Introduction to Machine Learning

Brown University CSCI 1950-F, Spring 2012
Prof. Erik Sudderth

Lecture 6:
Decision Theory for Continuous Estimation
Bayesian Model Selection
Directed Graphical Models

Many figures courtesy Kevin Murphy's textbook,
Machine Learning: A Probabilistic Perspective

Decision Theory

$y \in \mathcal{Y}$ ⟶ unknown hidden state of "nature"

$x \in \mathcal{X}$ ⟶ observed data

$a \in \mathcal{A}$ ⟶ set of possible actions we can take

$L(y, a)$ ⟶ real-valued loss function: the price we pay if we choose action *a*, and *y* is the true hidden state

- Goal: Choose the action which minimizes the expected loss

$$\delta(\mathbf{x}) = \underset{a \in \mathcal{A}}{\operatorname{argmin}} \, \mathbb{E}\left[L(y, a)\right] \qquad \delta : \mathcal{X} \to \mathcal{A}$$

  - Some averaging is necessary because we don't know *y*
  - Two notions of expectation: Bayesian versus frequentist
- Some communities speak of maximizing expected utility, which is equivalent if utility equals negative loss

Losses for Continuous Estimation

$y \in \mathbb{R}^d \longrightarrow$ unknown continuous latent variable

$x \in \mathcal{X} \longrightarrow$ observed data, can take values in any space

$\mathcal{A} = \mathcal{Y} \longrightarrow$ action is to estimate value of the latent variable

$L(y, a) \longrightarrow$ function giving loss for all possible mistakes

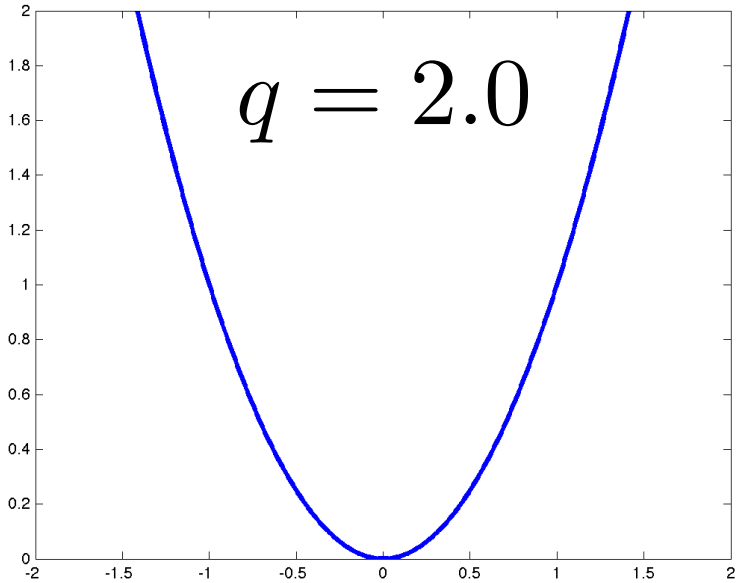
- Common choices for continuous loss functions:

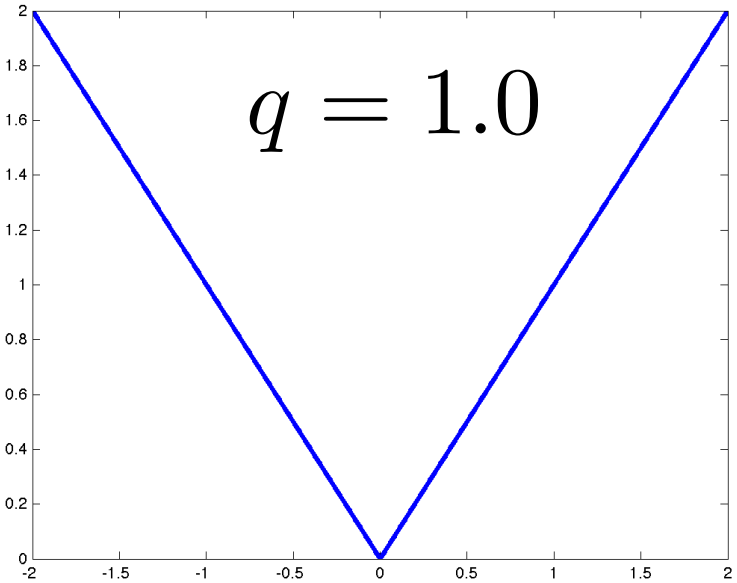
$$L(y, a) = (y - a)^2 \qquad \ell_2 \text{ loss, squared error}$$

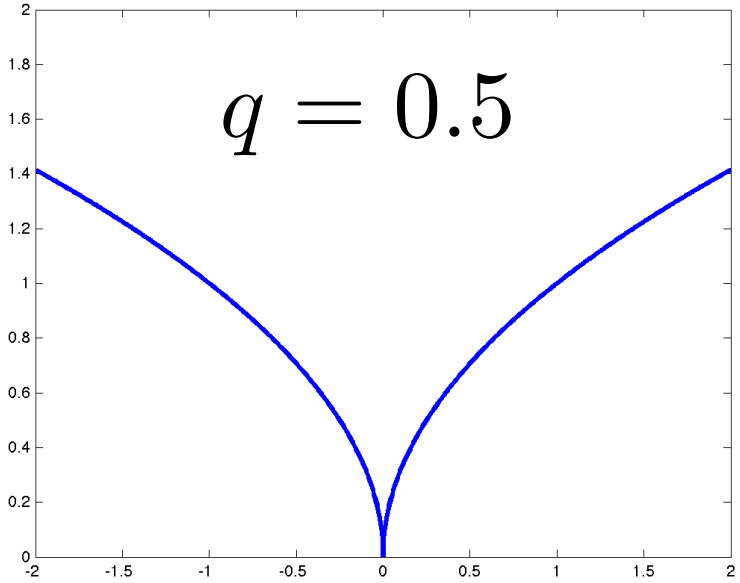
$$L(y, a) = |y - a| \qquad \ell_1 \text{ loss, absolute error}$$

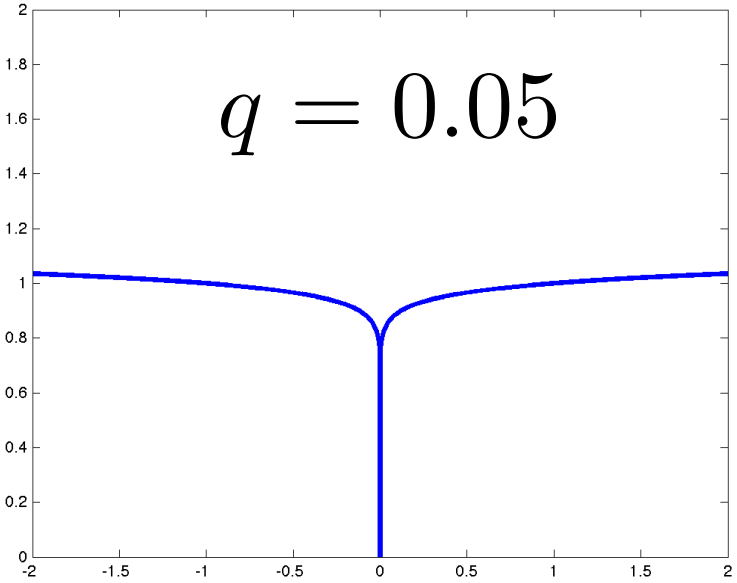
$$L(y, a) = |y - a|^q \qquad q > 0 \text{ tunable parameter}$$

