Model Selection under Covariate Shift*

Masashi Sugiyama[1] and Klaus-Robert Müller[2]

[1] Tokyo Institute of Technology, Tokyo, Japan
sugi@cs.titech.ac.jp http://sugiyama-www.cs.titech.ac.jp/~sugi/
[2] Fraunhofer FIRST.IDA, Berlin, and University of Potsdam, Potsdam, Germany
klaus@first.fhg.de http://ida.first.fraunhofer.de/~klaus/

Abstract. A common assumption in supervised learning is that the training and test input points follow the same probability distribution. However, this assumption is not fulfilled, e.g., in interpolation, extrapolation, or active learning scenarios. The violation of this assumption—known as the covariate shift—causes a heavy bias in standard generalization error estimation schemes such as cross-validation and thus they result in poor model selection. In this paper, we therefore propose an alternative estimator of the generalization error. Under covariate shift, the proposed generalization error estimator is unbiased if the learning target function is included in the model at hand and it is asymptotically unbiased in general. Experimental results show that model selection with the proposed generalization error estimator is compared favorably to cross-validation in extrapolation.

1 Introduction

Let us consider a regression problem of estimating an unknown function $f(\boldsymbol{x})$ from training examples $\{(\boldsymbol{x}_i, y_i) \mid y_i = f(\boldsymbol{x}_i)+\epsilon_i\}_{i=1}^n$, where $\{\epsilon_i\}_{i=1}^n$ are i.i.d. random noise with mean zero and unknown variance σ^2. Using a linear regression model

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{p} \alpha_i \varphi_i(\boldsymbol{x}), \tag{1}$$

where $\{\varphi_i(\boldsymbol{x})\}_{i=1}^p$ are fixed linearly independent functions and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_p)^\top$ are parameters, we would like to learn the parameter $\boldsymbol{\alpha}$ such that the squared test error expected over all test input points (or the *generalization error*) is minimized. Suppose the test input points independently follow a probability distribution with density $p_t(\boldsymbol{x})$ (> 0). Then the generalization error is expressed as

$$J = \int \left(\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2 p_t(\boldsymbol{x}) d\boldsymbol{x}. \tag{2}$$

A common assumption in this supervised learning is that the training *input* points $\{\boldsymbol{x}_i\}_{i=1}^n$ independently follow the *same* probability distribution as the

* The authors would like to thank Dr. Motoaki Kawanabe and Dr. Gilles Blanchard for their valuable comments. We acknowledge the Alexander von Humboldt Foundation and from the PASCAL Network of Excellence (EU #506778) for financial support.

test input points [4]. However, this assumption is not fulfilled, for example, in *interpolation* or *extrapolation* scenarios: only few (or no) training input points exist in the regions of interest, implying that the test distribution is significantly different from the training distribution. *Active learning* also corresponds to such cases because the locations of training input points are designed by users while test input points are provided from the environment [1]. The situation where the training and test distributions are different is referred to as the situation under the *covariate shift* [3] or the *sample selection bias* [2]. Let $p_x(\boldsymbol{x})$ (> 0) be the probability density function of training input points $\{\boldsymbol{x}_i\}_{i=1}^n$. An example of an extrapolation problem where $p_x(\boldsymbol{x}) \neq p_t(\boldsymbol{x})$ is illustrated in Figure 1.

When $p_x(\boldsymbol{x}) \neq p_t(\boldsymbol{x})$, two difficulties arise in a learning process. The first difficulty is parameter learning. The ordinary least-squares learning, given by

$$\min_{\boldsymbol{\alpha}} \left[\sum_{i=1}^{n} \left(\hat{f}(\boldsymbol{x}_i) - y_i \right)^2 \right], \tag{3}$$

tries to fit the data well in the region with high training data density. This implies that the prediction can be inaccurate if the region with high test data density has low training data density. Theoretically, it is known that when the training and test distributions are different and the true function is not *realizable* (i.e., the learning target function is included in the model at hand), least-squares learning is no longer *consistent* (i.e., the learned parameter does not converge to the optimal one even when the number of training examples goes to infinity). This problem can be overcome by using a least-squares learning weighted by the *ratio* of test and training data densities[3] [3].

$$\min_{\boldsymbol{\alpha}} \left[\sum_{i=1}^{n} \frac{p_t(\boldsymbol{x}_i)}{p_x(\boldsymbol{x}_i)} \left(\hat{f}(\boldsymbol{x}_i) - y_i \right)^2 \right]. \tag{4}$$

