이학석사 학위논문

A Viewpoint of Algebraic Geometry in Singular Learning Theory

(대수기하학적 관점에서의 특이학습이론)

2018년 2월

서울대학교 대학원

수 리 과 학 부

한 지 혜

A Viewpoint of Algebraic Geometry in Singular Learning Theory

(대수기하학적 관점에서의 특이학습이론)

지도교수 김 영 훈

이 논문을 이학석사 학위논문으로 제출함

2018년 2월

서울대학교 대학원

수 리 과 학 부

한 지 혜

한지혜의 이학석사 학위논문을 인준함

2017년 12월

위 원 장 현 동 훈 인

부 위 원 장 김 영 훈 인

위 원 김 다 노 인

A Viewpoint of Algebraic Geometry in Singular Learning Theory

by

HAN JI HYE

A DISSERTATION

Submitted to the faculty of the Graduate School
in partial fulfillment of the requirements
for the degree Master of Science
in the Department of Mathematics
Seoul National University
February 2018

Abstract

In this thesis, the goal of singular learning theory and its methods are described. To resolve the singularity problem, we will introduce resolution of singularities, the notion in algebraic geometry. In addition, we will go over some examples of singular learning theory (focus on strictly singular model) and show how to resolve singularities by computation. In particular, in the reduced rank regression model, the process of resolution will be described by taking blow-ups. Finally, we will introduce a determinantal variety, and examine the possibility of connection with the reduced rank regression model.

Keywords : Statistical Learning Theory, Resolution of Singularities
Student number : 2014-22355

Contents

Chapter 1

Introduction

1.1 Main Results

Recently, many people have interest on artificial intelligence. AlphaGo shocked the human community by defeating Lee Se-dol in the amazing GO match. As technology of artificial intelligence advances, knowledge of statistical learning theory has become very important. In this thesis, we will approach the statistical learning theory from viewpoint of algebraic geometry.

In Bayesian estimation, a typical method in statistical learning theory, we want to minimize the Bayes generalization error. When we compute this error, an invariant, called the "learning coefficient" λ, is very important. In regular learning theory, it is known that λ is just $d/2$, where d is dimension of parameter space ([11]). However, for strictly singular learning theory, it is not necessarily that the learning coefficient is $d/2$. Therefore, Watanabe computes the learning coefficient by showing that it is equal to the real log canonical threshold. To prove this, he uses the Hironaka's theorem -

resolution of singularities. In [11], he covers the reduced rank regression model, one kind of the neural networks, and computes λ for a case : the number of inputs and outputs are both 2, the number of hidden nodes is 2, and the rank of true distribution $r = 0$.

Therefore, in chapter 4, I will compute λ for the following three cases.

1. $M = H = N = n$ and $r = 0$;
2. $M = H = N = n$ and $r = 1$;
3. $M = m, \ \ H = 1, \ \ N = n$, and $r = 0$.

Hence, we will get the main results in this thesis as follows.

Proposition 1.1.1. *In the reduced rank regression model, the real log canonical threshold λ can be obtained as*

1. *If $M = H = N = n$ and $r = 0$, $\lambda = \frac{n(n-1)+1}{2}$;*
2. *If $M = H = N = n$ and $r = 1$, $\lambda = \frac{n^2}{2}$;*
3. *If $M = m, \ \ H = 1, \ \ N = n$ and $r = 0$, $\lambda = \frac{\min\{m,n\}}{2}$.*

Finally, in chapter 5, we will look over an object of algebraic geometry named determinantal variety and describe important propositions in [4], [7]. In [4], the authors suggest an resolution of singularities of M_k,

$$\tilde{M_k} = \tilde{M_k}(m, n) = \{(A, \Lambda) \in M \times G(n-k, n) | A\Lambda = 0\},$$

where M_k is set of matrices of rank at most k. Therefore, we will conclude this thesis by showing two computations for the real log canonical threshold ;

1. By using blow-ups used in chapter 4 ;
2. By using the resolution described in [4].

We will do this for the case of $M_2 \subset M = M_{3,3}$.

1. For $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix}$ and $B = \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{pmatrix}$. By taking blow-ups and isomorphism, $2K(w)$ can be equivalent to

$$2K_1(w) = a_{11}^2 b_{11}'^2 [1 + F[\tilde{b_{12}'}, \tilde{b_{13}'}, \tilde{a_{22}''}, \tilde{a_{32}''}]],$$

where $F[\tilde{b_{12}'}, \tilde{b_{13}'}, \tilde{a_{22}''}, \tilde{a_{32}''}]$ is a polynomial with $\tilde{b_{12}'}, \tilde{b_{13}'}, \tilde{a_{22}''}, \tilde{a_{32}''}$. Therefore, the real log canonical threshold is λ is $\frac{5}{2}$.

2. $$X = \begin{pmatrix} x_1 & x_2 & x_3 \\ x_4 & x_5 & x_6 \\ x_7 & x_8 & x_9 \end{pmatrix}, \quad det(X) = 0.$$

Consider an resolution

$$\tilde{M_2} = \{(X, \Lambda) \in M \times G(1,3) | X\Lambda = 0\}.$$

Since $G(1,3) \cong \mathbb{P}^2$, we can rewrite by

$$\tilde{M_2} = \{ \left(\begin{pmatrix} x_1 & x_2 & x_3 \\ x_4 & x_5 & x_6 \\ x_7 & x_8 & x_9 \end{pmatrix}, \begin{pmatrix} a \\ b \\ c \end{pmatrix} \right) \}$$

$with \quad ax_1 + bx_2 + cx_3 = 0, \; ax_4 + bx_5 + cx_6 = 0, ax_7 + bx_8 + cx_9 = 0.$

On $\{a = 1\}$, $2K(w)$ becomes

$$\begin{aligned} 2K(w) &= {x_5}^2[(b{x_2}' + c{x_3}')^2 + {x_2'}^2 + {x_3'}^2 \\ &+ (b + cx_6')^2 + 1 + {x_6'}^2 \\ &+ (bx_8' + cx_6'x_8')^2 + {x_8'}^2 + {x_6'}^2{x_8'}^2]. \end{aligned}$$

Thus, the real log canonical threshold λ is $\frac{5}{2}$.

1.2 Introduction

The term "machine learning" was firstly defined in the Arthur Samuel's thesis in 1959, and 8 years later, Frank Rosenblatt designed the first neural network, perceptron. Finally, in 1990s, scientists began developing computer programs to analyze a lot of data and to "learn" from results, so machine learning became reality. Recent advances of machine learning can be approached from viewpoint of statistical learning theory. The learning theory consists of regular learning theory and strictly singular learning theory. The term "regular" means the following two conditions hold;

1. The map, which maps a parameter $w \in W$ to $p(\ |w)$, is one-to-one ;
2. The Fisher information matrix is positive definite.

If a statistical learning theory is not regular, it is called "strictly singular" learning theory.

Definition 1.2.1 (Fisher information matrix, [11])**.** *For a given learning machine $p(x|w)$, where $x \in \mathbb{R}^N$ and $w \in \mathbb{R}^d$, we can define the "Fisher information matrix" by*

$$I(w) = \{I_{jk}(w)\}$$

where

$$I_{jk}(w) = \int (\frac{\partial}{\partial w_j} \log p(x|w))(\frac{\partial}{\partial w_k} \log p(x|w))p(x|w)dx \ (1 \leq j, k \leq d).$$

The Fisher information matrix is a real symmetric matrix. Therefore, if it is positive definite, we can define an inner product, hence it provides the Riemannian metric on the parameter space $\{w \in W\}$ by defining $g_{ij} =< \frac{\partial}{\partial w_i}, \frac{\partial}{\partial w_j} >$. From such an idea, Shun-ichi Amari developed the information geometry by taking probability distribution for a statistical model as a point in a Riemannian manifold in [2]. He applied techniques in differential geometry to the area of probability theory. However, for the strictly singular learning theory, it is either non-identifiable or not positive definite. For the latter case, since we cannot define an inner product, such an argument cannot be applied.

In the statistical learning theory, we need a numerical measure that tells the difference between two probability density functions.

Definition 1.2.2 (Kullback-Leibler distance, [11])**.** *Let U be an open subset in $\mathbb{R}^N$. For given two probability density functions $p(x)$, $q(x)$ on U, the "Kullback-Leibler distance" is defined by*

$$K(q||p) = \int_U q(x) \log \frac{q(x)}{p(x)} dx.$$

For the parameter set W, let $W_0 := \{w \in W | K(w) = 0\}$. If W_0 is not one point set, the distribution of the maximum likelihood estimators does not converge to the normal distribution ([11]). Hence, in strictly singular learning theory, the Bayesian estimation is more appropriate than the maximum likelihood method ([1]). The Bayes estimation minimizes the Bayes generalization error which is defined by the Kullback-Leibler distance between the true distribution $q(x)$ and the predictive distribution.

Definition 1.2.3 ([11]). *Let $D_n = \{X_1, X_2, \cdots, X_n\}$ be a set of random variables and $\varphi(w)$ be a priori probability density function. Then for given statistical model $p(x|w)$,*

1. *The "a posteriori probability density function $p(w|D_n)$ with the inverse temperature $\beta > 0$ " is defined by*

$$p(w|D_n) = \frac{1}{Z_n}\varphi(w)\prod_{i=1}^{n} p(X_i|w)^{\beta},$$

where Z_n is normalizing factor so that $p(w|D_n)$ is a probability density function of w. In fact, this Z_n is called the "evidence".

2. *The "predictive distribution" $p(x|D_n)$ is defined by*

$$p(x|D_n) = \int p(x|w)p(w|D_n)dw.$$

Then we can define the normalized evidence by $Z_n^0 := \frac{Z_n}{\prod_{i=1}^{n} q(X_i)^{\beta}}$, and the normalized stochastic complexity F_n^0 by $-\log Z_n^0$. Then the following theorem shows importance of this normalized stochastic complexity F_n^0.

Theorem 1.2.1 ([11]). *The Bayes generalization error with $\beta = 1$ is same as the increase of the normalized stochastic complexity. In order words,*

$$B_g = E_{X_{n+1}}[F_{n+1}^0] - F_n^0.$$

In [11], the asymptotic expansion of the stochastic complexity is proven as follows.

Theorem 1.2.2 (Convergence of stochastic complexity, [11]). *Let $-\lambda$ and m be the largest pole and its order of the zeta function*

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

where $K(w)$ is the Kullback-Leibler distance between $q(x)$ and $p(x|w)$. Then the stochastic complexity F_n has the asymptotic expansion

$$F_n = n\beta S_n + \lambda \log n - (m-1)\log\log n + F^R(\xi) + o_p(1),$$

where $F^R(\xi)$ is a random variable, $o_p(1)$ is a random variable whch converges to 0 in probability, and S_n will be described in chapter 2.

Due to the above theorem, the learning coefficient λ and its order m are very important since they asymptotically determine the stochastic complexity. Moreover, when a priori probability density function $\varphi(w)$ is positive at singularities, this learning coefficient is equal to the real log canonical threshold, which can be proved by Hironaka's theorem.

Definition 1.2.4 (Real log canonical threshold, [11])**.** *Let $f(x)$ be a real analytic function on open set $O \subset \mathbb{R}^d$. Let C be a compact set which is contained in O. For each $P \in C$ with $f(P) = 0$, there exists a triple (W, U, g) obtained by the Hironaka's theorem such that*

$$f(g(u) - P) = S{u_1}^{k_1}{u_2}^{k_2}\cdots{u_d}^{k_d},$$

$$g'(u) = a(u){u_1}^{h_1}{u_2}^{h_2}\cdots{u_d}^{h_d},$$

where $(k_1, \cdots, k_d)$ and $(h_1, \cdots, h_d)$ depend on P and triple (W, U, g). Then we can define the "real log canonical threshold" for a given compact set C by

$$\lambda(C) := \inf_{P\in C} \min_{1\le j\le d} \left(\frac{h_j+1}{k_j}\right),$$

where if $k_j = 0$, define $(h_j+1)/k_j = \infty$.

Therefore, we can resolve singularity problem in strictly singular learning theory by using the theorem in algebraic geometry. Furthermore, to

compute the real log canonical threshold, we should replace singularities to images of normal crossing singularities by taking recursive blow-ups. As mentioned before, the main result of this thesis is clarifying this process for specific examples. To obtain main result, we will go over basic notions and facts in the probability theory and statistical learning theory in the following chapter. In addition, the Hironaka's theorem in algebraic geometry, which is a key to resolve the problem of singular learning theory, will be described. In chapter 3, I will introduce examples of strictly singular learning theory such as

1. Bayesian networks ;
2. Hidden Markov models ;
3. Mixtures of statistical models ;
4. Layered neural networks ;
5. Boltzmann machines,

and interpret them as statistical sense. For some examples, I will describe explicit computation for resolving singularities. And then in chapter 4, we will focus on the reduced rank regression model, which is an example of neural networks. The table in the following page represents our examples which will be considered in chapter 3. When it is impossible to compute the Kullback-Leibler distance clearly, I leave it blank. The title of the last column "Pos. def." means positive definite.

		P(x| w)	K(w)	Parameter	Identifiable / Pos. def.
Bayesian Network	Gaussian	$\frac{1}{\sqrt{2\pi}} \exp(\frac{-(y-\sum_{i=1}^{k} a_i x_i)^2}{2})$	$\frac{1}{2} \sum_{i=1}^{k} (a_i - b_i)^2$	$\{(a_1, \cdots, a_k)\}$	O / O
	Discrete Exp.	$\frac{1}{Z} \exp(ax_1 y + bx_2 y + cx_3 y)$	$\frac{1}{2}(a^2b^2 + b^2c^2 + c^2a^2 + \cdots)$	{(a,b,c)}	O / X
Hidden Markov		$\frac{\prod_{i=1}^{t} p_{i-1,i} \prod_{i=1}^{t} \int_{[a_i, b_i]} \frac{1}{\sqrt{2\pi}} \exp - \frac{1}{2} x^2 dx}{P(o_1, \cdots, o_n)}$		$\{(p_{ij}, a_i, b_i)\}$	? / X
Mixture Model	Gaussian 1	$P(x\|\mu_1, \mu_2) = \frac{1}{2\sqrt{2\pi}} \left[\exp - \frac{(x-\mu_1)^2}{2} + \exp - \frac{(x-\mu_2)^2}{2}\right]$		$\{(\mu_1, \mu_2)\}$	X / ?
	Gaussian 2	$P(x\|c, \mu) = \frac{1}{\sqrt{2\pi}} [(1-c) \exp - \frac{x^2}{2} + c \exp - \frac{(x-\mu)^2}{2}]$		$\{(c, \mu)\}$	? / X
	Poisson 1	$P(x\|\lambda_1, \lambda_2) = \frac{1}{2} \left[\frac{\exp(-\lambda_1)\lambda_1^x}{x!} + \frac{\exp(-\lambda_2)\lambda_2^x}{x!}\right]$		$\{(\lambda_1, \lambda_2)\}$	X / ?
	Poisson 2	$P(x\|c, \lambda) = c \frac{\exp(-1)}{x!} + (1-c) \frac{\exp(-\lambda)\lambda^x}{x!}$		$\{(c, \lambda)\}$	? / X
Neural Network	2-layered	$\frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(y - as(bx) - cx)^2)$	$\frac{1}{2}(ab + c)^2 + \frac{3}{2}a^2b^4$	$\{(a, b, c)\}$	? / X
	3-layered	$p(x, y\|a, b, c, d) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(y - a\sigma(bx) - c\sigma(dx))^2$	$\frac{1}{2} \int (a\sigma(bx) + c\sigma(dx))^2 q(x) dx$	$\{(a, b, c, d)\}$	X / X
	Gaussian RBM	$\frac{1}{Z} \exp(-E(v, h))$ (E : associated energy)		$\{w_{ij}\}$	? / X

Definition 1.2.5 (The reduced rank regression, [11])**.** *Let M be the number of inputs and N be that of outputs. Assume there exist H hidden variables, and r means the rank of the true distribution. The reduced rank regression is described by a conditional probability density function as follows;*

$$p(y|x,w) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2}|y - xAB|^2\right),$$

where $x \in \mathbb{R}^M$, $y \in \mathbb{R}^N$, A is a $M \times H$ matrix, B is a $H \times N$ matrix, and σ is a constant. Hence, the parameter set $W = \{(A,B)|A \in \mathfrak{M}_{M,H},\ B \in \mathfrak{M}_{H,N}\}$.

