Statistical Learning Theory

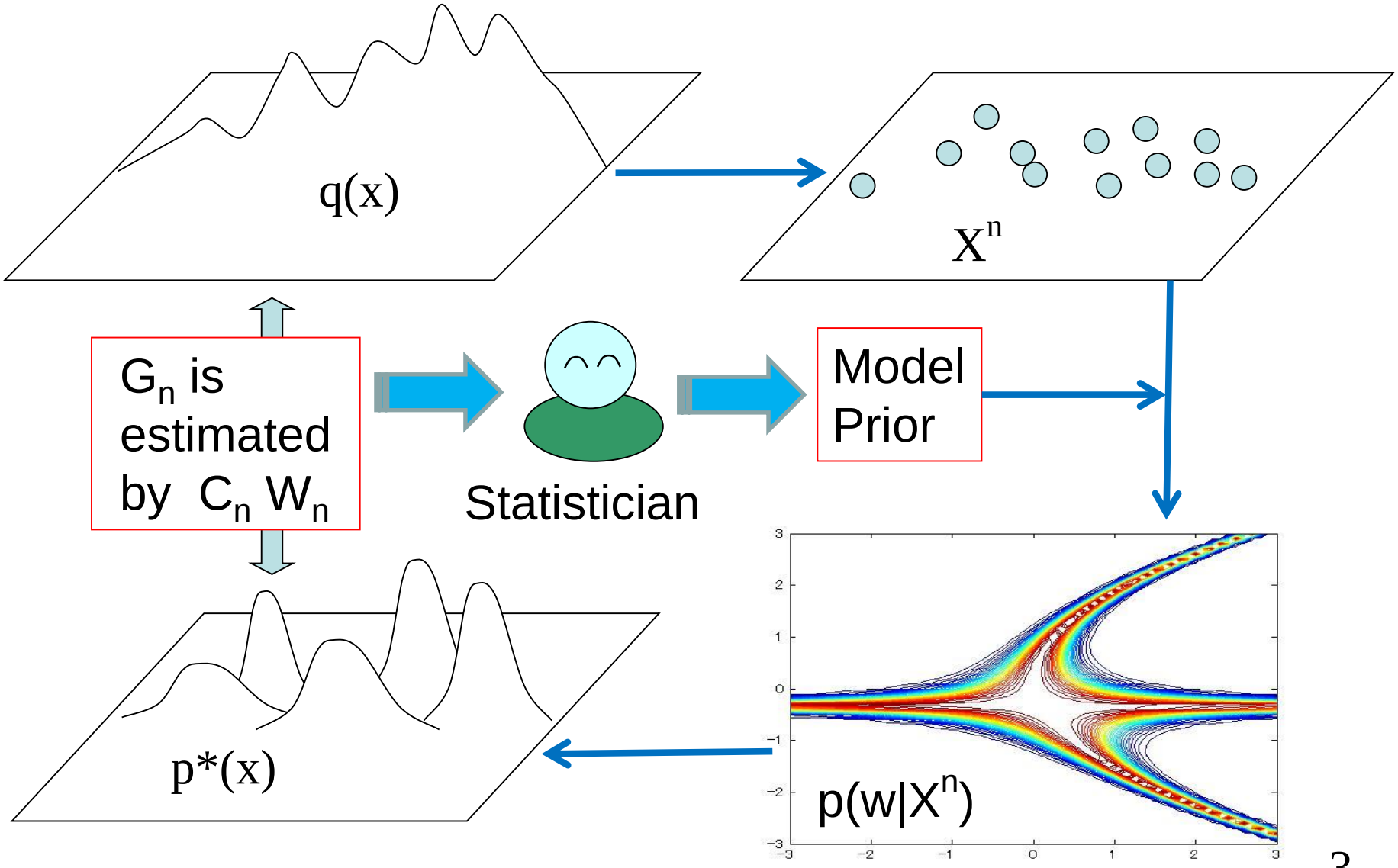
13 Cross Validation,
Information Criterion,
and Phase Transition

Sumio Watanabe

Tokyo Institute of Technology

1 Two birational Invariants

Mathematical Structure of Statistical Estimation



Review : Statistical Learning Theory

Let $(q(x), p(x|w), \varphi(w))$ be a triple of a true distribution, a statistical model, and a prior. Let w_0 be the parameter that minimizes $K(q(x)||p(x|w))$.

Using an i.i.d. sample X^n, whose p.d.f. is $q(x)$, we define

(1) posterior distribution $p(w|X^n)$,

(2) predictive distribution $p(x|X^n)$.

(3) and generalization loss.

Notations

Several expectations are given by

E : expectation over X^n (sample, or training set),

$\mathbf{E}_x$: expectation over X (test sample),

$\mathbf{E}_w$: expectation by the posterior average (= $\mathbf{E}_u$),

$< \quad | \xi >$: expectation by the renormalized posterior

$\mathbf{E}_\xi$: expectation over Gaussian process ξ.

Log Density Ratio Function

The log density ratio function is defined by

$$f(x,w) = \log (p(x|w_0) / p(x|w)),$$

which is equivalent to

$$p(x|w) = p(x|w_0) \exp(-f(x,w)).$$

Average by Posterior Distribution

The average of $\Psi(w)$ by the posterior distribution is

$$\mathbf{E}_w[\,\Psi(w)\,] = \frac{\int \Psi(w)\,\varphi(w)\,\Pi_i\, p(X_i|w)\,dw}{\int \varphi(w)\,\Pi_i\, p(X_i|w)\,dw},$$

which is equal to

$$\mathbf{E}_u[\,\Psi(g(u))\,] = \frac{\int \Psi(g(u))\,\varphi(g(u))\exp(-nK_n(g(u)))\,|g'(u)|\,du}{\int \varphi(g(u))\exp(-nK_n(g(u)))\,|g'(u)|\,du}.$$

Hironaka Theorem and Emprical Process

By using Hironaka theorem, $K(g(u)) = u^{2k}$ and

$f(x,g(u)) = u^k a(x,u)$ on each local coordinate, hence

$$n K_n(g(u)) = n u^{2k} - n^{1/2} u^k \xi_n(u),$$

where $\xi_n(u)$ is an empirical process which converges to a Gaussian process $\xi(u)$ in distribution on $C(W)$.

Renormalization of Posterior Distribution

Concentration of the posterior distribution is expressed by the renormalized posterior distribution,

$$\Pi_i\, p(X_i|w)\, \varphi(w)\, dw = \exp(-nL_n)\, \exp(-nK_n(w))\, \varphi(w)\, dw$$

$$= \exp(-nL_n)\, \exp(-nK_n(g(u)))\, \varphi(g(u))\, |g'(u)|\, du$$

$$= \exp(-nL_n)\, \exp(-n u^{2k} - n^{1/2} u^k \xi_n(u)))\, b(u)\, |u^h|\, du$$

$$= \exp(-nL_n)\, (\log n)^{m-1}/n^{\lambda} \int t^{\lambda-1} D(u) \exp(-t + t^{1/2}\xi_n(u))\, du\, dt$$

+ smaller order,

where λ and m are RCLT and its multiplicity, respectively.

Renormalization

Posterior ~ Renormalized Posterior

$\Pi_i\, p(X_i|w)\, \varphi(w)\, dw$

$\sim \exp(-nL_n)\, (\log n)^{m-1}/n^{\lambda}$ $\int t^{\lambda-1} D(u) \exp(- t + t^{1/2}\xi(u))\, du\, dt$

does not depend on n.

