Asymptotic Bayesian Generalization Error When Training and Test Distributions Are Different

Keisuke Yamazaki k-yam@pi.titech.ac.jp
Precision and Intelligence Laboratory, Tokyo Institute of Technology, R2-5, 4259 Nagatsuta, Midori-ku, Yokohama, 226-8503 Japan

Motoaki Kawanabe motoaki.kawanabe@first.fraunhofer.de
Fraunhofer FIRST, IDA, Kekuléstr. 7, 12489 Berlin, Germany

Sumio Watanabe swatanab@pi.titech.ac.jp
Precision and Intelligence Laboratory, Tokyo Institute of Technology

Masashi Sugiyama sugi@cs.titech.ac.jp
Department of Computer Science, Tokyo Institute of Technology, 2-12-1, O-okayama, Meguro-ku, Tokyo, 152-8552 Japan

Klaus-Robert Müller Klaus@first.fraunhofer.de
Fraunhofer FIRST, IDA, & Technical University of Berlin, Computer Science, Franklinstr. 28/29 10587 Berlin, Germany

Abstract

In supervised learning, we commonly assume that training and test data are sampled from the *same* distribution. However, this assumption can be violated in practice and then standard machine learning techniques perform poorly. This paper focuses on revealing and improving the performance of Bayesian estimation when the training and test distributions are different. We formally analyze the asymptotic Bayesian generalization error and establish its upper bound under a very general setting. Our important finding is that lower order terms—which can be ignored in the absence of the distribution change—play an important role under the distribution change. We also propose a novel variant of stochastic complexity which can be used for choosing an appropriate model and hyper-parameters under a particular distribution change.

Appearing in *Proceedings of the 24th International Conference on Machine Learning*, Corvallis, OR, 2007.

1. Introduction

The goal of supervised learning is to infer an underlying relation between input x and output y from training data. This allows us to predict the output value of an unseen test input point. A common assumption in the supervised learning scenario is that the test data is sampled from the *same* underlying distribution as the training data. However, this assumption is not often fulfilled in practice, e.g., when the data generation mechanism is non-stationary or the data sampling process has time or cost constraint. If the joint distribution $p(x, y)$ is totally different between training and test phases, we may not be able to extract any information about the test data from training data. Therefore, the change of distribution needs to be restricted in a reasonable way.

One of the most interesting types of distribution change would be the situation called the *covariate shift* (Shimodaira, 2000): the input distribution $p(x)$ varies but the functional relation $p(y|x)$ remains unchanged. For data from many applications such as off-policy reinforcement learning (Shelton, 2001), bioinformatics (Baldi et al., 1998) or brain-computer interfacing (Wolpaw et al., 2002), the covariate shift phenomenon is conceivable. Sample selection bias (Heckman, 1979) in econometrics may also include a particular form of

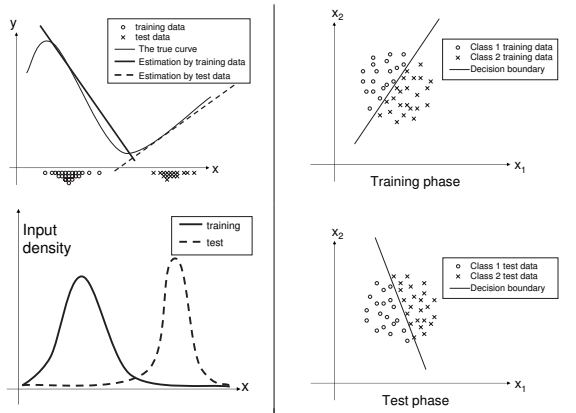

Figure 1. Schematic illustration of the distribution change: Left block: Covariate shift in a regression problem. Input distributions are different between the training and test phases, while the target function remains unchanged. Right block: Functional relation change in a classification problem. The target decision boundary changes, while the input distribution stays unchanged.

covariate shift. Another possible type of distribution change is the *functional relation change*, where $p(y|x)$ changes between the training and test phases. Under the classification scenarios, the situation called the *class prior change*—the class prior probability $p(y)$ is different for training and test data, can be often observed. See Fig.1 for illustration.

Standard supervised learning techniques are not designed to work appropriately under the distribution change. So far, several frequentist's methods have been developed to improve the performance, e.g., in the covariate shift scenarios (Shimodaira, 2000; Sugiyama & Müller, 2005; Sugiyama et al., 2007) and when the class-prior change occurs (Lin et al., 2002). However, it seems that a Bayesian perspective under such distribution changes is still an open research issue.

In this paper, we therefore investigate the behavior of Bayesian estimation in the presence of the distribution change. Our primal result is that lower order terms which can be ignored in the absence of the distribution change play an important role under the distribution change. Note that this result is derived without assuming the regularity condition (White, 1982) and is applicable to non-regular statistical models such as multi-layer perceptrons, Gaussian mixtures, and hidden Markov models. However, precisely investigating the lower order terms may only be possible in some limited cases. To cope with this problem, we derive an upper bound of the Bayesian generalization error for more general analysis.

