Representation, Inference and Learning of Bayesian Networks as Conjugate Exponential Family Models

January 26, 2015

Abstract

1 Introduction

Defining the data structure of a Bayesian network is not a straightforward problem. The definition of the data structure of a DAG is not complex when compared to the definition of the data structure of the conditional probability distributions encoded in the BN. The DAG is an *homogeneous* data structure, in the sense that it is only composed by nodes and directed edges. However, the set of different conditional distributions is not limited at all. For example, the data structure for representing a Multinomial distribution is, in a first look, quite different from the data structure needed to represent a Normal distribution. In the former case, we need to store the probability of each one of the cases of the multinomial variable, while in the latter case we need to store, for example, the mean and the variance of the Normal distribution. If we consider a Poisson, an Exponential, or a MoTBF, etc, the data structures needed to represent these distributions are completely different.

But these are only examples of unidimensional distributions. When defining the data structure of the conditional distributions things become much more complex. For example, the data structure for representing the conditional distribution of a Normal distribution given a set of normally distributed variables is, in a first look, totally different from the data structure needed to represent a conditional distribution of a Multinomial variable given a set of Multinomial variables. In the former case, under the conditional linear Gaussian framework, we need to store the coefficients for the linear combination of the parents variables plus the variance of the main variable. While the data structure for the multinomial given multinomial case is usually defined using a big probability table. If we want to allow the combination of Normal, Multinomial, Poison, Exponential, etc, in the same framework the number of data structures needed to represent all the possible combinations quickly explode.

It is also challenging the problem of making inferences and learn from data with BNs with different kinds of conditional distributions. For example, the maximum likelihood of a Normal distribution is obtained by computing the sample mean and variance, while the maximum likelihood of a Multinomial distribution is obtained by normalizing the sample *counts* of each one of the sates. Alternative methods are required for the different possible conditional probabilities, which means that the addition of new family of variables, i.e. Normal, Poission, or Exponential, etc, implies to define, code and test from scratch new maximum likelihood methods.

In the case of inference, things are even worse. For example, the combination and marginalization operations over probability potentials belonging to different distribution families is in general non-closed and, in principle, involves quite different approaches. I.e., the combination or multiplication of two multinomial potentials or distributions involve completely different methods than the combination or product of two Normal distributions. And similarly for the marginalization operation. So, defining and coding all of these operations for different family of distributions can become a daunting task.

In this technical report, we argue that if we restrict ourselves to the so-called conjugate exponential family models, we can avoid most of the above problems. Firstly, all the conditional probability distributions inside this family can be represented using the same data structure, which is simply composed by two n-dimensional vectors (the so-called natural and moment parameters) and two n-dimensional functions (the so-called sufficient statistics and log-normalizer functions). Moreover, we show how many learning and inference algorithms can be directly implemented on top of this general and unique representation. The result is a suitable framework for coding a toolbox which aims to deal with the problem of representing, making inference and learning general Bayesian networks from data.

2 Background and notation

Bayesian networks

Let $\mathbf{X} = \{X_1, \ldots, X_N\}$ denote the set of stochastic random variables defining our domain problem and $\mathbf{x}$ an observation vector. A Bayesian network defines a joint distribution $P(\mathbf{X})$ in the following form:

$$p(\mathbf{X}) = \prod_{i=1}^{N} p(X_i|Pa(X_i))$$

where $Pa(X_i) \subset \mathbf{X} \backslash X_i$ represents the so-called *parent variables* of X_i. Bayesian networks can be graphically represented by a directed acyclic graph (DAG). Each node, labelled X_i in the graph, is associated with a factor or conditional probability $p(X_i|Pa(X_i))$.

Additionally, for each parent $X_j \in Pa(X_i)$, the graph contains one directed edge pointing from X_j to the *child* variable X_i.

Exponential family models

A Bayesian network defines a joint probability in the exponential family with a natural (or canonical) parametrization if the joint distribution can be functionally expressed as follows,

$$p(\mathbf{x}|\theta) = h(x)exp(\theta^T s(\mathbf{x}) - A(\theta))$$

where θ is the so-called natural parameter, which belongs to the so-called *natural parameter space* $\Theta \equiv \{\theta \in \Re^K : \int_{\mathbf{x}} h(\mathbf{x})exp(\theta^T s(\mathbf{x}) - A(\theta))d\mathbf{x} < \infty\}$, $s(\mathbf{x})$ is the vector of sufficient statistics belonging to $\mathcal{S} \subseteq \Re^K$, A is the log partition function and $h(x)$ is the base measure.

The so-called *expectation or moment parameters* $\mu \in \mathcal{S}$ can also be used to parameterize probability distributions of the exponential family. This *expectation parameter* μ is defined as the expected vector of sufficient statistics with respect to θ:

$$\mu \triangleq E[s(\mathbf{x})|\theta] = \int s(\mathbf{x})p(\mathbf{x}|\theta)d\mathbf{x} \tag{2.1}$$

The *natural parameter* θ associated to an *expectation parameter* μ is obtained by solving the following optimization problem.

$$\theta(\mu) = arg\max_{\theta \in \Theta} \theta^T \mu - A(\theta) \tag{2.2}$$

Given the natural parameters, the inverse step of updating the moment parameters is not trivial in the case of conditional distributions, since as we will see in Section 2 the transformation requires the joint probability distribution of the children and the parents.

Regular Exponential Family

A Bayesian network belongs to the linear exponential family if Θ is an open and convex set, otherwise it belongs to the more general curved exponential family. Additionally, an exponential family is said to be minimal if there is non-zero constant vector α, such that $\alpha^T s(\mathbf{x})$ is equal to a constant for all $\mathbf{x}$. We will consider the regular exponential family to be the linear exponential one with minimal representation.

The regular exponential family with a minimal representation has been widely studied in the literature. The distributions in this family enjoy many useful properties. The following two properties are some of the most relevant ones: i) The transformation between θ and μ parameters is a one-to-one correspondence, i.e. μ is a dual set of

the model parameter θ; and ii) the moment parameters equal the gradient of the log-normalizer (the proof of this equality is (will be!) included in the appendix),

$$\mu = \frac{\partial A(\theta)}{\partial \theta} \tag{2.3}$$

However any BN which contains immoralities does not induce a regular exponential family. The limitations of this is that the capability to efficiently transformation from natural to moment parameters through this equation is lost. In the following sections we will describe how to handle this in the curved exponential family as well.

Conditional distributions as exponential family models

A conditional distribution $p(X|Pa(X))$ is in the exponential family if it can be written in the following functional form,

$$p(x|Pa(X)) = h(x)exp(\theta(Pa(X))^T s(x) - A(\theta(Pa(X)))) \tag{2.4}$$

where $\theta(Pa(X))$ denotes that the natural parameters are now a function of the parent variables of X.

