# Cross Validation and WAIC in Layered Neural Networks

Deep learning : Theory, Algorithms, and Applications

2018 March 19th-22nd, Tokyo, Riken AIP.

Sumio Watanabe

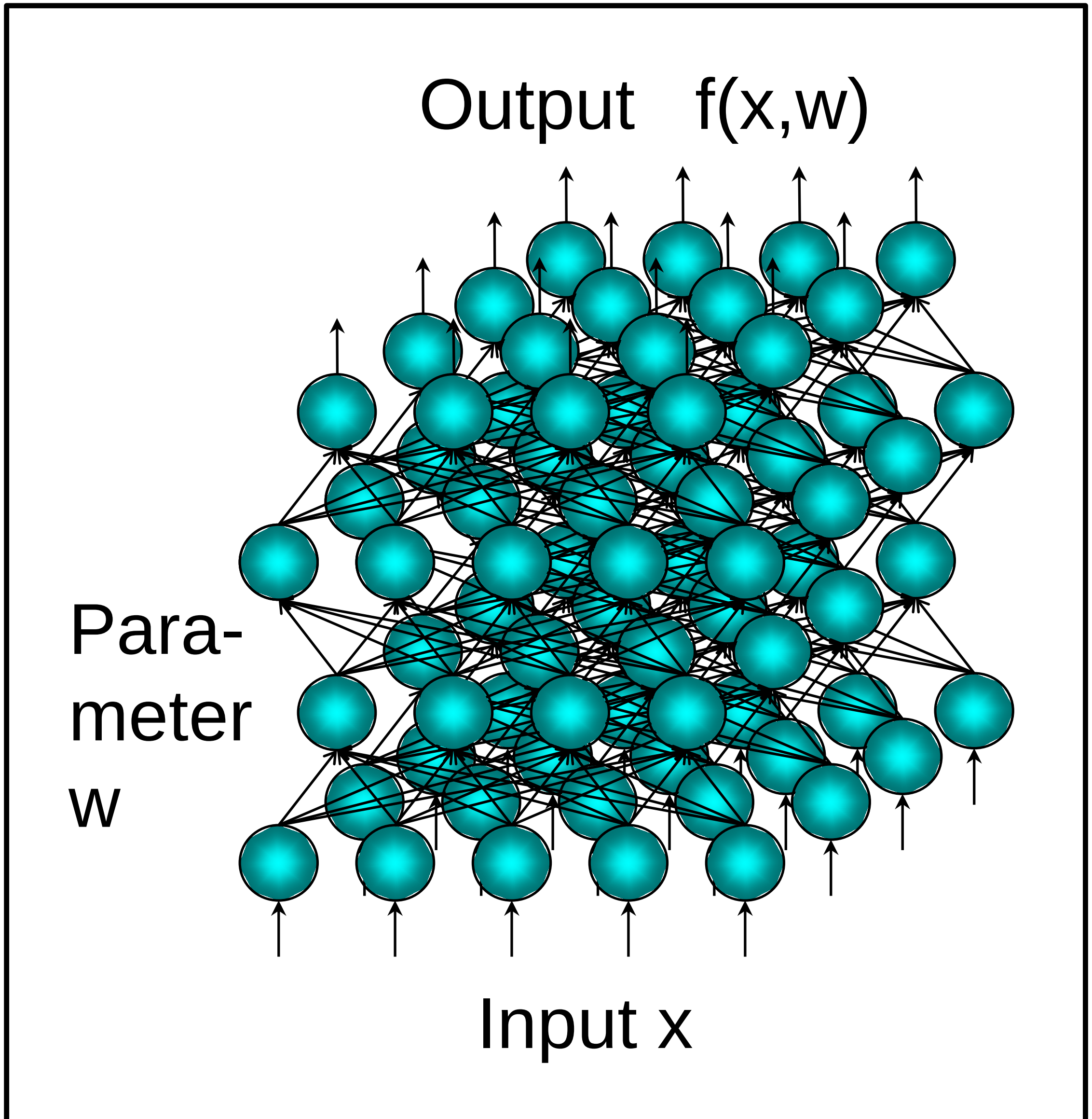
Tokyo Institute of Technology

# *CONTENTS*

## 1 Posterior of NN is highly singular

*Let's see the true posterior.*

