Probabilistic graphical models
CPSC 532c (Topics in AI)
Stat 521a (Topics in multivariate analysis)

Lecture 4

Kevin Murphy

Wedneday 21 September, 2004

Administrivia

- Mark Crowley will hold a regular discussion section on Fridays 1-2pm, CICSR 304. He will discuss HW1 and give a Matlab tutorial in the first meeting.

Course outline

- Representation
  - M Sep 13. Intro (ch 1)
  - W Sep 15. Bayes nets (ch 3)
  - M Sep 20. Markov nets (ch 5)
  - W Sep 22. Markov nets (ch 5); CPDs (ch 4)
- Exact inference in discrete state-spaces
  - M Sep 27. Gaussian BNs (ch 4); Intro to inference (ch 6)
  - W Sep 29. Variable elimination (ch 7)
  - M Oct 4. Variable elimination (ch 7)
  - W Oct 6. Junction tree (ch 8)
  - M Oct 11. Thanksgiving
  - W Oct 13. Guest lecture?
  - M Oct 18. Belief propagation (ch 8, handout)

- Learning
  - W Oct 20. Parameter learning in BNs (ch 12, 13)
  - M Oct 25. EM (ch 15)
  - W Oct 27. Parameter learning in MNs (handout)
  - M Nov 1. **Project proposals due**. Structure learning (ch 14).
  - W Nov 3. Structure learning (ch 14).
- Approximate inference
  - Mon Nov 8. Sampling (ch 9, handout)
  - Wed Nov 10. Sampling (ch 9, handout)
  - Mon Nov 15. Deterministic approx (ch 10, handout)
  - Wed Nov 17. Deterministic approx (ch 10, handout)
  - Mon Nov 22. Hybrid BNs (ch 11)
  - Wed Nov 24
  - Mon Nov 29
  - Wed Dec 1. **Last class**.

– Mon Dec 6. Project presentations

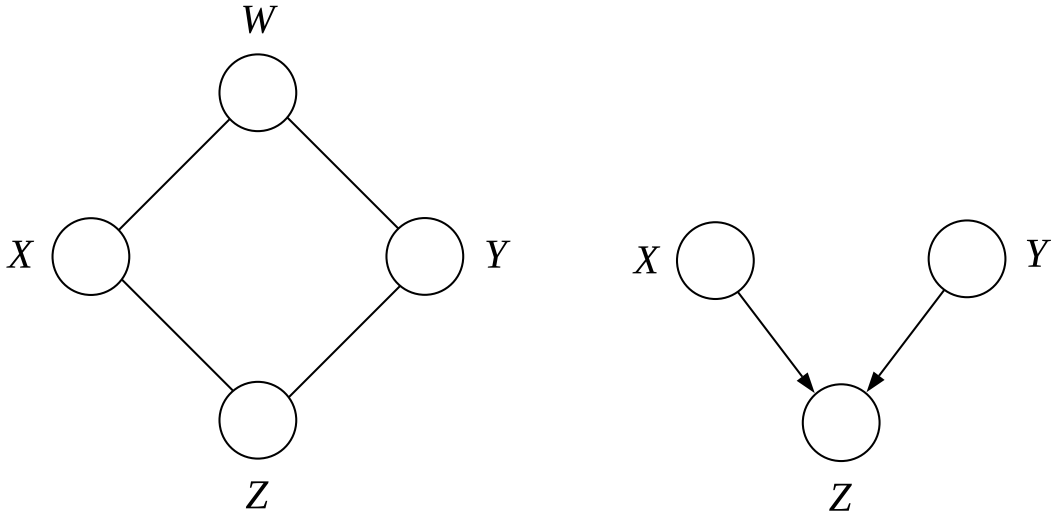
– Wed Dec 8. Project presentations

Review: independence properties

- Directed graphical models were defined in terms of local Markov property, from which we derived global Markov property (d-separation).
- Undirected graphical models were defined in terms of global Markov property (simple separation), from which we derived local Markov property.
- We can always represent any distribution by a DAG or an UG, by adding enough edges (i.e., reducing the size of $I(G)$ until it is inside $I(P)$).
- Some distributions can be represented perfectly by a DAG, others can be represented perfectly by an undirected graph, and others cannot be represented perfectly by either.

Review: Expressive Power

- Can we always convert directed ↔ undirected?
- No.



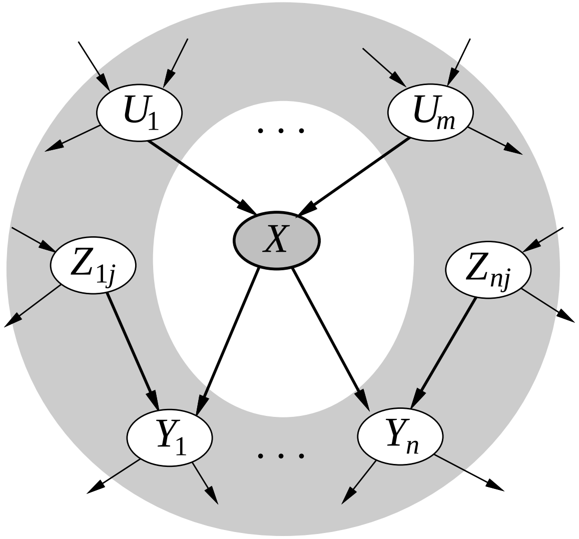
(a) (b)

