Singular Learning Theory

1 Outline of Lecture

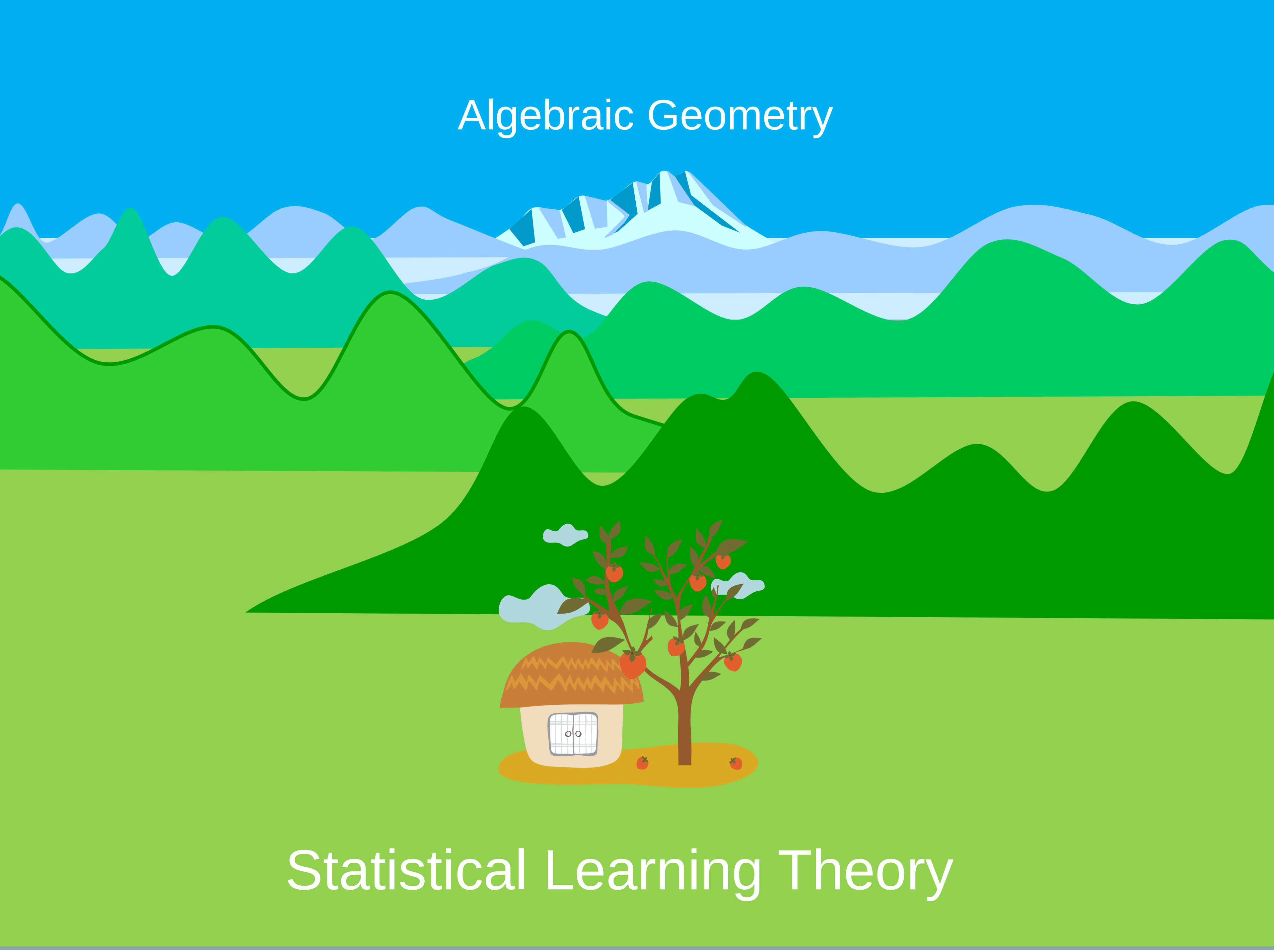
Sumio Watanabe
Tokyo Institute of Technology

1 Guidance of Lecture

A Journey from Math to AI.

This lecture is a journey from mathematics
to artificial intelligence.

First, we go to algebraic geometry
in a far world from statistical learning theory.



