



---

# **MACHINE LEARNING**

## **CHAPTER 1: PRELIMINARY**

---

# Outlines

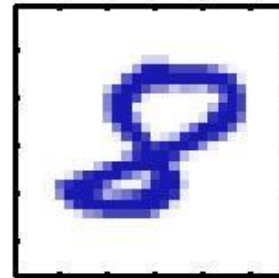
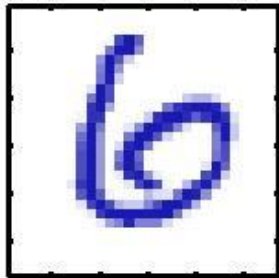
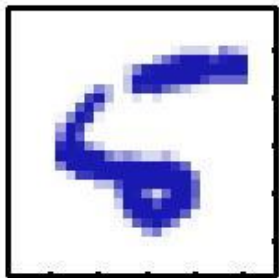
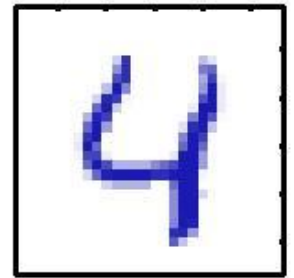
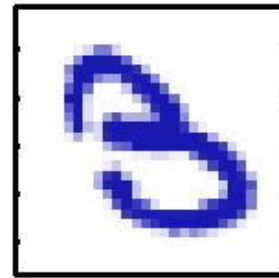
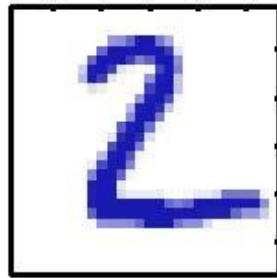
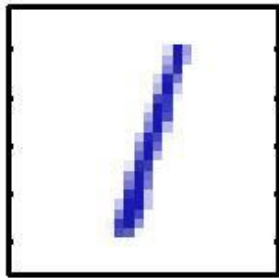
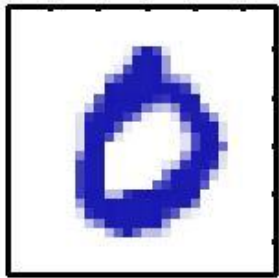
---

- Pattern Recognition
  - Curve Fitting and Regularization
  - Probabilities and Gaussian Distributions
  - Bayesian Inferences (ML and MAP)
  - Curse of Dimensionality
  - Decision Theory
  - Entropy and Information
-

# Example

---

## Handwritten Digit Recognition



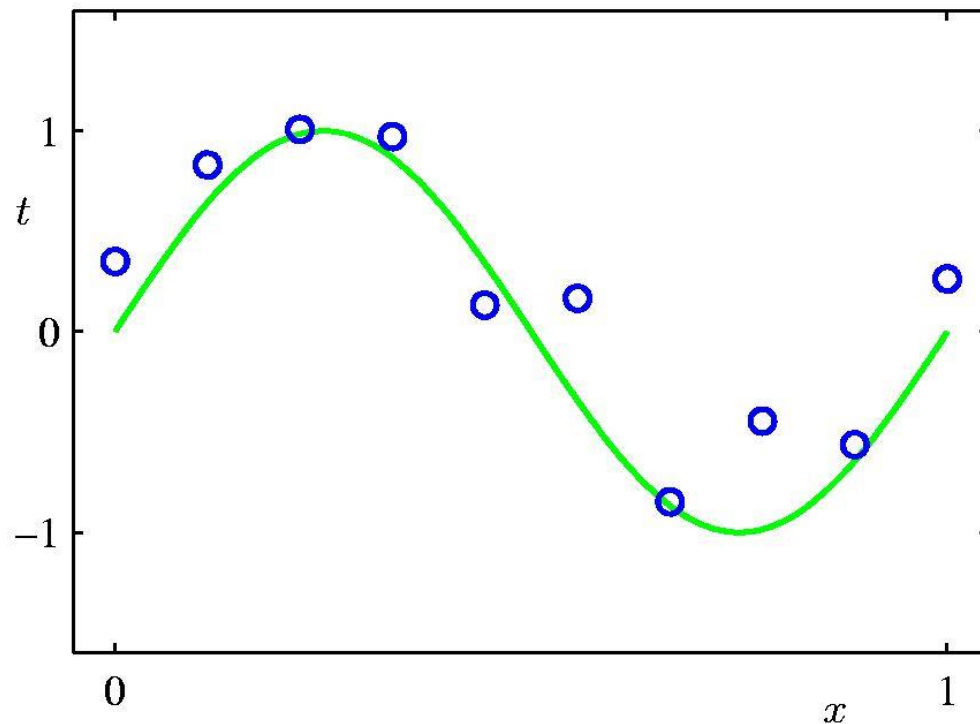
# Outlines

---

- Pattern Recognition
  - Curve Fitting and Regularization
  - Probabilities and Gaussian Distributions
  - Bayesian Inferences (ML and MAP)
  - Curse of Dimensionality
  - Decision Theory
  - Entropy and Information
-

# Polynomial Curve Fitting

---

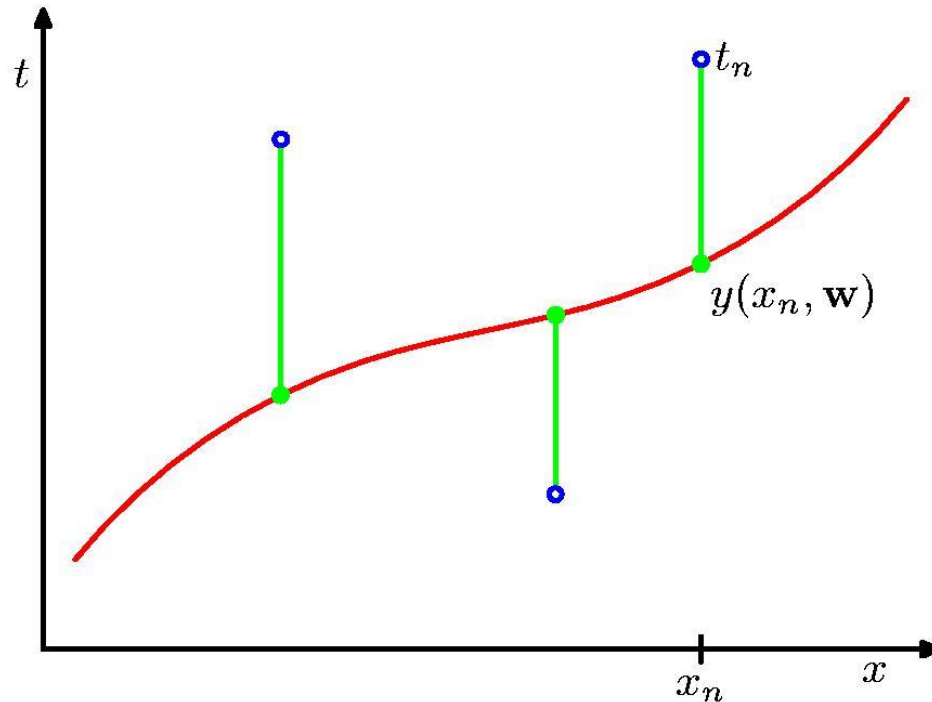


$$y(x, \mathbf{w}) = w_0 + w_1x + w_2x^2 + \dots + w_Mx^M = \sum_{j=0}^M w_jx^j$$