Continuous Loss Functions









Minimizing Expected Loss

$y \in \mathbb{R}^d$ ⟶ unknown continuous latent variable

$x \in \mathcal{X}$ ⟶ observed data, can take values in any space

$\mathcal{A} = \mathcal{Y}$ ⟶ action is to estimate value of the latent variable

$L(y, a)$ ⟶ function giving loss for all possible mistakes

- The *posterior expected loss* of taking action *a* is

$$\rho(a \mid x) = \mathbb{E}[L(y, a) \mid x] = \int_{\mathcal{Y}} L(y, a) p(y \mid x)\, dy$$

- The optimal *Bayes decision rule* is then

$$\delta(\mathbf{x}) = \arg\min_{a \in \mathcal{A}} \rho(\mathbf{a}|\mathbf{x})$$

- Bayesian estimation requires *both* model and loss

Optimal Bayesian Estimators

$$\rho(a \mid x) = \mathbb{E}[L(y,a) \mid x] = \int_{\mathcal{Y}} L(y,a) p(y \mid x)\, dy$$

$$L(y,a) = (y-a)^2 \qquad \ell_2 \text{ loss, squared error}$$

Posterior Mean

$$\hat{y} = \mathbb{E}[y \mid x] = \int_{\mathcal{Y}} y p(y \mid x)\, dy$$

$$L(y,a) = |y-a| \qquad \ell_1 \text{ loss, absolute error}$$

Posterior Median

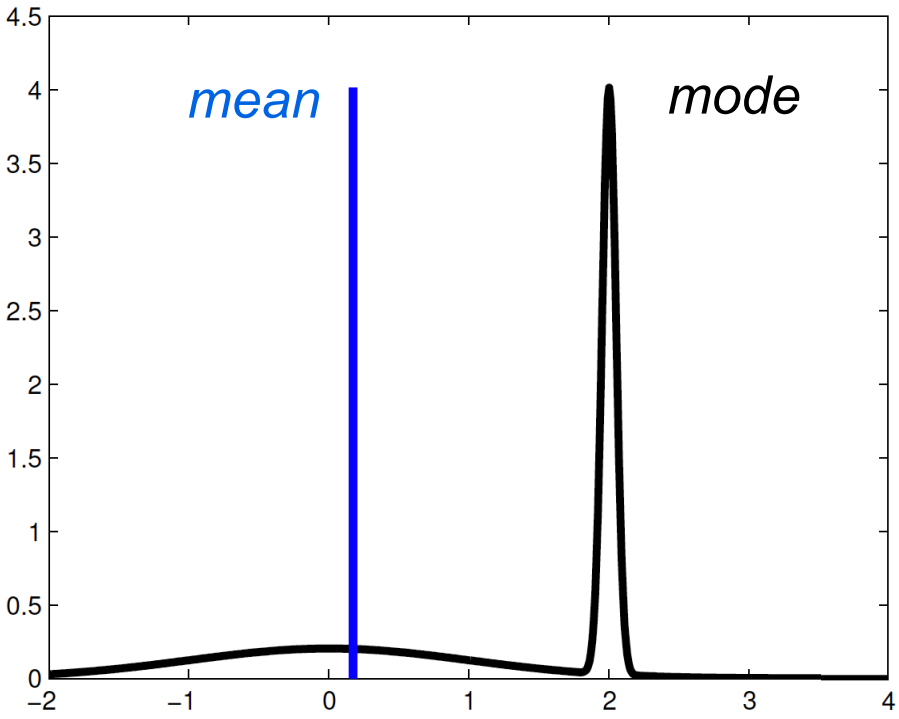
$$\int_{-\infty}^{\hat{y}} p(y \mid x)\, dy = \int_{\hat{y}}^{\infty} p(y \mid x)\, dy$$

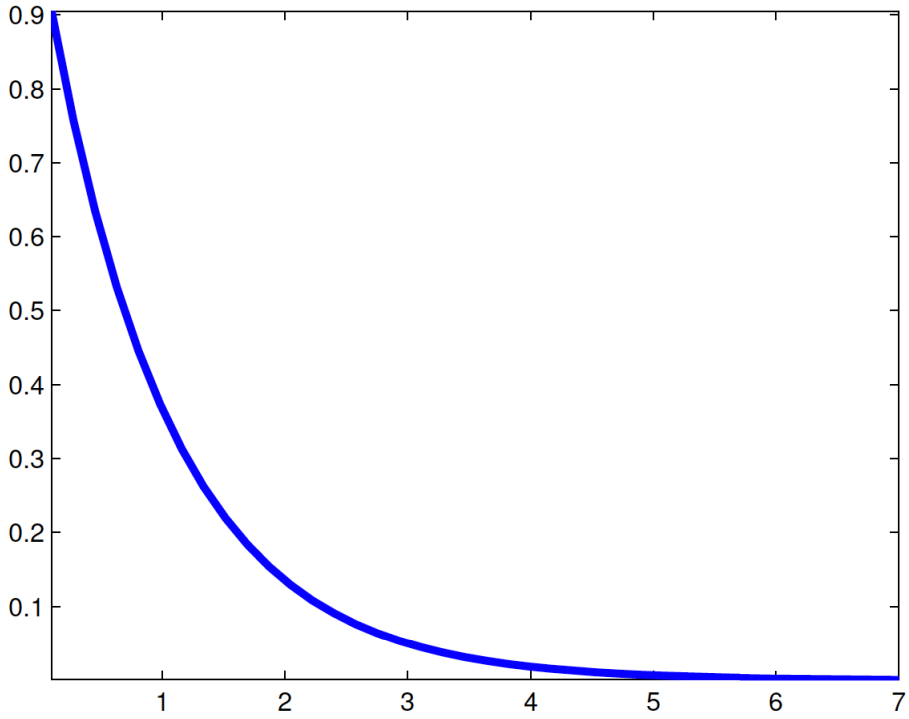
$$L(y,a) = |y-a|^q \qquad q > 0 \text{ tunable parameter}$$

