Generalization Error of Ensemble Estimators

Naonori UEDA and Ryohei NAKANO

NTT Communication Science Laboratories
Hikaridai Seika-cho Soraku-gun Kyoto 619-02 Japan
Tel: +81 774 95 1823 Fax: +81 774 95 1839
e-mail: ueda@cslab.kecl.ntt.jp

ABSTRACT

It has been empirically shown that a better estimate with less generalization error can be obtained by averaging outputs of multiple estimators. This paper presents an analytical result for the generalization error of ensemble estimators. First, we derive a general expression of the *ensemble generalization error* by using factors of interest (bias, variance, covariance, and noise variance) and show how the generalization error is affected by each of them. Then, some special cases are investigated. Next, the result of a simulation is shown to verify our analytical result. A practically important problem of the ensemble approach, *ensemble dilemma*, is also discussed.

1. Introduction

One of the most important issues to learning systems is to construct estimators with high generalization performance. The most popular way is to find the best model for a target task. In general, however, it is difficult to estimate the best model with given finite samples. On the other hand, as compared to that obtained by a single estimator [1]-[4], combining multiple regression estimators has been shown to improve generalization error. This combining approach has recently attracted major concern in the neural network community because of its simplicity and theoretical interest. This approach has also been succesfully applied to pattern classification tasks [5][6].

The output of the combined regression estimator for some input is defined as the linear combination of outputs of multiple estimators. Recent work has provided optimal combination weights under some assumptions [3]. The most popular way to combine multiple estimators is the simple averaging of outputs of multiple estimators. Several theoretical studies on generalization error of these ensemble estimators have recently been made. Perrone was the first to show that the simple averaging of M estimators reduces the generelation error by a factor of M under uncorrelated and unbiased assumptions [2].

In this paper, we provide more general and explicit expressions and show how the generalization error of the ensemble estimator is affected by each of factors of interest (variance, covariance, bias, and noise variance). Some special cases where the models of estimators are the same and/or the outputs of estimators are uncorrelated are also investigated. Based on these expressions, we will provide some important observations for improving the generalization error due to the averaging. We verify our analytical result by showing the result of a simulation, using standard feedforward neural networks as regression estimators. An important open problem of the ensemble approach is discussed at the end of the paper.

2. Generalization Error of Single Estimator

The purpose of regression, particularly for nonlinear regressions such as feedforward neural networks, is to construct an estimator $f(\boldsymbol{x};\Theta)$ that approximates an unknown target function $g(\boldsymbol{x})$. This is done by using a given set of N training samples $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_N, y_N)\}$, where $\boldsymbol{x}_i \in \mathcal{R}^d, y_i \in \mathcal{R}$. Θ is a parameter vector and, in the case of a feedforward neural network, Θ corresponds to the weight vector. We assume that there is a functional relationship between the training pair $\boldsymbol{x}_i$ and y_i: $y_i = g(\boldsymbol{x}_i) + \epsilon$, where ϵ is the additive noise with zero mean ($\mathrm{E}\{\epsilon\} = 0$) and the finite variance ($\mathrm{Var}\{\epsilon\} = \sigma^2 < \infty$). For convenience, let $z^N = \{z_1, \ldots, z_N\}$ denote the training set, where $z_i = (\boldsymbol{x}_i, y_i)$. z^N is a realization of a random sequence

0-7803-3210-5/96 $4.00

$Z^N = \{Z_1, \ldots, Z_N\}$, whose ith component consists of a random vector $Z_i = (X_i, Y_i)$. In other words, each z_i is an independent and identically distributed (i.i.d.) sample from an unknown joint probability distribution $p(\boldsymbol{x}, y)$. Usually, given z^N, the parameter is estimated as

$$\hat{\Theta}(z^N) = \text{argmin}_\Theta \sum_{i=1}^{N} (y_i - f(\boldsymbol{x}_i; \Theta))^2 / N. \tag{1}$$

Since the estimate $\hat{\Theta}$ depends on given z^N, we write $\hat{\Theta}(z^N)$ to clarify the dependency of z^N. Therefore, an output value of the estimator f for an input $\boldsymbol{x}$ should be written as $f(\boldsymbol{x}; \hat{\Theta}(z^N))$; however for simplicity we write the output as $f(\boldsymbol{x}; z^N)$. Note that $f(\boldsymbol{x}; z^N)$ is also a realization of random variable $f(\boldsymbol{x}; Z^N)$.

