Bias, variance and the combination of estimators; The case of linear least squares

Ronny Meir
Department of Electrical Engineering
Technion, Haifa 32000
Israel
rmeir@ee.technion.ac.il

Abstract

We consider the effect of combining several least squares estimators on the expected performance of a regression problem. Computing the exact bias and variance curves as a function of the sample size we are able to quantitatively compare the effect of the combination on the bias and variance separately, and thus on the expected error which is the sum of the two. First, we show that by splitting the data set into several independent parts and training each estimator on a different subset, the performance can in some cases be significantly improved. We find three basic regions of interest. For a small number of noisy samples the estimation quality is dramatically improved by combining several independent estimators. For intermediate sample sizes, however, the effect of combining estimators can in fact be deletarious, tending to increase the bias too much. For large sample sizes both the single and the combined estimator approach the same limit. Our results are derived analytically for the case of linear least-squares regression, and are valid for systems of large input dimensions. A definite conclusion of our work is that substantial improvement in the quality of least-squares estimation is possible by decreasing the variance at the cost of an increase in bias. This gain is especially pronounced for small and noisy data sets. We stress, however, that the approach of estimator combination is *not* a panacea for constructing improved estimators and must be applied with care.

1 Introduction

Many of the problems related to supervised learning (and statistical inference in general) can be boiled down to the question of balancing bias and variance. Roughly speaking, the bias measures the distance between the average estimator and a 'target' function while the variance quantifies the spread of the estimator with respect to the data distribution (see for example [3]). While reducing bias can usually be accomplished quite easily by simply increasing the complexity of the class of models studied, this usually comes at the expense of increasing the variance in such a way that the overall expected error (which is simply the sum of the bias and the variance) is often increased. Thus, many efforts have been devoted to the issue of decreasing variance, while attempting to keep the concomitant increase in bias as small as possible. One of the methods which has become popular recently in the neural network community is variance reduction by combining estimators. It should be noted however that the idea of combining estimators has been around in the statistics and econometrics literature at least since the late sixties [2]. Nevertheless, it seems that not much analytic work has been devoted to a detailed study of the effect of *finite* sample size and noise on the bias/variance balance, most of the exact results known to this author assuming knowledge of the underlying probability distributions and thus being effectively asymptotic in nature. It is the explicit goal of this paper to study in detail a simple problem of linear regression, where the full bias/variance curve can be computed exactly for *any* sample size and noise level. We believe that this simple and exactly solvable model can afford us insight into more complex non-linear problems, which are at the heart of much of the recent work in neural networks.

A further aspect of our work is related to the question of the partitioning of the data set between the various estimators. Thus, while most studies assume the each estimator is trained on the complete data set, it is possible to envisage a situation where the data set is broken up into several subsets, using each subset of data to form a different estimator. While such a scheme seems wasteful from the bias point of view, we will see that in fact it produces superior forecasts in some situations. This, perhaps suprising, result is due to a large decrease in variance resulting from the independence of the estimators, in the case where the data subsets are independent. Ideally, one would like to have an automatic and adaptive scheme for breaking up the data, perhaps in the spirit of the recent work of Jordan and Jacobs [5]. An analysis of this latter situation seems at this point rather difficult and will not be attempted here.

The remainder of the paper is organized as follows. In section 2 we introduce the basic ideas of the combination of estimators, giving explicit expressions for the bias and variance in this case. Section 3 then reviews some old and new results on least squares estimation for the case of a single linear estimator, while section 4 then extends this study to the case of a combination of linear least-squares estimators, specializing to the case of estimators trained on non-overlapping data subsets as well as estimators formed from the complete data sets. We conclude and summarize our findings in section 5, mentioning some open questions for future research.

2 On the combination of estimators

The basic objective of regression is the following: given a finite training set, D, composed of n input/output pairs, $D = \{(\mathbf{x}_\mu, y_\mu)\}_{\mu=1}^n$, drawn according to an *unkown* distribution $P(\mathbf{x}, y)$, find a function ('estimator'), $f(\mathbf{x}; D)$, which 'best' approximates y. Using the commonly used mean-squared error criterion this is equivalent to minimizing

$$\mathcal{E}(\mathbf{x}; D) = E\left[(y - f(\mathbf{x}; D))^2 \,|\mathbf{x}, D\right] = E\left[(y - E[y|\mathbf{x}])^2 \,|\mathbf{x}, D\right] + (f(\mathbf{x}; D) - E[y|\mathbf{x}])^2 , \quad (1)$$

where the expectation $E[\cdot]$ is taken with respect to the (unknown) probability distribution $P(y|\mathbf{x})$. Since the first term in the above equation does *not* depend on the data we focus in what follows on the second data-dependent term, which measures the distance between the regression function, $E[y|\mathbf{x}]$ and the estimator $f(\mathbf{x};D)$. Taking the expectation with respect to the data one obtains the well-known separation of the expected error into bias and variance [3],

$$\mathcal{E}(\mathbf{x}) = E_D\left[(f(\mathbf{x};D)-E[y|\mathbf{x}])^2\right] = \overbrace{(E_D f(\mathbf{x};D)-E(y|\mathbf{x}))^2}^{\text{“Bias”}} + \overbrace{E_D\left[(f(\mathbf{x};D)-E_D f(\mathbf{x};D))^2\right]}^{\text{“Variance”}}, \tag{2}$$

where we have defined $\mathcal{E}(\mathbf{x}) = E_D[\mathcal{E}(\mathbf{x};D)]$.

At this point it is useful to define the *empirical error* which is a finite sample version of the expected error, namely

$$\mathcal{E}_{\text{emp}}(f;D) = \frac{1}{n}\sum_{\mu=1}^{n}\left[y_\mu - f(\mathbf{x}_\mu;D)\right]^2 . \tag{3}$$

Many estimation methods in common use are based on minimization of the immediately available empirical error, which however is a noisy version of the true expected error.

We consider a data source of the form

$$y = g(\mathbf{x}) + \eta, \tag{4}$$

where the 'target' function $g(\mathbf{x})$ is an unknown (and potentially non-linear) function and η is a Gaussian random variable with zero mean and variance σ^2. Clearly then

$$E[y|\mathbf{x}] = g(\mathbf{x}). \tag{5}$$

In the usual scenario for parameter estimation one uses the complete data set, D, to form an estimator $f(\mathbf{x};D)$. In this paper we consider the case where the data set D is broken up into K subsets (not necessarily disjoint), such that $D = \cup_{k=1}^{K} D^{(k)}$, and a separate estimator is found for each subset. The full estimator is then given by the linear combination [4]

$$f(\mathbf{x};D) = \sum_{k=1}^{K} b_k f_k(\mathbf{x};D^{(k)}) . \tag{6}$$

We note in passing that unless $\sum_k b_k = 1$ the combined estimator may be biased (i.e. $E_D f(\mathbf{x};D) \neq g(\mathbf{x})$) even in the case where each estimator f_k is unbiased. For simplicity, we assume that the data set is broken up before training into subsets, each estimator is trained on its data set, and the linear combination is then taken. Although many ways of breaking up the data can be considered, we focus here on two extremes, for which exact results are possible: (i) $D^{(k)} \cap D^{(k')} = \emptyset$ - i.e. non-overlapping training sets, and (ii) $D^{(k)} = D$, i.e. fully-overlapping data sets. Intermediate situations are also possible but will not be pursued here.