The prediction performance of Bayesian estimation can be improved by properly choosing the model structure and hyper-parameters. For this purpose, the stochastic complexity (Rissanen, 1986) is often used as an evaluation criterion. The stochastic complexity corresponds to the probability of having the current training data given a model and hyper-parameters. Therefore, employing the stochastic complexity under the distribution change may not be suitable when the training and test data follow different distributions. In this paper, we propose a novel variant of stochastic complexity that can appropriately compensate for the effect of covariate shift.

2. Bayesian estimation without distribution change

In this section, we briefly introduce the standard Bayesian estimation procedure without distribution change and review asymptotic forms of the generalization error and the stochastic complexity.

2.1. Bayesian inference

Let $\{X^n, Y^n\} = \{X_1, Y_1, \ldots, X_n, Y_n\}$ be a set of training samples that are independently and identically generated following the true distribution $r(y|x)q(x)$. Let $p(y|x, w)$ be a learning machine and $\varphi(w)$ be the *a priori* distribution of parameter w. Then the *a posteriori* distribution is defined by

$$p(w|X^n, Y^n) = \frac{1}{Z(X^n, Y^n)} \prod_{i=1}^{n} p(Y_i|X_i, w)\varphi(w),$$

where

$$Z(X^n, Y^n) = \int \prod_{i=1}^{n} p(Y_i|X_i, w)\varphi(w)dw. \qquad (1)$$

The Bayesian predictive distribution is given by

$$p(y|x, X^n, Y^n) = \int p(y|x, w)p(w|X^n, Y^n)dw.$$

We evaluate the generalization error by the average Kullback divergence from the true distribution to the predictive distribution:

$$G(n) = E_{X^n,Y^n}\left[\int r(y|x)q(x)\log\frac{r(y|x)}{p(y|x, X^n, Y^n)}dxdy\right].$$

The stochastic complexity (Rissanen, 1986) is defined by

$$\overline{F}(X^n, Y^n) = -\log Z(X^n, Y^n), \qquad (2)$$

which can be used for selecting an appropriate model or hyper-parameters. When analyzing the behavior of the stochastic complexity, the following function plays an important role:

$$F(n) = E_{X^n,Y^n}\left[F(X^n, Y^n)\right], \qquad (3)$$

where $E_{X^n,Y^n}[\cdot]$ stands for the expectation value over all sets of training samples and

$$F(X^n, Y^n) = \overline{F}(X^n, Y^n) + \sum_{i=1}^{n} \log r(Y_i|X_i).$$

The generalization error and the stochastic complexity are linked by the following equation (Watanabe, 1999):

$$G(n) = F(n+1) - F(n). \quad (4)$$

2.2. Asymptotic generalization error

Watanabe (2001a) developed an algebraic geometrical approach to analyzing the asymptotic generalization error of Bayesian estimation. This approach includes as a special case the well-known result by Schwarz (1978), but is substantially more general—the generalization error of non-regular statistical models such as multi-layer perceptrons can also be analyzed.

When the learning machine $p(y|x,w)$ can attain the true distribution $r(y|x)$, i.e., there exists a parameter w^* such that $p(y|x,w^*) = r(y|x)$, the asymptotic expansion of $F(n)$ is given as follows (Watanabe, 2001a).

$$F(n) = \alpha \log n - (\beta - 1) \log\log n + O(1), \quad (5)$$

where the rational number $-\alpha$ and natural number β are the largest pole and its order of

$$J(z) = \int H(w)^z \varphi(w) dw.$$

$H(w)$ is defined by

$$H(w) = \int r(y|x) q(x) \log \frac{r(y|x)}{p(y|x,w)} dxdy. \quad (6)$$

Combining Eqs.(5) and (4) immediately gives

$$G(n) = \frac{\alpha}{n} - \frac{\beta - 1}{n \log n} + o\left(\frac{1}{n \log n}\right),$$

when $G(n)$ has an asymptotic form. The coefficients α and β indicate the speed of convergence of the generalization error when the number of training samples is sufficiently large.

When the learning machine cannot attain the true distribution (i.e., the model is misspecified), the stochastic complexity has an upper bound of the following asymptotic expression (Watanabe, 2001b).

$$F(n) \leq nC + \alpha \log n - (\beta - 1) \log\log n + O(1), \quad (7)$$

where C is a non-negative constant. When the generalization error has an asymptotic form, combining Eqs.(7) and (4) gives

$$G(n) \leq C + \frac{\alpha}{n} - \frac{\beta - 1}{n \log n} + o\left(\frac{1}{n \log n}\right), \quad (8)$$

where C is the bias.

3. Analysis of the Bayesian generalization error with distribution change

In this section, we analyze the generalization error of Bayesian estimation under the distribution change.