---

# Sum-of-Squares Error Function

---

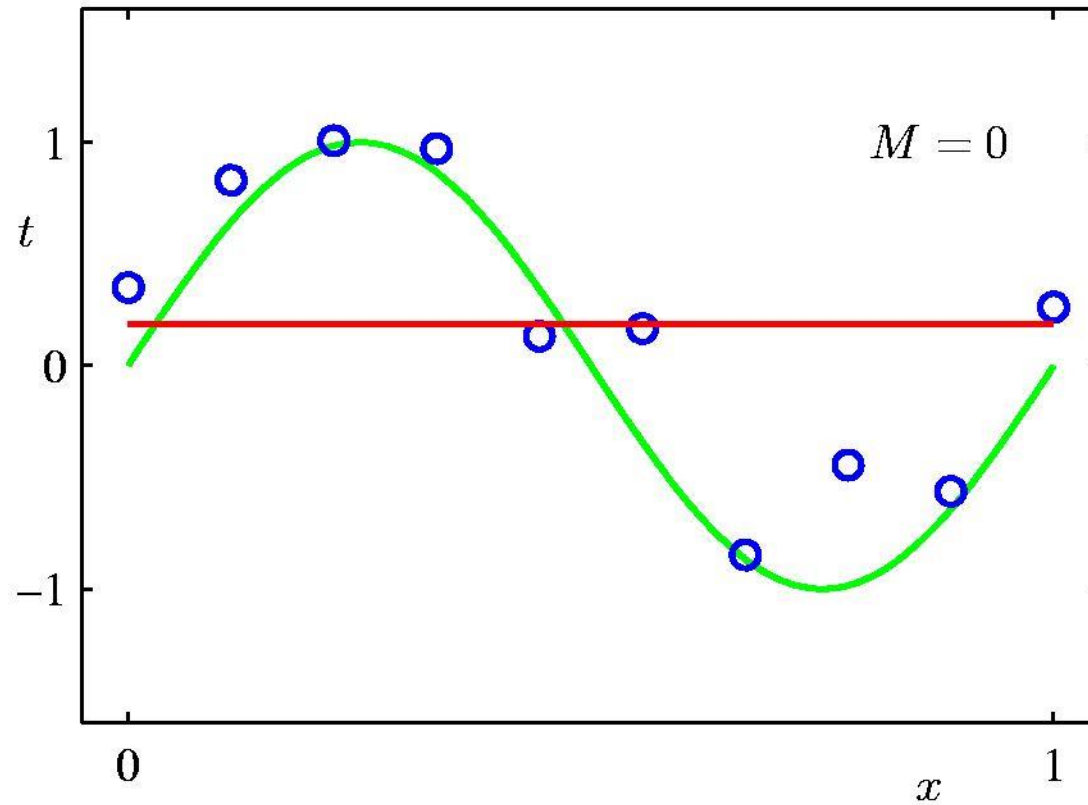


$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2$$

---

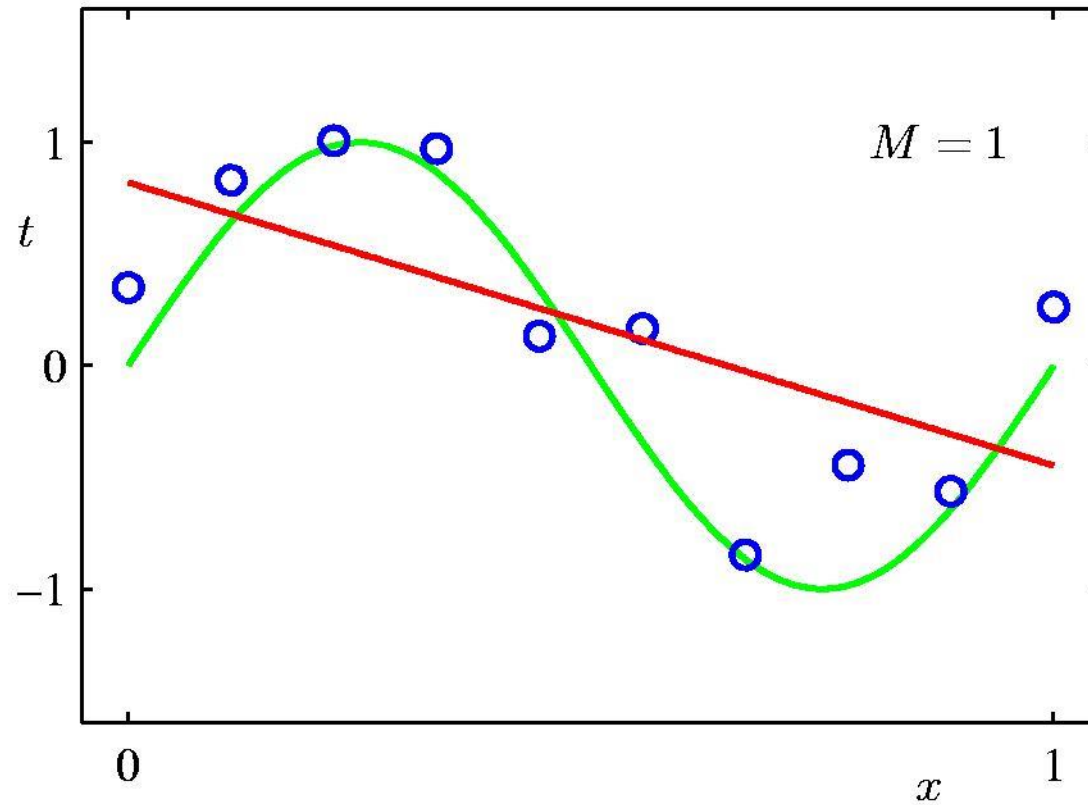
# 0<sup>th</sup> Order Polynomial

---



# 1<sup>st</sup> Order Polynomial

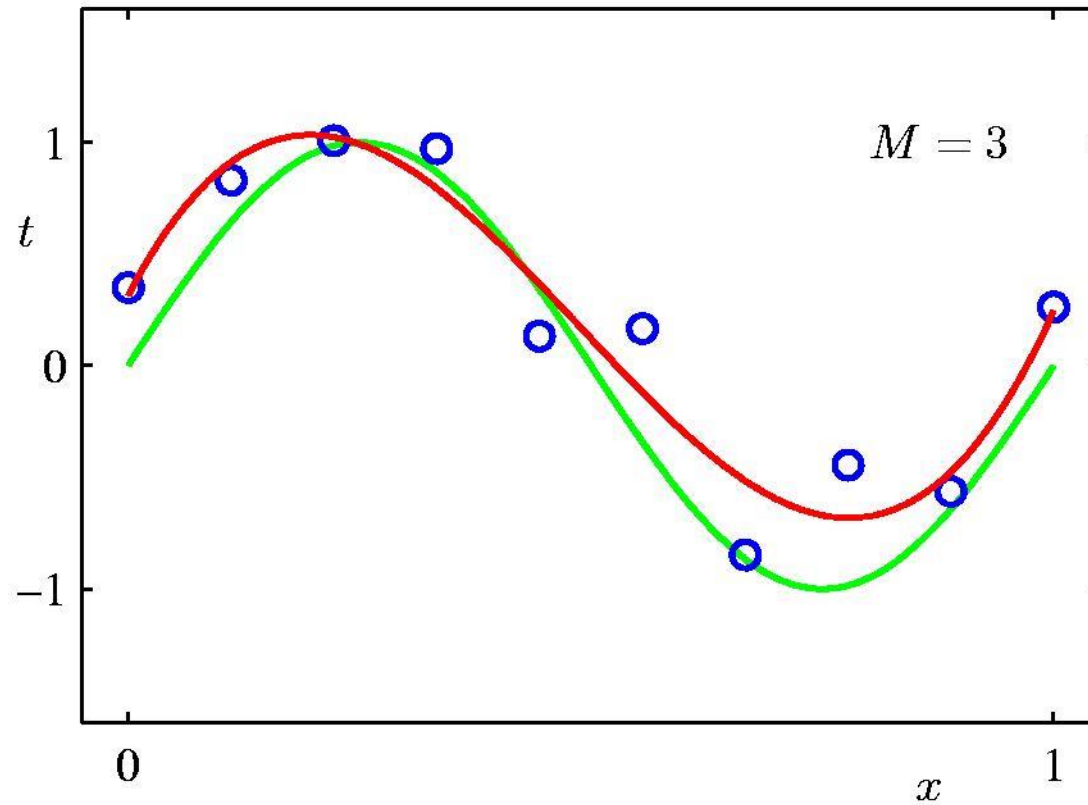
---





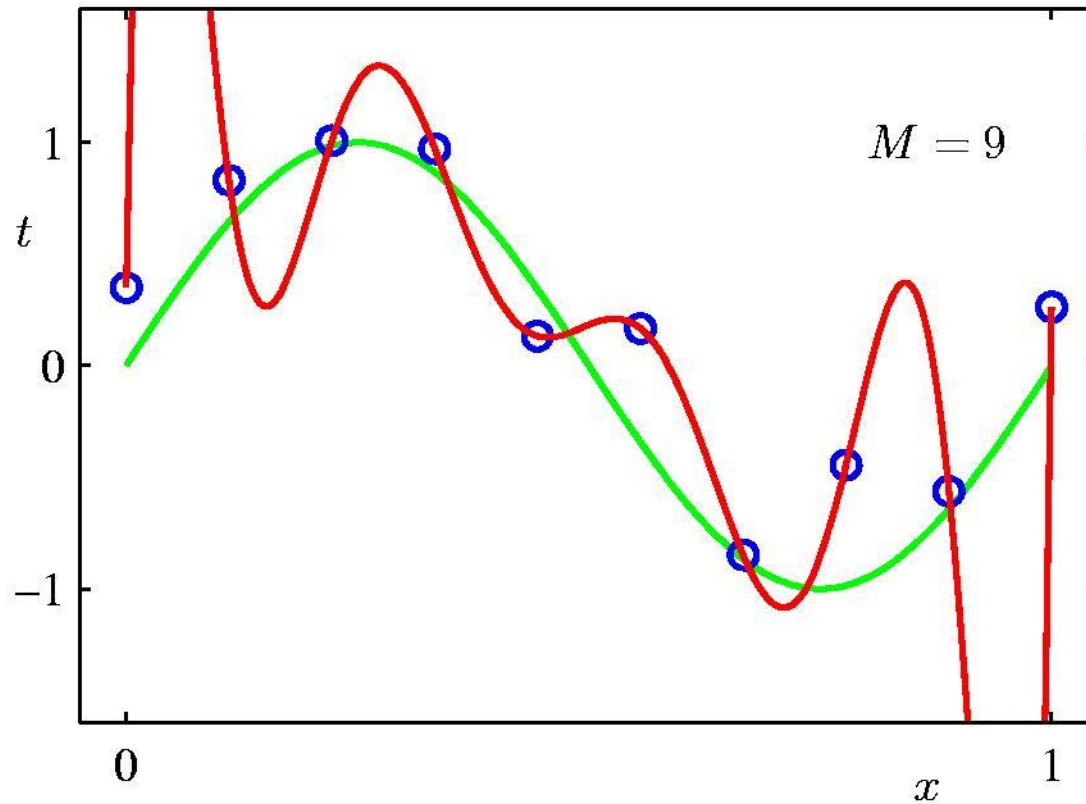
# 3<sup>rd</sup> Order Polynomial

---



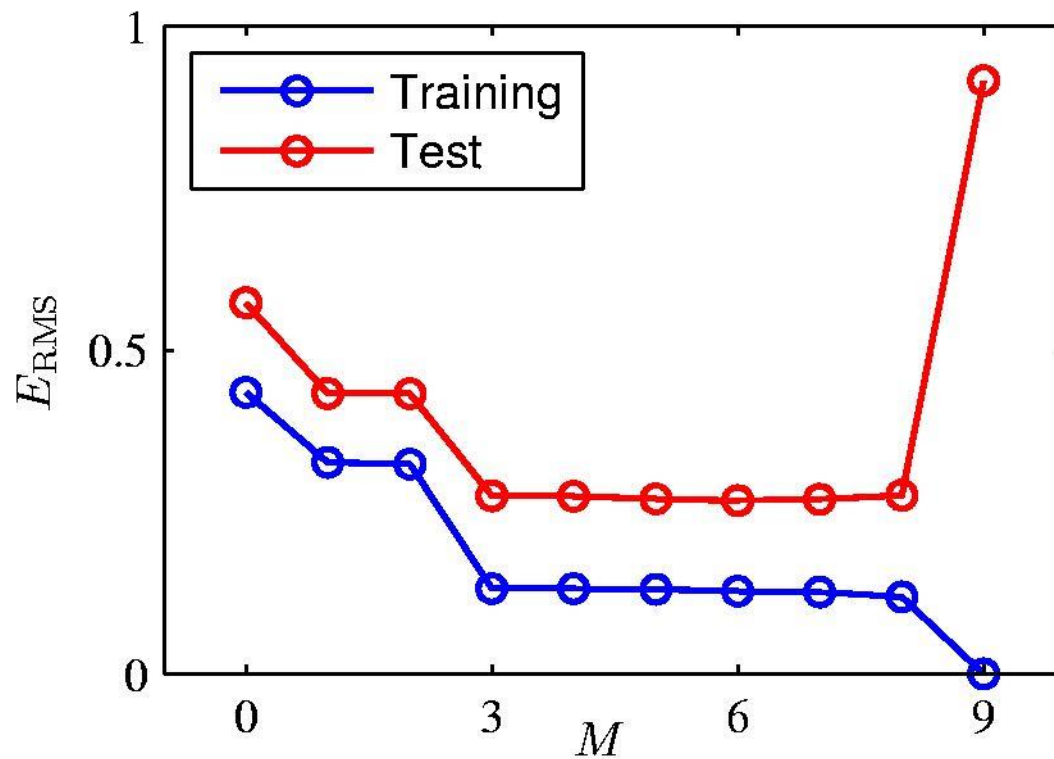