No directed model can represent these and only these independencies.
$\mathbf{x} \perp \mathbf{y} \mid \{\mathbf{w}, \mathbf{z}\}$
$\mathbf{w} \perp \mathbf{z} \mid \{\mathbf{x}, \mathbf{y}\}$

No undirected model can represent these and only these independencies.
$\mathbf{x} \perp \mathbf{y}$

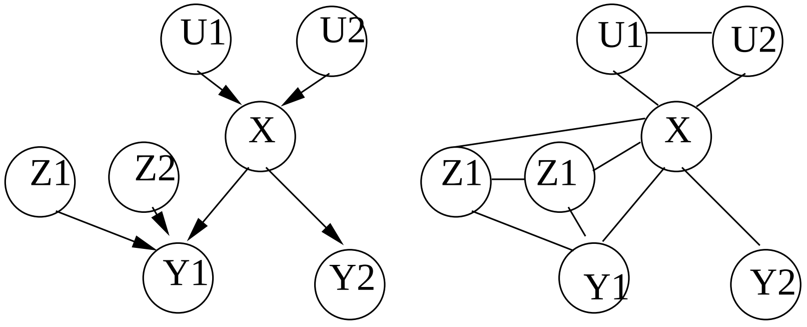
Converting Bayes nets to Markov nets

- Defn: A Markov net H is an I-map for a Bayes net G if $I(H) \subseteq I(G)$.
- We can construct a minimal I-map for a BN by finding the minimal Markov blanket for each node.
- We need to block all active paths coming into node X, from parents, children, and co-parents; so connect them all to X.



Moralization

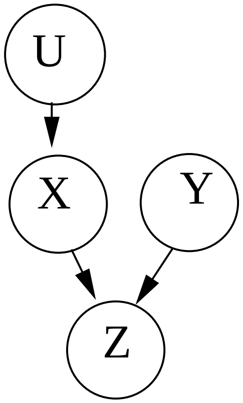
- Defn: the moral graph $H(G)$ of a DAG is constructed by adding undirected edges between any pair of disconnected ("unmarried") nodes X,Y that are parents of a child Z, and then dropping all remaining arrows.
- Thm 5.7.5: The moral graph $H(G)$ is the minimal I-map for Bayes net G.

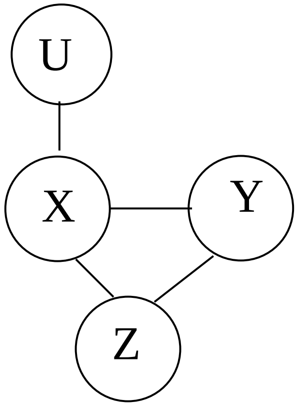


Bayes net to Markov net

- We assign each CPD to one of the clique potentials that contains it, e.g.

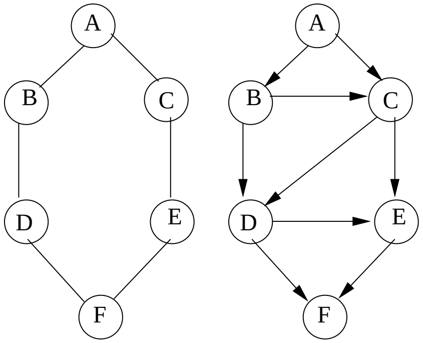
$$\begin{aligned} P(U,X,Y,Z) &= \frac{1}{Z}\psi(U,X) \times \psi(X,Y,Z) \\ &= \frac{1}{1}P(U)P(X|U) \times P(Y)P(Z|X,Y) \\ &= P(X,U) \times P(Z|X,Y)P(Y) \end{aligned}$$





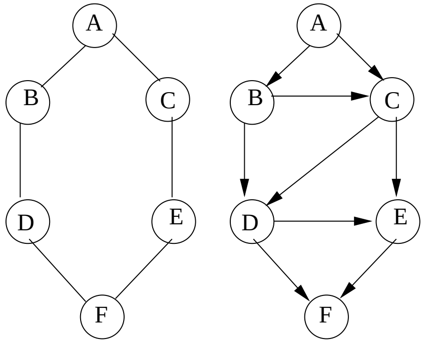
From Markov nets to Bayes nets

- Defn: A Bayes net G is an I-map for a Markov net H if $I(G) \subseteq I(H)$.
- We can construct a directed I-map by choosing a node ordering, and then picking the parents of node X_i as the subset U that renders X_i independent of its other predecessors $X_1, \ldots, X_{i-1}$.
- e.g., when we add C, the ancestors are A, B; since $C \not\perp B|A$, we need to add an edge from B to C.



- Different orderings may induce different edges.

