathcal{X}} h(x) \exp\left\{\eta^T T(x)\right\} d\mu(x) \tag{19.3}$$

Proof *From the definition of an exponential family we have that*

$$\begin{aligned} 1 &= \int_{\mathcal{X}} p(x;\eta) d\mu(x) \\ &= e^{-A(\eta)} \int_{\mathcal{X}} \exp\left\{\eta^T T(x)\right\} d\mu(x) \end{aligned}$$

This leads us immediately to our desired result. ■

Note that any *single* distribution $p(x)$ may be put trivially into an exponential family by letting $h(x) = p(x)$ and using $\eta = 0$ for any $T(x)$. Thus it does not make sense to say that a single distribution is in the exponential family or not. All we can talk meaningfully about is whether a family of distributions can be an exponential family or not (see below for examples of families which are not exponential families).

19.1.2 Flat vs Canonical Exponential Families

We will see later that derivatives of the log-partition function $A(\eta)$ yield sequential cumulants of the distribution itself. In particular, we will see that the second derivative of $A(\eta)$, $\frac{\partial^2 A(\eta)}{\partial \eta^2}$, turns out to be the variance of the distributions in the family. We've seen in earlier lectures that the variance is non-negative, which implies that $\frac{\partial^2 A(\eta)}{\partial \eta^2} \geq 0$, and $A(\eta)$ is convex. As the domain of the family Ω is defined by the finite elements of $\mathbb{R}^p$ according to $A(\cdot)$, we have that Ω itself is a convex set.

We now proceed with definitions of flat and canonical exponential families.

Definition 19.8 *A **flat** exponential family is an exponential family whose domain Ω is convex.*

[3]The terminology of partition function has its roots in statistical mechanics. It refers, generally, to the normalization constant of a probability distribution. For further context on this terminology, see https://en.wikipedia.org/wiki/Partition_function_(statistical_mechanics)#Connection_to_probability_theory

Let us consider a reparameterization of a subset of a flat family by defining a new transformation $\eta : \Theta \to \Omega$. η is our new parameterization function, and it maps from the new parameter space Θ to the domain of the family Ω. Using such a parameterization, we end up with

$$p(x;\theta) \triangleq p(x;\eta(\theta)) \;\; \forall \theta \in \Theta$$

We are now ready to define curved exponential families.

Definition 19.9 *Let $\eta : \Theta \to \Omega$ be a function mapping a new parameter space Θ to the domain of a family Ω. A **curved** exponential family is an exponential family for which $\eta(\Theta)$ yields a curved manifold in Ω.*

To see an example we could look to a Gaussian parameterized as $\mathcal{N}(\mu, \mu^2)$. If we plotted with μ on the x-axis and σ^2 on the y-axis for this family of distributions, we would get a parabola. This indicates a curved manifold in Ω, and as such a curved exponential family. (Note that the canonical parameter for a 1d-Gaussian is the precision, i.e. the inverse variance, but this would not change the fact that $\eta(\Theta)$ is still a curved manifold in this case)

19.1.3 Examples of Families which are not Exponential Families

Before proceeding further, we provide a few examples of families of distributions which turn out not to be exponential. First, although products of exponential families turn out to be exponential families, the same cannot be said for mixtures of exponential families. Take a mixture of two exponential families, we have (with $0 \leq \alpha \leq 1$)

$$\alpha h_1(x) \exp\left(\eta_1^T T_1(x) - A_1(\eta_1)\right) + (1-\alpha) h_2(x) \exp\left(\eta_2^T T_2(x) - A_2(\eta_2)\right)$$

This generally cannot be factorized into a new family of the form $h_3(x) \exp\left(\eta_3^T T_3(x) - A_3(\eta_3)\right)$. A classic example of this is the mixture of gaussian distributions.

