2009 International Symposium on Nonlinear Theory and its Applications
NOLTA'09, Sapporo, Japan, October 18-21, 2009

Phase Transition of Generalization Errors in Variational Bayes Learning

Shinji Oyama† and Sumio Watanabe‡

†Department of Computational Intelligence and Systems Science
‡Precision and Intelligence Laboratory
Tokyo Institute of Technology
Mailbox R2-5, 4259 Nagatsuta, Midori-ku, Yokohama, 226-8503, Japan
Email: s.oyama@cs.pi.titech.ac.jp

Abstract— In a mixture of exponential probability distributions, the mean field approximation is used in wide applications because it provides the fast learning algorithm. The mean field approximation in Bayesian learning is called the variational Bayes method. It was clarified by Kazuho Watanabe et. al. that there exists the phase transition of the variational free energy with respect to the hyperparameter. In this paper, we study the generalization errors of variational Bayes learning, and experimentally show the following facts. (1) The generalization error also has the phase transition. (2) At ordinary point, the generalization error strongly depends on the condition that the true distribution is contained in the learning machine or not, whereas, at the critical point, the generalization error does not depend on the condition.

1. Introduction

The variational Bayes learning is applied to a lot of information systems because it gives us the fast training algorithm by the mean field approxiamtion of Bayes *a posteriori* distribution. It is important to clarify the generalization performance of the variational Bayes in order to establish the optimal design method. However, almost all learning machines to which we can apply the varitational Bayes are not regular but singular statistical models, resulting that the conventional statistical theory can not be employed.

In Bayes learning, the mathematical foundation was clarified [10]. The free energy F is asymptotically equal to

$$F = S_n + \lambda \log n - (m-1)\log\log n + R,$$

where n is the number of training samples, S_n is the empirical entrpoy, and R is a random variable. The constants λ and m are determined by the zeta function of the learning theory. Let G be the generalization error which is defined by the Kullback-Leibler distance from the true distribution to the estimated distriution. The generalization error of Bayes learning is equal to

$$G = \frac{\lambda}{n} + o(\frac{1}{n}).$$

On the other hand, neither the free energy nor the generalization error was clarified in the vatiational Bayes learning. The variational Bayes free energy $\hat{F}$ satifies the inequality,

$$F \leq \hat{F}.$$

Moreover, it was clarified in [7, 8] that $\hat{F}$ has a phase transition with respect to the hyperpramater of the *a priori* distribution at $\phi_0 = (M+1)/2$, where M be the dimension of the data. However, in the variational Bayesian learning, the behavior of the generalization error has not yet been clarified.

In this paper, we experimentally show that the variational generalization error $\hat{G}$ also has the phase transition at $\phi_0 = (M+1)/2$, and that, if the learning machine is more complex than the true distribution, then the generalization error is largest at the critical point, whereas, if otherwise, the generalization error is smallest at the critical point.

2. Variational Bayes Learning

In this paper, we study a normal mixture on M dimensional Euclidean space $\mathbf{R}^M$,

$$p(x|w) = \sum_{k=1}^{K} \frac{a_k}{\sqrt{2\pi}^M} \exp(-\frac{1}{2}\|x - b_k\|^2), \qquad (1)$$

where K is the number of mixtured normal distributions. The parameter w is given by

$$w = (a, b) = \{(a_k, b_k); k = 1, 2, ..., K\},$$

which satisfies

$$0 \leq a_k \leq 1, \quad a_1 + a_2 + \cdots + a_K = 1$$

and $b_k \in \mathbf{R}^M$. In variational Bayesian learning, the conjugate prior is employed.

$$\begin{aligned}
\varphi(w) &= \varphi_1(a)\varphi_2(b), \\
\varphi_1(a) &= \frac{\Gamma(K\phi_0)}{\Gamma(\phi_0)^K}\delta\Big(\sum_{k=1}^{K} a_k - 1\Big)\prod_{k=1}^{K} a_k^{\phi_0 - 1}, \\
\varphi_2(b) &= \Big(\frac{\beta_0}{2\pi}\Big)^{KM/2}\prod_{k=1}^{K}\exp(-\frac{\beta_0}{2}\|b_k - b_0\|^2),
\end{aligned}$$

where $\phi_0 > 0$, $\beta_0 > 0$, and $b_0 \in \mathbf{R}^M$ are hyperparameters. The hyperparameter ϕ_0 is important because it determines

the balance of the mixture ratio. In this paper we study the phase transition of generalization error with respect to the hyperparameter ϕ_0.