A conditional distribution $p(X|Y)$ is said to be *conjugate* to a child distribution $p(W|X)$ if $p(X|Y)$ has the same functional form, with respect to X, as $p(W|X)$. An important property of the exponential conjugate conditional distributions is that they can also be expressed in the following functional form,

$$p(x|\mathbf{y}) = h(x)exp(\theta^T s(x, \mathbf{y}) - B(\theta)) \tag{2.5}$$

where $B(\theta)$ is the conditional log-normalizer.

A Bayesian network is said to be a conjugate exponential (CEF) model if all its conditional distributions belong to the exponential family and are conjugate. Importantly enough, the use of conjugate distributions allows that the posterior for each distribution has the same form as the prior, which means that only the values of the parameters change, but not the functional form of the distribution.

3 Bayesian networks as CEF models

3.1 Representation

In this section we show why conjugate-exponential family Bayesian networks (CEF-BNs) are quite amenable to be represented (and coded) as exponential family models.

By using Equation 2.5, a conjugate-exponential BN can be represented in the following way,

$$\begin{aligned}
\ln p(X_1, \ldots, X_n) &= \sum_{i=1}^{n} \ln p(X_i | Pa(X_i)) \\
&= \sum_{i=1}^{n} \theta_i(Pa(X_i))^T s_i(X_i) - B(\theta(Pa(X_i))) \\
&= \sum_{i=1}^{n} \theta_i^T s_i(X_i, Pa(X_i)) - B(\theta_i) \\
&= \begin{pmatrix} \theta_1 \\ \ldots \\ \theta_n \end{pmatrix}^T \begin{pmatrix} s_1(X_1, Pa(X_1)) \\ \ldots \\ s_n(X_n, Pa(X_n)) \end{pmatrix} - \sum_{i=1}^{n} B_i(\theta_i) \qquad (3.1)
\end{aligned}$$

The above expression show us that we can represent a CEF-BN by using the local representations of the conditional distributions:

- The natural parameters are formed by the composition of the local natural parameters of each conditional distribution.
- The sufficient statistics are formed the composition of the local sufficient statistics of each conditional distribution.
- The log-normalizer is the sum of the local conditional log-normalizer of each conditional distribution.

So, in order to represent a BN as an exponential family model we only have to worry about the local representation of each conditional distribution as in Equation 2.5. The global representation is just obtained by composing these local representations.

Let us notice that without the assumption of conjugacy for the conditional exponential distributions, the above representation would have not been possible.

3.2 From Natural to Moment Parameters

We now look at the transformation from to natural to moment parameters in a CEF-BN. Following Equation 2.1., we have that the vector of moment parameters in a CEF-BN model decomposes as follows,

$$\begin{aligned}
\mu &= E\left[s(X_1, \ldots, X_n)|\theta\right] \\
&= \int s(X_1, \ldots, X_n) p(X_1, \ldots, X_n|\theta) d\mathbf{X} = \qquad (3.2) \\
&= (\mu_1, \ldots, \mu_n)
\end{aligned}$$

where $\mu_i = \int s(X_i, Pa(X_i))p(X_i, Pa(X_i))d\mathbf{X}$. I.e. the moments parameter of a CEF-BN locally decomposes in moment parameters associated to local marginal probabilities. Let us note that to compute the local marginal probability $p(X_i, Pa(X_i))$ we need to perform inference over the whole BN.

3.3 From Moment to Natural Parameters

The transformation from moment to natural parameters also decomposes. Let us start by expanding Equation 2.2,

$$\begin{aligned} \theta(\mu) &= \arg\max_{\theta\in\Theta} \theta^T\mu - A(\theta) \\ &= \arg\max_{(\theta_1,\ldots,\theta_n)\in\Theta} \sum_{i=1}^{n} \theta_i^T\mu_i - B(\theta_i) \end{aligned} \tag{3.3}$$

Because the θ_i parameters are independent, the above maximization problem fully decomposes in a local maximization problem for each conditional probability distribution,

$$\theta_i(\mu_i) = \arg\max_{\theta_i\in\Theta_i} \theta_i^T\mu_i - B(\theta_i) \tag{3.4}$$

and the global solution is just the composition of the local solutions,

$$\theta(\mu) = (\theta_1(\mu_1), \ldots, \theta_n(\mu_n))$$

Let us note that, as oppose to the previous case, the transformation from moment to natural parameters can be performed locally at each conditional distribution. The global solution is just an aggregation of the local solutions.

4 Conditional distributions with multinomial parents in exponential form

In this subsection we try to exploit the commonalities in terms of structure that are present in many conditional distributions in order to ease its representation in exponential form and, also, the associated parameter transformations. The structure we try to exploit is the presence of multinomial variables in the parent set of a conditional distribution. The advantage of this representation is that if we know how to represent in a exponential form some distribution (ie. Poison distribution, a Normal distribution, a Multinomial distribution, a Normal distribution with Normal parents, etc), we

can directly derive the corresponding distribution conditioned to multinomial parents (i.e. Poison given Multinomial parents , Normal given Multinomial parents, Multinomial given Multinomial parents, Normal given Normal and Multinomial parents, etc). Similarly, we also show that for these conditional distributions the transformation from moment to natural parameters can be further decomposed.

4.1 Representation

Let $(\mathbf{Z}, \mathbf{Y})$ be the set of parents of a variable X, where $\mathbf{Y}$ is a set of multinomial variables and $\mathbf{Z}$ a set of continuous variables[1]. Let q denote the total number of parental configurations for the variables in $\mathbf{Y}$, and let $\mathbf{y}$ denote the l-th parental configuration $1 \leq l \leq q$.

The log-conditional probability of X given its parent-nodes $\mathbf{Z}$ and $\mathbf{Y}$ decomposes as follows:

$$\begin{aligned}
\ln p(X \mid \mathbf{Z}, \mathbf{Y}) &= \sum_{l=1}^{q} I(\mathbf{Y} = \mathbf{y}^l) \cdot \ln p(X|\mathbf{Z}, \mathbf{y}^l) \\
&= \sum_{l=1}^{q} I(\mathbf{Y} = \mathbf{y}^l) \cdot \Big(\theta_l^T s_l(X, \mathbf{Z}) - B_l(\theta_l)\Big) \\
&= \sum_{l=1}^{q} \theta_l^T I(\mathbf{Y} = \mathbf{y}^l) s(X, \mathbf{Z}) - \sum_{l=1}^{q} I(\mathbf{Y} = \mathbf{y}^l) B_l(\theta_l)
\end{aligned}$$

where θ_l, $s_l(X, \mathbf{Z})$ and $B_l(\theta_l)$ are provided when the local conditional distribution $p(X|\mathbf{Z}, \mathbf{y}^l)$ is expressed in exponential form.

