Introduction to Machine Learning

Linear Models for Regression

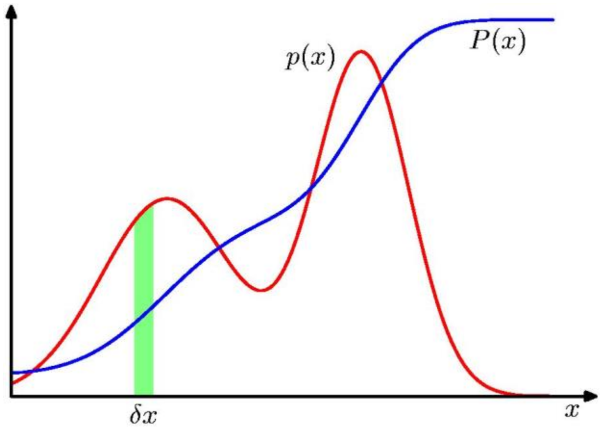
SHENG-JYH WANG

NATIONAL YANG MING CHIAO TUNG UNIVERSITY, TAIWAN

SPRING, 2024

Prerequisite Knowledge

Probability Density Function



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

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

$$p(x) \geqslant 0$$

probability density function (pdf)

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

cumulative distribution function (cdf)

Sum & Product Rules

- Joint pdf $p(X, Y)$

- Sum Rule
$$p(X) = \int p(X, Y) dY$$
$$p(X) = \sum_Y p(X, Y)$$

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

Expectation

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

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

Conditional Expectation

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

Variance & Covariance

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

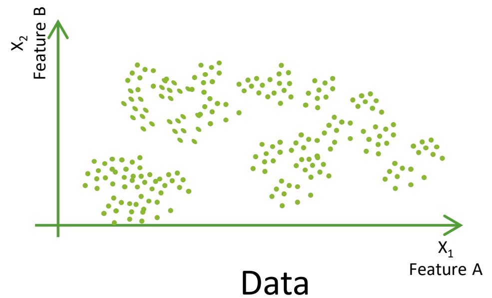
$$\mathrm{var}[f] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2.$$

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

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

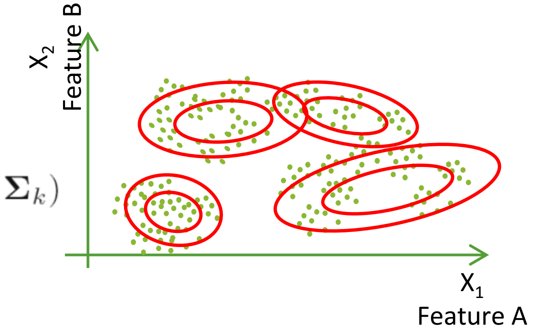
$$\begin{aligned} \mathrm{cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\}\{\mathbf{y}^{\mathrm{T}} - \mathbb{E}[\mathbf{y}^{\mathrm{T}}]\}\right] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^{\mathrm{T}}].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

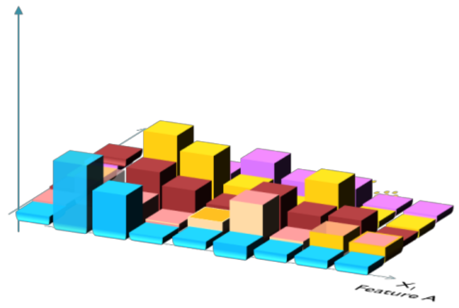
Probability Distribution



Data

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$





- Parameterized Probability Distributions
  - ✓ have a fixed number of parameters.
  - ✓ Have a few assumptions about the distribution form

- Non-parametric Models
  Not based on parameterized families of probability distributions

Commonly Used Parameterized Distributions

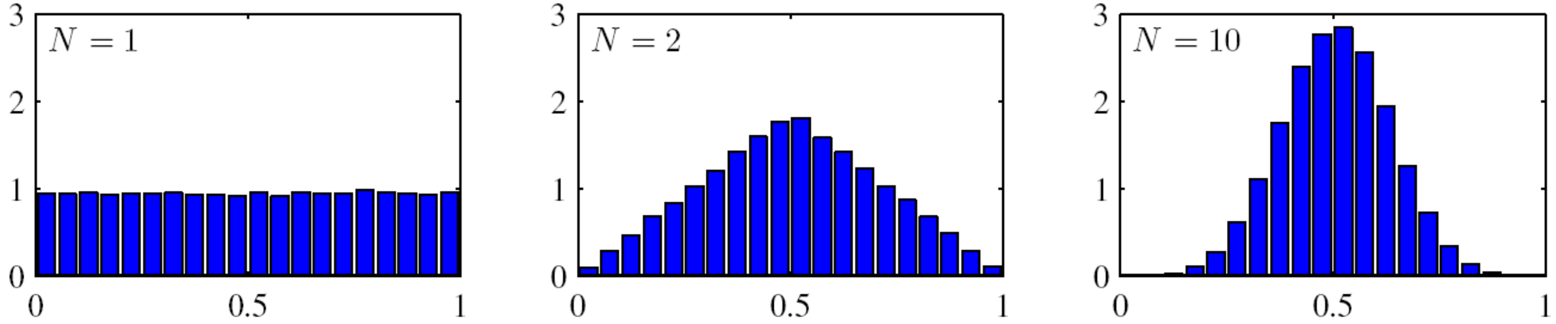
- **Regression Problem**
  - ✓ **Gaussian** Distribution
  - ✓ Conjugate Prior: **Gaussian** Distribution, **Wishart** Distribution, **Gaussian-Wishart** Distribution, **Gamma** Distribution
  - ✓ Related Distribution: **Student's t-distribution**
- **Binary Classification Problem**
  - ✓ **Bernoulli Distribution**, **Binomial Distribution**
  - ✓ Conjugate Prior: **Beta Distribution**
- **Multi-class Classification Problem**
  - ✓ **Multinomial Distribution**
  - ✓ Conjugate Prior: **Dirichlet Distribution**

Central Limit Theorem

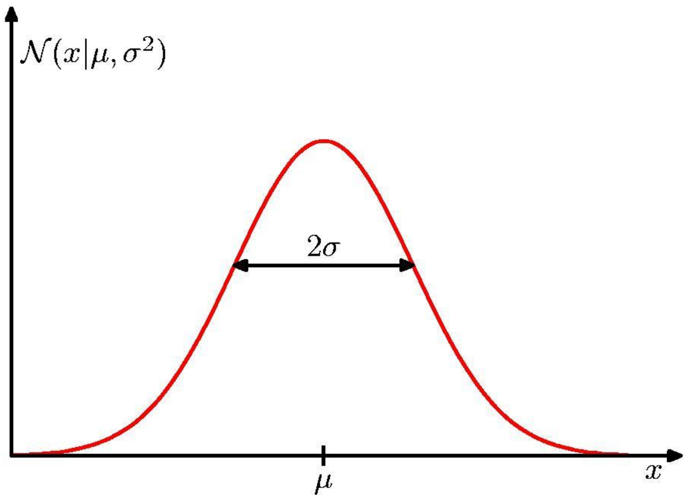
Subject to certain mild conditions, the sum of a set of random variables has a distribution that becomes increasingly Gaussian as the number of random variables increases.

$$Y = X_1 + X_2 + \cdots + X_K$$

Example: Average of N uniformly distributed random variables.



Gaussian Distribution (1/6)



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

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

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

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

Gaussian Distribution (2/6)

D-dimensional Multivariate Gaussian

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

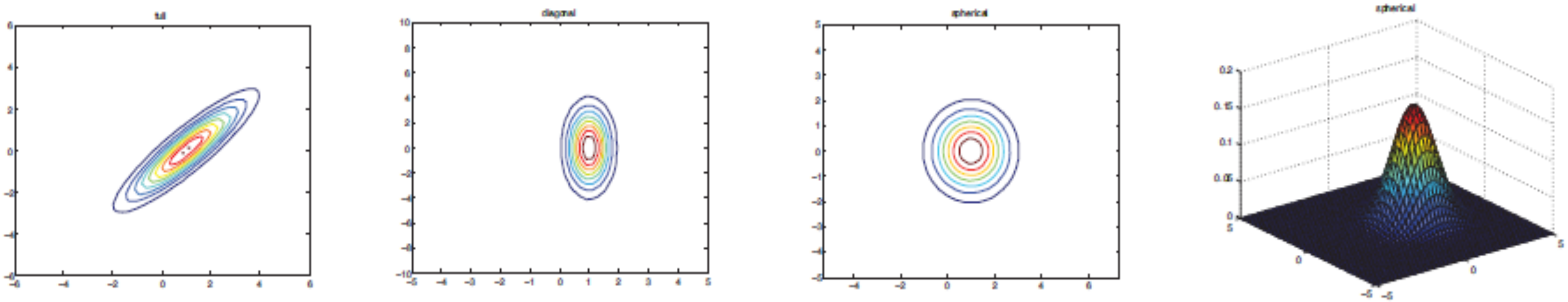
Remarks:

1. $\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$

   Δ: the *Mahalanobis distance* from μ to ***x*.**

2. Σ: **Covariance Matrix** $\Lambda = \Sigma^{-1}$: **Precision Matrix**

Gaussian Distribution (3/6)



(Ref: Murphy, “Machine Learning: A Probabilistic Perspective”)

Gaussian Distribution (4/6)

Moments of Multivariate Gaussian

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$$

$$\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] = \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma}.$$

$$\begin{aligned}\operatorname{cov}[\mathbf{x}] &= \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^{\mathrm{T}}\right] \\ &= \boldsymbol{\Sigma}\end{aligned}$$

Gaussian Distribution (5/6)

Suppose **x** is a *D*-dimensional vector with Gaussian distribution $N(\mathbf{x} \mid \mu, \Sigma)$ and we partition **x** into two disjoint subsets $\mathbf{x}_a$ and $\mathbf{x}_b$.

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}$$

$$\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix} \qquad \begin{aligned} \boldsymbol{\Lambda}_{aa} &= (\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba})^{-1} \\ \boldsymbol{\Lambda}_{ab} &= -(\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba})^{-1}\boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1} \end{aligned}$$

$$p(x_a \mid x_b) = N(x_a \mid \mu_{a|b}, \Sigma_{a|b})$$ **Conditional Gaussian Distribution**

$$\begin{aligned} \mu_{a|b} &= \Sigma_{a|b}\{\Lambda_{aa}\mu_a - \Lambda_{ab}(x_b - \mu_b)\} \\ &= \mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(x_b - \mu_b) \\ &= \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b) \end{aligned} \qquad \Sigma_{a|b} = \Lambda_{aa}^{-1} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$$

Gaussian Distribution (6/6)

Marginal Gaussian Distribution

$$p(\mathbf{x}_a) = \int p(\mathbf{x}_a, \mathbf{x}_b)\,\mathrm{d}\mathbf{x}_b \qquad \begin{aligned} \mathbb{E}[\mathbf{x}_a] &= \boldsymbol{\mu}_a \\ \operatorname{cov}[\mathbf{x}_a] &= \boldsymbol{\Sigma}_{aa} \end{aligned}$$

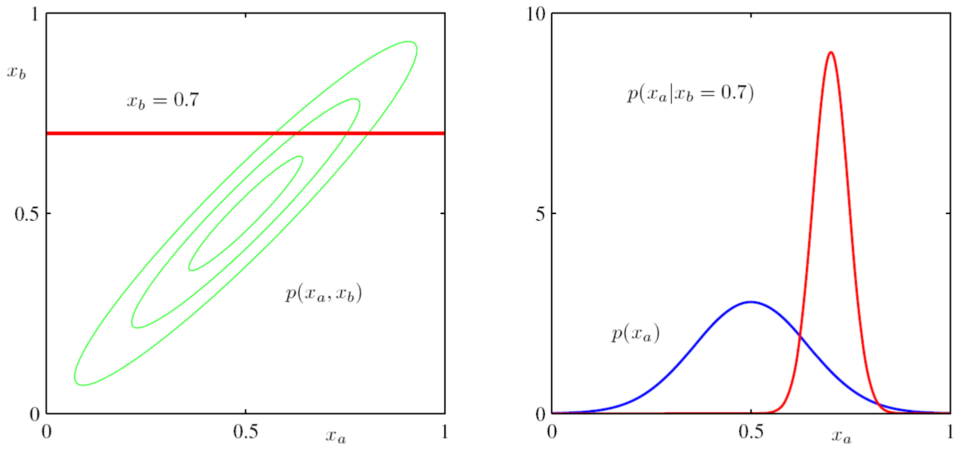
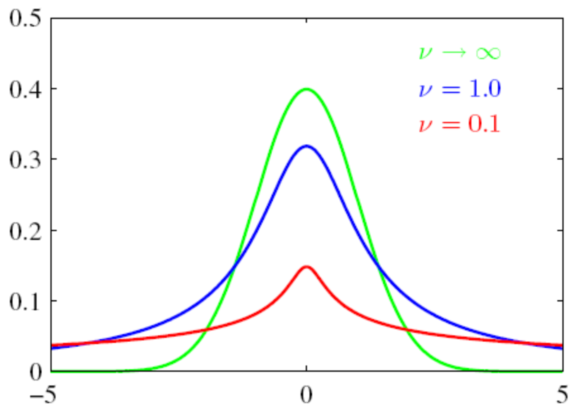


Figure 2.9 The plot on the left shows the contours of a Gaussian distribution $p(x_a, x_b)$ over two variables, and the plot on the right shows the marginal distribution $p(x_a)$ (blue curve) and the conditional distribution $p(x_a|x_b)$ for $x_b = 0.7$ (red curve).