Let $Y = (Y^1, Y^2, ..., Y^K)$ be the competitive hidden variable. That is to say, Y is a random variable which is defined on the set,

$$C = \{(1,0,0,...,0),\ (0,1,0,...,0),\ ...,\ (0,0,0,...,1)\}.$$

In other words, only one of Y^k is chosen to be one then the others are equal to zero. The simultaneous distribition of (X, Y) is defined by

$$p(x,y|w) = \prod_{k=1}^{K} \Big(\frac{a_k}{\sqrt{2\pi}^M} \exp(-\frac{\|x-b_k\|^2}{2}) \Big)^{y^k},$$

where $y = (y^1, y^2, ..., y^K) \in C$. Note that the marginal distribution of $p(x, y|w)$ is equal to $p(x|w)$,

$$p(x|w) = \sum_{y \in C} p(x,y|w),$$

where $\sum_{y \in C}$ shows the sum over C. Therefore, the learning machine $p(x|w)$ can be understood to be marginalized $p(x, y|w)$, where y is a hidden variable.

Let d_n and h_n be the set of training samples and the corresponding hidden variables respectively,

$$\begin{aligned} d_n &= \{x_i \in \mathbf{R}^M; i = 1,2,...,n\}, \\ h_n &= \{y_i \in C; i = 1,2,...,n,\}, \end{aligned}$$

where $d_n \in \mathbf{R}^{Mn}$ and $h_n \in C^n$. Then the simultaneous probability density function on (d_n, h_n, w) is equal to

$$P(d_n, h_n, w) = \varphi(w) \prod_{i=1}^{n} p(x_i, y_i|w). \tag{2}$$

For the given set of training samples d_n, the probability distribution on (h_n, w) is equal to

$$P(h_n, w|d_n) = \frac{1}{Z_n} P(d_n, h_n, w),$$

where Z_n is the partition function defined by

$$\begin{aligned} Z_n &= \sum_{h_n \in C^n} \int dw\ P(d_n, h_n, w) \\ &= \int dw\ \varphi(w) \prod_{i=1}^{n} p(x_i|w). \end{aligned}$$

The partition function is called the evidence or the marginal likelihood of $p(x|w)$ and $\varphi(w)$.

In the variational Bayesian learning, the probability distribution $P(h_n, w|d_n)$ is approximated by $q(h_n)r(w)$. The probability distributions $q(h_n)$ and $r(w)$ are optimized by the minimization of the Kullback-Leibler distance between them,

$$\mathcal{K}(q,r) = \sum_{h_n \in C^n} \int dw\ q(h_n) r(w) \log \frac{q(h_n) r(w)}{P(h_n, w|d_n)}.$$

Let S be the set of independent probability distributions on $C^n \times \mathbf{R}^d$,

$$\mathcal{S} = \{q(h_n) r(w)\}.$$

The probability distribution $P(h_n, w|d_n)$ is not contained in $\mathcal{S}$ in general, however, the optimal distribution is found in this set. The minimization of $\mathcal{K}(q, r)$ is equivalent to the minimization of the functional free energy,

$$\mathcal{F}(q,r) = \sum_{h_n \in C^n} \int dw\ q(h_n) r(w) \log \frac{q(h_n) r(w)}{P(h_n, w, d_n)}.$$