The estimation of the parameters b_k can be easily obtained if the data distribution is known, by simply minimizing the mean-squared error, thus obtaining the optimal mixture coefficients [4]. In the more typical case where this distribution is unkown, one may resort to other schemes such as least-squares fitting for the parameter vector $\mathbf{b} = \{b_1, \ldots, b_K\}$, namely

$$\mathbf{b}^* = \arg\min_{\mathbf{b}} \left\{ \sum_\mu \left[y_\mu - \sum_k b_k f_k(\mathbf{x}_\mu; D^{(k)}) \right]^2 \right\}. \tag{7}$$

The bias and variance of the combined estimator can be simply expressed in this case, and are given by

$$\begin{aligned} B(\mathbf{x}; g) &= \left[\sum_{k=1}^K b_k \left(E_{D^{(k)}} f_k(\mathbf{x}; D^{(k)}) \right) - g(\mathbf{x}) \right]^2 & (8) \\ V(\mathbf{x}; g) &= E_D \left[\sum_{k=1}^K b_k f_k(\mathbf{x}; D^{(k)}) - \sum_{k=1}^K b_k E_D f_k(\mathbf{x}; D^{(k)}) \right]^2 \\ &= \sum_k b_k^2 \left[E_{D^{(k)}} (f_k^2(\mathbf{x}; D^{(k)})) - (E_{D^{(k)}} f_k(\mathbf{x}; D^{(k)}))^2 \right] \\ &+ \sum_{k \neq k'} b_k b_{k'} \left\{ E_D[f_k(\mathbf{x}; D^{(k)}) f_{k'}(\mathbf{x}; D^{(k')})] - (E_{D^{(k)}} f_k(\mathbf{x}; D^{(k)}))(E_{D^{(k')}} f_{k'}(\mathbf{x}; D^{(k')})) \right\} \end{aligned}$$

For future purposes we have explicitly noted the dependence of the bias and variance on the input $\mathbf{x}$ and target function $g(\cdot)$. It is immediately apparent that the variance term is composed of two contributions. The first term, simply computes a weighted average of the single estimator variances, while the second term measures the average covariance between the different estimators. While the first term in the variance can be seen to decay as $1/K$ in the case where all the weights b_k are of the same order of magnitude, the second term is finite *unless* the covariances between estimators are very small. It would thus seem beneficial to attempt to make the estimators as weakly correlated as possible in order to decrease the variance. Observe that in the extreme case where the data sets are independent of each other, the second term in the variance vanishes identically. Note that the bias term depends only on single estimator properties and can thus be computed from the theory of the single estimator. As mentioned above, however, the second term in the variance expression explicitly depends on correlations between the different estimators, and thus requires the computation of quantities beyond those of single estimators. We shall return to these computational issues in section 4.

Note that the bias and variance expressions above depend explicitly on the 'test' input $\mathbf{x}$. In order to remove this explicit dependence, in the remainder of the paper we take expectations of $B(\mathbf{x})$ and $V(\mathbf{x})$ with respect to the probability distribution $p(\mathbf{x})$ according to which $\mathbf{x}$ is assumed to be drawn. We denote expectation values with respect to $p(\mathbf{x})$ by the symbol $E_p[\cdot]$.

One final comment is in place before proceeding to a specific model. We focus in this work on the *parametric* case where the the function $f(\mathbf{x}; D)$ belongs to some parametric class of functions (specifically linear functions). Thus, in this case $f(\mathbf{x}; D)$ is replaced by $f(\mathbf{x}; \hat{\mathbf{w}})$ where $\hat{\mathbf{w}} = \hat{\mathbf{w}}(D)$ is the parameter vector estimated from the data.

3 The single linear estimator

Before considering the case of a combination of estimators, we first review the case of a single *linear* estimator, given by

$$f(\mathbf{x}; D) = \hat{\mathbf{w}}^T \cdot \mathbf{x}, \tag{9}$$

where $\hat{\mathbf{w}}$ is estimated from the data set D. Following [1] we further assume that the data arises through an equation similar to (4) with $g = g(\mathbf{w}_0^T \cdot \mathbf{x})$. Looking back at Eq. (8) it is clear that both the bias and the variance are explicit functions of the weight vector $\mathbf{w}_0$. Again, in order to remove the explicit dependence on this vector we assume that $\mathbf{w}_0$ is drawn according to some distribution, $p_0(\mathbf{w}_0)$, and denote expectations with respect to this distribution by $E_0[\cdot]$. Thus, we define the averaged bias and variance by

$$B = E_0 E_p[B(\mathbf{x}; \mathbf{w}_0)] \quad , \quad V = E_0 E_p[V(\mathbf{x}; \mathbf{w}_0)]. \tag{10}$$

In this work we consider *least-squares* estimation which corresponds to minimizing the empirical error,

$$\mathcal{E}_{\text{emp}}(\mathbf{w}, D) = \|X\mathbf{w} - Y\|^2 = \sum_{\mu=1}^{n} \left[\mathbf{w}^T \cdot \mathbf{x}_\mu - y_\mu\right]^2, \tag{11}$$

where X is the $n \times d$ data matrix whose rows are the vectors $\mathbf{x}_\mu^T$, Y is the $n \times 1$ output vector and $\mathbf{w}$ is a $d \times 1$ weight vector. The components of the 'target' vector Y are given by $y_\mu = g(\mathbf{w}_0^T \cdot \mathbf{x}_\mu) + \eta_\mu$ where η_μ are i.i.d normal random variables with zero mean and variance σ^2. Note that while we take the estimator itself to be linear we allow the target function $g(\cdot)$ to be non-linear. This is meant to model the common situation where the model we are trying to fit is inadequate, since the correct model (even it exists) is usually unkown.

Thus, the least squares estimator is given by

$$\hat{\mathbf{w}} \in \arg\min_{\mathbf{w}} \mathcal{E}_{\text{emp}}(\mathbf{w}, D). \tag{12}$$

Since in this case the error-function, Eq. (11), is quadratic it possesses either a unique global minimum or a degenerate manifold of minima, in the case where the Hessian matrix, $X^T X$, is singular.

