Exponential Families and Friends: Learning the Parameters of the a Fully Observed BN

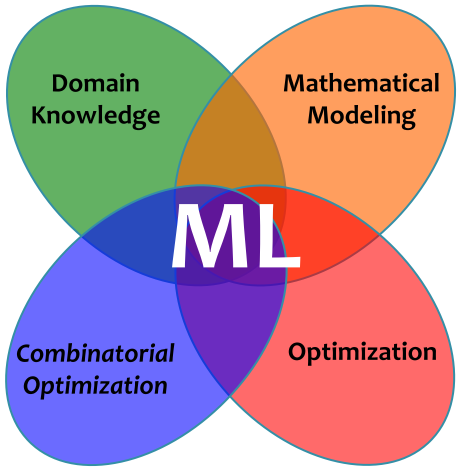
Kayhan Batmanghelich

Machine Learning

The **data** inspires the structures we want to predict

Our **model** defines a score for each structure

It also tells us what to optimize



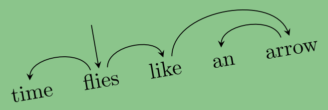
Inference finds {best structure, marginals, partition function} for a new observation

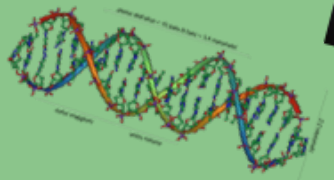
(**Inference** is usually called as a subroutine in learning)

Learning tunes the parameters of the model

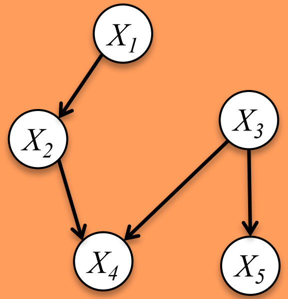
Machine Learning

Data

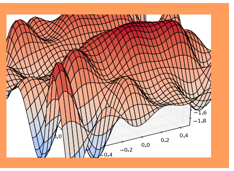




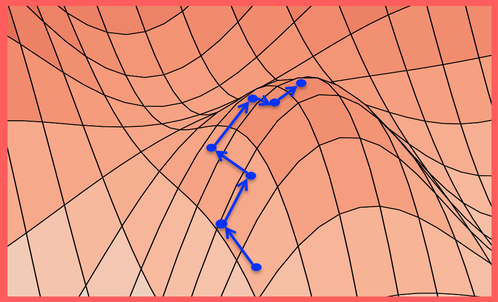
Model



Objective



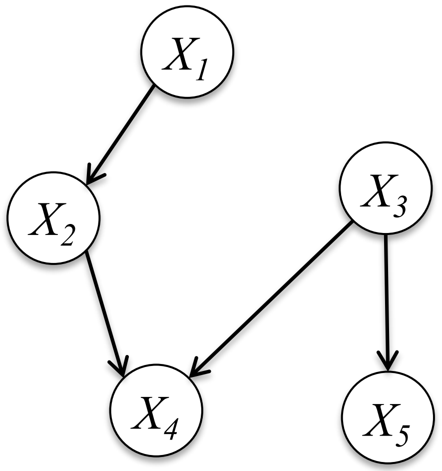
Learning



Inference

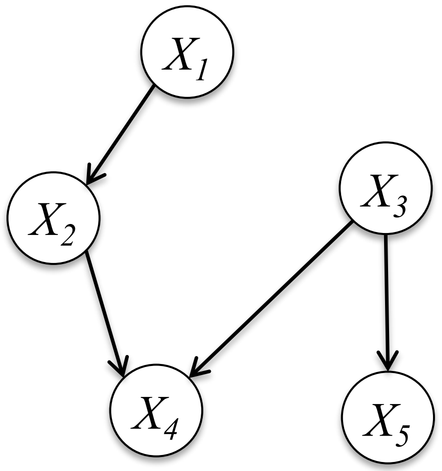
(**Inference** is usually called as a subroutine in learning)

Today's Lecture



$$p(X_1, X_2, X_3, X_4, X_5) =$$
$$p(X_5|X_3)p(X_4|X_2, X_3)$$
$$p(X_3)p(X_2|X_1)p(X_1)$$

