IEICE Transactions on Information and Systems
vol.E90-D, no.11, pp.1779-1786, 2007.

A New Meta-Criterion for Regularized Subspace Information Criterion

Yasushi Hidaka
Department of Computer Science
Tokyo Institute of Technology
2-12-1 O-okayama, Meguro-ku, Tokyo 152-8552, Japan

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

Abstract

In order to obtain better generalization performance in supervised learning, model parameters should be determined appropriately, i.e., they should be determined so that the generalization error is minimized. However, since the generalization error is inaccessible in practice, the model parameters are usually determined so that an estimator of the generalization error is minimized. The regularized subspace information criterion (RSIC) is such a generalization error estimator for model selection. RSIC includes an additional regularization parameter and it should be determined appropriately for better model selection. A meta-criterion for determining the regularization parameter has also been proposed and shown to be useful in practice. In this paper, we show that there are several drawbacks in the existing meta-criterion and give an alternative meta-criterion that can solve the problems. Through simulations, we show that the use of the new meta-criterion further improves the model selection performance.

Keywords

supervised learning, generalization capability, model selection, unbiased estimator, regularized subspace information criterion.

1 Introduction

Supervised learning is the problem of estimating an underlying function from samples [23]. If the underlying function is accurately learned, the output values for unseen input points can be estimated. This is called the generalization capability. The level of the generalization capability is evaluated by the 'closeness' between the learned function and the underlying function, i.e., the generalization error. The goal of supervised learning is to obtain the function with the minimum generalization error. The learned function usually depends on model parameters such as the regularization parameter. Therefore, in order to obtain a better function, the model parameters should be chosen appropriately, i.e., so that the generalization error is minimized.

However, since the true learning target function is unknown, the generalization error can not be directly calculated. For this reason, we usually determine the model parameters such that an estimate of the generalization error is minimized [11, 1, 19]. The subspace information criterion (SIC) is such a generalization error estimator for model selection and is shown to be useful for model selection [19, 18]. However, the goodness of SIC is guaranteed in the sense of unbiasedness, implying that the variance of SIC can be large, e.g., when the noise level is very high.

To cope with this problem, the regularized subspace information criterion (RSIC) has been proposed [17]. RSIC is no longer unbiased, but has smaller variance and is more stable than SIC. RSIC includes an additional tuning parameter in the generalization error estimator itself. In order to successfully perform model selection with RSIC, this tuning parameter should be determined appropriately. The paper [17] gave a useful meta-criterion for determining the tuning parameter. Since the meta-criterion includes the unknown generalization error, we use an unbiased estimator of the meta-criterion for optimizing the tuning parameter. RSIC as well as the unbiased estimator of the meta-criterion include the noise variance, which is practically replaced by its estimator.

In this paper, we first show that there are five drawbacks in the conventional RSIC. The problem (a) is that the meta-criterion does not directly evaluate the goodness of RSIC used in practice (i.e., the one with the noise variance replaced by its unbiased estimator). The problem (b) is that RSIC is not necessarily a good approximation to the single-trial generalization error since the meta-criterion evaluates an error from the expected generalization error. The problem (c) is that the goodness of other generalization error estimators can not be measured since the meta-criterion is a goodness measure specialized for RSIC. The problem (d) is that the computational cost of the estimator of the meta-criterion could be unnecessarily large since an unnecessary constant term is also estimated in the estimator. The problem (e) is that replacing the noise variance by its unbiased estimator breaks the unbiasedness of the estimator of the meta-criterion. Then we propose an alternative meta-criterion that can solve all the above problems. Through simulations, we show that the use of the new meta-criterion further improves the model selection performance.

2 Formulation of Supervised Learning

In this section, we formulate the supervised learning problem.

Let us consider the problem of approximating a function from training samples. Let

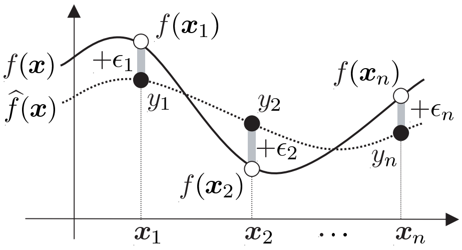

Figure 1: Supervised learning problem.

$f(\boldsymbol{x})$ be the learning target function, which is a real-valued function defined on $\mathcal{D} \subset \mathbb{R}^d$. We assume that $f(\boldsymbol{x})$ belongs to a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ [3, 23, 24]. Note that $\mathcal{H}$ is generally infinite dimensional. We denote the reproducing kernel of $\mathcal{H}$ by $K(\boldsymbol{x}, \boldsymbol{x}')$. Let $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ be training samples, where $\boldsymbol{x}_i \in \mathcal{D}$ is an input point and $y_i \in \mathbb{R}$ is an output value. We assume that the output value y_i is degraded by i.i.d. Gaussian noise ϵ_i with mean zero and variance σ^2:

$$y_i = f(\boldsymbol{x}_i) + \epsilon_i. \tag{1}$$

The training input points $\{\boldsymbol{x}_i\}_{i=1}^n$ could be either random or deterministic. The above formulation is summarized in Figure 1.

Let $\widehat{f}(\boldsymbol{x})$ be a learned function obtained from the training samples. The goal of supervised learning is to obtain the best approximation to the target function. To this end, we need to define the "goodness" measure of $\widehat{f}(\boldsymbol{x})$. In this paper, we measure the goodness of $\widehat{f}(\boldsymbol{x})$ by

$$\|\widehat{f} - f\|^2, \tag{2}$$

where $\|\cdot\|$ is the norm in the reproducing kernel Hilbert space $\mathcal{H}$. This quantity can be decomposed as

$$\|\widehat{f} - f\|^2 = \|\widehat{f}\|^2 - 2\langle \widehat{f}, f\rangle + \|f\|^2, \tag{3}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product in $\mathcal{H}$. Since the third term $\|f\|^2$ is a constant and it does not depend on $\widehat{f}(\boldsymbol{x})$, we ignore it and define the rest by G:

$$G = \|\widehat{f}\|^2 - 2\langle \widehat{f}, f\rangle. \tag{4}$$

We call G the *generalization error.*

Now our goal is formalized: we want to learn $\widehat{f}(\boldsymbol{x})$ from the training samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ such that the generalization error G is minimized. To this end, we need to define a search space for $\widehat{f}(\boldsymbol{x})$. The broadest choice would be the function space $\mathcal{H}$ itself, but it is hard to deal with since $\mathcal{H}$ is generally infinite dimensional. To alleviate this problem, we employ the following kernel model for learning[1] [9, 13, 15].

$$\widehat{f}(\boldsymbol{x}) = \sum_{i=1}^n \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i), \tag{5}$$

[1] All the discussion in this paper is still valid if we use a subset of $\{K(\boldsymbol{x}, \boldsymbol{x}_i)\}_{i=1}^n$ as basis functions.

where $\{\alpha_i\}_{i=1}^n$ are parameters to be learned. Note that this form is known to be a minimizer of some regularized functional in $\mathcal{H}$ [9].

Let

$$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_n)^\top, \tag{6}$$

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top, \tag{7}$$

where $^\top$ denotes the transpose. In this paper, we focus on the cases where the parameter vector $\boldsymbol{\alpha}$ is learned in a linear fashion, i.e., $\boldsymbol{\alpha}$ is obtained by

$$\boldsymbol{\alpha} = \boldsymbol{L}\boldsymbol{y}, \tag{8}$$

where $\boldsymbol{L}$ is an n-dimensional matrix which is independent of the noise $\{\epsilon_i\}_{i=1}^n$. We call $\boldsymbol{L}$ the *learning matrix.*

Consequently, the problem of learning $\widehat{f}(\boldsymbol{x})$ is converted into the problem of learning $\boldsymbol{L}$. Since the generalization error G includes the unknown learning target function $f(\boldsymbol{x})$, we can not directly learn $\boldsymbol{L}$ so that G is minimized. A standard approach to coping with this problem is to employ an accessible estimator of the unknown generalization error G. In the next section, we review existing methods for estimating G.

