COURSE INTRODUCTION

J. Elder

- The goal of this course is to provide a framework and computational tools for modeling visual inference, motivated by interesting examples from the recent literature.
- Models may be realized as algorithms to solve computer vision problems, or may constitute theories of visual processing in biological systems.
- The foundation of the course is a treatment of visual processing as a problem of statistical estimation and inference, grounded in the ecological statistics of the visual world.

Topics

- Bayesian decision theory
- Principal components and factor analysis
- Graphical Models
  - Markov Random Fields
  - Conditional Random Fields
  - Belief Propagation
- Clustering
  - Mean Shift
  - Expectation Maximization
  - Spectral Methods (Graph Cuts)
- Sampling
  - Gibbs Sampling
  - Markov Chain Monte Carlo
- Classifiers
  - Support Vector Machines
- Neural Networks

Course Format

- Each week will consist of two 1.5 hour meetings:
  - Meeting 1. A lecture by the instructor on a specific computational tool or approach
  - Meeting 2. A discussion, led by a specified student, of a selected computational vision paper in which this approach is applied to a specific problem.

Evaluation

- In addition to student presentations of short computational vision papers, two short MATLAB assignments will be collected and graded. The final project will involve application and possibly extension of a technique studied in the class to a problem chosen by the student.
  - Class Participation 10%
  - Paper Presentation 20%
  - Assignment 1 20%
  - Assignment 2 20%
  - Final Project 30%

Main Texts

- C.M. Bishop *Pattern Recognition and Machine Learning*. New York: Springer, 2006.
- S.J.D. Prince *Computer Vision Models*. Available in draft form at
  - http://computervisionmodels.blogspot.com/

k	Date	Topic	Required Readings	Additional Readings	Application Paper
	M Sept 13 W Sept 15	Probability & Bayesian Inference Probability Distributions & Parametric Modeling	Bishop Ch 1.1-1.2.5 (29 pages) Bishop Ch 2.1-2.3 (skip 2.3.5) (43 pages)	Pearl Ch 1.4-1.6, 2 Howson & Urbach 199 Prince Ch 1-4 Duda Ch 3.1-3.5	
	M Sept 20 W Sept 22	Probability Distributions & Parametric Modeling (cntd. Non-Parametric Modeling	Bishop Ch 2.5 (7 pages)	Duda Ch 4.1-4.5	Comaniciu & Meer 2002 (Mean Shift)
	M Sept 27 W Sept 29	Expectation Maximization	Prince Ch 5 (11 pages) Prince Ch 6.1-6.5, 6.8 (24 pages	Bishop Ch 9	
	M Oct 4 W Oct 6	Linear Subspace Models	Prince Ch 6.6-6.7, 6.9 (12 pages Bishop Ch 12 (40 pages)	Duda Ch 10.13-10.14	
	M Oct 11 W Oct 13	Reading Week			
	M Oct 18 W Oct 20	Linear Regression	Bishop Ch 3 (36 pages)	Prince Ch 7.1-7.2	
	M Oct 25 W Oct 27	Linear Classifiers	Bishop Ch 4.1-4.3 (34 pages)	Duda 5.1-5.8	
	M Nov 1 W Nov 3	Non-Linear Regression & Classification	Bishop Ch 6 (29 pages)	Prince Ch 7.3-7.4	
	M Nov 8 W Nov 10	Sparse Kernel Machines	Bishop 7.1 (20 pages)		
	M Nov 15 W Nov 17	Graphical Models: Introduction	Bishop Ch 8.1-8.3 (34 pages)		
	M Nov 22 W Nov 24	Graphical Models: Inference	Bishop Ch 8.4 (25 pages)		
	M Nov 29 W Dec 1	Graphical Models: Applications	Prince Ch 10-11 (56 pages)		
	M Dec 6 W Dec 8	Sampling Methods	Bishop Ch 11 (32 pages)		

Approximate Schedule

k	Date	Topic	Required Readings	Additional Readings	Application Paper
	M Sept 13 W Sept 15	Probability & Bayesian Inference Probability Distributions & Parametric Modeling	Bishop Ch 1.1-1.2.5 (29 pages) Bishop Ch 2.1-2.3 (skip 2.3.5) (43 pages)	Pearl Ch 1.4-1.6, 2 Howson & Urbach 199 Prince Ch 1-4 Duda Ch 3.1-3.5	
	M Sept 20 W Sept 22	Probability Distributions & Parametric Modeling (cntd.) Non-Parametric Modeling	Bishop Ch 2.5 (7 pages)	Duda Ch 4.1-4.5	Comaniciu & Meer 2002 (Mean Shift)
	M Sept 27 W Sept 29	Expectation Maximization	Prince Ch 5 (11 pages) Prince Ch 6.1-6.5, 6.8 (24 pages)	Bishop Ch 9	
	M Oct 4 W Oct 6	Linear Subspace Models	Prince Ch 6.6-6.7, 6.9 (12 pages) Bishop Ch 12 (40 pages)	Duda Ch 10.13-10.14	
	M Oct 11 W Oct 13	Reading Week			
	M Oct 18 W Oct 20	Linear Regression	Bishop Ch 3 (36 pages)	Prince Ch 7.1-7.2	
	M Oct 25 W Oct 27	Linear Classifiers	Bishop Ch 4.1-4.3 (34 pages)	Duda 5.1-5.8	
	M Nov 1 W Nov 3	Non-Linear Regression & Classification	Bishop Ch 6 (29 pages)	Prince Ch 7.3-7.4	
	M Nov 8 W Nov 10	Sparse Kernel Machines	Bishop 7.1 (20 pages)		
	M Nov 15 W Nov 17	Graphical Models: Introduction	Bishop Ch 8.1-8.3 (34 pages)		
	M Nov 22 W Nov 24	Graphical Models: Inference	Bishop Ch 8.4 (25 pages)		
	M Nov 29 W Dec 1	Graphical Models: Applications	Prince Ch 10-11 (56 pages)		
	M Dec 6 W Dec 8	Sampling Methods	Bishop Ch 11 (32 pages)		

PROBABILITY AND BAYESIAN INFERENCE

Credits

- Some of these slides were sourced and/or modified from:
  - Christopher Bishop, Microsoft UK
  - Simon Prince, UCL

INTRODUCTION: VISION AS BAYESIAN INFERENCE

J. Elder

CSE 6390/PSYC 6225 Computational Modeling of Visual Perception

Helmholtz

- Recognized ambiguity of images.
- Knowledge of scene properties and image formation used to resolve ambiguity and infer object properties.
- "Vision as Unconscious Inference"
- Can be formalized as Bayesian Decision Theory.

Hermann von Helmholtz

Helmholtz' Likelihood Principle

- Claim 1: The world is uncertain (to the observer)
- Claim 2: Vision is ill-posed
- Claim 3: Observers have evolved (are built) to perform valuable tasks well
- Conclusion: Vision is probabilistic inference

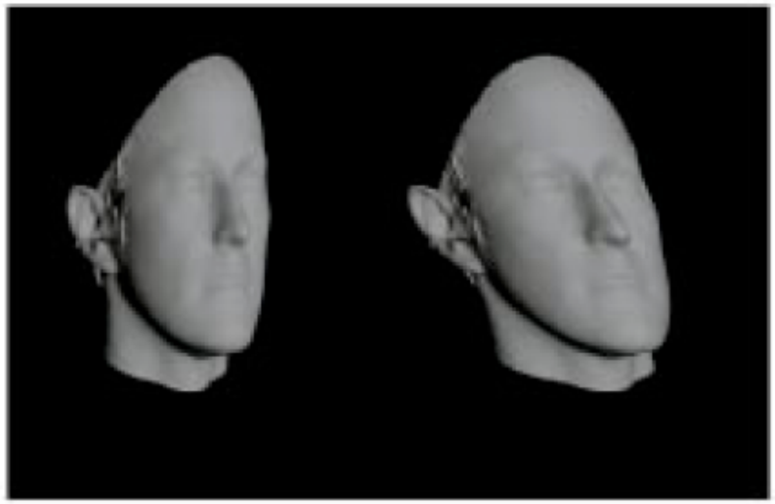
Vision is Ill-Posed

- Noise
  - "surface noise"
  - atmospheric effects
  - photon noise
  - neural noise
- Dimensionality
  - 1D → 2D
  - 2D → 3D
- Composition
  - e.g. Bilinear problem of colour (lightness) constancy:

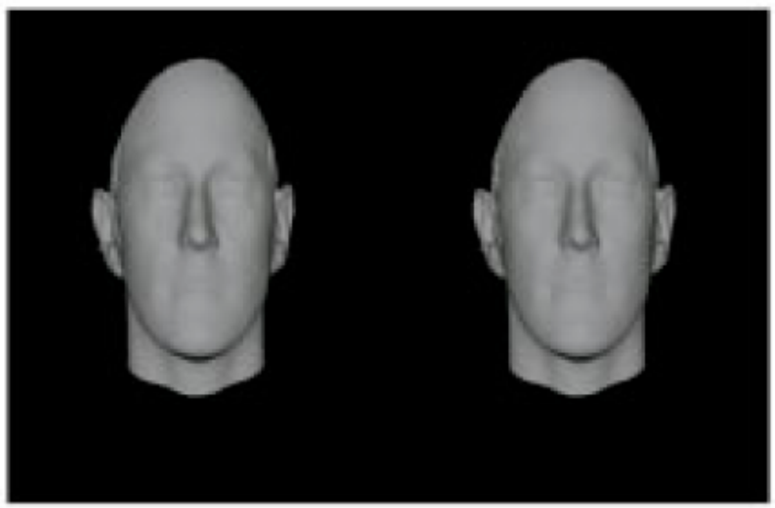
$$I = L \bullet R$$

Vision is Ill-Posed 2D→ 3D (N:1 Mapping)

Different Objects



Similar Images



From Kersten et al., 2004

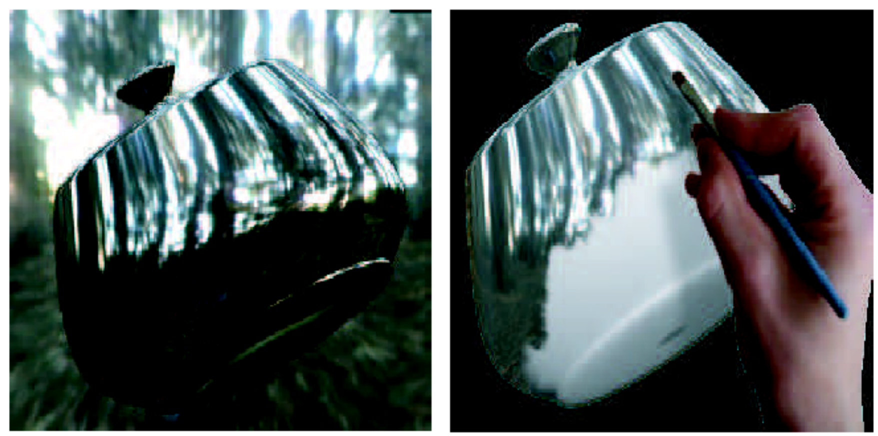
Vision is ill-posed (bilinearity of image)

C.

1:N Mapping

N:1 Mapping

D.



