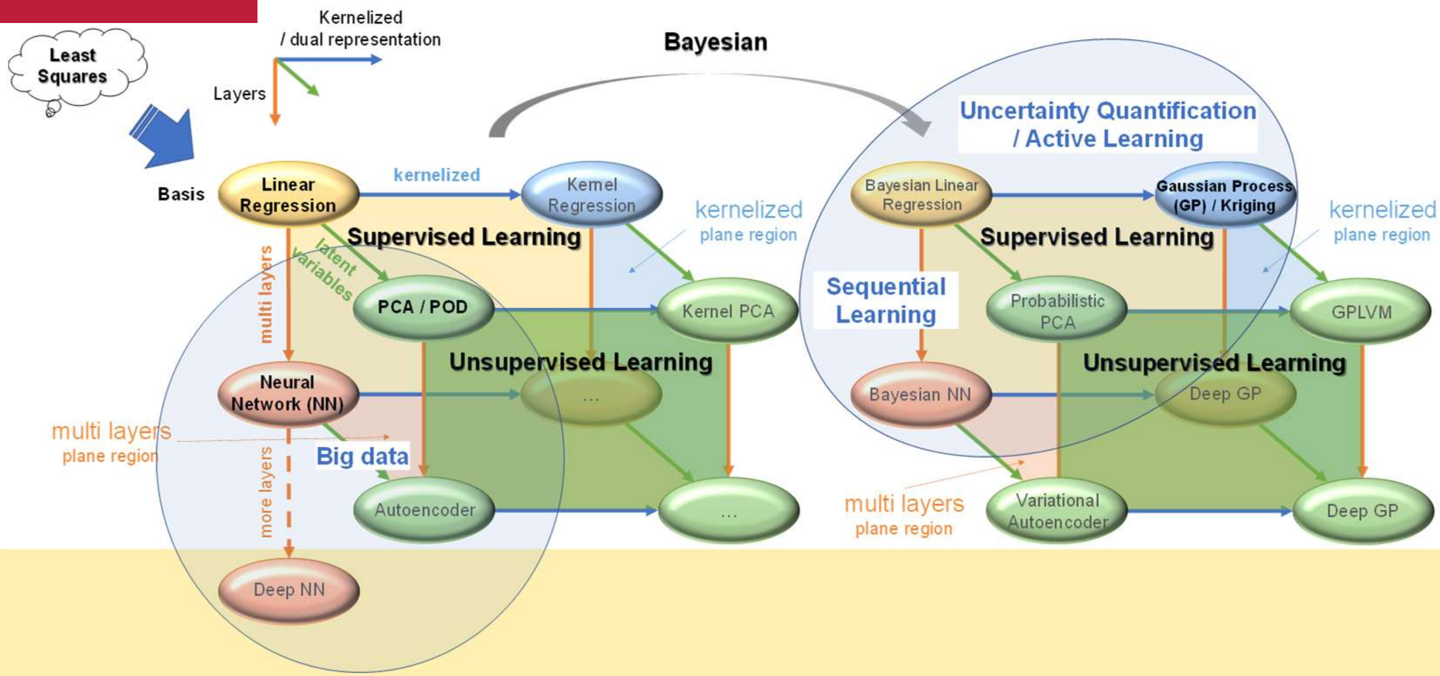
Technische Universität Braunschweig

IFL



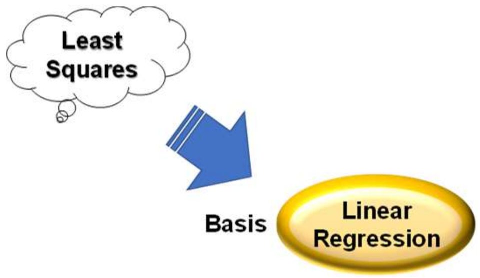
Scientific Machine Learning

Lecture 4: Linear Regression

Dr. Daigo Maruyama

Prof. Dr. Ali Elham

Current Position



Today at Lecture 4

Lecture content

- Gaussian Distribution
- Sampling Methods (Design of Experiments)
- Linear Regression

continued from Lecture 3

The lecture of this time basically follows the 2nd and 3rd chapters of the book:
Christopher M. Bishop "Pattern Recognition And Machine Learning" Springer-Verlag (2006)
The name of this book is shown as “PRML” when it is referred in the slides.

Gaussian Distribution (Normal Distribution)

Carl Friedrich Gauss
(1777-1855)

Born in Braunschweig

Collegium Carolinum at TUBS

Some important topics
related to Gaussian distributions

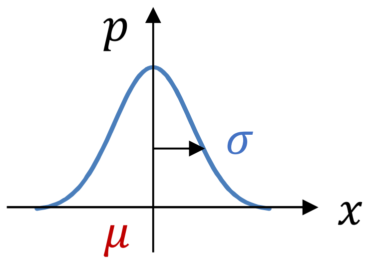
- Least square method
- Central limit theorem
- Gaussian Process

Gaussian Distribution (Normal Distribution)

Used in Regression

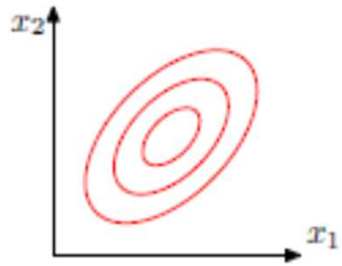
Gaussian distribution

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



Multivariate Gaussian distribution

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



Some useful properties
- conditional distribution
- marginal distribution

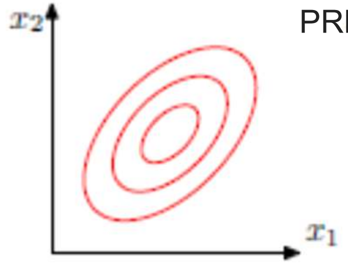
→ also Gaussian distributions

Gaussian Distribution (Normal Distribution)

Multivariate Gaussian distribution

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

PRML, p. 84



Contour map of a multivariate Gaussian distribution when $D = 2$

$\boldsymbol{x} = (x_1, \cdots, x_D)^{\mathrm{T}}$

$\boldsymbol{\mu} = (\mu_1, \cdots, \mu_D)^{\mathrm{T}}$

D parameters

$\boldsymbol{x}$: random variable (vector)

$\boldsymbol{\mu}$: mean (vector)

$\boldsymbol{\Sigma}$: covariance (matrix)

$$\boldsymbol{\Sigma} = \begin{pmatrix} var[1] & \cdots & cov[1,D] \\ \vdots & \ddots & \vdots \\ cov[D,1] & \cdots & var[D] \end{pmatrix} = \begin{pmatrix} {\sigma_1}^2 & \cdots & \sigma_{1,D} \\ \vdots & \ddots & \vdots \\ \sigma_{D,1} & \cdots & {\sigma_D}^2 \end{pmatrix}$$

$\frac{D\ (D\ +1)}{2}$ parameters

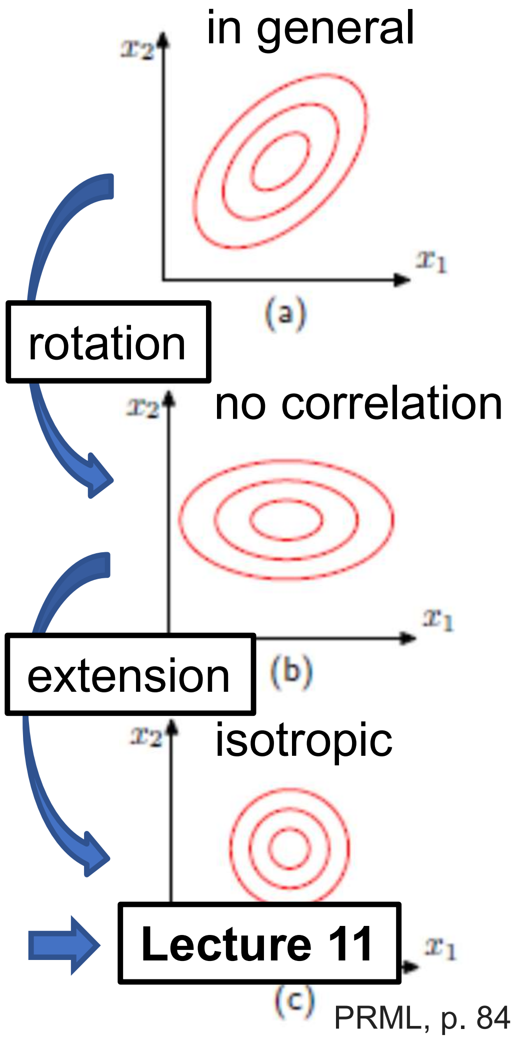
$\Rightarrow$ MLE?

$$cov[i,j] = cov[j,i]$$

Important properties of the covariance matrix $\boldsymbol{\Sigma}$:

- Symmetric
- Positive (semi)definite

Gaussian Distribution (Normal Distribution)