3 Generalization Error Estimators

In this section, we briefly review the generalization error estimators called the *subspace information criterion* (SIC) [19, 18] and its extension the *regularized SIC* (RSIC) [17].

3.1 Subspace Information Criterion

Let $\mathcal{S}$ be the subspace of $\mathcal{H}$ spanned by $\{K(\boldsymbol{x}, \boldsymbol{x}_i)\}_{i=1}^n$. Let $g(\boldsymbol{x})$ be the orthogonal projection of $f(\boldsymbol{x})$ onto $\mathcal{S}$. Note that, in the sense of Eq.(4), $g(\boldsymbol{x})$ is the optimal approximation to $f(\boldsymbol{x})$ in $\mathcal{S}$. (see Figure 2). Since $g(\boldsymbol{x})$ belongs to $\mathcal{S}$, it is expressed as

$$g(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i^* K(\boldsymbol{x}, \boldsymbol{x}_i), \tag{9}$$

where $\{\alpha_i^*\}_{i=1}^n$ are unknown optimal parameters. Let

$$\boldsymbol{\alpha}^* = (\alpha_1^*, \alpha_2^*, \ldots, \alpha_n^*)^\top. \tag{10}$$

Then the expectation of the generalization error can be expressed as follows [18].

$$\mathbb{E}_{\boldsymbol{\epsilon}} G[\boldsymbol{L}] = \mathbb{E}_{\boldsymbol{\epsilon}} \langle \boldsymbol{K}\boldsymbol{L}\boldsymbol{y}, \boldsymbol{L}\boldsymbol{y} \rangle - 2\mathbb{E}_{\boldsymbol{\epsilon}} \langle \boldsymbol{K}\boldsymbol{L}\boldsymbol{y}, \boldsymbol{\alpha}^* \rangle, \tag{11}$$

where $\mathbb{E}_{\boldsymbol{\epsilon}}$ is the expectation over the noise $\{\epsilon_i\}_{i=1}^n$ and $\boldsymbol{K}$ is the *kernel matrix*, i.e., the (i,j)-th element is given by

$$K_{i,j} = K(\boldsymbol{x}_i, \boldsymbol{x}_j). \tag{12}$$

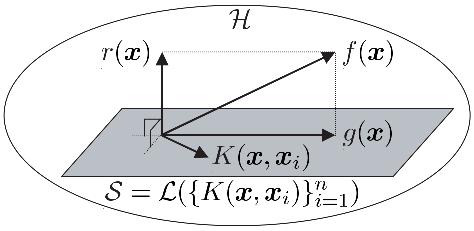
$\mathcal{H}$
$r(\boldsymbol{x})$
$f(\boldsymbol{x})$
$g(\boldsymbol{x})$
$K(\boldsymbol{x},\boldsymbol{x}_i)$
$\mathcal{S}=\mathcal{L}(\{K(\boldsymbol{x},\boldsymbol{x}_i)\}_{i=1}^n)$

Figure 2: Decomposition of the learning target function $f(\boldsymbol{x})$.

Since $\boldsymbol{\alpha}^*$ is unknown in Eq.(11), we replace it by a linear unbiased estimator $\widehat{\boldsymbol{\alpha}}_u$. (see Figure 3). Namely, with some n-dimensional matrix $\boldsymbol{R}_u$, $\widehat{\boldsymbol{\alpha}}_u$ is given as

$$\widehat{\boldsymbol{\alpha}}_u = \boldsymbol{R}_u \boldsymbol{y}, \tag{13}$$

which satisfies

$$\mathbb{E}_{\boldsymbol{\epsilon}} \widehat{\boldsymbol{\alpha}}_u = \boldsymbol{\alpha}^*. \tag{14}$$

Note that the subscript 'u' in the above equations stands for 'unbiased'. It is known that such $\boldsymbol{R}_u$ is given as follows [18].

$$\boldsymbol{R}_u = \boldsymbol{K}^\dagger, \tag{15}$$

where † denotes the Moore-Penrose generalized inverse [2].

Using $\widehat{\boldsymbol{\alpha}}_u$, we can express $\mathbb{E}_{\boldsymbol{\epsilon}} G$ as

$$\mathbb{E}_{\boldsymbol{\epsilon}} G[\boldsymbol{L}] =\mathbb{E}_{\boldsymbol{\epsilon}} \langle \boldsymbol{K}\boldsymbol{L}\boldsymbol{y}, \boldsymbol{L}\boldsymbol{y}\rangle - 2\mathbb{E}_{\boldsymbol{\epsilon}} \langle \boldsymbol{K}\boldsymbol{L}\boldsymbol{y}, \boldsymbol{R}_u\boldsymbol{y}\rangle + 2\sigma^2 \mathrm{tr}(\boldsymbol{K}\boldsymbol{L}\boldsymbol{R}_u^\top). \tag{16}$$

The subspace information criterion (SIC) is defined as the right-hand side of Eq.(16) with the expectation operator $\mathbb{E}_{\boldsymbol{\epsilon}}$ removed:

$$\mathrm{SIC}[\boldsymbol{L}] =\langle \boldsymbol{K}\boldsymbol{L}\boldsymbol{y}, \boldsymbol{L}\boldsymbol{y}\rangle - 2\langle \boldsymbol{K}\boldsymbol{L}\boldsymbol{y}, \boldsymbol{R}_u\boldsymbol{y}\rangle + 2\sigma^2 \mathrm{tr}(\boldsymbol{K}\boldsymbol{L}\boldsymbol{R}_u^\top). \tag{17}$$

For any $\boldsymbol{L}$, SIC is an unbiased estimator of $\mathbb{E}_{\boldsymbol{\epsilon}} G$.

$$\mathbb{E}_{\boldsymbol{\epsilon}} \mathrm{SIC}[\boldsymbol{L}] = \mathbb{E}_{\boldsymbol{\epsilon}} G[\boldsymbol{L}]. \tag{18}$$

The papers [19, 18] proposed choosing the learning matrix $\boldsymbol{L}$ that minimizes SIC from a set $\boldsymbol{\mathcal{L}}$ of candidates of $\boldsymbol{L}$:

$$\widehat{\boldsymbol{L}} = \mathop{\mathrm{argmin}}_{\boldsymbol{L}\in\boldsymbol{\mathcal{L}}} \mathrm{SIC}[\boldsymbol{L}]. \tag{19}$$

3.2 Regularized Subspace Information Criterion

It is reported that a good learning matrix $\boldsymbol{L}$ can be obtained by SIC [19, 18]. However, the goodness of SIC is only guaranteed in the sense of unbiasedness. This implies that the variance of SIC can be large, e.g., when the noise level is very high. In such cases, learning with SIC can be unstable. To cope with this problem, the regularized SIC (RSIC) has been proposed [17]. Below, we briefly review RSIC.

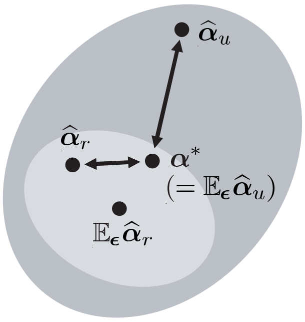
$\widehat{\boldsymbol{\alpha}}_u$
$\widehat{\boldsymbol{\alpha}}_r$
$\boldsymbol{\alpha}^*$
$(=\mathbb{E}_{\boldsymbol{\epsilon}}\widehat{\boldsymbol{\alpha}}_u)$
$\mathbb{E}_{\boldsymbol{\epsilon}}\widehat{\boldsymbol{\alpha}}_r$

Figure 3: Replacing unknown $\boldsymbol{\alpha}^*$ by an unbiased estimator $\widehat{\boldsymbol{\alpha}}_u$ (SIC) or by a regularized estimator $\widehat{\boldsymbol{\alpha}}_r$ (RSIC).

