MACHINE LEARNING

CHAPTER 1: PRELIMINARY

Learning Objectives

1、What are pattern recognition and machine learning?

2、What are curve fitting and regularization?

3、What are ML and MAP Bayesian inferences?

4、How to deal with the curse of dimensionality?

5、What is the relationship between decision theory and machine learning?

6、What are generative and discriminative models?

7、How to use entropy、KL divergence and mutual information for machine learning?

Outlines

➢ Pattern Recognition and Machine Learning

➢ 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
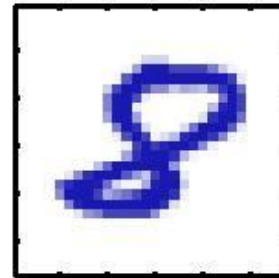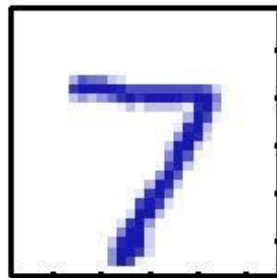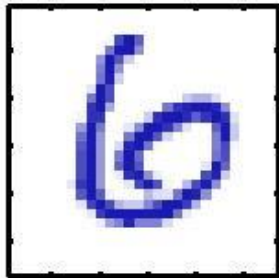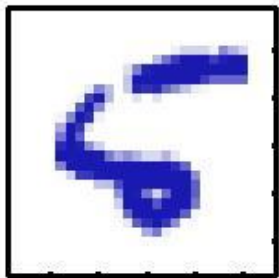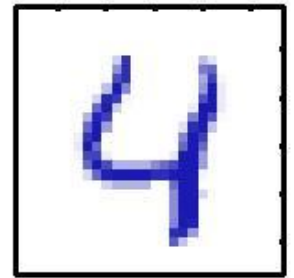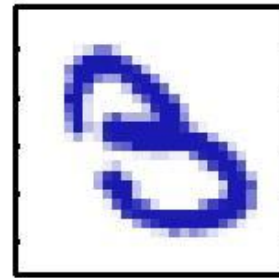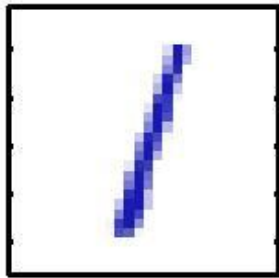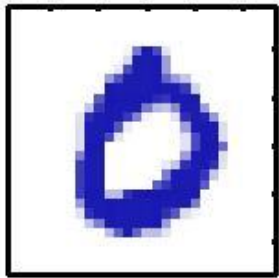
Machine Learning Pipeline

Dimension Reduction   Feature Modeling   Decision

| Data Set | → | Vector Space | → | Feature Space | → | Function Space | → | Decision Space |

Feature Selection   Boundary Modeling   Prediction
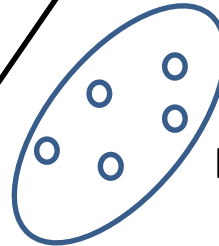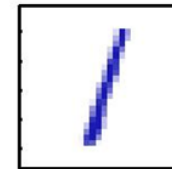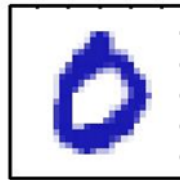
Feature Model

Boundary Model
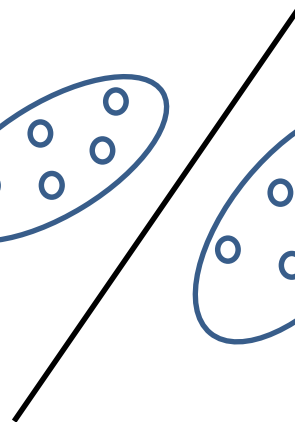
Deterministic Machine Learning

Process of Machine Learning:

$$\theta^* = \min_{\theta} E(f_\theta(x), y) + \lambda \Omega(\theta)$$

θ: model parameters
x: input data
y: output data
λ: weights
$f_\theta(\bullet)$: model
$E(\bullet)$: cost function
$\Omega(\theta)$: model constraints

Process of Prediction:

$$y' = f_{\theta*}(x')$$

x': new input data
y': new output data
$f_{\theta*}(\bullet)$: learned model

Probabilistic Machine Learning

Goal of Machine Learning：$p(\ real\ data\ ) \approx p(generated\ data\ |\ model,\ \theta\ )$

we assume：$p(\ training\ data\ ) \approx p(\ real\ data\ )$

Operation of Machine Learning：$p(training\ data\ ) \approx p(generated\ data\ |\ model,\ \theta\ )$

Process of Machine Learning：$\theta^* \approx \max_{\theta} p(training\ data\ |\ model,\ \theta)$

Process of Machine Learning：$\theta^* \approx \max_{\theta} p(\theta\ /training\ data,\ model)$

Process of prediction：$p\ (testing\ data\ |\ model,\ \theta^*)$

Bayesian Machine Learning

Process of Machine Learning:

$$p(\theta|training\ data, model) \propto p(training\ data\ /\ model, \theta)\ p_0(\theta)$$

posterior   likelihood   prior

Process of Prediction:

$$p(testing\ data\ |\ training\ data, model) =$$

$$\int p(testing\ data\ /\ model, \theta)\ p(\theta\ |\ training\ data, model)d\theta$$

Outlines

- Pattern Recognition and Machine Learning

- Curve Fitting and Regularization

- Probabilities and Gaussian Distributions

- Bayesian Inferences (ML and MAP)

- Curse of Dimensionality

- Decision Theory

- Entropy and Information

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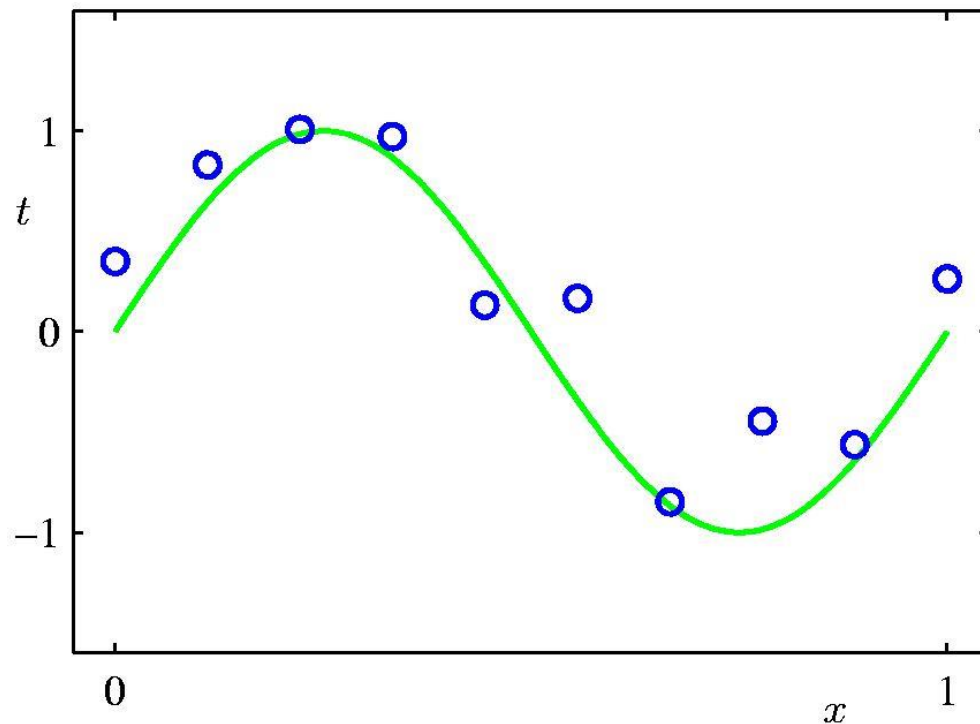
$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