Now, we introduce a new random vector $Z_0 = (X_0, Y_0) \in \mathcal{R}^{d+1}$, which has a distribution identical to that of Z_i, but is independent of Z_i for all i. Then, the *generalization error* (GErr) of the estimator f can be defined as the following mean squared error (MSE) averaged over all possible realizations of Z^N and Z_0:

$$\text{GErr}(f) = \text{E}_{Z^N} \left\{ \text{E}_{Z_0} \{ [Y_0 - f(X_0; Z^N)]^2 \} \right\}. \tag{2}$$

Here $\text{E}_{Z_0}\{\}$ and $\text{E}_{Z^N}\{\}$ represent expectation [1] with respect to the distribution of Z_0 and Z^N, respectively. From the above definition, it is clear that $\text{GErr}(f)$ depends on neither particular training set z^N nor an unknown sample z_0, and only depends on the sample size N and the model of the estimator f. With simple algebra, (2) can also be expressed by the following familiar "bias/variance" decomposition [7].

$$\text{GErr}(f) = \text{E}_{X_0} \left\{ \text{Var}\{f|X_0\} + \text{Bias}\{f|X_0\}^2 \right\} + \sigma^2, \tag{3}$$

where $\text{Var}\{f|X_0 = \boldsymbol{x}_0\}$ and $\text{Bias}\{f|X_0 = \boldsymbol{x}_0\}$ are conditional variance and conditional bias given $X_0 = \boldsymbol{x}_0$:

$$\begin{aligned} \text{Var}\{f|X_0\} &= \text{E}_{Z^N} \left\{ \left(f(X_0; Z^N) - \text{E}_{Z^N}\{f(X_0; Z^N)\} \right)^2 \right\}, \\ \text{Bias}\{f|X_0\} &= \text{E}_{Z^N}\{f(X_0; Z^N)\} - g(X_0). \end{aligned} \tag{4}$$

Note that $\text{Var}\{f|X_0\}$ and $\text{Bias}\{f|X_0\}$ are random variables dependent on the random variable[2] X_0.

3. Generalization Error of Ensemble Estimator

Let $f_1, \ldots, f_M$ denote M estimators, where the mth estimator is separately trained on $z^N_{(m)}$, $m = 1, \ldots, M$. For simplicity, we assume that the sample size of each training set is uniformly N. Note that training set $z^N_{(m)}$ is a realization of a random sequence $Z^N_{(m)}$ and that $Z^N_{(m)}, m = 1, \ldots, M$, have the same distribution $p(\boldsymbol{x}, y)$; however, they cannot always be assumed to be mutually independent. The independent case will be discussed later as a special case.

The output of the ensemble estimator for some input $\boldsymbol{x}$ is defined as the simple average of outputs of M estimators for $\boldsymbol{x}$ after they have been separately trained. Specifically,

$$f^{(M)}_{ens}(\boldsymbol{x}) = \frac{1}{M} \sum_{m=1}^{M} f_m(\boldsymbol{x}; z^N_{(m)}). \tag{5}$$

Then, we have the following theorem about the generalization error of the ensemble estimator.

Theorem 1 (Generalization error in general case)

Let $\text{GErr}(f^{(M)}_{ens})$ *denote the generalization error of the ensemble estimator given by (5). We have*

$$\text{GErr}(f^{(M)}_{ens}) = \text{E}_{X_0} \left\{ \frac{1}{M} \overline{\text{Var}}(X_0) + \left(1 - \frac{1}{M}\right) \overline{\text{Cov}}(X_0) + \overline{\text{Bias}}(X_0)^2 \right\} + \sigma^2, \tag{6}$$

[1] In order to clarify the random variable, we will use notation such as $\text{E}_{Z^N}\{\}$ instead of simply $\text{E}\{\}$.