So, the conditional distribution $p(X \mid \mathbf{Z}, \mathbf{Y})$ can be written in exponential form as follows,

$$\begin{aligned}
\ln p(X \mid \mathbf{Z}, \mathbf{Y}) &= \theta^T s(X, \mathbf{Y}) - B(\theta) \\
&= \begin{pmatrix} -B_1(\theta_1) \\ \vdots \\ -B_q(\theta_q) \\ \theta_1 \\ \vdots \\ \theta_q \end{pmatrix}^T \begin{pmatrix} I(\mathbf{Y} = \mathbf{y}^1) \\ \vdots \\ I(\mathbf{Y} = \mathbf{y}^q) \\ s_1(X, \mathbf{Z}) \cdot I(\mathbf{Y} = \mathbf{y}^1) \\ \vdots \\ s_q(X, \mathbf{Z}) \cdot I(\mathbf{Y} = \mathbf{y}^q) \end{pmatrix} - 0
\end{aligned} \tag{4.1}$$

The corresponding proof can be found (will be!) in the appendix.

[1] $\mathbf{Z}$ can be empty.

As can be seen, the exponential representation of a conditional probability with multinomial parents can be expressed as the composition of the exponential representation of the conditional distributions restricted to each one of the possible configurations of the multinomial parents.

4.2 From Natural to Moment Parameters

By using the decomposition of the sufficient statistics vector of this conditional probability distribution, the moment parameter vector associated to a natural parameter vector θ can be expressed as a combination of the moment parameters of the marginal probabilities $p(\mathbf{Y}|\theta)$ and $p(X, \mathbf{Z})$ as follows:

$$\begin{pmatrix} \mu_1^I \\ \vdots \\ \mu_q^I \\ \mu_1^I \mu_1^{local} \\ \vdots \\ \mu_q^I \mu_q^{local} \end{pmatrix}$$

where μ_l^I is the l-th component (i.e. a scalar) of the moment vector associated to the marginal distribution $p(\mathbf{Y}|\theta)$, $\mu_l^I = \int I(\mathbf{Y} = \mathbf{y}_l) p(\mathbf{Y}|\theta) d\mathbf{Y} = p(\mathbf{y}_l|\theta)$, and μ_l^{local} is the "local" moment parameter associated to the marginal distribution $p(X, \mathbf{Z}|\mathbf{y} = l)$, $\mu_l = \int s_l(X, \mathbf{Z}) p(X, \mathbf{Z}|\mathbf{y} = l, \theta) dX\mathbf{Z}$.

Again, we can see that the moment parameters of this conditional distribution can be expressed as a composition of the local moment parameters of each of the distributions conditioned to each configuration of the multinomial parents variables.

4.3 From Moment to Natural Parameters

We now look at how to transform from moment to natural parameters. Initially, the dimension of the optimization problem is $2q$ (i.e the dimension of the natural parameter vector), however we can exploit the structure present in the natural space and see that it boils down to a q dimensional problem,

$$\arg \max_{(\theta_1, \ldots, \theta_q) \in \Theta} \sum_{l=1}^{q} \theta_l^T \mu_{q+l} - \mu_l B_l(\theta_l)$$

Furthermore, the above optimization problem decomposes in a set of q independent optimization problems,

$$\arg\max_{\theta_l \in \Theta_l} \theta_l^T \mu_{q+l} - \mu_l B_l(\theta_l)$$

The solution of the above optimization problem is not affected if the optimized expression is divided by μ_l, which is a scalar[2]. So this problem can be transformed as follows,

$$\arg\max_{\theta_l \in \Theta_l} \theta_l^T \mu'_l - B_l(\theta_l) \tag{4.2}$$

where $\mu'_l = \frac{1}{\mu_l}\mu_{q+l}$, i.e.the element-wise division of the vector μ_{q+l} by the scalar μ_l.

The above problem simply corresponds to the transformation of the moment parameters μ'_l to their corresponding natural parameters θ_l for the conditional distribution $p(X|\mathbf{Z}, \mathbf{y}_l)$. So again, we see how the transformation from moment to natural decomposes in a series of local transformations.

5 Maximum Likelihood

In this section we look at the maximum likelihood problem in CEF-BNs. Our aim is to show how the solution to this multi-dimensional optimization problem has a common characterization in terms of exponential family representation and, moreover, boils down to smaller local problems for each conditional distribution[3].

Let us assume that we are given a set of m i.i.d. data samples $D = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}$ indexed by j. The maximum likelihood problem can be stated as follows:

$$\begin{aligned}
\theta^\star &= \arg\max_{\theta \in \Theta} \sum_{j=1}^{m} \ln p(\mathbf{x}^{(j)}|\theta) \\
&= \arg\max_{\theta \in \Theta} \sum_{j=1}^{m} \theta^T s(\mathbf{x}^{(j)}) - A(\theta) \\
&= \arg\max_{\theta \in \Theta} \theta^T \Big(\sum_{j=1}^{m} s(\mathbf{x}^{(j)}) \Big) - mA(\theta) \\
&= \arg\max_{\theta \in \Theta} \theta^T \Big(\frac{1}{m} \sum_{j=1}^{m} s(\mathbf{x}^{(j)}) \Big) - A(\theta)
\end{aligned}$$

[2] If $\mu_l = 0$ it would imply that $p(\mathbf{Y} = \mathbf{y}_t) = 0$, so it does not make sense to solve the problem.

[3] We point out that the following derivation is made without assuming that our models belongs to the regular exponential family and, in consequence, not using the equality of Equation 2.3.

where the last part is achieved by diving the optimized equation by the number of samples m, what does not affect the result of the optimization.

As widely known, the maximum likelihood is equivalent to a transformation form moment to natural parameters as stated in Equation 2.2,

$$\theta^{\star} = \theta\Big(\frac{1}{m}\sum_{j=1}^{m} s(\mathbf{x}^{(j)})\Big)$$

As shown in Section **??**, this problem decomposes for each conditional probability distribution. The above formation can be expressed in terms of local transformations defined in Equation 3.4,

$$\theta_i^{\star} = \theta_i\Big(\frac{1}{m}\sum_{j=1}^{m} s(x_i^{(j)}, \mathbf{pa}_i^{(j)})\Big)$$

To better understand the above decomposition, we should notice that $\theta_i(\mu_i)$ is directly related to maximum likelihood estimation of the conditional distribution $p(X_i|Pa(X_i))$,

$$\begin{aligned}
\theta^{\star} &= \arg\max_{\theta_i\in\Theta_i} \sum_{j=1}^{m} \ln p(x_i^{(j)}|\mathbf{pa}_i^{(j)}, \theta) \\
&= \arg\max_{\theta_i\in\Theta_i} \sum_{j=1}^{m} \Big(\theta_i^T s(x_i^{(j)}, \mathbf{pa}_i^{(j)}) - B_i(\theta_i)\Big) \\
&= \arg\max_{\theta_i\in\Theta_i} \theta_i^T \Big(\sum_{j=1}^{m} s(x_i^{(j)}, \mathbf{pa}_i^{(j)})\Big) - mB_i(\theta) \\
&= \arg\max_{\theta_i\in\Theta_i} \theta_i^T \Big(\frac{1}{m}\sum_{j=1}^{m} s(x_i^{(j)}, \mathbf{pa}_i^{(j)})\Big) - B_i(\theta)
\end{aligned}$$

where $\mathbf{pa}_i^{(j)}$ corresponds to the assignment to the parents of X_i according the to the j-th data sample $\mathbf{x}^{(j)}$.