A key idea of this weighted version is that the training data density is adjusted to the test data density by the density ratio, which is similar in spirit to *importance sampling*. Although the consistency becomes guaranteed by this modification, the weighted least-squares learning tends to have large variance. Indeed, it is no longer *asymptotically efficient* even when the noise is Gaussian. Therefore, in practical situations with finite samples, a stabilized estimator, e.g.,

$$\min_{\boldsymbol{\alpha}} \left[\sum_{i=1}^{n} \left(\frac{p_t(\boldsymbol{x}_i)}{p_x(\boldsymbol{x}_i)} \right)^{\lambda} \left(\hat{f}(\boldsymbol{x}_i) - y_i \right)^2 \right] \quad \text{for } 0 \leq \lambda \leq 1 \tag{5}$$

would give more accurate estimates[4]. Note that $\lambda = 0$ corresponds to the ordinary least-squares learning (3), while $\lambda = 1$ corresponds to consistent weighted

[3] In theory, we assume that $p_x(\boldsymbol{x})$ and $p_t(\boldsymbol{x})$ are known. Later in experiments, they are estimated from the data and we evaluate the practical usefulness of the theory.

[4] The learned parameter $\hat{\boldsymbol{\alpha}}_\lambda$ obtained by the weighted least-squares learning (5) is given by $\hat{\boldsymbol{\alpha}}_\lambda = \boldsymbol{L}_\lambda \boldsymbol{y}$, where $\boldsymbol{L}_\lambda = (\boldsymbol{X}^\top \boldsymbol{D}^\lambda \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{D}^\lambda$, $\boldsymbol{X}_{i,j} = \varphi_j(\boldsymbol{x}_i)$, $\boldsymbol{D}$ is the diagonal matrix with the i-th diagonal element $p_t(\boldsymbol{x}_i)/p_x(\boldsymbol{x}_i)$, and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top$.

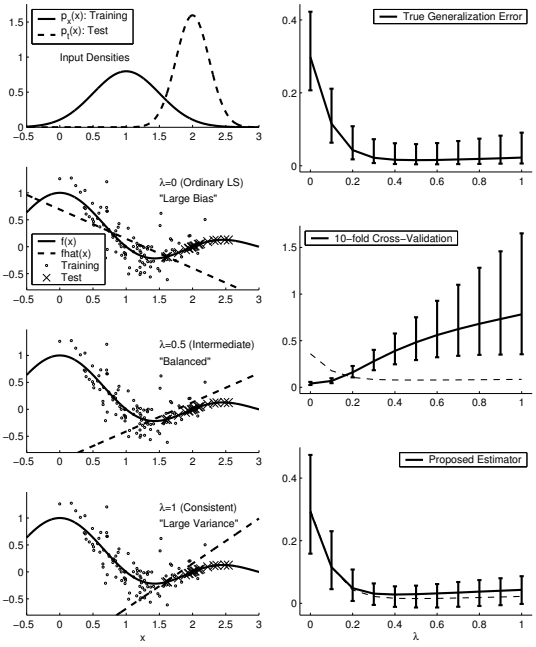

Fig. 1. An illustrative example of extrapolation by fitting a linear function $\hat{f}(\boldsymbol{x}) = \alpha_1 + \alpha_2 x$. [Left column]: The top graph depicts the probability density functions of the training and test input points, $p_x(x)$ and $p_t(x)$. In the bottom three graphs, the learning target function $f(x)$ is drawn by the solid line, the noisy training examples are plotted with $\circ$'s, a learned function $\hat{f}(x)$ is drawn by the dashed line, and the (noiseless) test examples are plotted with $\times$'s. Three different learned functions are obtained by weighted least-squares learning with different tuning parameter λ. $\lambda = 0$ corresponds to the ordinary least-squares learning (small variance but large bias), while $\lambda = 1$ gives an consistent estimate (small bias but large variance). With finite samples, an intermediate λ, say $\lambda = 0.5$, often provides better results. [Right column]: The top graph depicts the mean and standard deviation of the generalization error over 300 independent trials, as a function of λ. The middle and bottom graphs depict the means and standard deviations of the estimated generalization error obtained by the standard 10-fold cross-validation (10CV) and the proposed method. The dotted lines are the mean of the true generalization error. 10CV is heavily biased because of $p_x(x) \neq p_t(x)$, while the proposed estimator is almost unbiased with reasonably small variance.

least-squares learning (4). Thus, the parameter learning problem is now relocated to the model selection problem of choosing λ.

However, the second difficulty when $p_x(\boldsymbol{x}) \neq p_t(\boldsymbol{x})$ is model selection itself. Standard unbiased generalization error estimation schemes such as cross-validation are heavily biased, because the generalization error is over-estimated in the high training data density region and it is under-estimated in the high test data density region.

