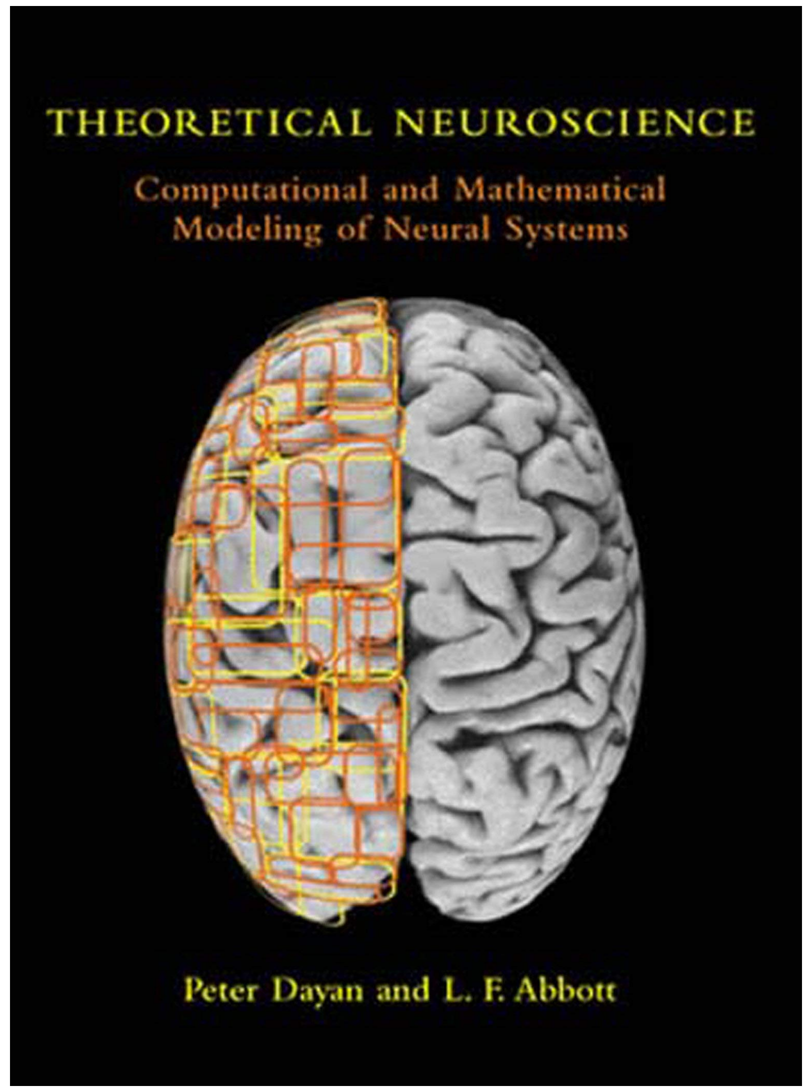
# Graphical Models

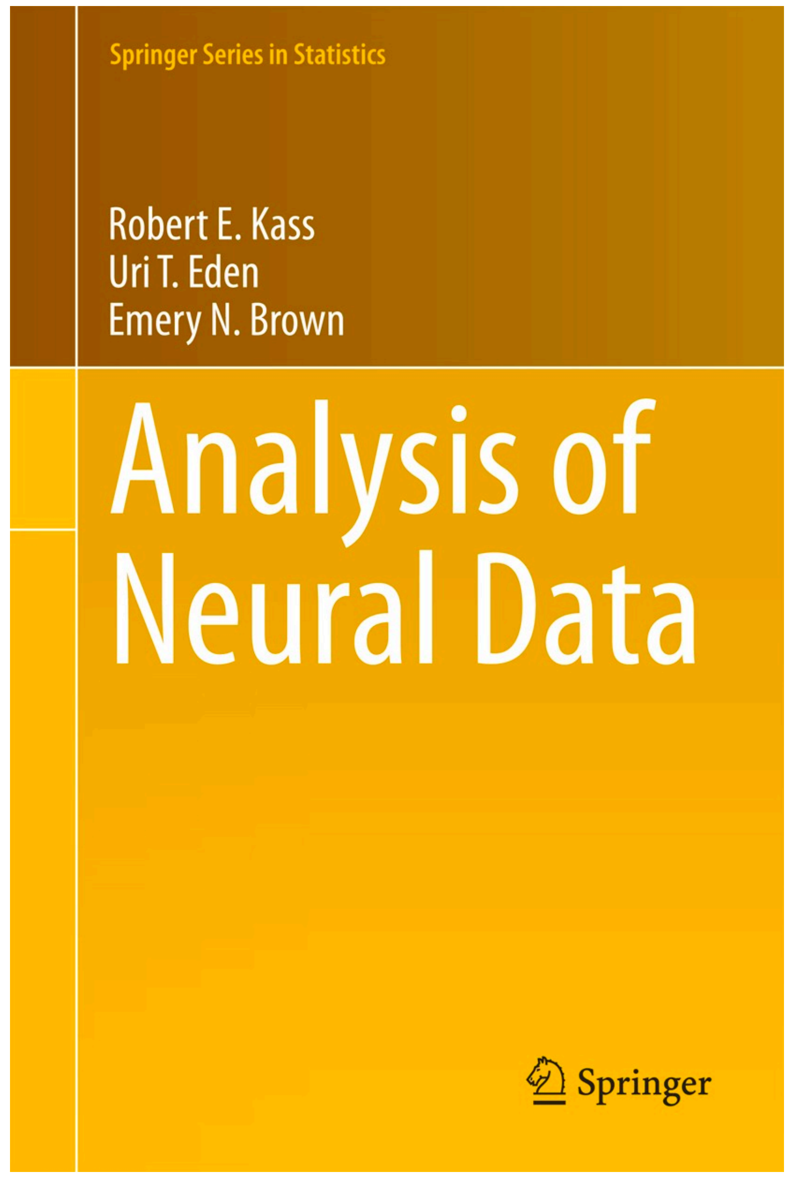
Yuanning Li

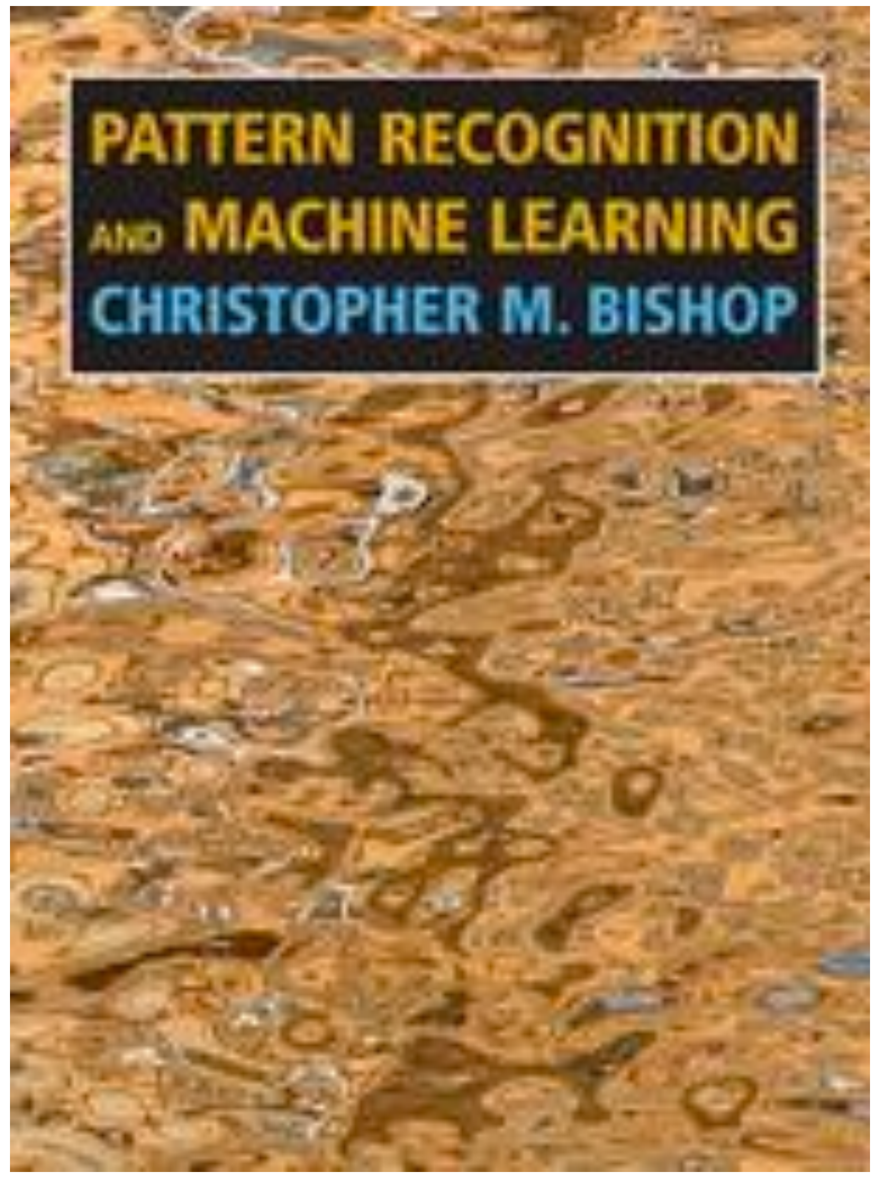
BME 2111
Neural Signal Processing and Data Analysis
2023 Fall

# Roadmap

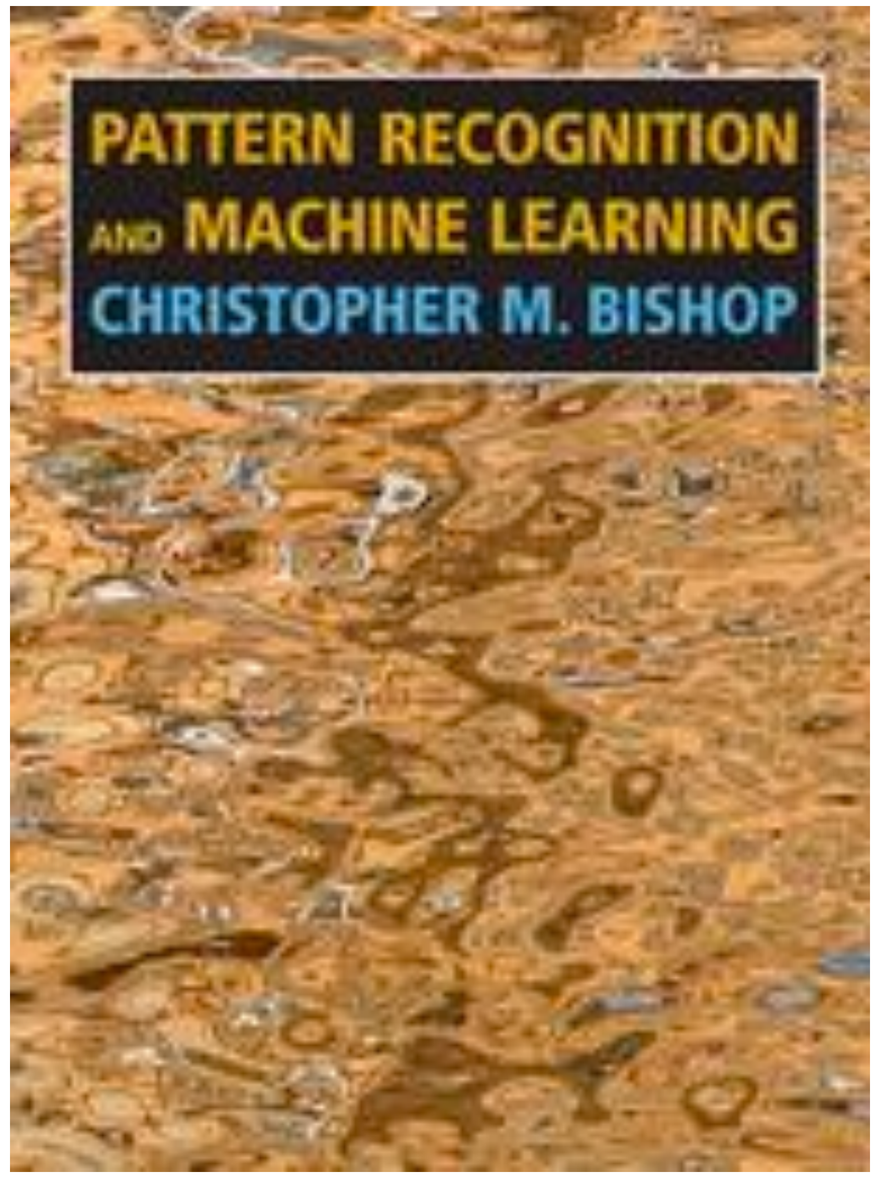
- Traditional neural signal processing methods
  - *Theoretical Neuroscience*, Chapter 1
- State-of-the-art neural signal processing methods
  - *Patter Recognition and Machine Learning*
  - *Analysis of Neural Data*