# Layered Neural Network is Nonidentifiable

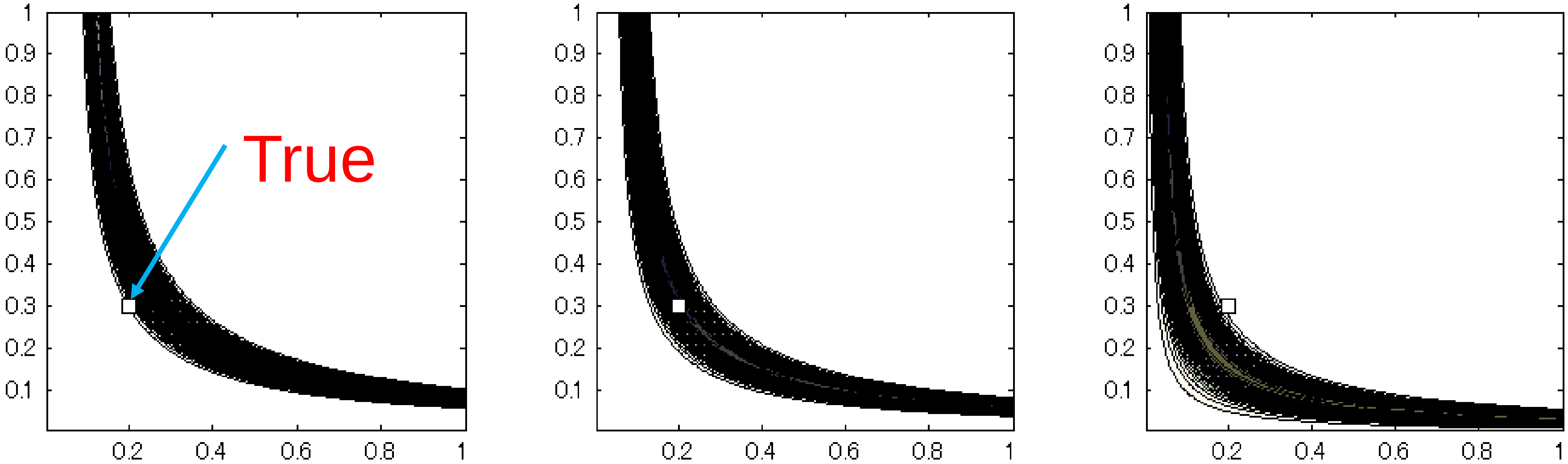



$w \mapsto f( \quad ,w)$
is not injective

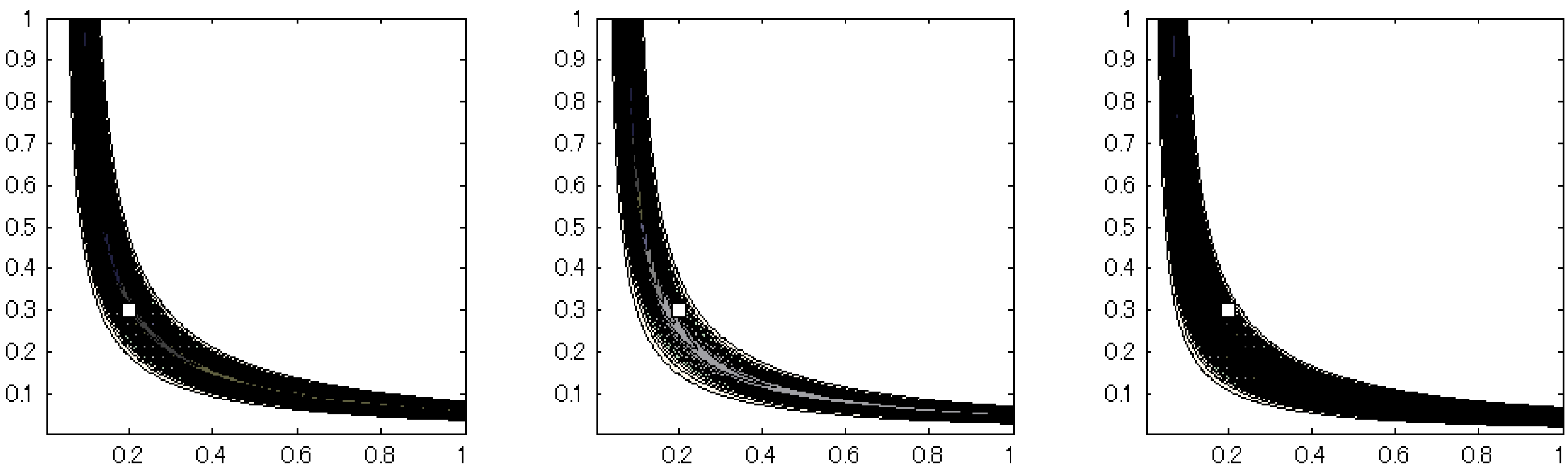
$\{ \partial_{w_j} f(x,w) \}$ is linearly dependent

Mathematical method was not established.

