Polyshrink: An Adaptive Variable Selection Procedure That Is Competitive with Bayes Experts

Dean P. Foster and Robert A. Stine *

The University of Pennsylvania

Department of Statistics
The Wharton School of the University of Pennsylvania
Philadelphia, PA 19104-6340
E-mail:{foster,stine}@wharton.upenn.edu

AbstractWe propose an adaptive shrinkage estimator for use in regression problems charaterized by many predictors, such as wavelet estimation. Adaptive estimators perform well over a variety of circumstances, such as regression models in which few, some or many coefficients are zero. Our estimator, *PolyShrink*, adaptively varies the amount of shrinkage to suit the estimation task. Whereas hard thresholding using the risk inflation criterion is optimal for models with few predictors, *PolyShrink* obtains a broader competitive optimality vis-a-vis the best Bayes expert. A Bayes expert is the predictive distribution implied by a prior distribution for the unknown coefficients. We derive non-asymptotic upper and lower bounds for the expected log-loss, or divergence, of Bayes experts whose prior is unimodal and symmetric about zero. Our bounds hold for any sample size and are pointwise in the sense that they hold for any values of the unknown parameters. These bounds allow us to show that *PolyShrink* produces a fitted model whose divergence lies within a constant factor of the divergence obtained by the best Bayes expert. In a simulation of four frequently considered wavelet estimation problems, *PolyShrink* obtains smaller mean squared error than hard thresholding, which is not adaptive, and several other adaptive estimators.

Keywords and phrases: divergence, empirical Bayes, thresholding, relative entropy, wavelet.

1. Introduction

Consider the familiar problem of choosing the predictors in a regression model that is to be used to predict future observations. Suppose that the response vector Y holds n independent observations $Y = (Y_1, \ldots, Y_n)$. To model Y, we have access to a large collection of $p \leq n$ orthonormal predictors arranged as the columns of an $n \times p$ matrix $X = [X_1, X_2, \ldots, X_p]$ with $X^t X = I_p$. Assume that these predictors affect the response through the standard linear model,

$$Y_i = x_i^t \beta + \sigma \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} N(0,1), \quad i = 1, \ldots, n, \tag{1}$$

for arbitrary column vectors β, $x_i \in \mathbb{R}^p$, where x_i^t is the i^{th} row of X. The problem is to predict accurately independent realizations of Y for known values of the predictors.

*The authors appreciate the helpful insights provided by the associate editor and referee.

TABLE 1

Average mean squared error of reconstructions of four test functions obtained in a simulation of hard thresholding at $\sqrt{2\log p}$, SureShrink, empirical Bayes thresholding (using exponential and Cauchy priors), and the proposed adaptive estimator. The standard error of each estimate is about 0.001, and differences within a column are statistically significant.

Estimator	Blocks	Bumps	HeaviSine	Doppler
Hard($\sqrt{2\log p}$)	0.255	0.286	0.064	0.131
SureShrink	0.219	0.247	0.064	0.127
Empirical Bayes (exp)	0.181	0.208	0.053	0.106
Empirical Bayes (cau)	0.176	0.199	0.053	0.102
PolyShrink	0.172	0.193	0.052	0.100

Our interest focuses on problems in which p is roughly equal to n. This context is the standard situation in, for example, harmonic analysis using a basis of sines and cosines or wavelets. This context is also common, albeit with substantial collinearity, in data mining. In many of these applications, the coefficients are sparse, "nearly black" in the language of Johnstone and Silverman [15]. In these cases, most of the slopes $\beta_j = 0$, implying the predictor X_j does not affect the distribution of Y. Indeed, much of the motivation for wavelets owes to their ability to produce sparse representations. Though it creates no bias to include predictors for which $\beta_j = 0$, adding these increases the variability of the fit. It should be remembered, however, that specification searches, such as modern versions of stepwise regression, produce selection bias. Miller [16] provides a recent survey of these results.

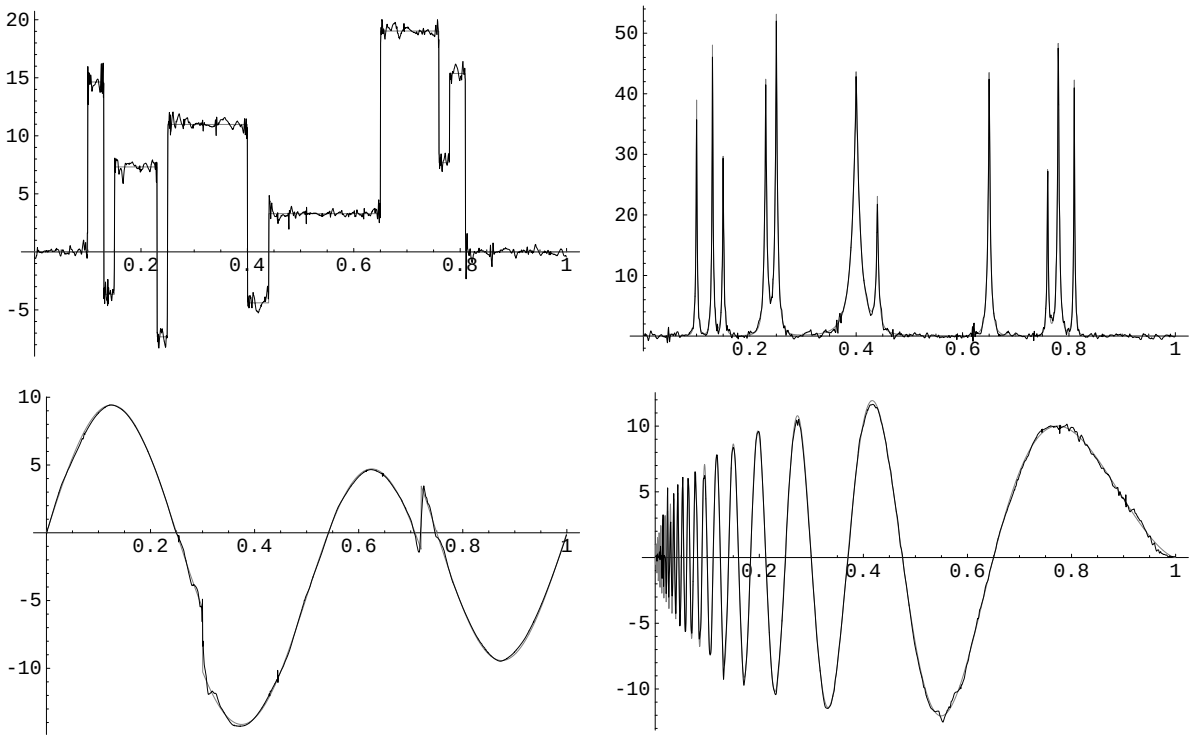
If the underlying model has *known* sparsity, then a variety of arguments motivate the following thresholding estimator. Assume that $\sigma^2 = 1$ is given and denote the least-squares estimators as $b_j = X_j^t Y \sim N(0,1)$. Our normalization of the regressors implies that the b_j are standardized. The hard thresholding estimator $\hat{\beta}^{H(\tau)}(b)$ sets to zero those b_j which are smaller in size than a threshold $\tau > 0$,

$$\hat{\beta}_j^{H(\tau)} = b_j I\{|b_j| \geq \tau\} \,,$$

where $I\{S\}$ denotes the indicator function of the set S. Oracle inequalities [9], risk inflation [11], data compression [12], and the classical Bonferroni rule all lead to the estimator $\hat{\beta}^{H(\sqrt{2\log p})}$ (or asymptotically similar soft-thresholding rules). This estimator includes in the model only those predictors for which $b_j^2 > 2\log p$, basically those for which the associated p-value is less than $1/p$. The problem with hard thresholding, of course, is that it is "too hard" when the underlying signal spreads over more than a handful of basis elements.

Suppose, for example, that it is known that the p^γ of the β_j are zero. Rather than hard thresholding at $\tau = \sqrt{2\log p}$, this knowledge might suggest setting the threshold to $\tau_\gamma = \sqrt{2\,\gamma\log p}$. Abramovich et al. [1] observe that using the incorrect threshold $\tau_{1/2}$ when $\gamma = 1/4$ results in a sixfold increase in the quadratic risk over that obtained by using $\tau_{1/4}$. To accommodate problems in which β is sparse or dense, one would prefer an estimator that adapts to the problem, for example by tuning the threshold.

FIGURE 1. *Adaptive* PolyShrink *reconstructions capture the signal in the four test functions introduced by Donoho and Johnstone [9] with small MSE in all cases. Clockwise from the upper left, the test functions are Blocks, Bumps, Doppler, and Heavisine.*



To illustrate the benefits of adaptive estimation, consider estimating the four familiar test functions introduced in Donoho and Johnstone [9]. These four functions have relatively sparse representations in a wavelet basis, but are visually distinct. One test function (called blocks) is piecewise constant whereas another (heavisine) is rather smooth but for two discontinuities. Another has a varying periodic structure (doppler), and the fourth has sharp spikes (bumps). Figure 1 shows reconstructions using $n = 2048$ equally spaced observations of these signals. For each simulated replication, we added random Gaussian noise with standard deviation $\sigma = 1$ to the underlying signal. We scaled the test functions so that the standard deviation of the underlying signal equals 7, as in Donoho and Johnstone [9]. We obtained the estimates in the figure by using the adaptive *PolyShrink* estimator that we define and study in the rest of this paper. The underlying signals appear in gray, but the fits are so accurate that they obscure the test functions.

The accuracy of the *PolyShrink* reconstructions shown in Figure 1 is not accidental. *PolyShrink* does well in general: it obtains smaller mean squared error than available alternatives for these four test signals. Table 1 summarizes the simulated MSE over 250 replications obtained by five estimators. The simplest, oldest and least accurate of these is the non-adaptive hard thresholding estimator $\hat{\beta}^{H(\sqrt{2\log n})}$ using the universal threshold $\sqrt{2\log n} \approx 3.9$. Whereas hard thresholding uses a fixed threshold, the other four estimators use the least squares estimates to gauge the best amount of thresholding to employ. In addition to *PolyShrink* we show results for two other adaptive estimators:

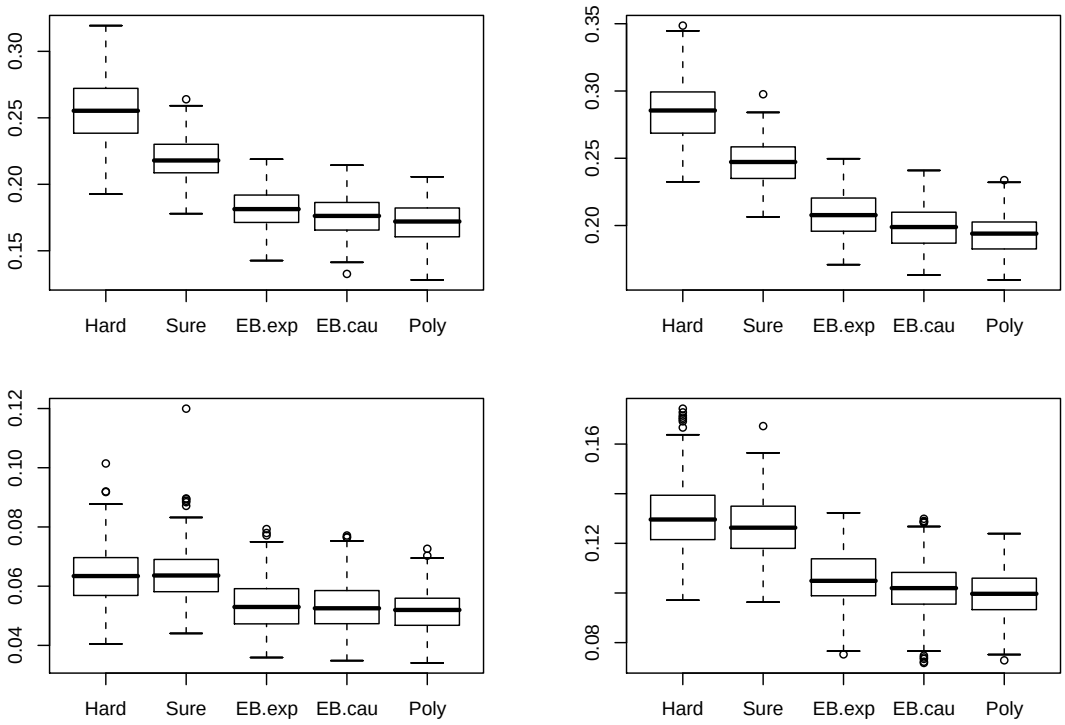
SureShrink, which uses the soft-threshold that minimizes Stein's unbiased estimate of risk [SURE, 8];

EbayesThresh, an empirical Bayes estimator that uses either a double exponential or Cauchy prior [15];

The boxplots in Figure 2 compare the MSE obtained by the five estimators over the 250 simulated samples. Though the boxplots overlap, the differences between mean values for each test function are rather significant. For example when estimating blocks, even though the MSE of *PolyShrink* is only 0.004 less than that obtained by the empirical Bayes estimator using a Cauchy prior, this difference produces a t-statistic larger than 15 because of the dependence between the estimates. (Each estimator was applied to the same 250 realizations.) Of the five estimators, only hard thresholding, the worst performing method, sets a common threshold over all resolutions of the wavelet transform. The benefits of adaptation are clear. Unless the signal is very sparse, an estimator that adapts its threshold to the signal at hand has smaller squared error.

Remark. For these computations, we used version 2.2.1 of **R** running on a macintosh computer. We used the **R** packages `waveslim` to obtain the wavelet decompositions and `ebayesthresh` to compute the empirical Bayes estimator. We chose the default values for parameters used by `ebayesthresh`.