# Topics we will cover in PRML




*Chap. 4: Classification. Naive Bayes.*
*Neuroscience application: discrete neural decoding*

*Chap. 8: Graphical models.*

*Chap. 9: Mixture models. Expectation-maximization.*
*Neuroscience application: spike sorting*

*Chap. 12: Principal components analysis. Factor analysis.*
*Neuroscience applications: spike sorting, dimensionality reduction*

*Chap. 13: Kalman filter.*
*Neuroscience application: continuous neural decoding*

# Probabilistic graphical models

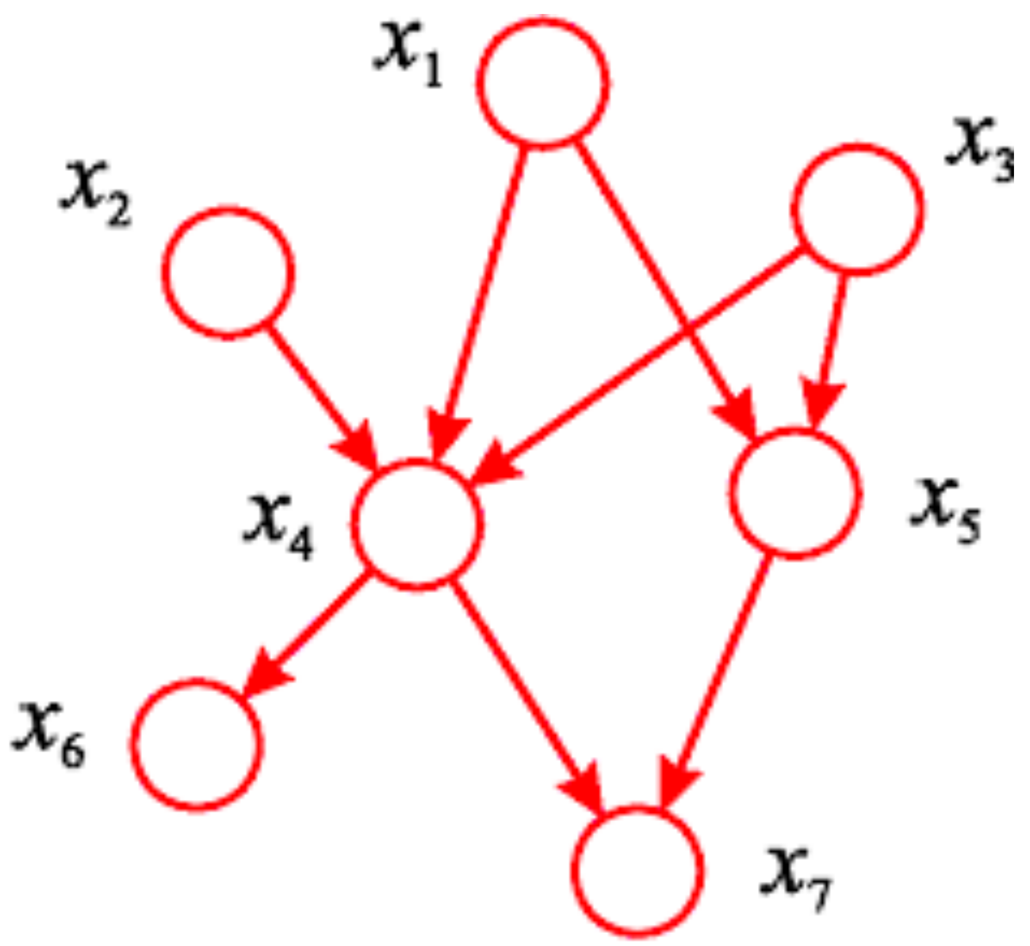
- Motivating question:
  - In statistical machine learning, we are often dealing with multivariate likelihood $P(x_1, x_2, \ldots, x_n)$ that describe distribution over a set of random variables $\{x_1, \ldots, x_n\}$
  - *Recall*:
    - Last time in classification, we maximize the likelihood of the observed data w.r.t. model parameters.

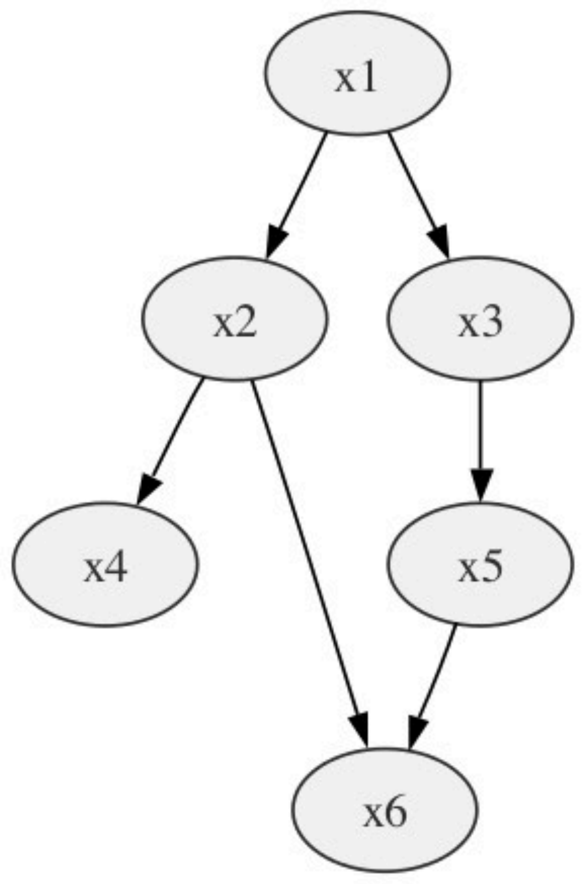
$$\arg\max_{\Theta} P_{\Theta}(\{\mathbf{x_1}, C_1\}, \{\mathbf{x_2}, C_2\}, \ldots, \{\mathbf{x_N}, C_N\})$$

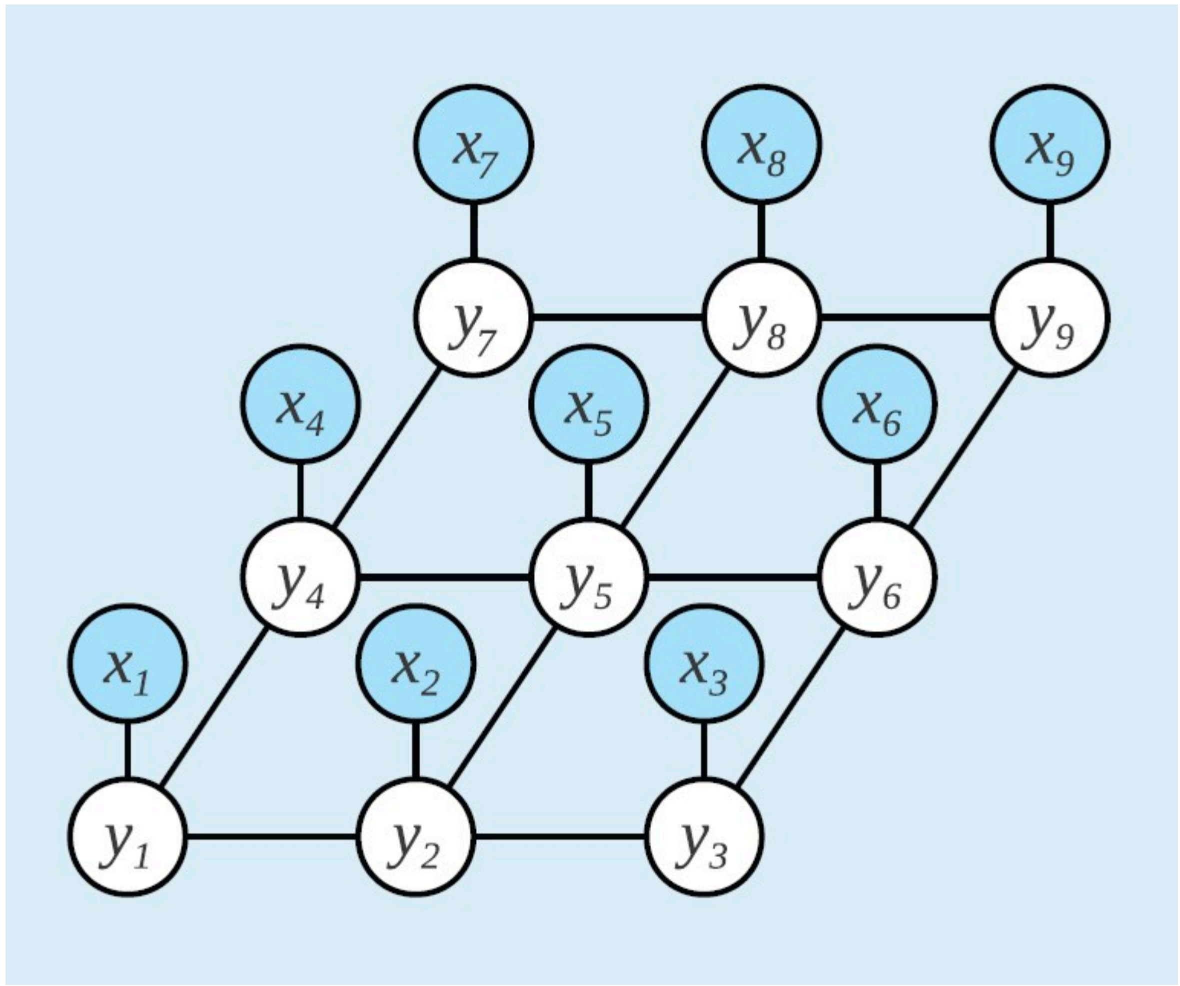
    This can be further decomposed into multiplication of PDFs

# Probabilistic graphical models

- Motivating question:
    - In statistical machine learning, we are often dealing with multivariate likelihood $P(x_1, x_2, \ldots, x_n)$ that describe distribution over a set of random variables $\{x_1, \ldots, x_n\}$
    - In modern statistics, we use probabilistic graphical models as a way to describe statistical models.