Renormalized Posterior Average

The average of $\Psi(u,t)$ by the renormalized posterior distribution is defined by

$$< \Psi(u,t) | \xi > = \left(\frac{\Sigma \int_{[0,1]^d} \int_0^{\text{infinity}} \Psi(u,t)\, t^{\lambda-1} D(u) \exp(- t + t^{1/2} \xi(u)) du dt}{\Sigma \int_{[0,1]^d} \int_0^{\text{infinity}} t^{\lambda-1}\, D(u)\, \exp(- t + t^{1/2} \xi\,(u)) du dt} \right).$$

It follows that, for an arbitrary $\alpha>0$,

$$\mathbf{E}_u[\, u^{\alpha k}\, F(u)\,] = n^{-\alpha/2} < F(u)\, t^{\alpha/2} \mid \xi_n > + o_p(n^{-\alpha/2}),$$

$$\mathbf{E}\, \mathbf{E}_u[\, u^{\alpha k}\, F(u)\,] = n^{-\alpha/2}\, \mathbf{E}_\xi < F(u)\, t^{\alpha/2} \mid \xi > + o(n^{-\alpha/2}).$$

Definition of Singular Fluctuation

Definition. The singular fluctuation is defined by using the renormalized posterior distribution,

$$\mathrm{Fluc}(\xi) = \mathbf{E}_x[\, <t\, a(X,u)^2 |\, \xi > - <t^{1/2} a(X,u) \mid \xi >^2 \,] \,,$$

$$\nu \;=\; (1/2)\, \mathbf{E}_{\xi}[\, \mathrm{Fluc}(\xi) \,] \,.$$

The meaning of singular fluctuation

We can prove the convergence as n tends to infinity,

$$\mathbf{E} \sum_{i=1}^{n} \mathbf{V}_w[\ \log p(X_i|w)\ ] \quad \text{goes to} \quad 2\nu.$$

Thus the singular fluctuation shows the limit value of the average variance of $\log p(X_i|w)$ by the posterior distribution.

Two Birational Invariants

The singular fluctuation ν and RLCT λ are birational invariants.

If $q(x)=p(x|w_0)$, $\varphi(w_0)>0$, and the Hessian matrix of $K(q(x)||p(x|w))$ at $w=w_0$ is positive definite, then $\lambda=\nu=d/2$, where d is the dimension of w.

In general, they are different from each other.

2 Cross Validation and Information Criterion

Definition of Generalization Loss

Definition. The generalization loss is defined by

$$G_n = - \mathbf{E}_x \log p(X | X^n)$$

$$= - \mathbf{E}_x \log \mathbf{E}_w[p(X | w)].$$

Remark. G_n is small if and only if $K(q(x)||p(x|X^n))$ is small. In practical applications, we want to minimize G_n, however, we cannot know it because the true q(x) is unknown.

Definition of Training Loss

Definition. The training loss is defined by

$$T_n = -(1/n) \sum_{i=1}^{n} \log p(X_i \mid X^n)$$

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

Remark. In practical applications, we can calculate the training loss without any information of the true distribution. However, it is different from G_n . Minimizing T_n does not minimize G_n in general.

Leave-One-Out Cross Validation Loss

Definition. The leave-one-out cross validation (LOOCV) loss is defined by

Set minus

$$C_n = -(1/n) \sum_{i=1}^{n} \log p(X_i \mid X^n - X_i)$$

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

Remark. For any n>1, $\mathbf{E}[C_n] = \mathbf{E}[G_{n-1}]$. In practical applications, we can calculate C_n even if the true distribution is unknown.

Definition of Information Criterion

Definition. The widely applicable information criterion (WAIC) is defined by

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

where $\mathbf{V}_w$ is the variance by the posterior distribution.

Remark. It is proved that W_n is asymptotically equivalent to C_n.

Asymptotic Losses

Main Theorem. Asymptotic behaviors are given by

$$G_n = L_0 + (1/2n)\{ 2\lambda + \langle t^{1/2}\xi_n(u)\rangle - \mathrm{Fluc}(\xi_n)\} + o_p(1/n),$$
$$T_n = L_n + (1/2n)\{ 2\lambda - \langle t^{1/2}\xi_n(u)\rangle - \mathrm{Fluc}(\xi_n)\} + o_p(1/n),$$
$$C_n = L_n + (1/2n)\{ 2\lambda - \langle t^{1/2}\xi_n(u)\rangle + \mathrm{Fluc}(\xi_n)\} + o_p(1/n),$$
$$W_n = L_n + (1/2n)\{ 2\lambda - \langle t^{1/2}\xi_n(u)\rangle + \mathrm{Fluc}(\xi_n)\} + o_p(1/n).$$

Their averages are given by

$$\mathbf{E}[G_n] = L_0 + \lambda / n + o(1/n),$$
$$\mathbf{E}[T_n] = L_0 + (\lambda - 2\nu) / n + o(1/n),$$
$$\mathbf{E}[C_n] = L_0 + \lambda / n + o(1/n),$$
$$\mathbf{E}[W_n] = L_0 + \lambda / n + o(1/n).$$

Relations among Asymptotic Losses

Corollary 1. From the Main Theorem, we obtain

$$\mathbf{E}[\, G_n \,] = \mathbf{E}[C_n] + o(1/n),$$

$$\mathbf{E}[\, G_n \,] = \mathbf{E}[W_n] + o(1/n),$$

$$(G_n - L_0) + (C_n - L_n) = 2\lambda/n + o_p(1/n),$$

$$(G_n - L_0) + (W_n - L_n) = 2\lambda/n + o_p(1/n).$$

If there exists w_0 such that $q(x)=p(x|w_0)$,

then $L_0 = S = -\mathbf{E}_x[\log q(x)]$, and $L_n = S_n = -(1/n) \Sigma \log q(X_i)$.

Inverse Correlation

Both the cross validation loss and the information criterion loss estimate the generalization loss.

They have the asymptotically same averages as the generalization loss, but have the inverse correlation.

Several researchers do not know this fact.

$$(G_n - L_0) + (C_n - L_n) = 2\lambda/n + o_p(1/n),$$

average	λ/n	λ/n	$2\lambda/n$
random	large	small	$2\lambda/n$
random	small	large	$2\lambda/n$

Difference between C_n and W_n

(1) If a sample is i.i.d., then C_n and W_n are asymptotically equivalent as a random variable.

(2) For an arbitrary n>1, $\mathbf{E}[G_{n-1}]=\mathbf{E}[C_n]$ holds, whereas $\mathbf{E}[C_n] = \mathbf{E}[W_n]+o(1/n)$ holds for large n.

(3) In an estimation of conditional probability q(y|x), C_n requires independency of $\{(X_i,Y_i)\}$, whereas W_n requires independecy of $\{(Y_i|X_i)\}$. W_n can be used even if $\{X_i\}$ is not independent.

