· Exponential Family

Once we choose 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

let $\mathcal{X}$ be a set of variables. An exponential family $\mathcal{P}$ over $\mathcal{X}$ is specified by 4 components

- A sufficient statistics function $\tau(\cdot)$ from assignments to $\mathcal{X}$ to $\mathbb{R}^k$ — sufficient statistics function
- A parameter space that is a convex set $\Theta \subseteq \mathbb{R}^m$ of legal parameters — parameter space
- A natural parameter function $t(\cdot)$ $t: \mathbb{R}^m \mapsto \mathbb{R}^k$ — legal parameter
- An auxiliary measure A over $\mathcal{X}$

each parameter $\theta \in \Theta$ specifies a distribution P_θ in this family as

$$P_\theta(\xi) = \frac{1}{Z(\theta)} A(\xi) \exp\left[\langle t(\theta), \tau(\xi) \rangle \right]$$

$$Z(\theta) = \sum_\xi A(\xi) \exp\left[\langle t(\theta), \tau(\xi) \rangle \right]$$ the partition function, must be finite

The parametric family $\mathcal{P}$ is defined as

$$\mathcal{P} = \{ P_\theta \mid \theta \in \Theta \}$$

The exponential family is a concise representation of a class of probability distribution that share a similar function form. A member of this family is specified by a parameter vector $\theta \in \Theta$

The sufficient statistic function $\tau(\cdot)$ summarizes the aspects of an instance that are relevant for assigning it a probability

The function $t(\cdot)$ maps the parameters to space of the sufficient statistics

A assigns additional preferences among instances that do not depend on the parameter (A is constant in most cases)

eg. Bernoulli

$$\tau(X) = \begin{bmatrix} 1\{X=1\} \\ 1\{X=0\} \end{bmatrix} \qquad t(\theta) = \begin{bmatrix} \ln(\theta) \\ \ln(1-\theta) \end{bmatrix}$$

$$\exp[\langle \tau(X), t(\theta) \rangle] = \exp\left[1\{X=1\} \cdot \ln(\theta) + 1\{X=0\} \ln(1-\theta) \right] = \begin{cases} \theta & X=1 \\ 1-\theta & X=0 \end{cases}$$

eg. Gaussian

$$p(X) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(X-\mu)^2}{2\sigma^2} \right]$$

$$\tau(X) = \begin{bmatrix} X \\ X^2 \end{bmatrix} \qquad t(\mu, \sigma^2) = \begin{bmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{bmatrix} \qquad Z(\mu, \sigma^2) = \sqrt{2\pi}\,\sigma \exp\left(\frac{\mu^2}{2\sigma^2}\right)$$

· Linear Exponential Family

(parameters have the same dimension as data's representation $T(x)$)

θ is called the natural parameters for the given sufficient statistic function

$$P_\theta(x) = \frac{1}{Z(\theta)} \exp\{\langle \theta, T(x) \rangle\}$$

require that each $\theta \in \Theta$ gives a legal distribution

define the set of allowable natural parameters, the parameter space, to be

$$\Theta = \left\{ \theta \in \mathbb{R}^k : \int \exp(\langle \theta, T(x) \rangle)\, dx < \infty \right\}$$

(every choice of θ leads to a valid distribution if X is discrete)

An exponential family over the natural parameter parameter space, and for which the natural parameter space is open and convex, is called a linear exponential family

(we need to define only the function $T(\cdot)$ to specify such a family)

eg. Bernoulli

in the previous example, $t(\theta) = \begin{bmatrix} \ln \theta \\ \ln(1-\theta) \end{bmatrix}$ is not a convex set

let $T(x) = \mathbb{1}\{x=1\}$ $\quad t(\theta) = \ln \frac{\theta}{1-\theta}$

$\exp(\langle t(\theta), T(x) \rangle) = \exp\left[\ln \frac{\theta}{1-\theta} \cdot \mathbb{1}\{x=1\}\right]$

$$\left.\begin{aligned} \exp[\langle t(\theta), T(1) \rangle] &= \frac{\theta}{1-\theta} \\ \exp[\langle t(\theta), T(0) \rangle] &= 1 \end{aligned}\right\} Z(\theta) = \frac{\theta}{1-\theta} + 1 = \frac{1}{1-\theta}$$

