Submitted to the Annals of Statistics

ON MINIMAX OPTIMALITY OF SPARSE BAYES PREDICTIVE DENSITY ESTIMATES

BY GOURAB MUKHERJEE

gourab@usc.edu

AND

BY IAIN M. JOHNSTONE

imj@stanford.edu

University of Southern California and Stanford University

MAY 5, 2021

We study predictive density estimation under Kullback-Leibler loss in ℓ_0-sparse Gaussian sequence models. We propose proper Bayes predictive density estimates and establish asymptotic minimaxity in sparse models. Fundamental for this is a new risk decomposition for sparse, or spike-and-slab priors.

A surprise is the existence of a phase transition in the future-to-past variance ratio r. For $r < r_0 = (\sqrt{5}-1)/4$, the natural discrete prior ceases to be asymptotically optimal. Instead, for subcritical r, a 'bi-grid' prior with a central region of reduced grid spacing recovers asymptotic minimaxity. This phenomenon seems to have no analog in the otherwise parallel theory of point estimation of a multivariate normal mean under quadratic loss.

For spike-and-uniform slab priors to have any prospect of minimaxity, we show that the sparse parameter space needs also to be magnitude constrained. Within a substantial range of magnitudes, such spike-and-slab priors can attain asymptotic minimaxity.

1. Introduction and main results. Predictive density estimation is a fundamental problem in statistical prediction analysis [1, 9]. Here, it is studied in a high dimensional Gaussian setting under sparsity assumptions on the unknown location parameters. Fuller references and background for the problem are given after a formulation of our main results.

We consider a simple Gaussian model for high dimensional prediction:

$$\boldsymbol{X} \sim N_n(\boldsymbol{\theta}, v_x \boldsymbol{I}), \qquad \boldsymbol{Y} \sim N_n(\boldsymbol{\theta}, v_y \boldsymbol{I}), \qquad \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} | \boldsymbol{\theta}. \tag{1}$$

Our goal is to predict the distribution of a future observation $\boldsymbol{Y}$ on the basis of the 'past' observation vector $\boldsymbol{X}$. In this model, the past and future observations are independent, but are linked by the common mean parameter $\boldsymbol{\theta}$ which is assumed to be unknown. The variances v_x and v_y may differ and are assumed to be known.

The true probability densities of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are denoted by $p(\boldsymbol{x}|\boldsymbol{\theta}, v_x)$ and $p(\boldsymbol{y}|\boldsymbol{\theta}, v_y)$ respectively. We seek estimators $\hat{p}(\boldsymbol{y}|\boldsymbol{x})$ of the future observation density $p(\boldsymbol{y}|\boldsymbol{\theta}, v_y)$, and study their risk properties under sparsity assumptions on $\boldsymbol{\theta}$ as dimension n increases to ∞.

To evaluate the performance of such a predictive density estimator (prde), we use Kullback-Leibler 'distance' as loss function:

$$L(\boldsymbol{\theta}, \hat{p}(\cdot|\boldsymbol{x})) = \int p(\boldsymbol{y}|\boldsymbol{\theta}, v_y) \log \frac{p(\boldsymbol{y}|\boldsymbol{\theta}, v_y)}{\hat{p}(\boldsymbol{y}|\boldsymbol{x})} d\boldsymbol{y}.$$

MSC2020 subject classifications: Primary 62C12; Secondary 62C25, 62F10, 62J07.

Keywords and phrases: Predictive density, Asymptotic Minimaxity, Proper Bayes Rule, Sparsity, High-dimensional, Least Favorable Prior, Spike and Slab.

The corresponding KL risk function follows by averaging over the distribution of the past observation:

$$\rho(\boldsymbol{\theta}, \hat{p}) = \int L(\boldsymbol{\theta}, \hat{p}(\cdot|\boldsymbol{x}))\, p(\boldsymbol{x}|\boldsymbol{\theta}, v_x) d\boldsymbol{x}.$$

Now, given a prior measure $\pi(d\boldsymbol{\theta})$, the average or integrated risk is

$$B(\pi, \hat{p}) = \int \rho(\boldsymbol{\theta}, \hat{p}) \pi(d\boldsymbol{\theta}). \tag{2}$$

For any prior measure $\pi(d\boldsymbol{\theta})$, proper or improper, such that the posterior $\pi(d\boldsymbol{\theta}|\boldsymbol{x})$ is well defined, the Bayes predictive density is given by

$$\hat{p}_\pi(\boldsymbol{y}|\boldsymbol{x}) = \int p(\boldsymbol{y}|\boldsymbol{\theta}, v_y) \pi(d\boldsymbol{\theta}|\boldsymbol{x}). \tag{3}$$

The Bayes predictive density in (3) minimizes both the posterior expected loss $\int L(\boldsymbol{\theta}, \hat{p}(\cdot|\boldsymbol{x}))\pi(d\boldsymbol{\theta}|\boldsymbol{x})$ and the integrated risk $B(\pi, \hat{p})$, when the latter is finite, among all density estimates. The minimum is the Bayes KL risk:

$$B(\pi) := \inf_{\hat{p}} B(\pi, \hat{p})\ . \tag{4}$$

We study the predictive risk $\rho(\boldsymbol{\theta}, \hat{p})$ in a high dimensional setting under an ℓ_0-sparsity condition on the parameter space. This 'exact sparsity' condition has been widely used in statistical estimation problems, e.g. [19, Ch. 8]. With $\|\boldsymbol{\theta}\|_0 = \#\{i : \theta_i \neq 0\}$, consider the parameter set:

$$\Theta_n[s] = \{\boldsymbol{\theta} \in \mathbb{R}^n : \|\boldsymbol{\theta}\|_0 \leq s\}.$$

The minimax KL risk for estimation over Θ is given by

$$R_N(\Theta) = \inf_{\hat{p}} \sup_{\boldsymbol{\theta} \in \Theta}\ \rho(\boldsymbol{\theta}, \hat{p}), \tag{5}$$

the infimum being taken over *all* predictive density estimators $\hat{p}(\boldsymbol{y}|\boldsymbol{x})$. We often write prde for predictive density estimate. The notation $a_n \sim b_n$ denotes $a_n/b_n \to 1$ as $n \to \infty$ and $a_n = O(b_n)$ denotes $|a_n/b_n|$ is bounded for all large n.

1.1. *Main Results.* Henceforth, we assume $v_x = 1$. As the problem is scale equivariant, results for general v_x will easily follow. A key parameter is the future-to-past variance ratio

$$r = v_y/v_x = v_y, \qquad v = (1 + r^{-1})^{-1}. \tag{6}$$

Here v is the 'oracle variance' which would be the variance of the UMVUE for $\boldsymbol{\theta}$, if both $\boldsymbol{X}$ and $\boldsymbol{Y}$ were observed. The variance ratio r determines not only the magnitude of the minimax risk but also the construction of minimax optimal prdes. In our asymptotic model, the dimension $n \to \infty$ and the sparsity $s = s_n$ may depend on n, but the variance ratio r remains fixed.

In the sparse limit $\eta_n = s_n/n \to 0$, for any fixed $r \in (0, \infty)$, Mukherjee and Johnstone [34] evaluated the minimax risk to be:

$$R_N(\Theta_n[s_n]) \sim \frac{1}{1+r} s_n \log(n/s_n) = \frac{1}{1+r} n\eta_n \log \eta_n^{-1}, \tag{7}$$

and a thresholding based prde was shown to attain the minimax risk.

By their nature, thresholding rules are not smooth functions of the data. This paper develops proper Bayes prdes – necessarily smooth functions – that are asymptotically minimax in sparse regimes. Our constructions begin with sparse *univariate* symmetric priors

$$\pi[\eta] = (1 - \eta)\delta_0 + \tfrac{1}{2}\eta(\nu^+ + \nu^-), \tag{8}$$

where δ_0 is unit mass at 0, and $\eta \in [0,1]$ is the sparsity parameter, while ν^+ is a probability measure on $(0,\infty)$ and ν^- is its reflection on $(-\infty,0)$.

For such sparse priors, we introduce a new risk decomposition, Theorem 2.1, that takes the degenerate prior δ_0 as starting point, instead of the more commonly used uniform prior. This risk decomposition is fundamental for all proofs in the paper.

Priors on vector $\boldsymbol{\theta}$ are built from i.i.d. draws:

$$\pi_n(d\boldsymbol{\theta}) = \prod_{i=1}^{n} \pi[\eta_n](d\theta_i), \tag{9}$$

where $\eta_n = s_n/n$ relates the multivariate sparsity s_n to the univariate parameter η_n. The Bayes prde based on prior π_n is the product density estimate:

$$\hat{p}_\pi(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^{n} \hat{p}_\pi(y_i|x_i). \tag{10}$$

The notation often drops the data suffixes and uses $\hat{p}_\pi$ for both the univariate and the multivariate Bayes predictive density when the context is clear.

We begin with a discrete 'grid prior' ν_{G}^+ in which the support points have equal spacing

$$\lambda = \lambda(\eta) = \sqrt{2\log \eta^{-v}}, \tag{11}$$

and geometric mass decay at rate $\eta^v = e^{-\lambda^2/2}$. More precisely,

$$\nu_{\mathsf{G}}^+ = c_{\mathsf{G}} \sum_{j=1}^{\infty} \eta^{(j-1)v} \delta_{\lambda j}, \qquad c_{\mathsf{G}} = 1 - \eta^v.$$

The corresponding sparse grid prior $\pi_{\mathsf{G}}[\eta]$ built via (8) has a schematic illustration in Figure 1. Such "Mallows" discrete priors are a natural starting point for our predictive setting given their optimality properties in point estimation, recalled in the next subsection.

The choice π_{G} can also be motivated directly with three observations. The first, stated precisely in Section 3.5, is that among symmetric univariate three point priors with $\nu^+ = \delta_{a\lambda}, a > 0$, only the choice $a = 1$ is asymptotically least favorable. Second, the convex hull of $\mathrm{supp}(\nu^+)$ must be unbounded, lest the risk function of $\hat{p}_\pi$ grow without bound for large $\boldsymbol{\theta}$. Third, the probability decay rate $\eta^v = \exp(-\lambda^2/2)$ as a function of spacing λ is similar to the geometric decay used in [18] for minimax sparse point-estimation using discrete priors. Among discrete univariate priors, then, the grid prior π_{G} is perhaps the simplest choice compatible with these remarks.

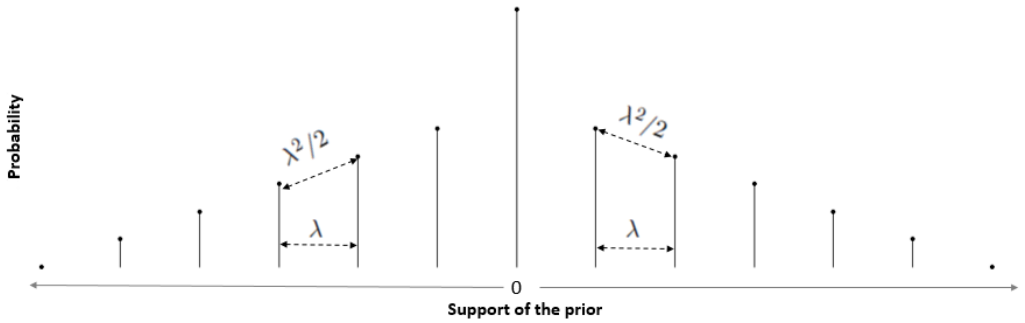

FIG 1. *Schematic for the grid prior. The uniform spacing λ between the support points is shown on the x-axis. The probabilities of the support points are shown on the y-axis using a logarithmic scale, hence the decay appears linear.*

Our first result gives a precise description of the first order asymptotic maximum risk of the Bayes prde $\hat{p}_{\mathsf{G}}$ based on the multivariate product prior $\pi_{\mathsf{G},n}(d\boldsymbol{\theta}) = \prod_{i=1}^{n} \pi_{\mathsf{G}}[\eta_n](\theta_i)$, where $\eta_n = s_n/n$. Define

$$h_r = (1+2r)(1+r)^{-2}(1-2r-4r^2)/4 \;\le\; 1/4 \qquad (12)$$
$$h_r^+ = \max(h_r, 0).$$

Let $r_0 = (\sqrt{5}-1)/4$ be the positive root of the equation $4r^2+2r-1=0$, and note that $h_r^+ > 0$ iff $r < r_0$.

THEOREM 1.1. *As $\eta_n = s_n/n \to 0$, for any fixed $r \in (0,\infty)$ we have*

$$\sup_{\Theta_n[s_n]} \rho(\boldsymbol{\theta}, \hat{p}_{\mathsf{G}}) = R_N(\Theta_n[s_n])\big(1 + h_r^+ + o(1)\big) \text{ as } n \to \infty\,.$$

Thus for all $r \ge r_0$, $\hat{p}_{\mathsf{G}}$ is exactly minimax optimal, while for all $r < r_0$, it is minimax suboptimal but still attains the minimax rate, and has maximum risk at most 1.25 times the minimax value, whatever be the value of r.