# 9<sup>th</sup> Order Polynomial

---



# Over-fitting

---



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

---

# Polynomial Coefficients

---

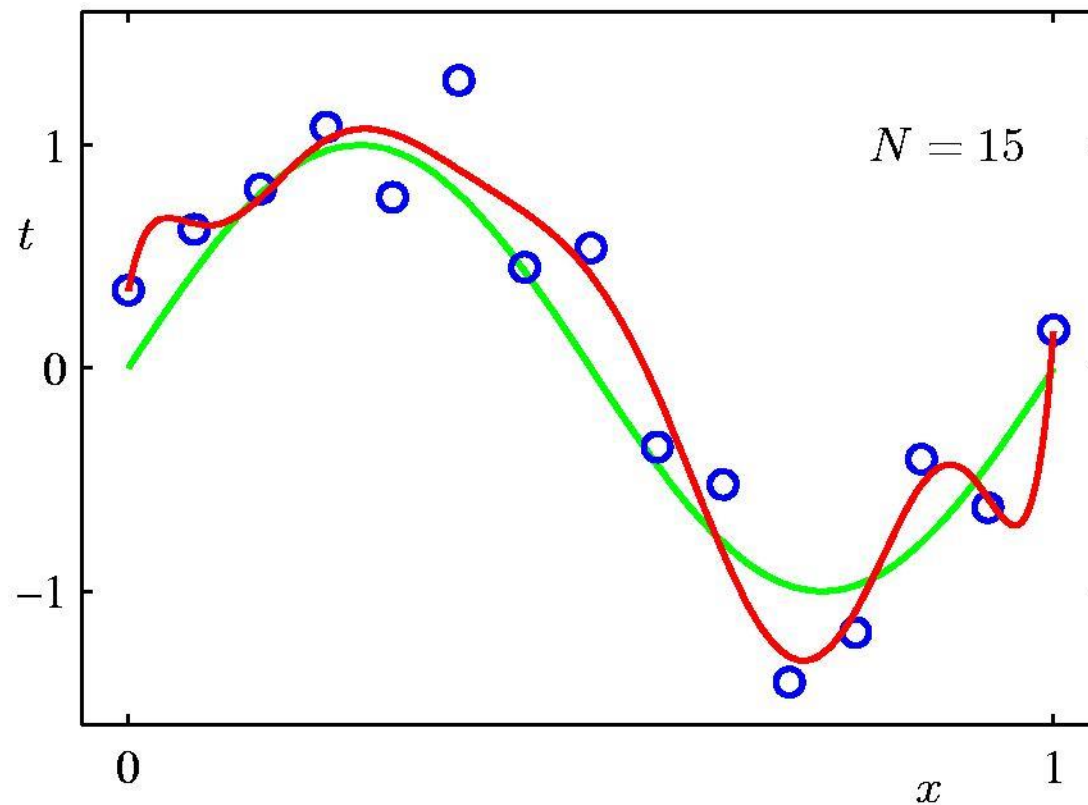
	$M = 0$	$M = 1$	$M = 3$	$M = 9$
$w_0^*$	0.19	0.82	0.31	0.35
$w_1^*$		-1.27	7.99	232.37
$w_2^*$			-25.43	-5321.83
$w_3^*$			17.37	48568.31
$w_4^*$				-231639.30
$w_5^*$				640042.26
$w_6^*$				-1061800.52
$w_7^*$				1042400.18
$w_8^*$				-557682.99
$w_9^*$				125201.43