The following table represents the main result aforementioned in the previous section.

	M=H=N=n, r=0	M=H=N=n, r=1	M=m, N=n, H=1, r=0
The real log canonical threshold	$\frac{n(n-1)+1}{2}$	$\frac{n^2}{2}$	$\frac{\min\{m,n\}}{2}$

Note that the last case considers a matrix which can be represented by product of two vectors and this is equivalent to a matrix is of rank at most 1. Thus, in chapter 5, we will look over important propositions related to the determinantal variety, described in [4], [7].

Proposition 1.2.1 ([7])**.** *M_k is an irreducible algebraic subvariety of M of codimension $(m-k)(n-k)$.*

Proposition 1.2.2 ([7])**.** *The singular locus of M_k is equal to M_{k-1}.*

In [4], in the proof of the first proposition, the authors suggest an resolution of singularities of M_k,

$$\tilde{M_k} = \tilde{M_k}(m,n) = \{(A,\Lambda) \in M \times G(n-k,n)|A\Lambda = 0\}.$$

Therefore, we will compute the real log canonical threshold by two ways described in the previous section, and obtain the same result. It shows us possibilities of connection with the reduced rank regression.

Chapter 2

Preliminaries

In this chapter, we will present elementary notions in probability theory and introduce a markov model. This refers to [12]. And then, we will discuss the goal of statistical learning theory, one of main objects in this thesis. This refers to [11]. Finally, we will conclude this chapter by covering singularity theory in algebraic geometry and describing the Hironaka's theorem, which is our key to resolve singularity problem.

2.1 Elementary Probability Theory

Only a man coming from the future may be exactly able to know what event will happen in the future. It seems impossible that a cat lying in front of a monitor types the letter "Meow". However, if someone, with pulling a knife on you, asks you whether the probability of that event must be zero, you wouldn't answer "yes" with confidence. From this point of view, we can say that full of uncertainty exists in the world where we live. Although every moment has uncertainty, people desire to find some regularity and make the

best or the most efficient decision with systematic tool. For example, we naturally have the probability in mind even when making a slight bet with friends. Therefore, the "Probability Theory" is the theory that helps us understand uncertain moments with numerical terms and makes a rational decision with reasonable grounds. In order to achieve this goal, it is required to introduce some terminologies.

Definition 2.1.1 (Probability Space, [11])**.** *Let Ω be a metric space and a set $\mathfrak{B}$ be a sigma algebra which consists of subsets in Ω. A function $P : \mathfrak{B} \to [0, 1]$ is called a "probability measure" if it satisfies*

1. $P(\Omega) = 1$
2. *For disjoint subsets $\{A_k\}$, $P(\cup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} P(A_k)$.*

Then we call a triple of a metric space, a sigma algebra, and a probability measure $(\Omega, \mathfrak{B}, P)$ a "probability space".

Definition 2.1.2 (Random Variable, [11])**.** *Let $(\Omega, \mathfrak{B}, P)$ be a probability space. For a measurable space $(\Omega_1, \mathfrak{B}_1)$, a function*

$$X : \Omega \ni \omega \mapsto X(\omega) \in \Omega_1$$

is called "random variable" if it is measurable.

Remark 2.1.1. *If we define*

$$\mu(A) := P(X^{-1}(A)),$$

μ becomes a probability measure on $(\Omega_1, \mathfrak{B}_1)$. Then $(\Omega_1, \mathfrak{B}_1, \mu)$ is a probability space. Here, the probability measure μ is called a "probability distribution of the random variable X". For convenience, we will denote μ by P_X.

Definition 2.1.3 (Expectation, [12]). *Let X be a random variable from the probability space $(\Omega, \mathfrak{B}, P)$ to $(\Omega_1, \mathfrak{B}_1)$ with the probability distribution P_X. Then we can define the "Expectation of X" as follows;*

$$E[X] = \int X(\omega)P(d\omega) = \int xP_X(dx)$$

if the integration is finite in Ω_1.

There are well-known probability distributions, such as Bernoulli distribution, exponential distribution and so on, used to explain specific circumstance. Particularly, the "Poisson distribution" can be used in microbiology, since the Poisson process is an example of the counting process. When we conduct a experiment about conjugation between Hfr strain and F- strain, we need to count the Colony forming unit (CFU). Since we already know the value of CFU over time is subject to the Poisson distribution, we can calculate the CFU more effectively.

In case Ω is not finite space, many differences can be found compared to the finite case. However, if the sequence of random variables satisfies the good condition, we can deduce a lot of meaningful results such as the law of large numbers and Central limit theorem.

Definition 2.1.4 (i.i.d. sequence, [12]). *A sequence of random variables $\{X_i\}$ is called a "independent and identically distributed (shortly, i.i.d.)" if each X_i has the same probability distribution as the others and all are mutually independent.*

Unfortunately, when trying to make a model from real circumstances, there are few cases which can be explained by i.i.d. sequence. Instead, let's consider more weaker condition preferred to apply to many cases.

Definition 2.1.5 (Markov process, [12])**.** *Let $\{X_n|n \geq 0\}$ be a sequence of random variables. $\{X_n|n \geq 0\}$ is called the "Markov process" if it satisfies the following condition;*

$$P(X_{n+1} \in A|X_0 = x_0, X_1 = x_1, \cdots, X_n = x_n) = P(X_{n+1} \in A|X_n = x_n)$$

for arbitrary $x_0, \cdots, x_n$ and a set A.

In that case, we say $\{X_n\}$ has the Markov property and if the random variables are discrete, it is called "Markov chain".

Remark 2.1.2. *In Markov chain, the range of random variables is called "state space" and its element becomes "state". If the state space is the set of integers, the process is completely described by p_{ij}, where $p_{ij} := P(X_{n+1} = j|X_n = i)$. Thus, we call a matrix $P = (p_{ij})$ with entries p_{ij} "transition matrix".*

If $p_{ii} = 1$, i is called "absorbing state".

Example 2.1.1. *Obviously, if $\{X_n\}$ are mutually independent, it satisfies the Markov property.*

Now, let's introduce a binary relation on the state space.

Definition 2.1.6 ([12])**.** *If there exists $n \geq 0$ such that $p_{ij}^{(n)} > 0$, we say "i is accessible from j" and denote $i \to j$. If $i \to j$ and $j \to i$, denote $i \leftrightarrow j$ and we can check it is an equivalence relation. In other words, we can divide the state space under this relation.*

Definition 2.1.7 ([12])**.** *Let $\{X_n\}$ be a Markov chain and N_i be a number of visiting i state.*

1. *If $P(N_i < \infty|X_0 = i) = 1$, i is called "transient" state.*

2. *If* $P(N_i = \infty | X_0 = i) = 1$, *$i$ is called "recurrent" state.*

Definition 2.1.8 (First return time, [12])**.** *If we let* $R_i := \min\{n \geq 1 | X_n = i\}$, *$R_i$ means the first time to return i from i. Then we can define two kinds of recurrent states.*

1. *If* $E(R_i | X_0 = i) < \infty$, *$i$ is called "positively recurrent state" ;*

2. *If* $E(R_i | X_0 = i) = \infty$, *$i$ is called "null-recurrent state".*

We have already known weather prediction is one of the most used examples of Markov chain. However, there can be some interesting applications in medical area.

Example 2.1.2. *Let's suppose there is a patient who thinks carefully which method is the best for his health among drug treatments and having surgery. If he takes drug treatments, several results are possible: no reaction (remain original state), death because of fatal side effects, and becoming healthy. On the other hand, if he takes a surgery, there are possible results: becoming healthy, death because of hemorrhage. Thus, in this example, there exist 3 states: with disease, healthy, dead. Since it is impossible that the deceased can be neither one with disease nor healthy, we can say "dead" is absorbing state. In this case, Markov chain is useful to explain the principle of this situation. Thus, we can describe this situation as in Figure 2.1.*

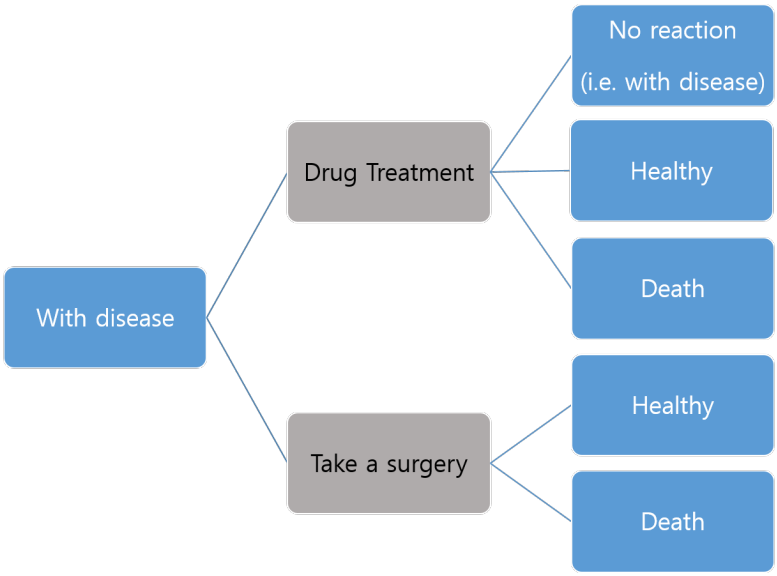

Figure 2.1: Decision of efficient treatment

In the above example, the patient and doctors should decide what process would be the best choice for patient's recovery. Since it is possible that a person who got better after treatment (drug or surgery) can come with the disease again, "healthy $\rightarrow$ with disease" holds with the use of the aforementioned symbol. Therefore, if we consider all possible cases, there will be a infinite tree. At this point, it is natural to inquire about what will be the limit of state distribution.

Definition 2.1.9 (Stability condition, [12])**.** *We say "a given Markov chain $\{X_n | n \geq 0\}$ has the stability condition" if the following holds;*

For arbitrary i, j, there exists π_j which is independent to i and satisfies $\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j$ and $\sum_j \pi_j = 1$.

Remark 2.1.3. *The above stability condition can be restated by*

$\pi = \pi P,$

$\sum_j \pi_j = 1$, $\pi_j \geq 0.$

For convenience, let's put the above two equations "(*) condition".

The following theorem shows us when a given Markov chain converges to the stationary Markov chain.

Theorem 2.1.1 ([12])**.** *For a given irreducible(under relation $\leftrightarrow$) and non-periodic Markov chain, the following holds;*

1. *If all the states are positively recurrent, the chain has the stability condition. In other words, for i, j, there exists $\pi_j = \lim_{n\to\infty} p_{ij}^{(n)}$ which satisfies (*) condition. And this is the unique solution and π_j is equal to the expectation of the first return time of j.*

2. *Conversely, if there exists a solution of equations in (*), for arbitrary i, j,*
$$\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j$$
holds and all states are positively recurrent.

3. *If all states are transient or null-recurrent, for arbitrary i, j,*
$$\lim_{n\to\infty} p_{ij}^{(n)} = 0.$$

4. *If there is no solution of equations in (*), for arbitrary i, j,*
$$\lim_{n\to\infty} p_{ij}^{(n)} = 0.$$

Remark 2.1.4. *Therefore, in the example 2.1.2, we can calculate the rate of healthy state for each treatment respectively. By comparing these results, it would be possible to decide what treatment is the best choice.*

2.2 Statistical Learning Theory

In our world, understanding the given data is essential in various fields. For example, in the medical area, we can consider the situation that researchers should determine whether a particular drug causes severe adverse reaction when it is taken. In this situation, we can let $\{X_n\}$ be characteristics of a patient's blood sample and Y be the patient's risk of side effect. If we can predict Y from $\{X_n\}$, it is possible to avoid prescribing the medicine to patients whose estimate of Y is high. As in the aforementioned, modeling and prediction are ultimate goals of statistical learning theory. Now, let's study some basic concepts in statistical learning theory, which refers to [11].

Definition 2.2.1 (Random samples, [11])**.** *For $(\Omega, \mathfrak{B}, P)$ a probability space, let $X : \Omega \to \mathbb{R}^N$ be a random variable which is subject to a probability distribution $q(x)dx$. Here, we will call $q(x)$ a "true probability density function", and random variables $\{X_n\}$ subject to the same probability distribution of X are called "random samples".*

In this paper, we will denote D_n by the set of random samples $\{X_n\}$.

Definition 2.2.2 (Learning machine, [11])**.** *Let W be the set of parameters. For a given $w \in W$, a conditional probability density function $p(x|w) = p_w(x)$ for $x \in \mathbb{R}^N$ is called a "learning machine" or "statistical model". Hence, $w \mapsto p_w$ becomes a map from the parameter to the possibility density function.*

In statistical learning theory, one of goals is the development of a method to find a probability density function $p^*(x)$ from D_n by using $p(x|w)$. Here, p^* is called the estimated probability density function. To determine whether an approximation is good, we need a numerical measure.

Definition 2.2.3 (Kullback-Leibler distance, [11])**.** *Let U be an open subset in $\mathbb{R}^N$. For given two probability density functions $p(x)$, $q(x)$ on U, the "Kullback-Leibler distance" is defined by*

$$K(q||p) = \int_U q(x) \log \frac{q(x)}{p(x)} dx.$$

Now, we can measure the difference between the true density function $q(x)$ and the estimated function $p^*(x)$.

Definition 2.2.4 (Generalization error and Training error, [11])**.** *From the above definition, $K(q||p^*)$ is called the generalization error of statistical estimation $D_n \mapsto p^*$.*

The "training error" is defined by

$$K_n(q||p^*) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{q(X_i)}{p^*(X_i)},$$

In statistical learning, it is important to find a mathematical relation between these two errors. If the generalization error can be estimated from the training error, we can choose the best model among many possible models.