For those conditional distributions with multinomial parents, the problem further decomposes as shown in Section 4.3.

6 EM algorithms in CEF-BNs

7 Variational Inference in CEF-BNs

8 Expectation Propagation Inference in CEF-BNs

APPENDIX

A A binary child given a binary parent

Let X and Y be two binary variables. The log-conditional probability of the child-node X given its parent-node Y is expressed as follows:

$$\begin{aligned}\ln p(X \mid Y) = I(X = x^1)I(Y = y^1) \ln p_{x^1|y^1} + I(X = x^2)I(Y = y^1) \ln p_{x^2|y^1} \\ +I(X = x^1)I(Y = y^2) \ln p_{x^1|y^2} + I(X = x^2)I(Y = y^2) \ln p_{x^2|y^2}\end{aligned}$$

This conditional probability distribution can be expressed in different exponential forms as follows:

- **First form**:

$$\begin{aligned}\ln p(X \mid Y) &= \theta^T s(X, Y) - A(\theta) \\ &= \begin{pmatrix} \ln p_{x^1|y^1} \\ \ln p_{x^2|y^1} \\ \ln p_{x^1|y^2} \\ \ln p_{x^2|y^2} \end{pmatrix}^T \begin{pmatrix} I(X = x^1)I(Y = y^1) \\ I(X = x^2)I(Y = y^1) \\ I(X = x^1)I(Y = y^2) \\ I(X = x^2)I(Y = y^2) \end{pmatrix} - 0 \\ &= \begin{pmatrix} \theta_{11} \\ \theta_{21} \\ \theta_{12} \\ \theta_{22} \end{pmatrix}^T \begin{pmatrix} I(X = x^1)I(Y = y^1) \\ I(X = x^2)I(Y = y^1) \\ I(X = x^1)I(Y = y^2) \\ I(X = x^2)I(Y = y^2) \end{pmatrix} - 0\end{aligned}$$

- **Second form**:

$$
\begin{aligned}
\ln p(X \mid Y) &= \theta(Y)^T s(X) - A(Y) \\
&= \begin{pmatrix} I(Y = y^1) \ln p_{x^1|y^1} + I(Y = y^2) \ln p_{x^1|y^2} \\ I(Y = y^1) \ln p_{x^2|y^1} + I(Y = y^2) \ln p_{x^2|y^2} \end{pmatrix}^T \begin{pmatrix} I(X = x^1) \\ I(X = x^2) \end{pmatrix} - 0 \\
&= \begin{pmatrix} m_1^Y \cdot \theta_{11} + m_2^Y \cdot \theta_{12} \\ m_1^Y \cdot \theta_{21} + m_2^Y \cdot \theta_{22} \end{pmatrix}^T \begin{pmatrix} I(X = x^1) \\ I(X = x^2) \end{pmatrix} - 0
\end{aligned}
$$

- **Third form**:

$$
\begin{aligned}
\ln p(X \mid Y) &= \theta(X)^T s(Y) - A(X) \\
&= \begin{pmatrix} I(X = x^1) \ln p_{x^1|y^1} + I(X = x^2) \ln p_{x^2|y^1} \\ I(X = x^1) \ln p_{x^1|y^2} + I(X = x^2) \ln p_{x^2|y^2} \end{pmatrix}^T \begin{pmatrix} I(Y = y^1) \\ I(Y = y^2) \end{pmatrix} - 0 \\
&= \begin{pmatrix} m_1^X \cdot \theta_{11} + m_2^X \cdot \theta_{21} \\ m_1^X \cdot \theta_{12} + m_2^X \cdot \theta_{22} \end{pmatrix}^T \begin{pmatrix} I(Y = y^1) \\ I(Y = y^2) \end{pmatrix} - 0
\end{aligned}
$$

B A multinomial child given a set of multinomial parents

Let X be a multinomial variable with k possible values such that $k \geq 2$, and let $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ denote the set of parents of X, such that all of them are multinomial. Each parent Y_i, $1 \geq i \geq n$, has r_i possible values or states such that $r_i \geq 2$. A parental configuration for the child-node X is then a set of n elements $\{Y_1 = y_1^v, \ldots, Y_i = y_i^v, \ldots, Y_n = y_n^v\}$ such that y_i^v denotes a potential value of variable Y_i such that $1 \leq v \leq r_i$. Let $q = r_1 \times \ldots \times r_n$ denote the total number of parental configurations, and let $\mathbf{y}^l$ denote the l^{th} parental configuration such that $1 \leq l \leq q$.

The log-conditional probability of the child-node X given its parent-nodes $\mathbf{Y}$ can be expressed as follows:

$$\ln p(X \mid \mathbf{Y}) = \sum_{j=1}^{k} \sum_{l=1}^{q} I(X = x^j) I(\mathbf{Y} = \mathbf{y}^l) \ln p_{x^j|\mathbf{y}^l}$$

Similarly the above log-conditional probability can be expressed in the following exponential forms:

- **First form**:

$$\begin{aligned}
\ln p(X \mid \mathbf{Y}) &= \theta^T s(X, \mathbf{Y}) - A(\theta) \\
&= \begin{pmatrix} \ln p_{x^1|\mathbf{y}^1} \\ \vdots \\ \ln p_{x^1|\mathbf{y}^q} \\ \vdots \\ \ln p_{x^k|\mathbf{y}^1} \\ \vdots \\ \ln p_{x^k|\mathbf{y}^q} \end{pmatrix}^T \begin{pmatrix} I(X = x^1) I(\mathbf{Y} = \mathbf{y}^1) \\ \vdots \\ I(X = x^1) I(\mathbf{Y} = \mathbf{y}^q) \\ \vdots \\ I(X = x^k) I(\mathbf{Y} = \mathbf{y}^1) \\ \vdots \\ I(X = x^k) I(\mathbf{Y} = \mathbf{y}^q) \end{pmatrix} - 0 \\
&= \begin{pmatrix} \theta_{11} \\ \vdots \\ \theta_{1q} \\ \vdots \\ \theta_{k1} \\ \vdots \\ \theta_{kq} \end{pmatrix}^T \begin{pmatrix} I(X = x^1) I(\mathbf{Y} = \mathbf{y}^1) \\ \vdots \\ I(X = x^1) I(\mathbf{Y} = \mathbf{y}^q) \\ \vdots \\ I(X = x^k) I(\mathbf{Y} = \mathbf{y}^1) \\ \vdots \\ I(X = x^k) I(\mathbf{Y} = \mathbf{y}^q) \end{pmatrix} - 0
\end{aligned}$$