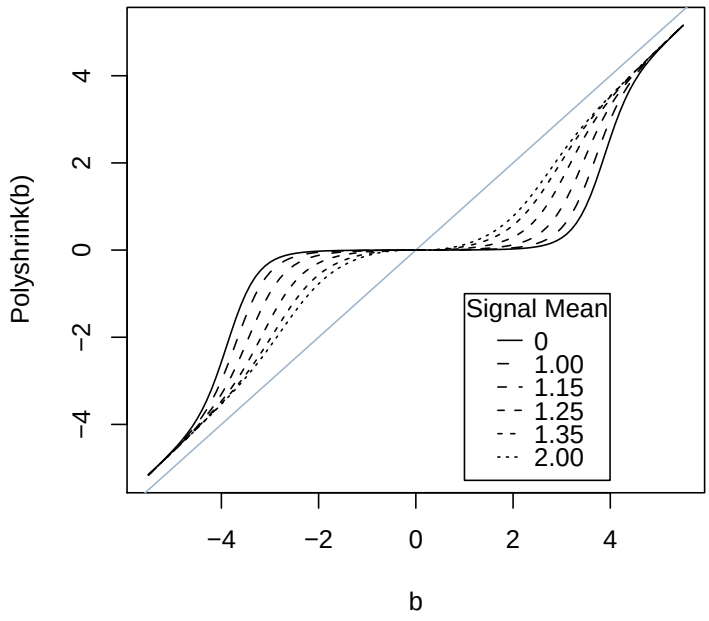
FIGURE 2. *These boxplots compare the mean squared error of five estimators of the four test functions (clockwise, from upper left, are results for estimating blocks, bumps, doppler, and heavisine).*



For filtering these wavelet coefficients (and comparison to other published simulations), we applied thresholding only to the 6 levels of the LA8 wavelet coefficients (2016 coefficients with most localized support). In practice, it may be quite hard to decide how many levels of the wavelet transform to filter. The *Polyshrink* wavelet estimator has no tuning parameters, and we applied it to all but the two, top-level wavelet coefficients. To avoid issues of various estimates of scale, for every estimator we made use of the fact that $\sigma = 1$. An **R** package of our software is available at www-stat.wharton.upenn.edu/∼stine and through the CRAN archive (cran.r-project.org).

The *PolyShrink* estimator $\tilde{\beta}(b)$ adaptively thresholds its estimate of each coordinate. To convey the nature of the adaption, Figure 3 shows $\tilde{\beta}_1(b)$, the estimator of β_1, as a function of the first element in the least squares estimator b. For this figure, we set the dimension $p = 65$. The various curves in Figure 3 characterize how $\tilde{\beta}_1(b)$ changes with the size of the *other* coordinates $b_2, \ldots, b_{65}$. *PolyShrink* uses these other coordinates to adjust the amount of shrinkage when estimating β_1 from b_1. To suggest the effect of signal in the other coordinates, we set $b_{j+1} = m + \Phi^{-1}(j/65), j = 1, \ldots, 64$, and varied $m = 0, 1, 1.15, 1.25, 1.35, 2.0$, as indicated in the legend of the figure. At one extreme (the solid curve in the figure), there is no signal in the other coordinates; their values are as though one has observed an idealized sample of pure Gaus-

FIGURE 3. *The adaptive* Polyshrink *estimator resembles hard thresholding with a large threshold when there is little signal and resembles soft thresholding with a small threshold when there is substantial signal. The curves show one coordinate* $\tilde{\beta}_1(b)$ *for varying levels of signal spread over* $p = 65$ *coefficients.*



sian noise. Such sparsity is commonly encountered when estimating the coefficients of wavelet basis elements having most localized support. Heuristically, the absence of signal in the other coefficients makes $\tilde{\beta}_1(b)$ "suspicious" of any signal found in b_1, and so $\tilde{\beta}_1(b)$ heavily shrinks b_1 toward zero. Indeed, the plot of $\tilde{\beta}_1(b)$ resembles a smoothed version of the hard thresholding estimator with a threshold slightly more than 2. As m increases, the other coordinates indicate the presence of signal, and $\tilde{\beta}_1(b)$ offers less and less shrinkage. Once $m = 2$, $\tilde{\beta}_1(b)$ resembles a soft thresholding estimator. Though not apparent in Figure 3, however, $\tilde{\beta}_1(b)$ returns to the diagonal as $|b|$ increases. Asymptotically in b_1, the deviation is of order $O(1/b_1)$:

$$b_1 - \tilde{\beta}(b_1) \approx \frac{2b_1}{1 + b_1^2} \ .$$

The expressions that we use to compute $\tilde{\beta}$ offer further insight into how the *Polyshrink* estimator adapts. Although our theorems do not explicitly provide an estimator, an appendix describes the conversion from the existential style of the theory to a more standard estimator. Assume as in the orthogonal regression (1) that we observe observe a p-element vector of estimates b with mean vector β and variance 1, $b \sim N(\beta, I_p)$. For wavelets, b denotes the standardized "raw" wavelet transform at a given resolution, and β denotes the corresponding wavelet coefficients of the signal. The *PolyShrink* estimator uses a collection of mixtures of the form

$$g_{\epsilon_k}(z) = (1 - \epsilon_k)\phi(z) + \epsilon_k \psi(z) \ , \quad \epsilon_k = 2^{k-(K+1)}, \quad k = 1, \ldots, K, \tag{2}$$

where $K = 1 + \lfloor \log_2 p \rfloor$, $\phi(z) = \exp(-z^2/2)/\sqrt{2\pi}$ is the standard normal density and $\psi(z) = 1/(\sqrt{2}\pi(1 + z^2/2))$ is a Cauchy density with scale $\sqrt{2}$. The notation $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x. The mixtures g_{ϵ_k} place larger weight on the Cauchy density as ϵ_k increases. We extend g_ϵ to vector arguments by multiplication,

$$g_\epsilon(b) = g_\epsilon(b_1) g_\epsilon(b_2) \cdots g_\epsilon(b_p) \ .$$

With this notation in hand, the *PolyShrink* estimator is readily defined and computed. (An implementation in **R** requires about 20 lines of straightforward code.) Let $w_k = 2^{-k} g_{\epsilon_k}(b)$ and normalize these weights

$$\tilde{w}_k = w_k / \sum_{i=1}^{K} w_i \ , \quad \tilde{w} = (\tilde{w}_1, \ldots, \tilde{w}_K)^t \ .$$

The normalized weights $\tilde{w}$ allow $\tilde{\beta}_j$ to *borrow strength* from other coordinates when choosing an appropriate amount of shrinkage. The *PolyShrink* estimator can then be written in the classical style of Stein [18],

$$\tilde{\beta} = b + S\,\tilde{w} \ , \tag{3}$$

where the $p \times K$ matrix S holds the scores

$$S_{jk} = \frac{d}{dz} \log g_{\epsilon_k}(z)|_{b_j} \ . \tag{4}$$

Thus $\tilde{\beta}$ blends the score functions associated with mixtures g_ϵ. When there is little signal (illustrated by the solid curve in Figure 3), the weights $\tilde{w}$ concentrate in $\tilde{w}_1$, placing the most emphasis on the score function with the largest Gaussian component. In the calculations producing Figure 3, $\tilde{w}_1 = 0.62$ when $m = 0$. With signal present, the weights shift toward mixtures with a larger Cauchy component. For example, with $m = 1$ this weight decreases to $\tilde{w}_1 = 0.46$. With $m = 2$, the estimator puts virtually all of its weight on the score function of $g_{1/2}$, setting $\tilde{w}_1 \approx 0$ and $\tilde{w}_7 = 0.999$.

The remainder of this paper develops the theoretical properties of the *PolyShrink* estimator. The following section presents a more formal introduction that describes the origin of our estimator and its properties. In particular, we explain our choice of a different type of loss function that allows us to prove that the associated risk of $\tilde{\beta}$ lies within a constant factor of the smallest risk obtained by a class of Bayesian estimators that we call *Bayes experts*. Following in Section 3, we offer some additional motivation by pointing out further connections to other estimators. In Section 4, we derive lower bounds for the risk of Bayes experts, and in Section 5, we describe the adaptive estimator. We prove a key result, Theorem 2, in Section 6, and relegate further details of the proofs to the second appendix. The first appendix describes the conversion from our theorems to the *PolyShrink* estimator (3).

2. Competitive Analysis of Estimators

Consider the problem of predicting the next value of the response, Y_{n+1}. The *PolyShrink* estimator described in the introduction arises from the following competitive analysis. The goal is to find an estimator that predicts as well the best *expert* who builds a prediction using side information that would not be known to the statistician. This approach produces an adaptive estimator because the estimator must do well as well as the best expert, regardless of the unknown parameters. For example, one might seek a predictor $\widehat{Y}_{n+1} = x_{n+1}\widehat{\beta}$ whose expected squared prediction error $E_1^{n+1}(Y_{n+1} - \widehat{Y}_{n+1})^2$ approaches that obtained by an expert who knows more than the data reveal about the form of the model. (The expectation E_1^{n+1} is over the joint distribution of the observable response $Y_1, \ldots, Y_n$ and the future value Y_{n+1}.) The extra information available to the expert might come from an oracle that identifies the predictors that have non-zero coefficients as in Donoho and Johnstone [9] or, less informatively, an oracle that reveals the number of non-zero coefficients.

For our analysis, the scope of the extra information lies somewhere between these alternatives: an oracle provides a prior distribution $\pi(\beta)$ for the regression coefficients. Given π, a Bayesian can express uncertainty about the future value Y_{n+1} with the predictive distribution that encompasses the posterior distribution of β. Denote the scalar density of the response associated with predictors in the row vector x as

$$P_\beta(y) = \phi\left((y - x^t\beta)/\sigma\right) \ .$$

To simplify the notation, we omit x from the arguments of this function and leave this dependence

implicit. We denote the predictive distribution of Y_{n+1} given Y by $P_\pi^{n+1}(y \mid Y)$. This predictive distribution is a *Bayes expert* in our terminology. In general, the predictive distribution for Y_{i+1} given $Y_1, \ldots, Y_i$ is

$$P_\pi^{i+1}(y \mid Y_1, \ldots, Y_i) = \int P_\beta(y)\pi(\beta|Y_1, \ldots, Y_i)\, d\beta\ , \quad i = 2, \ldots, n, \tag{5}$$

where

$$\pi(\beta \mid Y_1, \ldots, Y_i)$$

denotes the posterior distribution of β given $Y_1, \ldots, Y_i$. For the initial case that lacks prior data, we define P_π^1 as

$$P_\pi^1(y) = \int P_\beta(y)\pi(\beta)\, d\beta\ .$$

It may seem that Bayes experts are little more than a collection of prior distributions, and so one might question the need to introduce this terminology. While our notion of a Bayes expert does employ a prior, our use of the word "expert" is meant to convey something more. An *expert* is *any* prediction methodology, a map from the data to the space of new observations, that benefits from information outside the scope of what would typically be available for estimation. We should also point out that the objectives of our analysis differ from those that one would typically associate with Bayesian analysis of a collection of priors. When confronted by a collection of priors, a Bayesian might seek a method to elicit the best prior or to find a robust prior. We have a different objective. Rather than seek the best Bayes estimator, we use Bayes estimators as a yardstick with which we can judge the performance of adaptive variable selection.

We structure our competitive analysis around Bayes experts whose prior is unimodal and symmetric. We denote the class of priors by $\mathcal{M}$. More precisely, the scalar distribution $\pi \in \mathcal{M}$ iff (1) π is symmetric and (2) π assigns decreasing probability to intervals as the center of the interval moves away from zero. Thus, if $\pi \in \mathcal{M}$ and

$$p_w(c) = \int_{c-w}^{c+w} d\pi\ , \quad w > 0, \tag{6}$$

then $p_w(c) = p_w(-c)$ and $p_w(c)$ is monotone decreasing on $[0, \infty]$. In essence, the class $\mathcal{M}$ consists of symmetric, unimodal densities that may possess an atom of probability at the origin. The assumed orthogonality of the estimation problem allows us to extend π to vectors by multiplication, $\pi(\beta) = \pi(\beta_1) \cdots \pi(\beta_p)$. Thus $\pi(0)$ reveals the expected number of elements of β that are zero. The rest of π indicates how the signal distributes over the other coordinates. The choice of experts with $\pi \in \mathcal{M}$ avoids artificial problems that would be caused by competing against super-efficient estimators. For example, a realizable estimator cannot compete with an expert using a point-mass prior $\delta_3(x)$ if in fact $\beta_j = 3$. The only point mass prior in $\mathcal{M}$ is δ_0.

