Information Theory and Mixing Least-Squares Regressions

Gilbert Leung, Member, IEEE, and Andrew R. Barron, Senior Member, IEEE

Abstract—For Gaussian regression, we develop and analyze methods for combining estimators from various models. For squared-error loss, an unbiased estimator of the risk of the mixture of general estimators is developed. Special attention is given to the case that the component estimators are least-squares projections into arbitrary linear subspaces, such as those spanned by subsets of explanatory variables in a given design. We relate the unbiased estimate of the risk of the mixture estimator to estimates of the risk achieved by the components. This results in simple and accurate bounds on the risk and its estimate, in the form of sharp and exact quadratic inequalities. That is, without advance knowledge of which model is best, the resulting performance is comparable to or perhaps even superior to what is achieved by the best of the individual models. We also provide simulations showing that the performance of these mixture estimators is better than that of a related model-selection estimator which picks a model with the highest weight.

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. It is common to employ a two-stage practice which first examines the data and picks a best model based on some 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, it is well-known that 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 necessarily follow from the selection procedure, and therefore can be overly optimistic.

Combining estimators from different models is an alternative to model selection in regression. In particular, we may take each model to be a linear subspace of the full space and the corresponding estimator for such a model is the least-squares projection of the observations onto that subspace. 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. Furthermore, the risk of the resulting mixture is not much more than an idealized target defined by the risk achieved by the best of the various estimators (one for each model considered). This is what Yang [1] calls combining for adaptation and the risk target is called the mixture projection target by Tsybakov [2] since a upper-bounds the risks of all model selection-based estimators. The general sharp risk bounds, or oracle inequalities, shown in this paper are obtained by a single inequality, which in turn is a direct consequence of the analysis in Section III. Moreover, the resulting mixture estimator outperforms better in simulations than a related model-selection estimator, which picks the estimate corresponding to the highest-weighted model. A primary motivation is that mixing often improves the risk in regression estimation as we shall see. Another motivation for such mixtures comes from consideration of Bayes estimation which is well-known to possess desirable properties in any statistical decision problem. With squared-error loss, the Bayes procedures are convex combinations of estimators weighted by the corresponding models' posterior probability (see Hoeing et al. [7] and the references cited therein), not model selection procedures.

A key tool in our analysis is the unbiased estimate of risk by Stein [4] for estimation of normal mean under squared-error loss. We adapt it to provide risk assessment for mixtures of general estimators and to produce risk bounds for the mixture of least-squares estimators in linear models. Mixture of shrinkage estimators, which are non-linear, is analyzed in a separate paper using similar techniques (see also [6]).

A. Overview

In regression and function estimation, we have observations $Y_i$ of response values plus independent Gaussian noise for $i = 1, 2, \ldots, n$. One estimator, under squared-error loss, the unknown mean $\mu$ of the random vector $Y \in \mathbb{R}^n$ distributed $\text{Normal}(\mu, \sigma^2 I)$, where for simplicity $\sigma^2$ is known. A linear subspace of $\mathbb{R}^n$ is the least-squares projection of the observations onto that subspace. In this paper, $X_m$ is a basis of $\mathbb{R}^n$, denoted by $X_m$ for which the mean $\mu$ is modeled as $X_m \beta$ for some unknown $\beta \in \mathbb{R}^m$. Let $\hat{\mu}_m$ be the least-squares projection of the observed $Y$ for each model $m$.

We form a convex combination of these estimators

$\hat{\mu} = \sum_{m=1}^{M} \omega_m \hat{\mu}_m,$

where the data-determined weights $\omega_m$ are chosen to give emphasis to models assessed to be better. In particular, for each model $m$, let $r_m$ be an unbiased estimate of the risk of $\hat{\mu}_m$,

$r_m = [Y - \mu_m]'^2 + \sigma^2(2d_m - n), \tag{1}$

Manuscript received April 2004.
G. Leung and A. R. Barron are with the Statistics Department of Yale University. E-mail: gilbert.leung@yale.edu, andrew.barron@yale.edu
in accordance with Akaike [7], [8], Mallows [9], or Stein [4], [5]. The unbiasedness means that \( E \hat{\mu} = E [\mu - \hat{\mu}^2] \) for each \( \mu \), where the expectation is taken with respect to the sampling distribution \( Y \) given \( \mu \). Then we define the weights to be
\[
\omega_n \propto \exp \left\{ -\frac{\beta \hat{\mu}^2}{2} \right\}, \quad \beta > 0,
\]
normalized to have sum unit across \( n \in M \). The tuning parameter \( \beta \) adjusts the degree of concentration of the weights on the models with small risk estimates. The two extremes are \( \beta \to 0 \), which gives the uniform distribution on \( M \), and \( \beta \to \infty \), which assigns non-zero weights to only the models with minimal estimated risk. Typical values are \( \beta = 1 \), which gives the weighted mixture a Bayes interpretation, and \( \beta = 1/2 \), which leads to the main risk bounds.

We will show that, the w-averaged risk estimate \( \sum_{\omega \in M} \omega_n \hat{\rho}_n \) is a crucial part of an unbiased risk estimate \( \hat{\rho} \) of the mixture \( \hat{\mu} \). In fact, for \( \beta \leq 1/2 \), it is an upper-bound,
\[
\hat{\rho} \leq \sum_{\omega \in M} \omega_n \hat{\rho}_n \leq \hat{\rho} \leq \sum_{\omega \in M} \omega_n \hat{\rho}_n + 4e^2 \log M,
\]
equation added above.

Taking expectation, we see that the risk satisfies
\[
E[|\hat{\mu} - \mu|] \leq \min \{ E[|\hat{\mu}^2 - \mu^2|] + 4e^2 \log M \}, \tag{5}
\]
The risk of the projection estimator and risk factor \( r \) can be presented in the usual squared bias and variance decomposition via the Pythagorean identity
\[
r_m = E[|\hat{\mu}_m - \mu|]^2 = |\hat{\mu}_m - \mu|^2 + \Delta_m^2, \tag{6}
\]
where \( \hat{\mu}_m \) is the projection of the true mean \( \mu \) into the subspace \( m \). Define the risk target \( r_m = \min_{m \in M} r_m \) as the minimum of these risks of the individual estimators,
\[
r = \min_{m \in M} E[|\hat{\mu}_m - \mu|^2]. \tag{7}
\]
It corresponds to the model with the optimal bias and variance tradeoff. This \( r \) is the main term in the bound (5) for the risk of the combined estimator. Indeed, \( |\hat{\mu} - \mu|^2 = \sum_{m \in M} (|\hat{\mu}_m - \mu|^2) \) in a sum of \( n \) terms, so typically \( r \) is much larger than the log \( M \) term in (5) (unless one has the surprisingly good fortune \( \beta \) is in one of the subspaces considered with dimension not much more than \( \log M \)).

It is sometimes useful to incorporate a deterministic factor \( r_0 \) in the model weight \( w_0 \) to take into account model complexity or model preferences in addition to risk assessment. Suppose such factors \( r_0 \) are assigned, where expressing them in the form \( r_0 = \exp(-C_0 \phi) \) and requiring that they sum to at most one endows \( C_0 \) with the complexity interpretation. Thus, models with low complexities are favoured. The new weights become
\[
\omega_n \propto \omega_n \exp \left\{ -\frac{\beta \hat{\mu}^2}{2} \right\}, \tag{2}
\]
where these combined weights \( \omega \) are again normalized to have unit sum. As before, the choice \( \beta = 1 \) has a Bayes interpretation, and \( \beta = 1/2 \) leads to the main risk bounds.