$P_\theta(x) = (1-\theta) \exp\left[\ln \frac{\theta}{1-\theta} \cdot \mathbb{1}\{x=1\}\right]$

$P_\theta(1) = \theta \qquad P_\theta(0) = 1-\theta$

eg. Multinomial

$t(\theta) = \langle \ln \theta_1, \ln \theta_2, \cdots \ln \theta_k \rangle$

$T(x) = \langle \mathbb{1}\{x=1\}, \mathbb{1}\{x=2\}, \mathbb{1}\{x=k\} \rangle$

$P_\theta(x) = \exp[\langle t(\theta), T(x) \rangle]$

$t(\theta)$ is not a convex set

$t(\theta) = \langle \ln \frac{\theta_1}{\theta_k}, \ln \frac{\theta_2}{\theta_k}, \cdots, \ln \frac{\theta_{k-1}}{\theta_k} \rangle$

$T(x) = (\mathbb{1}\{x=1\}, \mathbb{1}\{x=2\}, \ldots, \mathbb{1}\{x=k-1\})$

$Z(\theta) = \sum_x \exp(\langle t(\theta), T(x) \rangle) = \frac{\theta_1}{\theta_k} + \frac{\theta_2}{\theta_k} + \cdots \frac{\theta_{k-1}}{\theta_k} + 1 = \frac{1-\theta_k}{\theta_k} + 1 = \theta_k$

$$P_\theta(X=i) = \begin{cases} \frac{\theta_i}{\theta_k} \cdot \theta_k = \theta_i \\ \theta_k \cdot 1 = \theta_k \end{cases}$$

· Factored Exponential Family

the log-linear model is

$$P(X_1, X_2 \ldots X_n) \propto \exp\left\{\sum_{i=1}^{k} \theta_i f_i(D_i)\right\}$$

it's a linear exponential family $\theta = (\theta_1, \cdots \theta_k)$ $T(X) = \langle f_1(X_1), \cdots, f_k(X_k)\rangle$

By choosing appropriate features, we can devise a log-linear model to represent a given discrete Markov network structure

discrete Markov networks are linear exponential families

· Factored Exponential Family : Product Distribution

An (unnormalized) exponential factor families Φ is defined by A, T, t, Δ and Θ

A factor in this family is

$$\phi_\theta(X) = A(X) \exp\{\langle t(\theta), T(X)\rangle\}$$

Let $\Phi_1, \cdots, \Phi_k$ be exponential families. the family composition of $\Phi_1, \cdots \Phi_k$ is the family $\Phi_1 \times \cdots \times \Phi_k$, parametrized by $\theta = \theta_1 \circ \theta_2 \cdots \circ \theta_k \in \Theta_1 \times \Theta_2 \times \cdots \Theta_k$

$$P_\theta(X) \propto \prod_i \phi_{\theta_i}(X) = \left(\prod_i A_i(X)\right) \exp\left\{\sum_i \langle t_i(\theta_i) \cdot T_i(X)\rangle\right\}$$

where ϕ_{θ_i} is a factor in family Φ_i

The composition of exponential factors is an exponential family

$$T(X) = T_1(X) \circ T_2(X) \cdots \circ T_k(X) \qquad t(\theta) = t_1(\theta_1) \circ t_2(\theta_2) \cdots \circ t_k(\theta_k)$$