In this paper, we therefore propose a *new* generalization error estimator. Under covariate shift, the proposed estimator is proved to be exactly unbiased with finite samples in realizable cases and asymptotically unbiased in general. Furthermore, the proposed generalization error estimator is shown to be able to accurately estimate the *difference* of the generalization error, which is a useful property in model selection.

For simplicity, we focus on the problem of choosing the tuning parameter λ in the following. Note, however, that the proposed theory can be easily extended to general model selection of choosing basis functions or regularization constant.

2 A New Generalization Error Estimator

Let us decompose the learning target function $f(\boldsymbol{x})$ into $f(\boldsymbol{x}) = g(\boldsymbol{x}) + r(\boldsymbol{x})$, where $g(\boldsymbol{x})$ is the orthogonal projection of $f(\boldsymbol{x})$ onto the span of $\{\varphi_i(\boldsymbol{x})\}_{i=1}^p$ and the residual $r(\boldsymbol{x})$ is orthogonal to $\{\varphi_i(\boldsymbol{x})\}_{i=1}^p$, i.e., $\int r(\boldsymbol{x})\varphi_i(\boldsymbol{x})p_t(\boldsymbol{x})d\boldsymbol{x} = 0$. Since $g(\boldsymbol{x})$ is included in the span of $\{\varphi_i(\boldsymbol{x})\}_{i=1}^p$, it is expressed by $g(\boldsymbol{x}) = \sum_{i=1}^p \alpha_i^* \varphi_i(\boldsymbol{x})$, where $\boldsymbol{\alpha}^* = (\alpha_1^*, \alpha_2^*, \ldots, \alpha_p^*)^\top$ are unknown optimal parameters.

Let $\boldsymbol{U}$ be a p-dimensional matrix with the (i,j)-th element $\boldsymbol{U}_{i,j} = \int \varphi_i(\boldsymbol{x})\varphi_j(\boldsymbol{x})p_t(\boldsymbol{x})d\boldsymbol{x}$, which is assumed to be accessible in the current setting. Then the generalization error J is expressed as

$$\begin{aligned} J(\lambda) &= \int \hat{f}_\lambda(\boldsymbol{x})^2 p_t(\boldsymbol{x})d\boldsymbol{x} - 2\int \hat{f}_\lambda(\boldsymbol{x})f(\boldsymbol{x})p_t(\boldsymbol{x})d\boldsymbol{x} + \int f(\boldsymbol{x})^2 p_t(\boldsymbol{x})d\boldsymbol{x} \\ &= \langle \boldsymbol{U}\hat{\boldsymbol{\alpha}}_\lambda, \hat{\boldsymbol{\alpha}}_\lambda \rangle - 2\langle \boldsymbol{U}\hat{\boldsymbol{\alpha}}_\lambda, \boldsymbol{\alpha}^* \rangle + C, \end{aligned} \tag{6}$$

where $C = \int f(\boldsymbol{x})^2 p_t(\boldsymbol{x})d\boldsymbol{x}$. In Eq.(6), the first term $\langle \boldsymbol{U}\hat{\boldsymbol{\alpha}}_\lambda, \hat{\boldsymbol{\alpha}}_\lambda \rangle$ is accessible and the third term C does not depend on λ. Therefore, we focus on estimating the second term "$-2\langle \boldsymbol{U}\hat{\boldsymbol{\alpha}}_\lambda, \boldsymbol{\alpha}^* \rangle$".

Hypothetically, let us suppose that the following two quantities are available.

(i) A matrix $\boldsymbol{L}_u$ which gives a linear unbiased estimator of the unknown true parameter $\boldsymbol{\alpha}^*$: $\mathbb{E}_{\boldsymbol{\epsilon}} \boldsymbol{L}_u \boldsymbol{y} = \boldsymbol{\alpha}^*$, where $\mathbb{E}_{\boldsymbol{\epsilon}}$ denotes the expectation over the noise $\{\epsilon_i\}_{i=1}^n$.
(ii) An unbiased estimator σ_u^2 of the noise variance σ^2: $\mathbb{E}_{\boldsymbol{\epsilon}} \sigma_u^2 = \sigma^2$.