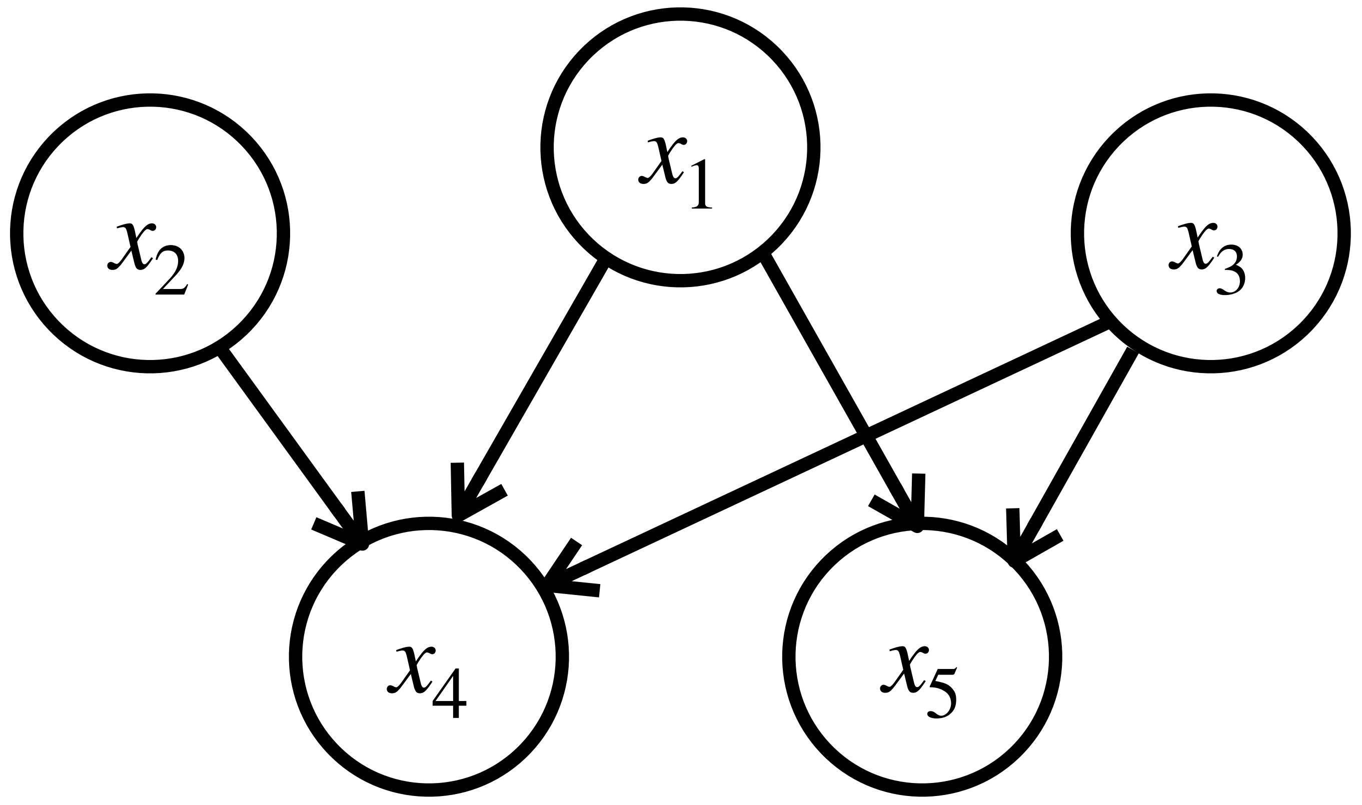



# Probabilistic graphical models

- What are they?
  - Diagrammatic representations of probability distributions.
- Why do we use them?
  - They provide a simple way to visualize the structure of a probabilistic model.
  - Properties of the model, such as conditional independence, can be obtained by inspection of the graph.
- Components of a graphical model
  - Each <u>node</u> represents a random variable
  - Each <u>link</u> represents a probabilistic relationship between variables

# Directed Graphical Models

- Also known as Bayesian Networks
- Example:




$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2)P(x_3)P(x_4 \,|\, x_1, x_2, x_3)P(x_5 \,|\, x_1, x_3)$$

- Relationship between directed graph and joint probability distribution:

$$P(x_1, \ldots, x_K) = \prod_{k=1}^{K} P(x_k \,|\, \{\text{parents of } x_k\})$$

# Fully connected graphs

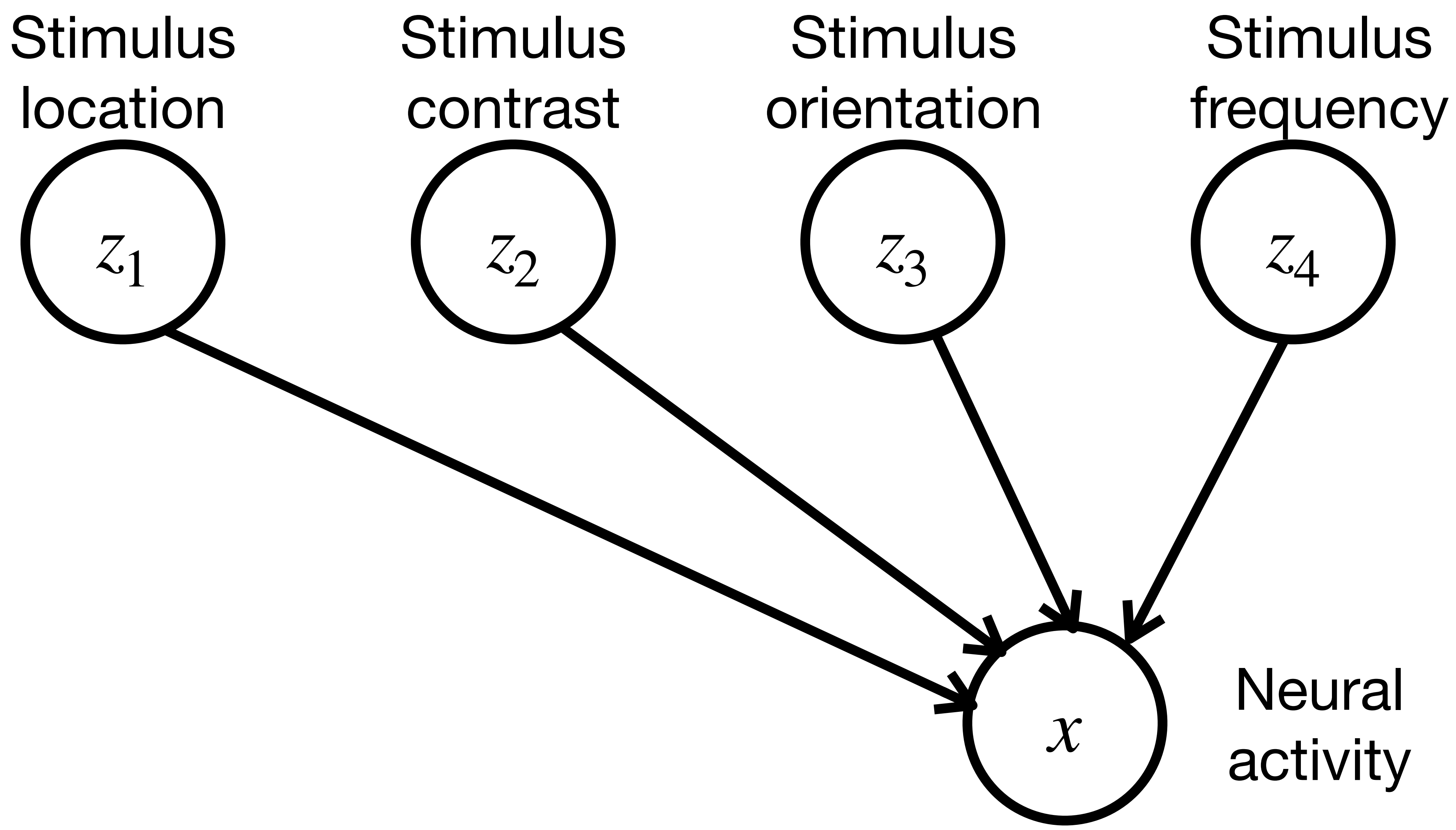
- For any joint distribution $P(x_1, \ldots, x_K)$, we can write:

$$P(x_1, \ldots, x_K) = P(x_1)P(x_2 | x_1)P(x_3 | x_1, x_2) \cdots P(x_K | x_1, \ldots, x_{K-1})$$

- This corresponds to a fully connected graph.
- It is the absence of links that conveys interesting properties of probability distributions.

# How are graphical models used in neuroscience?

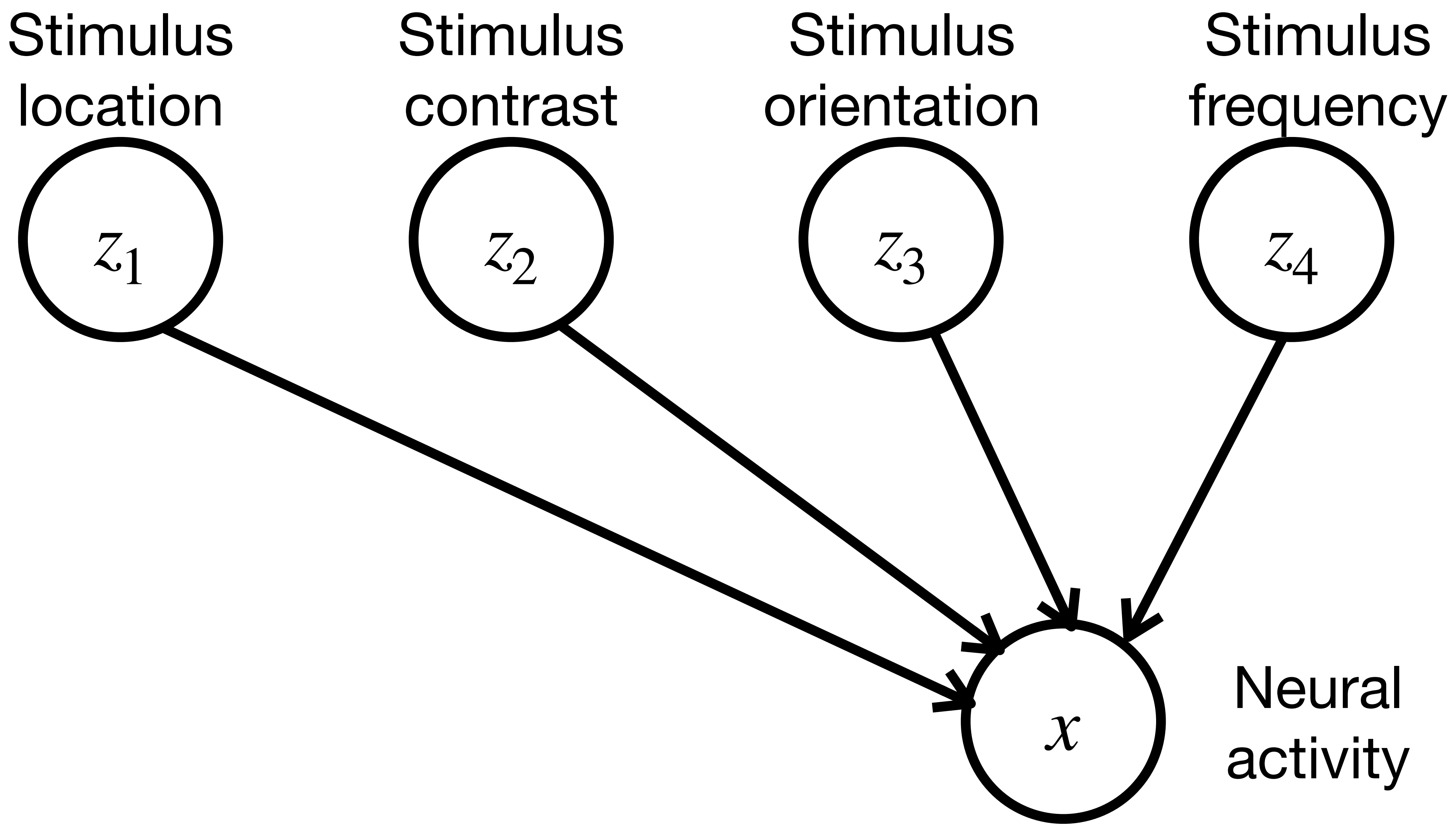
- We record neural activity $x$ and want to explain the activity in terms of variables $z_1, \ldots, z_M$
- Example:




- Based on this graph, how can we factor the joint distribution $P(z_1, z_2, z_3, z_4, x)$?