PRML, p. 84

$$\boldsymbol{\Sigma}_{(a)} = \begin{pmatrix} {\sigma_1}^2 & \cdots & \sigma_{1,D} \\ \vdots & \ddots & \vdots \\ \sigma_{D,1} & \cdots & {\sigma_D}^2 \end{pmatrix}$$

$\mathcal{N}(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$\frac{D\ (D\ +1)}{2}$ parameters

$$\boldsymbol{\Sigma}_{(b)} = \begin{pmatrix} {\sigma_1}^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & {\sigma_D}^2 \end{pmatrix}$$

$\mathcal{N}(y|\mu, {\sigma_i}^2) \quad i = 1, \cdots, M$

for each axis

D parameters

$$\boldsymbol{\Sigma}_{(c)} = \sigma^2 \mathbf{I} = \begin{pmatrix} \sigma^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2 \end{pmatrix}$$

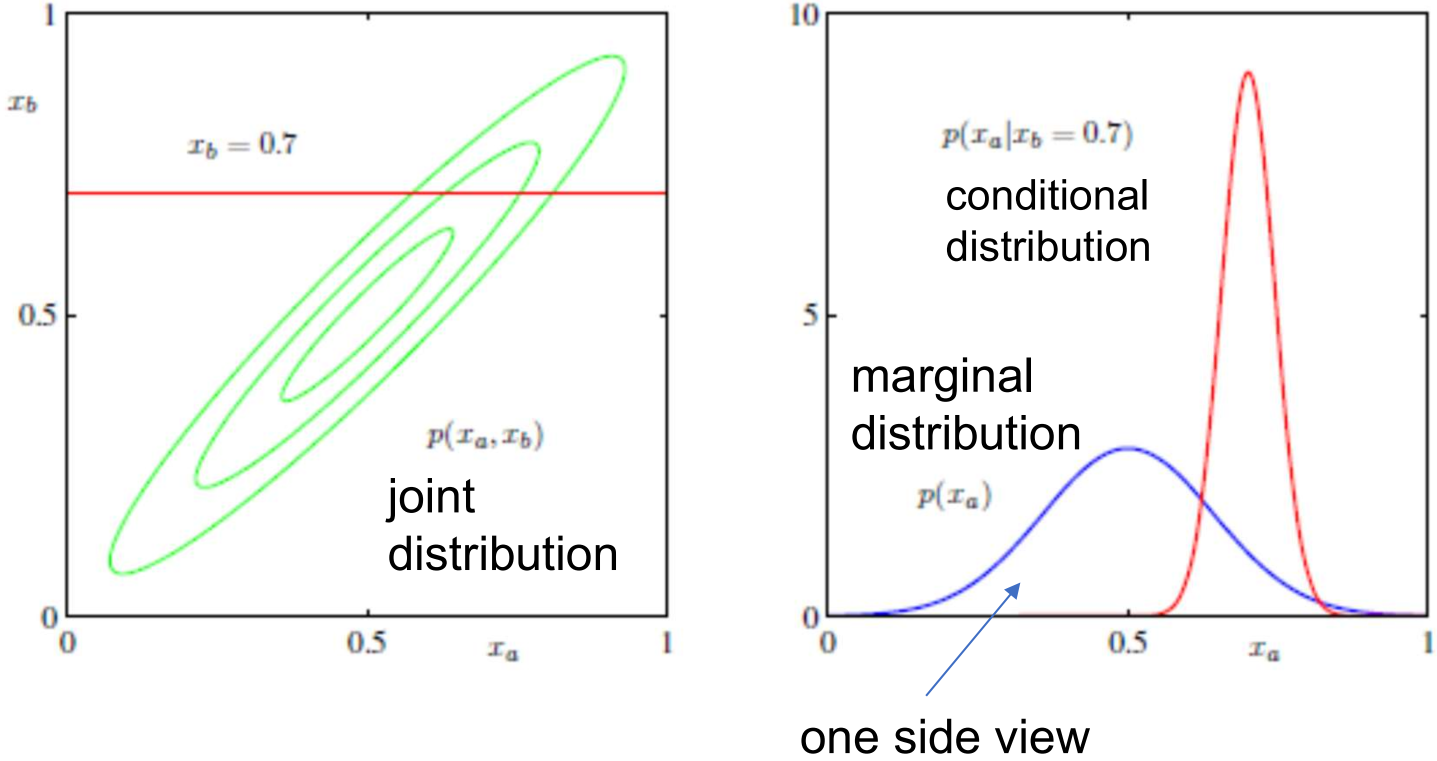
$\mathcal{N}(y|\mu, \sigma^2)$

for each axis

1 parameter

Gaussian Distribution (Normal Distribution)

Please imagine a mountain from the top view and a side view
(but the view has to be scaled to satisfy the area=1)



PRML, p. 90

Important property: Both of them become Gaussian distributions.

Gaussian Distribution (Normal Distribution)

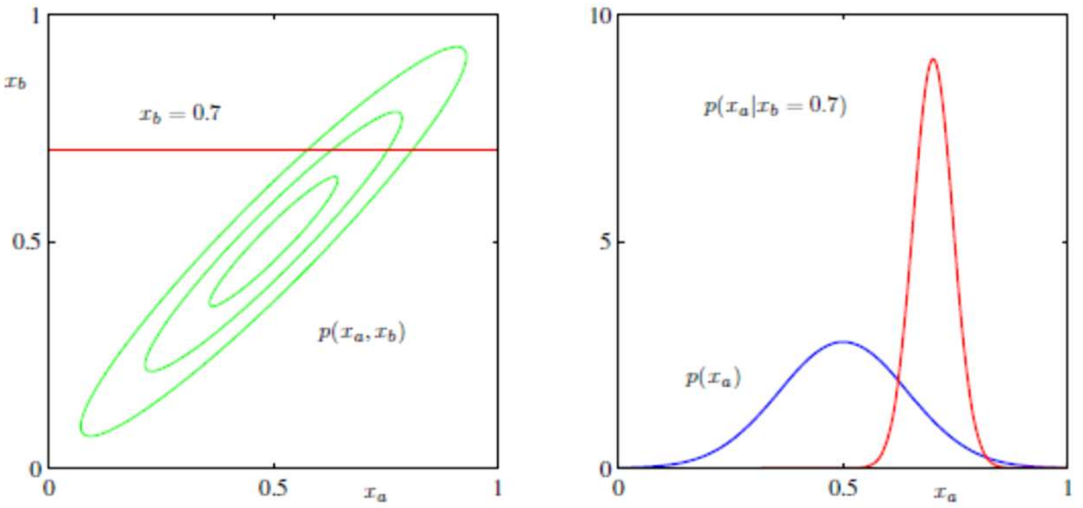
We have now a multivariate Gaussian distribution:

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

When we partition $\boldsymbol{x}$ into two disjoint subsets $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$:

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

b



a

Gaussian Distribution (Normal Distribution)

We have now a multivariate Gaussian distribution:

$$p(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$$

When we partition $\boldsymbol{x}$ into two disjoint subsets $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$:

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

What if :

- **Marginal distribution** $p(\boldsymbol{x}_a)$
- **Conditional distribution** $p(\boldsymbol{x}_a|\boldsymbol{x}_b)$
- (**Joint distribution** $p(\boldsymbol{x}) = p(\boldsymbol{x}_a,\boldsymbol{x}_b) = p(\boldsymbol{x}_a|\boldsymbol{x}_b)p(\boldsymbol{x}_b)$)

In general, there is no guarantee that $p(\boldsymbol{x}_a)$ and $p(\boldsymbol{x}_a|\boldsymbol{x}_b)$ can be represented by the same distributions $p(\boldsymbol{x})$.

Gaussian Distribution (Normal Distribution)

We have now a multivariate Gaussian distribution:

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

When we partition $\boldsymbol{x}$ into two disjoint subsets $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$:

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

Marginal distribution $p(\boldsymbol{x}_a)$

$$p(\boldsymbol{x}_a) = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa})$$

simple and intuitive

pretending that we did not see any information from the subset $\boldsymbol{x}_b$.