Another example of a family of distributions not in the exponential family is the continuous uniform distribution $Uniform(0, \theta)$. To see this, observe that for any exponential family, the support of distributions of the family must be defined only in terms of $h(x)$. Recall that the density of a distribution in a (canonical) exponential family is $h(x) \exp\left\{\eta^T T(x) - A(\eta)\right\}$. Since the exponential function is generally non-zero, the support of a distribution in an exponential family must necessarily be defined as $\text{supp}(p) = \{x : h(x) > 0\}$. However, for $Uniform(0, \theta)$ the support depends on the parameter θ which contradicts our requirement for exponential families. Hence $Uniform(0, \theta)$ is not an exponential family.

19.1.4 Properties of $A(\eta)$

As mentioned before $A(\eta)$ is often called the cumulant generating function. This is because successive derivatives of $A(\eta)$ yield respectively ordered cumulants of the distribution. We'll

first prove that the first derivative of the log-partition function is the first cumulant – the mean, i.e., $A(\eta) = \mathbb{E}_{p(x;\eta)}[T(x)]$. We have the following

$$\begin{aligned}
\nabla_\eta A(\eta) &= \frac{\partial A(\eta)}{\partial \eta} \\
&= \frac{\partial}{\partial \eta}\left(\log \int \exp\left\{\eta^T T(x)\right\} h(x) d\mu(x)\right) \\
&= \frac{\int T(x) \exp\{\eta^T T(x)\} h(x) d\mu(x)}{\int \exp\{\eta^T T(x)\} h(x) d\mu(x)} \\
&= \int T(x) \exp\left\{\eta^T T(x) - A(\eta)\right\} h(x) d\mu(x) \\
&= \mathbb{E}_{p(x;\eta)}[T(x)]
\end{aligned}$$

We can follow a similar line of reasoning to show that the second derivative of $A(\eta)$ is equal to the distribution's second cumulant – the variance.

$$\begin{aligned}
\frac{\partial^2 A}{\partial \eta^2} &= \int T(x) \exp\{\eta^T T(x) - A(\eta)\}(T(x) - a'(\eta)) h(x) d\mu(x) \\
&= \int T(x) \exp\{\eta^T T(x) - A(\eta)\}(T(x) - \mathbb{E}_{p(x;\eta)}[T(x)]) h(x) d\mu(x) \\
&= \int T(x)^2 \exp\{\eta^T T(x) - A(\eta)\} h(x) d\mu(x) - \mathbb{E}_{p(x;\eta)}[T(x)] \int T(x) \exp\{\eta^T T(x) - A(\eta)\} h(x) d\mu(x) \\
&= \mathbb{E}_{p(x;\eta)}[T(x)^2] - \mathbb{E}_{p(x;\eta)}[T(x)]^2 \\
&= Var_{p(x;\eta)}[T(x)]
\end{aligned}$$

Further derivates of A will lead to further cumulants. The third derivative will yield the third central moment, and so on (though, it's useful to note that further cumulants for $n > 3$ don't necessarily equal the nth central moment).

As we mentioned before, the above result means that $A(\eta)$ is convex (as if the second derivative of a function is non-negative, then the function is convex, and the variance is non-negative). Since $\Omega = \{\eta : A(\eta) < \infty\}$, we also have that Ω is convex.

Given the above results, we call the define the gradient of A as

$$\nabla_\eta A(\eta) = \mathbb{E}_{p(x;\eta)}[T(x)] \triangleq \mu(\eta)$$

$\mu(\eta)$ exists for η in the interior of Ω, and we call $\mu(\eta)$ the *moment vector*. It turns out, as well, that the hessian of A is the covariance of $T(x)$. The proof of this is left as an exercise to the reader, but the final result shows that

$$\left(\frac{\partial^2 A(\eta)}{\partial \eta_i \partial \eta_j}\right)_{i,j} = \mathbb{E}_{p(x;\eta)}\left[(T(x) - \mu(\eta))(T(x) - \mu(\eta))^T\right] = Cov(T(x))$$

Note that, for exponential families, the moment matching estimate is *not* always equivalent to the maximum likelihood estimate (it's only equivalent when you use the $T(x)$ for the moments). For example, consider the gamma distribution $\Gamma(\alpha, \beta)$. The sufficient statistic for this distribution is $T(x) = \begin{bmatrix} \log x \\ x \end{bmatrix}$. Running moment matching with $\tilde{T}(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$ will end up giving a different estimate than would maximum likelihood.