Example.1 Reduced Rank Regression

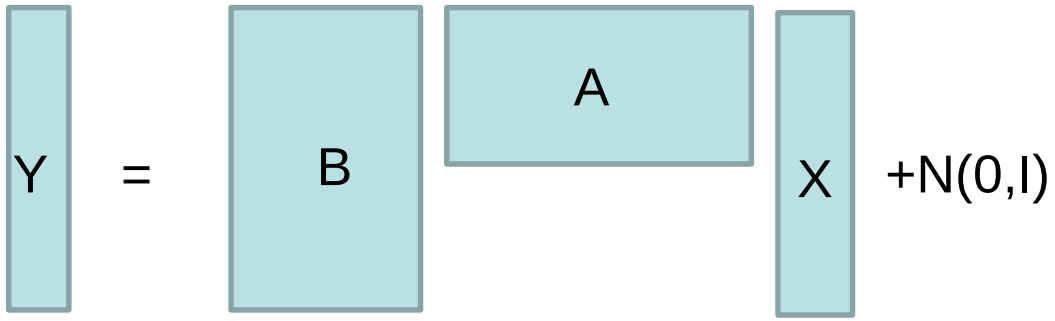
X in $\mathbf{R}^M$, Y in $\mathbf{R}^N$, B in $\mathbf{R}^{NH}$, A in $\mathbf{R}^{HM}$

A statistical model of reduced rank regression is defined by

model $p(y|x,A,B) = (1/2\pi)^{N/2} \exp(-(1/2)|| y-BAx||^2)$

prior $\varphi(A,B)$: positive on sufficiently large set.

true $q(y|x)=p(y|x,A_0,B_0)$, $q(x) =N(x,0,I)$.



RLCT of Reduced Rank Regression

Proposition.1 (Aoyagi's theorem)

Let H_0 be the rank of B_0A_0. Then the real log canonical threshold (RLCT) of reduced rank regression is given by

(1) If $M+N+H+H_0$ is even,

$\lambda=(1/8)(2(H+H_0)(M+N)-(M-N)^2-(H+H_0)^2)$, and $m=1$.

(2) If $M+N+H+H_0$ is odd,

$\lambda=(1/8)(2(H+H_0)(M+N)-(M-N)^2-(H+H_0)^2+1)$, and $m=2$.

An Experiment in Example.1

A sample (X^n, Y^n) (n=100) was generated from

$$p(y|x, B_0A_0).$$

We study a case when M=6, N=5, H=3, and

$$H_0 = \text{rank}(B_0A_0)=2.$$

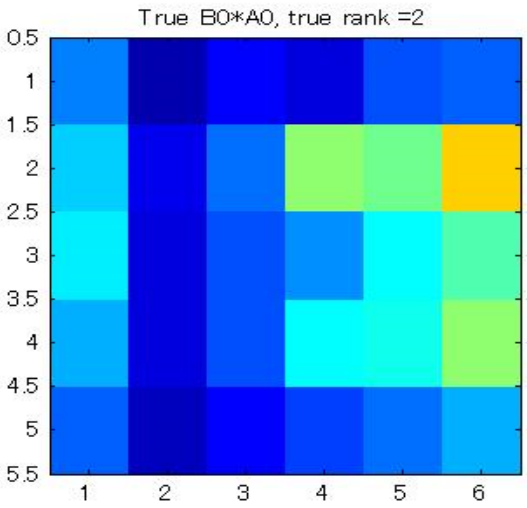
Since $M+N+H+H_0$ is even,

$$\lambda = (1/8)(2(H+H_0)(M+N)-(M-N)^2-(H+H_0)^2)=10.5.$$

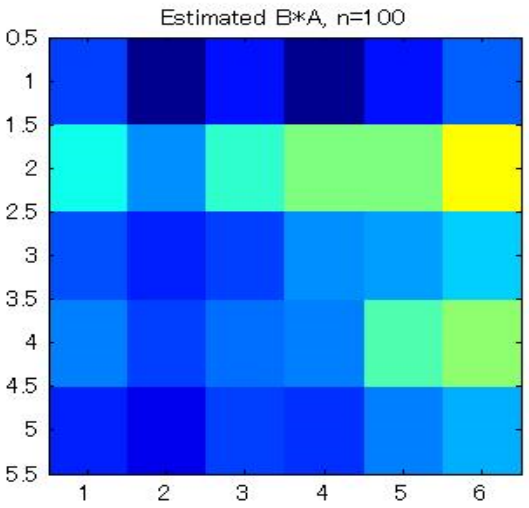
Hence the average is

$$\mathbf{E}[G_n]-S = \mathbf{E}[C_n-S_n] + o(1/n)=\lambda/n+o(1/n) = 0.105 + o(1/n).$$

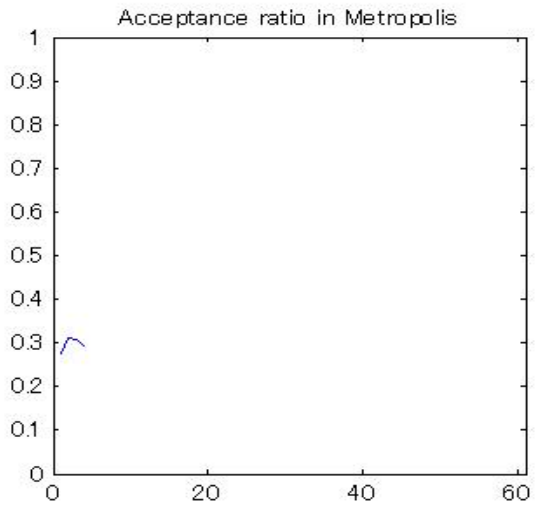
True : B_0A_0 5 times 6
rank(B_0A_0)=2



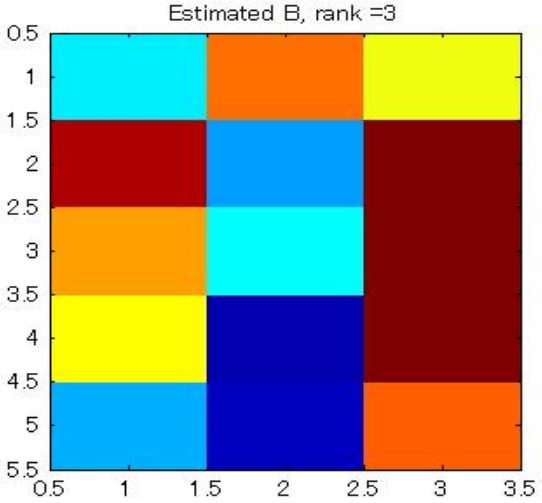
Posterior
Average (BA)



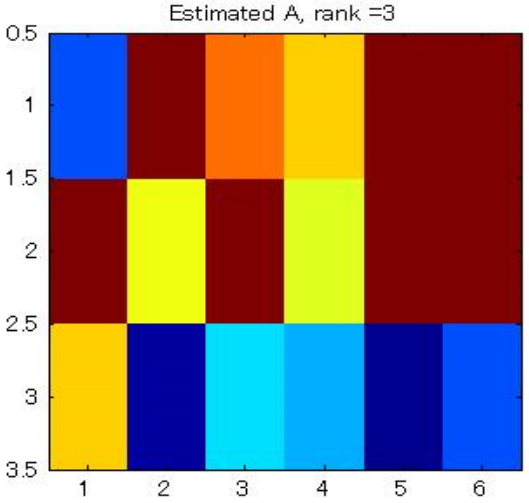
Acceptance
Prob in MCMC



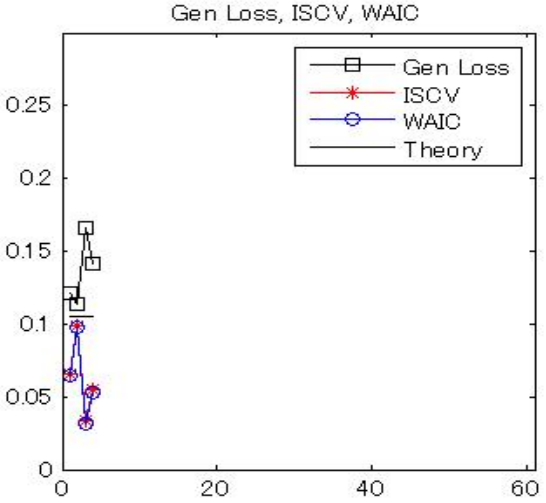
B(5 times 3)