---

# Data Set Size: $N = 15$

---

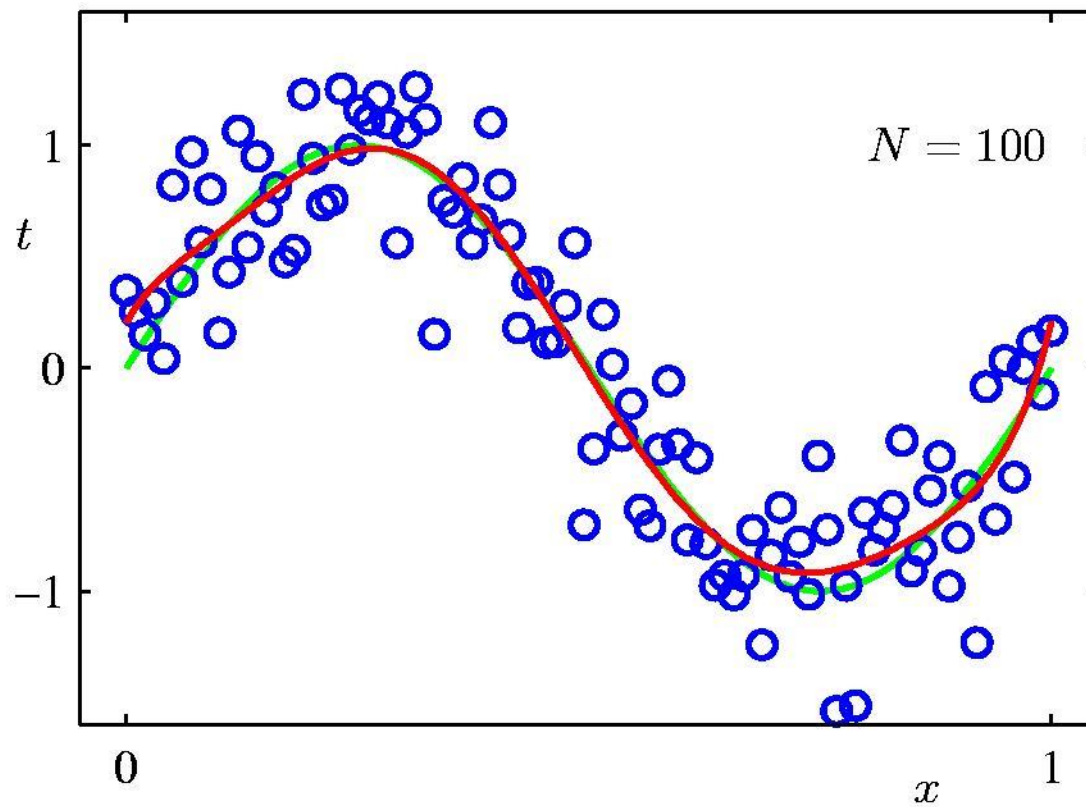
9<sup>th</sup> Order Polynomial



# Data Set Size: $N = 100$

---

9<sup>th</sup> Order Polynomial



# Regularization

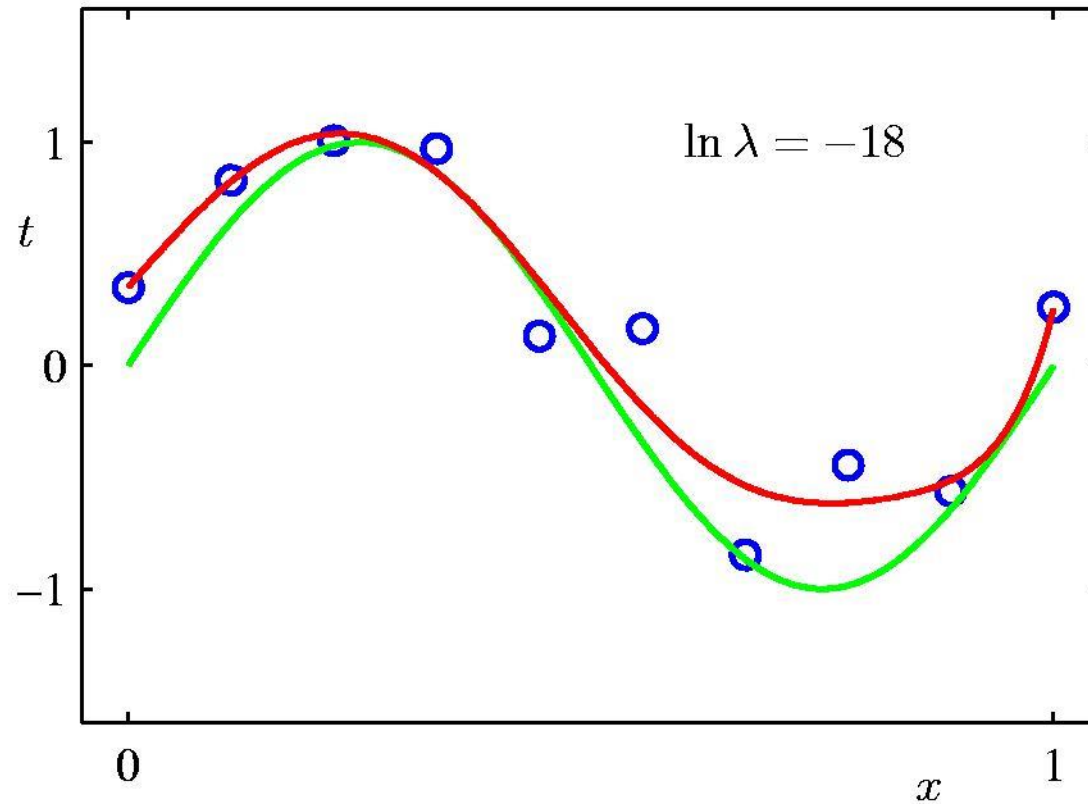
---

Penalize large coefficient values

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

# Regularization: $\ln \lambda = -18$

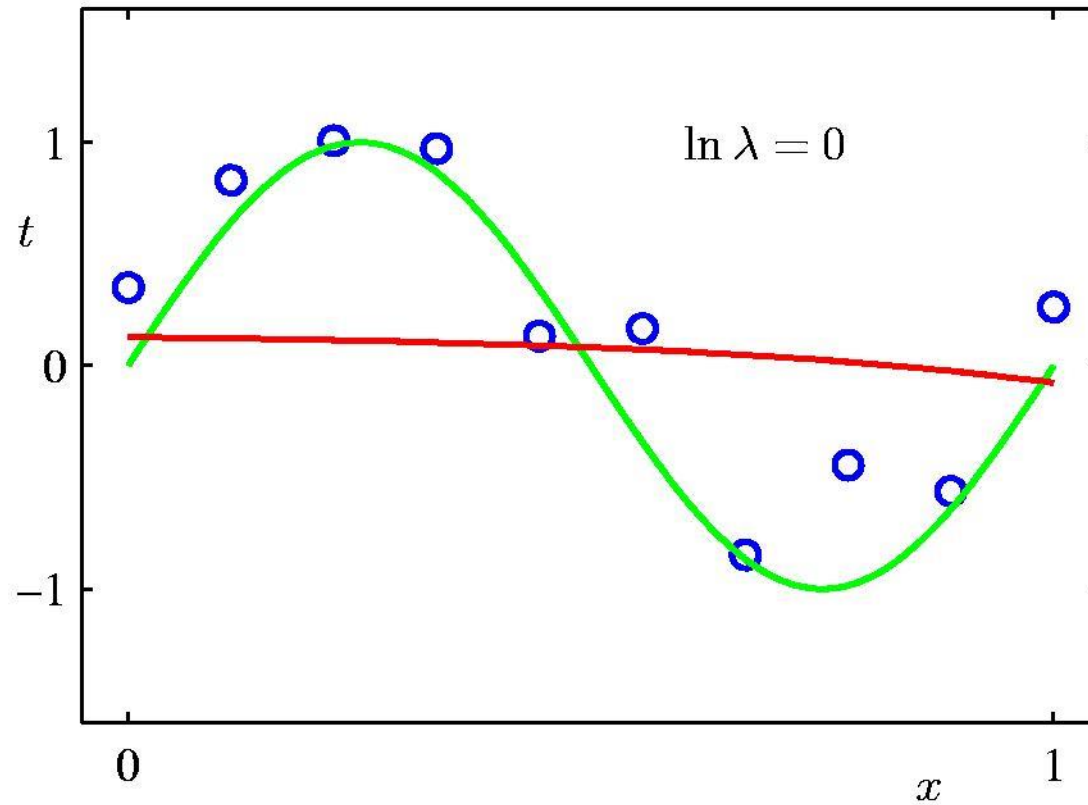
---





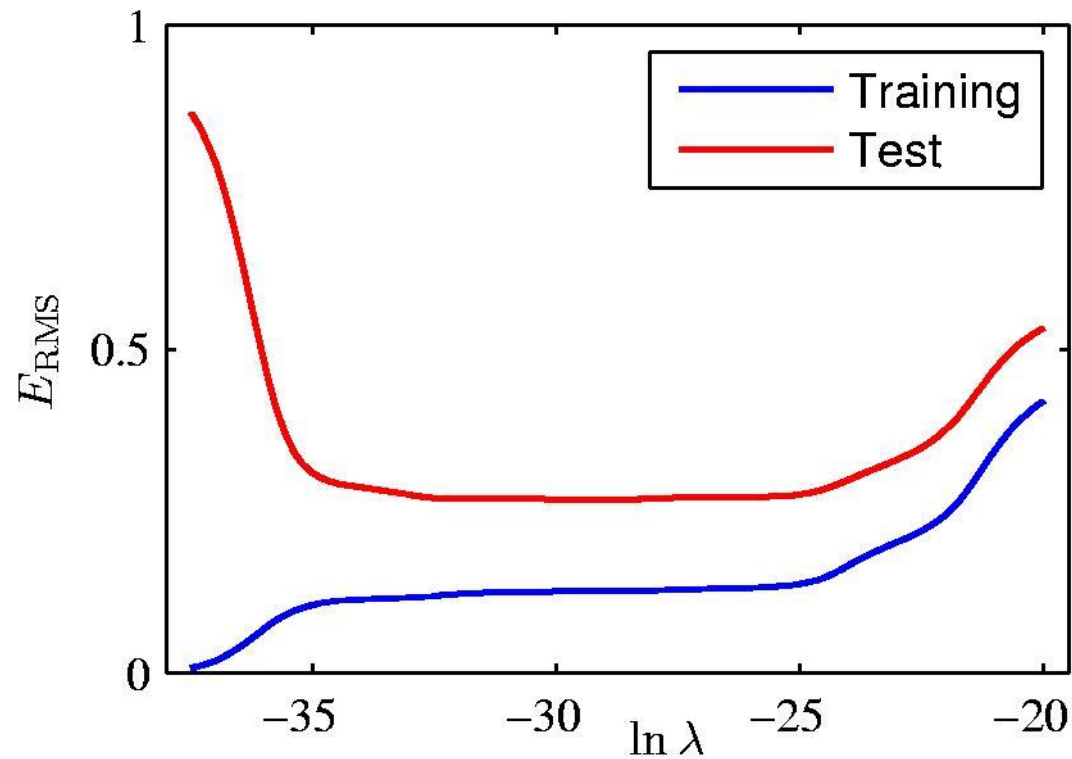
# Regularization: $\ln \lambda = 0$

---



# Regularization: $E_{\text{RMS}}$ vs. $\ln \lambda$

---



# Polynomial Coefficients

---

	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
$w_0^*$	0.35	0.35	0.13
$w_1^*$	232.37	4.74	-0.05
$w_2^*$	-5321.83	-0.77	-0.06
$w_3^*$	48568.31	-31.97	-0.05
$w_4^*$	-231639.30	-3.89	-0.03
$w_5^*$	640042.26	55.28	-0.02
$w_6^*$	-1061800.52	41.32	-0.01
$w_7^*$	1042400.18	-45.95	-0.00
$w_8^*$	-557682.99	-91.53	0.00
$w_9^*$	125201.43	72.68	0.01

---

# Outlines

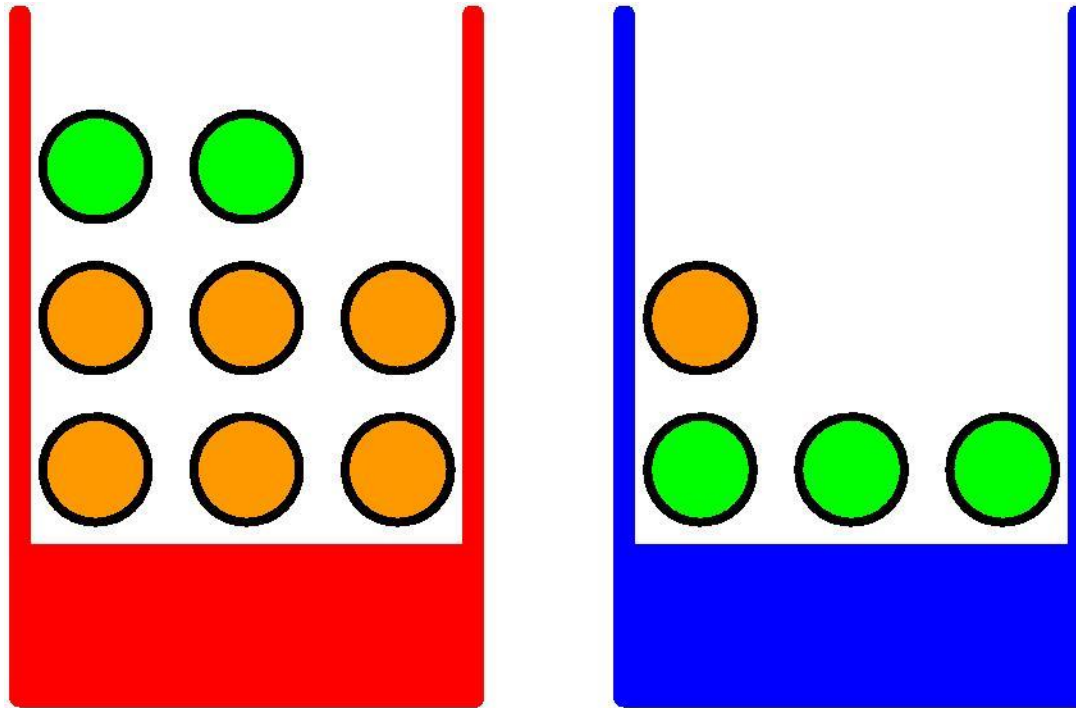
---

- Pattern Recognition
  - Curve Fitting and Regularization
  - Probabilities and Gaussian Distributions
  - Bayesian Inferences (ML and MAP)
  - Curse of Dimensionality
  - Decision Theories
  - Entropy and Information
-

# Probability Theory

---

Apples and Oranges



# Probability Theory

---

$y_j$			$n_{ij}$	
			$x_i$	

Marginal Probability

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

Joint Probability

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

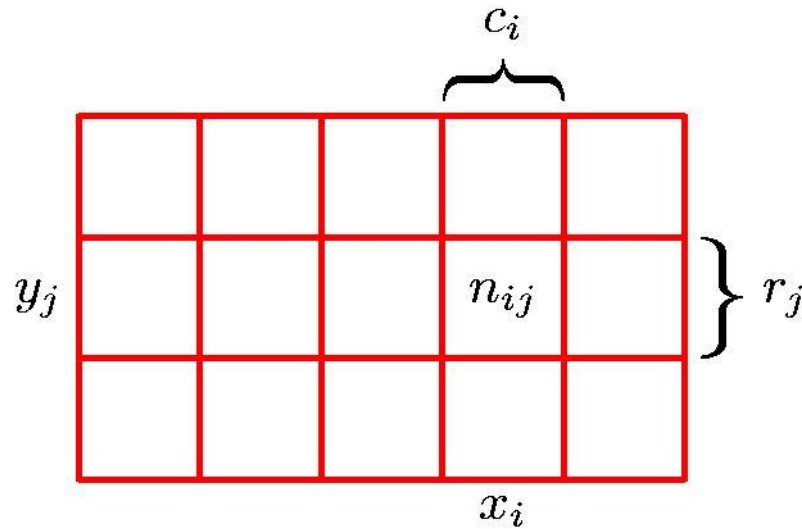
Conditional Probability

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

---

# Probability Theory

---



Sum Rule

$$\begin{aligned} p(X = x_i) &= \frac{c_i}{N} = \frac{1}{N} \sum_{j=1}^L n_{ij} \\ &= \sum_{j=1}^L p(X = x_i, Y = y_j) \end{aligned}$$

Product Rule

$$\begin{aligned} p(X = x_i, Y = y_j) &= \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} \\ &= p(Y = y_j | X = x_i) p(X = x_i) \end{aligned}$$