[2] Strictly speaking, they should be defined by using a conditional expectation like $\text{E}_{Z^N|X_0}\{\varphi(X_0, Z^N)|X_0\}$. However, since Z^N is independent of X_0, $\text{E}_{Z^N|X_0}\{\varphi(X_0, Z^N)|X_0\}$ becomes $\text{E}_{Z^N}\{\varphi(X_0, Z^N)\}$.

where $\overline{\mathrm{Var}}(X_0)$, $\overline{\mathrm{Cov}}(X_0)$, *and* $\overline{\mathrm{Bias}}(X_0)$ *are average conditional variance, conditional covariance, and conditional bias, averaged over* M *estimators, respectively. That is,*

$$\begin{cases} \overline{\mathrm{Var}}(X_0) = \frac{1}{M}\sum_{m=1}^{M} \mathrm{Var}\{f_m|X_0\}, \\ \overline{\mathrm{Cov}}(X_0) = \frac{1}{M(M-1)}\sum_{m}\sum_{m'\neq m} \mathrm{Cov}\{f_m, f_{m'}|X_0\}, \\ \overline{\mathrm{Bias}}(X_0) = \frac{1}{M}\sum_{m=1}^{M} \mathrm{Bias}\{f_m|X_0\}. \end{cases} \tag{7}$$

Here f_m *denotes* $f_m(X_0; Z^N_{(m)})$. *Using the following average generalization error averaged over* M *estimators,*

$$\overline{\mathrm{GErr}} = \frac{1}{M}\sum_{m=1}^{M}\left(\mathrm{E}_{X_0}\{\mathrm{Var}\{f_m|X_0\} + \mathrm{Bias}\{f_m|X_0\}^2\} + \sigma^2\right),$$

we can also present the relationship between $\mathrm{GErr}(f^{(M)}_{ens})$ *and* $\overline{\mathrm{GErr}}$ *as follows.*

$$\mathrm{GErr}(f^{(M)}_{ens}) = \frac{1}{M}\overline{\mathrm{GErr}} + \left(1-\frac{1}{M}\right)\sigma^2 + \mathrm{E}_{X_0}\left\{\left(1-\frac{1}{M}\right)\overline{\mathrm{Cov}}(X_0) + \frac{1}{M^2}\sum_{m}\sum_{m'\neq m}\mathrm{Bias}\{f_m|X_0\}\mathrm{Bias}\{f_{m'}|X_0\}\right\}. \tag{8}$$

Proof Using $f^{(M)}_{ens}$ instead of f in (3), we have

$$\mathrm{GErr}(f^{(M)}_{ens}) = \mathrm{E}_{X_0}\left\{\mathrm{Var}\{f^{(M)}_{ens}|X_0\} + \mathrm{Bias}\{f^{(M)}_{ens}|X_0\}^2\right\} + \sigma^2, \tag{9}$$

Since $f^{(M)}_{ens}$ depends on $z^N_{(1)}, \ldots, z^N_{(M)}$, $\mathrm{Var}\{f^{(M)}_{ens}|X_0\}$ and $\mathrm{Bias}\{f^{(M)}_{ens}|X_0\}$ can be given by extending (4) as

$$\begin{aligned} \mathrm{Var}\{f^{(M)}_{ens}|X_0\} &= \mathrm{E}_{Z^N_{(1)},\ldots,Z^N_{(M)}}\left\{\left[\frac{1}{M}\sum_{m=1}^{M} f_m - \mathrm{E}_{Z^N_{(1)},\ldots,Z^N_{(M)}}\{\frac{1}{M}\sum_{m=1}^{M} f_m\}\right]^2\right\} \\ &= \frac{1}{M^2}\sum_{m=1}^{M}\mathrm{E}_{Z^N_{(m)}}\left\{[f_m - \mathrm{E}_{Z^N_{(m)}}\{f_m\}]^2\right\} + \\ &\qquad \frac{1}{M^2}\sum_{m}\sum_{m'\neq m}\mathrm{E}_{Z^N_{(m)},Z^N_{(m')}}\left\{[f_m - \mathrm{E}_{Z^N_{(m)}}\{f_m\}][f_{m'} - \mathrm{E}_{Z^N_{(m')}}\{f_{m'}\}]\right\} \\ &= \frac{1}{M}\overline{\mathrm{Var}}(X_0) + \left(1-\frac{1}{M}\right)\overline{\mathrm{Cov}}(X_0) \end{aligned} \tag{10}$$