From Kersten et al., 2004

Julian Beever

Julian Beever

Julian Beever

Liu Bolin

Liu Bolin

Liu Bolin

Bayes' Rule

Scene Property To Be Inferred

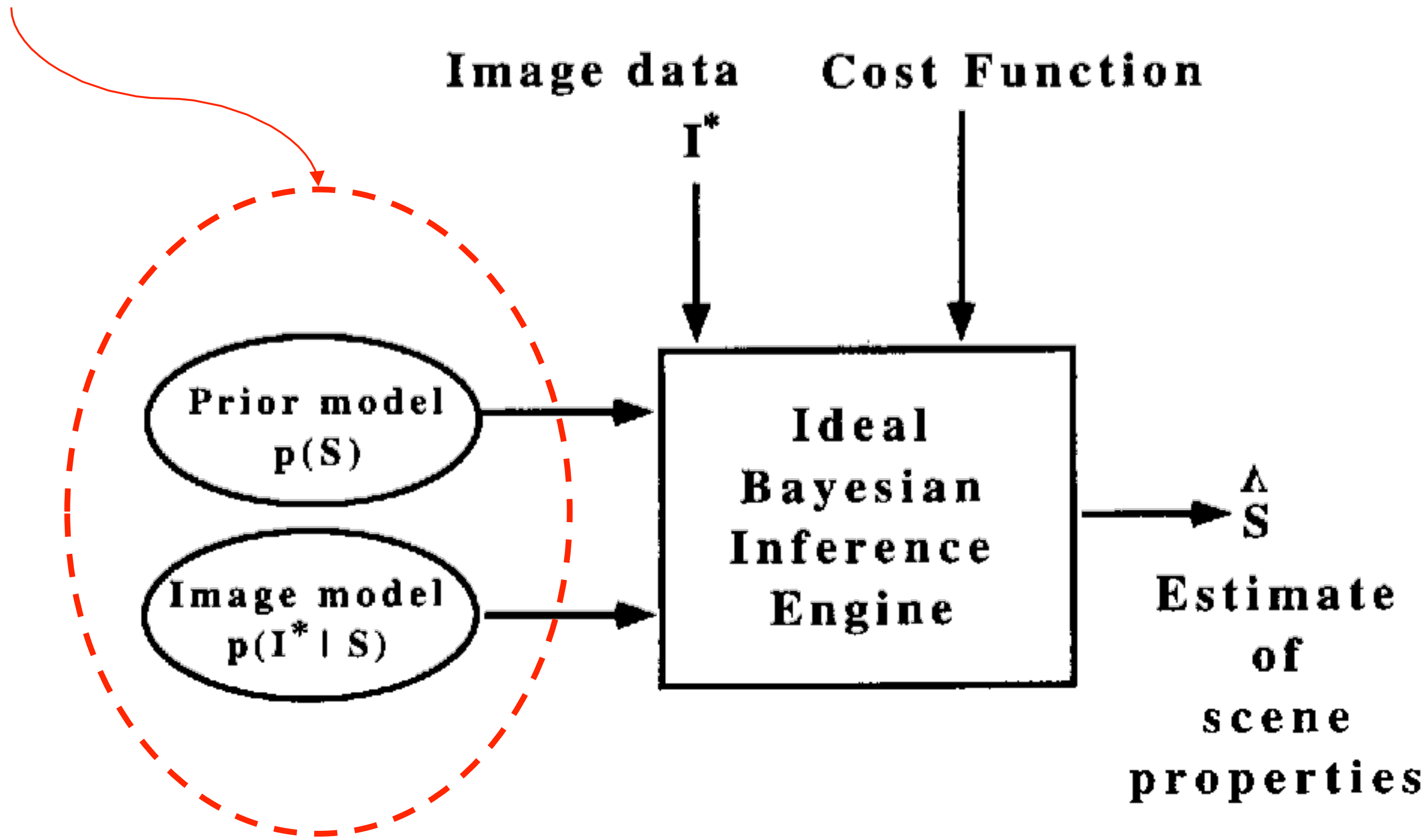
Image Observation

$$p(S \mid I) \propto p(I \mid S)\,p(S)$$

Posterior $\propto$ Likelihood $\times$ Prior

Generative Models and Ideal Observers

Generative Model: $p(S,I) = p(I \mid S)p(S)$



From Kersten et al., 2004

TOPIC 1.
PROBABILITY & BAYESIAN INFERENCE

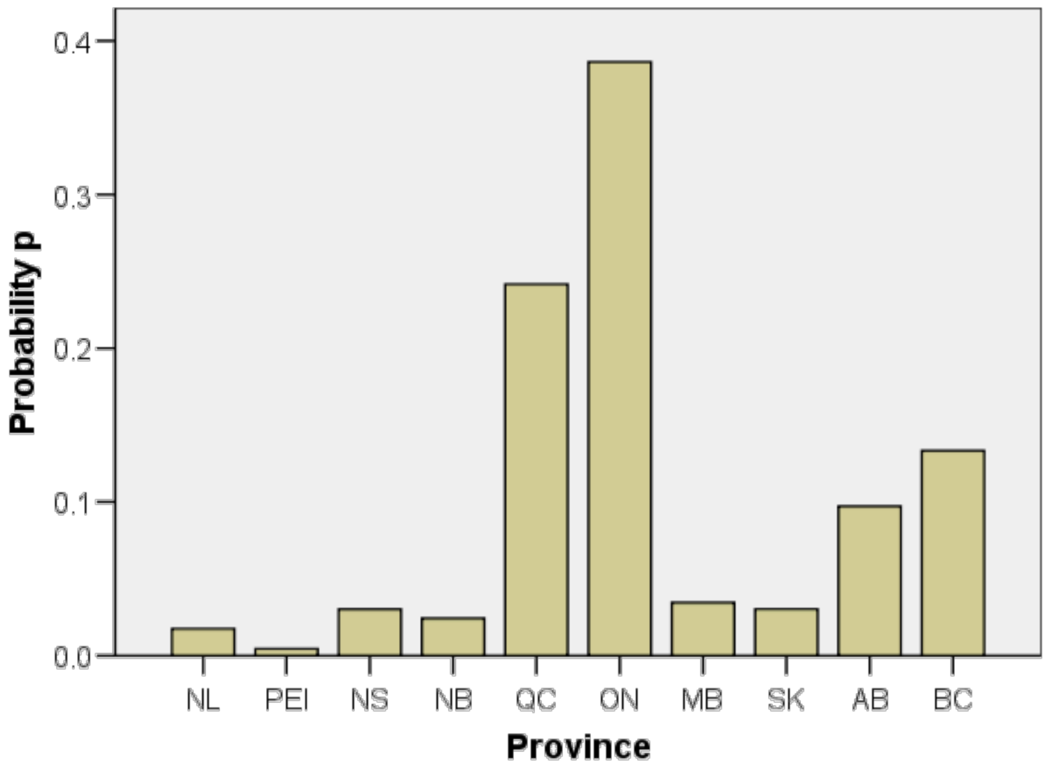
J. Elder

Random Variables

- A **random variable** is a variable whose value is uncertain.
- For example, the height of a randomly selected person in this class is a random variable – I won't know its value until the person is selected.
- Note that we are not completely uncertain about most random variables.
  - For example, we know that height will probably be in the 5'-6' range.
  - In addition, 5'6" is more likely than 5'0" or 6'0".
- The function that describes the probability of each possible value of the random variable is called a **probability distribution**.

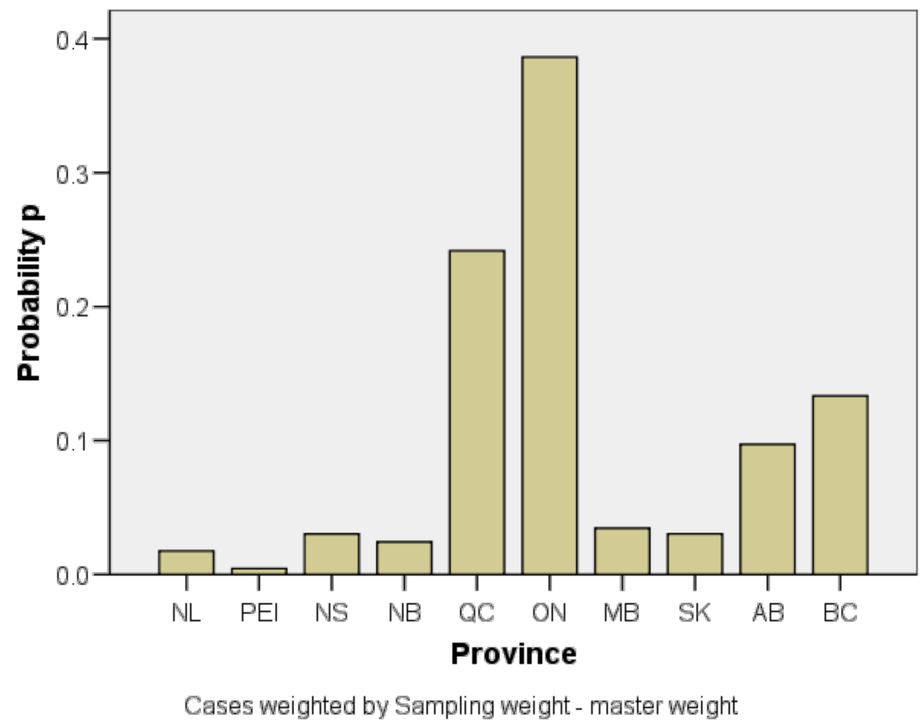
Probability Distributions

- For a **discrete** distribution, the probabilities over all possible values of the random variable must **sum** to 1.



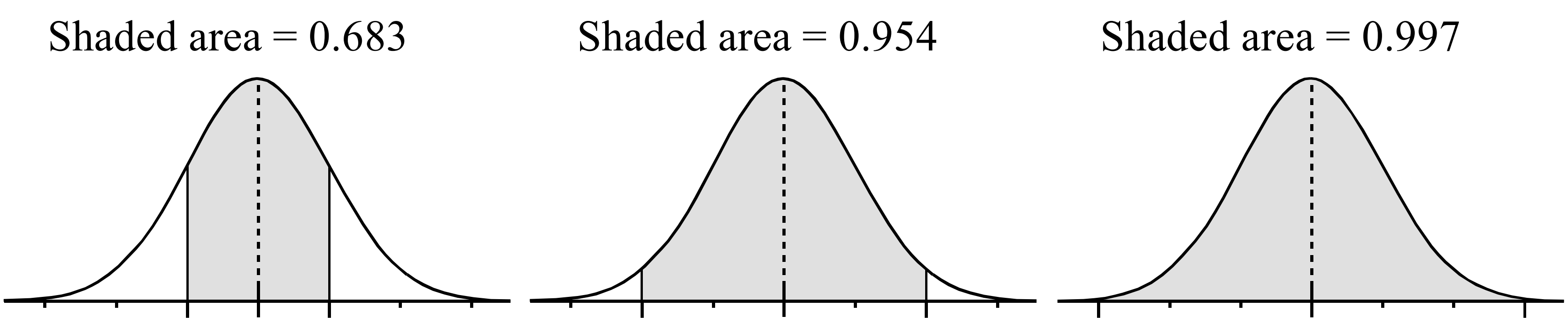
Probability Distributions

- For a **discrete** distribution, we can talk about the probability of a particular score occurring, e.g., p(Province = Ontario) = 0.36.
- We can also talk about the probability of any one of a subset of scores occurring, e.g., p(Province = Ontario or Quebec) = 0.50.
- In general, we refer to these occurrences as **events**.



Probability Distributions

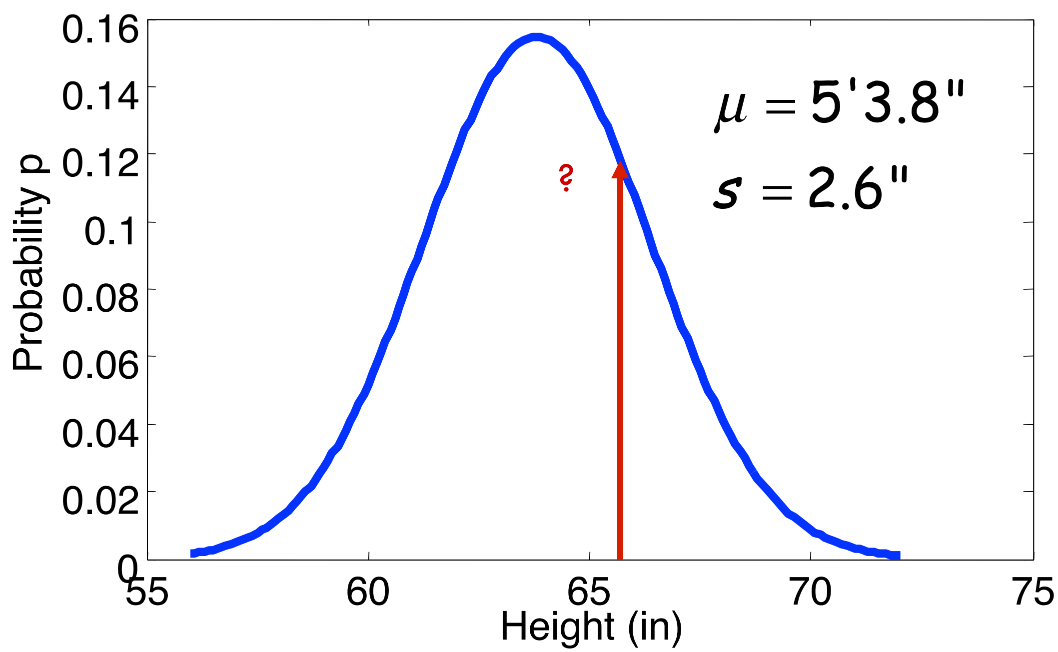
- For a **continuous** distribution, the probabilities over all possible values of the random variable must **integrate** to 1 (i.e., the area under the curve must be 1).
- Note that the height of a continuous distribution can exceed 1!



Continuous Distributions

- For continuous distributions, it **does not** make sense to talk about the probability of an exact score.
  - e.g., what is the probability that your height is exactly 65.485948467… inches?