3.1. Notations

In the following, let us denote the training distribution with subscript 0 and the test distribution with subscript 1. Then the covariate shift situation is expressed by

$$r_0(y|x) = r_1(y|x) \quad \text{and} \quad q_0(x) \neq q_1(x),$$

while the functional relation change is described as

$$r_0(y|x) \neq r_1(y|x) \quad \text{and} \quad q_0(x) = q_1(x).$$

For $i = 0, 1$, let

$$G^i(n) = E^i_{X_{n+1}, Y_{n+1}} E^0_{X^n, Y^n} \left[\log \frac{r_i(Y_{n+1}|X_{n+1})}{p(Y_{n+1}|X_{n+1}, X^n, Y^n)} \right],$$

where $E^i_{X,Y}$ stands for the expectation over X and Y taken from $r_i(y|x)q_i(x)$. Note that the functions $G^0(n)$ and $G^1(n)$ correspond to the generalization errors without and with distribution change, respectively. Our primal goal in this section is to reveal the asymptotic form of $G^1(n)$.

3.2. Asymptotic expansion of generalization error

Let us define a function,

$$U^i(n+1) = E^i \left[-\log \int \exp\left(-\sum_{j=1}^{n} \log \frac{r_0(Y_j|X_j)}{p(Y_j|X_j, w)} - \log \frac{r_i(Y_{n+1}|X_{n+1})}{p(Y_{n+1}|X_{n+1}, w)} \right) \varphi(w) dw \right],$$

for $i = 0, 1$, where $E^i \equiv E^i_{X_{n+1}, Y_{n+1}} E^0_{X^n, Y^n}$. Note that $U^0(n) = F(n)$ (see Eq.(3)). We assume

(A1) $G^i(n)$ has an asymptotic expansion and $G^i(n) \to B_i$ as $n \to \infty$, where B_i is a constant.

(A1') $U^i(n)$ has the following asymptotic expansion,

$$U^i(n) = \underbrace{a_i n + b_i \log n + \cdots}_{T^i_H(n)} + \underbrace{c_i + \frac{d_i}{n} + o\left(\frac{1}{n}\right)}_{T^i_L(n)} \quad (9)$$

in the descending order with respect to n, where a_i, b_i, c_i, and d_i are constants. independent of n.

Note that **(A1')** is not essential but only for notational convenience.

Lemma 1 *The generalization error $G^1(n)$ is expressed as*

$$G^1(n) = U^1(n+1) - U^0(n). \quad (10)$$

Theorem 1 *Under the assumptions* **(A1)** *and* **(A1')**, *the asymptotic expansion of $G^1(n)$ is expressed by*

$$G^1(n) = a_0 + (c_1 - c_0) + \frac{b_0 + (d_1 - d_0)}{n} + o\left(\frac{1}{n}\right),$$

and $T_H^1(n) = T_H^0(n)$.

The proof is given in Appendix A.

In the standard case without the distribution change, it is straightforward to show that

$$G^0(n) = a_0 + \frac{b_0}{n} + o\left(\frac{1}{n}\right).$$

Thus an important finding from the above theorem is that the lower order terms in $T_L^i(n)$ which do not appear in the asymptotic expansion of $G^0(n)$ can not be ignored in the asymptotic expansion of $G^1(n)$.

Example 1 *Let the true distribution and learning model be*

$$r(y|x) = \frac{1}{\sqrt{2\pi}\sigma_2}\exp\left(-\frac{y^2}{2\sigma_2^2}\right),$$
$$p(y|x,a) = \frac{1}{\sqrt{2\pi}\sigma_2}\exp\left(-\frac{(y-ax)^2}{2\sigma_2^2}\right),$$

where the parameter is only $a \in R^1$. *Note that the learning model* $p(y|x,a)$ *can attain the true distribution* $r(y|x)$ *by* $a = 0$. *We assume a Gaussian prior,*

$$\varphi(a) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{a^2}{2\sigma^2}\right).$$

The training and test input distributions q_0 *and* q_1 *are respectively defined by*

$$q_i(x) = \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{x^2}{2\sigma_i^2}\right).$$

The coefficients in Theorem 1 are as follows:

$$a_0 = a_1 = 0, \;\; b_0 = b_1 = 1/2,$$
$$c_1 - c_0 = 0, \;\; d_1 - d_0 = (\sigma_1/\sigma_0 - 1)/2.$$

Then the generalization errors are written as

$$G^1(n) = \frac{\sigma_1^2}{2n\sigma_0^2} + o\left(\frac{1}{n}\right), \;\; G^0(n) = \frac{1}{2n} + o\left(\frac{1}{n}\right).$$

We omit the derivation because of lack of space; but we note that, due to the lower order terms, the derivation is not straightforward despite its simplicity at first glance.