Let $\widehat{\boldsymbol{\alpha}}_r$ be some linear regularized estimator of $\boldsymbol{\alpha}^*$:

$$\widehat{\boldsymbol{\alpha}}_r = \boldsymbol{R}\boldsymbol{y}, \tag{20}$$

where $\boldsymbol{R}$ is an n-dimensional matrix which is independent of the noise $\{\epsilon_i\}_{i=1}^n$. We call $\boldsymbol{R}$ the reference matrix since $\widehat{\boldsymbol{\alpha}}_r$ is used as a reference.

A major reason why SIC can have large variance would be the instability of $\widehat{\boldsymbol{\alpha}}_u$. A basic idea of RSIC is to replace the unbiased estimator $\widehat{\boldsymbol{\alpha}}_u$ with a biased but more stable estimator $\widehat{\boldsymbol{\alpha}}_r$: (see Figure 3 again):

$$\mathrm{RSIC}[\boldsymbol{L};\boldsymbol{R}] =\langle \boldsymbol{KLy},\boldsymbol{Ly}\rangle - 2\langle \boldsymbol{KLy},\boldsymbol{Ry}\rangle + 2\sigma^2\mathrm{tr}(\boldsymbol{KLR}^\top), \tag{21}$$

where the notation $\mathrm{RSIC}[\boldsymbol{L};\boldsymbol{R}]$ means that it is a functional of $\boldsymbol{L}$ with a 'parameter' matrix $\boldsymbol{R}$.

In RSIC, the parameter matrix $\boldsymbol{R}$ should be determined appropriately[2]. To this end, we need a goodness measure of $\boldsymbol{R}$. The paper [17] proposed using the following criterion.

$$J[\boldsymbol{R};\boldsymbol{L}] = (\mathrm{RSIC}[\boldsymbol{L};\boldsymbol{R}] - \mathbb{E}_{\boldsymbol{\epsilon}}G[\boldsymbol{L}])^2, \tag{22}$$

where the notation $J[\boldsymbol{R};\boldsymbol{L}]$ means that it is a functional of $\boldsymbol{R}$ with a parameter matrix $\boldsymbol{L}$. Now we want to determine $\boldsymbol{R}$ so that the above J is minimized. However, J includes unknown G so it can not be directly calculated. Let $\boldsymbol{B}$ and $\boldsymbol{C}$ be

$$\boldsymbol{B} =2\boldsymbol{R}_u^\top\boldsymbol{KL} - 2\boldsymbol{R}^\top\boldsymbol{KL}, \tag{23}$$

$$\boldsymbol{C} =\boldsymbol{L}^\top\boldsymbol{KL} - 2\boldsymbol{R}^\top\boldsymbol{KL}. \tag{24}$$

Then an unbiased estimator of $\mathbb{E}_{\boldsymbol{\epsilon}}J$ is given as follows [17].

$$\begin{aligned}\widehat{J}[\boldsymbol{R};\boldsymbol{L}] = &\left\{\langle \boldsymbol{By},\boldsymbol{y}\rangle - \sigma^2\mathrm{tr}(\boldsymbol{B})\right\}^2\\ &- \sigma^2\|(\boldsymbol{B}+\boldsymbol{B}^\top)\boldsymbol{y}\|^2 + \sigma^4\mathrm{tr}(\boldsymbol{B}^2+\boldsymbol{BB}^\top)\\ &+ \sigma^2\|(\boldsymbol{C}+\boldsymbol{C}^\top)\boldsymbol{y}\|^2 - \sigma^4\mathrm{tr}(\boldsymbol{C}^2+\boldsymbol{CC}^\top),\end{aligned} \tag{25}$$

[2] If we have a good $\boldsymbol{L}$, it may be appropriate to use it as $\boldsymbol{R}$. However, obtaining a good $\boldsymbol{L}$ is the goal here and thus we need to search for a good $\boldsymbol{R}$.

which satisfies, for any $\boldsymbol{R}$ and $\boldsymbol{L}$,

$$\mathbb{E}_{\boldsymbol{\epsilon}}\widehat{J}[\boldsymbol{R};\boldsymbol{L}] = \mathbb{E}_{\boldsymbol{\epsilon}}J[\boldsymbol{R};\boldsymbol{L}]. \tag{26}$$

The paper [17] proposed using the above $\widehat{J}$ instead of J for determining $\boldsymbol{R}$.

Learning $\boldsymbol{L}$ based on RSIC and $\widehat{J}$ is carried out as follows. First, a set $\mathcal{L}$ of candidates of $\boldsymbol{L}$ and a set $\mathcal{R}$ of candidates of $\boldsymbol{R}$ are prepared. For each $\boldsymbol{L} \in \mathcal{L}$, $\boldsymbol{R}$ is optimized within $\mathcal{R}$:

$$\widehat{\boldsymbol{R}}_{\boldsymbol{L}} = \operatorname*{argmin}_{\boldsymbol{R}\in\mathcal{R}} \widehat{J}[\boldsymbol{R};\boldsymbol{L}]. \tag{27}$$

Then, using $\widehat{\boldsymbol{R}}_{\boldsymbol{L}}$, we optimize $\boldsymbol{L}$ within $\mathcal{L}$:

$$\widehat{\boldsymbol{L}} = \operatorname*{argmin}_{\boldsymbol{L}\in\mathcal{L}} \mathrm{RSIC}[\boldsymbol{L};\widehat{\boldsymbol{R}}_{\boldsymbol{L}}]. \tag{28}$$

3.3 Learning Methods

When we learn $\boldsymbol{L}$ using RSIC, we have to determine the set $\mathcal{L}$ from which $\boldsymbol{L}$ is searched and the set $\mathcal{R}$ from which $\boldsymbol{R}$ is searched. The largest possible set is $\mathbb{R}^n$, but it is generally too broad to be searched from. Conventionally, we form the set $\mathcal{L}$ and the set $\mathcal{R}$ based on some learning criterion. For example, in the case of *ridge learning* [8, 22, 13], the parameter $\boldsymbol{\alpha}$ is determined so that the regularized squared error is minimized[3].

$$\sum_{i=1}^{n}\left(\widehat{f}(\boldsymbol{x}_i) - y_i\right)^2 + \lambda\|\boldsymbol{\alpha}\|^2, \tag{29}$$

where λ is a non-negative scalar called the ridge parameter. A minimizer of the above regularized squared error is given by

$$\boldsymbol{L} = (\boldsymbol{K}^2 + \lambda\boldsymbol{I})^{\dagger}\boldsymbol{K}, \tag{30}$$

where $\boldsymbol{I}$ is the identity matrix. In the following, we focus on ridge learning. Then the problem of choosing the learning matrix $\boldsymbol{L}$ is reduced to the problem of choosing the ridge parameter λ.

When RSIC is employed, we have to optimize the reference matrix $\boldsymbol{R}$ in addition to the learning matrix $\boldsymbol{L}$. Below, we focus on using ridge learning also for $\boldsymbol{R}$:

$$\boldsymbol{R} = (\boldsymbol{K}^2 + \gamma\boldsymbol{I})^{\dagger}\boldsymbol{K}, \tag{31}$$

where γ is a non-negative scalar. Now the problem of choosing $\boldsymbol{R}$ and $\boldsymbol{L}$ is reduced to the problem of choosing γ and λ.

The paper [17] proposed determining the ridge parameter λ based on RSIC as follows. First, a finite set of candidate values of λ and a finite set of candidate values of γ are prepared. For each λ, γ is optimized based on $\widehat{J}$. Then λ is optimized based on RSIC using the chosen γ.