19.1.5 Examples of Families which are Exponential Families

To illustrate our definition of an exponential family more clearly, we'll now show some examples of families which are exponential families.

Example 19.1.1 ***Multinoulli Distribution***

Let X be a random variable distributed as $X \sim Multinoulli(\pi)$. We have that $\mathcal{X} = \{0, 1\}^k$, and that the sample space Ω_X is the set of one-hot encodings, i.e., $\Omega_X = \Delta_k \cap \mathcal{X}$. We have that $\pi \in \Delta_k$ and that $\pi_i \geq 0 \; \forall i$. We will view π as our parameter for the exponential family θ and try to factorize the family according to the exponential family form. To do so, we'll use the trick of applying the exponential to the logarithm of the likelihood. This trick turns out to be widely applicable when trying to show that a family is an exponential family, and it'll turn out to be so here. Let's jump in,

$$\begin{aligned} p(x; \pi) &= \prod_{j=1}^{k} \pi_j^{x_j} \\ &= \exp\left(\sum_{j=1}^{k} x_j \log \pi_j \right) \\ &= \exp\left(\left[\sum_{j=1}^{k} x_j \log \pi_j \right] - 0 \right) \end{aligned}$$

From this, we can effectively read off the values for $h(x), \eta, T(x), d\mu(x)$, and $A(\eta)$. We find that

1. $\eta_j(\pi) = \log \pi_j$
2. $T(x) = x$
3. $d\mu(x)$ is the counting measure on $\mathcal{X}$
4. $h(x) = \mathbb{1}\{x \in \Omega_X\} = \mathbb{1}\{x \in \Delta_k \cap \mathcal{X}\}$
5. $A(\eta(\pi)) = 0 \;\; \forall \pi \in \Theta$

At first glance this looks great! But if we look closer, we see that $A(\eta(\pi)) = 0$, which is rather bad news as we know that successive derivatives of $A(\eta(\pi))$ yield the moments for distributions in this family. $A(\eta(\pi)) = 0$ suggests that all the moments of the distributions are 0 as well! This doesn't quite feel right. Perhaps we can refactorize the distribution another way such that $A(\eta(\pi)) \neq 0$.

Observe that $\Theta = Int(\Delta_k)$ where $Int(\cdot)$ is denotes the interior of a space. This means that Θ is of dimension $k-1$, that $\eta(\Theta)$ is of dimension $k-1$, and Ω_X is of dimension k. Recalling our definition of a minimal exponential family, we can see that one dimension of our sufficient statistics will be linearly dependent and as such we do not have a minimal exponential family. Indeed, for any x such that $h(x) > 0$ we can write $T_k(x) = 1 - \sum_{j=1}^{k-1} T_j(x)$. This suggests that we could reparameterize one of the sufficient statistics $T_j(x)$ in terms of all the other $T_i(x) s.t. i \neq j$. So this means that there are multiple η's which give rise to the exact same distribution (in particular, $\eta + \alpha\mathbf{1}$ for any α yields the same distribution as η). Given this, we can see that a less redundant sufficient statistic is

$$T(x) = \begin{pmatrix} x_1 \\ \vdots \\ x_{k-1} \end{pmatrix}$$

Further, the partition function $Z(\eta)$ is

$$Z(\eta) = \sum_{x \in \Omega_X} \exp\left(\eta^T T(x)\right) = \sum_{j=1}^{k-1} \exp\left(\eta_j\right) + 1$$

This allows us to write the likelihood $p(x;\eta)$ as

$$p(x;\eta) = \exp\left\{ \sum_{j=1}^{k-1} \eta_j x_j - \log\left(1 + \sum_{j=1}^{k-1} e^{\eta_j}\right) \right\} \tag{19.4}$$

This shows us that $A(\eta) = \log\left(1 + \sum_{j=1}^{k-1} e^{\eta_j}\right)$. This gives us our *minimal* exponential family.