Graph triangulation



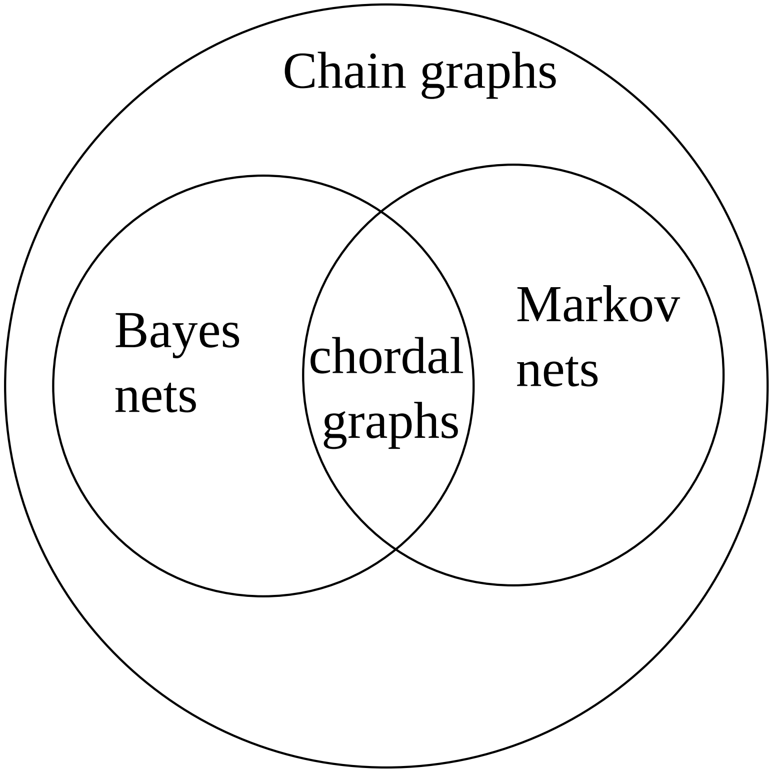
- The example above showed how we added extra edges to the DAG so that the largest loop was a 3-cycle (triangle).
- Defn: An undirected graph is called **chordal** or **triangulated** if every loop $X_1 - X_2 \cdots X_k - X_1$ for $k \geq 4$ has a chord, i.e., an edge connecting X_i and X_j for i, j non-adjacent.
- Defn: a directed graph is chordal if its underlying undirected graph is chordal.
- Thm 5.7.15: If G is a minimal I-map for Markov net H, then G is chordal.

Chordal graphs

- Converting a Bayes net to a Markov net adds extra moralization arcs.
- Converting a Markov net to a Bayes net adds extra triangulation arcs.
- Q: When can we convert a BN to a MN or vice versa without having to add extra arcs?
- A: when the graph is chordal.
- Thm 5.7.18 (if): Let H be a chordal Markov net. Then there is a Bayes net G s.t. $I(H) = I(G)$.
- Thm 5.7.16 (only-if): Let H be a non-chordal Markov net. Then there is no Bayes net G s.t. $I(H) = I(G)$.

Chordal graphs

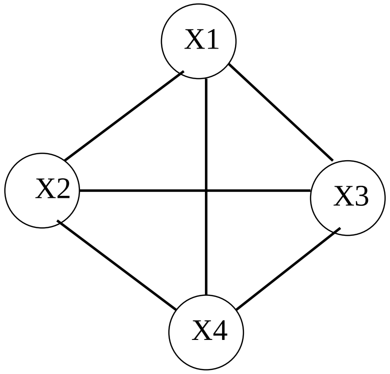
- Chordal graphs encode independencies that can be exactly represented by either directed or undirected graphs.
- Chain graphs combine directed and undirected graphs and represent a larger set of distributions.



Local structure

- So far, we have mostly studied independence properties that follow from the graph structure.
- Now we look at structure within the potentials/ CPDs of a model.
- Local structure often reduces the number of parameters in the model (so less data is needed for learning).
- Local structure can sometimes be exploited to speed up inference.

Factorizing clique potentials



- Sometimes a clique potential can be written as a product of subclique potentials.
- Max clique version

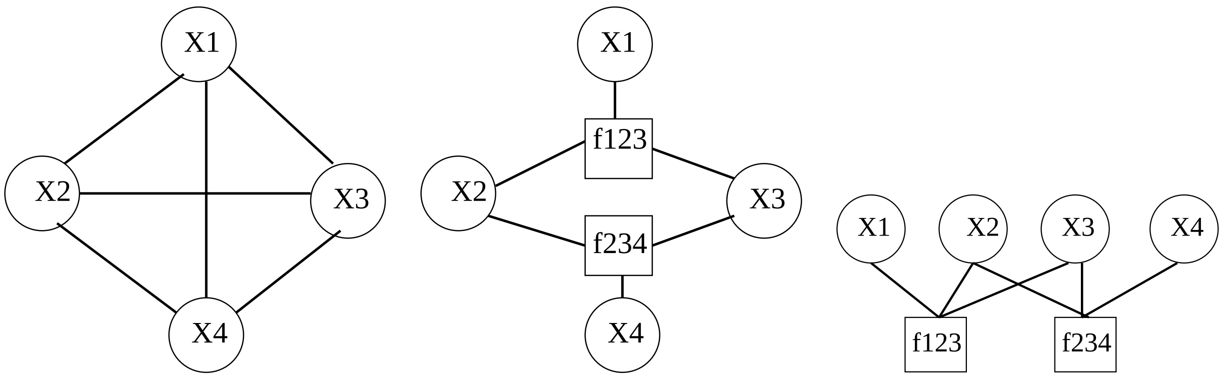
$$P(X_{1:4}) = \frac{1}{Z}\psi_{1234}(X_{1234})$$

- One possible sub clique version

$$P(X_{1:4}) \ = \ \frac{1}{Z}\psi_{123}(x_{123})\psi_{234}(x_{234})$$

Factor graphs

- Factorized potentials can be represented graphically using a factor graph.
- Defn: a factor graph is undirected bipartite graph with two kinds of nodes. Round nodes represent variables, square nodes represent factors (potentials), and there is an edge from each variable to every factor that mentions it.
- eg if $\psi_{1234} = \psi_{123} \times \psi_{234}$.