[3] This type of learning method is also referred to as *least squares support vector machines* [21] or *kernel regularized least squares* [4]; *Kernel Fisher discriminant analysis* [12] may also be regarded as the same type.

4 New Criterion for Determining $\boldsymbol{R}$

In this section, we first point out the drawbacks of the conventional RSIC-based model selection method and then propose a new method that can systematically overcome the problems.

4.1 Drawbacks in Existing Approach

RSIC as well as $\widehat{J}$ include the noise variance σ^2 in their definitions. However, since the noise variance σ^2 is generally unknown, it is practically replaced by an estimator, for example, an unbiased estimator $\widehat{\sigma}^2$.

$$\widehat{\sigma}^2 = \frac{\langle \boldsymbol{V}\boldsymbol{y}, \boldsymbol{y} \rangle}{\mathrm{tr}(\boldsymbol{V})}, \tag{32}$$

where $\boldsymbol{V} = \boldsymbol{I} - \boldsymbol{K}\boldsymbol{K}^{\dagger}$and which satisfies[4]

$$\mathbb{E}_{\boldsymbol{\epsilon}} \widehat{\sigma}^2 = \sigma^2. \tag{33}$$

Let $\widehat{J}'$ be the criterion with σ^2 in $\widehat{J}$ replaced by $\widehat{\sigma}^2$.

$$\begin{aligned} \widehat{J}'[\boldsymbol{R}; \boldsymbol{L}] = & \left\{ \langle \boldsymbol{B}\boldsymbol{y}, \boldsymbol{y} \rangle - \widehat{\sigma}^2 \mathrm{tr}(\boldsymbol{B}) \right\}^2 \\ & - \widehat{\sigma}^2 \|(\boldsymbol{B} + \boldsymbol{B}^\top)\boldsymbol{y}\|^2 + \widehat{\sigma}^4 \mathrm{tr}(\boldsymbol{B}^2 + \boldsymbol{B}\boldsymbol{B}^\top) \\ & + \widehat{\sigma}^2 \|(\boldsymbol{C} + \boldsymbol{C}^\top)\boldsymbol{y}\|^2 - \widehat{\sigma}^4 \mathrm{tr}(\boldsymbol{C}^2 + \boldsymbol{C}\boldsymbol{C}^\top). \end{aligned} \tag{34}$$

Similarly, let RSIC$'$ be the criterion with σ^2 in RSIC replaced by $\widehat{\sigma}^2$.

$$\mathrm{RSIC}'[\boldsymbol{L}; \boldsymbol{R}] = \langle \boldsymbol{K}\boldsymbol{L}\boldsymbol{y}, \boldsymbol{L}\boldsymbol{y} \rangle - 2\langle \boldsymbol{K}\boldsymbol{L}\boldsymbol{y}, \boldsymbol{R}\boldsymbol{y} \rangle + 2\widehat{\sigma}^2 \mathrm{tr}(\boldsymbol{K}\boldsymbol{L}\boldsymbol{R}^\top). \tag{35}$$

There are five problems in this conventional approach.

(a) J evaluates the goodness of RSIC, but RSIC$'$ is used for model selection in reality.

(b) Minimizing J means deciding RSIC so that it is close to the expected generalization error $\mathbb{E}_{\boldsymbol{\epsilon}} G$. This implies that in each single trial, RSIC is not necessarily a good approximation to G.

(c) J is a goodness measure specialized for RSIC. It is nice to employ a more general goodness measure that is applicable to a wider class of estimators of G.

[4] In some RKHSs such as the Gaussian RKHS, the kernel matrix always has full rank theoretically if $\{\boldsymbol{x}_i\}_{i=1}^n$ are all distinct [15]. In such cases, $\widehat{\sigma}^2$ can not be defined since the denominator is zero. However, in practice, $\boldsymbol{K}$ is numerically degenerated and $\widehat{\sigma}^2$ may still be used (see also Section 5.2).

(d) Eq.(22) can be decomposed as

$$J = \mathrm{RSIC}^2 - 2\mathrm{RSIC} \cdot \mathbb{E}_{\boldsymbol{\epsilon}} G + (\mathbb{E}_{\boldsymbol{\epsilon}} G)^2. \tag{36}$$

The third term in the right-hand side of Eq.(36) is a constant and it does not depend on RSIC. Since the constant does not affect the choice of $\boldsymbol{R}$, we do not need to estimate it. This implies that the conventional method also estimates the irrelevant term, which simply increases the computational cost (although we should admit the difference is subtle in practice).

(e) Replacing σ^2 by $\widehat{\sigma}^2$ generally breaks the unbiasedness, i.e.,

$$\mathbb{E}_{\boldsymbol{\epsilon}} \widehat{J}' \neq \mathbb{E}_{\boldsymbol{\epsilon}} J. \tag{37}$$

The purpose of this paper is to systematically solve the above five problems.

4.2 Proposed Meta-Criterion

We propose an alternative method that can settle the above drawbacks.

The problem (a) can be solved by replacing RSIC with RSIC$'$ in J:

$$(\mathrm{RSIC}'[\boldsymbol{L}; \boldsymbol{R}] - \mathbb{E}_{\boldsymbol{\epsilon}} G[\boldsymbol{L}])^2. \tag{38}$$

The problem (b) can be solved by using the squared error between RSIC$'$ and G, i.e., $\mathbb{E}_{\boldsymbol{\epsilon}} G$ in Eq.(38) is replaced by G:

$$(\mathrm{RSIC}'[\boldsymbol{L}; \boldsymbol{R}] - G[\boldsymbol{L}])^2. \tag{39}$$

The problem (c) can be eased by using a general form of the estimator of G. Let $\widetilde{G}$ be a *quadratic* estimator of G:

$$\widetilde{G}[\boldsymbol{L}; \boldsymbol{R}] = \langle \boldsymbol{H}\boldsymbol{y}, \boldsymbol{y} \rangle, \tag{40}$$

where $\boldsymbol{H}$ is some n-dimensional matrix. Note that $\widetilde{G}$ includes RSIC$'$ as a special case; indeed putting

$$\boldsymbol{H} = \boldsymbol{L}^\top \boldsymbol{K} \boldsymbol{L} - 2\boldsymbol{R}^\top \boldsymbol{K} \boldsymbol{L} + \frac{2\mathrm{tr}(\boldsymbol{R}^\top \boldsymbol{K} \boldsymbol{L})}{\mathrm{tr}(\boldsymbol{V})} \boldsymbol{V} \tag{41}$$

yields $\widetilde{G} = \mathrm{RSIC}'$. For $\widetilde{G}$, Eq.(39) is expressed as

$$(\widetilde{G}[\boldsymbol{L}; \boldsymbol{R}] - G[\boldsymbol{L}])^2. \tag{42}$$

The problem (d) can be avoided by ignoring the constant term included in Eq.(42). Thus we propose the following criterion for measuring the goodness of $\boldsymbol{R}$.

$$J_{new}[\boldsymbol{R}; \boldsymbol{L}] = (\widetilde{G}[\boldsymbol{L}; \boldsymbol{R}] - G[\boldsymbol{L}])^2 - (G[\boldsymbol{L}])^2. \tag{43}$$