Student's t-Distribution (1/2)

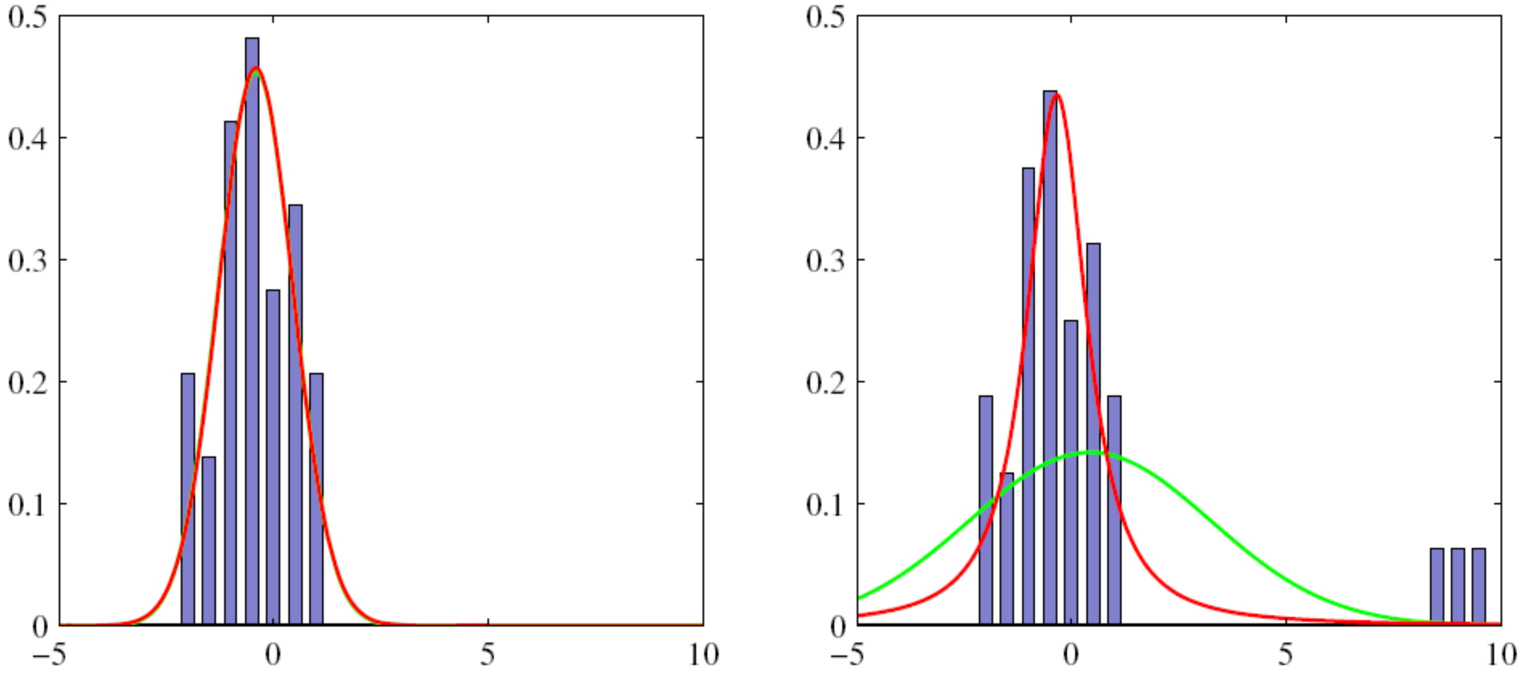
- or simply called **the t-distribution**
- developed by William Sealy Gosset under the pseudonym "Student".
- If we have a univariate Gaussian *N*(*x*|μ, τ^{-1}) together with a Gamma prior Gam(τ/*a*,*b*),

$$
\begin{aligned}
p(x|\mu, a, b) &= \int_0^\infty \mathcal{N}(x|\mu, \tau^{-1})\mathrm{Gam}(\tau|a, b)\,\mathrm{d}\tau \qquad (2.158)\\
&= \int_0^\infty \frac{b^a e^{(-b\tau)}\tau^{a-1}}{\Gamma(a)} \left(\frac{\tau}{2\pi}\right)^{1/2} \exp\left\{-\frac{\tau}{2}(x-\mu)^2\right\} \mathrm{d}\tau \\
&= \frac{b^a}{\Gamma(a)} \left(\frac{1}{2\pi}\right)^{1/2} \left[b + \frac{(x-\mu)^2}{2}\right]^{-a-1/2} \Gamma(a+1/2)
\end{aligned}
$$



Student’s t-Distribution (2/2)

Student’s t-distribution has longer tails than a Gaussian. $\Rightarrow$ Less sensitive to outliers.



Bayes' theorem

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

w: parameters, *D*: data

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}$$

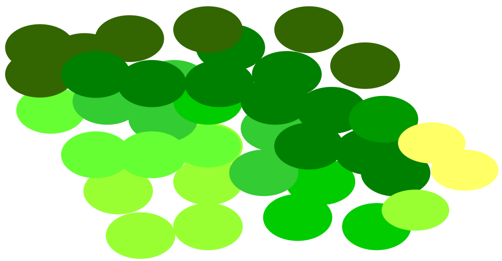
$p(\mathbf{w})$: prior probability

$p(D|\mathbf{w})$: likelihood function of **w**

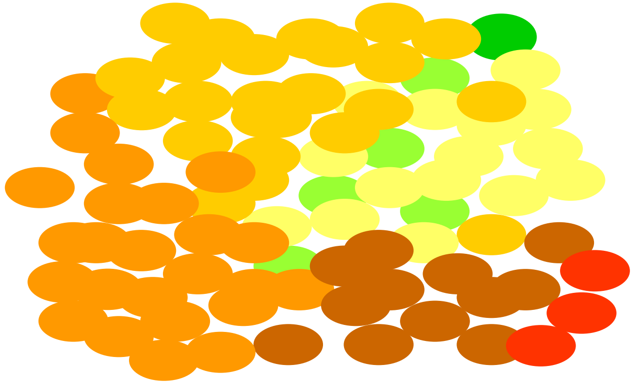
$p(\mathbf{w}|D)$: posterior probability

posterior ∝ likelihood × prior

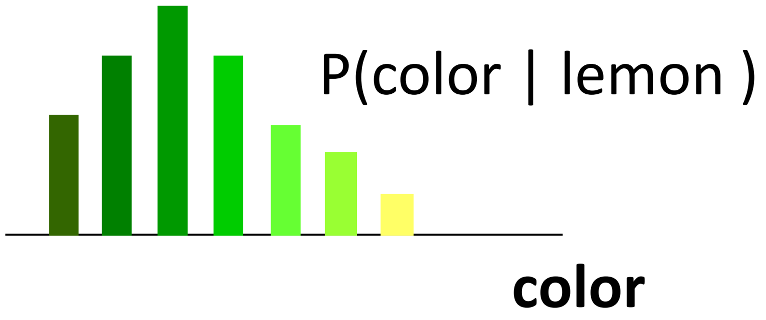
Bayesian Inference (1/5)

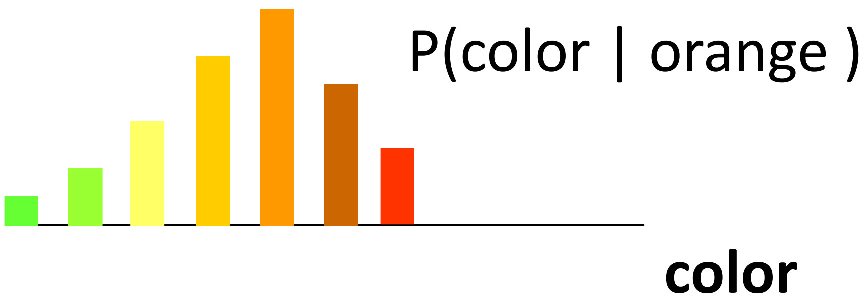


lemon



orange



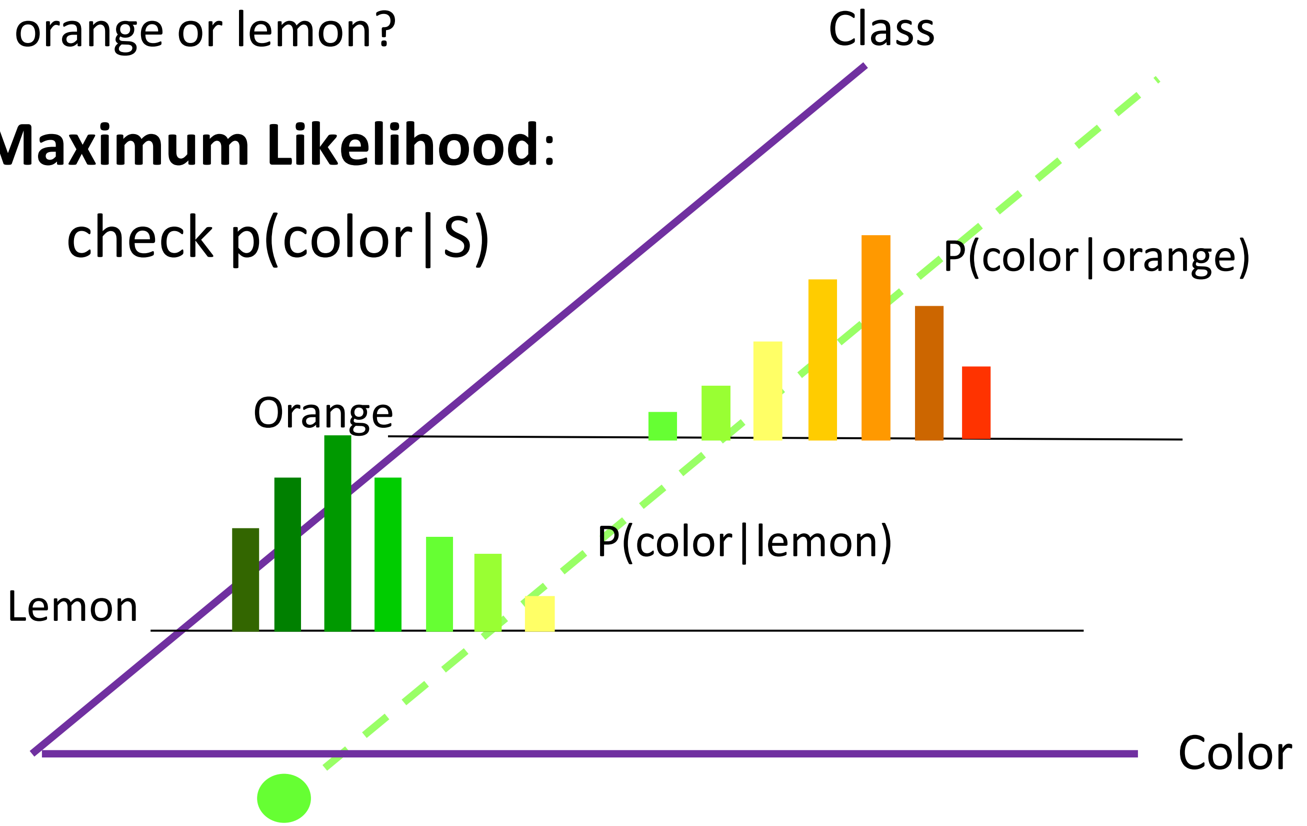


Bayesian Inference (2/5)

color:
S: orange or lemon?