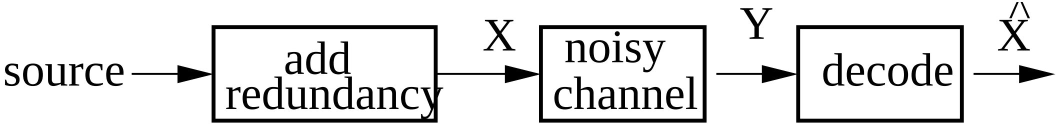


Application of factor graphs: error-control codes

- In ECC, we transmit a message (sequence of bits) X over a noisy channel. The receiver receives a noisy signal Y and has to estimate the original message:

$$\hat{X} = \arg\max_x P(y|x)P(x)$$

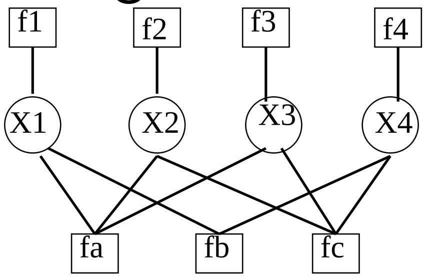
where $P(y|x)$ is the noise model and $P(x)$ is the source model.



- This is equivalent to inference in a probabilistic model.

Low-density parity check codes (LDPC)

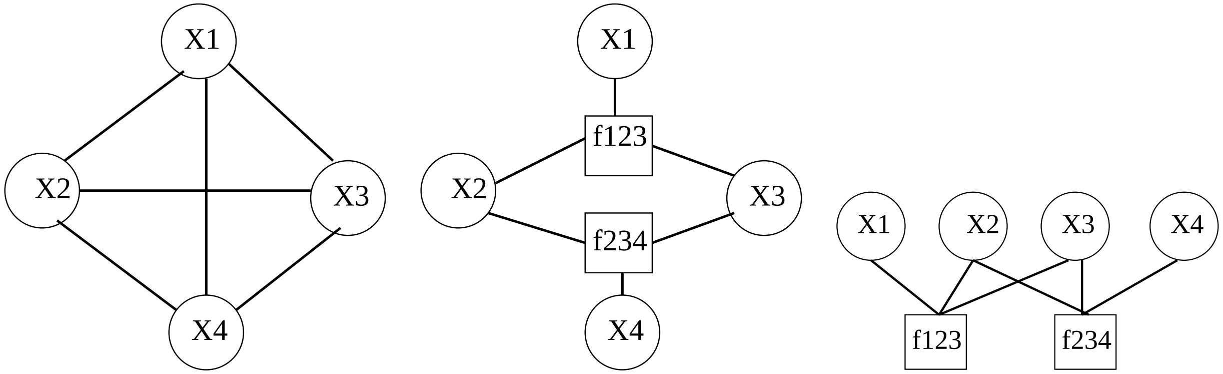
- A parity check code adds parity bits which are 1 iff an even number of the checked variables is 1, e.g.



  where $f_i = p(y_i|x_i)$ and f_a, f_b, f_c are parity check factors.

- Assigns 0 probability to settings of $\vec{x}$ that violate the parity constraints.
- If we impose an upper bound on the degree of the message nodes and the parity nodes, the graph is low-density.
- In an LDPC, the degree of the nodes is chosen from some distribution.
- This construction comes closer to the Shannon limit than any other code!

Representing the factors



- How do we parameterize the factors themselves?
- If each variable X_i has K possible discrete values, We can represent $f(X_1, X_2, X_3)$ as $K \times K \times K$ table.
- What do we do if the number of states K is large?
- e.g., consider a model of spelling which looks at all overlapping triples of letters, so $X_i \in \{a, b, \ldots, z\}$. We cannot afford 26^3 parameters!

Log-linear models

- We can parameterize each clique potential (factor) $\psi_c(x_c)$ as follows.
- Define a feature function $f_i(x_{c_i})$, where $C_i \subseteq C$ is a subset of the variables in C.
- Associate a scalar weight θ_i with each such feature.
- Then define

$$\psi_c(x_c) = \exp\left(\sum_{i \in I_C} \theta_i f_i(x_{c_i})\right)$$

- e.g., for the spelling model, $f_1(x_1, x_2, x_3) = \delta(x_{1:3} = \text{ing})$, $f_2(x_1, x_2, x_3) = \delta(x_{1:2} = \text{qu})$, etc.

Log-linear models

- Overall distribution is just a log-linear model (exponential family)

$$
\begin{aligned}
P(x|\theta) &= \frac{1}{Z(\theta)} \prod_{c \in C} \psi_c(x_c) \\
&= \frac{1}{Z(\theta)} \prod_{c \in C} \exp \left(\sum_{i \in I_C} \theta_i f_i(x_{c_i}) \right) \\
&= \frac{1}{Z(\theta)} \exp \left(\sum_{c \in C} \sum_{i \in I_C} \theta_i f_i(x_{c_i}) \right) \\
&= \frac{1}{Z(\theta)} \exp \left(\sum_{i \in I} \theta_i f_i(x_{c_i}) \right)
\end{aligned}
$$

- We can infer the graph structure from the features by connecting all the variables that are mentioned in the same function.