The solution to the least squares problem is well known [6], and will be briefly summarized. When the number of examples, n, is smaller than the input dimension, d, the problem is underdetermined and there are many solutions with *zero* empirical error. The solutions can be written out explicitly in the form

$$\hat{\mathbf{w}} = X^T(XX^T)^{-1}Y + \left(I - X^T(XX^T)^{-1}X\right)V \qquad (n < d), \tag{13}$$

where V is an *arbitrary* d-dimensional vector. We note in passing that the solutions (13) are biased, namely $E_D[\hat{\mathbf{w}}^T \cdot \mathbf{x}] \neq g(\mathbf{w}_0^T \cdot \mathbf{x})$, *even* in the case of a linear function $g(\cdot)$. We give an explicit expression for the bias and variance of these estimators in Eq. (20) below. It should be noted that any vector $\mathbf{w}$ satisfying this equation (and thus any least-squares estimator) becomes singular as n approaches d from below, since the matrix XX^T becomes singular. One can show that the minimal norm solution is given in this case by the choice $V = 0$. We follow common practice in referring to this latter solution as the *Moore-Penrose* solution. It is common in the literature to neglect the study of the underdetermined regime since the solution is not unique in this case. We, however, will pay specific attention to this case, corresponding to the often prevalent situation where the amount of data is small, attempting to show that the combination of estimators approach can significantly improve the quality of predictors in this regime. Moreover, many important inverse problems in signal processing fall into this category [8].

In the overdetermined case, $n > d$ (assuming the matrix X to be of full rank), a zero error solution is possible only if the function $g(\cdot)$ is linear and there is no noise, namely $E[\eta^2] = 0$. In any other case, the problem becomes *unrealizable* and the minimum error is non-zero. In any event, in this regime the *unique* solution minimizing the empirical error is given by

$$\hat{\mathbf{w}} = (X^T X)^{-1} X^T Y \qquad\qquad (n > d). \tag{14}$$

It is eay to see that this estimator is unbiased for linear $g(\cdot)$.

In order to compute the bias and variance for this model we use Eq. (8) with $K = 1$ and $b_k = 1$. The expected bias is then given by

$$B = E_0 E_p \left[\mathbf{x}^T \cdot \overline{\mathbf{w}} - g(\mathbf{x}^T \cdot \mathbf{w}_0)\right]^2 . \tag{15}$$

Similarly we can compute the variance term, which is explicitly given by

$$V = E_0 E_p \left[\overline{(\mathbf{x}^T \cdot \hat{\mathbf{w}})^2} - \left(\mathbf{x}^T \cdot \overline{\mathbf{w}}\right)^2\right] . \tag{16}$$

For the ease of notation we have denoted expectations with respect to the data distribution by overbars, thus for example $E_D[\hat{\mathbf{w}}]$ is replaced by $\overline{\mathbf{w}}$. In order to actually compute the expectations with respect to $\mathbf{x}$ and the weight vector $\mathbf{w}_0$. We assume in what follows that the random vector $\mathbf{x}$ is distributed according to a multi-dimensional normal distributions of zero mean and covariance matrix $(1/d)I$. The vector $\mathbf{w}_0$ is similarly distributed with unit covariance matrix. The reason for the particular scaling chosen for the covariance matrices will become clear below. In the remainder of the paper we will be concerned with exact calculations in the so called *thermodynamic limit*

$$d \to \infty, \qquad n \to \infty, \qquad \alpha = n/d \;\text{ finite.} \tag{17}$$

This limit is particularly useful in that the central limit theorem allows one to make precise statements about the behavior of the system, for an *effectively finite* sample size, α. As we show below (see figure 2) the reults obtained in this limit produce very good results even for systems with as few as $d = 100$ input variables. We note in passing that in the thermodynamic limit, $d \to \infty$, we have $\sum_i x_i^2 \to 1$ with probability 1 and similarly for $(1/d)\sum_i w_{0i}^2$. Using these simple distributions we can directly compute the bias and variance. Instead of performing the direct computation we make the following observation (due to [1]), which becomes especially useful when studying non-linear models. Since the variables x_i are i.i.d random variables and $\mathbf{x}$ is uncorrelated to $\hat{\mathbf{w}}$, it is clear that each of the variables $u = \mathbf{x}^T \cdot \overline{\mathbf{w}}$, $v = \mathbf{x}^T \cdot \hat{\mathbf{w}}$ and $z = \mathbf{x}^T \cdot \mathbf{w}_0$ is normally distributed with zero mean and variance $\frac{1}{d}\|\overline{\mathbf{w}}\|^2$, $\frac{1}{d}\|\hat{\mathbf{w}}\|^2$ and $\frac{1}{d}\|\mathbf{w}_0\|^2$ respectively. The respective covariances are given by

$$E_p[uv] = \frac{1}{d}\overline{\mathbf{w}}^T \cdot \hat{\mathbf{w}} \quad , E_p[uz] = \frac{1}{d}\overline{\mathbf{w}}^T \cdot \mathbf{w}_0 \quad , E_p[vz] = \frac{1}{d}\hat{\mathbf{w}}^T \cdot \mathbf{w}_0. \tag{18}$$

Denoting

$$R = E_0[\overline{\mathbf{w}}^T \cdot \mathbf{w}_0] \quad , \quad r = E_0\|\overline{\mathbf{w}}\|^2 \quad , \quad Q = E_0\overline{\|\hat{\mathbf{w}}\|^2}, \tag{19}$$

it is easy to see that the bias and variance are given by

$$\begin{aligned} B &= r - 2\overline{ug}R + \overline{g^2} \\ V &= Q - r. \end{aligned} \tag{20}$$

In the above equations we have used $\overline{g^2} = \int Dug^2(u)$ and $\overline{ug} = \int Du\ ug(u)$ where the Gaussian measure Du is defined by $Du = (e^{-u^2/2}/\sqrt{2\pi})du$. Note that if the target function g is odd, one finds $E_0[\overline{\mathbf{w}}] = 0$ for the Moore-Penrose solution (as can easily be verified from Eq. (13) and (14)), although one still finds that $r = E_0\|\overline{\mathbf{w}}\|^2 \neq 0$. We note in passing that the same result is obtained for any i.i.d variables, x_i, with zero mean and variance $1/d$. We thus note that a complete calculation of the expected bias and variance requires the explicit computation of the variables R, r and Q defined above.

In principle, with the explicit expressions (13) and (14) at hand one may proceed to compute all the quantities relevant to the evaluation of the bias and variance. Unfortunately, it turns out that a direct computation of r, R and Q using these expressions is a rather difficult task in the theory of random matrices, keeping in mind the potential non-linearity of the function $g(\cdot)$. A way to solve the problem can be undertaken via a slightly indirect route, using tools from statistical physics. This problem has been recently studied and fully solved in [1]. We summarize here the main conclusions, with the slight modification of including noise in the data, as in Eq. (4). The interested reader should consult the above reference for more details on the calculation. It turns out that a complete characterization of the solution requires the definition of one further quantity. Keeping in mind that for $\alpha < 1$ there are many least squares solutions with zero empirical error, it is not surprising to find that an important quantity is the expected overlap between two different solutions, giving rise to a further variable

$$q = \frac{1}{d}E_0[\overline{\hat{\mathbf{w}}^T \cdot \hat{\mathbf{w}}'}] \qquad (q < Q), \tag{21}$$

where $\hat{\mathbf{w}}$ and $\hat{\mathbf{w}}'$ correspond to two different solutions (estimators) of the *same* norm, Q. Note that for $n > d$ a single solution exists and thus $Q = q$, while for $n < d$ any solution with $q < Q$ is possible.