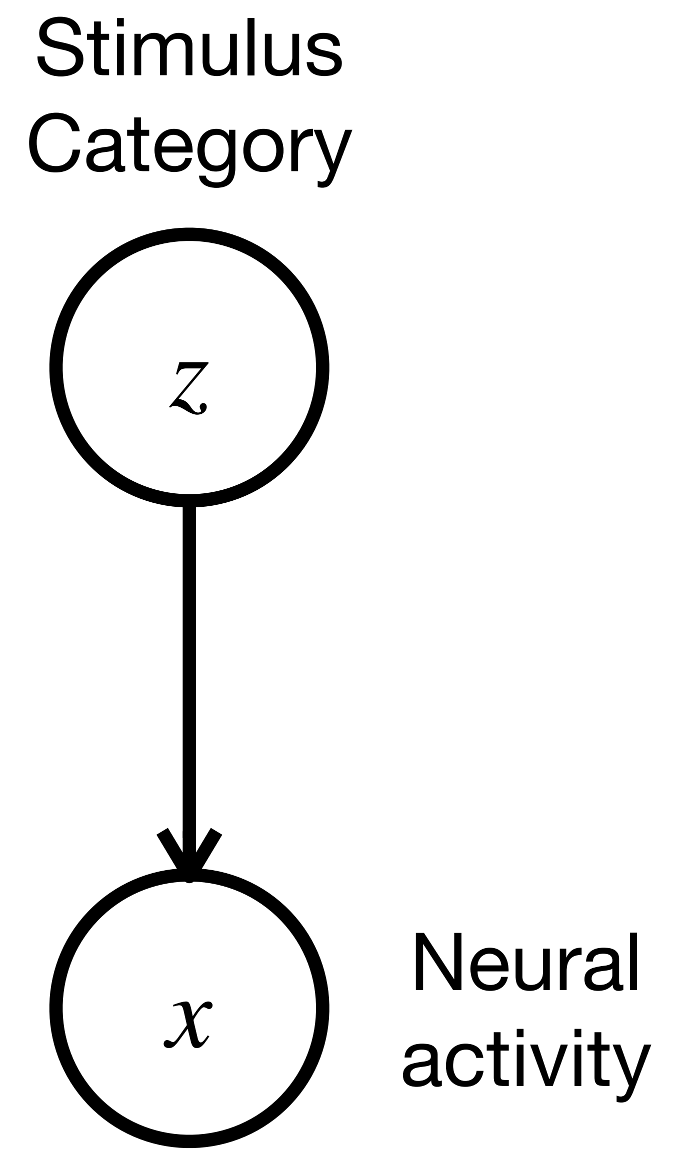
# Generative models

- Graphical model provides a picture of the causal process by which the data arose.
- Graphical model provides an intuitive way of generating synthetic data from joint distribution.
- Example:
  - Assume a generalized linear model
  - $\mu = w_1 z_1 + w_2 z_2 + w_3 z_3 + w_4 z_4$
  - $x \sim \mathcal{N}(\mu, \sigma^2)$



# Generative models

- Graphical model provides a picture of the <u>causal process</u> by which the data arose.
- Graphical model provides an intuitive way of generating synthetic data from joint distribution.
- Example:
  - Probabilistic generative model for classification
  - $\mathbf{x} \mid z \sim \mathcal{N}(\mu_z, \Sigma_z^2)$