- **Maximum Likelihood**:

  check p(color|S)



Bayesian Inference (3/5)

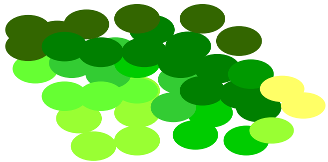
color: 

S: orange or lemon?

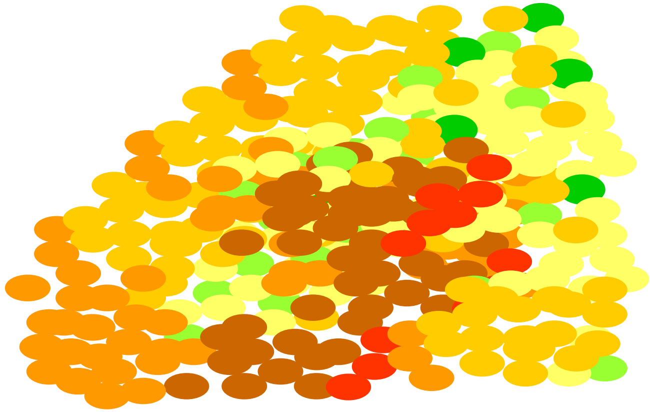
- **Maximum A Posteriori:**

check $p(S|\text{color}) = \frac{p(S,\text{color})}{p(\text{color})} = \frac{p(\text{color}|S)p(S)}{p(\text{color})}$

$\propto p(\text{color}|S)p(S)$



lemon



orange

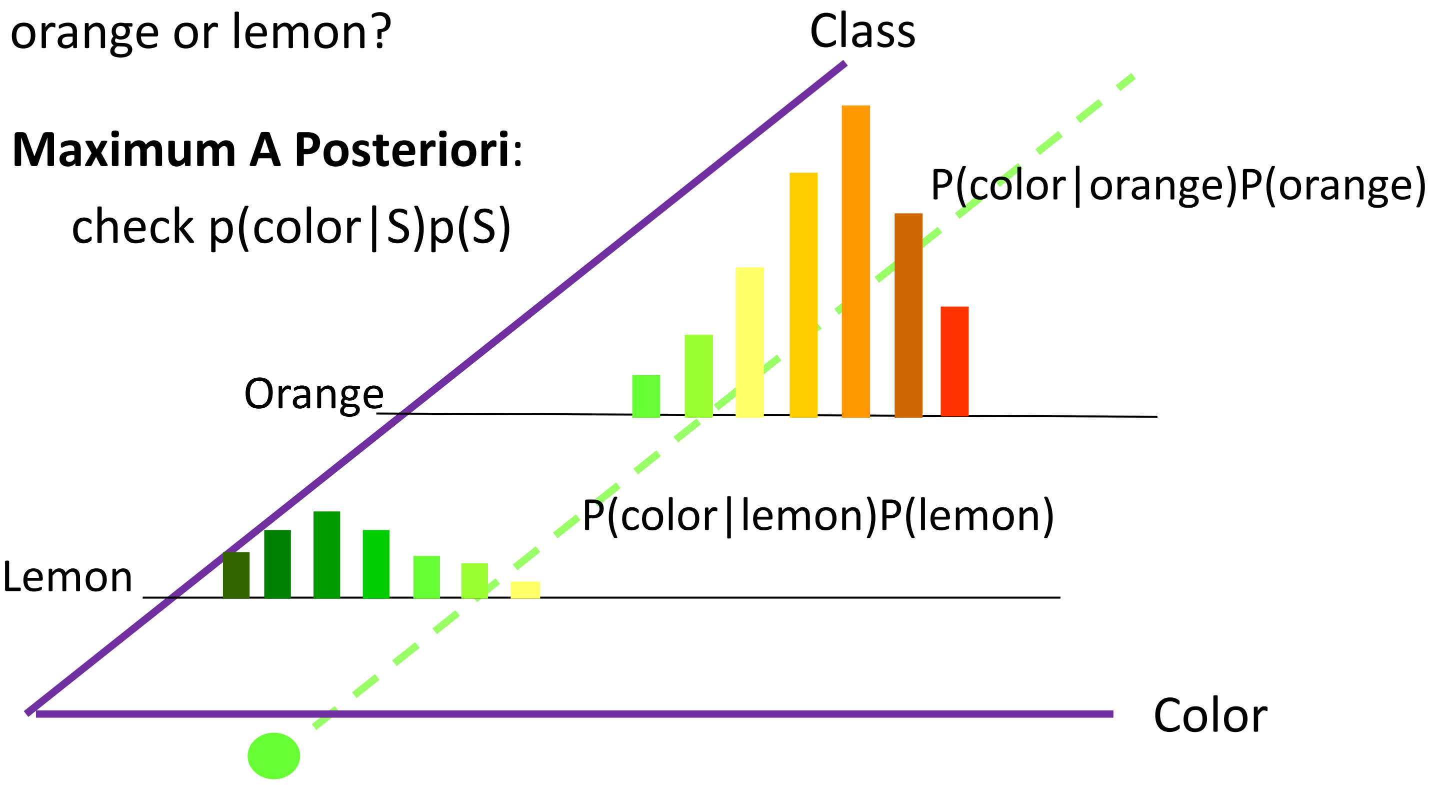
Bayesian Inference (4/5)

color:

S: orange or lemon?

- **Maximum A Posteriori**:

  check p(color|S)p(S)



Bayesian Inference (5/5)

Assume we have a model $p(X|\theta)$ and we have a set of observation $\{X_1, X_2, \ldots., X_N\}$.

$$p(\theta|X_1, X_2, \ldots., X_N) = \frac{p(X_1, X_2, \ldots., X_N|\theta)p(\theta)}{p(X_1, X_2, \ldots., X_N)}$$

Posterior Probability

$$\propto p(X_1, X_2, \ldots., X_N|\theta)p(\theta)$$

Likelihood Function Prior Probability

$$p(X_1, X_2, \ldots., X_N|\theta) = \prod_{n=1}^{N} p(X_n|\theta)$$

Bayes' Theorem for Gaussian Variables

Given

$$
\begin{aligned}
p(\mathbf{x}) &= \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right) \\
p(\mathbf{y}|\mathbf{x}) &= \mathcal{N}\left(\mathbf{y}|\mathbf{A}\mathbf{x}+\mathbf{b}, \mathbf{L}^{-1}\right)
\end{aligned}
$$

we have

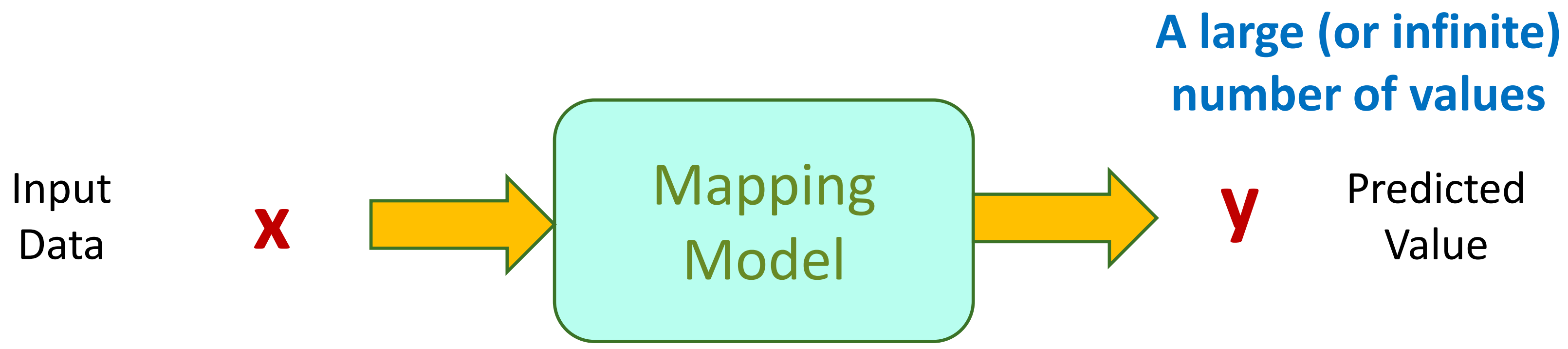
$$
\begin{aligned}
p(\mathbf{y}) &= \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu}+\mathbf{b}, \mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}) \\
p(\mathbf{x}|\mathbf{y}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y}-\mathbf{b})+\boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma})
\end{aligned}
$$

where

$$
\boldsymbol{\Sigma} = (\boldsymbol{\Lambda}+\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}
$$

Regression Problem

Linear Model for Regression

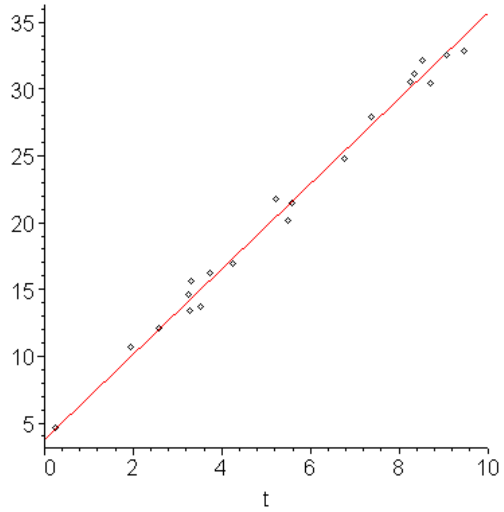


Linear Model

Linear Model for Regression

Given a training data set comprising N observations $\{\mathbf{x}_n\}$ and the corresponding target values $\{t_n\}$, the goal is to predict the value of t for a new value of $\mathbf{x}$.