Today's Lecture

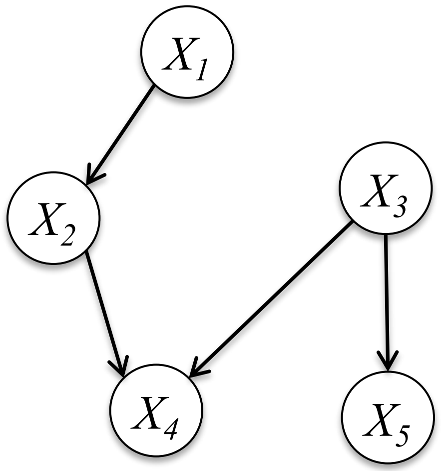


$$p(X_1, X_2, X_3, X_4, X_5) =$$

$$p(X_5|X_3)p(X_4|X_2, X_3)$$

$$p(X_3)p(X_2|X_1)p(X_1)$$

Today's Lecture



$$p(X_1, X_2, X_3, X_4, X_5) =$$
$$p(X_5|X_3)p(X_4|X_2, X_3)$$
$$p(X_3)p(X_2|X_1)p(X_1)$$

How do we define and learn these conditional and marginal distributions for a Bayes Net?

Today's Lecture

1. **Exponential Family Distributions**

   A candidate for marginal distributions, $p(X_i)$

2. **Generalized Linear Models**

   Convenient form for conditional distributions, $p(X_j \mid X_i)$

3. **Learning Fully Observed Bayes Nets**

   Easy thanks to decomposability

A candidate for marginal distributions, $p(X_i)$

1. EXPONENTIAL FAMILY

Why the Exponential Family?

1. **Pitman-Koopman-Darmois theorem:** it is the only family of distributions with **sufficient statistics that do not grow** with the size of the dataset
2. Only family of distributions for which **conjugate priors** exist (see Murphy textbook for a description)
3. It is the distribution that is closest to uniform (i.e. **maximizes entropy**) – subject to moment matching constraints
4. Key to **Generalized Linear Models** (next section)
5. Includes some of your favorite distributions!

Adapted from Murphy (2012) textbook

Whiteboard

- Definition of multivariate exponential family

Whiteboard

- Example 1: Categorical distribution

Whiteboard

- Example 2: Multivariate Gaussian distribution

Moments and the Partition Function

$$p(x;\theta) = \exp\left[x^T\theta - A(\theta)\right] h(x)$$

Moments and the Partition Function

$$p(x;\theta) = \exp\left[\theta^T T(x) - A(\theta)\right] h(x)$$

$A(\theta)$ convex

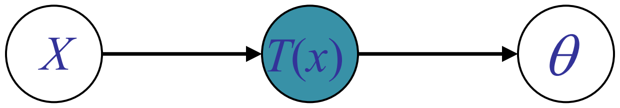
$$\nabla_\theta A(\theta) = \mathbb{E}[T(x)]$$

$T(x) \in \mathbb{R}^d$

$$\nabla^2_\theta A(\theta) = \underbrace{\mathbb{E}[T(x)T(x)^T] - \mathbb{E}[T(x)]\mathbb{E}[T(x)]^T}_{\mathrm{Cov}(T(x)) \in \mathbb{R}^{d \times d}} \succeq 0$$

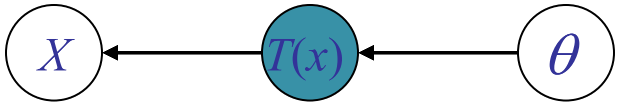
Sufficiency

- For $p(x;\theta)$, $T(x)$ is *sufficient* for θ if there is no information in X regarding θ beyond that in $T(x)$.
  - We can throw away X for the purpose of inference w.r.t. θ.
  - Bayesian view



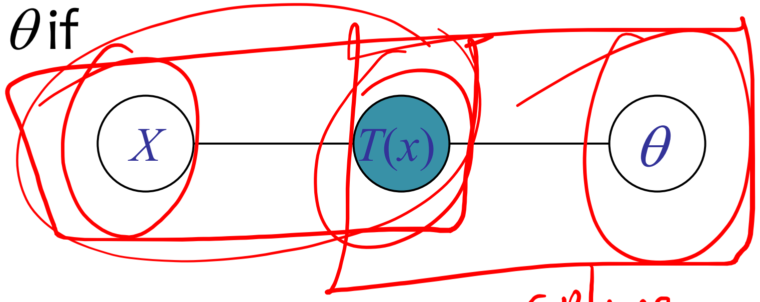
$$p(\theta \mid T(x), x) = p(\theta \mid T(x))$$