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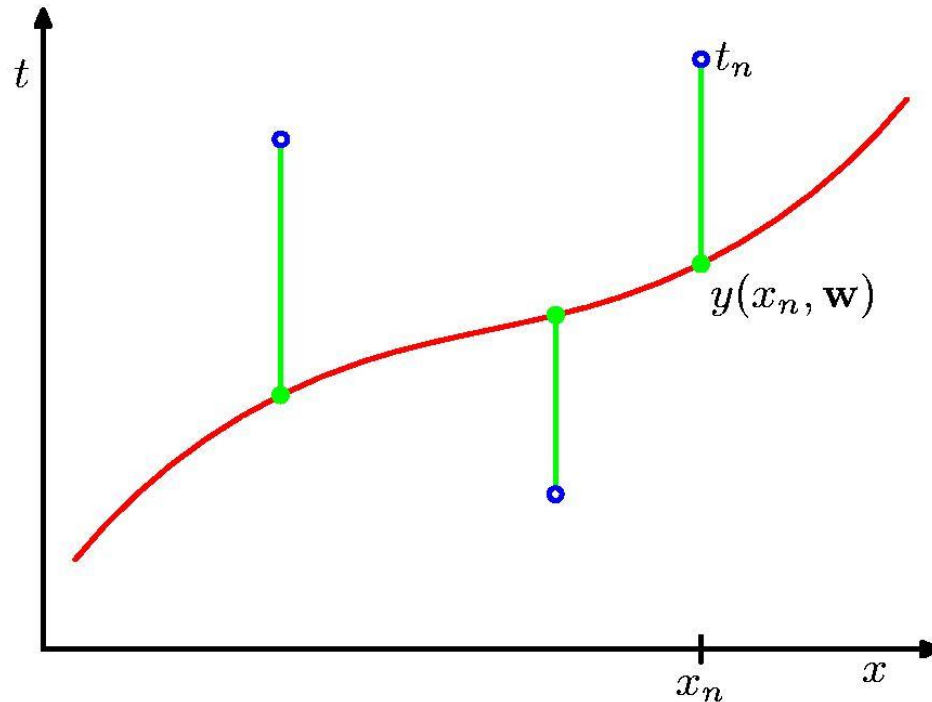
$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2$$

0th Order Polynomial

1ˢᵗ Order Polynomial
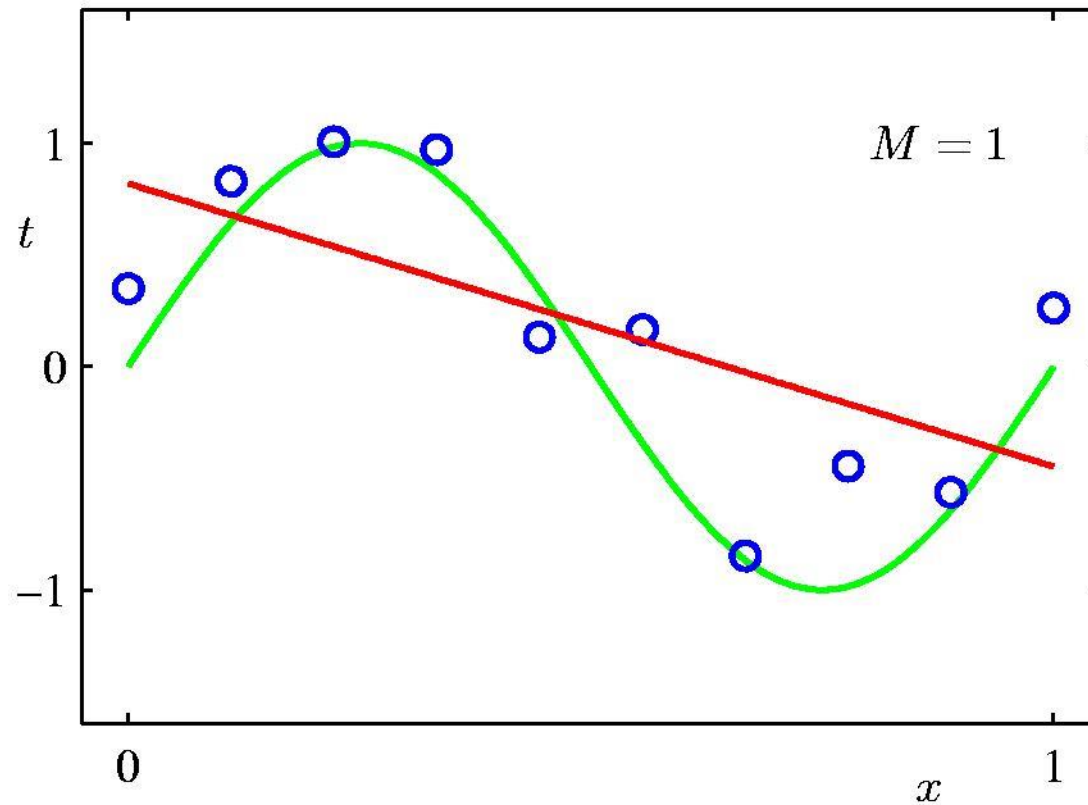
3rd Order Polynomial
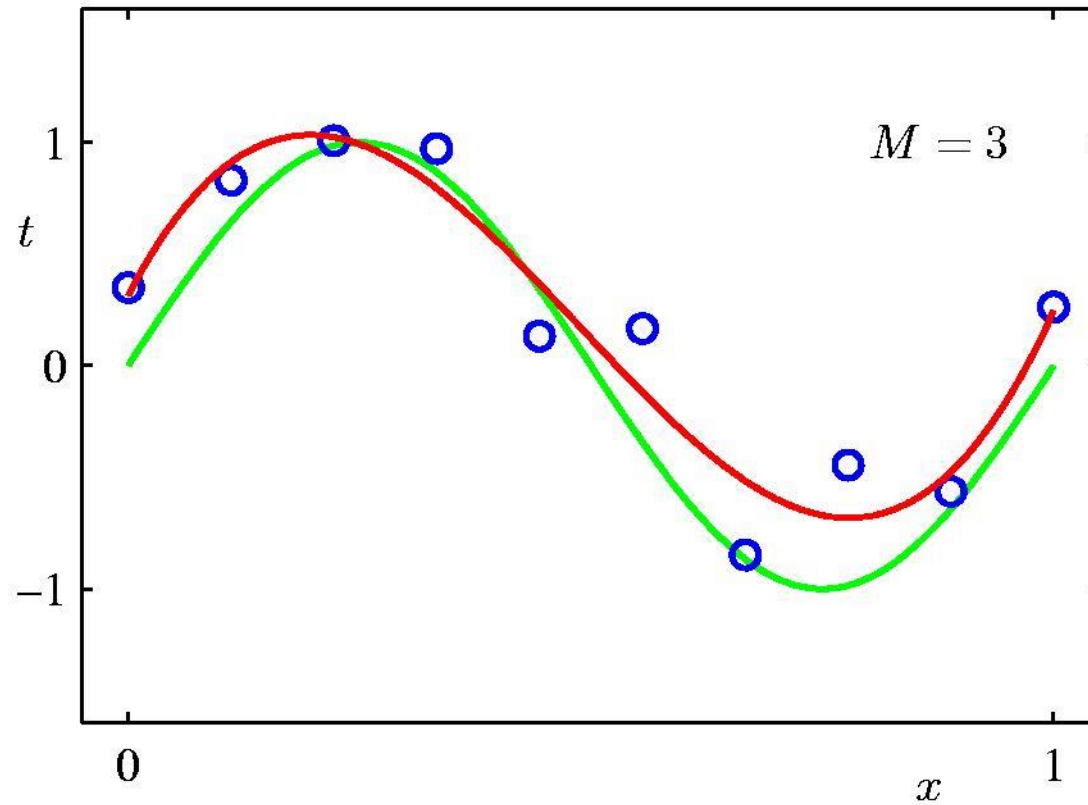
9th Order Polynomial
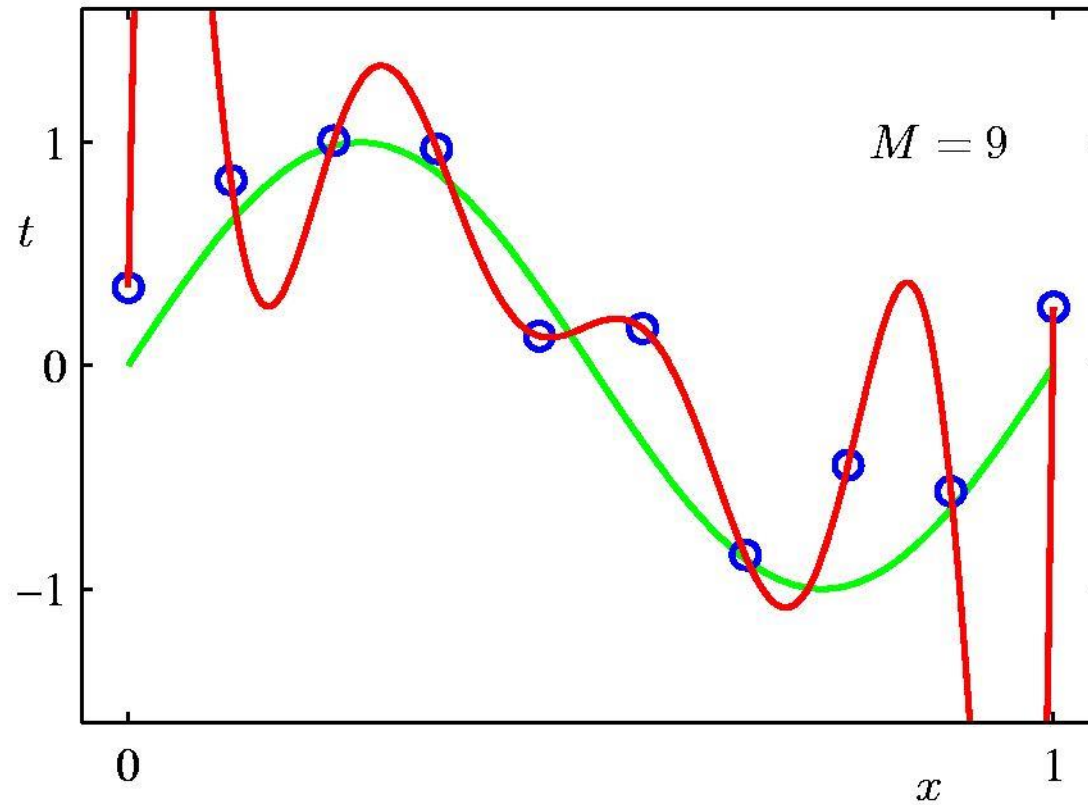
Over-fitting
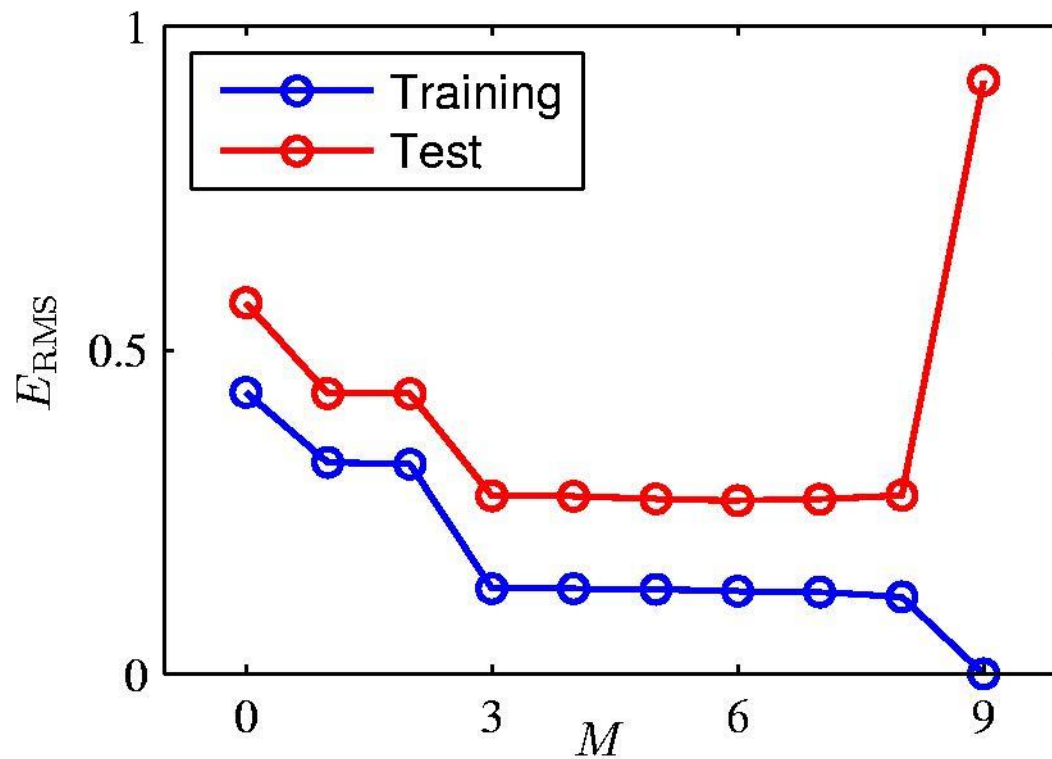
Root-Mean-Square (RMS) Error: $E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^\star)/N}$