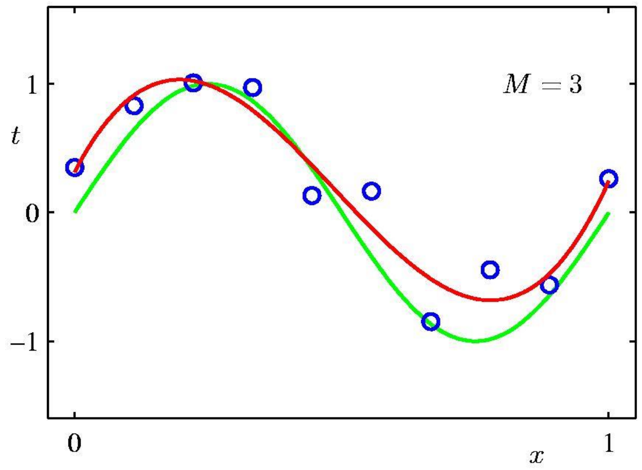
Examples:



$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \ldots + w_D x_D$$

Linear Fitting

Linear Model for Regression



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

Polynomial Fitting

Linear Model for Regression

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$$

$$\mathbf{w} = (w_0, \ldots, w_{M-1})^{\mathrm{T}}$$

$$\boldsymbol{\phi} = (\phi_0, \ldots, \phi_{M-1})^{\mathrm{T}}$$

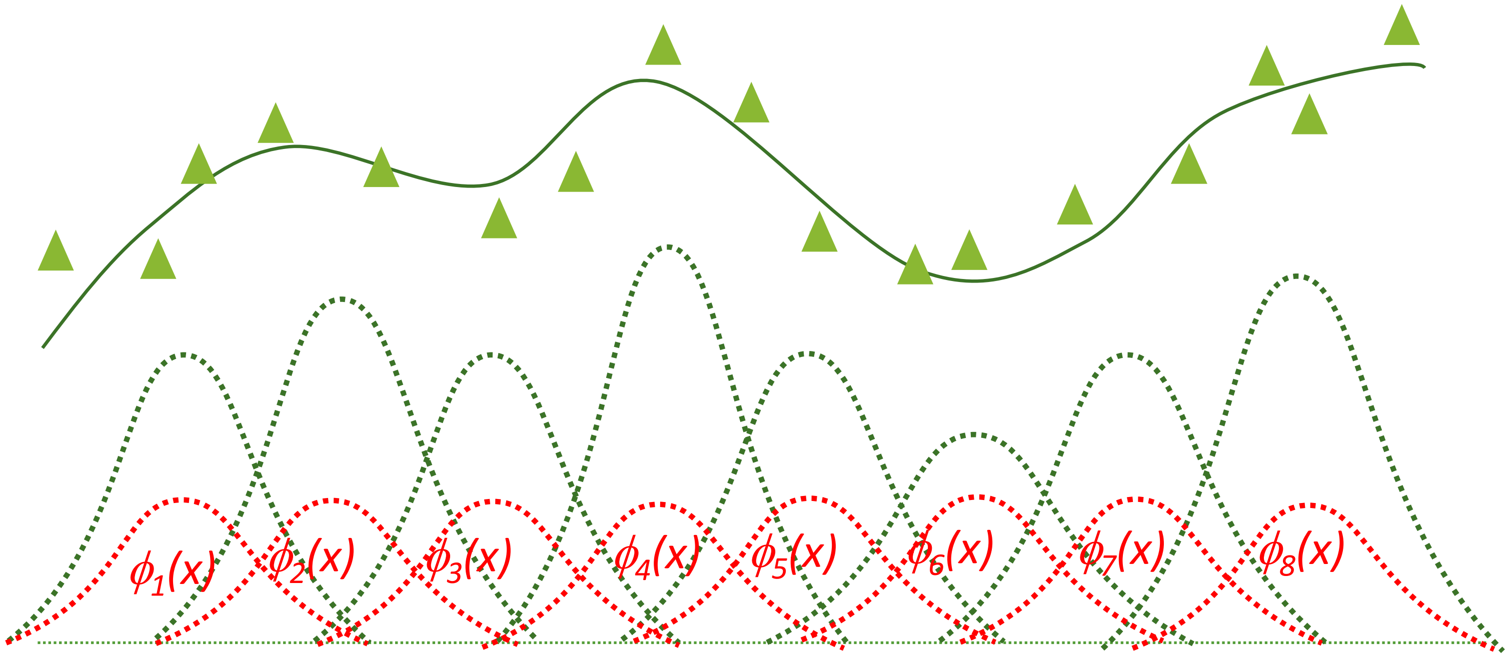
$\phi_j(\mathbf{x})$: basis functions

Remarks:

1. $\phi_j(\mathbf{x})$ are known as basis functions.
2. Typically, we define $\phi_0(\mathbf{x}) = 1$ and w_0 is the bias parameter.

Linear Model for Regression

$$y = w_0 + w_1\phi_1(\mathbf{x}) + w_2\phi_2(\mathbf{x}) + \ldots + w_{M-1}\phi_{M-1}(\mathbf{x})$$



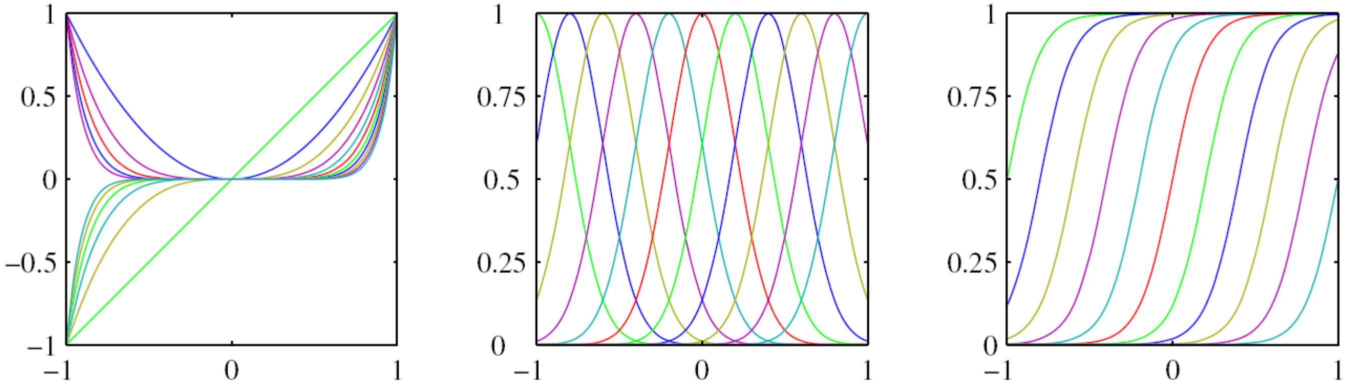
Linear Model for Regression

Examples of Basis Functions:

$$\phi_j(x) = x^j$$ Polynomial

$$\phi_j(x) = \exp\left\{-\frac{(x-\mu_j)^2}{2s^2}\right\}$$ Gaussian

$$\phi_j(x) = \sigma\left(\frac{x-\mu_j}{s}\right)$$ where $$\sigma(a) = \frac{1}{1+\exp(-a)}$$ Logistic Sigmoid

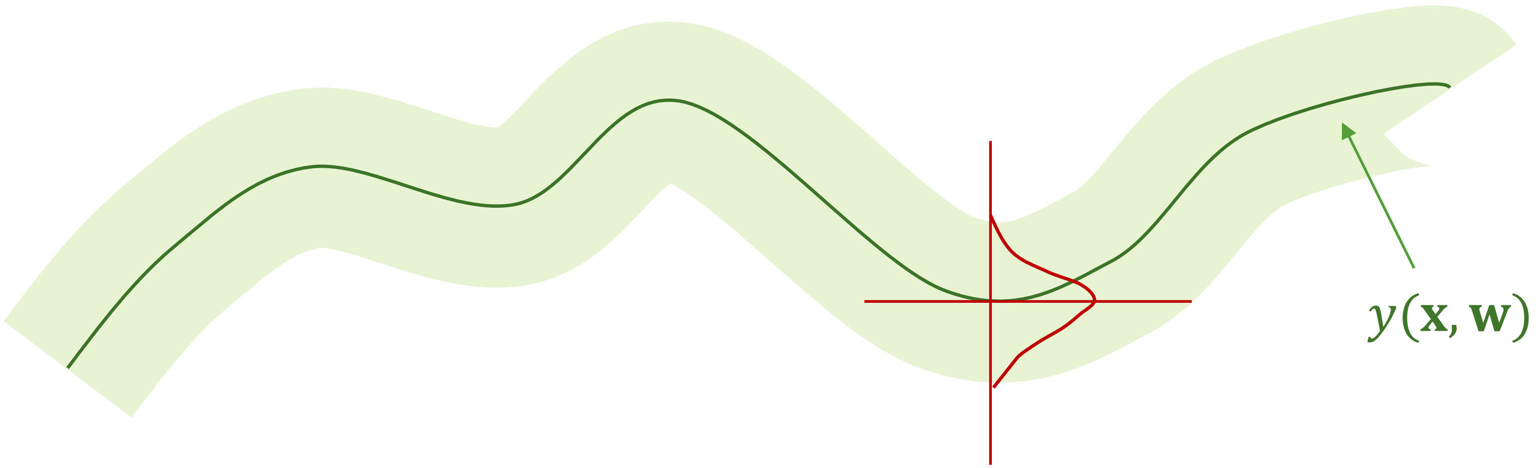


Probabilistic Perspective of Linear Regression Model (1/7)

Assume observations from a deterministic function with added Gaussian noise:

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon \qquad \text{where} \quad p(\varepsilon \mid \beta) = N(\varepsilon \mid 0, \beta^{-1})$$

$$p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1}) \qquad \text{where } y(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T \phi(\mathbf{x})$$

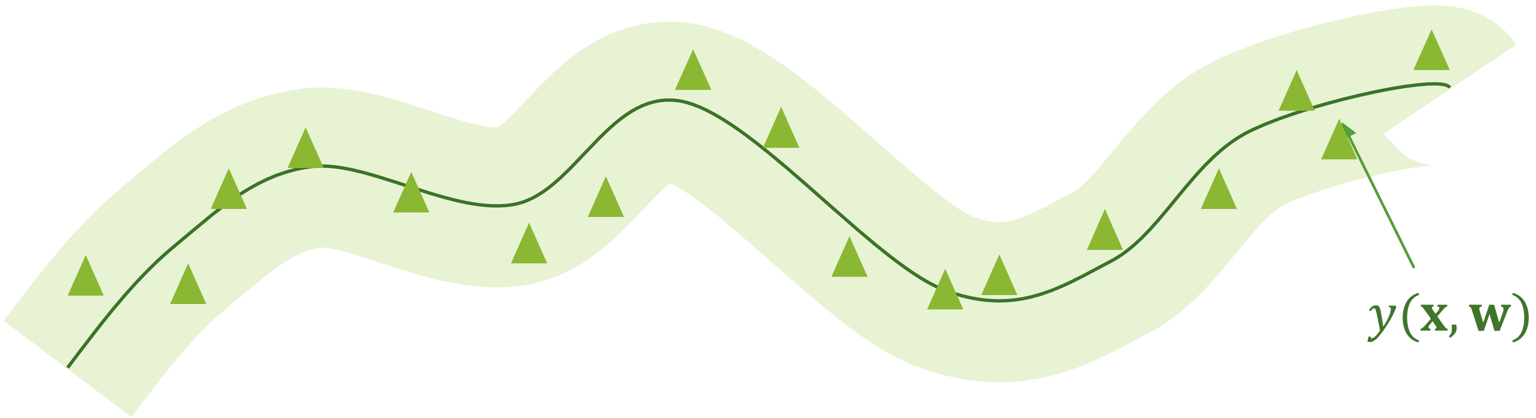


Probabilistic Perspective of Linear Regression Model (2/7)

Given a data set of inputs $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ with corresponding target values $t_1, \ldots, t_N$, we have

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1})$$

Likelihood Function



Probabilistic Perspective of Linear Regression Model (3/7)

Given a data set of inputs $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ with corresponding target values $t_1, \ldots, t_N$, we have

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1})$$

Likelihood Function

$$\begin{aligned}\ln p(\mathbf{t}|\mathbf{w}, \beta) &= \sum_{n=1}^{N} \ln \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}) \\ &= \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \beta E_D(\mathbf{w})\end{aligned}$$