Let's verify that this formulation is correct. Recall that the first derivative of $A(\eta)$ yields the first moment of the family of distributions, i.e., $\nabla A(\eta) = \mathbb{E}_{p(x;\eta)}[T(x)]$ for $\eta \in Int(\Omega)$. We have that

$$\begin{aligned} \frac{\partial A}{\partial \eta_j} &= \frac{1}{z(\eta)} e^{\eta_j} \\ &= p(x = j \mid \eta) \\ &= \mathbb{E}_{p(x;\eta)}[T_j(x)] \end{aligned}$$

as desired.

Example 19.1.2 ***1-D Gaussian:***

Consider $X \sim \mathcal{N}(\mu, \sigma^2)$, $\mathcal{X} = \mathbb{R}$. Here $\theta = (\mu, \sigma^2)$ ("moment parameterization").

$$p(x; (\mu, \sigma^2)) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right) \tag{19.5}$$

$$= \exp\left(-\frac{x^2}{2}\left[\frac{1}{\sigma^2}\right] + x\left[\frac{\mu}{\sigma^2}\right] - \left[\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)\right] \right) \tag{19.6}$$

$$\tag{19.7}$$

Drawing parallels with the exponential family equation 19.3, we get:

$$T(x) = \begin{bmatrix} x \\ -x^2/2 \end{bmatrix}, \eta(\theta) = \begin{bmatrix} \mu/\sigma^2 \\ 1/\sigma^2 \end{bmatrix} = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} \tag{19.8}$$

Here, $\eta_2 = \frac{1}{\sigma^2} > 0$ (precision), and $\eta_1 = \eta_2 \mu$. The domain is $\Omega = \{(\eta_1, \eta_2) : \eta_2 > 0, \eta_1 \in \mathbb{R}\}$. Further, $h(x) = 1$ (although some people use $h(x) = \frac{1}{\sqrt{2\pi}}$ for the Gaussian).

Example 19.1.3 ***Multivariate Gaussian:***

Now consider $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathcal{X} = \mathbb{R}^d$. Here $\theta = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

$$p(\boldsymbol{x}; \theta = (\boldsymbol{\mu}, \boldsymbol{\Sigma})) = \frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \right) \tag{19.9}$$

$$= \exp\left(-\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \left[\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \frac{1}{2}\log(2\pi|\boldsymbol{\Sigma}|)\right] \right) \tag{19.10}$$

$$= \exp\left(\mathrm{tr}\left(\left\langle \boldsymbol{\Sigma}^{-1}, -\frac{\boldsymbol{x}^\top \boldsymbol{x}}{2} \right\rangle \right) + \boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \left[\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \frac{1}{2}\log(2\pi|\boldsymbol{\Sigma}|)\right] \right) \tag{19.11}$$

$$\tag{19.12}$$

Similar to the 1-D Gaussian example, here we can consider

$$\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1} \text{ (precision)}$$
$$\boldsymbol{\eta} = \boldsymbol{\Lambda}\boldsymbol{\mu} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$$
$$T(x) = \begin{bmatrix} \boldsymbol{x} \\ -\boldsymbol{x}\boldsymbol{x}^\top/2 \end{bmatrix}$$

Example 19.1.4 ***Discrete UGM***

Let $p \in \mathcal{L}(G)$, (G is an undirected graph), with the potential functions $\psi_C(x_C) > 0 \; \forall \; C, x_C$, where C denotes cliques in G and x_C denotes a particular assignment of values to the

variables in clique C. Define $\mathcal{X}_C \coloneqq \{(y_i)_{i \in C} : \text{s.t. } y \in \mathcal{X}\} = \bigtimes_{i \in C} \mathcal{X}_i$, which is the set of all possible (joint) assignments of values to the variables in clique C.