Polynomial Coefficients

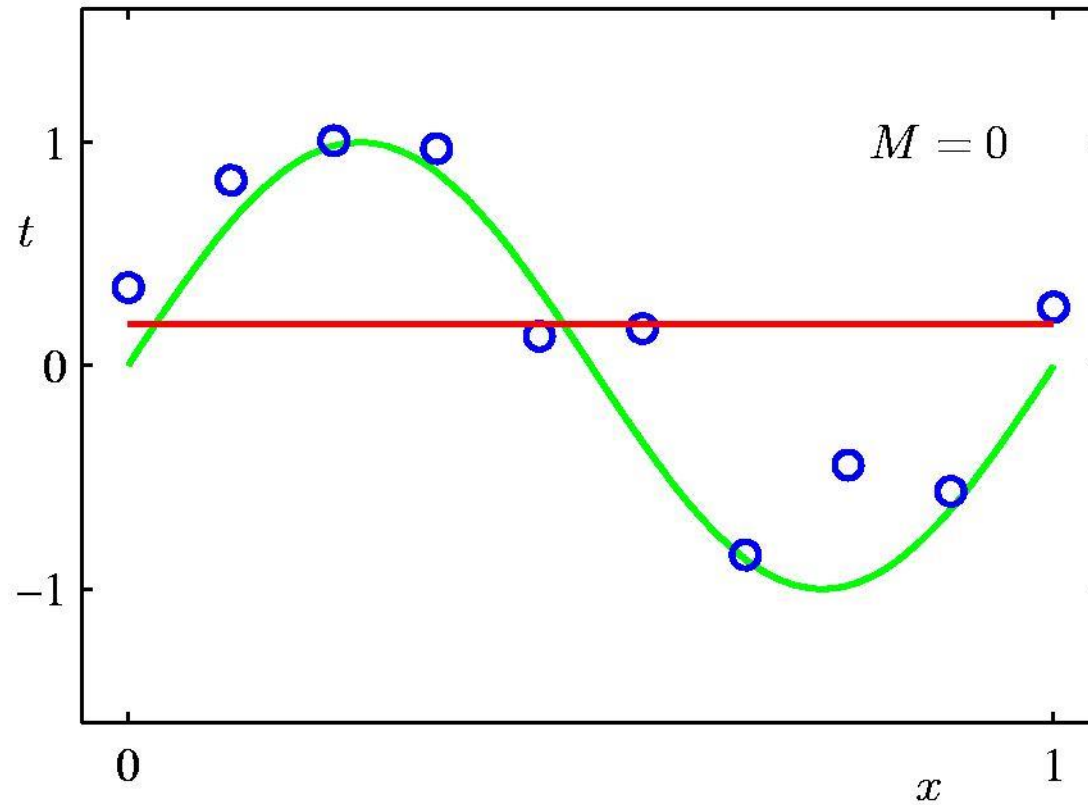|	$M = 0$	$M = 1$	$M = 3$	$M = 9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$			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