- **Second form**:

$$\begin{aligned}
\ln p(X \mid \mathbf{Y}) &= \theta(\mathbf{Y})^T s(X) - A(\mathbf{Y}) \\
&= \begin{pmatrix} I(\mathbf{Y} = \mathbf{y}^1) \ln p_{x^1|\mathbf{y}^1} + \ldots + I(\mathbf{Y} = \mathbf{y}^q) \ln p_{x^1|\mathbf{y}^q} \\ \vdots \\ I(\mathbf{Y} = \mathbf{y}^1) \ln p_{x^k|\mathbf{y}^1} + \ldots + I(\mathbf{Y} = \mathbf{y}^q) \ln p_{x^k|\mathbf{y}^q} \end{pmatrix}^T \begin{pmatrix} I(X = x^1) \\ \vdots \\ I(X = x^k) \end{pmatrix} - 0 \\
&= \begin{pmatrix} \mathbf{m}_1^{\mathbf{Y}} \cdot \theta_{11} + m_q^{\mathbf{Y}} \cdot \theta_{1q} \\ \vdots \\ \mathbf{m}_1^{\mathbf{Y}} \cdot \theta_{k1} + m_q^{\mathbf{Y}} \cdot \theta_{kq} \end{pmatrix}^T \begin{pmatrix} I(X = x^1) \\ \vdots \\ I(X = x^k) \end{pmatrix} - 0
\end{aligned}$$

such that $\mathbf{m}_1^{\mathbf{Y}} = \prod_{i=1}^n I(Y_i = y_i^1) = \prod_{i=1}^n m_1^{Y_i}$ denotes the expected sufficient statistics for the first parental configuration, and $\mathbf{m}_q^{\mathbf{Y}} = \prod_{i=1}^n I(Y_i = y_i^{r_i}) = \prod_{i=1}^n m_{r_i}^{Y_i}$ denotes the expected sufficient statistics for the last parental configuration.

- **Third form**:

$$\begin{aligned}
\ln p(X \mid \mathbf{Y}) &= \theta(X)^T s(\mathbf{Y}) - A(X) \\
&= \begin{pmatrix} I(X = x^1) \ln p_{x^1|\mathbf{y}^1} + \ldots + I(X = x^k) \ln p_{x^k|\mathbf{y}^1} \\ \vdots \\ I(X = x^1) \ln p_{x^1|\mathbf{y}^q} + \ldots + I(X = x^k) \ln p_{x^k|\mathbf{y}^q} \end{pmatrix}^T \begin{pmatrix} I(\mathbf{Y} = \mathbf{y}^1) \\ \vdots \\ I(\mathbf{Y} = \mathbf{y}^q) \end{pmatrix} - 0 \\
&= \begin{pmatrix} m_1^X \cdot \theta_{11} + \ldots + m_k^X \cdot \theta_{k1} \\ \vdots \\ m_1^X \cdot \theta_{1q} + \ldots + m_k^X \cdot \theta_{kq} \end{pmatrix}^T \begin{pmatrix} I(\mathbf{Y} = \mathbf{y}^1) \\ \vdots \\ I(\mathbf{Y} = \mathbf{y}^q) \end{pmatrix} - 0
\end{aligned}$$

$$
\begin{aligned}
\ln p(X \mid \mathbf{Y}) &= \theta(X, \mathbf{Y}')^T s(Y_i) - A(X) \text{ such that } \mathbf{Y}' = \mathbf{Y} \setminus Y_i \\
&= \begin{pmatrix} m_1^X I(\mathbf{Y}' = \mathbf{y}'^1) \ln p_{x^1|\mathbf{y}'^1} + \ldots + m_k^X I(\mathbf{Y}' = \mathbf{y}'^1) \ln p_{x^k|\mathbf{y}'^1} \\ \vdots \\ m_1^X I(\mathbf{Y}' = \mathbf{y}'^{q'}) \ln p_{x^1|\mathbf{y}'^{q'}} + \ldots + m_k^X I(\mathbf{Y}' = \mathbf{y}'^{q'}) \ln p_{x^k|\mathbf{y}'^{q'}} \end{pmatrix}^T \begin{pmatrix} I(Y_i = y_i^1) \\ \vdots \\ I(Y_i = y_i^{r_i}) \end{pmatrix} - 0 \\
&= \begin{pmatrix} m_1^X \cdot \mathbf{m}_1^{\mathbf{Y}'} \cdot \theta'_{11} + \ldots + m_k^X \cdot \mathbf{m}_1^{\mathbf{Y}'} \cdot \theta'_{k1} \\ \vdots \\ m_1^X \cdot \mathbf{m}_{q'}^{\mathbf{Y}'} \cdot \theta'_{1q'} + \ldots + m_k^X \cdot \mathbf{m}_{q'}^{\mathbf{Y}'} \cdot \theta'_{kq'} \end{pmatrix}^T \begin{pmatrix} I(Y_i = y_i^1) \\ \vdots \\ I(Y_i = y_i^{r_i}) \end{pmatrix} - 0
\end{aligned}
$$

where $\mathbf{m}_1^{\mathbf{Y}'} = I(\mathbf{Y}' = \mathbf{y}'^1) = I(Y_1 = y_1^1) \cdot \ldots I(Y_{i-1} = y_{i-1}^1) \cdot I(Y_{i+1} = y_{i+1}^1) \cdot \ldots I(Y_n = y_n^1)$ denotes the expected sufficient statistics for the first configuration of the parent set $\mathbf{Y}' = \mathbf{Y} \setminus Y_i$, and $\mathbf{m}_{q'}^{\mathbf{Y}'} = I(\mathbf{Y}' = \mathbf{y}'^{q'}) = I(Y_1 = y_1^{q'}) \cdot \ldots I(Y_{i-1} = y_{i-1}^{q'}) \cdot I(Y_{i+1} = y_{i+1}^{q'}) \cdot \ldots I(Y_n = y_n^{q'})$ denotes the expected sufficient statistics for the last configuration of the parent set $\mathbf{Y}'$, with $q' = q/r_i$ denotes the total number of configurations of the parent set $\mathbf{Y}'$.

