Combining Least-Squares Regressions: An Upper-Bound on Mean-Squared Error

Gilbert Leung
Qualcomm, Incorporated
5775 Morehouse Drive
San Diego, CA 92121, USA
gleung@qualcomm.com

Andrew R. Barron
Statistics Dept., Yale University
P.O. Box 208290
New Haven, CT 06520, USA
andrew.barron@yale.edu

***Abstract*— For Gaussian regression, we develop and analyse methods for combining estimators from various models. For squared-error loss, an unbiased estimator of the risk of a mixture of general estimators is developed. Special attention is given to the case that the components are least-squares projections into arbitrary linear subspaces. We relate the unbiased risk estimate for the mixture estimator to estimates of the risks achieved by the components. This results in accurate bounds on the risk and its unbiased estimate — without advance knowledge of which model is best, the resulting performance is comparable to what is achieved by the best of the individual models.**

I. Introduction

Regression problems in statistics concern estimating some functional relation between a response variable and explanatory variables. Often there are multiple models describing such a relation. For example, each model can be a linear subspace of the full model space and the corresponding estimator for such a model to be the least-squares projection of the observations into that subspace. It is common to employ a two-stage practice which first picks a good model based on some data-dependent model assessment criterion, and then uses an appropriate regression estimator for that model. This is useful especially when a parsimonious model for explanation of the response is desired. However, model selection procedures can be unstable, as small changes in the data often lead to a significant change in model choice. Moreover, the inference done with the estimator for the chosen model does not account for model uncertainty from the selection procedure, and therefore can be overly optimistic.

An alternative to model selection-based estimation is to combine estimators from different models. As Bayes estimators are well-known to possess desirable statistical properties, a motivation for such mixtures comes from Bayesian philosophy. With squared-error loss, Bayes procedures never select a model, but rather, are convex combinations of estimators weighted by the corresponding models' posterior probability.

In this paper we study properties of the statistical risk (mean-squared error) of the combined estimator. An information-theoretic characterization of an unbiased estimate of its risk is provided. Connection with Bayes estimators is discussed. Furthermore, the risk of the resulting mixture is not much more than an idealized target defined by the minimum of risks achieved by the various estimators (one for each model considered). The general sharp risk bounds in this paper are obtained by choosing certian types of weights that adapt to the data. Moreover, the resulting mixture estimator often performs better in simulations [1] than a related model-selection estimator which picks the estimate corresponding to the highest-weighted model.

Problem Setting

One often reduces regression and function estimation problems into the following simplified canonical form. We start with observations

$$Y_i = f_i(X,\theta) + \varepsilon_i, \qquad i = 1,2,\ldots,n \tag{1}$$

of response values f plus independent Gaussian noise $\varepsilon \sim \mathcal{N}(0,\sigma^2 I)$, where X are dependent variables and θ are nuisance parameters. One estimates, under squared-error loss, the unknown mean $\mu = f(X,\theta)$ of the random vector Y. For simplicity σ^2 is known, and taken to be 1 henceforth.

The simple estimator Y can be obtained by maximizing likelihood or by least-squares (for μ in the entire parameter space $\mathbb{R}^n$), and has a mean-squared error of n. It is well-known in statistics that, for $n \geq 3$, there exist estimators with mean-squared errors below n without assuming any restriction of the parameter space. The type often studied is of the form

$$\hat{\mu}_i = c_i Y_i$$

where $0 \leq c_i \leq 1$ and may depend on Y. The main application of the present paper is an estimator of this form.

A linear regression model m is a d_m-dimensional linear subspace of $\mathbb{R}^n$ where the mean vector μ may reside. Let $\hat{\mu}^m = \mathcal{P}_m Y$ be the estimator that projects Y onto m. With $\|\cdot\|$ the Euclidean norm, the Pythagorean identity decomposes its risk into bias and variance:

$$r_m \stackrel{\text{def}}{=} \mathbb{E}\,\|\hat{\mu}^m - \mu\|^2 = \|\mathcal{P}_m\mu - \mu\|^2 + d_m, \tag{2}$$

where the expectation is taken with respect to the sampling distribution Y given μ. Thus, if μ is close to m and $d_m < n$, then $\hat{\mu}^m$ will have small risk, perhaps much smaller than n.