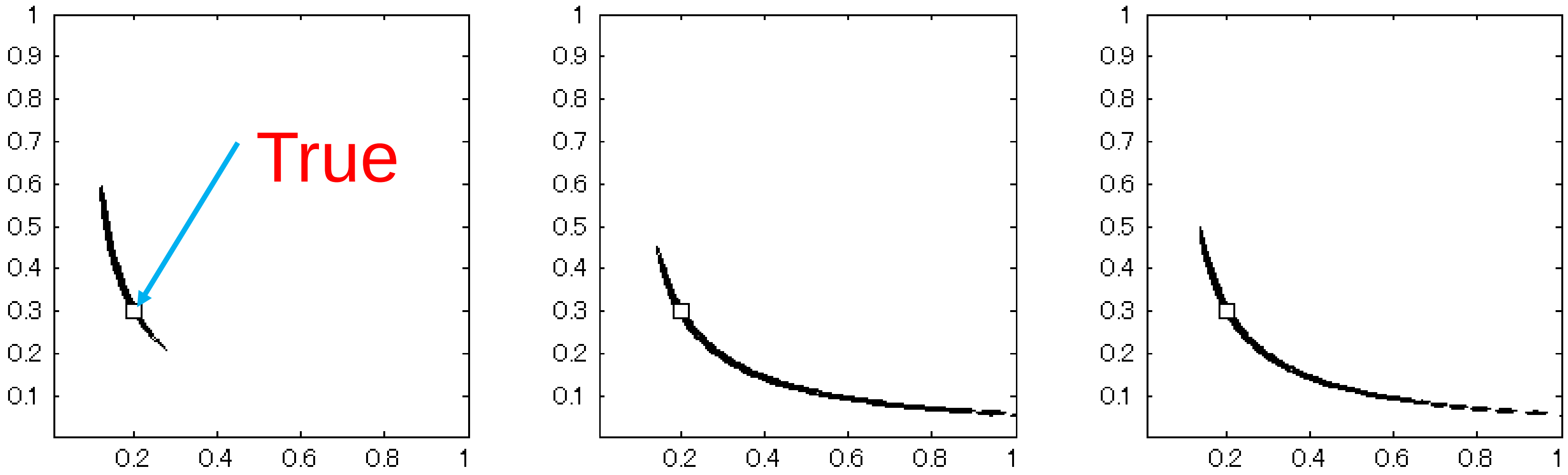
# Posterior of $(y\text{-}a\tanh(bx))^2$ for n=100