The basic concept of marginal distributions

Gaussian Distribution (Normal Distribution)

We have now a multivariate Gaussian distribution:

$$p(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$$

When we partition $\boldsymbol{x}$ into two disjoint subsets $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$:

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

Conditional distribution $p(\boldsymbol{x}_a|\boldsymbol{x}_b)$

$$p(\boldsymbol{x}_a|\boldsymbol{x}_b) = \mathcal{N}\left(\boldsymbol{x}_a\middle|\boldsymbol{x}_b, \boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b}\right)$$

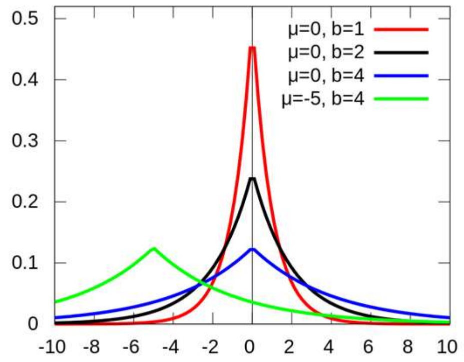
where,

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}{}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b)$$

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}{}^{-1}\boldsymbol{\Sigma}_{ba}$$

No need to remember but the concept is important in Gaussian Processes (Lectures 6-8)

Other Probability Distributions



wikipedia

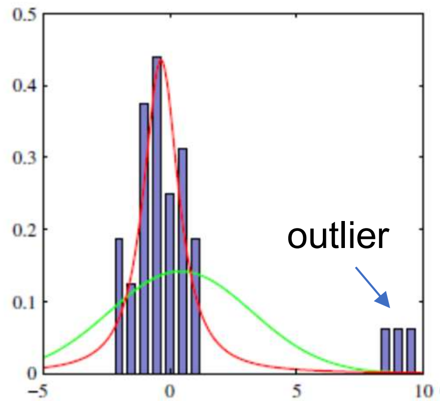
- **Laplace distribution**

$$p(x|\mu, b) = \frac{1}{2b}\exp\left\{-\frac{|x-\mu|}{b}\right\}$$

- **Cauchy distribution** (a special case of **Student's t-distribution**)

$$p(x|x_0, \gamma) = \frac{1}{\pi}\frac{\gamma}{(x-x_0)^2+\gamma^2}$$

used to treat outlier



PRML, p. 104

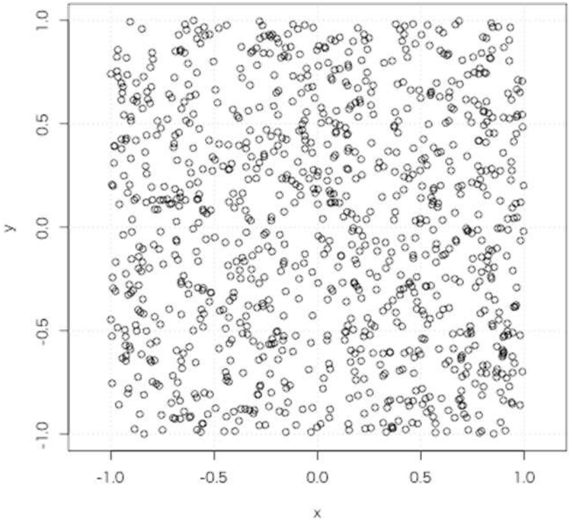
red: Gaussian
Green: student's-t

Most of the introduced probability distributions are categorized in Exponential Family.

$$p(\boldsymbol{x}|\boldsymbol{\eta}) = h(\boldsymbol{x})g(\boldsymbol{\eta})\exp\{\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{u}(\boldsymbol{x})\}$$

Lecture content

3. Sampling Methods (Design of Experiments)



Curse of Dimensionality

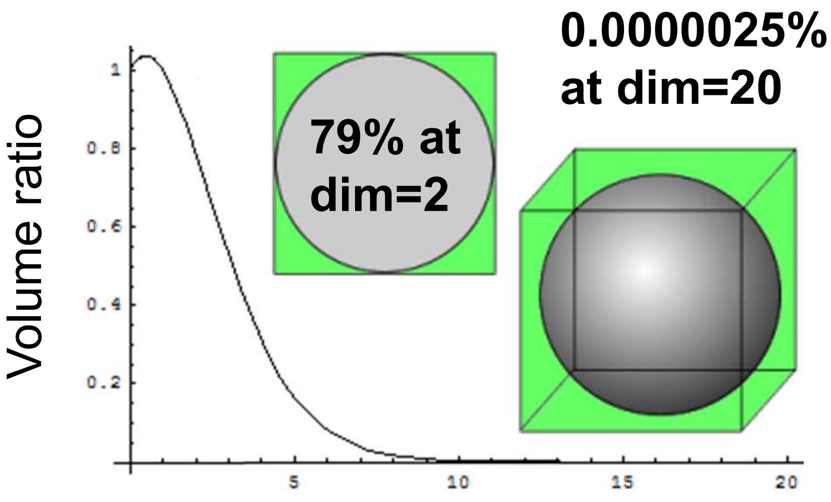
In high-dimensional space

The volume ratio between the cube and the sphere is counterintuitive.

almost skin (tiny volume)

$$V_D = \frac{\pi^{D/2}}{(D/2)!} = \frac{\pi^{D/2}}{\Gamma(D/2+1)}$$

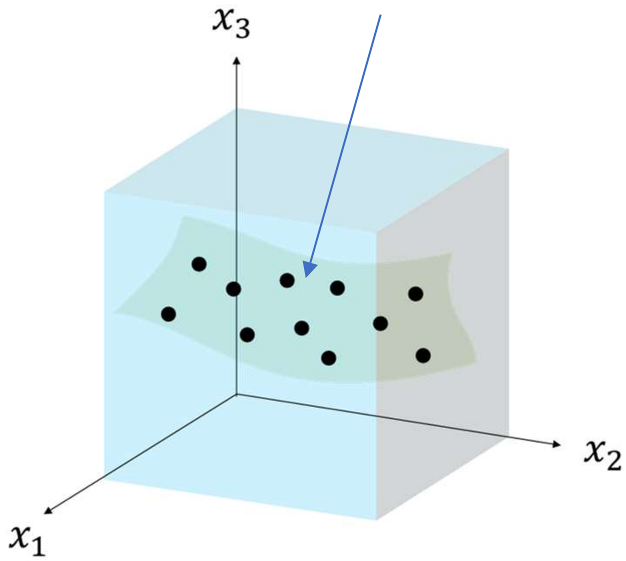
$$\frac{volume\ of\ hypersphere}{volume\ of\ hypercube} = \frac{V_D}{2^D} = \frac{\pi^{D/2}}{2^D (D/2)!} \longrightarrow 0$$



D: dimensionality

Curse of Dimensionality

a low-dimensional space - manifold



only two input actually:
- Rotation
- Translation

10,000 dims to 2 dims

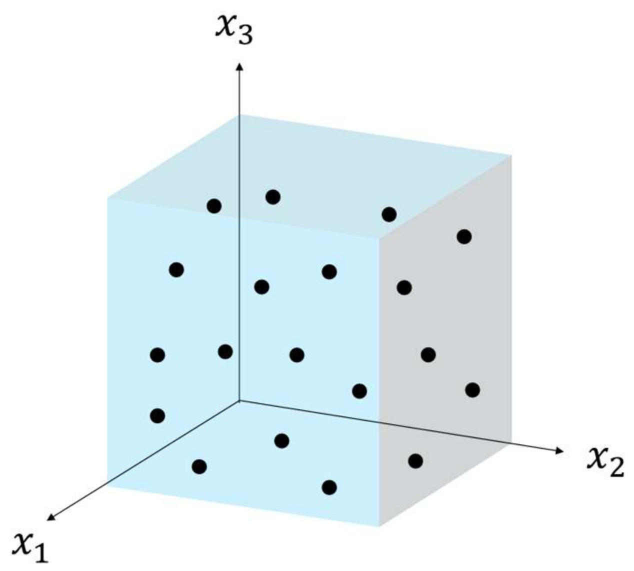
➡ **Lecture 11**

<u>Dimensionality Reduction</u> (data is lying on a low-dimensional space - manifold)
- Big data: plentiful data in hand
- **Traditional engineering design: data is produced ➡ Design or Experiments**

The parameters (design variables) are selected from engineering viewpoints.

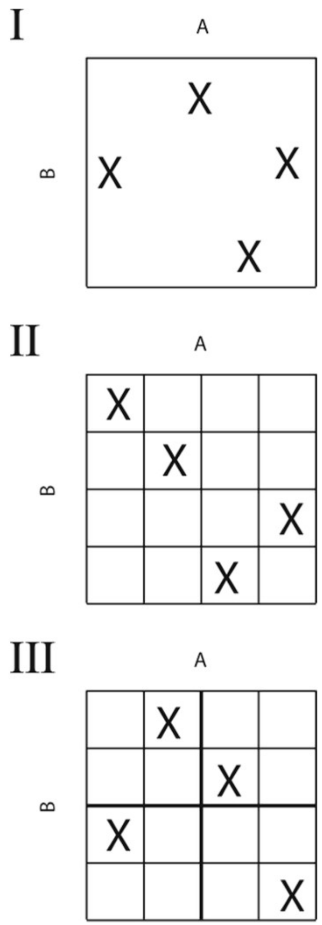
Design of Experiments (DoE)

Uniform distributions (**Design of Experiments**)
- Monte Carlo (MC) sampling method
- Latin Hypercube sampling (LHS) method
- Quasi Monte Carlo (QMC) sampling method
    - Halton sequence
    - Sobol sequence



Arbitrary distributions (**Lecture 12**)
- Markov-Chain Monte Carlo (MCMC)