As the future-to-past variance ratio r decreases, the difficulty of the predictive density estimation problem increases, as we have to estimate the future observation density based on increasingly noisy past observations. Theorem 1.1 shows that rules which are minimax optimal for higher values of r can be sub-optimal for lower values of r. This phenomenon was seen with threshold density estimates in [34, Sec. S.2, Lemma S.2.1] as well as in the recent work of [29] on non-sparse prediction.

To obtain asymptotic minimaxity for all r, we need to modify the prior. The Bi-grid π_{B} prior is obtained from π_{G} by selecting an 'inner zone' on which the spacing of the prior atoms is reduced from λ to $b\lambda$, where

$$b = \min\{4r(1+r)(1+2r)^{-1}, 1\}. \qquad (13)$$

Note that $b < 1$ iff $r < r_0$. The decay ratio in the inner zone is increased from $\eta^v = e^{-\lambda^2/2}$ to $\eta^{vb^2} = e^{-b^2\lambda^2/2}$. See Figure 2 for a schematic depiction. Section 3.3 explains why the reduced spacing in the inner zone is needed. In brief, the narrower grid 'pulls down' the maximum risk of the Bayes rule for π_{B} to the asymptotically minimax level.

More precisely, $\pi_B[\eta]$ is a univariate sparse symmetric prior of form (8) with

$$\nu_B^+ = c_{\mathsf{B}}\Big[\sum_{k=1}^{K} \eta^{(k-1)vb^2}\delta_{\nu_k} + \eta^{(K-1)vb^2}\sum_{j=1}^{\infty}\eta^{jv}\delta_{\mu_j}\Big].$$

The normalization $c_{\mathsf{B}} = c_{\mathsf{B}}(\eta)$ is at (36). The support points fall in two zones:

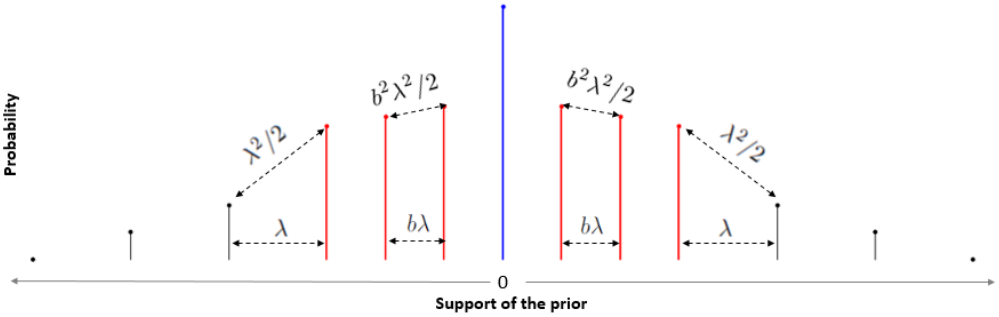

FIG 2. *Schematic for the bi-grid prior. The x-axis now shows the two spacings, and the y-axis (again on a logarithmic scale) the two different rates of log-linear decay of the prior probabilities.*

(i) Inner zone: $\nu_k = \lambda + (k-1)b\lambda$ for $k = 1, \dots, K$
(ii) Outer zone: $\mu_j = \nu_K + j\lambda$ for $j = 1, 2, \dots$

The cardinality of the inner zone is

$$K = 1 + \lceil 2b^{-3/2} \rceil. \tag{14}$$

In fact, any integer $K \in [1 + \lceil 2b^{-3/2} \rceil, \infty]$ works, see Section 3.6. For definiteness, and minimal departure from π_{G}, we use (14).

How do the mass distributions of π_{B} and π_{G} compare? A crude continuous approximation (see supplementary material) says that the 'density ratio' $d\pi_{\mathsf{B}}/d\pi_{\mathsf{G}}(x)$ increases exponentially in x in the inner zone. In the outer zone it is a constant greater than one, i.e. π_{B} has more mass in the tails.

A main result of the paper is that the Bayes predictive density estimate $\hat{p}_{\mathsf{B}}$ based on the product prior $\pi_{\mathsf{B},n}(d\boldsymbol{\theta}) = \prod_{i=1}^{n} \pi_{\mathsf{B}}[\eta_n](d\theta_i)$ is asymptotically minimax optimal.

THEOREM 1.2. *For each fixed $r \in (0, \infty)$, as $\eta_n = s_n/n \to 0$, we have*

$$\sup_{\Theta_n[s_n]} \rho(\boldsymbol{\theta}, \hat{p}_{\mathsf{B}}) = R_N(\Theta_n[s_n])(1 + o(1)) \qquad \text{as } n \to \infty.$$

The following theorem shows that the bi-grid prior $\pi_{\mathsf{B},n}$ is also asymptotically least favorable.

THEOREM 1.3. *If $s_n \to \infty$ and $s_n/n \to 0$, then*

$$B(\pi_{\mathsf{B},n}) = R_N(\Theta_n[s_n]) \cdot (1 + o(1)).$$

Unlike Theorem 1.2 we need the assumption that $s_n \to \infty$. It ensures that $\pi_{\mathsf{B},n}$ actually concentrates on $\Theta_n[s_n]$, namely that $\pi_{\mathsf{B},n}(\Theta_n[s_n]) \to 1$ as $n \to \infty$. For the case where s_n does not diverge to ∞ an asymptotically least favorable prior can be constructed from a sparse prior built from 'independent blocks'. The construction is discussed in Section 3.4.

1.2. *Discussion.* A fully Bayesian approach is a natural route to prdes with good properties [15, 2], with advantages over 'plug-in' or thresholding based density estimates. Indeed, a coordinatewise threshold rule $\hat{p}_{\mathrm{T}}(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^{n} \hat{p}_{\mathrm{T}}(y_i|x_i)$ is typically built from univariate prdes which combine two Bayes prdes – for example based on uniform $\hat{p}_{\mathrm{U}}$ and cluster priors $\hat{p}_{\mathrm{CL}}$, as in [34, Eq. (14)]:

$$\hat{p}_{\mathrm{T}}(y_i|x_i) = \begin{cases} \hat{p}_{\mathrm{U}}(y_i|x_i) & \text{if } |x_i| > v^{-1/2}\lambda \\ \hat{p}_{\mathrm{CL}}(y_i|x_i) & \text{if } |x_i| \le v^{-1/2}\lambda. \end{cases}$$

This is manifestly discontinuous as a function of the data $\boldsymbol{x}$.

The bi-grid Bayes rule achieves the same purposes as the hybrid $\hat{p}_{\mathrm{T}}$. Indeed, the close spacing $b\lambda$ in the inner section of π_{B} yields the same risk control as the (unevenly spaced) cluster prior for small and moderate θ, while the uniform λ spacing in the outer section of π_{B} controls risk for large θ in the same way as the uniform prior.

Decision theoretic parallels between predictive density estimation and the point estimation of a Gaussian mean under quadratic loss have been established by [10, 13, 11, 4, 23, 21, 43, 14] for unconstrained $\boldsymbol{\theta}$, and by [42], [7], [26] and [34] for various constraint sets Θ.

The phase transition seen in Theorems 1.1 and 1.2 seems however to have no parallel in point estimation. Indeed, it follows from [18] that a first order minimax rule for quadratic loss in the sparse setting is derived from the Mallows prior [28], with $\nu_{\mathsf{Q}}^{+} =$

$(1-\eta)\sum_{j=1}^{\infty}\eta^{j-1}\delta_{\lambda_e j}$. Here $\lambda_e = \sqrt{2\log\eta^{-1}} = v^{-1/2}\lambda$ so that the predictive setting involves a reduced spacing in the prior. More significantly, there is no analog in point estimation of the inner section with its further reduced spacing for $r < r_0$.

Our main technical contribution lies in sharp methods for bounding the global KL risk for general bi-grid priors, see Lemmas 3.1 and 3.2, and for spike-and-slab priors, Section 4. The sharp predictive risk bounds established here provide new asymptotic perspectives in the information geometric framework of [22, 24, 44] and augment new sparse prediction techniques for general multivariate predictive density estimation theory developed in [10, 4, 23, 25, 27, 30, 32].

1.3. *Minimax risk of Spike and Slab priors.* Some of the most popular Bayesian variable selection techniques are built on the "spike and slab" priors [31, 12, 16]. Such priors and their computationally tractable extensions have found success in variable selection in high-dimensional sparse regression models, e.g. [37, 3, 40, 38, 39, 17] and the references therein. While this is a well established methodological research area [36], optimality of their respective predictive density estimates has so far not been studied.

Here, we consider simple "spike and slab" prior distributions in the flavor of the foundational paper [31]. Begin with a sparse univariate prior, a special case of (8),

$$\pi_{\mathsf{S}}[\eta,\ell] = (1-\eta)\delta_0 + \eta/(2\ell)I\{\mu\in[-\ell,\ell]\}\,d\mu\,. \tag{15}$$

In parallel with (9), build a multivariate product prior $\pi_{\mathsf{S},n}$ from n i.i.d. copies of $\pi_{\mathsf{S}}[\eta_n,\ell]$, where as before $\eta_n = s_n/n$. We might consider multivariate Bayes predictive densities $\hat{p}_S[\ell]$ based on $\pi_{\mathsf{S},n}$.

It is intuitively clear that such Bayes prdes are necessarily asymptotically sub-optimal: for any fixed $\ell\in[0,\infty)$, for all $s_n > 0$, we have

$$\left\{\sup_{\Theta_n[s_n]}\rho(\boldsymbol{\theta},\hat{p}_S[\ell])\right\}\Big/ R_N(\Theta_n[s_n]) = \infty \tag{16}$$

for each fixed n. Indeed, the support of π_{S} is restricted to $[-\ell,\ell]$, and the corresponding prde has large risk away from the support. A formal proof follows Lemma 4.1.

Consider therefore bounded subsets of the sparse parameter sets $\Theta_n[s_n]$:

$$\Theta_n[s,t] = \{\boldsymbol{\theta}\in\mathbb{R}^n : \|\boldsymbol{\theta}\|_0 \le s \text{ and } |\theta_i|\le t \text{ for all } i = 1,\dots,n\}.$$

We allow $t = t_n$ to increase with n, and note next that the increase must be at least as fast as $\lambda_n = \lambda(\eta_n)$, cf. (11), to have minimax risk equivalent to $\Theta_n[s_n]$.

LEMMA 1.4. *For all t_n there is a simple bound*

$$R_N(\Theta_n[s_n,t_n]) \le s_n t_n^2/(2r).$$

If $t_n > \lambda_n = \sqrt{2\log\eta_n^{-v}}$, then

$$R_N(\Theta_n[s_n,t_n]) \sim s_n\lambda_n^2/(2r) \sim R_N(\Theta_n[s_n]). \tag{17}$$

The following result exhibits a substantial range of magnitude constraints t_n for which $\hat{p}_{\mathsf{S}}[t_n]$ is asymptotically minimax over $\Theta_n[s_n,t_n]$. All proofs for this subsection, along with a figure and high-level overview of the strategy, appear in Section 4.

THEOREM 1.5. *As $\eta_n = s_n/n \to 0$, suppose that $t_n/(\log\eta_n^{-1})^{1/2}\to\infty$ but $\log t_n/(\log\eta_n^{-1})\to 0$. Then as $n\to\infty$,*

$$\sup_{\Theta_n[s_n,t_n]}\rho(\boldsymbol{\theta},\hat{p}_{\mathsf{S}}[t_n]) \sim R_N(\Theta_n[s_n,t_n]).$$

Note that if $t_n \to \infty$ at a rate slower than $(\log \eta_n^{-1})^{1/2}$ then, by Lemma 1.4, $R_N(\Theta_n[s_n, t_n])$ is no longer equivalent to $R_N(\Theta_n[s_n])$ as $n \to \infty$. At the other extreme, we show next that if t_n grows at rate $\eta_n^{-\beta}$ or higher for any $\beta > 0$, then no spike and uniform slab procedure can be minimax optimal.

THEOREM 1.6. *If $\eta_n = s_n/n \to 0$ and $\log t_n = \beta \log \eta_n^{-1}$ for some $\beta > 0$, then*

$$\min_{\ell > 1} \sup_{\Theta[s_n, t_n]} \rho(\boldsymbol{\theta}, \hat{p}_S[\ell]) \geq (1+\beta) R_N(\Theta_n[s_n, t_n])(1 + o(1)) \text{ as } n \to \infty.$$

We emphasize that Theorem 1.6 shows that, even for true parameters within the support of the uniform slab, risk can exceed the minimax bound. Informally, the proof shows that if the slab is small, $\log \ell < \beta \lambda_n^2$, then the risk at $\theta = t_n$ is unacceptably large, while if it is large, $\log \ell \geq \beta \lambda_n^2$, there is poor risk at $\theta = \sqrt{1+\beta}\lambda_n$.

1.4. *Organization of the Paper.* Section 2 presents the fundamental risk decomposition, its proof and some discussion. Section 3 presents the risk properties of the Grid and Bi-grid prior based prdes and proofs of the main results. Section 4 proves the spike-and-slab results. Section 5 compares the performance of the prdes through simulation experiments. The Appendix and Sections 1, 2 of the supplement contain the proofs of the lemmas.

Notations. The standard normal density and cumulative distribution are denoted by ϕ and Φ; $\tilde{\Phi} = 1 - \Phi$. For sequences $a_n \sim b_n$ means $\lim_{n\to\infty} a_n/b_n = 1$.