We want to determine $\boldsymbol{R}$ so that the above J_{new} is minimized. However, J_{new} includes the unknown G so it should be estimated. The remaining issue, the problem (e), could be solved by defining $\widehat{J}_{new}$ with the estimation error of σ^2 taken into account:

$$\begin{aligned}\widehat{J}_{new}[\boldsymbol{R};\boldsymbol{L}] =&\langle \boldsymbol{H}\boldsymbol{y},\boldsymbol{y}\rangle^2 + 2\langle \boldsymbol{H}\boldsymbol{y},\boldsymbol{y}\rangle\langle(\boldsymbol{S}-\boldsymbol{T})\boldsymbol{y},\boldsymbol{y}\rangle\\ &- 2\widehat{\sigma}^2\langle(\boldsymbol{H}+\boldsymbol{H}^\top)\boldsymbol{S}\boldsymbol{y},\boldsymbol{y}\rangle - 2\widehat{\sigma}^2\mathrm{tr}(\boldsymbol{S})\langle\boldsymbol{H}\boldsymbol{y},\boldsymbol{y}\rangle\\ &+ 4\widehat{\sigma}^4\frac{\mathrm{tr}(\boldsymbol{V}(\boldsymbol{H}+\boldsymbol{H}^\top)\boldsymbol{S}) + \mathrm{tr}(\boldsymbol{S})\mathrm{tr}(\boldsymbol{V}\boldsymbol{H})}{\mathrm{tr}(\boldsymbol{V})+2},\end{aligned} \tag{44}$$

where

$$\boldsymbol{S} = 2\boldsymbol{R}_u^\top\boldsymbol{K}\boldsymbol{L}, \tag{45}$$

$$\boldsymbol{T} = \boldsymbol{L}^\top\boldsymbol{K}\boldsymbol{L}. \tag{46}$$

Then we have the following theorem.

Theorem 1 *For any $\boldsymbol{R}$ and any $\boldsymbol{L}$, we have*

$$\mathbb{E}_{\boldsymbol{\epsilon}}\widehat{J}_{new}[\boldsymbol{R};\boldsymbol{L}] = \mathbb{E}_{\boldsymbol{\epsilon}}J_{new}[\boldsymbol{R};\boldsymbol{L}]. \tag{47}$$

A proof of the above theorem is given in A. The above theorem shows that $\widehat{J}_{new}$ is an unbiased estimator of $\mathbb{E}_{\boldsymbol{\epsilon}}J_{new}$. Thus, the use of $\widehat{J}_{new}$ can resolve all the five problems of the existing approach listed in Section 4.1. We propose using $\widehat{J}_{new}$ for determining $\boldsymbol{R}$.

5 Simulations

In this section, we experimentally compare the generalization performance of the existing and proposed methods.

5.1 Illustrative Example

Let the learning target function be

$$f(x) = \mathrm{sinc}(x). \tag{48}$$

We employ the Gaussian reproducing kernel Hilbert space [15] as $\mathcal{H}$, where the reproducing kernel is given by

$$K(x,x') = \exp\left(-\frac{(x-x')^2}{2c^2}\right) \tag{49}$$

with $c = 1$. Note that the sinc function is included in the above Gaussian reproducing kernel Hilbert space [5]. We draw 10 points $\{x'_i\}_{i=1}^{10}$ independently from the uniform distribution on $(-\pi,\pi)$. Let the training input points $\{x_i\}_{i=1}^{20}$ be $x_i = x_{i+10} = x'_i$ for $i = 1,2,\ldots,10$, i.e., we duplicate the training input points twice[5]. Noise $\{\epsilon_i\}_{i=1}^{20}$ are drawn

[5] This setting guarantees that $\widehat{\sigma}^2$ defined by Eq.(32) is unbiased.

independently from the normal distribution with mean zero and variance σ^2. Training output values $\{y_i\}_{i=1}^{20}$ are created as

$$y_i = \mathrm{sinc}(x_i) + \epsilon_i. \tag{50}$$

We test $\sigma^2 = 0.04$ and 0.16. For each σ^2, we repeat the simulation 500 times by changing $\{x_i\}_{i=1}^{n}$ and $\{\epsilon_i\}_{i=1}^{n}$. In the experiments, σ^2 is treated as an unknown variable and is estimated by Eq.(32).

Among the five drawbacks of existing approach listed in Section 4.1, the problems (a), (b), and (e) affect the generalization performance. Here, we compare the performance of the following methods.

Existing: none of the problems (a), (b), and (e) are resolved.

Proposed: all the problems (a), (b), and (e) are resolved.

Reference (a): only the problem (a) is resolved.

Reference (b): only the problem (b) is resolved.

Reference (e): only the problem (e) is resolved.

Thus the purposes of the experiments are to numerically evaluate whether the proposed method works better than the existing method and which improvement is the most crucial. λ and γ are chosen from the following sets Λ and Γ:

$$\Lambda = \Gamma = \{10^{-3}, 10^{-2.5}, 10^{-2}, \ldots, 10^{3}\}. \tag{51}$$

So far, we called G the generalization error, where $\|f\|^2$ is ignored (see Eq.(3)). With some abuse, we call the following $\overline{G}$ the generalization error through this section.

$$\overline{G} = \|\widehat{f} - f\|^2 = G + \|f\|^2. \tag{52}$$

In the simulation, we approximately compute the value of $\overline{G}$ by replacing f with $\widetilde{f}$, where $\widetilde{f}$ is an approximation of f obtained using the kernel regression model with the same Gaussian kernel (49) and a large number of artificially generated noiseless samples. This approximation seems to be accurate enough for the current experiments.

The mean and standard deviation of the generalization error obtained by each method over 500 runs are described in the upper half of Table 1. All the values are normalized by the mean generalization error of the existing method for better comparison. The better method by the *Wilcoxon signed rank test* [7] at the significance level 5% is indicated by '$*$'. The table shows that the proposed method is significantly better than the existing method when $\sigma^2 = 0.16$ and they are comparable when $\sigma^2 = 0.04$. For this illustrative simulation, the reference (b) also works very well.

Table 1: Mean generalization errors over repetitions. The numbers in the bracket are standard deviations. All the values are normalized by the mean generalization error of the existing method for better comparison. The best method and comparable ones by the Wilcoxon signed rank test at the significance level 5% is indicated by '∗'.

Data Set	Existing	Proposed	Reference(a)	Reference(b)	Reference(e)
Toy ($\sigma^2 = 0.04$)	∗1.0000(0.3930)	∗0.9941(0.3825)	∗1.0058(0.3936)	∗0.9914(0.3838)	1.2133(0.3765)
Toy ($\sigma^2 = 0.16$)	1.0000(0.3450)	∗0.9864(0.3450)	1.0022(0.3465)	∗0.9833(0.3475)	1.0503(0.2527)
Kin-8fm	∗1.0000(0.1794)	1.0316(0.1911)	1.0242(0.1824)	∗0.9999(0.1811)	2.5880(0.3747)
Kin-8nm	1.0000(0.1087)	∗0.9987(0.1076)	∗0.9982(0.1080)	1.0004(0.1088)	1.0902(0.0922)
Kin-8fh	∗1.0000(0.0931)	∗1.0017(0.0925)	1.0025(0.0925)	∗0.9997(0.0920)	1.2275(0.1719)
Kin-8nh	1.0000(0.1148)	∗0.9883(0.1084)	0.9923(0.1102)	1.0011(0.1155)	1.0196(0.1175)
Pumadyn-8fm	∗1.0000(0.0880)	∗1.0017(0.0891)	1.0036(0.0910)	∗0.9995(0.0884)	1.2386(0.1927)
Pumadyn-8nm	1.0000(0.1078)	0.9971(0.1076)	0.9975(0.1076)	0.9998(0.1075)	∗0.9387(0.0725)
Pumadyn-8fh	1.0000(0.1132)	∗0.9883(0.1045)	∗0.9896(0.1079)	0.9995(0.1120)	1.0212(0.1168)
Pumadyn-8nh	1.0000(0.1297)	0.9830(0.1293)	0.9853(0.1303)	1.0007(0.1291)	∗0.9056(0.0910)

5.2 Real Data Sets

Next we investigate the effectiveness of the proposed method using real data sets. We use 8 practical data sets provided by DELVE [14]: *Kin-8fm*, *Kin-8nm*, *Kin-8fh*, *Kin-8nh*, *Pumadyn-8fm*, *Pumadyn-8nm*, *Pumadyn-8fh*, and *Pumadyn-8nh*.