Log-linear models

$$P(x|\theta) = \frac{1}{Z(\theta)} \exp\left(\sum_{i \in I} \theta_i f_i(x_{c_i})\right)$$

- This form is completely general. By defining one indicator feature for every possible value of x_{c_i}, we can associate a separate parameter with each cell in the multi-dimensional array representing ψ_{c_i}.
- For Gaussians, we can use features $f_{ij}(x_i, x_j) = x_i \times x_j$ for every pair of connected nodes, $f_i(x_i) = x_i$ for every single node, and $f_0 = 1$ as a constant term:

$$\begin{aligned} P(x_{1:n}) &= \frac{1}{Z} e^{-H(x)} \\ H(x) &= \sum_{ij} V_{ij} x_i x_j + \sum_i \alpha_i x_i + C \end{aligned}$$

Structured CPDs

- So far we have discussed how to represent potentials/factors using a number of parameters that is less than exponential in the number of nodes in the clique.
- Now we will examine analogous techniques for compact representations of conditional probability distributions.
- We will start by examining compact representations for *un*conditional probability distributions, i.e., nodes with no parents.

Exponential Family

- For a numeric random variable $\mathbf{x}$

$$\begin{aligned} p(\mathbf{x}|\eta) &= h(\mathbf{x}) \exp\{\eta^\top T(\mathbf{x}) - A(\eta)\} \\ &= \frac{1}{Z(\eta)} h(\mathbf{x}) \exp\{\eta^\top T(\mathbf{x})\} \end{aligned}$$

  is an exponential family distribution with *natural parameter* η.

- Function $T(\mathbf{x})$ is a *sufficient statistic*.
- Function $A(\eta) = \log Z(\eta)$ is the log normalizer.
- Key idea: all you need to know about the data in order to estimate parameters is captured in the summarizing function $T(\mathbf{x})$.
- Examples: Bernoulli, binomial/geometric/negative-binomial, Poisson, gamma, multinomial, Gaussian, ...

Poisson

- For an integer count variable with *rate* λ:

$$\begin{aligned} p(x|\lambda) &= \frac{\lambda^x e^{-\lambda}}{x!} \\ &= \frac{1}{x!}\exp\{x\log\lambda - \lambda\} \end{aligned}$$

- Exponential family with:

$$\begin{aligned} \eta &= \log\lambda \\ T(x) &= x \\ A(\eta) &= \lambda = e^{\eta} \\ h(x) &= \frac{1}{x!} \end{aligned}$$

- e.g. number of photons $\mathbf{x}$ that arrive at a pixel during a fixed interval given mean intensity λ
- Other count densities: (neg)binomial, geometric.

Bernoulli Distribution

- For a binary random variable $x = \{0, 1\}$ with $p(x = 1) = \pi$:

$$
\begin{aligned}
p(x|\pi) &= \pi^x (1-\pi)^{1-x} \\
&= \exp\left\{ \log\left(\frac{\pi}{1-\pi}\right) x + \log(1-\pi) \right\}
\end{aligned}
$$

- Exponential family with:

$$
\begin{aligned}
\eta &= \log \frac{\pi}{1-\pi} \\
T(x) &= x \\
A(\eta) &= -\log(1-\pi) = \log(1+e^{\eta}) \\
h(x) &= 1
\end{aligned}
$$

- The *logistic* or *sigmoid* function links natural parameter and chance of heads

$$
\pi = \frac{1}{1+e^{-\eta}} = \frac{e^{\eta}}{1+e^{\eta}} = \mathrm{logistic}(\eta) = \sigma(\eta)
$$

Multinomial

- For a categorical (discrete), random variable taking on K possible values, let π_k be the probability of the k^{th} value. We can use a binary vector $\mathbf{x} = (x_1, x_2, \ldots, x_k, \ldots, x_K)$ in which $x_k{=}1$ if and only if the variable takes on its k^{th} value. Now we can write,

  $$p(\mathbf{x}|\pi) = \pi_1^{x_1}\pi_2^{x_2}\cdots\pi_K^{x_K} = \exp\left\{\sum_i x_i \log \pi_i\right\}$$

  Exactly like a probability table, but written using binary vectors.

- If we observe this variable several times $\mathbf{X} = \{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^N\}$, the (iid) probability depends on the *total observed counts* of each value:

$$p(\mathbf{X}|\pi) = \prod_n p(\mathbf{x}^n|\pi) = \exp\left\{\textstyle\sum_i \left(\sum_n x_i^n\right) \log \pi_i\right\} = \exp\left\{\textstyle\sum_i c_i \log \pi_i\right\}$$

Multinomial as Exponential Family

- The multinomial parameters are constrained: $\sum_i \pi_i = 1$.
  Define (the last) one in terms of the rest: $\pi_K = 1 - \sum_{i=1}^{K-1} \pi_i$

$$
\begin{aligned}
p(x|\pi) &= \exp\left(\sum_{i=1}^{K} x_i \log \pi_i\right) \\
&= \exp\left(\sum_{i=1}^{K-1} x_i \log \pi_i + \left(1 - \sum_{i=1}^{K-1} x_i\right) \log \pi_K\right) \\
&= \exp\left(\sum_{i=1}^{K-1} x_i \log \pi_i - \sum_{i=1}^{K-1} x_i \log \pi_K + \log \pi_K\right) \\
&= \exp\left(\sum_{i=1}^{K-1} x_i \log \frac{\pi_i}{\pi_K} + \log \pi_K\right)
\end{aligned}
$$

Multinomial as Exponential Family

$$
\begin{aligned}
p(x|\pi) &= \exp\left(\sum_{i=1}^{K-1} x_i \log\frac{\pi_i}{\pi_K} + \log\pi_K\right)\\
\eta_i &= \log\frac{\pi_i}{\pi_K}, \quad \eta_K = 0\\
T(x_i) &= x_i\\
h(\mathbf{x}) &= 1\\
A(\eta) &= -\log\left(1 - \sum_{i=1}^{K-1}\pi_i\right) = \log\left(\sum_{i=1}^{K} e^{\eta_i}\right)
\end{aligned}
$$