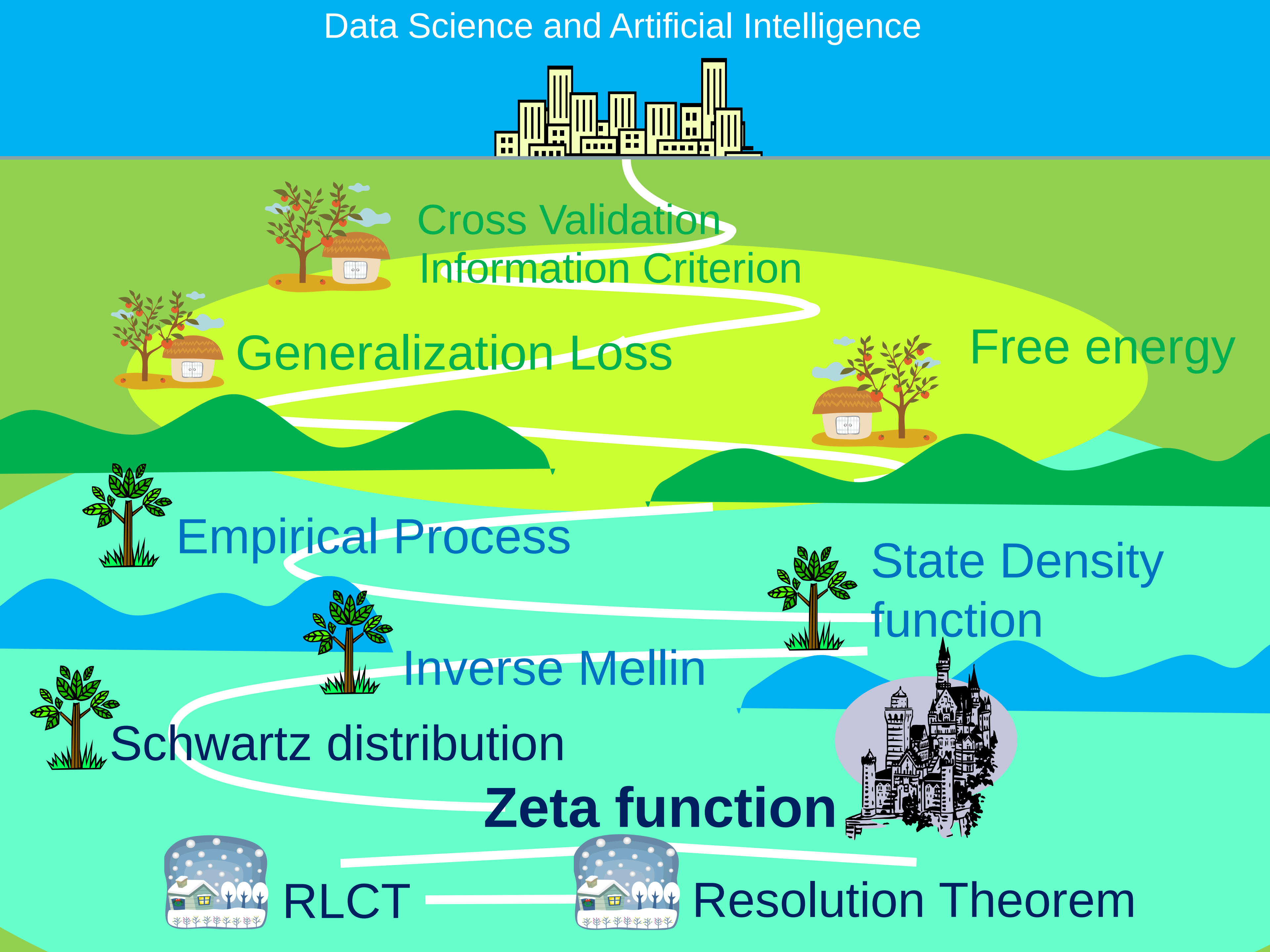
Algebraic Geometry

Statistical Learning Theory

Make a way back to artificial intelligence

Second we make our way back from algebraic geometry to statistical learning theory.

New research field will be opened.



Data Science and Artificial Intelligence

Cross Validation

Information Criterion

Generalization Loss

Free energy

Empirical Process

State Density function

Inverse Mellin

Schwartz distribution

Zeta function

RLCT

Resolution Theorem

Check Prerequisite Knowledge by yourself

If you do not know the following concepts, you had better learn them in undergraduate course.

(1) set theory and topology,
(2) probability space,
(3) central limit theorem,
(4) ring and ideal,
(5) meromorphic function,
(6) analytic continuation,
(7) metric space C[0,1],
(8) delta function $\delta(x)$,

and basic mathematical concepts.

2 Outline of Lecture

Outline of the lecture

In the first day of this lecture, the outline is explained.

In the lecture after the second day, definitions, theorems, and proofs will be given. However, in the first day, they are omitted, and summary is explained.

You are recommended to decide whether you attend this lecture or not, based on this outline.

2 Outline of Lecture

2.1 Definition of Learning Curve

True distribution and sample

Let X be an $\mathbf{R}^N$ valued random variable whose probability density function (p.d.f) is q(x).

Let $X^n=(X_1, X_2, \ldots, X_n)$ be a set of independent random variables which have the same p.d.f. as X.

Then q(x) is called a true distribution and n is referred to as a sample size (or the number of training examples).

Also q(x) is called a data-generating distribution which is unknown uncertainty in real problems.

Statistical model and prior

A probability density function $p(x|w)$ of x in $\mathbf{R}^N$ for a given parameter w in $\mathbf{R}^d$ is called a statistical model or a learning machine.

A probability density function on $\mathbf{R}^d$, $\varphi(w)$, is called a prior.

Note. Both a model and a prior are fictional candidates, because uncertainty is unknown.

Remark. In supervised learning, $p(y|x,w)$ is employed instead of $p(x|w)$, but the same mathematical theory holds.

Model Setting

When a statistical model $p(x|w)$ and a prior distribution $\varphi(w)$ are determined, they are represented by

$$w \sim \varphi(w),$$

$$X^n \sim \prod_{i=1}^{n} p(X_i|w).$$

If this is a generating process of X^n, the simultaneous distribution of X^n does not depend on the order of $\{X_i\}$.
In such a case, X^n is called exchangeable.

Existence of True Distribution

If X^n is exchangeable, then by de Finetti’s theorem, there exist both a probability distribution q(x) and a functional probability distribution Q(q) such that

$$q(x) \sim Q(q),$$

$$X^n \sim \Pi\, q(X_i).$$

In other words, unknown true distributions Q(q) and q(x) mathematically exist for a person, an agent, or
an artificial intelligence who prepares p(x|w) and φ(w).

Purpose of Statistical Leaning Theory

A person, an agent, or an artificial intelligence makes a statistical model $p(x|w)$ and a prior $\varphi(w)$.

Of course, $p(x|w)$ and $\varphi(w)$ are fictional candaidates, which may or may not be appropriate for the unknown uncertainty $q(x)$.

The main purpose of the statistical learning theory is to establish a mathematical foundation to study learning phenomena for an arbitrary triple $(q(x), p(x|w), \varphi(w))$.

Definition : Posterior distribution

For a given sample X^n , the posterior distribution is defined by

$$p(w|X^n) = \ (1/Z)\ \varphi(w) \prod_{i=1}^{n} p(X_i|w),$$

where Z is a normalizing constant,

$$Z= \int \varphi(w) \prod_{i=1}^{n} p(X_i|w)\ dw,$$

which is referred to as the marginal likelihood.
The expectation value and the variance by the posterior distribution are denoted by $\mathbf{E}_w[\ ]$ and $\mathbf{V}_w[\ ]$, respectively.

Definition: Predictive distribution

The predictive distribution is defined by

$$p^*(x) = p(x|X^n) = \mathbf{E}_w[p(x|w)],$$

by which the true distribution q(x) is estimated.

The generalization error is defined by

$$G_n^{(0)} = \int q(x) \log \{ q(x) / p^*(x) \} dx.$$

Note: $G_n^{(0)}$ is a random variable, since it depends on X^n.

Why is statistical learning theory necessary ?

Both a statistical model p(x|w) and a prior φ(w) are made by a person, an agent, or an artificial intelligence who does not know the unknown uncertainty q(x).

Hence both the posterior and predictive distributions are not appropriate for q(x) in general.

We need to examine whether the predictive distribution p(x|X^n) is appropriate or not according to unknown q(x).

Can we compare p(x|X^n) with unknown uncertainty ?