C A normal child given a set of normal parents

Let X be a normal variable and $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ denote the set of parents of X, such that all of them are normal.

The log-conditional probability of X given its parents $\mathbf{Y}$ can be expressed as follows:

$$\ln p(X|Y_1, \ldots, Y_n) = \ln \left(\frac{1}{\sigma\sqrt{2(\beta_0 + \sum_i^n \beta_i Y_i)}} \mathrm{e}^{-\frac{(y-(\beta_0+\sum_i^n \beta_i Y_i))^2}{2\sigma^2}} \right)$$

Similarly the above log-conditional probability can be expressed in the following exponential forms:

- **First form - Joint suff. stat. (Maxim. Likelihood)**:

$$\begin{aligned}
\ln p(X \mid \mathbf{Y}) &= \theta^T s(X, \mathbf{Y}) - A(\theta) + h(\mathbf{Y}) \\
&= \begin{pmatrix}
\frac{-1}{2\sigma^2} &=& \theta_{\text{-}1} \\
\frac{\beta_0}{\sigma^2} &=& \theta_0 \\
\frac{\beta_1}{\sigma^2} &=& \theta_1 \\
\vdots & & \\
\frac{\beta_n}{\sigma^2} &=& \theta_n \\
\frac{-\beta_0\beta_1}{2\sigma^2} &=& \theta_{01} \\
\vdots & & \\
\frac{-\beta_0\beta_n}{2\sigma^2} &=& \theta_{0n} \\
\frac{-\beta_1^2}{2\sigma^2} &=& \theta_{1^2} \\
\vdots & & \\
\frac{-\beta_n^2}{2\sigma^2} &=& \theta_{n^2} \\
\frac{-\beta_1\beta_2}{2\sigma^2} &=& \theta_{12} \\
\vdots & & \\
\frac{-\beta_1\beta_n}{2\sigma^2} &=& \theta_{1n} \\
\vdots & & \\
\frac{-\beta_{n-1}\beta_n}{2\sigma^2} &=& \theta_{n\text{-}1n}
\end{pmatrix}^T
\begin{pmatrix}
X^2 &=& m_{X^2} \\
X &=& m_X \\
XY_1 &=& m_{XY_1} \\
\vdots & & \\
XY_n &=& m_{XY_n} \\
Y_1 &=& m_{Y_1} \\
\vdots & & \\
Y_n &=& m_{Y_n} \\
Y_1^2 &=& m_{Y_1^2} \\
\vdots & & \\
Y_n^2 &=& m_{Y_n^2} \\
Y_1Y_2 &=& m_{Y_1Y_2} \\
\vdots & & \\
Y_1Y_n &=& m_{Y_1Y_n} \\
\vdots & & \\
Y_{n\text{-}1}Y_n &=& m_{Y_{n\text{-}1}Y_n}
\end{pmatrix}
- \left(\frac{\beta_0^2}{2\sigma^2} + \ln\sigma \right) + \frac{1}{\ln\sqrt{2\mu_{X|Y}}}
\end{aligned}$$

  where $\mu_{X|Y} = \beta_0 + \sum_i^n \beta_i Y_i$

  - **From moment to natural parameters: (matrix representation)**

$$
\begin{aligned}
\ln p(X \mid \mathbf{Y}) &= \theta^T s(X, \mathbf{Y}) - A(\theta) + h(\mathbf{Y}) \\
&= \begin{pmatrix} \beta_0(\sigma^2)^{-1} & = & \theta_0 \\ -\beta_0\beta^T(2\sigma^2)^{-1} & = & \theta_{\beta_0\beta^T} \\ -(2\sigma^2)^{-1} & = & \theta_{\text{-}1} \\ \beta(\sigma^2)^{-1} & = & \theta_\beta \\ -\beta'\beta^T(2\sigma^2)^{-1} & = & \theta_{\beta\beta^T} \end{pmatrix}^T \begin{pmatrix} X & = & \mathrm{E}(X) \\ Y & = & \mathrm{E}(Y) \\ XX^T & = & \mathrm{E}(XX^T) \\ YX^T & = & \mathrm{E}(YX^T) \\ YY^T & = & \mathrm{E}(YY^T) \end{pmatrix} \\
&- \left(\frac{\beta_0^2}{2\sigma^2} + \ln\sigma\right) + \frac{1}{\ln\sqrt{2\mu_{X|Y}}}
\end{aligned}
$$

where

$$
\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} \qquad Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}
$$

* FIRST STEP:

$$
\begin{aligned}
\mu_X &= E(X) \\
\mu_{\mathbf{Y}} &= \mathrm{E}(Y) \\
\Sigma_{XX} &= \mathrm{E}(XX^T) - \mathrm{E}(X)\,\mathrm{E}(X)^T \\
\Sigma_{\mathbf{YY}} &= \mathrm{E}(YY^T) - \mathrm{E}(Y)\,\mathrm{E}(Y)^T \\
\Sigma_{X\mathbf{Y}} &= \mathrm{E}(YX^T)^T - \mathrm{E}(X)\,\mathrm{E}(Y)^T \\
\Sigma_{\mathbf{Y}X} &= \mathrm{E}(YX^T) - \mathrm{E}(Y)\,\mathrm{E}(X)
\end{aligned}
$$

* SECOND STEP (Theorem 7.4 in page 253, Koller & Friedman):

$$
\begin{aligned}
\beta_0 &= \mu_X - \Sigma_{X\mathbf{Y}}\Sigma_{\mathbf{YY}}^{-1}\mu_{\mathbf{Y}} \\
\beta &= \Sigma_{X\mathbf{Y}}\Sigma_{\mathbf{YY}}^{-1} \\
\sigma^2 &= \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}
\end{aligned}
$$

All natural parameters θ can now be calculated considering these equations.

– **From natural to moment parameters:** Via inference.