Definition 2.2.5 (Likelihood function and Log likelihood ratio function, [11])**.** *Let $q(x)$ be a true distribution and $p(x|w)$ be a parametric model. Then we can define the following;*

1. *The "likelihood function" $L_n(w)$ is defined by*

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

2. *The "log density ratio function" $f(x, w)$ is defined by*

$$f(x, w) = \log \frac{q(x)}{p(x|w)}.$$

3. *The "Kullback-Leibler distance" $K(w)$ is defined by*

$$K(w) = \int q(x) f(x, w) dx.$$

4. *The "log likelihood ratio function" $K_n(w)$ is defined by*

$$K_n(w) = \frac{1}{n} \sum_{i=1}^{n} f(X_i, w).$$

Remark 2.2.1. *If we consider the empirical entropy $S_n = -\frac{1}{n}\sum_{i=1}^{n} \log q(X_i)$, we can deduce the following equality;*

$$-\frac{1}{n} \log L_n(w) = K_n(w) + S_n.$$

Since S_n is independent to w, maximization of $L_n(w)$ is equivalent to minimization of $K_n(w)$.

Remark 2.2.2. *When $E[K(w)]$ is finite, by the law of big numbers,*

$$K_n(w) \mapsto K(w)$$

converges in probability for $w \in W$.

If $E[K(w)^2]$ is also finite, by the central limit law, $\sqrt{n}(K_n(w) - K(w))$ converges in distribution to the normal distribution for $w \in W$. Thus, if we let $S := -\int q(x) \log q(x) dx$, we can deduce that

$$-\frac{1}{n} \log L_n(w) \mapsto K(w) - S$$

holds for $w \in W$.

Anyone who reads the above remark would hope that minimization of $K_n(w)$ is equivalent to that of $K(w)$ because once it holds, then maximization of $L_n(w)$ would be the best method in statistical estimation. Unfortunately, since minimization and expectation are not commutative, we cannot

obtain such a result. To analyze the difference between $K(w)$ and $K_n(w)$, we need the notion of convergence in a functional space.

Definition 2.2.6 (Fisher information matrix, [11])**.** *For a given learning machine $p(x|w)$, where $x \in \mathbb{R}^N$ and $w \in \mathbb{R}^d$, we can define the "Fisher information matrix" by*

$$I(w) = \{I_{jk}(w)\}$$

where

$$I_{jk}(w) = \int (\tfrac{\partial}{\partial w_j} \log p(x|w))(\tfrac{\partial}{\partial w_k} \log p(x|w))p(x|w)dx \ (1 \leq j, k \leq d).$$

Remark 2.2.3. *By the definition, the Fisher information matrix is always symmetric. Besides, if we let $X = (\partial_1 \log p(x|w), \cdots, \partial_d \log p(x|w))$, I becomes $E[XX^t]$. Thus, for non-zero column vector u,*

$$u^t I u = u^t E[XX^t]u = E[u^t X X^t u] = E[||X^t u||^2] \geq 0,$$

hence the matrix is always positive semi-definite.

Remark 2.2.4. *Since $\int p(x|w)dx = 1$, we can deduce that*

$$I_{jk}(w) = -\int (\frac{\partial^2}{\partial w_j \partial w_k} \log p(x|w))p(x|w)dx.$$

Therefore, for the true parameter w_0 which satisfies $q(x) = p(x|w_0)$, then

$$I_{jk}(w_0) = \frac{\partial^2}{\partial w_j \partial w_k} K(w_0)$$

i.e, the Fisher information matrix is same to the Hessian matrix of the Kullback-Leibler distance at the true parameter.

Definition 2.2.7 (Identifiability, [11])**.** *A learning machine is called "identifiable" if the map*

$$W \ni w \mapsto p(\ |w)$$

is one-to-one. A model is called "nonidentifiable" if it is not identifiable.

Definition 2.2.8 (Positive definite metric, [11])**.** *A learning machine $p(x|w)$ is said to "have a positive definite metric" if its Fisher information matrix $I(\omega)$ is positive definite for each $w \in W$.*

Now, we are ready to define singular statistical model.

Definition 2.2.9 (Singular statistical models, [11])**.** *Suppose that the support of $p(x|w)$ is independent to w. A learning machine $p(x|w)$ is said to be "regular" if it is identifiable and has a positive definite metric. If it is not regular, then it is called "strictly singular". The set of singular statistical models consists of both regular and strictly singular models.*

Remark 2.2.5. *In this remark, we will discuss the reason why the "positive definite" condition for regular learning machine is important. The Fisher information matrix is real symmetric matrix, so when it is positive definite, we can define an inner product. If we define $g_{ij} =< \frac{\partial}{\partial w_i}, \frac{\partial}{\partial w_j} >$, it gives a Riemannian metric $g =< \ , \ >$ on $\{w \in W\}$. Therefore, for regular learning machine, we can consider the parameter space as a Riemannian manifold.*

From such an idea, Shun-ichi Amari developed the information geometry by taking probability distributions for a statistical model as a point in Riemannian manifold in [2]. He applied the techniques in differential geometry to the area of probability theory. However, for the strictly singular learning theory, such an argument cannot be applied.

Remark 2.2.6. *If we introduce an equivalence relation $\sim$ on the set of parameters W;*

$$w_1 \sim w_2 \Longleftrightarrow p(x|w_1) = p(x|w_2) \text{ for all } x.$$

Then the map $(W/\sim) \ni w \to p_w$ *becomes one-to-one. However, the quotient set* $(W/\sim)$ *is neither the Euclidean space nor a manifold, so it is difficult to construct statistical learning theory on* $(W/\sim)$.

For a strictly singular learning machine, the Bayesian estimation is more appropriate than the maximum likelihood method.([1])

Definition 2.2.10 ([11])**.** *Let* $D_n = \{X_1, X_2, \cdots, X_n\}$ *be a set of random variables and* $\varphi(w)$ *be a priori probability density function. Then for given statistical model* $p(x|w)$,

1. *The "a posteriori probability density function* $p(w|D_n)$ *with the inverse temperature* $\beta > 0$*" is defined by*

$$p(w|D_n) = \frac{1}{Z_n}\varphi(w)\prod_{i=1}^{n} p(X_i|w)^{\beta},$$

   where Z_n *is normalizing factor such that* $p(w|D_n)$ *is a probability density function of* w*. In fact, this* Z_n *is called the "evidence".*
   In particular, for $\beta = 1$, $p(w|D_n)$ *is called a "strict Bayes a posteriori density".*

2. *The "predictive distribution"* $p(x|D_n)$ *is defined by*

$$p(x|D_n) = \int p(x|w)p(w|D_n)dw.$$

Now, we are ready to define the Bayes estimation.

Definition 2.2.11 ([11])**.** *For the same as above notations,*

1. *The "Bayes estimation" is the map*

$$D_n \mapsto p^*(x) := p(x|D_n).$$

2. *The "Bayes generalization error" B_g is defined by the Kullback-Leibler distance between $q(x)$ and $p^*(x)$, i.e.,*

$$B_g = \int q(x) \log \frac{q(x)}{p(x|D_n)} dx.$$

3. *The "Bayes training error" B_t is defined by*

$$B_t = \frac{1}{n} \sum_{i=1}^{n} \log \frac{q(X_i)}{p(X_i|D_n)}.$$

4. *The "stochastic complexity" F_n is defined by*

$$F_n = -\log Z_n.$$

Remark 2.2.7. *The "normalized evidence" is defined by*

$$Z_n^0 := \frac{Z_n}{\prod_{i=1}^{n} q(X_i)^\beta}.$$

Likewise, the "normalized stochastic complexity" is defined by

$$F_n^0 := -\log Z_n^0.$$

Then we can find an important relation between this F_n^0 and the Bayes generalization B_g when $\beta = 1$.

Theorem 2.2.1 ([11])**.** *The Bayes generalization error with $\beta = 1$ is same as the increase of the normalized stochastic complexity. In order words,*

$$B_g = E_{X_{n+1}}[F_{n+1}^0] - F_n^0.$$

As we can see in the above theorem, the stochastic complexity is crucial thing in Bayes learning theory. In [11], the asymptotic expansion of the stochastic complexity is proven as follows.

Theorem 2.2.2 (Convergence of stochastic complexity, [11])**.** *Let $-\lambda$ and m be the largest pole and its order of the zeta function*

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

where $K(w)$ is the Kullback-Leibler distance between $q(x)$ and $p(x|w)$. Then the stochastic complexity F_n has the asymptotic expansion

$$F_n = n\beta S_n + \lambda \log n - (m-1) \log \log n + F^R(\xi) + o_p(1),$$

where $F^R(\xi)$ is a random variable, $o_p(1)$ is a random variable which converges to 0 in probability, and S_n is defined as in Remark 2.2.1.

Remark 2.2.8. *In the following chapter, the constant λ is an birational invariant and it is real log canonical threshold if $\varphi(w)$ is positive at singularities. Thus, the above theorem tells us that the stochastic complexity F_n is asympototically determined by the algebraic geometrical birational invariant ([11]). For this reason, in Chapter 3 and 4, we will describe concrete computation of this λ for specific cases, which are main results of this thesis.*

2.3 Singularity Theory

For a strictly singular statistical model, the set of true parameters W_0 is not one point but a real analytic set. Since W_0 contains complicated singularities, it is difficult to predict its behavior in a neighborhood of W_0. In order to solve this problem, we should introduce a powerful theorem

in algebraic geometry, called Hironaka's theorem. Before introducing this theorem, let's study basic singularity theory, which refers to [11].

Definition 2.3.1 (Singularities of a set, [11])**.** *Let A be a nonempty set in $\mathbb{R}^d$.*

1. *A point $P \in A$ is called "nonsingular" if there exists open sets U, $V \subset \mathbb{R}^d$ and an analytic isomorphism $f : U \to V$ such that*

$$f(A \cap U) = \{(x_1, \cdots, x_r, 0, \cdots, 0) | x_i \in \mathbb{R}\} \cap V,$$

   where r is a nonnegative integer.

2. *If a point P is not nonsingular, it is called a "singular point of A". The set of all singularities of A is called the "singular locus of A" and is denoted by $Sing(A)$.*

Then we can find the sufficient condition for a nonsingular point as follows.

Theorem 2.3.1 ([11])**.** *Let U be an open set in $\mathbb{R}^d$, and $f_1(x), \cdots, f_k(x)$ be a set of analytic functions $(1 \leq k \leq d)$. Consider a real analytic set defined by*

$$A = \{x \in U | f_1(x) = f_2(x) = \cdots = f_k(x) = 0\}.$$

If a point $P \in A$ satisfies $detJ(P) \neq 0$, then P is a nonsingular point of A, where $J(P)$ is Jacobian matrix at P.

As we mentioned before, it is difficult to analyze a function in a neighborhood of singular point. However, any neighborhood of a real analytic set can be considered as an image of normal crossing singularities as follows.

Theorem 2.3.2 (Resolution of singularities, [11])**.** *Let $f(x)$ be a nonconstant real analytic function from a neighborhood of the origin in $\mathbb{R}^d$ to $\mathbb{R}$ with $f(0) = 0$. Then there exists a triple (W, U, g) where W is open in $\mathbb{R}^d$ which contains 0, U is a d-dimensional real analytic manifold, and $g : U \to W$ is a real analytic map, which satisfies the following;*

1. *The map g is proper.*
2. *Let $W_0 := \{x \in W | f(x) = 0\}$ and $U_0 := \{u \in U | f(g(u)) = 0\}$, then g is a real analytic isomorphism of $U \backslash U_0$ and $W \backslash W_0$.*
3. *For all $P \in U_0$, there exists a local coordinate $u = (u_1, \cdots, u_d)$ of U in which P is the origin and*

$$f(g(u)) = S{u_1}^{k_1}{u_2}^{k_2} \cdots {u_d}^{k_d},$$

where $S = 1$ or -1, k_i's are nonnegative integers, and the Jacobian determinant of $x = g(u)$ is

$$g'(u) = a(u){u_1}^{h_1}{u_2}^{h_2} \cdots {u_d}^{h_d},$$

where $a(u)$ is a real analytic function, and h_i's are nonnegative integers.

Remark 2.3.1. *The above theorem states that any analytic function can be represented by a normal crossing function by choosing an appropriate manifold.*

Definition 2.3.2 (Real log canonical threshold, [11])**.** *Let $f(x)$ be a real analytic function on open set $O \subset \mathbb{R}^d$. Let C be a compact set which is contained in O. For each $P \in C$ with $f(P) = 0$, there exists a triple (W, U, g) described in the above theorem such that*

$$f(g(u) - P) = S{u_1}^{k_1}{u_2}^{k_2} \cdots {u_d}^{k_d},$$

$$g'(u) = a(u){u_1}^{h_1}{u_2}^{h_2} \cdots {u_d}^{h_d},$$

where $(k_1, \cdots, k_d)$ *and* $(h_1, \cdots, h_d)$ *depend on* P *and triple* (W, U, g)*. Then we can define the "real log canonical threshold" for a given compact set* C *by*

$$\lambda(C) := \inf_{P \in C} \min_{1 \leq j \leq d} \left(\frac{h_j + 1}{k_j}\right),$$

where if $k_j = 0$*, define* $(h_j + 1)/k_j = \infty$*.*

Remark 2.3.2. *We can prove that the real log canonical threshold does not depend on the triple* (W, U, g)*. Such a number that does not depend on the triple is called a birational invariant. Thus, the real log canonical threshold is one example of birational invariants.*

Definition 2.3.3 (Normal crossing, [11])**.** *Let* $U \subset \mathbb{R}^d$ *be an open set. A real analytic function* $f(x)$ *on* U *is called "normal crossing at* $x^* = (x_1^*, \cdots, x_d^*)$*", if there exists an open* $U' \subset U$ *such that*

$$f(x) = a(x) \prod_{j=1}^{d} (x_j - x_j^*)^{k_j}$$

for $x \in U'$*, where* $a(x)$ *is a real analytic function and* k_i*'s are nonnegative integers.*

From now on, we will present important tools in algebraic geometry, which are used to solve singularity problems. In particular, we will introduce the specific method to find a resolution map described before. In order to do this, we should study blow-up.

At first, we should introduce the criterion for nonsingularity.

Definition 2.3.4 (Dimension of a real algebraic set, [11])**.** *Let V be a nonempty real algebraic set in $\mathbb{R}^d$. Let $I(V) =< f_1, \cdots, f_r >$.*
The "dimension of the real algebraic set V" is defined by the maximum value of the rank of the Jacobian matrix, i.e.,

$$d_0 = \max_{x \in V} rank J(x).$$

Theorem 2.3.3 ([11])**.** *Let V be a nonempty real algebraic set in $\mathbb{R}^d$ whose dimension is d_0.*
Then $x \in V$ is a nonsingular point of V if and only if $rank J(x) = d_0$.
In other words, $x \in V$ is a singularity if and only if $rank J(x) < d_0$.

There are a lot of learning machines which contain singularities in parameter space. In order to obtain statistical learning theory, we should resolve the complexity of singularities. The notion of blow-up is the key to resolve this problem. Any singularities can be desingularized by finitely many recursive blow-ups.

Definition 2.3.5 (Blow-up of a real algebraic set, [11])**.** *Let r be an integer satisfying $2 \leq r \leq d$. Let V be defined by as follows;*

$$V := \{x \in \mathbb{R}^d | x_1 = x_2 = \cdots = x_r = 0\}.$$

Let W be a real algebraic set containing V. Then, the "blow-up of W with center V" is defined by

$$B_V(W) := \overline{\{(x, (x_1 : \cdots : x_r)) | x \in W \backslash V\}}.$$

Definition 2.3.6 (Strict and total transform, exceptional set, [11])**.** *Let $\pi : B_V(W) \to W$ be a map defined by $\pi((x, y)) = x$. Then the following*

relation holds;

$$B_V(W) \subset \pi^{-1}(W).$$

The set $B_V(W)$ is called a "strict transform of W" while $\pi^{-1}(W)$ is a "total transform".