Prerequisite : Kullback Leibler divergence

Let q(x), p(x) be p.d.f.s on $\mathbf{R}^N$. KL divergence is defined by

$$K(q||p) = \int q(x) \log (q(x)/p(x))\, dx.$$

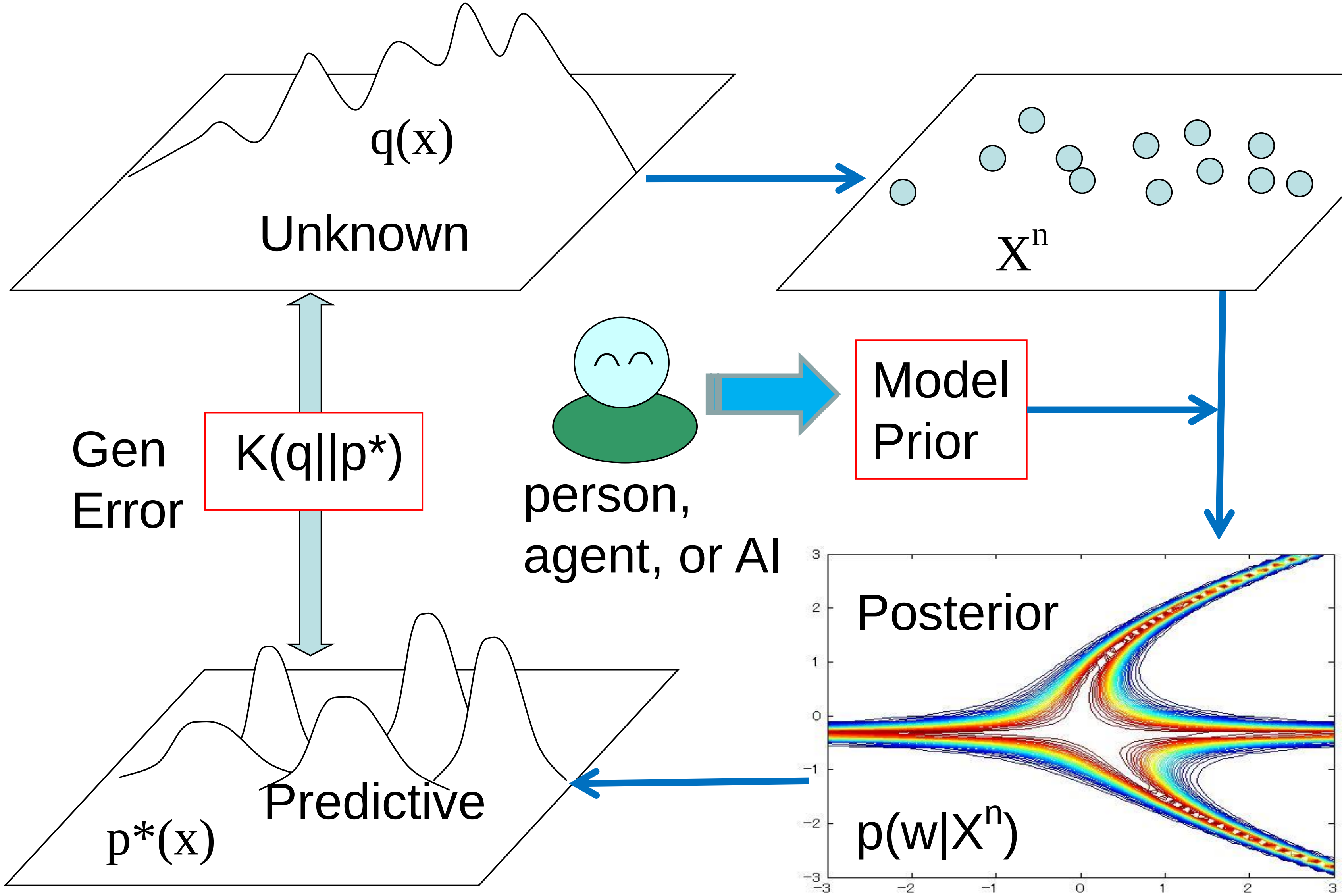
Then, for arbitrary continuous p.d.f.s q(x), p(x)>0,

(1) K(q||p)>=0.

(2) K(q||p)=0 if and only if q(x)=p(x) (for all x).

The generalization error satisfies $G_n^{(0)}$ =K(q||p*), which represents the difference between true and estimation.

Mathematical Structure of Statistical Learning



q(x)

Unknown

X^n

Model
Prior

Gen
Error

K(q||p*)