Each of the *Bank*, *Kin*, and *Pumadyn* data family consists of four different data sets. They are labeled as 'fm', 'nm', 'fh', and 'nh', where 'f' or 'n' signifies 'fairly linear' or 'non-linear', respectively, and 'm' or 'h' signifies 'medium unpredictability/noise' or 'high unpredictability/noise', respectively. Each data set includes 8192 samples, each of which consists of 8-dimensional input and 1-dimensional output data. For convenience, every input attribute is normalized to $[0,1]$. 100 randomly selected samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{100}$ are used for training. In the real data set, we can not measure the generalization error by Eq.(52) since the true function f is totally unknown. Instead, we evaluate the generalization performance by the mean squared test error defined by

$$\frac{1}{1000}\sum_{i=1}^{1000}\left(\widehat{f}(\boldsymbol{x}'_i) - y'_i\right)^2, \tag{53}$$

where $\{(\boldsymbol{x}'_i, y'_i)\}_{i=1}^{1000}$ denote the randomly chosen test samples which are not used for training. A Gaussian kernel with width $c = 1$ is again employed (see Eq.(49)). λ and γ are chosen from the following values:

$$\lambda, \gamma \in \{10^{-4}, 10^{-3.5}, 10^{-3}, \ldots, 10^{4}\}. \tag{54}$$

The simulation is repeated 500 times, randomly selecting the training set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{100}$ in each trial. In this simulation, the computation of the Moore-Penrose generalized inverse was rather unstable. To avoid numerical troubles, we discarded eigenvalues less than 10^{-2}.

The mean and standard deviation of the generalization error obtained by each method over 500 runs are described in the lower half of Table 1. The table shows that the proposed method tends to outperform the existing method particularly for the data sets with high noise. None of the reference methods is comparable to the proposed method. This implies that each single improvement is not enough, but combining all three improvements together can result in good performance.

6 Conclusions

In this paper, we first pointed out five drawbacks of the existing meta-criterion for RSIC and proposed an alternative one that can solve all the drawbacks. We experimentally showed that the proposed method improves the accuracy of RSIC especially in the high noise level cases.

The original RSIC was experimentally shown to compare favorably with cross validation and an empirical Bayesian method [17] (see also the paper [16] for theoretical discussions). Therefore, the proposed method could be regarded as a useful model selection method in terms of the accuracy. However, since RSIC includes an additional tuning parameters $\boldsymbol{R}$, it may have higher computational costs than leave-one-out cross validation, given the fact that the leave-one-out cross validation score can be analytically and thus efficiently computed for kernel ridge regression [24]. Therefore, our important future work is to improve the computational cost of the new RSIC—since the expression of RSIC is simpler than that of the analytic form of the leave-one-out cross validation score, we expect that the optimal $\boldsymbol{L}$ and $\boldsymbol{R}$ are analytically obtained, e.g., following the lines of [20, 6].

Acknowledgments

This work was supported by MEXT Grant-in-Aid for Young Scientists 17700142 and Grant-in-Aid for Scientific Research (B) 18300057.

References

[1] H. Akaike, “A new look at the statistical model identification,” IEEE Transactions on Automatic Control, vol.AC-19, no.6, pp.716–723, 1974.

[2] A. Albert, Regression and the Moore-Penrose Pseudoinverse, Academic Press, New York and London, 1972.

[3] N. Aronszajn, “Theory of reproducing kernels,” Transactions of the American Mathematical Society, vol.68, pp.337–404, 1950.

[4] T. Evgeniou, M. Pontil, and T. Poggio, “Regularization networks and support vector machines,” Advances in Computational Mathematics, vol.13, no.1, pp.1–50, 2000.

[5] F. Girosi, “An equivalence between sparse approximation and support vector machines,” Neural Computation, vol.10, no.6, pp.1455–1480, 1998.

[6] S. Gokita, M. Sugiyama, and K. Sakurai, “Analytic optimization of adaptive ridge parameters based on regularized subspace information criterion.” submitted.

[7] R.E. Henkel, Tests of Significance, SAGE Publication, Beverly Hills, 1979.

[8] A.E. Hoerl and R.W. Kennard, “Ridge regression: Biased estimation for nonorthogonal problems,” Technometrics, vol.12, no.3, pp.55–67, 1970.

[9] G.S. Kimeldorf and G. Wahba, “Some results on Tchebycheffian spline functions,” Journal of Mathematical Analysis and Applications, vol.33, no.1, pp.82–95, 1971.

[10] E.L. Lehmann, Theory of Point Estimation, Wiley, New York, 1983.

[11] C.L. Mallows, “Some comments on C_P,” Technometrics, vol.15, no.4, pp.661–675, 1973.

[12] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, A. Smola, and K.R. Müller, "Constructing descriptive and discriminative nonlinear features: Rayleigh coefficients in kernel feature spaces," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.25, no.5, pp.623–628, 2003.

[13] T. Poggio and F. Girosi, "Networks for approximation and learning," Proceedings of the IEEE, vol.78, no.9, pp.1481–1497, 1990.

[14] C.E. Rasmussen, R.M. Neal, G.E. Hinton, D. van Camp, M. Revow, Z. Ghahramani, R. Kustra, and R. Tibshirani, "The DELVE manual," 1996.

[15] B. Schölkopf and A.J. Smola, Learning with Kernels, MIT Press, Cambridge, MA, 2002.

[16] M. Sugiyama, "Supervised learning under covariate shift," The Brain & Neural Networks, vol.13, no.3, pp.111–118, 2006. in Japanese.

[17] M. Sugiyama, M. Kawanabe, and K.R. Müller, "Trading variance reduction with unbiasedness: The regularized subspace information criterion for robust model selection in kernel regression," Neural Computation, vol.16, no.5, pp.1077–1104, 2004.

[18] M. Sugiyama and K.R. Müller, "The subspace information criterion for infinite dimensional hypothesis spaces," Journal of Machine Learning Research, vol.3, pp.323–359, Nov. 2002.

[19] M. Sugiyama and H. Ogawa, "Subspace information criterion for model selection," Neural Computation, vol.13, no.8, pp.1863–1889, 2001.

[20] M. Sugiyama and K. Sakurai, "Analytic optimization of shrinkage parameters based on regularized subspace information criterion," IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol.E89-A, no.8, pp.2216–2225, 2006.

[21] J.A.K. Suykens, T.V. Gestel, J.D. Brabanter, B.D. Moor, and J. Vandewalle, Least Squares Support Vector Machines, World Scientific Pub. Co., Singapore, 2002.

[22] A.N. Tikhonov and V.Y. Arsenin, Solutions of Ill-Posed Problems, V. H. Winston, Washington DC, 1977.

[23] V.N. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.

[24] G. Wahba, Spline Model for Observational Data, Society for Industrial and Applied Mathematics, Philadelphia and Pennsylvania, 1990.