No general closed form, but approaches MAP as $q \to 0$

$$\hat{y} = \arg\max_y p(y \mid x)$$

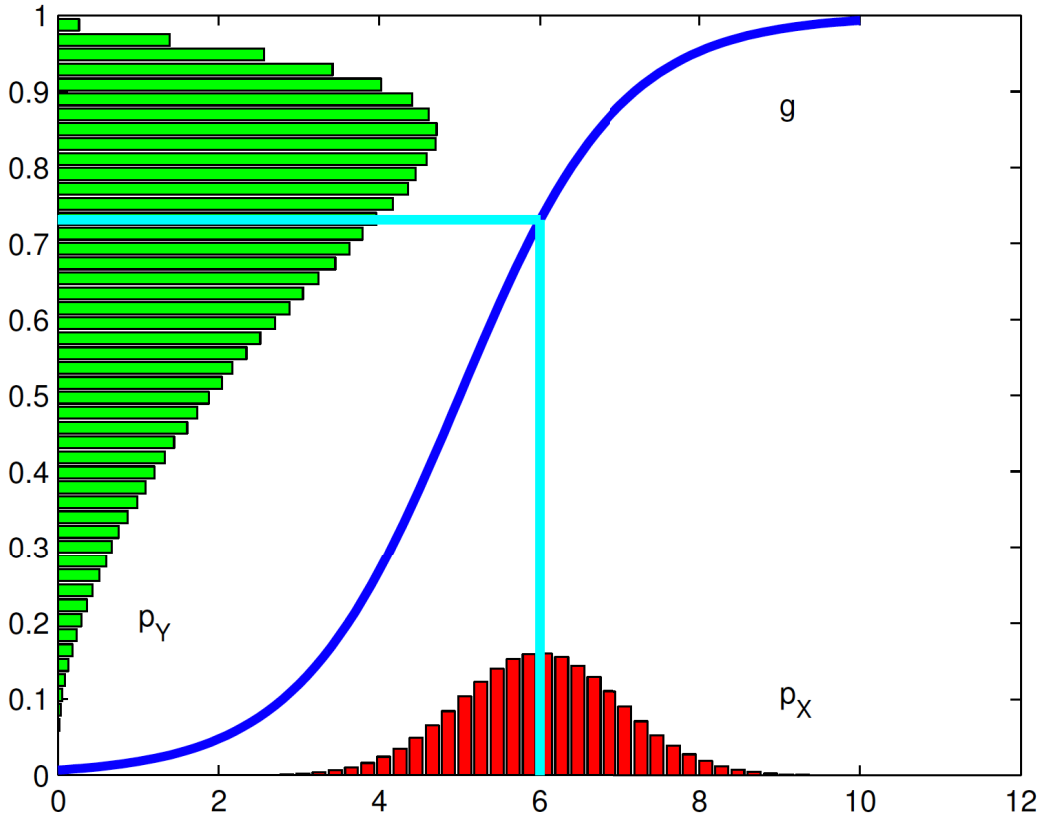
Warning: MAP may be atypical





The MAP pseudo-loss penalizes all errors equally,
but continuous MAP estimates are incorrect with probability 1

Warning: MAP not invariant to reparameterization



$$y = g(x)$$

$$p_y(y) = p_x(x)\left|\frac{dx}{dy}\right|$$

$$\hat{x} = \text{argmax}_x\, p_x(x)$$

$$\hat{y} = \text{argmax}_y\, p_y(y)$$

$$\neq g(\hat{x}) \quad \textit{in general}$$

ML estimates are invariant to reparameterization, as are Bayesian estimates based on non-degenerate losses.

What are Good Loss Functions?

Bayesian color constancy

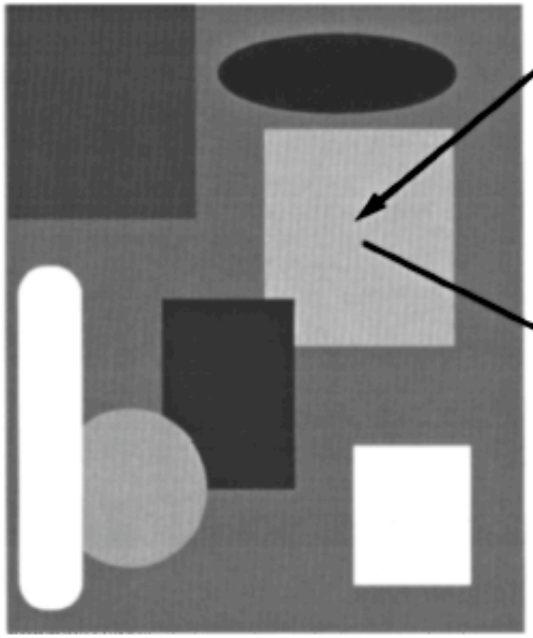
Journal of the Optical Society of America A, July 1997

David H. Brainard

Department of Psychology, University of California, Santa Barbara, California 93106

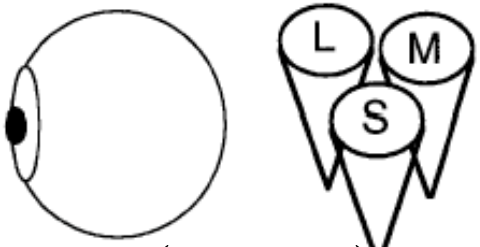
William T. Freeman

MERL, a Mitsubishi Electric Research Laboratory, Cambridge, Massachusetts 02139



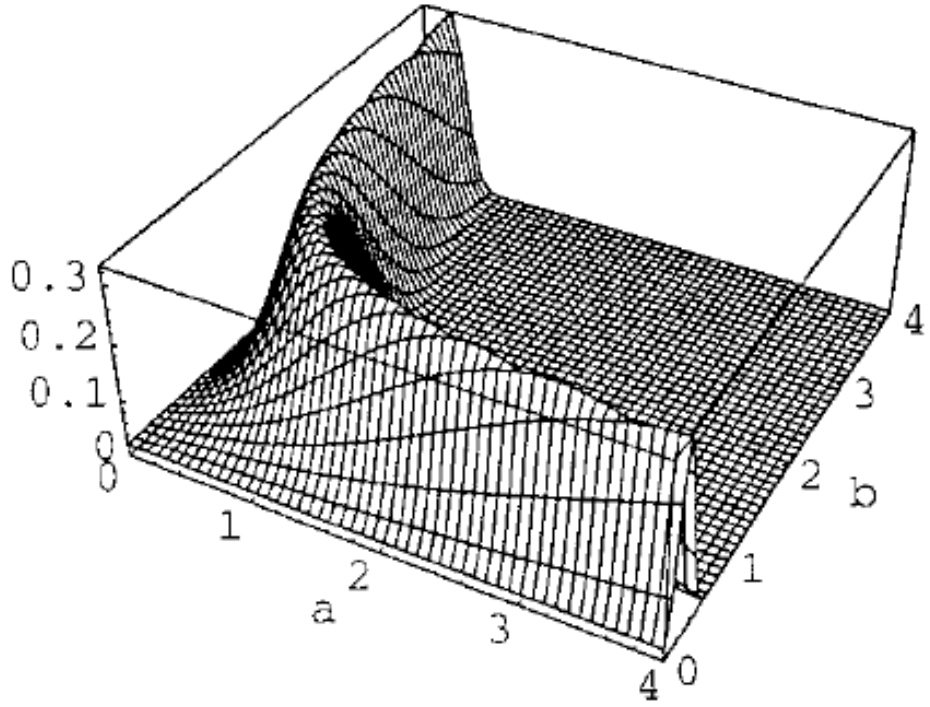
Illuminant $e \in \mathbb{R}^m$

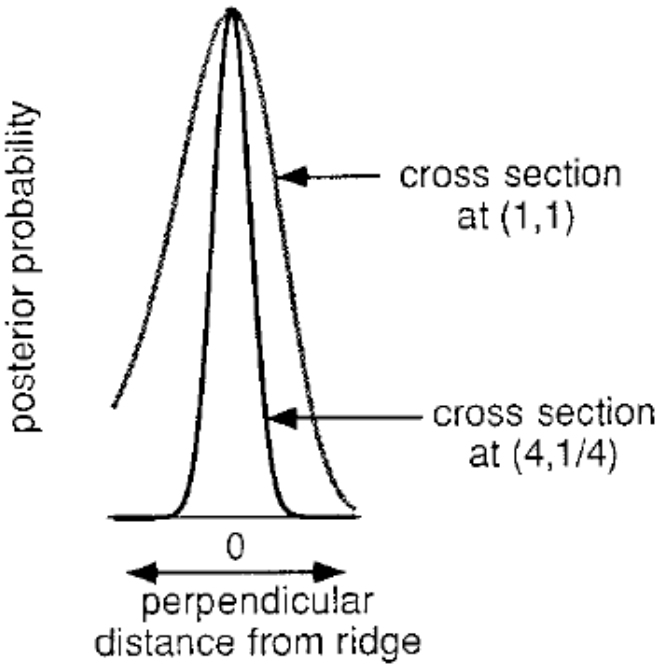
Reflectance at location *j*:
$s_j \in \mathbb{R}^m$