- The *softmax* function relates moment and natural (canonical) parameters:

$$\pi_i = \frac{e^{\eta_i}}{\sum_j e^{\eta_j}}$$

Gaussian (normal)

- For a continuous univariate random variable:

$$\begin{aligned} p(x|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \\ &= \frac{1}{\sqrt{2\pi}}\exp\left\{\frac{\mu x}{\sigma^2}-\frac{x^2}{2\sigma^2}-\frac{\mu^2}{2\sigma^2}-\log\sigma\right\} \end{aligned}$$

- Exponential family with:

$$\begin{aligned} \eta &= [\mu/\sigma^2\ ;\ -1/2\sigma^2] \\ T(x) &= [x\ ;\ x^2] \\ A(\eta) &= \log\sigma + \mu^2/2\sigma^2 \\ h(x) &= 1/\sqrt{2\pi} \end{aligned}$$

- Note: a univariate Gaussian is a two-parameter distribution with a two-component vector of sufficient statistics.

Multivariate Gaussian Distribution

- For a continuous vector random variable:

$$p(x|\mu,\Sigma) = |2\pi\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^\top\Sigma^{-1}(\mathbf{x}-\mu)\right\}$$

- Exponential family with:

$$\begin{aligned}\eta &= [\Sigma^{-1}\mu\,;\ -1/2\Sigma^{-1}]\\ T(x) &= [\mathbf{x}\,;\ \mathbf{x}\mathbf{x}^\top]\\ A(\eta) &= \log|\Sigma|/2 + \mu^\top\Sigma^{-1}\mu/2\\ h(x) &= (2\pi)^{-n/2}\end{aligned}$$

- Note: a d-dimensional Gaussian is a d+d^2-parameter distribution with a d+d^2-component vector of sufficient statistics (but because of symmetry and positivity, parameters are constrained)

Moments

- We can easily compute moments of any exponential family distribution by taking the derivatives of the log normalizer $A(\eta)$.
- The q^{th} derivative gives the q^{th} centred moment.

$$\begin{aligned} \frac{dA(\eta)}{d\eta} &= \text{mean} \\ \frac{d^2A(\eta)}{d\eta^2} &= \text{variance} \\ &\dots \end{aligned}$$

- When the sufficient statistic is a vector, partial derivatives need to be considered.

First moment

$$\begin{aligned}
\int p(\mathbf{x}|\eta)dx &= \int h(\mathbf{x})\exp\{\eta T(\mathbf{x}) - A(\eta)\}dx = 1 \\
Z(\eta) &= \int h(\mathbf{x})\exp\{\eta T(\mathbf{x})\}dx \\
A(\eta) &= \log Z(\eta) \\
\frac{dA}{d\eta} &= \frac{d}{d\eta}\log Z(\eta) = \frac{\frac{d}{d\eta}Z(\eta)}{Z(\eta)} \\
&= \frac{\int T(\mathbf{x})h(\mathbf{x})\exp\{\eta T(\mathbf{x})\}}{Z(\eta)} \\
&= ET(X)
\end{aligned}$$

Second moment

$$\begin{aligned}\frac{d^2A}{d\eta^2} &= \ldots \\ &= ET^2(X) - (ET(X))^2 \\ &= VarT(X)\end{aligned}$$

Example: 1D Gaussian

- Exponential family with:

$$\begin{aligned}\eta &= [\mu/\sigma^2\,;\ -1/2\sigma^2] \\ T(x) &= [x\,;\ x^2] \\ A(\eta) &= \log\sigma + \mu^2/2\sigma^2 = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-2\eta_2) \\ h(x) &= 1/\sqrt{2\pi}\end{aligned}$$

- First moment

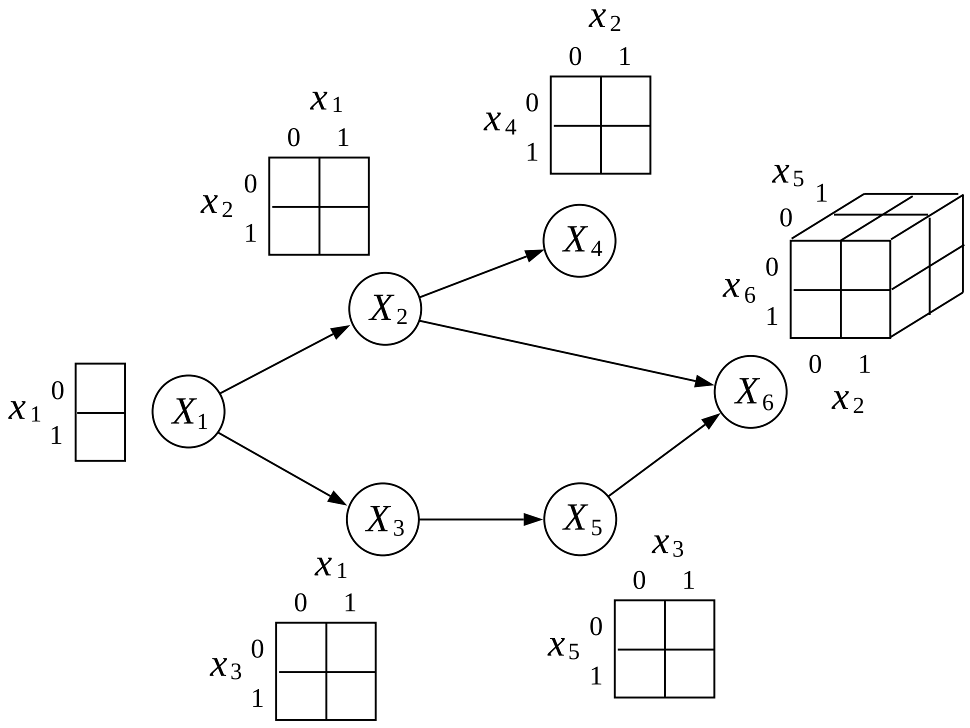
$$\frac{\partial A}{\partial \eta_1} = \frac{\eta_1}{2\eta_2} = \frac{\mu/\sigma^2}{1/\sigma^2} = \mu$$