Log Likelihood Function

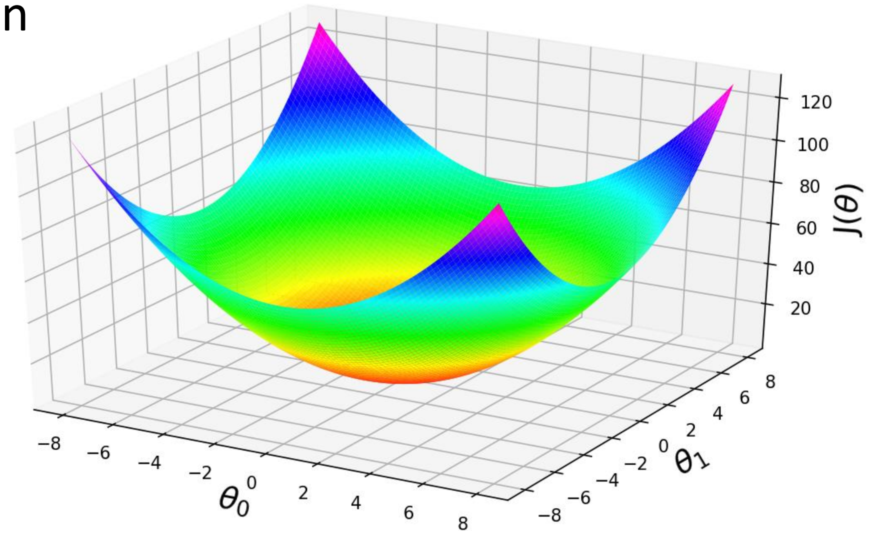
where $$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2.$$

Probabilistic Perspective of Linear Regression Model (4/7)

- To find the optimal solution of $\ln p(\mathbf{t}|\mathbf{w}, \beta)$

In general, we use **gradient descent** to find the solution

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta \nabla \underbrace{\ln p(\mathbf{t}|\mathbf{w}, \beta)}_{J(\mathbf{w})}$$



For quadratic cost functions, we may find $\mathbf{w}_{optimal}$ by directly solving the equation

$$\nabla \ln p(\mathbf{t}|\mathbf{w}, \beta) = 0$$

https://machinelearningspace.com/a-comprehensive-guide-to-gradient-descent-algorithm/

Probabilistic Perspective of Linear Regression Model (5/7)

$$\nabla \ln p(\mathbf{t}|\mathbf{w}, \beta) = \sum_{n=1}^{N} \left\{ t_n - \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n) \right\} \boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}}$$

$$\mathbf{w}_{\mathrm{ML}} = \left(\boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi} \right)^{-1} \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{t} \quad \Rightarrow \quad \hat{\mathbf{y}} = \Phi \mathbf{w}_{\mathbf{ML}} = \Phi(\Phi^T \Phi)^{-1} \Phi^T \mathbf{t}$$

Normal Equations for the least squares problem.

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}$$ **Design Matrix**

Remark: $\Phi^{+} \equiv (\Phi^{T}\Phi)^{-1}\Phi^{T}$ is known as the ***Moore-Penrose*** pseudo-inverse of the matrix Φ

Similarly,

$$\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N} \sum_{n=1}^{N} \{ t_n - \mathbf{w}_{\mathrm{ML}}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n) \}^2$$

$$\nabla \ln p(\mathbf{t}|\mathbf{w},\beta) = \beta\sum_{n=1}^{N}\left\{t_n - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)\right\}\phi(\mathbf{x}_n)^{\mathrm{T}} = 0 \equiv \sum_{n=1}^{N} t_n \phi(\vec{x}_n) = \sum_{n=1}^{N} w^T\phi(\vec{x}_n)\phi(\vec{x}_n) \equiv \Phi^T \vec{t} = \Phi^T \Phi w \equiv \vec{w} = (\Phi^T\Phi)^{-1}\Phi^T\vec{t}$$

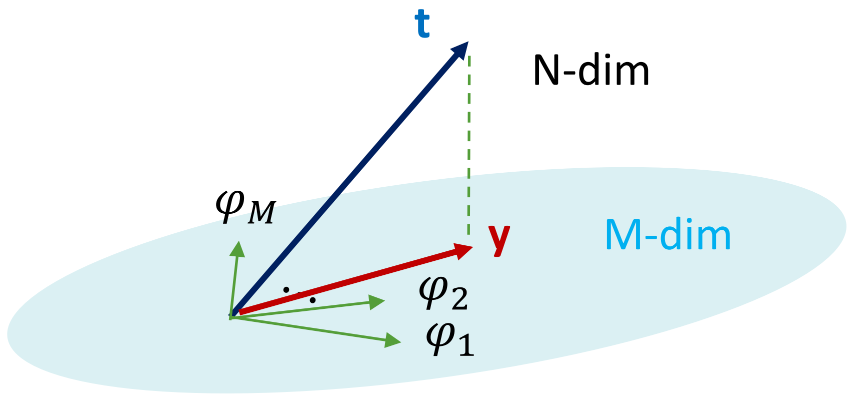
$$\begin{bmatrix}\phi(\vec{x}_1) & \phi(\vec{x}_2) & \cdots\end{bmatrix}\begin{bmatrix} t_0 \\ t_1 \\ \vdots \end{bmatrix} \qquad \begin{bmatrix}\phi(\vec{x}_1) & \phi(\vec{x}_2) & \cdots\end{bmatrix}\underline{\begin{bmatrix} w^T\phi(\vec{x}_1) \\ w^T\phi(\vec{x}_2) \\ \vdots \end{bmatrix}}$$

$$\left(\begin{bmatrix}\phi^T(\vec{x}_1) \\ \phi^T(\vec{x}_2) \\ \vdots\end{bmatrix}\right)\begin{bmatrix}\vec{w}\end{bmatrix}$$

design matrix Φ

Probabilistic Perspective of Linear Regression Model (6/7)

Geometric Interpretation



$$\boldsymbol{y} = \boldsymbol{\Phi}\mathbf{W}_{ML} = [\varphi_1, \ldots, \varphi_M]\mathbf{W}_{ML}$$

y lives in an M-dimensional subspace ***S*** of the N-dimensional space.

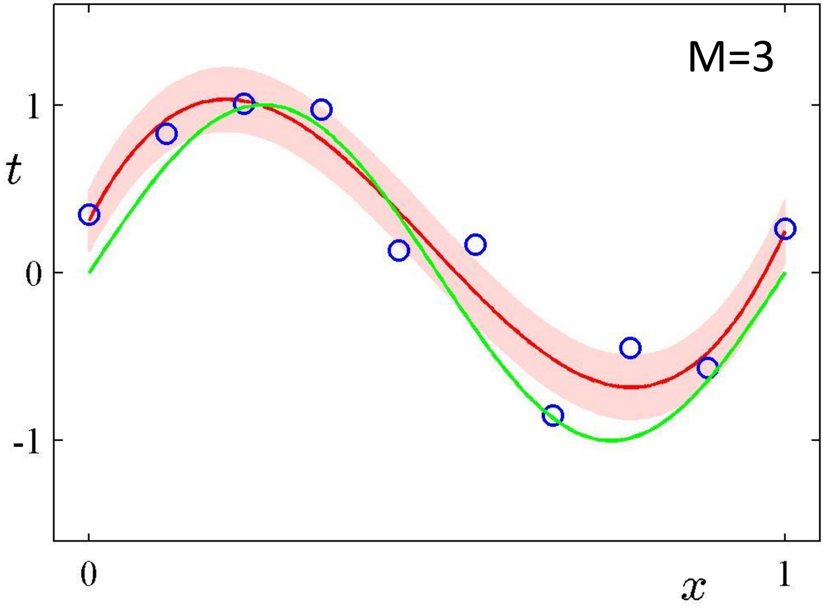
$\Rightarrow$ $\mathbf{w}_{\mathbf{ML}}$ minimizes the distance between ***t*** and its orthogonal projection on ***S***.

Probabilistic Perspective of Linear Regression Model (7/7)

Having determined **w** and β, we can make predictions for new values of *x*.

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

ML Estimation



Sequential Learning (On-line Learning)

Stochastic (sequential) gradient descent

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \nabla E_n$$

For the case of the sum-of-squares error function,

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} + \eta(t_n - \mathbf{w}^{(\tau)\mathrm{T}}\boldsymbol{\phi}_n)\boldsymbol{\phi}_n$$

Least-mean-squares (LMS) algorithm

where $\phi_n = \phi(\boldsymbol{x}_n)$.

Remark: The value of η needs to be chosen with care to ensure that the algorithm converges.

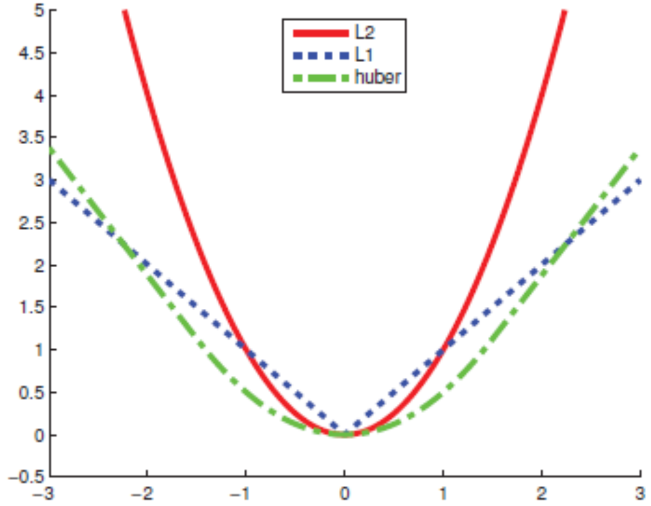
Robust Linear Regression

Replace the Gaussian distribution with a distribution that has heavy tails, like Laplace distribution or Student's t distribution.

Example: Laplace distribution

$$p(t|\mathbf{x}, \mathbf{w}, b) \propto \exp(-\frac{1}{b}|t - \mathbf{w}^T\varphi(\mathbf{x})|)$$

$$E(\mathbf{w}) \propto \sum_{n=1}^{N} |t_n - \mathbf{w}^T\varphi(\mathbf{x}_n)|$$



Ref: K.P. Murphy, "Machine Learning: A Probabilistic Perspective")

Regularized Least Squares (1/5)

Consider the error function

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w})$$

λ: regularization coefficient

data term *regularization term*

With the sum-of-squares error function and a quadratic regularizer, we have

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

Ridge Regression

$$\Rightarrow \quad \mathbf{w} = \left(\lambda\mathbf{I} + \mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}$$

Regularized Least Squares (2/5)

With a more general regularizer, we have

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$$

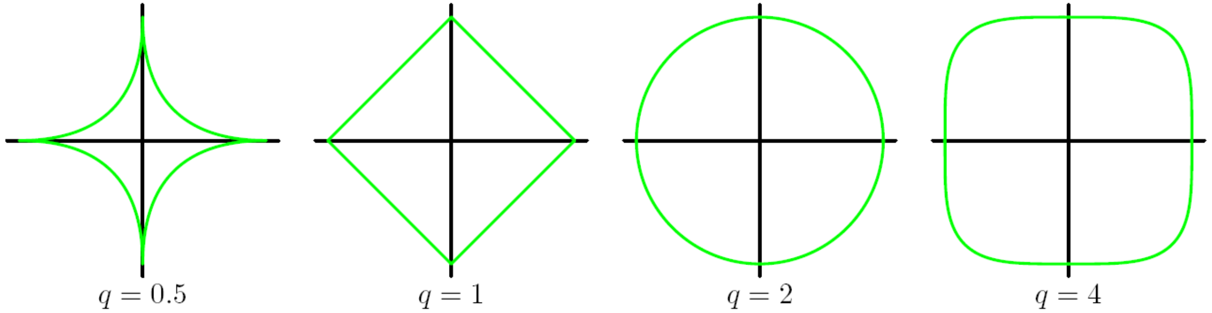


Figure 3.3 Contours of the regularization term in (3.29) for various values of the parameter q.

Remark: For the case of q = 1 (named ***Lasso*** in statistics), it tends to generate sparser solutions than a quadratic regularizer.

LASSO: Least Absolute Shrinkage and Selection Operator

Regularized Least Squares (3/5)

Probabilistic Perspective

Add a prior distribution over the polynomial coefficient **w**.