2. A risk decomposition for spike and slab priors. *Univariate problem.* We focus on priors with i.i.d. components (9), so that the predictive density then has product form (10). The predictive risk is then additive

$$\rho(\boldsymbol{\theta}, \hat{p}_\pi) = \sum_{i=1}^{n} \rho(\theta_i, \hat{p}_\pi). \tag{18}$$

[We use $\hat{p}_\pi$ for both univariate and multivariate Bayes predictive densities: the context will make clear which is used.]

For our sparse parameter sets $\Theta_n[s]$ and $\Theta_n[s,t]$, there is an easy reduction of the maximum multivariate risk of a product rule to a univariate risk maximum. Indeed, (18) yields

$$s_n \sup_{|\theta| \leq t_n} \rho(\theta, \hat{p}) \leq \sup_{\Theta_n[s_n,t_n]} \rho(\boldsymbol{\theta}, \hat{p}) \leq n(1-\eta_n)\rho(0, \hat{p}) + s_n \sup_{\theta \in \mathbb{R}} \rho(\theta, \hat{p}). \tag{19}$$

Sparse priors. Now suppose that $X|\theta \sim N(\theta, 1)$ and $Y|\theta \sim N(\theta, r)$ and that the past and future observations X, Y are independent given θ. Consider a sparse proper prior of the form

$$\pi(d\mu) = (1-\eta)\delta_0 + \eta\nu(d\mu), \tag{20}$$

for probability measure ν on $\mathbb{R}$ and $\eta \in [0,1]$. The associated (univariate) Bayes predictive density estimate $\hat{p}_\pi$ is given by (3).

The following risk decomposition is fundamental. It will be applied to study discrete priors in Section 3 and uniform slab priors in Section 4.

THEOREM 2.1. *With the preceding definitions, let $Z \sim \mathcal{N}(0,1)$ and $v = (1 + r^{-1})^{-1}$. For a sparse prior (20),*

$$\begin{aligned} \rho(\theta, \hat{p}_\pi) &= \rho(\theta, \hat{p}_{\delta_0}) - \mathbb{E} \log N_{\theta,v}(Z) + \mathbb{E} \log D_\theta(Z), \\ &= \theta^2/(2r) - \mathbb{E} \log N_{\theta,v}(Z) + \mathbb{E} \log D_\theta(Z), \end{aligned} \tag{21}$$

where $D_\theta(Z) = N_{\theta,1}(Z)$ *and*

$$N_{\theta,v}(Z) = 1 + \frac{\eta}{1-\eta}\int \exp\left\{\frac{\mu Z}{\sqrt{v}} + \frac{\mu\theta}{v} - \frac{\mu^2}{2v}\right\}\nu(d\mu)\,. \tag{22}$$

Decomposition (21) takes the degenerate prior δ_0 as starting point for comparison of the risk $\rho(\theta,\hat{p}_\pi)$ of a Bayes prde. This is natural for sparse priors (20) and might be contrasted with the representation George, Liang and Xu [10, Lemma 2], which takes the uniform prior prde as point of departure.

PROOF. The decomposition (21) compares $\rho(\theta,\hat{p}_\pi)$ to $\rho(\theta,\hat{p}_{\delta_0}) = \theta^2/(2r)$, the Bayes risk of $\hat{p}_{\delta_0}(y|x) = \phi(y|0,r)$ corresponding to $\pi = \delta_0$ and $\eta = 0$. Accordingly, using (3), write the Bayes predictive density as

$$\hat{p}_\pi(y|x) = \frac{\int \phi(y|\mu,r)\phi(x-\mu)\pi(d\mu)}{\int \phi(x-\mu)\pi(d\mu)} = \phi(y|0,r)\frac{N(x,y)}{D(x)}, \tag{23}$$

after rewriting numerator and denominator in the first ratio respectively as

$$\pi_0\phi(y|0,r)\phi(x)N(x,y), \qquad \text{and} \qquad \pi_0\phi(x)D(x),$$

where $\pi_0 = \pi(\{0\}) = 1-\eta$. After simple algebra, we find

$$N(x,y) = \int \exp\left\{\mu\left(x+\frac{y}{r}\right) - \frac{\mu^2}{2}\left(1+\frac{1}{r}\right)\right\}\frac{\pi(d\mu)}{\pi_0} \tag{24}$$

and $D(x)$ is analogous, but without terms in y and r. Note also that

$$\rho(\theta,\hat{p}_{\delta_0}) = \mathbb{E}_\theta \log\left(\frac{\phi(Y|\theta,r)}{\phi(Y|0,r)}\right) = \mathbb{E}_\theta\left[\frac{\theta Y}{r} - \frac{\theta^2}{2r}\right] = \frac{\theta^2}{2r}.$$

Hence, from (23) and the definition of predictive loss

$$L(\theta,\hat{p}_\pi(\cdot|x)) = \mathbb{E}_\theta \log\left(\frac{\phi(Y|\theta,r)}{\hat{p}_\pi(Y|x)}\right) = \frac{\theta^2}{2r} - \mathbb{E}_\theta \log N(x,Y) + \log D(x).$$

To obtain $\rho(\theta,\hat{p}_\pi)$, take expectation also over $X \sim N(\theta,1)$. Since $Y \sim N(\theta,r)$ independently of X, the random variable $X + Y/r \sim \mathcal{N}(\theta/v, 1/v)$ may be expressed in the form $\theta/v + Z/\sqrt{v}$. Recalling the sparse prior form $\pi(d\mu) = (1-\eta)\delta_0 + \eta\nu$, we get

$$N(X,Y) \overset{\mathcal{D}}{=} 1 + \frac{\eta}{1-\eta}\int \exp\left\{\frac{\mu Z}{\sqrt{v}} + \frac{\mu\theta}{v} - \frac{\mu^2}{2v}\right\}\nu(d\mu) = N_{\theta,v}(Z).$$

Similarly, $D(X) \overset{\mathcal{D}}{=} D_\theta(Z)$ and the lemma follows from the previous two displays. □

Clearly $N_{\theta,v}(Z), D_\theta(Z) \geq 1$, and so we have the simple but useful "basic lower" and "basic upper" risk bounds

$$\frac{\theta^2}{2r} - \mathbb{E}\log N_{\theta,v}(Z) \leq \rho(\theta,\hat{p}_\pi) \leq \frac{\theta^2}{2r} + \mathbb{E}\log D_\theta(Z). \tag{25}$$

From Jensen's inequality,

$$\mathbb{E}\log N_{\theta,v}(Z) \leq \log\big(\mathbb{E}N_{\theta,v}(Z)\big), \tag{26}$$

and since $\mathbb{E}\exp(\zeta Z) = \exp(\zeta^2/2)$, by Fubini's theorem

$$\mathbb{E}N_{\theta,v}(Z) = 1 + \frac{\eta}{1-\eta}\int \exp\left(\frac{\mu\theta}{v}\right)\nu(d\mu), \tag{27}$$

and, in particular,

$$\mathbb{E}D_0(Z) = \mathbb{E}N_{0,1}(Z) = (1-\eta)^{-1}.$$

Consequently, from the right side of (25), then (26) (for $v = 1$) and the previous display,

$$\rho(0,\hat{p}_\pi) \le \log(1-\eta)^{-1} = \eta(1+o(1)) \quad \text{as} \quad \eta \to 0. \tag{28}$$

3. Risk properties for discrete priors. The bulk of this section is devoted to the proof of Theorems 1.1 and 1.2. We first outline the approach. First, return to the univariate reduction (19). From (28) it is clear that $n\rho(0,\hat{p}_1) \le n\eta_n(1+o(1)) = s_n(1+o(1))$. So for the minimaxity results of Theorems 1.1 (for $r > r_0$), 1.2 and 1.5, it suffices to show the univariate bound

$$\sup_{\theta\in\mathbb{R}} \rho(\theta,\hat{p}_1) \le \lambda_n^2/(2r) + o(\lambda_n^2), \tag{29}$$

for then, with $\Theta_n = \Theta_n[s_n]$ or $\Theta_n[s_n,t_n]$,

$$\sup_{\Theta_n} \rho(\boldsymbol{\theta},\hat{p}_{\pi_n}) \le s_n[\lambda_n^2/(2r) + o(\lambda_n^2)].$$

To establish (29), we use the key risk decomposition of Proposition 2.1. For this we introduce a class of discrete sparse priors that includes both grid and bi-grid priors. We develop lower and upper bounds respectively for $\mathbb{E}\log N_{\theta,v}(Z)$ and $\mathbb{E}\log D_\theta(Z)$ in (21). These bounds are combined to yield an upper estimate

$$\rho(\theta,\hat{p}_{\mathsf{D}}) \le (2r)^{-1}\lambda^2\sigma(\theta) + O(\lambda),$$

for some function σ. In Section 3.3 we first provide heuristics—Figure 3—and then a formal proof of conditions under which $\sigma(\theta) \le 1$ for all θ, establishing Theorems 1.1 and 1.2.

3.1. *A class of discrete sparse priors.* For $0 < b \le 1$ and $r > 0$, let

$$\pi_{\mathsf{D}}[\eta;b,r] = \sum_{j\in\mathbb{Z}} \pi_j\delta_{\mu_j} \tag{30}$$

where $\mu_{-j} = -\mu_j$, $\pi_{-j} = \pi_j$. The support points satisfy $\mu_0 = 0$ and $\mu_j = \lambda\alpha_j$ for $j > 0$, where the piecewise linear spacing function

$$\alpha_j = \begin{cases} 1 + b(j-1) & 1 \le j \le K \\ \alpha_K + j - K & j > K \end{cases} \tag{31}$$

has increments $\dot{\alpha}_j = \alpha_{j+1} - \alpha_j = b$ or 1 according as $j \le K$ or $j > K$. Set $\zeta = \eta^v$. The prior masses are given by

$$\pi_0 = 1-\eta, \qquad \pi_j = c(\eta)\eta\zeta^{\beta_j - 1}, \tag{32}$$

for $j \ge 1$. The decay function in the prior probabilities

$$\beta_j = \begin{cases} 1 + b^2(j-1) & 1 \le j \le K \\ \beta_K + j - K & j > K \end{cases} \tag{33}$$

has the same form as α_j with b replaced by b^2. This choice is crucial for Lemma 3.1 below and its consequent risk bounds. In particular, note that $\beta_j \le \alpha_j$ and that the increments $\dot\beta_j = \beta_{j+1} - \beta_j$ satisfy

$$\dot\beta_j = \dot\alpha_j^2 \qquad \text{all } j \ge 1. \tag{34}$$

In addition, $l \to g_l = \alpha_l^2 - \beta_l$ is increasing for $l \ge 1$, as

$$g_{l+1} - g_l = \dot\alpha_l(\alpha_{l+1} + \alpha_l) - \dot\beta_l = 2\dot\alpha_l\alpha_l > 0, \tag{35}$$

The normalizing constant $c(\eta) = c_{\mathsf{B}}(\eta)/2$, where,

$$\frac{1}{c_{\mathsf{B}}(\eta)} = \sum_{j=1}^{\infty} \zeta^{\beta_j - 1} = \frac{1-\eta^{b^2vK}}{1-\eta^{b^2v}} + \frac{\eta^{b^2v(K-1)+v}}{1-\eta^v}. \tag{36}$$

3.2. *Risk Component Bounds for discrete priors.* Since π_{D} is a sparse prior, we may apply the decomposition of predictive risk given in Proposition 2.1. Inserting the discrete measure (30), we obtain

$$N_{\theta,v}(Z) = 1 + \sum_{j\ne 0} N_j, \tag{37}$$

$$N_j = \pi_0^{-1}\pi_j \exp\{v^{-1/2}\mu_j Z + v^{-1}(\mu_j\theta - \tfrac{1}{2}\mu_j^2)\} \tag{38}$$

In the special case $v = 1$, it will be helpful to write $D_\theta(Z) = N_{\theta,1}(Z)$ as

$$D_\theta(Z) = 1 + \sum_{j\ne 0} D_j, \tag{39}$$

$$D_j = \pi_0^{-1}\pi_j \exp\{\mu_j Z + \mu_j\theta - \tfrac{1}{2}\mu_j^2\}. \tag{40}$$

The probability ratio π_j/π_0 can also be written in exponential form. To this end, introduce $c_1(\eta) = c(\eta)(1-\eta)^{-1}$. Recall that $v^{-1} = 1 + r^{-1}$ and $\zeta = \eta^v = \exp(-\lambda^2/2)$ and then rewrite $\eta = \zeta^{v^{-1}} = \exp\{-\frac{1}{2}\lambda^2(1+r^{-1})\}$. Using (32), we arrive at

$$\pi_0^{-1}\pi_j = c_1(\eta)\exp\{-\tfrac{1}{2}\lambda^2(\beta_j + r^{-1})\}. \tag{41}$$

We can therefore, for example, rewrite

$$\begin{aligned} D_j &= c_1(\eta)\exp\{\mu_j Z - G(\mu_j;\theta)\} \\ G(\mu_j;\theta) &= \tfrac{1}{2}\mu_j^2 - \mu_j\theta + \tfrac{1}{2}\lambda^2(\beta_j + r^{-1}). \end{aligned} \tag{42}$$

To obtain an upper bound for $\rho(\theta, \hat p_{\mathsf{D}})$ we use (22). It turns out to be enough to focus on (the logs of) two consecutive terms N_j, N_{j+1} in (37); ignoring all other terms trivially yields a lower bound for $N_{\theta,v}$. For the upper bound for D_θ, a single (suitably chosen) term D_j in (39) suffices, but more care is needed to show that the neglected terms are negligible.