- Second moment

$$\frac{\partial^2 A}{\partial \eta_1^2} = -\frac{1}{2\eta_2} = \sigma^2$$

Nodes with parents

- For discrete (categorical) variables, the most basic parametrization is the probability table which lists $p(x = k^{th} \text{ value})$.
- Since PTs must be nonnegative and sum to 1, for k-ary nodes there are $k-1$ free parameters.
- If a discrete node has discrete parent(s) we make one table for each setting of the parents: this is a *conditional probability table* or CPT.



Generalized Linear Models

- Consider the CPD for Y with parent X.
- A GLM is when $p(\mathbf{y}|\mathbf{x})$ is exponential family with conditional mean $\mu_i = f_i(\theta^\top \mathbf{x})$.
- The choice of exponential family member is dictated by the *type* of Y:
  - Class labels: Bernoulli or Multinomial
  - Counts: Poisson
  - Real valued: Gaussian

Canonical link

- We saw earlier that for an exponential family, $\mu = \frac{dA(\eta)}{d\eta}$.
- This mapping is invertible (since $\frac{d^2A(\eta)}{d\eta^2} = VarT(X) > 0$, so $A(\eta)$ is convex).
- Call this invertible mapping from moment parameters to canonical parameters $\eta = \psi(\mu)$.
- A GLM is when $p(\mathbf{y}|\mathbf{x})$ is exponential family with conditional mean $\mu_i = f_i(\theta^\top \mathbf{x})$.
- The function f is called the *response function*.
- If $f = \psi^{-1}$, then it is called the *canonical response function* or *canonical link*:
- Example: logistic function is canonical link for Bernoulli variables; softmax function is canonical link for multinomials

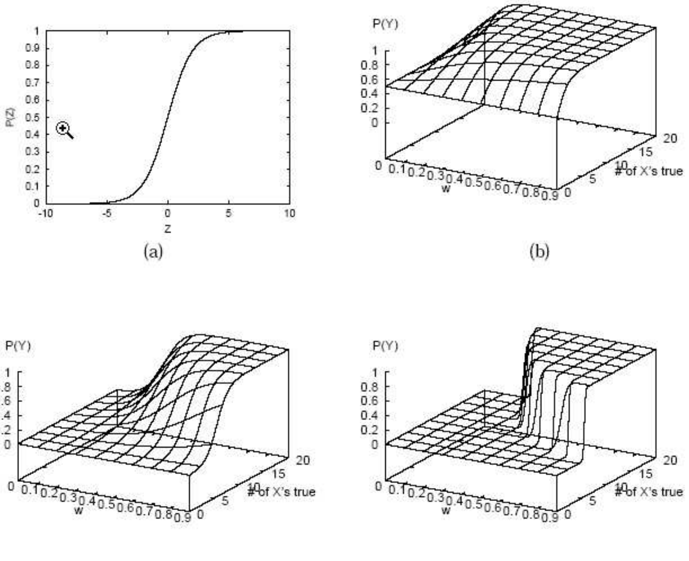
Canonical CPDs for $X \rightarrow Y$

X	Y	$p(Y \vert X)$
$\mathbb{R}^n$	$\mathbb{R}^m$	$\text{Gauss}(Y; WX + \mu, \Sigma)$
$\mathbb{R}^n$	$\{0, 1\}$	$\text{Bernoulli}(Y; p = \frac{1}{1+e^{-\theta^T x}})$
$\{0, 1\}^n$	$\{0, 1\}$	$\text{Bernoulli}(Y; p = \frac{1}{1+e^{-\theta^T x}})$
$\mathbb{R}^n$	$\{1, \ldots, K\}$	$\text{Multinomial}(Y; p_i = \text{softmax}(x, \theta))$

Sigmoid function

$$P(Y=1|X_1,\ldots,X_n)=\sigma(w_0+\sum_{i=1}^{n}w_iX_i)$$

$P(Y=1)$ vs number of X's that are on vs w



(a) (b)

(c) (d)

- a: 1D sigmoid
- b: $w_0 = 0$
- c: $w_0 = -5$
- d: w and w_0 are multiplied by 10

Log-odds

- We can interpret the parameters of a sigmoid in terms of how they affect the log-odds:

$$\frac{P(Y=1|X_{1:n})}{P(Y=0|X_{1:n})} = \frac{e^Z/(1+e^Z)}{1/(1+e^Z)} = e^Z$$

  where $Z = w_0 + \sum_i w_i X_i$.