person,
agent, or AI

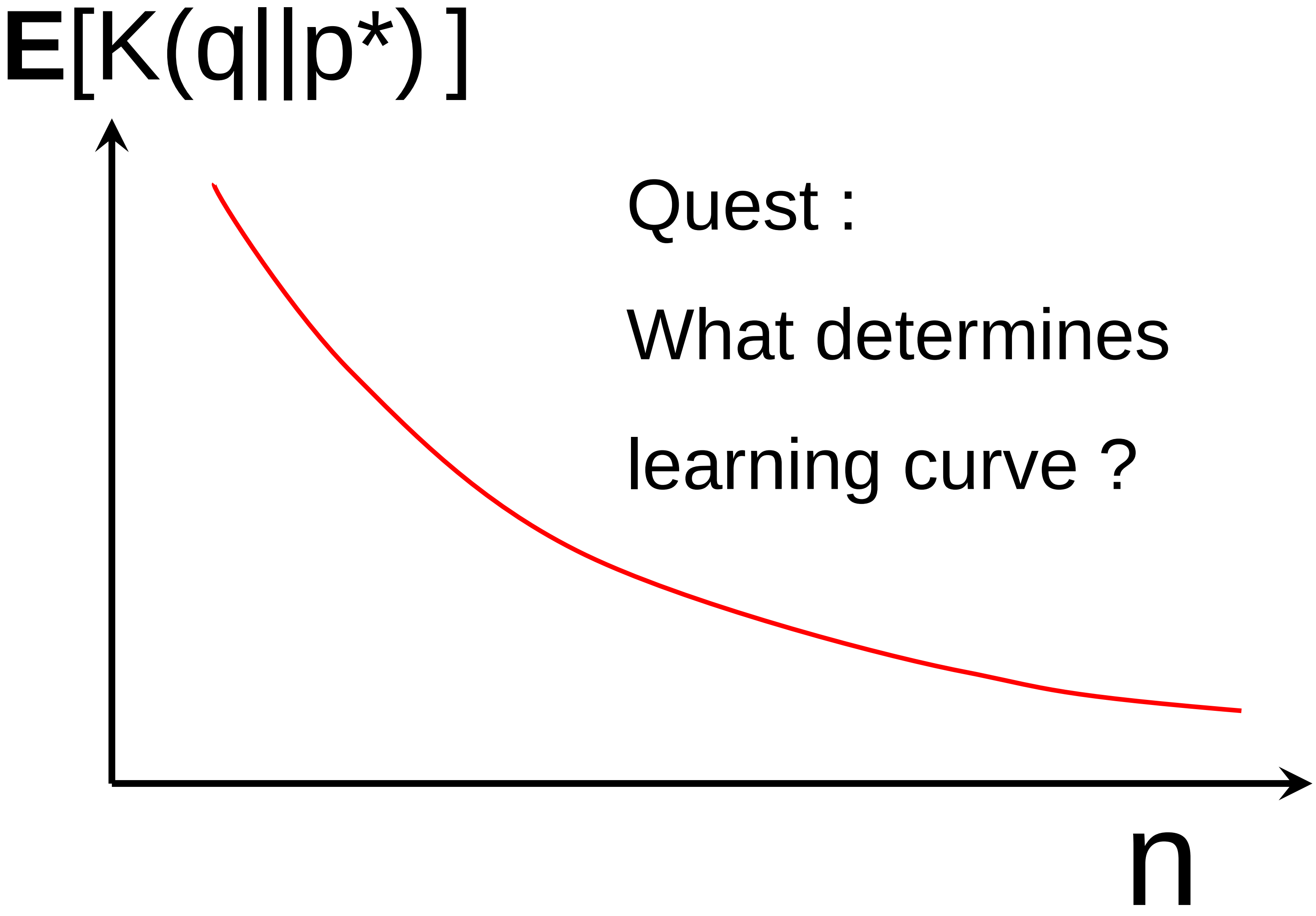
Posterior

$p(w|X^n)$

Predictive

p*(x)

Learning Curve

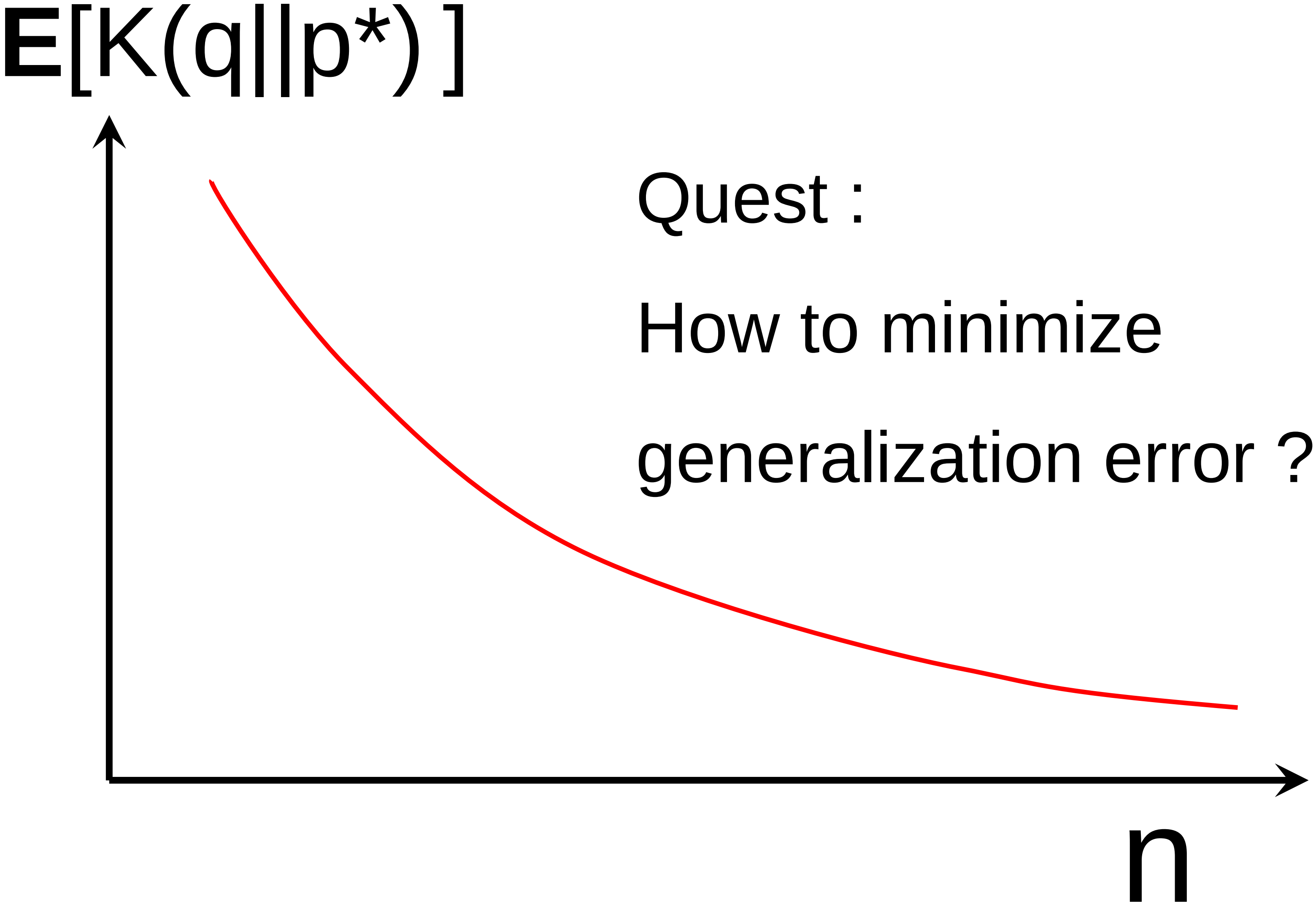


Quest :

What determines

learning curve ?

Minimize Generalization Error



Quest :

How to minimize

generalization error ?

Mathematical Learning Theory

In order to study the learning curve for an arbitrary $(q(x),p(x|w),\varphi(w))$, we needs almost all mathematical concepts.

2 Outline of Lecture

2.2 Mathematical Theory of Learning Curve

Hessian Matrix

Let K(w) be a function of w in $\mathbf{R}^d$.

The Hessian matrix at $w=w_0$ is defined by

$$J_{ij}(w_0)=(\partial^2/\partial w_i \partial w_j)\, K(w_0).$$

Example. Let $K(a,b)= a^2+b^4$. Hessian matrix at (a,b)=(1,1) is

$$J(1,1) = \begin{pmatrix} 2 & 0 \\ 0 & 12 \end{pmatrix}$$

Regular case

Let $K(w)=K(q(x)||p(x|w))$ and let w_0 be the parameter that minimizes $K(w)$.