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

Posterior Probability

Likelihood Function

Prior Probability

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

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1})$$

Regularized Least Squares (4/5)

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

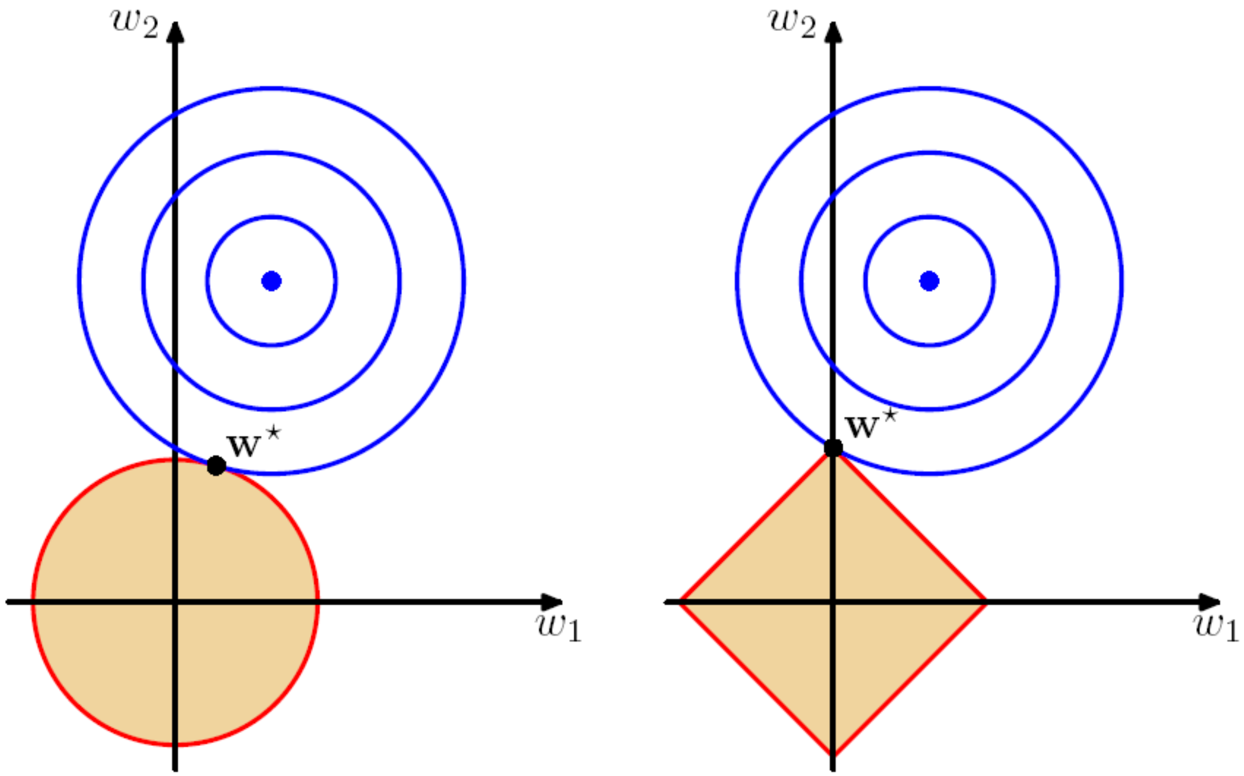
where $\ln p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta) = \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \beta E_D(\mathbf{w})$

$$\ln p(w|\alpha) = \frac{M+1}{2}\ln(\frac{\alpha}{2\pi}) - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}$$

$$\Longrightarrow \quad -\ln p(\mathbf{w}|\mathbf{x},t,\alpha,\beta) \propto \quad \frac{\beta}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^T\varphi(\mathbf{x}_n)\}^2 + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}$$

or $\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^T\varphi(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}$ where $\lambda = \frac{\alpha}{\beta}$

Regularized Least Squares (5/5)



q = 2 q = 1

Likelihood	Prior	Name
Gaussian	Uniform	Least Squares
Laplace	Uniform	Robust Regression
Student	Uniform	Robust Regression
Gaussian	Gaussian	Ridge Regression
Gaussian	Laplace	Lasso Regression

Extension to Multiple Outputs

Use the same set of basis functions to model all of the components of the target vector

$$\mathbf{y}(\mathbf{x}, \mathbf{w}) = \mathbf{W}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}) \qquad\qquad p(\mathbf{t}|\mathbf{x}, \mathbf{W}, \beta) = \mathcal{N}(\mathbf{t}|\mathbf{W}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}), \beta^{-1}\mathbf{I})$$

Given observed inputs, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, and targets, $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_N]^{\mathrm{T}}$, the log likelihood function is given by

$$\begin{aligned} \ln p(\mathbf{T}|\mathbf{X}, \mathbf{W}, \beta) &= \sum_{n=1}^{N} \ln \mathcal{N}(\mathbf{t}_n | \mathbf{W}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\mathbf{I}) \\ &= \frac{NK}{2} \ln\left(\frac{\beta}{2\pi}\right) - \frac{\beta}{2} \sum_{n=1}^{N} \left\| \mathbf{t}_n - \mathbf{W}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n) \right\|^2 \end{aligned}$$

$$\Rightarrow \quad \mathbf{W}_{\mathrm{ML}} = \left(\boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{T}$$

For each single target variable $\mathbf{t}_k$, we have

$$\mathbf{w}_k = \left(\boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{t}_k = \boldsymbol{\Phi}^{\dagger} \mathbf{t}_k$$

where $\mathbf{t}_k = [t_{1k}, t_{2k}, \ldots, t_{Nk]}^{\mathrm{T}}$

Bayesian Linear Regression (1/8)

For the likelihood function

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n), \beta^{-1}) = N(\mathbf{t} \mid \mathbf{\Phi}\mathbf{w}, \beta^{-1}\mathbf{I})$$

we define a conjugate prior over **w**

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$$

$$\Rightarrow \quad p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

Posterior Probability Function

where

$$\begin{aligned}\mathbf{m}_N &= \mathbf{S}_N\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}\right) \\ \mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}.\end{aligned}$$

$$\mathbf{w}_{\mathrm{MAP}} = \mathbf{m}_N$$

MAP (Maximum A Posteriori) Estimation

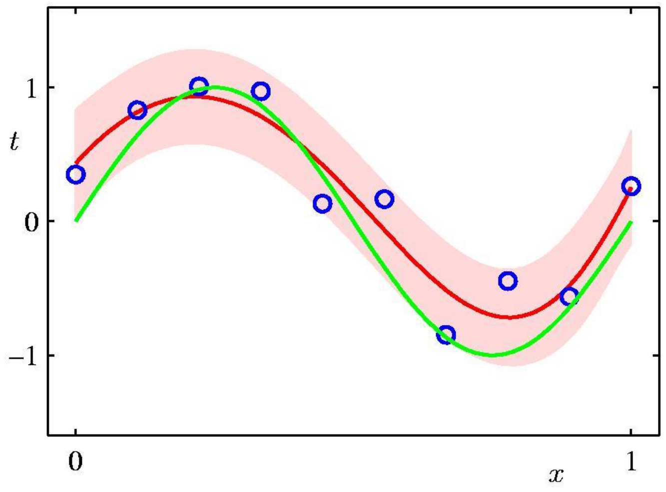
Bayesian Linear Regression (2/8)

Example:

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$$

$$\Rightarrow \quad \begin{aligned} \mathbf{m}_N &= \beta \mathbf{S}_N \mathbf{\Phi}^{\mathrm{T}} \mathbf{t} \\ \mathbf{S}_N^{-1} &= \alpha \mathbf{I} + \beta \mathbf{\Phi}^{\mathrm{T}} \mathbf{\Phi} \end{aligned}$$

$$\ln p(\mathbf{w}|\mathbf{t}) = -\frac{\beta}{2} \sum_{n=1}^{N} \{t_n - \mathbf{w}^{\mathrm{T}} \phi(\mathbf{x}_n)\}^2 - \frac{\alpha}{2} \mathbf{w}^{\mathrm{T}} \mathbf{w} + \text{const}$$



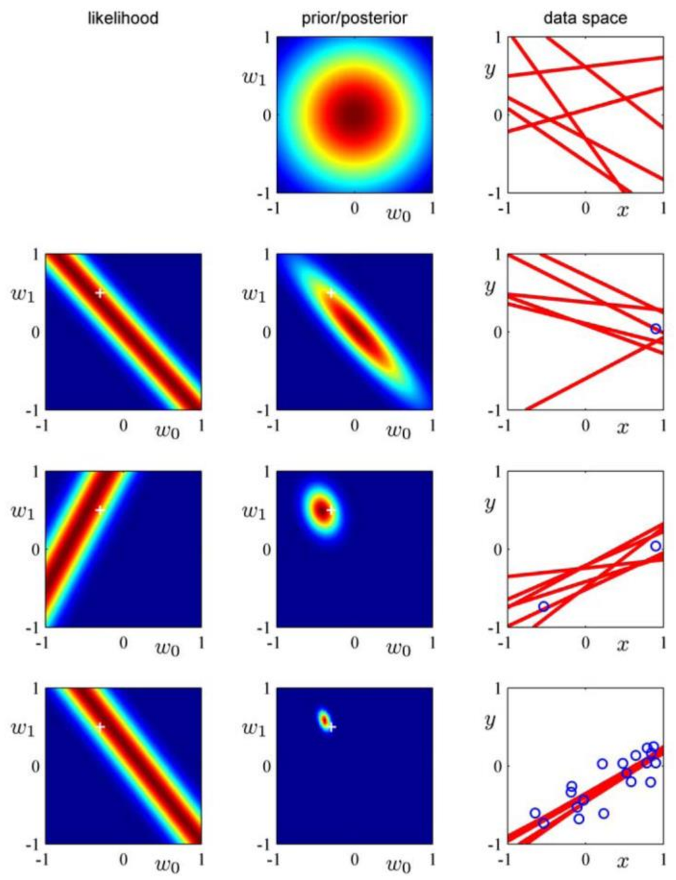
Determine $\mathbf{w}_{\text{MAP}}$ by minimizing regularized sum of squares error.

Bayesian Linear Regression (3/8)

Example: Sequential Bayesian learning for straight-line fitting

$$y(x, \mathbf{w}) = w_0 + w_1 x$$

Ground Truth: $a_0 = -0.3$, $a_1 = 0.5$



Bayesian Linear Regression (4/8)

Other forms of prior

$$p(\mathbf{w}|\alpha) = \left[\frac{q}{2}\left(\frac{\alpha}{2}\right)^{1/q}\frac{1}{\Gamma(1/q)}\right]^M \exp\left(-\frac{\alpha}{2}\sum_{j=1}^{M}|w_j|^q\right)$$

Finding the maximum of the posterior distribution over **w** corresponds to minimization of the regularized error function

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$$

Bayesian Linear Regression (5/8)

Predictive Distribution

$$p(t|\mathbf{t}, \alpha, \beta) = \int p(t|\mathbf{w}, \beta) p(\mathbf{w}|\mathbf{t}, \alpha, \beta) \, \mathrm{d}\mathbf{w}$$

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

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

$$\begin{aligned} \mathbf{m}_N &= \mathbf{S}_N \left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{t} \right) \\ \mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \beta \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi}. \end{aligned}$$