---

# The Rules of Probability

---

Sum Rule

$$p(X) = \sum_Y p(X, Y)$$

Product Rule

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

---



# Bayes' Theorem

---

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

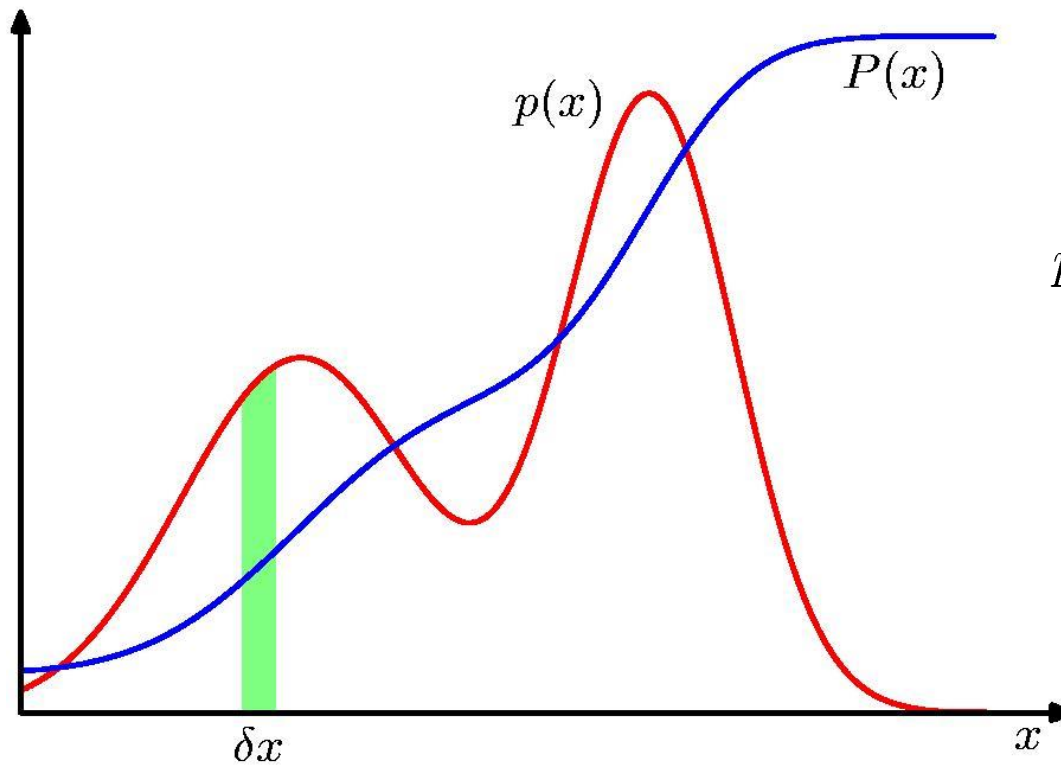
$$p(X) = \sum_Y p(X|Y)p(Y)$$

posterior  $\propto$  likelihood  $\times$  prior

---

# Probability Densities

---



$$p(x \in (a, b)) = \int_a^b p(x) dx$$

$$P(z) = \int_{-\infty}^z p(x) dx$$

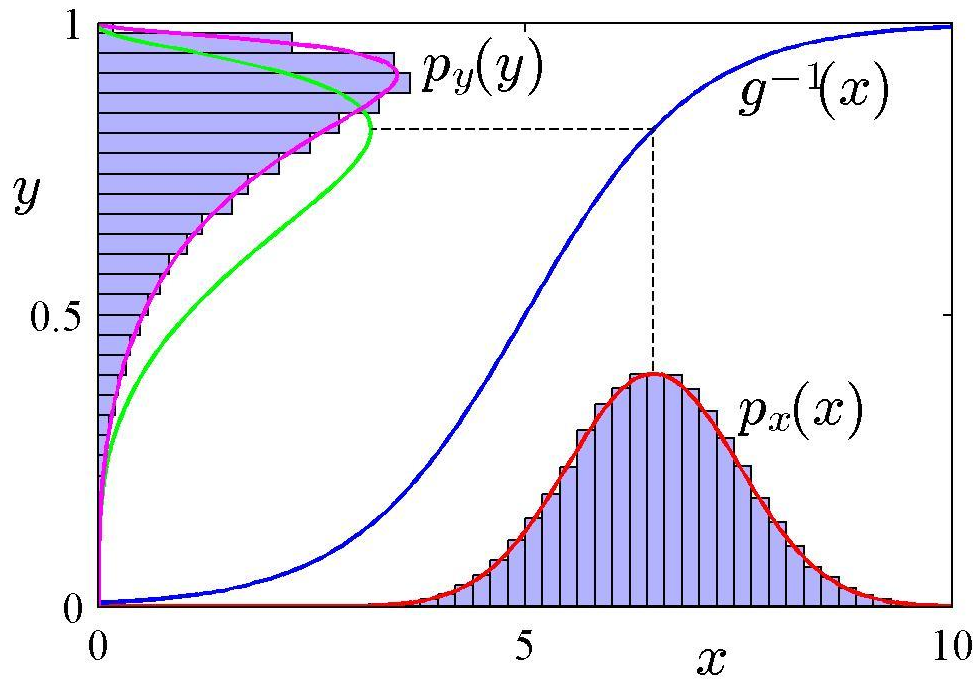
$$p(x) \geq 0$$

$$\int_{-\infty}^{\infty} p(x) dx = 1$$

---

# Transformed Densities

---




$$\begin{aligned} p_y(y) &= p_x(x) \left| \frac{dx}{dy} \right| \\ &= p_x(g(y)) |g'(y)| \end{aligned}$$

# Expectations

---

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

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

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


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

---

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

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

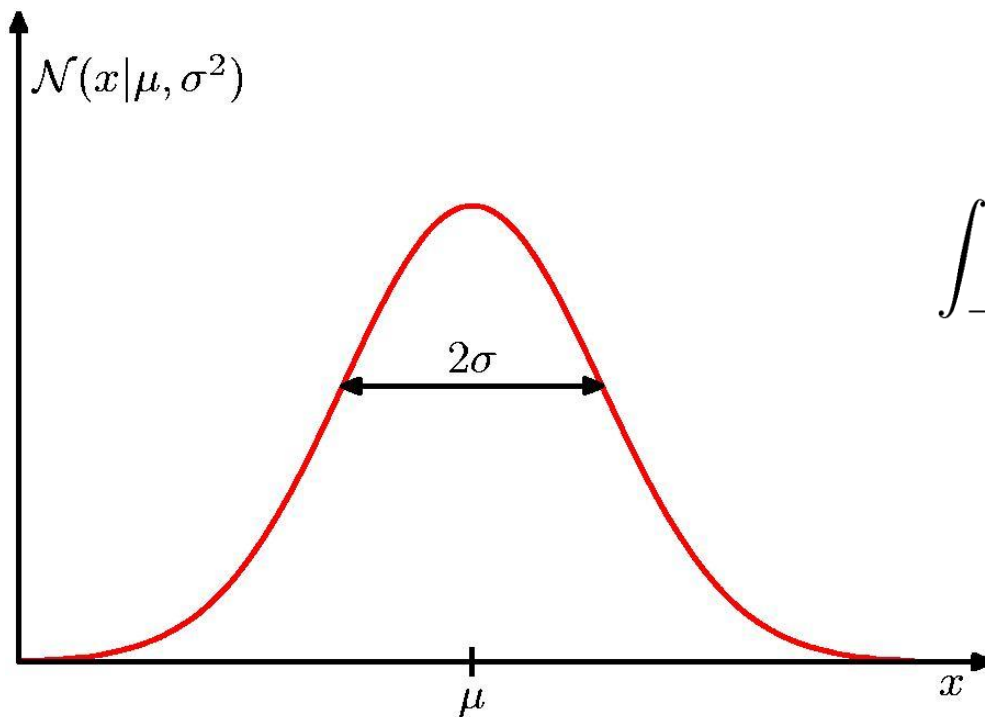
$$\begin{aligned} \text{cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x}, \mathbf{y}} [\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\} \{\mathbf{y}^T - \mathbb{E}[\mathbf{y}^T]\}] \\ &= \mathbb{E}_{\mathbf{x}, \mathbf{y}}[\mathbf{x}\mathbf{y}^T] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^T] \end{aligned}$$

---

# The Gaussian Distribution

---

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



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

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

# Gaussian Mean and Variance

---

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

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

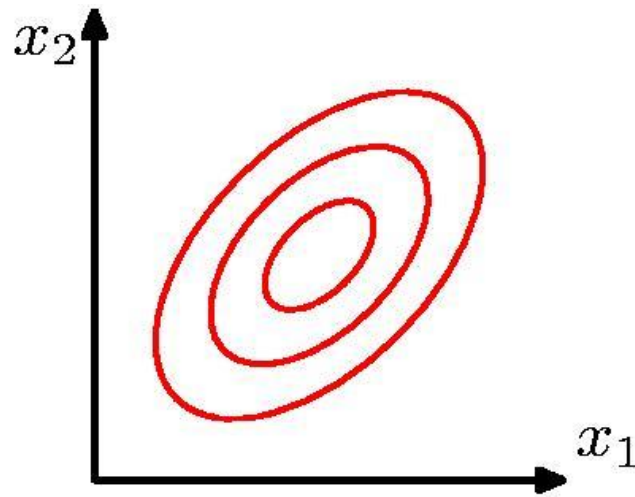
$$\text{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$$

---

# The Multivariate Gaussian

---

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





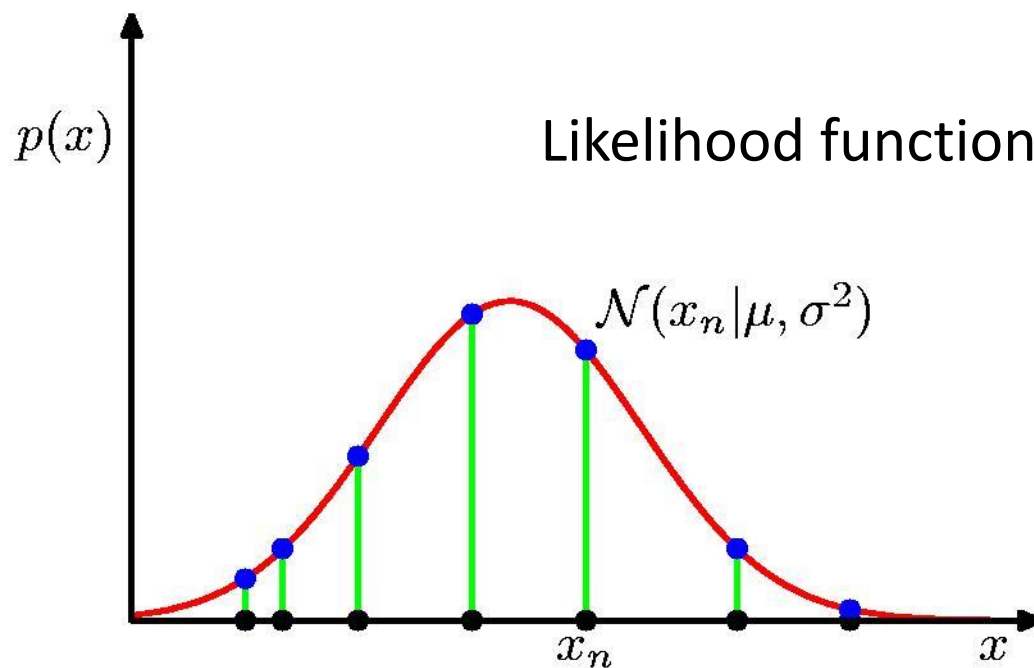
# Outlines

---

- Pattern Recognition
  - Curve Fitting and Regularization
  - Probabilities and Gaussian Distributions
  - Bayesian Inferences (ML and MAP)
  - Curse of Dimensionality
  - Decision Theories
  - Entropy and Information
-

# Gaussian Parameter Estimation

---



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

# Maximum (Log) Likelihood

---

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

$$\mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N x_n \qquad \sigma_{\text{ML}}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2$$