$$r_j = A(e \otimes s_j)$$
$$A \in \mathbb{R}^{3 \times M}$$

Toy Example

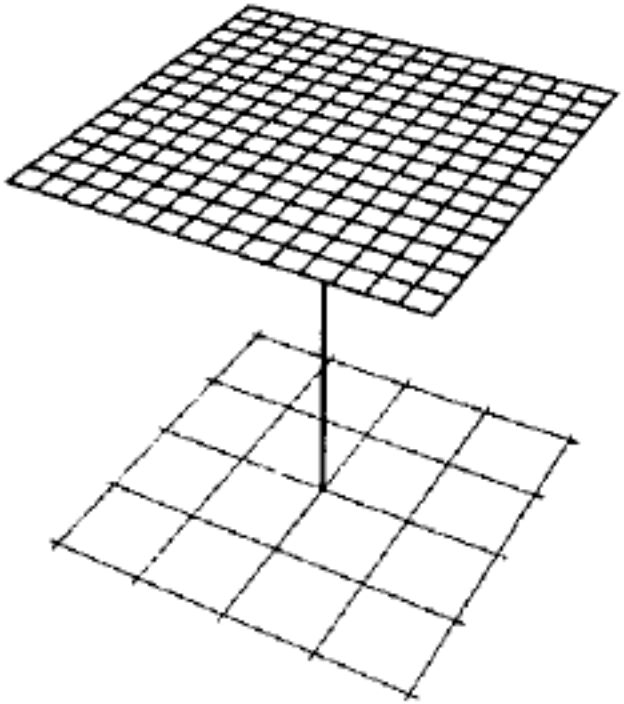




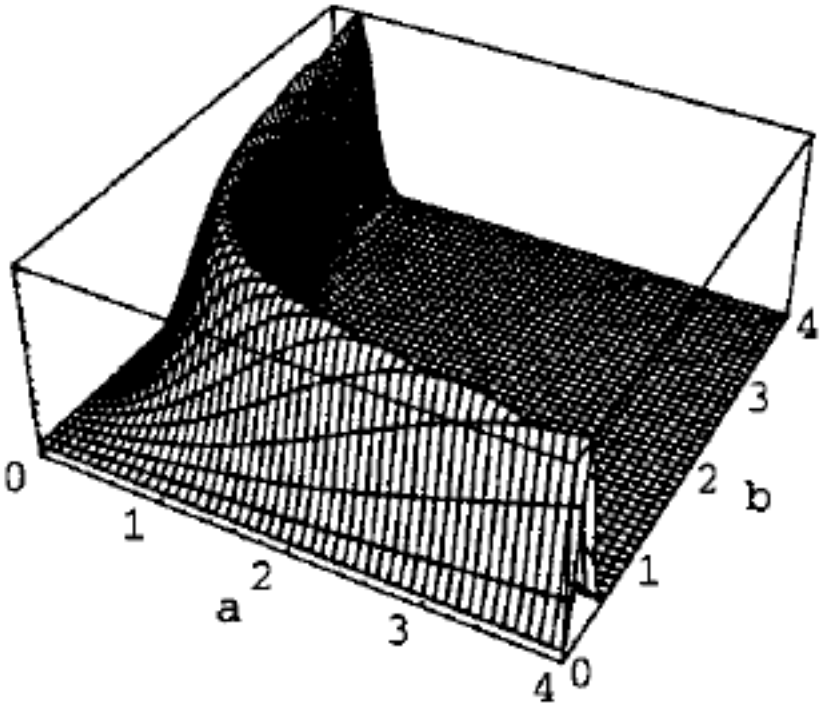
$$(a, b) \sim \text{Unif}([0, 4] \times [0, 4])$$