Data Set Size: $N = 15$

9th Order Polynomial
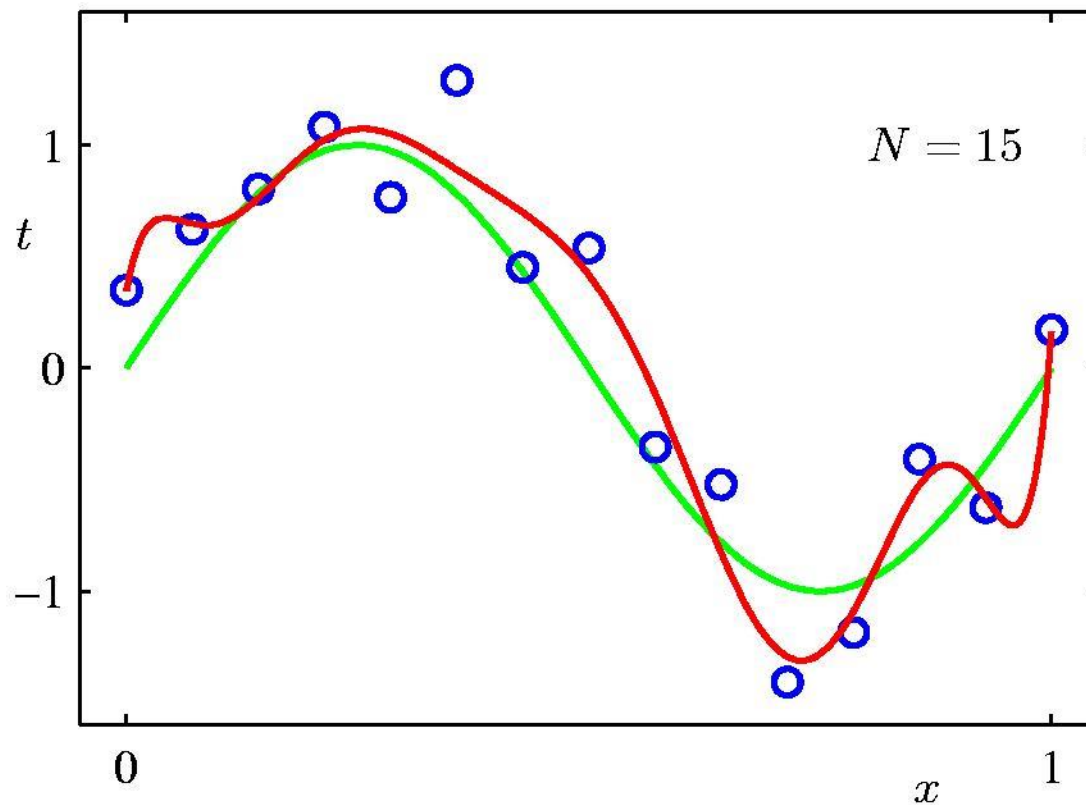
Data Set Size: $N = 100$

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: $\ln \lambda = -18$
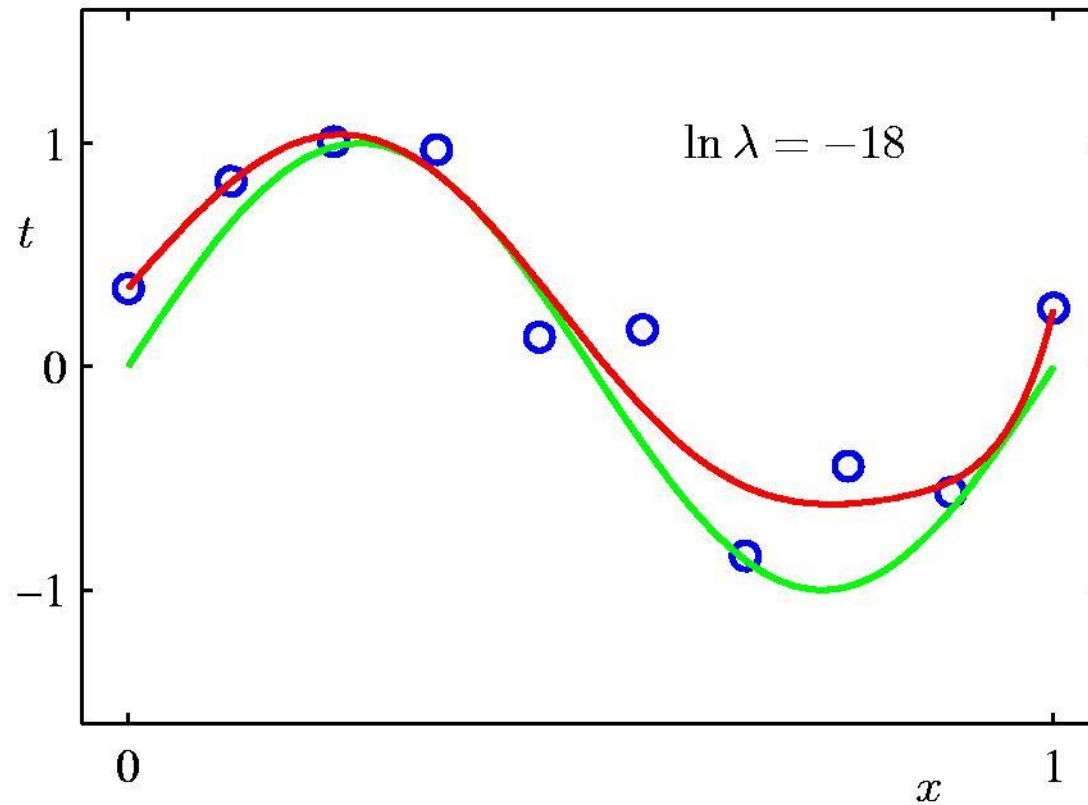
Regularization: $\ln \lambda = 0$

$\ln \lambda = 0$
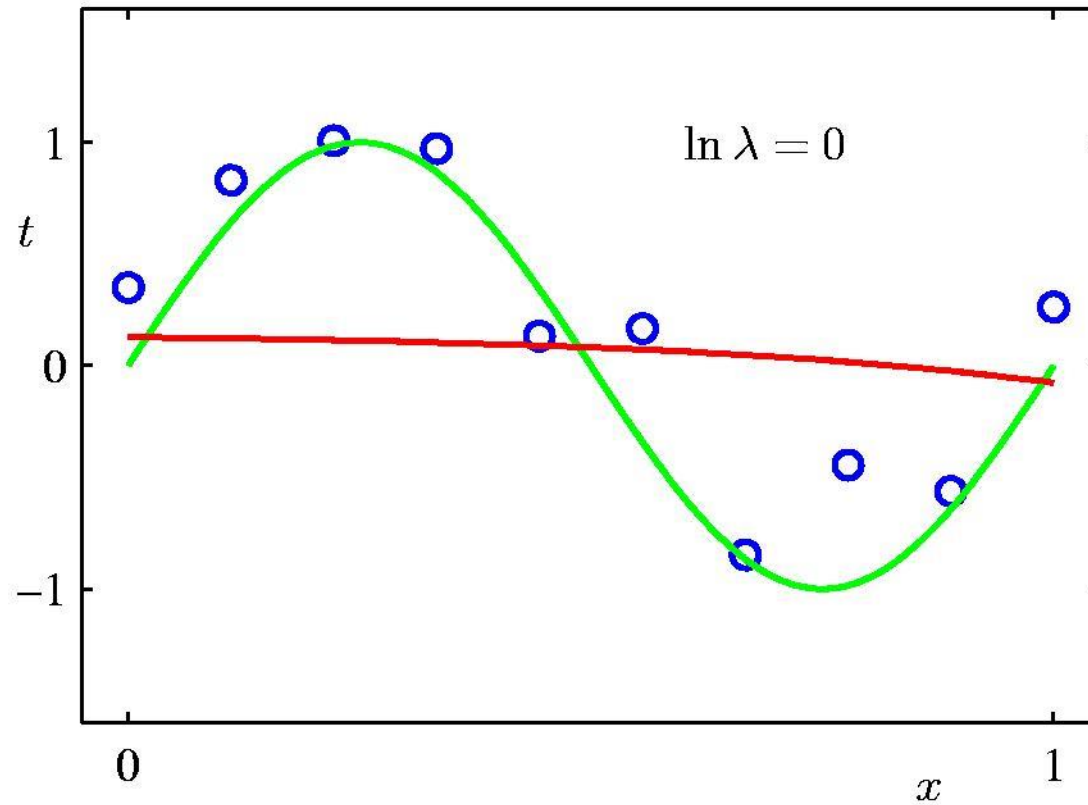
Regularization: E_{RMS} vs. $\ln \lambda$
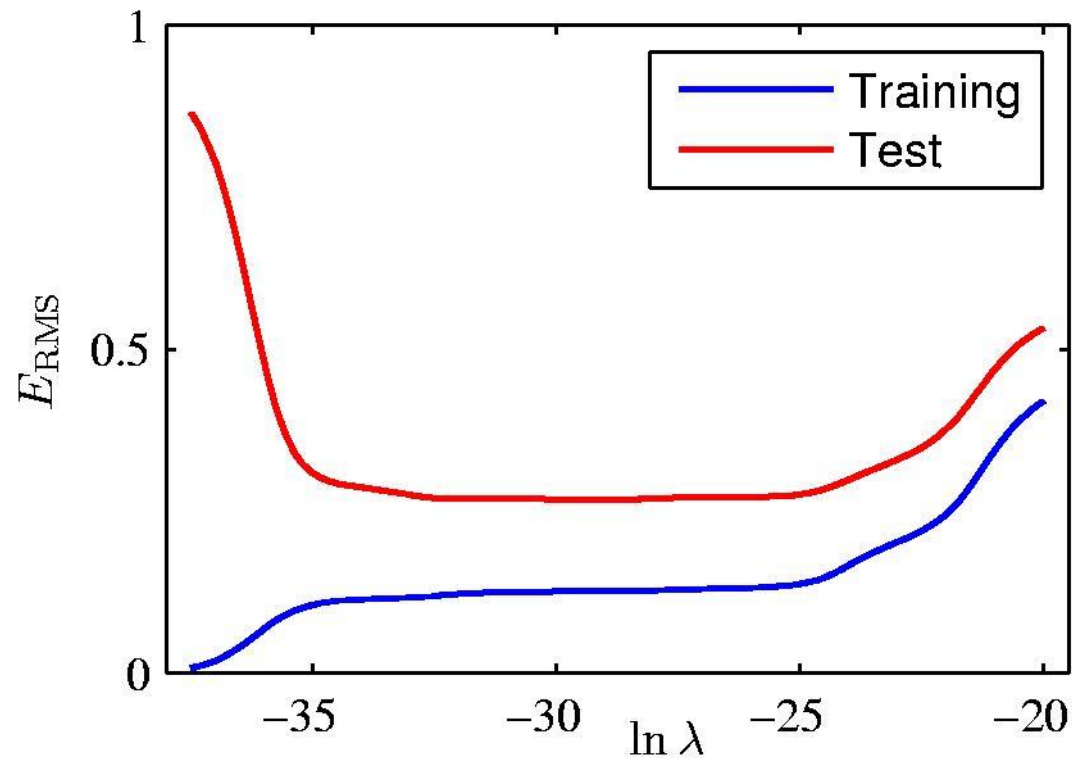
Polynomial Coefficients