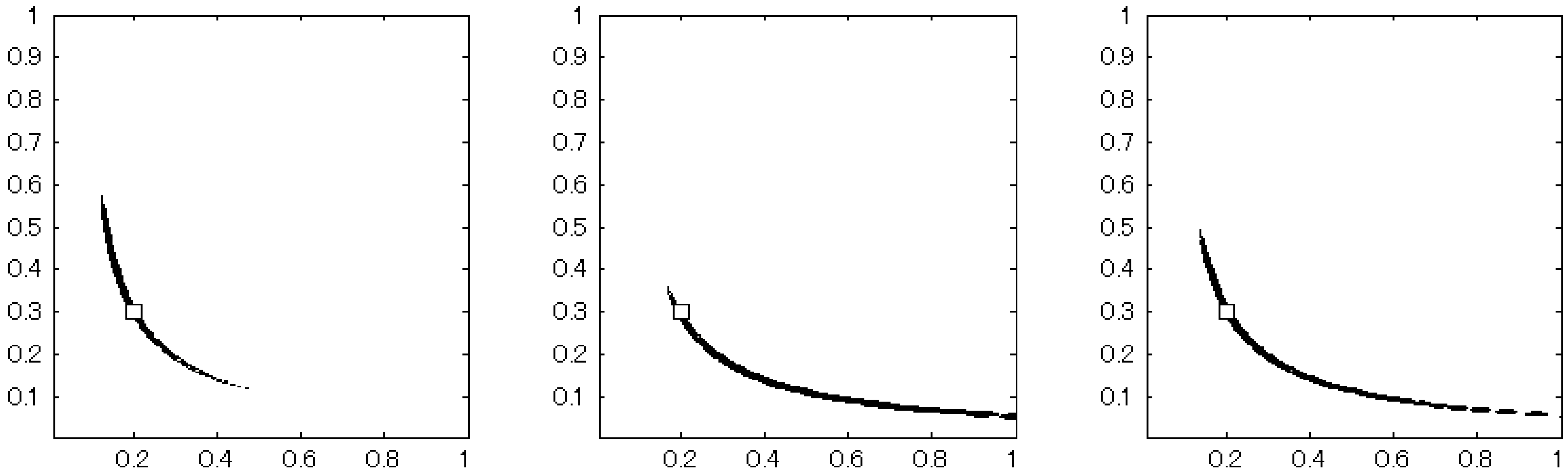

*Even if the true is regular, the posterior is singular.*



# Posterior of $(y - a\tanh(bx))^2$ for n=10000




*Even for n=10000, the posterior is singular.*



# 2 Bayesian Learning

*For singular learning machines, Bayes makes the generalization loss smaller.*

# Bayesian learning

(1) $\{X_i, Y_i \,;\, i=1,2,\ldots n\} \sim q(x)q(y|x)$

(2) Learning machine $p(y|x,w)$

(3) Prior $\varphi(w)$

In a regression case, $p(y|x,w) \propto \exp(\, -C(y-f(x,w))^2\, )$

Minus log likelihood

$$H(w) = -\Sigma \log p(Y_i|X_i,w)$$

# Posterior and Predictive

Posterior

$$E_w[\quad] = \frac{\int (\quad)\, \exp(-H(w))\, \varphi(w)\, dw}{\int \exp(-H(w))\, \varphi(w)\, dw}$$

Predictive $p^*(y|x) = E_w[\, p(y|x,w)\, ]$

estimates

True $q(y|x)$

# Training and Generalization Losses

Generalization Loss $G = -E_{(X,Y)}[\log p^*(Y|X)]$

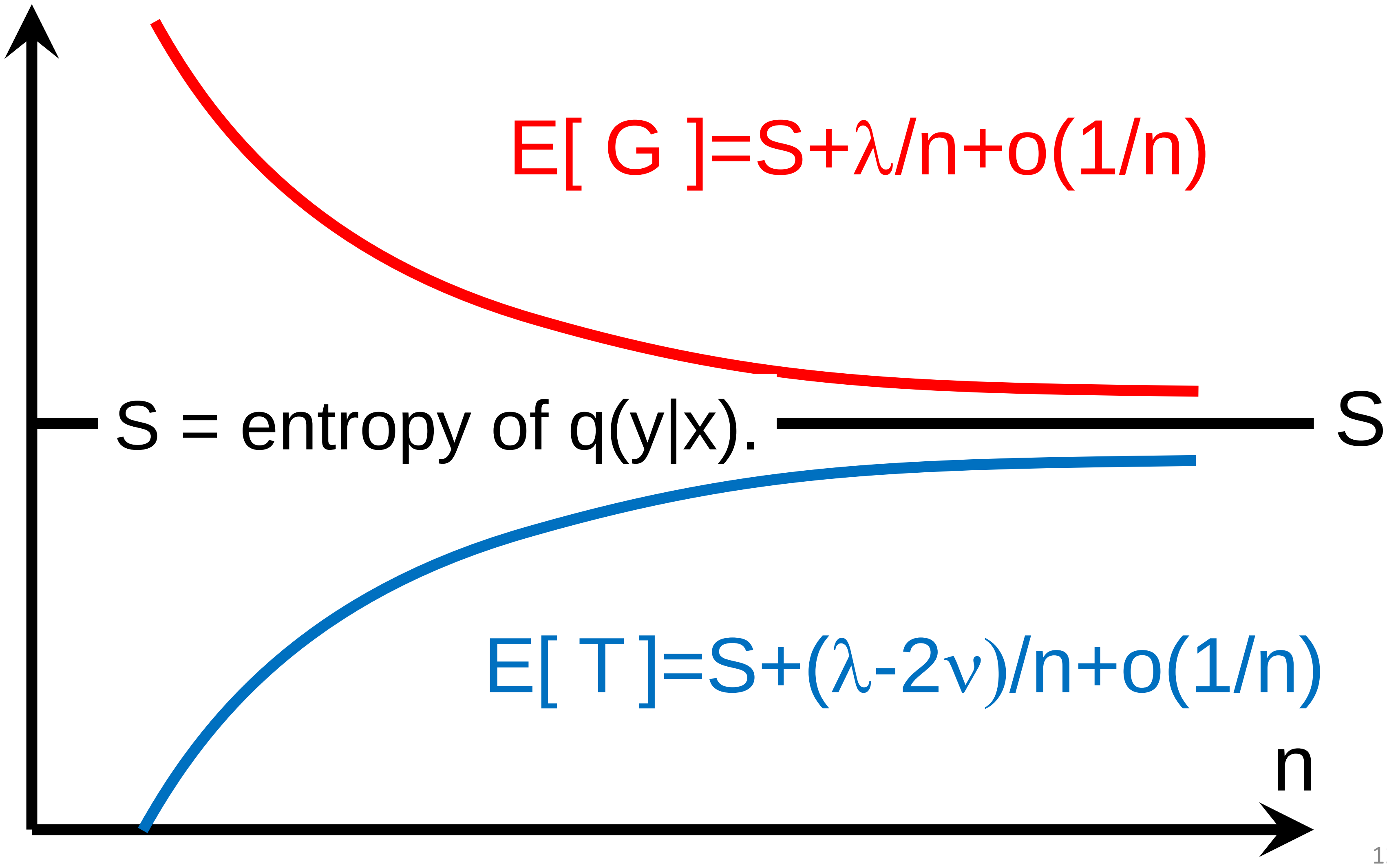
Training Loss $T = -(1/n) \sum_{i=1}^{n} \log p^*(Y_i | X_i)$