The "exceptional set" is defined by the closure of $\pi^{-1}(W) \backslash B_V(W)$.

Definition 2.3.7 (General blow-up in Euclidean space, [11])**.** *Let r be an integer satisfying $2 \leq r \leq d$. V and W are real algebraic sets in $\mathbb{R}^d$ such that $V \subset W$. Let $I(V) =< f_1, \cdots, f_r >$. Then the "blow-up of W with center V", $B_V(W)$, is defined by*

$$B_V(W) := \overline{\{(x, (f_1 : \cdots : f_r)) | x \in W \backslash V\}}.$$

The following two theorems describe the process of desingularization.

Theorem 2.3.4 (Hironaka's theorem 1)**.** *For real algebraic set V, there exists a sequence of real algebraic varieties $V_0(= V), V_1, \cdots, V_n$ which satisfies the following conditions.*

1. *V_n is nonsingular*
2. *For $i = 1, 2, \cdots, n$, $V_i = B_{C_{i-1}}(V_{i-1})$, where V_i is a blow-up of V_{i-1} with center C_{i-1}.*
3. *C_i is a nonsingular real algebraic variety which is contained in $Sing(V_i)$.*

Theorem 2.3.5 (Hironaka's theorem 2)**.** *Let $f(x) \in \mathbb{R}[x_1, x_2, \cdots, x_d]$. Then there exists a sequence of pairs of real algebraic varieties $(V_0, W_0), \cdots, (V_n, W_n)$ which satisfies the following;*

1. *$V_i \subset W_i$ $(1 \leq i \leq n)$.*

2. *$V_0 = V(f)$, $W_0 = \mathbb{R}^d$.*

3. *$\{W_i | i = 0, \cdots, n\}$ are nonsingular algebraic varieties.*

4. *V_n is defined by a normal crossing polynomial on each local coordinate of W_n.*

5. *For $i = 1, 2, \cdots, n$, $W_i = B_{C_{i-1}}(W_{i-1})$, where W_i is a blow-up of W_{i-1} with center C_i.*

6. *Let $\pi_i : W_i \to W_{i-1}$ be a projection map defined in $B_{C_{i-1}}(W_{i-1})$. Then V_i is the total transform of π_i.*

7. *The center C_i of each blow-up is a nonsingular real algebraic variety which is contained in the set of critical points of $f \circ \pi_1 \circ \pi_2 \circ \cdots \circ \pi_i$.*

Remark 2.3.3. *The theorem 1 states that any real algebraic set can be understood as an image of a nonsingular real algebraic variety. Here, exceptional sets are not contained in the blow-ups, while the theorem 2 are. Therefore, in the theorem 2, any singularities of a real algebraic set are images of normal crossing singularities since exceptional sets are contained.*

Chapter 3

Examples of Singular Learning Theory

In this chapter, we will look over some examples of singular learning theory. If possible, we will compute the Kullback-Leibler distance and obtain the real canonical threshold. Before we consider each of the cases, the following table represents all the examples which will be described in this chapter. When it is impossible to compute the Kullback-Leibler distance clearly, I leave it blank. The title of the last column "Pos. def." means positive definite.

		P(x| w)	K(w)	Parameter	Identifiable / Pos. def.
Bayesian Network	Gaussian	$\frac{1}{\sqrt{2\pi}} \exp(\frac{-(y-\sum_{i=1}^{k} a_i x_i)^2}{2})$	$\frac{1}{2} \sum_{i=1}^{k} (a_i - b_i)^2$	$\{(a_1, \cdots, a_k)\}$	O / O
	Discrete Exp.	$\frac{1}{Z} \exp(ax_1 y + bx_2 y + cx_3 y)$	$\frac{1}{2}(a^2b^2 + b^2c^2 + c^2a^2 + \cdots)$	{(a,b,c)}	O / X
Hidden Markov		$\frac{\prod_{i=1}^{t} p_{i-1,i} \prod_{i=1}^{t} \int_{[a_i, b_i]} \frac{1}{\sqrt{2\pi}} \exp - \frac{1}{2} x^2 dx}{P(o_1, \cdots, o_n)}$		$\{(p_{ij}, a_i, b_i)\}$	? / X
Mixture Model	Gaussian 1	$P(x\|\mu_1, \mu_2) = \frac{1}{2\sqrt{2\pi}} \left[\exp - \frac{(x-\mu_1)^2}{2} + \exp - \frac{(x-\mu_2)^2}{2}\right]$		$\{(\mu_1, \mu_2)\}$	X / ?
	Gaussian 2	$P(x\|c, \mu) = \frac{1}{\sqrt{2\pi}} [(1-c) \exp - \frac{x^2}{2} + c \exp - \frac{(x-\mu)^2}{2}]$		$\{(c, \mu)\}$	? / X
	Poisson 1	$P(x\|\lambda_1, \lambda_2) = \frac{1}{2} \left[\frac{\exp(-\lambda_1)\lambda_1^x}{x!} + \frac{\exp(-\lambda_2)\lambda_2^x}{x!}\right]$		$\{(\lambda_1, \lambda_2)\}$	X / ?
	Poisson 2	$P(x\|c, \lambda) = c \frac{\exp(-1)}{x!} + (1-c) \frac{\exp(-\lambda)\lambda^x}{x!}$		$\{(c, \lambda)\}$	? / X
Neural Network	2-layered	$\frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(y - as(bx) - cx)^2)$	$\frac{1}{2}(ab + c)^2 + \frac{3}{2}a^2b^4$	$\{(a, b, c)\}$	? / X
	3-layered	$p(x, y\|a, b, c, d) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(y - a\sigma(bx) - c\sigma(dx))^2$	$\frac{1}{2} \int (a\sigma(bx) + c\sigma(dx))^2 q(x) dx$	$\{(a, b, c, d)\}$	X / X
	Gaussian RBM	$\frac{1}{Z} \exp(-E(v, h))$ (E : associated energy)		$\{w_{ij}\}$	? / X

3.1 Bayesian Networks

There are many cases of which variables are dependent. A Bayesian network represents conditional dependencies between variables. Thus, it is commonly used to describe cause and effect. Formal definitions refer to [10].

Definition 3.1.1 ([10]). *A "directed graph" is a pair (V, E), where V is a finite, nonempty set whose elements are called nodes, and E is a set of ordered pairs of distinct elements of V. Elements of E are called directed edges, and if there exists an edge from X to Y, denote $(X, Y) \in E$.*

In particular, a directed graph is called "directed acylic graph (in short, DAG)" if it contains no cycles.

Remark 3.1.1. *Let $G = (V, E)$ be a DAG. For nodes X, Y in V, if $(X, Y) \in E$, X is called a "parent" of Y and Y is called a "descendent" of X, respectively. If Y is neither a descendent of X nor a parent of X, Y is called "nondescendent" of X.*

Definition 3.1.2 (Bayesian network, [10]). *A joint probability distribution P of random variables in set V and a DAG $G = (V, E)$ are given. We say (G, P) satisfies the "Markov condition" if for each $X \in V$, X is conditionally independent of the set of all its nondescendents given the set of all its parents. If (G, P) satisfies the Markov condition, it is called a "Bayesian network".*

Since singular model contains regular model as a special case, we will go over one example of regular model and then focus only on strictly singular model.

Example 3.1.1 (Gaussian Bayesian Networks). *Suppose the number of antibodies in our body becomes* 10 *times more than before a day when some*

virus infected. Let's assume that viral life cycle would be subject to normal distribution with mean 10 *and standard deviation* σ*. However, if that virus infected before, the number of antibodies increases faster due to the effect of memory cells. Therefore, the other influences should be considered, and they are assumed to be normally distributed with mean* 0 *and standard deviation* σ*. Let two random variables* X *be the days under virus influence and* Y *be the number of antibodies in our body. That is, the number of antibodies will be* $10x$ *if the virus survives* x *days. However, antibodies may be made more or less than this based on the other influences. In other words,*

$$Y = 10X + \epsilon_Y,$$

where $\epsilon_Y \sim N(0, \sigma^2)$*. Since the expectation of those other influences is* 0*, we can deduce that* $E[Y|x] = 10x$*, and since the variance of those other influences is* σ^2*,* $V[Y|x] = \sigma^2$*. Consequently,* Y *is distributed conditionally as follows;*

$$P(y|x) = NormalDen(10x, \sigma^2),$$

where $NormalDen(\cdot, \cdot)$ *means the probability density function of normal distribution with given mean and variance.*

Generally, we can consider the case Y *is a linear function of its parents plus an error term* ϵ_Y *which is subject to normal distribution with mean* 0 *and variance* σ^2*. Let* $X_1, \cdots, X_k$ *be the parents of* Y*, then we can describe the situation by*

$$y = a_1x_1 + a_2x_2 + \cdots + a_kx_k + \epsilon_Y,$$

where $P(\epsilon_Y) = NormalDen(0, \sigma^2)$*, and* Y *is distributed conditionally as follows;*

$$P(y|x) = NormalDen(a_1x_1 + a_2x_2 + \cdots + a_kx_k, \sigma^2).$$

Remark 3.1.2. *In the above example, consider the case $\sigma = 1$ and $\{X_n\}$ is a i.i.d sequence, each X_n is subject to standard normal distribution. Then by the argument in previous example, we can deduce*

$$P(x, y|a_1, \cdots, a_k) = \frac{1}{\sqrt{2\pi}} \exp \frac{-(y - a_1x_1 - \cdots - a_kx_k)^2}{2}$$

If the true distribution is given by $(b_1, \cdots, b_k)$, then the Kullback-Leibler distance $K(a_1, \cdots, a_k)$ can be computed as follows;

$$K(a_1, \cdots, a_k) = \frac{1}{2} \int ((a_1 - b_1)x_1 + \cdots + (a_k - b_k)x_k)^2 q(x, y)dxdy.$$

If we let $Z := (a_1 - b_1)x_1 + \cdots + (a_k - b_k)x_k$, $K(a_1, \cdots, a_k)$ becomes $\frac{1}{2}E[Z^2]$, which is equal to $\frac{1}{2}Var[Z]$ because expectation of Z is 0.
The condition i.i.d. implies $Var[(a_1 - b_1)x_1 + \cdots + (a_k - b_k)x_k] = (a_1 - b_1)^2 Var[x_1] + \cdots + (a_k - b_k)^2 Var[x_k]$, so we can conclude

$$K(a_1, \cdots, a_k) = \frac{1}{2}((a_1 - b_1)^2 + \cdots + (a_k - b_k)^2).$$

Note that $K(a_1, \cdots, a_k) = 0$ if and only if $a_i = b_i$ for all $1 \leq i \leq k$. The Fisher information matrix

$$I(a_1, \cdots, a_k) = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & 1 \end{pmatrix}$$

$k \times k$ identity matrix, hence it is positive definite for an arbitrary $(a_1, \cdots, a_k)$. Therefore, it is an example of regular statistical model.

However, strictly singular case can also be found in Bayesian network.

Example 3.1.2 ([11]). *(Discrete exponential Bayesian network) Let X_1, X_2, X_3, and Y are random variables which take values $\{-1, 1\}$. Consider a Bayesian network which is defined by the probability distribution*

$$P(x, y|a, b, c) = \frac{1}{Z} \exp(ax_1 y + bx_2 y + cx_3 y),$$

where $Z = Z(a, b, c)$ is normalizing constant. Hence, the probability distri-

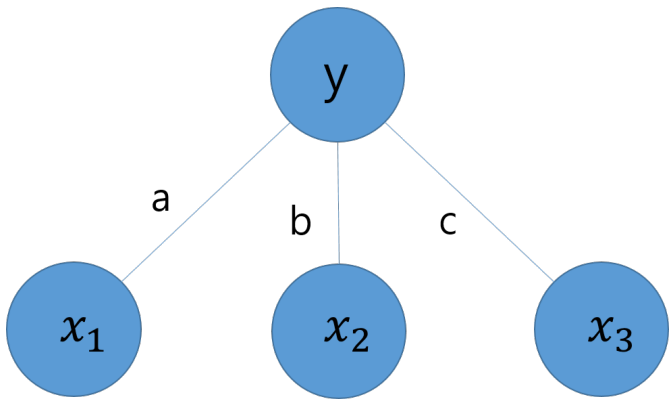

Figure 3.1: Discrete exponential Bayesian network

bution of X is given by

$$\begin{aligned} P(x|a, b, c) &= \frac{1}{Z} \sum_{y=\pm 1} \exp(ax_1 y + bx_2 y + cx_3 y) \\ &= \frac{1}{2Z} \cosh(ax_1 + bx_2 + cx_3). \end{aligned}$$

Note that $\tanh(ax_i) = \tanh(a)x_i$ *for* $x_i = \pm 1$, *and*

$$\begin{aligned} \cosh(u + v) &= \cosh(u)\cosh(v) + \sinh(u)\sinh(v), \\ \sinh(u + v) &= \sinh(u)\cosh(v) + \cosh(u)\sinh(v). \end{aligned}$$

Thus, we can deduce that

$$P(x|a,b,c) = \frac{1}{8}[1+\tanh(a)\tanh(b)x_1x_2+\tanh(b)\tanh(c)x_2x_3+\tanh(c)\tanh(a)x_3x_1].$$

If the true parameters $(a_0, b_0, c_0) = (0,0,0)$

$$K(a,b,c) = \frac{1}{2}(a^2b^2 + b^2c^2 + c^2a^2) + \cdots,$$

and the Fisher information matrix is equal to zero at $(0,0,0)$*. Thus, this is an example of strictly singular statistical model.*

3.2 Hidden Markov Models

In the Markov model introduced in chapter 2, we know all the states and transition matrix. However, if we have little knowledge about a given word, can we deduce it exactly? To be specific, given information is just the number of intersection points when we draw a vertical line on middle of alphabet. When we observe a sequence $3, 3, 2$, we can try to predict what is the original word. This is one example of Hidden Markov models.

