ASYMPTOTIC FREE ENERGY OF VARIATIONAL BAYESIAN DEEP LEARNING

Hiroshi Wakimori & Tikara Hosino
Nihon Unisys, Ltd.
1-1-1 Toyosu Koto-ku, Tokyo 135-8560 Japan

{hiroshi.wakimori,chikara.hoshino}@unisys.co.jp

ABSTRACT

The Bayesian deep learning method is promising for its theoretical foundation. Especially, it was probed that free energy can asymptotically identify the structure of the true distribution without random fluctuations. In this paper, we derive the expected asymptotic variational free energy which is the expectation of the objective function of the approximated free energy, in the case of Gaussian trial posterior. The results show that the variance of the posterior reflects the relative structure of the true distribution and the learning model. This result clarifies the theoretical insights of model selection or model distillation in variational approximation of Bayesian methods.

1 INTRODUCTION

Bayesian learning of deep neural networks has several advantages over maximum likelihood method or maximum posterior method (MAP). For example, the Bayesian method avoids over-fitting of models and resolves model selection problems in a theoretically established way (Watanabe, 2009). Since accurate calculation of posterior distribution using MCMC is computationally inefficient, approximation methods such as variational method are widely used. The variational method approximates the posterior distribution of parameters $p(\theta|x^n)$ where x^n is given n samples and θ is model parameters by the trial distribution $q(\theta)$ using the variational principle on free energy.

$$\begin{aligned}\text{Free energy} \equiv -\log p(x^n) &= -\log \int p(x^n|\theta)p(\theta)d\theta = -\log \int p(x^n|\theta)p(\theta)\frac{q(\theta)}{q(\theta)}d\theta \\ &\leq E[-\log p(x^n|\theta)]_{q(\theta)} + KL(q(\theta)||p(\theta)),\end{aligned}$$

where $p(x|\theta)$ is learning model, $p(\theta)$ is prior distribution of parameters. The simple trial posterior distribution is the product of univariate Gaussian distributions $q(\theta) = \prod_{i=1}^{|\theta|} \mathcal{N}(\theta_i|\mu_i, \sigma_i^2)$ (Hinton & Van Camp, 1993; Graves, 2011). This approximation is implemented by Bayesian deep learning frameworks such as Edward (Tran et al., 2016). In mixture of exponential family and reduced rank regression case, the mean field type of variational free energy is clarified (Watanabe & Watanabe, 2006; Nakajima & Watanabe, 2007). Despite being widely used in practice, the theoretical properties of this approximation in deep learning have not been clarified.

The contributions of the paper are as follows,

- We derive the expected asymptotic free energy in the variational Bayesian deep learning.
- We show that the variational asymptotic free energy of deep learning is substantially different from the regular statistical models.
- We show that the variance of the posterior distribution reflects the relative structure of the true distribution and the learning model.

2 SIMPLE EXAMPLE

In this section, we present a simple example of the main points of analysis in the general case. We consider the following learning model of which input and output is (x, y).

$$p(y|x,a,b,c) = \frac{1}{(2\pi)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(y - af(bx) - cx)^2\right), p(a), p(b), p(c) \sim \mathcal{N}(0,1)$$

where $f(x)$ is the activate function, let $f(x) = x + x^2$. To examine the behavior of the redundant part of the learning model, we assume x and y are irrelevant and independently generated from $p_0(x), p_0(y) \sim \mathcal{N}(0,1)$.

For parameters, we introduce trial diagonal posterior $q(a) \sim \mathcal{N}(\mu_a, \sigma_a^2), q(b) \sim \mathcal{N}(\mu_b, \sigma_b^2), q(c) \sim \mathcal{N}(\mu_c, \sigma_c^2)$. Then, the expectation of the variational free energy $F_{vb}(n)$ is given by

$$\begin{aligned} F_{vb}(n) &\equiv E[E[-\log p(y^n|x,a,b,c)]_{q(a),q(b),q(c)}]_{p_0(x),p_0(y)} + \sum_{i\in(a,b,c)} KL(q(i)||p(i)) \\ &= nS_0 + \frac{n}{2}\left((\mu_a\mu_b + \mu_c)^2 + \mu_a^2\sigma_b^2 + \mu_b^2\sigma_a^2 + \sigma_a^2\sigma_b^2 + \sigma_c^2 + 3(\mu_a^2 + \sigma_a^2)(3\sigma_b^4 + 6\mu_b^2\sigma_b^2 + \mu_b^4)\right) \\ &+ \frac{1}{2}\log\frac{1}{\sigma_a^2} + \frac{\sigma_a^2}{2} + \frac{\mu_a^2}{2} - \frac{1}{2} + \frac{1}{2}\log\frac{1}{\sigma_b^2} + \frac{\sigma_b^2}{2} + \frac{\mu_b^2}{2} - \frac{1}{2} + \frac{1}{2}\log\frac{1}{\sigma_c^2} + \frac{\sigma_c^2}{2} + \frac{\mu_c^2}{2} - \frac{1}{2}, \end{aligned}$$

where $S_0 = \frac{1}{2}(2\pi + 1)$ is entropy of $p_0(y)$. We consider the minimization of this equation in the asymptotic case ($n \to \infty$). In its minimization, $\mu_a^2, \mu_b^2, \mu_c^2$ is obviously 0.