If q(y|x) is realizable by p(y|x,w), then G and T converge to S (entropy of the true).

# 3 Learning Curve is Given by Birational Invariants

*To study singular learning machines, algebraic geometry is necessary.*

# Learning Curves are given by Algebraic Geometry



# Birational Invariants

$\lambda$ and $\nu$ are birational invariants.

$\lambda$ is the real log canonical threshold.

$\nu$ is the singular fluctuation.

Cf. If $\{ \partial_{w_j} f(x,w) \}$ is linearly independent, then

$\lambda = \nu = d/2$, where d is the dimension of w.

# Cross Validation

Theorem (Gelfand 1998). Importance sampling CV.

$$C = (1/n) \Sigma_i \log E_w[\ 1/p(Y_i|X_i,w)\ ]$$

$$E[G] = E[\ C\ ] + O(1/n^2)$$

Epifani (2008) proved that, if a leverage sample point is contained, then $E_w[\ 1/p\ ]$ does not exist.

Leverage sample point : a sample point that affects the statistical estimation result strongly.

Vehtari and Gelman (2015) proposed approximation of importance by Pareto distribution.

# Information Criterion

Theorem. Widely Applicable Information Criterion

$$W = T + (1/n)\ \Sigma_i\ V_w[\ \log p(Y_i|X_i,w)\ ]$$

$$E[G] = E[\ W\ ] + O(1/n^2)$$

Cf. This is a generalized version of AIC.
If $\{\ \partial_{w_j} f(x,w)\ \}$ are linearly independent,

$$E[G] = E[\ T\ ] + d/n + o(1/n)$$

In this case CV and WAIC are equivalent in higher order $(1/n^2)$ (2015).

# Cross Validation and Information Criteria

Cross validation requires that $\{X_i, Y_i\}$ is independent.

AIC and WAIC do that $\{Y_i|X_i\}$ is independent.

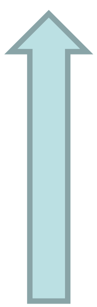
# 4 Generalization Loss can be Estimated by CV and WAIC.