For a Bayes expert, a natural measure of accuracy is the expected log-probability loss or divergence (relative entropy). We use $D(f \parallel g)$ to denote the divergence between the true density f and

some other density g,

$$D(f \parallel g) = E_f \ \log \frac{f(X)}{g(X)} = \int f(x) \log \frac{f(x)}{g(x)} dx \ . \tag{7}$$

The divergence is well-defined if the support of g contains the support of f, and the assumed normality of the errors in our model produces density functions with infinite support. The divergence of a Bayes expert, the predictive distribution P_π^{n+1}, from the distribution of Y_{n+1} is (suppressing the dependence on x_{n+1})

$$\begin{aligned} L(\beta, P_\pi^{n+1}) &= D(P_\beta \parallel P_\pi^{n+1}) \\ &= \int \log \frac{P_\beta(y)}{P_\pi^{n+1}(y \mid Y)} P_\beta(y) \, dy \ . \end{aligned} \tag{8}$$

Divergence has a long history in model selection. For example, *AIC* arose as an unbiased estimate of the divergence [2]. The divergence can also resemble more familiar loss functions. For example, if the posterior distribution concentrates near a point mass at, say, $\hat{\beta}$ then the divergence is roughly quadratic,

$$L\left(\beta, P_\pi^{n+1}(y|Y)\right) \approx \frac{\left(x_{n+1}(\hat{\beta}(Y) - \beta)\right)^2}{2\sigma^2} \ .$$

We depart from this approach to obtain our results on the competitive accuracy of Bayes experts. The loss function defined by (8) is a marginal loss in the sense that it describes the error when predicting Y_{n+1} given x_{n+1} and information in the prior n observations. We obtain a more analytically tractable quantity by replacing the marginal loss with the accumulated loss. Rather than judge the expert by how well P_π^{n+1} matches P_β, we instead sum the deviations of P_π^i from P_β for $i = 1, \ldots, n$. Although one prefers an estimator that predicts well in the future over one that fits well in the past ("Prediction is difficult, particularly about the future."), we have found it easier to prove theorems about the accumulated loss than the marginal loss. This change also removes the dependence on the unknown x_{n+1}, replacing it with the observed values of the predictors. The product of these predictive distributions defines the prequential likelihood [5, 6]

$$P_\pi(Y_1, \ldots, Y_n) = P_\pi^1(Y_1) \prod_{i=2}^{n} P_\pi^i(Y_i \mid Y_1, \ldots, Y_{i-1}) \ .$$

The *cumulative log-loss* of a Bayes expert with prior π is then

$$L_n(\beta, P_\pi) = \sum_{i=1}^{n} \log \frac{P_\beta(Y_i)}{P_\pi^i(y \mid Y_1, \ldots, Y_{i-1})} \ . \tag{9}$$

The expected value of L_n is the divergence of the prequential likelihood for Y. Thus, we define the *divergence risk* obtained over the sequence of n predictions as

$$\begin{aligned} R_n(\beta, P_\pi) &\equiv E_1^n \left(L_n(\beta, P_\pi)\right) \\ &= D(P_\beta(Y) \parallel P_\pi(Y)) \ . \end{aligned} \tag{10}$$

The success of the *PolyShrink* estimator summarized in the introduction suggests that the shift from marginal to accumulated loss nonetheless produces a useful estimator.

We can now summarize our results for the *PolyShrink* estimator. Our first theorem provides non-asymptotic lower bounds for divergence risk of Bayes experts whose prior distribution $\pi \in \mathcal{M}$, the class of symmetric, unimodal priors. Theorem 1 also shows that, given a tuning constant $\epsilon(\pi)$ that depends on the prior, there exists an an estimator $g_{\epsilon(\pi)}(\pi)$ whose divergence is within a constant factor of that obtained by the best Bayes expert.

Theorem 1. *For $\epsilon > 0$, define the bounding function*

$$B(\beta, \epsilon) = \sum_{j=1}^{p} \min\left(\beta_j^2 + \epsilon, \frac{1}{\beta_j^2} + \log\frac{\beta_j}{\epsilon}\right) . \tag{11}$$

There exists a real-valued functional $\epsilon(\pi)$ such that, under the model (1) with $p \le n$ orthonormal predictors, the divergence risk (10) of any Bayes expert in the class $\mathcal{M}$ (see equation (6)) is bounded below as follows:

$$(\forall n, \forall\beta, \forall\pi \in \mathcal{M}) \qquad R_n(\beta, P_\pi) \ge c\, B(\beta, \epsilon(\pi)) , \tag{12}$$

for some positive constant $c < 1$. Further, there exists an estimator $g_{\epsilon(\pi)}$ such that

$$(\forall n, \forall\beta, \forall\pi \in \mathcal{M}) \quad R_n(\beta, g_{\epsilon(\pi)}) \le 2\, B(\beta, \epsilon(\pi)) \tag{13}$$

We have numerically estimated the constant $c \approx 1/10$, but it may be possible to close the gap between (12) and (13) by using methods developed in Foster et al. [13].

The bounds in Theorem 1 reduce the problem of choosing a prior $\pi \in \mathcal{M}$ to the choice of a number, ϵ. In fact, even the problem of picking ϵ can be eliminated as the following theorem shows. Our second theorem describes an estimator $\tilde{g}$ which in turn defines the *PolyShrink* estimator $\tilde{\beta}$ defined in (3) and described in the introduction. This estimator avoids the need for the functional $\epsilon(\pi)$ in Theorem 1 by mixing g_ϵ over several values of ϵ. The weights in this mixture do not depend on the prior π.

Theorem 2. *There exists a distribution $\tilde{g}$ (namely that implying the* Polyshrink *estimator) whose divergence risk is a linear function of the best obtained by any Bayes expert whose prior $\pi \in \mathcal{M}$ (see equation (6)). In particular,*

$$(\forall n, \forall\beta) \qquad R_n(\beta, \tilde{g}) \le \omega_0 + \omega_1 \inf_{\pi\in\mathcal{M}} R_n(\beta, P_\pi) . \tag{14}$$

3. Connections

Thresholding estimators possess a form of competitive optimality in the style of our theorems. Let $Q_n(\beta, \hat{\beta}) = E_1^n \sum_j (\beta_j - \hat{\beta}_j(Y))^2$ denote the quadratic risk of an estimator $\hat{\beta}$. Donoho and Johnstone [9] and Foster and George [11] show that the quadratic risk of the hard thresholding

estimator $\hat{\beta}^{H(\tau)}$ using the universal threshold $\tau^2 = 2\log p$ is bounded by a logarithmic factor of the best possible. Abbreviating $\hat{\beta}^H = \hat{\beta}^{H(\sqrt{2\log p})}$ then

$$Q_n(\beta, \hat{\beta}^H) \leq (2\log p + 1) \inf_{\hat{\beta}\in\mathcal{S}} Q_n(\beta, \hat{\beta}) \;, \tag{15}$$

where $\mathcal{S}$ denotes the class of least-squares estimators based on selecting a subset of the predictors. Donoho and Johnstone [9] introduce the notion of an oracle to describe the optimality of $\hat{\beta}^H$. Suppose that a statistician can consult an oracle that indicates which elements of β to estimate. Then the bound (15) shows that the quadratic risk of $\hat{\beta}^H$ lies within a factor of $2\log p$ of that obtained by consulting the best of all such coordinate oracles. Furthermore, this competitive performance is the best possible vis-a-vis coordinate oracles. If we label the class of estimators based on these oracles $\mathcal{C}$, the class of *coordinate experts*, then we can state this result as:

$$(\forall n, \forall\beta) \qquad Q_n(\beta, \hat{\beta}^H) \leq \omega_0^{Q(\mathcal{C})} + \omega_1^{Q(\mathcal{C})} \inf_{\hat{\beta}\in\mathcal{C}} Q_n(\beta, \hat{\beta}) \;.$$

where $\omega_1^{Q(\mathcal{C})} = (1 + 2\log p)$. No variable selection procedure can do better than come within a factor of $\log p$ of the quadratic risk attainable by coordinate experts.

We can contrast this with our result using the divergence risk and Bayes experts. A coordinate expert knows which predictors to use in the regression and does not have to rely on data to decide which coordinates to estimate. In contrast, Bayes experts are not so well-informed; each has only a prior distribution for β from the class $\mathcal{M}$. While useful, the prior is not so informative as knowing which β_j to estimate. In a sense, competing with Bayes experts resembles the comparison of two estimators rather than the comparison of an estimator to the truth. Consequently, we can find a data-driven procedure which is more competitive with Bayes experts than with coordinate experts:

$$(\forall n, \forall\beta) \qquad R_n(\beta, \tilde{g}) \leq \omega_0 + \omega_1 \inf_{\pi\in\mathcal{M}} R_n(\beta, P_\pi) \;,$$

where we obtain an $O(1)$ for ω_1 bound instead of the $O(\log p)$ bound for $\omega_1^{Q(\mathcal{C})}$.

This comparison suggests the use of a different "ruler" to judge and compare estimators. No data-driven estimator can come close to the performance of coordinate experts, so the alternatives look equally good. Because we can come within a constant factor of the predictive risk of the Bayes experts, however, we can hopefully use these as a more finely graduated ruler to separate better estimators from others. Since Bayes estimators are admissible, a trivial lower bound for ratio of risks is 1, and so our estimator lies within a constant of being optimal. Without the $\log p$ factor inherent in the coordinate-based approach, we can begin the process of finding estimators that reduce this constant.

An alternative approach to finding adaptive estimators considers their minimax properties. Our theorems are point-wise in the sense we bound the divergence for *each* β with the divergence of the best expert for that β. One obtains a uniform bound for *every* β by finding the minimax risk

over some class of estimators and parameter spaces. Rather than do well for every β, a minimax estimator need only be competitive for the hardest problems. Because the unconstrained minimax quadratic risk for the regression model (1) is p, obtained by fitting every predictor, a non-trivial minimax analysis must constrain the parameter space. Abramovich et al. [1] argue that a natural constraint is to consider sparse problems in which few $\beta_j \neq 0$ and restrict β to an ℓ_d ball for d near zero. They show that this constraint leads to penalized estimators that do well over a range of sparse problems. For example, let

$$\Theta_n^0 = \{\beta : \|\beta\|_0 = n^{-\delta}\}, \quad \text{where} \quad \|\beta\|_0 = \sum_j I\{\beta_j \neq 0\}, \quad \delta > 0,$$

denote a nearly black parameter space in which β has few non-zero elements. The ℓ_0 norm $\|x\|_0$ counts the number of non-zero values in x. Let

$$\mathcal{R}(\Theta_n^0) = \inf_{\hat{\beta}} \sup_{\beta \in \Theta_n^0} E\|\beta - \hat{\beta}\|_0$$

denote the corresponding minimax risk over all estimators $\hat{\beta}$. Abramovich et al. [1] show that an estimator $\hat{\beta}^F$ derived from the false discovery rate of Benjamini and Hochberg [3] is asymptotically minimax:

$$\sup_{\beta \in \Theta_n^0} E\|\beta - \hat{\beta}^F\|_0 \leq \mathcal{R}(\Theta_n^0)(1 + o(1)),$$

as $n \to \infty$ for a suitable choice of the tuning parameter that controls the estimator. Whereas Theorem 2 shows that the divergence risk of $\tilde{\beta}$ is competitive *for every* β for all n, these results show that the ℓ_0 risk of $\hat{\beta}^F$ is asymptotically minimax.

Johnstone and Silverman [15] also consider the estimation problem in nearly black parameter spaces. They show that an empirical Bayes estimator does well in both sparse problems as well as when β is more dense. To this end, they consider empirical Bayes estimators with mixture priors of the form $h(x) = (1 - w)\delta_0(x) + w\,\gamma(x)$ where $\delta_0(x)$ denotes a point mass at zero and $\gamma(x)$ is a symmetric, unimodal distribution, implying that the prior $h(x) \in \mathcal{M}$. [14, consider empirical Bayes model selection in a more parametric setting.] As an estimator, they use the median of the posterior distribution rather than a predictive distribution. The resulting estimator is thus very similar to the *PolyShrink* estimator. For example, plots of their estimator [see Figure 5 of 15] resemble those of the *PolyShrink* estimator in Figure 3. Also, the mean squared errors in the simulation reported in the introduction are very similar. In the style of Abramovich et al. [1], they show that their estimator has uniformly bounded risk over a wide range of parameter spaces and asymptotically obtains the minimax risk. With an emphasis on the problem of estimating wavelets, Zhang [19] extends and refines these results to a wider class of prior distributions and obtains a more precise, though complex, collection of asymptotic properties.

4. Lower Bounds for the Bayes Divergence

The results in this section and those that follow hold for the normal location model. These apply to variable selection in regression since we can use an orthogonal rotation to convert the the slopes of the orthonormal regression (1) into a vector of means. Let $X = [X_1, X_2, \ldots, X_p]$ denote the $n \times p$ design matrix of the predictors. If we pre-multiply both sides of (1) by the transpose X^t, we obtain

$$\tilde{Y} = X^t Y \sim N(\beta, \sigma^2 I_n) \; . \tag{16}$$

The question of which predictors to include then becomes one of deciding how to estimate the mean vector β. Since σ^2 is assumed known, we abuse our notation and further reduce the problem to the following canonical form:

$$Y_j \sim N(\mu_j,\ 1) \; , \quad j = 1, \ldots, p \; , \text{ independently.} \tag{17}$$

This reduction of a regression to a location problem is common in the analysis of variable selection methods [e.g. 10]. The independence among the Y_j allows us to work in the scalar setting.