$$p(x) = \frac{1}{\mathcal{Z}} \prod_{C \in \mathcal{C}} \psi_C(x_C) = \exp\left(\sum_{C \in \mathcal{C}} \log \psi_C(x_C) - \log \mathcal{Z}\right) \tag{19.13}$$

$$= \exp\left(\sum_{C \in \mathcal{C}} \sum_{y_C \in \mathcal{X}_C} \mathbb{1}\{y_C = x_C\} \log \psi_C(x_C) - \log \mathcal{Z}\right) \tag{19.14}$$

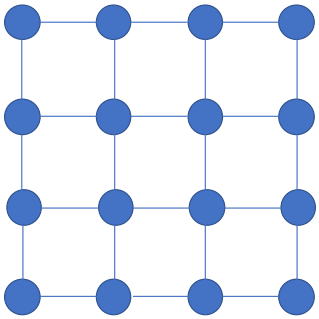
So, $T(x) = \begin{bmatrix} \vdots \\ \mathbb{1}\{x_C = y_C\} \\ \vdots \end{bmatrix}$ and $\eta(\theta) = \begin{bmatrix} \vdots \\ \log \psi_C(x_C) \\ \vdots \end{bmatrix}$. Thus, the sufficient statistics is denoted by $T_{C,y_C}(x)$ and the canonical parameter for it is η_{C,y_C}, where $C \in \mathcal{C}$ and $y_C \in \mathcal{X}_C$. Of course, $A(\eta) = \log \mathcal{Z}$ as well. Note that this is not a minimal representation.

Remark 19.1.2 ***Multinoulli(π)*** *is a special case where G is the complete graph (hence, there is one single clique). Note that K (the number of possible values for the multinouilli) is exponential in the size of the graph in a UGM (for example if each variable is binary, $K = 2^{|V|}$).*

Remark 19.1.3 ***Feature perspective****: Instead of using all indicators $\mathbb{1}\{y_C = x_C\}$, only a relevant subset could be chosen based on the task at hand. For example, consider modelling a sentence as x, where x_i denote words in the sentence. Usually the vocabulary size is huge (let's say 50000). Thus, all possible assignments in a complete graph (nodes being words) is clearly intractable (even though it can be represented as a single huge clique for exponential family representation, message passing over UGM is intractable). One possibility is take (consecutive) word pairs (x_i, x_{i-1}) as cliques and use features $\mathbb{1}\{y_C = x_C\}$ to represent if a particular word pair is present in the sentence (say, x_i = 'ran' and x_{i-1} = 'dog'). However, in this case the number of word pairs is huge (50000*50000), making the feature dimensionality high. A more compact feature representation would, therefore, be wrt POS tags, say, $\mathbb{1}\{x_i$ is a verb and x_{i-1} is a noun$\}$.*

Remark 19.1.4 ***Binary Ising Model****:*

Here the variables $x_i \in \{0, 1\}$ are binary corresponding to two directions of spins of an atom. The greatest clique size is 2 (i.e. $|C| \leq 2$). Suppose, we use nodes and node pairs (edges) as cliques in the graph (considering node potentials and edge potentials in UGM), we have dimension of $T(x)$ as $2|V| + 4|E|$ (this is an overparameterized exponential family – similar to the multinoulli case – as $\sum_{y_C} T_{C,y_C}(x) = 1$). *Can we find a minimal exponential family representation?*



A minimal representation: $T(x) = \begin{bmatrix} (x_i)_{i\in V} \\ (x_i, x_j)_{\{i,j\}\in E} \end{bmatrix} = \begin{bmatrix} \vdots \\ \mathbb{1}\{x_i = 1\} \\ \vdots \\ \mathbb{1}\{x_i = 1, x_j = 1\} \\ \vdots \end{bmatrix}$

This has dimensionality $|V| + |E|$. This representation makes sense as given p_i, p_j and p_{ij}, we can recover joint probability distribution of x_i and x_j (as shown below).

		x_i: 0	1
x_j:	0	$1 - p_i - p_j + p_{ij}$	$p_i - p_{ij}$
	1	$p_j - p_{ij}$	p_{ij}

More generally, we have that $\mathbb{E}[T(x)] = \begin{pmatrix} (\mu_i)_{i\in V} \\ (\mu_{i,j})_{\{i,j\}\in E} \end{pmatrix}$.