Now consider $\mathcal{M}$ a finite class of linear models m. Since we do not know which model is best, we form a convex

combination of these estimators

$$\hat{\mu} = \sum_{m \in \mathcal{M}} w_m \hat{\mu}^m, \tag{3}$$

where the data-determined weights w_m are chosen to give emphasis to models assessed to be better. This will be the setting we focus on in Section II-C, although the result in Section II-A applies to mixing any class of estimators $\{\hat{\mu}^m : m \in \mathcal{M}\}$, and some special forms of weights are suggested in II-B. The goal is to derive upper-bounds to the risk $r = \|\hat{\mu} - \mu\|^2$ of (3), given in Section IV using information-theoretic techniques in III.

II. Unbiased Estimate of Mean-Squared Error

A. General Mixtures

A key tool in our analysis is the unbiased estimate, or assessment, of risk. We use the notation $a \cdot b = \sum_{i=1}^n a_i b_i$ for the inner product and ∇ for the gradient $(\nabla_i)_{i=1}^n$ where $\nabla_i = \partial/\partial Y_i$. Assume that each $\hat{\mu}^m$, when expressed as a function of Y, is almost differentiable (that is, each of its coordinates can be represented by a directional integral, which is implied by continuity together with piecewise differentiability) with almost-everywhere derivatives $\nabla_i \hat{\mu}_i^m$.

Then for each m, in accordance with Akaike [2], Mallows [3], or Stein [4], we have

$$\hat{r}_m \stackrel{\text{def}}{=} \|\hat{\mu}^m - Y\|^2 + 2\sum_{i=1}^n \nabla_i \hat{\mu}_i^m - n. \tag{4}$$

as an unbiased estimate for the risk of $\hat{\mu}^m$, meaning $\mathbb{E}\,\hat{r}_m = \mathbb{E}\,\|\mu - \hat{\mu}^m(Y)\|^2$ for each $\mu \in \mathbb{R}^n$. We will give explicit formulae for the case of linear models m and least-squares estimation in section II-C. But here, we only assume that such an unbiased risk assessment $\hat{r}_m$ exists for each m, so both m and its corresponding estimator $\hat{\mu}^m$ are quite general. In effect, m serves here merely as an index to a collection $\mathcal{M}$ of arbitrary estimators $\hat{\mu}^m$.

Our first main result relates the unbiased assessment of the risk of $\hat{\mu}$ to unbiased assessments of the risks of the individual estimators $\hat{\mu}^m$.

Theorem 1: For each $m \in \mathcal{M}$, assume that $\hat{\mu}^m$ is almost differentiable in Y, and $\mathbb{E}\,|\nabla_i \hat{\mu}_i^m| < \infty$ for each i. Consider the mixture (3) with non-negative almost differentiable weights $w_m(Y)$ that sum to one, and satisfy $\mathbb{E}\,|(\nabla_i w_m)\hat{\mu}_i^m| < \infty$ for each i. Then an unbiased estimate of the risk $r = \mathbb{E}\,\|\hat{\mu} - \mu\|^2$ of (3) is given by

$$\hat{r} = \sum_{m \in \mathcal{M}} w_m \Big[\hat{r}_m - \|\hat{\mu}^m - \hat{\mu}\|^2 + 2(\nabla \log w_m) \cdot (\hat{\mu}^m - \hat{\mu})\Big]. \tag{5}$$

In addition, if

$$w_m(Y) = \frac{\exp(-\rho_m)\pi_m}{\sum_{m'} \exp(-\rho_{m'})\pi_{m'}} \tag{6}$$

for almost differentiable $\rho_m = \rho_m(Y)$ and arbitrary constants π_m, then

$$\hat{r} = \sum_{m \in \mathcal{M}} w_m \Big[\hat{r}_m - \|\hat{\mu}^m - \hat{\mu}\|^2 + 2(\nabla \rho_m) \cdot (\hat{\mu} - \hat{\mu}^m)\Big]. \tag{7}$$

Remark: One can adjust $\rho_m(Y)$ by adding any function of Y that does not depend on m without changing either the value of w_m or the validity of (7).