If $P(h_n, w|d_n)$ is contained in the set $\mathcal{S}$, then the minimum value of $\mathcal{F}(q, r)$ is equal to the true free energy $-\log Z_n$. The variational free energy is defined by

$$\hat{F} \equiv \min_{(q,r) \in \mathcal{S}} \mathcal{F}(q,r).$$

By the definition, the variational free energy is not smaller than the true free energy,

$$\hat{F} \geq F.$$

In variational Bayes learning, it is shown that $q(h_n)$ and $r(w)$ should satisfy the relations,

$$q(h_n) = \frac{1}{C_1} \exp\Big(E_r[\log P(d_n, h_n, w)] \Big), \tag{3}$$

$$r(w) = \frac{1}{C_2} \exp\Big(E_q[\log P(d_n, h_n, w)] \Big), \tag{4}$$

where $E_r[\ \ ]$ and $E_q[\ \ ]$ are expectations over $r(w)$ and $q(h_n)$, respectively. Here $C_1, C_2 > 0$ are normalizing constants. The learning algorithm of the variational Bayes learning is defined by the recursive procedure of these two equations. If the expectation value of the hidden varibale $\overline{y}_i^k = E_q[y_i^k]$ is given, then

$$r(a,b) \propto \prod_{k=1}^{K} (a_k)^{S_k - 1} \exp(-\frac{T_k}{2} \|b_k - U_k\|^2),$$

where $S_k, T_k\ U_k$ are defined by

$$\begin{aligned} S_k &= \sum_{i=1}^{n} \overline{y}_i^k + \phi_0, \\ T_k &= \sum_{i=1}^{n} \overline{y}_i^k + \beta_0, \\ U_k &= \frac{1}{T_k} \{ \sum_{i=1}^{n} \overline{y}_i^k x_i + \beta_0 b_0 \}. \end{aligned}$$

Conversely, if the foregoing values are given, then

$$q(h_n) \propto \prod_{i=1}^{n} \prod_{k=1}^{K} \exp(y_i^k L_i^k),$$

where

$$L_i^k = \psi(S_k) - \frac{1}{2}\{\frac{M}{T_k} + \|x_i - U_k\|^2\},$$
$$\bar{y}_i^k = \frac{\exp(L_i^k)}{\sum_{j=1}^K \exp(L_i^j)}.$$

Finally we obtained the variational Bayes learning algorithm,

$$(S_k, T_k, U_k) \Leftarrow \bar{y}_i^k,$$
$$\bar{y}_i^k \Leftarrow (S_k, T_k, U_k).$$

3. Variational Bayes Theory

The Bayes *a posteriori* distribution is defined by

$$p(w|D_n) = \frac{1}{Z_n}\sum_{h_n} P(h_n, w|D_n),$$

and the Bayes predictive distribution is defined by

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

Let the generalization error of Bayes learning G is defined by the Kullback-Leibler distance from the true distribution $q(x)$ to the Bayes predictive distribution $p(x|d_n)$,

$$G = EE_X\Big[\log\frac{q(X)}{p(X|d_n)}\Big],$$

where E shows the expectation value over the training samples d_n and $E_X[\ ]$ that over the testing samples. In Bayes learning it is well known that, for an arbitrary n,

$$E[G] = E[F_{n+1}] - E[F_n],$$

where F_n is the free energy for the given n samples.

The variational generalization error is defined by

$$\hat{G} = EE_X\Big[\log\frac{q(X)}{p(X|\hat{w})}\Big],$$

where $\hat{w}$ is the parameter that is optimized by the variational Bayes learning. However, the variational Bayes generalization error is not equal to the increase of the free energy.

$$E[\hat{G}] \neq E[\hat{F}_{n+1}] - E[\hat{F}_n].$$

The following theorem is well known in variational Bayes learning [7, 8].