Design of Experiments (DoE)



wikipedia:
Latin hypercube sampling

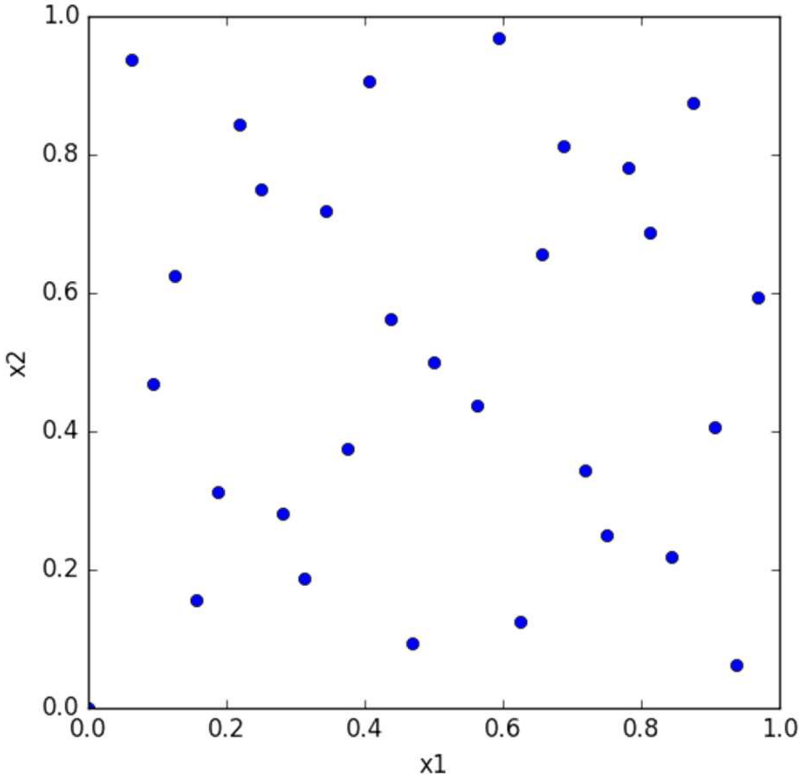
Latin hypercube sampling

Properties in practical use:

- The partition has to be defined first (the sample size defined).
- New sample points cannot be added.

Design of Experiments (DoE)

```
[[ 0.          0.        ]
 [ 0.5         0.5       ]
 [ 0.75        0.25      ]
 [ 0.25        0.75      ]
 [ 0.375       0.375     ]
 [ 0.875       0.875     ]
 [ 0.625       0.125     ]
 [ 0.125       0.625     ]
 [ 0.1875      0.3125    ]
 [ 0.6875      0.8125    ]
 [ 0.9375      0.0625    ]
 [ 0.4375      0.5625    ]
 [ 0.3125      0.1875    ]
 [ 0.8125      0.6875    ]
 [ 0.5625      0.4375    ]
 [ 0.0625      0.9375    ]
 [ 0.09375     0.46875   ]
 [ 0.59375     0.96875   ]
 [ 0.84375     0.21875   ]
 [ 0.34375     0.71875   ]
 [ 0.46875     0.09375   ]
 [ 0.96875     0.59375   ]
 [ 0.71875     0.34375   ]
 [ 0.21875     0.84375   ]
 [ 0.15625     0.15625   ]
 [ 0.65625     0.65625   ]
 [ 0.90625     0.40625   ]
 [ 0.40625     0.90625   ]
 [ 0.28125     0.28125   ]
 [ 0.78125     0.78125   ]
 [ 0.53125     0.03125   ]
 [ 0.03125     0.53125   ]
 ...,
 [ 0.05273438  0.23242188]
 [ 0.55273438  0.73242188]
 [ 0.80273438  0.48242188]
 [ 0.30273438  0.98242188]
 [ 0.42773438  0.35742188]
 [ 0.92773438  0.85742188]
 [ 0.67773438  0.10742188]
 [ 0.17773438  0.60742188]
 [ 0.24023438  0.41992188]
 [ 0.74023438  0.91992188]
 [ 0.99023438  0.16992188]
 [ 0.49023438  0.66992188]]
```

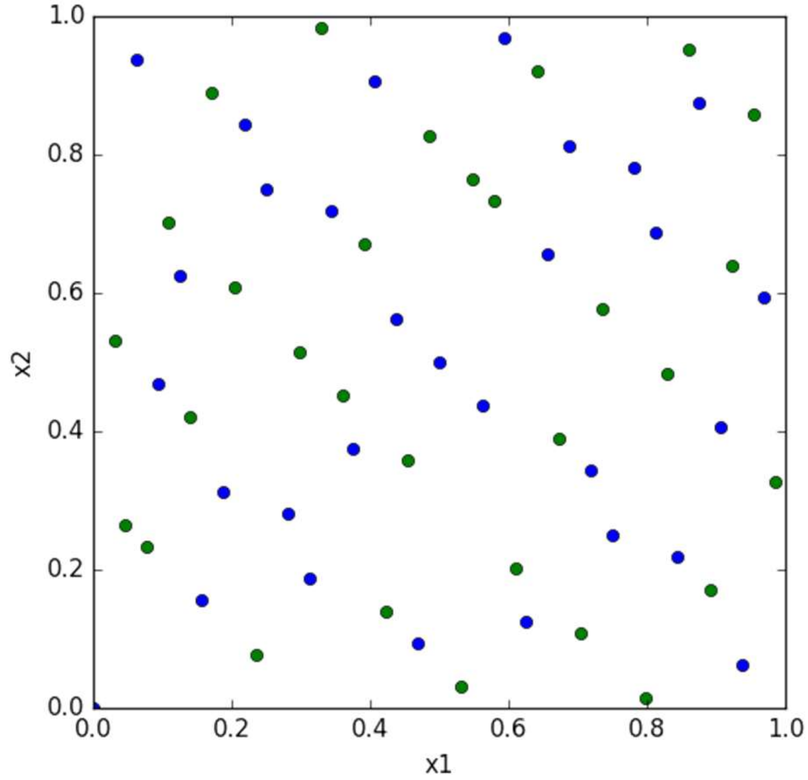


Sobol sequence (Quasi Monte Carlo sampling)

- Low-discrepancy sequence (uniformity)
- Reproducibility
- The uniformity is kept in high dimensions (especially Sobol).

Design of Experiments (DoE)

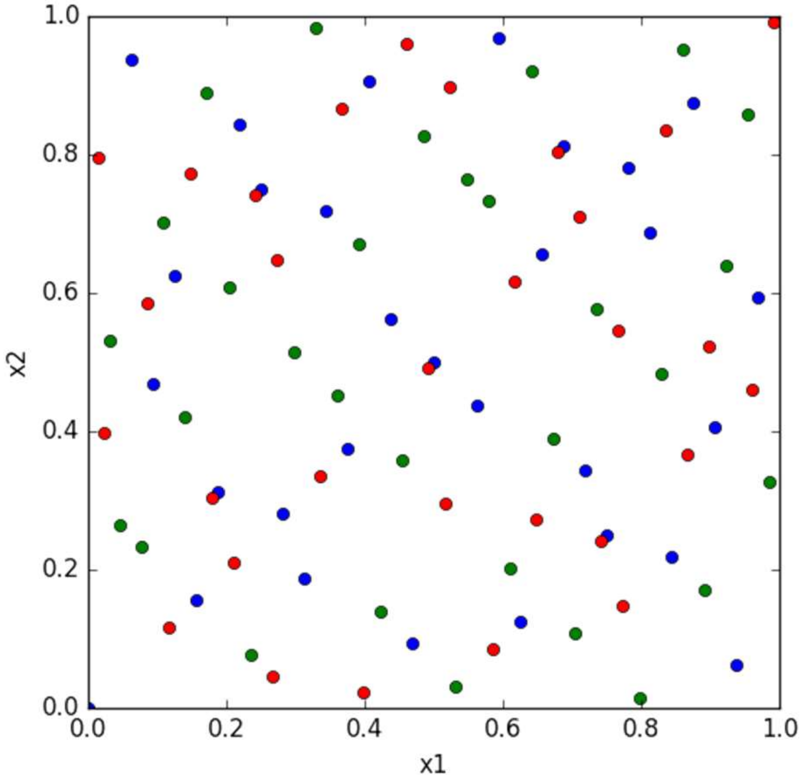
```
[[ 0.         0.        ]
 [ 0.5        0.5       ]
 [ 0.75       0.25      ]
 [ 0.25       0.75      ]
 [ 0.375      0.375     ]
 [ 0.875      0.875     ]
 [ 0.625      0.125     ]
 [ 0.125      0.625     ]
 [ 0.1875     0.3125    ]
 [ 0.6875     0.8125    ]
 [ 0.9375     0.0625    ]
 [ 0.4375     0.5625    ]
 [ 0.3125     0.1875    ]
 [ 0.8125     0.6875    ]
 [ 0.5625     0.4375    ]
 [ 0.0625     0.9375    ]
 [ 0.09375    0.46875   ]
 [ 0.59375    0.96875   ]
 [ 0.84375    0.21875   ]
 [ 0.34375    0.71875   ]
 [ 0.46875    0.09375   ]
 [ 0.96875    0.59375   ]
 [ 0.71875    0.34375   ]
 [ 0.21875    0.84375   ]
 [ 0.15625    0.15625   ]
 [ 0.65625    0.65625   ]
 [ 0.90625    0.40625   ]
 [ 0.40625    0.90625   ]
 [ 0.28125    0.28125   ]
 [ 0.78125    0.78125   ]
 [ 0.53125    0.03125   ]
 [ 0.03125    0.53125   ]
 ...,
 [ 0.05273438  0.23242188]
 [ 0.55273438  0.73242188]
 [ 0.80273438  0.48242188]
 [ 0.30273438  0.98242188]
 [ 0.42773438  0.35742188]
 [ 0.92773438  0.85742188]
 [ 0.67773438  0.10742188]
 [ 0.17773438  0.60742188]
 [ 0.24023438  0.41992188]
 [ 0.74023438  0.91992188]
 [ 0.99023438  0.16992188]
 [ 0.49023438  0.66992188]]
```



New sample points can be added keeping the uniformity.