- **Second form**:

$$\begin{aligned}
\ln p(X \mid \mathbf{Y}) &= \theta(\mathbf{Y})^T s(X) - A\big(\theta(\mathbf{Y})\big) + h(\mathbf{Y}) \\
&= \begin{pmatrix} \frac{\mu_{X|Y}}{\sigma^2} \\ \frac{-1}{2\sigma^2} \end{pmatrix}^T \begin{pmatrix} X \\ X^2 \end{pmatrix} - \left(\frac{\mu_{X|Y}^2}{2\sigma^2} + \ln \sigma \right) + \ln \frac{1}{\sqrt{2\mu_{X|Y}}} \\
&= \begin{pmatrix} \theta_0 + \sum_i^n \theta_i m^{Y_i} \\ \theta_{\text{-}1} \end{pmatrix}^T \begin{pmatrix} X \\ X^2 \end{pmatrix} - \left(\frac{\ln(2\theta_{\text{-}1})}{2} - \theta_{\text{-}1} \left(\theta_0 + \sum_i^n \theta_i m^{Y_i} \right)^2 \right) \\
&+ \ln \frac{1}{\sqrt{2(\theta_0 + \sum_i^n \theta_i m^{Y_i})}}
\end{aligned}$$

- **Third form**:

$$
\begin{aligned}
\ln p(X \mid \mathbf{Y}) &= \theta(X)^T s(\mathbf{Y}) - A\big(\theta(X)\big) + h(\mathbf{Y}) \\
&= \begin{pmatrix} -\frac{\beta_1^2}{2\sigma^2} \\ \cdots \\ -\frac{\beta_n^2}{2\sigma^2} \\ \frac{\beta_1(X-\beta_0)}{\sigma^2} \\ \cdots \\ \frac{\beta_n(X-\beta_0)}{\sigma^2} \\ -\frac{\beta_1\beta_2}{\sigma^2} \\ \cdots \\ -\frac{\beta_1\beta_n}{\sigma^2} \\ \cdots \\ -\frac{\beta_{n-1}\beta_n}{\sigma^2} \end{pmatrix}^T \begin{pmatrix} Y_1^2 \\ \cdots \\ Y_n^2 \\ Y_1 \\ \cdots \\ Y_n \\ Y_1Y_2 \\ \cdots \\ Y_1Y_n \\ \cdots \\ Y_{n-1}Y_n \end{pmatrix} - \left(\frac{(X-\beta_0)^2}{\sigma^2} + \ln\sigma \right) + \frac{1}{\ln\sqrt{2\mu_{X|Y}}} \\
&= \begin{pmatrix} \theta_{1^2} \\ \cdots \\ \theta_{n^2} \\ \theta_1 m^X + \theta_{01} \\ \cdots \\ \theta_n m^X + \theta_{0n} \\ \theta_{12} \\ \cdots \\ \theta_{1n} \\ \cdots \\ \theta_{n\text{-}1n} \end{pmatrix}^T \begin{pmatrix} Y_1^2 \\ \cdots \\ Y_n^2 \\ Y_1 \\ \cdots \\ Y_n \\ Y_1Y_2 \\ \cdots \\ Y_1Y_n \\ \cdots \\ Y_{n-1}Y_n \end{pmatrix} - \left(\frac{X^2 - 2X\beta_0 + \beta_0^2}{\sigma^2} + \ln\sigma \right) + \frac{1}{\ln\sqrt{2\mu_{X|Y}}} \\
&= \begin{pmatrix} \theta_{1^2} \\ \cdots \\ \theta_{n^2} \\ \theta_1 m^X + \theta_{01} \\ \cdots \\ \theta_n m^X + \theta_{0n} \\ \theta_{12} \\ \cdots \\ \theta_{1n} \\ \cdots \\ \theta_{n\text{-}1n} \end{pmatrix}^T \begin{pmatrix} Y_1^2 \\ \cdots \\ Y_n^2 \\ Y_1 \\ \cdots \\ Y_n \\ Y_1Y_2 \\ \cdots \\ Y_1Y_n \\ \cdots \\ Y_{n-1}Y_n \end{pmatrix} - \left((-2\beta_{\text{-}1} m^{X^2} - 2m^X\beta_0 - \frac{1}{2}\beta_0^2\beta_{\text{-}1}^{-1}) + \frac{\ln(2\theta_{\text{-}1})}{2} \right) \\
&+ \ln\frac{1}{\sqrt{2(\theta_0 + \sum_i^n \theta_i m^{Y_i})}}
\end{aligned}
$$

D A base distribution given a binary parent

Let X be any base distribution variable, and let Y be a binary variable. The log-conditional probability of the child-node X given its binary parent-node Y is expressed as follows:

$$\begin{aligned}
&\ln p(X \mid Y) = I(Y = y^1) \ln p_{X|y^1} + I(Y = y^2) \ln p_{X|y^2} \\
&= I(Y = y^1)\Big(\theta_{X1} \cdot s(X) - A(\theta_{X1})\Big) + I(Y = y^2)\Big(\theta_{X2} \cdot s(X) - A(\theta_{X2})\Big) \\
&= I(Y = y^1) \cdot \theta_{X1} \cdot s(X) - I(Y = y^1) \cdot A(\theta_{X1}) + I(Y = y^2) \cdot \theta_{X2} \cdot s(X) - I(Y = y^2) \cdot A(\theta_{X2})
\end{aligned}$$

This conditional probability distribution can be expressed in different exponential forms as follows:

- **First form**:

$$\begin{aligned}
\ln p(X \mid Y) &= \theta^T s(X, Y) - A(\theta) \\
&= \begin{pmatrix} -A(\theta_{X1}) \\ -A(\theta_{X2}) \\ \theta_{X1} \\ \theta_{X2} \end{pmatrix}^T \begin{pmatrix} I(Y = y^1) \\ I(Y = y^2) \\ s(X) \cdot I(Y = y^1) \\ s(X) \cdot I(Y = y^2) \end{pmatrix} - 0
\end{aligned}$$

- **Second form**:

$$\begin{aligned}
\ln p(X \mid Y) &= \theta(Y)^T s(X) - A(Y) \\
&= \begin{pmatrix} I(Y = y^1) \\ I(Y = y^2) \\ I(Y = y^1) \cdot \theta_{X1} \\ I(Y = y^2) \cdot \theta_{X2} \end{pmatrix}^T \begin{pmatrix} -A(\theta_{X1}) \\ -A(\theta_{X2}) \\ s(X) \\ s(X) \end{pmatrix} - 0 \\
&= \begin{pmatrix} m_1^Y \\ m_2^Y \\ m_1^Y \cdot \theta_{X1} \\ m_2^Y \cdot \theta_{X2} \end{pmatrix}^T \begin{pmatrix} -A(\theta_{X1}) \\ -A(\theta_{X2}) \\ s(X) \\ s(X) \end{pmatrix} - 0
\end{aligned}$$

- **Third form**:

$$
\begin{aligned}
\ln p(X \mid Y) &= \theta(X)^T s(Y) - A(X) \\
&= \begin{pmatrix} -A(\theta_{X1}) \\ -A(\theta_{X2}) \\ s(X) \cdot \theta_{X1} \\ s(X) \cdot \theta_{X2} \end{pmatrix}^T \begin{pmatrix} I(Y = y^1) \\ I(Y = y^2) \\ I(Y = y^1) \\ I(Y = y^2) \end{pmatrix} - 0
\end{aligned}
$$

E A base distribution given a set of multinomial parents

Let X be any base distribution, and let $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ denote the set of parents of X, such that all of them are multinomial. Each parent Y_i, $1 \geq i \geq n$, has r_i possible values or states such that $r_i \geq 2$. A parental configuration for the child-node X is then a set of n elements $\{Y_1 = y_1^v, \ldots, Y_i = y_i^v, \ldots, Y_n = y_n^v\}$ such that y_i^v denotes a potential value of variable Y_i such that $1 \leq v \leq r_i$. Let $q = r_1 \times \ldots \times r_n$ denote the total number of parental configurations, and let $\mathbf{y}^l$ denote the l^{th} parental configuration such that $1 \leq l \leq q$.