· Factored Exponential Family : Bayesian Network

if we have a set of CPDs from an exponential family, then their product is also in the exponential family

$$P_{\theta_i}(X_i \mid X_{i-1}) \propto A_i(X_i) \cdot \exp\{\langle t_i(\theta_i), T_i(X_i, X_{i-1})\rangle\}$$

$$P_\theta(X_1, \cdots X_k) \propto \prod_i A_i(X_i) \cdot \exp\left\{\sum_i \langle t_i(\theta_i), T_i(X_i, X_{i-1})\rangle\right\}$$

$$= \prod_i A_i(X_i) \cdot \exp\{\langle t(\theta), T(X)\rangle\}$$

A Bayesian network with exponential CPDs defines an exponential family

eg table - CPD $P(X|U)$

$$T_{X|U}(X, U) = \langle \mathbb{1}\{X = x, U = u\} : x \in Val(X), u \in Val(U)\rangle$$

$$t_{X|U}(\theta) = \langle \ln P(x|u) : x \in Val(X), u \in Val(U)\rangle$$

$$p(x|u) = \exp\{\langle t_{X|U}(\theta), T_{X|U}(\theta)\rangle\}$$

can use similar representation to capture any table CPD / tree CPD eg. linear Gaussian

$X = \beta_0 + \beta_1 u_1 + \cdots + \beta_k u_k + \epsilon \qquad \epsilon \sim N(0, \sigma^2)$

$p(x|u) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{ -\frac{1}{2\sigma^2} (x - (\beta_0 + \beta_1 u_1 + \cdots \beta_k u_k))^2 \}$

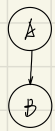
$T_{x|u}(x) = \langle 1, x, u_1, \cdots, u_k, xu_1, \cdots xu_k, u_1^2, u_1u_2, \cdots u_k^2 \rangle$

A Bayesian network that is the product of exponential CPDs defines an exponential family

We cannot construct a Bayesian network from a product of unnormalized exponential factors

eg. A CPD $P(B|A)$ with both A and B binary

A	θ_A
a^0	θ_{a^0}
a^1	θ_{a^1}



A	B	$\theta_{B|A}$
a^0	b^0	$\theta_{a^0|b^0}$
a^0	b^1	$\theta_{a^0|b^1}$
a^1	b^0	$\theta_{a^1|b^0}$
a^1	b^1	$\theta_{a^1|b^1}$

A ~ Bernoulli

$T(A) = 1\{A=a^1\} \qquad t(\theta_{a^1}) = \ln\frac{\theta_{a^1}}{1-\theta_{a^1}}$

$\exp\{\langle t(\theta_{a^1}), T(a^1)\rangle\} = \frac{\theta_{a^1}}{1-\theta_{a^1}}$

$\exp\{\langle t(\theta_{a^1}), T(a^0)\rangle\} = 1$

$Z_A(\theta_{a^1}) = 1 + \frac{\theta_{a^1}}{1-\theta_{a^1}} = \frac{1}{1-\theta_{a^1}}$

$P_{\theta_{a^1}}(a^1) = \theta_{a^1} \qquad P_{\theta_{a^1}}(a^0) = 1-\theta_{a^1}$

$t_A(\theta_{a^1}) = \frac{\theta_{a^1}}{\theta_{a^0}}$

$T_A(A) = 1\{A=a^1\}$

B|A ~ Bernoulli

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

$T(B,A) = \langle T(B|a^1), T(B|a^0) \rangle$

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

$t(\theta) = \langle t(\theta|a^1), t(\theta|a^0) \rangle$

$\exp(\langle T(b^0,a^0), t(\theta)\rangle) = 1$

$\exp(\langle T(b^1,a^0), t(\theta)\rangle) = \frac{\theta_{b^1|a^0}}{\theta_{b^0|a^0}}$

$Z_{B|a^0}(\theta) = 1 + \frac{\theta_{b^1|a^0}}{\theta_{b^0|a^0}} = \frac{1}{1-\theta_{b^1|a^0}}$

$P(b^1|a^0) = \theta_{b^0|a^0} \cdot \frac{\theta_{b^1|a^0}}{\theta_{b^0|a^0}} = \theta_{b^1|a^0}$

$P(b^0|a^0) = \theta_{b^0|a^0}$

$t_{B|A}(\theta_{b|a}) = [\ln\frac{\theta_{b^1|a^1}}{\theta_{b^0|a^1}}, \ln\frac{\theta_{b^1|a^0}}{\theta_{b^0|a^0}}]$