The computation of the variables R, Q and q has been performed for the single estimator case by [1] using the replica method of statistical mechanics. More recently Opper and Kinzel [7] have been able to derive the same results using purely probabilistic arguments, which circumvent some of the mathematical problems of the replica approach. The results obtained by these authors are summarized by the following equations:

$$\begin{aligned} \alpha &< 1: \qquad R = \alpha\overline{ug}, \qquad q = \frac{\alpha}{1-\alpha}\left(\overline{g^2}+\sigma^2-\alpha\overline{ug}^2\right) \qquad (Q > q), \\ \alpha &> 1: \qquad R = \overline{ug}, \qquad q = \frac{1}{\alpha-1}\left(\overline{g^2}+\sigma^2-(2-\alpha)\overline{ug}^2\right) \qquad (Q = q). \end{aligned} \tag{22}$$

Since the above authors were interested only in the expected error, which is the sum of the bias and the variance, they did not actually need to compute the variable r which is seen to drop out of the calculation in this case as can be seen by referring to Eq. (20). However, it is not difficult to extend the results of Bös et al. to the calculation of r. While the calculation can be done within the replica approach, we present in the appendix a direct evaluation of r, following the method used recently by Opper and Kinzel [7]. One obtains the simple result

$$r = R^2. \tag{23}$$

Using these results, the bias and variance are given explicitly in this case by

$$\begin{aligned} \alpha < 1: \quad B &= \overline{g^2} - \alpha(2-\alpha)\overline{ug}^2 \quad , \qquad V = Q - \alpha^2\overline{ug}^2 \\ \alpha > 1: \quad B &= \overline{g^2} - \overline{ug}^2 \qquad\qquad , \qquad V = \frac{1}{\alpha-1}\left[\overline{g^2}+\sigma^2-\overline{ug}^2\right] \end{aligned} \tag{24}$$

We see from this solution that for $\alpha > 1$ the bias is constant, while the variance is monotonically decreasing with the sample size α. For $\alpha < 1$, there are of course multiple solutions corresponding to different normalizations Q. It is immediatley apparent, however, that the Moore-Penrose solution of minimal norm, Q, gives rise to the smallest variance of all least-squares estimators (the bias is unaffected by the normalization of the solution), specifically

$$V_{\mathrm{MP}} = \frac{\alpha}{1-\alpha}\left[\overline{g^2}+\sigma^2-\alpha(2-\alpha)\overline{ug}^2\right] \qquad (\alpha < 1). \tag{25}$$

The expected (or generalization) error is given simply by $\mathcal{E} = B + V$, and is thus smallest for the Moore-Penrose solution. Note that this latter result is independent of whether the function $g(\cdot)$ is linear or not. We note in passing that in the simple case where the target function $g(\cdot)$ is *linear* and the data is noise-free ($\sigma^2 = 0$) one obtains the particularly simple result $\mathcal{E} = 1 - \alpha$ for $\alpha < 1$ and $\mathcal{E} = 0$ above $\alpha = 1$. Note however that in any other case the expected error is a non-linear function of the normalized sample size α.

We plot in figure 1 (solid curve) the bias and variance versus α for the Moore-Penrose solution in the case where the target function $g(\cdot)$ is linear and the noise variance is $\sigma^2 = 0.4$. The dashed curve will be discussed in the next section, in the context of combining multiple estimators. The results for the expected error, $\mathcal{E}$, are plotted in figure 2 together with the results from computer simulations performed using $d = 100$. As observed previously, the Moore-Penrose solution becomes singular at $\alpha = 1$. While the bias is a non-increasing function of the sample size, it displays no singularity, the singularity in the expected error resulting solely from the variance term. Keeping in mind the remarks in the previous section we note that the expected error of the Moore-Penrose solution is lowest among all solutions minimizing the empirical error and thus the singularity at $\alpha = 1$ will be manifest in all solutions which attempt to minimize this error. As we show in the next section, one can achieve lower expected error by combining estimators, which has the effect of decreasing the variance at the cost of increasing the bias. We note however that no singularity exists in the expected error in the case of zero noise (recall that in this case $\mathcal{E} = 1 - \alpha$ for $\alpha < 1$). However, since most real data sets contain noise of some form or another, it is expected that the singular case is the more realistic one. The behavior of the bias, variance and expected error in the case of a non-linear function g is similar. For example, we plot in figure 3 results for the case where the function $g(\cdot)$ is the (highly non-linear) sign function, namely $g(u) = \mathrm{sgn}(u)$ and the noise level is zero. Note the in this case the singularity persists even in the absense of noise.

4 Combining linear estimators

Having summarized the situation in the case of a single estimator, we proceed to the case of K linear estimators. In this case we assume that the complete data set, D, is broken up into K subsets $D^{(k)}$ of size $n_k = \alpha_k d$ each. In particular we consider two extreme situations: (i) The data sets are independent of each other, namely $D^{(k)} \cap D^{(k')} = \emptyset$ for all k, and (ii) $D^{(k)} = D$ for all k. We refer to the first situation as the non-overlapping case, while to the second case where all estimators are trained on the same data set as the fully-overlapping case. Denoting by $\hat{\mathbf{w}}^{(k)}$ the k'th least-squares estimator, based on training set $D^{(k)}$, we define the following quantities analagous to Eq. (19),

$$
\begin{aligned}
R^{(k)} &= E_0\left[\overline{\hat{\mathbf{w}}^{(k)T}} \cdot \mathbf{w}_0\right] , \\
r^{(k,k')} &= E_0\left[\overline{\hat{\mathbf{w}}^{(k)T}} \cdot \overline{\hat{\mathbf{w}}^{(k')}}\right] \qquad r^{(k)} = r^{(k,k)} , \\
\rho^{(k,k')} &= E_0\left[\overline{\hat{\mathbf{w}}^{(k)T} \cdot \hat{\mathbf{w}}^{(k')}}\right] \qquad Q^{(k)} = \rho^{(k,k)} .
\end{aligned}
\tag{26}
$$

Making use of Eqs. (8) and the probability distribution of $\mathbf{w}_0$ one then straighforwardly finds for the mixture estimator

$$
\begin{aligned}
B_{\mathrm{K}} &= \sum_{k,k'} b_k b_{k'} r^{(k,k')} - 2\overline{ug} \sum_k b_k R^{(k)} + \overline{g^2}, \\
V_{\mathrm{K}} &= \sum_k b_k^2 \left[Q^{(k)} - r^{(k)}\right] + \sum_{k \neq k'} b_k b_{k'} \left[\rho^{(k,k')} - r^{(k,k')}\right] .
\end{aligned}
\tag{27}
$$