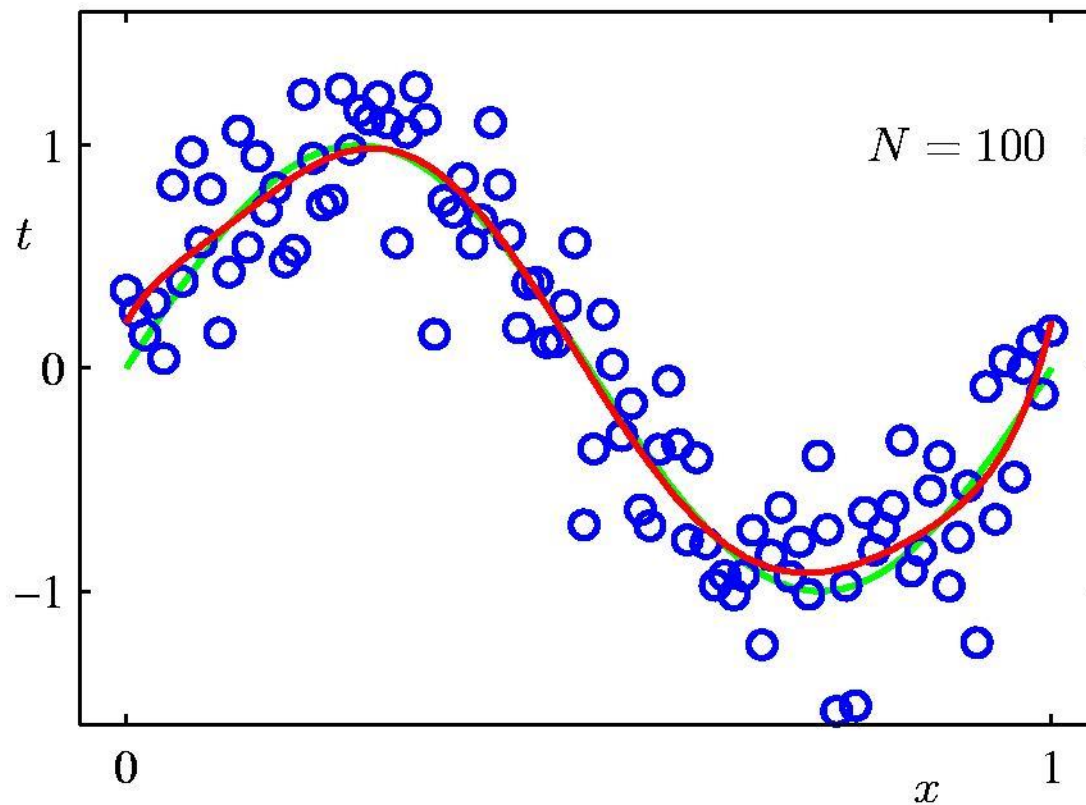|	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
$w_0^\star$	0.35	0.35	0.13
$w_1^\star$	232.37	4.74	-0.05
$w_2^\star$	-5321.83	-0.77	-0.06
$w_3^\star$	48568.31	-31.97	-0.05
$w_4^\star$	-231639.30	-3.89	-0.03
$w_5^\star$	640042.26	55.28	-0.02
$w_6^\star$	-1061800.52	41.32	-0.01
$w_7^\star$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

Outlines

➢ Pattern Recognition and Machine Learning

➢ 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
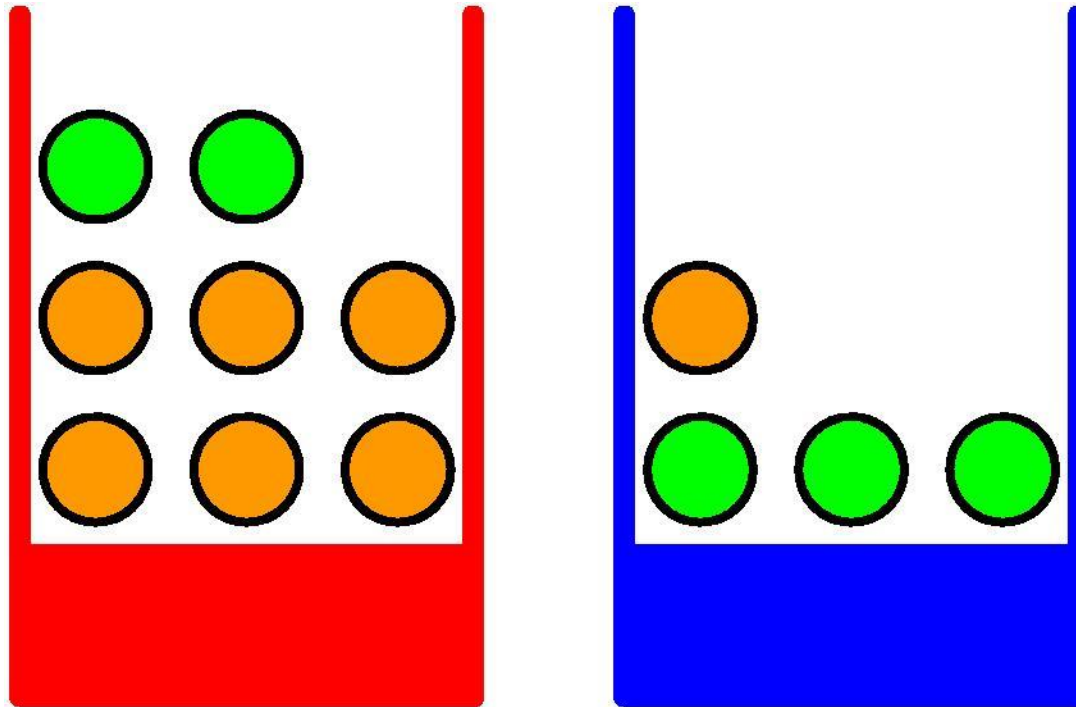
Probability Theory
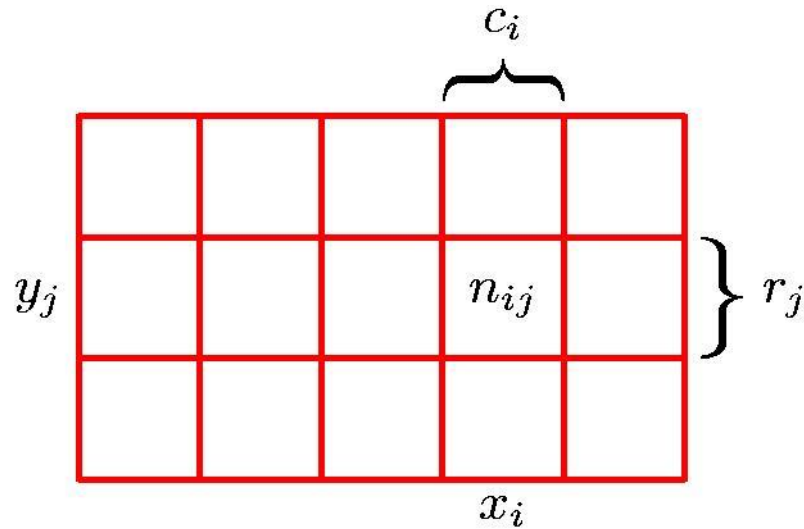
Marginal Probability