Note that $\boldsymbol{L}_u$ does not depend on $\boldsymbol{L}_\lambda$. Then it holds that

$$\mathbb{E}_{\boldsymbol{\epsilon}} \langle \boldsymbol{U}\hat{\boldsymbol{\alpha}}_\lambda, \boldsymbol{\alpha}^* \rangle = \langle \mathbb{E}_{\boldsymbol{\epsilon}} \boldsymbol{U}\boldsymbol{L}_\lambda \boldsymbol{y}, \mathbb{E}_{\boldsymbol{\epsilon}} \boldsymbol{L}_u \boldsymbol{y} \rangle = \mathbb{E}_{\boldsymbol{\epsilon}} [\langle \boldsymbol{U}\boldsymbol{L}_\lambda \boldsymbol{y}, \boldsymbol{L}_u \boldsymbol{y} \rangle - \sigma_u^2 \mathrm{tr}(\boldsymbol{U}\boldsymbol{L}_\lambda \boldsymbol{L}_u^\top)], \tag{7}$$

which implies that we can construct an unbiased estimator of $\mathbb{E}_{\boldsymbol{\epsilon}} \langle \boldsymbol{U}\hat{\boldsymbol{\alpha}}_\lambda, \boldsymbol{\alpha}^* \rangle$ if $\boldsymbol{L}_u$ and σ_u^2 are available. However, in general, neither $\boldsymbol{L}_u$ nor σ_u^2 may not be available. So we use the following approximations instead:

$$\widehat{\boldsymbol{L}}_u = (\boldsymbol{X}^\top \boldsymbol{D}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{D} \quad \text{and} \quad \widehat{\sigma_u^2} = \|\boldsymbol{G}\boldsymbol{y}\|^2 / \mathrm{tr}(\boldsymbol{G}), \tag{8}$$

where $\boldsymbol{G} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$. Actually, $\widehat{\boldsymbol{L}}_u$ corresponds to Eq.(4), which implies that $\widehat{\boldsymbol{L}}_u$ exactly fulfills the requirement (i) in realizable cases and asymptotically satisfies it in general [3]. On the other hand, it is known that the above $\widehat{\sigma_u^2}$ exactly fulfills the requirement (ii) in realizable cases [1]. Although, in general cases, $\widehat{\sigma_u^2}$ does not satisfy the requirement (ii) even asymptotically, it turns out that the asymptotic unbiasedness of $\widehat{\sigma_u^2}$ is not needed in the following.

Based on the above discussion, we define the following estimator $\hat{J}$ of the generalization error J.

$$\hat{J}(\lambda) = \langle \boldsymbol{U}\boldsymbol{L}_\lambda \boldsymbol{y}, \boldsymbol{L}_\lambda \boldsymbol{y} \rangle - 2\langle \boldsymbol{U}\boldsymbol{L}_\lambda \boldsymbol{y}, \widehat{\boldsymbol{L}}_u \boldsymbol{y} \rangle + 2\widehat{\sigma_u^2}\mathrm{tr}(\boldsymbol{U}\boldsymbol{L}_\lambda \widehat{\boldsymbol{L}}_u^\top). \tag{9}$$

Let $B_{\boldsymbol{\epsilon}}$ be the bias of $\hat{J}$: $B_{\boldsymbol{\epsilon}} = \mathbb{E}_{\boldsymbol{\epsilon}}[\hat{J} - J] + C$. Then we have the following theorem (proof is omitted because of lack of space).

Theorem 1 *If* $r(\boldsymbol{x}_i) = 0$ *for* $i = 1, 2, \ldots, n$, $B_{\boldsymbol{\epsilon}} = 0$. *If* $\delta = \max\{|r(\boldsymbol{x}_i)|\}_{i=1}^n$ *is sufficiently small,* $B_{\boldsymbol{\epsilon}} = \mathcal{O}(\delta)$. *If* n *is sufficiently large,* $B_{\boldsymbol{\epsilon}} = \mathcal{O}_p(n^{-\frac{1}{2}})$.

This theorem implies that, except for the constant C, $\hat{J}$ is exactly unbiased if $f(\boldsymbol{x})$ is strictly realizable, it is almost unbiased if $f(\boldsymbol{x})$ is almost realizable, and it is asymptotically unbiased in general. We can also prove that the above $\hat{J}$ can estimate the *difference* of the generalization error among different models. However, because of lack of space, we omit the detail.

3 Numerical Examples

Figure 1 shows the numerical results of an illustrative extrapolation problem. The curves in the right column show that the proposed estimator gives almost unbiased estimates of the generalization error with reasonably small variance (note that the target function is not realizable in this case).