Now, the computation of the parameters $Q^{(k)}$ and $R^{(k)}$ is identical to that for the case of a single estimator and can be directly read from the results of [1] with the single modification of rescaling the sample size α to α_k, the sample size of the subset used by the k'th estimator. The only interaction between estimators arises from the terms $\rho^{(k,k')}$ and $r^{(k,k')}$ which couple two different estimators. Note however that $r^{(k,k')}$ is independent of the degree of overlap between the data sets $D^{(k)}$ since each estimator is averaged independently. Thus, using the result of the appendix one finds that in general $r^{(k,k')} = R^{(k)}R^{(k')}$. For the case where the data sets are independent of each other, namely $D^{(k)} \cap D^{(k')} = \emptyset$ it is obvious that $\rho^{(k,k')} = r^{(k,k')}$. One then finds

$$\rho^{(k,k')} = r^{(k,k')} = R^{(k)}R^{(k')} \qquad \text{(independent data sets).} \tag{28}$$

In the case where the data sets are identical, $D^{(k)} = D$ for all k, it is easy to convince oneself that $\rho^{(k,k')}$ is just the variable q defined in Eq. (21) and computed by [1]. At this point we proceed by discussing individually the two data generation scenarios described in section 1.

4.1 Non-overlapping data sets

As shown above, since the data sets are independent of each other one has in this case $\rho^{(k,k')} = r^{(k,k')}$, implying that the second term in the expression for the variance in Eq. (27) vanishes. Specializing now to the case where the sizes of the data sets are equal, namely $\alpha_k = n_k/d = \alpha/K$ for all k, we expect the variables $R^{(k)}$ and $Q^{(k)}$ to be independent of the specific index k. In the particular case of a *uniform* mixture ($b_k = 1/K$ for all k) we obtain

$$\begin{aligned} B_{\mathrm{K}}^{(u)} &= R_K^2 - 2\overline{ug}R_K + \overline{g^2} \\ V_{\mathrm{K}}^{(u)} &= (Q_K - R_K^2)/K, \end{aligned} \tag{29}$$

where the subscript K indicates that the values of the corresponding variables must be taken with respect to $\alpha_K = \alpha/K$. The specific values of the parameters R_K and Q_K can be found directly from the solution of [1] by replacing α by α/K in Eqs. (22). We thus have the following explicit results for the Moore-Penrose solution

$$\begin{aligned} \frac{\alpha}{K} &< 1: \quad B_{\mathrm{K}}^{(u)} = \overline{g^2} - \frac{\alpha}{K}(2-\frac{\alpha}{K})\overline{ug}^2 \quad , \quad V_{\mathrm{K}}^{(u)} = \frac{1}{K}\left(\frac{\frac{\alpha}{K}}{1-\frac{\alpha}{K}}\right)\left[\overline{g^2} + \sigma^2 - \frac{\alpha}{K}(2-\frac{\alpha}{K})\overline{ug}^2\right] \\ \frac{\alpha}{K} &> 1: \quad B_{\mathrm{K}}^{(u)} = \overline{g^2} - \overline{ug}^2 \quad , \quad V_{\mathrm{K}}^{(u)} = \frac{1}{K}\left(\frac{1}{\frac{\alpha}{K}-1}\right)\left[\overline{g^2} + \sigma^2 - \overline{ug}^2\right] \end{aligned} \tag{30}$$

Note that in this case the equation for the bias is simply given by the result for the single estimator case, evaluated at a rescaled (and lower) value of α. Since the bias is a non-increasing function α (as can be seen for example in figure 1, or from the explicit expression, Eq. (24)), it is always increased (or at best unchanged) by combining estimators from non-overlapping data subsets. It can also be seen from the above equation that as K increases the variance can be made arbitrarily small, albeit at the expense of increasing the bias. It can thus be expected that for any sample size α there is an optimal number of estimators minimizing the expected error, $\mathcal{E} = B+V$. In fact, for small α one may expand the equations for the bias and variance obtaining for the expected error

$$\mathcal{E}_{\mathrm{K}}^{(u)} = \overline{g^2} - \frac{\alpha}{K}\left[2\overline{ug}^2 - \frac{1}{K}(\overline{g^2}+\sigma^2)\right] + O(\alpha^2), \tag{31}$$

from which it is easy to see that the optimal value of K is given by $K^* = (\overline{g^2} + \sigma^2)/\overline{ug}^2$ (or more precisely the integer closest to this number). As could be expected we see that the optimal value of K scales with the noise level σ^2, since the effect of combining multiple independent estimators is to reduce the effective noise variance by a factor of K. Another interesting situation concerns the case $\alpha > K$. A simple comparison of Eqs. (24) and (30) shows that in this case the single estimator always yields a lower expected error. Thus, while we have shown that for small sample sizes the effect of splitting the data set into independent subsets is helpful, this is longer the case if the sample size is sufficiently large, in which case a single estimator based on the complete data set is superior. For $\alpha \to \infty$, however, one finds that all uniform mixtures converge (to leading order) at the same rate, namely

$$\mathcal{E}_K \approx \mathcal{E}_\infty + (\mathcal{E}_\infty + \sigma^2)/\alpha, \tag{32}$$

where $\mathcal{E}_\infty = \overline{g^2} - \overline{ug}^2$. For finite values of α, however, the value of K has a strong effect on the quality of the stimator, as can be seen in the figures below. In figure 1 we have plotted (dashed line) the results for the bias and variance for the combined estimator, while figure 2 presents the same results for the expected error in the case of a linear target function $g(\cdot)$ and a noise variance $\sigma^2 = 0.4$. In this case we see that the decrease in variance is sufficiently large to overcome the concomitant increase in bias. Very similar behavior can be observed in figure 3, corresponding to the unrealizable case where $g(\cdot)$ is the sign function and the noise level is zero. In all cases the combined estimator diverges at $\alpha = K$, for the very same reason that the single-estimator does so at $\alpha = 1$. However, in the former case the divergence is suppressed by a further factor of $1/K$, as is evident from Eq. (30). It should be noted however that in the case where the noise is absent and the target function is linear, no reduction in expected error is possible with respect to the Moore-Penrose solution. Thus, the combination of estimators is seen to be particularly useful for noisy and unrealizable problems.