As in the case without \( \tau \), the w-averaged risk estimate is an upper-bound for the unbiased risk estimate \( \hat{\rho} \) of the mixture \( \hat{\mu} \) (with the new weights \( \omega \)) when \( \beta \leq 1/2 \), with equality at \( \beta = 1/2 \). Information theory also elucidates the risk analysis of mixing with this more general form of weights. The average risk estimate admits the following analogous representation
\[
\sum_{\omega \in M} \omega_n \hat{\rho}_n = \hat{\rho}_n + \frac{3}{4 \beta} \left( C_0 \phi \right) + \log \omega_0, \tag{3}
\]
where
\[
\tau = \arg \min_{\omega \in M} \left\{ \omega_n + \frac{3}{4 \beta} \left( C_0 \phi \right) \right\}.
\]

Moreover, the following risk bound is shown to hold:
\[
E[|\hat{\mu} - \mu|^2] \leq \min_{\omega \in M} \left\{ E[|\hat{\mu}_n - \mu|^2 + 4e^2 C_0 \phi] \right\}. \tag{4}
\]

The right side, expressed via (6) as
\[
\min_{\omega \in M} \left\{ |\hat{\mu}_n - \mu|^2 + C_0 \phi + 4e^2 C_0 \phi \right\}
\]
is an index of resolvability of \( \mu \) by the model class \( M \) which calibrates the mixture estimator by the best tradeoff in approximation, dimensionality, and complexity (corresponding to the three terms respectively) among the models in \( M \).

We note that the log \( M \) terms in (4) and (5) (except beyond the minimum) are now weighted under the \( C_0 \) term in (4) and (5). Not surprisingly, the latter recovers the former when the uniform model weights \( \hat{\rho} = 1/2M \) are used, but in this case, we will show tighter bounds due to a technical refinement.

B. Background

Bayesian interpretations of least-squares regressions date back to ideas of Bayes, Laplace, and Gauss. In particular, the linear least-squares projections in Gaussian models arise as the Bayesian estimators with (improper) uniform prior on the coefficients of linear combinations. Each associated posterior weight for such a model is proportional to \( \exp(-\|\cdot\|^2/2\sigma^2) \),
A. Risk Assessment for General Mixture

We use the notation \( a \sim \sum_{i=1}^{n} a_i b_i \) for the inner product and \( \nabla \) for the gradient: \( \nabla f_{\lambda}(\theta) \) where \( \nabla \theta = \partial f_{\lambda}(\theta) / \partial \theta \).

Suppose each estimator \( \mu_\lambda \) when expressed as a function of \( \lambda \) in \( \mathbb{R}^p \) is almost differentiable (that is, each of its coordinates can be represented by a directional integral, which is implied by continuity and piecewise differentiability) and that the associated almost-everywhere derivatives \( \nabla \mu_\lambda \) has finite absolute expectation: \( E[|\nabla \mu_\lambda|^2] < \infty \) for each \( \lambda \).

Then Stein [4, 5] gives an unbiased estimate \( \hat{\lambda}_n \) of the risk \( \lambda = E \left[ \frac{1}{n} \sum_{i=1}^{n} (\mu_\lambda - \mu)^2 \right] \), where \( \mathbb{E}[\hat{\lambda}_n] = \lambda \).

Our goal is to give an unbiased risk estimate for the mixture

\[ \hat{\lambda} = \sum_{\lambda \in \Lambda} \pi_\lambda \hat{\lambda}_n \]

where the weights \( \pi_\lambda \) are non-negative, sum to one, and almost differentiable. We further assume that \( E[|\nabla \pi_\lambda|^2] \) and \( E[|\nabla \hat{\lambda}_n|^2] \) are finite. We also suppose \( \Lambda \) is finite (though under mild conditions, the conclusions can be extended for infinite \( \Lambda \)).

The following theorem relates the unbiased assessment of the risk of \( \lambda \) to unbiased assessments of the risks of the individual estimators \( \mu_\lambda \).

Theorem I: With the above assumptions, an unbiased estimate of the risk of \( \lambda = E \left[ \hat{\lambda} - \lambda \right] \) of the mixture \( \lambda = \sum_{\lambda \in \Lambda} \pi_\lambda \mu_\lambda \) is given by

\[ \hat{\phi} = \sum_{\lambda \in \Lambda} \pi_\lambda \hat{\phi}_n(a, \nabla \lambda) \]

(8)

In addition, if

\[ \hat{\phi}_n(a, \nabla \lambda) = \sum_{\lambda \in \Lambda} \pi_\lambda \hat{\phi}_n(a, \nabla \mu_\lambda) \]

(9)

for almost differentiable \( \pi_\lambda \) and arbitrary constants \( \theta \) then

\[ \hat{\phi} = \sum_{\lambda \in \Lambda} \pi_\lambda \hat{\phi}_n(a, \nabla \lambda) \]

(10)

Remark: One can adjust \( \hat{\phi}_n(a, \nabla \lambda) \) by adding any function of \( Y \) that does not depend on \( m \) without changing either the value of \( \hat{\phi} \) or the validity of (10).

This unbiased estimate (8) of risk has three terms. The principal term \( \sum_{\lambda \in \Lambda} \pi_\lambda \hat{\phi}_n(a, \nabla \lambda) \) is the weighted average of the individual risk estimates. This average is a crude risk assessment, possibly biased. However, with suitable design of the weights, it becomes an upper-bound for the unbiased risk assessment \( \hat{\phi} \) for the mixture of least-squares estimators, as we shall see. Also, an information-theoretic representation of this term yields the conclusion that it is not much larger than \( \min_{\lambda \in \Lambda} \hat{\phi}_n(a, \nabla \lambda) \).

The second term \( \sum_{\lambda \in \Lambda} \pi_\lambda \mu_\lambda \) wondrously illustrates an advantage of model mixing. If the estimates \( \mu_\lambda \) vary with \( m \) that is, if the fits are different for different \( m \), then averaging them (with weights \( \pi_\lambda \)) leads to a reduction in the unbiased risk assessment given by the weighted average of the squared distance of the \( \mu_\lambda \) from their centroid \( \hat{\mu} \).

A nice feature of the unbiased risk estimator is that it clearly reveals this reduction based on variability of estimates (as \( m \) varies), rather than based on the classical variance of the estimators.
(which addresses variability with changes in the sample, not changing in the estimates with m for a given sample).

The third term \(-2 \sum_{m \in M} [w_m (Y \log w_m) - (\mu - \mu_m)]^2\) quantifies the effect of the data-sensitivity of the weight (through their gradient with respect to the data Y). Constant weights would make this term zero, but would not permit means to adapt to the fit to the models that have smaller \(f_m\). Finally, the exponential form of weights (9) gives a particularly clean mixture risk estimate (10) 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").

However, in the corollary after the proof of the theorem, we will show how to design weights such that this third term can be cancelled with the second term.