$$p(y \mid a, b) = \text{Norm}(y \mid ab, \sigma^2)$$

MAP Loss Function

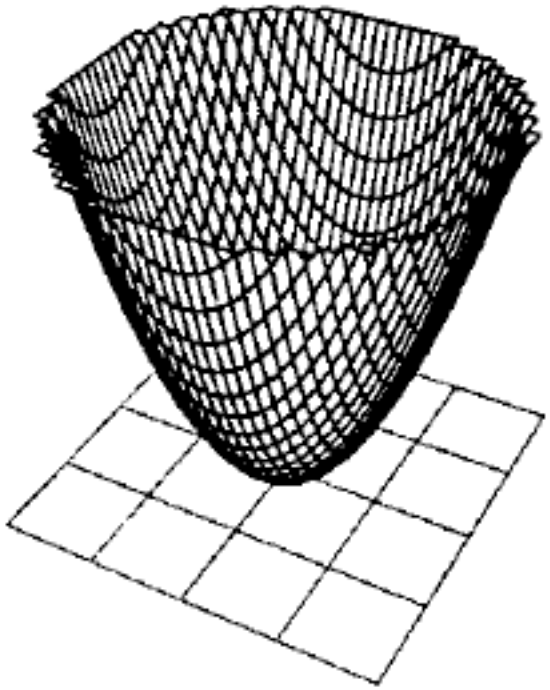


(a) MAP loss function

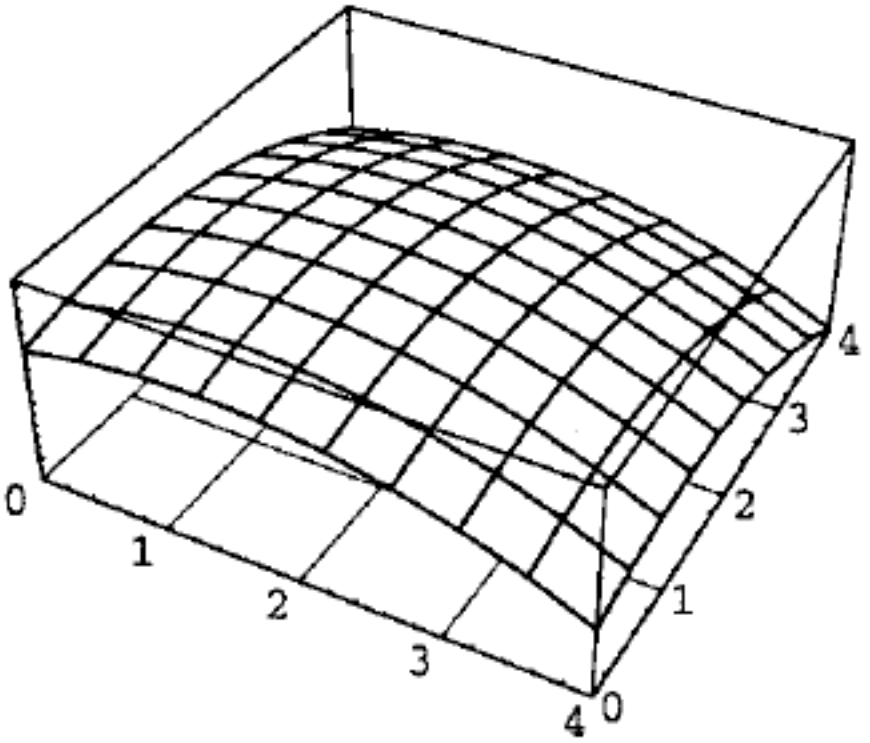


(d) (minus) MAP expected loss

Quadratic Loss Function

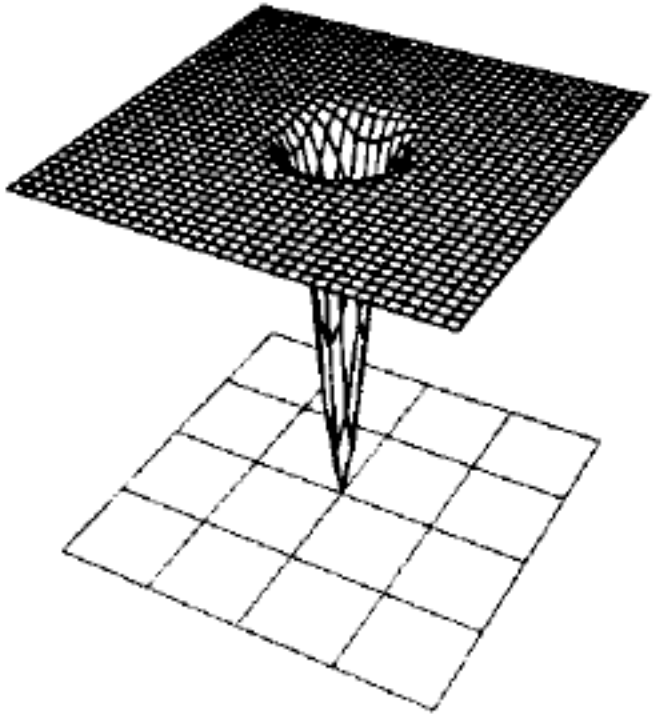


(b) MMSE loss function

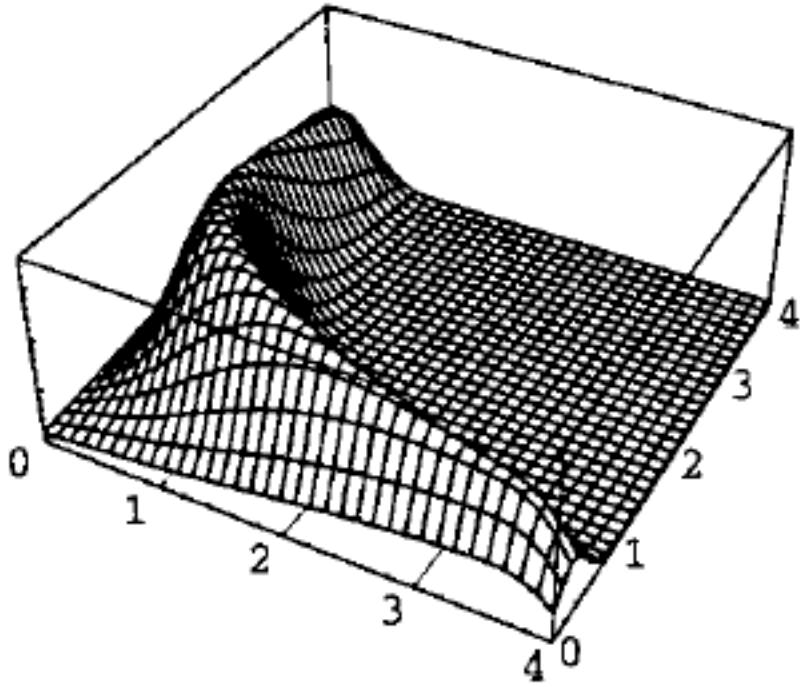


(e) (minus) MMSE expected loss

Local Mass Loss Function

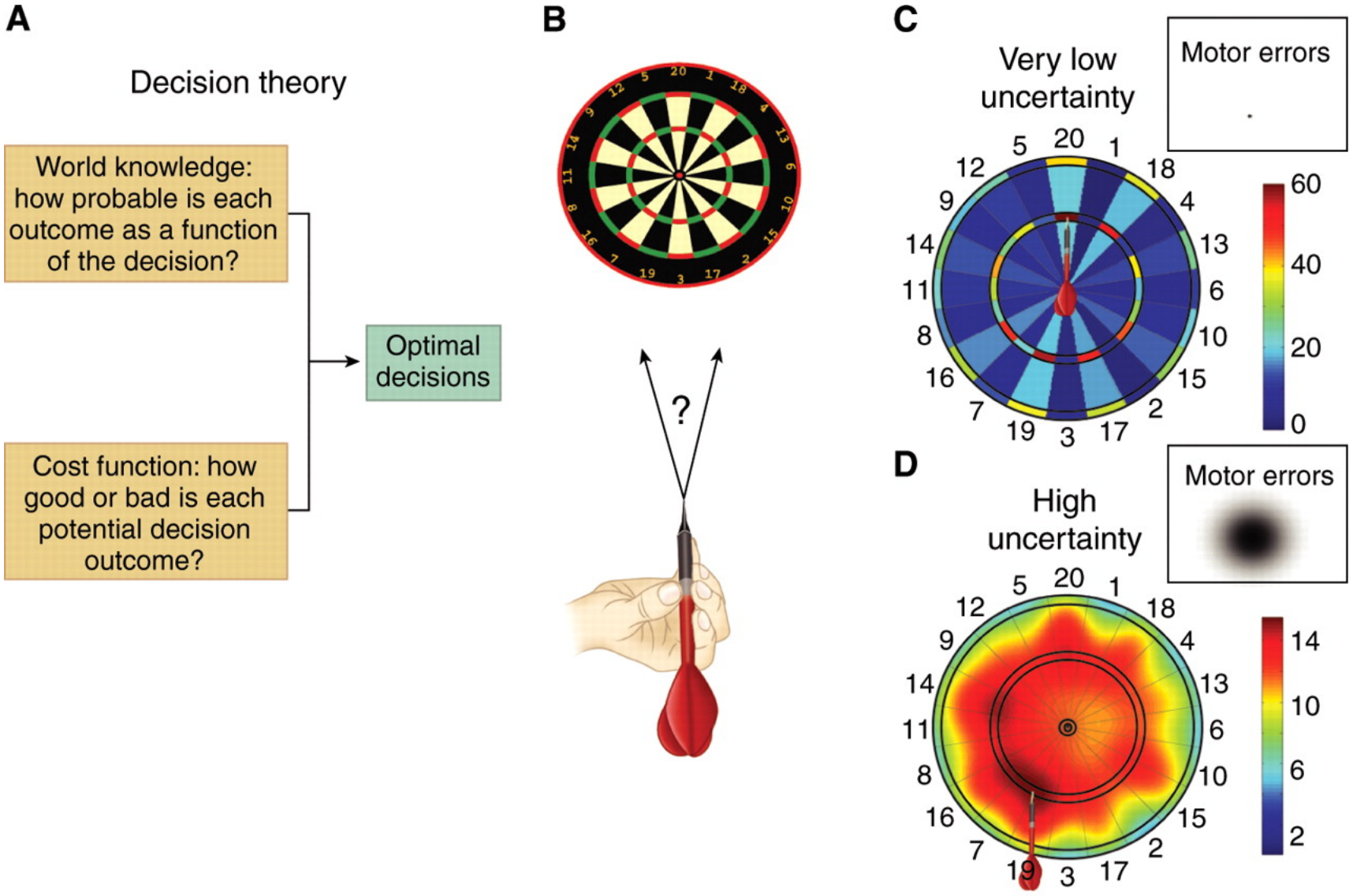


(c) MLM loss function



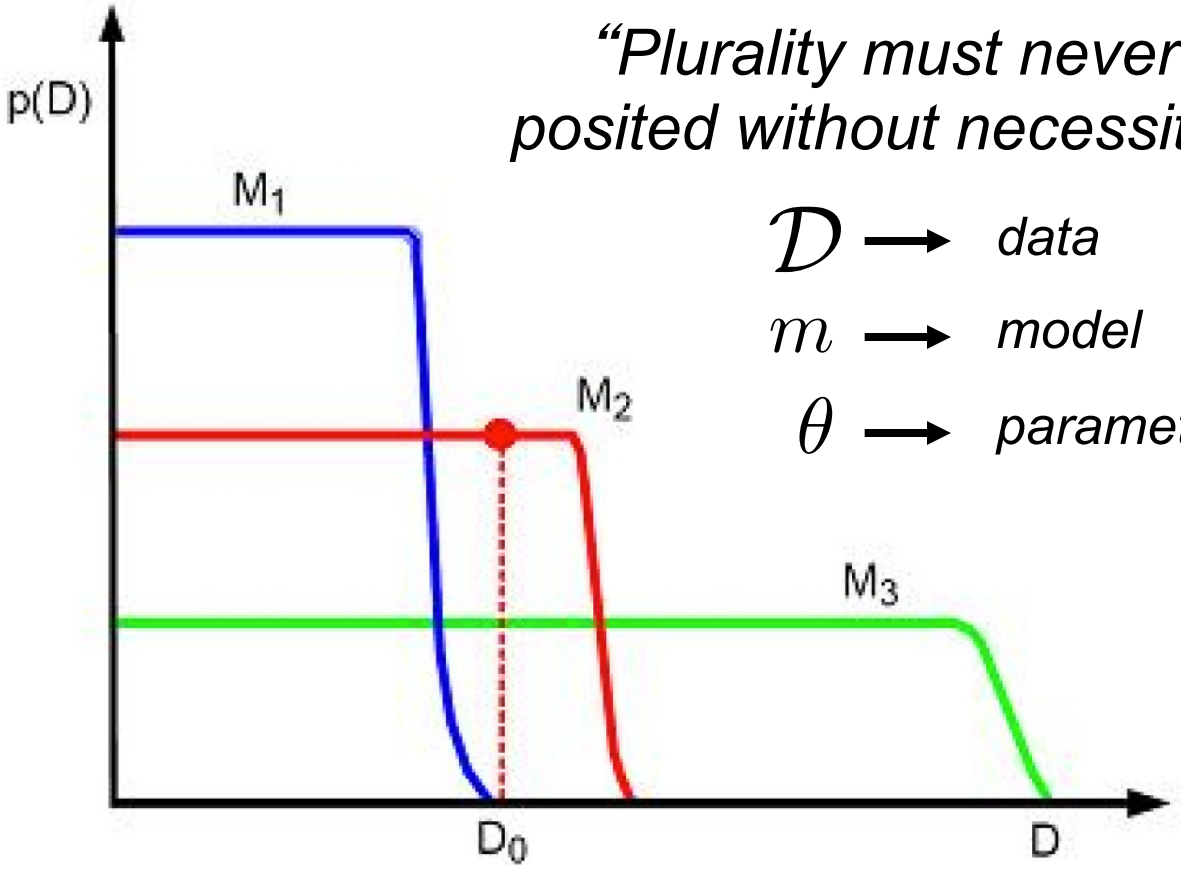
(f) (minus) MLM expected loss

Modeling Human Decisions



Koerding, Science Magazine, Oct. 2007

Bayesian Ockham's Razor



"Plurality must never be posited without necessity."

$\mathcal{D}$ → *data*

m → *model*

θ → *parameters*

William of Ockham

$$p(m|\mathcal{D}) = \frac{p(\mathcal{D}|m)p(m)}{\sum_{m \in \mathcal{M}} p(m, \mathcal{D})}$$