Design of Experiments (DoE)

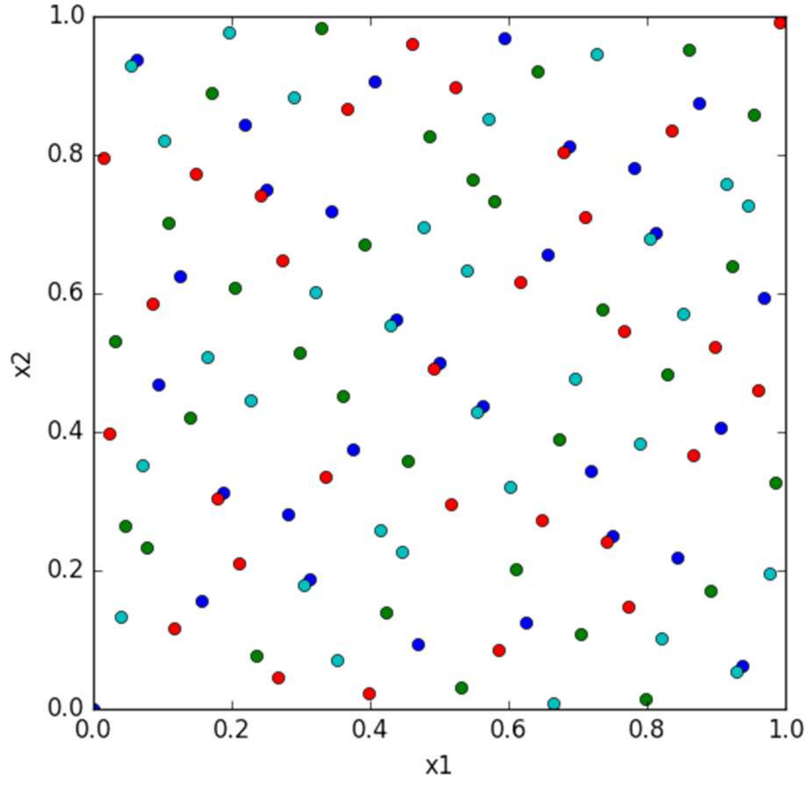
```
[[ 0.          0.        ]
 [ 0.5         0.5       ]
 [ 0.75        0.25      ]
 [ 0.25        0.75      ]
 [ 0.375       0.375     ]
 [ 0.875       0.875     ]
 [ 0.625       0.125     ]
 [ 0.125       0.625     ]
 [ 0.1875      0.3125    ]
 [ 0.6875      0.8125    ]
 [ 0.9375      0.0625    ]
 [ 0.4375      0.5625    ]
 [ 0.3125      0.1875    ]
 [ 0.8125      0.6875    ]
 [ 0.5625      0.4375    ]
 [ 0.0625      0.9375    ]
 [ 0.09375     0.46875   ]
 [ 0.59375     0.96875   ]
 [ 0.84375     0.21875   ]
 [ 0.34375     0.71875   ]
 [ 0.46875     0.09375   ]
 [ 0.96875     0.59375   ]
 [ 0.71875     0.34375   ]
 [ 0.21875     0.84375   ]
 [ 0.15625     0.15625   ]
 [ 0.65625     0.65625   ]
 [ 0.90625     0.40625   ]
 [ 0.40625     0.90625   ]
 [ 0.28125     0.28125   ]
 [ 0.78125     0.78125   ]
 [ 0.53125     0.03125   ]
 [ 0.03125     0.53125   ]
 ...,
 [ 0.05273438  0.23242188]
 [ 0.55273438  0.73242188]
 [ 0.80273438  0.48242188]
 [ 0.30273438  0.98242188]
 [ 0.42773438  0.35742188]
 [ 0.92773438  0.85742188]
 [ 0.67773438  0.10742188]
 [ 0.17773438  0.60742188]
 [ 0.24023438  0.41992188]
 [ 0.74023438  0.91992188]
 [ 0.99023438  0.16992188]
 [ 0.49023438  0.66992188]]
```



New sample points can be added keeping the uniformity.

Design of Experiments (DoE)

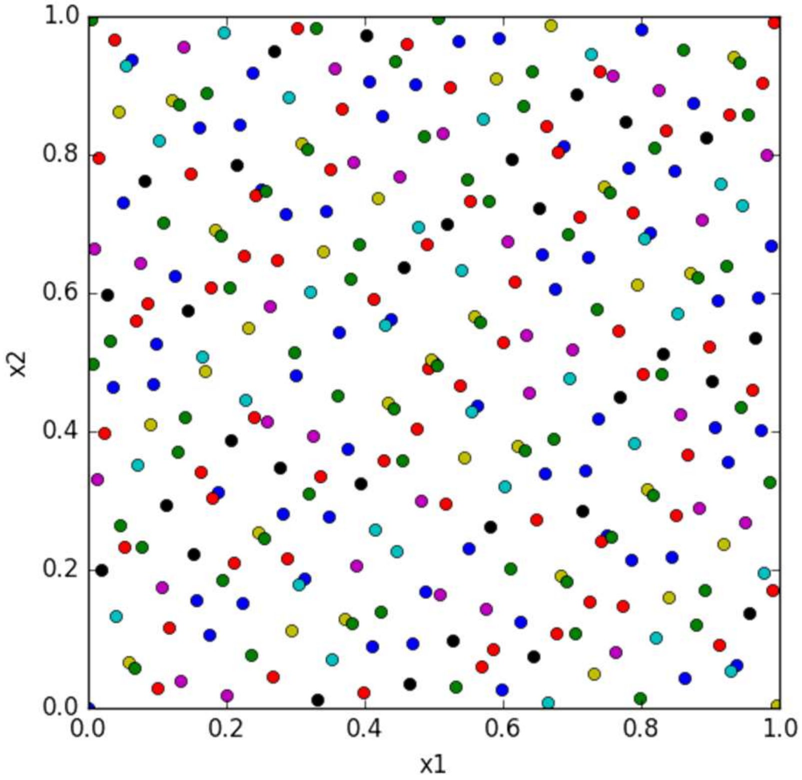
```
[[ 0.          0.        ]
 [ 0.5         0.5       ]
 [ 0.75        0.25      ]
 [ 0.25        0.75      ]
 [ 0.375       0.375     ]
 [ 0.875       0.875     ]
 [ 0.625       0.125     ]
 [ 0.125       0.625     ]
 [ 0.1875      0.3125    ]
 [ 0.6875      0.8125    ]
 [ 0.9375      0.0625    ]
 [ 0.4375      0.5625    ]
 [ 0.3125      0.1875    ]
 [ 0.8125      0.6875    ]
 [ 0.5625      0.4375    ]
 [ 0.0625      0.9375    ]
 [ 0.09375     0.46875   ]
 [ 0.59375     0.96875   ]
 [ 0.84375     0.21875   ]
 [ 0.34375     0.71875   ]
 [ 0.46875     0.09375   ]
 [ 0.96875     0.59375   ]
 [ 0.71875     0.34375   ]
 [ 0.21875     0.84375   ]
 [ 0.15625     0.15625   ]
 [ 0.65625     0.65625   ]
 [ 0.90625     0.40625   ]
 [ 0.40625     0.90625   ]
 [ 0.28125     0.28125   ]
 [ 0.78125     0.78125   ]
 [ 0.53125     0.03125   ]
 [ 0.03125     0.53125   ]
 ...,
 [ 0.05273438  0.23242188]
 [ 0.55273438  0.73242188]
 [ 0.80273438  0.48242188]
 [ 0.30273438  0.98242188]
 [ 0.42773438  0.35742188]
 [ 0.92773438  0.85742188]
 [ 0.67773438  0.10742188]
 [ 0.17773438  0.60742188]
 [ 0.24023438  0.41992188]
 [ 0.74023438  0.91992188]
 [ 0.99023438  0.16992188]
 [ 0.49023438  0.66992188]]
```



New sample points can be added keeping the uniformity.