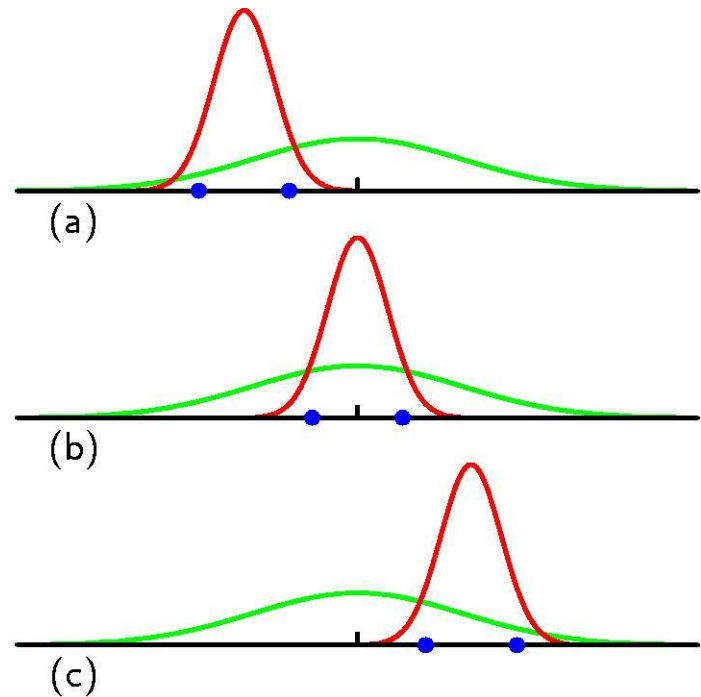
# Properties of $\mu_{\text{ML}}$ and $\sigma_{\text{ML}}^2$

---

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

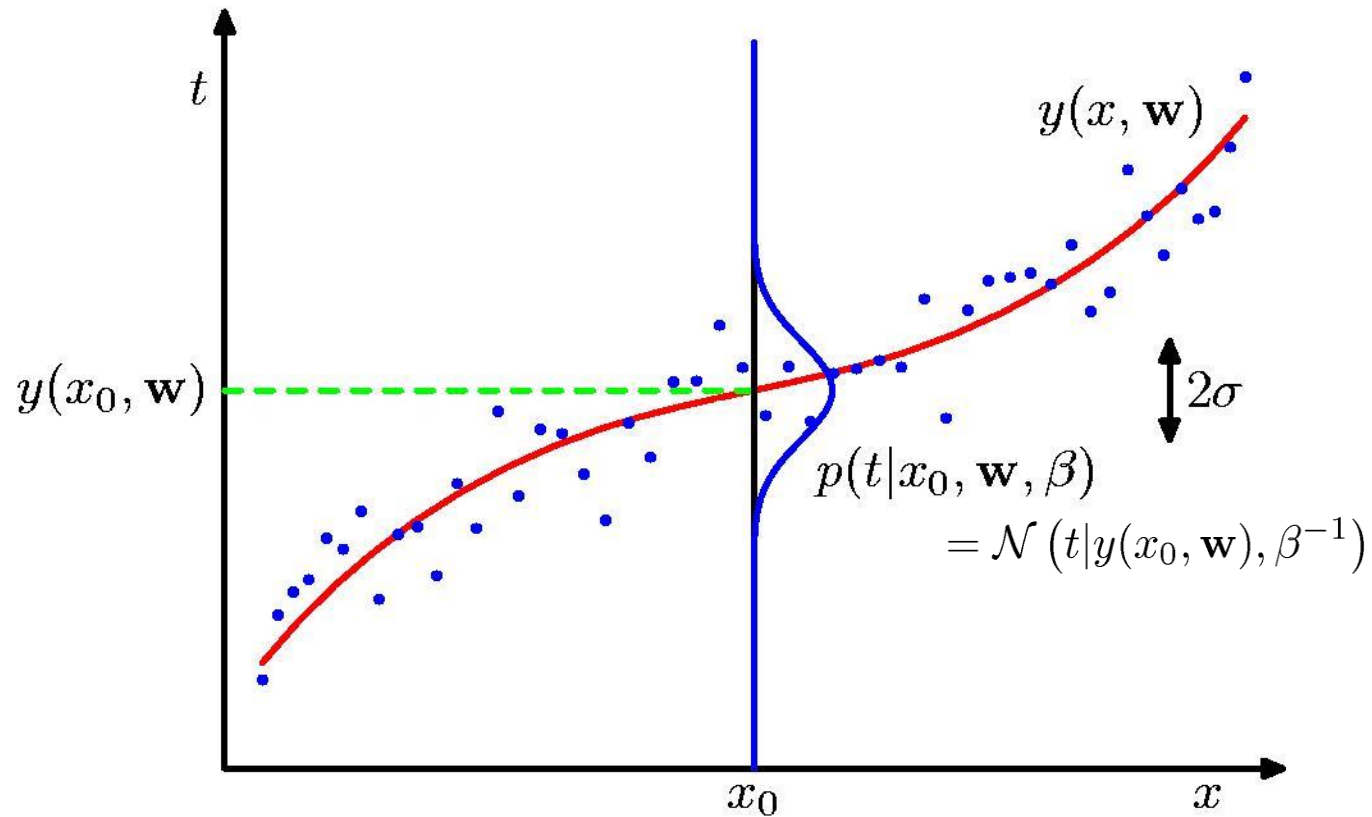
$$\mathbb{E}[\sigma_{\text{ML}}^2] = \left(\frac{N-1}{N}\right) \sigma^2$$

$$\begin{aligned}\tilde{\sigma}^2 &= \frac{N}{N-1} \sigma_{\text{ML}}^2 \\ &= \frac{1}{N-1} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2\end{aligned}$$



# Curve Fitting Re-visited

---



# Maximum Likelihood

---

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

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

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

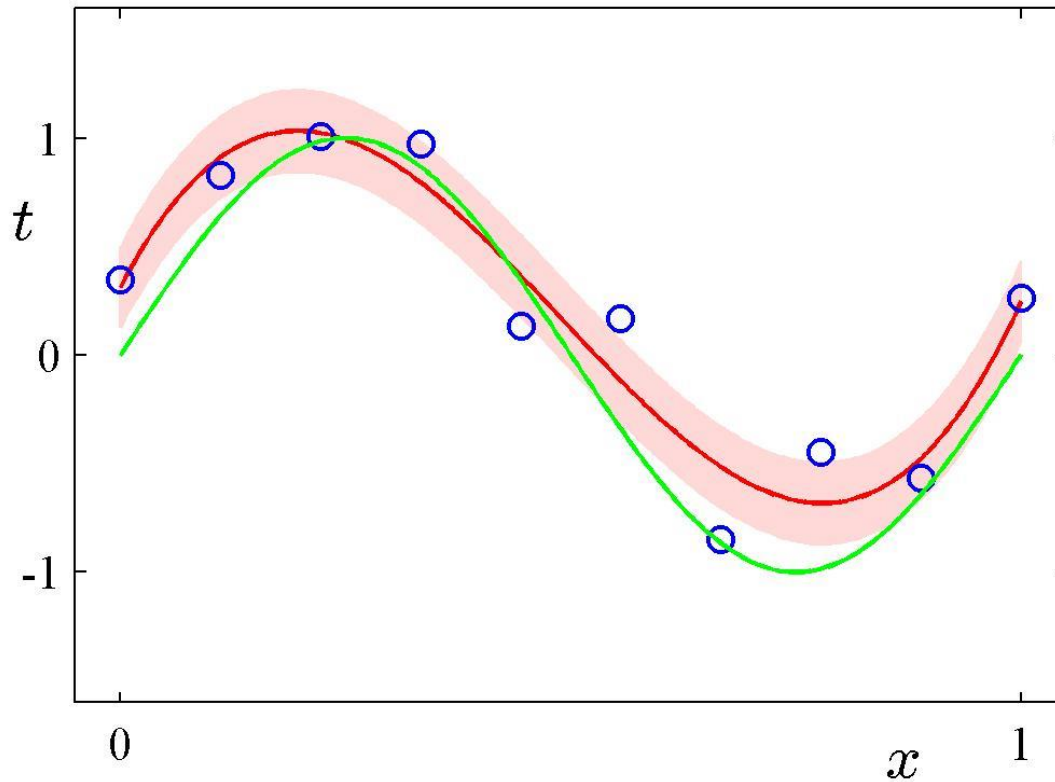
$$\frac{1}{\beta_{\text{ML}}} = \frac{1}{N} \sum_{n=1}^N \{y(x_n, \mathbf{w}_{\text{ML}}) - t_n\}^2$$

---

# Predictive Distribution

---

$$p(t|x, \mathbf{w}_{\text{ML}}, \beta_{\text{ML}}) = \mathcal{N}(t|y(x, \mathbf{w}_{\text{ML}}), \beta_{\text{ML}}^{-1})$$



# MAP: A Step towards Bayes

---

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

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

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

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

---



# Bayesian Curve Fitting

---

$$p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|x, \mathbf{w})p(\mathbf{w}|\mathbf{x}, \mathbf{t}) d\mathbf{w} = \mathcal{N}(t|m(x), s^2(x))$$

$$m(x) = \beta \phi(x)^T \mathbf{S} \sum_{n=1}^N \phi(x_n) t_n \qquad s^2(x) = \beta^{-1} + \phi(x)^T \mathbf{S} \phi(x)$$

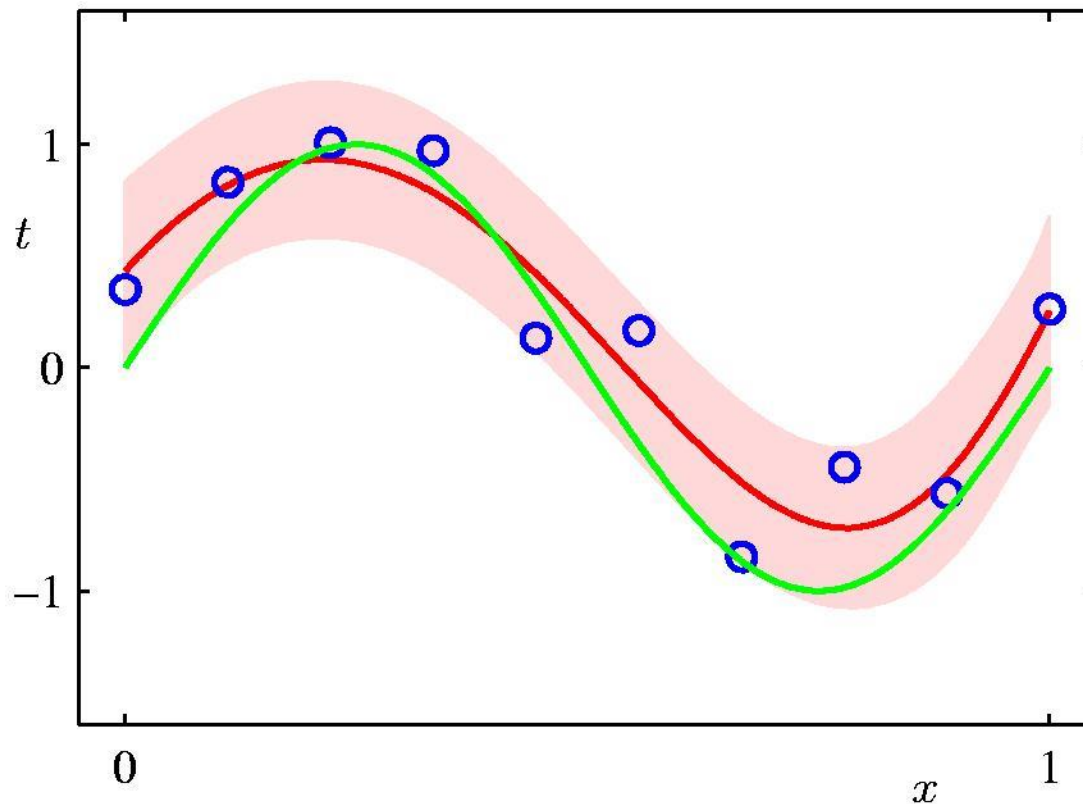
$$\mathbf{S}^{-1} = \alpha \mathbf{I} + \beta \sum_{n=1}^N \phi(x_n) \phi(x_n)^T \qquad \phi(x_n) = (x_n^0, \dots, x_n^M)^T$$