Normal Approximation to probability distribution for height of Canadian females
(parameters from General Social Survey, 1991)

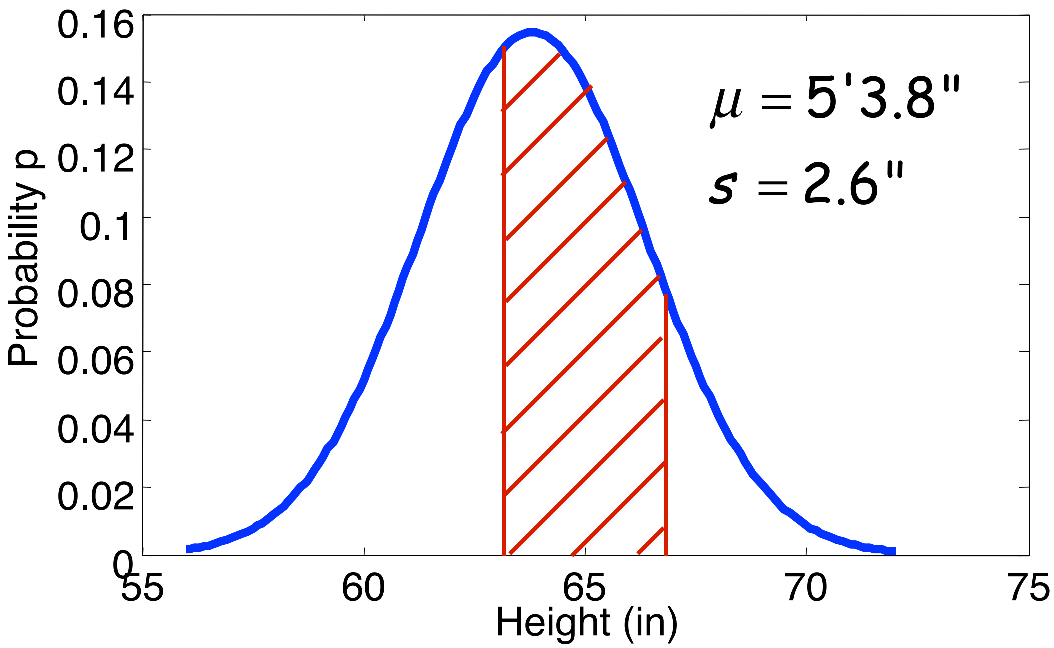


Continuous Distributions

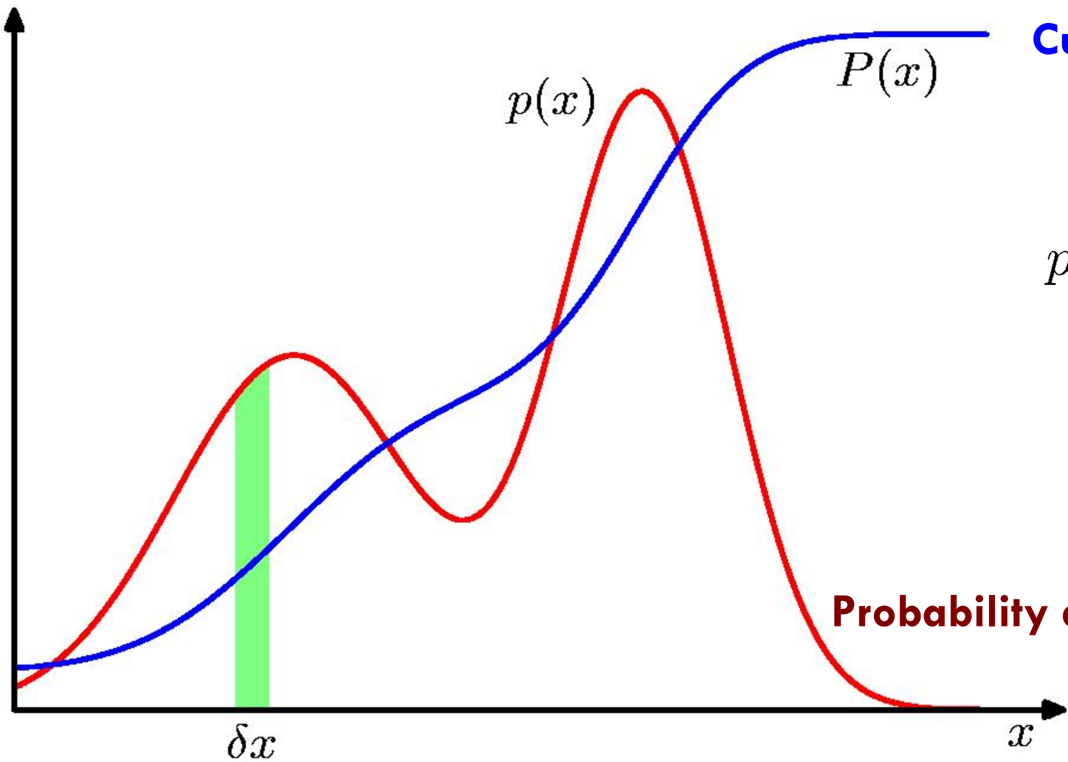
- It **does** make sense to talk about the probability of observing a score that falls within a certain range
  - e.g., what is the probability that you are between 5'3" and 5'7"?
  - e.g., what is the probability that you are less than 5'10"?

  Valid events

Normal Approximation to probability distribution for height of Canadian females
(parameters from General Social Survey, 1991)



Probability Densities



Cumulative distribution (CDF)

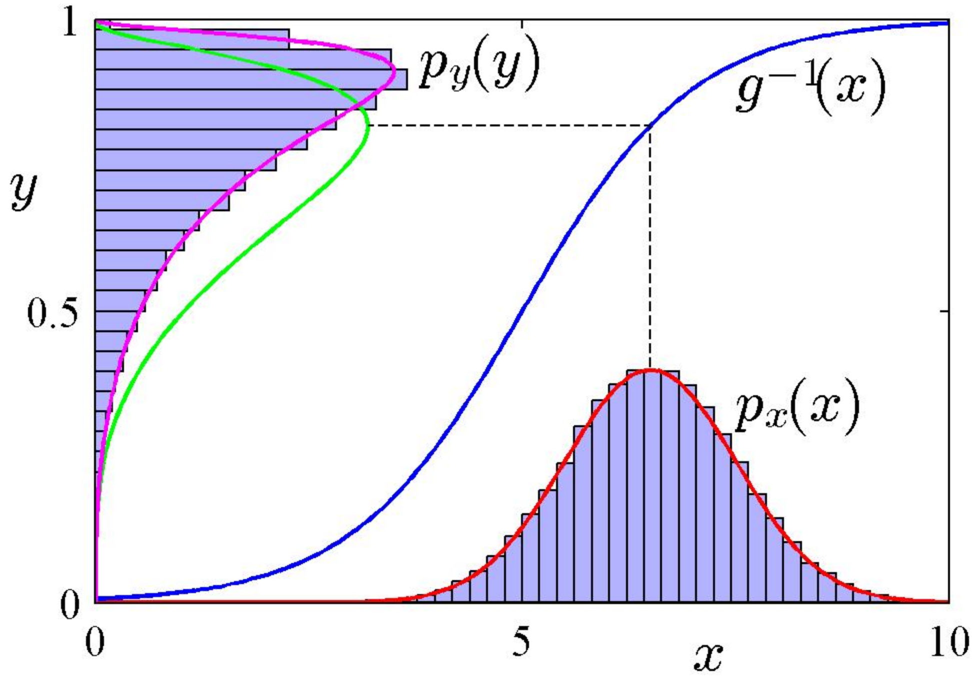
$$p(x \in (a, b)) = \int_a^b p(x) \, \mathrm{d}x$$

$$P(z) = \int_{-\infty}^{z} p(x) \, \mathrm{d}x$$

Probability density (PDF)

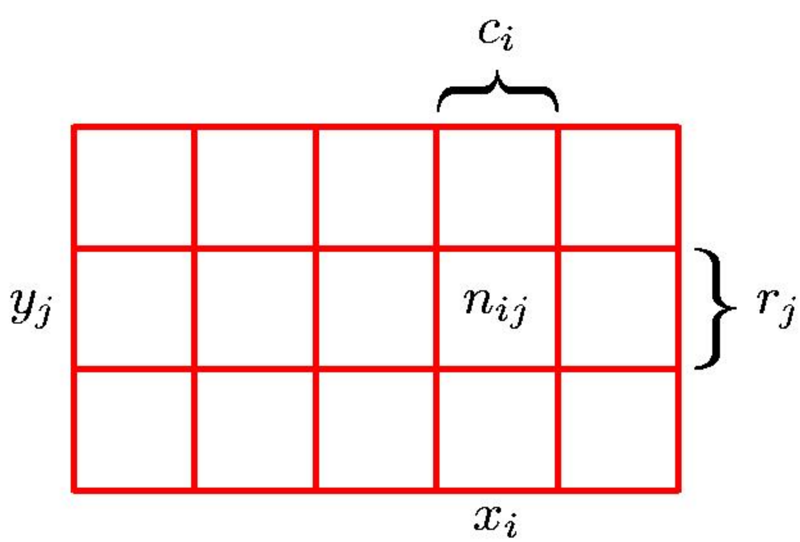
$$p(x) \geqslant 0 \qquad \int_{-\infty}^{\infty} p(x) \, \mathrm{d}x = 1$$

Transformed Densities



$$p_y(y) = p_x(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right|$$
$$= p_x(g(y))\,|g'(y)|$$