This unbiased estimate (5) of risk has three terms. The principal term $\sum_m w_m \hat{r}_m$ is the weighted average of the individual risk estimates. This average is a crude risk assessment, possibly biased. An information-theoretic representation of this term yields the conclusion that it is not much larger than $\min_m \hat{r}_m$.

The second term $-\sum_m w_m \|\hat{\mu} - \hat{\mu}^m\|^2$ wonderfully illustrates an advantage of mixing estimators. If the estimates $\hat{\mu}^m$ vary with m (that is, if the fits are different for different m), then averaging them (with weights w_m) leads to a reduction in the unbiased risk assessment given by the weighted average of the squared distances of the $\hat{\mu}^m$ from their centroid $\hat{\mu}$.

The third term $2\sum_m [w_m(\nabla \log w_m) \cdot (\hat{\mu}^m - \hat{\mu})]$ quantifies the effect of the data-sensitivity of the weights (through their gradients with respect to the data Y). Constant weights would make this term zero, but would not permit means to adapt the fit to the models that have smaller $\hat{r}_m$. Finally, the exponential form of weights (6) gives a particularly clean mixture risk estimate (7) that depends on the weights via the gradient of the exponents in the relative weighting only and not the normalization.

If our weights focus on models assessed to be good, then our intuition says that the third term quantifies the price one pays for making the mixture estimator adaptive, so it should have a positive expectation (otherwise, mixing offers a "free lunch").

Proof of Theorem 1: See p.16 of [1] for a proof of the first claim (5). For the second, $\nabla \log w_m(Y)$ equals $-\nabla \rho_m(Y)$ minus a function (the gradient of $\log \sum_k \exp(-\rho_k)\pi_k$) which does not depend on m. Now since $\hat{\mu} - \hat{\mu}^m$ has w-average being the null vector $\underline{0}$, its inner product with a quantity not depending on m averages to 0 under the weights w, so that we are left with the $\nabla \rho_m(Y)$ term. This proves (7). □

Given a collection of models and its corresponding estimators, one way to use Theorem 1 is to design data-determined weights w_m to make the unbiased estimate of risk (5) for the mixture small. The weights (6) offer a tractable start, and we can further simplify (7) in certain cases laid out in Section II-B. Our risk bounds developed later belong to this category.

A second application is evaluation of model classes and their respective mixture estimators, as there can be multiple model classes that meaningfully decompose a common parameter space into scientifically reasonable models. For example, two classes may consist of different linear models, and a third class may contain curved models. Provided that we have the component estimators in each model class and know how to weight them (again Section II-B offers some suggestions), we evaluate how effective each model class explains the data by using the unbiased estimate of risk for the corresponding mixture estimator of the class, i.e. as a test of goodness of fit. One can go further to use this model class assessment to aid model (and model class) design.

B. Special Forms of Weights (6)

Bayes mixtures use weights (6), where w_m is the posterior probability $p(m\,|\,Y)$ of model m. The exponent $-\rho_m$ is the log marginal density $\log p(Y\,|\,m)$ of Y for model m obtained by integrating the Gaussian likelihood $p(Y\,|\,\mu)$ with respect to a specified prior $p(\mu\,|\,m)$ for μ restricted to m. The prior may be improper (an infinite measure) so long as it yields $p(Y\,|\,m) < \infty$ for each Y and m, such that w_m is defined. The constants π_m are of course the prior probability for model m. Here, the component estimator $\hat{\mu}^m$ for the model m is posterior Bayes, i.e. the expectation of μ under the posterior density $p(\mu\,|\,Y,m)$ (with support in m). It is easy to show that (e.g. see [5])

$$\hat{\mu}^m = Y - \nabla\rho_m(Y). \tag{8}$$

This implies the assumption of the following corollary with $\beta = 1$.