**Proof of Theorem 1:** According to [4], [5], an unbiased estimate of the risk of an estimator \(\hat{\beta}\) is given by

\[
\hat{\tau} = \|\hat{\mu} - Y\|^2 + 2 \sum_{i=1}^n \nabla \hat{\mu}_i - n,
\]

as long as each \(\nabla \hat{\mu}_i\) has finite absolute expectation, but our assumptions are sufficient to ensure this. Now with a variance calculation using the weights \(w_i\) as a distribution on \(\mathcal{M}\), summed over each of the coordinates, we rewrite the first term above as

\[
\|\hat{\mu} - Y\|^2 = \sum_{m \in M} w_m \|\hat{\mu}_m - Y\|^2 = \|\|\mu - \mu_m\|^2\|.
\]

The second term can be expanded via differentiation under the summation sign,

\[
\nabla_i \sum_{m \in M} w_m \hat{\mu}_m = \sum_{m \in M} w_m \nabla_i \hat{\mu}_m + \sum_{m \in M} (\nabla_m w_m) \hat{\mu}_m
\]

and we recognize in these components the terms of

\[
f_m = \|\|\mu - \mu_m\|^2\| + 2 \sum_{i=1}^n \nabla_i \hat{\mu}_m - n.
\]

such that

\[
\hat{\tau} = \sum_{m \in M} w_m \left[ f_m - (\mu - \mu_m)\right]^2 + 2 \sum_{i=1}^n \sum_{m \in M} (\nabla_m w_m) \hat{\mu}_m
\]

after exchanging the order of summation over \(m\) and \(i\). The last term here is the same as

\[
2 \sum_{i=1}^n \sum_{m \in M} (\nabla_m w_m) \hat{\mu}_m
\]

because \(\sum_{i=1}^n (\nabla_m w_m) = \left[ \nabla \sum_{i=1}^n w_m \right] \mu_m = 0\) (as the weights \(w_m\) sum to a constant). The upper display equals

\[
2 \nabla (\tau_{\mu_m}) = 2 (\|\mu - \mu_m\| - \mu_m)
\]

by exchanging the order of summation again and the first claim (8) follows.

For the second claim, \(\nabla \log p_m\) equals \(-\nabla \log p_m\) minus a function (the gradient of \(\log \sum_{m \in M} \exp(-p_m) w_m\)) which does not depend on \(m\). Now since \(\mu - \mu_m\) has \(w\)-average being the null vector 0, its inner product with a quantity not depending on \(m\) averages to 0 under the weights \(w_m\) so that we are left with the \(\nabla \log p_m\) (Y) term. This proves (10).

Now we show the cancellation of the third term by the second.

**Corollary 2:** If \(p_m(Y)\) has gradient \(\beta(Y - \mu_m)\) for some \(\beta \geq 0\), then

\[
\hat{\tau} = \sum_{m \in M} w_m \left[ f_m - (1 - 2\beta)\|\mu_m - \mu\|^2\right].
\]

(12)

In addition, if \(\beta \leq \frac{1}{\|\hat{\mu}\|^2}\), the risk estimate can be bounded by

\[
\hat{\tau} \leq \sum_{m \in M} w_m f_m,
\]

with equality when \(\beta = \frac{1}{\|\hat{\mu}\|^2}\).

**Remark:** The parameter \(\beta\) in the weights \(w_m\) controls the relative importance of averaging across models (small \(\beta\)) and picking out the one that was empirically best (large \(\beta\)). When \(\beta\) is strictly less than \(\frac{1}{\|\hat{\mu}\|^2}\), we see risk improvement due to averaging independent of the observation — the unbiased risk estimate is strictly less than \(\sum f_m w_m\).

**Proof:** From the stated assumption of the form of \(p_m(Y)\), we see that after adding a function not depending on \(m\), \(\nabla \log p_m\) (Y) matches a multiple of \(\mu - \mu_m\) so the first claim follows from (10). The rest of the corollary is a special case of (12).

**Remark:** In any Bayes mixture, for which \(w_m(Y)\) is the posterior probability of model \(m\), the expected \(-\log p_m\) arises as the log of the marginal density of \(Y\) for model \(m\) under any general prior for \(\mu_m\) which yields a posterior. The posterior Bayes estimator \(\hat{\mu}_m\) for each model \(m\) satisfies \(\hat{\mu}_m = Y - \nabla \log p_m\) (Y) (e.g. see [20]). Thus Corollary 2 applies to Bayes mixtures with \(\beta = 1\).

In this paper, we focus attention on weights \(w_m\) that emphasize models with small risk estimates \(f_m\). We do this by taking the exponent \(w_m\) to equal \(\hat{\tau}_{f_m}/2\). (We will comment more about the choice of \(\beta\) later.) In the case of linear models, we shall see that with such weights the condition of Corollary 2 is satisfied.

**B. Mixing Least-Squares Estimators**

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 procedure, or via the explicit projection matrix \(\hat{\mu}_m = X_m (X_m' X_m)^{-1} X_m' Y\), such that \(\nabla \log p_m\) (Y) equals \(-\nabla \log p_m\) (Y) minus a function (the gradient of \(\log \sum_{m \in M} \exp(-p_m) w_m\)) which does not depend on \(m\). This exists an orthonormal basis for \(\mathbb{R}^n\) that is natural for \(m\) such that the first \(d_m\) elements of this basis span \(m\). A
point $Y$ in $R^n$ can be represented by a linear combination of these basis elements, whose coefficients are obtained by inner products of $Y$ with the basis elements. In other words, there exists an orthonormal matrix $Q = Q_0$, such that any $Y \in R^n$ has representation $QY$ with coefficients obtained as $Z = Q^*Y$, which is natural for $m$. The said basis consists of the columns of $Q$ (which can be obtained by, say, Gram-Schmidt). Moreover, $Z \sim \text{Normal}(0, I)$, with $\theta = Q^*\mu$ and $\theta^m = (Z_1, \ldots, Z_m, 0, \ldots, 0)'$ is the corresponding least-squares projection in the new coordinate system which simply retains the first $d_m$ elements of $Z$. Similarly, the projection $\theta^m$ of $\theta$ has the representation

$$(0, \ldots, 0, d_m, 0, \ldots, 0)'$$

in this system. Then, since the norm is preserved by orthonormal transformations, the risk of $\mu^m$ is

$$r_m(\mu^m) = \mathbb{E} [|\theta^m - \mu|^2] = \mathbb{E} [\theta^m - \theta]^2 = \sum_{k=d_m}^{m} \theta_k^2 + d_m.$$  

Now with $\theta^m$ as the projection of $\theta$ into $m$, the first term above equals

$$\sum_{k=d_m}^{m} \theta_k^2 = |\theta - \theta^m|^2 = |\mu - \mu^m|^2.$$  

Thus, we re-established the Pythagorean identity (6) for the risk,

$$r_m = |\theta - \theta^m|^2 + d_m.$$  

Similarly, the unbiased risk estimate $r_m$ is more easily computed in the new coordinate system. From (13), and the unbiasedness of $Z_k^2 - 1$ for $\theta_k^2$ for each $k$, we deduce that the following is an unbiased estimate for $r_m$:

$$r_m = \sum_{k=d_m}^{m} Z_k^2 + 2d_m - n,$$

which can also be obtained by (11). More succinctly,

$$r_m = |Z - \theta^m|^2 + 2d_m - n = |\mu - \mu^m|^2 + 2d_m - n.$$  

To obtain the gradient of $r_m$ with respect to $Y$, it is convenient to compute the gradient first with respect to $Z$ and then to account for the derivatives $dZ_k/dY$. In accordance with the multivariate chain rule,

$$d r_m = \sum_{k=d_m}^{m} Z_k d Z_k = \sum_{k=d_m}^{m} (dZ_k/dY) dY.$$  

Now

$$dZ_k = 2Z_k d(Y_k) = 2(Z_k - \theta_k).$$

Thus the gradient with respect to $Z$ of $r_m$ is $2(Z - \theta^m)$. Multiplying this by $Q$ on the left gives $2(Y - \mu^m)$ as the gradient with respect to $Y$ of $r_m$, as claimed.