Bring in a co-ordinate system (l,ω) for θ: each $\theta \ge 0$ can be uniquely written in the form

$$\theta = \lambda(\alpha_l + \omega), \qquad l \in \mathbb{N},\ \omega \in [0, \dot\alpha_l).$$

We can therefore write $l = l(\theta)$ and $\omega = \omega(\theta)$.

We argue heuristically that $l(\theta)$ is an appropriate choice of index for our bounds. Indeed, from (38) and (41),

$$\mathbb{E}\log N_j = c - \tfrac{1}{2}\{(\mu_j - \theta)^2/v - \lambda^2\beta_j\} \tag{43}$$

after collecting terms not involving j into c. Hence, for $\theta \in [\mu_l, \mu_{l+1})$, the choice $j = l$ or $l+1$ will minimize or nearly minimize the quadratic, and these suffice for the lower bound. For D_θ, we have from (42) that $\mathbb{E}\log D_j = \log c_1(\eta) - G(\mu_j;\theta)$. We show in the Appendix (in the proof of Lemma 3.1) that $j \to G(\mu_j;\theta)$ is indeed minimized at $j = l$ for *each* $\theta \in [\mu_l, \mu_{l+1})$.

Focus therefore on the terms $N_{l(\theta)}$ and $D_{l(\theta)}$. When $\theta = \lambda(\alpha_l + \omega)$,

$$\mu_j\theta - \tfrac{1}{2}\mu_j^2 = \tfrac{1}{2}\lambda^2(2\alpha_j(\alpha_l+\omega) - \alpha_j^2).$$

Combining this with (41), for $j = l, l+1$, we can write

$$\begin{aligned} N_l &= c_1(\eta)\exp\{\tfrac{1}{2}\lambda^2 n(l,\omega) + \alpha_l\lambda Z/\sqrt{v}\} \\ N_{l+1} &= c_1(\eta)\exp\{\tfrac{1}{2}\lambda^2 \check{n}(l,\omega) + \alpha_{l+1}\lambda Z/\sqrt{v}\} \\ D_l &= c_1(\eta)\exp\{\tfrac{1}{2}\lambda^2 d(l,\omega) + \alpha_l\lambda Z\} \end{aligned} \tag{44}$$

in terms of three linear functions of ω:

$$\begin{aligned} n(l,\omega) &= v^{-1}(\alpha_l^2 + 2\alpha_l\omega) - \beta_l - r^{-1} \\ d(l,\omega) &= \qquad \alpha_l^2 + 2\alpha_l\omega - \beta_l - r^{-1}. \end{aligned} \tag{45}$$

and, corresponding to N_{l+1},

$$\check{n}(l,\omega) = n(l,\omega) + 2v^{-1}\dot{\alpha}_l\omega - (1+v^{-1})\dot{\alpha}_l^2. \tag{46}$$

We now state our key uniform bounds on the risk components of (21).

LEMMA 3.1. *For any fixed $r \in (0,\infty)$ and $b \in (0,1]$, with λ defined in (11), uniformly in $\theta = \lambda(\alpha_l + \omega) \geq \lambda$, we have the following bounds:*

$$\begin{aligned} \mathbb{E}\log N_{\theta,v}(Z) &\geq \tfrac{1}{2}\lambda^2(n \vee \check{n})(l,\omega) + O(1), \\ \mathbb{E}\log D_\theta(Z) &\leq \tfrac{1}{2}\lambda^2 d^+(l,\omega) + O(\lambda), \end{aligned}$$

For $0 \leq \theta < \lambda$ we just have $\mathbb{E}\log N_{\theta,v}(Z) \geq 0$, and $\mathbb{E}\log D_\theta(Z) \leq O(\lambda)$.

The proof is given in the appendix. The appearance of the positive part of $d(l,\omega)$ in the upper bound may be understood this way: if $d(l,\omega) < 0$, we cannot expect the term D_l to dominate $D_0 = 1$ in (39).

In the reverse direction, we need only a bound for θ lying in a subset of $[\mu_1, \mu_2)$ in our proofs of theorems 1.1 and 1.2.

LEMMA 3.2. *For any fixed $r \in (0,\infty)$, $b \in (0,1]$, with λ defined in (11) and setting $\omega_1 = b(1+v)/2$, uniformly in $\theta \in \lambda[\alpha_1, \alpha_1 + \omega_1]$, we have*

$$\mathbb{E}\log N_{\theta,v}(Z) \leq \tfrac{1}{2}\lambda^2 n(1,\omega) + O(\lambda).$$

3.3. *Proof of Theorems 1.1 and 1.2.* Inserting the bounds of Lemma 3.1 in risk decomposition (21), we get

$$\begin{aligned} \rho(\theta, \hat{p}_{\mathrm{D}}) &\leq (2r)^{-1}\lambda^2\sigma(l,\omega) + O(\lambda) \\ \sigma(l,\omega) &= \begin{cases} \omega^2 & \text{if } l = 0, \\ (\alpha_l+\omega)^2 - r(n \vee \check{n})(l,\omega) + rd^+(l,\omega) & \text{if } l \geq 1. \end{cases} \end{aligned} \tag{47}$$

Our task is to bound $\sigma(l,\omega)$; more specifically for Theorem 1.1 to show that $\sigma(l,\omega) \leq 1 + h_r^+$ for the grid prior and for Theorem 1.2 to ensure that $\sigma(l,\omega) \leq 1$ for the bigrid prior with b in (13). Figure 3 shows the idea of the main part of the proof. We argue below that the

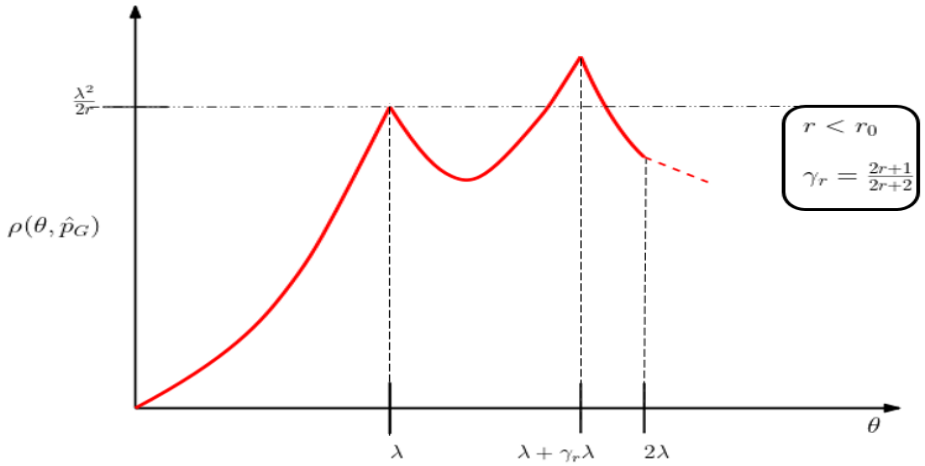

FIG 3. *Schematic for the risk bound (47) for $\theta \to \rho(\theta, \hat{p}_G)$ for the grid prior; being asymptotically minimax when the second peak is no higher than the first.*

maximum of $\sigma(\theta) = \sigma(l, \omega)$ falls in the interval $[\lambda, \lambda\alpha_2]$, which in the case of the uniform grid prior is just $[\lambda, 2\lambda]$. The function $\sigma(\theta)$ is argued to be piecewise quadratic with

$$\max_{\lambda \le \theta \le 2\lambda} \sigma(\theta) = \max\{1, 1 + \gamma_r(\gamma_r - 2r)\}.$$

The second maximum is attained at $\theta_* = \lambda + \gamma_r\lambda$, with $\gamma_r = (2r+1)/(2r+2)$. It will then follow that the grid prior estimate $\hat{p}_G$ is asymptotically minimax if and only if $\gamma_r \le 2r$, which translates to $r \ge r_0 = (\sqrt{5} - 1)/4$.

For $r < r_0$, the maximum risk can be 'pulled down' by reducing the spacing between λ and the next support point $\lambda + b\lambda$ (we set $\dot{\alpha}_l = b$). For the bi-grid prior, the second maximum then satisfies

$$\theta_* = \lambda + \gamma_r b\lambda, \qquad \sigma(\theta_*) = 1 + \gamma_r b(\gamma_r b - 2r) \le 1$$

exactly when b is no larger than the value (13).

To begin the proof, observe first that by symmetry we can reduce to $\theta \ge 0$. For $l = 0$, control on the risk is immediate from (25), and so, from now on consider $l \ge 1$. We make some observations on $\sigma(l, \omega)$. When $d(l, \omega) \ge 0$, from (45), $r(n - d) = \alpha_l^2 + 2\alpha_l\omega$, and so

$$\sigma(l, \omega) \le (\alpha_l + \omega)^2 - r(n - d) = \omega^2 \le 1. \tag{48}$$

Back in the general case, from (46), both $n(l, \omega)$ and $\check{n}(l, \omega)$ are linear for $\omega \in [0, \dot{\alpha}_l]$, intersecting at $\omega_l = \dot{\alpha}_l(1 + v)/2 < \dot{\alpha}_l$. Now $n(l, 0) > \check{n}(l, 0)$ while $\check{n}$ has a larger positive slope. Hence $n \vee \check{n}$ equals n on $[0, \omega_l]$ and $\check{n}$ on $[\omega_l, \dot{\alpha}_l]$. A calculation shows that

$$\begin{aligned} \check{n}(l, \dot{\alpha}_l) &= v^{-1}(\alpha_l^2 + 2\alpha_l\dot{\alpha}_l + \dot{\alpha}_l^2) - \dot{\alpha}_l^2 - \beta_l - r^{-1} \\ &= v^{-1}\alpha_{l+1}^2 - \beta_{l+1} - r^{-1} = n(l+1, 0). \end{aligned}$$

Similarly $d(l, \dot{\alpha}_l) = d(l+1, 0)$ and so $\theta \to d(\theta)$ is piecewise linear, continuous and strictly increasing from $d(1, 0) = -r^{-1} < 0$ to $+\infty$ as $\theta \to \infty$. Consequently there is a unique $\theta_* = (l_*, \omega_*)$ at which $d(\theta_*) = 0$.

From these remarks, it follows that $\sigma(l, \omega)$ is piecewise quadratic and convex for $\omega \in [\alpha_l, \alpha_{l+1}]$. Hence its maxima can only occur among the join points $\omega = 0, \omega_l, 1$ and ω_* in the

single case $l = l_*$. However, since $d(\theta_*) = 0$, it follows from (48) that $\sigma(l_*, \omega_*) \le 1$, so we can safely ignore this case. Consequently, from (47) and noting that $\sigma(0, 1-) = 1$, we have

$$\|\sigma\|_\infty := \max_{\theta \ge 0, \theta = (l,\omega)} \sigma(l,\omega) = 1 \vee \max_{l \ge 1} \{\sigma(l,0), \sigma(l,\omega_l), \sigma(l,\dot\alpha_l)\}. \tag{49}$$

Now suppose that $0 \le \omega \le \omega_l$ and that $d(l,\omega) \le 0$. In this case, since $n \le \check n$ and $1 - rv^{-1} = -r$, we have

$$\begin{aligned} \sigma(l,\omega) &= (\alpha_l + \omega)^2 - rn(l,\omega) \\ &= \omega^2 + (\alpha_l^2 + 2\alpha_l\omega)(1 - rv^{-1}) + r\beta_l + 1 \\ &= 1 + \omega^2 + r(\beta_l - \alpha_l^2) - 2r\alpha_l\omega \\ &\le 1 + \omega^2 - 2r\omega \end{aligned} \tag{50}$$

say, where we used $\alpha_l \ge 1$ and $\alpha_l^2 - \beta_l \ge \alpha_1^2 - \beta_1 = 0$, from (35).

In particular $\sigma(l,0) \le 1$, and combining with (48), this holds for *all* l. Also, for $l \in \mathcal{L} = \{l : d(l,\omega_l) < 0\}$, we have $\sigma(l,\omega_l) \le 1 + \omega_l(\omega_l - 2r)$, while for $l \notin \mathcal{L}$, again from (48), $\sigma(l,\omega_l) \le 1$. Now, (49) simplifies to

$$\|\sigma\|_\infty \le 1 + \max_{l \in \mathcal{L}} \omega_l(\omega_l - 2r)_+. \tag{51}$$

For the grid prior, $b = 1$. We have $\omega_l = (1+v)/2 = (2r+1)/(2r+2)$, and

$$\omega_l(\omega_l - 2r) = (1+2r)(1+r)^{-2}(1 - 2r - 4r^2)/4 = h_r\ .$$

and we have established the upper bound in Theorem 1.1.