If the Hessian matrix $J_{ij}(w_0) =(\partial^2/\partial w_i \partial w_j) K(w_0)$ is positive definite (i.e. regular case), then

$$\mathbf{E}[G_n^{(0)}] = K(w_0)+ d/(2n) +o(1/n),$$

where d is the dimension of the parameter.

Remark. If $q(x)=p(x|w_0)$, then $K(w_0)=0$ and

$J_{ij}(w_0) =$ Fisher information matrix.

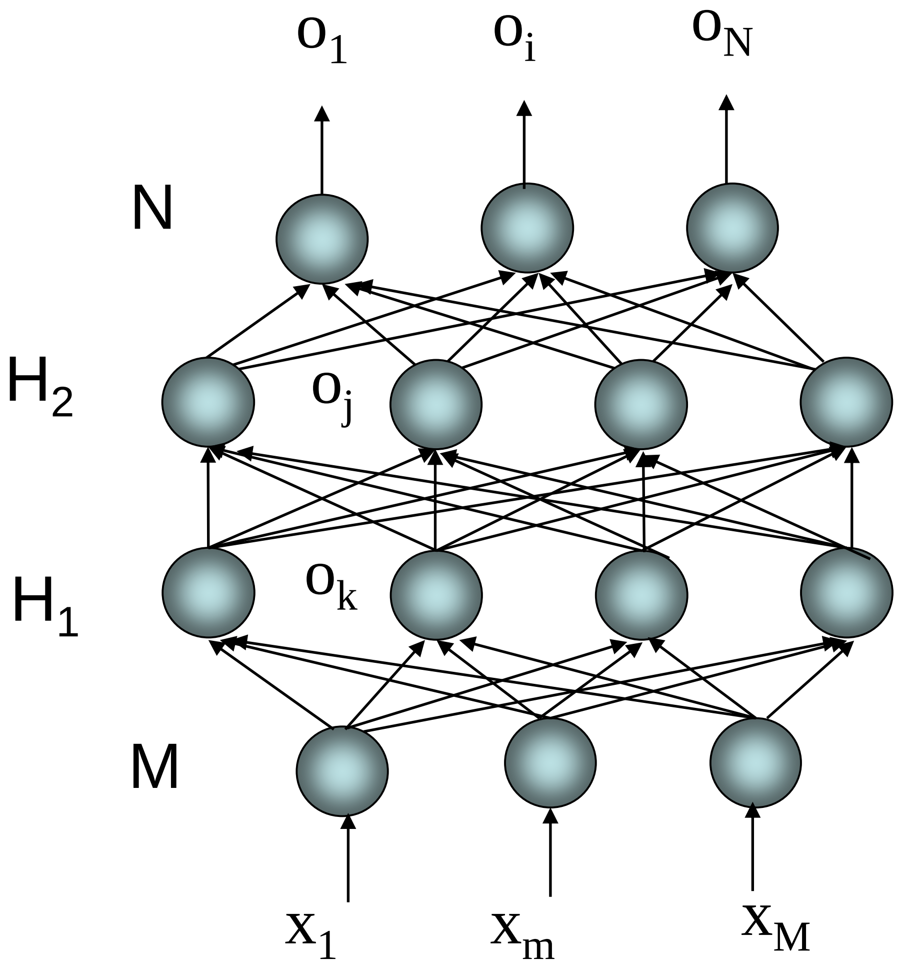
Almost all learning mchines are singular

If the Hessian matrix contains zero eigenvalue, such a case (or a model) is called singular.

If a statistical model has hierarchical structure or hidden variables, it becomes singular.

For example, neural networks, normal mixtures, Boltzmann machines, hidden Markov models, Matrix factorizations, Latent Dirichlet allocations, …, are singular.