# Estimation of Generalization Loss

True: $x=(x_1,x_2)$

$g(x) = \exp( -x_1^2-x_2^2-x_1x_2)$

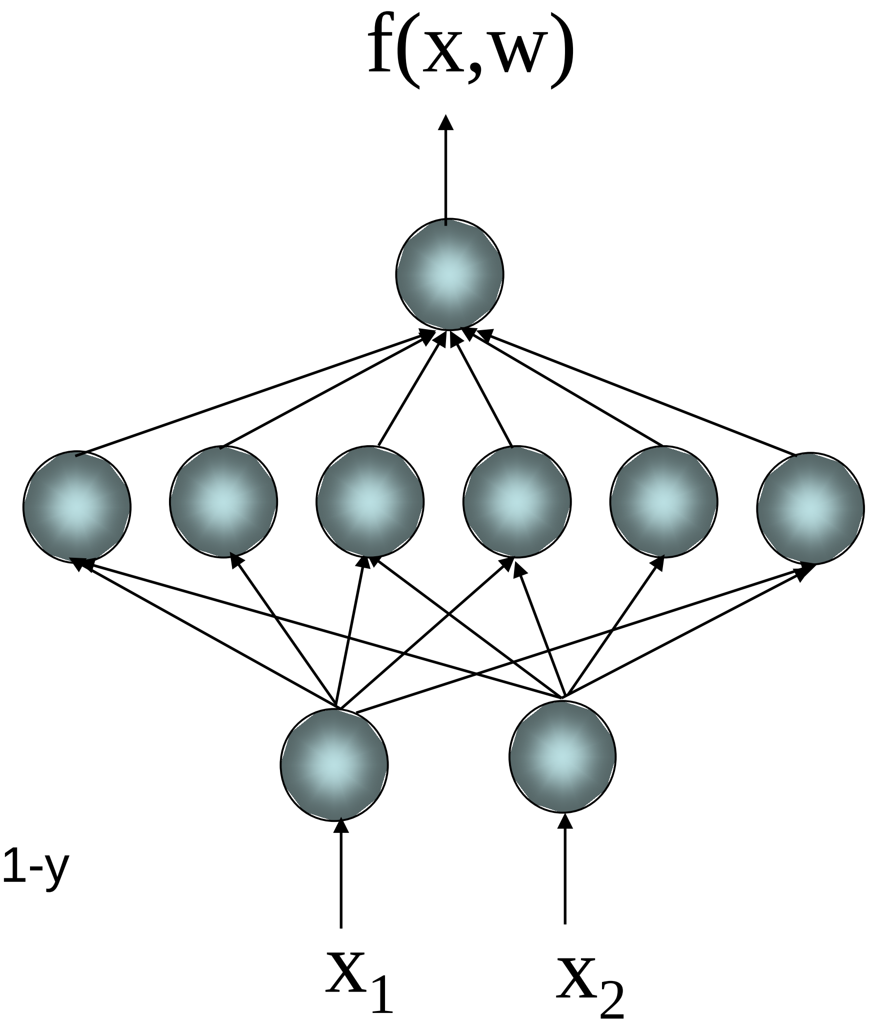
$q(y|x) = g(x)^y (1-g(x))^{1-y}$



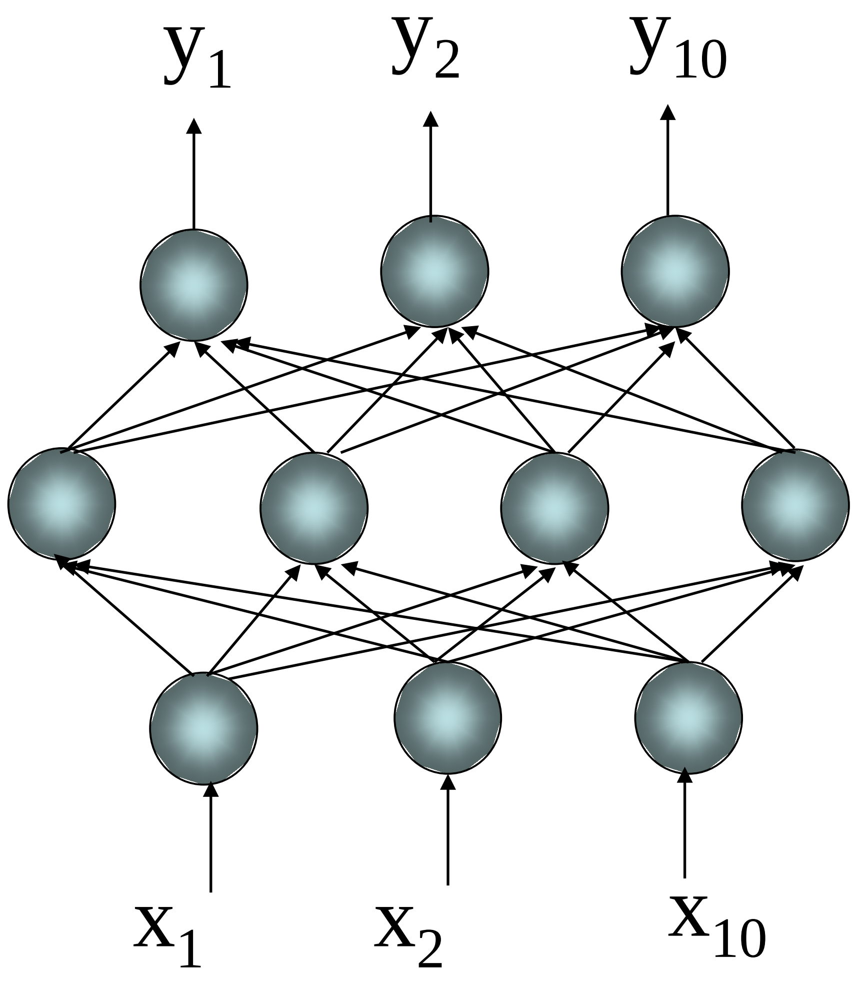
Learner :