4.2 Overlapping Data

We focus now on the case where all estimators are formed from the same data set D, namely $D^{(k)} = D$ for all k. Since there is a unique solution to the least-squares estimation problem for $\alpha > 1$, all least-squares estimators coincide in this regime. Thus, we focus here on the case $\alpha < 1$, where multiple least-squares estimators coexist. We further assume that only mixtures of estimators of the same norm Q are allowed. It is clear in this case that the cross-correlation parameter $\rho^{(k,k')}$, defined in Eq. (26), is given by the variable q of Eq. (21). Using this fact we obtain for the uniform mixture

$$\begin{aligned} B_{\mathrm{K}}^{(u)} &= R^2 - 2\overline{ug}R + \overline{g^2} \\ V_{\mathrm{K}}^{(u)} &= (Q - R^2) - (Q - q)(1 - \frac{1}{K}) \qquad (\alpha < 1). \end{aligned} \tag{33}$$

Clearly the expression for the bias in this case is identical to that obtained for the single estimator, since all estimators are based on the same data set and the bias term depends only on single estimator properties. The variance term, however, is modified due to the correlation between the estimators expressed through the variable $\rho^{(k,k')}$. Since the variance for the case of a single estimator is $Q - R^2$ and since $q \leq Q$ it is clear in this case that the variance is reduced while the bias remains unchanged. Thus we conclude that the mixture of estimators in this case indeed produces superior performance to that of the single estimator. However, it can be seen that in the case of the Moore-Penrose solution, corresponding to choosing the smallest possible norm Q, the expected error is minimal. To see this note that the variance can also be written as $V_{\mathrm{K}}^{(u)} = q - R^2 + \frac{1}{K}(Q - q)$ which is minimized for $Q = q$, *independently* of the number of estimators in the mixture. We thus conclude that for $\alpha < 1$ the Moore-Penrose pseudo-inverse solution yields the lowest expected error, and this *cannot* be improved on by combining least-squares estimators obtained from the full data set D. We should stress, however, that we do not claim that this estimator is optimal, only that it is optimal among all least-squares estimators. For example, it has recently been shown that better estimators can be obtained by early stopping methods, which do *not* attempt to minimize the empirical error ([1][10]).

Recall that we have shown in the previous section that (for small and noisy data sets) combining estimators formed using non-overlapping data subsets produced results superior to those of any single estimator trained on the complete data set. An interesting conclusion of these results is that splitting the data set into non-overlapping subsets is a better strategy than training each estimator with the full data. As mentioned previously, the basic reason for this is the independence of the estimators formed in this fashion, which helps to reduce the variance term more drastically than in the case where the estimators are dependent (having been exposed to overlapping data sets).

5 Conclusions

In this paper we have studied the effect of combining different estimators on the performance of linear regression. In particular we have focused on the case of linear least-squares estimation, computing exactly the full bias and variance curves for the case where the input dimension is very large (the so called thermodynamic limit). While we have focused specifically on the case of linear estimators, it should not be hard to extend these results to simple non-linear functions of the form $f(\mathbf{w}^T \cdot \mathbf{x})$ (see section 2). The case of a combination of more complex estimators (such as multi-layered neural networks) is much more demanding, as even the case of a single such network is rather difficult to analyze.

Several positive conclusions we can draw from our study are the following. First, the general claim that combining experts is always helpful is clearly fallacious. While we have shown that combining estimators is beneficial in some cases (such as small noisy data sets), this is not the case in general. Second, we have shown that in some situations (specifically unrealizable rules and small sample size) it is advantageous to split the data into several non-overlapping subsets. It turns out that in this case the decrease in variance resulting from the independence of the different estimators, is larger than the concomitant increase in bias. It would be interesting to try to generalize our results to the case where the data is split in a more efficient manner, perhaps along the lines of the interesting results obtained recently by [9] in the context of classification. Third, our results agree with the general notion that when attempting to learn an unrealizable function (whether due to noise or to a mismatch with the target function) the best option is to learn with errors (see for example [1][10]).

While we have addressed in this paper the simplest scenario (namely linear least squares estimation) for the combination of estimators, there are quite a few open avenues for extending this work. Since we have shown that for relatively small and noisy data sets it is advantageous to break up the data into several independent subsets, it would be interesting to ask whether this is a general result or perhaps specific to our problem. While we cannot of course claim generality, we believe that the above result will hold in one form or another for other (non-linear) least squares problems. A further interesting line of research taken by several recent authors (see for example [5]) is that of adaptive mixtures of experts, where the coefficients b_k of the mixture in Eq. (6) are explicit functions of the current input. It would be interesting to attempt an analysis of this situation along the lines of this paper. Ultimately one would like to have a general theory for combining empirical estimators. While the analysis seems straighforward in the asymptotic regime (large sample size relative to the input dimension) we have seen that the situation for (effectively) finite sample size is rather complex.

Acknowledgement The author is grateful to Manfred Opper for a very helpful conversation and to the Ollendorff center of the Electrical Engineering Department at the Technion for partial financial support.

Appendix

In this appendix we compute the variable r, defined in Eq. (19), showing that it is equal to R^2 as defined in the same equation. The other variables may be computed along similar lines as will be indicated below. Although these variables can be computed straightforwardly using the replica approach (see [1]) we present here a direct calculation along the lines of [7]. We focus for the sake of illustration on the Moore-Penrose solution, Eq. (13), in the region $n < d$, given by

$$\mathbf{w} = X^T(XX^T)^{-1}Y. \tag{34}$$

Using the definition of r from Eq. (13) we find in this case

$$r = E_0 \sum_i \left\{ E_D \left[X^T \left(XX^T \right)^{-1} Y \right] \right\}_i^2 . \tag{35}$$

Writing out the squared term in the curly brackets explicitly, and noting that the only dependence on $\mathbf{w}_0$ is through the vector Y, we can express r as follows:

$$r = E_{D^{(1)}} E_{D^{(2)}} \left\{ \sum_{\mu\nu} \sum_{\mu'\nu'} E_0[y_\mu^{(1)} y_\nu^{(2)}] \left(X_1 X_1^T \right)^{-1}_{\mu\nu} \left(X_2 X_2^T \right)^{-1}_{\mu'\nu'} \left(X_1 X_2^T \right)_{\nu\nu'} \right\} , \tag{36}$$

where we have used the trivial identity $(E_D[f(X)])^2 = E_{D^{(1)}} E_{D^{(2)}}[f(X_1) f(X_2)]$. Here the matrices X_a, $a = 1, 2$, are given by $(X_a)_{\mu i} = x_{\mu i}^{(a)}$. The expectation with respect to $\mathbf{w}_0$ is given by

$$E_0[y_\mu^{(1)} y_{\mu'}^{(2)}] = E_0 \left[g \left(\mathbf{x}_\mu^{(1)T} \cdot \mathbf{w}_0 \right) g \left(\mathbf{x}_{\mu'}^{(2)T} \cdot \mathbf{w}_0 \right) \right] . \tag{37}$$