In this regression case, the learning model can attain the true distribution, and the true line ($y = 0$) does not change in the test phase (i.e., the covariate shift). When the test input distribution is wider than the training input distribution (i.e., $\sigma_1 > \sigma_0$), the generalization errors satisfy $G^1(n) > G^0(n)$. This is consistent with an intuition that the data far from the origin bring more information on the true function ($y = 0$).

3.3. Bound of generalization error

The above theorem gives an insight that clarifying the asymptotic form of $G^1(n)$ requires to compute the lower order terms. However, the algebraic geometrical approach (see Section 2.2) does not take account of the lower order terms and we need to directly investigate them. This is usually very hard—even for a simple case such as Example 1, the calculation of lower order terms is not straightforward.

Here we propose a different approach: deriving an upper bound on $G^1(n)$ in terms of $G^0(n)$. Since the algebraic geometrical method can be used for revealing $G^0(n)$, this approach allows us to deal with a broader class of models. We assume

(A2) The largest difference between the training and test distributions is finite, i.e.

$$M = \max_{x,y\sim r_0(y|x)q_0(x)}\left[\frac{r_1(y|x)q_1(x)}{r_0(y|x)q_0(x)}\right] < \infty.$$

Theorem 2 *Under the assumptions* **(A1)** *and* **(A2)**, *the generalization error* $G^1(n)$ *asymptotically has an upper bound,*

$$G^1(n) \le MG^0(n) + D_1 + D_2,$$
$$\text{where } D_1 = \int r_1(y|x)q_1(x)\log\frac{r_1(y|x)}{r_0(y|x)}dxdy,$$
$$D_2 = \begin{cases} 0 & r_1(y|x) = r_0(y|x), \\ 1 & \text{otherwise.} \end{cases}$$

The proof is given in Appendix B.

Example 2 *Let* $r_1(y|x) = r_0(y|x)$ *and the training input distribution* $q_0(x)$ *and the test input distribution* $q_1(x)$ *be Gaussians,*

$$q_i(x) = \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left[-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right].$$

Then,

$$M = \max_{x\sim q_0(x)}\left[\frac{q_1(x)}{q_0(x)}\right] = \frac{\sigma_0}{\sigma_1}\max_{x\sim q_0(x)}\exp\left[\frac{(\mu_0-\mu_1)^2}{2(\sigma_0^2-\sigma_1^2)}\right.$$
$$\left. - \frac{\sigma_0^2-\sigma_1^2}{2\sigma_0^2\sigma_1^2}\left(x - \frac{\sigma_0^2\mu_1 - \sigma_1^2\mu_0}{\sigma_0^2-\sigma_1^2}\right)^2\right].$$

In this case, **(A2)** *requires* $\sigma_0 > \sigma_1$ *independent of* μ_i *and then* $M = \frac{\sigma_0}{\sigma_1}\exp\left[\frac{(\mu_0-\mu_1)^2}{2(\sigma_0^2-\sigma_1^2)}\right]$.

3.4. Discussion of the theorems

Under the assumption **(A1)**, B_i represents the bias and S_i corresponds to the speed of convergence, i.e.,

$$G^i(n) = B_i + \frac{S_i}{n} + o\left(\frac{1}{n}\right).$$

Here we analyze the speed of convergence and bias of Bayesian estimation under the distribution change.

3.4.1. Bias

According to Theorem 1, $B_0 = a_0, \;\; B_1 = a_0 + (c_1 - c_0)$. Thus the constant terms of $U^i(n)$ induce the distinc-

Table 1. Upper bounds of B_1. CS and FRC denote the covariate shift and functional relation change, respectively.

	$B_0 = 0$	$B_0 \neq 0$
No dist. change	$B_1 = 0$	$B_1 = B_0$
CS	$B_1 = 0$	$B_1 \leq MB_0$
FRC	$B_1 = D_1$	$B_1 \leq MB_0 + D_1 + 1$
CS & FRC	$B_1 = D_1$	$B_1 \leq MB_0 + D_1 + 1$

tion of the bias. As for the bias B_1, the following corollary holds.

Corollary 1 *When the learning machine $p(y|x, w)$ can realize the true $r_0(y|x)$ (i.e., $B_0 = 0$), $B_1 = D_1$.*

The proof is given in Appendix B. Table 1 summarizes the upper bound of B_1 in each case based on Theorem 2 and Corollary 1. Note that $D_1 = 0$ under the covariate shift ($r_1(y|x) = r_0(y|x)$).

When the learning model $p(y|x, w)$ can realize the true $r_0(y|x)$ (the middle column of Table 1), $B_1 \geq B_0 (= 0)$ always holds. Therefore, the bias is generally larger than the case without the distribution change. However, when the learning model $p(y|x, w)$ can not realize the true $r_0(y|x)$ (the right column of Table 1), $B_1 < B_0$ can occur depending on the sign of $c_1 - c_0$.