A(3 times 6)



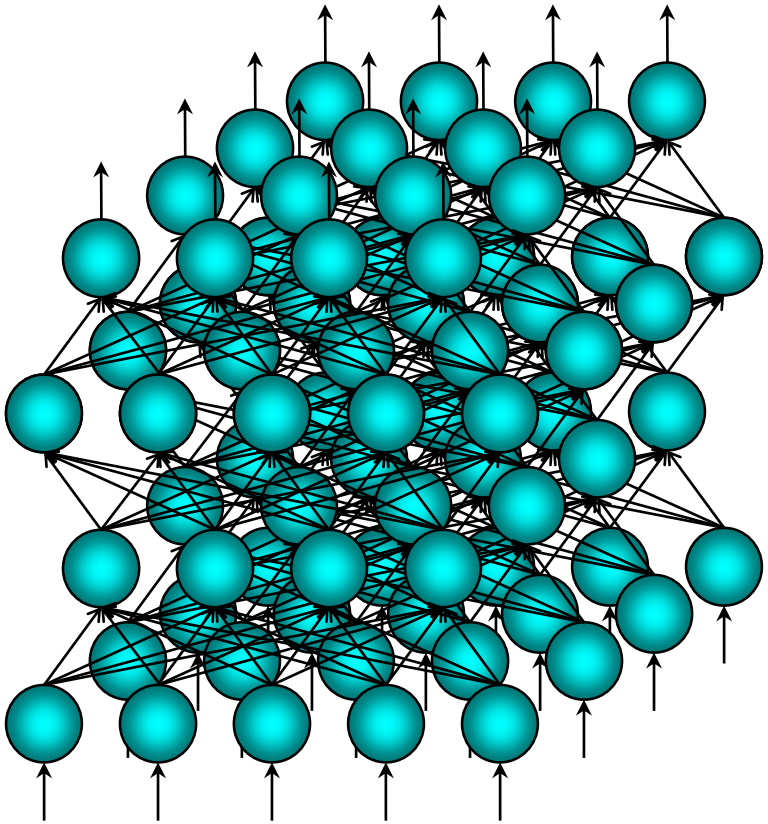
G, C, and W



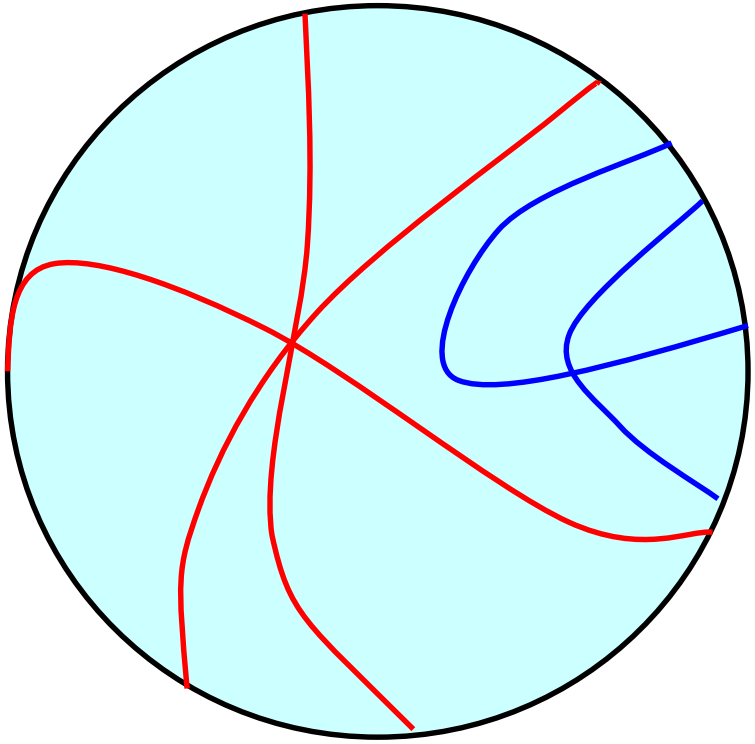
http://watanabe-www.math.dis.titech.ac.jp/users/swatanab/red_rank_reg.mp4