Since the variables $u_\mu^{(a)} = \mathbf{x}_\mu^{(a)T} \cdot \mathbf{w}_0$ are Gaussian random variables with zero mean and covariance matrix $E_0[(u_\mu^{(a)})^2] = 1$ and $E_0[u_\mu^{(1)} u_{\mu'}^{(2)}] = \left(X_1 X_2^T \right)_{\mu\mu'}$ we can express the average in Eq. (37) as a double integral over the correlated Gaussian random variables $u_\mu^{(a)}$. This intgeral can be simplified by defining two uncorrelated standard Gaussian variables x and y through

$$\begin{aligned} u_\mu^{(1)} &= x \\ u_{\mu'}^{(2)} &= cx + \sqrt{1-c^2}\, y, \end{aligned} \tag{38}$$

where we have defined $c = \left(X_1 X_2^T \right)_{\mu\mu'}$. The required expectation value is then given by

$$E_0[y_\mu^{(1)} y_{\mu'}^{(2)}] = \int Dx \int Dy \; g(x)\, g\left(cx + \sqrt{1-c^2} y \right) , \tag{39}$$

In order to evaluate the integral we recall that $c \sim O(1/d)$ in the limit of large d and thus expand the argument of the integral in powers of c, obtaining to leading order

$$E_0[y_\mu^{(1)} y_{\mu'}^{(2)}] = c \left[\int Dx x g(x) \right] \left[\int Dy \; g'(y) \right] .$$

Finally, an easy integration by parts shows that the last integral is equal to $\int Dyyg(y)$ from which we obtain

$$E_0[y_\mu^{(1)} y_{\mu'}^{(2)}] = \overline{ug}^2 \left(X_1 X_2^T\right)_{\mu\mu'} \tag{40}$$

where we have used $\overline{ug} = \int Duug(u)$. Note that in the case of a sharp target function, $g(\cdot) = \mathrm{sgn}(\cdot)$, the calculation above must be modified as the derivative is zero almost everywhere. However, a direct calculation in this case yields the same result, with $\overline{ug} = \sqrt{2/\pi}$ in this latter case. Thus we obtain

$$r = \frac{\overline{ug}^2}{d} \; E_{D^{(1)}} E_{D^{(2)}} \left\{ \sum_{\mu\nu} \sum_{\mu'\nu'} \left(X_1 X_2^T\right)_{\mu\mu'} \left(X_1 X_1^T\right)^{-1}_{\mu\nu} \left(X_2 X_2^T\right)^{-1}_{\mu'\nu'} \left(X_1 X_2^T\right)_{\nu\nu'} \right\}. \tag{41}$$

Now, the matrices $\left(X_a X_a^T\right)$ $(a = 1,2)$ are diagonally dominant since in the thermodynamic limit, $d \to \infty$, the diagonal elements approach unity with probability 1 while the off-diagonal elements are $O(1/\sqrt{d})$. Thus to leading order in $1/d$ we find

$$\left(X_a X_a^T\right)^{-1}_{\mu\nu} \approx \delta_{\mu\nu} - \left(\mathbf{x}_\mu^{(a)} \cdot \mathbf{x}_\nu^{(a)T}\right)(1 - \delta_{\mu\nu}).$$

Plugging this expression into Eq. (41) and noting that the off-diagonal terms of $\left(X_a X_a^T\right)^{-1}$ contribute vanishingly small terms to r when $d \to \infty$ one finds to leading order in $1/d$,

$$\begin{aligned} r &= \frac{\overline{ug}^2}{d} E_{D^{(1)}} E_{D^{(2)}} \left\{ \sum_{\mu\nu} \sum_{\mu'\nu'} \left(X_1 X_2^T\right)_{\mu\mu'} \delta_{\mu\nu} \delta_{\mu'\nu'} \left(X_1 X_2^T\right)_{\nu\nu'} \right\} \\ &= \frac{\overline{ug}^2}{d} E_{D^{(1)}} E_{D^{(2)}} \left\{ \sum_{\mu\mu'} \left(X_1 X_2^T\right)^2_{\mu\mu'} \right\}. \end{aligned} \tag{42}$$

Using $E[x_{\mu i}^{(a)} x_{\mu' j}^{(b)}] = \frac{1}{d} \delta_{ab} \delta_{\mu\mu'} \delta_{ij}$ valid for $\{a, b\} \in \{1, 2\}$ we finally obtain

$$r = \overline{ug}^2 \alpha^2, \tag{43}$$

which on comparing with Eq. (22) can be seen to equal R^2. The computation in the case $n > d$, where Eq. (14) is valid may be pursued in a similar fashion. The variables Q and R defined in Eq. (19) have been computed in [7] along similar lines.

References

[1] S. Bös, W. Kinzel and M. Opper, The generalization ability of perceptrons with continuous outputs, *Phys. Rev. A* 47:1384-1391, 1993.

[2] J.M. Bates and C.W.J. Granger, The combination of forecasts, *Operations Research Quartelry*, 20:319-325, 1969.

[3] S. Geman, E. Bienenstock and R. Dorsat, Neural networks and the bias/variance dilemma, *Neural Computation* 4:1-58, 1992.

[4] C.W.J. Granger and R. Ramanathan, Improved methods of combining forecasts, *J. of Forecast.* 3:197-204, 1984.

[5] M.I. Jordan and R.A. Jacobs, Hierarchical mixtures of experts and the EM algorithm, *Neural Computation*, 6:181-214, 1994.

[6] M. Kendall and A. Stuart, *The Advanced Theory of Statistics*, vol. 2, Fourth edition, McMillan Publishing, N.Y. 1979.

[7] M. Opper and W. Kinzel, Statistical mechanics of generalization, in *Physics of Neural networks*, van Hemmen, J.S., E. Domany and K. Schulten eds., Springer-Verlag, Berlin 1994.

[8] L.L. Scharf, *Statistical Signal Processing: Detection, Estimation and Time Series Analysis*, Addison-Wesley, Massachusetts 1991.

[9] C. Van den Broeck and J.M.R. Parrondo, Generalization error in a self-similar committee machine, *Phys. Rev. Lett.*, 71(15): 2355-2359, 1993.

[10] C. Wang., S.S. Venkatesh and J.S. Judd, Optimal stopping and effective machine complexity in learning, in *Advances in Neural Information Processing Systems 6*, eds. J. Cowan, G. Tesauro and J. Alspector, Morgan Kaufmann, San Mateo CA, 1994.