Thus by choosing $\mu^m$, proportional to $\mu_0 \exp(-\beta d_m/2)$ with linear least-squares estimators the condition for Corollary 2 is satisfied. With these weights at $\beta = \frac{1}{2}$, the resulting expression in (10) is only the $w$-average of the unbiased risk estimates $r_m$ of the individual models.

### III. INFORMATION-THEORETIC CHARACTERIZATION OF RISK ASSESSMENT

We analyse the average risk estimate $\sum m w_m r_m$ in this section. It is the primary term in the estimate for the risk of the mixture $\tilde{\mu}$, and for $\beta \leq \frac{1}{2}$, it has special roles in the unbiased risk estimate $\tilde{r}$ as concluded by Corollary 2.

**Remark:** When the unknown mean $\mu$ can be well-approximated by multiple models $m$, the resulting risk of the mixture at $\mu$ would not be very sensitive to the choice of $\beta$ around the values of interest at 1 (Bayes) and $\frac{1}{2}$ (clean bound).

Since the choice $\beta = \frac{1}{2}$ makes this average risk estimate unbiased for the risk of $\mu$, we will set it so in this section for a brisk exposition. The generalization to any $\beta > 0$ can be obtained by replacing 4 with $2/\beta$, though the average risk estimate will no longer be unbiased when $\beta \neq \frac{1}{2}$. This will be explicitly done in the next section where a tighter bound is proven for the case with weights (2).

### A. Sharp Bounds on Risk Estimate of Mixture

The following enunciates the relationship between the average $\sum m w_m r_m$ and the minimum $r_m$ for some minimizing $r_m \in M$.

**Theorem 4:** (i) For each $m \in M$, let $w_m = \frac{\exp(-r_m/4)}{\sum_{m \in M} \exp(-r_m/4)}$ (16) then with $r$ being any model achieving $r = \min_m r_m$, the unbiased risk estimate for $\mu = \sum_m w_m r_m$ satisfies

$$\tilde{r} = \sum_{m \in M} w_m r_m = r + 4 \left[ \frac{H(u)}{6} + \log w_m \right]$$

$$\leq r + 4 \frac{H(u)}{6}$$

(b) More generally, for each $m$, let $w_m = \frac{\exp(-r_m/4)}{\sum_{m \in M} \exp(-r_m/4)}$, where $w_m = \exp(-C_m)$ and $\sum_{m \in M} w_m \leq 1$. Then, here with $r$ being any model attaining $\min_m r_m + 4C_m$ the unbiased risk estimate for $\mu$ satisfies

$$\tilde{r} = \sum_{m \in M} w_m r_m = r + 4 \left[ C_m - D(u/2\pi) + \log w_m \right]$$

$$\leq \min_{m \in M} [r_m + 4C_m].$$

(19)
Proof: Part (a) is a special case of part (b) with \( m_0 = \frac{1}{3} \mu \). For part (b), observe that

\[
\hat{r}_m = 4 \log \frac{m_0}{\mu} - \log \sum_{m_0 \leq m} \text{exp}(-\hat{r}_m / 4) = \hat{r}_m + 4 \log \frac{C_m - \log \frac{m_0}{\mu} + \log \mu \mu_0}{\mu_0}.
\]

Thus, the equality follows by averaging over \( m \in M \) with weights \( \mu \). The inequality results since \( D \geq 0 \) and \( \mu_0 < 1 \) (the logarithm of the latter is strictly negative).

Therefore, for the first form of weights, the average risk estimate (17) is the minimum of the individual risk estimates plus a price for mixing, a function of the mixing weights \( \mu \). If the weights \( \mu \) are concentrated on mostly one model \( n_i \), then \( \hat{H}(n) \) is close to zero and the combined risk estimate is very close to the minimum \( \hat{r}_m \). Moreover, if in the distribution \( \mu \) there are several, say \( J \), values of \( m \) with nearly minimal risk estimates, then accounting for those \( J \) values in the sum on the right side of (20), one has a further reduction of about \( 4 \log J \) from \( \sum_{J} \mu J \), clearly showing the advantage of the mixing. In any case, since \( R \) is less than the log-cardinality, the average risk estimate cannot exceed \( \hat{r}_m \) by a relatively small amount \( 4 \log \mu \). This bound will be improved in the next section.

Mixing with the general weights including \( m_0 \) also has an interpretation. The average risk estimate is the minimum of the complexity-inflated risk estimate plus a reduction due to mixing, a function of \( \mu \) and \( \hat{r} \). If the data-dependent weights \( \mu \) differ little from the constant weights \( \mu \), then the quantity \( \sum_{m} \mu m \hat{r} \) would be close to its upper-bound \( C_m \). Moreover, if \( \mu \) is the distribution \( \mu \) there are, say \( J \) values of \( m \) with nearly minimal complexity-inflated risk estimates \( \mu_m + 4 C_m \), then again (20) shows a further reduction of about \( 4 \log J \) from the average risk estimate.

Remark: The condition \( \sum \mu \exp(-\hat{r}_m) \leq 1 \) is of course Kree's inequality [21] in base \( e \) and the model complexity is connected to the length of some codeword (in bits) that describes the model. However, our theory does not require such an interpretation.

Characterizing the average risk estimate by the minimum \( \hat{r}_m \) is useful as it leads directly to a risk bound.

b. Risk Bound for Mixing Least-Squares Regressions

Corollary 5: The risk \( r = \mathbb{E} [\| \hat{\mu} - \mu \|^2] \) of the mixture of least-squares regressions \( \hat{\mu} = \sum_{m} \mu_m m \hat{\mu}_m \) with weights (16) satisfies

\[
r \leq \min_{m \in M} \left( \hat{r}_m + 4 \log(\lambda M) \right),
\]

where \( \hat{r}_m = \mathbb{E} [\| \hat{\mu}_m - \mu \|^2] \), taking value (14), is the risk of \( \mu_m \) and \( \lambda M \) is the cardinality of \( M \). Mixing with weights (18) yields a risk that satisfies

\[
r \leq \min_{m \in M} \left( \hat{r}_m + 4 C_m \right),
\]

where \( C_m = \log(1/\mu_m) \).