For the lower bound it suffices to look at the risk at a single point. In view of Figure 3 and the discussion preceding (49), we try $\theta_1 = \lambda(1+\omega_1)$. Look at the risk at θ_1. Apply Lemma 3.2 using $n(1,\omega_1) = 2v^{-1}\omega_1$, to get from (25)

$$\rho(\theta_1, \hat p_{\mathsf{G}}) \ge (2r)^{-1}\lambda^2\{(1+\omega_1)^2 - 2rv^{-1}\omega_1\} + O(\lambda) = (2r)^{-1}\lambda^2(1 + h_r^+) + O(\lambda),$$

since the quantity in braces equals $1 + \omega_1^2 - 2r\omega_1 = \sigma(1,\omega_1) = 1 + h_r^+$. This completes the proof of Theorem 1.1.

We now turn our attention to proving Theorem 1.2. We first verify that if $b \le \min\{1, 4r\}$, then $l \ge 1 + \lceil 2b^{-3/2} \rceil$ necessarily implies $d(l,\omega_l) \ge 0$.

From the monotonicity (35), along with $\omega \ge 0$, we have

$$d(l,\omega_l) + r^{-1} \ge \alpha_l^2 - \beta_l \ge \alpha_K^2 - \beta_K \ge b^2(K-1)^2 \ge 4b^{-1} \ge r^{-1},$$

using $b \le 2$ for the third, $K \ge 1 + \lceil 2b^{-3/2} \rceil$ for the fourth and $b \le 4r$ for the fifth inequalities.

Now, return to (51): if $l \in \mathcal{L}$ then $d(l,\omega_l) < 0$ and so, from the previous paragraph necessarily $l < 1 + \lceil 2b^{-3/2} \rceil$ which by definition entails $\dot\alpha_l = b$ so long as $K \ge 1 + \lceil 2b^{-3/2} \rceil$. Now $\omega_l = b(1+v)/2 \le 2r$ is equivalent to $b \le 4r/(1+v)$. So, in this case, $\|\sigma\| = 1$, and so for *all* θ we have $\rho(\theta, \hat p_{\mathsf{D}}) \le (2r)^{-1}\lambda^2 + O(\lambda)$, which establishes (29) and hence Theorem 1.2.

3.4. *Proof of Theorem 1.3.* By Theorem 1.2 it suffice to prove a lower bound on the Bayes risk. As $\pi_{\mathsf{B},n}$ is i.i.d. and due to the product structure of the problem, its Bayes risk simplifies

$$B(\pi_{\mathsf{B},n}, \hat p_{\mathsf{B}}) = nB(\pi_{\mathsf{B}}, \hat p_{\mathsf{B}}).$$

For the univariate problem the Bayes risk of the prior π_{B} is

$$\begin{aligned} B(\pi_{\mathsf{B}}, \hat p_{\mathsf{B}}) &\ge \eta_n\, c(\eta_n)\, \{\rho(\lambda_n, \hat p_{\mathsf{B}}) + \rho(-\lambda_n, \hat p_{\mathsf{B}})\} \\ &= 2\eta_n\, c(\eta_n)\, \rho(\lambda_n, \hat p_{\mathsf{B}}) \ge 2\eta_n\, c(\eta_n)\, [\lambda_n^2/(2r) - \mathbb{E} \log N_{\lambda_n, v}(Z)], \end{aligned}$$

where the equality above follows by symmetry and the inequality by (25). From (36) we have $2c(\eta_n) = c_{\mathsf{B}}(\eta) \geq 1 - O(\eta_n^{b^2 v})$. Lemma 3.2 shows that $\mathbb{E} \log N_{\lambda,v}(Z) = O(\lambda)$ because $n(1,0)$ [defined in (45)] equals 0. Hence $B(\pi_{\mathsf{B}}, \hat{p}_{\mathsf{B}}) \geq \eta_n \lambda_n^2/(2r) \cdot (1 + o(1))$ and the proof is done.

3.5. *Three point priors.* Let $\pi_a = \pi_a[\eta]$ be a sparse symmetric three point prior given by (8) with $\nu^+ = \delta_{a\lambda}$ for $a > 0$. In Section 7, we prove:

LEMMA 3.3. *Let $\hat{p}_a$ be the prde corresponding to π_a. Then, as $\eta \to 0$,*

$$\rho(a\lambda, \hat{p}_a) \leq (2r)^{-1} \lambda^2 \tau(a) + O(\lambda) \tag{52}$$

$$\tau(a) = \begin{cases} a^2 & a^2 \leq 1, \\ [1 - r(a^2 - 1)]_+ & a^2 \geq 1. \end{cases} \tag{53}$$

In particular, as $\eta \to 0$, the prior π_a is least favorable only when $a = 1$:

$$B(\pi_a, \hat{p}_a) \sim \eta \rho(a\lambda, \hat{p}_a) \sim (2r)^{-1} \eta \lambda^2 \tau(a). \tag{54}$$

3.6. *Remarks.* 1. When $K = \infty$, the bigrid prior π_{B} has support points (in $\mathbb{R}^+$) separated by $(1, b, b, \ldots)$. We denote this special case $\pi_{\mathsf{B}'}$, and we emphasize that it is still a bigrid prior (unless $b = 1$), though it may be seen as simpler than π_{B}. The proof of Theorem 1.2 shows that with $b = 4r(1+r)(1+2r)^{-1}$, prior $\pi_{\mathsf{B}'}$ is asymptotically minimax for $r \leq r_0$.

However, there is no choice of b for which $\pi_{\mathsf{B}'}$ is asymptotically minimax for all r. Indeed, if b be fixed, simply choose r small enough that $b > 4r(1+r)(1+2r)^{-1} = 4r/(1+v)$, and then from (50), we have

$$\|\sigma\|_\infty \geq \sigma(1, \omega_1) = 1 + \omega_1(\omega_1 - 2r) > 1.$$

2. When s_n does not diverge to ∞, an 'independent blocks' sparse prior using π_{B} is asymptotically least favorable, along the lines of [19, Ch. 8.6]. Let $\pi_S(\tau; m)$ denote a single spike prior of scale τ on $\mathbb{R}^m$. This chooses an index $I \in \{1, \ldots, m\}$ at random and sets $\theta = \tau e_I$, where e_i is a unit length vector in the ith co-ordinate direction. We randomly draw τ from $(\nu_{\mathsf{B}}^+ + \nu_{\mathsf{B}}^-)/2$. However, instead of (11), we choose $\lambda = v^{1/2}(t_m - \log t_m)$ where $t_m = \sqrt{2 \log m}$. The independent blocks prior $\pi_{\mathsf{IB,n}}$ on $\Theta[s_n]$ is built by dividing $\{1, \ldots, n\}$ into s_n contiguous blocks B_j, each of length $m = m_n = [n/s_n]$. Independently for each block B_j, draw components according to $\pi_S(\cdot; m)$ and set $\theta_i = 0$ for the remaining $n - m_n s_n$ co-ordinates. This prior is supported on $\Theta[s_n]$ as any draw from $\pi_{\mathsf{IB,n}}$ has exactly s_n non-zero components. The proof that it is least favorable is then analogous to that of Theorem 6 in [34].

4. Risk properties of Spike and Slab procedures. We again use the risk decomposition provided by Lemma 2.1, now with the univariate spike and slab prior $\pi_S[\eta, \ell]$. We use $N_{\theta,v}^S(Z)$ and $D_\theta^S(Z)$ to denote the associated risk components of Lemma 2.1 for the spike and slab predictive density estimates $\hat{p}_S[\ell]$ based on the prior $\pi_S[\eta, \ell]$ for some $\ell > 0$ (the dependence on ℓ is kept implicit in the notations).

Figure 4 gives a schematic showing the strategy for the proof of Theorems 1.5 and 1.6. Separate risk bounds for $\hat{p}_S[\ell]$ are established below for θ lying in intervals roughly corresponding to $[0, \lambda_n], [\lambda_n, \lambda_e]$ and $[\lambda_e, \ell]$ where $\lambda_e = v^{-1/2}\lambda_n$; a threshold used in sparse point estimation. The critical interval is $[\lambda_n, \lambda_e]$, and the risk bound there suffices for asymptotic minimaxity if $\log \ell = o(\lambda_n^2)$, which leads to Theorem 1.5 if $\log t_n = o(\lambda_n^2)$ and we take $\ell = t_n$.

If, however, $\log t_n \sim \beta\lambda_n^2$, then no uniform slab width works: if $\log \ell \geq \beta\lambda_n^2/(2v)$, roughly, then the maximum at approximately $\theta = \sqrt{1+\beta}\lambda_n$ is too high, while for $\log \ell < \beta\lambda_n^2/(2v)$, the maximum risk is too large near the right endpoint, $\theta = t_n$.

To ease notation we often drop the suffixes from λ_n and η_n, particularly while discussing univariate *prde*s. Their risk functions are calculated in the regime $\lambda \to \infty$ as $\eta \to 0$.

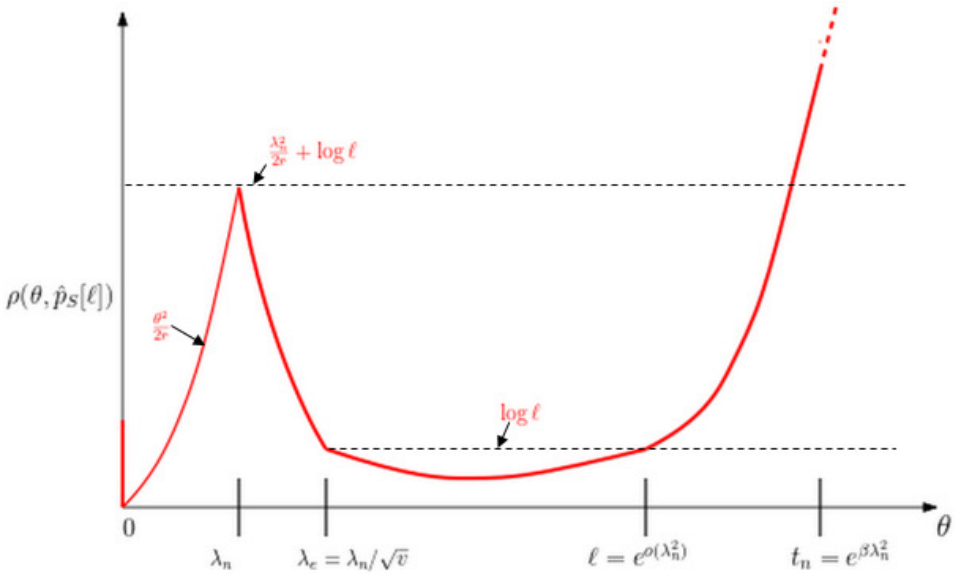

FIG 4. *Schematic for risk bounds* (56) *for uniform slab prior* $\pi_S[\eta, \ell]$ *and estimate* $\hat{p}_S[\ell]$.

Proof of Lemma 1.4. For the first upper bound, simply take $\nu = \delta_0$ in Lemma 2.1; the corresponding $\pi = \delta_0$ has $\rho(\theta, \hat{p}_{\delta_0}) = \theta^2/(2r)$. The bound now follows from (19). For the second statement, we claim that whenever $t_n > \lambda_n$, then as $n \to \infty$,

$$R_N(\Theta_n[s_n, t_n]) \sim R_N(\Theta_n[s_n]) \sim s_n\lambda_n^2/(2r). \tag{55}$$

Indeed, the independent blocks prior π_n^{IB} constructed in [34, Theorem 6] to show that $R_N(\Theta_n[s_n]) \sim s_n\lambda_n^2/(2r)$ is actually, by its very definition, supported on $\Theta_n[s_n, \nu_n]$, where $\nu_n < \sqrt{v}\sqrt{2\log[n/s_n]} \leq \lambda_n < t_n$. Since obviously $\Theta_n[s_n, \nu_n] \subset \Theta_n[s_n, t_n] \subset \Theta_n[s_n]$, the conclusion (55) follows. □.

For lower bounds on risk of its predictive density estimate, the following convexity inequality is helpful. It is proved in the supplement.

LEMMA 4.1. *If* $\eta \leq \frac{1}{2}$ *and* $\theta\ell/v \geq 1$, *then*

$$\mathbb{E} \log N_{\theta,v}^S(Z) \leq \theta\ell/v.$$

The *proof of* (16) follows easily from the above lemma. From the left side of (25) and Lemma 4.1,

$$\rho(\theta, \hat{p}_S[\ell]) \geq \frac{\theta^2}{2r} - \frac{\theta\ell}{v}, \qquad \text{for} \quad \theta \geq \frac{v}{\ell}.$$

Hence, from (18),

$$\sup_{\Theta_n[s_n]} \rho(\boldsymbol{\theta}, \hat{p}_S[\ell]) \geq s_n \sup_{\theta \in \mathbb{R}} \rho(\theta, \hat{p}_S[\ell]) = \infty,$$

while $R_N(\Theta_n[s_n])$ is finite for each n, e.g. [34], so (16) follows.