Corollary 2: If $\rho_m(Y)$ in (6) has gradient $\beta(Y - \hat{\mu}^m)$ for each $m \in \mathcal{M}$ and some $\beta \geq 0$, then

$$\hat{r} = \sum_{m\in\mathcal{M}} w_m \Big[\hat{r}_m - (1-2\beta)\|\hat{\mu}^m - \hat{\mu}\|^2\Big]. \tag{9}$$

In addition, if $\beta \leq {}^1\!/_2$, the risk estimate can be bounded by

$$\hat{r} \leq \sum_{m\in\mathcal{M}} w_m \hat{r}_m,$$

with equality when $\beta = {}^1\!/_2$.

Proof: From the stated assumption of the form of $\rho_m(Y)$, we see that after adding a function not depending on m, $\nabla\rho_m(Y)$ matches a multiple of $\hat{\mu} - \hat{\mu}^m$ so the first claim follows from (7). The next claim is a special case of (9). □

We can apply Theorem 1 to more general weights. Consider w_m emphasizing models with small risk estimates $\hat{r}_m$.

$$w_m = \frac{\pi_m \exp(-\beta\hat{r}_m/2)}{\sum_{m'} \pi_{m'} \exp(-\beta\hat{r}_{m'}/2)}, \quad \beta > 0, \tag{10}$$

where the positive constants π_m are a mechanism for assigning model preference. That is, we take $\rho_m = \beta\hat{r}_m/2$ in (6). The parameter β controls the relative importance of averaging across models (small β) and picking out the one that is empirically best (large β). The two extremes are $\beta \to 0$, which ignores the observations Y and weights the models by π only, and $\beta \to \infty$, which uses only the model(s) with minimal estimated risk.

Intuitive appeal aside, the main motivation for these weights is that, in the case of linear models m and least-squares estimation (to be explored next), they also yield further simplification of (7) via Corollary 2.

In particular, the connection between these two cases (Bayes and linear least-squares) arises when the choice of prior $p(\mu\,|\,m)$ is uniform (and improper) under each linear model m, such that least-squares $\hat{\mu}^m$ for m is also posterior Bayes, and the posterior weights take the form of (10) with $\beta = 1$.

C. Linear Least-Squares

Now we specialize to the case that each model $m \in \mathcal{M}$ is a linear subspace of $\mathbb{R}^n$. The estimator $\hat{\mu}^m$ under such a model is the least-squares projection of the observations Y into the d_m-dimensional linear space, the column space of a design matrix X_m of a subset of explanatory variables. This can be accomplished by Gram-Schmidt procedures, or explicitly via the projection matrix $\mathcal{P}_m = X_m(X_m'X_m)^{-1}X_m'$ such that $\hat{\mu}^m = \mathcal{P}_m Y$.

Lemma 3: For each linear model m, let $\hat{\mu}^m = \mathcal{P}_m Y$. Then

$$\hat{r}_m = \|Y - \hat{\mu}^m\|^2 + 2d_m - n, \tag{11}$$

is its unbiased risk estimate and has gradient $2(Y - \hat{\mu}^m)$.

Proof: Use Stein's identity (4), together with the fact that $\operatorname{tr}\mathcal{P}_m = d_m$ to show that $\hat{r}_m$ is unbiased. Then write

$$\|Y - \hat{\mu}^m\|^2 = Y'(I - \mathcal{P}_m)'(I - \mathcal{P}_m)Y = Y'(I - \mathcal{P}_m)Y,$$

where the last equality follows from the fact that $I - \mathcal{P}_m$ is symmetric and also a projection (onto the space orthogonal to m). Thus the gradient of (11) is $2(I - \mathcal{P}_m)Y = 2(Y - \hat{\mu}^m)$. □

Thus, for linear least-squares estimators, using weights (10) satisfies the condition for Corollary 2. The tuning parameter β adjusts the degree of concentration of the weights on the models with small risk estimates. Typical values are $\beta = 1$, which gives the weighted mixture a Bayes interpretation; and $\beta = {}^1\!/_2$, which leads to the main risk bounds below. When the unknown mean μ can be well-approximated by multiple models m, the resulting risk of the mixture at μ would not be very sensitive to the choice of β in $[{}^1\!/_2, 1]$.