$$\begin{aligned} \mathrm{Bias}\{f^{(M)}_{ens}|X_0\} &= \mathrm{E}_{Z^N_{(1)},\ldots,Z^N_{(M)}}\left\{\frac{1}{M}\sum_{m=1}^{M} f_m\right\} - g \\ &= \frac{1}{M}\sum_{m=1}^{M}\mathrm{E}_{Z^N_{(m)}}\{f_m - g\} = \overline{\mathrm{Bias}}(X_0). \end{aligned} \tag{11}$$

Substitution of (10) and (11) into (9) gives (6). Once (6) is proven, the derivation of (8) is trivial. ∎

One can see that when $M = 1$, (6) reduces to (3). Theorem 1 shows that the generalization error of the ensemble estimator, unlike that of a single estimator, is affected not only by variance, bias and noise variance, but also by covariance. This may be intuitively reasonable. Moreover, these explicit expressions yield much information about how the generalization error of the ensemble estimator is affected by each of factors of interest (Var, Cov, Bias, and σ^2), and they also quantify the improvement due to the ensemble.

Corollary 1 (Equal model case)

If models of M *estimators are the same,* [3] *then we have*

$$\mathrm{GErr}(f^{(M)}_{ens}) = \mathrm{E}_{X_0}\left\{\frac{1}{M}\mathrm{Var}\{f|X_0\} + \left(1-\frac{1}{M}\right)\overline{\mathrm{Cov}}(X_0) + \mathrm{Bias}\{f|X_0\}^2\right\} + \sigma^2, \tag{12}$$

[3] In feedforward neural networks, the same model means the same network structure.

An equivalent expression is

$$\mathrm{GErr}(f_{ens}^{(M)}) = \frac{1}{M}\mathrm{GErr}(f) + \left(1 - \frac{1}{M}\right) \mathrm{E}_{X_0}\left\{\mathrm{Bias}\{f|X_0\}^2 + \overline{\mathrm{Cov}}(X_0) + \sigma^2\right\}. \tag{13}$$

where $\mathrm{GErr}(f)$ *is the generalization error of a single estimator given in (3).*

Proof Since $Z_{(m)}^N, m = 1, \ldots, M$ have the same distribution, if models of M estimators are the same, then clearly variances (biases) of all estimators become identical. That is,

$$\mathrm{Var}\{f_1|X_0\} = \cdots = \mathrm{Var}\{f_M|X_0\} \equiv \mathrm{Var}\{f|X_0\}, \qquad \mathrm{Bias}\{f_1|X_0\} = \cdots = \mathrm{Bias}\{f_M|X_0\} \equiv \mathrm{Bias}\{f|X_0\}.$$

Then, (6) and (8) reduce to (12) and (13), respectively. ∎

Corollary 2 (Equal model & uncorrelated case)
If models of M estimators are the same and outputs of M estimators are mutually uncorrelated,[4] *then we have*

$$\mathrm{GErr}(f_{ens}^{(M)}) = \mathrm{E}_{X_0}\left\{\frac{1}{M}\mathrm{Var}\{f|X_0\} + \mathrm{Bias}\{f|X_0\}^2\right\} + \sigma^2. \tag{14}$$

An equivalent expression is

$$\mathrm{GErr}(f_{ens}^{(M)}) = \frac{1}{M}\mathrm{GErr}(f) + \left(1 - \frac{1}{M}\right) \mathrm{E}_{X_0}\left\{\mathrm{Bias}\{f|X_0\}^2 + \sigma^2\right\}. \tag{15}$$