4.1. *Proof of Theorem 1.5.* Recall that $\lambda = \sqrt{2v \log \eta^{-1}}$ and define $\tilde{\lambda} = \lambda/\sqrt{v} + \sqrt{2 \log \lambda}$. We will show a piecewise risk bound

$$\rho(\theta, \hat{p}_S[\ell]) \le \begin{cases} \theta^2/(2r) + O(\lambda \log \lambda) & 0 < \theta < \lambda \\ \lambda^2/(2r) + \log \ell + O(\lambda \log \lambda) & \lambda \le \theta < \tilde{\lambda} \\ \log \ell + O(\lambda) & \tilde{\lambda} \le \theta \le \ell. \end{cases} \tag{56}$$

For $0 < \theta < \lambda$, simply use the basic upper bound (25) along with the following bound for D_θ^S, shown in the supplement: for each $r > 0$,

$$\mathbb{E} \log D_\theta^S(Z) \le \begin{cases} O(\lambda \log \lambda) & 0 < \theta < \tilde{\lambda} \\ \theta^2/2 - \lambda^2/(2v) + O(\lambda) & \theta \ge \tilde{\lambda}. \end{cases} \tag{57}$$

For the remaining two cases, i.e. for $\theta > \lambda$, we use the full decomposition (21) of Lemma 2.1. To this end, an alternative representation for $N_{\theta,v}^S$ will be useful. Completing the square in (22), we get

$$N_{\theta,v}^S(Z) = 1 + c(\eta)\sqrt{v} \exp(\tfrac{1}{2} Z_{\theta,v}^2) \Phi_{\ell,v}, \tag{58}$$

where we have set $Z_{\theta,v} = Z + \theta/\sqrt{v}$ and

$$\Phi_{\ell,v} = \Phi(v^{-1/2}(\ell - \theta) - Z) - \Phi(v^{-1/2}(-\ell - \theta) - Z).$$

In the supplement, we show that, uniformly in $v \in (0,1)$, $\ell \ge 1$ and $|\theta| \le \ell$,

$$\mathbb{E} \log \Phi_{\ell,v} \ge a_0 := \log \phi(0) + 2/3. \tag{59}$$

The constant $c(\eta) = \eta(1-\eta)^{-1}\{2\ell\phi(0)\}^{-1}$ satisfies

$$-\log \ell - \lambda^2/(2v) \le \log\{(1-\eta)c(\eta)\} = \log \phi(0) - \log \ell - \lambda^2/(2v) \tag{60}$$

From the preceding three displays and $\mathbb{E} Z_{\theta,v}^2 = 1 + \theta^2/v$ we obtain

$$\begin{aligned} -\mathbb{E} \log N_{\theta,v}^S(Z) &\le -\log c(\eta) - \tfrac{1}{2} \log v - \tfrac{1}{2} \mathbb{E} Z_{\theta,v}^2 - \mathbb{E} \log \Phi_{\ell,v} \\ &\le \log \ell + \lambda^2/(2v) - \theta^2/(2v) + O(1). \end{aligned} \tag{61}$$

Now observe from (61) and $v^{-1} = r^{-1} + 1$ that

$$\theta^2/(2r) - \mathbb{E} \log N_{\theta,v}^S(Z) \le \lambda^2/(2r) - (\theta^2 - \lambda^2)/2 + \log \ell + O(1).$$

Combining this with the bounds in (57) yields the remaining two bounds.

For any $\ell \ge 1$ such that $\log \ell = o(\lambda^2)$, we conclude that as $\lambda \to \infty$,

$$\sup_{\theta \le \ell} \rho(\theta, \hat{p}_S[\ell]) \le \frac{\lambda^2}{2r}(1 + o(1)).$$

This completes the proof of (29) and, as remarked there, the proof of Theorem 1.5.

4.2. *Proof of Theorem 1.6.* We use the basic lower risk bound (25), and show that for suitable θ that $\mathbb{E} \log N_{\theta,v}^S Z$ cannot be large enough to offset the leading term $\theta^2/(2r)$. To obtain a result uniform over all slab widths ℓ, we need two different types of upper bound on $N_{\theta,v}^S$.

Define t_λ and $\tilde{t}_\lambda = o(t_\lambda)$ by setting $\log t_\lambda = \beta\lambda^2/(2v)$ and $\log \tilde{t}_\lambda = \log t_\lambda - \lambda$. We look first at large values of ℓ, using representation (58). Observe first that for $\ell > \tilde{t}_\lambda$, the right side of (60) yields

$$\sqrt{v} c(\eta) \le C \exp\{-\log \tilde{t}_\lambda - \lambda^2/(2v)\} = C \exp\{-\tilde{\theta}^2/(2v)\}$$

for a constant $C = C(v)$ if we set $\tilde{\theta}^2 = \lambda^2 + 2v \log \tilde{t}_\lambda$. Using now (58) and $\Phi_{\ell,v} < 1$, we have

$$\log N^S_{\tilde{\theta},v}(Z) \le \log\{1 + C\exp[-\tilde{\theta}^2/(2v) + (Z + \tilde{\theta}/\sqrt{v})^2/2]\}$$
$$\le \log 2 + \log(1 + C) + Z^2/2 + |Z|\tilde{\theta}/\sqrt{v}.$$

Consequently $\mathbb{E}\log N^S_{\tilde{\theta},v}(Z) \le k_1 + k_2\tilde{\theta}$ where $k_i = k_i(v)$. Hence, from the left side of risk bound (25),

$$\rho(\tilde{\theta}, \hat{p}_S[\ell]) \ge \frac{\tilde{\theta}^2}{2r} - k_1\tilde{\theta} - k_2.$$

Now observe from the definition of $\tilde{t}_\lambda$ that $\tilde{\theta}^2 = (1+\beta)\lambda^2 - 2v\lambda$ and that $\tilde{\theta} < t_\lambda$ for large λ. We conclude that for large λ,

$$\inf_{\ell > \tilde{t}_\lambda} \sup_{\theta \in [0, t_\lambda]} \frac{\rho(\theta, \hat{p}_S[\ell])}{\lambda^2/(2r)} \ge 1 + \beta + O(\lambda^{-1}). \tag{62}$$

For $\ell \le \tilde{t}_\lambda$, we set $\theta = t_\lambda$ and use the left side of (25), then Lemma 4.1:

$$\sup_{\theta \le t_\lambda} \rho(t_\lambda, p_S[\ell]) \ge \frac{t_\lambda^2}{2r} - \frac{t_\lambda \ell}{v} \ge \frac{t_\lambda^2}{2r} - \frac{t_\lambda \tilde{t}_\lambda}{v} \ge \frac{t_\lambda^2}{2r}(1 + o(1)),$$

where in the last inequality we used $\tilde{t}_\lambda = o(t_\lambda)$. Consequently,

$$\inf_{\ell \le \tilde{t}_\lambda} \sup_{\theta \in [0, t_\lambda]} \frac{\rho(\theta, \hat{p}_S[\ell])}{\lambda^2/(2r)} \ge \frac{t_\lambda^2}{\lambda^2}(1 + o(1)). \tag{63}$$

Combining (62) with (63) and then using (19) to go over to the multivariate problem, we obtain

$$\min_{\ell > 1} \sup_{\Theta_n[s_n, t_n]} \rho(\boldsymbol{\theta}, \hat{p}_S[\ell]) \ge (1+\beta) s_n \lambda_n^2/(2r)(1 + o(1)).$$

Theorem 1.6 now follows from (17) of Lemma 1.4.

5. Numerical Experiments. We turn to the numerical effectiveness of our asymptotic results under different levels of sparsity η_n, with special focus on moderate values. The product structure and the good bounds (19) relating maximal multivariate and univariate risks allow us to concentrate on the univariate prdes. We use a constrained prior space

$$\mathfrak{m}_\ell(\eta) = \{\pi \in \mathcal{P}(\mathbb{R}) : \pi(\theta = 0) \ge 1 - \eta,\ \pi(|\theta| > \ell) = 0\},$$

and set $\ell = 5\lambda = 5\sqrt{2\log \eta^{-v}}$. We consider three sparsity levels: (a) Moderate: $\eta = 0.1$, (b) High: $\eta = 0.001$, (c) Very High: $\eta = 10^{-10}$.

We compare the following prdes:

- Hard threshold Plug-in prde (H-Plugin): [34, Eqn. (31)]

$$\hat{p}_H(y|x) = p(y|\hat{\theta}_H, v_y) \text{ where } \hat{\theta}_H(x) = x\, I\{|x| > (v_x/v)^{1/2}\lambda\}\ .$$

- Cluster prior and Thresholding (C-Thresh) based asymptotically minimax prde $\hat{p}_T$ proposed in [34, Eqn. (12)-(14)]
- Bayes prdes based on the grid and bi-grid priors (Grid, Bi-Grid) rescaled on $\mathfrak{m}_\ell$: $\hat{p}_{\mathsf{G}}, \hat{p}_{\mathsf{B}}$
- Spike and Slab predictive density estimator (SS): $\hat{p}_S[\ell]$.

Sparsity	r	Asymp	H-Plugin	C-Thresh	Grid	Bi-Grid	SS
0.1	1	1.1513	120.4%	82.5%	88.3%	88.3%	105.8%
	0.5	1.5351	173.6%	108.8%	104.9%	104.9%	118.0%
	0.25	1.8421	278.5%	128.0%	127.0%	129.0%	132.3%
	0.1	2.0933	588.1%	145.2%	165.4%	155.9%	146.5%
0.001	1	3.4539	109.1%	70.7%	70.8%	70.8%	86.2%
	0.5	4.6052	162.1%	85.9%	84.6%	84.6%	96.9%
	0.25	5.5262	267.6%	89.9%	100.2%	96.8%	106.9%
	0.1	6.2798	582.8%	107.2%	115.6%	113.4%	118.0%
1E-10	1	11.5129	123.9%	150.4%	78.6%	78.6%	86.9%
	0.5	15.3506	185.4%	87.9%	87.1%	87.1%	93.9%
	0.25	18.4207	308.4%	94.6%	98.1%	96.3%	100.1%
	0.1	20.9326	677.0%	101.8%	110.5%	101.7%	106.3%

TABLE 1

Numerical evaluation of the maximum risk for the different univariate predictive densities over $[-\ell, \ell]$ as the degree of sparsity (η) and predictive difficulty r varies. Here, we have chosen $\ell = 5\lambda$, where λ is defined in (11). In 'Asymp' column we report the asymptotic minimax risk $\lambda^2/(2r)$. In the other columns, we report the maximum risk of the estimators as quotients of the 'Asymp' risk.

Table 1 reports the maximum value of the risk plots for these predictive estimators (supplement table 1 shows the locations of the respective maximas). Figure 5 plots $\theta \to \rho(\theta, \hat{p})$, showing however the rescaled value $\rho(0, \hat{p})(1-\eta)/\eta$ at $\theta = 0$. [The hard threshold plug-in density estimator $\hat{p}_H$ is omitted, as has poor maximum risk in Table 1 and confuses the plots.]

The tables and plots show that the Bi-grid prior Bayes prde $\hat{p}_{\mathsf{B}}$ and the C-Thresh prde $\hat{p}_T$ have similar worst case performance. For each r, the maximal risks of $\hat{p}_{\mathsf{B}}$ and $\hat{p}_T$ lie near or below the asymptotic level of $\log \eta^{-1}/(1+r)$ under high and very high sparsity, and at worst moderately above the asymptotic level for moderate sparsity. However $\hat{p}_T$ has substantially higher risk at the origin than the other prdes considered here, particularly for moderate sparsity. Differences in the performances of the grid and bi-grid prior based prdes appear under high sparsity; for further comparisons see supplementary figures 2 and 3. The maximal risk of the spike and slab procedure is higher than that of $\hat{p}_T$ or $\hat{p}_{\mathsf{B}}$ but does not exceed the asymptotic minimax level by much. Finally, the basic features of the risk plots are unchanged even under moderate sparsity.

6. Discussion and future work. Product priors based on infinite cluster priors $\pi_\infty[\eta, r]$ of [34, Sec. 6] will lead to minimax optimal Bayes prdes. Details, which do not follow directly from those for the bi-grid prior, are provided in [8].

Our discussion of spike and slab priors was confined to uniform slabs. Theorem 2.1 can be used to show that Gaussian slabs are sub-optimal, while Bayes prdes based on heavier tailed slabs in the range from Laplace to Cauchy are minimax optimal. The tools to bound the maximal risk of continuous priors differ from those used here and will be detailed separately [33].

Our results are based on known sparsity levels. We make a few remarks on adaptation to unknown sparsity from theoretical and computational perspectives. A manuscript in preparation considers adaptivity for continuous slabs with Laplace and Cauchy tails. Adaptation to minimax risk is possible up to multiplicative constants and an additive logarithmic term. Both exact sparsity (ℓ_0) and approximate sparsity (ℓ_p, $0 < p < 2$) are considered.

Recently, computationally tractable Bayesian methods which adapt to unknown sparsity levels and possibly dense signals have been developed for point estimation [5, 3, 40]. In our sequence model (1), under unknown sparsity level $\eta_n = s_n/n$, there exist fast procedures for estimating posteriors from spike-and-slab priors that are mixtures of a Dirac measure at 0 and a continuous distribution [20, 6, 41].