III. Information-theoretic Characterization of Average Risk Assessment

We analyse the average risk estimate $\sum_m w_m\hat{r}_m$ in this section using weights (10). It is the primary term in the estimate for the risk of the mixture $\hat{\mu}$; and for $\beta \leq {}^1\!/_2$, Corollary 2 says that it upper-bounds the unbiased risk estimate.

A. Upper-bound for Average (and Unbiased) Risk Estimate

Since the choice $\beta = {}^1\!/_2$ makes this average risk estimate unbiased for the risk of $\hat{\mu}$, we will set it so temporarily for a brisk exposition. The generalization to any $\beta > 0$ (done soon) can be obtained by replacing all occurrences of 4 below with $2/\beta$, though the average risk estimate will no longer be unbiased when $\beta \neq {}^1\!/_2$.

From now on, we write $\pi_m = \exp(-C_m)$, where C_m can be interpreted as the complexity for model m, giving preference to low-complexity models. For example, C_m may be chosen to increase with the dimension of m, and the resulting mixture estimator would weight the "small" models higher. We further impose that $\sum_m \pi_m \leq 1$.

Remark: This condition $\sum_m \exp(-C_m) \leq 1$ is of course Kraft's inequality [6] in base e and the model complexity is connected to the length of some codeword (in *nats*) that describes the model. However, our theory does not require

such coding interpretation. One can simply use the model preference π as a starting point to define $C_m = -\log \pi_m$.

Theorem 4: For each $m \in \mathcal{M}$, define weights w by putting $\beta = 1/2$ in (10). Then, with $\hat{m}$ being any model attaining $\min_m \{\hat{r}_m + 4C_m\}$ the unbiased risk estimate for $\hat{\mu} = \sum_m w_m \hat{\mu}^m$ satisifies

$$\hat{r} = \sum_{m \in \mathcal{M}} w_m \hat{r}_m = \hat{r}_{\hat{m}} + 4\Big[C_m - D(w\|\pi) + \log w_{\hat{m}}\Big] < \min_{m \in \mathcal{M}} \{\hat{r}_m + 4C_m\}, \tag{12}$$

where $D(w\|\pi) = \sum_m w_m \log(w_m/\pi_m)$ is the Kullback-Leibler divergence of π from w.

Proof: Observe that

$$\hat{r}_m = 4\Big[\log \frac{\pi_m}{w_m} - \log \textstyle\sum_{m'} \pi_{m'} \exp(-\hat{r}_{m'}/4)\Big] = \hat{r}_{\hat{m}} + 4\Big[C_{\hat{m}} - \log \frac{w_m}{\pi_m} + \log w_{\hat{m}}\Big]. \tag{13}$$

Thus, the equality follows by averaging over $m \in \mathcal{M}$ with weights w. The inequality results since $D \geq 0$ and $w_{\hat{m}} < 1$ (the logarithm of the latter is strictly negative). □

From now on, let $M = \#\mathcal{M}$ be the cardinality of $\mathcal{M}$. The following is immediate.

Corollary 5: Put $\pi_m = 1/M$. Here, with any model $\hat{m}$ achieving $\min_m \hat{r}_m \stackrel{\text{def}}{=} \hat{r}_*$, the unbiased risk estimate for $\hat{\mu}$ satisifies

$$\hat{r} = \sum_{m \in \mathcal{M}} w_m \hat{r}_m = \hat{r}_* + 4\Big[H(w) + \log w_{\hat{m}}\Big] < \hat{r}_* + 4\log M, \tag{14}$$

where $H(w) = \sum_m w_m \log(1/w_m)$ is the entropy of the distribution w. □

Therefore, for the special case $\pi_m = 1/M$, the average risk estimate (14) is the minimum of the individual risk estimates $\hat{r}_*$ plus a price for mixing, which is a function of the mixing weights w. If the weights w are concentrated on mostly one model $\hat{m}$, then both the entropy $H(w)$ and $\log w_{\hat{m}}$ are close to zero and the combined risk estimate is close to the minimum $\hat{r}_*$. This bound is useful when our model class contains an estimator with risk much larger than $4\log M$, which is typical for linear least-squares when n is large and the full model $\mathbb{R}^n$ is in the class. We now tighten this bound.