  - Frequentist view



$$p(x \mid T(x), \cancel{\theta}) = p(x \mid T(x))$$

  - The Neyman factorization theorem
    - $T(x)$ is *sufficient* for θ if



$$p(x, T(x), \theta) = \psi_1(T(x), \theta)\psi_2(x, T(x))$$

$$\Rightarrow p(x \mid \theta) = g(T(x), \theta)h(x, T(x))$$

Sufficiency

$$p(x;\theta) = \exp\left[\theta^T T(x) - A(\theta)\right] h(x)$$

$T(x) \in \mathbb{R}^d$

- Let's assume $\mathbf{x}_i \overset{iid}{\sim} p(x;\theta)$

$$p(\mathbf{x}_1, \cdots, \mathbf{x}_n; \theta) = \boxed{\left(\prod_{j=1}^{n} h(\mathbf{x}_j)\right)} \exp\left(\theta^T \boxed{\sum_j^n T(x_j)} - \boxed{nA(\theta)}\right)$$

$T \in \mathbb{R}^d$

$\max_{\eta} \; \eta^T T(x) - NA(\eta)$

$D = \{x_1, \ldots, x_n\}$

θ

MLE for Exponential Family

$\max_{\theta} P(D; \theta) = \max_{\theta} \prod_{i=1}^{n} P(x_i; \theta)$

- For *iid* data, the log-likelihood is

$T(x) \in \mathbb{R}^2$

$$\ell(\eta; D) = \log \prod_n h(x_n) \exp\{\eta^T T(x_n) - A(\eta)\}$$

$$= \cancel{\sum_n \log h(x_n)} + \left(\eta^T \boxed{\sum_n T(x_n)}\right) - NA(\eta)$$

param

$\arg\max$

$\arg\max_{\theta} \sum_{i=1}^{n} \log P(x_i; \theta)$

- Take derivatives and set to zero:

$$\frac{\partial \ell}{\partial \eta} = \sum_n T(x_n) - N \frac{\partial A(\eta)}{\partial \eta} = 0$$

$$\frac{\partial A(\eta)}{\partial \eta} = \frac{1}{N} \sum_n T(x_n)$$

$$\Rightarrow \quad \hat{\mu}_{MLE} = \frac{1}{N} \sum_n T(x_n)$$

- This amounts to **moment matching**.
- We can infer the canonical parameters using $\hat{\eta}_{MLE} = \psi(\hat{\mu}_{MLE})$

Examples

- Gaussian:

$$\eta = \left[\Sigma^{-1}\mu; -\tfrac{1}{2}\mathrm{vec}\left(\Sigma^{-1}\right)\right]$$
$$T(x) = \left[x; \mathrm{vec}\left(xx^T\right)\right]$$
$$A(\eta) = \tfrac{1}{2}\mu^T\Sigma^{-1}\mu + \tfrac{1}{2}\log|\Sigma|$$
$$h(x) = (2\pi)^{-k/2}$$

$$\Rightarrow \mu_{MLE} = \frac{1}{N}\sum_n T_1(x_n) = \frac{1}{N}\sum_n x_n$$

- Multinomial:

$$\eta = \left[\ln\left(\pi_k / \pi_K\right); 0\right]$$
$$T(x) = [x]$$
$$A(\eta) = -\ln\left(1 - \sum_{k=1}^{K-1}\pi_k\right) = \ln\left(\sum_{k=1}^{K} e^{\eta_k}\right)$$
$$h(x) = 1$$

$\eta =$

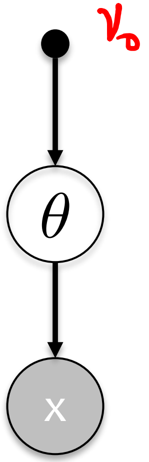
$$\Rightarrow \mu_{MLE} = \frac{1}{N}\sum_n x_n$$

- Poisson:

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

$$\Rightarrow \mu_{MLE} = \frac{1}{N}\sum_n x_n$$

Whiteboard



- Bayesian estimation of exponential family

$$p(x;\theta) = \exp\left[\theta^T T(x) - A(\theta)\right] h(x)$$

- We have observed iid samples and we are interested in

$$x_i \overset{i.i.d}{\sim} P(x;\theta)$$

$$P(\theta; \nu_0)$$

$$p(\theta | \underbrace{\{x_1, \cdots, x_n\}}_{\mathcal{D}})$$

likelihood

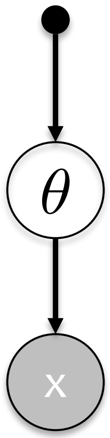
$$P(\theta | D) = \frac{\cancel{P(D;\theta)} \quad P(D|\theta)P(\theta)}{P(D)}$$

prior

data

$$\int_\theta P(D,\theta)\, d\theta$$

Posterior Mean Under Conjugate Prior



$$p(x;\theta) = \exp\left[\theta^T T(x) - A(\theta)\right] h(x)$$

$$p(\theta;\tau,n_0) = \exp\left(\tau^T\theta - n_0 A(\theta) - \tilde{A}(\tau,n_0)\right)$$

$$p(\theta|\mathcal{D}) = p(\theta;\tau + \sum_i T(x_i); n + n_0)$$

- Posterior mean of θ

MLE

param prior

$$\mathbb{E}[\theta|\mathcal{D}] = \frac{n}{n+n_0}\left(\frac{\sum_i T(x_i)}{n}\right) + \frac{n_0}{n_0+n}\left(\frac{\tau}{n_0}\right)$$