3 Phase Transition in Learning Process

Learning Machine and Parameter Set

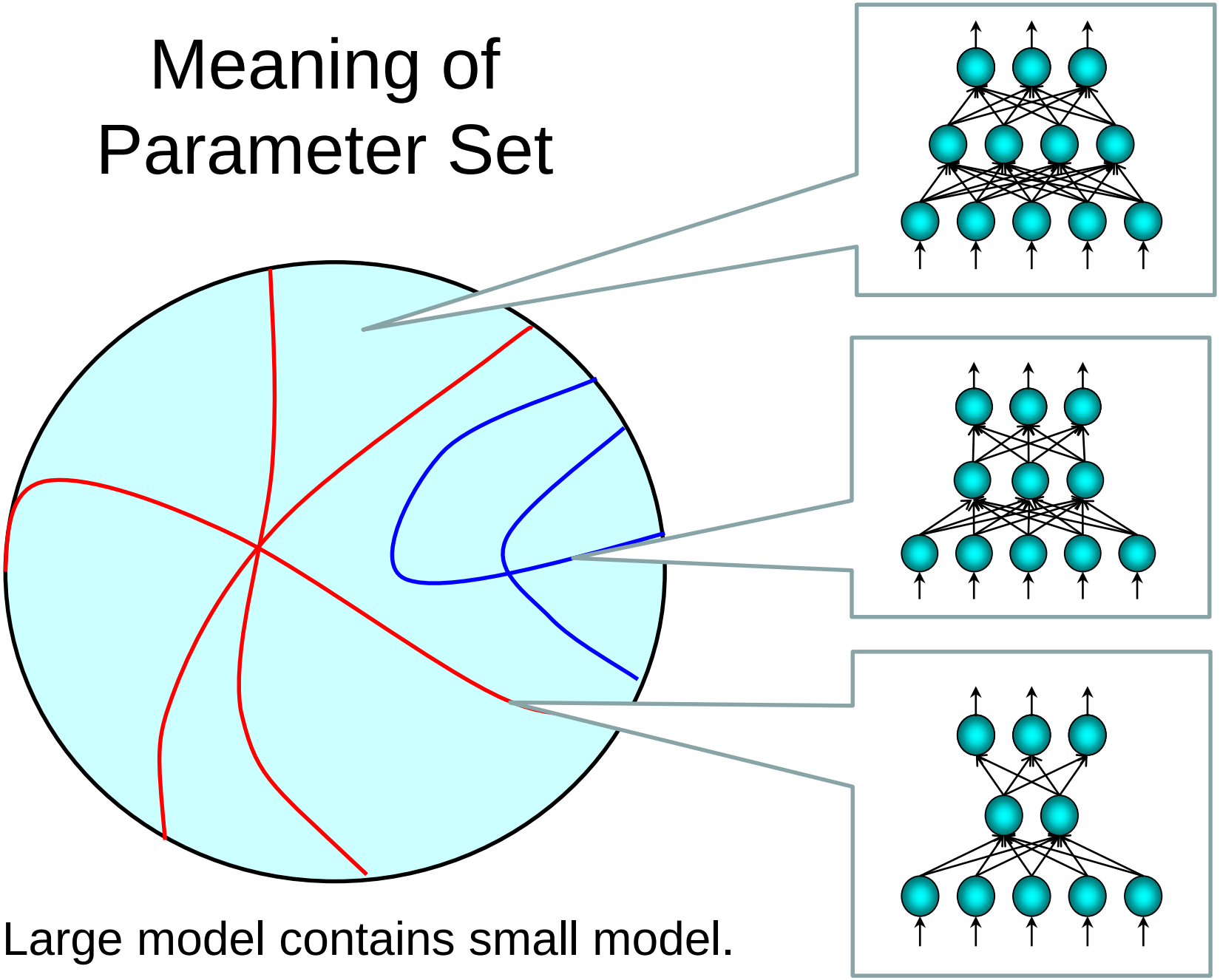


Statistical Model p(x|w)



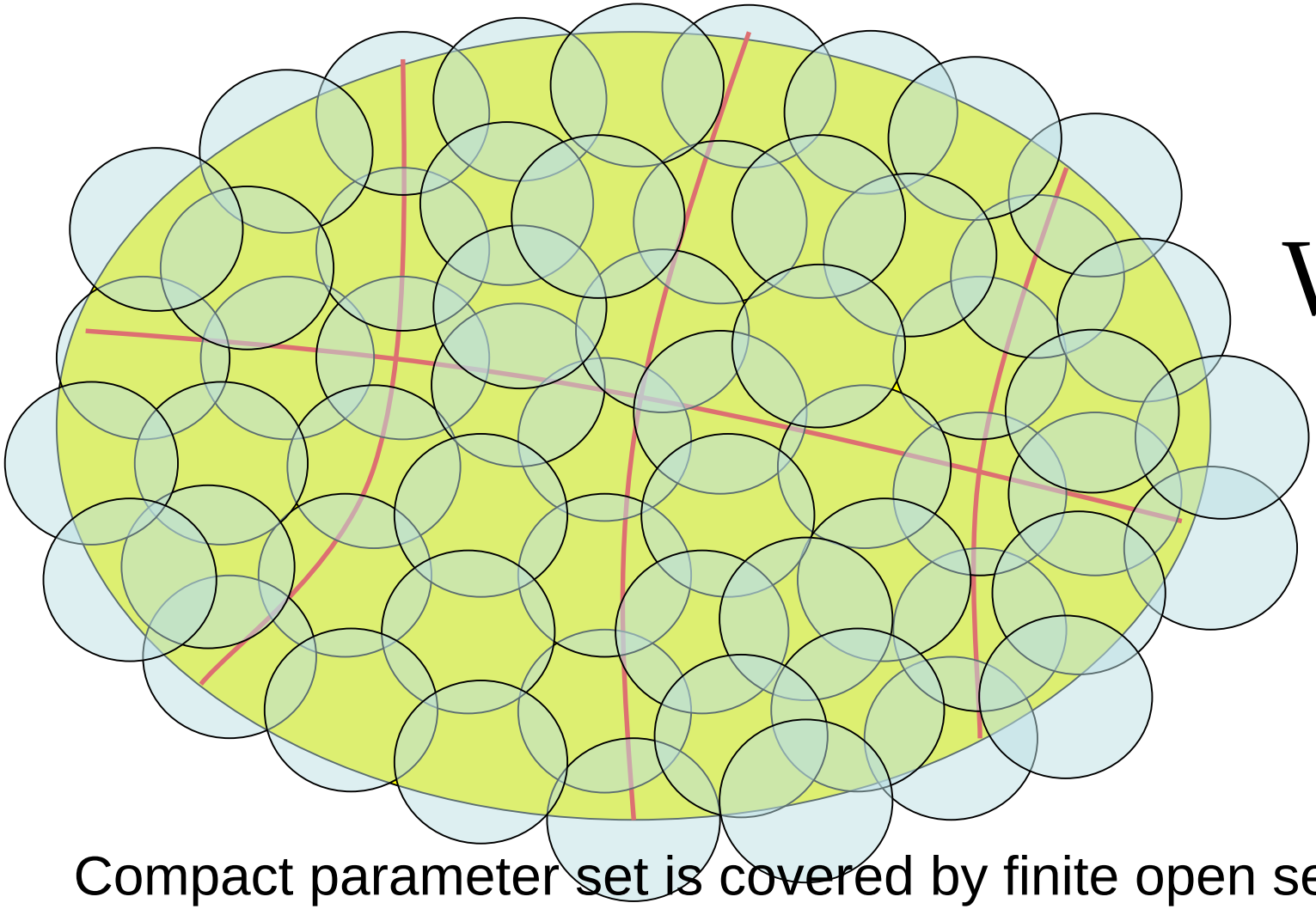
Parameter Set W

Meaning of Parameter Set



Large model contains small model.

Covering of the Parameter Set



Compact parameter set is covered by finite open set.

Decomposition of Free Energy

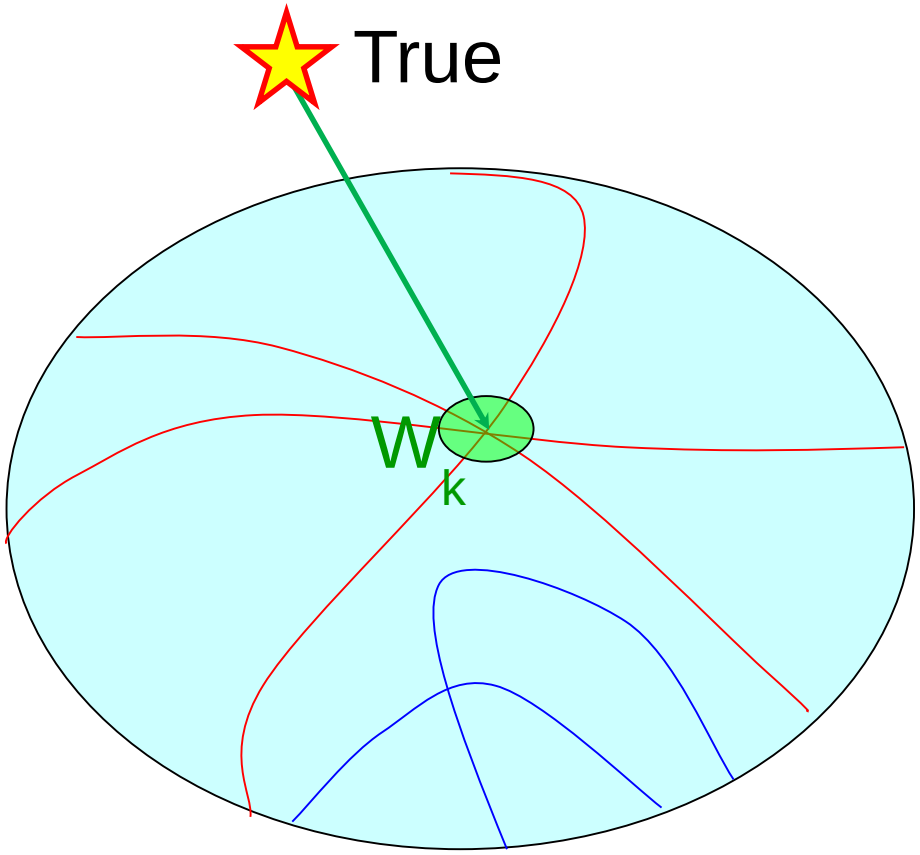
By using the partition of unity $\Sigma_k \varphi_k(w) = \varphi(w)$,