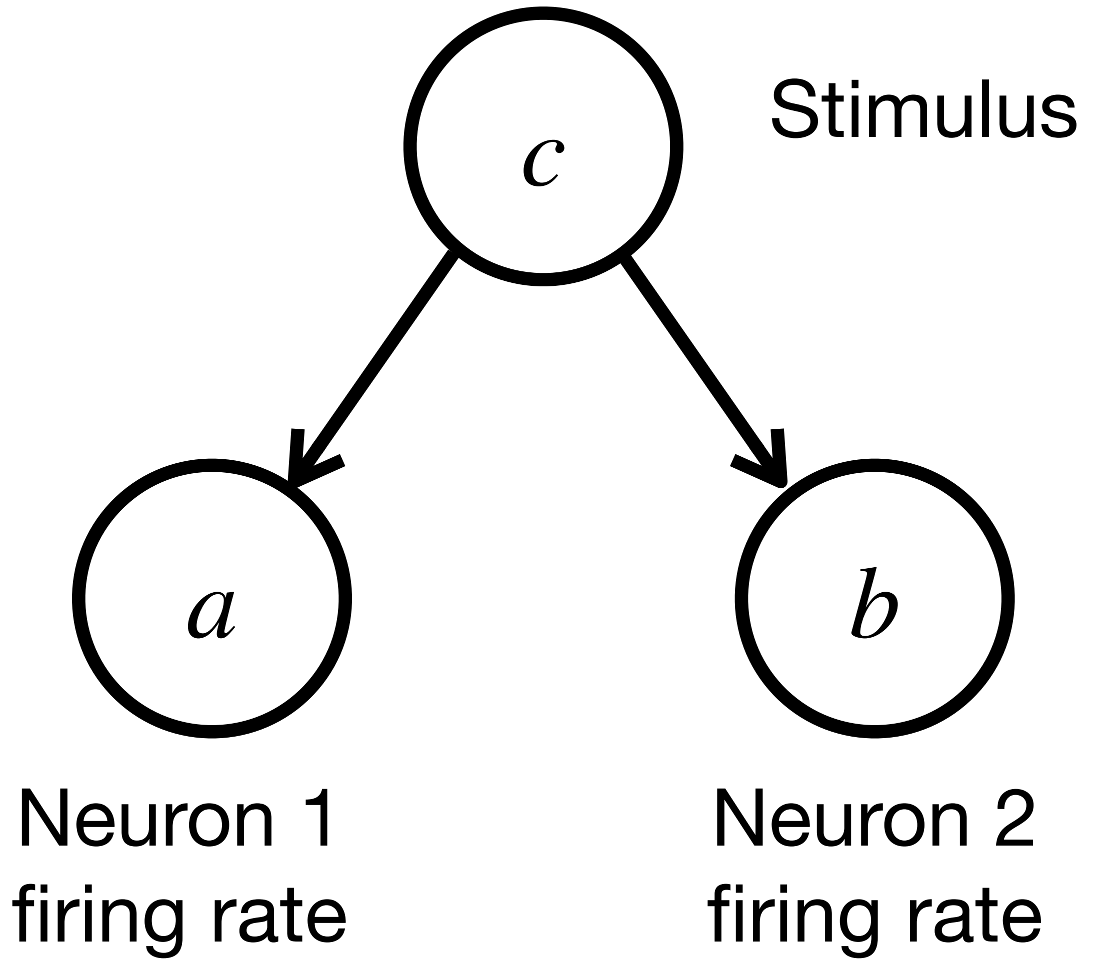


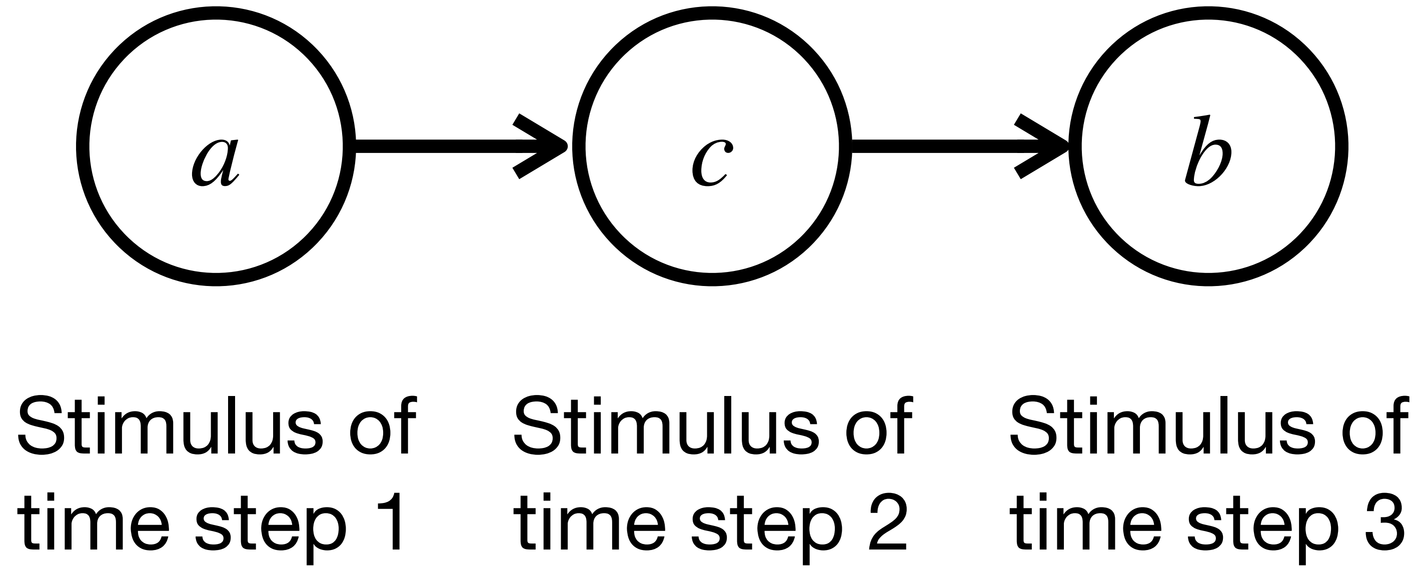
# Conditional independence

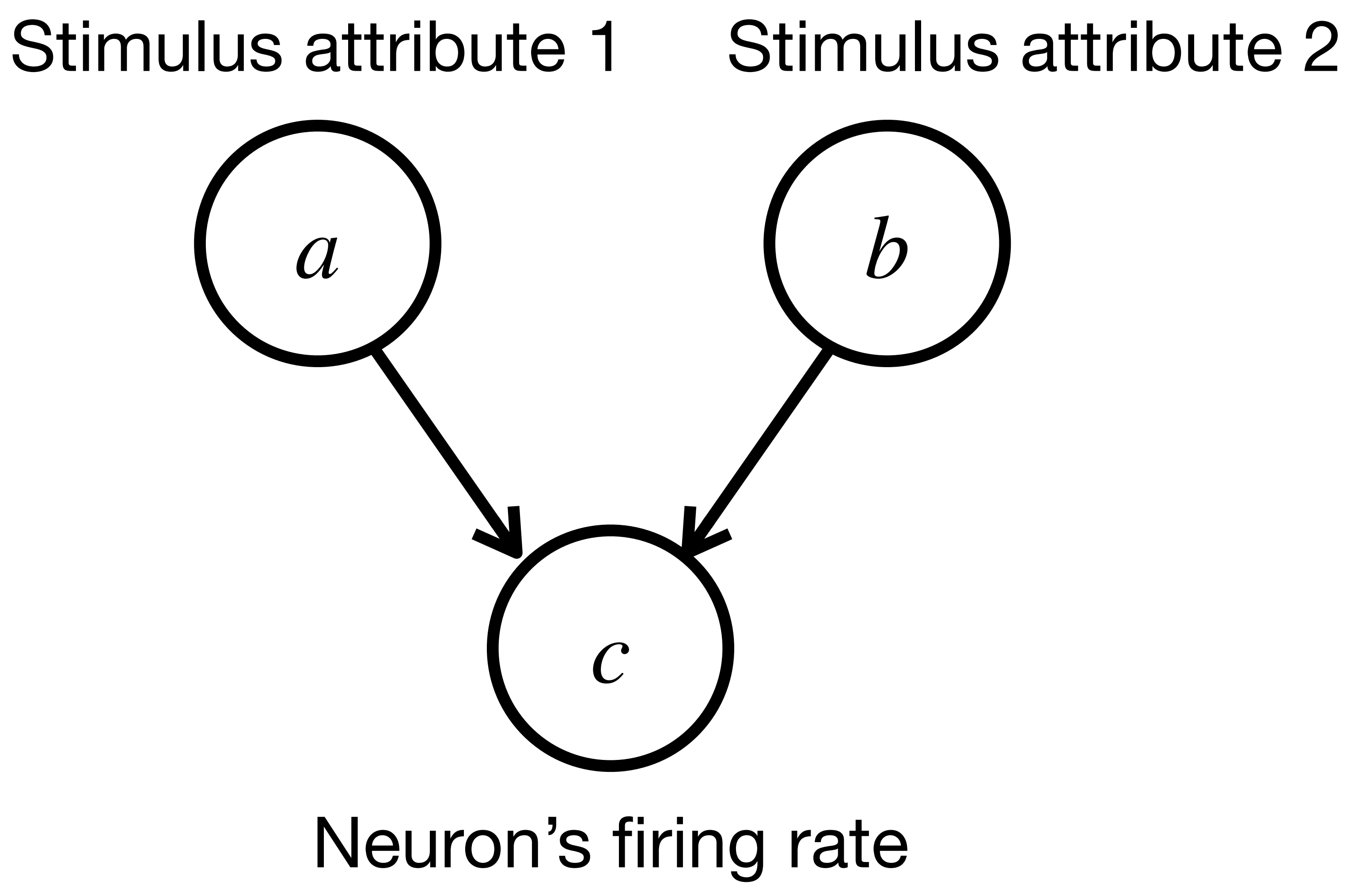
- Recall the definition of independence of $a$ and $b$:
  - $P(a|b) = P(a)$ or $P(a,b) = P(a)P(b)$
- Definition of conditional independence:
  - $P(a|b,c) = P(a|c)$ or $P(a,b|c) = P(a|c)P(b|c)$
- We would say “a and b are conditionally independent given c”

# Conditional independence

- For each of the following graphical models, let's ask:
  i) What is the factored form of $P(a, b, c)$?
  ii) Are $a$ and $b$ independent?
  iii) Are $a$ and $b$ conditionally independent given $c$?