Joint Distributions



Marginal Probability

$$p(X = x_i) = \frac{c_i}{N}.$$

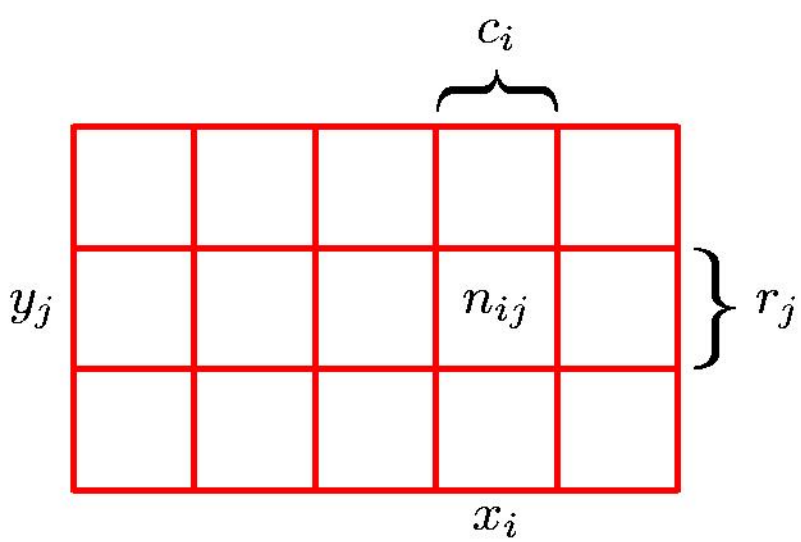
Joint Probability

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Conditional Probability

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

Joint Distributions



Sum Rule

$$p(X = x_i) = \frac{c_i}{N} = \frac{1}{N} \sum_{j=1}^{L} n_{ij}$$

$$= \sum_{j=1}^{L} p(X = x_i, Y = y_j)$$

Product Rule

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N}$$

$$= p(Y = y_j | X = x_i) p(X = x_i)$$

Joint Distributions: The Rules of Probability

- Sum Rule $$p(X) = \sum_Y p(X, Y)$$
- Product Rule $$p(X, Y) = p(Y|X)p(X)$$

END OF LECTURE 1
SEPT 13, 2010

Application Papers

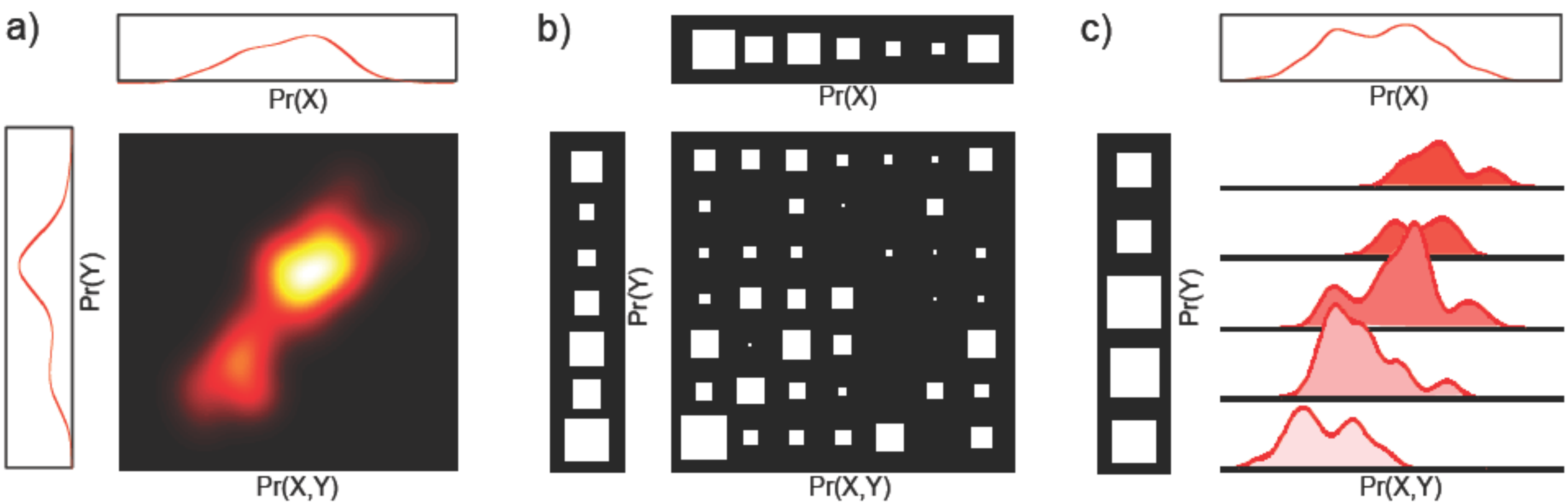
k	Date	Topic	Required Readings	Additional Readings	Application Papers
	M Sept 13 W Sept 15	Probability & Bayesian Inference Probability Distributions & Parametric Modeling	Bishop Ch 1.1-1.2.5 (29 pages) Bishop Ch 2.1-2.3 (skip 2.3.5) (43 pages)	Pearl Ch 1.4-1.6, 2 Howson & Urbach 199[illegible] Prince Ch 1-4 Duda Ch 3.1-3.5	
	M Sept 20 W Sept 22	Probability Distributions & Parametric Modeling (cntd.) Non-Parametric Modeling	Bishop Ch 2.5 (7 pages)	Duda Ch 4.1-4.5	Comaniciu & Meer 2002 (Mean Shift)
	M Sept 27 W Sept 29	Expectation Maximization	Prince Ch 5 (11 pages) Prince Ch 6.1-6.5, 6.8 (24 pages)	Bishop Ch 9	Stauffer & Grimson 1998 Weber & Perona 2000
	M Oct 4 W Oct 6	Subspace Models	Prince Ch 6.6-6.7, 6.9 (12 pages) Bishop Ch 12 (40 pages)	Duda Ch 10.13-10.14	Tenenbaum et al 2000 Roweis & Saul 2000
	M Oct 11 W Oct 13	Reading Week			
	M Oct 18 W Oct 20	Linear Regression	Bishop Ch 3 (36 pages)	Prince Ch 7.1-7.2	Moghaddam 2002 Cremers 2003
	M Oct 25 W Oct 27	Linear Classifiers	Bishop Ch 4.1-4.3 (34 pages)	Duda 5.1-5.8	Belhumeur et al 1997 Martin et al 2004
	M Nov 1 W Nov 3	Kernel Methods	Bishop Ch 6 (29 pages)	Prince Ch 7.3-7.4	Toyama & Blake 2001 Grochow et al 2004
	M Nov 8 W Nov 10	Sparse Kernel Machines Combining Models	Bishop 7.1 (20 pages) Bishop Ch 14 (20 pages)		Agarwal & Triggs 2006 Zhang et al 2007
	M Nov 15 W Nov 17	Graphical Models: Introduction	Bishop Ch 8.1-8.3 (34 pages)		Freeman et al 2000 Shi & Malik 2000
	M Nov 22 W Nov 24	Graphical Models: Inference	Bishop Ch 8.4 (25 pages)		Boykov & Funka-Lea 2006 He et al 2004
	M Nov 29 W Dec 1	Graphical Models: Applications	Prince Ch 10-11 (56 pages)		Frey & Jojic 2005 Szeliski et al 2008
	M Dec 6 W Dec 8	Sampling Methods	Bishop Ch 11 (32 pages)		Zhu 1999 Yuille & Kersten 2006

Marginalization

We can recover probability distribution of any variable in a joint distribution by integrating (or summing) over the other variables

$$Pr(X) = \int Pr(X,Y)dY$$
$$Pr(Y) = \int Pr(X,Y)dX$$

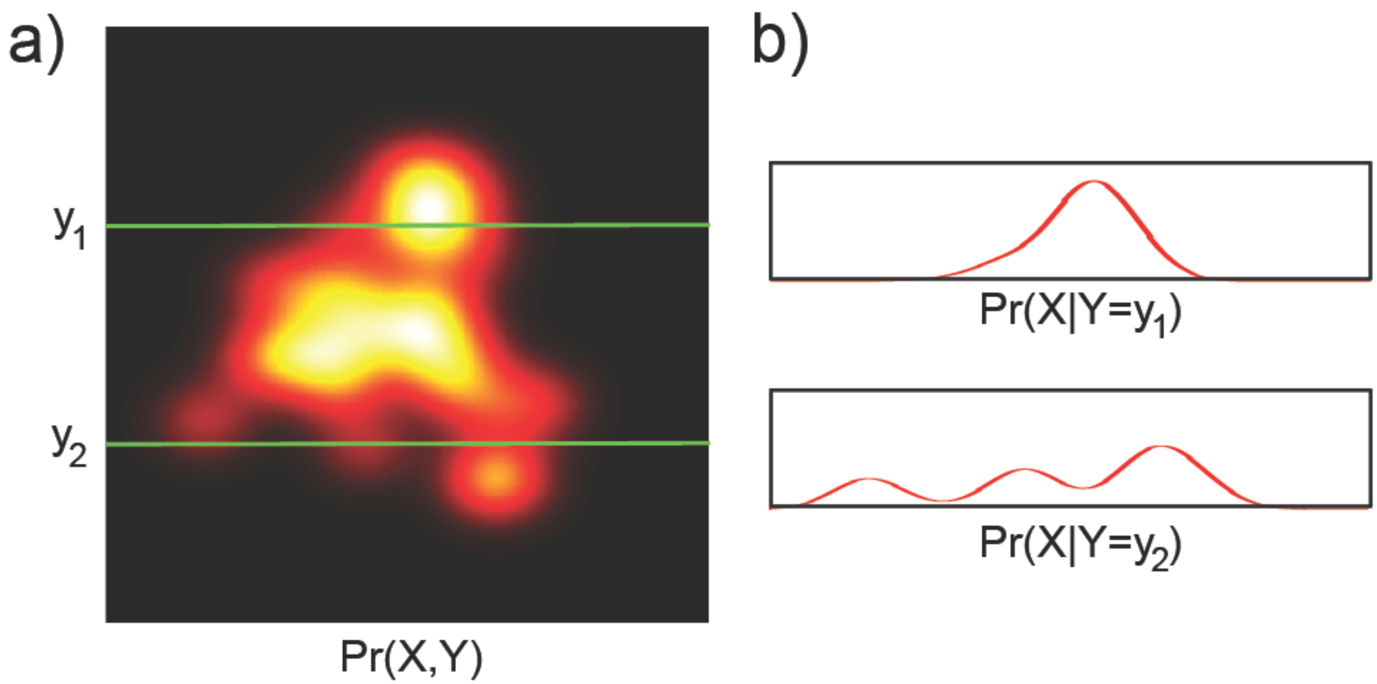
$$Pr(X,Y) = \sum_W \sum_Z Pr(W,X,Y,Z)$$



Conditional Probability

- Conditional probability of X given that Y=y* is relative propensity of variable X to take different outcomes given that Y is fixed to be equal to y*

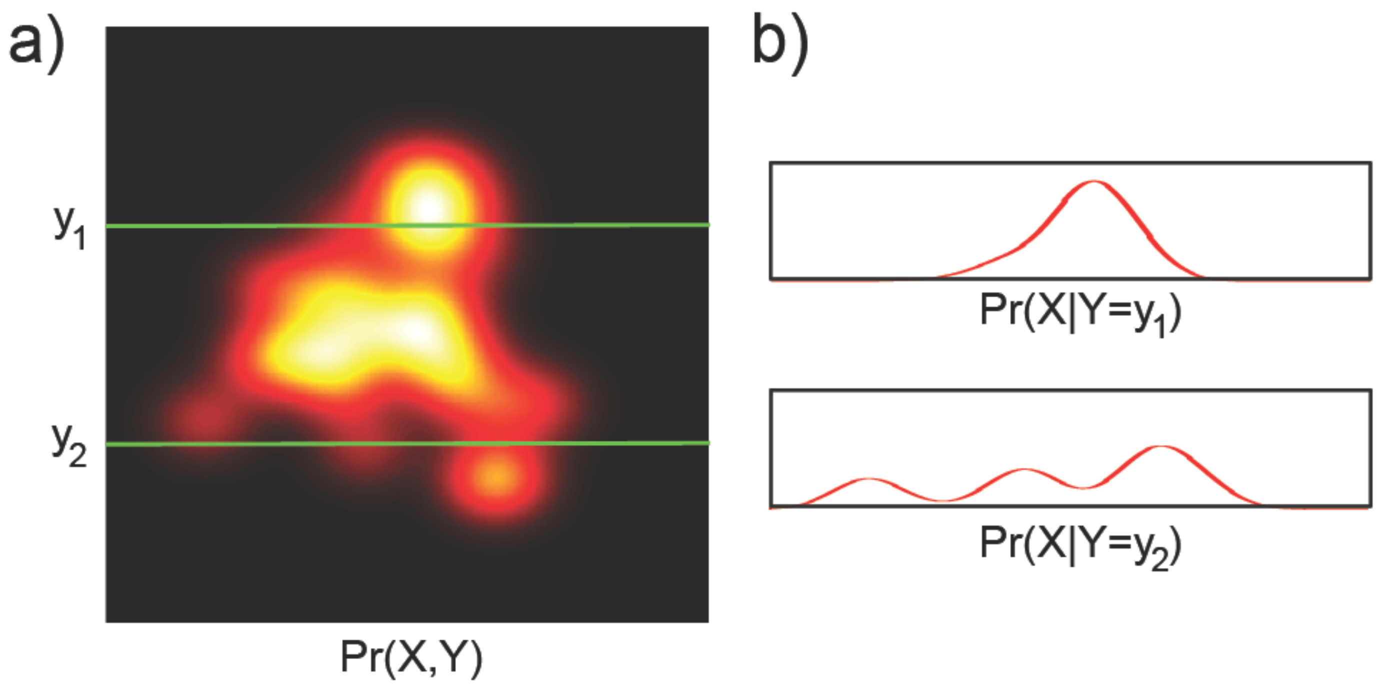
- Written as Pr(X|Y=y*)



Conditional Probability

- Conditional probability can be extracted from joint probability
- Extract appropriate slice and normalize

$$Pr(X|Y=y^*) = \frac{Pr(X,Y=y^*)}{\int(Pr(X,Y=y^*)dX} = \frac{Pr(X,Y=y^*)}{Pr(Y=y^*)}$$



Conditional Probability

$$Pr(X|Y = y^*) = \frac{Pr(X, Y = y^*)}{\int (Pr(X, Y = y^*)dX} = \frac{Pr(X, Y = y^*)}{Pr(Y = y^*)}$$