An example of learning machines



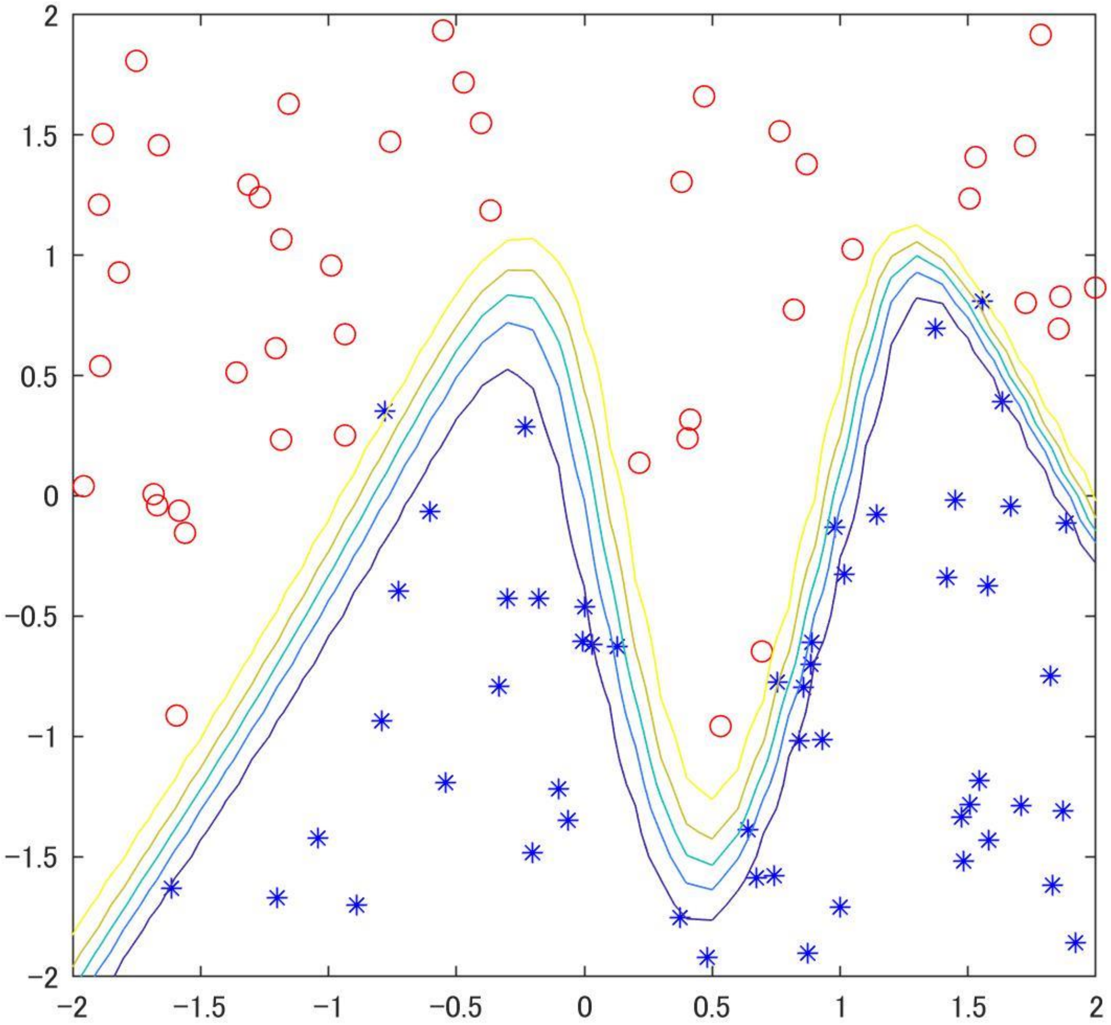
A neural network

is an example

of a singular

learning machines

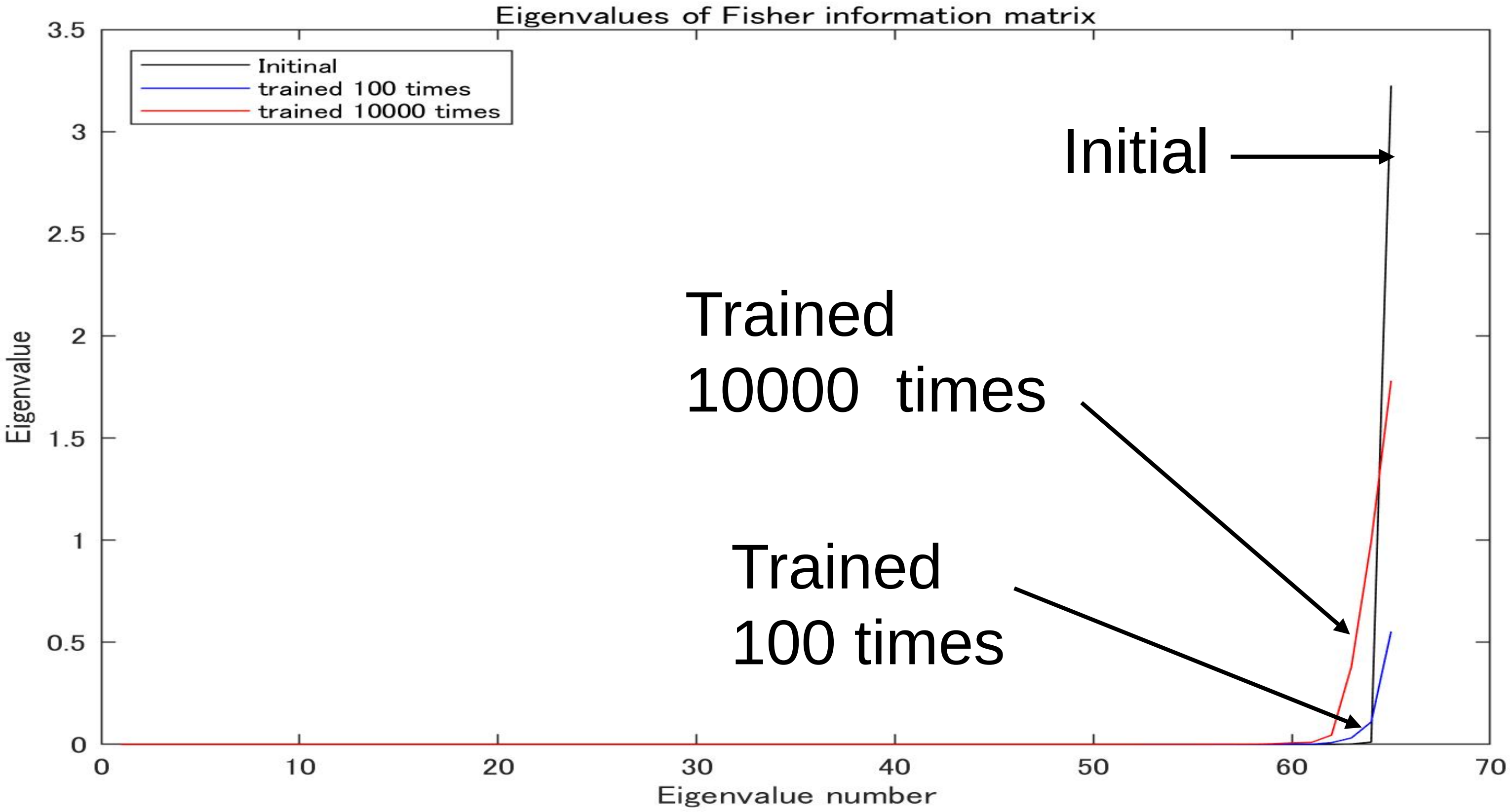
A simple classification by NN



A neural network which has 2-8-4-1 units were trained so that it classifies ‘o’ and ‘*’.

Here, n=100, trained cycle =10000.

Eigenvalues of Fisher information matrix in NN



Initial

Trained 10000 times

Trained 100 times

Eigenvalues sorted from small to large.

Main Theorem 1 : Singular learning curve

Main Theorem 1. Even if a statistical model is singular, there exists a constant λ >0, such that

$$\mathbf{E}[G_n^{(0)}] = K(w_0)+ \lambda /n +O(1/n).$$

The constant λ is called the real log canonical threshold (RLCT), whose definition is given by the following.

Zeta function and RLCT

The zeta function of a statistical model and prior is defined by

$$\zeta(z) = \int K(w)^z \varphi(w) dw.$$

Then $\zeta(z)$ is a holomorphic function in Re(z)>0. We can prove that it is uniquely and analytically continued to a meromorphic function onto the entire complex plane.