Design of Experiments (DoE)

```
[[ 0.         0.        ]
 [ 0.5        0.5       ]
 [ 0.75       0.25      ]
 [ 0.25       0.75      ]
 [ 0.375      0.375     ]
 [ 0.875      0.875     ]
 [ 0.625      0.125     ]
 [ 0.125      0.625     ]
 [ 0.1875     0.3125    ]
 [ 0.6875     0.8125    ]
 [ 0.9375     0.0625    ]
 [ 0.4375     0.5625    ]
 [ 0.3125     0.1875    ]
 [ 0.8125     0.6875    ]
 [ 0.5625     0.4375    ]
 [ 0.0625     0.9375    ]
 [ 0.09375    0.46875   ]
 [ 0.59375    0.96875   ]
 [ 0.84375    0.21875   ]
 [ 0.34375    0.71875   ]
 [ 0.46875    0.09375   ]
 [ 0.96875    0.59375   ]
 [ 0.71875    0.34375   ]
 [ 0.21875    0.84375   ]
 [ 0.15625    0.15625   ]
 [ 0.65625    0.65625   ]
 [ 0.90625    0.40625   ]
 [ 0.40625    0.90625   ]
 [ 0.28125    0.28125   ]
 [ 0.78125    0.78125   ]
 [ 0.53125    0.03125   ]
 [ 0.03125    0.53125   ]
 ...,
 [ 0.05273438  0.23242188]
 [ 0.55273438  0.73242188]
 [ 0.80273438  0.48242188]
 [ 0.30273438  0.98242188]
 [ 0.42773438  0.35742188]
 [ 0.92773438  0.85742188]
 [ 0.67773438  0.10742188]
 [ 0.17773438  0.60742188]
 [ 0.24023438  0.41992188]
 [ 0.74023438  0.91992188]
 [ 0.99023438  0.16992188]
 [ 0.49023438  0.66992188]]
```

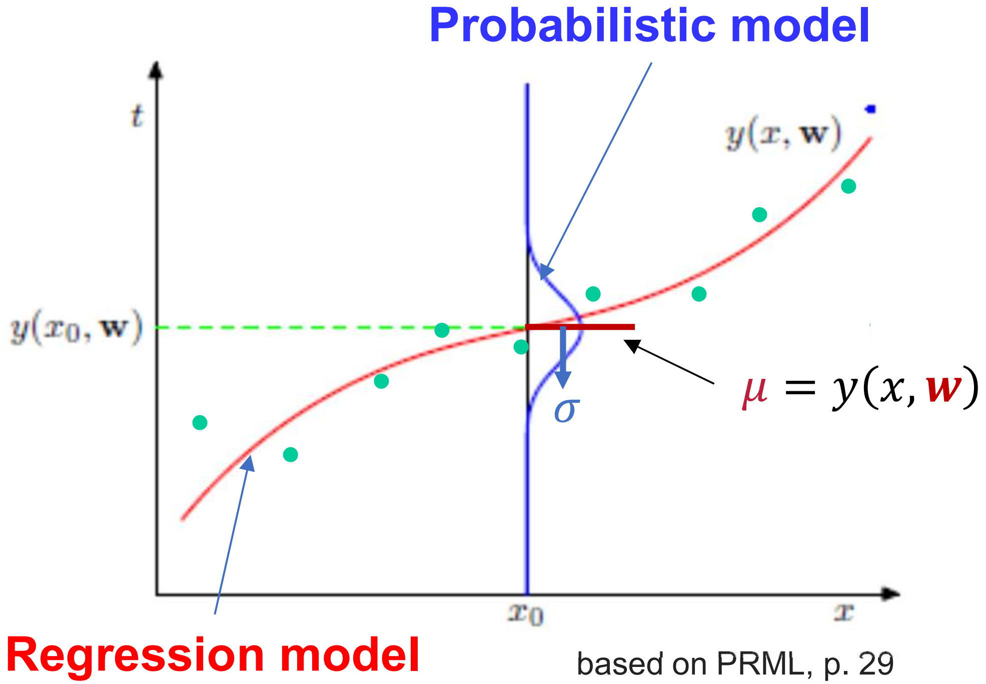


New sample points can be added keeping the uniformity.

3. Linear Regression

Regression Model

In the curve fitting problem, we defined a probabilistic model.



based on PRML, p. 29

Probabilistic model $p(t|x)$

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

e.g. $p(t|x)$ is a Gaussian distribution.

Regression model $E[t|x]$

$$\mu = y(x, \boldsymbol{w})$$

e.g. $y(x, \boldsymbol{w})$ is a polynomial function.

x: deterministic variable
t: random variable

The regression model is a part of the probabilistic model*.

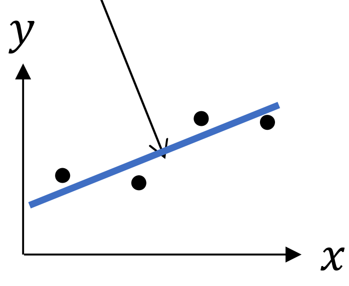
Regression model ∈ **Probabilistic model**

*There are a few exceptions (e.g. support vector machine).

Regression Model

The simplest linear regression

$$y = w_0 + w_1 x$$



$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{argmin}\, E(\boldsymbol{w})$$

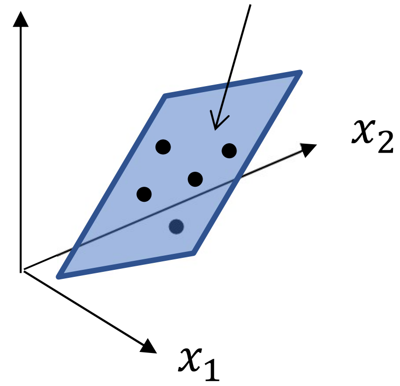
$$E(\boldsymbol{w}) = \sum_{i=1}^{N} \{y_i - y(x_i, \boldsymbol{w})\}^2$$

$$\begin{cases} \dfrac{\partial E(\boldsymbol{w})}{\partial w_0} = 0 \\ \dfrac{\partial E(\boldsymbol{w})}{\partial w_1} = 0 \end{cases}$$

$\Rightarrow$ $\boxed{\hat{\boldsymbol{w}} \text{ is analytically solved.}}$

Please confirm this by yourself.

$$y(x, \boldsymbol{w}) = w_0 + w_1 x_1 + w_2 x_2 = \sum_{i=0}^{2} w_i x_i$$



$D = 2$ in this case

$(D + 1)$ parameters

Parameters: dependent on the number of the input

The components of $\boldsymbol{x}$

$$\mathbf{X} = \begin{array}{c} \overbrace{\hphantom{x_{11} \cdots x_{1D}}}^{x_{i1}, \cdots, x_{iD}} \\ \begin{pmatrix} 1 & \boldsymbol{x}_{11} & \cdots & \boldsymbol{x}_{1D} \\ 1 & \vdots & \ddots & \vdots \\ 1 & \boldsymbol{x}_{N1} & \cdots & \boldsymbol{x}_{ND} \end{pmatrix} \end{array} \left.\vphantom{\begin{pmatrix}1\\1\\1\end{pmatrix}}\right\} \begin{array}{c} \text{sample} \\ \boldsymbol{x}_1, \cdots \boldsymbol{x}_N \end{array}$$

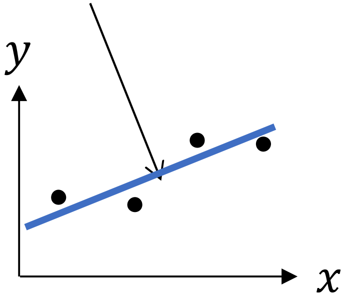
$$\boxed{\hat{\boldsymbol{w}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{t}}$$

analytically solved

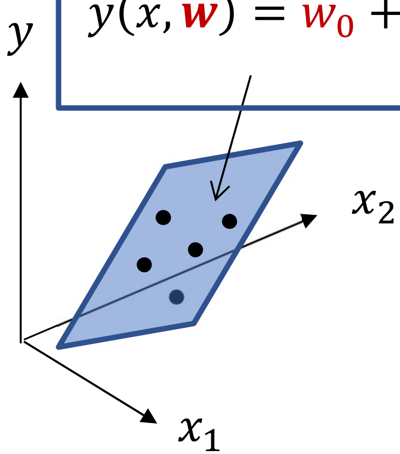
Regression Model

The simplest linear regression

$$y = w_0 + w_1 x$$



$$y(x, \boldsymbol{w}) = w_0 + w_1 x_1 + w_2 x_2 = \sum_{i=0}^{2} w_i x_i$$



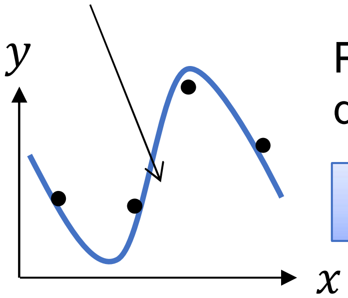
$D = 2$ in this case

$(D + 1)$ parameters

Parameters:
dependent on the number of the input

Nonlinear model (**still linear regression**)

$$y(x, \boldsymbol{w}) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 = \sum_{i=0}^{3} w_i x^i$$



Parameters:
dependent on your model

analytically solved?