Definition 3.2.1 (Hidden Markov Model, [8])**.** *Let* $\{S_1, \cdots, S_n\}$ *be a set of states and* p_{ij} *be a transition probability* $P(S_j|S_i)$*, and* B *is a set of outputs in a Markov model. An observation sequence is given by* $O = \{o_1, \cdots, o_t\}$*, where* $o_i \in B$ *for all* i*. Then the goal of hidden Markov model is compute* $P(S_1, \cdots, S_t|o_1, \cdots, o_t)$*. In other words, from observation, we want to find hidden states. Note that*

$$P(S_1, \cdots, S_t|o_1, \cdots, o_t) = \frac{P(o_1, \cdots, o_t|S_1, \cdots, S_t)P(S_1, \cdots, S_t)}{P(o_1, \cdots, o_t)}.$$

Here, by the Markov property, we can deduce

$$P(S_1, \cdots, S_t) = \prod_{i=1}^{t} P(S_i|S_{i-1}).$$

Besides, since each of observations are independent,

$$P(o_1, \cdots, o_t | S_1, \cdots, S_t) = \prod_{i=1}^{t} P(o_i | S_i).$$

Thus, if we let $\pi_i := P(S_i)$ *be initial state probabilities and* $b_i := P(o_i|S_i)$*, a hidden Markov model is parametrized by* $\{p_{ij}\}$*,* $\{b_i\}$*, and* $\{\pi_i\}$*.*

Example 3.2.1. *Consider the case when the emission probabilities are independently subject to the standard normal distribution. In addition, assume that* $\pi_i = \frac{1}{t}$ *for all* i*. For convenience, let* $o_i = [a_i, b_i] \subset \mathbb{R}$*. for* $i = 1, \cdots, t$*. Thus, for a given observation sequence* $O = \{o_1, \cdots, o_t\}$*, we can compute* $P(S_1, \cdots, St | o_1, \cdots, o_t)$ *as follows;*

$$P(S_1, \cdots, S_t | o_1, \cdots, o_t) = \frac{P(o_1, \cdots, o_t | S_1, \cdots, S_t) P(S_1, \cdots, S_t)}{P(o_1, \cdots, o_t)}.$$

By the argument in the above definition,

$$P(S_1, \cdots, S_t | o_1, \cdots, o_t) = \frac{\prod_{i=1}^{t} P(o_i|S_i) \prod_{i=1}^{t} P(S_i|S_{i-1})}{P(o_1, \cdots, o_t)}.$$

By substituting $P(S_i|S_{i-1}) = p_{i-1,i}$ *and* $P(o_i|S_i) = \int_{[a_i,b_i]} \frac{1}{\sqrt{2\pi}} \exp -\frac{1}{2}x^2 dx$*,*

$$P(S_1, \cdots, S_t | o_1, \cdots, o_t) = \frac{\prod_{i=1}^{t} p_{i-1,i} \prod_{i=1}^{t} \int_{[a_i,b_i]} \frac{1}{\sqrt{2\pi}} \exp -\frac{1}{2}x^2 dx}{P(o_1, \cdots, o_t)},$$

where the parameter space is $\{(p_{i,j}, a_i, b_i) | a_{i,j} \geq 0, \sum_j p_{ij} = 1, a_i, b_i \in \mathbb{R}, a_i \leq b_i\}$*. Therefore, we can deduce that*

$$\begin{aligned} & \frac{\partial}{\partial p_{i-1,i}} \log P(S_1, \cdots, S_t | o_1, \cdots, o_t) \\ = \quad & \frac{\prod_{j \neq i} p_{j-1,j}}{P(S_1, \cdots, S_t | o_1, \cdots, o_t)} \frac{F(a_i,b_i)P(O) - p_{i-1,i} F(a_i,b_i) P(O)'}{P(O)^2}, \end{aligned}$$

where $F(a_i, b_i) = \prod_{i=1}^{t} \int_{[a_i,b_i]} \frac{1}{\sqrt{2\pi}} \exp -\frac{1}{2}x^2 dx$. *On the other hand,*

$$\begin{aligned} & \frac{\partial}{\partial a_j} \log P(S_1, \cdots, S_t | o_1, \cdots, o_t) \\ = \; & \frac{\prod_{i=1}^{t} p_{i-1,i}}{P(S_1,\cdots,S_t|o_1,\cdots,o_t)} \frac{-\exp(-\frac{1}{2}a_j^2)P(O) - F(a_i,b_i)P(O)'}{P(O)^2}, \end{aligned}$$

and

$$\begin{aligned} & \frac{\partial}{\partial b_j} \log P(S_1, \cdots, S_t | o_1, \cdots, o_t) \\ = \; & \frac{\prod_{i=1}^{t} p_{i-1,i}}{P(S_1,\cdots,S_t|o_1,\cdots,o_t)} \frac{\exp(-\frac{1}{2}b_j^2)P(O) - F(a_i,b_i)P(O)'}{P(O)^2}, \end{aligned}$$

where $F(a_i, b_i) := \prod_{i=1}^{t} \int_{[a_i,b_i]} \frac{1}{\sqrt{2\pi}} \exp -\frac{1}{2}x^2 dx$ *and* $P(O) := P(o_1, \cdots, o_t)$. *Therefore, for parameters* $(p_{i,j}, a_i, b_i)$ *with* $p_{i-1,i} = 0$, *the Fisher information matrix is zero, hence it is a strictly singular statistical model.*

3.3 Mixtures of Statistical Models

There are many cases which are poorly suited for a single specific distribution. Hence, if we consider the mixture of several distributions, it will explain the case more adequately. It is possible to treat mixtures of any statistical model we already know, but Gaussian mixtures is the most popular example.

Definition 3.3.1 (Gaussian mixtures)**.** *The probability density function of "Gaussian mixtures" is given by weighted sum of* n *probability density functions of normal distribution. In other words, if we let* $P(x)$ *the probability density function of mixtures from* $N(\mu_1, \sigma_1^2), \cdots, N(\mu_n, \sigma_n^2)$,

$$P(x) = \sum_{i=1}^{n} \frac{c_i}{\sqrt{2\pi}\sigma_i} \exp[-\frac{1}{2}(\frac{x-\mu_i}{\sigma_i})^2],$$

where $\{c_i\}$ *is positive mixture weights, i.e.,* $\sum_{i=1}^{n} c_i = 1$.

Therefore, a Gaussian mixture model is parametrized by weights $\{c_i\}$, means $\{\mu_i\}$, and standard deviations $\{\sigma_i\}$.

At first, let's discuss some simple examples.

Example 3.3.1. *Consider the case of mixture of two normal distributions whose standard deviations are equally* 1*. Also, assume that both normal distributions influence equally to the mixture, i.e.,* $c_1 = c_2 = 1/2$*. Then, the probability density function is computed as;*

$$P(x|\mu_1, \mu_2) = \frac{1}{2\sqrt{2\pi}} \left[\exp[-\frac{(x-\mu_1)^2}{2}] + \exp[-\frac{(x-\mu_2)^2}{2}] \right],$$

with parameter space $\{(\mu_1, \mu_2)|\mu_i \in \mathbb{R}\}$*. Then for* (μ_1, μ_2) *with* $\mu_1 \neq \mu_2$*, if we interchange their roles, we can get same probability density function. Therefore, this model is not identifiable.*

Example 3.3.2 ([11])**.** *Here, let's consider the case with different parameter set. The probability density function is given by*

$$P(x|c, \mu) = \frac{1}{\sqrt{2\pi}}[(1-c)\exp[-\frac{x^2}{2}] + c\exp[-\frac{(x-\mu)^2}{2}]],$$

with parameter space $\{(c, \mu)|0 \leq c \leq 1, -\infty < \mu < \infty\}$*.*

Then we can deduce that

$$\frac{\partial}{\partial c} \log P(x|c, \mu) = \frac{1}{\sqrt{2\pi}} \frac{(\exp \frac{-(x-\mu)^2}{2} - \exp \frac{-x^2}{2})}{P(x|c, \mu)},$$

$$\frac{\partial}{\partial \mu} \log P(x|c, \mu) = \frac{c}{\sqrt{2\pi}} \frac{(exp \frac{-(x-\mu)^2}{2}(x-\mu))}{P(x|c, \mu)}.$$

Therefore, for $(c, \mu) = (0, 0)$*, the Fisher information matrix is zero, so this is a strictly singular statistical model.*

It is also possible to consider the mixture of discrete probability distributions.

Example 3.3.3. *(Poisson mixtures) For λ_1, λ_2, consider the case of mixture of two Poisson distribution with parameter λ_i. Assume both influence equally to the mixture, i.e., $c_1 = c_2 = 1/2$. In other words,*

$$P(x|\lambda_1, \lambda_2) = \frac{1}{2}\left[\frac{\exp(-\lambda_1)\lambda_1^x}{x!} + \frac{\exp(-\lambda_2)\lambda_2^x}{x!}\right].$$

Then by the same argument in example 3.3.1, this model is not identifiable, so it is an example of strictly singular model.

Example 3.3.4. *Consider the probability density function is given by*

$$P(x|c, \lambda) = c\frac{\exp(-1)}{x!} + (1-c)\frac{\exp(-\lambda)\lambda^x}{x!},$$

with parameter space $\{(c, \lambda)|0 \leq c \leq 1, \lambda > 0\}$. Then

$$\frac{\partial}{\partial c}\log P(x|c, \lambda) = \frac{\frac{1}{ex!} - \frac{\exp(-\lambda)\lambda^x}{x!}}{P(x|c, \lambda)},$$

and

$$\frac{\partial}{\partial \lambda}\log P(x|c, \lambda) = \frac{(1-c) \times \frac{\exp(-\lambda)\lambda^{x-1}(x-\lambda)}{x!}}{P(x|c, \lambda)}.$$

Therefore, for $(c, \lambda) = (1, 1)$, the Fisher information matrix is zero, so this is a strictly singular statistical model.

Then, it is natural that we want to find an relationship between a Kullback-Leibler distance of mixture model and that of components. There is an useful inequality between them as follows.

Lemma 3.3.1. *For $\{x_i, y_i \geq 0 | i = 1, \cdots, n\}$, the equality*

$$\sum_{i=1}^{n} x_i \log\frac{x_i}{y_i} \geq \left(\sum_{i=1}^{n} x_i\right)\log\frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} y_i},$$

with equality holds if and only if x_i/y_i are constant.

Proof. Without loss of generality, we can assume $x_i, y_i > 0$ for all i. Then we can get the following inequality

$$\sum c_i(t_i \log t_i) \geq (\sum c_i t_i) \log(\sum c_i t_i),$$

for $c_i \geq 0$ with $\sum_i c_i = 1$ by applying the Jensen's inequality to a convex function $t \log t$. The equality holds if and only if $t_1 = \cdots = t_n$. If we substantiate $c_i = \frac{y_i}{\sum_j y_j}$ and $t_i = \frac{x_i}{y_i}$, the above inequality becomes

$$\sum \frac{x_i}{\sum y_j} \log \frac{x_i}{y_i} \geq \sum \frac{x_i}{\sum y_j} \log \sum \frac{x_i}{\sum y_j}.$$

By multiplying $\sum y_j$ to both sides, we can prove lemma. □

Now, we can find an relationship between Kullback-Leibler distance of mixture and that of its components.

Theorem 3.3.1 ([6])**.** *Given two mixture densities $\sum_{i=1}^n c_i f_i$ and $\sum_{i=1}^n d_i g_i$. Then the Kullback-Leibler distance of these densities has a upper bound as follows;*

$$K(\sum_{i=1}^n c_i f_i || \sum_{i=1}^n d_i g_i) \leq K(\mathbf{c}||\mathbf{d}) + \sum_{i=1}^n c_i K(f_i||g_i),$$

where $\mathbf{c} = (c_1, \cdots, c_n)$ and $\mathbf{d} = (d_1, \cdots, d_n)$. The equality holds if and only if $\frac{c_i f_i}{\sum_i c_i f_i} = \frac{d_i g_i}{\sum_i d_i g_i}$ for all i.

Proof. By definition, $K(\sum_{i=1}^n c_i f_i || \sum_{i=1}^n d_i g_i) = \int (\sum_i c_i f_i) \log \frac{\sum_i c_i f_i}{\sum_i d_i g_i}$. Then by applying the above lemma,

$$\begin{aligned}
\int (\sum_i c_i f_i) \log \frac{\sum_i c_i f_i}{\sum_i d_i g_i} &\leq \int \sum_i c_i f_i \log \frac{c_i f_i}{d_i g_i} \\
&= \sum_i c_i \log \frac{c_i}{d_i} + \sum_i c_i \int f_i \log \frac{f_i}{g_i} \\
&= K(\mathbf{c}||\mathbf{d}) + \sum_i c_i K(f_i||g_i).
\end{aligned}$$

□

Example 3.3.5. *By using the above theorem, we can find an upper bound of Kullback-Leibler distance for Poisson mixture described in example 3.3.3. Let* (λ_1', λ_2') *be a true parameter. Then*

$$K\left(\sum_{i=1}^{2}\frac{1}{2}\frac{\exp(-\lambda_i')\lambda_i'^x}{x!}||\sum_{i=1}^{2}\frac{1}{2}\frac{\exp(-\lambda_i)\lambda_i^x}{x!}\right) \leq \sum_{i=1}^{2}\frac{1}{2}K(\frac{\exp(-\lambda_i')\lambda_i'^x}{x!}||\frac{\exp(-\lambda_i)\lambda_i^x}{x!}).$$

Note that we can compute the right side. Since $\log(\frac{\exp(-\lambda_i')\lambda_i'^x}{x!}/\frac{\exp(-\lambda_i)\lambda_i^x}{x!}) = x\log(\frac{\lambda_i'}{\lambda_i}) + \lambda_i - \lambda_i'$ *and the Kullback-Leibler distance is the expectation with respect to true distribution, we can deduce that*

$$K(\frac{\exp(-\lambda_i')\lambda_i'^x}{x!}||\frac{\exp(-\lambda_i)\lambda_i^x}{x!}) = \lambda_i'\log(\frac{\lambda_i'}{\lambda_i}) + \lambda_i - \lambda_i'.$$

3.4 Layered Neural Networks

Definition 3.4.1 (Neural Networks, [9])**.** *A "neural network" is a network of simple elements called "neurons", which receive input and change their state according to given input and an "activation function", then it produces "output". The network has a form of directed and weighted graph, where the neurons are the nodes and the connection between neurons are weighted directed edges, called "synapses".*

Example 3.4.1 (Perceptron, [9])**.** *Neural networks can be used to solve the pattern recognition problem. Specifically, by comparing the output with a given threshold, if the output is bigger then the input vector is assigned to class 0, and otherwise it belongs to class 1. The simplest example is a*

perceptron, which has only one unit. If $x_1, \cdots, x_d$ *are given inputs, by taking sign of* $x_1w_1 + \cdots + x_dw_d$*, where each* x_i *is connected to the unit with weight* w_i*, we can classify the input vector into two classes. In this case, the value of threshold would be* 0*.*

The neural networks is a method used to explain human brains. However, in the above definition, there exists one problem. If a loop exists in the network, it is difficult to decide the result of computation performed by network. Therefore, more adaptable notion should be needed.

Definition 3.4.2 (Layered Neural Networks, [9])**.** *The units in a layered neural network are organized in layers, with the output of neurons in a layer serving as the inputs to the next layer's neurons. Therefore, there are no loops.*

As we can see in perceptron, the classification rule is depend on the weights w_i's. Therefore, we need to choose weights which are well-fitted to the training example. In the case of a single unit, since the relation between weights and the output is simple, so we can find "good" weights by a simple method called perceptron convergence procedure.

Remark 3.4.1 (Perceptron Convergence Procedure, [9])**.** *Let* a *be a output and* t *be a target. Suppose that the output is not same to target, so we should adjust weight* w_j *by* $\triangle w_j$ *for each* j*. Let's assume* $t = 1$ *and* $a = -1$*, i.e.,* $x_1w_1 + \cdots + x_dw_d < 0$ *and we want it to be positive. If we replace* w_j *by* $w_j + \triangle w_j$*, the term* $(\triangle w_j)x_j$ *is added. If* $x_j > 0$*, by increasing the value* w_j*, we can make* $x_1w_1 + \cdots + x_dw_d < 0$ *positive, and otherwise, by decreasing the values* w_j*, we can reach a goal.*

As the above remark, we have a systematic way to find well-fitted weights. However, in the case of layered neural networks, there are prob-

lems. First, we no longer have a simple relationship between weights and the output. Moreover, if we use discontinuous threshold function like sign function, a small change in the weight induces drastic change of the output. Since outputs of units are treated as inputs of the next layer's neuron, so effect of changing weights would be tremendous. Therefore, we need a smooth threshold function.