$E[\theta]$ $(\theta \sim p(\theta|D))$

$n \to \infty$

Convenient form for conditional distributions, $p(X_j \mid X_i)$

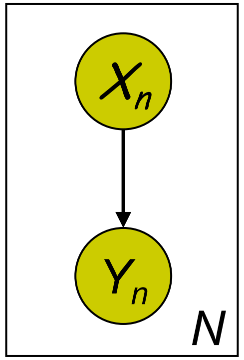
2. GENERALIZED LINEAR MODELS

Why Generalized Linear Models? (GLIMs)

1. Generalization of **linear regression, logistic regression, probit regression**, etc.
2. Provides a **framework for creating new conditional distributions** that come with some convenient properties
3. Special case: GLMs with canonical response functions are **easy to train** with MLE.
4. *No ~~Free~~ Lunch*: What about **Bayesian estimation of GLMs?** Unfortunately, we have to turn to approximation techniques since, in general, there isn't a closed form of the posterior.

Generalized Linear Models (GLMs)

- GLM
  - The observed input x is assumed to enter into the model via a linear combination of its elements $\xi = \theta^T x$
  - The conditional mean μ is represented as a function $f(\xi)$ of ξ, where f is known as the response function
  - The observed output y is assumed to be characterized by an exponential family distribution with conditional mean μ.

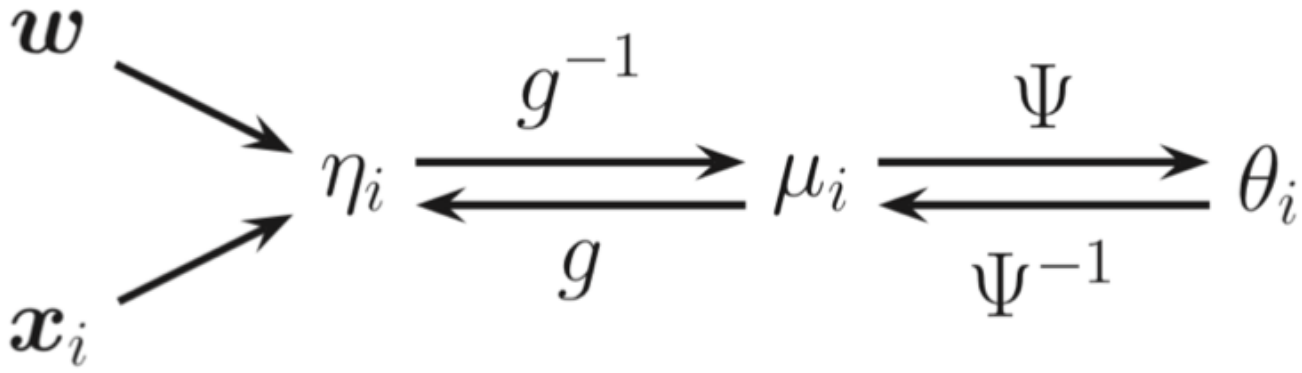


Whiteboard

- Constructive definition of GLMs
- Definition of GLMs with canonical response functions

Examples of the canonical response functions

Distrib.	Link $g(\mu)$	$\theta = \psi(\mu)$	$\mu = \psi^{-1}(\theta) = \mathbb{E}[y]$
$\mathcal{N}(\mu, \sigma^2)$	identity	$\theta = \mu$	$\mu = \theta$
$\mathrm{Bin}(N, \mu)$	logit	$\theta = \log(\frac{\mu}{1-\mu})$	$\mu = \mathrm{sigm}(\theta)$
$\mathrm{Poi}(\mu)$	log	$\theta = \log(\mu)$	$\mu = e^{\theta}$



Whiteboard

- MLE with GLM with Canonical response

MLE for GLMs with canonical response

$$\mathbf{X} = \begin{bmatrix} -- & \mathbf{x}_1 & -- \\ -- & \mathbf{x}_2 & -- \\ \vdots & \vdots & \vdots \\ -- & \mathbf{x}_n & -- \end{bmatrix} \quad \vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

- Log-likelihood

$$\mathcal{L}(w) = \sum_i \log h(y_i) + \sum_i \left(y_i w^T x_i - A(\eta_i)\right)$$