$$F_n = -\log \left(\int_W \prod p(X_i|w)\, \varphi(w)\, dw \right)$$

$$= -\log \left(\sum_k \int_{W_k} \prod p(X_i|w)\, \varphi_k(w)\, dw \right)$$

$$= -\log \left(\sum_k \exp(-F_{nk}) \right)$$

Local Free Energy



$\lambda_k \log n$: variance

$n(L_{nk} - S)$: Bias

$$\int_{W_k} \Pi_i \, p(X_i|w) \, \varphi_k(w) \, dw = (1/n)^{\lambda_k} \exp(- n (L_{nk}))$$

Local Free energy

Let λ_k and m_k be the local RCLT and its multiplicity of the coordinate k, then by using Hironaka theorem,

$$F_{nk} = -\log \int_{W_k} \Pi_i p(X_i|w)\, \varphi(w)\, dw$$

$$= -\log \{$$

$$\exp(-nL_{nk})\, (\log n)^{m_{k-1}}/n^{\lambda_k} \iint t^{\lambda-1} D(u) \exp(- t + t^{1/2}\xi_n(u))\, du\, dt \}$$

$$= nL_{nk} + \lambda_k \log n - (m_k-1) \log\log n + O_p(1).$$

Free Energy of Complex Model

$$F_n = -\log \left(\Sigma_k \exp(- nL_k - \lambda_k \log n \right)$$

$$= \min_k \left(n(L_k) + \lambda_k \log n \right) + \text{smaller}$$

The bias term $n(L_k - S)$ is small if model is complex.

The variance λ_k is small if model is simple.

For a given sample size n, the free energy is given by the smalllest $\{ n L_k + \lambda_k \log n \}$.

Phase Transition and Critical Point

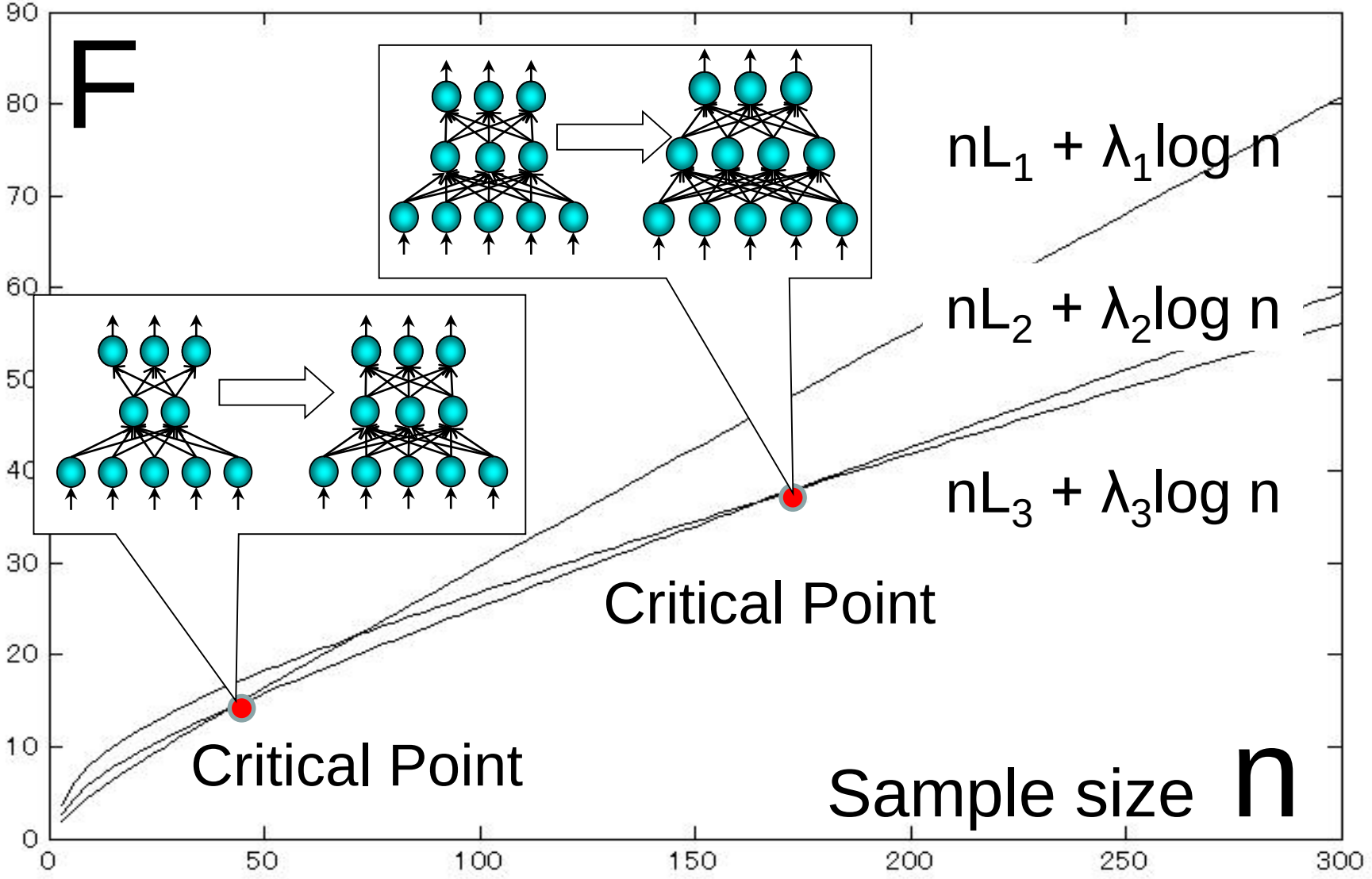
The integration of the local coordinate is in proportion to the probability of the local coordinate by the posterior distribution.

$$\mathrm{Prob}(W_k|X^n) = \frac{\int_{W_k} \Pi\ p(X_i|w)\,\varphi_k(w)\,dw}{Z_n}$$