Definition 3.4.3 ([9]). *A bounded function with "S" shape which is differentiable and increasing is called a "sigmoid function". The function* $f(x) := \frac{1}{1+\exp(-x)}$ *is the one of examples of sigmoid function.*

In the case of layered neural network, we also want to minimize the difference between target and output. Consider the error over training examples as a function of weights in the network, and let W be a set of all possible weights which is considered as a parameter space. For each $w \in W$, we can compute $(target - output)^2$, thus the set of all these values is called "error surface". Therefore, one goal in layered neural network can be rewritten by finding a choice of weights for which the error surface is as low as possible. The concrete process of this is described in [10].

Example 3.4.2 ([11]). *Consider a layered statistical model with weights* a, b, c *and a activation function* $s(x) := x + x^2$ *as we can see in Figure 3.2. Assume the output is normally distributed conditionally. In other words,*

$$p(y|x, w) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(y - as(bx) - cx)^2).$$

Then the Kullback-Leibler distance $K(a, b, c)$ *can be computed as follows;*

$$K(a, b, c) = \frac{1}{2} \int (as(bx) + cx)^2 q(x) dx.$$

Then if we let $Z = ab^2X^2 + (ab + c)X$, K *is equal to* $\frac{1}{2}E[Z^2]$, *which can be deduced from* $Var[Z]$ *and* $E[Z]$. *First,* $E[Z] = ab^2E[X^2] + (ab + c)E[X] =$

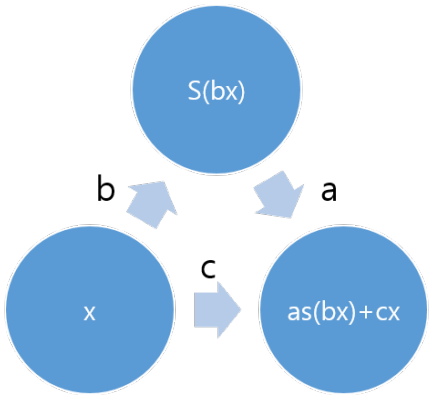

Figure 3.2: Layered neural network - example 3.4.2

$ab^2E[X^2] = ab^2(Var[X]) = ab^2$ since X is supposed to be subject to standard normal distribution. On the other hand, $Var[Z] = Var[ab^2X^2 + (ab + c)X] = a^2b^4Var[X^2] + (ab + c)^2Var[X] + 2ab^2(ab + c)Cov[X, X^2]$. Here, note that $Cov[X, X^2] = E[X(X^2 - 1)] = E[X^3] = 0$. So, the only thing to need is the value of $Var[X^2]$, which is equal to $E[X^4] - E[X^2]^2 = 3 - 1 = 2$. Therefore, $K = \frac{1}{2}E[Z^2] = \frac{1}{2}(Var[Z] + a^2b^4) = \frac{1}{2}(ab + c)^2 + \frac{3}{2}a^2b^4$.

Hence $K(a, b, c) = 0$ if and only if $ab = c = 0$. The Fisher information matrix is equal to zero at $(0, 0, 0)$, so it is a strictly singular statistical model.

Example 3.4.3 ([11])**.** *There are many situations explained by a layered neural network with one input unit X, one output unit Y, and the remain units in the second layer are all hidden. Let's consider the simplest one which has two units in the second layer with weights (a, b, c, d) described as in Figure 3.3.*

Assume Y is subject to standard normal distribution and an activation

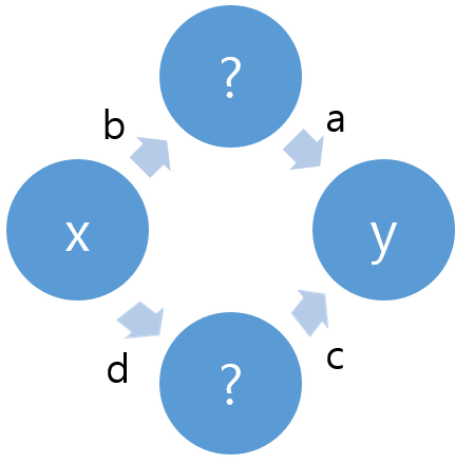

Figure 3.3: Layered network with one input, two hidden units, and one output - example 3.4.3

function $\sigma(x) := \exp(x) - 1$. *Then*

$$p(x, y|a, b, c, d) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(y - a\sigma(bx) - c\sigma(dx))^2.$$

Hence,

$$K(a, b, c, d) = \frac{1}{2} \int (a\sigma(bx) + c\sigma(dx))^2 q(x)dx.$$

If we use the Taylor expansion,

$$a\sigma(bx) + c\sigma(dx) = \sum_{k=1}^{\infty} \frac{x^k}{k!}(ab^k + cd^k),$$

$K(a, b, c, d) = 0$ *is equivalent to* $p_k := ab^k + cd^k$ *is all zero for* k. *However, we can check that* $p_n \in < p_1, p_2 >$ *for all* n. *So, we should focus on a function*

$$f(a, b, c, d) := (ab + cd)^2 + (ab^2 + cd^2)^2.$$

By recursive blow-ups, we can obtain the coordinate is given by

$$\begin{aligned} a &= a, \\ b &= b_1 d, \\ c &= a(b_1 - 1)b_1 c_5 d - ab_1. \\ d &= d. \end{aligned}$$

On this coordinate,

$$f = d^4 a^2 b_1^2 (b_1 - 1)^2 \{c_5^2 + (1 + c_5 d)^2\}.$$

Since $g'(u)$ is $ab_1(b_1 - 1)d^2$, the real log canonical threshold λ is $\frac{3}{4}$.

In particular, the case, when activation function is identity $\sigma(x) = x$, is called the "reduced rank regression", which is the main object of next chapter.

3.5 Boltzmann Machines

Definition 3.5.1 (Boltzmann Machine, [5])**.** *A "Boltzmann machine" is a stochastic recurrent neural network with an associated energy. It consists of not only visible variables but also hidden variables, which is a difference from Hopfield network. Each states have binary (0 or 1) values and every nodes are connected with weight. Thus, a boltzmann machine is parametrized by $\{U, V, W, b, c\}$, where U are visible-visible weights, V are hidden-hidden weights, and W are visible-hidden weights. b and c mean biases of visible and hidden units repectively.*

Remark 3.5.1 (Restricted Boltzmann machine, [5])**.** *In the above definition, if a Boltzmann machine is consdiered as a two-layered neural network (i.e. there are no connections between visible-visible and hidden-hidden.),*

we call it "restricted Boltzmann machine" (in short, RBM). Therefore, an RBM is a bi-partite graph of m visible and n hidden units, which is parametrized by $\{w_{ij}, b_j, c_i | 1 \leq j \leq m, 1 \leq i \leq n\}$. The associated energy $E(v,h)$ is defined as follows;

$$E(v,h) = -\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij}h_iv_j - \sum_{j=1}^{m} b_jv_j - \sum_{i=1}^{n} c_ih_i.$$

Then the joint distribution of visible and hidden units (v,h) is

$$P(v,h|w_{ij}, b_j, c_i) = \frac{\exp(-E(v,h))}{Z},$$

where Z is normalizing factor, i.e., $Z = \sum_{v\in\{0,1\}^m}\sum_{h\in\{0,1\}^n}\exp(-E(v,h))$.

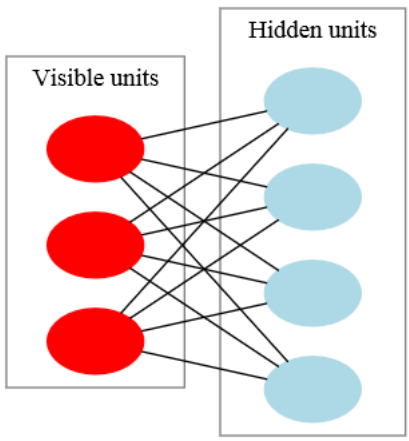

Figure 3.4: Structure of RBM

In the above definition, values of v and h are just 0 or 1, but we can consider a case when both v and h are subject to Gaussian distribution.

Definition 3.5.2 (Gaussian RBM, [3])**.** *A "Gaussian RBM" is an RBM with the associated energy is given by*

$$E(v,h) = \sum_{j=1}^{m} \frac{(v_j - b_j)^2}{2\sigma_v^2} + \sum_{i=1}^{n} \frac{(h_i - c_i)^2}{2\sigma_h^2} - \sum_{j=1}^{m}\sum_{i=1}^{n} \frac{(v_j - b_j)w_{ij}(h_j - c_i)}{\sigma_v \sigma_h}.$$

Example 3.5.1. *Consider a Gaussian RBM with mean vectors* $\mathbf{b}$*,* $\mathbf{c}$*, and* $\sigma_v = \sigma_h = 1$*. In other words, the joint distribution of* (v,h) *is given by*

$$\begin{aligned} P(v,h|w_{ij}) &= \frac{1}{Z}\exp(-\sum_{j=1}^{m}\frac{(v_j - b_j)^2}{2} - \sum_{i=1}^{n}\frac{(h_i - c_i)^2}{2} \\ &+ \sum_{j=1}^{m}\sum_{i=1}^{n}(v_j - b_j)w_{ij}(h_i - c_i)). \end{aligned}$$

Therefore, for parameters $\{w_{ij}\}$*, we can obtain*

$$\frac{\partial}{\partial w_{ij}} \log P = \frac{1}{Z^2}[-\exp(-E)(v_j - b_j)(h_i - c_i) \times Z - \exp(-E)\partial_{w_{ij}} Z],$$

$$\frac{\partial}{\partial b_j} \log P = \frac{1}{Z^2}[\exp(-E)(v_j - b_j) \times Z - \exp(-E)\partial_{b_j} Z],$$

$$\frac{\partial}{\partial c_i} \log P = \frac{1}{Z^2}[\exp(-E)(h_i - c_i) \times Z - \exp(-E)\partial_{c_i} Z].$$

Since $\partial_{w_{ij}} Z$ *has term* $(v_j - b_j)(h_i - c_i)$*,* $\partial_{b_j} Z$ *has term* $(v_j - b_j)$*, and* $\partial_{c_i} Z$ *has term* $(h_i - c_i)$*. Therefore, for* $(v = (v_j), h = (h_i)) = ((b_j),(c_i))$*, the Fisher information matrix is zero, hence it is a strictly singular model.*

Chapter 4

The Reduced Rank Regression

In this chapter, as I mentioned earlier, we will focus on the reduced rank regression, whose activation function is trivial.

Definition 4.0.1 (The reduced rank regression, [11])**.** *Let M be the number of inputs and N be that of outputs. Assume there exist H hidden variables, and r means the rank of the true distribution. The reduced rank regression is described by a conditional probability density function as follows;*

$$p(y|x,w) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2}|y - xAB|^2\right),$$

where $x \in \mathbb{R}^M$, $y \in \mathbb{R}^N$, A is a $M \times H$ matrix, B is a $H \times N$ matrix, and σ is a constant. Hence, the parameter set $W = \{(A,B)|A \in \mathfrak{M}_{M,H},\ B \in \mathfrak{M}_{H,N}\}$.

Remark 4.0.1. *In the above definition, if we let the true distribution $q(x,y|w)$ is given by (A_0, B_0), then the Kullback-Leibler distance is described*

as;

$$K(w) = \frac{1}{2}||AB - A_0B_0||^2,$$

where $||\ ||$ *is the matrix norm.*

4.1 The case when M=N=H and r=0

Example 4.1.1 (M=N=H=2 and r=0 case, [11])**.** *Let* $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ *and* $B = \begin{pmatrix} e & f \\ g & h \end{pmatrix}$. *Since* $r = 0$,

$$\begin{aligned} 2K(w) &= \left\| \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} e & f \\ g & h \end{pmatrix} \right\|^2 \\ &= (ae+bg)^2 + (af+bh)^2 + (ce+dg)^2 + (cf+dh)^2. \end{aligned}$$

For the coordinate given by

$$a = a, \ \ b = ab', \ \ c = ac', \ \ d = ad',$$

$$e' = e + b'g, \ \ f' = f + b'h, \ \ d'' = d' - b'c',$$

we can deduce that

$$2K(w) = a^2(e'^2 + f'^2 + (c'e' + d''g)^2 + (c'f' + d''h)^2).$$

Since this is equivalent to $a^2(e'^2 + f'^2 + d''^2g^2 + d''^2h^2)$, *we can take blow-up at the center* $< e', f', d'' >$. *In other words, for the coordinate*

$$\begin{aligned} e' &= e' \\ f' &= e'\tilde{f} \\ d'' &= e'\tilde{d}, \end{aligned}$$

$$2K(w) = a^2 e'^2 (1 + \tilde{f}^2 + \tilde{d}^2 g^2 + \tilde{d}^2 h^2).$$

Since $g'(u) = a^3 e'^2$, *the real log canonical threshold is* $\lambda = \frac{3}{2}$.

Example 4.1.2 (M=N=H=3 and r=0 case)**.** *Let* $A = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}$ *and* $B = \begin{pmatrix} j & k & l \\ m & n & o \\ p & q & r \end{pmatrix}$. *By the same argument of the previous example, we can obtain*

$$\begin{aligned} 2K(w) &= (aj + bm + cp)^2 + (ak + bn + cq)^2 + (al + bo + cr)^2 \\ &+ (dj + em + fp)^2 + (dk + en + fq)^2 + (dl + eo + fr)^2 \\ &+ (gj + hm + ip)^2 + (gk + hn + iq)^2 + (gl + ho + ir)^2. \end{aligned}$$

Then by using blow-up and isomorphism,

$$a = a, \;\; b = ab', \;\; c = ac', \;\; \cdots, i = ai',$$

$$j' = j + b'm + c'p, \;\; k' = k + b'n + c'q, \;\; l' = l + b'o + c'r,$$

we can rewrite $2K(w)$ *as follows;*

$$\begin{aligned} 2K(w) &= a^2[(j + b'm + c'p)^2 + (k + b'n + c'q)^2 + (l + b'o + c'r)^2] \\ &+ a^2[(d'j + e'm + f'p)^2 + (d'k + e'n + f'q)^2 + (d'l + e'o + f'r)^2] \\ &+ a^2[(g'j + h'm + i'p)^2 + (g'k + h'n + i'q)^2 + (g'l + h'o + i'r)^2] \\ &= a^2[j'^2 + k'^2 + l'^2] \\ &+ a^2[(d'j + e'm + f'p)^2 + (d'k + e'n + f'q)^2 + (d'l + e'o + f'r)^2] \\ &+ a^2[(g'j + h'm + i'p)^2 + (g'k + h'n + i'q)^2 + (g'l + h'o + i'r)^2]. \end{aligned}$$

Now, under isomorphism

$$e'' = e' - b'd', \;\; f'' = f' - c'd', \;\; h'' = h' - b'g', \;\; i'' = i' - c'g',$$

$$\begin{aligned} 2K(w) &= a^2[j'^2 + k'^2 + l'^2] \\ &+ a^2[(d'j' + e''m + f''p)^2 + (d'k' + e''n + f''q)^2 + (d'l' + e''o + f''r)^2] \\ &+ a^2[(g'j' + h''m + i''p)^2 + (g'k' + h''n + i''q)^2 + (g'l' + h''o + i''r)^2]. \end{aligned}$$

This is equivalent to

$$\begin{aligned} 2K_1(w) &= a^2[j'^2 + k'^2 + l'^2] \\ &+ a^2[(e''m + f''p)^2 + (e''n + f''q)^2 + (e''o + f''r)^2] \\ &+ a^2[(h''m + i''p)^2 + (h''n + i''q)^2 + (h''o + i''r)^2]. \end{aligned}$$

Therefore, by taking blow-up at the center $< j', k', l', e'', f'', h'', i'' >$ *as follows;*

$$j' = j', \;\; k' = \tilde{k}j', \;\; l' = \tilde{l}j', \;\; e'' = \tilde{e}j', \cdots, i'' = \tilde{i}j',$$

$$\begin{aligned} 2K_1(w) &= a^2 j'^2[1 + \tilde{k}^2 + \tilde{l}^2 \\ &+ (\tilde{e}m + \tilde{f}p)^2 + (\tilde{e}n + \tilde{f}q)^2 + (\tilde{e}o + \tilde{f}r)^2 \\ &+ (\tilde{h}m + \tilde{i}p)^2 + (\tilde{h}n + \tilde{i}q)^2 + (\tilde{h}o + \tilde{i}r)^2]. \end{aligned}$$

Since $g'(u) = a^8 j'^6$, *the real log canonical threshold* λ *is* $\frac{7}{2}$.