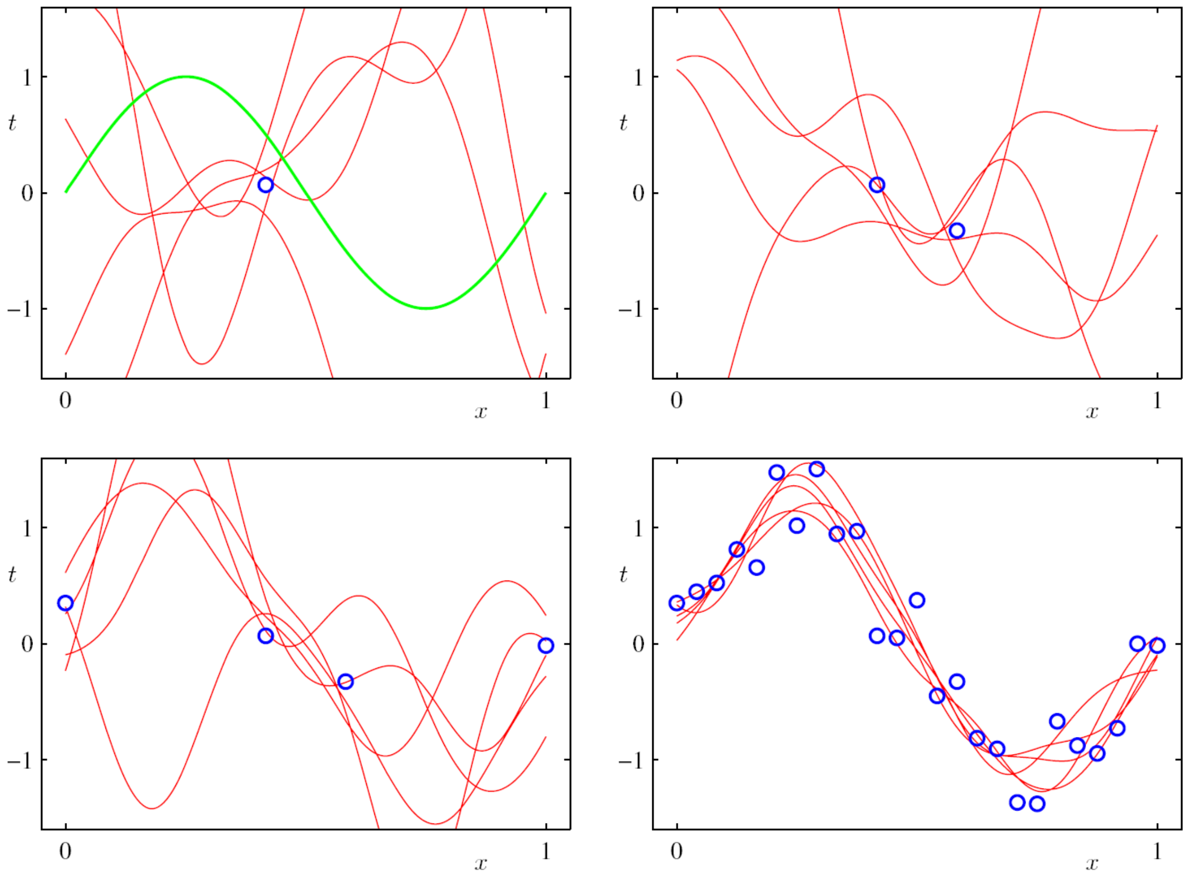
$$\Rightarrow \quad p(t|\mathbf{x}, \mathbf{t}, \alpha, \beta) = \mathcal{N}(t|\mathbf{m}_N^{\mathrm{T}} \phi(\mathbf{x}), \sigma_N^2(\mathbf{x}))$$

where $\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \phi(\mathbf{x})^{\mathrm{T}} \mathbf{S}_N \phi(\mathbf{x})$

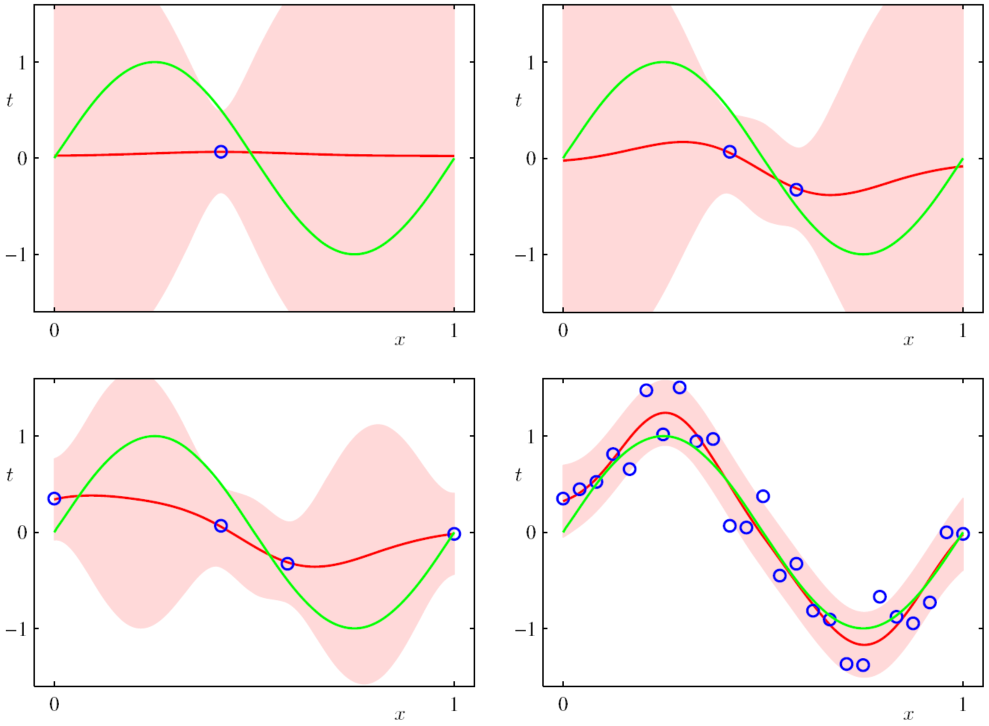
the noise on the data

the uncertainty associated with the parameters w

Bayesian Linear Regression (6/8)



Bayesian Linear Regression (7/8)



red curve: mean of the predictive distribution
red shaded region: one standard deviation span around the mean

Bayesian Linear Regression (8/8)

Given the model $p(t|x, \mathbf{w}, \beta) = \mathcal{N}\left(t|y(x, \mathbf{w}), \beta^{-1}\right)$

(1) **ML approach**: find the **w** that maximizes the likelihood function

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

$$p(t \mid x, D) = p(t \mid x, \mathbf{w}_{\mathbf{ML}}, \beta^{-1})$$

(2) **MAP approach**: find the **w** that maximizes the posterior probability

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

$$p(t \mid x, D) = p(t \mid x, \mathbf{w}_{\mathbf{MAP}}, \beta^{-1})$$

(3) **Bayesian Predictive Distribution**: consider all **w**'s

$$p(t \mid x, D) = p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|x, \mathbf{w})p(\mathbf{w}|\mathbf{x}, \mathbf{t})\, \mathrm{d}\mathbf{w}$$

Loss Function for Regression (1/2)

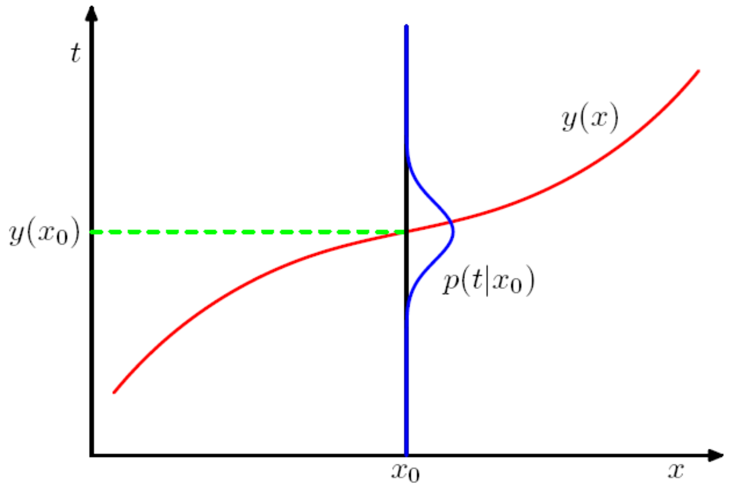
$$\mathbb{E}[L] = \iint L(t, y(\mathbf{x}))p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t.$$

Example: Squared Error

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

$$\frac{\delta \mathbb{E}[L]}{\delta y(\mathbf{x})} = 2\int \{y(\mathbf{x}) - t\}p(\mathbf{x}, t)\,\mathrm{d}t = 0$$

$$y(\mathbf{x}) = \frac{\displaystyle\int tp(\mathbf{x}, t)\,\mathrm{d}t}{p(\mathbf{x})} = \int tp(t|\mathbf{x})\,\mathrm{d}t = \mathbb{E}_t[t|\mathbf{x}]$$



Regression Function

Loss Function for Regression (2/2)

Another viewpoint

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

$$E[L] = \int [y(\boldsymbol{x}) - E[t|\boldsymbol{x}]\}^2 p(\boldsymbol{x}) d\boldsymbol{x} + \int Var[t|\boldsymbol{x}] d\boldsymbol{x}$$

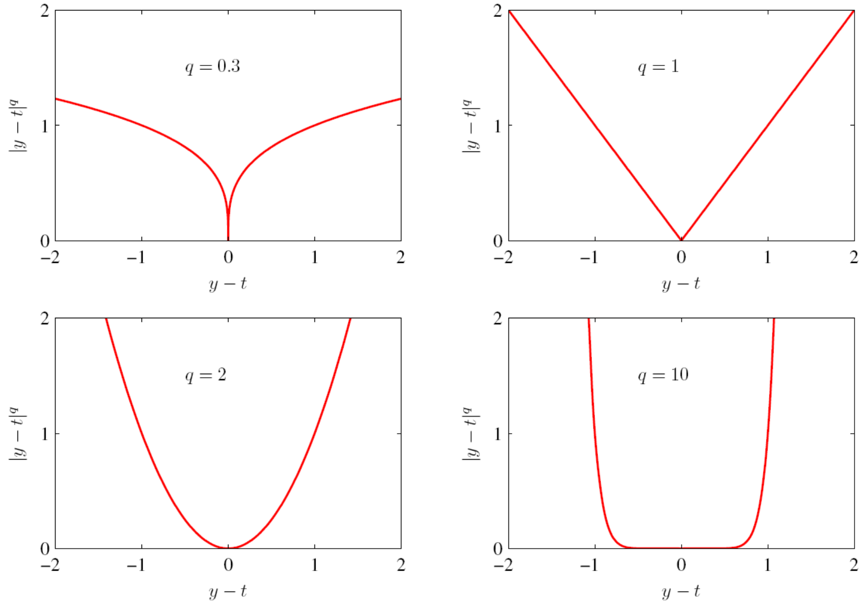
equal to zero when
y(**x**) = E[t|**x**]

intrinsic variability of
the target data

Minkowski Loss

$$\mathbb{E}[L_q] = \iint |y(\mathbf{x}) - t|^q p(\mathbf{x}, t)\, \mathrm{d}\mathbf{x}\, \mathrm{d}t$$

- The minimum of E[L_q] is given by
  the conditional mean for q = 2,
  the conditional median for q = 1, and
  the conditional mode for $q \rightarrow 0$.