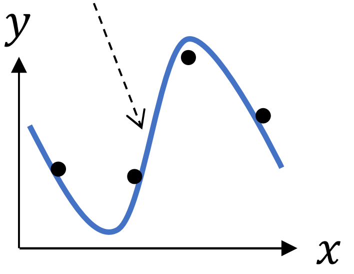
- D-dimensional input
- M-degrees

The parameters $\boldsymbol{w}$ increase exponentially.

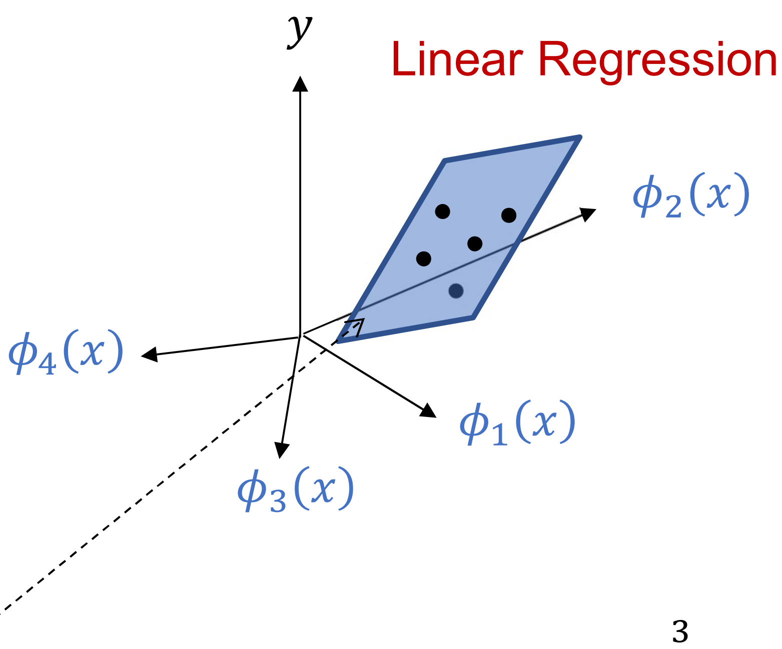
Linear Regression

Shifting to the simplest linear regression by mapping $\boldsymbol{\phi}$: $\boldsymbol{x} \rightarrow \boldsymbol{s}$ $\quad$ $\boldsymbol{s} = \boldsymbol{\phi}(\boldsymbol{x})$

$$y(x, \boldsymbol{w}) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 = \sum_{i=0}^{3} w_i x^i$$



from 1D to 4D
in this example



In this example:

$$\boldsymbol{\phi}(x) = \big(\phi_0(x), \phi_1(x), \phi_2(x), \phi_3(x)\big)^{\mathrm{T}}$$
$$= (x^0, x^1, x^2, x^3)^{\mathrm{T}}$$

$$y(\boldsymbol{s}, \boldsymbol{a}) = a_0 s_0 + a s_1 + a_2 s_2 + a_3 s_3 = \sum_{i=0}^{3} a_i s_i$$

Linear Regression

Examples of $\boldsymbol{\phi}(x)$:

$$\phi_i(x) = \exp\left\{-\frac{(x-\mu_i)^2}{2\sigma^2}\right\}$$

Gaussian function

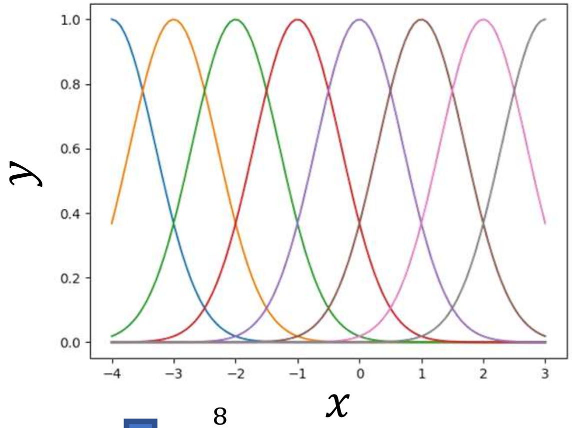
$\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_i, \cdots, \mu_8)$ and σ are user-defined.

We can use any nonlinear functions for $\boldsymbol{\phi}(\boldsymbol{x})$.

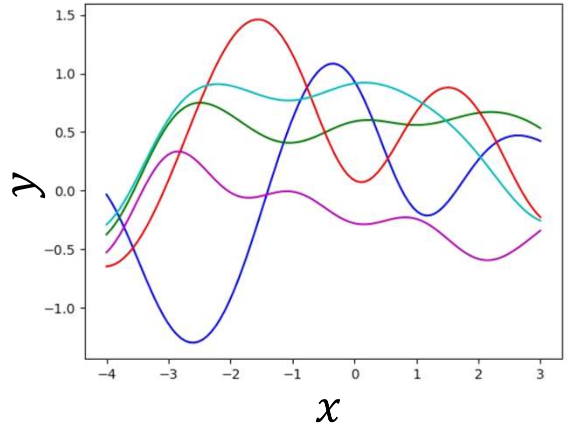
5 examples from randomly generated $\boldsymbol{w}$

$\boldsymbol{w}$ is composed of 8 elements.

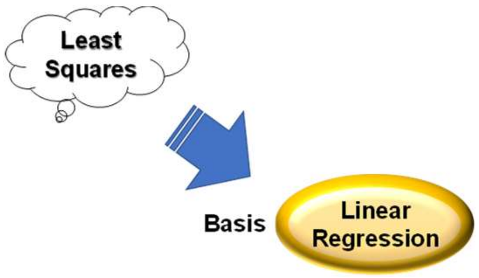
8 Gaussian functions $\boldsymbol{\phi}(x)$



$$\sum_{i=1}^{8} w_i \phi_i(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x})$$



Linear Regression



a linear function of the parameters $\boldsymbol{w}$

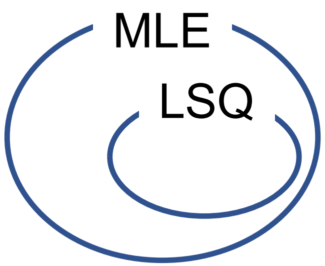
Linear regression (in general)

$$y(\boldsymbol{x},\boldsymbol{w}) = w_0 + w_1\phi_1(\boldsymbol{x}) + w_2\phi_2(\boldsymbol{x}) + \cdots + w_M\phi_M(\boldsymbol{x}) = \sum_{i=0}^{M} w_i\phi_i(\boldsymbol{x}) = \boxed{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x})}$$

M can be freely defined.

- $(M+1)$ parameters $\boldsymbol{w} = (w_0, w_1, w_2, \cdots, w_M)^{\mathrm{T}}$
- no matter how many dimensionality the original input $\boldsymbol{x}$ has

Then,



with a defined probabilistic model

Linear Regression

Likelihood function

$$L(\boldsymbol{w},\sigma) \equiv -\ln p(\mathbf{t}|\mathbf{x},\boldsymbol{w},\sigma^2) = \underbrace{\frac{1}{2\sigma^2}\sum_{i=1}^{N}\{y_i - \boxed{y(x_i,\boldsymbol{w})}\}^2}_{\text{the least square term}} + \frac{N}{2}\ln(2\pi\sigma^2)$$

negative log of the likelihood function

In Linear Regression: $y(\boldsymbol{x},\boldsymbol{w}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x})$

Maximum Likelihood Estimation (MLE)

$$\widehat{\boldsymbol{w}},\hat{\sigma} = \underset{\boldsymbol{w},\sigma}{argmin}\, L(\boldsymbol{w},\sigma)$$

Need optimization algorithm?