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

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
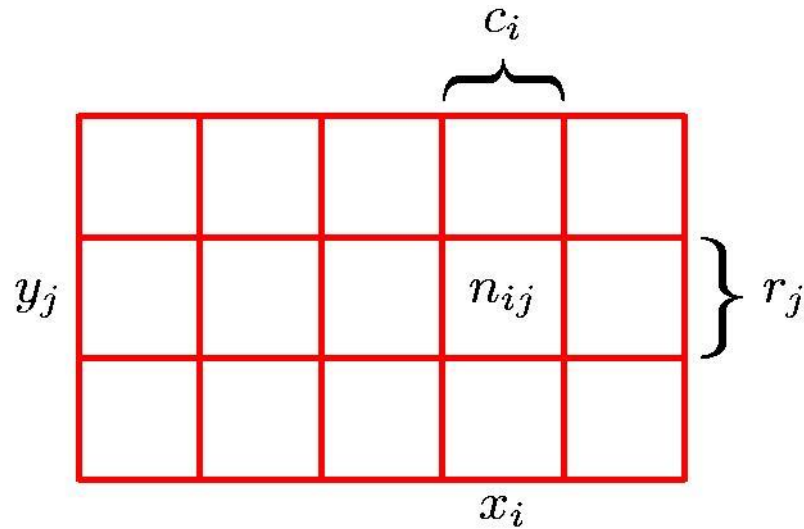
Sum Rule

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

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

Product Rule

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

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

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

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

$$p(X) = \sum_Y p(X|Y)p(Y) \quad : \text{normalization}$$

posterior $\propto$ likelihood $\times$ prior

$p(Y/X)$ $\qquad$ $p(X/Y)$ $\qquad$ $p(Y)$

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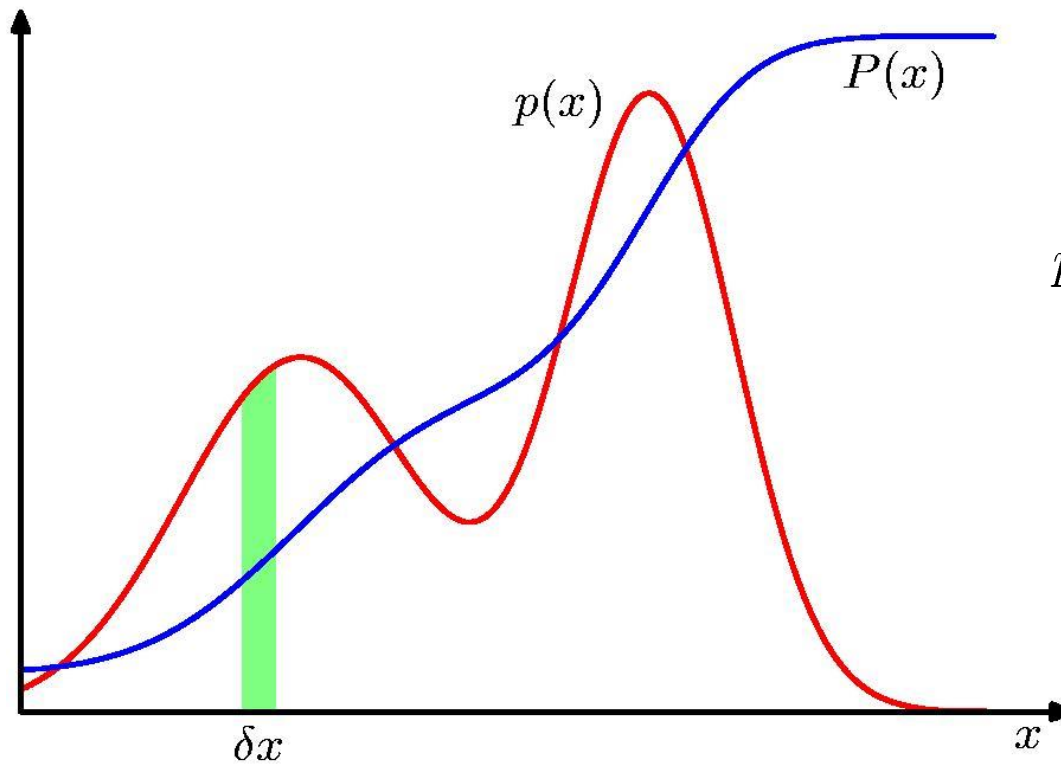
$$p(x \in (a, b)) = \int_a^b p(x)\, \mathrm{d}x$$

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

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

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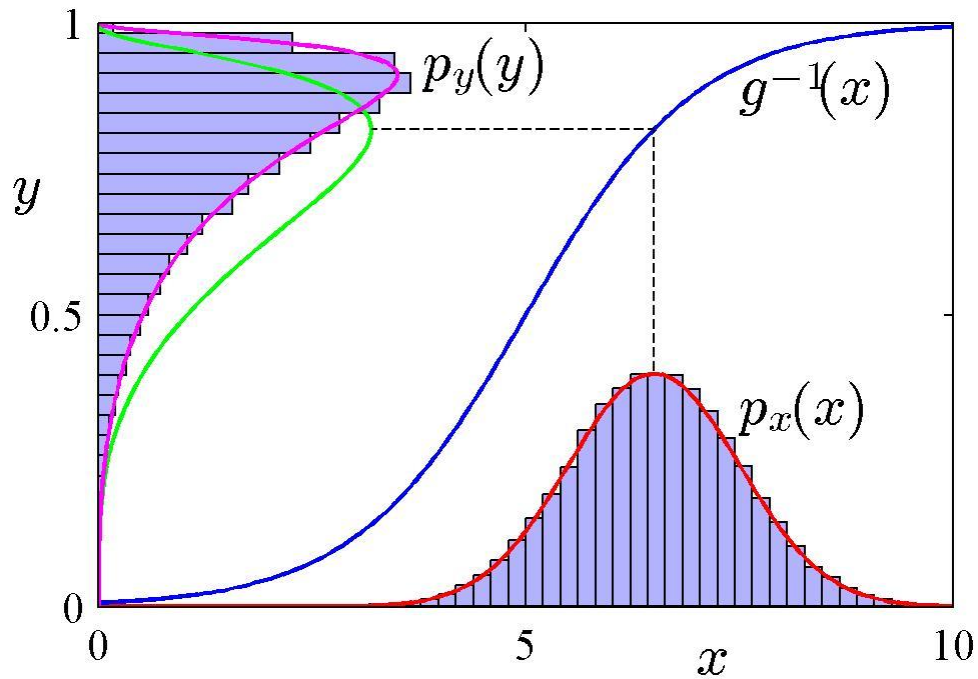
$$p_y(y) = p_x(x) \left| \frac{\mathrm{d}x}{\mathrm{d}y} \right|$$
$$= p_x(g(y)) \left| g'(y) \right|$$