$$p(\mathcal{D}|m) = \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|m)d\boldsymbol{\theta}$$

Even with uniform $p(m)$, marginal likelihood provides a model selection bias

Computing Marginal Likelihoods

$$p(\mathcal{D}|m) = \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|m)d\boldsymbol{\theta}$$

Monte Carlo Approximation

$$p(\mathcal{D} \mid m) \approx \frac{1}{S}\sum_{s=1}^{S} p(\mathcal{D} \mid \theta^{(s)}) \qquad \theta^{(s)} \sim p(\theta \mid m)$$

Conjugate Priors

Beta-Bernoulli, Dirichlet-Multinoulli, Gaussian-Gaussian, etc.

$$p(\theta) = \frac{1}{Z_0} q(\theta)$$

$$p(\theta \mid \mathcal{D}) = \frac{1}{p(\mathcal{D})} p(\theta) p(\mathcal{D} \mid \theta)$$

$$= \frac{1}{Z_N} q(\theta \mid \mathcal{D})$$

Difference of log normalization constants

$$\log p(\mathcal{D}) = \log Z_N - \log Z_0$$

Example: Is this coin fair?

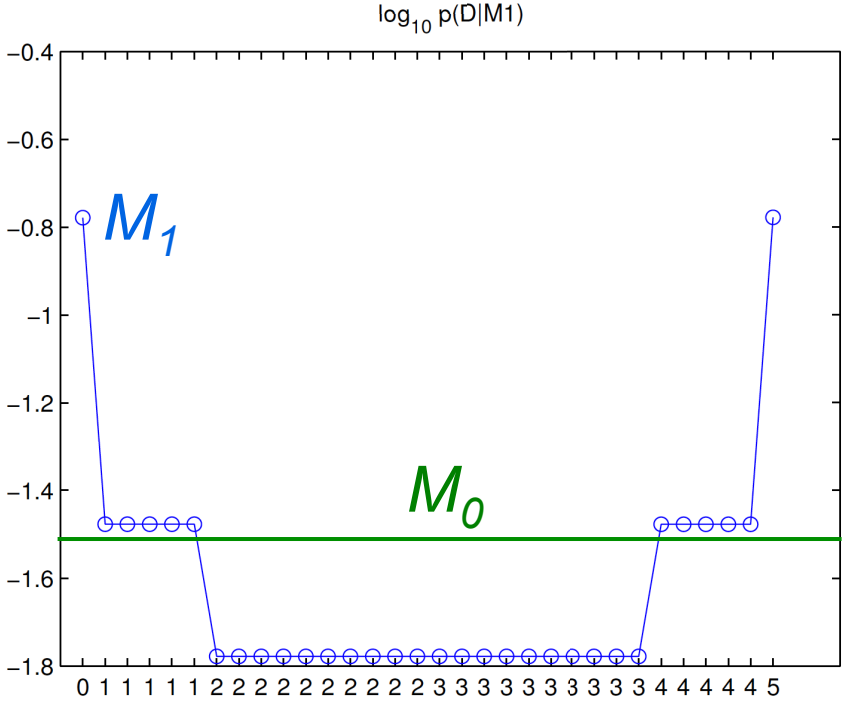
M_0: Tosses are from a fair coin: $\theta = 1/2$