- Derivative of Log-likelihood

$$\begin{aligned}\nabla_w \mathcal{L}(w) &= \sum_i \left(x_i y_i - \frac{dA(\eta_i)}{d\eta_i} \frac{d\eta_i}{\theta} \right) \\ &= \sum_i (y_i - \mu_i) x_i \\ &= \mathbf{X}^T(\mathbf{y} - \boxed{\mu})\end{aligned}$$

This is a function of *w*

- Online learning for canonical GLMs
  - Stochastic gradient ascent = least mean squares (LMS) algorithm:

$$w^{t+1} = w^t + \boxed{\rho}(y_i - \mu_i^t) x_i$$

Step length

Batch learning for canonical GLMs

- The Hessian matrix

$$\mathbf{H} = -\frac{1}{\sigma^2}\sum_{i=1}^{N} \boxed{\frac{d\mu_i}{d\theta_i}} \mathbf{x}_i \mathbf{x}_i^T = -\frac{1}{\sigma^2}\mathbf{X}^T \boxed{\mathbf{S}} \mathbf{X}$$

$$\mathbf{X} = \begin{bmatrix} -- & \mathbf{x}_1 & -- \\ -- & \mathbf{x}_2 & -- \\ \vdots & \vdots & \vdots \\ -- & \mathbf{x}_n & -- \end{bmatrix}$$

$$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Involves the second derivative of $A(\theta)$

$$\mathbf{S} = \mathrm{diag}\left(\frac{d\mu_1}{d\theta_1}, \ldots, \frac{d\mu_N}{d\theta_N}\right)$$

Iteratively Reweighted Least Squares (IRLS)

$$\nabla_w \mathcal{L}(w) = \mathbf{X}^T(\mathbf{y} - \mu)$$

$$\mathbf{H} = -\frac{1}{\sigma^2}\sum_{i=1}^{N}\frac{d\mu_i}{d\theta_i}\mathbf{x}_i\mathbf{x}_i^T = -\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{S}\mathbf{X}$$

Distrib.	Link $g(\mu)$	$\theta = \psi(\mu)$	$\mu = \psi^{-1}(\theta) = \mathbb{E}[y]$
$\mathcal{N}(\mu, \sigma^2)$	identity	$\theta = \mu$	$\mu = \theta$
$\text{Bin}(N, \mu)$	logit	$\theta = \log(\frac{\mu}{1-\mu})$	$\mu = \text{sigm}(\theta)$
$\text{Poi}(\mu)$	log	$\theta = \log(\mu)$	$\mu = e^{\theta}$

- Recall Newton-Raphson methods with cost function

$$w^{t+1} = w^t + H^{-1}(w^t)\nabla\mathcal{L}(w^t)$$

$$= \left(\mathbf{X}^T S(w^t)\mathbf{X}\right)^{-1}\left[\mathbf{X}^T \boxed{S(w^t)}\mathbf{X}w^t + \mathbf{X}^T(\mathbf{y} - \mu)\right]$$

$$= \left(\mathbf{X}^T S(w^t)\mathbf{X}\right)^{-1}\mathbf{X}^T S(w^t)\boxed{\mathbf{z}^t} \qquad \mathbf{z}^t = \mathbf{X}w^t + S(w^t)^{-1}(\mathbf{y} - \mu^t)$$

$S(w^t)$ $S(w^t)$

$S = \left[\ \frac{d\mu}{d\eta}\ \right.$

Iterative Reweighted Least Squares (IRLS)

$$\nabla_w \mathcal{L}(w) = \mathbf{X}^T(\mathbf{y} - \mu)$$

$$\mathbf{H} = -\frac{1}{\sigma^2}\sum_{i=1}^{N} \frac{d\mu_i}{d\theta_i}\mathbf{x}_i\mathbf{x}_i^T = -\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{S}\mathbf{X}$$

Distrib.	Link $g(\mu)$	$\theta = \psi(\mu)$	$\mu = \psi^{-1}(\theta) = \mathbb{E}[y]$
$\mathcal{N}(\mu, \sigma^2)$	identity	$\theta = \mu$	$\mu = \theta$
$\text{Bin}(N, \mu)$	logit	$\theta = \log(\frac{\mu}{1-\mu})$	$\mu = \text{sigm}(\theta)$
$\text{Poi}(\mu)$	log	$\theta = \log(\mu)$	$\mu = e^{\theta}$