We adopt the following notation. Write the normal density with mean μ by $\phi_\mu(y) = e^{-(y-\mu)^2/2}/\sqrt{2\pi}$ and abbreviate the standard normal density as $\phi_0 = \phi$. The cumulative normal is denoted $\Phi(x) = \int_{-\infty}^{x} \phi(t)dt$. With these conventions, the distribution of Y_i is

$$\phi_{\mu_i}(y) = \phi(y - \mu_i) \; .$$

Let $H(\phi)$ denote the entropy of the standard normal density,

$$H(\phi) = -\int \phi(y) \log \phi(y) dy = \tfrac{1}{2}(1 + \log 2\pi) \; . \tag{18}$$

The marginal distribution for Y_i implied by the prior $\pi(\mu)$ is

$$\phi_\pi(y) = \int \phi(y - \mu)\pi(\mu)d\mu \; . \tag{19}$$

The one-dimensional divergence, viewed as a function of a scalar parameter μ, is

$$d_\pi(\mu) \equiv D(\phi_\mu \parallel \phi_\pi) = \int \phi_\mu(x) \log \frac{\phi_\mu(x)}{\phi_\pi(x)} dx \; . \tag{20}$$

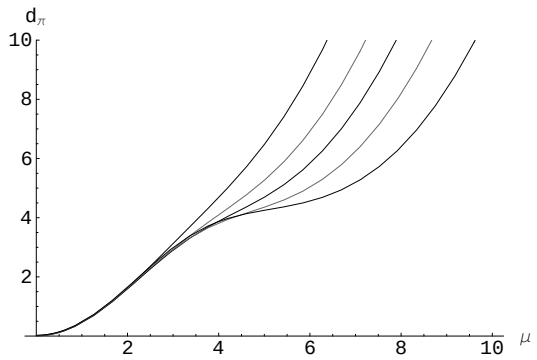
If we associate $\beta_j = \mu_j$ and note that the prior factors as $\pi(\beta) = \pi(\beta_1) \cdots \pi(\beta_p)$, then the divergence risk defined in (10) may be written as

$$R_n(\beta, P_\pi) = \sum_{j=1}^{p} d_\pi(\mu_j) \; . \tag{21}$$

The following related examples suggest the properties of the divergence attainable by Bayes experts. For $\mu \approx 0$, an expert with prior $\pi(x) = \delta_0(x)$ (the indicator function at 0) performs well. Its divergence is quadratic in μ,

$$d_{\delta_0}(\mu) \;\; = \;\; D(\phi_\mu \parallel \phi_0)$$

FIGURE 4. *Divergence of five spike-and-slab priors (23) with slab widths m that minimize the divergence d_π at $\mu = 1, 2, 3, 4, 5$ for fixed slab probability $p = 0.05$.*



$$\begin{aligned} &= \int \phi(y-\mu) \log \frac{\phi(y-\mu)}{\phi(y)} dy \\ &= E\, Y^2/2 + \tfrac{1}{2} \log 2\pi - H(\phi) \\ &= \mu^2/2 \; . \end{aligned}$$

For a coordinate with non-zero parameter, the expert will achieve a lower divergence by placing more probability away from zero. Let

$$U_m(x) = \mathbf{1}_{[-m,m]}/2m \tag{22}$$

denote the uniform density on $[-m, m]$. The divergence for the expert that uses a uniform prior out to $|\mu| + 1$ is

$$\begin{aligned} d_{U_{|\mu|+1}}(\mu) &= D(\phi_\mu \parallel \phi_{U_{|\mu|+1}}) \\ &\geq \log 2(\mu+1) - H(\phi) \; . \end{aligned}$$

Thus, the divergence for these experts is logarithmic. As a compromise, a spike-and-slab expert (prior) combines these two extremes,

$$\pi_{p,m}(\mu) = (1-p)\delta_0(\mu) + p\, U_m(\mu) \; . \tag{23}$$

Figure 4 shows the divergence obtained with a sequence of five spike-and-slab priors whose slab widths m minimize the divergence at $\mu = 1, 2, 3, 4, 5$ for fixed slab probability $p = 0.05$. The lower bounds for the divergence near zero grow quadratically, and then flatten to grow logarithmically as μ increases.

Our lower bounds for $d_\pi(\mu)$ require a single functional of the prior. This functional measures the ratio of how much probability ϕ_π puts into a tail relative to the standard normal. To define this functional, let $r_\pi(\tau)$ denote the ratio of tail integrals

$$r_\pi(\tau) = \frac{\int_\tau^\infty \phi_\pi(y) dy}{\int_\tau^\infty \phi_0(y) dy} \; . \tag{24}$$

$r_\pi(\tau)$ is monotone increasing for $\tau > 0$ because of the monotonicity of the likelihood ratio shown in Lemma 5. Hence, for an arbitrary constant $\kappa > 1$, we can define the threshold

$$\tau_\kappa(\pi) = \inf_\tau \{\tau : r_\pi(\tau) > \kappa\} \; . \tag{25}$$

Given $\tau_\kappa(\pi)$, we define the needed functional to be the associated tail probability of ϕ_π:

$$\epsilon_\kappa(\pi) = \int_{\tau_\kappa}^{\infty} \phi_\pi(y) dy \; . \tag{26}$$

Where it does not lead to ambiguity (since there is typically only one prior π under consideration), we abbreviate $\epsilon_\kappa(\pi)$ as ϵ_κ. We shall also need the related tail integral of ϕ_0:

$$\delta_\kappa(\pi) = \int_{\tau_\kappa}^{\infty} \phi_0(y) dy = \epsilon_\kappa(\pi)/\kappa \; . \tag{27}$$

The threshold $\tau_\kappa(\pi)$ and tail area $\epsilon_\kappa(\pi)$ are equivalent in the sense that we can find one from the other, without further knowledge of π, through the relationship

$$\tau_\kappa(\pi) = -\Phi^{-1}\left(\frac{\epsilon_\kappa(\pi)}{\kappa}\right) \approx \sqrt{2 \log \frac{\kappa}{\epsilon_\kappa(\pi)}} \; . \tag{28}$$

This approximation follows from the crude asymptotic equivalence $\Phi(-x) \approx e^{-x^2/2}$. Lemma 6 provides bounds on the approximation in (28) that are accurate to within 1 of the actual value. Since $r_\pi(0) = 1$, we require $\kappa > 1$ so that $\tau_\kappa > 0$ and $\epsilon_\kappa < 1/2$.

Results in the Appendix show that the divergence of any Bayes model with prior $\pi \in \mathcal{M}$ grows quadratically near the origin and eventually grows at a logarithmic rate. From the previous examples, we see that the divergence cannot be better than quadratic (as with a point mass at zero) or logarithmic (obtained by the slab). Figure 5 shows the lower bounds for $d_\pi(\mu)$ from Lemma 15 with the tuning parameter $\kappa = e^2$ and $\epsilon_\kappa = 0.001$. For reference, the figure denotes the location of the threshold τ_κ with a short vertical line near $\mu = 3.6$. The bounds shown in Figure 5 are, however, rather hard to work with, and so we use a simpler set of bounds for our proofs. In Lemma 16, we show that

$$\exists \lambda > 0 \text{ s.t. } \forall \pi \in \mathcal{M} : \quad \lambda \, d_\pi(\mu) \geq \min\left(\mu^2 + \epsilon_\kappa(\pi), \; \frac{1}{\mu^2} + \log \frac{\mu}{\epsilon_\kappa(\pi)}\right) \tag{29}$$

Though easier to work with, these simpler bounds retain the shape of the tighter bounds shown in Figure 5. Namely, the bounds are quadratic for μ near 0 and grow logarithmically as μ moves away from 0. Equation (29) implies the first statement of Theorem 1.

5. The Mixture Marginal

One estimator that achieves the upper bound given in Theorem 1 is a mixture of a normal density with a Cauchy density. Let

$$\psi(y) = \frac{1}{\sqrt{2}\,\pi(1 + y^2/2)} \tag{30}$$

FIGURE 5. *Lower bounds from Lemma 15 are approximately quadratic in μ near the origin and grow logarithmically in μ for larger values of the parameter. The figure includes (in gray) the segments from the components of the lower bound, with $\kappa = e^2$ and $\epsilon_\kappa = 0.001$. The vertical line near $\mu = 3.6$ locates the position of τ_κ.*

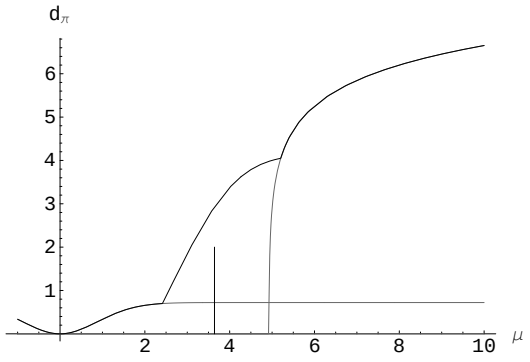
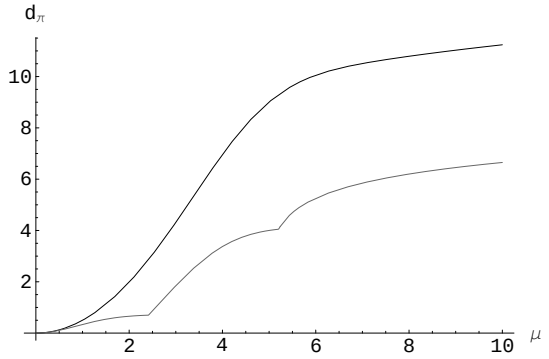


FIGURE 6. *Divergence of the Cauchy mixture $g_{0.001}$ and the lower bound for the divergence $d_\pi(\mu)$ attainable by a Bayes prior with $\epsilon_\kappa(\pi) = 0.001$.*



denote the Cauchy density with scale $\sqrt{2}$. This choice of scaling implies that the equivalent estimator $\hat{\beta}(b)$ shown in Figure 3 is monotone increasing. Define a marginal distribution as a one-parameter mixture of ϕ and ψ,

$$g_\epsilon(y) = (1-\epsilon)\,\phi(y) + \epsilon\,\psi(y)\ , \quad 0 \le \epsilon \le 1. \tag{31}$$

For $\epsilon < \frac{1}{2}$, Lemma 17 shows that the divergence of g_ϵ is bounded:

$$D(\phi_\mu \parallel g_\epsilon) \le 2\min\{\mu^2 + \epsilon, 1/\mu^2 + \log \mu/\epsilon\}\ . \tag{32}$$

Equation (32) implies statement 2 of Theorem 1.

Figure 6 shows the upper bound for the divergence of g_ϵ with $\epsilon = 0.001$. This plot includes the lower bound for the divergence attainable by a Bayes expert with prior $\pi \in \mathcal{M}$ as shown in Figure 5. The visual impression from the figure is that both bounds increase at comparable rates, as demonstrated in the proof of Theorem 2 that follows in the next section.

6. Proof of Theorem 2

This section provides a proof of Theorem 2, which we state here with more details than when it was originally stated:

Theorem 2. *Let $K = 1 + \lfloor \log_2 p \rfloor$. The sub-density*

$$\begin{aligned} g(\vec{y}) &= \tfrac{1}{2} g_{(2^{-K})}(\vec{y}) + \tfrac{1}{4} g_{(2^{-K+1})}(\vec{y}) + \cdots + 2^{-K+1} g_{(1/4)}(\vec{y}) + 2^{-K} g_{(1/2)}(\vec{y}) \\ &= \sum_{k=1}^{K} 2^{-k} g_{(2^{-(K+1-k)})}(\vec{y}) \ , \end{aligned}$$

where $g_\epsilon(\vec{y}) = \prod g_\epsilon(y_j)$, and $g_\epsilon(y_j) = (1-\epsilon)\,\phi(y_j) + \epsilon\,\psi(y_j)$ has a risk function that is linear in the best risk obtained by any unimodal prior. In particular,

$$(\forall n, \forall \vec{\mu}) \qquad \sum_{j=1}^{p} D(\phi_{\mu_j} \parallel g) \le \omega_0 + \omega_1 \inf_{\pi \in \mathcal{M}} \sum_{j=1}^{p} d_\pi(\mu_j) \ .$$

Where equation (6) and (20) define $\mathcal{M}$ and $d_\pi(\cdot)$ respectively.

We set the constant κ defined in (25) to $\kappa = e^2$ in the following derivations. This choice allows us to absorb several constants into a multiple of d_π itself.

Lemma 1. *Let $g_\delta(y) = (1-\delta)\,\phi(y) + \delta\,\psi(y)$. For any $\pi \in \mathcal{M}$, if $\epsilon_\kappa(\pi) < \delta \le 1/2$, where $\epsilon_\kappa(\pi)$ is defined in (26) then*

$$D(\phi_\mu \parallel g_\delta) \le 2\lambda\, d_\pi(\mu) + 2\delta \ ,$$

Proof. From equation (32) we see that for $\delta > \epsilon_\kappa(\pi)$,

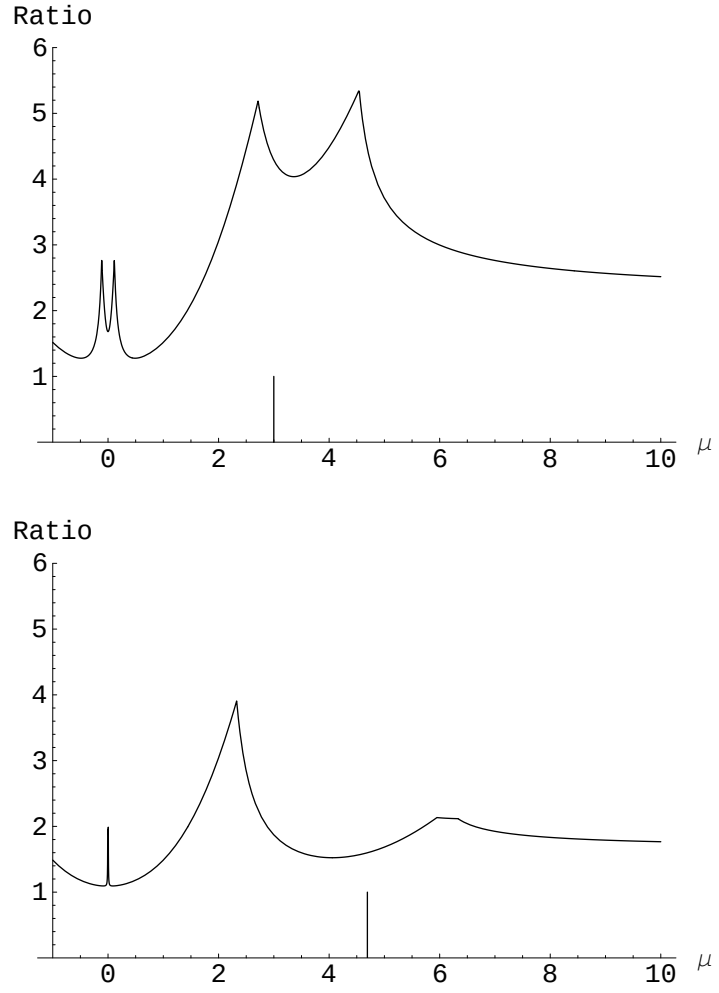
$$\begin{aligned} D(\phi_\mu \parallel g_\delta) &\le 2\min\{\mu^2 + \delta, 1/\mu^2 + \log \mu/\delta\} \\ &\le 2\delta + 2\min\{\mu^2, 1/\mu^2 + \log \mu/\delta\} \\ &\le 2\delta + 2\min\{\mu^2 + \epsilon_\kappa(\pi), 1/\mu^2 + \log \mu/\epsilon_\kappa(\pi)\} \\ &\le 2\delta + 2\lambda d_\pi(\mu) \ , \end{aligned}$$

where the last inequality comes from equation (29). □

To clarify the arguments of our proofs, we have shown analytically that the bound obtains for $2\lambda = 10000$. Numerical calculations suggest that the best coefficient is less than 6. To support this claim, Figure 7 shows the ratio of the upper bound for $D(\phi_\mu \parallel g_{\epsilon_\kappa})$ from Lemma 17 to the lower bound for $d_\pi(\mu)$ from Lemma 15 for two choices of ϵ_κ, (a) $\epsilon_\kappa = 0.01$ and (b) $\epsilon_\kappa = 0.00001$, both over the range $-1 \le \mu \le 10$. In all cases shown (and others not pictured here) the ratio of divergences $D(\phi_\mu \parallel g_{\epsilon_\kappa})/d_\pi(\mu) < 6$.