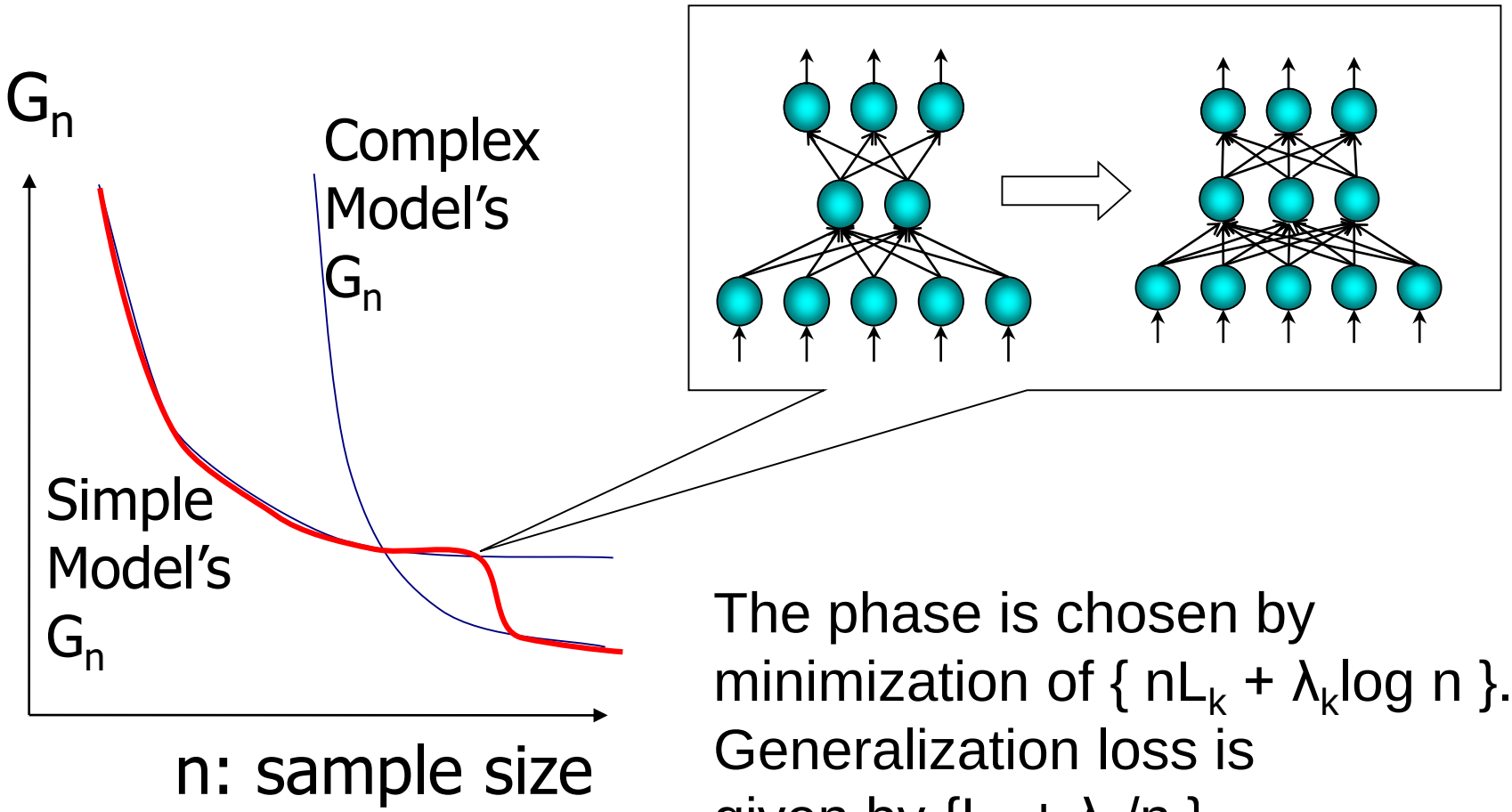
Hence the local coordinate that minimize the local free energy is chosen by the high probability.

As n is controlled from smaller to larger, the chosen parameter set changes, which is called phase transition, and the changing point is called the critical point.

Phase Transition and Critical Point



Behavior of Generalization Loss



The phase is chosen by minimization of $\{ nL_k + \lambda_k \log n \}$.
Generalization loss is given by $\{L_k + \lambda_k/n \}$

Learning Curve of Complex Model

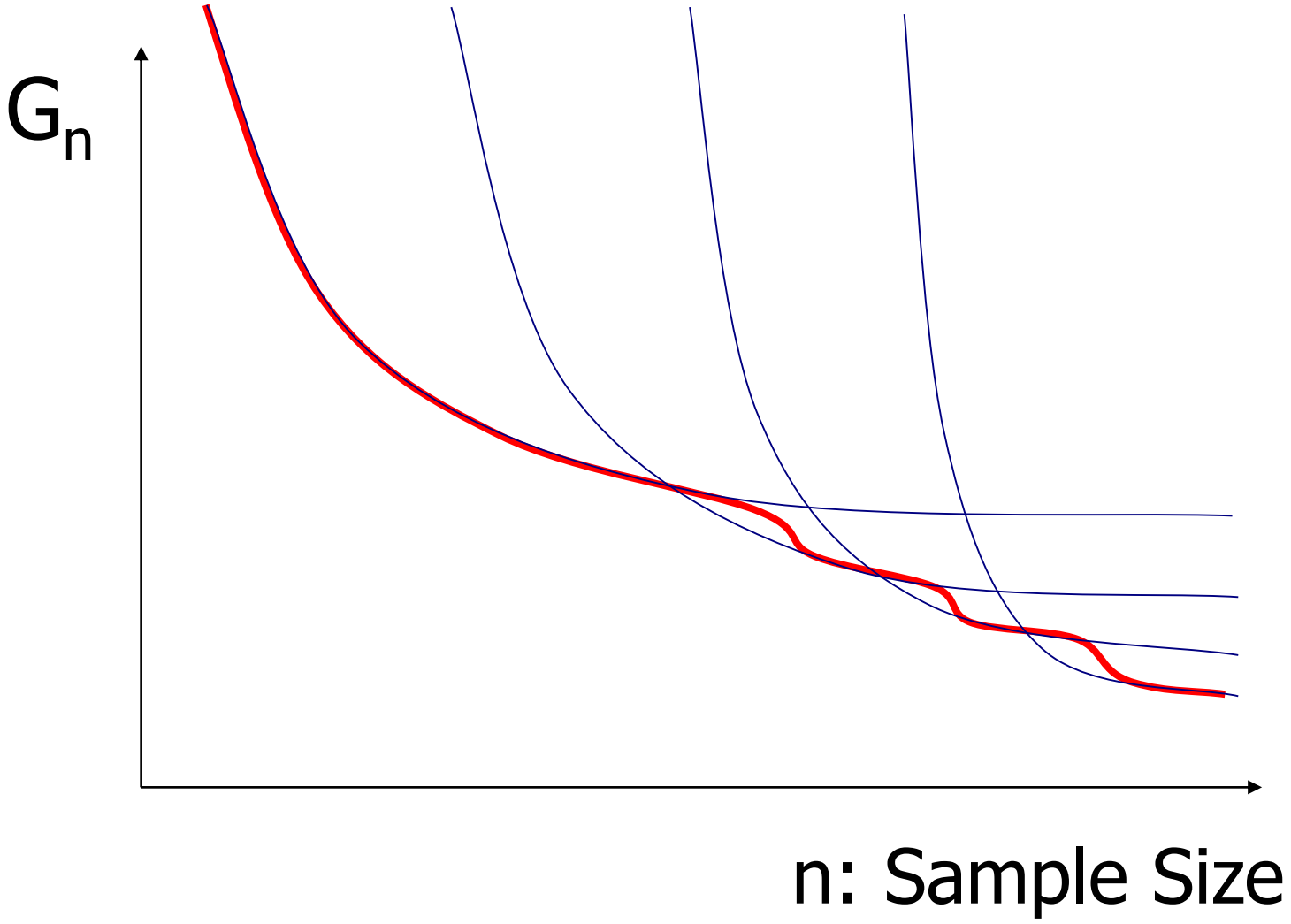
Theorem. The learning curve is given by

$$\mathbf{E}[G_n] = L_k + \lambda_k/n,$$

where $k = \text{argmin}\ \{ nL_k + \lambda_k \log n \}$.