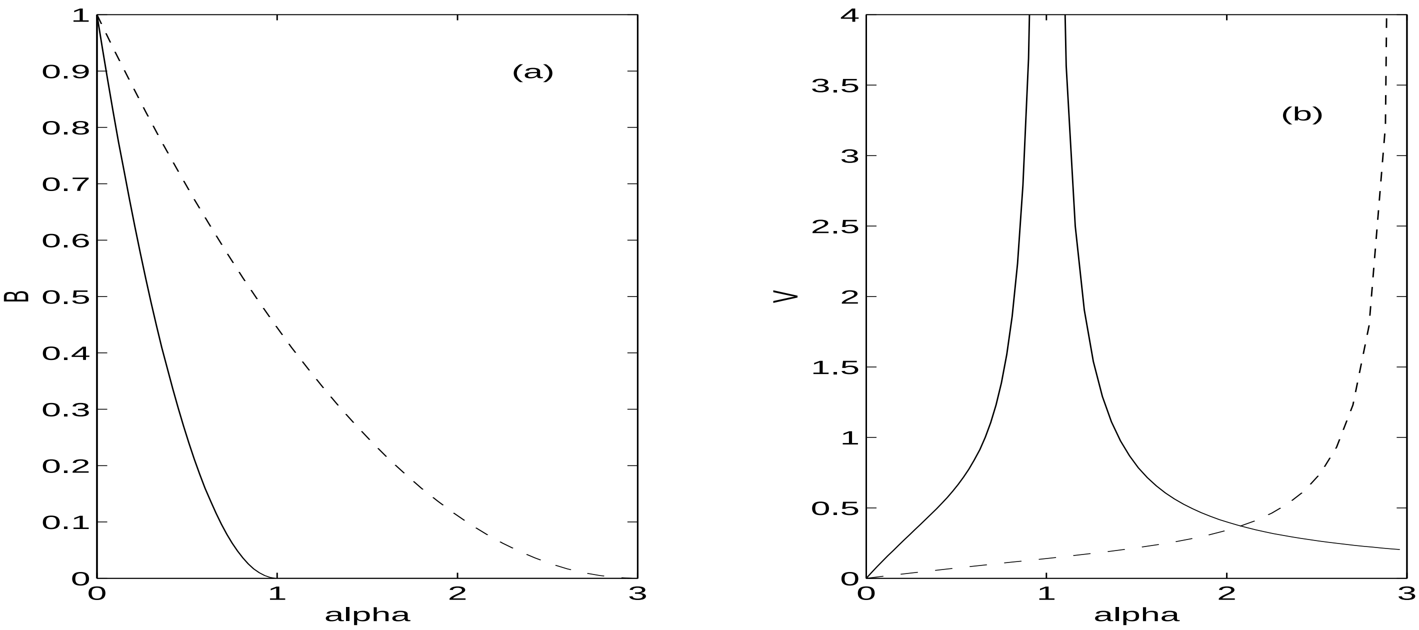

Figure 1: (a) The bias, B, plotted against the normalized sample size $\alpha = n/d$ for a linear target function $g(\cdot)$ and noise variance $\sigma^2 = 0.4$. (b) Same as (a) but with respect to the variance. The continuous curves represent the single estimator, while the dashed curve is for the combined estimator ($K = 3$).

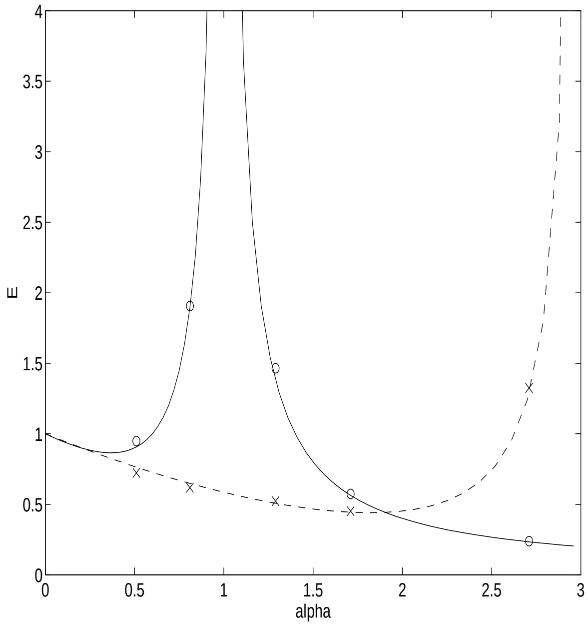

Figure 2: Expected error plotted against the normalized sample size for a linear target function $g(\cdot)$ and noise variance $\sigma^2 = 0.4$. The circles are the results of computer simulations with $d = 100$ for the single estimator case, while the crosses denote the combined estimator ($K = 3$).

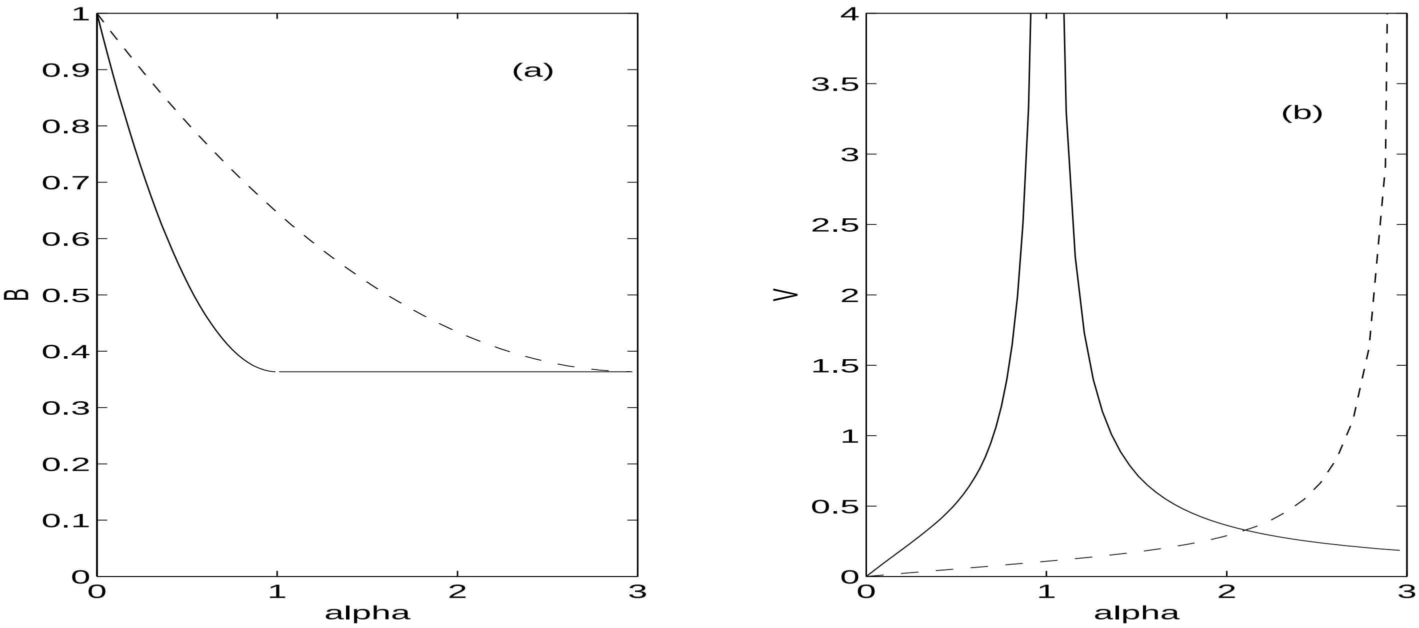

Figure 3: Same as figure 1, but for the case of a non-linear target function, $g(\cdot) = \mathrm{sgn}(\cdot)$, and zero noise.