To obtain practical results that apply in the multivariate problem, we have to remove δ from the bound. The competitive bound continues to hold so long as the parameter δ indexing the marginal distribution g_δ is suitably close to ϵ_κ. At the same time, we can remove 2δ from the bound for

FIGURE 7. *This plot shows our bounding function for the risk of Bayes experts for (a)* $\epsilon_\kappa = 0.01$ *and (b)* $\epsilon_\kappa = 0.00001$ *In these examples the ratio* $D\big(\phi_\mu \,\big\|\, g_{\epsilon_\kappa}\big)/d_\pi(\mu)$ *is bounded by 6.*



$D(\phi_\mu \parallel g_\delta)$ provided in Lemma 1 by increasing the size of the multiplier. For $\delta/2 \le \epsilon_\kappa \le \delta$ and $\kappa = e^2$, $\frac{1+\log\kappa}{\kappa} \le \frac{1}{2}$. Thus, the bound $L_0(\mu,\epsilon)$ of Lemma 15 becomes (see Lemma 11)

$$d_\pi(\mu) \ge \epsilon_\kappa(1 - \frac{1+\log\kappa}{\kappa}) \ge \epsilon_\kappa/2 \; .$$

Consequently, we have

$$\begin{aligned} D(\phi_\mu \parallel g_\delta) &\le 2\delta + 2\lambda d_\pi(\mu) \le 4\epsilon_\kappa + 2\lambda d_\pi(\mu) \\ &\le (2\lambda+8) d_\pi(\mu) \; . \end{aligned} \tag{33}$$

We handle separately the special case for ϵ_κ near zero where the previous bound no longer usefully applies. Here, p is a positive integer that corresponds to the number of parameters in the regression model. From Lemma 1, we obtain for $\epsilon_\kappa < \delta < 1/(2p)$ that

$$D(\phi_\mu \parallel g_\delta) \le 2\lambda\, d_\pi(\mu) + 1/p \; , \tag{34}$$

where λ is the universal constant from Lemma 1.

From these results, we can extend the bounds on the ratio to problems in which $Y_i \sim N(\mu_i, 1)$, $i = 1, \ldots, p$. There is some δ such that $1/2p < \delta = 2^{-k} \le .5$, and

$$\begin{aligned} D(\phi_{\mu_i} \parallel g_\delta) &\le \max[(2\lambda+8)\, d_\pi(\mu_i), 2\lambda d_\pi(\mu_i) + 2/p] \\ &\le (2\lambda+8)\, d_\pi(\mu_i) + 1/p \; , \end{aligned} \tag{35}$$

where we bound each such term for $i = 1, \ldots, p$ by either (33) or (34). Summing the individual terms gives

$$\sum_{j=1}^{p} D(\phi_{\mu_j} \parallel g_\delta) \le 1 + (2\lambda+8) \sum_j d_\pi(\mu_j) \; . \tag{36}$$

The final step is to remove δ and obtain a universal model. We obtain a bound for the the divergence of such a model in the following.

Lemma 2. *Let $Y_i \sim N(\mu_i, 1)$, $i = 1, \ldots, p$. There exists a density g such that*

$$(\forall \pi) \quad \sum_{i=1}^{p} D(\phi_{\mu_i} \parallel g) \le 2 + (2\lambda+12) \sum_{i=1}^{p} d_\pi(\mu_i) \; ,$$

where λ is the universal constant from Lemma 1.

Proof: Let $k = \lceil \log_2 2p \rceil$. Let

$$g(y) = \sum_{j=1}^{k} 2^{-j} g_{(2^{-(k+1-j)})}(y) \; .$$

Suppose $\epsilon_\kappa(\pi) < 1/(2p)$. First note that $2^{-k} \le 1/(2p)$. Clearly

$$g(y) \ge \tfrac{1}{2} g_{(2^{-k})}(y) \; . \tag{37}$$

Hence

$$\begin{aligned}\sum_i D(\phi_{\mu_i} \parallel g) &\leq \log 2 + \sum D(\phi_{\mu_i} \parallel g_{(2^{-k})}) \\ &\leq \log 2 + 1 + 2\lambda \sum_i d_\pi(\mu_i) \ ,\end{aligned}$$

from equation (34), and our desired bound follows.

Now suppose $\epsilon_\kappa(\pi) \geq 1/(2p)$. Let j be such that $\epsilon_\kappa(\pi) \leq 2^{-(k+1-j)} < 2\epsilon_\kappa(\pi)$. We can bound the divergence by

$$\begin{aligned}\sum D(\phi_{\mu_i} \parallel g) &\leq j \log 2 + \sum_i D(\phi_{\mu_i} \parallel g_{(2^{-(k+1-j)})}) \\ &\leq j \log 2 + (2\lambda + 8) \sum d_\pi(\mu_i) \ ,\end{aligned}$$

where the second inequality follows from (36). Recalling that $d_\pi(\mu) \geq \epsilon_\kappa/2$ we see that $j \log 2 < 2^j \leq 2p\,\epsilon_\kappa < 4 \sum d_\pi(\mu_i)$, and again we get the desired bound. □

When we translate these divergences into risk notation (21), this Lemma implies Theorem 2.

Appendix: Calculating the estimator

The adaptive *PolyShrink* estimator $\tilde{\beta}(b)$ shown in Figure 3 and used in the simulations approximates the expectation of the model parameters. Because our model for the data summarized by the function $\tilde{g}$ in Theorem 2 does not use a typical parameterization, we obtain an equivalent estimator by using a conditional expectation – a predictor. This predictor is constructed so that its expectation *is* one of the model parameters.

We begin with a simple scalar problem, and then generalize the approach. Consider the scalar estimation problem in which $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(\mu, 1)$. Although our procedure does not produce an estimator of μ per se, we obtain an approximate estimator as the conditional expectation

$$\hat{\mu} = E_{\hat{f}}(Y_1 \mid Y_2, \ldots, Y_n) \ . \tag{38}$$

The expectation that defines $\hat{\mu}$ is with respect to the estimated joint density $\hat{f}(y_1, \ldots, y_n)$ for the data implied by our model.

To compute this estimator, we make use of a convenient orthogonal rotation that concentrates μ into one coordinate. Let W denote an $n \times n$ orthogonal matrix that has the form

$$W = \begin{bmatrix} \frac{1}{\sqrt{n}} & \sqrt{\frac{n-1}{n}} & 0 & \cdots & 0 \\ \frac{1}{\sqrt{n}} & -\sqrt{\frac{1}{n(n-1)}} & w_{23} & \cdots & w_{2n} \\ \frac{1}{\sqrt{n}} & -\sqrt{\frac{1}{n(n-1)}} & w_{33} & \cdots & w_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{n}} & -\sqrt{\frac{1}{n(n-1)}} & w_{n3} & \cdots & w_{nn} \end{bmatrix}$$

Let $Z_1, Z_2, \ldots, Z_n$ denote the coordinates of $Z = W^tY$, and observe that $Z_1 = \sqrt{n}\,\overline{Y}$ and that Y_1 appears only in Z_1 and Z_2. Concentrating Y_1 into two of the rotated coordinates simplifies the integration that defines $\hat{\mu}$. The remaining elements of W are arbitrary up the constraint that $W^tW = I$ and the zeros in the first row. For use below, write the leading terms of the rotation as

$$\begin{aligned} Z_1 &= \frac{1}{\sqrt{n}}Y_1 + \sqrt{\frac{n-1}{n}}S \\ Z_2 &= \sqrt{\frac{n-1}{n}}Y_1 - \sqrt{\frac{1}{n}}S\,, \end{aligned} \tag{39}$$

where

$$S = \sum_2^n Y_i/\sqrt{n-1} \sim N(\sqrt{n-1}\,\mu,\, 1) \tag{40}$$

denotes the normalized sufficient statistic for μ given $Y_2, \ldots, Y_n$.

With $\tilde{g}$ as defined in Theorem 2, our estimator of the joint distribution of $Y_1, \ldots, Y_n$ is

$$\hat{f}(y_1, \ldots, y_n) = \tilde{g}(z_1) \prod_{i=2}^n \phi(z_i)\,.$$

The density g only applies to z_1. Because we know that the data have a fixed mean in this scalar context, this model implies that the remaining coordinates $z_2, \ldots, z_p$ are Gaussian noise and free of μ. The estimator of μ can thus be written as

$$\tilde{\mu} = \frac{\int y_1 \hat{f}(y_1, \ldots, y_n)dy_1}{\int \hat{f}(y_1, \ldots, y_n)dy_1} = \frac{\int y_1 \tilde{g}(z_1)\phi(z_2)dy_1}{\int \tilde{g}(z_1)\phi(z_2)dy_1}\,. \tag{41}$$

The term $\prod_{i=3}^n \phi(z_i)$ cancels in this calculation because these coordinates are free of y_1 because of the zeros forced into the first row of W. Although the integrals in the ratio (41) lack a closed form, numerical integration works well because of the smoothness and thin tails of the Gaussian components. The dashed curve in Figure 3 shows $\sqrt{n}\,\tilde{\mu}$ as a function of S, the standardized sufficient statistic for μ given $Y_2, Y_3, \ldots, Y_n$. The figure shows both the estimator and its argument on a standard error scale.

While numerical integration is fine for showing plots of $\tilde{\mu}$, it is far too slow for routine calculation and simulation. Fortunately, it is easy to find a rather accurate approximation to $\tilde{\mu}$. This approximation also reveals $\tilde{\mu}$ has a rather familiar form. We obtain this approximation by expanding $\tilde{g}$ in the integrals in the ratio (41) and removing terms of order $1/n$. Using the definitions in (39), $\sqrt{n}$ times the numerator of $\tilde{\mu}$ is approximately

$$\begin{aligned} \sqrt{n}\int y_1 \tilde{g}(z_1)\phi(z_2)dy_1 &= \sqrt{n}\int y_1 \left(\tilde{g}(s) + \tilde{g}'(s)\frac{y_1}{\sqrt{n}}\right) \phi(y_1 - s/\sqrt{n})\, dy_1 + O(1/\sqrt{n}) \\ &= s\,\tilde{g}(s) + \tilde{g}'(s) + O(1/\sqrt{n})\,. \end{aligned}$$

The same approach handles the denominator and we arrive at

$$\sqrt{n}\,\tilde{\mu} \approx S + \frac{\tilde{g}'(S)}{\tilde{g}(S)}\,,$$

with the partial sum S defined in (40). The remainder in the approximation is quite small. Once $n > 10$ or so, plots of the approximate estimator are indistinguishable from those obtained by numerical integration.

The case in which the regression parameter β is vector-valued provides an important insight into the adaptive character of our estimator. To obtain a characterization similar to that provided for the scalar estimator, we require some rather special assumptions on the basis. These conditions are only needed to provide this characterization and do not affect, nor limit, the use of the estimator. They simply allow us to isolate a single element of β. Assume now as in (1) that $Y_i \sim N(x_i^t\beta, I)$, where β denotes an unknown p-element vector and the n row vectors x_i^t combine to form an $n \times p$ orthogonal matrix X. The p columns of are the first p columns of the $n \times n$ orthogonal matrix W. Assume also that the first column of X is constant, so that the scalar example is a special case. Let $\mu = \beta_1$. Assume further for convenience that $x_{1,2} = \cdots = x_{1,p} = 0$. For example, one might have a sinusoidal basis in which the columns of X are given by the discrete Fourier basis, $x_{i,j} = \sin 2\pi(i-1)(j-1)/n$ for $j = 2, \ldots, p$. As in the scalar case, assume that the remaining noise coordinates confine the impact of Y_1 by setting $w_{1,p+2} = \cdots = w_{1,n} = 0$, with $w_{1,1} \neq 0$ and $w_{1,p+1} \neq 0$.