---

# Bayesian Predictive Distribution

---

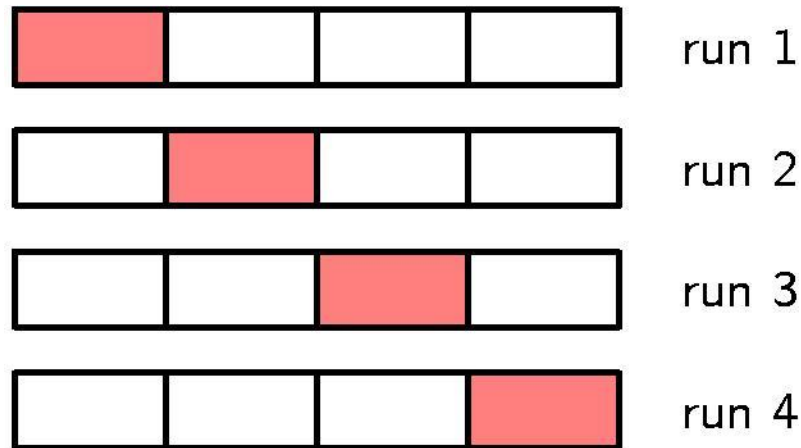
$$p(t|x, \mathbf{x}, \mathbf{t}) = \mathcal{N}(t|m(x), s^2(x))$$



# Model Selection

---

## Cross-Validation



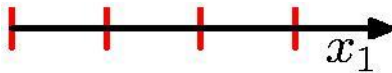
# Outlines

---

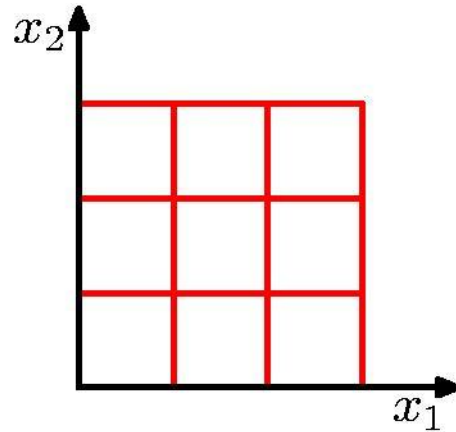
- Pattern Recognition
  - Curve Fitting and Regularization
  - Probabilities and Gaussian Distributions
  - Bayesian Inferences (ML and MAP)
  - Curse of Dimensionality
  - Decision Theory
  - Entropy and Information
-

# Curse of Dimensionality

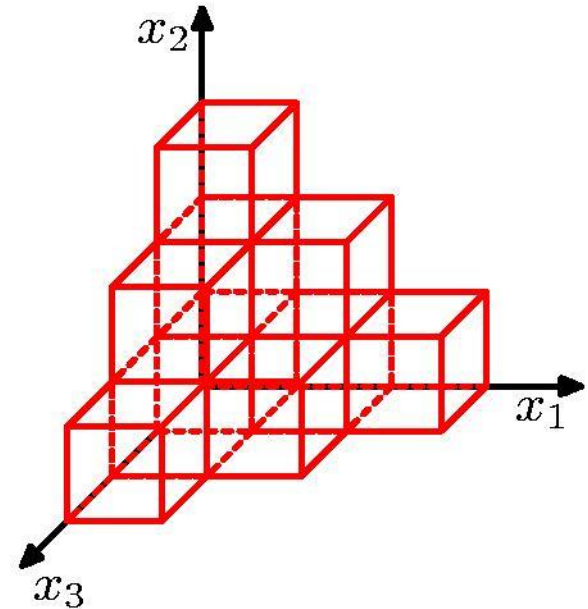
---



$D = 1$



$D = 2$



$D = 3$

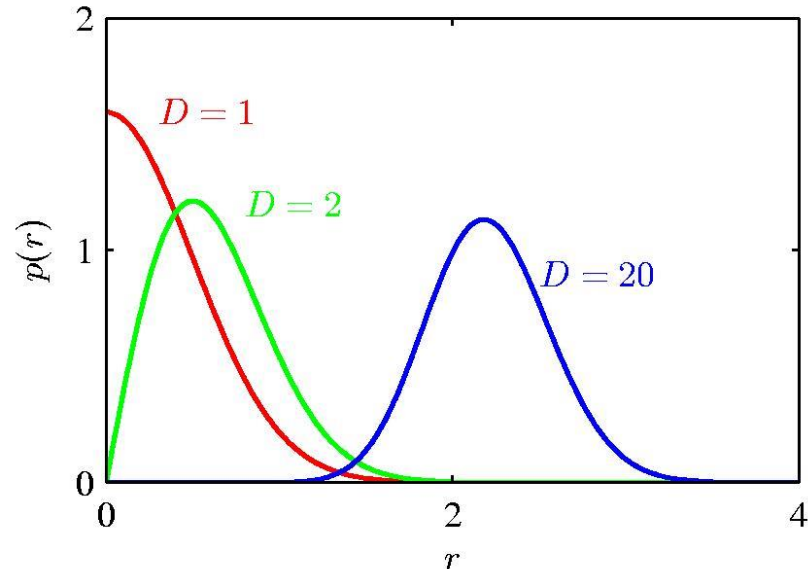
# Curse of Dimensionality

---

Polynomial curve fitting,  $M = 3$

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^D w_i x_i + \sum_{i=1}^D \sum_{j=1}^D w_{ij} x_i x_j + \sum_{i=1}^D \sum_{j=1}^D \sum_{k=1}^D w_{ijk} x_i x_j x_k$$

Gaussian Densities in  
higher dimensions



# Outlines

---

- Pattern Recognition
  - Curve Fitting and Regularization
  - Probabilities and Gaussian Distributions
  - Bayesian Inferences (ML and MAP)
  - Curse of Dimensionality
  - Decision Theory
  - Entropy and Information
-

# Decision Theory

---

Inference step

Determine either  $p(t|\mathbf{x})$  or  $p(\mathbf{x}, t)$ .

Decision step

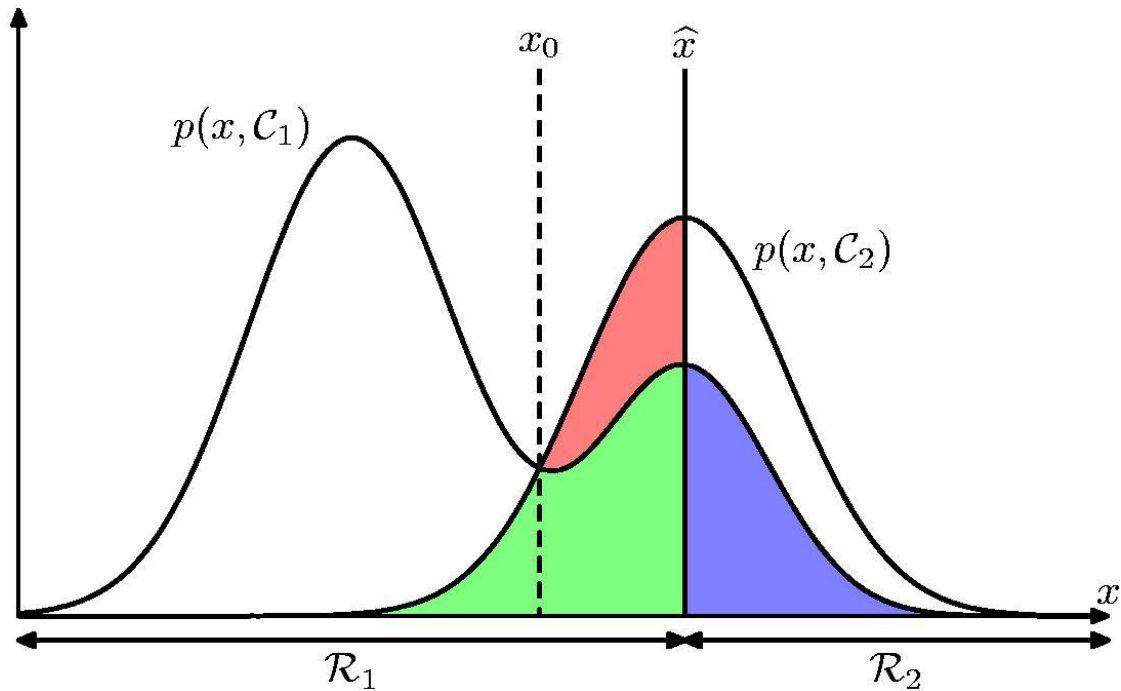
For given  $\mathbf{x}$ , determine optimal  $t$ .