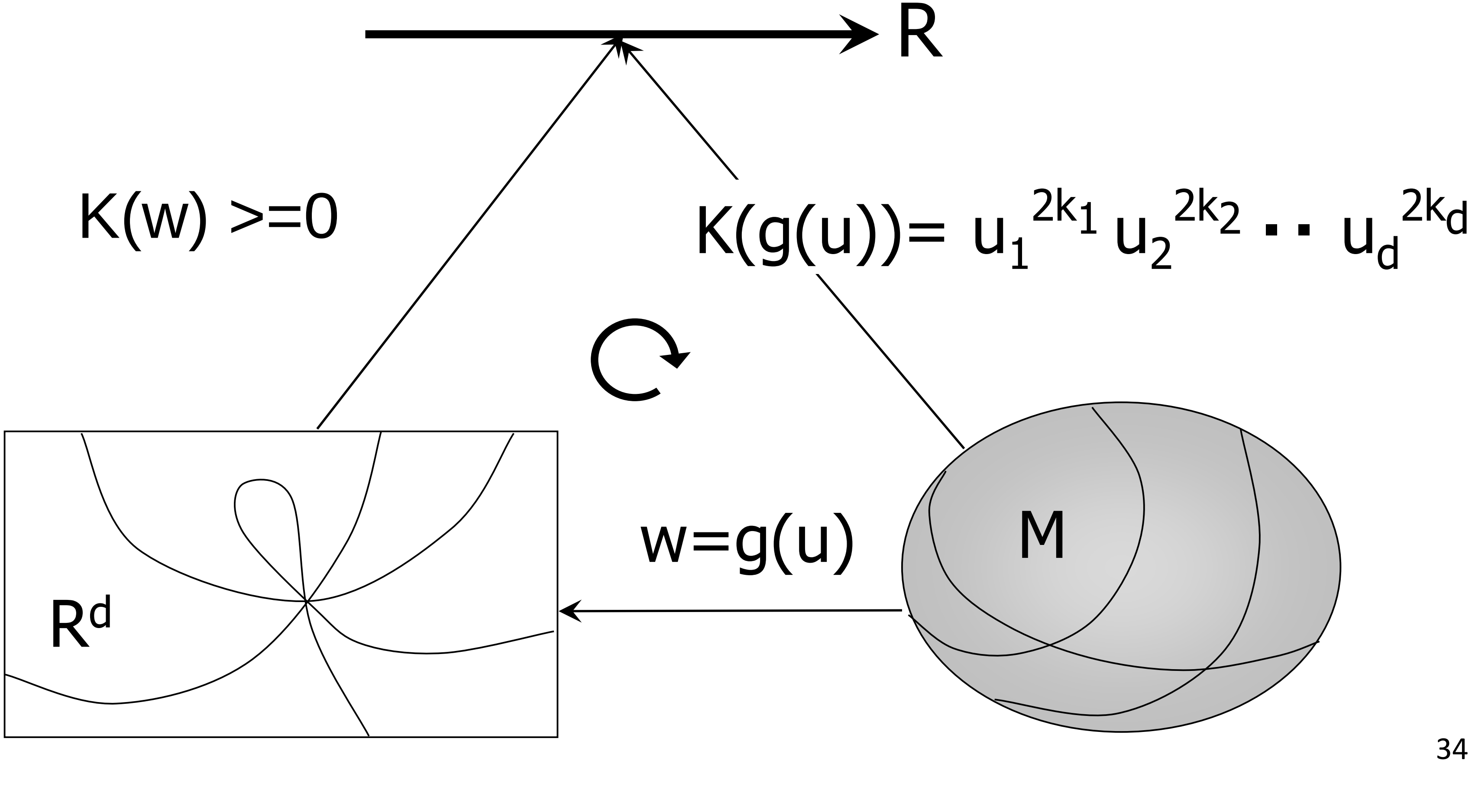
Zeta function and RLCT

Then we can prove that all poles of zeta function are real and negative. By the largest pole ($-\lambda$), RLCT is defined.

(Not prerequisite knowledge) RLCT is important in algebraic geometry. It can be discovered by Hironaka Resolution theorem. The resolution of singularities can be found by recursive blowups.

Hironaka Resolution Theorem

This theorem will be explained in the lecture.
This is not the prerequisite knowledge.



R

$K(w) >= 0$

$K(g(u))= u_1^{2k_1}\, u_2^{2k_2} \cdot\cdot\; u_d^{2k_d}$

$w=g(u)$

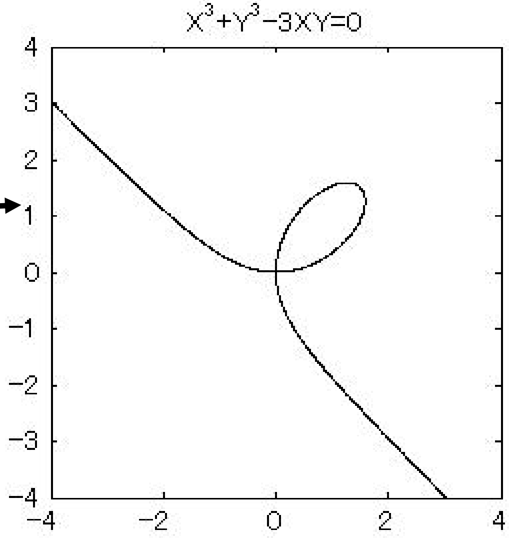
R^d

M

Example of blowup

This concept will be explained in the lecture.
This is not the prerequisite knowledge.