- More usually written in compact form

$$Pr(X|Y) = \frac{Pr(X, Y)}{Pr(Y)}$$

- Can be re-arranged to give

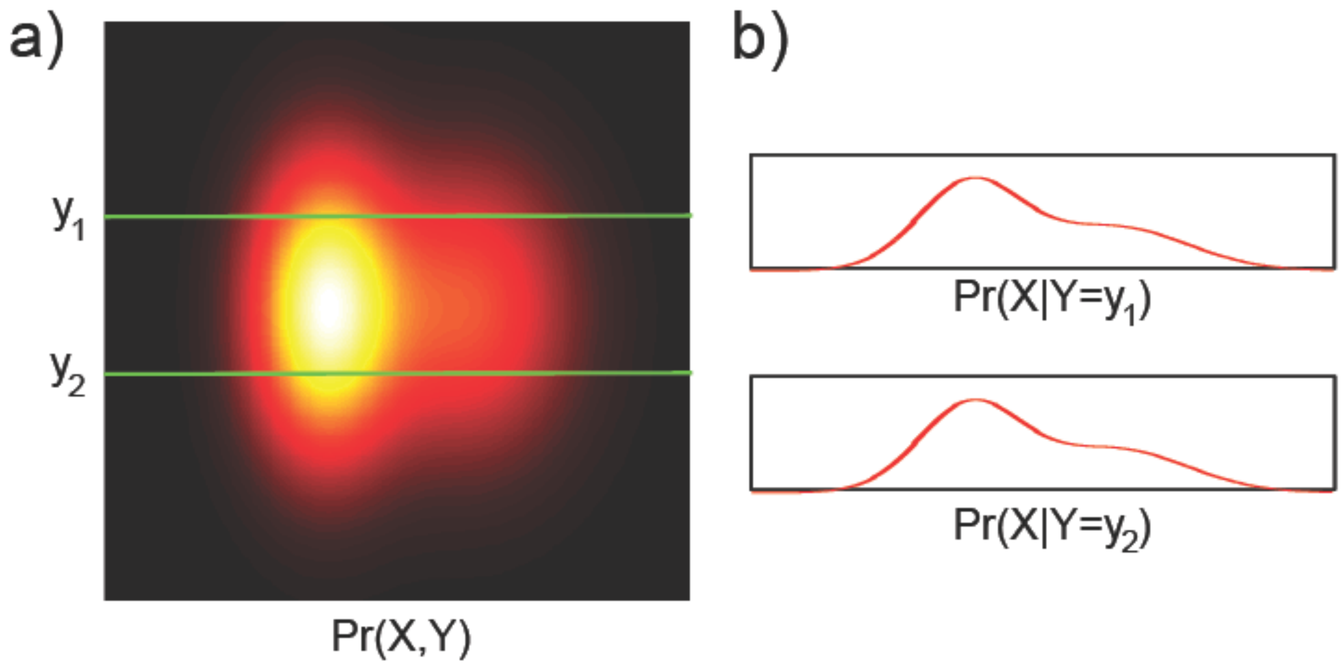
$$Pr(X, Y) = Pr(X|Y)Pr(Y)$$

$$Pr(X, Y) = Pr(Y|X)Pr(X)$$

Independence

- If two variables X and Y are independent then variable X tells us nothing about variable Y (and vice-versa)

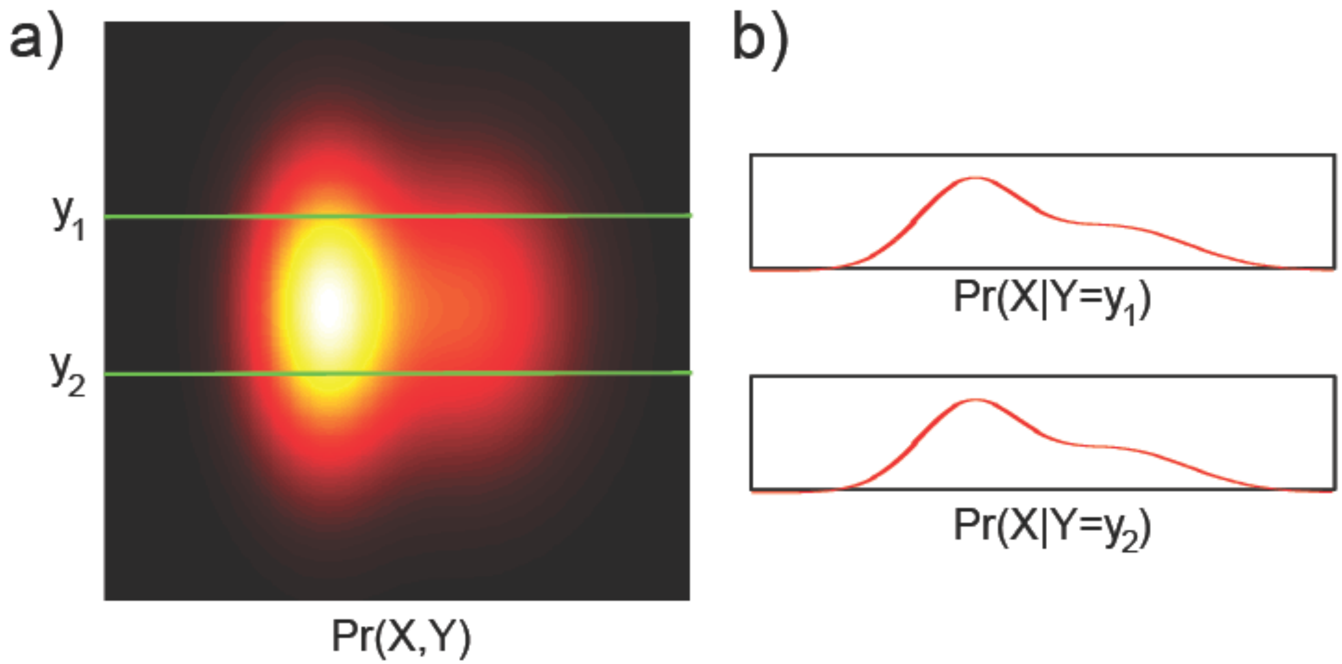
$$Pr(X|Y) = Pr(X)$$
$$Pr(Y|X) = Pr(Y)$$



Independence

- When variables are independent, the joint factorizes into a product of the marginals:

$$
\begin{aligned}
Pr(X,Y) &= Pr(X|Y)Pr(Y) \\
&= Pr(X)Pr(Y)
\end{aligned}
$$



Bayes' Rule

From before:

$$Pr(X,Y) = Pr(X|Y)Pr(Y)$$
$$Pr(X,Y) = Pr(Y|X)Pr(X)$$

Combining:

$$Pr(Y|X)Pr(X) = Pr(X|Y)Pr(Y)$$

Re-arranging:

$$\begin{aligned} Pr(Y|X) &= \frac{Pr(X|Y)Pr(Y)}{Pr(X)} \\ &= \frac{Pr(X|Y)Pr(Y)}{\int Pr(X,Y)dY} \\ &= \frac{Pr(X|Y)Pr(Y)}{\int Pr(X|Y)Pr(Y)dY} \end{aligned}$$

Bayes' Rule Terminology

Likelihood – propensity for observing a certain value of X given a certain value of Y

Prior – what we know about y before seeing x

$$Pr(Y|X) = \frac{Pr(X|Y)Pr(Y)}{Pr(X)}$$

Posterior – what we know about y after seeing x

Evidence –a constant to ensure that the left hand side is a valid distribution

Expectations

$$\mathbb{E}[f] = \sum_x p(x) f(x)$$

$$\mathbb{E}[f] = \int p(x) f(x) \, \mathrm{d}x$$

$$\mathbb{E}_x[f|y] = \sum_x p(x|y) f(x)$$

Conditional Expectation (discrete)

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^{N} f(x_n)$$

Approximate Expectation (discrete and continuous)

Variances and Covariances

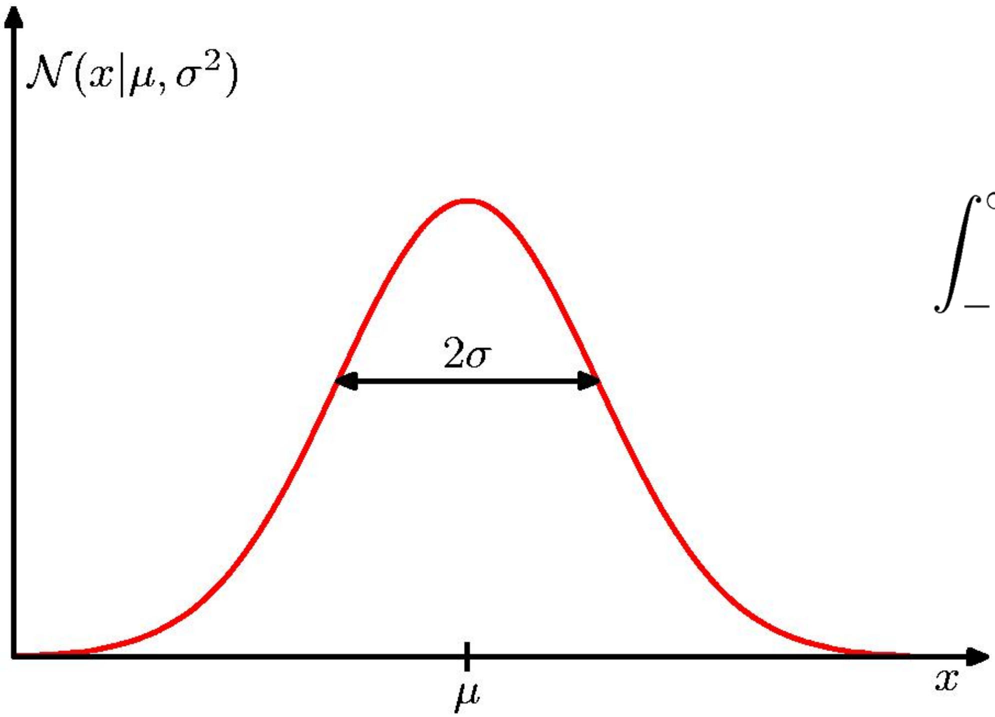
$$\mathrm{var}[f] = \mathbb{E}\left[\left(f(x) - \mathbb{E}[f(x)]\right)^2\right] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$

$$\begin{aligned}\mathrm{cov}[x, y] &= \mathbb{E}_{x,y}\left[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}\right] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]\end{aligned}$$

$$\begin{aligned}\mathrm{cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\}\{\mathbf{y}^{\mathrm{T}} - \mathbb{E}[\mathbf{y}^{\mathrm{T}}]\}\right] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^{\mathrm{T}}]\end{aligned}$$

The Gaussian Distribution

$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$



$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right)\,\mathrm{d}x = 1$$

Gaussian Mean and Variance

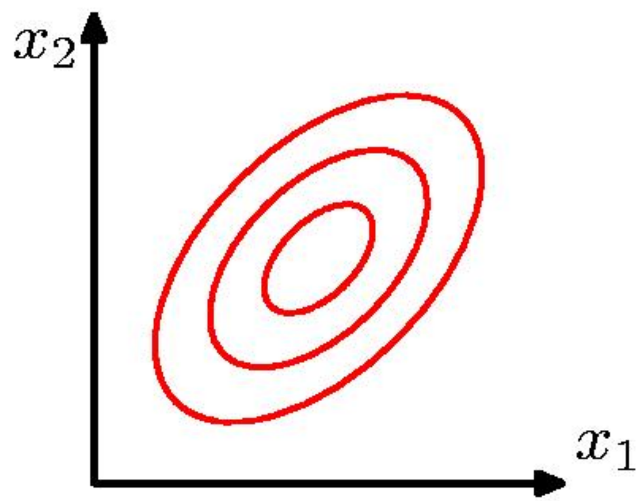
$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) x \, \mathrm{d}x = \mu$$

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) x^2 \, \mathrm{d}x = \mu^2 + \sigma^2$$

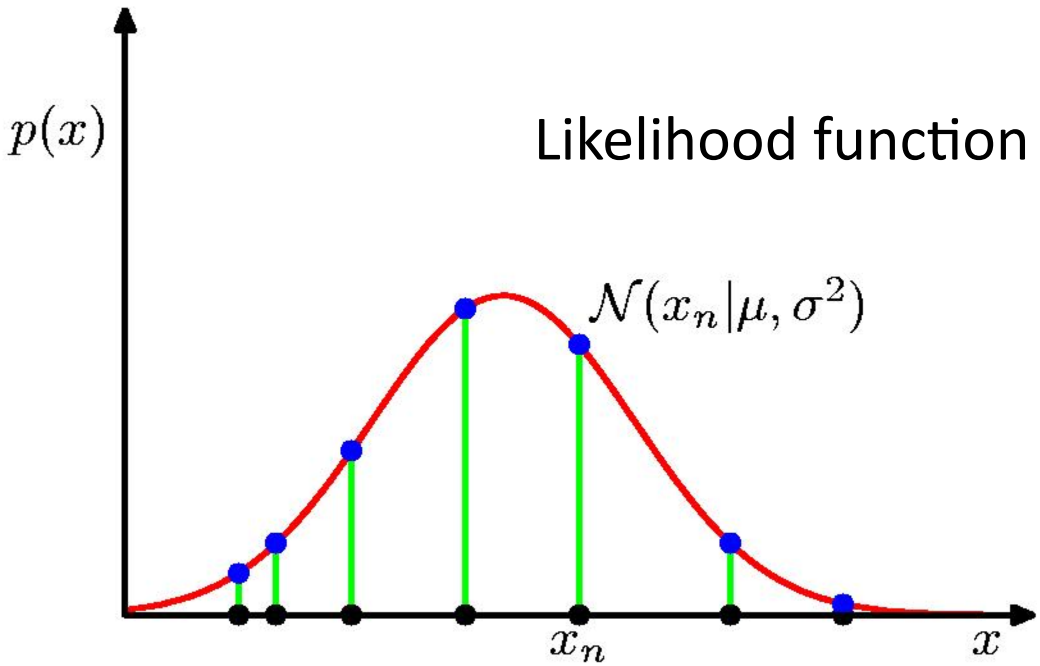
$$\operatorname{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$$