---



# Minimum Misclassification Rate

---



$$\begin{aligned} p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2) d\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1) d\mathbf{x}. \end{aligned}$$

---

# Minimum Expected Loss

---

Example: classify medical images as 'cancer' or 'normal'

		Decision	
		cancer	normal
Truth	cancer	0	1000
	normal	1	0

# Minimum Expected Loss

---

$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x}$$

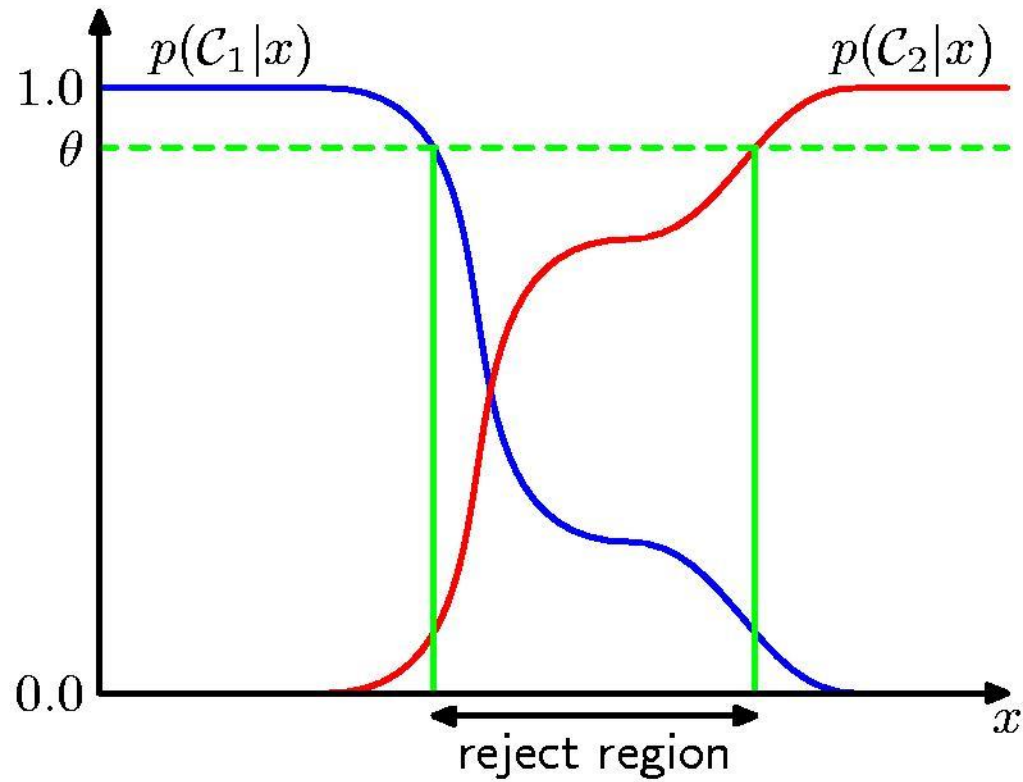
Regions  $\mathcal{R}_j$  are chosen to minimize

$$\mathbb{E}[L] = \sum_k L_{kj} p(\mathcal{C}_k | \mathbf{x})$$

---

# Reject Option

---



# Why Separate Inference and Decision?

---

- Minimizing risk (loss matrix may change over time)
- Reject option
- Unbalanced class priors
- Combining models

# Decision Theory for Regression

---

Inference step

Determine  $p(\mathbf{x}, t)$ .

Decision step

For given  $\mathbf{x}$ , make optimal prediction,  $y(\mathbf{x})$ , for  $t$ .

Loss function:  $\mathbb{E}[L] = \iint L(t, y(\mathbf{x}))p(\mathbf{x}, t) \, d\mathbf{x} \, dt$

---

# The Squared Loss Function

---

$$\mathbb{E}[L] = \iint \{y(\mathbf{x}) - t\}^2 p(\mathbf{x}, t) \, d\mathbf{x} \, dt$$

$$\begin{aligned} \{y(\mathbf{x}) - t\}^2 &= \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}] + \mathbb{E}[t|\mathbf{x}] - t\}^2 \\ &= \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 + 2\{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}\{\mathbb{E}[t|\mathbf{x}] - t\} + \{\mathbb{E}[t|\mathbf{x}] - t\}^2 \end{aligned}$$

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 p(\mathbf{x}) \, d\mathbf{x} + \int \text{var} [t|\mathbf{x}] p(\mathbf{x}) \, d\mathbf{x}$$

$$y(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}]$$

---

# Generative vs Discriminative

---

Generative approach:

Model  $p(t, \mathbf{x}) = p(\mathbf{x}|t)p(t)$

Use Bayes' theorem  $p(t|\mathbf{x}) = \frac{p(\mathbf{x}|t)p(t)}{p(\mathbf{x})}$

Discriminative approach:

Model  $p(t|\mathbf{x})$  directly

---



# Outlines

---

- Pattern Recognition
  - Curve Fitting and Regularization
  - Probabilities and Gaussian Distributions
  - Bayesian Inferences (ML and MAP)
  - Curse of Dimensionality
  - Decision Theories
  - Entropy and Information
-

# Entropy

---

$$H[x] = - \sum_x p(x) \log_2 p(x)$$

Important quantity in

- coding theory
- statistical physics
- machine learning

# Entropy

---

Coding theory:  $x$  discrete with 8 possible states; how many bits to transmit the state of  $x$ ?

All states equally likely

$$H[x] = -8 \times \frac{1}{8} \log_2 \frac{1}{8} = 3 \text{ bits.}$$

# Entropy

---

$x$	a	b	c	d	e	f	g	h
$p(x)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{64}$
code	0	10	110	1110	111100	111101	111110	111111

$$\begin{aligned} H[x] &= -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{4} \log_2 \frac{1}{4} - \frac{1}{8} \log_2 \frac{1}{8} - \frac{1}{16} \log_2 \frac{1}{16} - \frac{4}{64} \log_2 \frac{1}{64} \\ &= 2 \text{ bits} \end{aligned}$$

$$\begin{aligned} \text{average code length} &= \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{16} \times 4 + 4 \times \frac{1}{64} \times 6 \\ &= 2 \text{ bits} \end{aligned}$$

---

# Entropy

---

In how many ways can  $N$  identical objects be allocated  $M$  bins?

$$W = \frac{N!}{\prod_i n_i!}$$

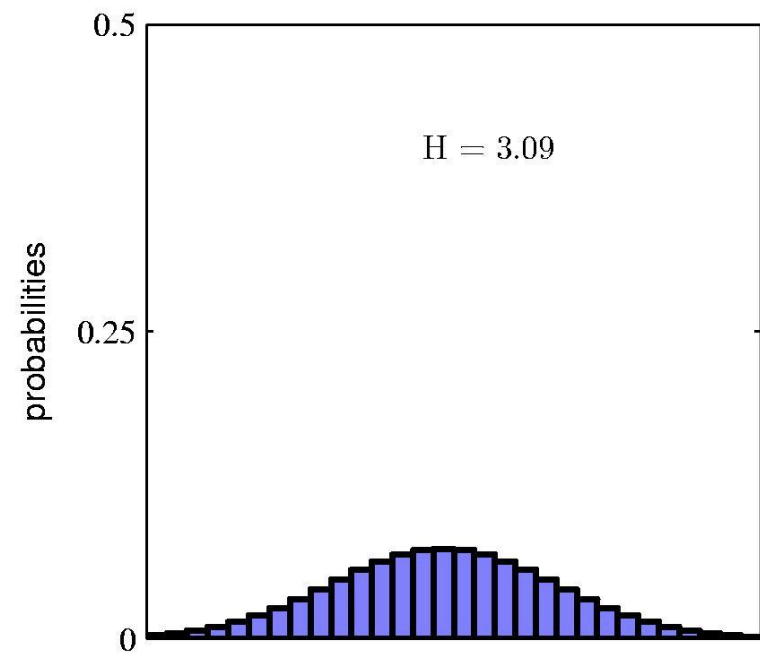
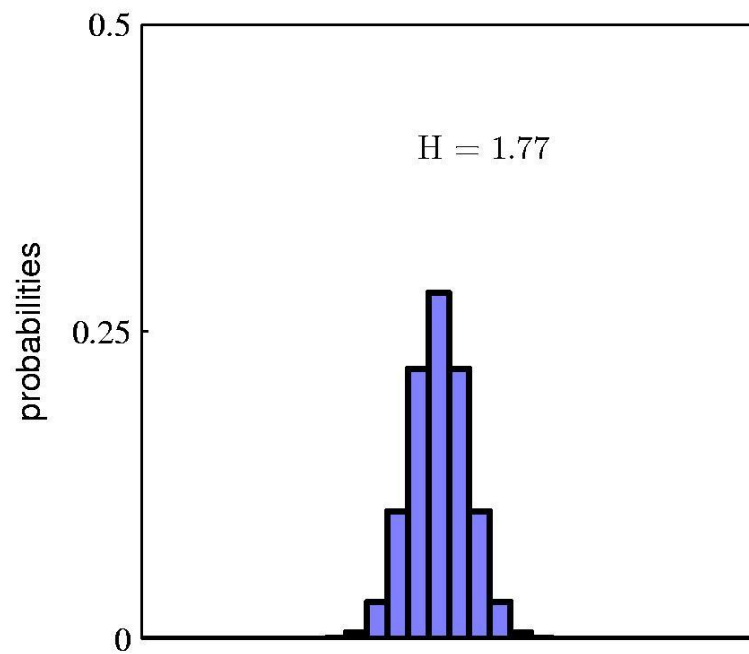
$$H = \frac{1}{N} \ln W \simeq - \lim_{N \rightarrow \infty} \sum_i \left( \frac{n_i}{N} \right) \ln \left( \frac{n_i}{N} \right) = - \sum_i p_i \ln p_i$$

Entropy maximized when  $\forall i : p_i = \frac{1}{M}$

---

# Entropy

---



# Differential Entropy

---

Put bins of width  $\Delta$  along the real line

$$\lim_{\Delta \rightarrow 0} \left\{ - \sum_i p(x_i) \Delta \ln p(x_i) \right\} = - \int p(x) \ln p(x) dx$$

Differential entropy maximized (for fixed  $\sigma^2$ ) when

$$p(x) = \mathcal{N}(x|\mu, \sigma^2)$$

in which case

$$H[x] = \frac{1}{2} \{ 1 + \ln(2\pi\sigma^2) \} .$$

---

# Conditional Entropy

---

$$H[\mathbf{y}|\mathbf{x}] = - \iint p(\mathbf{y}, \mathbf{x}) \ln p(\mathbf{y}|\mathbf{x}) \, d\mathbf{y} \, d\mathbf{x}$$

$$H[\mathbf{x}, \mathbf{y}] = H[\mathbf{y}|\mathbf{x}] + H[\mathbf{x}]$$

---



# The Kullback-Leibler Divergence

---

$$\begin{aligned}\text{KL}(p\|q) &= - \int p(\mathbf{x}) \ln q(\mathbf{x}) \, d\mathbf{x} - \left( - \int p(\mathbf{x}) \ln p(\mathbf{x}) \, d\mathbf{x} \right) \\ &= - \int p(\mathbf{x}) \ln \left\{ \frac{q(\mathbf{x})}{p(\mathbf{x})} \right\} \, d\mathbf{x}\end{aligned}$$

$$\text{KL}(p\|q) \simeq \frac{1}{N} \sum_{n=1}^N \{ -\ln q(\mathbf{x}_n|\boldsymbol{\theta}) + \ln p(\mathbf{x}_n) \}$$

$$\text{KL}(p\|q) \geq 0$$

$$\text{KL}(p\|q) \neq \text{KL}(q\|p)$$

---

# Mutual Information

---

$$\begin{aligned} I[\mathbf{x}, \mathbf{y}] &\equiv \text{KL}(p(\mathbf{x}, \mathbf{y}) \| p(\mathbf{x})p(\mathbf{y})) \\ &= - \iint p(\mathbf{x}, \mathbf{y}) \ln \left( \frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} \right) d\mathbf{x} d\mathbf{y} \end{aligned}$$

$$I[\mathbf{x}, \mathbf{y}] = H[\mathbf{x}] - H[\mathbf{x}|\mathbf{y}] = H[\mathbf{y}] - H[\mathbf{y}|\mathbf{x}]$$

---

# HW1

---

1.1 1.2 1.3 1.6 1.10 1.17 1.25 1.30 1.40 1.41

Due: Oct 18

---