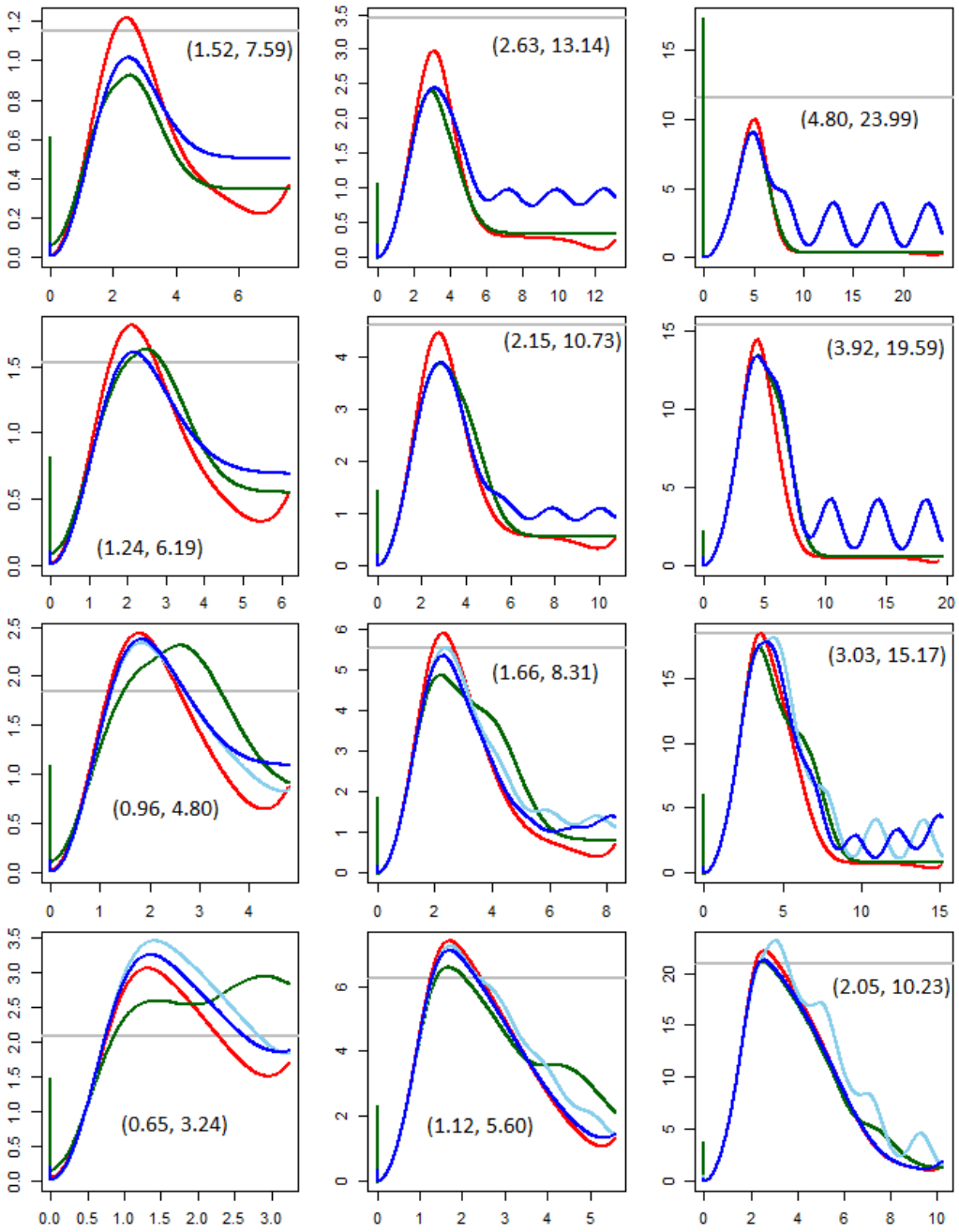

FIG 5. *Risk plots* $\rho(\theta,\cdot)\,\{(1-\eta)/\eta\}^{I\{\theta=0\}}$ *for univariate predictive density estimators* $\hat{p}_T$ *(dark green),* $\hat{p}_G$ *(skyblue),* $\hat{p}_B$ *(blue) and* $\hat{p}_S$ *(red) versus* $\theta \in [0,\ell]$*, for* $\ell = 5\lambda$*. Columns vary with moderate, high and very high sparsity,* $\eta = 0.1, 0.001, 10^{-10}$*, left to right. Rows vary* $r = 1, 0.5, 0.25$ *and* 0.1 *from top to bottom. The horizontal line shows the asymptotic univariate minimax risk of* $\log \eta^{-1}/(1+r) = \lambda^2/(2r)$*, with* $\lambda = \sqrt{2\log \eta^{-v}}$ *and* ℓ *shown in the insets. Note that,* $\hat{p}_G$ *(skyblue) and* $\hat{p}_B$ *(blue) overlap exactly in plots for the first two rows.*

7. Appendix. We present proofs of the risk bounds in Lemmas 3.1 and 3.3. The proof of Lemma 3.2 uses tools similar to Lemma 3.1 and appears in the supplement.

Proof of Lemma 3.1. We do the easy lower bounds involving $N_{\theta,v}(Z)$ first. Indeed, the

bound for $\theta < \lambda$ follows just from $N_{\theta,v}(Z) \geq 1$. For $\theta = \lambda(\alpha_l + \omega) \geq \lambda$, from (44) using $\mathbb{E}(Z) = 0$ we get:

$$\mathbb{E}\log N_l = \log c_1(\eta) + \tfrac{1}{2}\lambda^2 n(l,\omega), \text{ and,}$$
$$\mathbb{E}\log N_{l+1} = \log c_1(\eta) + \tfrac{1}{2}\lambda^2 \check{n}(l,\omega).$$

But $\log c_1(\eta) = \log c(\eta) - \log(1-\eta)^{-1} = O(1)$ as $\lambda \to \infty$. Hence, the proof of the lower bound is completed by using

$$\mathbb{E}\log N_{\theta,v}(Z) \geq \max\{\mathbb{E}\log N_l,\ \mathbb{E}\log N_{l+1}\}.$$

The proof of the upper bound on $\mathbb{E}\log D_\theta(Z)$ is more involved, and we first outline the approach. From (39) and $1 + x + y < (1+x)(1+y/x)$, we have

$$\log D_\theta(Z) \leq \log(1 + D_l) + \log(1 + \check{D}_l), \tag{64}$$

where we set $\check{D}_l = \sum_{i \notin \{0,l\}} D_i/D_l$. Henceforth in the proof, we make the choice $l = l(\theta)$ except that when $0 \leq \theta < \mu_1$ we set $l = 1$.

For the first term (henceforth we call it the main term) in (64) we will show

$$\mathbb{E}\log(1 + D_l) \leq \begin{cases} \frac{1}{2}\lambda^2 d^+(l,\omega) + O(\lambda) & \text{for } l \geq 1 \\ O(1) & \text{if } 0 \leq \theta < \lambda \end{cases} \tag{65}$$

with $O(\lambda)$ being uniform in l. For the other term in (64) we will show that it is $O(\lambda)$ for all l (and so, henceforth we call it the remainder term). For that purpose, we write $D_{i,l} = D_i/D_l$ and decompose

$$\check{D}_l = \sum_{k=1}^{\infty} D_{l+k,l} + \sum_{k=1}^{l-1} D_{l-k,l} + \sum_{j=1}^{\infty} D_{-j,l}.$$

We use the elementary inequality $\log(1 + \sum \gamma_m) \leq \sum \log(1+\gamma_m)$ to obtain that $\mathbb{E}\log(1 + \check{D}_l)$ is bounded above by

$$\mathbb{E}\log\Big(1 + \sum_{k=1}^{\infty} D_{l+k,l}\Big) + \mathbb{E}\log\Big(1 + \sum_{k=1}^{l-1} D_{l-k,l}\Big) + \mathbb{E}\log\Big(1 + \sum_{j=1}^{\infty} D_{-j,l}\Big).$$

Now, note that $D_{-j} \stackrel{\mathcal{D}}{=} D_j \exp\{-2\mu_j\theta\} \leq D_j$ since $\mu_j = -\mu_j, \pi_{-j} = \pi_j$ and $\mathcal{L}(Z)$ is symmetric. Hence

$$\sum_{j=1}^{\infty} D_{-j,l} \stackrel{\mathcal{D}}{\leq} \sum_{j=1}^{\infty} D_{j,l} = \sum_{k=1}^{l-1} D_{l-k,l} + 1 + \sum_{k=1}^{\infty} D_{l+k,l}.$$

Combining the above two displays and again using the aforementioned inequality on logsums we obtain

$$\mathbb{E}\log(1 + \check{D}_l) \leq 2\mathbb{E}\log\Big(1 + \sum_{k=1}^{\infty} D_{l+k,l}\Big) + \log 2 + 2\mathbb{E}\log\Big(1 + \sum_{k=1}^{l-1} D_{l-k,l}\Big). \tag{66}$$

We will later show that the two main right side terms are each $O(\lambda)$. This concludes the outline; we now turn to detailed analysis.

The Main term in (64). We first dispose of the case $0 \leq \theta < \lambda$. From (40) and (41),

$$D_1 = c_1(\eta)\exp\{\lambda Z + \lambda\theta - \tfrac{1}{2}\lambda^2(2 + r^{-1})\}.$$

Since $\theta < \lambda$ and $c_1(\eta) < (1-\eta)^{-1}$, and using $\log(1+x) \le \log 2 + (\log x)_+$,

$$\log(1+D_1) \le \log 2 + \log(1-\eta)^{-1} + \lambda\big(Z - 2^{-1}r^{-1}\lambda\big)_+$$

and hence $\mathbb{E}\log(1+D_1) \le O(1)$. This last bound uses an inequality we also need later: from the two term bound on Mills ratio (e.g. [19, Exercise 8.1]),

$$\mathbb{E}(Z-x)_+ = \phi(x) - x\tilde{\Phi}(x) \le x^{-2}\phi(x). \tag{67}$$

Now suppose that $\theta = \lambda(\alpha_l + \omega) \ge \lambda$ and use representation (44) for D_l. Abbreviating $\frac{1}{2}\lambda^2 d(l,\omega)$ as $d_{l\omega}$, we obtain

$$\begin{aligned}\mathbb{E}\log(1+D_l) &= \mathbb{E}\log D_l + \mathbb{E}\log(1+D_l^{-1}) \\ &= \log c(\eta) + \log(1-\eta)^{-1} + d_{l\omega} + \log 2 + \mathbb{E}(\log D_l^{-1})_+.\end{aligned}$$

Symmetry of $\mathcal{L}(Z)$ about 0 implies that $\log D_l^{-1} \stackrel{\mathcal{D}}{=} -\log c(\eta) + \log(1-\eta) + \mu_l Z - d_{l\omega}$. As $c(\eta) < 1$, we have

$$\mathbb{E}(\log D_l^{-1})_+ \le -\log c(\eta) + \mathbb{E}(\mu_l Z - d_{l\omega})_+.$$

From the previous two displays and $\log(1-\eta)^{-1} = O(\eta)$, we have

$$\mathbb{E}\log(1+D_l) \le d_{l\omega} + \mathbb{E}(\mu_l Z - d_{l\omega})_+ + O(1). \tag{68}$$

We now bound the expectation on the right side. Consider first those l for which $\alpha_l \le 2 + r^{-1}$ and thus $\mu_l \le (2+r^{-1})\lambda$. Noting that

$$\mathbb{E}(\mu_l Z - d_{l\omega})_+ \le -d_{l\omega} I\{d_{l\omega} \le 0\} + \mu_l \mathbb{E}Z_+,$$

we then conclude that

$$d_{l\omega} + \mathbb{E}(\mu_l Z - d_{l\omega})_+ \le (d_{l\omega})_+ + (2+r^{-1})\phi(0)\lambda.$$

Now consider the remaining l, with $\alpha_l \ge 2 + r^{-1}$, for which we claim that

$$\alpha_l^2 - \beta_l - r^{-1} \ge \tfrac{1}{2}\alpha_l^2. \tag{69}$$

We verify this via the equivalent form $\alpha_l^2 - 2\beta_l \ge 2r^{-1}$. Indeed, since $\beta_l \le \alpha_l$, we have

$$\alpha_l^2 - 2\beta_l \ge \alpha_l(\alpha_l - 2) \ge (2+r^{-1})r^{-1} \ge 2r^{-1}.$$

Since $\omega \ge 0$, we have from (45) and (69),

$$d_{l\omega} \ge \tfrac{1}{2}\lambda^2[\alpha_l^2 - \beta_l - r^{-1}] \ge \tfrac{1}{4}(\lambda\alpha_l)^2 = \tfrac{1}{4}\mu_l^2.$$

From the bound (67), we calculate

$$\mathbb{E}(\mu_l Z - d_{l\omega})_+ \le \mu_l \mathbb{E}(Z - \mu_l/4)_+ \le 16\frac{\phi(\mu_l/4)}{\mu_l}$$

uniformly in $\lambda \ge 1$ and l such that $\alpha_l \ge 2 + r^{-1}$. The right side is uniformly bounded in l. Combining the two cases with (68), we have proven the bound (65) on the first term of (64).