- Recall Newton-Raphson methods with cost function

$$\begin{aligned} w^{t+1} &= w^t + H^{-1}(w^t)\nabla\mathcal{L}(w^t) \\ &= \left(\mathbf{X}^T S(w^t)\mathbf{X}\right)^{-1}\left[\mathbf{X}^T S(w^t)\mathbf{X}w^t + \mathbf{X}^T(\mathbf{y} - \mu)\right] \\ &= \boxed{\left(\mathbf{X}^T S(w^t)\mathbf{X}\right)^{-1}\mathbf{X}^T S(w^t)\mathbf{z}^t} \end{aligned}$$

$$\mathbf{z}^t = \mathbf{X}w^t + S(w^t)^{-1}(\mathbf{y} - \mu^t)$$

It looks like $(X^TX)^{-1}X^Ty$

Iteratively Reweighted Least Squares (IRLS)

$$\nabla_w \mathcal{L}(w) = \mathbf{X}^T(\mathbf{y} - \mu)$$

$$\mathbf{H} = -\frac{1}{\sigma^2}\sum_{i=1}^{N} \frac{d\mu_i}{d\theta_i}\mathbf{x}_i\mathbf{x}_i^T = -\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{S}\mathbf{X}$$

Distrib.	Link $g(\mu)$	$\theta = \psi(\mu)$	$\mu = \psi^{-1}(\theta) = \mathbb{E}[y]$
$\mathcal{N}(\mu, \sigma^2)$	identity	$\theta = \mu$	$\mu = \theta$
$\text{Bin}(N, \mu)$	logit	$\theta = \log(\frac{\mu}{1-\mu})$	$\mu = \text{sigm}(\theta)$
$\text{Poi}(\mu)$	log	$\theta = \log(\mu)$	$\mu = e^\theta$

- Recall Newton-Raphson methods with cost function

$$w^{t+1} = w^t + H^{-1}(w^t)\nabla\mathcal{L}(w^t)$$

$$= \left(\mathbf{X}^T S(w^t)\mathbf{X}\right)^{-1}\left[\mathbf{X}^T S(w^t)\mathbf{X}w^t + \mathbf{X}^T(\mathbf{y} - \mu)\right]$$

$$= \left(\mathbf{X}^T S(w^t)\mathbf{X}\right)^{-1}\mathbf{X}^T S(w^t)\mathbf{z}^t \qquad \mathbf{z}^t = \mathbf{X}w^t + S(w^t)^{-1}(\mathbf{y} - \mu^t)$$

- This can be understood as solving the following " Iteratively reweighted least squares " problem

$$w^{t+1} = \arg\max_w (z^t - \mathbf{X}w)^T S(w^t)(z^t - \mathbf{X}w)$$

Examples

$$\nabla_w \mathcal{L}(w) = \mathbf{X}^T(\mathbf{y} - \mu)$$

$$\mathbf{H} = -\frac{1}{\sigma^2}\sum_{i=1}^{N}\frac{d\mu_i}{d\theta_i}\mathbf{x}_i\mathbf{x}_i^T = -\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{S}\mathbf{X}$$

Distrib.	Link $g(\mu)$	$\theta = \psi(\mu)$	$\mu = \psi^{-1}(\theta) = \mathbb{E}[y]$
$\mathcal{N}(\mu, \sigma^2)$	identity	$\theta = \mu$	$\mu = \theta$
$\mathrm{Bin}(N, \mu)$	logit	$\theta = \log(\frac{\mu}{1-\mu})$	$\mu = \mathrm{sigm}(\theta)$
$\mathrm{Poi}(\mu)$	log	$\theta = \log(\mu)$	$\mu = e^\theta$

- Recall Newton-Raphson methods with cost function

$$w^{t+1} = w^t + H^{-1}(w^t)\nabla\mathcal{L}(w^t)$$

$$= \left(\mathbf{X}^T S(w^t)\mathbf{X}\right)^{-1}\left[\mathbf{X}^T S(w^t)\mathbf{X}w^t + \mathbf{X}^T(\mathbf{y} - \mu)\right]$$

$$= \left(\mathbf{X}^T S(w^t)\mathbf{X}\right)^{-1}\mathbf{X}^T S(w^t)\mathbf{z}^t$$

$$\mathbf{z}^t = \mathbf{X}w^t + S(w^t)^{-1}(\mathbf{y} - \mu^t)$$

$$w^{t+1} = \arg\max_w (z^t - \mathbf{X}w)^T S(w^t)(z^t - \mathbf{X}w)$$

Practical Issues

- It is very common to use *regularized* maximum likelihood.

$$p(y=\pm 1|x,\theta)=\frac{1}{1+e^{-y\theta^T x}}=\sigma(y\theta^T x)$$

$$p(\theta)\sim \mathrm{Normal}(0,\lambda^{-1}I)$$

$$l(\theta)=\sum_n \log\left(\sigma(y_n\theta^T x_n)\right)-\frac{\lambda}{2}\theta^T\theta$$

  - IRLS takes $O(Nd^3)$ per iteration, where N= number of training cases and d= dimension of input x.
  - Quasi-Newton methods, that approximate the Hessian, work faster.
  - Conjugate gradient takes $O(Nd)$ per iteration, and usually works best in practice.
  - Stochastic gradient descent can also be used if N is large c.f. perceptron rule.