The log-conditional probability of the child-node X given its parent-nodes $\mathbf{Y}$ can be expressed as follows:

$$\begin{aligned}
\ln p(X \mid Y) &= \sum_{l=1}^{q} I(\mathbf{Y} = \mathbf{y}^l) \cdot \ln p_{X|\mathbf{y}^l} \\
&= \sum_{l=1}^{q} I(\mathbf{Y} = \mathbf{y}^l) \cdot \Big(\theta_{Xl} \cdot s(X) \cdot A(\theta_{Xl})\Big) \\
&= \sum_{l=1}^{q} I(\mathbf{Y} = \mathbf{y}^l) \cdot \theta_{Xl} \cdot s(X) - I(\mathbf{Y} = \mathbf{y}^l) \cdot A(\theta_{Xl})
\end{aligned}$$

This conditional probability distribution can be expressed in different exponential forms as follows:

- **First form**:

$$\begin{aligned}
\ln p(X \mid \mathbf{Y}) &= \theta^T s(X, \mathbf{Y}) - A(\theta) \\
&= \begin{pmatrix} -A(\theta_{X1}) \\ \vdots \\ -A(\theta_{Xq}) \\ \theta_{X1} \\ \vdots \\ \theta_{Xq} \end{pmatrix}^T \begin{pmatrix} I(\mathbf{Y} = \mathbf{y}^1) \\ \vdots \\ I(\mathbf{Y} = \mathbf{y}^q) \\ s(X) \cdot I(\mathbf{Y} = \mathbf{y}^1) \\ \vdots \\ s(X) \cdot I(\mathbf{Y} = \mathbf{y}^q) \end{pmatrix} - 0
\end{aligned}$$

- **Second form**:

$$
\begin{aligned}
\ln p(X \mid \mathbf{Y}) &= \theta(\mathbf{Y})^T s(X) - A(\mathbf{Y}) \\
&= \begin{pmatrix} I(\mathbf{Y} = \mathbf{y}^1) \\ \vdots \\ I(\mathbf{Y} = \mathbf{y}^q) \\ I(\mathbf{Y} = \mathbf{y}^1) \cdot \theta_{X1} \\ \vdots \\ I(\mathbf{Y} = \mathbf{y}^q) \cdot \theta_{Xq} \end{pmatrix}^T \begin{pmatrix} -A(\theta_{X1}) \\ \vdots \\ -A(\theta_{Xq}) \\ s(X) \\ \vdots \\ s(X) \end{pmatrix} - 0 \\
&= \begin{pmatrix} \mathbf{m}_1^{\mathbf{Y}} \\ \vdots \\ \mathbf{m}_q^{\mathbf{Y}} \\ \mathbf{m}_1^{\mathbf{Y}} \cdot \theta_{X1} \\ \vdots \\ \mathbf{m}_q^{\mathbf{Y}} \cdot \theta_{Xq} \end{pmatrix}^T \begin{pmatrix} -A(\theta_{X1}) \\ \vdots \\ -A(\theta_{Xq}) \\ s(X) \\ \vdots \\ s(X) \end{pmatrix} - 0
\end{aligned}
$$

- **Third form**:

$$
\begin{aligned}
\ln p(X \mid \mathbf{Y}) &= \theta(X)^T s(\mathbf{Y}) - A(X) \\
&= \begin{pmatrix} -A(\theta_{X1}) \\ \vdots \\ -A(\theta_{Xq}) \\ s(X) \cdot \theta_{X1} \\ \vdots \\ s(X) \cdot \theta_{Xq} \end{pmatrix}^T \begin{pmatrix} I(\mathbf{Y} = \mathbf{y}^1) \\ \vdots \\ I(\mathbf{Y} = \mathbf{y}^q) \\ I(\mathbf{Y} = \mathbf{y}^1) \\ \vdots \\ I(\mathbf{Y} = \mathbf{y}^q) \end{pmatrix} - 0
\end{aligned}
$$

$$
\begin{aligned}
\ln p(X \mid \mathbf{Y}) &= \theta(X, \mathbf{Y}')^T s(Y_i) - A(X) \text{ such that } \mathbf{Y}' = \mathbf{Y} \setminus Y_i \\
&= \begin{pmatrix} -A(\theta_{X1}) \\ \vdots \\ -A(\theta_{Xq}) \\ s(X) \cdot \mathbf{m}_1^{\mathbf{Y}'} \cdot \theta'_{X1} + \ldots + s(X) \cdot \mathbf{m}_1^{\mathbf{Y}'} \cdot \theta'_{X1} \\ \vdots \\ s(X) \cdot \mathbf{m}_{q'}^{\mathbf{Y}'} \cdot \theta'_{Xq'} + \ldots + s(X) \cdot \mathbf{m}_{q'}^{\mathbf{Y}'} \cdot \theta'_{Xq'} \end{pmatrix}^T \begin{pmatrix} I(Y_i = y_i^1) \\ \vdots \\ I(Y_i = y_i^{r_i}) \\ I(Y_i = y_i^1) \\ \vdots \\ I(Y_i = y_i^{r_i}) \end{pmatrix} - 0
\end{aligned}
$$

Notations

The list below presents a summary of the used notations:

X	Child variable
k	Range of possible values of a multinomial variable X
j	Index over X values, i.e., $1 \geq j \geq k$
Y	One parent variable
$\mathbf{Y}$	Set of parent variables
n	Number of parent variables
i	Index over parent variables, i.e., $1 \geq i \geq n$
r_i	Range of possible values of a multinomial variable Y_i
q	Total number of configurations of a multinomial parent set $\mathbf{Y}$
l	Index over the possible parental configuration values, i.e., $1 \geq l \geq q$
$\mathbf{y}^l$	The l^{th} configuration of a multinomial parent set $\mathbf{Y}$
θ_{jl}	Equal to $\ln p_{x^j \mid \mathbf{y}^l}$, denoting the log-conditional probability of X in its state j given the l^{th} parent configuration
θ_{Xl}	Equal to $\ln p_{X \mid \mathbf{y}^l}$, denoting the log-conditional probability of a base distribution variable X given the l^{th} parent configuration
p	Probability distribution
m	Expected sufficient statistics
s	Sufficient statistics