$$F_{vb}(n) = nS_0 + \frac{n}{2}\left(\sigma_a^2\sigma_b^2 + \sigma_c^2 + 9\sigma_a^2\sigma_b^4\right) + \frac{1}{2}\left(\log\frac{1}{\sigma_a^2}\frac{1}{\sigma_b^2}\frac{1}{\sigma_c^2}\right) + \frac{\sigma_a^2}{2} + \frac{\sigma_b^2}{2} + \frac{\sigma_c^2}{2} - \frac{3}{2}$$

The leading-order term is determined by the variance of the posterior. Thus, for estimation of the posterior order, we assume $\sigma_i^2 = \frac{1}{n^{\alpha_i}}$,

$$\begin{aligned} F_{vb}(n) &= nS_0 + \frac{1}{2}\left(\frac{n}{n^{\alpha_a+\alpha_b}} + \frac{n}{n^{\alpha_c}} + 9\frac{n}{n^{\alpha_a+2\alpha_b}}\right) \\ &+ \frac{1}{2}(\alpha_a + \alpha_b + \alpha_c)\log n + \frac{1}{2}\left(\frac{1}{n^{\alpha_a}} + \frac{1}{n^{\alpha_b}} + \frac{1}{n^{\alpha_c}}\right) - \frac{3}{2} \end{aligned}$$

This is the optimization problem under $0 \le \alpha_a, \alpha_b, \alpha_c \le 1$.

- $O(n)$ terms : $\alpha_a + \alpha_b \ge 1$ and $\alpha_c \ge 1$ (The variance of the posterior is upper bounded.)
- $O(\log n)$ terms : $\min \alpha_a + \alpha_b + \alpha_c$ (Maximize the entropy of the posterior.)
- $O(1)$ terms : $\min \frac{1}{n^{\alpha_a}} + \frac{1}{n^{\alpha_b}} = \max\min A$ (As possible, make the variances equally.) where A is the set of α not determined by the above conditions.

At the first condition, we ignored $\alpha_a + 2\alpha_b \ge 1$ because it is satisfied when $\alpha_a + \alpha_b \ge 1$ is satisfied. It is noted that, in the same manner, the higher order terms than σ_i^2 can be ignored by the characteristics of the moments of the Gaussian distributions. We use this fact in the next section. It is understood that it is sufficient to solve as strong constraint conditions in order from the top. In this example, the solution is $\alpha_a = \alpha_b = \frac{1}{2}, \alpha_c = 1$ and

$$F_{vb}(n) = nS_0 + \log n + O(1)$$

3 GENERAL CASE

As the general case, we assume that: First, the true distribution and the learning model are written by computational graphs. Second, the learning model includes true distribution $p(y|x)$ and the true distribution p_0 is parameterized by minimum number of parameters which satisfy the condition $KL(p_0(y|x)||p(y|x)) = 0$ and we define the entropy of the true distribution $S \equiv E[-\log p_0(y|x)]_{p_0(y|x)p_0(x)}$. Third, the activation function $f(x)$ has Taylor expansion and the expansion has x term (such as $tanh$, $sigmoid$, elu (Clevert et al., 2015), $swish$ (Ramachandran et al., 2017)).

The general case can be solved in the same way. For the set of all paths including redundant parameters on the computational graph ($Path$) and all parameters (θ), we solve the optimization problem:

- $\alpha_i = 1$ if θ_i belongs to the true distribution.
- $\forall \quad path_j \in Path \quad \sum_{\alpha_i \in path_j} g(\alpha_i) \geq 1$
  where $g(\alpha_i)$ returns 0 if θ_i belongs to the true distribution, otherwise α_i.
- $\min \alpha_1 + \cdots + \alpha_{|\theta|}$
- $\min \frac{1}{n^{\alpha_1}} + \cdots + \frac{1}{n^{\alpha_{|\theta|}}} = \max \min A$
  where A is the set of α not determined by the above conditions.

Then, using the coefficient $\lambda = \alpha_1 + \cdots + \alpha_{|\theta|}$, the variational free energy is

$$F_{vb}(n) = nS + \frac{\lambda}{2} \log n + O(1)$$

In this case, the mean of redundant parameters μ_i is 0.