Today's Lecture

1. **Exponential Family Distributions**

   A candidate for marginal distributions, $p(X_i)$

2. **Generalized Linear Models**

   Convenient form for conditional distributions, $p(X_j \mid X_i)$

3. **Learning Fully Observed Bayes Nets**

   Easy thanks to decomposability

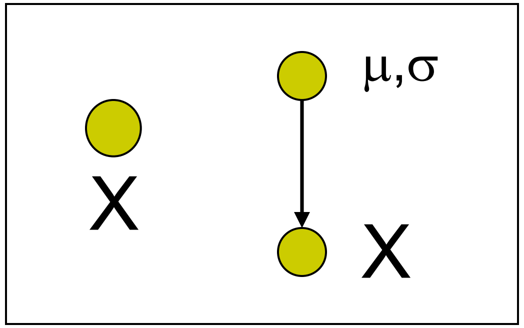
Easy thanks to decomposability

3. LEARNING FULLY OBSERVED BNS

Simple GMs are the building blocks of complex BNs

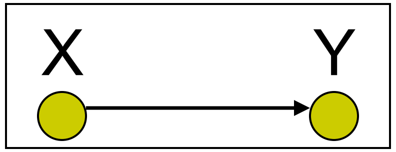
Density estimation

Parametric and nonparametric methods



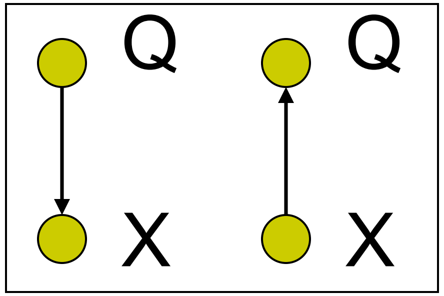
Regression

Linear, conditional mixture, nonparametric



Classification

Generative and discriminative approach



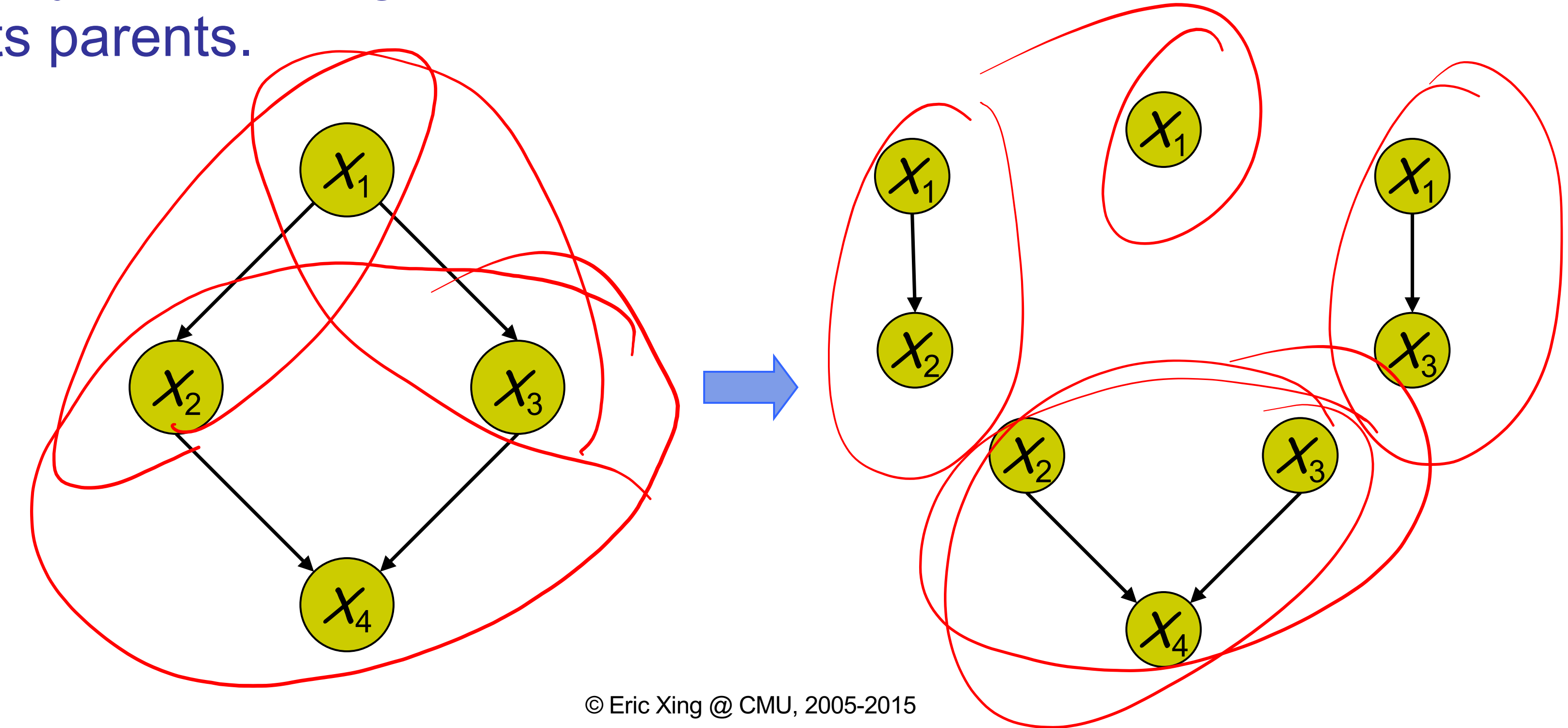
Decomposable likelihood of a BN

Recall from Lecture 2…

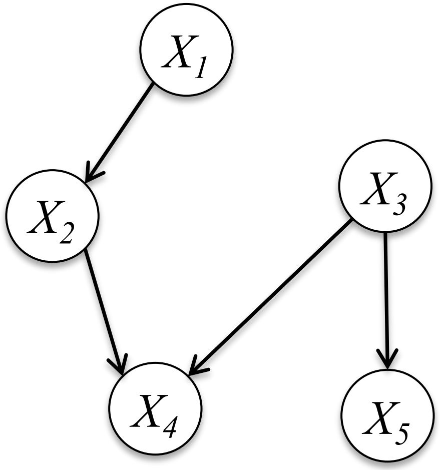
- Consider the distribution defined by the directed acyclic GM:

$$p(x \mid \theta) = p(x_1 \mid \theta_1) p(x_2 \mid x_1, \theta_2) p(x_3 \mid x_1, \theta_3) p(x_4 \mid x_2, x_3, \theta_4)$$

- This is exactly like learning four separate small BNs, each of which consists of a node and its parents.



Learning Fully Observed BNs



$$
\begin{aligned}
\boldsymbol{\theta}^* &= \operatorname*{argmax}_{\boldsymbol{\theta}} \log p(X_1, X_2, X_3, X_4, X_5) \\