Proof Since f_m and $f_{m'}$ are mutually independent, $\mathrm{Cov}\{f_m, f_{m'}|X_0\} = 0$. Hence, setting $\overline{\mathrm{Cov}} \equiv 0$ in (12) and (13) lead to (14) and (15), respectively. ∎

These results can easily be extended to the *general ensemble estimators* defined by

$$f_{gens}^{(M)}(\boldsymbol{x}) = \sum_{m=1}^{M} \alpha_m f_m(\boldsymbol{x}; z_{(m)}^N), \qquad \text{where} \quad \alpha_m > 0, \sum_{m=1}^{M} \alpha_m = 1. \tag{16}$$

Due to lack of space, only the result will be provided in the appendix.

Remarks At this point, the following important observations can be made :

- If f_m and $f_{m'}$ are *negatively* correlated, then the correlation contributes to a decrease in the generalization error. Conversely, if f_m and $f_{m'}$ are *positively* correlated, then the correlation increases the generalization error.
- Under the uncorrelated assumption, from (10), $\mathrm{Var}\{f_{ens}^{(M)}|X_0\}$ reduces to $\overline{\mathrm{Var}}(X_0)/M$ due to the averaging, and this reduction decreases the generalization error[5]. On the other hand, since $\mathrm{Bias}\{f_{ens}^{(M)}|X_0\}$ is a simple average of M biases $(\mathrm{Bias}\{f_m|X_0\}, m = 1, \ldots, M)$, the bias term basically does not contribute to a decrease in the generalization error. Therefore, combining estimators that each has a higher ratio of variance to bias is more useful for decreasing generalization error than combining estimators that each has a lower ratio of variance to bias.
- Even if we increase the number of ensembles M, the generalization error cannot be less than $\mathrm{E}_{X_0}\left\{\mathrm{Bias}\{f|X_0\}^2\right\} + \sigma^2$. That is, the amount gives the lower bound of $\mathrm{GErr}(f_{ens}^{(M)})$.

4. Simulation

In order to verify the theoretical results presented so far, we performed a simulation for the case of Corollary 2 that is easily simulated. We used standard feedforward neural networks with one input, ten hidden units and one output (1-10-1) as regression estimators. Figure 1 shows a target function

[4] Uncorrelated estimators can be obtained by training each estimator on a different (independent) training set.

[5] This variance reduction by a factor of M has been already mentioned by Perrone [2]. Perrone's result for basic ensemble agrees with (15) when Bias $\equiv 0$ and $\sigma^2 \equiv 0$.

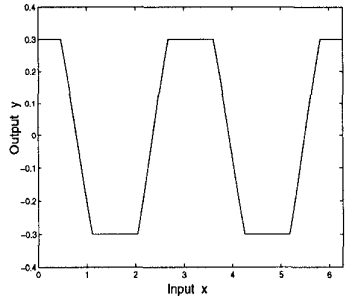

Fig. 1: Target function.

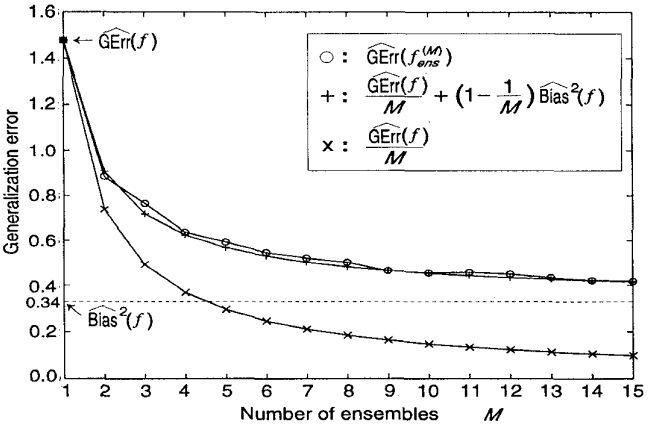

Fig. 2: Number of ensembles vs. generalization error.