B. Refinements

Now we bring to the fore the role of an arbitrary $\beta > 0$ in mixing estimators, and improve upon Corollary 5.

Definition 6: Let $\psi = \psi(M)$ be a function in $M \geq 2$ defined by the solution to

$$\psi = \log \frac{M-1}{\psi} - 1. \qquad \square$$

Note that $\psi(M)$ is increasing in M. Also, for each $K > 0$,

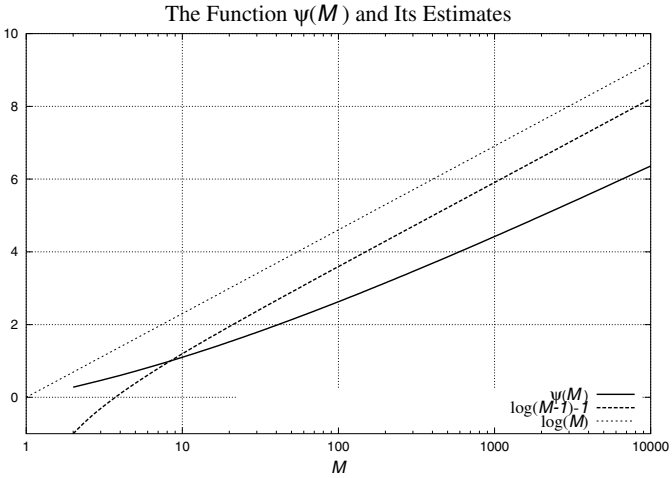
$$\psi \leq \max\Big\{K, \log \frac{M-1}{K} - 1\Big\}, \tag{15}$$

by considering separately whether $\psi \leq K$ or not. The proof of the following lemma is straightforward.

Lemma 7: $\psi(M) < \log M$ for $M \geq 2$. □

An upper-bound for ψ is provided by (15) using $K = 1$.

$$\psi_1(M) \stackrel{\text{def}}{=} \max\{1, \log(M-1) - 1\}. \tag{16}$$



The following theorem is interesting in its own right, as it gives a tight bound for the average $\sum_m w_m \hat{r}_m$ for arbitrary values $\hat{r}_m$ (not necessarily any risk estimates, and possibly negative). Nonetheless, it is mostly a technical refinement, since for large M, the reduction of $\psi(M)$ from $\log M$ is only of "secondary" order, $\log(\log M)$.

Theorem 8: For any real values $\{\hat{r}_m : m \in \mathcal{M}\}$ with $M = \#\mathcal{M} < \infty$, and weights (10) with $\pi_m = 1/M$, we have

$$\sum_{m \in \mathcal{M}} w_m \hat{r}_m \leq \min_{m \in \mathcal{M}} \hat{r}_m + \frac{2}{\beta}\psi(M). \tag{17}$$

Proof: See Lemma 2.20b in [1]. □

Remark: One proof using the idea of separating out the case $m = \hat{m}$ from the rest is reminiscent of Fano's inequality [6].

Characterizing the average risk estimate by the minimum is useful as it leads directly to a risk bound.