$T_{B|A}(B,A) = [1\{B=b^1, A=a^1\}, 1\{B=b^1, A=a^0\}]$

$t_{AB}(\theta) = [\ln\frac{\theta_{a^1}}{\theta_{a^0}}, \ln\frac{\theta_{b^1|a^1}}{\theta_{b^0|a^1}}, \ln\frac{\theta_{b^1|a^0}}{\theta_{b^0|a^0}}]$

$T_{AB}(A,B) = [1\{A=a^1\}, 1\{B=b^1, A=a^1\}, 1\{B=b^1, A=a^0\}]$

$Z(\theta)$ is differen for 2 CPDs $P(B|A=a^0)$ $P(B|A=a^1)$

$\exp\{\langle t_{AB}(\theta), T_{AB}(a^0,b^0)\rangle\} = 1 \qquad P(a^0,b^0) = \frac{1}{Z_A(\theta_{a^1})} \cdot \frac{1}{Z_{B|a^0}(\theta)} \exp\{\} = \theta_{b^0|a^0} \cdot \theta_{a^0}$

$\exp\{\langle t_{AB}(\theta), T_{AB}(a^0,b^1)\rangle\} = \frac{\theta_{b^1|a^0}}{\theta_{b^0|a^0}} \qquad P(a^0,b^1) = \theta_{a^0} \cdot \theta_{b^0|a^0} \cdot \frac{\theta_{b^1|a^0}}{\theta_{b^0|a^0}} = \theta_{b^1|a^0} \cdot \theta_{a^0}$

$\exp\{\langle t_{AB}(\theta), T_{AB}(a^1,b^0)\rangle\} = \frac{\theta_{a^1}}{\theta_{a^0}} \qquad P(a^1,b^0) = \frac{1}{Z_A(\theta_{a^1})} \cdot \frac{1}{Z_{B|a^1}(\theta)} \exp\{\} = \theta_{a^0}\, \theta_{b^0|a^1} \cdot \frac{\theta_{a^1}}{\theta_{a^0}} = \theta_{b^0|a^1} \cdot \theta_{a^1}$

$\exp\{\langle t_{AB}(\theta), T_{AB}(a^1,b^1)\rangle\} = \frac{\theta_{a^1}\, \theta_{b^1|a^1}}{\theta_{a^0} \cdot \theta_{b^0|a^1}} \qquad P(a^1,b^1) = \theta_{a^0}\, \theta_{b^0|a^1} \cdot \frac{\theta_{a^1}\, \theta_{b^1|a^1}}{\theta_{a^0} \cdot \theta_{b^0|a^1}} = \theta_{a^1} \cdot \theta_{b^1|a^1}$

to have an exponential representation of a Bayesian network, need to ensure that each CPD is locally normalized

Can have an extra 1 in $T(x)$ and $\log\frac{1}{Z(\theta)}$ in $t(\theta)$

eg $A \sim \text{Bernoulli}(\theta_{a'})$

$T(A) = \mathbb{1}\{A=a'\}$ $\quad t(\theta_{a'}) = \ln\frac{\theta_{a'}}{1-\theta_{a'}}$

$\exp\{\langle T(a'), t(\theta_{a'})\rangle\} = \frac{\theta_{a'}}{1-\theta_{a'}}$

$\exp\{\langle T(a^0), t(\theta_{a'})\rangle\} = 1$

$Z_A(\theta_{a'}) = 1 + \frac{\theta_{a'}}{1-\theta_{a'}} = \frac{1}{1-\theta_{a'}}$

$\Rightarrow$ $T_A(A) = \langle \mathbb{1}\{A=a'\}, 1 \rangle$

$t_A(\theta_{a'}) = \langle \ln\frac{\theta_{a'}}{1-\theta_{a'}}, \ln(1-\theta_{a'}) \rangle$

$\exp\{\langle T(a'), t(\theta_{a'})\rangle\} = \frac{\theta_{a'}}{1-\theta_{a'}} \cdot (1-\theta_{a'}) = \theta_{a'}$

$\exp\{\langle T(a^0), t(\theta_{a'})\rangle\} = 1 \cdot (1-\theta_{a'}) = 1-\theta_{a'}$