19.2 Estimation of parameters for PGM

19.2.1 DGM

For the parametric family:

$$\mathcal{P}_\Theta = \left\{ p_\theta(x) = \prod_i p(x_i | x_{\pi_i}, \underbrace{\theta_i}_{\text{independent parameterization}}) : \theta = \left(\theta_1, \cdots, \theta_{|v|}\right) \in \Theta = \Theta_1 \times \cdots \times \Theta_{|v|} \right\} \tag{19.15}$$

$\Rightarrow$ MLE decouples in $|v|$ independent problems. Assume we have a dataset $\{x^{(i)}\}_{i=1}^n$. We have that

$$p\left(\text{data}|\theta\right) = \prod_{i=1}^{n} p\left(x^{(i)}|\theta\right) = \prod_{i=1}^{n}\prod_{j=1}^{|v|} p\left(x^{i}|x^{i}_{\pi_j}, \theta_j\right) \tag{19.16}$$

$$\log\left[\cdot\right] = \sum_{j=1}^{|v|} \underbrace{\left[\sum_{i=1}^{n} \log p\left(x_j(i)|x^{i}_{\pi_j}, \theta_j\right)\right]}_{f_j(\theta_j)} \tag{19.17}$$

Example 19.2.1 ***DGM with Discrete Random Variables***

In this case, we effectively have that the MLE, θ_j^{MLE}, is the proportion of observations. We have that

$$\hat{p}(x_j = k \mid x_{\pi_j} = \textit{stuff}) = \frac{\#(x_j = k, x_{\pi_j} = \textit{stuff})}{\#(x_{\pi_j} = \textit{stuff}}$$

Note that if we have latent variables (i.e., unobserved variable) in our DGM, we should use Expected Maximization (EM) as we did for Hidden Markov Models (HMM) previously.

19.2.2 UGM

Let's now consider parametric family subset of a UGM by using exponential families (as saw before). Let the likelihood $p(x|\eta)$ be defined as

$$p(x|\eta) = \exp\left(\sum_{C}\langle\eta_C, T_C(x_C)\rangle - A(\eta)\right)$$

The log-likelihood is:

$$\frac{1}{n}\sum_{i=1}^{n}\log p(x^{(i)}|\eta) = \sum_{c}\eta_c^T \underbrace{\left(\frac{1}{n}\sum_{j=1}^{n} T_c(x_c^{(i)})\right)}_{\mu_c} - \frac{\cancel{n}A(\eta)}{\cancel{n}} \tag{19.18}$$

Now, let's take the gradient with respect to η_C.

$$\nabla_{\eta_C}\left[\frac{1}{n}\sum_{i=1}^{n}\log p(x^{(i)}|\eta)\right] = \hat{\mu}_C - \mu_C(\eta)$$

We find ourselves in a bit of trouble now, however, as $\mu_C(\eta) = \mathbb{E}_{p(x;\eta)}[T_C(x_C)]$, and to compute this we need to use inference. For example, take the Ising model, for which we have $T_{i,j}(x_i, x_j) = x_i \cdot x_j$ for $x_i \in \{0, 1\}$, and as such the expectation is

$$\begin{aligned}\mathbb{E}[T_{i,j}(x_i, x_j)] &= \mu_{i,j} \\ &= p(x_i = 1, x_j = 1 \mid \eta)\end{aligned}$$

To get the value of $p(x_i = 1, x_j = 1 \mid \eta)$ we'll have to use some sort of approximate inference as the treewidth is linear in the side of the grid for a grid UGM, so exact inference is intractable. This might be sampling (e.g., Gibbs sampling), or a variational method (e.g., mean field approximation). We will cover sampling starting in the next lecture.