4 Fully Connected Deep Learning

For actual example, we consider fully connected deep neural networks. We assume the network has L layers and each layer i has n_i nodes. Let the number of the nodes which belong to the true distribution is $t_i > 0$ and that of the other redundant nodes is $r_i \geq 0$. The output layer satisfies $t_L = n_L, r_L = 0$. In this case the variational free energy F_{vb} (under the $n \to \infty$) is given by

$$F_{vb}(n) = nS + \frac{\lambda}{2} \log n + O(1), \ \lambda = \sum_{i=j}^{L-2} (n_i - n_{i+2}) r_{i+1} + \sum_{i=1}^{L-1} n_i t_{i+1}$$

where $j = \min_{j=1}^{L-2} \sum_{i=j}^{L-2} (n_i - n_{i+2}) r_{i+1}$.

Specific examples are shown in Figure 1 and 2. Figure 1 is an example in which the numbers of nodes in all layers are the same. All α of connections between redundant nodes are $\frac{1}{L-1} = \frac{1}{4}$. On the other hand, Figure 2 is an example in which the number of nodes in a layer is smaller than other layers. Around this layer, the symmetry breaks and α of connections between redundant nodes are larger than other layers. In both examples, α of connections between true distribution nodes are 1. Thus, in Bayesian learning, it is theoretically possible to distinguish redundant nodes using α (the variance behavior against n).

In these examples, we show that the order of the variance is not uniform but automatically determined by the relative structure of the true distribution and the learning model and depends on the depth (Takamatsu et al., 2006). This behavior is different from the regular statistical models whose variance of the posterior is uniformly $O(\frac{1}{n})$ and coefficient λ is the number of the model parameters. Moreover, exact Bayes free energy is different from variational approximation because of higher order terms in the posterior which is Gaussian approximation ignored (Aoyagi, 2013). Finally, in these examples, the mean of the parameters μ_i is the same as MAP solution. The difference is occurred when the objective function is not expected by the true distribution. For this problem, analyzing the difference between $\min_{q(\theta)} E_{q(\theta)}[-\log(x^n|\theta)]$ and $\min_\theta[-\log(x^n|\theta)]$ is the future work.

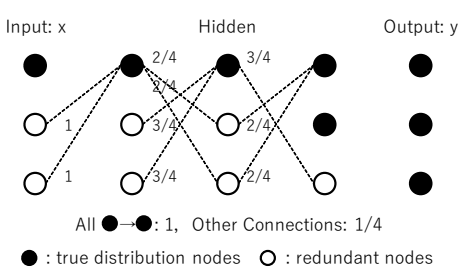

Figure 1: α of the example that every layer has the same number of nodes.

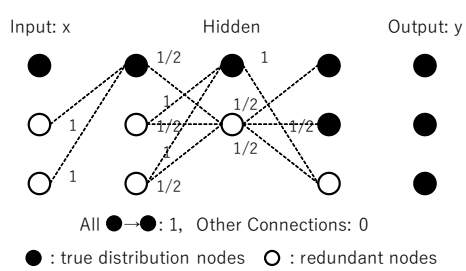

Figure 2: α of the example that a layer has small number of nodes.

REFERENCES

Miki Aoyagi. Consideration on singularities in learning theory and the learning coefficient. *Entropy*, 15(9):3714–3733, 2013.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

Alex Graves. Practical variational inference for neural networks. In *Advances in Neural Information Processing Systems*, pp. 2348–2356, 2011.

Geoffrey E Hinton and Drew Van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pp. 5–13. ACM, 1993.

Shinichi Nakajima and Sumio Watanabe. Variational bayes solution of linear neural networks and its generalization performance. *Neural Computation*, 19(4):1112–1153, 2007.

Prajit Ramachandran, Barret Zoph, and Quoc V Le. Swish: a self-gated activation function. *arXiv preprint arXiv:1710.05941*, 2017.

Shingo Takamatsu, Shinichi Nakajima, and Sumio Watanabe. Localized bayes estimation for non-identifiable models. In *International Conference on Neural Information Processing*, pp. 650–659. Springer, 2006.

Dustin Tran, Alp Kucukelbir, Adji B. Dieng, Maja Rudolph, Dawen Liang, and David M. Blei. Edward: A library for probabilistic modeling, inference, and criticism. *arXiv preprint arXiv:1610.09787*, 2016.

Kazuho Watanabe and Sumio Watanabe. Variational bayesian stochastic complexity of mixture models. In *Advances in neural information processing systems*, pp. 1465–1472, 2006.

Sumio Watanabe. *Algebraic geometry and statistical learning theory*. Cambridge University Press, 2009.