$$x = g(y)$$

Expectations

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

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

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

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

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

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

The Gaussian Distribution

$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$
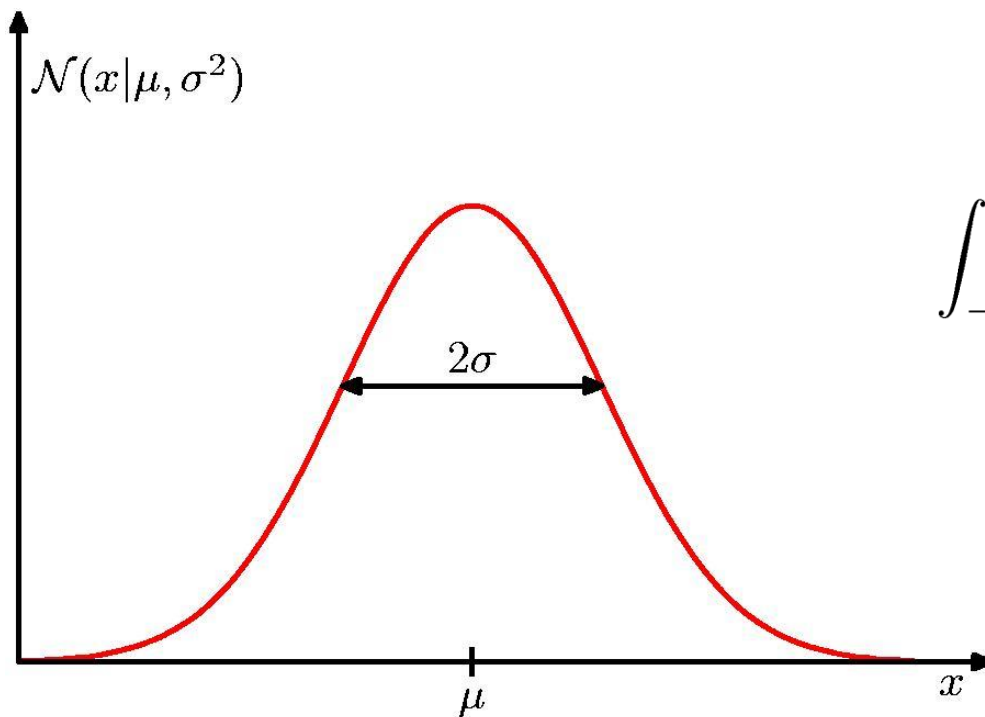
$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

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

Gaussian Mean and Variance

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

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

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

The Multivariate Gaussian

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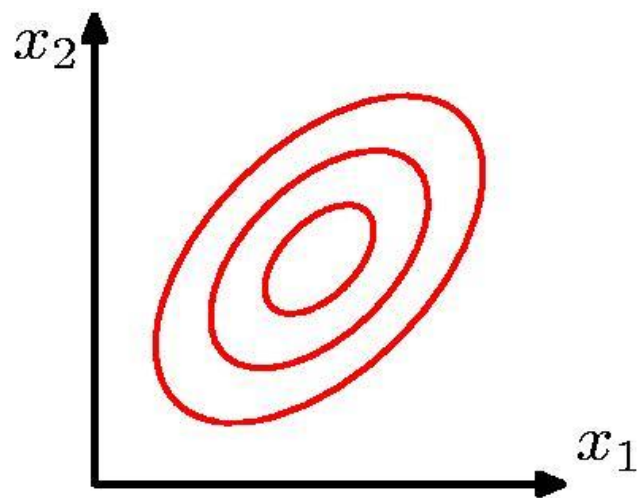
Outlines

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

Gaussian Parameter Estimation

Likelihood function
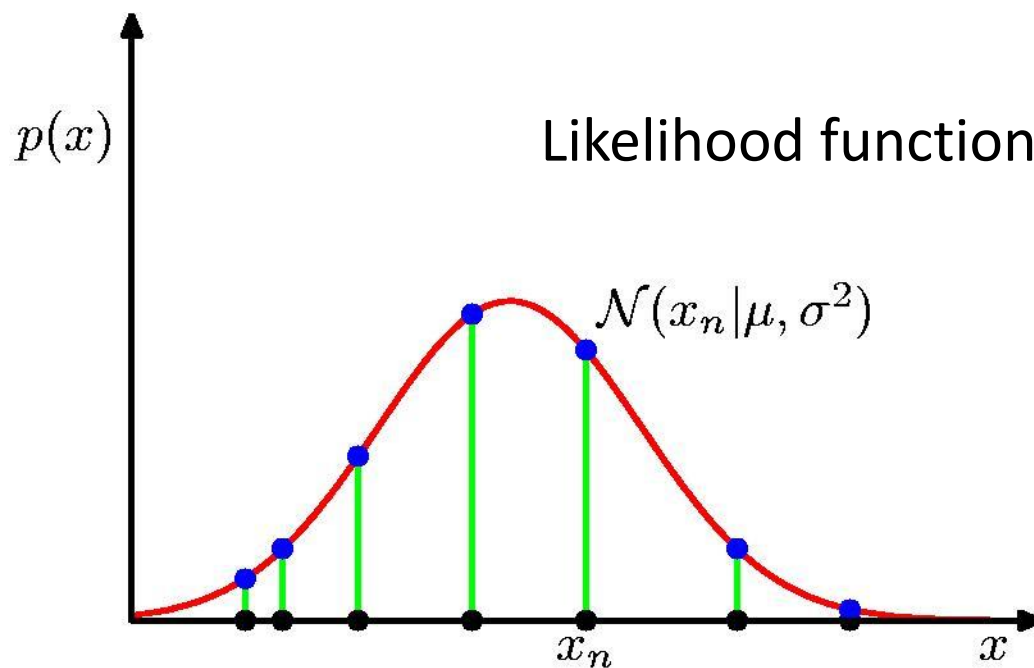
$$N(x_n|\mu, \sigma^2)$$

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

Maximum (Log) Likelihood

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

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

Properties of μ_{ML} and σ^2_{ML}

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

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

$$\widetilde{\sigma}^2 = \frac{N}{N-1}\sigma^2_{\mathrm{ML}}$$

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

(a)

(b)