3.4.2. Speed of convergence

According to Theorem 1, $S_0 = b_0$, $S_1 = b_0 + (d_1 - d_0)$. Note that S_0 corresponds to α in Eq.(8). The sign of $d_1 - d_0$, which determines the magnitude relation between S_0 and S_1, depends on the setting. In Example 1, the both variances σ_0^2 and σ_1^2 affect the sign.

We recall that a faster convergence does not necessarily imply a lower generalization error due to the bias term. However, the speed term is dominant when the model can attain the true distribution under the covariate shift. In this case, $B_0 = B_1 = 0$ and $S_1 \leq MS_0$, $M = \max_{x \sim q_0} q_1(x)/q_0(x)$. In the above equation, the quantity M appears as a maximum factor of speeding-down. This would be natural since M represents the amount of difference in the training and test distributions. For example, if the support of the test distribution is not included in the support of the training distributions, M becomes infinity and no conclusion is derived from the bound. In such a case, the explicit computation as Example 1 is required.

3.5. Numerical example

Let us illustrate the error bounds using a simple regression problem borrowed from Sugiyama and Müller (2005): $y = \mathrm{sinc}(x) + \varepsilon$, where the noise ε follows $N(0, \sigma^2)$ with $\sigma^2 = (1/4)^2$. We assume the noise variance σ^2 is known. The training and the test inputs are subject to $N(1, (1/2)^2)$ and $N(2, (1/4)^2)$, respectively (see Fig 2). We use the linear regression model $\widehat{f}(x) = w_1 + w_2 x$, and estimate the parameter $w = (w_1, w_2)$ in the Bayesian framework with the prior distribution $\varphi(w)$ being

$$N(\mu, \lambda^{-1} I_2). \tag{11}$$

μ and λ are hyper-parameters and are determined based on the stochastic complexity.

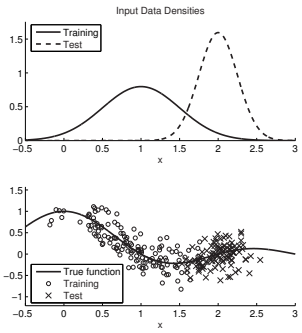

Figure 2. Illustrative example of covariate shift.

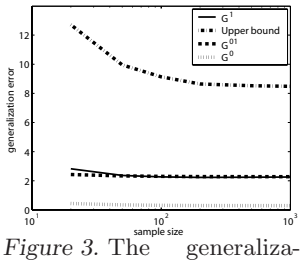

Figure 3. The generalization error $G^1(n)$ and its upper bound. The comparative errors $G^0(n)$ and $G^{01}(n)$ are also plotted.

Under this setting, the posterior distribution $p(w|X^n, Y^n)$ is Gaussian $N(\mu_n, \Lambda_n)$, where

$$\mu_n = \Lambda_n(\sigma^{-2}\Psi^n Y^n + \lambda\mu),$$
$$\Lambda_n = (\sigma^{-2}\Psi^n(\Psi^n)^\top + \lambda I_2)^{-1},$$
$$\Psi^n = (\psi(X_1), \ldots, \psi(X_n)), \quad \psi(x) = (1, x)^\top.$$

The predictive distribution $p(y|x, X^n, Y^n)$ is also Gaussian $N(m_n(x), v_n(x))$, where

$$m_n(x) = \psi^\top(x)\mu_n, \quad v_n(x) = \psi^\top(x)\Lambda_n\psi(x) + \sigma^2.$$

The generalization errors are expressed as

$$G^i(n) = \frac{1}{2} E^0_{X^n, Y^n}\left\{ \int q_i(x) \left[\frac{\{\mathrm{sinc}(x) - m_n(x)\}^2}{v_n(x)} + \frac{\sigma^2}{v_n(x)} - 1 - \log\frac{\sigma^2}{v_n(x)} \right] \mathrm{d}x \right\},$$

for $i = 0, 1$. The maximum ratio M defined by **(A2)** is $2\exp(8/3)$. Note that M is finite according to Example 2. The generalization errors G^0 and G^1 converge to $B_0 \doteq 0.2939$ and $B_1 \doteq 2.2818$, respectively, where we used the fact that μ_n and Λ_n converge to $E^0[\psi(x)\psi^\top(x)]^{-1}E^0[\psi(x)f(x)]$ and $0_{2\times 2}$, respectively.

Fig.3 depicts the generalization error $G^1(n)$ and its upper bound as functions of the sample size n. The values are averages over 400 (for $n \leq 100$) and 100 (for $n > 100$) realizations. For comparison, the generalization error $G^0(n)$ in the absence of the distribution change and its shift defined by $G^{01}(n) = G^0(n) - B_0 + B_1$ are also plotted. This shifted error has the same bias as $G^1(n)$ and the same convergence speed term as $G^0(n)$. In this specific example, the bias term of $G^1(n)$ is larger than $G^0(n)$ and the speed term does not seem so different.