The equivalent adaptive estimator of μ is again a conditional expectation, only one with a more complex model for the distribution of Y,

$$\tilde{\mu}_p = E_{\tilde{f}_p}[Y_1|Y_2, \ldots, Y_n] \; . \tag{42}$$

To define $\tilde{f}_p$, again let $Z = W^tY$. Then

$$\tilde{f}_p(y_1, \ldots, y_n) = \tilde{g}(z_1, \ldots, z_p) \prod_{i=p+1}^{n} \phi(z_i) \; .$$

Because of the convenient choice of the basis, the expectation simplifies to a ratio of one-dimensional integrals,

$$\tilde{\mu}_p = \frac{\int y_1 \tilde{g}(z_1, \ldots, z_p)\phi(z_{p+1})dy_1}{\int \tilde{g}(z_1, \ldots, z_p)\phi(z_{p+1})dy_1} \; . \tag{43}$$

Noting that $\tilde{g}$ is a weighted sum of mixtures, set $\epsilon_j = 2^{-j}$ and write

$$\begin{aligned}
\tilde{g}(z_1, \ldots, z_p) &= \sum_{j=1}^{k} \epsilon_j \tilde{g}_{\epsilon_{k+1-j}}\left(\sqrt{\frac{n-1}{n}}\, s + y_1/\sqrt{n}, z_2, \ldots, z_p)\right) \\
&= \sum_{j=1}^{k} \epsilon_j \tilde{g}_{\epsilon_{k+1-j}}\left(\sqrt{\frac{n-1}{n}}\, s + y_1/\sqrt{n}\right) \prod_{i=2}^{p} \tilde{g}_{\epsilon_{k+1-j}}(z_i)
\end{aligned}$$

Because of the position of zeros in the first row of W, the integration simplifies. Using the previous expression for $\tilde{g}(z_1, \ldots, z_p)$ and the arguments that provide the scalar estimator, we obtain (with $k = \lceil \log n \rceil$)

$$\sqrt{n} \quad \int y_1 \tilde{g}(z_1, \ldots, z_p)\phi(z_{p+1})dy_1$$

$$
\begin{aligned}
&= \sqrt{n}\sum_{j=1}^{k} \epsilon_j \tilde{g}_{\epsilon_{k+1-j}}(z_2,\ldots,z_p)\int y_1 \tilde{g}_{\epsilon_{k+1-j}}(z_1)\phi(z_{p+1})dy_1 \\
&= \sqrt{n}\sum_{j=1}^{k} \epsilon_j \tilde{g}_{\epsilon_{k+1-j}}(z_2,\ldots,z_p)\int y_1 \tilde{g}_{\epsilon_{k+1-j}}(s+y_1/\sqrt{n})\phi(y_1-S/\sqrt{n})dy_1 + O(1/\sqrt{n}) \\
&= \sum_{j=1}^{k} \epsilon_j \tilde{g}_{\epsilon_{k+1-j}}(z_2,\ldots,z_p)\left(s\,\tilde{g}_{\epsilon_{k+1-j}}(s) + \tilde{g}'_{\epsilon_{k+1-j}}(s)\right) + O(1/\sqrt{n}) \\
&= \sum_{j=1}^{k} \epsilon_j \tilde{g}_{\epsilon_{k+1-j}}(s,z_2,\ldots,z_p)\left(s + \tilde{g}'_{\epsilon_{k+1-j}}(s)/\tilde{g}_{\epsilon_{k+1-j}}(s)\right) + O(1/\sqrt{n})\ .
\end{aligned}
$$

Similarly, the denominator of $\tilde{\mu}_p$ reduces to

$$\int \tilde{g}(z_1,\ldots,z_p)\phi(z_{p+1})dy_1 = \sum_{j=1}^{k} \epsilon_j \tilde{g}_{\epsilon_{k+1-j}}(s,\, z_2,\ldots,z_p) + O(1/n).$$

Thus we obtain the approximation

$$\tilde{\mu}_p \approx S + \sum_{j=1}^{k} \omega_j \,\frac{\tilde{g}'_{\epsilon_{k+1-j}}(S)}{\tilde{g}_{\epsilon_{k+1-j}}(S)} \Big/ \sum_j \omega_j\ , \tag{44}$$

where the weights are $\omega_j = \epsilon_j\, \tilde{g}_{\epsilon_{k+1-j}}(S, z_2,\ldots,z_p)$.

This approximation shows how the estimator adapts to the presence of more signal in other components of the least-squares fit. If $z_2,\ldots,z_p$ are near zero, then the weights ω_j put most of the mass at $j=k$. That is, most of the weight is assigned to mixture components that downweight the Cauchy density. As the other coordinates move away from zero – suggesting a problem in which there is substantial signal – the ω_j shift to put more mass on components with larger weights on the Cauchy density. These weights then resemble an empirical Bayes procedure that chooses a prior that conforms to the estimated signal.

Appendix: Proofs

There are three subsections to this appendix. The first proves some Lemmas of general utility. The second section proves the lower bound statement of Theorem 1. The third section proves the upper bound of Theorem 1.

6.1. Preliminaries

We begin by deriving an obvious property of the Bayes estimator: for a prior $\pi \in \mathcal{M}$, the Bayes estimator of μ lies between 0 and the observed value. The case of $y<0$ is analogous.

Lemma 3. *Suppose $y \geq 0$. Then*

$$0 \leq E(\mu|Y=y) \leq y\ ,$$

where $Y \sim \phi(\mu,1)$, with a prior $\mu \sim \pi$ such that $\pi \in \mathcal{M}$, and $\mathcal{M}$ is defined in (6).

Proof:

Write $\phi = \phi_0$. The lower bound is obvious by partitioning the expectation as

$$\begin{aligned}
\phi_\pi(y)E(\mu|y) &= \int \mu\phi(y-\mu)\pi(\mu)\,d\mu \\
&= \int_{-\infty}^{0} \mu\phi(y-\mu)\pi(\mu)\,d\mu + \int_0^\infty \mu\phi(y-\mu)\pi(\mu)\,d\mu \\
&= \int_0^\infty -\mu\phi(y+\mu)\pi(\mu)\,d\mu + \int_0^\infty \mu\phi(y-\mu)\pi(\mu)\,d\mu \\
&= \int_0^\infty \mu\left(\phi(y-\mu)-\phi(y+\mu)\right)\pi(\mu)\,d\mu \\
&\geq 0\,.
\end{aligned}$$

The upper bound comes from a similar partitioning argument. The idea is to remove most of the integral by use of the symmetry around 0 of ϕ and ϕ_π. For this purpose, define

$$\tilde{\pi}(\mu) = \begin{cases} \pi(\mu) & \mu > y\,, \\ \pi(2y-\mu) & \mu \leq y\,, \end{cases}$$

which is symmetric around y. Thus, $\int (y-\mu)\phi(y-\mu)\tilde{\pi}(\mu)\,d\mu = 0$. With this definition, we have

$$\begin{aligned}
\phi_\pi(y)E(y-\mu|Y=y) &= \int (y-\mu)\phi(y-\mu)\pi(\mu)\,d\mu \\
&= \int (y-\mu)\phi(y-\mu)\left(\pi(\mu)-\tilde{\pi}(\mu)\right)\,d\mu \\
&= \int_{-\infty}^{y} (y-\mu)\phi(y-\mu)\left(\pi(\mu)-\tilde{\pi}(\mu)\right)\,d\mu \\
&\geq 0
\end{aligned}$$

since all of the terms in the integral are positive for $\mu < y$. □

The second preliminary result that we show establishes the monotonicity of the divergence of the Bayes mixture.

Lemma 4. *The divergence $d_\pi(\mu)$ (defined in equation (20)) is monotone increasing as μ moves away from 0. That is, for $|\mu_0| < |\mu_1|$,*

$$d_\pi(|\mu_0|) < d_\pi(|\mu_1|)\,.$$

where $d_\pi(\cdot)$ is

Proof:

Since $d_\pi(\mu) = d_\pi(-\mu)$, we need only consider the case $0 < \mu_0 < \mu_1$. Define the midpoint $\zeta = (\mu_0+\mu_1)/2$. Then

$$\begin{aligned}
d_\pi(\mu_1) - d_\pi(\mu_0) &= \int \phi(y-\mu_0)\log\phi_\pi(y)\,dy - \int \phi(y-\mu_1)\log\phi_\pi(y)\,dy \\
&= \int \phi(u)\log\phi_\pi(u+\mu_0)\,du - \int \phi(u)\log\phi_\pi(u+\mu_1)\,du \\
&= \int_\zeta^\infty \phi(u)\log\frac{\phi_\pi(u+\mu_0)}{\phi_\pi(u+\mu_1)}du - \int_{-\infty}^{\zeta} \phi(u)\log\frac{\phi_\pi(u+\mu_1)}{\phi_\pi(u+\mu_0)}du
\end{aligned}$$

$$\begin{aligned} &= \int_0^\infty \phi(u-\zeta)\log\frac{\phi_\pi(u+\mu_0-\zeta)}{\phi_\pi(u+\mu_1-\zeta)}du - \int_0^\infty \phi(-u-\zeta)\log\frac{\phi_\pi(\mu_1-\zeta-u)}{\phi_\pi(\mu_0-\zeta-u)}du \\ &= \int_0^\infty (\phi(\zeta-u)-\phi(\zeta+u))\log\frac{\phi_\pi(u+(\mu_0-\mu_1)/2)}{\phi_\pi(u+(\mu_1-\mu_0)/2)}du \\ &> 0\,, \end{aligned}$$

since both factors in the final integral are positive over the range of integration. □

We will also need the following property of the likelihood ratio.

Lemma 5. *The likelihood ratio of $\phi_\pi(y)/\phi_0(y)$ is monotone increasing for $y > 0$:*

$$\frac{d}{dy}\,\frac{\phi_\pi(y)}{\phi_0(y)} > 0\,. \tag{45}$$

Proof. Assume $y > 0$. Differentiating under the integral, we have

$$\begin{aligned} \frac{d}{dy}\frac{\phi_\pi(y)}{\phi_0(y)} &= \frac{\phi_0(y)\phi_\pi'(y) - \phi_\pi(y)\phi_0'(y)}{\phi_0^2(y)} \\ &= \frac{\phi_0(y)\left(\int \mu\phi(y-\mu)\pi(\mu)d\mu - y\phi_\pi(y)\right) + y\phi_0(y)\phi_\pi(y)}{\phi_0^2(y)} \\ &= \frac{\int \mu\phi(y-\mu)\pi(\mu)d\mu}{\phi_0(y)} \\ &> 0\,, \end{aligned}$$

because of the monotonicity of the Bayes estimator shown in Lemma 3. □

The following Lemma offers a more precise version of the approximation given in (28). We do not exploit the accuracy of this approximation and include it here for completeness.

Lemma 6.

$$\sqrt{-2\log\epsilon_\kappa + \log(2/\pi) - 2\log\kappa} - 1 \le \tau_\kappa \le \sqrt{-2\log\epsilon_\kappa + \log(2/\pi) - 2\log\kappa}$$

and:

$$\frac{2\kappa e^{-(\tau_\kappa+1)^2/2}}{\sqrt{2\pi}} \le \epsilon_\kappa \le \frac{2\kappa e^{-\tau_\kappa^2/2}}{\tau\sqrt{2\pi}}$$

Proof:

$$\epsilon_\kappa = 2\kappa\Phi(-\tau) \le 2\kappa\phi_0(\tau)/\tau \le \frac{2\kappa e^{-\tau^2/2}}{\tau\sqrt{2\pi}}\,.$$

The other direction is

$$\epsilon_\kappa = 2\kappa\Phi_0(-\tau) \ge 2\kappa\phi_0(\tau+1) \ge \frac{2\kappa e^{-(\tau+1)^2/2}}{\sqrt{2\pi}}\,.$$

For $\tau_\kappa > 1$, these show that

$$\begin{aligned} \log\epsilon_\kappa &\le \tfrac{1}{2}\log(2/\pi) + \log\kappa - \tau^2/2 \\ \log\epsilon - \tfrac{1}{2}\log(2/\pi) - \log\kappa &\le -\tau^2/2 \\ -2\log\epsilon + \log(2/\pi) + 2\log\kappa &\ge \tau^2 \\ \sqrt{-2\log\epsilon + \log(2/\pi) + 2\log\kappa} &\ge \tau \end{aligned}$$

For the other bound, we have:

$$\begin{aligned}
\log\epsilon &\geq \tfrac{1}{2}\log(2/\pi) + \log\kappa - (\tau+1)^2/2 \\
2\log\epsilon &\geq \log(2/\pi) + 2\log\kappa - (\tau+1)^2 \\
2\log\epsilon - \log(2/\pi) - 2\log\kappa &\geq -(\tau+1)^2 \\
\sqrt{-2\log\epsilon + \log(2/\pi) - 2\log\kappa} - 1 &\leq \tau
\end{aligned}$$

□

6.2. Lower Bounds

This section proves equation (12) which is the lower bound statement of Theorem 1.

Our first Lemma is a direct consequence of the monotonicity of d_π and the well-known relationship between the L_1 norm and the divergence. This Lemma shows that d_π (defined in equation (20)) lies above a simple function of the cumulative normal distribution Φ. This bound is useful for $d_\pi(\mu)$ for values of $\mu \approx 1$. Since the bound provided by this Lemma goes to 0 as $\mu \to 0$ and is itself bounded by $1/(2\log 2)$, we need to improve it for both small and large values of the parameter.

Lemma 7.

$$d_\pi(\mu) \geq \frac{(2\Phi(\mu)-1)^2}{2\log 2} \; .$$

Proof. It is well known that the divergence is bounded below by a multiple of the L_1 norm [for example, see 4]

$$D(f \parallel g) \geq \frac{1}{2\log 2}\|f-g\|_1^2 \; . \tag{46}$$

The monotonicity of d_π established in Lemma 4 shows that

$$\begin{aligned}
d_\pi(\mu) &\geq \max(d_\pi(\mu), d_\pi(0)) \\
&= \max(D(\phi_\mu \parallel \phi_\pi), D(\phi_0 \parallel \phi_\pi)) \; .
\end{aligned}$$

Now using the L_1 bound from (46), we obtain

$$\begin{aligned}
d_\pi(\mu) &\geq \frac{1}{2\log 2}\max\left(\|\phi_\mu - \phi_\pi\|_1^2, \|\phi_0 - \phi_\pi\|_1^2\right) \\
&\geq \frac{1}{8\log 2}\|\phi_\mu - \phi_0\|_1^2 \; ,
\end{aligned}$$

where the last step uses the triangle inequality for the L_1 norm,

$$\|a-c\|_1 = \|a-b+b-c\|_1 \leq \|a-b\|_1 + \|b-c\|_1 \leq 2\max(\|a-b\|_1, \|b-c\|_1) \; .$$

The expression in the statement of the Lemma comes from noting that the L_1 distance between two normal densities is

$$\|\phi_\mu - \phi_\theta\| = 2\left(2\Phi\left(\frac{|\mu-\theta|}{2}\right) - 1\right) \; .$$

□

Our next step in providing lower bounds for ϕ_π is to find a lower bound for $d_\pi(0)$. Our basis for this bound is the following Lemma. This Lemma bounds for the divergence based on the difference in probabilities assigned to a set. It is similar in result and proof to the well-known L_1 lower bound for the divergence provided, for example, in [4]. This version is specialized to our setting.