Theorem Let the true probability distribution be given by eq.(1) with K_0. Let $M^* = (M+1)/2$. The variational free energy $\hat{F}_n$ satisfies the following inequality,

$$S_n + \lambda_1 \log n + nK_n(\hat{w}) + c_1 < \hat{F}_n < S_n + \lambda_2 \log n + c_2,$$

where S_n is the empirical entropy, $K_n(\hat{w})$ is the empirical Kullback-Leibler distance using the variational estimator $\hat{w}$, c_1, c_2 are constants, and both λ_1 and λ_2 satisfy

$$\lambda_1 = \begin{cases} (K-1)\phi_0 + M/2 & (\phi_0 \leq M^*) \\ (MK+K-1)/2 & (\phi_0 > M^*) \end{cases}$$
$$\lambda_2 = \begin{cases} (K-K_0)\phi_0 + (MK_0+K_0-1)/2 & (\phi_0 \leq M^*) \\ (MK+K-1)/2 & (\phi_0 > M^*) \end{cases}.$$

This theorem shows that, if the learning machine is more complex than the true distribution, the variational free energy has the phase transition at $\phi_0 = (M+1)/2$.

4. Experiments

In the experiments, we adopoted the condition, $M = 2$, $n = 200$, $K = 2$, $b_0 = 0$, and $\beta_0 = 0.000001$. In this case the critical point of the variational free energy is $\phi_0 = (M+1)/2 = 1.5$. In the true distribution, the mixture ratios were set as $(0.5, 0.5)$, $(0.7, 0.3)$, and $(0.9, 0.1)$, which are shown in Figures 1, 2, and 3 respectively. In each cases, the distance between the average of two normal distributions were set as $D = 0, 1, 2, 3$. From Figure 1 to 3, the horizontal and longitudinal lines respectively show the hyperparameter ϕ_0 and the generalization error.

In experiments $D = 0, 1$, there were phase transitions of the generalization error at $\phi_0 = (M+1)/2$. That is to say, the learning machine was more complex than the true distribution, the generalization error had also the phase transition. In such cases, the generalization errors were largest at the critical point.

In experiments $D = 2, 3$, it seems that the generalizaton errors were smallest at the critical point.

5. Discussion

From the information engineering point of view, the method to determine the hyperparameter is important. If the hyperparameter ϕ_0 is chosen as the critical point, then the generalization error does not depend on the condition of the true distribution. If it is chosen far from the critical point, then the generalization error becomes smaller or larger than that of the critrical point.

In the real world applications, the hyperparameter is sometimes optimized by the minimization of the variational free energy. It is the future study to clarify the generalization errors in such an optimization method.

6. Conclusion

We clarified by experiments that the variational generalization error has the phase transition whose critical point is $\phi_0 = (M+1)/2$. If the learning machine is more complex than the true distribution, then the generalization error is largest at the critical point, whereas, if otherwise, then the generalization error is smallest at the critical point.

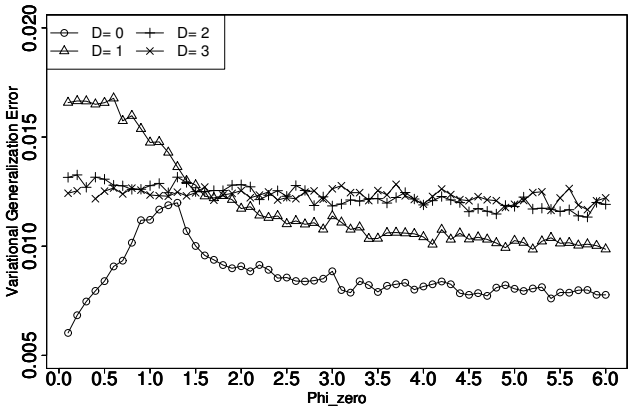

Figure 1: Hyperparameter ϕ_0 and generalization error ,the true mixture ratios are $(0.5, 0.5)$