IV. Risk Bounds For Mixing Least-Squares

Corollary 9: If $\hat{\mu}^m$ are least-squares regressions with risk estimates $\hat{r}_m$ in (11), then with $\hat{r}_* = \min_m \hat{r}_m$, the unbiased risk estimate for the mixture of least-squares regressions $\hat{\mu} = \sum_m w_m \hat{\mu}^m$ using weights (10) with $\pi_m = 1/M$ satisfies

$$\hat{r} \leq \hat{r}_* + \frac{2}{\beta}\psi(\#\mathcal{M}) < \hat{r}_* + \frac{2}{\beta}\log(\#\mathcal{M})$$

for each $\beta \leq 1/2$. Hence, with r_m as the risks (2) of the individual estimators, the risk $r = \mathbb{E}\,\|\hat{\mu} - \mu\|^2$ satisfies

$$r \leq \min_{m \in \mathcal{M}} r_m + \frac{2\psi(\#\mathcal{M})}{\beta} < \min_{m \in \mathcal{M}} r_m + \frac{2}{\beta}\log(\#\mathcal{M}).$$

Proof: Corollary 2 implies that the unbiased risk estimate for $\hat{\mu}$ is upper-bounded by the average risk estimate for this range of β, which in turn is bounded as in (17). Recalling $\psi(M) < \log(M)$ proves the first claim. The second conclusion

follows from taking the expected value of each side of (17) and using $\mathbb{E}\min_m \hat{r}_m \le \min_m \mathbb{E}\,\hat{r}_m$. □

Again, the best bound occurs at $\beta = 1/2$. We compare by tabulating below these terms ψ and (16) of order $\log M$ in these bounds.

$M = \#\mathcal{M}$	2	5	10	20	40	100	1000
$4\log M$	2.8	6.4	9.2	12.0	14.8	18.4	27.6
$4\psi_1(M)$	4.0	4.0	4.8	7.8	10.7	14.4	23.6
$4\psi(M)$	1.1	2.9	4.4	6.1	7.9	10.5	17.7

And we see that the improved bound of order $\psi(M)$ is twice as tight as that of order $\log M$ for $M \le 10$ and the approximate upper-bound with $\psi_1(M)$ is very good for $5 < M \le 20$.

Corollary 10: The risk r of the mixture (3) of least-squares estimators $\hat{\mu}^m$ with weights (10), restricting β to $(0, 1/2]$, satisfies

$$r \le \min_{m\in\mathcal{M}} \left\{ r_m - \frac{2}{\beta}\log\pi_m \right\},$$

where r_m, taking value (2), is the risk of $\hat{\mu}^m$.

Proof: Starting with (12), this is another application of $\mathbb{E}\min_m \hat{r}_m \le \min_m \mathbb{E}\,\hat{r}_m$. □

Thus, with $C_m = -\log\pi_m$ and formula (2) for r_m, this upper-bound can be interpreted as an index of resolvability,

$$\min_{m\in\mathcal{M}} \left\{ \|\mathcal{P}_m\mu - \mu\|^2 + d_m + \frac{2}{\beta}C_m \right\},$$

an idealized trade-off among approximation error, dimension, and complexity of the models considered. This is a calibration of the error our model class $\mathcal{M}$ provides for $\hat{\mu}$ even if μ were known.

With the refinement for the case with π uniform ($C_m = \log M$), Corollary 9 gives sharper bounds than Corollary 10 does, but the latter allows us to control model complexities.

V. Conclusion and Discussion

We have developed an unbiased estimator for the risk of mixture estimators. For the case of linear models and least-squares components, this results in simple and accurate risk bounds. The mixture estimator adapts to the models considered by emphasizing those assessed to have low risks. The tight risk bounds, with explicit constants of 1 and 4 for the target $\min_m r_m$ and the order $\log M$ term, respectively, represent improved oracle inequalities for regression problems. Although the theory is offered in a somewhat restricted setting, it gives insight to understanding practical procedures in more realistic settings by careful comparison with the mixtures and risk bounds considered here.

Our bounds comply with the model selection bound of order $\log M$ in [7]. For leading-term models [8], the means μ_i correspond to ordered coefficients of orthogonalized regressors or basis functions. The mixture has the form $\hat{\mu}_i = c_i Y_i$ with c_i decreasing in i. Estimators of this type have good minimax properties [9], [10], and can adapt to functions in Sobolev classes via the Gaussian sequence models (1). Adaptation to more complex (e.g. Besov) classes is possible by considering all subsets of basis functions with model complexity control. More discussion along this line will appear in a journal paper.