Proof: To show the second inequality, we take the expected value of each side of (19). This recovers the risk \( r \) by the unbiasedness of \( \hat{r} \) on the left. Applying

\[
\mathbb{E} \left[ \min_{m \in M} \left( \hat{r}_m + 4 C_m \right) \right] \leq \mathbb{E} \left[ \min_{m \in M} \left( \hat{r}_m + 4 C_m \right) \right]
\]

for the right side yields the second statement. The proof for the first statement is the same.

Note that the mixture \( \mu \), its risk estimate \( \hat{r}_m \) and its risk \( r \) all change with the weights, e.g. from (16) to (18). But the risks for the individual models (14)

\[
r_m(\mu) = \mathbb{E} [\| \hat{\mu}_m - \mu \|^2] = \mathbb{E} [\| \hat{\mu} - \mu \|^2] + \lambda d_m,
\]

where \( \mu_m \) is the projection of \( \mu \) onto the linear space \( m \), and hence, the risk target, depend on neither the estimator nor the weights.

The risk is bounded by an idealized trade-off among error of approximation \( \| \hat{\mu}_m - \mu \|^2 \), dimension \( d_m \), and complexity \( C_m \) of the models considered, expressed via an index of reusability

\[
\min_{m \in M} \left[ \frac{\| \hat{\mu}_m - \mu \|^2}{d_m + 4 C_m} \right].
\]

This is a calibration of the error our model class \( M \) provides for \( \hat{\mu} \) even if \( \mu \) were known. The first term \( \| \hat{\mu}_m - \mu \|^2 \) entails a sum \( \sum_{m} \mu_m (\| \hat{\mu}_m - \mu \|^2) \) of squared approximation errors for the \( n \) means \( \mu_m \). For large \( n \), this sum is naturally the dominant term among the three, unless the unknown \( \mu \) is extrinsically in, or close to one of the linear spaces considered. Which of the remaining terms, \( d_m \) and \( 4 C_m \), is larger depends on the model that yields the best overall trade-off. We will discuss this trade-off at greater length in Section VI.

IV. Refinements

A. Tighter Bounds

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

\[
\psi \log M - 1 = 1.
\]

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

\[
\psi \leq \max \left\{ K, \log \frac{M - 1}{K - 1} \right\}.
\]

by considering separately whether \( \psi \leq K \) or not.

Theorem 7: Given the risk estimates \( \hat{r}_m \) for a finite collection \( m \in M \) of estimators \( \hat{\mu}_m \), the average risk estimate using weights

\[
w_m = \frac{\text{exp}(-\beta \hat{r}_m)}{\sum \text{exp}(-\beta \hat{r}_m / 2)}
\]

with any \( \beta > 0 \) satisfies

\[
\sum \mu_m w_m \hat{r}_m \leq \min_{m \in M} \left( \hat{r}_m + 2 \beta \log (\lambda M) \right).
\]

Proof: First, observe that

\[
\hat{r}_m \geq \frac{1}{\beta} \log \frac{1}{w_m} - \log \sum \mu \text{exp}(-\beta \hat{r}_m / 2).
\]
which, upon averaging with \( w \) over \( m \), yields

\[
\sum_{m \in M} w_m \epsilon_m = \epsilon_t + \frac{2}{\beta} [H(w) + \log w_0] \tag{24}
\]

where \( \epsilon_t \) is a model achieving the minimum risk estimate \( \epsilon_t = \min_m \epsilon_m \). Consider \( m \) a random variable on the probability space \((M, w)\) and define an indicator \( A = \{m \neq \hat{m}\} \) on the same space. We can then use chain rule for the entropy of \( m \) conditional on \( A \) (Ch. Fano’s inequality),

\[
H(m) = H(m | A) P_A + H(A),
\]

where \( H(A) = H(B \cap A) = H(1 - w_0) \) is the binary entropy of the weight \( w_0 \). Then the above can be expressed as

\[
H(w) = (1 - w_0) H(\beta) + H(w_0) + \log w_0,
\]

where \( \{\epsilon_m : m \neq \hat{m}\} \) are the the normalized weights on \( M/\{\hat{m}\} \). Thus, (24) becomes

\[
\sum_{m \in M} w_m \epsilon_m - \epsilon_t = \frac{2}{\beta} (1 - w_0) H(\beta) + H(w_0) + \log w_0.
\]

Hence, the bracketed terms on the right are upper-bounded by

\[
(1 - w_0) \log(\beta M - 1) + H(w_0) + \log w_0
\]

which is concave in \( w_0 \) and equals

\[
(1 - w_0) \left[ \log(\beta M - 1) - \log \frac{1}{w_0} \right].
\]

Setting to zero the first derivative of (25) with respect to \( w_0 \), we see that the maximum of the bound occurs at \( w_0 = w_t \) satisfying

\[
\log(\beta M - 1) - \log \frac{1}{w_t} = \frac{1}{w_t - 1}.
\]

In terms of the odds of the event \( A \), i.e. \( (1 - w_0)/w_0 \), the maximum of the bound occurs at

\[
O_t = \frac{1 - w_t}{w_t} = \frac{1}{w_t - 1} = \log \beta M - 1.
\]

Multiply (26) by \( w_t \) and substitute this back in (25), which incidentally becomes the right side in the above display. The desired bound thus follows from the definition of \( \psi \).

An upper-bound for \( \psi(\beta M) \) is provided by (21) using

\[
K = \log(\beta M - 1) - 1,
\]

which is not a bad approximation when \( \beta M \) is large. To see this, first we note that a numerical method for obtaining \( \psi \) is to iterate

\[
K_{\text{new}} = \log \frac{\beta M - 1}{K_{\text{old}}} - 1
\]

until the fixed point for \( K \) is reached. Then, with the above \( K \) as an initial value, one iteration subtracts from the old \( K \) the value log \( K \). The latter is roughly log \( \log(\beta M) \) which is much smaller than log \( \beta M \) for large \( \beta M \) so the iteration will not affect the value much relatively speaking (i.e. \( K \) is already near the fixed point). Perhaps a picture is worth a thousand iterations, so refer to Figure 1 for the comparison of \( \psi \) with its estimates.

Now we are ready for the refined risk bound.

![The Function \( \psi(M) \) and its Estimates](image_url)

**Fig. 1.** The Function \( \psi(M) \) and its Estimates

**Corollary 8:** If \( \beta \) \( \mu \) least-squares regressions with risk estimates \( \epsilon_m \) in (1), then the mixture estimator \( \hat{\mu} = \sum w_m \hat{\mu}_m \) using weights (22) has a risk that satisfies

\[
E[|\hat{\mu} - \mu|^2] \leq \min_{m \in M} \epsilon_m + \frac{2k(\beta M)}{\beta}
\]

for each \( \beta \leq \frac{1}{\epsilon_t} \), where the risks of the individual estimates \( \epsilon_m \) are given in (14).

**Proof:** Corollary 2 implies that the unbiased risk estimate for \( \hat{\mu} \) is upper-bounded by the average risk estimate for this range of \( \beta \), which in turn is bounded as in (23). Thus the conclusion follows from taking the expected value of each side of (23) and using \( E \min_m \epsilon_m \leq \min_m E \epsilon_m \).