A Proof of Theorem 1

We first show three lemmas which will be used for proving Theorem 1. Let

$$\boldsymbol{z} = (f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), \ldots, f(\boldsymbol{x}_n))^\top. \tag{55}$$

Lemma 1 For any matrix $\boldsymbol{P}$ and $\boldsymbol{Q}$, it holds that

$$\mathbb{E}_{\boldsymbol{\epsilon}}\langle \boldsymbol{P\epsilon}, \boldsymbol{\epsilon}\rangle\langle \boldsymbol{Q\epsilon}, \boldsymbol{\epsilon}\rangle = \sigma^4 \mathrm{tr}(\boldsymbol{P})\mathrm{tr}(\boldsymbol{Q}) + \sigma^4 \mathrm{tr}(\boldsymbol{PQ} + \boldsymbol{P}^\top \boldsymbol{Q}). \tag{56}$$

(Proof of Lemma 1) It holds that

$$\mathbb{E}_{\boldsymbol{\epsilon}}\langle \boldsymbol{P\epsilon}, \boldsymbol{\epsilon}\rangle\langle \boldsymbol{Q\epsilon}, \boldsymbol{\epsilon}\rangle = \mathbb{E}_{\boldsymbol{\epsilon}} \sum_{i,j,k,l=1}^{n} P_{i,j} Q_{k,l} \epsilon_i \epsilon_j \epsilon_k \epsilon_l, \tag{57}$$

where $P_{i,j}$ and $Q_{i,j}$ denote the (i,j)-th elements of $\boldsymbol{P}$ and $\boldsymbol{Q}$, respectively. It is known that when the random variable ϵ_i is drawn from the normal distribution with mean zero and variance σ^2, it holds that $\mathbb{E}_{\boldsymbol{\epsilon}}\epsilon_i^4 = 3\sigma^4$ (e.g., [10]). This implies that all terms in $\mathbb{E}_{\boldsymbol{\epsilon}} \sum_{i,j,k,l=1}^{n} P_{i,j} Q_{k,l} \epsilon_i \epsilon_j \epsilon_k \epsilon_l$ vanish except four cases: $i = j = k = l$, $i = j \neq k = l$, $i = k \neq j = l$, and $i = l \neq j = k$. Therefore, we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\epsilon}}\langle \boldsymbol{P\epsilon}, \boldsymbol{\epsilon}\rangle\langle \boldsymbol{Q\epsilon}, \boldsymbol{\epsilon}\rangle =& \mathbb{E}_{\boldsymbol{\epsilon}}\big(\sum_{i} P_{i,i}Q_{i,i}\epsilon_i^4 + \sum_{i\neq k} P_{i,i}Q_{k,k}\epsilon_i^2\epsilon_k^2 + \sum_{i\neq j} P_{i,j}Q_{i,j}\epsilon_i^2\epsilon_j^2 + \sum_{i\neq j} P_{i,j}Q_{j,i}\epsilon_i^2\epsilon_j^2\big) \\
=& 3\sigma^4 \sum_{i} P_{i,i}Q_{i,i} + \sigma^4 \sum_{i\neq j} P_{i,i}Q_{j,j} + \sigma^4 \sum_{i\neq j} P_{i,j}Q_{i,j} + \sigma^4 \sum_{i\neq j} P_{i,j}Q_{j,i} \\
=& \sigma^4\big(\sum_{i} P_{i,i}Q_{i,i} + \sum_{i\neq j} P_{i,i}Q_{j,j}\big) + \sigma^4\big(\sum_{i} P_{i,i}Q_{i,i} + \sum_{i\neq j} P_{i,j}Q_{i,j}\big) \\
&+ \sigma^4\big(\sum_{i} P_{i,i}Q_{i,i} + \sum_{i\neq j} P_{i,j}Q_{j,i}\big) \\
=& \sigma^4 \sum_{i,j} P_{i,i}Q_{j,j} + \sigma^4 \sum_{i,j} P_{i,j}Q_{i,j} + \sigma^4 \sum_{i,j} P_{i,j}Q_{j,i} \\
=& \sigma^4 \mathrm{tr}(\boldsymbol{P})\mathrm{tr}(\boldsymbol{Q}) + \sigma^4 \mathrm{tr}(\boldsymbol{P}^\top\boldsymbol{Q}) + \sigma^4 \mathrm{tr}(\boldsymbol{PQ}),
\end{aligned} \tag{58}$$

which yields Eq.(56). ■

Lemma 2 It holds that

$$\sigma^4 = \mathbb{E}_{\boldsymbol{\epsilon}} \widehat{\sigma}^4 \frac{\mathrm{tr}(\boldsymbol{V})}{\mathrm{tr}(\boldsymbol{V}) + 2}. \tag{59}$$

(Proof of Lemma 2) From the definition of $\boldsymbol{V}$, we have $\boldsymbol{V}^\top = \boldsymbol{V}$, $\boldsymbol{V}^2 = \boldsymbol{V}$ and $\boldsymbol{V}\boldsymbol{z} = 0$. Then $\mathbb{E}_{\boldsymbol{\epsilon}}\widehat{\sigma}^4$ is expressed as follows.

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\epsilon}}\widehat{\sigma}^4 &= \frac{\mathbb{E}_{\boldsymbol{\epsilon}}\langle \boldsymbol{V}\boldsymbol{y}, \boldsymbol{y}\rangle^2}{\mathrm{tr}(\boldsymbol{V})^2} \\ &= \frac{\mathbb{E}_{\boldsymbol{\epsilon}}\big(\langle \boldsymbol{V}\boldsymbol{z}, \boldsymbol{z}\rangle + \langle (\boldsymbol{V}+\boldsymbol{V}^\top)\boldsymbol{z}, \boldsymbol{\epsilon}\rangle + \langle \boldsymbol{V}\boldsymbol{\epsilon}, \boldsymbol{\epsilon}\rangle\big)^2}{\mathrm{tr}(\boldsymbol{V})^2} \\ &= \frac{\mathbb{E}_{\boldsymbol{\epsilon}}\langle \boldsymbol{V}\boldsymbol{\epsilon}, \boldsymbol{\epsilon}\rangle^2}{\mathrm{tr}(\boldsymbol{V})^2}. \end{aligned} \tag{60}$$

From Lemma 1, we have

$$\mathbb{E}_{\boldsymbol{\epsilon}}\langle \boldsymbol{V}\boldsymbol{\epsilon}, \boldsymbol{\epsilon}\rangle^2 = \sigma^4\mathrm{tr}(\boldsymbol{V})^2 + 2\sigma^4\mathrm{tr}(\boldsymbol{V}). \tag{61}$$

Eqs.(60) and (61) yield Eq.(59). ■

Lemma 3 For any matrix $\boldsymbol{P}$, it holds that

$$\sigma^2\langle \boldsymbol{P}\boldsymbol{z}, \boldsymbol{z}\rangle = \mathbb{E}_{\boldsymbol{\epsilon}}\widehat{\sigma}^2\langle \boldsymbol{P}\boldsymbol{y}, \boldsymbol{y}\rangle - \mathbb{E}_{\boldsymbol{\epsilon}}\widehat{\sigma}^4\frac{\mathrm{tr}(\boldsymbol{V})\mathrm{tr}(\boldsymbol{P}) + 2\mathrm{tr}(\boldsymbol{V}\boldsymbol{P})}{\mathrm{tr}(\boldsymbol{V})+2}. \tag{62}$$

(Proof of Lemma 3) $\mathbb{E}_{\boldsymbol{\epsilon}}\widehat{\sigma}^2\langle \boldsymbol{P}\boldsymbol{y}, \boldsymbol{y}\rangle$ is expressed as

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\epsilon}}\widehat{\sigma}^2\langle \boldsymbol{P}\boldsymbol{y}, \boldsymbol{y}\rangle &= \frac{\mathbb{E}_{\boldsymbol{\epsilon}}\langle \boldsymbol{V}\boldsymbol{y}, \boldsymbol{y}\rangle\langle \boldsymbol{P}\boldsymbol{y}, \boldsymbol{y}\rangle}{\mathrm{tr}(\boldsymbol{V})} \\ &= \frac{\mathbb{E}_{\boldsymbol{\epsilon}}\langle \boldsymbol{V}\boldsymbol{\epsilon}, \boldsymbol{\epsilon}\rangle\langle \boldsymbol{P}\boldsymbol{y}, \boldsymbol{y}\rangle}{\mathrm{tr}(\boldsymbol{V})} \\ &= \sigma^2\langle \boldsymbol{P}\boldsymbol{z}, \boldsymbol{z}\rangle + \frac{\mathbb{E}_{\boldsymbol{\epsilon}}\langle (\boldsymbol{P}+\boldsymbol{P}^\top)\boldsymbol{z}, \boldsymbol{\epsilon}\rangle\langle \boldsymbol{V}\boldsymbol{\epsilon}, \boldsymbol{\epsilon}\rangle}{\mathrm{tr}(\boldsymbol{V})} + \frac{\mathbb{E}_{\boldsymbol{\epsilon}}\langle \boldsymbol{V}\boldsymbol{\epsilon}, \boldsymbol{\epsilon}\rangle\langle \boldsymbol{P}\boldsymbol{\epsilon}, \boldsymbol{\epsilon}\rangle}{\mathrm{tr}(\boldsymbol{V})}. \end{aligned} \tag{63}$$