&= \operatorname*{argmax}_{\boldsymbol{\theta}} \log p(X_5 | X_3, \theta_5) + \log p(X_4 | X_2, X_3, \theta_4) \\
&\qquad + \log p(X_3 | \theta_3) + \log p(X_2 | X_1, \theta_2) \\
&\qquad + \log p(X_1 | \theta_1)
\end{aligned}
$$

$$\theta_1^* = \operatorname*{argmax}_{\theta_1} \log p(X_1 | \theta_1)$$

$$\theta_2^* = \operatorname*{argmax}_{\theta_2} \log p(X_2 | X_1, \theta_2)$$

$$\theta_3^* = \operatorname*{argmax}_{\theta_3} \log p(X_3 | \theta_3)$$

$$\theta_4^* = \operatorname*{argmax}_{\theta_4} \log p(X_4 | X_2, X_3, \theta_4)$$

$$\theta_5^* = \operatorname*{argmax}_{\theta_5} \log p(X_5 | X_3, \theta_5)$$

Summary

1. **Exponential Family Distributions**
   - A candidate for marginal distributions, $p(X_i)$
   - Examples: Multinomial, Dirichlet, Gaussian, Gamma, Poisson
   - MLE has closed form solution
   - Bayesian estimation easy with conjugate priors
   - Sufficient statistics by inspection
2. **Generalized Linear Models**
   - Convenient form for conditional distributions, $p(X_j \mid X_i)$
   - Special case: GLIMs with canonical response
     - Output $\boldsymbol{y}$ follows an exponential family
     - Input $\boldsymbol{x}$ introduced via a linear combination
   - MLE for GLIMs with canonical response by SGD
   - In general, Bayesian estimation relies on approximations
3. **Learning Fully Observed Bayes Nets**
   - Easy thanks to decomposability