We compare the risk bound with \( \psi \) and its approximation \( r_t + \psi(\beta M) \leq r_t + 4 \max(1, (\log M - 1)) \), as tabulated below. Let \( \psi(\beta M) = \max(1, (\log M - 1)) \) be the an approximation to \( \psi \).

<table>
<thead>
<tr>
<th>( M )</th>
<th>( \log M )</th>
<th>( 4 \log M )</th>
<th>( 4 \psi(\beta M) )</th>
<th>( 4 \psi(\beta M) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2.8</td>
<td>6.4</td>
<td>9.2</td>
<td>12.0</td>
</tr>
<tr>
<td>3</td>
<td>4.0</td>
<td>4.0</td>
<td>4.8</td>
<td>7.8</td>
</tr>
<tr>
<td>4</td>
<td>1.1</td>
<td>2.9</td>
<td>4.4</td>
<td>6.1</td>
</tr>
</tbody>
</table>

**B. Worst Case Sequences**

Suppose we have a sequence of orthonormal basis vectors \( \phi_1, \ldots, \phi_n \) and consider the leading-term models \( m \) spanned by \( \phi_1, \ldots, \phi_n \), where we conveniently abuse notation to identify \( m \) with its dimension \( d_m \). Let \( X_{d_1}, X_{d_2}, \ldots, X_{d_m} \) be the coefficients in the representation of \( Y \) using these basis functions. We also include the zero-dimensional model which ascribes \( \mu = 0 \). Thus with \( 0 \leq \mu \leq 5 \) there is a collection of \( M = n + 1 \) models. The least-squares estimator zeroes out the coefficient past \( m \). We ask for what \( Z \) sequence is the excess of \( \psi(T) \) maximal.

Mixing the leading-term estimators with \( \beta = \frac{1}{\epsilon_t} \) gives

\[
T(H) = r_t + 4[H(w) + \log w_0]
\]

In addition, since \( \beta M = n + 1 \), this can be bounded by

\[
T < r_t + 4 \max(1, (\log n) - 1).
\]
For large $n$, the implied bound

$$H(w) + \log w_n \leq \log n$$

is tight in the sense that, there exists a sequence of $Z$ (induced by the dimension $n \to \infty$) such that

$$\lim_{n \to \infty} \frac{H(w) + \log w_n}{\log n} = 1.$$  \hspace{1cm} (27)

Indeed, for each $n$, we pick a worst-case $Z$ as follows. Let $k > 0$ and

$$|Z_k|^2 = \begin{cases} k + 2 & \text{if } i = n \\ 2 & \text{if } i < n \end{cases}$$

Then

$$r_n = \frac{2n - n}{2n - 1} + (k + 2 - n) = n + k$$

Thus $r_n = r_n = n$ and $w_n$ is roughly uniform over the first $n$ models ($m < n$) with $r_n \sim \log n$ for $n$ large.

As the difference $k = r_n - r_n$ between the risk estimates for the minimal model $m = n$ and other models grows large, the weights $w_n$ concentrates more at $m = n$. We can put $k = n$ such that $w_n$ tends to 1 and $\log w_n$ tends to 0. This demonstrates the limit in (27). The sequence $Z$ used here is highly unlikely. Taking expectations, the corresponding risk bound with $E[H(w)]$ need not be tight.

In our simulations (shown in the next section) using $\beta = 1$, the computed risks for various choices of $\beta$ suggest that the excess beyond $r_n$ is usually less than $\log n$, better than the upper bounds of $4 \log n$ or $4\log(n+1)$ given.

The nested subset favours the leading coordinate at the detriment of the trailing ones. In general, if the trailing observations are large, as $Z_m$ is in this toy example, the fact that it is left out under most of the leading-term models contributes to the overall high risk. It is expected that if the models considered "cove" the variables $Z_m$ more uniformly, the maximum risk of mixture estimator can be reduced.

V. SIMULATIONS

In this section, we will show simulations based on the leading-term models for $n = 20$. Besides the mixture estimator proposed, we will examine the AIC estimator in this nested model setting as well for comparison.

To learn something from the simulations, we impose structure on $\beta$, and the function estimation context is chosen for such considerations. Thus, $\theta$, the $\theta_i$ is considered as the coefficient of an orthonormal basis expansion of the function, or signal, to be estimated. Here we have simulation results divided into three scenarios for the underlying signal $\beta$.

1) Constant One-Block: The first case describes $\beta$ when $\beta$ is indeed a low-pass signal.
2) Gradual Decay: The second is probably quite typical in reality when $\beta$ decays as the reciprocal of $i$.
3) Odd or Even Function: The third describes an odd or even function parameterized by the Fourier basis, so only every other coordinate of $\beta$ is non-zero.

In all of the following plots, we have taken the variance in each dimension $\sigma^2$ to be 1. The least-square estimator $\hat{\beta}_{LS}$, a risk of $n = 20$. Our bound is $4.46$ is in excess beyond the risk target. But the simulations below show that the true cost for mixing is probably around $\log n = 5$.

In the set of simulation results, we have used $\beta = 1$ in our mixture estimator, because this corresponds to a Bayesian procedure. But we have also tried $\beta = 2/3$ and $\beta = 2$. The performance of the mixture estimator is not very sensitive to the choice of $\beta$ in this region.

For the following plots, the risk target is directly computed, while the risk $r(\theta)$ of our mixture estimator and the risk of AIC model selection estimator are computed by averaging the losses of these estimators for a very large number of independent draws of samples, each of size $20$, from the Normal $N(0, \sigma^2)$ and the specified coefficients $\theta_i$.

Note that the $x$-axes of the following plot do not have a common scale.

1) Constant One-Block: The risk target is

$$r(\theta) = \min \{\beta_1^2, 10\}.$$  

First, it is clear that the target excludes the trailing ten coefficients since they are in fact zero. So we only need to consider the first ten coordinates of $\beta$ together since they are constant. If $\beta_1^2 < 10$, we see better off leaving them out since the bias so incurred is less than the variance of 10 if we include them. The kink of the target above at $\beta_1^2 = 10$ is due to this minimum operation.

![Fig. 2. Risks and Target: Vanishing One-Block. $\theta^2 = \{0, 10\}$](image)

Figure 2 shows that the AIC mixture estimator performs only about 2 worse than the risk target at small and large $\beta$, but misbehaves, and even beats the target around $\beta_1^2 = 10$. The AIC leading-term selection estimator is uniformly worse than our mixture, but these two are close when $\beta_1^2$ is large. This is expected since the $\beta$ in this case is represented exactly by the model $m = 10$ such that AIC picks it correctly when the signal-to-noise ratio is high; the adaptive weights in our mixture give strong emphasis on the right model as well.
2) Gradual Decay: For this scenario, an integral approximation shows that the tail sum
\[
\sum_{i=0}^{\infty} r_i^2 \approx \|\theta\|^2 \left( 1 - \frac{\log n}{\log \pi} \right)
\]
and the minimum of \( r_n \) occurs at roughly \( m_i = \min \{ n, \|\theta\|^2 / \log n \} \). Note that now not any one of the component estimators considered in our model class truly describes the underlying \( \theta \). Also, the risk target increases up to \( n \) as \( \|\theta\|^2 \) increases.