M_1: Tosses are from a coin of unknown bias: $\theta \sim \mathrm{Unif}(0,1)$

Marginal Likelihoods

$$p(\mathcal{D}|M_0) = \left(\frac{1}{2}\right)^N \qquad p(\mathcal{D}|M_1) = \int p(\mathcal{D}|\theta)p(\theta)d\theta = \frac{B(\alpha_1 + N_1, \alpha_0 + N_0)}{B(\alpha_1, \alpha_0)}$$

$$B(a,b) \triangleq \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$



Number of heads in N=5 tosses

Very unbalanced counts are much more likely to have been caused by a biased coin

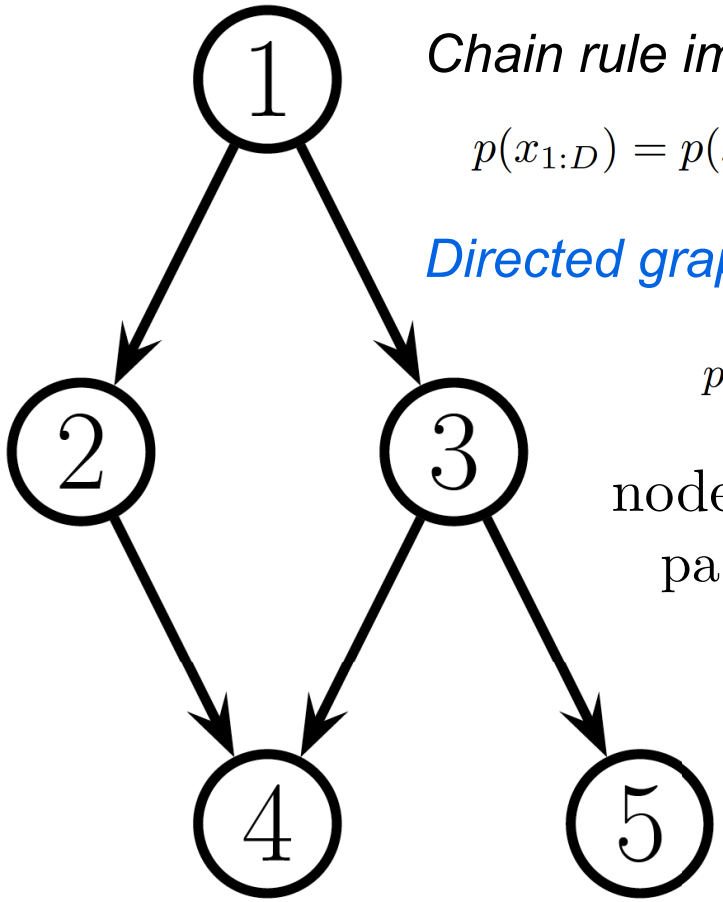
Model Selection: Bayes' Factors

$$BF_{1,0} := \frac{p(\mathcal{D}|M_1)}{p(\mathcal{D}|M_0)}$$

Bayes factor $BF(1,0)$	Interpretation
$B < \frac{1}{100}$	Decisive evidence for H_0
$B < \frac{1}{10}$	Strong evidence for H_0
$\frac{1}{10} < B < \frac{1}{3}$	Moderate evidence for H_0
$\frac{1}{3} < B < 1$	Weak evidence for H_0
$1 < B < 3$	Weak evidence for H_1
$3 < B < 10$	Moderate evidence for H_1
$B > 10$	Strong evidence for H_1
$B > 100$	Decisive evidence for H_1

As suggested by Jeffreys. Caveats: Can exhibit sensitivity to choice of priors for each model's parameters.
Most reliable when comparing pairs of "similar" models.

Directed Graphical Models



Chain rule implies that any joint distribution equals:

$$p(x_{1:D}) = p(x_1)p(x_2|x_1)p(x_3|x_2,x_1)p(x_4|x_1,x_2,x_3)\ldots p(x_D|x_{1:D-1})$$

Directed graphical model implies a restricted factorization:

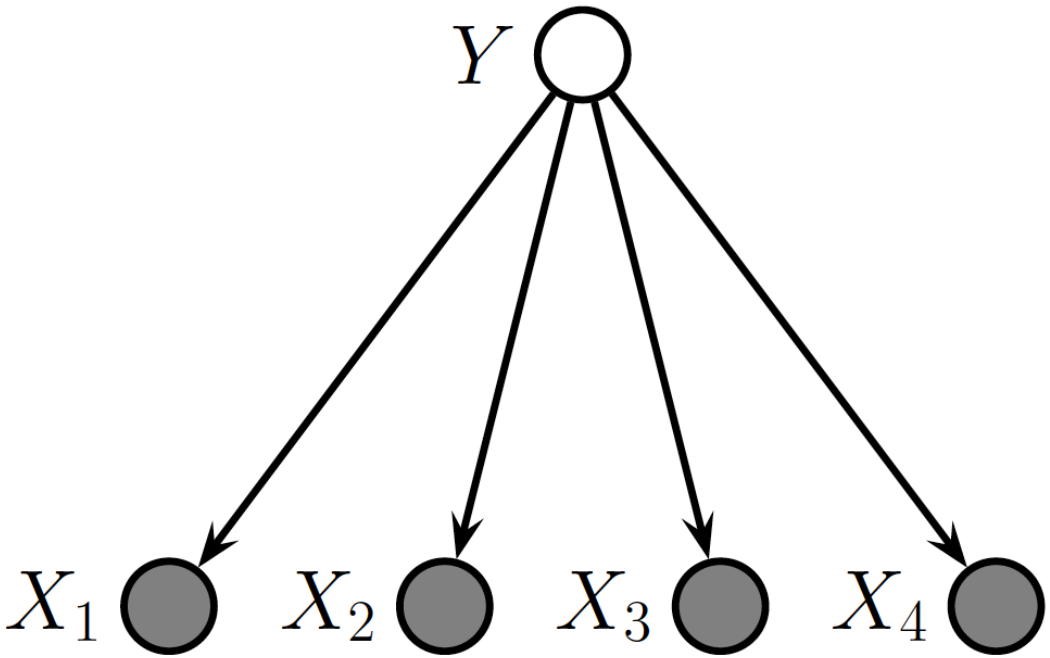
$$p(\mathbf{x}_{1:D}|G) = \prod_{t=1}^{D} p(x_t|\mathbf{x}_{\mathrm{pa}(t)})$$

$$\begin{aligned} \text{nodes} &\rightarrow \text{random variables} \\ \mathrm{pa}(t) &\rightarrow \text{parents with edges pointing to node } t \end{aligned}$$

Valid for any directed acyclic graph (DAG): equivalent to dropping conditional dependencies in standard chain rule

$$\begin{aligned} p(\mathbf{x}_{1:5}) &= p(x_1)p(x_2|x_1)p(x_3|x_1,\cancel{x_2})p(x_4|\cancel{x_1},x_2,x_3)p(x_5|\cancel{x_1},\cancel{x_2},x_3,\cancel{x_4}) \\ &= p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2,x_3)p(x_5|x_3) \end{aligned}$$