Example 4.1.3 (M=N=H=n and r=0 case)**.** *For* $A = (a_{ij})$ *and* $B = (b_{ij})$ $(1 \leq i, j \leq n)$, *we can generalize the value of the real log canonical threshold by the argument used in above examples. By using blow-up and isomorphism,*

$$a_{11} = a_{11}, \;\; a_{ij} = a_{11}a'_{ij} \;\; ((i,j) \neq (1,1)),$$

$$b'_{1k} = b_{1k} + \sum_{l=2}^{n} a'_{1l} b_{lk},$$
$$a''_{ik} = a'_{ik} - a'_{1k} a'_{i1} \quad (2 \leq i, k \leq n),$$

we can deduce $2K(w)$ *is equivalent to* $K_1(w)$ *expressed as follows;*

$$\begin{aligned} 2K_1(w) &= a_{11}^2 [\sum_{i=1}^{n} b_{1i}'^2 \\ &+ \sum_{i=1}^{n} (\sum_{j=2}^{n} a''_{2j} b_{ji})^2 \\ &+ \sum_{i=1}^{n} (\sum_{j=2}^{n} a''_{3j} b_{ji})^2 \\ &+ \cdots \\ &+ \sum_{i=1}^{n} (\sum_{j=2}^{n} a''_{nj} b_{ji})^2]. \end{aligned}$$

Then by taking blow-up at the center $< b_{1i}', a_{2j}'', \cdots, a_{nj}''; 1 \leq i \leq n, 2 \leq j \leq n >$*, on one coordinate,*

$$2K_1(w) = a_{11}^2 {b'_{11}}^2 [1 + F[\tilde{b_{1i}}, \tilde{a_{2i}}, \cdots, \tilde{a_{ni}}]],$$

where $F[\tilde{b_{1i}}, \tilde{a_{2i}}, \cdots, \tilde{a_{ni}}]$ *is a nonnegative polynomial with* $\tilde{b_{1i}}, \tilde{a_{2i}}, \cdots, \tilde{a_{ni}}$ $(2 \leq i \leq n)$*. Since the number of elements in* $\{b_{1i}', a_{2j}'', \cdots, a_{nj}''; 1 \leq i \leq n, 2 \leq j \leq n\}$ *is* $n + (n-1)^2$*,* $g'(u) = {b'_{11}}^{n+(n-1)^2-1} = {b'_{11}}^{n(n-1)}$*. Therefore, for the case when* $M = N = H = n$ *with rank zero, the real log canonical threshold* λ *is* $\frac{n(n-1)+1}{2}$*.*

Therefore, we can completely compute Kullback-Leibler distance, its resolution of singularities, and the real log canonical threshold for the case when the number of inputs, outputs, and hidden variables are same and rank of true distribution $r = 0$. Then, let's consider the case of $r = 1$.

4.2 Other cases

Example 4.2.1 (M=H=N=2 and r=1 case)**.** *For the case $r \neq 0$, the Kullback-Leibler distance is no longer expressed as $\frac{1}{2}||AB||^2$. Let A and B be as described in the example 4.1.1, and since A_0B_0 for true parameter (A_0, B_0) is 2×2 matrix with rank 1, let*

$$A_0B_0 = \begin{pmatrix} i & ki \\ j & kj \end{pmatrix}$$

Thus, $K(w)$ is computed as follows;

$$\begin{aligned} 2K(w) &= \left\| \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} e & f \\ g & h \end{pmatrix} - \begin{pmatrix} i & ki \\ j & kj \end{pmatrix} \right\|^2 \\ &= (ae + bg - i)^2 + (af + bh - ki)^2 + (ce + dg - j)^2 + (cf + dh - kj)^2. \end{aligned}$$

Thus, by taking blow-up and isomorphism

$$a = a, \;\; b = ab', \;\; c = ac', \;\; d = ad', \;\; i = ai', \;\; j = aj',$$

$$e' = e + b'g - i', f' = f + b'h - ki', d'' = d' - b'c', \;\; j'' = j' - c'i',$$

it follows that

$$\begin{aligned} 2K(w) &= a^2[(e + b'g - i')^2 + (f + b'h - ki')^2 + (c'e + d'g - j')^2 + (c'f + d'h - kj')^2] \\ &= a^2[e'^2 + f'^2 + (c'e' + d''g - j'')^2 + (c'f' + d''h - kj'')^2], \end{aligned}$$

which is equivalent to

$$2K_1(w) = a^2[e'^2 + f'^2 + (d''g - j'')^2 + (d''h - kj'')^2].$$

By taking blow-up at the center $< e', f', d'', j'' >$, then for the coordinate

$$e' = e', \;\; f' = \tilde{f}e', \;\; d'' = \tilde{d}e', \;\; j'' = \tilde{j}e',$$

$$2K_1(w) = a^2 e'^2[1 + F(\tilde{f}, \tilde{d}, \tilde{j})],$$

where F is a nonnegative polynomial on $\tilde{f}, \tilde{d}, \tilde{j}$. Since $g'(u) = a^5 e'^3$, the real log canonical threshold λ is $\frac{4}{2} = 2$.

Example 4.2.2. *(M=H=N=n and r=1 case) For true parameter (A_0, B_0), let*

$$A_0 B_0 = (\mathbf{c}, k_1\mathbf{c}, \cdots, k_{n-1}\mathbf{c}),$$

where $\mathbf{c} = (c_1, c_2, \cdots, c_n)^t$. Then it becomes

$$\begin{aligned}
2K(w) &= [(\sum_{i=1}^n a_{1i}b_{i1} - c_1)^2 + (\sum_{i=1}^n a_{1i}b_{i2} - k_1c_1)^2 + \cdots + (\sum_{i=1}^n a_{1i}b_{in} - k_{n-1}c_1)^2] \\
&+ [(\sum_{i=1}^n a_{2i}b_{i1} - c_2)^2 + (\sum_{i=1}^n a_{2i}b_{i2} - k_1c_2)^2 + \cdots + (\sum_{i=1}^n a_{2i}b_{in} - k_{n-1}c_2)^2] \\
&+ \cdots \\
&+ [(\sum_{i=1}^n a_{ni}b_{i1} - c_n)^2 + (\sum_{i=1}^n a_{ni}b_{i2} - k_1c_n)^2 + \cdots + (\sum_{i=1}^n a_{ni}b_{in} - k_{n-1}c_n)^2].
\end{aligned}$$

By taking blow-up and isomorphism,

$$a_{11} = a_{11}, \;\; a_{ij} = a_{11}a'_{ij} \;\; ((i,j) \neq (1,1)),$$

$$b'_{1k} = b_{1k} + \sum_{l=2}^n a'_{1l}b_{lk} - c'_k,$$

$$a''_{ik} = a'_{ik} - a'_{1k}a'_{i1} \;\; (2 \leq i, k \leq n), \;\; c''_k = c'_k - a'_{k1}c'_1 \;\; (2 \leq k \leq n),$$

we can deduce $2K(w)$ is equivalent to $K_1(w)$ expressed as follows;

$$\begin{aligned} 2K_1(w) &= \sum_{i=1}^{n} b_{1i}'^2 \\ &+ \sum_{i=1}^{n}(\sum_{j=2}^{n} a_{2j}'' b_{ji} - k_{i-1}c_2'')^2 \\ &+ \sum_{i=1}^{n}(\sum_{j=2}^{n} a_{3j}'' b_{ji} - k_{i-1}c_3'')^2 \\ &+ \cdots \\ &+ \sum_{i=1}^{n}(\sum_{j=2}^{n} a_{nj}'' b_{ji} - k_{i-1}c_n'')^2, \end{aligned}$$

where k_0 is 1. Then by taking blow-up at the center $< b_{1i}', a_{2j}'', \cdots, a_{nj}'', c_k''; 1 \leq i \leq n, 2 \leq j \leq n, 2 \leq k \leq n >$, on one coordinate,

$$2K_1(w) = a_{11}^2 b_{11}'^2 [1 + F[\tilde{b_{1i}}, \tilde{a_{2i}}, \cdots, \tilde{a_{ni}}, \tilde{c_i}]],$$

where $F[\tilde{b_{1i}}, \tilde{a_{2i}}, \cdots, \tilde{a_{ni}}, \tilde{c_i}]$ is a nonnegative polynomial with $\tilde{b_{1i}}, \tilde{a_{2i}}, \cdots, \tilde{a_{ni}}, \tilde{c_i}$ $(2 \leq i \leq n)$. Since the number of elements in $\{b_{1i}', a_{2j}'', \cdots, a_{nj}'', c_j'' | 1 \leq i \leq n, 2 \leq j \leq n\}$ is $n + (n-1)^2 + (n-1) = n^2$, the real log canonical threshold λ is $\frac{n^2}{2}$.

We can also consider a extreme case.

Example 4.2.3 (M=m, N=n, H=1, r=0)**.** *In this case matrices A and B become vectors, so we can write $A = (a_1, \cdots, a_m)^T$ and $B = (b_1, \cdots, b_n)$.*

Thus, we can compute the Kullback-Leilbler distance as follows;

$$\begin{aligned} 2K(w) &= \left\| \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix} \begin{pmatrix} b_1 & b_2 & \cdots & b_n \end{pmatrix} \right\|^2 \\ &= \sum_{1 \le i \le m, 1 \le j \le n} (a_i b_j)^2. \end{aligned}$$

Thus, by taking blow-up

$$a_1 = a_1, \;\; a_2 = a_1 a_2', \;\; \cdots, \;\; a_m = a_1 a_m',$$

It can be rewritten as

$$2K(w) = {a_1}^2 (\sum_{1 \le j \le n} {b_j}^2)[1 + \sum_{2 \le i \le m} {a_i'}^2].$$

By taking blow-up

$$b_1 = b_1, \;\; b_2' = b_1 b_2', \;\; \cdots, \;\; b_n = b_1 b_n',$$

$$2K(w) = a_1^2 {b_1}^2 [1 + \sum_{2 \le i \le m} {a_i'}^2][1 + \sum_{2 \le j \le n} {b_j'}^2].$$

Since $g'(u) = {a_1}^{m-1} {b_1}^{n-1}$, *the real log canonical threshold* λ *is* $\frac{\min\{m,n\}}{2}$.

Chapter 5

Determinantal Variety

In this chapter, we will discuss an object of algebraic geometry, determinantal variety. Some definitions refer to [7].

Definition 5.0.1. *(Generic Determinantal Variety, [7]) Let M be the projective space $\mathbb{P}^{mn-1}$ which is associated to the vector space of $m \times n$ matrices. For a positive integer k, $M_k \subset M$ is defined by the subset of matrices of rank at most k. This is called a "generic determinantal variety".*

Remark 5.0.1. *By the definition, a generic determinantal variety is the common zero locus of all $(k+1) \times (k+1)$ minor determinants.*

Example 5.0.1. *Let A be a $m \times n$ matrix. Then we can deduce that A is of at most rank 1 if and only if there exist two vectors $v = (v_1, \cdots, v_m)$ and $w = (w_1, \cdots, w_n)$ such that $A = v^t w$. Therefore, the M_1 is just Segre variety, the image of the Segre map $\sigma : \mathbb{P}^{m-1} \times \mathbb{P}^{n-1} \to \mathbb{P}^{mn-1}$.*

Then we have following propositions described in [4], [7].

Proposition 5.0.1 ([7])**.** *M_k is an irreducible algebraic subvariety of M of codimension $(m-k)(n-k)$.*

Proof. Let $\tilde{M}_k$ be defined by

$$\tilde{M}_k = \tilde{M}_k(m,n) = \{(A,\Lambda) \in M \times G(n-k,n) | A\Lambda = 0\}.$$

Then we can consider $\tilde{M}_k$ as an algebraic vector bundle over $G(n-k,n)$ of rank $mk-1$ by projection to $G(n-k,n)$. Thus, $\tilde{M}_k$ is connected and has dimension $dim(G(n-k,n)) + (mk-1) = k(m+n-k)-1$. For the projection $\pi : M \times G(n-k,n) \to M$, π maps $\tilde{M}_k$ onto M_k, hence M_k is an irreducible algebraic subvariety of codimension $(m-k)(n-k)$. □

Proposition 5.0.2 ([7])**.** *The singular locus of M_k is equal to M_{k-1}.*

Proof. For a point $A \in M_l - M_{l-1}$, i.e., a matrix of rank exactly l, we can choose bases for K^m and K^n so that A can be represented as follows;

$$\begin{pmatrix} 1 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & & \ddots & & & & \vdots \\ 0 & \cdots & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & & & & & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \end{pmatrix}$$

Then in affine neighborhood U in A given by $\{X_{11} \neq 0\}$, for the coordinates

$\{x_{ij} := X_{ij}/X_{11}\}$, we can write an element of U as follows;

$$\begin{pmatrix} 1 & x_{1,2} & x_{1,3} & \cdots & \cdots & \cdots & x_{1,m} \\ x_{2,1} & 1+x_{2,2} & x_{2,3} & \cdots & \cdots & \cdots & x_{2,m} \\ \vdots & & \ddots & & & & \vdots \\ x_{l,1} & \cdots & \cdots & 1+x_{l,l} & x_{l,l+1} & \cdots & x_{l,m} \\ x_{l+1,1} & \cdots & \cdots & x_{l+1,l} & x_{l+1,l+1} & \cdots & x_{l+1,m} \\ \vdots & & & & & \ddots & \vdots \\ x_{n,1} & \cdots & \cdots & \cdots & \cdots & \cdots & x_{n,m} \end{pmatrix},$$

For $l = k$, we should identify the linear terms of the $(k+1)\times(k+1)$ minors of this matrix. The only such minors with nonzero differential at origin A are those involving the first k rows and columns, so their linear terms are exactly the coordinates $x_{i,j}$ with $i, j > k$. But the number of such $x_{i,j}$ is $(m-k)(n-k)$, we can deduce that M_k is smooth at a point of $M_k - M_{k-1}$, so $Sing(M_k) \subset M_{k-1}$.
On the other hand, for $l < k$, there is no $(k+1)\times(k+1)$ minor which has any linear terms. Since $(k+1)\times(k+1)$ minors generate the ideal of M_k, projective tangent space to M_k at $A \in M_{k-1}$ is all of M. Therefore, M_{k-1} is exactly the singular locus of M_k.