The Bias-Variance Decomposition (1/3)

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

intrinsic noise on the data

where $$h(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}] = \int t p(t|\mathbf{x})\,\mathrm{d}t.$$

$$\begin{aligned}
&\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] + \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 \\
&\quad = \{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2 + \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 \\
&\qquad + 2\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}.
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - h(\mathbf{x})\}^2\right] \\
&\quad = \underbrace{\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2}_{(\text{bias})^2} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right]}_{\text{variance}}
\end{aligned}$$

The Bias-Variance Decomposition (2/3)

$$\text{expected loss} = (\text{bias})^2 + \text{variance} + \text{noise}$$

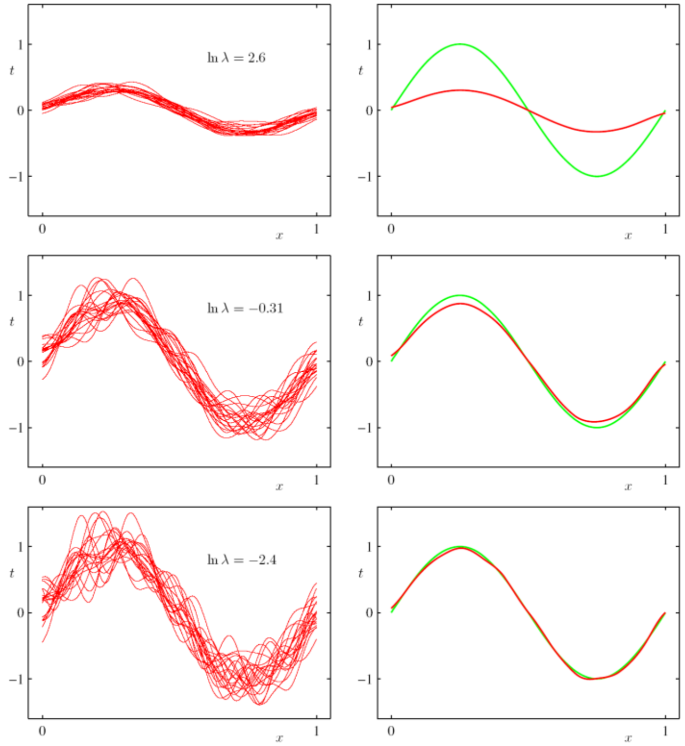
where

$$\begin{aligned}
(\text{bias})^2 &= \int \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
\text{variance} &= \int \mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right] p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
\text{noise} &= \int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x},t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t
\end{aligned}$$

Remarks:

1. There is a trade-off between bias and variance.
   - ✓ flexible models: low bias and high variance
   - ✓ rigid models: high bias and low variance
2. The model with the optimal predictive capability leads to the best balance between bias and variance.

The Bias-Variance Decomposition (3/3)



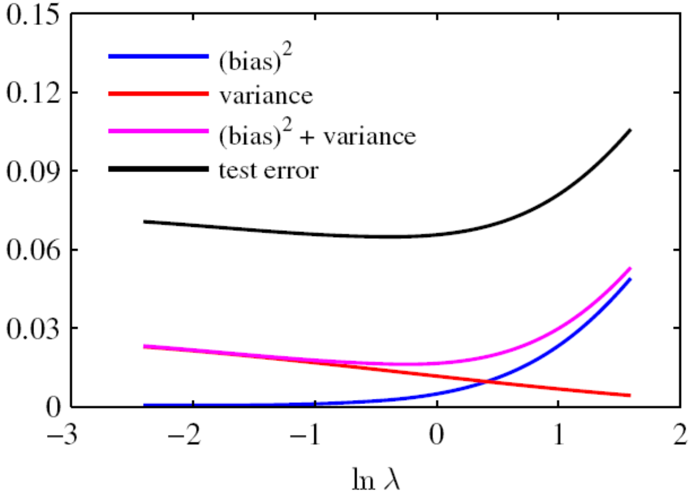


Figure 3.5 Illustration of the dependence of bias and variance on model complexity, governed by a regularization parameter λ, using the sinusoidal data set from Chapter 1. There are $L = 100$ data sets, each having $N = 25$ data points, and there are 24 Gaussian basis functions in the model so that the total number of parameters is $M = 25$ including the bias parameter. The left column shows the result of fitting the model to the data sets for various values of $\ln \lambda$ (for clarity, only 20 of the 100 fits are shown). The right column shows the corresponding average of the 100 fits (red) along with the sinusoidal function from which the data sets were generated (green).

Appendix

Conjugate Prior (1/2)

Sequential view of the inference problem.

$$p(\boldsymbol{\theta}|\mathbf{x}_1) \propto p(\mathbf{x}_1|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

$$p(\boldsymbol{\theta}|\mathbf{x}_1, \mathbf{x}_2) \propto p(\mathbf{x}_1, \mathbf{x}_2|\boldsymbol{\theta})p(\boldsymbol{\theta}) = p(\mathbf{x}_2|\boldsymbol{\theta})p(\mathbf{x}_1|\boldsymbol{\theta})p(\boldsymbol{\theta}) \propto p(\mathbf{x}_2|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{x}_1)$$

$$p(\boldsymbol{\theta}|\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) \propto p(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3|\boldsymbol{\theta})p(\boldsymbol{\theta}) = p(\mathbf{x}_3|\boldsymbol{\theta})p(\mathbf{x}_2|\boldsymbol{\theta})p(\mathbf{x}_1|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

$$\propto p(\mathbf{x}_3|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{x}_1, \mathbf{x}_2)$$

$$\vdots$$

$$p(\boldsymbol{\theta}|\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N) \propto p(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N|\boldsymbol{\theta})p(\boldsymbol{\theta}) \propto p(\mathbf{x}_N|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{N-1})$$

- The posterior obtained after observing N-1 data points becomes the prior when we observe the Nth data point.

Conjugate Prior (2/2)

If the posterior distributions p(θ|x) and the prior probability distribution p(θ) are in the same family, the prior and posterior are called **conjugate distributions**. In this case, the prior is called a **conjugate prior** for the likelihood function.

$$Posterior \propto Likelihood \times Prior$$

Same Distribution

Conjugate Priors for the Gaussian (1/6)

Case 1: σ^2 is known, but μ is unknown.

$$p(\mathbf{X}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}$$

Likelihood function (a function of μ.)

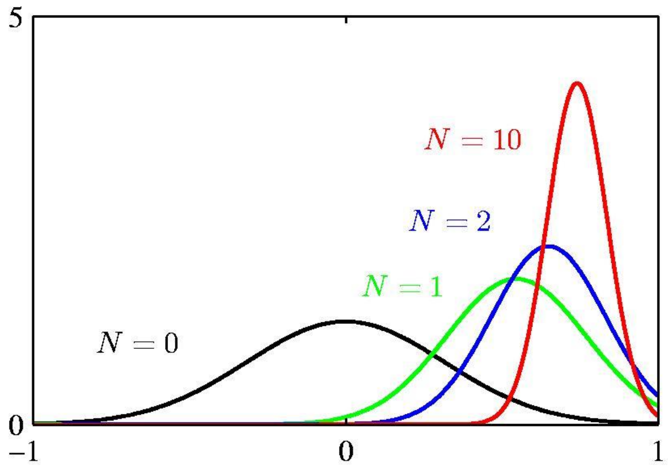
$$p(\mu) = \mathcal{N}\left(\mu|\mu_0, \sigma_0^2\right)$$ Conjugate Prior

$$\Rightarrow \quad p(\mu|\mathbf{X}) = \mathcal{N}\left(\mu|\mu_N, \sigma_N^2\right)$$ Posterior distribution

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{\mathrm{ML}}$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

where $\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$



Conjugate Priors for the Gaussian (2/6)

Case 2: μ is known, but σ^2 is unknown.

Suppose that the mean is known and we wish to infer the variance.

Let $\lambda \equiv 1/\sigma^2$

$$p(\mathbf{X}|\lambda) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \lambda^{-1}) \propto \lambda^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n - \mu)^2\right\}$$

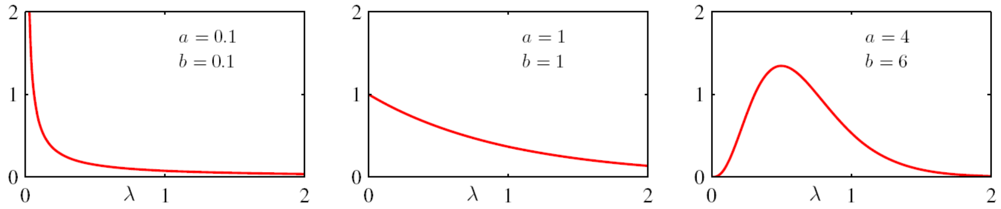
This has a Gamma shape as a function of λ.

Remark: **Gamma distribution**

$$\mathrm{Gam}(\lambda|a,b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)$$

$$\mathbb{E}[\lambda] = \frac{a}{b}$$

$$\mathrm{var}[\lambda] = \frac{a}{b^2}$$



Conjugate Priors for the Gaussian (3/6)

$\Rightarrow$ Conjugate prior: Gamma distribution.
Consider a prior distribution Gam(λ/a_0,b_0).

$$p(\lambda|\mathbf{X}) \propto \lambda^{a_0-1}\lambda^{N/2}\exp\left\{-b_0\lambda - \frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}$$

$p(\lambda|\mathbf{X})$ = Gam($\lambda|a_N,b_N$)

$$\begin{aligned} a_N &= a_0 + \frac{N}{2} \\ b_N &= b_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2 = b_0 + \frac{N}{2}\sigma^2_{\mathrm{ML}} \end{aligned}$$

Conjugate Priors for the Gaussian (4/6)

Case 3: both μ and σ² are unknown.

$$p(\mathbf{X}|\mu,\lambda) = \prod_{n=1}^{N}\left(\frac{\lambda}{2\pi}\right)^{1/2}\exp\left\{-\frac{\lambda}{2}(x_n-\mu)^2\right\}$$

$$\propto \left[\lambda^{1/2}\exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^N \exp\left\{\lambda\mu\sum_{n=1}^{N}x_n - \frac{\lambda}{2}\sum_{n=1}^{N}x_n^2\right\}$$

Conjugate prior

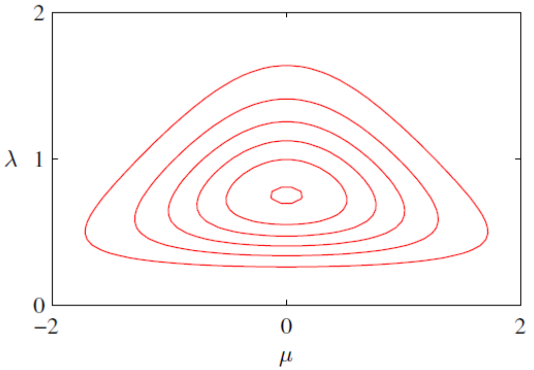
$$p(\mu,\lambda) \propto \left[\lambda^{1/2}\exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^\beta \exp\left\{c\lambda\mu - d\lambda\right\}$$

$$= \exp\left\{-\frac{\beta\lambda}{2}(\mu - c/\beta)^2\right\}\lambda^{\beta/2}\exp\left\{-\left(d-\frac{c^2}{2\beta}\right)\lambda\right\}$$

$$p(\mu,\lambda) = \mathcal{N}(\mu|\mu_0,(\beta\lambda)^{-1})\mathrm{Gam}(\lambda|a,b)$$

normal-gamma or *Gaussian-gamma* distribution

Figure 2.14 Contour plot of the normal-gamma distribution (2.154) for parameter ; $\mu_0 = 0$, $\beta = 2$, $a = 5$ and



Conjugate Priors for the Gaussian (5/6)

Remarks:

For multivariate Gaussian distribution $N(\mathbf{x}|\boldsymbol{\mu},\Lambda^{-1})$

Case 1: unknown mean, known precision matrix

Conjugate prior distribution – Gaussian distribution

Case 2: known mean, unknown precision matrix

Conjugate prior distribution – Wishart distribution

$$\mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W},\nu) = B|\boldsymbol{\Lambda}|^{(\nu-D-1)/2}\exp\left(-\frac{1}{2}\mathrm{Tr}(\mathbf{W}^{-1}\boldsymbol{\Lambda})\right)$$

where $$B(\mathbf{W},\nu) = |\mathbf{W}|^{-\nu/2}\left(2^{\nu D/2}\,\pi^{D(D-1)/4}\prod_{i=1}^{D}\Gamma\left(\frac{\nu+1-i}{2}\right)\right)^{-1}$$

ν : *degrees of freedom* of the distribution,

W: D×D scale matrix

Conjugate Priors for the Gaussian (6/6)

Case 3: unknown mean and precision matrix

Conjugate prior distribution – Gaussian-Wishart distribution

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{\mu}_0, \beta, \mathbf{W}, \nu) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0, (\beta\boldsymbol{\Lambda})^{-1})\, \mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W}, \nu)$$