4. Importance-weighted stochastic complexity

In the previous section, we analyzed the asymptotic generalization performance of Bayesian estimation under the distribution change. In this section, we shift our focus toward more practical aspects and show how the generalization performance could be improved under the covariate shift scenarios.

4.1. Definition

In Bayesian inference, the stochastic complexity is often used for selecting the model structure and hyper-parameters. However, as seen in Eqs.(1) and (2), the original stochastic complexity is computed from the marginal likelihood of *training* data. Under the covariate shift, we need to select the model structure and hyper-parameters in terms of the likelihood of the *test* data (both inputs and outputs). However, the test data is not available during the training phase.

To cope with this problem, we propose a variant of stochastic complexity called the *importance-weighted stochastic complexity (IWSC)*, which is defined as follows.

$$\overline{F}^{IW}(X^n, Y^n) = -\log \int \exp\left[l^{IW}(w)\right] \varphi(w) dw, \quad (12)$$

$$\text{where } l^{IW}(w) = \sum_{i=1}^{n} W(X_i) \log p(Y_i|X_i, w),$$

and $W(x) = q_1(x)/q_0(x)$. Here, we assume that $W(x)$ is known; when it is unknown, we may use an estimate (e.g., (Huang et al., 2007)).

4.2. Example of IWSC

Here we illustrate the behavior of IWSC using a toy example. Let us consider the following setting: the model $p(y|x, a)$ is a Gaussian $N(af(x), \sigma_4)$, where a is the parameter. The prior is the same as Example 1, $N(\mu, \sigma)$. The true distribution $r(y|x)$ is also a Gaussian $N(g(x), \sigma_2)$. Then IWSC is rewritten as

$$\overline{F}^{IW}(X^n, Y^n|\mu) = \left\{\sum_{i=1}^{n} W(X_i)\right\} \log(\sqrt{2\pi}\sigma_4)$$
$$+ \frac{1}{2} \log\left[1 + \frac{\sigma^2}{\sigma_4^2} \sum_{i=1}^{n} W(X_i) f^2(X_i)\right] + \frac{\mu^2}{2\sigma^2}$$
$$- \frac{1}{2\sigma^2\sigma_4^2} \frac{(\sigma^2 \sum_{i=1}^{n} W(X_i) f(X_i) Y_i + \sigma_4^2 \mu)^2}{\sigma^2 \sum_{i=1}^{n} W(X_i) f^2(X_i) + \sigma_4^2}. \quad (13)$$

First, let us analyze the behavior of the average IWSC. Considering the order of n, we can prove that

$$E^0_{X^n,Y^n}\left[\overline{F}^{IW}(X^n, Y^n|\mu)\right]$$
$$= n\left(\log(\sqrt{2\pi}\sigma_4) - \frac{\left\langle fg\right\rangle_1^2}{2\sigma_4^2\left\langle f^2\right\rangle_1}\right) + \frac{1}{2}\log n + O(1),$$

where μ is a hyper-parameter and $\left\langle f\right\rangle_1 = E^1_x[f(x)]$. This implies that the average IWSC is expressed in terms of the expectation over the data from the *test* input distribution, not on the *training* input distribution, and thus the application of the importance weight $W(x)$ would be reasonable.

Next we focus on the optimization of the hyper-parameter μ. We can show that the hyper-parameter that minimizes Eq.(13) is given as

$$\hat{\mu} \equiv \arg\min \overline{F}^{IW}(X^n, Y^n|\mu) = \frac{\sum_{i=1}^{n} W(X_i) f(X_i) Y_i}{\sum_{i=1}^{n} W(X_i) f^2(X_i)}.$$

This result claims that IWSC selects a reasonable hyper-parameter because we can prove that $\hat{\mu}$ on average converges to

$$E^0_{X^n,Y^n}[\hat{\mu}] = \frac{\left\langle Wfg\right\rangle_0}{\left\langle Wf^2\right\rangle_0} = \frac{\left\langle fg\right\rangle_1}{\left\langle f^2\right\rangle_1},$$

where $\left\langle f\right\rangle_0 = E^0_x[f(x)]$. The convergent point is the same as the hyper-parameter selected by the ordinary stochastic complexity when the training input distribution agrees with the test input distribution.

4.3. Experimental results

We report a result of a simple numerical example to illustrate how IWSC actually works. We again used the same toy regression problem used in Section 3.5.