$(y = g(x), x, y \in \mathcal{R})$. We set $N = 40$. That is, $z_{(m)}^{40} = \{(x_1, y_1), \ldots, (x_{40}, y_{40})\}$, $m = 1, \ldots, M$ are artificially and independently generated so that each (x_i, y_i) pair satisfies $y_i = g(x_i)$. This means $\sigma^2 \equiv 0$. A test set (z_0^{1000}) containing 1000 samples is also generated in the same way. For $M = 1, \ldots, 15$, the estimate of $\mathrm{GErr}(f_{ens}^{(M)})$, denoted as $\widehat{\mathrm{GErr}}(f_{ens}^{(M)})$, is calculated by the following approximation: $\widehat{\mathrm{GErr}}(f_{ens}^{(M)}) = \frac{1}{1000}\sum_{i=1}^{1000}\frac{1}{100}\sum_{j=1}^{100}(y_i^{(0)} - f_{ens}^{(M)}(x_i^{(0)}; z_{(j)}^{40}))^2$. That is, $\mathrm{E}_{Z^{40}}$ is approximated by 100 sample sets. Similarly, the estimate of $\mathrm{E}_{X_0}\{\mathrm{Bias}\{f|X_0\}^2\}$, denoted as $\widehat{\mathrm{Bias}}^2(f)$, is calculated as: $\widehat{\mathrm{Bias}}^2(f) = \frac{1}{1000}\sum_{i=1}^{1000}(\frac{1}{100}\sum_{j=1}^{100} f(x_i^{(0)}; z_{(j)}^{40}) - y_i^{(0)})^2$. Clearly, $\widehat{\mathrm{GErr}}(f)$ can be obtained as $\widehat{\mathrm{GErr}}(f_{ens}^{(1)})$.

We summarize the above simulation result in Fig. 2. In Fig. 2, 'o' indicates $\widehat{\mathrm{GErr}}(f_{ens}^{(M)})$ and '+' indicates the theoretical result obtained by substituting values of $\widehat{\mathrm{GErr}}(f)$ and $\widehat{\mathrm{Bias}}^2(f)$ and $\sigma^2 = 0$ into (15). For comparison, $\widehat{\mathrm{GErr}}(f)/M, M = 1, \ldots, 15$ are also plotted in Fig. 2 as '×'. One can see that the experimental curve ('o') fits into the theoretical one (far from $\widehat{\mathrm{GErr}}(f)/M$) and that both curves approach $\widehat{\mathrm{Bias}}^2(f) = 0.34$ as M increases.

5. Discussion

We should notice that these results do not always support the usefulness of the ensemble approach. In the case of Corollary 2, one can see that an increase in M necessitates an increase in the total number of training samples. In other words, if fixed N samples are available,[6] we must divide the data set into M disjoint sets. Each set contains N/M samples such that M sets are mutually uncorrelated. In this case, it is natural to ask, "Why don't you train a single estimator by using all N training samples at once? Does combining estimators, each of which is trained on divided N/M samples, still outperform the single estimator trained on N samples?" In general, both bias and variance depend on the sample size and the dimensionality of the parameter to be estimated. This means that bias (variance) obtained by N training samples is definitely less than that obtained by N/M training samples. Hence, recalling that the generalization error of a single estimator is expressed by the sum of the variance and the squared bias, it is quite possible to develop the situation where a single estimator outperforms an ensemble estimator. An interesting result is presented in Table 1, which shows that a signle estimator really outperforms an ensemble estimator.

In Table 1, $\widehat{\mathrm{GErr}}(f)$ and $\widehat{\mathrm{Bias}}^2(f)$ were obtained from a single estimator trained on N samples. For example, $N = 400$ corresponds to $M = 10$, because each of M estimators is trained on 40 samples in this

[6] In many practical situations, only one small sample set is available, because collecting samples often involves a large cost and is sometimes impossible.