$f(x,w)$ : Neural Network

$p(y|x,w) = f(x,w)^y (1-f(x,w))^{1-y}$



# Model Selection




True:

$10 \to 5 \to 10$

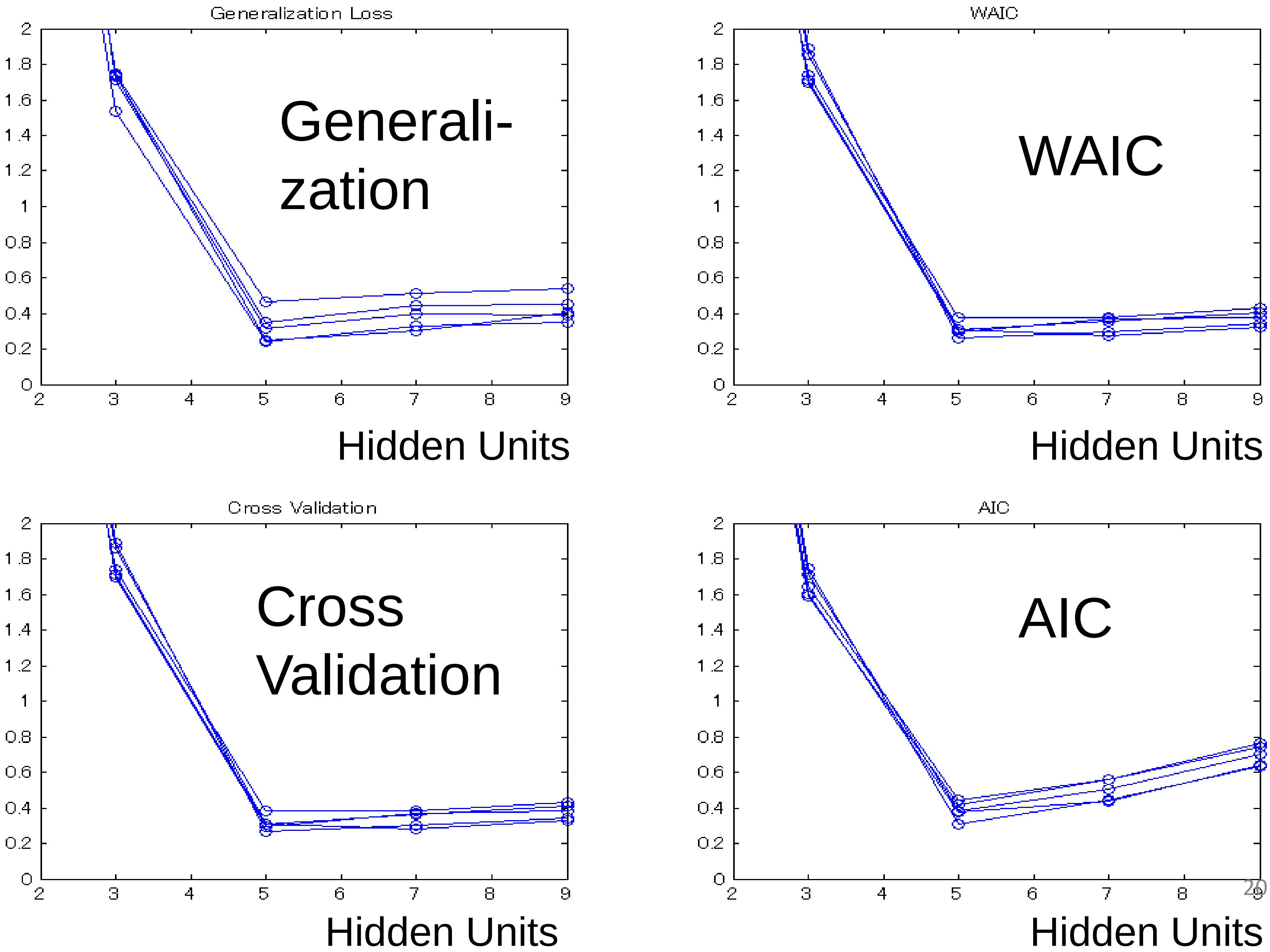
Candidates:

$10 \to (1, 3, 5, 7, 9)$
$\to 10$

n =200
n_test=1000

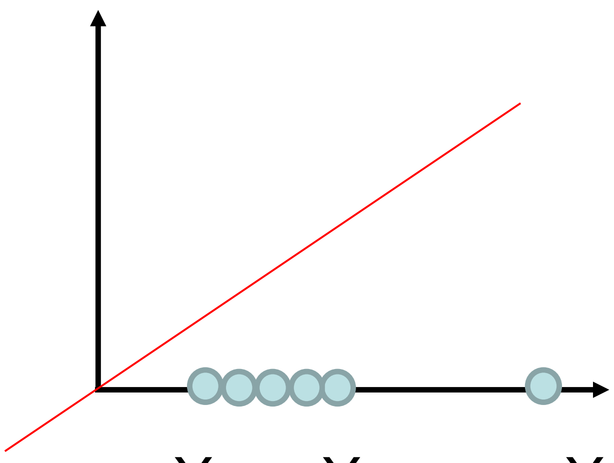
Posterior was approximated by Langevin equation.

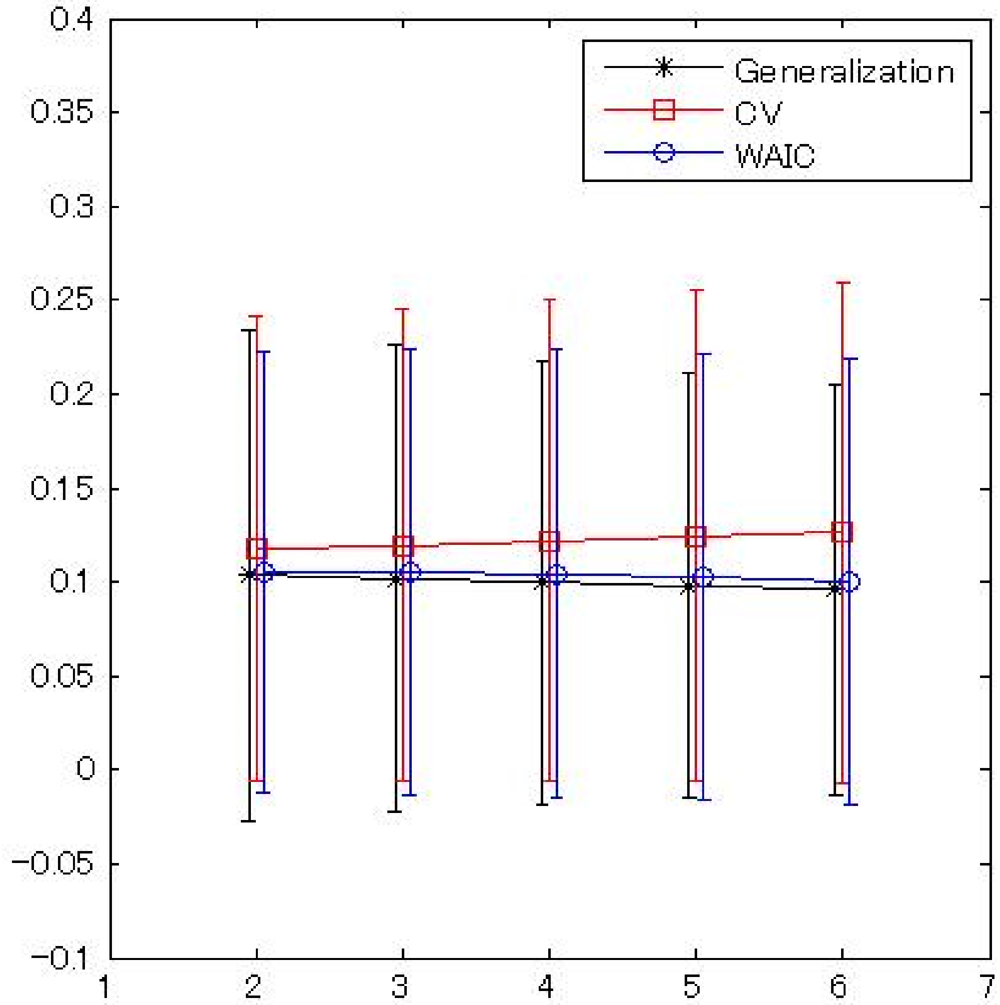
# An experiment: Random 10 trials



# Difference between CV and WAIC in Regression.

A leverage sample point was controlled. WAIC and CV were compared with the generalization loss.




Place of a Leverage point $X_{10}$ .

# Conclusion

(1) Posterior of NN is singular. Learning curves are given by birational invariants.

(2) Generalization losses are estimated by cross validation and WAIC.

# Future Study

To construct MCMC for large networks.