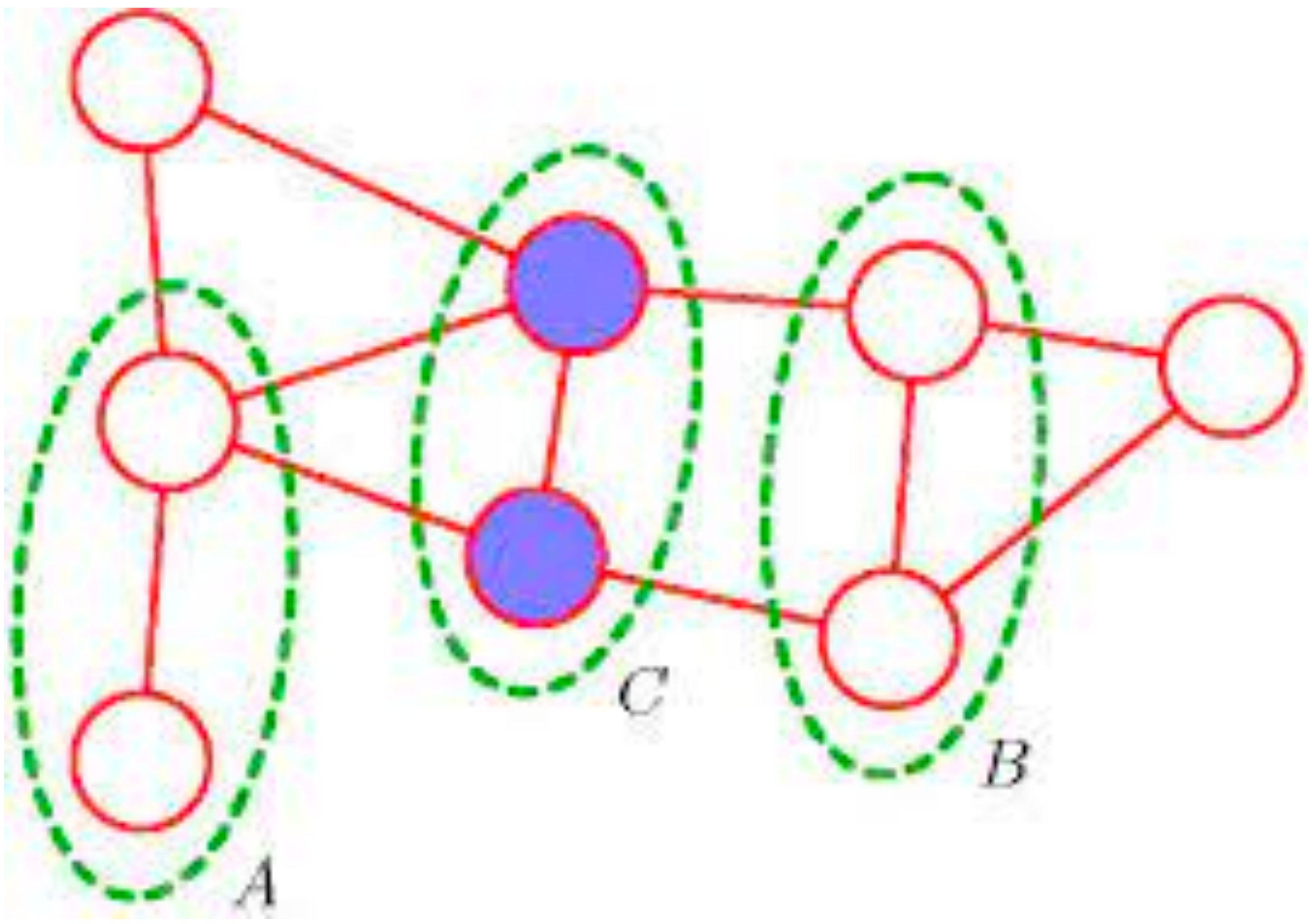


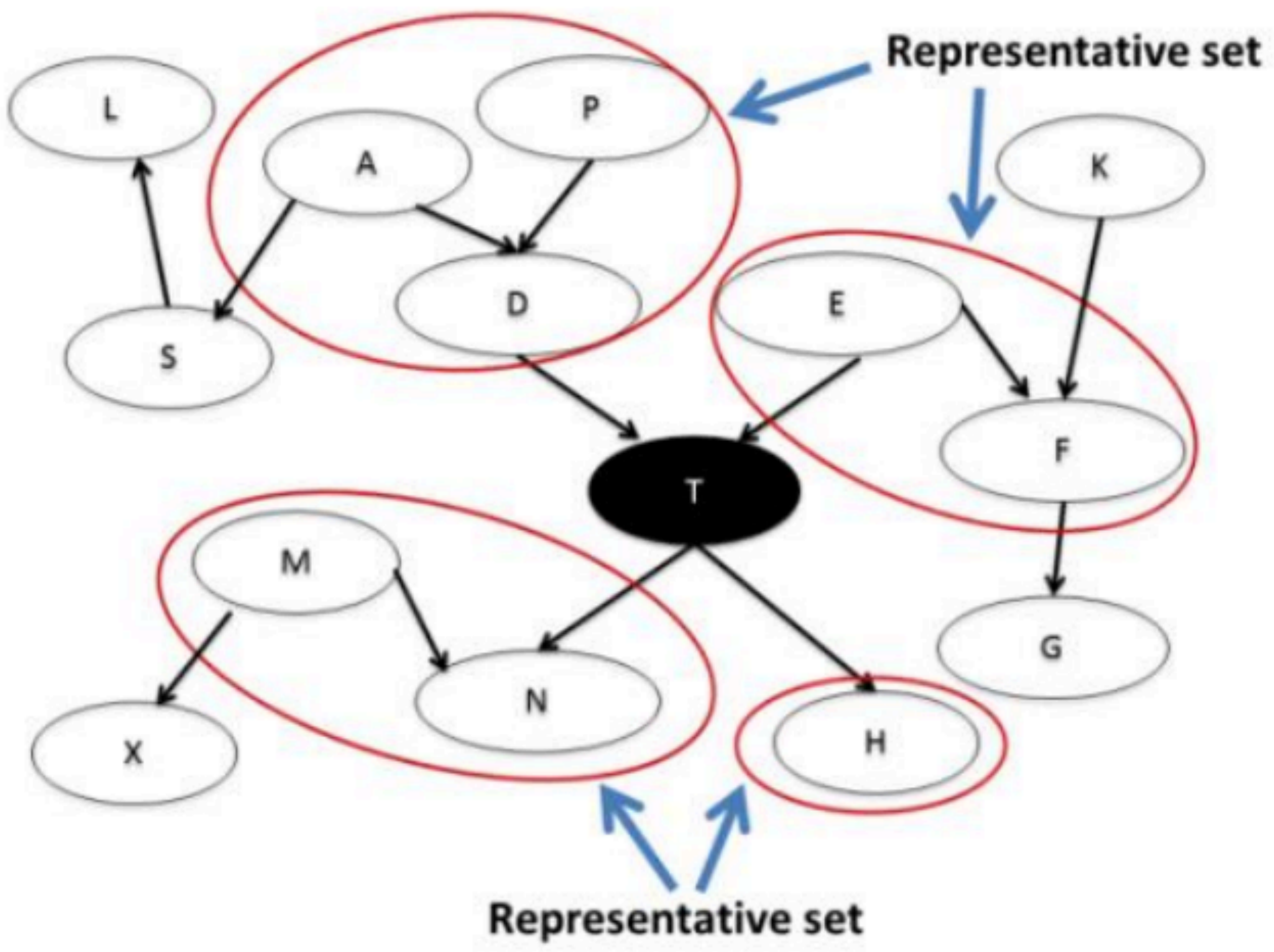



# Conditional independence

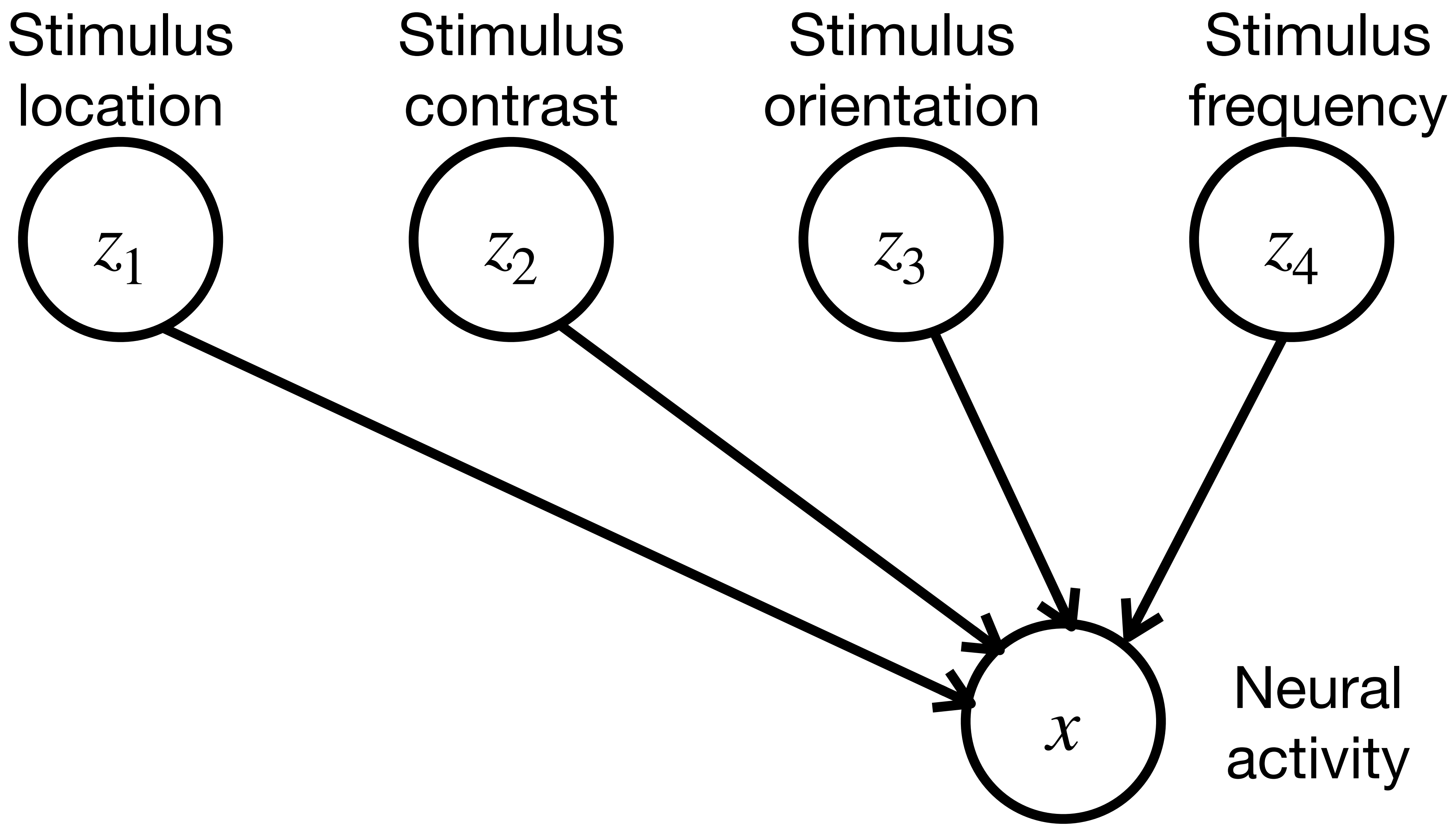
- $a, b, c$ can be one variable, or a set of (non-overlapping) variables



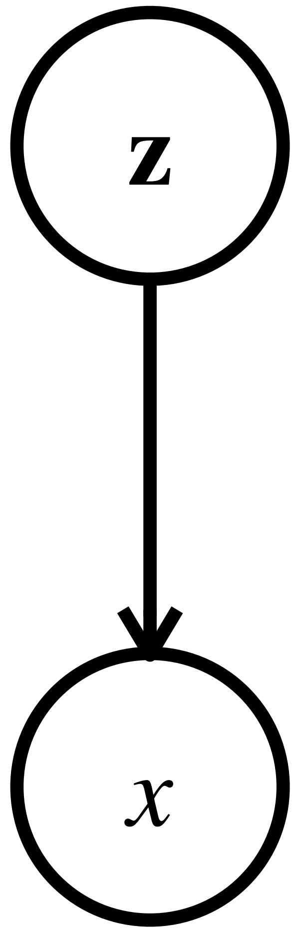



# Generative models

- Graphical model provides a picture of the <u>causal process</u> by which the data arose.
- Graphical model provides an intuitive way of generating synthetic data from joint distribution.
- Example:
  - Assume a generalized linear model
  - $\mu = w_1 z_1 + w_2 z_2 + w_3 z_3 + w_4 z_4$
  - $x \sim \mathcal{N}(\mu, \sigma^2)$
    - What is this?



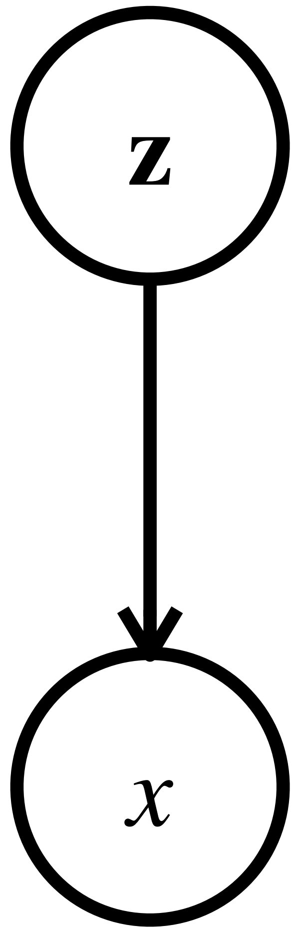
# Generalized linear model: linear regression




- Assume $x = \mathbf{w}^T\mathbf{z} + \epsilon$, and $\epsilon \sim \mathcal{N}(0,\sigma^2)$
- Then we have $P(x_i\,|\,\mathbf{z}_i) = \dfrac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\dfrac{(x_i - \mathbf{w}^T\mathbf{z}_i)^2}{2\sigma^2}\right)$
- Maximum likelihood (ML) is equivalent to least mean squares (LMS) minimization

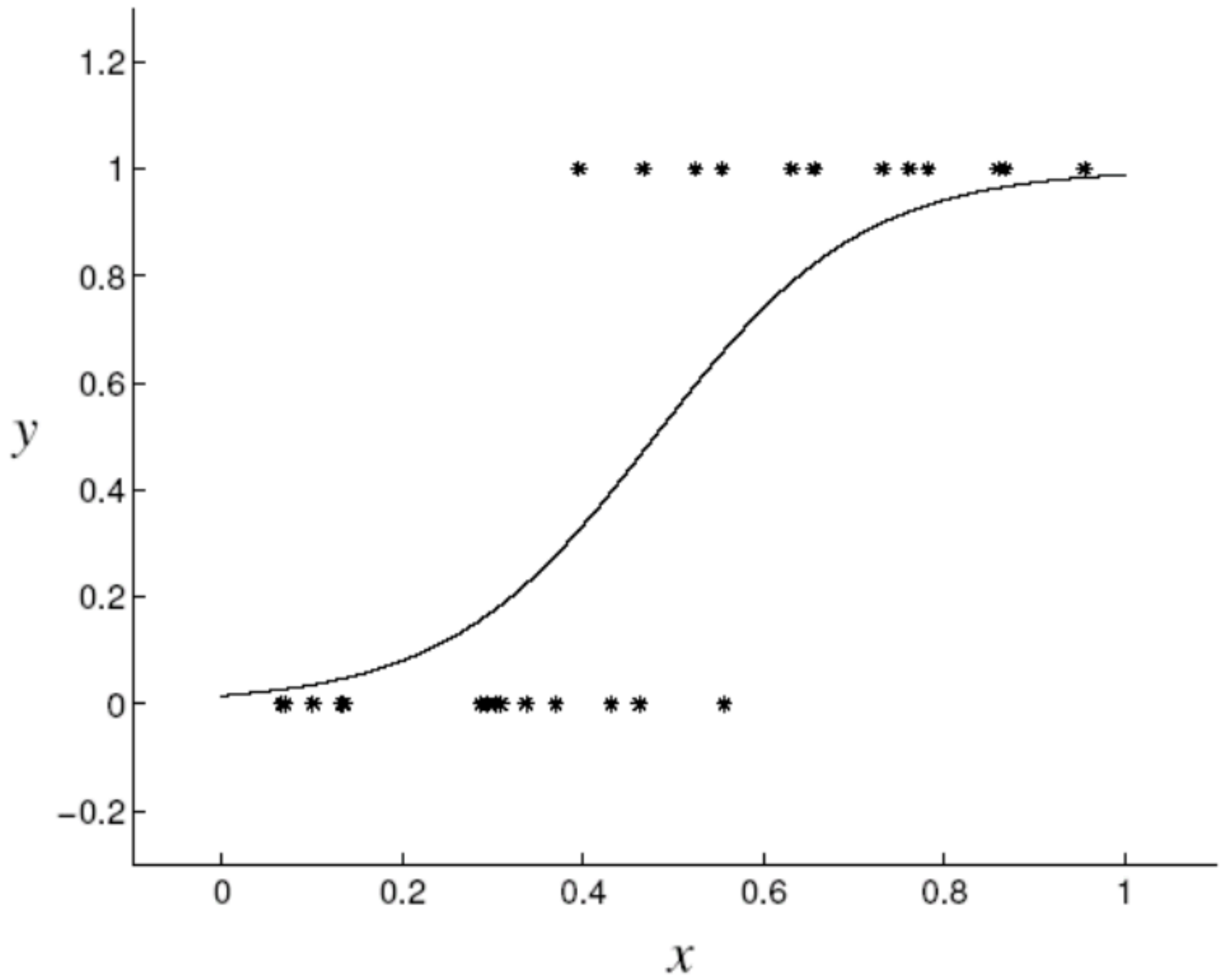
$$\arg\max_{\mathbf{w}} \prod_{i=1}^{N} P(x_i\,|\,\mathbf{z}_i) \Leftrightarrow \arg\min_{\mathbf{w}} \sum_{i=1}^{N} (x_i - \mathbf{w}^T\mathbf{z}_i)^2$$

# Generalized linear model: logistic regression




- Assume conditional distribution to be Bernoulli
  - $P(x_i \mid \mathbf{z}_i) = \mu(\mathbf{z})^x (1 - \mu(\mathbf{z}))^{(1-y)}$

  where $\mu$ is a logistic function
  - $\mu(\mathbf{z}) = \dfrac{1}{1 + \exp(-\mathbf{w}^T \mathbf{z})}$



# Generalized linear model: exponential family

- For a numeric random variable $x$:

$$p(x|\eta) = h(x)\exp\left\{\eta^T T(x) - A(\eta)\right\} = \frac{1}{Z(\eta)} h(x)\exp\left\{\eta^T T(x)\right\}$$

  Is an exponential family distribution with natural (canonical) parameter $\eta$

- Function $T(x)$ is a sufficient statistic
- Function $A(\eta) = \log Z(\eta)$ is the log normalizer
- Examples: Bernoulli, multinomial, Gaussian, Poisson, gamma

# Generalized linear model: exponential family

- Example: multivariate Gaussian

$$p(x\,|\,\mu,\Sigma) = \frac{1}{(2\pi)^{k/2}\,|\,\Sigma\,|^{1/2}}\exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$
$$= \frac{1}{(2\pi)^{k/2}}\exp\left\{-\frac{1}{2}\mathrm{Tr}(\Sigma^{-1}xx^T)+\mu^T\Sigma^{-1}x-\frac{1}{2}\mu^T\Sigma^{-1}\mu-\log|\,\Sigma\,|\right\}$$