![Risks & Target: Gradual Decay](image)

Fig. 3. Risks and Target: Gradual Decay. \( \theta_i^2 = 1 / i \)

Figure 3 says that our mixture estimator tracks the risk target very well, a remarkable fact since the \( \theta \) is not one of the leading-term type. In fact, for moderately sized \( \theta \), the mixture performs slightly better than the risk target. The AIC selected estimator, however, starts out being roughly the same as the mixture when \( \theta = 0 \), but ultimately performs worse uniformly by 2 or 3 as \( \theta \) is non-zero.

3) Odd or Even Function: The risk target for this case is
\[
r_n = \min \{ \|\theta\|^2, 19 \},
\]
and it switches from the quadratic to constant at \( \|\theta\|^2 = 19 \). Since the leading-term models cannot exclude individual coordinates in a leading block, this kind of \( \theta \) essentially resembles a single constant block to the target, and hence the form of \( r_n \) analogous to that in the constant one-block case, with the only change of the saturation point of 19 instead of 10 for the cardinality of the block.

As before, the mixture has an overhead of 2 at \( \theta = 0 \) but then it matches the target risk at around the saturation point of \( \|\theta\|^2 = 19 \). Its risk creeps beyond \( n = 20 \) a little right after \( \|\theta\|^2 \) reaches 19 before it tops down to almost the risk target (up to an offset of 0.5) as \( \theta \) becomes large.

The behavior of the AIC estimator starts off being roughly the same as that of our mixture when \( \theta \) is small, but then it overshoots the target, by quite a bit, and at around \( \|\theta\|^2 = 19 \), then it asymptotes to the risk of the mixture again as \( \theta \) becomes large.

![Risks & Target](image)

Fig. 4. Risks and Target: Odd or Even Function. \( \theta_i^2 = i \cdot \log(i) \)

VI. COMPLEXITY, APPROXIMATION, AND ADAPTATION
In this section, to facilitate discussion of approximation and risk on a standardized scale, we shall use the average squared error \( \|\theta\|^2 = \sum_{i=0}^{n-1} r_i^2 / n \). Thus the sum of squared error loss is divided by \( n \) and the corresponding risk \( r \) of the combined estimator is bounded by the index of resolvability
\[
r_0 \leq \min \left\{ \|\theta\|^2, \sum_{m=1}^{\infty} \frac{4C_m}{m} \right\}.
\]
Thus we have a trade-off between approximation error, dimension relative to sample size, and complexity relative to sample size.

Suppose a dictionary \( \mathcal{D} \) of candidate basis vectors \( \varphi \in \mathbb{R}^P \) is available. These may arise in practical regression problems from a possibly large number of explanatory variables or transformation thereof. In function modelling, they take the form \( \varphi = (\varphi(x_1), \ldots, \varphi(x_n))^T \) by evaluation of a function \( \phi(x) \) at given input values \( x_1, \ldots, x_n \). Typical choices include Fourier polynomial, spline, wavelet, or artificial neural network (sigmoidal) basis functions. Note that in some of these settings the basis vectors in \( \Phi \) are not to be orthogonal (indeed, there are often too many legitimate candidate vectors placed in the dictionary to permit orthogonality as vectors in \( \mathbb{R}^P \)). From the dictionary we form a collection of linear models which we index as \( m \in \mathcal{M} \).

We consider two types of models to combine: those corresponding to general subsets of particular sizes from \( \Phi \) (permitting the mixture of subsets in the combined estimator) and those that include all terms in some prescribed form up to particular orders (permitting the mixture of model orders).

A General Subsets
Consider first general subsets. Suppose the dictionary is of size \( D \). Then for arbitrary subsets \( m \) of size \( d_m \),
\[
C_m = \log \left( \frac{D}{d_m} \right) + \log(n + 1)
\]
is a valid code-length for $d_n$, satisfying $0 \leq d_n \leq \min(0, n)$ where the empty subset model with $d_n = 0$ is $\emptyset$. The main term in this complexity is the top cardinality of models of each dimension $d_n$, which to first order is $C_{d_n} \sim \omega_1 H(d_n, D)$ which is approximately $d_n \log(D/d_n)$ for $d_n$, of smaller order than $D$. Suppose the dictionary size $D$ is greater than or equal to the sample size $n$ and not more than a polynomial in $n$, and consider $d_n$ less than a fractional power of $D$. Then the complexity of these subset models is proportional to $d_n \log(n)$ and substantially larger than the dimension $d_n$ in the resolvability bound. This larger cost with the lower factor is required to achieve adaptation to arbitrary subsets of $\Phi$. The other term $\log(n-1)$ corresponds to the length of a description of $d$.

The variable subsets can be crucial for achieving small approximation error with an appropriate number of terms. Convex subset classes as discussed in [2], [14], [22]-[27] provide a setting in which the flexibility of variable subsets is required. Suppose the mean response vector $\mu$ has representation $\mu = \sum \beta_i \phi_i$ summing over a possibly large number of vectors in the dictionary, with coefficients satisfying $\|\beta_i\|_1 \leq c$ for some $c > 0$. That is, $\mu_l$ is in the convex hull of $\Phi^l = \{\phi_i : \phi_i \in \Phi\}$. Suppose also that there is a bound on elements of the dictionary so that $\|\phi_i\|_1 \leq 1$ for all $\phi_i \in \Phi$. Then, as shown in [22], [28], for each such target $\mu$ and each number of terms $k$ there is a subset $\Phi_k \subseteq \Phi$ of size $d_n = k$ for which a $k$-term approximation $\mu^k = \sum_i \alpha_i \phi_i$ achieves $\|\mu - \mu^k\|_2 \leq \frac{C}{\sqrt{k}}$.

Here adaptation of the subset to the target $\mu$ is essential. There is no fixed $k$-dimensional space such that linear projection of $\mu$ into the space can achieve such favourable rates simultaneously for all $k$ in the convex hull class [22] (Lemma 6 and Theorem 6).

The mixture over subsets estimator $\hat{\mu}$ achieves a risk bounded by $\min_{0.5} \left( C^2 \sqrt{d} \log^{-1}(d/\delta) \right)^{1/2}$ (up to negligible terms), which when optimized over $d$ reveals a rate close to $1/\sqrt{n}$, to within a logarithmic factor, suitable for the convex hull class. For more detailed expressions of upper and lower bounds on minimax risk in convex hull classes, see Yang et al [14], [26] and Tretyakov [2]. Note that here our mixture is taken over subsets and over the dimension $d$, and so in formulating the estimator no knowledge of particular of possible regularity is presumed. We could just as well have captured the sparsity via other $\ell_0$ norm conditions $\sum_i |\beta_i| \leq c$ with $0 < c < 2$ to which the mixture will also adapt.

Other cases that require variable subsets include piecewise constant models studied in [29] as lower bounds showed that the log factor in the risk is bound necessary, and wavelet models (see [29] and the references cited therein), where natural conditions on the coefficients (via $\ell_0$ norms $\sum_i |\beta_i|_0$ on each level $l$) correspond to certain Besov spaces. The work yields advantages in risk properties of modeling mixtures in such settings.