Our weights (10) have been used explicitly with $\beta = 1$ by others. Buckland et al [11] offers numerical evaluations for the case with $\pi_m = 1/M$. See Hartigan [12] for resolving model weight ambiguity via hypothesis testing.

Demonstration of detailed risk properties of weighted regressions has been challenging. Analogous information-theoretic bounds for Bayes predictive density estimation have been developed in [13]–[17], with extensions to regression by Catoni and Yang. George [18], [19] also studied mixing estimators, with emphasis on shrinkage estimators, and provided an expression for the risk estimate of the mixture using Stein's result. Also, the presentation here for least-squares has a shrinkage estimator analogue [1].

References

[1] G. Leung, "Improving Regression through Model Mixing," Ph.D. dissertation, Yale University, 2004, Available at http://people.qualcomm.com/gleung/thesis.

[2] H. Akaike, "Information Theory and an Extension of the Maximum Likelihood Principle," in *Second Intl. Symp. Inform. Theory*, B. N. Petrov and F. Csáki, Eds., Budapest: Akadémia Kiado, 1973, pp. 267–281.

[3] C. Mallows, "Some comments on C_p," *Technometrics*, vol. 15, pp. 661–675, 1973.

[4] C. Stein, "Estimation of the mean of a multivariate normal distribution," in *Proc. Prague Symp. Asymptotic Statist.*, 1973, pp. 345–381.

[5] E. L. Lehmann and G. Casella, *Theory of point estimation*, 2nd ed. New York: Springer, 1998, (originally by E.L. Lehmann, 1983).

[6] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[7] A. B. Tsybakov, "Optimal rates of aggregation," in *Computational Learning Theory and Kernel Machines: Lecture Notes in Artificial Intelligence*, B.Scholkopf and M.Warmuth, Eds., vol. 2777. Heidelberg: Springer, 2003, pp. 303–313.

[8] G. Leung and A. R. Barron, "Information Theory, Model Selection and Model Mixing for Regression," in *Proc. Conf. Inform. Sci. Systems*, Princeton, 2004.

[9] M. S. Pinsker, "Optimal Filtering of Square Integrable Signals in Gaussian White Noise," *Problems in Information Transmission*, vol. 16, pp. 120–133, 1980, translated from Russian.

[10] R. Beran and L. Dümbgen, "Modulation of estimators and confidence sets," *Ann. Statist.*, vol. 26, no. 5, pp. 1826–1856, 1998.

[11] S. T. Buckland, K. P. Burnham, and N. H. Augustin, "Model Selection: An Integral Part of Inference," *Biometrics*, vol. 53, pp. 603–618, 1997.

[12] J. A. Hartigan, "Bayesian regression using Akaike priors," Yale University, New Haven, Preprint, 2002.

[13] A. R. Barron, "Are Bayes rules consistent in information?" in *Open Problems in Communication and Computation*, T. Cover and B. Gopinath, Eds. Springer, 1987.

[14] O. Catoni, "Mixture approach to universal model selection," Laboratoire de Mathématiques de l'Ecole Normale Supérieure, Paris, Preprint 30, 1997.

[15] ——, "'Universal' aggregation rules with exact bias bounds," Laboratoire de Probabilités et Modèles Aléatoires, CNRS, Paris, Preprint 510, 1999.

[16] Y. Yang, "Combining different regression procedures for adaptive regression," *J. Multivar. Anal.*, vol. 74, pp. 135–161, 2000.

[17] ——, "Combining forecasting procedures: some theoretical results," *Econometric Theory*, vol. 20, pp. 176–222, 2004.

[18] E. I. George, "Minimax multiple shrinkage estimation," *Ann. Statist.*, vol. 14, pp. 188–205, 1986.

[19] ——, "Combining minimax shrinkage estimators," *J. Amer. Statist. Assoc.*, vol. 81, pp. 431–445, 1986.