The Multivariate Gaussian

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$



Gaussian Parameter Estimation



$$p(\mathbf{x}|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(x_n|\mu, \sigma^2\right)$$

Maximum (Log) Likelihood

$$\ln p\left(\mathbf{x}|\mu,\sigma^2\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

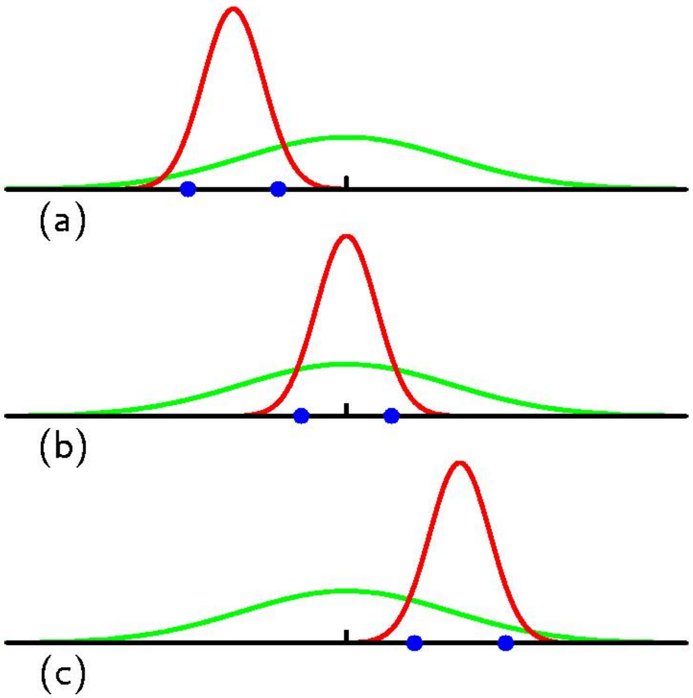
$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}x_n \qquad \sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2$$

Maximum likelihood estimates of normal parameters

$$\mathbb{E}[\mu_{\mathrm{ML}}] = \mu$$

$$\mathbb{E}[\sigma^2_{\mathrm{ML}}] = \left(\frac{N-1}{N}\right)\sigma^2$$

$$\begin{aligned}\widetilde{\sigma}^2 &= \frac{N}{N-1}\sigma^2_{\mathrm{ML}} \\ &= \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2\end{aligned}$$

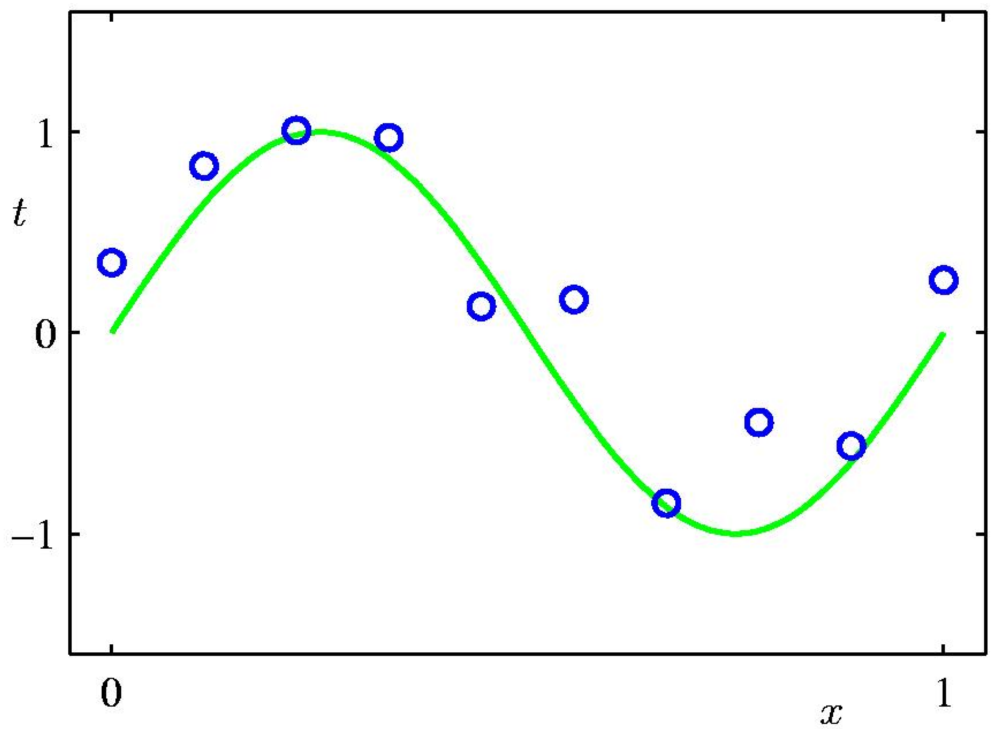


APPLYING PROBABILITY THEORY TO INFERENCE

J. Elder

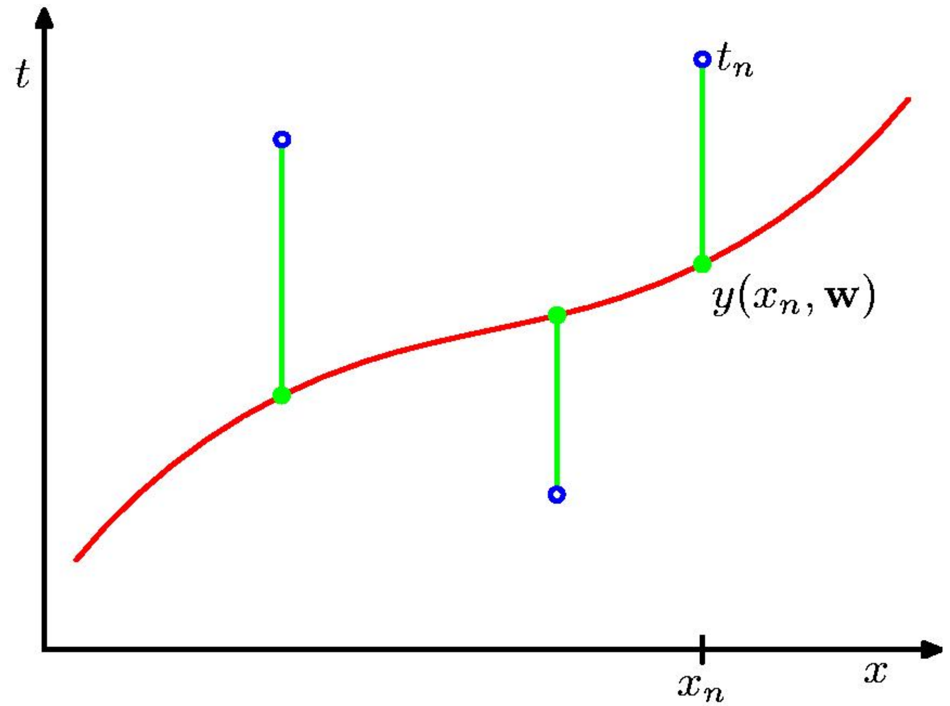
CSE 6390/PSYC 6225 Computational Modeling of Visual Perception

Polynomial Curve Fitting



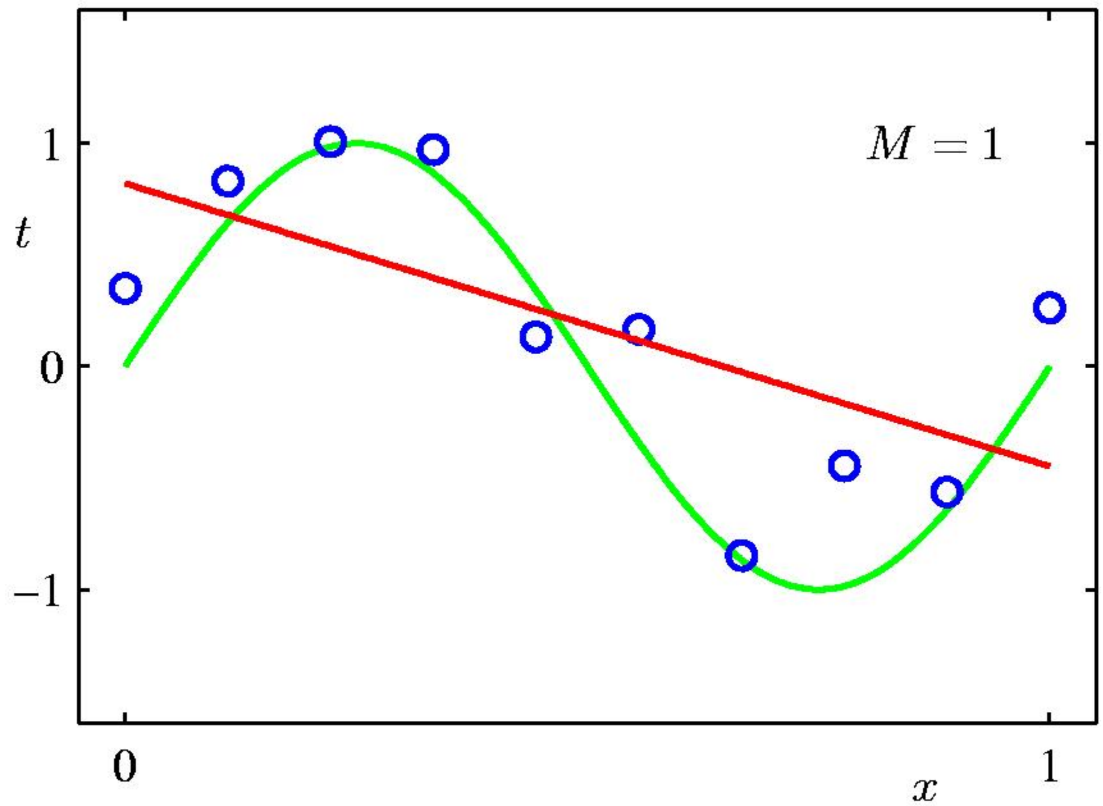
$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Sum-of-Squares Error Function