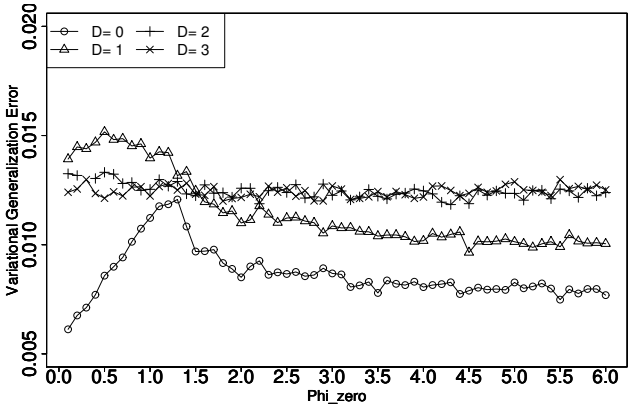

Figure 2: Hyperparameter ϕ_0 and generalization error ,the true mixture ratios are $(0.7, 0.3)$

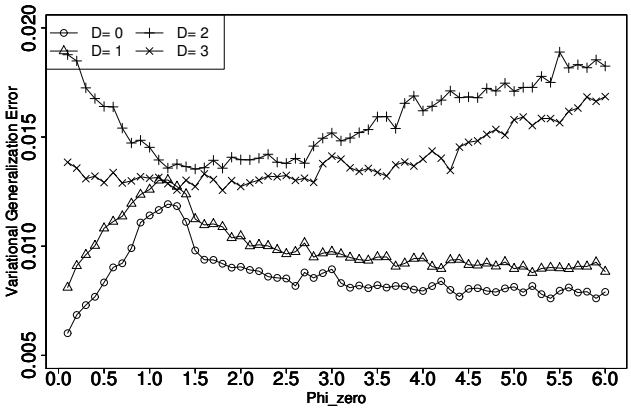

Figure 3: Hyperparameter ϕ_0 and generalization error ,the true mixture ratios are $(0.9, 0.1)$

Acknowledgement

This research was partially supported by the Ministry of Education, Science, Sports and Culture in Japan, Grant-in-Aid for Scientific Research 18079007.

References

[1] S.-i. Amari, H. Park, and T. Ozeki, "Singularities Affect Dynamics of Learning in Neuromanifolds," Neural Comput., 18(5),pp.1007 - 1065,2006.

[2] H. Hironaka, "Resolution of singularities of an algebraic variety over a field of characteristic zero," Ann. of Math., Vol.79, 109-326,1964.

[3] T. Hosino, K. Watanabe, S. Watanabe, "Stochastic complexity of Hidden Markov Models on the Variational Bayesian Learning," to appear in IEICE Transactions (D-II).

[4] K. Nagata, S. Watanabe, "The Exchange Monte Carlo Method for Bayesian Learning in Singular Learning Machines," Proc. of WCCI2006,(Canada, Vancouver), 2006.

[5] S. Nakajima, S. Watanabe, "Variational Bayes Solution of Linear Neural Networks and its Generalization Performance," Neural Computation, to appear.

[6] S. Oyama, S. Watanabe, "On the effect of hyperparameter to the generalization error in variational Bayes learning," IEICE Technical Report, NC2008-87, pp.31-36, 2009.

[7] K. Watanabe, S. Watanabe, "Stochastic Complexities of Gaussian Mixtures in Variational Bayesian Approximation," Journal of Machine Learning Research, Vol.7, (Apr), pp. 625–644, 2006.

[8] K. Watanabe, S. Watanabe, "Stochastic complexities of general mixture models in variational Bayesian leaning," Neural Networks, to appear.

[9] S. Watanabe, "Generalized Bayesian framework for neural networks with singular Fisher information matrices," Proc. of International Symposium on Nonlinear Theory and Its applications, (Las Vegas), pp.207-210, 1995.

[10] S. Watanabe, "Algebraic Analysis for Nonidentifiable Learning Machines," Neural Computation, Vol.13, No.4, pp.899-933, 2001.

[11] S. Watanabe, "Algebraic geometry of singular learning machines and symmetry of generalization and training errors," Neurocomputing, Vol.67, pp.198-213, 2005.