To summarize the story for general subsets of basis vectors from a dictionary, we have in this case that the complexity $C_n$, essential to the risk bound, is larger than the dimension of the models $d_n$, metric entropy evaluation in Yang et al [26] which determine minimum risks show that this larger complexity is intrinsic for corresponding classes of target vectors. In these cases small approximation error by models of moderate dimension requires adaptation of subsets, and one achieves good approximations in optimal balance with complexity by mixing estimators over these models.

B. Structured Subsets

Next we consider models that include all terms in prescribed contexts. These models have complexity smaller than dimension and are also important in theory and applications. Among the simplest such models are those of nested leading-term type such as polynomials (of adjustable degree) and truncated Fourier series (of adjustable maximal frequency). These linear models are indexed by $m = (1, 2, \ldots, k)$ with dimension $d_m = k$ where $k \leq n$. The models can be described using either $C_n = \log d_m$ to slightly favour small models or $C_n = \log n$ which gives an equal weight factor. Here $\log d_m$ is a code-length for set to $\log d_m = \log(d_m + 1) \log \log(d_m + 1)$ for $d_m > 0$ and it is negligible in comparison to the dimension.

We do not need to restrict the list of models to be nested. For instance polynomial splines on equal spaced knots provide a sequence of models indexed by $m = (k, r)$, where $k$ is the number of knots and $r$ is the degree of the local polynomials and we may set $C_n = \log k + \log r$.

In these cases, as one increase dimension (to improve approximation error), the complexity seems to be seen of smaller dimension than the order. Now when the complexity is negligibly small compared to the dimension, the interpretation of the resolvability simplifies. It becomes simply an expression of the optimal trade-off between squared bias and variance among the linear models. This negligible complexity situation is preferred for cases in which a good approximation trade-off is achieved without taking subsets of terms.

For example, suppose $\phi_{il} \in \Phi$ are orthonormal basis functions and that the mean $\mu = \sum \phi_i \beta_i$ is in an ellipsoid (also called Sobolev) class $E_k^d$ which is the collection of points in $\mathbb{R}^d$ such that $\sum_i |\beta_i|^2 q_k^2 \leq \beta_i$, where $q_k$ is an increasing sequence. Now the leading-term model which stops at dimension $m$ provides an approximation $\mu^m = \sum_i \alpha_i \phi_i$ for which the approximation error (the squared bias of the projection estimator for model model $m$) $\|\mu - \mu^m\|_2^2 = \sum_i \alpha_i^2 + \sum_i \alpha_i^2 q_k^2$ is bounded by $\|\phi_l\|^2 \sum_{k} q_k^{-2}$ uniformly for points in $E_k^d$ with equality attainable when $q_k$ is non-zero only on the $m + 1$-term. Adding the variance term $m/n$ and minimizing over $m$ yields a risk minimum $R^* = m/n + m/n \log n$, which is known to be the minimax rate over all possible estimators for each such ellipsoid $E_k^d$ (see e.g. [29]). For example when $q_k^2 = k^2$ (as arise in characterizing Sobolev classes using Fourier series), we have rate $O(\sqrt{\log(n)/n}) + O(\log^2(n))$. Note that in the construction of the mixture there is no presumption of any particular structure sequence $\phi_l$, smoothness index $\kappa$, or size of ball $h$. The mixture across model dimensions $m$ is adaptive in that, in providing risk bounded by the risk of the best linear model, for each $m$, it will be simultaneously minimax rate-optimal for all ellipsoids $E_k^d$ for all $a$ and $h$. 
Let us discuss finite sample adaptation properties in the general setting of a sequence of linear models $m$ (not necessarily nested and not necessarily built from orthornormal terms). The closeness of the resolvability bound, with constant multiplier of 1, to the squared form and dimension term, provides an oracle inequality that exhibits already in finite samples the type of optimality previously studied in asymptotic settings. Previously, Shi, Braverman, and Li [31] and others have shown that estimators based on certain model selection criteria achieve a ratio of risk relative to the minimum of risks over all size models which converges to $1 + 2$ for fixed sequences of means $\mu$ (provided the sequence is such that $\sigma^2(\mu) \to \infty$ as $n \to \infty$, and provided that the log-cardinality of models of each dimension is $O(\log n)$ order than $d$, but that convergence is not uniform in $\mu$. A corresponding finite sample statement we can give here concerns the risk $r(\mu)$ of our combined estimator which mixes across all models in the list. For any constant $\gamma$ (or sequence $\gamma_n$) consider the minimum of risks $r^*_\gamma(\mu) = \min_r \{r(m, \mu) : \gamma_n \geq \gamma r(m, \mu) \}$ where the dimension exceeds a multiple of the complexity, and where $r(m, \mu) = \|m^2 - \mu^2\| + \frac{d_m}{n} \ln n$ is the risk for the estimator for model $m$. Then the combined estimator achieves (via the inequality $r(\mu) \leq \min_r \{r(m, \mu) + C \gamma_n/\gamma\}$ simultaneously for all $\gamma$ a ratio $r^*_\gamma(\mu) \leq \frac{1 + 4}{\gamma}$ uniformly in $\mu$. In this analysis one recovers limit properties by picking $\gamma_n$ tending to infinity at various rates so that the factor in the denominator includes models of appropriate dimension. Thus the risk of the combined estimator is never worse than the best risk among the models for which the complexity is not too small compared to the dimensions. This can be quantified not only through multiplicative constants made arbitrarily close to 1, but more precisely using multiplicative constants of the form $1 + O(1/n)$ for the complexity relative to the size in the finite sample bound $r^*_\gamma(\mu) \leq \min_r \{\|m^2 - \mu^2\| + \frac{d_m}{n} + C \gamma_n/\gamma\}$. For a similar spirit of oracle inequalities but with larger multiplicative constants, see Barrenetxea et al. [29], Birge and Massart [32], for model selection in least-squares regression, Dudoit and Johnstone [33-36] for shrinkage estimation in orthogonal basis, Birge and Massart [37], Devroye and Lugosi [38], and Yang et al. [39] for density estimation, and Wegkamp et al. [39] for $L_2$ risk regression and Juditsky et al. [27], Nemirovsky [30] and Trybakov [2] for function aggregation in regression. To summarize, whether it is better to use variable subsets or complete models of various orders in regression, depends on the nature of the unknown target $\mu$. If coefficients $\mathfrak{a}$ in complete representation are scattered throughout a dictionary of possibilities then the target requires variable subsets associated with sparse approximations (which are proven to adapt under $L_2$ norm conditions for $\epsilon < 1$), achieving good risk properties when mixing with suitable model complexity. On the other hand if the coefficients are typical of those in ellipsoid cluster with an ordered sequence of indices reflecting decreasing widths of the axes as we increase $k$, then mixtures of nested leading-term projections can achieve suitable trade-off in approximation and dimension (with negligible model complexity penalty). If one does not know in advance in which of the two settings is more appropriate for a case at hand, then they may be combined (adding only a $(\log 2/\sigma^2)$/p ratio to the complexity term). The combined estimator achieves risk corresponding to the best radii in approximation, distortion, and complexity.

ACKNOWLEDGEMENT

G. Lourie would like to thank J. Tavakoli, H. Leeb, and D. Pollard for useful comments for his PhD dissertation.

REFERENCES