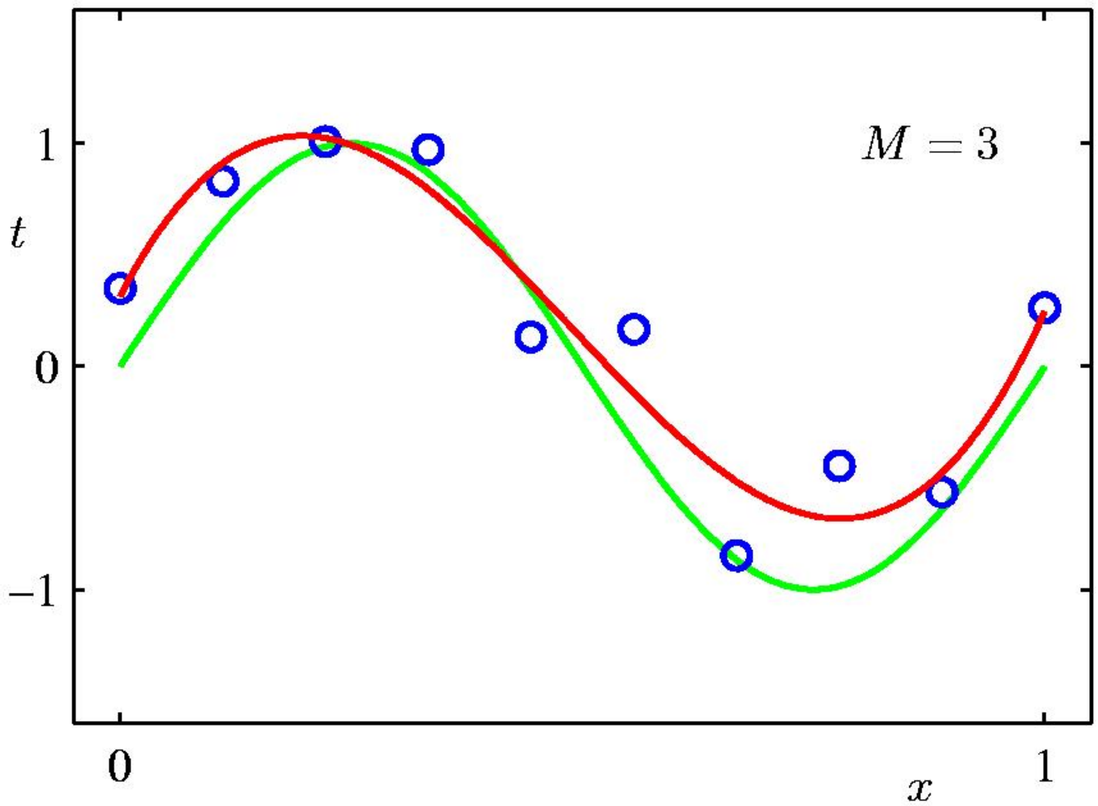


$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2$$

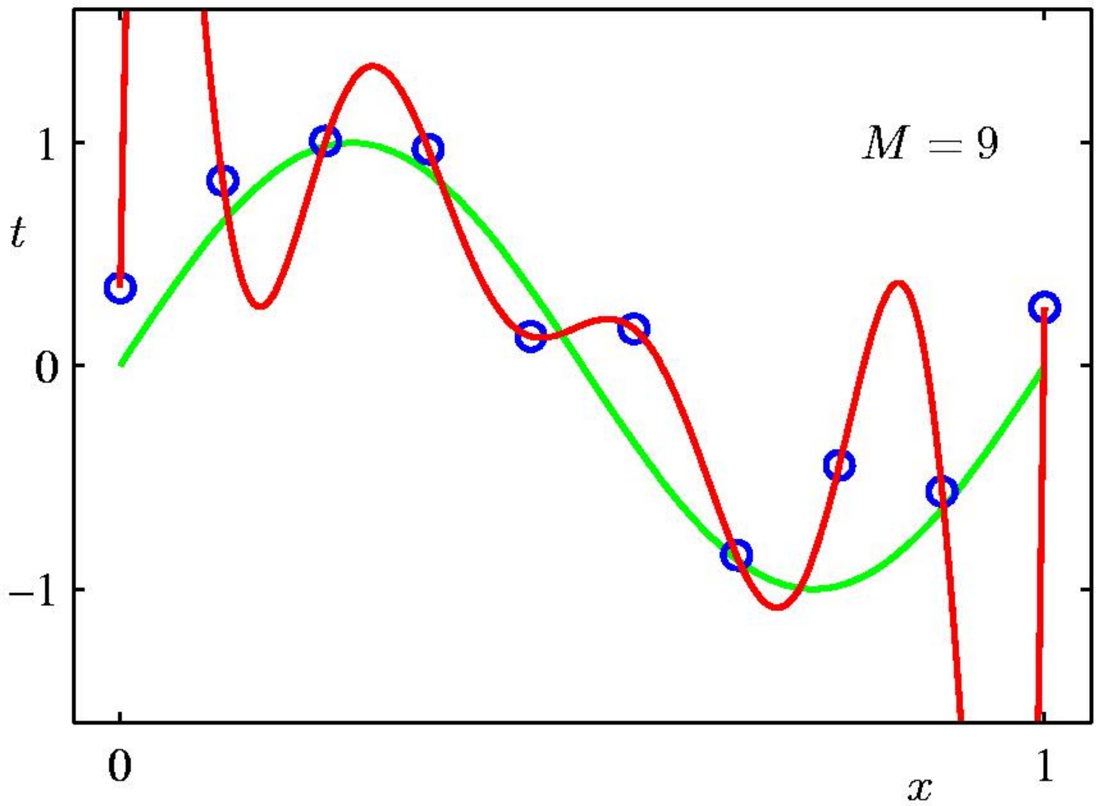
1^{st} Order Polynomial



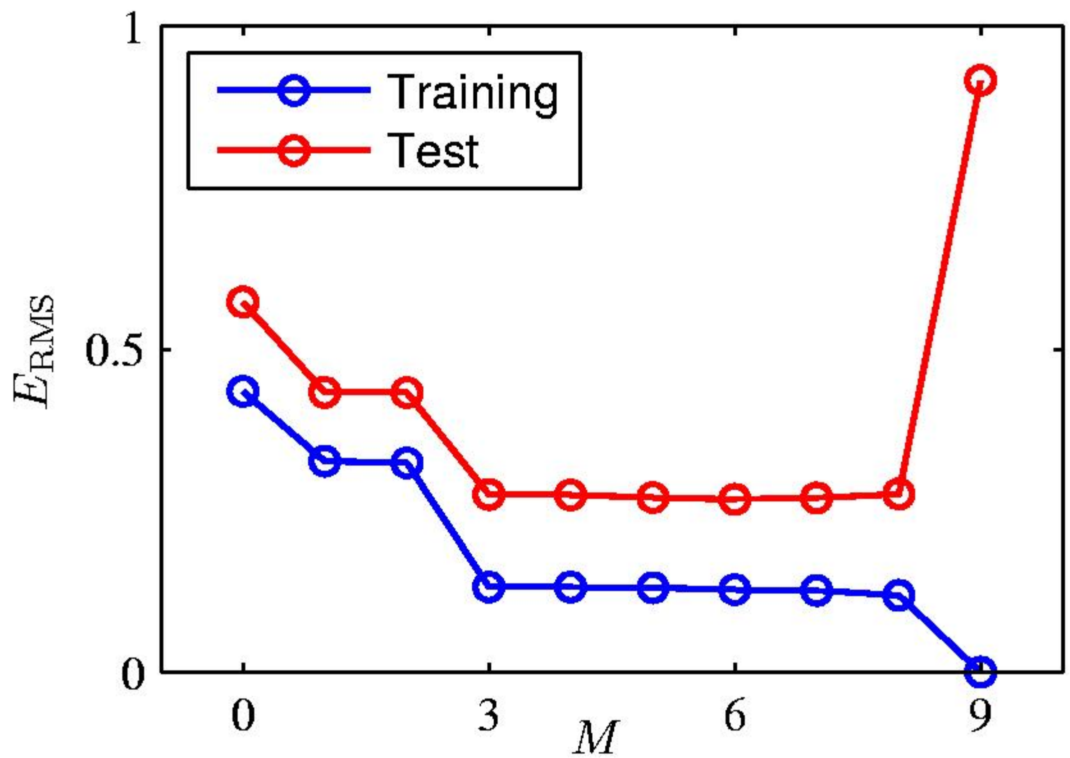
3rd Order Polynomial



9th Order Polynomial



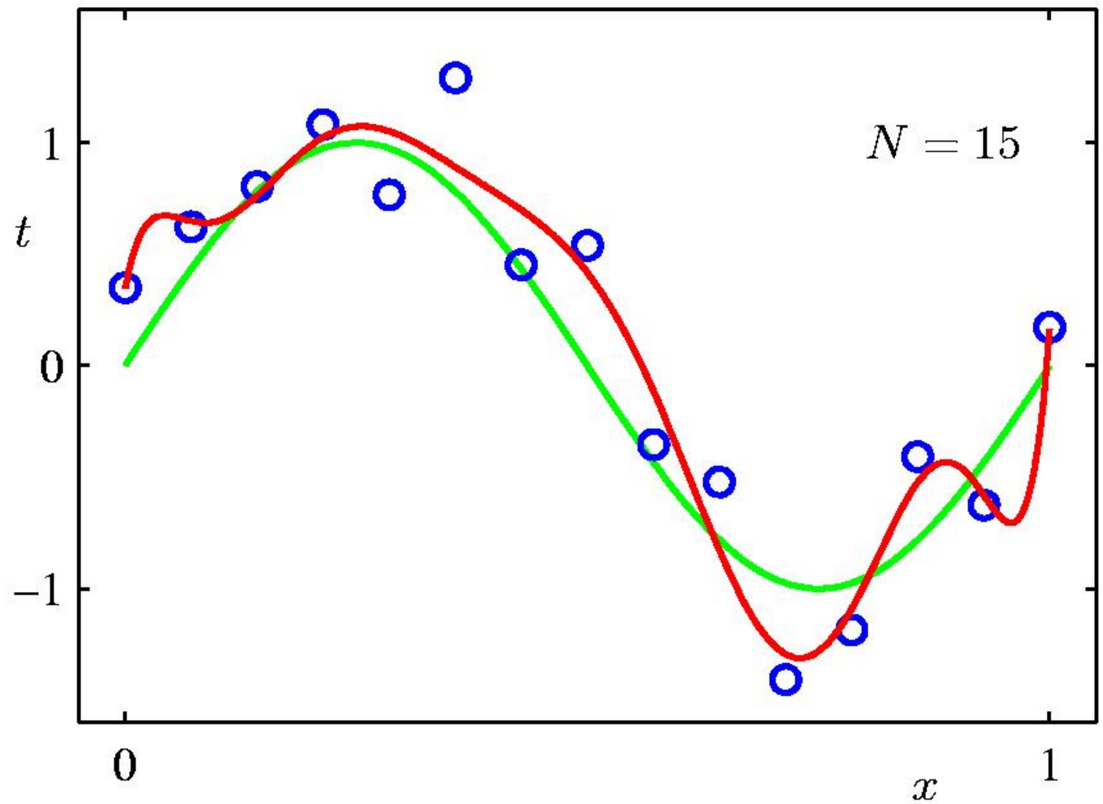
Over-fitting



Root-Mean-Square (RMS) Error: $E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^{\star})/N}$