□

Remark 5.0.2 (Tangent space to determinantal variety, [7])**.** *For a smooth point $A \in M_k - M_{k-1}$, the tangent space to M_k at A can be described as follows;*

$$\mathbb{T}_A(M_k) = \mathbb{P}\{\varphi \in Hom(K^n, K^m) | \varphi(Ker(A)) \subset Im(A)\}.$$

Proof. In terms of the coordinates $\{x_{i,j}\}$, the tangent space to M_k at A can be identified as the space of matrices whose lower right $(m-k)\times(n-k)$

submatrix is zero. Note that the bases $\{e_i\}, \{f_j\}$ for K^n and K^m satisfy that the kernel of A is spanned by $\{e_{k+1}, \cdots, e_n\}$ and the image of A is spanned by $\{f_1, \cdots, f_k\}$. Thus, we have proved that the projective tangent space to M_k at A corresponds the linear maps $\varphi : K^n \to K^m$ mapping the kernel of A into the image of A. $\square$

Remark 5.0.3 (Tangent cone to determinantal variety, [7])**.** *Similarly, we can also describe the tangent cones to the determinantal varieties. Let A be a point of $M_l - M_{l-1}$. As in the proposition 5.0.2, an element of U, affine neighborhood in A can be expressed as*

$$\begin{pmatrix} 1 & x_{1,2} & x_{1,3} & \cdots & \cdots & \cdots & x_{1,m} \\ x_{2,1} & 1+x_{2,2} & x_{2,3} & \cdots & \cdots & \cdots & x_{2,m} \\ \vdots & & \ddots & & & & \vdots \\ x_{l,1} & \cdots & \cdots & 1+x_{l,l} & x_{l,l+1} & \cdots & x_{l,m} \\ x_{l+1,1} & \cdots & \cdots & x_{l+1,l} & x_{l+1,l+1} & \cdots & x_{l+1,m} \\ \vdots & & & & & \ddots & \vdots \\ x_{n,1} & \cdots & \cdots & \cdots & \cdots & \cdots & x_{n,m} \end{pmatrix},$$

for coordinates $\{x_{ij} := X_{ij}/X_{11}\}$. We should find the leading terms of the $(k+1)\times(k+1)$ minors. They must be the terms in the expansions of minors involving the diagonal entries from the first $l \times l$ block. Therefore, they are exactly $(k+1-l) \times (k+1-l)$ minors of the lower right $(m-l) \times (n-l)$ submatrix. It implies that the tangent cone to M_k at A is contained in the space of matrices whose lower right $(m-l)\times(n-l)$ block has rank $\leq k-l$. However, this locus is irreducible and has dimension $((m-l)-(k-l))((n-l)-(k-l)) = (m-k)(n-k)$, so the above inclusion becomes equality.
As in the previous remark, note that the bases $\{e_i\}$ and $\{f_j\}$ for K^n and K^m satisfy that $Ker(A)$ is spanned by $\{e_{l+1}, \cdots, e_n\}$ and $Im(A)$ is span

of $\{f_1, \cdots, f_l\}$. *Hence, the lower right* $(m-l) \times (n-l)$ *submatrix of* B *represents the composition;*

$$B' : Ker(A) \hookrightarrow K^m \underbrace{\rightarrow}_{B} K^n \rightarrow K^n/Im(A).$$

Therefore,

$$\mathcal{T}_A(M_k) = \{B \in Hom(K^m, K^n) | rank(B' : Ker(A) \rightarrow Coker(A)) \leq k - l\}.$$

In chapter 4, we computed the real log canonical threshold λ for some cases in the reduced rank regression. Among them, we obtained a result $\lambda = \frac{\min\{m,n\}}{2}$ in the case of $M = m$, $N = n$, $H = 1$, and $r = 0$. In that case, $2K(w) = \|AB\|^2$ and AB is product of two vectors. Since a matrix with rank ≤ 1 if and only if it can be expressed as a product of two vectors, this case can be regarded as an example of determinantal variety, introduced in the beginning of this chapter. Precisely, it corresponds to M_1. From aforementioned facts, we have known

$$\{0\} = M_0 \subset M_1 \subset M_2 \subset \cdots,$$

M_{k-1} is the singular locus M_k.
Therefore, we can compute λ by using the resolution of singularities described in Proposition 5.0.1.

Remark 5.0.4 (M=m$\geq$ 2 ,N=2)**.** *Consider a space of* $m \times 2$ *matrices of rank at most 1. Then* $\tilde{M_k}$ *becomes*

$$\tilde{M_k} = \{(X, \Lambda) \in M \times G(1,2); X\Lambda = 0\}.$$

We can let

$$X = \begin{pmatrix} x_1 & x_2 \\ x_3 & x_4 \\ \vdots & \vdots \\ x_{2m-1} & x_{2m} \end{pmatrix}, \quad \Lambda = (y_1, y_2)^T,$$

with relations $x_{2i-1}y_1 + x_{2i}y_2 = 0$ *for all* $i = 1, 2, \cdots, m$. *On* $\{y_1 = 1\}$, *it becomes* $x_{2i-1} = -x_{2i}y_2$.
Therefore, X *becomes*

$$X = \begin{pmatrix} -x_2y_2 & x_2 \\ -x_4y_2 & x_4 \\ \vdots & \vdots \\ -x_{2m}y_2 & x_{2m} \end{pmatrix}.$$

On the other hand, in chapter 4,

$$AB = \begin{pmatrix} a_1b_1 & a_1b_2 \\ a_2b_1 & a_2b_2 \\ \vdots & \vdots \\ a_mb_1 & a_mb_2 \end{pmatrix},$$

and we took blow-up $b_1 = b_2b_1'$ *and* $b_2 = b_2$, *so it becomes*

$$AB = \begin{pmatrix} a_1b_2b_1' & a_1b_2 \\ a_2b_2b_1' & a_2b_2 \\ \vdots & \vdots \\ a_mb_2b_1' & a_mb_2 \end{pmatrix},$$

which is of form similar to X.
Therefore, it would be possible that problems in the reduced rank theorem can be solved by using geometric facts in determinantal variety.

Example 5.0.2 (M=N=3, H=2, and r=0 case)**.** *At first, we will compute the real log canonical threshold in this case by using blow-up used in chapter 4. Note that a* 3×3 *matrix is of rank at most 2 if and only if it can be represented by a product of* 3×2 *matrix and* 2×3 *matrix. Hence for* $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix}$ *and* $B = \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{pmatrix}$*, let's make* $2K(w)$ *have normal crossing singularities.*

$$\begin{aligned} 2K(w) &= \left\| \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{pmatrix} \right\|^2 \\ &= \sum_{i=1}^{3} (a_{11}b_{1i} + a_{12}b_{2i})^2 \\ &+ \sum_{i=1}^{3} (a_{21}b_{1i} + a_{22}b_{2i})^2 \\ &+ \sum_{i=1}^{3} (a_{31}b_{1i} + a_{32}b_{2i})^2 \end{aligned}$$

By taking blow-up and isomorphism;

$$a_{11} = a_{11}, \quad a_{ij} = a_{11}a'_{ij} \ ((i,j) \neq (1,1)),$$

$$b'_{1i} = b_{1i} + a'_{12}b_{2i},$$

$$a''_{22} = a'_{22} - a'_{21}a'_{12}, \quad a''_{32} = a'_{32} - a'_{31}a'_{12},$$

we can deduce that

$$2K(w) = a_{11}^2[b_{11}'^2 + b_{12}'^2 + b_{13}'^2 + \sum_{i=1}^{3}(a'_{21}b'_{1i} + a''_{22}b_{2i})^2 + \sum_{i=1}^{3}(a'_{31}b'_{1i} + a''_{32}b_{2i})^2].$$

This is equivalent to

$$2K_1(w) = a_{11}^2[b_{11}'^2 + b_{12}'^2 + b_{13}'^2 + (b_{21}^2 + b_{22}^2 + b_{23}^2)(a_{22}''^2 + a_{32}''^2)].$$

Therefore, by taking blow-up at the center $< b_{11}', b_{12}', b_{13}', a_{22}'', a_{23}'' >$, *on one coordinate,*

$$2K_1(w) = a_{11}^2 b_{11}'^2[1 + F[\tilde{b_{12}'}, \tilde{b_{13}'}, \tilde{a_{22}''}, \tilde{a_{32}''}]],$$

where $F[\tilde{b_{12}'}, \tilde{b_{13}'}, \tilde{a_{22}''}, \tilde{a_{32}''}]$ *is a nonnegative polynomial with* $\tilde{b_{12}'}, \tilde{b_{13}'}, \tilde{a_{22}''}, \tilde{a_{32}''}$. *Therefore, the real log canonical threshold is* λ *is* $\frac{5}{2}$.

Now, let's compute the same case by using the resolution of singularities described in [4]. Let X *be a* 3×3 *matrix is of rank at most 2. Then*

$$X = \begin{pmatrix} x_1 & x_2 & x_3 \\ x_4 & x_5 & x_6 \\ x_7 & x_8 & x_9 \end{pmatrix}, \quad det(X) = 0.$$

By using same notation in the beginning of this chapter,

$$\tilde{M_2} = \{(X, \Lambda) \in M \times G(1,3) | X\Lambda = 0\}.$$

Since $G(1,3) \cong \mathbb{P}^2$, *we can rewrite by*

$$\tilde{M_2} = \{ \left(\begin{pmatrix} x_1 & x_2 & x_3 \\ x_4 & x_5 & x_6 \\ x_7 & x_8 & x_9 \end{pmatrix}, \begin{pmatrix} a \\ b \\ c \end{pmatrix} \right) \}$$

$$with \quad ax_1 + bx_2 + cx_3 = 0, \; ax_4 + bx_5 + cx_6 = 0, ax_7 + bx_8 + cx_9 = 0.$$

On $\{a = 1\}$, *it becomes*

$$x_1 = -bx_2 - cx_3, \; x_4 = -bx_5 - cx_6, x_7 = -bx_8 - cx_9.$$

Therefore,

$$\begin{aligned} 2K(w) &= {x_1}^2 + {x_2}^2 + \cdots + {x_9}^2 \\ &= (bx_2 + cx_3)^2 + {x_2}^2 + {x_3}^2 \\ &+ (bx_5 + cx_6)^2 + {x_5}^2 + {x_6}^2 \\ &+ (bx_8 + cx_9)^2 + {x_8}^2 + {x_9}^2. \end{aligned}$$

By taking blow-up, then on $\{x_2x_6 - x_3x_5 = 1\}$*, we may assume* $x_5x_9 - x_6x_8 = 0$*. It's because that if* $x_5x_9 - x_6x_8 = k(\neq 0)$*, we can take isomorphism as follows ;*

1. $x'_8 = \frac{x_8}{k} + x_2$,
2. $x'_9 = \frac{x_9}{k} + x_3$.

Since

$$\begin{aligned} x_5x'_9 - x_6x'_8 &= x_5(\frac{x_9}{k} + x_3) - x_6(\frac{x_8}{k} + x_2) \\ &= \frac{(x_5x_9 - x_6x_8)}{k} - (x_2x_6 - x_3x_5) \\ &= 1 - 1 = 0, \end{aligned}$$

if we replace x_8 *by* x'_8*, and* x_9 *by* x'_9*, it becomes* $x_5x_9 - x_6x_8 = 0$*. Therefore, by taking blow-up at center* $< x_2, x_3, x_5, x_6, x_8, x_9 >$*, on one chart,*

$$\begin{aligned} 2K(w) &= {x_5}^2[(b{x_2}' + c{x_3}')^2 + {x'_2}^2 + {x'_3}^2 \\ &+ (b + cx'_6)^2 + 1 + {x'_6}^2 \\ &+ (bx'_8 + cx'_9)^2 + {x'_8}^2 + {x'_9}^2] \\ &= {x_5}^2[(b{x_2}' + c{x_3}')^2 + {x'_2}^2 + {x'_3}^2 \\ &+ (b + cx'_6)^2 + 1 + {x'_6}^2 \\ &+ (bx'_8 + cx'_6x'_8)^2 + {x'_8}^2 + {x'_6}^2{x'_8}^2]. \end{aligned}$$

Thus, we can obtain the same result $\lambda = \frac{5}{2}$*, so it would be possible to compute* λ *for general case by using the resolution of singularities of determinantal variety.*

Bibliography

[1] H. Akaike. : *Likelihood and Bayes procedure. Bayesian Statistics* University Press, pp. 143-166. (1980).

[2] S. Amari. : *Methods of Information Geometry.* Oxford University Press (2000).

[3] S. Amari. : *Information Geometry and Its Applications.* Springer (2016).

[4] E. Arbarello, M. Cornalba, P. Griffiths, and J. Harris. : *Geometry of Algebraic Curves.* Springer (1985).

[5] MA. Cueto, J. Morton, B. Sturmefels. : *Geometry of the restricted Boltzmann machine.* arxiv:0908.4425 (2009).

[6] Minh N. Do. : *Fast Approximation of Kullback-Leibler Distance for Dependence Trees and Hidden Markov Models* IEEE Signal Processing Letters, 10 (4), pp. 115-118. (2003).

[7] J. Harris. : *Algebraic Geometry A First Course.* Springer (1993).

[8] A. Hoover. : *Lecture Note of Analysis of tracking systems - Lecture 1. Introduction to Markov Models.* (2017).

[9] S. Kulkarni and G. Harman. : *An Elementary Introduction to Statistical Learning Theory.* Wiley (2011).

[10] E. Richard. : *Probabilistic methods for bioinformatics : with an introduction to Bayesian networks* Morgan Kaufmann. (2009).

[11] S. Watanabe. : *Algebraic Geometry and Statistical Learning Theory.* Cambridge University Press (2009).

[12] 전종우, 손건태 : *확률의 개념 및 응용* 자유아카데미 (2005).

국 문 초 록

이 논문에서는 특이 학습이론의 목표와 그 방법들을 소개한다. 특이점 문제를 해소하기 위해 대수기하에서 쓰이는 특이점 해결방법을 제시한다. 특이 학습이론의 몇 가지 예시들을 살펴보고 계산을 통해 특이점을 해소하는 과정을 제시한다. 특히, reduced rank regression 모델에 대해서는 어떤 블로우업을 통해 특이점을 해소하는지 자세한 과정을 제시한다. 끝으로 determinantal variety에 대해 간략히 소개한 뒤, 그것과 reduced rank regression 모델의 연결가능성을 확인한다.

주요어휘 : 통계적 학습이론, 특이점 해결
학번 : 2014-22355