Table: 1: Sample size effect on generalization error

N	40	80	120	160	200	280	400
$\widehat{\mathrm{GErr}}(f)$	1.48	0.40	0.35	0.16	0.14	0.09	0.08
$\widehat{\mathrm{Bias}}^2(f)$	0.34	0.17	0.11	0.10	0.08	0.06	0.05

simulation. As previously mentioned, bias monotonically decreases as N increases. Comparing $\widehat{\mathrm{GErr}}(f_{ens}^{(M)})$ in Fig. 2 with corresponding $\widehat{\mathrm{GErr}}(f)$ in Table 1, one can see that the generalization error of the single estimator trained on larger samples is much smaller than that of the ensemble estimator. For example, when $M = 10$, $\widehat{\mathrm{GErr}}(f_{ens}^{(10)}) = 0.46$, while the corresponding $\widehat{\mathrm{GErr}}(f)$ of a single estimator is 0.08 ($N = 400$). In this experiment, the single estimator was superior to the ensemble estimator for all M. These results suggest that when we train each estimator, we should utilize as many samples as possible. The straightforward way is to divide samples into overlapped ones. However, this will produce positive correlations among these sample sets, and increase the generalization error as mentioned before. This problem may be called "ensemble dilemma." Future work will focus on solving this problem.

Appendix

Let $\mathrm{GErr}(f_{gens}^{(M)})$ denote the generalization error of the general ensemble estimator given by (16). Then, we have

$$\mathrm{GErr}(f_{gens}^{(M)}) = \mathrm{E}_{X_0}\left\{\sum_{m=1}^{M}\sum_{l=1}^{M}\alpha_m^*\alpha_l^* R_{ml}\right\}, \qquad \text{where} \quad \alpha_i^* = \sum_j R_{ij}^{-1} / \sum_k \sum_j R_{kj}^{-1}. \tag{17}$$

α_i^* denotes the optimal weight that minimizes the generalization error of the general ensemble estimator given in (16). R_{ij}^{-1} indicates the i,j the component of the inverse matrix of $\mathbf{R}$. The i,jth component of matrix $\mathbf{R}$ is given by

$$R_{ij} = \begin{cases} \mathrm{Var}\{f_i|X_0\} + \mathrm{Bias}\{f_i|X_0\}^2, & \text{if} \quad i = j \\ \mathrm{Cov}\{f_i, f_j|X_0\} + \mathrm{Bias}\{f_i|X_0\}\mathrm{Bias}\{f_j|X_0\}, & \text{otherwise.} \end{cases} \tag{18}$$

Note that if M estimators have the same model, then α_i^* reduces to $1/M$.

References

[1] Wolpert D. H.: "Stacked generalization," *Neural Networks*, **5**, 2, pp. 241–259, 1992.

[2] Perrone M. P.: "Improving regression estimates: Averaging methods for variance reduction with extensions to general convex measure optimization," *PhD Thesis*, Brown University, 1993.

[3] Tresp V. and Taniguchi M.: "Combining estimators using non-constant weighting functions," In Tesauro G. *et al.* eds., *Advances in Neural Information Processing Systems 7*, MIT Press, Cambridge MA, 1995.

[4] Krogh A. and Vedelsby J.: "Neural network ensembles, cross validation, and active learning," In Tesauro G. *et al.* eds., *Advances in Neural Information Processing Systems 7*, MIT Press, Cambridge MA, 1995.

[5] Hansen L. K. and Salamon P.: "Neural network ensembles," *IEEE Trans. Pattern Anal. Machince Intell.*, **PAMI-12**, 10, pp. 993–1001, 1990.

[6] Hampshire J. B. and Waibel A.: "The meta-pi network: Building distributed knowledge representations for robust multisource pattern recogintion," *IEEE Trans. Pattern Anal. Machince Intell.*, **PAMI-14**, 7, pp. 751–769, 1992.

[7] Geman S., Bienenstock E. and Cooper L. N.: "Neural networks and the bias/variance dilemma," *Neural Computation*, **4**, 1, pp. 1–58, 1992.