Remark. The k that is chosen by the minimum free energy is smaller the model tha minimizes the generalization loss.

Learning Curve of Complex Model



Singularities and Phase Transition

Singularities determines the local free energy by its local RLCT and bias term.

The phase that maximizes the posterior probability is chosen by the posterior distribution, which minimizes the local free energy.

As n tends to infinity, the chosen model changes from simple to complex. This phenomenon is caused by singularities.

Appendix

Proof of Main Theorem

Applications to Learning Theory

We will use the following relations,

$$(1)\ \mathbf{E}_w[\, f(X,w)\,] = (1/n^{1/2}) < a(X,u) t^{1/2}\ |\xi_n > + o_p(1/n^{1/2}),$$

$$(2)\ \mathbf{E}_w[\, f(X,w)^2\,] = (1/n) < a(X,u)^2\ t\ |\xi_n > + o_p(1/n),$$

whose averages satisfy

$$(1)\ \mathbf{E}\mathbf{E}_X\mathbf{E}_w[\, f(X,w)\,] = (1/n)\, \mathbf{E}_\xi < t\ |\xi > + o(1/n),$$

$$(2)\ \mathbf{E}\mathbf{E}_X\mathbf{E}_w[\, f(X,w)^2\,] = (1/n)\, \mathbf{E}\mathbf{E}_X < a(X,u)^2\ t\ |\xi > + o(1/n).$$

Preparation of Proof

By using $p(x|w)=p(x|w_0)\exp(-f(x,w))$, losses are written as the following forms,

Generalization $\quad G_n = L_0 - \mathbf{E}_x[\, \log \mathbf{E}_w[\, \exp(\, -f(X,w)\,)]\,]$,

Training $\quad T_n = L_n - (1/n)\,\Sigma \log \mathbf{E}_w[\, \exp(\, -f(X_i,w))\,]$,

Cross Validation $\quad C_n = L_n + (1/n)\,\Sigma \log \mathbf{E}_w[\, \exp(\, f(X_i,w))\,]$,

WAIC $\quad W_n = T_n + (1/n)\,\Sigma\, \mathbf{V}_w[\, f(X_i,w)\,]$.

Cumulant Generating Function

Definition. Cumulant generating function is defined by

$$\Phi(\alpha) = - (1/n) \Sigma_{i=1}^{n} \log \mathbf{E}_w[p(X_i|w)^\alpha]$$

$$= \alpha L_n - (1/n) \Sigma \log \mathbf{E}_w[\exp(- \alpha f(X_i,w))].$$

It follows that

$$T_n = \Phi(1),$$

$$C_n = - \Phi(-1),$$

$$W_n = \Phi(1) - \Phi''(0).$$

Lemma 1 and Proof

Lemma.1 $< t \,|\xi> = \lambda + (1/2) < t^{1/2}\xi \,|\xi>$.

(Proof) By the definition,

$$< t \mid \xi > = \frac{\Sigma \int\int t\, t^{\lambda-1} D(u) \exp(- t + t^{1/2}\xi(u)) du dt}{\Sigma \int\int t^{\lambda-1} D(u) \exp(- t + t^{1/2}\xi(u)) du dt} .$$

By using the partial integration,

$$\int \exp(-t)\, t^{\lambda} \exp(t^{1/2}\xi(u)) dt = [\, (-\exp(-t))\, t^{\lambda} \exp(t^{1/2}\xi(u)) \,]_0^{\text{infinity}}$$

$$+ \int \exp(- t)\, t^{\lambda-1} \exp(t^{1/2}\xi(u)) \{ \lambda + (1/2) t^{1/2}\xi(u) \} dt.$$

Applying this to $< t \,|\xi>$. (Lemma 1, Q.E.D.)

Lemma.2 and Proof

Lemma.2 $\mathbf{E}_\xi \langle t^{1/2}\xi \,|\xi\rangle = 2\nu$.

From Theorem.1, the following equality is derived,

$\lambda = \mathbf{E}_\xi \langle t|\xi\rangle - (1/2)\, \mathbf{E}_\xi \mathbf{E}_X \{ \langle a(X,u)^2 t\,|\xi\rangle - \langle a(X,u) t^{1/2}\,|\xi\rangle^2 \}$.

By Lemma.1, $\mathbf{E}_\xi \langle t\,|\xi\rangle = \lambda + (1/2)\, \mathbf{E}_\xi \langle t^{1/2}\xi\,|\xi\rangle$,

hence

$\mathbf{E}_\xi \langle t^{1/2}\xi\,|\xi\rangle = \mathbf{E}_\xi \mathbf{E}_X \{ \langle a(X,u)^2 t\,|\xi\rangle - \langle a(X,u) t^{1/2}\,|\xi\rangle^2 \}$,

which is equal to 2ν. (Lemma 2, Q.E.D.)

Proof of main Theorem (1)

The equations about the generalization loss and its expectation were proved in Theorem 1. Here we prove the other equations.

By using $\Phi(\alpha)= \alpha L_n - (1/n) \Sigma \log E_w[\exp(-\alpha f(X_i,w))]$,

$T_n = \Phi(1) = \Phi(0) + \Phi'(0) + \Phi''(0)/2 + O_p(1/n^{3/2})$,

$C_n = -\Phi(-1) = -\Phi(0) + \Phi'(0) - \Phi''(0)/2 + O_p(1/n^{3/2})$,

$W_n = \Phi(1) - \Phi''(0) = \Phi(0) + \Phi'(0) - \Phi''(0)/2 + O_p(1/n^{3/2})$.

Note that $\Phi(0)= 0$.

Proof of Main Theorem (2)

By using $\Phi(\alpha)= \alpha L_n - (1/n)\,\Sigma \log E_w[\exp(-\alpha f(X_i,w))]$,

$$\Phi'(0) = L_n + \mathbf{E}_w[(1/n)\,\Sigma f(X_i,w)] = L_n + \mathbf{E}_w[K_n(w)]$$

$$= L_n + \mathbf{E}_u[u^{2k} - u^k \xi_n(u)/n^{1/2}]$$

$$= L_n + (1/n)\langle t - t^{1/2}\xi_n \rangle = L_n + \lambda - (1/2n)\langle t^{1/2}\xi_n \rangle,$$

where we used Lemma.1.

$$\Phi''(0) = -(1/n)\,\Sigma\, \mathbf{V}_w[f(X_i,w)]$$

$$= -(1/n)\,\Sigma\,\{\mathbf{E}_u[(a(X_i,u)\,u^k)^2] - \mathbf{E}_u[a(X_i,u)\,u^k]^2\}$$

$$= -\mathrm{Fluc}(\xi_n)/n + o_p(1/n)\ ,$$

which shows the first half of the theorem. The latter half is proved by their expectations and Lemma.2. （Q.E.D.）