Lemma 8. *Let f and g denote two scalar probability measures, with the support of f contained in the support of g. Let A denote a measurable set with f measure $\epsilon = f(A)$ such that*

$$\delta = g(A) < f(A) = \epsilon \; .$$

Then

$$D(g \parallel f) > h(\delta, \epsilon) = \delta \log \frac{\delta}{\epsilon} + (1-\delta) \log \frac{1-\delta}{1-\epsilon} \; . \tag{47}$$

Proof. To make the notation more descriptive, define the divergence between two random variables as the divergence of their distributions. Using this notation, if X and Y denote random variables with distributions defined by f and g, respectively, then $D(g \parallel f) = D(Y \parallel X)$. Now let $1_A(x)$ denote the indicator function for the set A. Conditioning allows us to write the divergence in two ways by making use of the chain rule for divergence (*e.g.*, [4]),

$$\begin{aligned} D((Y_1, Y_2) \parallel (X_1, X_2)) &= D(Y_1 \parallel X_1) + D(Y_2|Y_1 \parallel X_2|X_1) \\ &= D(Y_2 \parallel X_2) + D(Y_1|Y_2 \parallel X_1|X_2) \end{aligned} \tag{48}$$

Since $1_A(X)$ is constant given X (and similarly for $1_A(Y)$ given Y), $D(1_A(Y)|Y \parallel 1_A(X)|X) = 0$, and (48) implies

$$\begin{aligned} D(Y \parallel X) &= D(1_A(Y) \parallel 1_A(X)) + D(Y|1_A(Y) \parallel X|1_A(X)) \\ &\geq D(1_A(Y) \parallel 1_A(X)) \\ &= \delta \log \frac{\delta}{\epsilon} + (1-\delta) \log \frac{1-\delta}{1-\epsilon} \; , \end{aligned}$$

where the inequality obtains because the divergence is non-negative. □

This Lemma trivially leads to a global lower bound for $d_\pi(\mu)$. In particular, if in equation (47) we set $\epsilon = \epsilon_\kappa(\pi)$ (from equation (26)) and $\delta = \delta_\kappa(\pi) = \epsilon_\kappa/\kappa$ from (27), then we obtain following lower bound.

Lemma 9.

$$d_\pi(\mu) \geq (\epsilon_\kappa/\kappa) \log \frac{1}{\kappa} + (1 - \epsilon_\kappa/\kappa) \log \frac{1 - \epsilon_\kappa/\kappa}{1 - \epsilon_\kappa} \; .$$

We have two useful approximations of the lower bound provided in Lemma 9. The first is better for small ϵ_κ.

Lemma 10. *For all μ and $\kappa > 1$,*

$$d_\pi(\mu) \geq \epsilon_\kappa \left(1 - \frac{1 + \log \kappa}{\kappa}\right) \; .$$

Proof. Using $\delta_\kappa(\pi)$ from (27) in $h(\delta, \epsilon)$ defined in (47), we have

$$\begin{aligned} d_\pi(\mu) \geq h(\delta_\kappa(\pi), \epsilon_\kappa) &= (1 - \frac{\epsilon_\kappa}{\kappa}) \log(\frac{1}{1 - \frac{\epsilon_\kappa(1-1/\kappa)}{1-\frac{\epsilon_\kappa}{\kappa}}}) + \epsilon_\kappa/\kappa \log(1/\kappa) \\ &\geq (1 - \frac{\epsilon_\kappa}{\kappa}) \frac{\epsilon_\kappa(1-1/\kappa)}{1-\frac{\epsilon_\kappa}{\kappa}} - \epsilon_\kappa/\kappa \log(\kappa) \\ &= \epsilon_\kappa(1-1/\kappa) - \epsilon_\kappa/\kappa \log(\kappa) \\ &= \epsilon_\kappa(1 - \frac{1+\log\kappa}{\kappa}) \; . \end{aligned}$$

□

The second bound relies on the property of the Bayes estimator from Lemma 3 along with Jensen's inequality.

Lemma 11. *For all μ and $\kappa > 1$,*

$$d_\pi(\mu) \geq \frac{\epsilon_\kappa(1-1/\kappa)^2}{2(1-\epsilon_\kappa)} \; .$$

Proof. We again use $h(\delta, \epsilon)$ from (47) and define

$$q(\delta, \epsilon) = \frac{1}{2} \, \frac{(\delta - \epsilon)^2}{\epsilon(1-\epsilon)} \; .$$

Write the difference as $r(\delta, \epsilon) = h(\delta, \epsilon) - q(\delta, \epsilon)$. Now observe that $r(\epsilon, \epsilon) = 0$ and that the partial derivative $(\partial/\partial\delta) r(\delta, \epsilon) = 0$ if $\delta = \epsilon$. The second partial derivative of r is positive for $\delta < \epsilon$,

$$\frac{\partial^2}{\partial \, \delta^2} r(\delta, \epsilon) = \frac{1}{\delta(1-\delta)} - \frac{1}{\epsilon(1-\epsilon)} > 0 \; ,$$

since the maximum value of the denominator occurs at 1/2. It follows that $r(\delta, \epsilon)$ is convex for $0 < \delta < \epsilon < 1/2$ and lies above the tangent x-axis, and so $r(\delta, \epsilon) > 0$ under these conditions. The bound obtains for d_π since $\epsilon_\kappa < 1/2$ by definition.

Now define the set A of Lemma 8 as $A = \{y : |y| > \tau_\kappa\}$. From (27), $\epsilon_\kappa = \kappa \, \delta_\kappa(\pi)$. Plugging these into the bound on the divergence given by Lemma 8, we obtain

$$\begin{aligned} d_\pi(0) &\geq \frac{\epsilon_\kappa^2(1-1/\kappa)^2}{2\epsilon_\kappa(1-\epsilon_\kappa)} \\ &= \frac{\epsilon_\kappa(1-1/\kappa)^2}{2(1-\epsilon_\kappa)} \end{aligned}$$

The bound holds everywhere because of the monotonicity of d_π established in Lemma 4. □

For larger values of $|\mu|$, we have the following two Lemmas. The first provides a useful lower bound for $d_\pi(\mu)$ for $\mu \approx \tau_\kappa$.

Lemma 12. *Let $Y \sim \phi_\mu$. Then*

$$d_\pi(\mu) \geq \mu^2/2 + \tfrac{1}{2} \left((\tau_\kappa^2 - 1 - \mu^2) P(|Y| \geq \tau_\kappa) + \phi_\mu(\tau_\kappa)(|\mu| - \tau - e^{2|\mu|\tau_\kappa}(|\mu| + \tau)) \right) - \log \kappa$$

Proof. As in the prior proof, let $A = \{y : |y| < \tau_\kappa\}$ and let A^c denote its complement. The proof of this Lemma is a direct evaluation of the divergence, with the inequality obtained by exploiting the monotonicity of $\phi_\pi(|y|)$ and the definition of τ_κ in (25). This definition implies $\phi_\pi(y) < \kappa\phi_0(y)$ for $|y| < \tau_\kappa$.

$$\begin{aligned} d_\pi(\mu) &= \int \log \frac{1}{\phi_\pi(y)} \phi_\mu(y) dy - H(\phi) \\ &\geq \int_A \log \frac{1}{\kappa\phi_0(y)} \phi_\mu(y) dy + \int_{A^c} \log \frac{1}{\kappa\phi_0(\tau_\kappa)} \phi_\mu(y) dy - H(\phi) \ , \end{aligned}$$

since $\phi_\pi(\tau_\kappa) = \kappa\phi_0(\tau_\kappa)$. Collecting terms, we have

$$\begin{aligned} d_\pi(\mu) &\geq \tfrac{1}{2} \int \min(y^2, \tau_\kappa^2) \phi_\mu(y) dy + \tfrac{1}{2} \log 2\pi - H(\phi) - \log \kappa \\ &= \tfrac{1}{2} \left(\int_A y^2 \phi_\mu(y) dy + \tau_\kappa^2 P(|Y| \geq \tau_\kappa) \right) - \tfrac{1}{2} - \log \kappa \ , \end{aligned}$$

since $H(\phi) - \frac{1}{2}\log 2\pi = 1/2$. Solving the integral in somewhat closed form and collecting constants give the lower bound for $d_\pi(\mu)$ as claimed in the Lemma. □

We use a separate result to bound $d_\pi(\mu)$ for larger $|\mu|$. The proof of this Lemma is an application of the so-called data-processing inequality to the random variable $Y \vee \mu - \delta$ where $Y \sim \phi_\pi$.

Lemma 13. *For arbitrary δ, set $\alpha = \Phi(-\delta)$. Then for $|\mu| \geq \delta$,*

$$d_\pi(\mu) \geq (1-\alpha) \log \frac{1}{\phi_\pi(|\mu| - \delta)} + \alpha \log \alpha - H(\phi) \ . \tag{49}$$

Proof. Assume $\mu > \delta$ and split the divergence $d_\pi(\mu)$ into two parts,

$$\begin{aligned} d_\pi(\mu) &= \int \phi_\mu(y) \log \frac{\phi_\mu(y)}{\phi_\pi(y)} dy \\ &= \int_{y \leq \mu-\delta} \phi_\mu(y) \log \frac{\phi_\mu(y)}{\phi_\pi(y)} dy + \int_{y > \mu-\delta} \phi_\mu(y) \log \frac{\phi_\mu(y)}{\phi_\pi(y)} dy \ . \end{aligned}$$

Let A denote the event $\{Y \leq \mu - \delta\}$. Let $\phi_\mu(y|A)$ and $\phi_\pi(y|A)$ denote the conditional probability measures, and define $\phi_\pi(A) = \int_A \phi_\pi(x) dx$. The first term of the divergence can be analyzed using a conditioning argument.

$$\begin{aligned} \int_{y \leq \mu-\delta} \phi_\mu(y) \log \frac{\phi_\mu(y)}{\phi_\pi(y)} dy &= \int_{y \leq \mu-\delta} \alpha\phi_\mu(y|A) \log \frac{\alpha\, \phi_\mu(y|A)}{\phi_\pi(y|A)\phi_\pi(A)} dy \\ &= \alpha \log \frac{\alpha}{\phi_\pi(A)} + \alpha \int_{y \leq \mu-\delta} \phi_\mu(y|A) \log \frac{\phi_\mu(y|A)}{\phi_\pi(y|A)} dy \\ &\geq \alpha \log \frac{\alpha}{\phi_\pi(A)} \\ &\geq \alpha \log \alpha \ . \end{aligned} \tag{50}$$

The second term can be analyzed more directly using the monotonicity of ϕ_π:

$$\int_{y > \mu-\delta} \phi_\mu(y) \log \frac{\phi_\mu(y)}{\phi_\pi(y)} dy \geq -H(\phi) + \int_{y > \mu-\delta} \phi_\mu(y) \log \frac{1}{\phi_\pi(y)} dy$$

$$\begin{aligned} &\geq -H(\phi) + \int_{y>\mu-\delta} \phi_\mu(y) \log \frac{1}{\phi_\pi(\mu-\delta)} dy \\ &= -H(\phi) - (1-\alpha) \log \phi_\pi(\mu-\delta) \ . \end{aligned} \tag{51}$$

With the bounds (50) and (51) combined, we arrive at the lower bound as stated. □

To make use of this Lemma, we remove the Bayes mixture ϕ_π from (49) and replace it with a simple function of ϵ_κ defined in (26). Assume $\mu > \tau_\kappa + \delta$. Then we can bound $\phi_\pi(\mu-\delta)$ from above by

$$\phi_\pi(\mu-\delta) < \frac{\epsilon_\kappa}{\mu-\delta-\tau_\kappa} \ . \tag{52}$$

because the tail area ϵ_κ is greater than the area of a rectangle of height $\phi_\pi(\mu-\delta)$ and width $2(\mu-\delta-\tau_\kappa)$. Putting the bound for ϕ_π from (52) into (49) gives the following Lemma:

Lemma 14. *For $|\mu| > \tau_\kappa + \delta$, define $\alpha = \Phi(-\delta)$. Then*

$$d_\pi(\mu) \geq (1-\alpha) \log \frac{|\mu|-\delta-\tau_\kappa}{\epsilon_\kappa} + \alpha \log \alpha - H(\phi) \ . \tag{53}$$

The choice of δ remains. Clearly one would like to choose δ so as to maximize the bound in (53). Although we can numerically find a value of δ that maximizes this bound, the expression as offered is too complex to allow an explicit solution. However, if we assume δ is of moderate size, we can approximate the bound in (53) as

$$h(\delta) = \Phi(\delta) \log(\mu-\delta-\tau_\kappa) \ . \tag{54}$$

If we then set $(\partial/\partial\delta)h(\delta) = 0$ and drop constants and terms of smaller size, we arrive at an approximately optimal choice of δ. If we write the solution as a function of μ and the threshold, then the approximate optimal choice is $\hat{\delta}(\mu, \tau_\kappa)$ where

$$\hat{\delta}(\mu,\tau) = \sqrt{2\log(\mu-\tau)} \quad \text{for } \mu > \tau + 1 \ , \tag{55}$$

and zero otherwise. The following Lemma collects several bounds for $d_\pi(\mu)$ from Lemma 7, Lemma 9 and Lemma 14.