Linear Regression

M : number of the parameter $\boldsymbol{w}$
N : sample size

$$y(\boldsymbol{x}, \boldsymbol{w}) = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x})$$

$$\frac{\partial L(\boldsymbol{w},\sigma)}{\partial \boldsymbol{w}} = 0 \text{ or } \frac{\partial E(\boldsymbol{w})}{\partial \boldsymbol{w}} = 0$$

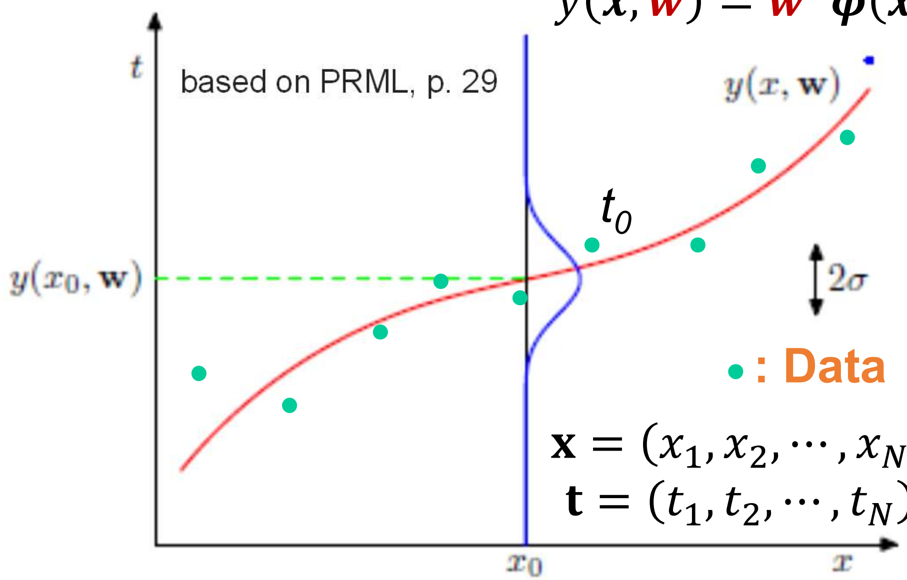
analytically solved

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

$M \times M$ square matrix

$$\boldsymbol{\Phi}^{*-1} \equiv \left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}$$

$\boldsymbol{\Phi}^{*}\hat{\boldsymbol{w}} = \mathbf{t}$ a linear system



$$\mathbf{x} = (x_1, x_2, \cdots, x_N)$$
$$\mathbf{t} = (t_1, t_2, \cdots, t_N)$$

The components of $\boldsymbol{\phi}(\boldsymbol{x})$

$\phi_1(\boldsymbol{x}), \cdots, \phi_M(\boldsymbol{x})$

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_1(\boldsymbol{x}_1) & \cdots & \phi_M(\boldsymbol{x}_1) \\ \vdots & \ddots & \vdots \\ \phi_1(\boldsymbol{x}_N) & \cdots & \phi_M(\boldsymbol{x}_N) \end{pmatrix}$$

sample

$\boldsymbol{x}_1, \cdots \boldsymbol{x}_N$

$N \times M$ matrix

Pseudo inverse matrix

Linear Regression

Important properties of Linear Regression

- The regression model is a linear function of the parameters $\boldsymbol{w} = (w_0, w_1, w_2, \cdots, w_M)^{\mathrm{T}}$
- The function $\boldsymbol{\phi}(\boldsymbol{x}) = \left(\phi_0(\boldsymbol{x}), \phi_1(\boldsymbol{x}), \phi_2(\boldsymbol{x}), \cdots, \phi_M(\boldsymbol{x})\right)^{\mathrm{T}}$ can be nonlinear of the input $\boldsymbol{x}$.

⇨
- The number of the parameters $\boldsymbol{w}$ is M.
- Therefore, $\hat{\boldsymbol{w}}$ can be obtained analytically.

(Under the assumption that the probabilistic model is an isotropic Gaussian distribution)

Linear Regression

Regularization

Penalty on the parameter $\boldsymbol{w}$ to avoid overfitting

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmin}}\, E(\boldsymbol{w})$$

$$E(\boldsymbol{w}) = \sum_{i=1}^{N} \{y_i - \underbrace{y(x_i, \boldsymbol{w})}_{\boxed{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x})}}\}^2$$

s.t. $\|\boldsymbol{w}\|^2 \leq \eta$

Review:
The error function $E(\boldsymbol{w})$
= the simplest expression of the negative log likelihood.

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmin}}\, E_{reg}(\boldsymbol{w}) = \left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \lambda \mathbf{I}\right)^{-1} \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}$$

still analytically solved

where, $E_{reg}(\boldsymbol{w}) = E(\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|^2$

regularization term

Linear Regression

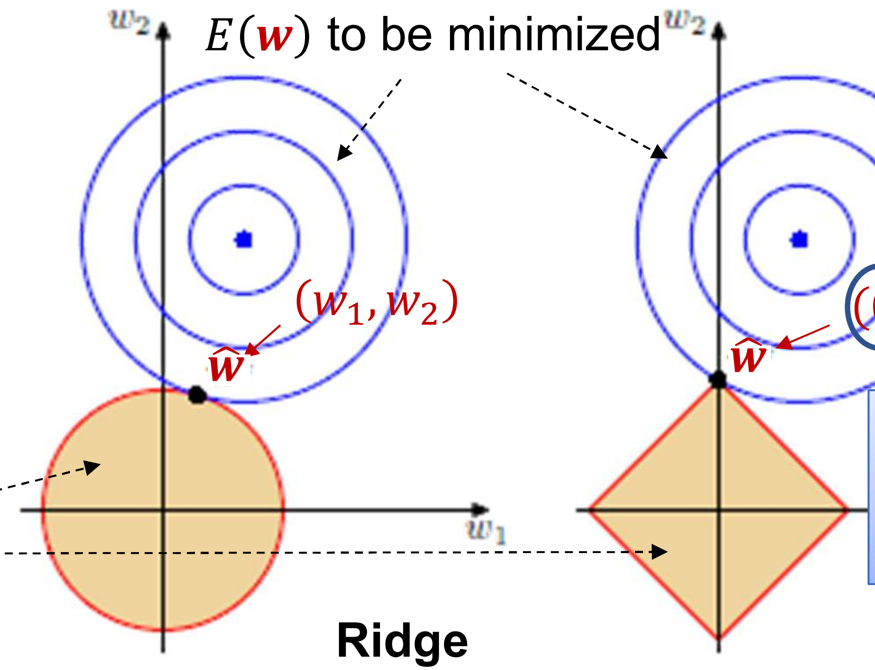
Other regularization techniques

regularization term

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmin}}\, E_{reg}(\boldsymbol{w}) \qquad \text{where,} \qquad E_{reg}(\boldsymbol{w}) = E(\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|^{q}$$

parameter $\boldsymbol{w}$ space

PRML, p. 146



A sparse model can be obtained!

But $\hat{\boldsymbol{w}}$ cannot be analytically obtained any more.

the domain $\boldsymbol{w}$ where can move

$\|\boldsymbol{w}\|^2 \leq \eta$

when $q = 2$ when $q = 1$

Summary until Linear Regression (1/2)

A larger framework than the least square method was introduced based on perspectives of the probability theory
(Some examples will be introduced in other lectures).

The concept (a procedure):

1. **Define a probabilistic model (using probability distributions)**
2. **Maximize the likelihood function determined by dataset**
3. **(if prior distributions are set, maximize the posterior distribution)**

Compute $\underbrace{\text{the log of the likelihood/posterior}}_{\text{error function}}$ to avoid numerical errors

error function the least square is one of the error functions.

- The concept is further extended to be generalized until the lectures of "Bayes".

Summary until Linear Regression (2/2)

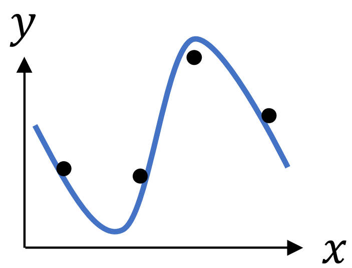
A regression model can be defined in the process of the definition of the probabilistic model.

- Linear regression model is one of the regression models.
  - The model is a linear function of the parameters $\boldsymbol{w}$

The 2nd process: MLE becomes easy (analytically obtained).

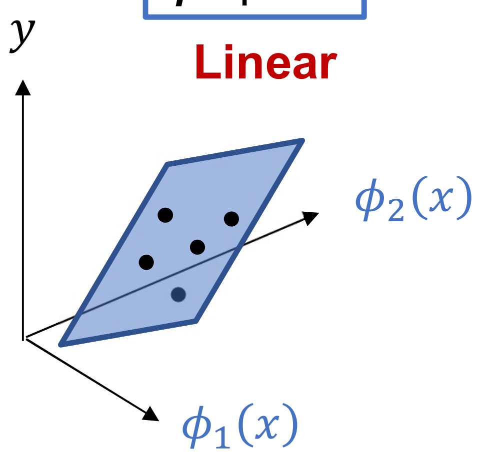
$\boldsymbol{x}$ space

Complex nonlinear function



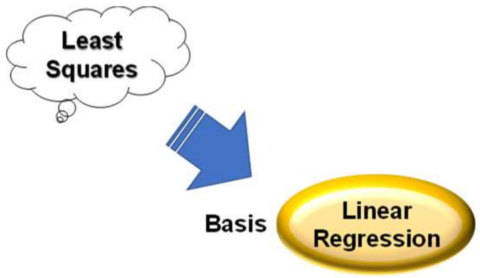
$\boldsymbol{\phi}$ space

Linear



This concept will be important.

Current Position to Next



We now reached to the “**origin**”.

The theory is extended by introducing Bayesian perspective by from the next lecture using two slots of lectures

The required tools will be all equipped.