Linear models were learned with 200 training samples by the Bayesian procedure. The noise variance σ^2 is assumed to be known for simplicity and the hyper-parameters μ and λ in Eq.(11) were selected based on the stochastic complexity (2) or IWSC (12). The results are depicted in Fig.4. The solid lines in the left-most and second-left graphs show the mean of the prior, i.e. $\mu_1 + \mu_2 x$, while the dashed lines indicate the regions within three times of the standard deviations determined by the dispersion parameter λ. '$\circ$' are training samples and '$\times$' are noiseless test samples. The result of IWSC (see the second-left graph) predicts the output values in the test region very well, while SC only captures the training samples (see the left-most graph). We remark that the hyper-parameters in the second-left panel were obtained only from the training samples ('$\circ$') through IWSC. The profiles of SC and IWSC over the mean hyper-parameter μ are depicted in the second-right and right-most graphs, showing that both surfaces are smooth with a unique

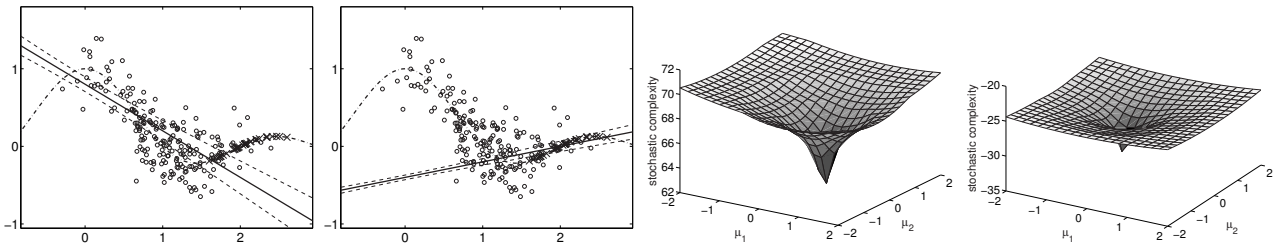

Figure 4. Learned functions obtained based on SC (left-most) and IWSC (second-left). SC (second-right) and IWSC (right-most) over the mean hyper-parameter μ (λ was optimally chosen by SC and IWSC, respectively).

minimum but at the different points.

4.4. Related works

The importance-weight has been widely used in the context of frequentist's approach. Maximum likelihood estimation is no longer consistent under the covariate shift when the model is misspecified; instead, the maximizer of the importance-weighted log-likelihood is consistent (Shimodaira, 2000).

$$\max \sum_{i=1}^{n} W(X_i) \log p(Y_i|X_i, w). \qquad (14)$$

However, this is not efficient and is rather unstable in practical situations with finite samples. To cope with this problem, an adaptive variant was proposed (Shimodaira, 2000):

$$\max \sum_{i=1}^{n} W(X_i)^{\lambda} \log p(Y_i|X_i, w), \qquad (15)$$

where $0 \leq \lambda \leq 1$. λ controls the trade-off between consistency and efficiency and it needs to be chosen appropriately for better estimation. Note that any empirical error based methods could be extended similarly.

The task of choosing λ is the model selection problem. Standard model selection methods such as Akaike's information criterion (Akaike, 1974) and cross-validation (Stone, 1974) are not designed to work well under the covariate shift. To cope with this problem, a modified information criterion has been developed (Shimodaira, 2000), where the importance-weight plays an essential role. Similarly, an importance-weighted model selection criterion specialized for linear regression (Sugiyama & Müller, 2005) and an importance-weighted cross-validation method (Sugiyama et al., 2007) have been developed and have shown to work well in real-world problems.

In the above importance-weighting framework, it is theoretically assumed that the importance is known a priori. However, this may not be the case in practice. To cope with this problem, a method of directly estimating the importance in a non-parametric way has been developed (Huang et al., 2007), which effectively makes use of the kernel trick (Schölkopf & Smola, 2002) in a class of reproducing kernel Hilbert spaces.

Experimental design where the training input distribution is designed by users is a relevant situation since it naturally induces the covariate shift. A standard approach to experimental design in least-squares regression often ignores the bias of the estimator and design the training input distribution so that the variance of the estimator is minimized (Fedorov, 1972). However, when the model is misspecified—which is a usual case in practice—the bias may not be ignored because of the covariate shift. Instead, the importance-weighted least squares method produces an asymptotic unbiased estimator and its use allows us to apply the variance-only approach also in the experimental design of approximately linear regression (Wiens, 2000; Sugiyama, 2006). Furthermore, an experimental design method for totally misspecified models has been developed (Kanamori & Shimodaira, 2003), where the importance-weight plays an essential role in establishing the consistency.

5. Conclusions

This paper clarified the asymptotic behavior of the Bayesian generalization error under the distribution change. Our result gave an interesting insight that the lower order terms which are ignored in the standard asymptotic theory play important roles under the distribution change. We also established an upper bound of the asymptotic generalization error in terms of the generalization error in the absence of the distribution change. In order to improve the prediction performance, we proposed a variant of stochastic complexity which can be used for choosing an appropriate model and hyper-parameters under the covariate shift.

Our future study will focus on investigating and improving the tightness of the bound. Similar to IWSC, the likelihood term in the posterior distribution can also be modified by the importance weight $W(X_i)$ for compensating for the change in the input distributions (Shimodaira, 2000). A promising direction in this line would be to combine these procedures into a single framework and analyze the generalization performance.