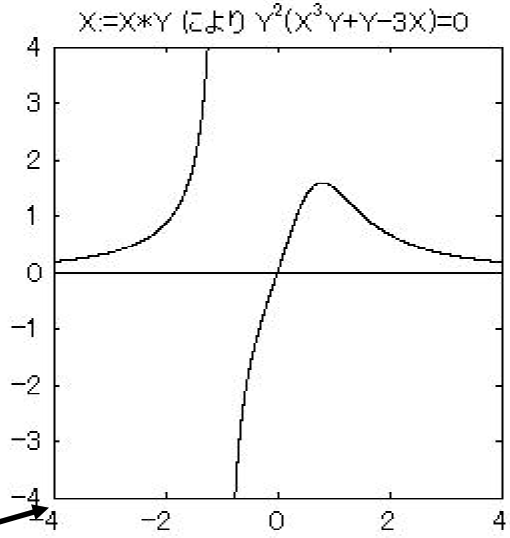
$x^3+y^3-3xy=0$


$X^3+Y^3-3XY=0$

$x=x_1y_1$
$y=y_1$

Then

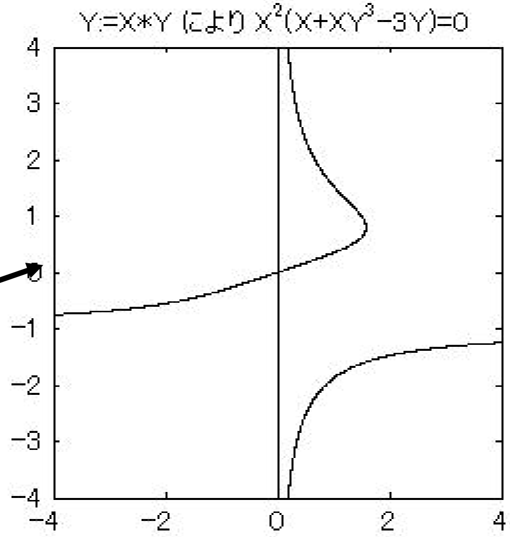
$y_1^2(x_1^3y_1+y_1-3x_1)=0$


X:=X*Y により $Y^2(X^3Y+Y-3X)=0$

$x=x_2$
$y=x_2y_2$

Then

$x_2^2\,(x_2+x_2y_2^3-3y_2)=0$


Y:=X*Y により $X^2(X+XY^3-3Y)=0$

2 Outline of Lecture

2.3 Mathematical Theory of Estimating Learning Curve

Generalization loss and error

The generalization loss is defined by

$$G_n = - \mathbf{E}_x[\log p^*(X)].$$

Then by using the definition of entropy $S=- \mathbf{E}_x[\log q(X)]$,

$$G_n = G_n^{(0)} + S.$$

In the real world, G_n is estimated instead of $G_n^{(0)}$.

Cross validation and Information Criterion

The leave-one-out cross validation is defined by

$$C_n = (1/n) \sum_{i=1}^{n} \log \mathbf{E}_w[1/p(X_i|w)].$$

The widely applicable information criterion is defined by

$$W_n = - (1/n) \sum_{i=1}^{n} \log \mathbf{E}_w[\, p(X_i|w)\,]$$
$$+(1/n) \sum_{i=1}^{n} \log \mathbf{V}_w[\, \log p(X_i|w)\,].$$

Main Theorem 2 :

Main Theorem2. For an arbitrary triple (q(x),p(x|w),φ(w)),

we can prove

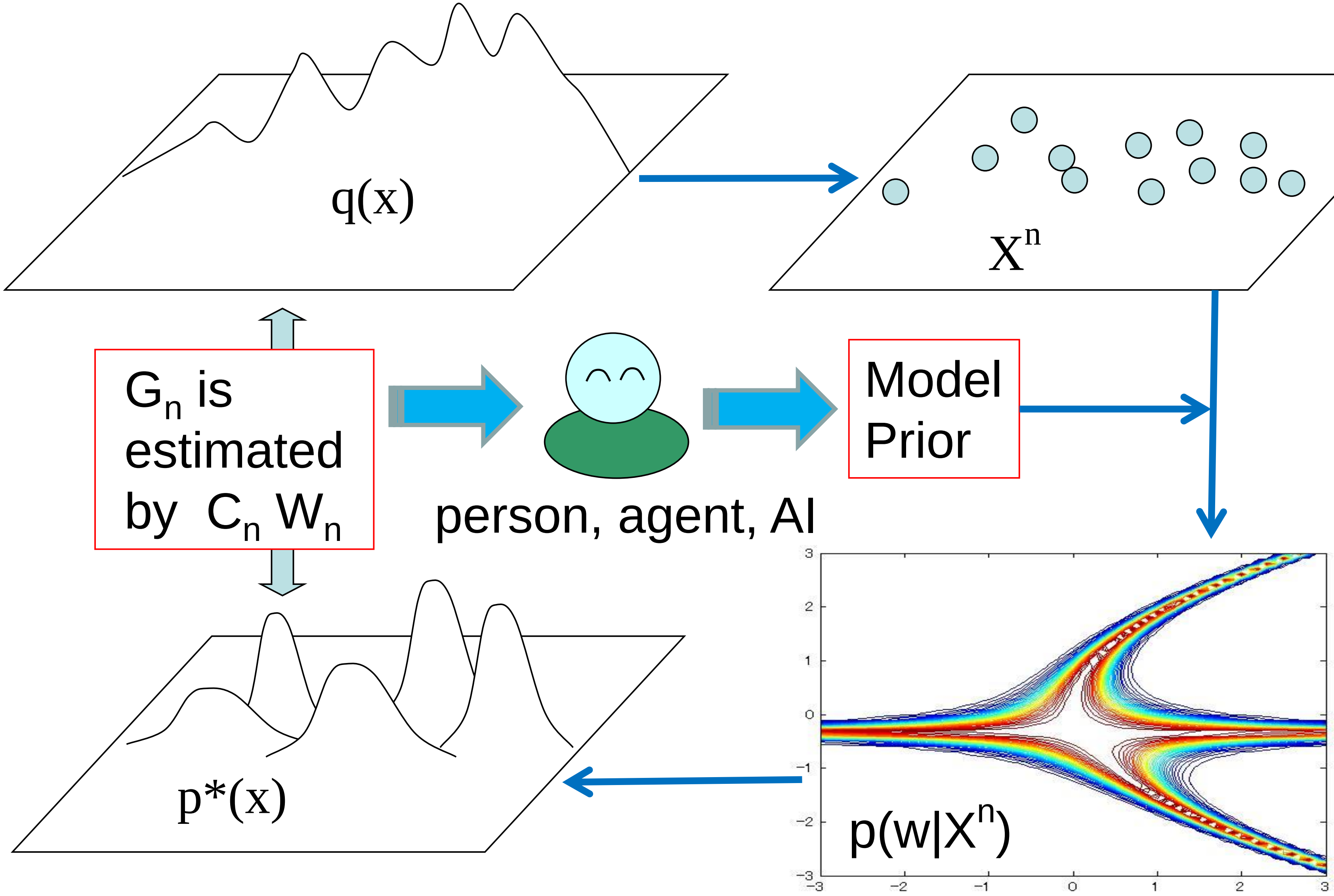
$$E[G_n] = E[C_n] + O(1/n^2),$$

$$E[G_n] = E[W_n] + O(1/n^2).$$

By using these equations, (p(x|w),φ(w)) can be optimized

so that it makes the average generalization loss smallest.

Mathematical Structure of Statistical Estimation



$q(x)$

X^n

G_n is estimated by C_n W_n

person, agent, AI

Model Prior

$p(w|X^n)$

$p^*(x)$

Applications of LOOCV and WAIC

Many applications of LOOCV and WAIC have been reported in real world problems, which can be found by searching internet. Also there are many software which implement LOOCV and WAIC.

However, this lecture is devoted to mathematical foundation of them, hence applications are not explained in the lecture.