Lemma 15.

$$d_\pi(\mu) \geq L(\mu, \epsilon_\kappa)$$

where

$$L(\mu,\epsilon) = \max\{L_0(\mu,\epsilon),\ L_1(\mu),\ L_2(\mu,\epsilon),\ L_3(\mu,\epsilon)\} \ , \tag{56}$$

where the bounding functions are

$$L_0(\mu,\epsilon) = (\epsilon_\kappa/\kappa) \log \frac{1}{\kappa} + (1-\epsilon_\kappa/\kappa) \log \frac{1-\epsilon_\kappa/\kappa}{1-\epsilon_\kappa} \ ,$$

$$L_1(\mu) = \frac{(2\Phi(\mu)-1)^2}{2\log 2} \ ,$$

$$L_2(\mu,\epsilon) = \tfrac{1}{2}\left(\int_A y^2\phi_\mu(y)dy + \tau_\kappa^2 P(|Y|\ge\tau_\kappa) - 1\right) - \log\kappa$$

and

$$L_3(\mu,\epsilon) = (1-\alpha)\log\frac{|\mu|-\delta-\tau_\kappa}{\epsilon_\kappa} + \alpha\log\alpha - H(\phi),$$

where in L_3 *we assume* $|\mu| > \delta+\tau_\kappa$ *and set* $\alpha = \Phi(-\delta)$.

A simpler form of these bounds facilitates our comparison of the expert divergence d_π to that obtained by the predictive model g. Though the bounds are not so tight as those just laid out, they are adequate to show the existence of finite ratios.

Lemma 16. *For* $\kappa = e^2$*:*

$$\exists\lambda>0 \quad s.t. \quad \forall\pi\in\mathcal{M}: \lambda\, d_\pi(\mu) \ge \min\left(\mu_i^2+\epsilon_\kappa(\pi), \frac{1}{\mu_i^2} + \log\frac{\mu_i}{\epsilon_\kappa(\pi)}\right)$$

Proof. We proceed by simplifying the bounds for various ranges of μ, starting with those near 0.

Case A: $(\mu^2 \le \epsilon_\kappa)$ For $\kappa = e^2$, $\frac{1+\log\kappa}{\kappa} \le \frac{1}{2}$. Thus, Lemma 11 gives

$$\begin{aligned} d_\pi(\mu) &\ge \epsilon_\kappa(1-\frac{1+\log\kappa}{\kappa}) \ge \epsilon_\kappa/2 \ge \mu^2/4 + \epsilon_\kappa/4 \\ &= (\mu^2+\epsilon_\kappa)/4\,. \end{aligned} \tag{57}$$

So any $\lambda > 4$ will work.

Case B: $(\epsilon_\kappa \le \mu^2 \le 9)$ For μ in this range, $(2\Phi(\mu)-1)^2 \ge \mu^2/10$. Thus, the L_1 bound from Lemma 7 gives

$$\begin{aligned} d_\pi(\mu) &\ge \frac{(2\Phi(\mu)-1)^2}{2\log 2} \ge \frac{\mu^2}{20\log 2} \ge \frac{\mu^2/2+\epsilon_\kappa/2}{20\log 2} \\ &\ge (\mu^2+\epsilon_\kappa)/(40\log 2) \end{aligned} \tag{58}$$

So any $\lambda > 40\log 2$ will work.

Case C: $(9 \le \mu^2 \le \tau_\kappa^2/2)$ We simplify the $L_2(\mu,\epsilon)$ bound by replacing the expectation $E_\mu(Y^2\wedge\tau_\kappa^2)$ by something more manageable, here a truncated linear function that is tangent to the quadratic. Recalling $\log\kappa = 2$, we obtain

$$\begin{aligned} L_2(\mu,\epsilon) &= \tfrac{1}{2}\left(\int_A y^2\phi_\mu(y)dy + \tau_\kappa^2 P(|Y|\ge\tau_\kappa) + \log 2\pi\right) - H(\phi) - \log\kappa \\ &\ge \tfrac{1}{2}\int (y\wedge\mu)^2\phi_\mu(y)dy - 5/2 \\ &\ge \tfrac{1}{2}\int \max[0,\min(\tau_\kappa^2,\mu(2y-\mu))]\phi_\mu(y)dy - 5/2 \\ &\ge \tfrac{\mu}{2}\int \max[-\tau_\kappa^2/\mu,\min(\tau_\kappa^2/\mu,(2y-\mu))]\phi_\mu(y)dy - 5/2 \\ &= \mu^2/2 - 5/2 \end{aligned}$$

$$\geq \quad \mu^2/2 - 5/2 \ ,$$

since $\max(-\tau_\kappa^2/\mu, \min(\tau_\kappa^2/\mu, (2y-\mu)))$ is an odd function centered at μ.

For $\mu > 3$ we see that

$$(\mu^2 - 5)/2 > \mu^2/10 + 1 > \mu^2/10 + \epsilon/10$$

so our bounds apply with $\lambda = 10$.

Case D: $(\tau_\kappa^2/2 \leq \mu^2 \leq 72\tau_\kappa^2)$ By monotonicity of d_π (see Lemma 4), we know that for μ in this range $d_\pi(\mu) > d_\pi(\tau_\kappa/\sqrt{2}) = (\tau_\kappa^2 + \epsilon)/20$. At the upper limit for this range, we must have

$$72\tau_\kappa^2 + \epsilon \leq \lambda(\tau_\kappa^2 + \epsilon)/20 \ ,$$

so $\lambda \geq 1440$ will work.

Notice that if τ_κ is less than 3 (and hence the previous case holds vacuously) we still achieve this bound by appealing to the case before that. This means $\lambda \geq 72 * 40 \log 2 = 5760 \log 2$ will suffice.

Case E: $(72\tau_\kappa^2 \leq \mu^2)$ First note that $\tau_\kappa > \sqrt{2}$ for $\kappa = e^2$. If we set $\delta = \Phi^{-1}(.9) \approx 1.28$ in the bound $L_3(\mu, \epsilon)$ given in Lemma 15, then for this range of μ:

$$\begin{aligned}
d_\pi(\mu) &\geq 0.9 \log \frac{\mu - \tau_\kappa - \delta}{\epsilon_\kappa} + 0.1 \log 0.1 - H(\phi) \\
&\geq 0.1 \log \frac{\mu - \tau_\kappa - \delta}{\epsilon_\kappa} + 0.8 \log[2(6\sqrt{2}-1)\sqrt{2} - \delta] + 0.1 \log 0.1 - H(\phi) \\
&\geq 0.1 \log \frac{\mu - \tau_\kappa - \delta}{\epsilon_\kappa} + 0.8 \log 18 + 0.1 \log 0.1 - H(\phi) \\
&\geq 0.1 \log 1/\epsilon_\kappa + 0.1 \log(\mu - \tau_\kappa - \delta) + 0.66 \\
&\geq 0.1 \log 1/\epsilon_\kappa + 0.1 \log \mu + 0.1 \log \frac{\mu - \tau_\kappa - \delta}{\mu} + 0.66 \\
&\geq 0.1 \log 1/\epsilon_\kappa + 0.1 \log \mu + 0.1 \log \frac{(6\sqrt{2}-1)\tau_\kappa - \delta}{6\sqrt{2}\tau_\kappa} + 0.66 \\
&\geq 0.1 \log 1/\epsilon_\kappa + 0.1 \log \mu + 0.6 \qquad (59)
\end{aligned}$$

Since for $\mu \geq 1$ we have that $.6 \geq .1/\mu^2$, we can write the lower bound as $d_\pi(\mu) \geq 0.1 \log \mu/\epsilon_\kappa + 0.1/\mu^2$ since μ is bigger than 1. Thus, any $\lambda \geq 10$ will work for this case.

In summary, any $\lambda \geq 5760 \log 2$ will prove our desired result. □

6.3. Upper Bounds

Lemma 17. *For* $g_\epsilon(y) = (1-\epsilon)\phi(y) + \epsilon\psi(y)$

$$D(\phi_\mu \parallel g_\epsilon) \leq \min\{\mu^2/2 - \log(1-\epsilon), \log \mu^2\pi + 3/\mu^2 - \log(\epsilon) - H(\phi)\} \ .$$

Proof. Begin by bounding the divergence as the minimum of the divergence of either density used to define g_ϵ:

$$D(\phi_\mu \parallel g_\epsilon) \quad = \quad D(\phi_\mu \parallel (1-\epsilon)\phi_0 + \epsilon\psi)$$

$$\leq \quad \min(D(\phi_\mu \parallel (1-\epsilon)\phi_0), D(\phi_\mu \parallel \epsilon\,\psi)) \; .$$

The first term is easy:

$$D(\phi_\mu \parallel (1-\epsilon)\phi_0) = D(\phi_\mu \parallel \phi_0) - \log(1-\epsilon) = \mu^2/2 - \log(1-\epsilon) \; .$$

The second piece is less straightforward. First write

$$D(\phi_\mu \parallel \epsilon\,\psi) = D(\phi_\mu \parallel \psi) - \log(\epsilon) \; .$$

We can bound the divergence between the Gaussian and the scaled Cauchy, $\psi(z) = 1/(\sqrt{2}\pi(1 + z^2/2))$, as follows,

$$\begin{aligned} D(\phi_\mu \parallel \psi) &= \int \phi_\mu(y) \log \frac{\phi_\mu(y)}{\psi(y)} dy \\ &= \int \phi_\mu(y) \log(\pi\sqrt{1/2}(2+y^2)) dy - H(\phi) \\ &= \log(\mu^2+a)\pi/\sqrt{2} + \int \phi(z) \log\left(1 + \frac{2\mu z + z^2 + 2 - a}{\mu^2 + a}\right) dy - H(\phi) \end{aligned}$$

Using the bound $\log(1+x) \leq x$ and $a = 3$ we get a bound on the divergence of $D(\phi_\mu \parallel \psi) \leq \log(\mu^2+3)\pi\sqrt{2} - H(\phi) \leq \log \mu^2 + \log \pi\sqrt{2} + 3/\mu^2 - H(\phi)$. □

The following Lemma is weaker than the above bound but simpler to work with theoretically. It proves the second statement of Theorem 1.

Lemma 18. *If $\epsilon < .5$, then for $g_\epsilon(y) = (1-\epsilon)\phi(y) + \epsilon\psi(y)$*

$$D(\phi_\mu \parallel g_\epsilon) \leq 2\min\{\mu^2 + \epsilon, 1/\mu^2 + \log \mu/\epsilon\} \; .$$

Proof. We show this by bounding each term in Lemma 17 separately. The first term is bounded above by

$$\begin{aligned} \mu^2/2 - \log(1-\epsilon) &\leq \mu^2/2 + (3/2)\epsilon \\ &\leq 2\mu^2 + 2\epsilon \; , \end{aligned}$$

where the first inequality follows from the restriction that $\epsilon \leq .5$. The second term is bounded as:

$$\begin{aligned} \log \mu^2\pi + 3/\mu^2 - \log(\epsilon) - H(\phi) &= 2\log\mu + \log\pi + 3/\mu^2 - \log(\epsilon) - (\log 2\pi e)/2 \\ &= 2\left(\log\frac{\mu}{\epsilon} + 1/\mu^2\right) + \left(1/\mu^2 + \log\epsilon + \frac{(\log\pi/2e)}{2}\right) \\ &\leq 2\left(\log\frac{\mu}{\epsilon} + 1/\mu^2\right) + \left(1/\mu^2 + \log\epsilon\right) \end{aligned}$$

If $\mu \geq \sqrt{\log 2}$ and$\epsilon \leq 1/2$ then $1/\mu^2 + \log\epsilon$ is less than zero and the bound follows. If $\mu \leq 1$, then $\mu^2 + \epsilon$ is the binding term and we have our desired result. Further, if $\epsilon < e^{-4}$, and $1 \leq \mu \leq 2$ the $\mu^2 + \epsilon$ is again binding. For the region $e^{-4} \leq \epsilon \leq 1/2$ and $1 \leq \mu \leq 2$ a look at a graph generated by numerical calculation is sufficient to see that the bound holds over that region. □

References

[1] Abramovich, F., Benjamini, Y., Donoho, D. and Johnstone, I. (2000). Adapting to unknown sparsity by controlling the false discovery rate. Tech. Rep. 2000–19, Dept. of Statistics, Stanford University, Stanford, CA.

[2] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csàki, eds., *2nd International Symposium on Information Theory*. Akad. Kiàdo, Budapest.

[3] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statist. Soc., Ser. B* **57** 289–300.

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

[5] Dawid, A. P. (1984). Present position and positional developments: some personal views, statistical theory, the prequential approach. *Journal of the Royal Statist. Soc., Ser. A* **147** 278–292.

[6] Dawid, A. P. (1992). Prequential analysis, stochastic complexity and bayesian inference. In J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds., *Bayesian Statistics 4*. Oxford University Press, Oxford.

[7] Donoho, D. (2002). Kolmogorov complexity. Tech. rep., Stanford University, Stanford, CA.

[8] Donoho, D. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the Amer. Statist. Assoc.* **90** 1200–1224.

[9] Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455.

[10] Efron, B. (2001). Selection criteria for scatterplot smoothing. *Annals of Statistics* **29** 470–504.

[11] Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *Annals of Statistics* **22** 1947–1975.

[12] Foster, D. P. and Stine, R. A. (1996). Variable selection via information theory. Tech. Rep. Discussion Paper 1180, Center for Mathematical Studies in Economics and Management Science, Northwestern University, Chicago.

[13] Foster, D. P., Stine, R. A. and Wyner, A. J. (2002). Universal codes for finite sequences of integers drawn from a monotone distribution. *IEEE Trans. on Info. Theory* **48** 1713–1720.

[14] George, E. I. and Foster, D. P. (2000). Calibration and empirical bayes variable selection. *Biometrika* **87** 731–747.

[15] Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *Annals of Statistics* **32** 1594–1649.

[16] Miller, A. J. (2002). *Subset Selection in Regression (Second Edition).* Chapman& Hall, London.

[17] Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *Annals of Statistics* **11** 416–431.

[18] Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *AS* **9** 1135–1151.

[19] Zhang, H.-C. (2005). General empirical Bayes wavelet methods and exactly adaptive minimax estimation. *Annals of Statistics* **33** 54–100.