It holds that

$$\mathbb{E}_{\boldsymbol{\epsilon}}\langle (\boldsymbol{P}+\boldsymbol{P}^\top)\boldsymbol{z}, \boldsymbol{\epsilon}\rangle\langle \boldsymbol{V}\boldsymbol{\epsilon}, \boldsymbol{\epsilon}\rangle = \mathbb{E}_{\boldsymbol{\epsilon}}\sum_{i,j,k,l=1}^{n}(P_{i,j}+P_{j,i})V_{k,l}z_i\epsilon_j\epsilon_k\epsilon_l. \tag{64}$$

It holds that $\mathbb{E}_{\boldsymbol{\epsilon}}\epsilon_i^3 = 0$ (e.g., [10]). This implies that all terms in $\mathbb{E}_{\boldsymbol{\epsilon}}\sum_{i,j,k,l=1}^{n}(P_{i,j}+P_{j,i})V_{k,l}z_i\epsilon_j\epsilon_k\epsilon_l$ vanish, i.e.,

$$\mathbb{E}_{\boldsymbol{\epsilon}}\langle (\boldsymbol{P}+\boldsymbol{P}^\top)\boldsymbol{z}, \boldsymbol{\epsilon}\rangle\langle \boldsymbol{V}\boldsymbol{\epsilon}, \boldsymbol{\epsilon}\rangle = 0. \tag{65}$$

On the other hand, from Lemma 1, it holds that

$$\mathbb{E}_{\boldsymbol{\epsilon}}\langle \boldsymbol{V}\boldsymbol{\epsilon}, \boldsymbol{\epsilon}\rangle\langle \boldsymbol{P}\boldsymbol{\epsilon}, \boldsymbol{\epsilon}\rangle = \sigma^4\mathrm{tr}(\boldsymbol{V})\mathrm{tr}(\boldsymbol{P}) + 2\sigma^4\mathrm{tr}(\boldsymbol{V}\boldsymbol{P}). \tag{66}$$

Eqs.(63), (65), (66) and (59) yield Eq.(62). ■

(Proof of Theorem 1) Eq.(43) is expressed as

$$\begin{aligned} J_{new}[\boldsymbol{R};\boldsymbol{L}] =&(\widetilde{G}[\boldsymbol{L};\boldsymbol{R}])^2 - 2\widetilde{G}[\boldsymbol{L};\boldsymbol{R}]G[\boldsymbol{L}] \\ =&\langle \boldsymbol{H}\boldsymbol{y},\boldsymbol{y}\rangle^2 + 2\langle \boldsymbol{H}\boldsymbol{y},\boldsymbol{y}\rangle\langle(\boldsymbol{S}-\boldsymbol{T})\boldsymbol{y},\boldsymbol{y}\rangle - 2\langle \boldsymbol{H}\boldsymbol{y},\boldsymbol{y}\rangle\langle \boldsymbol{S}\boldsymbol{y},\boldsymbol{\epsilon}\rangle. \end{aligned} \tag{67}$$

From Lemmas 1, 2 and 3, we have

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\epsilon}}\langle \boldsymbol{H}\boldsymbol{y},\boldsymbol{y}\rangle\langle \boldsymbol{S}\boldsymbol{y},\boldsymbol{\epsilon}\rangle =&\sigma^2\langle(\boldsymbol{H}+\boldsymbol{H}^\top)\boldsymbol{S}\boldsymbol{z},\boldsymbol{z}\rangle + \sigma^2\mathrm{tr}(\boldsymbol{S})\langle \boldsymbol{H}\boldsymbol{z},\boldsymbol{z}\rangle \\ &+ \sigma^4\mathrm{tr}(\boldsymbol{H})\mathrm{tr}(\boldsymbol{S}) + \sigma^4\mathrm{tr}(\boldsymbol{H}\boldsymbol{S}+\boldsymbol{H}^\top\boldsymbol{S}) \\ =&\mathbb{E}_{\boldsymbol{\epsilon}}\widehat{\sigma}^2\langle(\boldsymbol{H}+\boldsymbol{H}^\top)\boldsymbol{S}\boldsymbol{y},\boldsymbol{y}\rangle \\ &- \mathbb{E}_{\boldsymbol{\epsilon}}\widehat{\sigma}^4\frac{\mathrm{tr}(\boldsymbol{V})\mathrm{tr}((\boldsymbol{H}+\boldsymbol{H}^\top)\boldsymbol{S}) + 2\mathrm{tr}(\boldsymbol{V}(\boldsymbol{H}+\boldsymbol{H}^\top)\boldsymbol{S})}{\mathrm{tr}(\boldsymbol{V})+2} \\ &+ \mathbb{E}_{\boldsymbol{\epsilon}}\widehat{\sigma}^2\mathrm{tr}(\boldsymbol{S})\langle \boldsymbol{H}\boldsymbol{y},\boldsymbol{y}\rangle - \mathbb{E}_{\boldsymbol{\epsilon}}\widehat{\sigma}^4\frac{\mathrm{tr}(\boldsymbol{S})\big(\mathrm{tr}(\boldsymbol{V})\mathrm{tr}(\boldsymbol{H}) + 2\mathrm{tr}(\boldsymbol{V}\boldsymbol{H})\big)}{\mathrm{tr}(\boldsymbol{V})+2} \\ &+ \mathbb{E}_{\boldsymbol{\epsilon}}\widehat{\sigma}^4\frac{\mathrm{tr}(\boldsymbol{V})\mathrm{tr}(\boldsymbol{H})\mathrm{tr}(\boldsymbol{S})}{\mathrm{tr}(\boldsymbol{V})+2} + \mathbb{E}_{\boldsymbol{\epsilon}}\widehat{\sigma}^4\frac{\mathrm{tr}(\boldsymbol{V})\mathrm{tr}(\boldsymbol{H}\boldsymbol{S}+\boldsymbol{H}^\top\boldsymbol{S})}{\mathrm{tr}(\boldsymbol{V})+2} \\ =&\mathbb{E}_{\boldsymbol{\epsilon}}\widehat{\sigma}^2\langle(\boldsymbol{H}+\boldsymbol{H}^\top)\boldsymbol{S}\boldsymbol{y},\boldsymbol{y}\rangle + \mathbb{E}_{\boldsymbol{\epsilon}}\widehat{\sigma}^2\mathrm{tr}(\boldsymbol{S})\langle \boldsymbol{H}\boldsymbol{y},\boldsymbol{y}\rangle \\ &- 2\mathbb{E}_{\boldsymbol{\epsilon}}\widehat{\sigma}^4\frac{\mathrm{tr}(\boldsymbol{V}(\boldsymbol{H}+\boldsymbol{H}^\top)\boldsymbol{S}) + \mathrm{tr}(\boldsymbol{V}\boldsymbol{H})\mathrm{tr}(\boldsymbol{S})}{\mathrm{tr}(\boldsymbol{V})+2}. \end{aligned} \tag{68}$$

Eqs.(67), (68) and (44) yield Eq.(47). ■