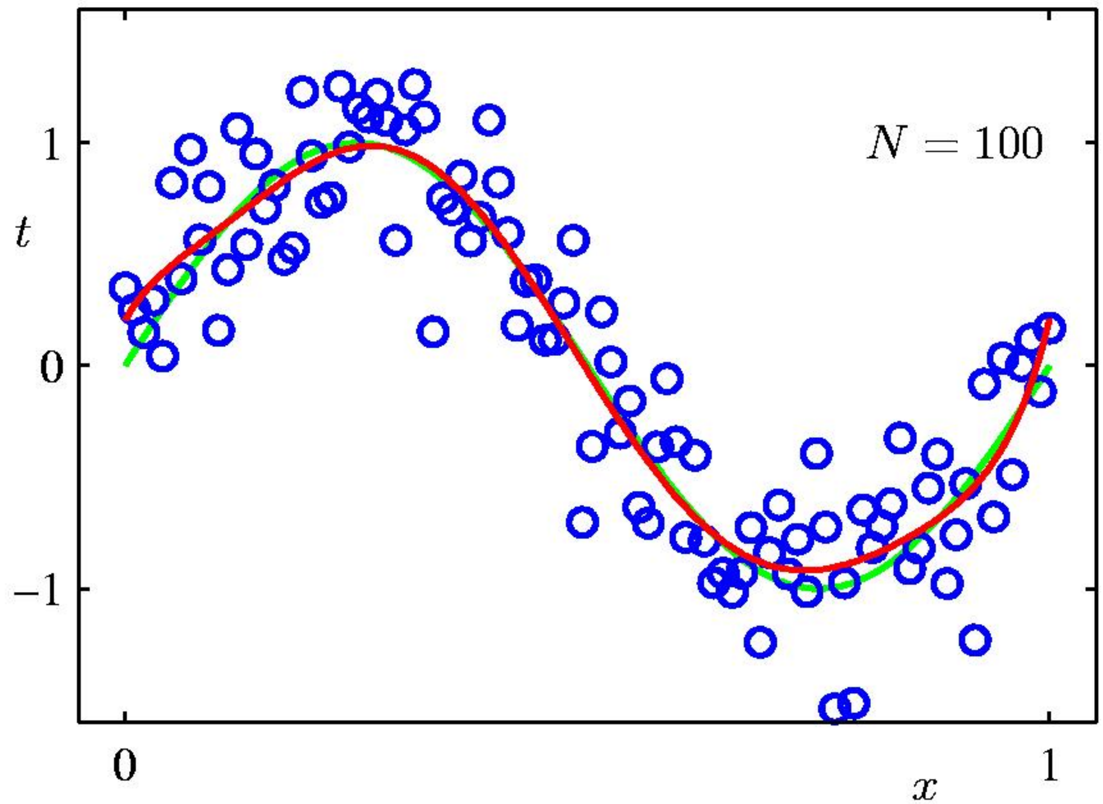
Overfitting and Sample Size

9th Order Polynomial



Overfitting and Sample Size

9th Order Polynomial



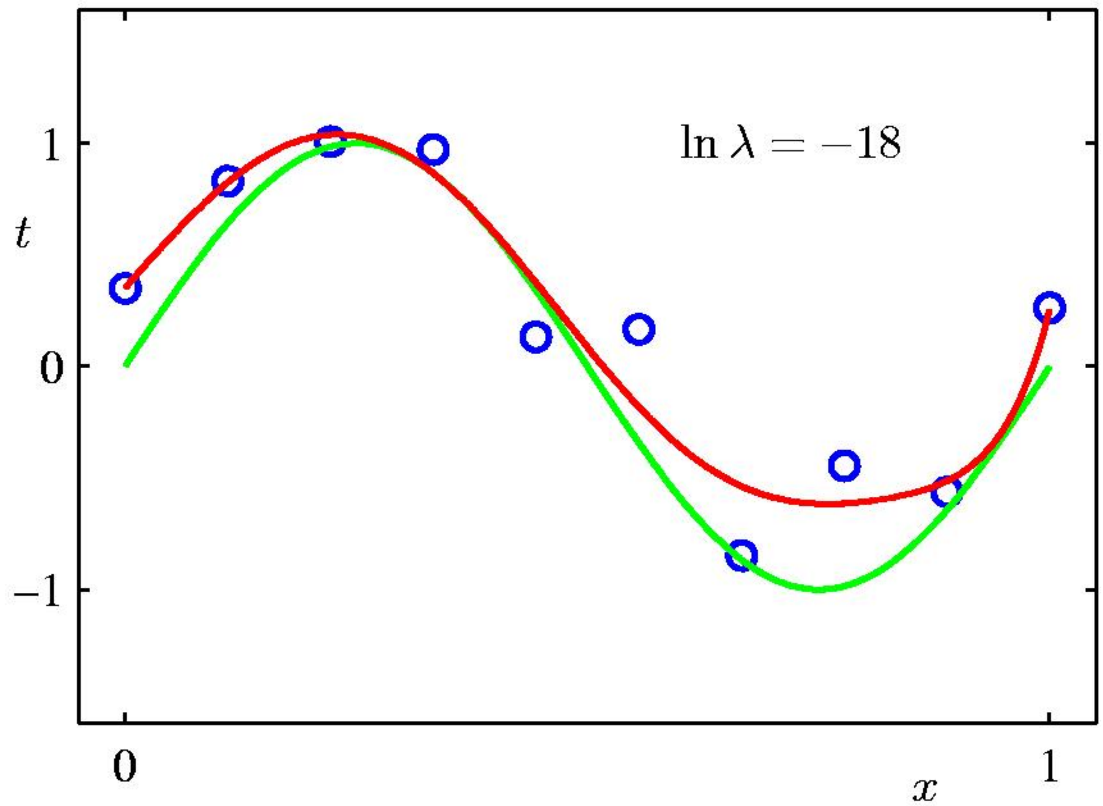
Regularization

- Penalize large coefficient values

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

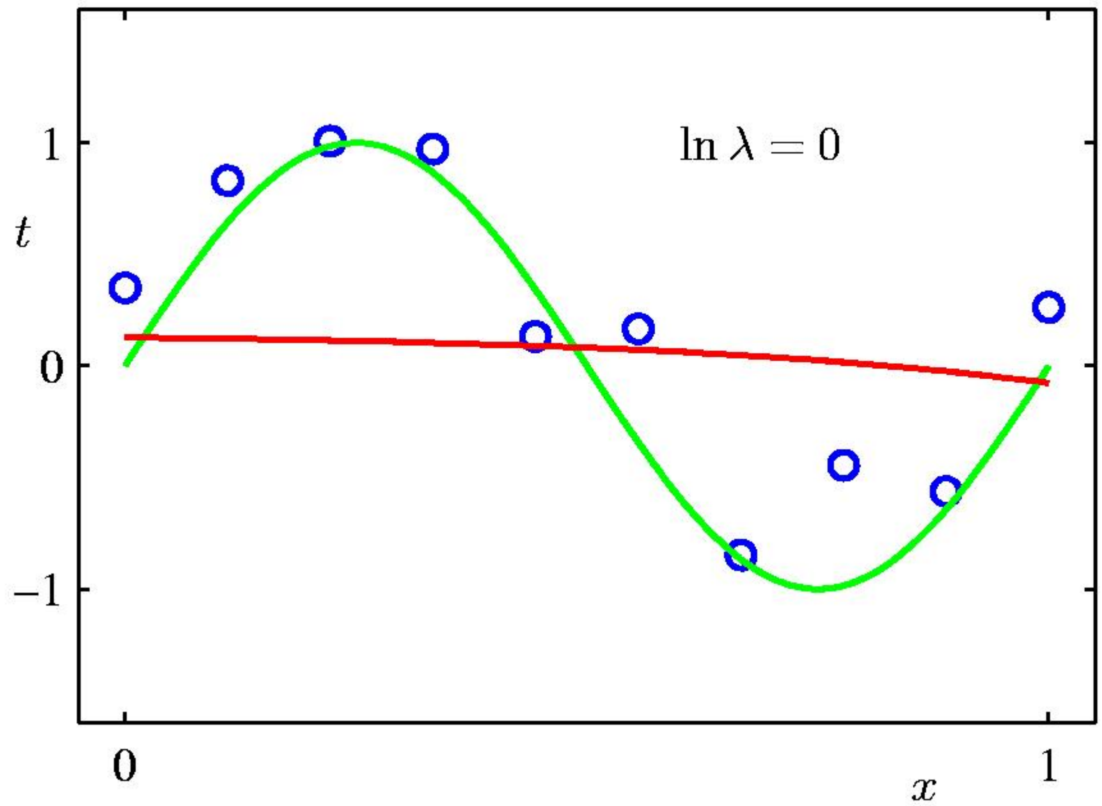
Regularization

9th Order Polynomial



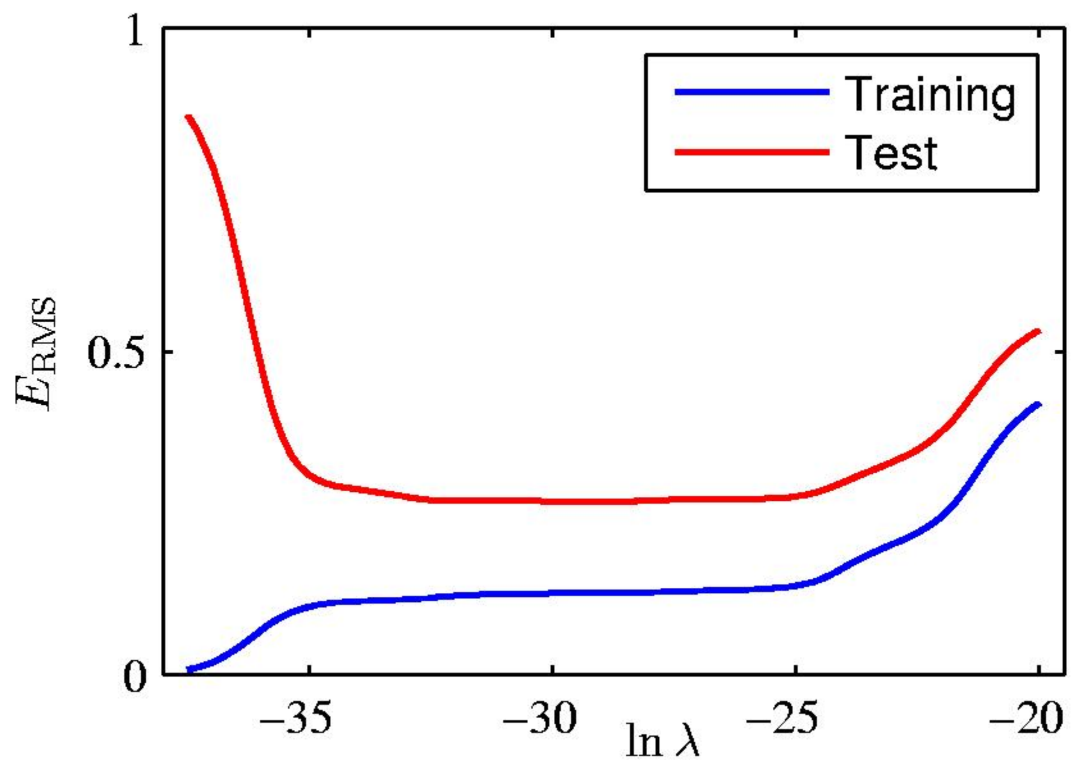
Regularization

9th Order Polynomial

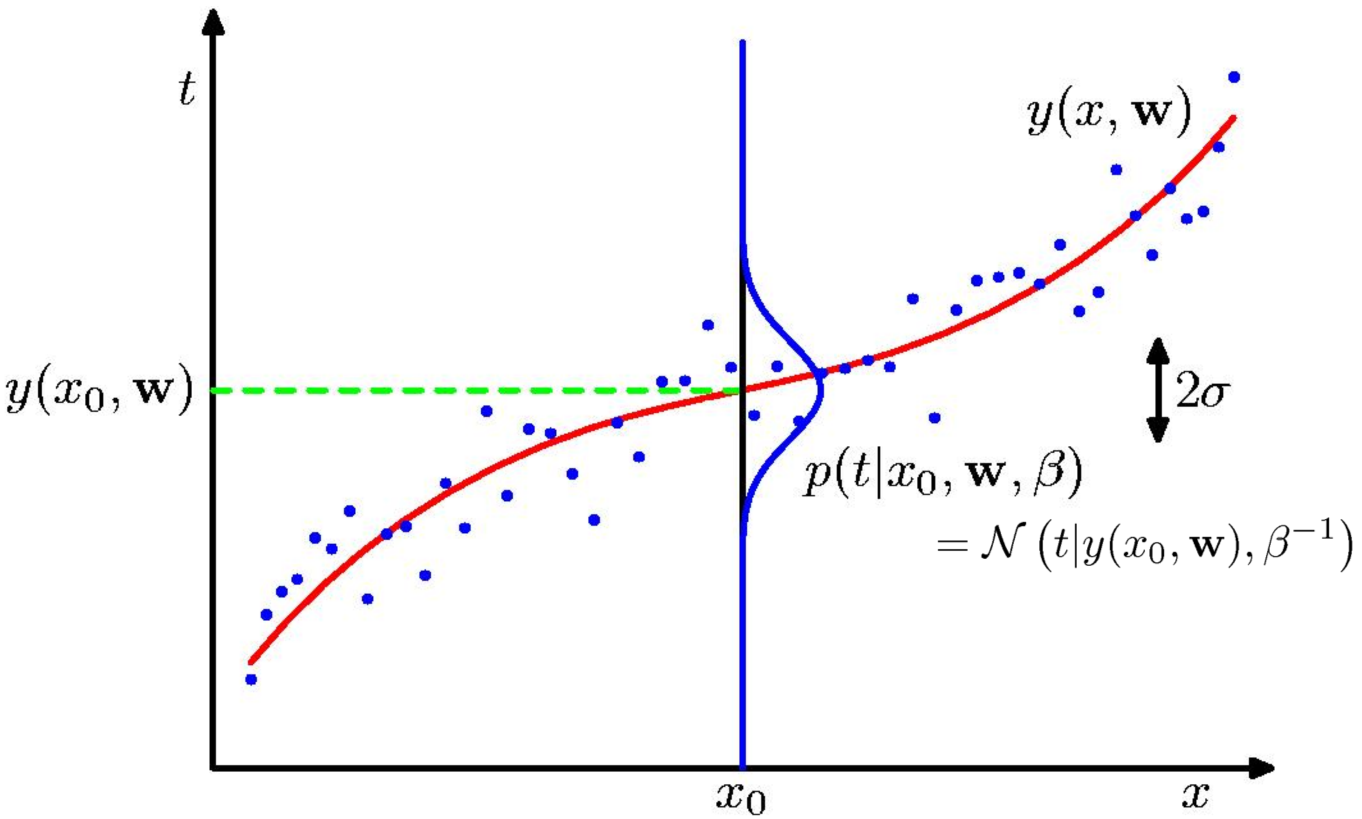


Regularization

9th Order Polynomial



Probabilistic View of Curve Fitting



Maximum Likelihood

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n | y(x_n, \mathbf{w}), \beta^{-1}\right)$$

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \underbrace{-\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2}_{\beta E(\mathbf{w})} + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi)$$

Determine $\mathbf{w}_{\mathrm{ML}}$ by minimizing sum-of-squares error, $E(\mathbf{w})$.

$$\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}_{\mathrm{ML}}) - t_n\}^2$$

MAP: A Step towards Bayes

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\}$$

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$$

$$\beta\widetilde{E}(\mathbf{w}) = \frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

Determine $\mathbf{w}_{\mathrm{MAP}}$ by minimizing regularized sum-of-squares error, $\widetilde{E}(\mathbf{w})$.

Some Key Ideas

- Change of variables and transformed densities
- Derivation of sum and product rules of probability
- Maximum likelihood and bias
- Least-squares as optimal probabilistic modeling