$B|A \sim$ Bernoulli

$T(A,B) = \langle \mathbb{1}\{B=b', A=a'\}, \mathbb{1}\{B=b', A=a^0\} \rangle$

$t(\theta_{b'|a'}, \theta_{b'|a^0}) = \langle \ln\frac{\theta_{b'|a'}}{\theta_{b^0|a'}}, \ln\frac{\theta_{b'|a^0}}{\theta_{b^0|a^0}} \rangle$

$$\begin{cases} \exp(\langle t(\theta), T(b^0,a^0)\rangle) = 1 \\ \exp(\langle t(\theta), T(b',a^0)\rangle) = \frac{\theta_{b'|a^0}}{\theta_{b^0|a^0}} \\ Z_{B|a^0}(\theta) = 1 + \frac{\theta_{b'|a^0}}{\theta_{b^0|a^0}} = \frac{1}{1-\theta_{b'|a^0}} \end{cases}$$

$$\begin{cases} \exp(\langle t(\theta), T(b^0,a')\rangle) = 1 \\ \exp(\langle t(\theta), T(b',a')\rangle) = \frac{\theta_{b'|a^0}}{\theta_{b^0|a'}} \\ Z_{B|a'}(\theta) = 1 + \frac{\theta_{b'|a'}}{\theta_{b^0|a'}} = \frac{\theta_{b'|a'}}{1-\theta_{b'|a'}} \end{cases}$$

$T_{B|A}(A,B) = \langle \mathbb{1}\{B=b', A=a'\}, \mathbb{1}\{A=a'\}, \mathbb{1}\{B=b', a=a^0\}, \mathbb{1}\{A=a^0\} \rangle$

$t_{B|A}(\theta_{b'|a'}, \theta_{b'|a^0}) = \langle \ln\frac{\theta_{b'|a'}}{\theta_{b^0|a'}}, \ln(1-\theta_{b'|a'}), \ln\frac{\theta_{b'|a^0}}{\theta_{b^0|a^0}}, \ln(1-\theta_{b'|a^0}) \rangle$

$\exp\{\langle t_{B|A}(\cdot), T_{B|A}(a^0, b^0)\rangle\} = 1-\theta_{b'|a^0} = \theta_{b^0|a^0}$

$\exp\{\langle t_{B|A}(\cdot), T_{B|A}(a^0, b')\rangle\} = \frac{\theta_{b'|a^0}}{\theta_{b^0|a^0}} \cdot (1-\theta_{b'|a^0}) = \theta_{b'|a^0}$

$\exp\{\langle t_{B|A}(\cdot), T_{B|A}(a', b^0)\rangle\} = 1-\theta_{b'|a'} = \theta_{b^0|a'}$

$\exp\{\langle t_{B|A}(\cdot), T_{B|A}(a', b')\rangle\} = \frac{\theta_{b'|a'}}{\theta_{b^0|a'}} \cdot (1-\theta_{b'|a'}) = \theta_{b'|a'}$

$T_{AB}(A,B) = \langle \mathbb{1}\{A=a'\}, 1, \mathbb{1}\{B=b', A=a'\}, \mathbb{1}\{A=a'\}, \mathbb{1}\{B=b', A=a^0\}, \mathbb{1}\{A=a^0\} \rangle$

$t_{AB}(\theta_{a'}, \theta_{b'|a'}, \theta_{b'|a^0}) = \langle \ln\frac{\theta_{a'}}{\theta_{a^0}}, \ln(1-\theta_{a'}), \ln\frac{\theta_{b'|a'}}{\theta_{b^0|a'}}, \ln(1-\theta_{b'|a'}), \ln\frac{\theta_{b'|a^0}}{\theta_{b^0|a^0}}, \ln(1-\theta_{b'|a^0}) \rangle$

$P_{AB}(A,B) = \exp\{\langle T_{AB}(A,B), t_{AB}(\theta_{a'}, \theta_{b'|a'}, \theta_{b'|a^0})\rangle\}$

Though a Bayesian network with suitable CPDs defines an exponential family, this family is not generally a linear one.

Any Bayesian network that contain immoralities does not induce a linear exponential family