(c)
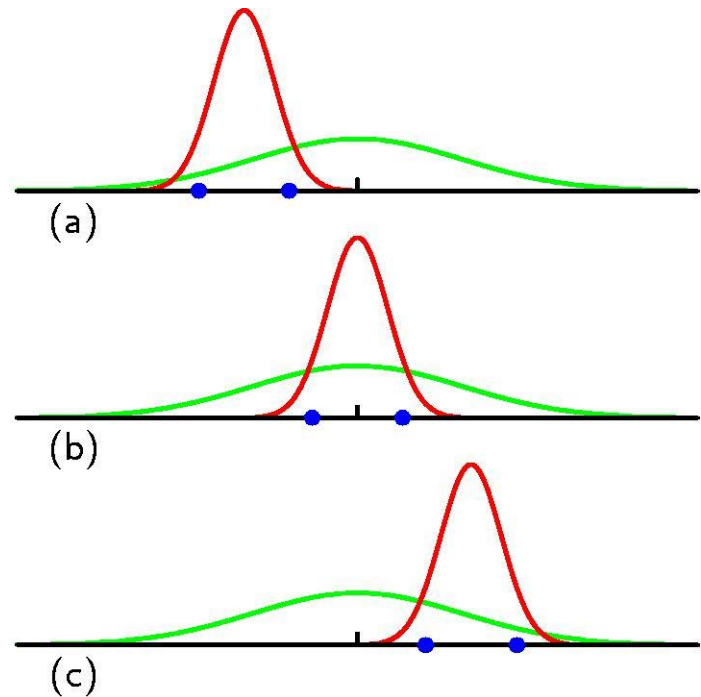
Curve Fitting Re-visited

(t, x): training data $\Rightarrow w, \beta$ (w, β, x_0): $\Rightarrow p(t/ x_0, w, \beta)$
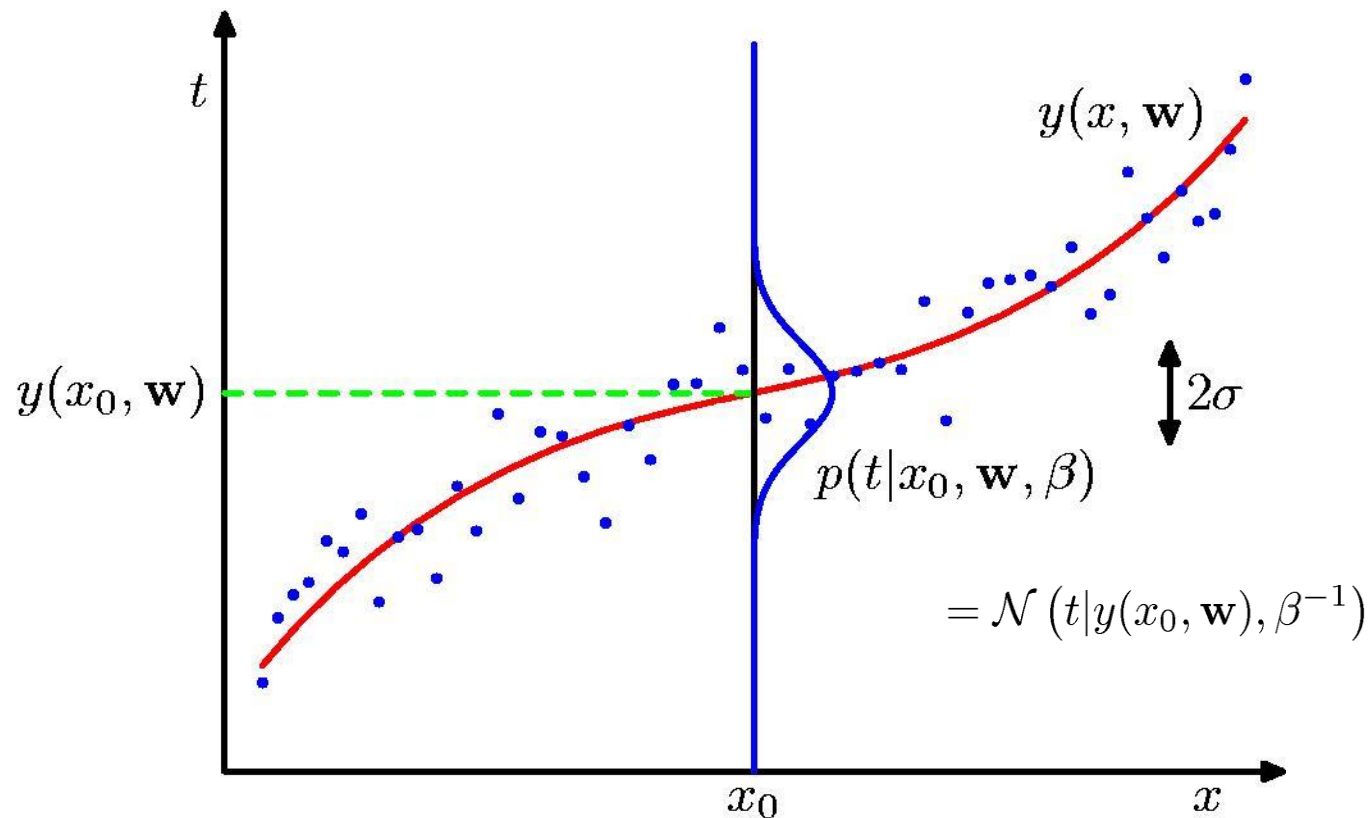
Maximum Likelihood

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

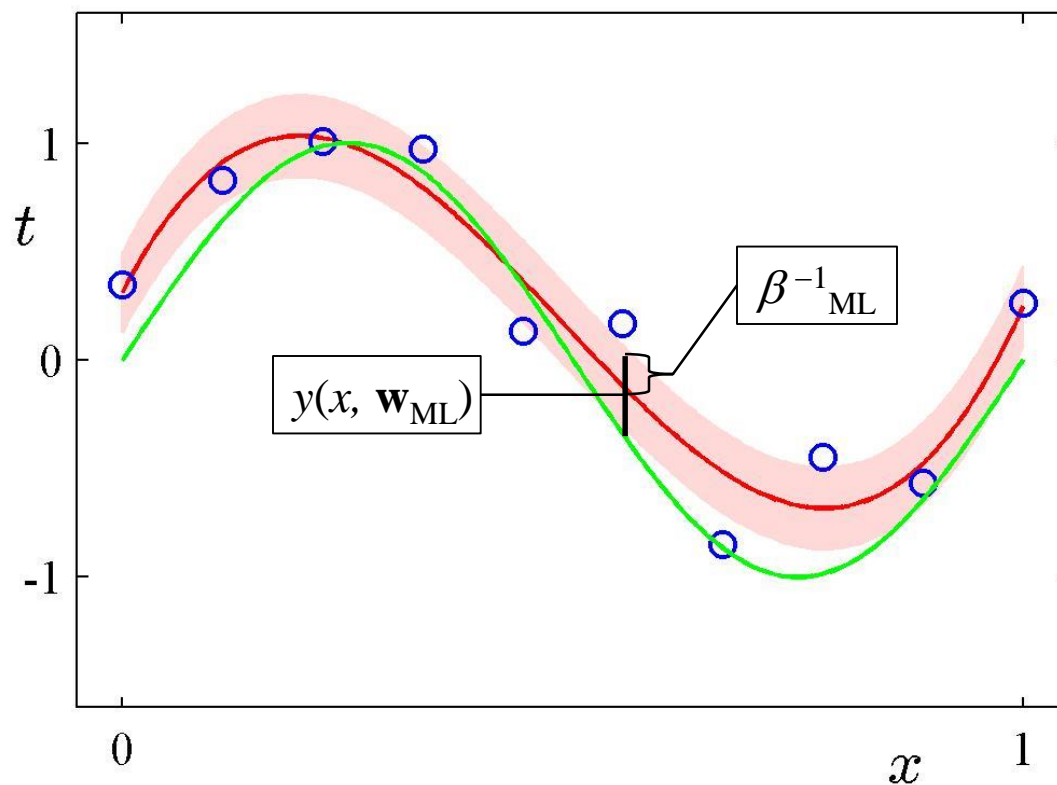$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = -\underbrace{\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2}_{\beta E(\mathbf{w})} + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi)$$

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

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

Predictive Distribution

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

MAP: A Step towards Bayes

MAP: Maximum *A Posteriori*

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

$$\boxed{posteriori} \longrightarrow \quad p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha) \longleftarrow \boxed{priori}$$

$$\boxed{likelihood}$$

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

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

Bayesian Curve Fitting

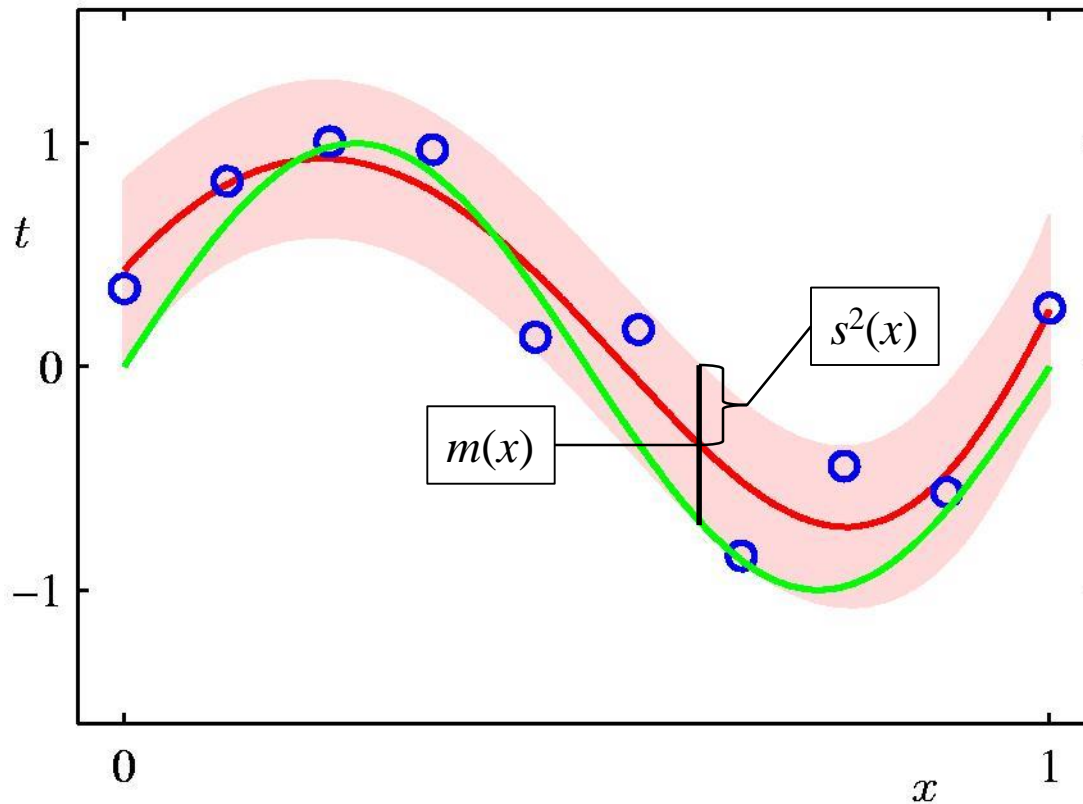$$p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|x, \mathbf{w})p(\mathbf{w}|\mathbf{x}, \mathbf{t})\,\mathrm{d}\mathbf{w} = \mathcal{N}\left(t|m(x), s^2(x)\right)$$

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

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

We will go through more details in a later lecture.

Bayesian Predictive Distribution

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

Model Selection and Evaluation

Cross-Validation