This research was partly supported by the Alexander von Humboldt Foundation, MEXT 18079007, 17700142 and 18300057, the Okawa Foundation, and Microsoft IJARC.

A. Proof of Lemma 1 and Theorem 1

In the same way to derive Eq.(4), we can obtain $G^i(n) = U^i(n+1) - U^0(n)$. The asymptotic expansions of $U^i(n)$ and $G^1(n)$ are immediately derived based on this relation, **(A1)**, and **(A1')**. If a coefficient of $T_H^1(n)$ in $U^1(n)$ is different from that of $T_H^0(n)$, the assumption **(A1)** is violated. For example, if $a_1 \neq a_0$, $G^1(n)$ has the term $(a_1 - a_0)n$. This means $G^1(n) \to \infty$ as $n \to \infty$. Therefore, it must hold that $T_H^1(n) = T_H^0(n)$.

B. Proof of Theorem 2 and Corollary 1

Define that

$$D_3 = 1 - E^0_{X^n,Y^n}\left[\int \frac{r_1(y|x)p(y|x,X^n,Y^n)}{r_0(y|x)} q_1(x)dxdy\right].$$

According to $S(x) \equiv e^{-x} - 1 + x$,

$$G^1(n) = E^0_{X^n,Y^n}\left[\int r_1(y|x)q_1(x)\log\frac{r_0(y|x)}{p(y|x,X^n,Y^n)}dxdy\right] + D_1 = D_4(n) + D_1 + D_3, \quad (16)$$

$$\text{where } D_4(n) = E^0_{X^n,Y^n}\left[\int \frac{r_1(y|x)q_1(x)}{r_0(y|x)q_0(x)} r_0(y|x)q_0(x) \times S\left(\log\frac{r_0(y|x)}{p(y|x,X^n,Y^n)}\right)dxdy\right].$$

When $r_1(y|x) = r_0(y|x)$, $D_3 = D_2(= 0)$. Otherwise, $D_3 \leq D_2(= 1)$. Since $S(x) \geq 0$, $G^1(n) \leq MG^0(n) + D_1 + D_2$, which completes the proof of Theorem 2. Next, we prove Corollary 1. When $n \to \infty$ and $B_0 = 0$,

$$p(y|x,X^n,Y^n) \to r_0(y|x).$$

Therefore, $D_4(n)$ and D_3 in Eq.(16) asymptotically goes to zero, which means $B_1 = D_1$.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *AC-19*, 716–723.

Baldi, P., Brunak, S., & Stolovitzky, G. A. (1998). *Bioinformatics: The machine learning approach.* Cambridge: MIT Press.

Fedorov, V. V. (1972). *Theory of optimal experiments.* New York: Academic Press.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*, 153–162.

Huang, J., Smola, A., Gretton, A., Borgwardt, K. M., & Schölkopf, B. (2007). Correcting sample selection bias by unlabeled data. In B. Schölkopf, J. Platt and T. Hoffman (Eds.), *Advances in neural information processing systems 19.* Cambridge, MA: MIT Press.

Kanamori, T., & Shimodaira, H. (2003). Active learning algorithm using the maximum weighted log-likelihood estimator. *Journal of Statistical Planning and Inference*, *116*, 149–162.

Lin, Y., Lee, Y., & Wahba, G. (2002). Support vector machines for classification in nonstandard situations. *Machine Learning*, *46*, 191–202.

Rissanen, J. (1986). Stochastic complexity and modeling. *Annals of Statistics*, *14*, 1080–1100.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels.* Cambridge, MA: MIT Press.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6 (2)*, 461–464.

Shelton, C. R. (2001). Importance sampling for reinforcement learning with multiple objectives. *PhD thesis.* Massachusetts Institute of Technology.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, *90*, 227–244.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B*, *36*, 111–147.

Sugiyama, M. (2006). Active learning in approximately linear regression based on conditional expectation of generalization error. *Journal of Machine Learning Research*, *7*, 141–166.

Sugiyama, M., Krauledat, M., & Müller, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, *8*.

Sugiyama, M., & Müller, K.-R. (2005). Input-dependent estimation of generalization error under covariate shift. *Statistics & Decisions*, *23*, 249–279.

Watanabe, S. (1999). Algebraic analysis for singular statistical estimation. *Lecture Notes on Computer Science Springer*, *1720*, 39–50.

Watanabe, S. (2001a). Algebraic analysis for non-identifiable learning machines. *Neural Computation*, *13 (4)*, 899–933.

Watanabe, S. (2001b). Algebraic information geometry for learning machines with singularities. *Advances in Neural Information Processing Systems*, *14*, 329–336.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*, 1–25.

Wiens, D. P. (2000). Robust weights and designs for biased regression models: Least squares and generalized M-estimation. *Journal of Statistical Planning and Inference*, *83*, 395–412.

Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., & Vaughan, T. M. (2002). Brain-computer interfaces for communication and control. *Clinical Neurophysiology*, *113*, 767–791.