- Exponential family representation:

$$\eta = \left[\Sigma^{-1}\mu;-\frac{1}{2}\mathrm{vec}(\Sigma)^{-1}\right] = [\eta_1,\mathrm{vec}(\eta_2)],\quad \eta_1 = \Sigma^{-1}\mu, \eta_2 = -\frac{1}{2}\Sigma^{-1}$$
$$T(x) = [x;\mathrm{vec}(xx^T)]$$
$$A(\eta) = \frac{1}{2}\mu^T\Sigma^{-1}\mu + \log|\,\Sigma\,| = -\frac{1}{2}\mathrm{Tr}(\eta_2\eta_1\eta_1^T) - \frac{1}{2}\log(-2\eta_2)$$
$$h(x) = (2\pi)^{-k/2}$$

# Generalized linear model: exponential family

- Example: Poisson distribution

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

- Exponential family representation:

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

# Why exponential family?

- Moment generating property

$$
\begin{aligned}
\frac{dA}{d\eta} &= \frac{d}{d\eta}\log Z(\eta) = \frac{1}{Z(\eta)}\frac{d}{d\eta}Z(\eta) \\
&= \frac{1}{Z(\eta)}\frac{d}{d\eta}\int h(x)\exp\{\eta^T T(x)\}dx \\
&= \int T(x)\frac{h(x)\exp\{\eta^T T(x)\}}{Z(\eta)}dx \\
&= \mathbb{E}[T(x)]
\end{aligned}
$$

$$
\begin{aligned}
\frac{d^2A}{d\eta^2} &= \int T^2(x)\frac{h(x)\exp\{\eta^T T(x)\}}{Z(\eta)}dx - \int T(x)\frac{h(x)\exp\{\eta^T T(x)\}}{Z(\eta)}dx\frac{1}{Z(\eta)}\frac{d}{d\eta}Z(\eta) \\
&= \mathbb{E}[T^2(x)] - \mathbb{E}^2[T(x)] \\
&= \mathrm{Var}[T(x)]
\end{aligned}
$$

# Moment estimation

- 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 centered moment
- $\frac{dA(\eta)}{d\eta} =$ mean, $\frac{d^2A(\eta)}{d\eta^2} =$ variance…
- When the sufficient statistic is a stacked vector, partial derivatives need to be considered.

# Moment vs canonical parameters

- The moment parameter $\mu$ can be derived from the natural (canonical) parameter

$$\frac{dA(\eta)}{d\eta} = \mathbb{E}(T(x)) \overset{\text{def}}{=} \mu$$

- $A(\eta)$ is convex since

$$\frac{d^2A(\eta)}{d\eta^2} = \text{Var}[T(x)] > 0$$

- Hence we can invert the relationship and infer the canonical parameter from the moment parameter (1-to-1):

$$\eta \overset{def}{=} \Psi(\mu)$$

A distribution in the exponential family can be parameterized not only by $\eta$ the canonical parameterization, but also by $\mu$ the moment parameterization.

# MLE for Exponential Family

- For i.i.d. data, the log-likelihood is

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

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

- Take derivatives and set to zero:

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

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

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

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

# Sufficiency

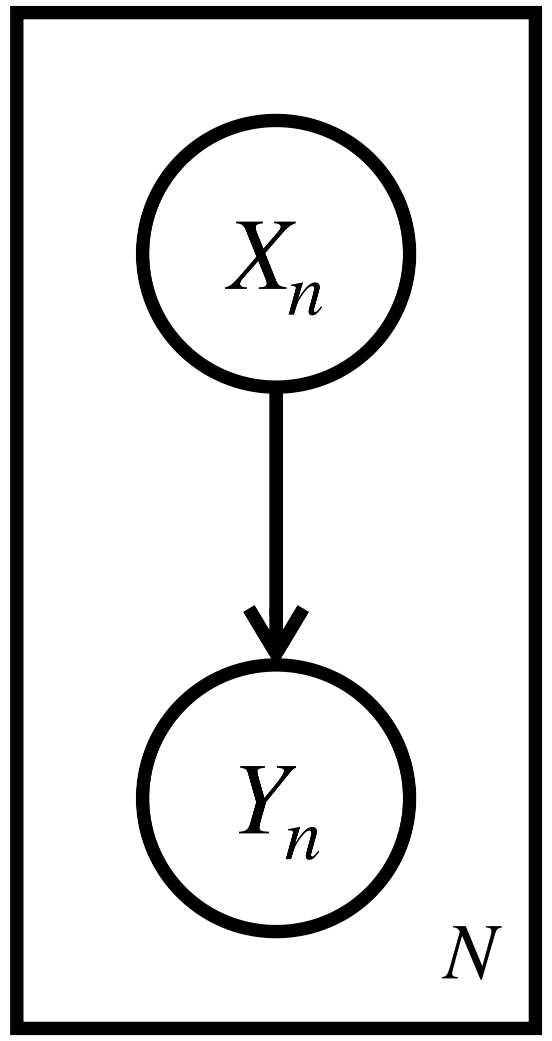
- For $p(x|\theta)$, $T(x)$ is sufficient for $\theta$ if there is no information in $X$ regarding $\theta$ beyond that in $T(x)$.
  - We can throw away $X$ for the purpose of inference w.r.t. $\theta$
  - Bayesian view $\quad X \rightarrow T(x) \rightarrow \theta \quad$ $p(\theta | T(x), x) = p(\theta | T(x))$
  - Frequentist view $\quad X \leftarrow T(x) \leftarrow \theta \quad$ $p(x | T(x), \theta) = p(x | T(x))$
  - The Neyman factorization theorem:
    - $T(x)$ is sufficient for $\theta$ if

$$p(x, T(x), \theta) = \Psi_1(T(x), \theta)\Psi_2(x, T(x))$$

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

$$X - T(x) - \theta$$

# Generalized Linear Models (GLMs)




- The graphical model
  - Linear regression
  - Discriminative linear classification
  - Commonality:
    - model $\mathbb{E}_p(Y) = \mu = f(\theta^T X)$
    - What is $p()$? The conditional distribution of $Y$
    - What is $f()$? The response function
- 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 $\mu$ is represented as a function $f(\xi)$ of $\xi$, 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 $\mu$

# GLM

$$\begin{tikzcd}[row sep=small]
\theta \arrow[dr] & & & & \\
& \xi \arrow[r, "f"] & \mu \arrow[r, "\psi"] & \eta \arrow[r, "EXP"] & y \\
x \arrow[ur] & & & &
\end{tikzcd}$$

$$p(y\,|\,\eta) = h(y)\exp\left\{\eta^T(x)y - A(\eta)\right\}$$

- $\Rightarrow p(y\,|\,\eta,\phi) = h(y,\phi)\exp\left\{\frac{1}{\phi}\left(\eta^T(x)y - A(\eta)\right)\right\}$
- The choice of exp family is constrained by the nature of the data $Y$
  - Example: $y$ is a continuous vector $\rightarrow$ multivariate Gaussian
    $y$ is a class label $\rightarrow$ Bernoulli or multinomial
- The choice of the response function
  - Following some mild constrains, e.g. [0, 1]. Positivity…
  - Canonical response function $f = \Psi^{-1}(\,\cdot\,)$
    - In this case $\theta^T x$ directly corresponds to canonical parameter $\eta$

# Example canonical response functions

| Model | Canonical response function |
|---|---|
| Gaussian | $\mu = \eta$ |
| Bernoulli | $\mu = 1/(1 + e^{-\eta})$ |
| multinomial | $\mu_i = \eta_i / \sum_j e^{\eta_j}$ |
| Poisson | $\mu = e^{\eta}$ |
| gamma | $\mu = -\eta^{-1}$ |

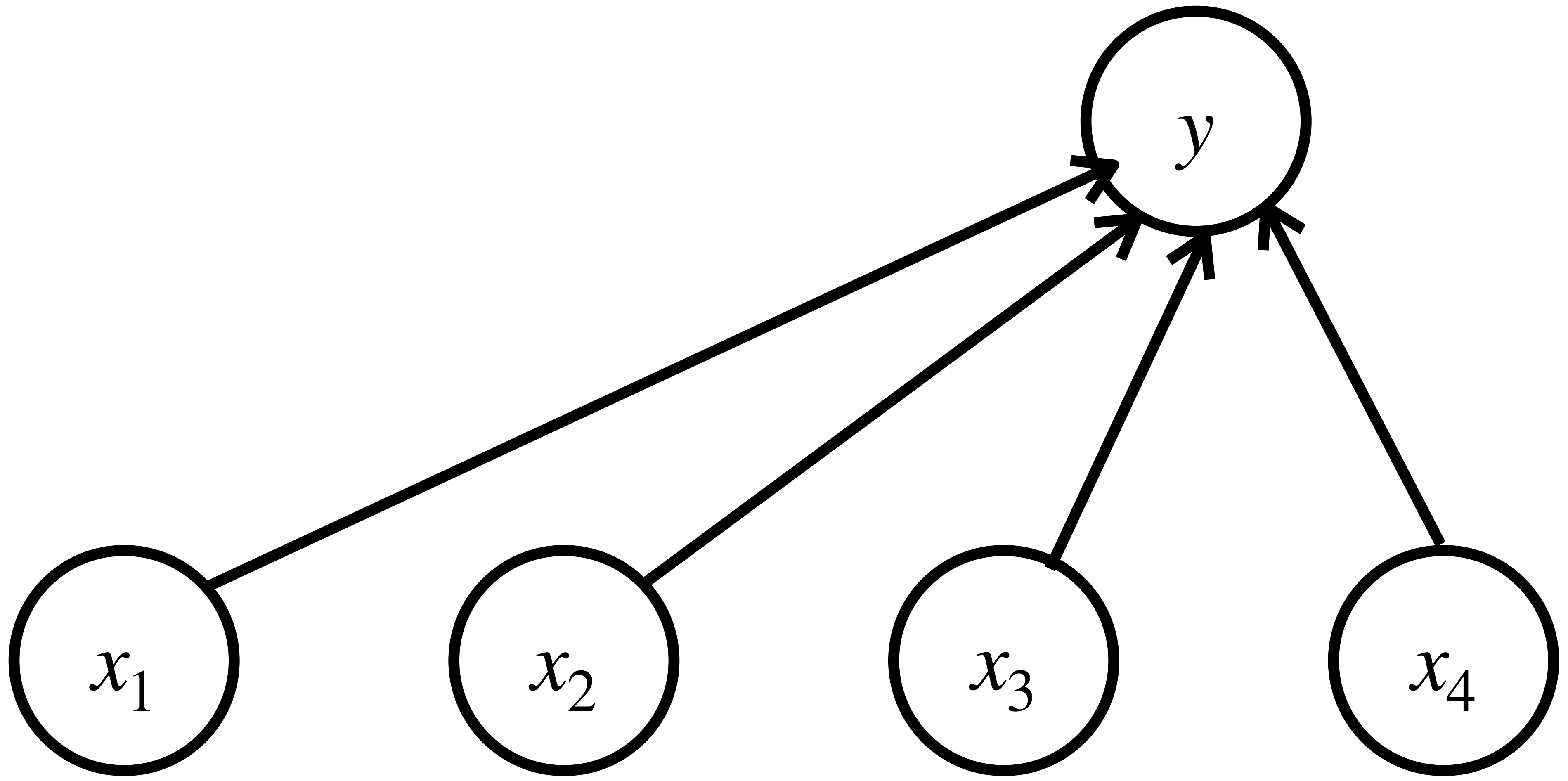
# MLE for GLMs with natural response

- Log-likelihood

$$\ell = \sum_n \log h(y_n) + \sum_n \left(\theta^T x_n y_n - A(\eta_n)\right)$$

- Derivative of Log-likelihood

$$\frac{d\ell}{d\theta} = \sum_n \left(x_n y_n - \frac{dA(\eta_n)}{d\eta_n}\frac{d\eta_n}{d\theta}\right)$$
$$= \sum_n (y_n - \mu_n)x_n$$
$$= X^T(y - \mu)$$