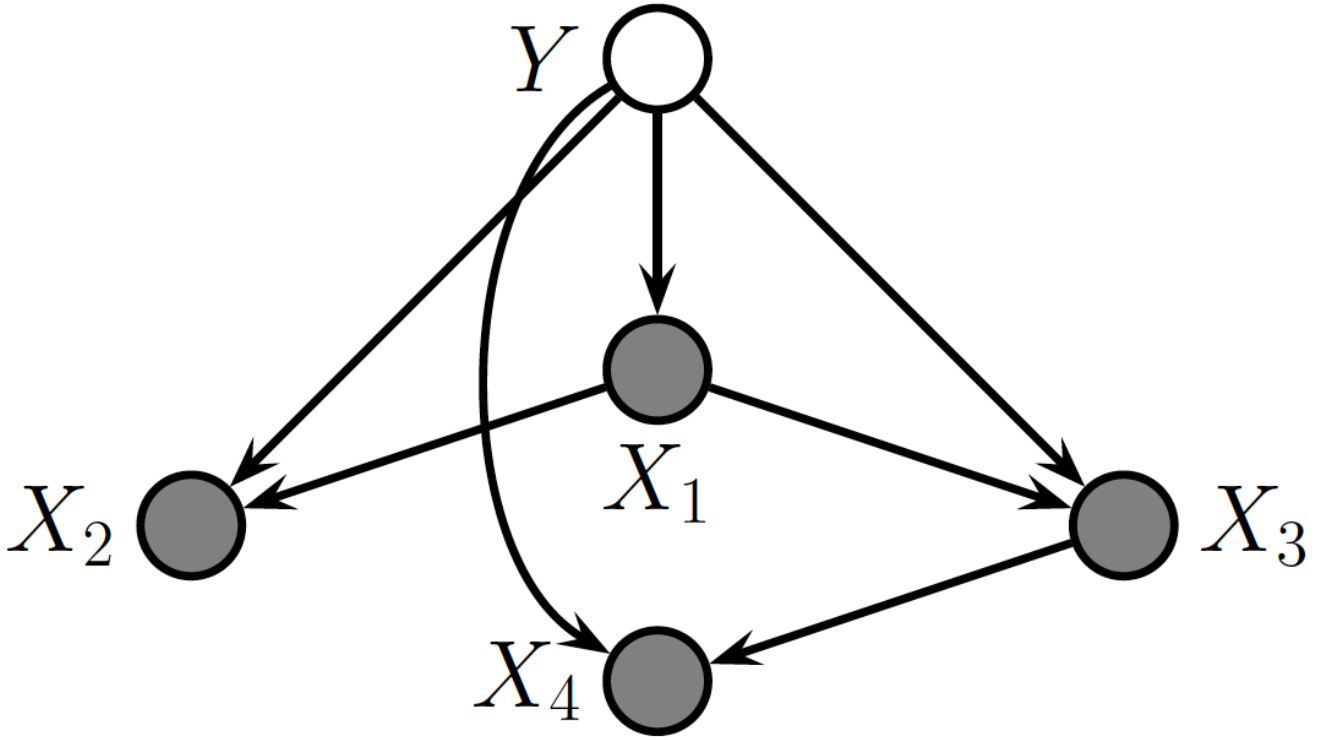
Name That Model



Naïve Bayes:

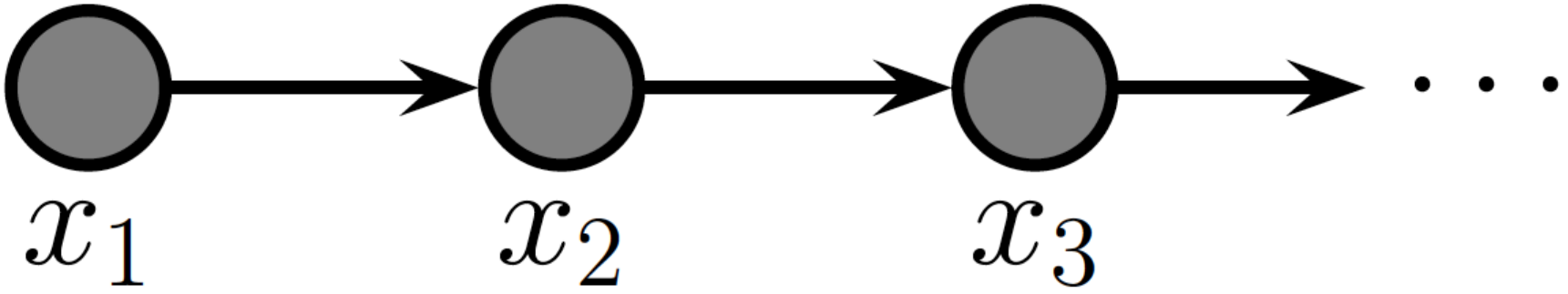
$$p(y, \mathbf{x}) = p(y) \prod_{j=1}^{D} p(x_j|y)$$

Name That Model



Tree-augmented Naïve Bayes

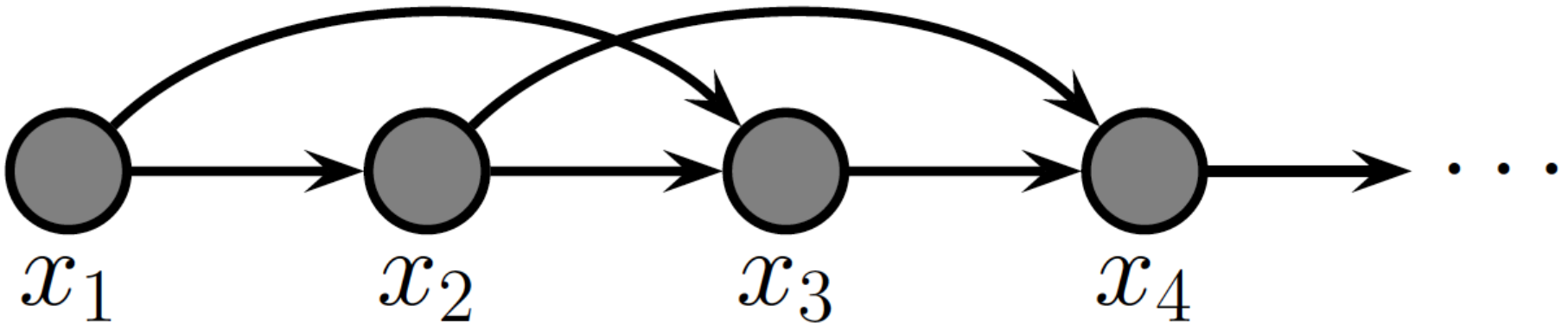
Name That Model



Markov chain:

$$p(\mathbf{x}_{1:D}) = p(x_1) \prod_{t=1}^{D} p(x_t|x_{t-1})$$

Name That Model



Second-order Markov chain:

$$p(\mathbf{x}_{1:T}) = p(x_1, x_2)p(x_3|x_1, x_2)p(x_4|x_2, x_3)\ldots = p(x_1, x_2)\prod_{t=3}^{T} p(x_t|x_{t-1}, x_{t-2})$$