We also applied the proposed method to *Abalone* data set available from the UCI repository. It is a collection of 4177 samples, each of which consists of 8 input variables (physical measurements of abalones) and 1 output variable (the age of abalones). The first input variable is qualitative (male/female/infant) so it was ignored, and the other input variables were normalized to $[0, 1]$ for convenience. From the population, we randomly sampled n abalones for training and 100 abalones for testing. Here, we considered a biased sampling: the sampling of the 4-th input variable (weight of abalones) has negative bias for training and positive bias for testing. That is, the weight of training abalones tends to be small while that for the test abalones tends to be large. We used multi-dimensional linear basis functions for learning. Here we suppose that the test input points are known (i.e., the setting corresponds to transductive inference [4]) and the density functions $p_x(\boldsymbol{x})$ and $p_t(\boldsymbol{x})$ were estimated from the training input points and test input points, respectively, using a kernel density estimation method.

Figure 2 depicts the mean values of each method over 300 trials for $n = 50, 200$, and 800. The error bars are omitted because they were excessive and deteriorated the graphs. Note that the true generalization error J is calculated using the test examples. The proposed $\hat{J}$ seems to give reasonably good curves and its minimum roughly agrees with the minimum of the true test error. On the other hand, irrespective of n, the minimizer of 10CV tend to be small.

We chose the tuning parameter λ by each method and estimated the age of the test abalones by using the chosen λ. The mean squared test error for all test abalones were calculated, and this procedure was repeated 300 times. The mean and standard deviation of the test error of each method are described in the left half of Table 1. It shows that $\hat{J}$ and 10CV work comparably for $n = 50, 200$, while $\hat{J}$ outperforms 10CV for $n = 800$. Hence, the proposed method overall compares favorably to 10CV.

We also carried out similar simulations when the sampling of the 6-th input variable (weight of gut after bleeding) is biased. The results described in the right half of Table 1 showed similar trends to the previous ones.

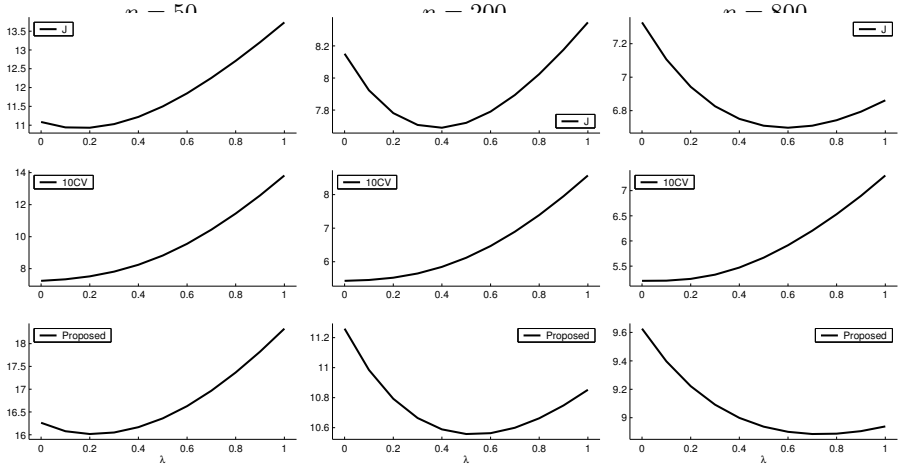

Fig. 2. Extrapolation of the 4-th variable in the Abalone dataset. The mean of each method is described. Each column corresponds to each n.

Table 1. Extrapolation of the 4-th variable (left) or the 6-th variable (right) in the Abalone dataset. The mean and standard deviation of the test error obtained with each method are described. The better method and comparable one by the t-test at the significance level 5% are described with boldface.

n	$\hat{J}$	10CV
50	$\mathbf{11.67 \pm 5.74}$	$\mathbf{10.88 \pm 5.05}$
200	$\mathbf{7.95 \pm 2.15}$	$\mathbf{8.06 \pm 1.91}$
800	$\mathbf{6.77 \pm 1.40}$	7.23 ± 1.37

n	$\hat{J}$	10CV
50	$\mathbf{10.67 \pm 6.19}$	$\mathbf{10.15 \pm 4.95}$
200	$\mathbf{7.31 \pm 2.24}$	$\mathbf{7.42 \pm 1.81}$
800	$\mathbf{6.20 \pm 1.33}$	6.68 ± 1.25

4 Conclusions

In this paper, we proposed a new generalization error estimator under covariate shift. The proposed estimator is shown to be unbiased with finite samples in realizable cases and asymptotically unbiased in general. Experimental results showed that model selection with the proposed generalization error estimator is compared favorably to the standard cross-validation in extrapolation scenarios.

References

1. V. V. Fedorov. *Theory of Optimal Experiments*. Academic Press, New York, 1972.
2. J. J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–162, 1979.
3. H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.
4. V. N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, Inc., New York, 1998.