- Online learning for canonical GLMs
  - Stochastic gradient ascent = least mean squares (LMS) algorithm
    - $\theta^{t+1} = \theta^t + \rho(y_n - \mu_n^t)x_n$
    - where $\mu_n^t = \left(\theta^t\right)^T x_n$ and $\rho$ is a step size

**This is called back-propagation when applied to neural networks**

# Batch learning for canonical GLMs

- The Hessian matrix

$$H = \frac{d^2\ell}{d\theta d\theta^T} = \frac{d}{d\theta^T}\sum_n (y_n - \mu_n)x_n = \sum_n x_n \frac{d\mu_n}{d\theta^T}$$

$$= -\sum_n x_n \frac{d\mu_n}{d\eta_n}\frac{d\mu_n}{d\theta^T}$$

$$= -\sum_n x_n \frac{d\mu_n}{d\eta_n} x_n^T \quad \text{since } \eta_n = \theta^T x_n$$

$$= -X^T W X$$

- Where $X = [x_n^T]$ is the design matrix and

$$W = \text{diag}\left(\frac{d\mu_1}{d\eta_1}, \ldots, \frac{d\mu_N}{d\eta_N}\right)$$

which can be computed by calculating the 2nd derivative of $A(\eta_n)$

# Recall LMS

- Cost function in matrix form:

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{n}(x_i^T\theta - y_i)^2$$
$$= \frac{1}{2}(X\theta - y)^T(X\theta - y)$$

- To minimize $J(\theta)$, take derivative and set to zero:

$$\nabla_\theta = \frac{1}{2}\nabla_\theta \mathrm{Tr}\left(\theta^T X^T X\theta - \theta^T X^T y - y^T X\theta + y^T y\right)$$
$$= \frac{1}{2}\left(\nabla_\theta \mathrm{Tr}(\theta^T X^T X\theta) - 2\,\nabla_\theta \mathrm{Tr}(y^T X\theta) + \nabla_\theta(y^T y)\right)$$
$$= \frac{1}{2}(X^T X\theta + X^T X\theta - 2X^T y)$$
$$= X^T X\theta - X^T y = 0$$

$$X = \begin{bmatrix} -- & x_1 & -- \\ -- & x_2 & -- \\ \vdots & \vdots & \vdots \\ -- & x_n & -- \end{bmatrix}_{n\times p}$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1}$$

$$\Rightarrow$$

$$X^T X\theta = X^T y$$

$$\theta^* = (X^T X)^{-1}X^T y$$

# Iterative Reweighted Least Squares (IRLS)

- Recall Newton-Raphson methods with cost function $J$

$$\theta^{t+1} = \theta^t - H^{-1} \nabla_\theta J$$

- We now have

$$\nabla_\theta J = X^T(y - \mu)$$
$$H = -X^T W X$$

- Now

$$\begin{aligned}
\theta^{t+1} &= \theta^t + H^{-1} \nabla_\theta \ell \\
&= \left(X^T W^t X\right)^{-1} [X^T W^t X \theta^t + X^T(y - \mu^t)] \\
&= \left(X^T W^t X\right)^{-1} X^T W^t z^t
\end{aligned}$$

- Where the adjusted response is $z^t = X\theta^t + \left(W^t\right)^{-1}(y - \mu^t)$
- This can be understood as solving the following "iteratively reweighed least squares" problem:

$$\theta^{t+1} = \arg\max_\theta (z - X\theta)^T W (z - X\theta)$$

# Logistic regression

- Assume conditional distribution to be Bernoulli

  $P(y|x) = \mu(x)^y (1-\mu(x))^{(1-y)}$

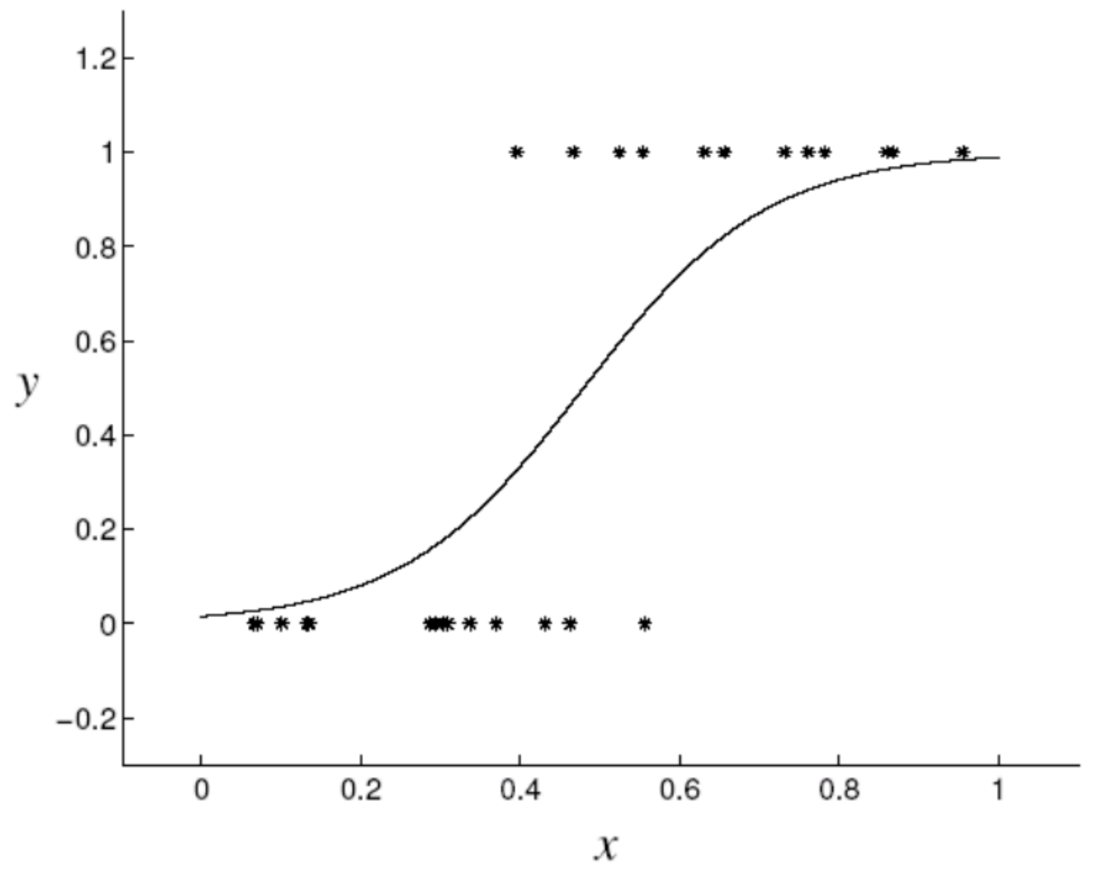
  where $\mu$ is a logistic function

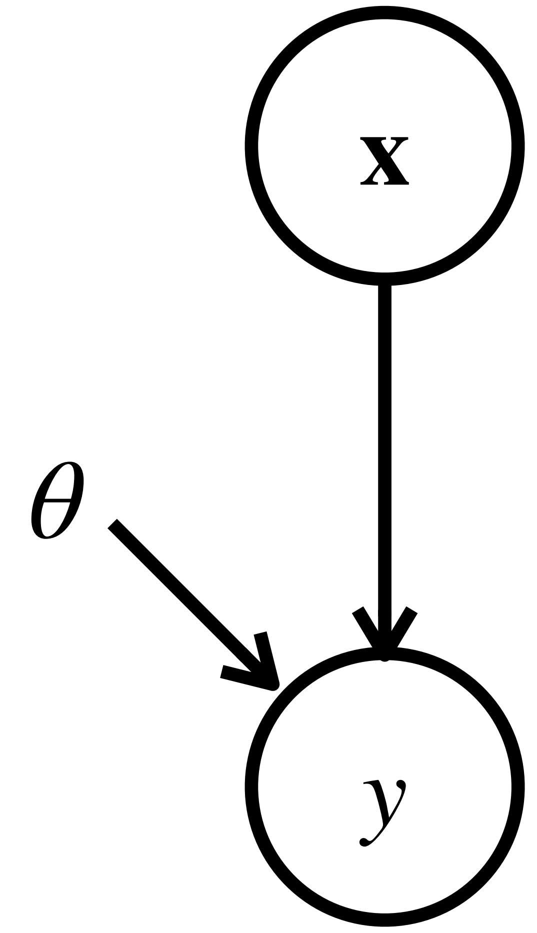
  $$\mu(x) = \frac{1}{1+\exp(-\eta(x))}$$

- $p(y|x)$ is an exponential family function with
  - Mean $E[y|x] = \mu = \dfrac{1}{1+\exp\{-\eta(x)\}}$
  - Canonical response function $\eta = \xi = \theta^T x$
- IRLS

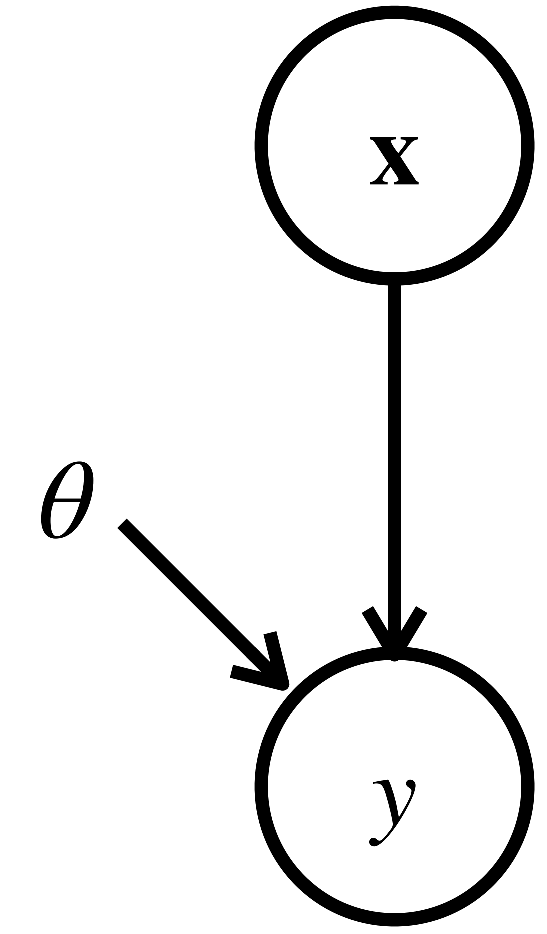
  $$\frac{d\mu}{d\eta} = \mu(1-\mu)$$

  $$W = \begin{pmatrix} \mu_1(1-\mu_1) 0 & & \cdots 0 \\ 0 & \mu_2(1-\mu_2) & \cdots 0 \\ 0 & 0 & \cdots 0 \\ 0 & 0 & \cdots \mu_N(1-\mu_N) \end{pmatrix}$$



# Logistic regression: practical issues




- It is very common to use regularized maximum likelihood

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

$$p(\theta) \sim \mathcal{N}(0, \lambda^{-1}\mathbf{I})$$

What if $p(|\theta|) \sim \text{Exp}(\lambda)$?

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

- 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:

$$\nabla_\theta \ell = \left(1 - \sigma(y_n\theta^T x_n)\right) y_n x_n - \lambda\theta$$

# Simple GMs are the building blocks of complex Bayes networks

- Density estimation
  - Parametric and nonparametric methods
- Regression
  - Linear, conditional mixture, nonparametric
- Classification
  - Generative and discriminative approach