- Consider the effect as X_j changes from 0 to 1:

$$\frac{P(Y=1|X_{1:n})}{P(Y=0|X_{1:n})} = \frac{\exp(w_0 + \sum_{i \neq j} w_i X_i + w_j)}{\exp(w_0 + \sum_{i \neq j} w_i X_i)} = e^{w_j}$$

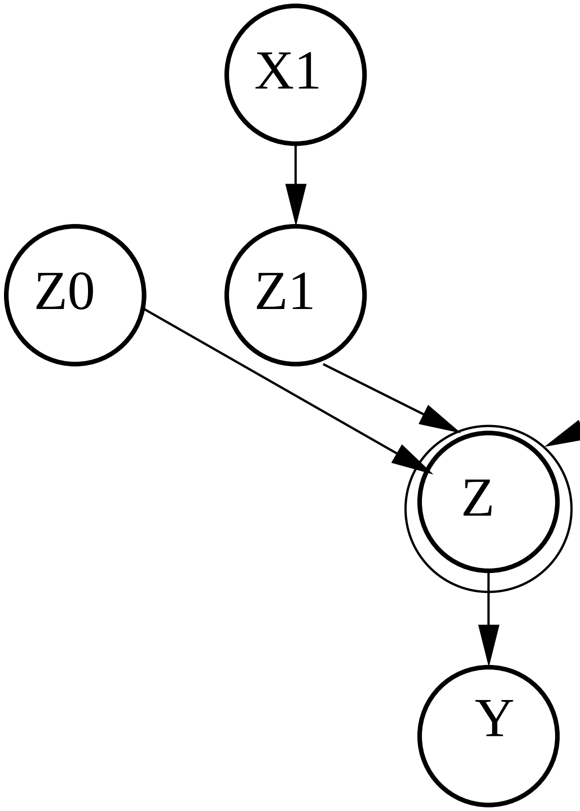
- If $w_j > 0$ then $e^{w_j} > 1$ so it increases the probability of $P(Y=1)$. Conversely if $w_j < 0$.
- If $w_j = 0$, then X_j is irrelevant (feature selection).

Other CPDs for $X \rightarrow Y$

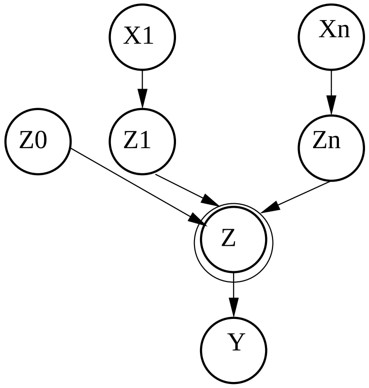
X	Y	$p(Y\|X)$
$\mathbb{R}^n$	$\mathbb{R}$	regression-box$(Y; X)$
$\mathbb{R}^n$	$\{1, \ldots, K\}$	classification-box$(Y; x)$
$\{1, \ldots, L\}$	$\mathbb{R}^n$	Gauss$(Y; \mu_X, \Sigma_X)$
$\{1, \ldots, L\}^n$	$\mathbb{R}$	regression-tree$(Y; X)$
$\{1, \ldots, L\}$	$\{1, \ldots, K\}$	$L \times K$ CPT
$\{1, \ldots, L\}^n$	$\{1, \ldots, K\}$	classification-tree$(Y; X)$
$\{0, 1\}^n$	$\{0, 1\}$	noisy-or

Independence of causal influence

- A CPD $P(Y|X_{1:n})$ exhibits ICI if it can be represented as a mini Bayes net as shown below, where Z is a deterministic function of the Z_i's.



Noisy-or



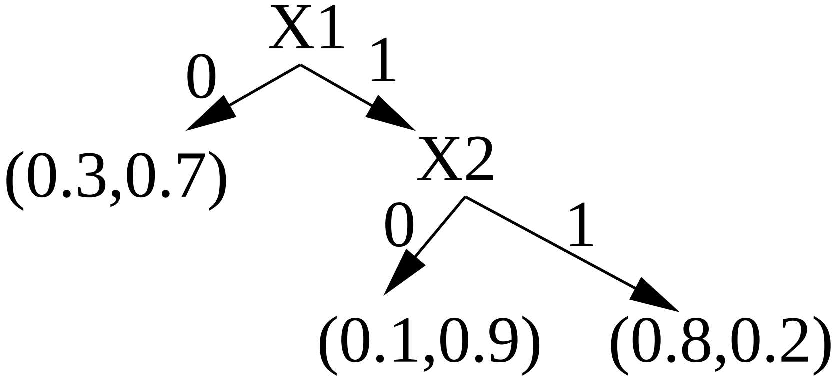
- $Y = Z$, Z is deterministic OR of Z_i's, but the link from X_i to Z_i flips 1's to 0's w.p. q_i. $Z_0 = 1$ is always on (leak node). Hence

$$P(Y = 0|X_{1:n}) = q_0 \prod_{i:X_i=1} q_i = q_0 \prod_i q_i^{X_i} = q_0 \sum_i e^{X_i \log q_i}$$

- Similar to sigmoid, but parameters are constrained $q_i \in [0, 1]$.
- Can be used to speed up inference.
- Cognitively plausible.

Context-specific independence

- CSI is when some links in the graph can be removed depending on the values of certain variables.
- eg. $P(Y|X_1, X_2)$ is represented as this decision tree:



- If $X_1 = 1$, then the link from $X_2 \rightarrow Y$ can be removed.
- This property arises in data association problems: let Z determine the identity of the observation; then $P(Y|Z = i, X_{1:n}) = f(Y, X_i)$.
- This property can be exploited in inference (condition on Z and the graph becomes sparser).