We turn now to bounding the remainder (66). This depends on the decay between successive terms D_j, so we start by using (42) to derive a useful representation for D_{j+1}/D_j. Indeed, using $\mu_j = \lambda\alpha_j$ and $\theta = \lambda(\alpha_l + \omega)$, we define

$$\begin{aligned}\Delta_j = \Delta(j;l,\omega) &= (2/\lambda^2)[G(\mu_{j+1};\theta) - G(\mu_j;\theta)] \\ &= \dot{\alpha}_j[\alpha_{j+1} + \alpha_j - 2\alpha_l - 2\omega] + \dot{\beta}_j\end{aligned}$$

and arrive at, for $j \geq 1$,

$$\frac{D_{j+1}}{D_j} = \exp\{\lambda \dot{\alpha}_j Z - \tfrac{1}{2}\lambda^2 \Delta_j\}. \tag{70}$$

We now show that Δ_j crosses zero at $j = l$, meaning $\Delta_j \geq 0$ for $j \geq l$ and $\Delta_j \leq 0$ for $j < l$. This will also verify the claim in Section 2 that $j \to G(\mu_j; \theta)$ is minimized at $j = l(\theta)$ for each $\theta \in [\mu_l, \mu_{l+1})$. The argument splits into two largely parallel cases.

Suppose first that $j \geq l$, so that $j = l + k$ for $k \geq 0$. Using $\alpha_l + \omega < \alpha_{l+1}$, then $\dot{\beta}_{l+k} = \dot{\alpha}^2_{l+k}$ and finally $\dot{\alpha}_{l+k} + \alpha_{l+k} = \alpha_{l+k+1}$, we have for any $k \geq 0$,

$$\Delta_{l+k} > \dot{\alpha}_{l+k}[\alpha_{l+k+1} + \alpha_{l+k} - 2\alpha_{l+1}] + \dot{\alpha}^2_{l+k} = 2\dot{\alpha}_{l+k}(\alpha_{l+k+1} - \alpha_{l+1}) \geq 0, \tag{71}$$

with the last inequality being strict for $k \geq 1$.

Suppose now that $j < l$, so that $j = l - k - 1$ for $k \geq 0$. Using $\alpha_l + \omega \geq \alpha_l$, then $\dot{\beta}_{l-k-1} = \dot{\alpha}^2_{l-k-1}$ and finally $\dot{\alpha}_{l-k-1} + \alpha_{l-k-1} = \alpha_{l-k}$, we have

$$\begin{aligned}\Delta_{l-k-1} &\leq \dot{\alpha}_{l-k-1}[\alpha_{l-k} + \alpha_{l-k-1} - 2\alpha_l] + \dot{\alpha}^2_{l-k-1} \\ &= 2\dot{\alpha}_{l-k-1}(\alpha_{l-k} - \alpha_l) \leq 0,\end{aligned}$$

with strict inequality when $k \geq 1$.

As final preparation, we record a useful bound whose proof is provided in the supplement.

LEMMA 7.1. *If $a_1, a_2, \ldots$ are positive, then for each $n \geq 1$,*

$$\log\left(1 + \sum_{k=1}^{n+1} a_k\right) < \log(1 + a_1) + \sum_{k=1}^{n} \frac{a_{k+1}}{a_k}. \tag{72}$$

We next concentrate on bounding the *first term of* (66). Noting that D_js are positive, use (72) with $a_k = D_{l+k}/D_l$ and $\log(1 + a_1) \leq \log 2 + (\log a_1)_+$ to write

$$\mathbb{E}\log\left(1 + \sum_{k=1}^{\infty} D_{l+k,l}\right) \leq \log 2 + \mathbb{E}\Big(\log \frac{D_{l+1}}{D_l}\Big)_+ + \mathbb{E}\Big\{\sum_{k=1}^{\infty} \frac{D_{l+k+1}}{D_{l+k}}\Big\}. \tag{73}$$

In (70) with $j = l$, we have seen that $\Delta_l \geq 0$ and so

$$\mathbb{E}\Big(\log \frac{D_{l+1}}{D_l}\Big)_+ \leq \lambda \dot{\alpha}_l \, \mathbb{E} Z_+ \leq \lambda \phi(0).$$

When $j = l + k$, observe from (71) that $\Delta_{l+k} \geq 2\dot{\alpha}^2_{l+k} + 2\dot{\alpha}_{l+k}(\alpha_{l+k} - \alpha_{l+1})$. From (70), now with $j = l + k$ for $k \geq 1$,

$$\begin{aligned}\mathbb{E}\Big\{\frac{D_{l+k+1}}{D_{l+k}}\Big\} &= \exp\{\tfrac{1}{2}\lambda^2[\dot{\alpha}^2_{l+k} - \Delta_{l+k}]\} \\ &\leq \exp\{-\tfrac{1}{2}\lambda^2[\dot{\alpha}^2_{l+k} + 2\dot{\alpha}_{l+k}(\alpha_{l+k} - \alpha_{l+1})]\} \\ &\leq \exp\{-\tfrac{1}{2}\lambda^2 b^2 - \lambda^2 b^2 (k-1)\},\end{aligned}$$

so that the right side of (73) is $O(\lambda) + O(e^{-\lambda^2 b^2/2}) = O(\lambda)$. The last inequality in the above display uses $j \to \dot{\alpha}_j$ is increasing and $\dot{\alpha}_j \geq b$.

Second term of (66). Now use (72) with $a_k = D_{l-k}/D_l$:

$$\mathbb{E}\log\left(1+\sum_{k=1}^{l-1} D_{l-k,l}\right) \le \log 2 + \mathbb{E}\Big(\log\frac{D_{l-1}}{D_l}\Big)_+ + \mathbb{E}\Big\{\sum_{k=1}^{l-2}\frac{D_{l-k-1}}{D_{l-k}}\Big\}. \tag{74}$$

In (70) with $j = l-1$, we have seen that $\Delta_{l-1} \le 0$ and so

$$\mathbb{E}\Big(\log\frac{D_{l-1}}{D_l}\Big)_+ \le \lambda\dot\alpha_{l-1}\mathbb{E}Z_+ \le \lambda\phi(0).$$

From (70), now with $j = l-k-1$,

$$\begin{aligned}\mathbb{E}\Big\{\frac{D_{l-k-1}}{D_{l-k}}\Big\} &= \mathbb{E}\{\exp\{-\lambda\dot\alpha_{l-k-1}Z + \tfrac{1}{2}\lambda^2\Delta_{l-k-1}\}\}\\ &\le \exp\{\tfrac{1}{2}\lambda^2\dot\alpha_{l-k-1}[\dot\alpha_{l-k-1} + 2(\alpha_{l-k}-\alpha_l)]\},\end{aligned}$$

as $\Delta_{l-k-1} \le 2\dot\alpha_{l-k-1}(\alpha_{l-k}-\alpha_l)$. Again, using $j \to \dot\alpha_j$ is increasing and $\dot\alpha_j \ge b$, we have

$$\dot\alpha_{l-k-1} + 2(\alpha_{l-k}-\alpha_l) \le \dot\alpha_{l-k} - 2(\alpha_l - \alpha_{l-k+1}) - 2\dot\alpha_{l-k} \le -b - 2(k-1)b.$$

Using $\dot\alpha_{l-k-1} \ge b$ again, we conclude that

$$\mathbb{E}\Big\{\sum_{k=1}^{l-2}\frac{D_{l-k-1}}{D_{l-k}}\Big\} \le \sum_{k=1}^{\infty}\exp\{-\tfrac{1}{2}\lambda^2 b^2 - \lambda^2 b^2(k-1)\} = O(e^{-\lambda^2 b^2/2}).$$

Thus, we have proved the desired bound on the second term. This completes the proof of the lemma.

Proof of Lemma 3.3. The argument borrows some steps from the proof of Lemma 3.1, but is simpler, though not a special case. The three point prior corresponds, in (30) to choices $\pi_0 = 1-\eta, \pi_1 = \eta/2, \mu_1 = a\lambda$. From (37)-(38), we have $N_{a\lambda,v}(Z) = 1 + N_1 + N_{-1}$, with

$$N_1 = c_1(\eta)\exp\{v^{-1/2}a\lambda Z + (2v)^{-1}a^2\lambda^2 - \tfrac{1}{2}\lambda^2(1+r^{-1})\}, \tag{75}$$

where $c_1(\eta) = 2^{-1}(1-\eta)^{-1}$ and $N_{-1} = N_1\exp(-2v^{-1/2}a\lambda Z - 2v^{-1}a^2\lambda^2)$. Correspondingly $D_\theta(Z) = 1 + D_1 + D_{-1}$, where D_1 and D_{-1} are obtained from N_1 and N_{-1} by replacing v with 1. From Theorem 2.1, and $\log(1+D_1+D_{-1}) \le \log(1+D_1) + \log(1+D_{-1}/D_1)$,

$$\begin{aligned}\rho(a\lambda,\hat p_a) &= (2r)^{-1}a^2\lambda^2 - \mathbb{E}\log(1+N_1+N_{-1}) + \mathbb{E}\log(1+D_1+D_{-1})\\ &\le (2r)^{-1}a^2\lambda^2 - \mathbb{E}\log(1+N_1) + \mathbb{E}\log(1+D_1) + \mathbb{E}\log(1+D_{-1}/D_1)\\ &\le (2r)^{-1}a^2\lambda^2 - (\mathbb{E}\log N_1)_+ + 2\log 2 + \mathbb{E}(\log D_1)_+ + \mathbb{E}[\log(D_{-1}/D_1)]_+.\end{aligned}$$

From (75), and its analog for D_1, we have, on setting $\epsilon(\eta) = \log c_1(\eta) < 0$, recalling that $rv^{-1} = r+1$, and using (67)

$$\begin{aligned}(\mathbb{E}\log N_1)_+ &= [\epsilon(\eta) + (2v)^{-1}(a^2-1)\lambda^2]_+ \ge \epsilon(\eta) + (2r)^{-1}\lambda^2(r+1)(a^2-1)_+\\ \mathbb{E}(\log D_1)_+ &\le \epsilon(\eta) + a\lambda\mathbb{E}Z_+ + (2r)^{-1}\lambda^2(ra^2-r-1)_+\\ \mathbb{E}[\log(D_{-1}/D_1)]_+ &= 2a\lambda\mathbb{E}(Z-a\lambda)_+ = 2\phi(a\lambda)/a\lambda = O(\lambda).\end{aligned}$$

Combine the last four displays to get

$$\rho(a\lambda,\hat p_a) \le (2r)^{-1}\lambda^2\tilde\tau(a) + O(\lambda),$$

where

$$\tilde\tau(a) = a^2 - (r+1)(a^2-1)_+ + (ra^2-r-1)_+ = \tau(a).$$

Acknowledgements. The authors thank Associate Editor and three referees for especially stimulating comments that improved the presentation. GM was supported in part by the Zumberge individual award from the University of Southern California's James H. Zumberge faculty research and innovation fund and by NSF DMS 1811866. IMJ was supported in part by NSF DMS 1407813, 1418362 and 1811614 and thanks the Australian National University for hospitality while working on this paper.

SUPPLEMENTARY MATERIAL

Supplementary Materials to: On Minimax Optimality of Sparse Bayes Predictive Density Estimates
(url). The supplement [35] proves Lemma 3.2 and all the inequalities and lemmas used in Section 4. It also contains results from additional numerical experiments and further discussions on the risk properties of prdes.

REFERENCES

[1] AITCHISON, J. and DUNSMORE, I. R. (1975). *Statistical prediction analysis*. Cambridge University Press, Cambridge. MR0408097 (53 ##11864)
[2] ASLAN, M. (2006). Asymptotically minimax Bayes predictive densities. *Ann. Statist.* **34** 2921–2938. MR2329473 (2008g:62093)
[3] BHATTACHARYA, A., PATI, D., PILLAI, N. S. and DUNSON, D. B. (2015). Dirichlet–Laplace priors for optimal shrinkage. *Journal of the American Statistical Association* **110** 1479–1490.
[4] BROWN, L. D., GEORGE, E. I. and XU, X. (2008). Admissible predictive density estimation. *Ann. Statist.* **36** 1156–1170. MR2418653 (2009i:62023)
[5] CARVALHO, C. M., POLSON, N. G. and SCOTT, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* 465-480.
[6] CASTILLO, I., VAN DER VAART, A. et al. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics* **40** 2069–2101.
[7] FOURDRINIER, D., MARCHAND, É., RIGHI, A. and STRAWDERMAN, W. E. (2011). On improved predictive density estimation with parametric constraints. *Electron. J. Stat.* **5** 172–191. MR2792550 (2012h:62088)
[8] GANGOPADHYAY, U. and MUKHERJEE, G. (2021). On Discrete Priors and Sparse Minimax Optimal Predictive Densities. *Electronic Journal of Statistics* **15**.
[9] GEISSER, S. (1993). *Predictive inference*. *Monographs on Statistics and Applied Probability* **55**. Chapman and Hall, New York. An introduction. MR1252174 (95k:62006)
[10] GEORGE, E. I., LIANG, F. and XU, X. (2006). Improved minimax predictive densities under Kullback-Leibler loss. *Ann. Statist.* **34** 78–91. MR2275235 (2008h:62034)
[11] GEORGE, E. I., LIANG, F. and XU, X. (2012). From minimax shrinkage estimation to minimax shrinkage prediction. *Statist. Sci.* **27** 82–94. MR2953497
[12] GEORGE, E. I. and MCCULLOCH, R. E. (1997). Approaches for Bayesian variable selection. *Statistica sinica* 339–373.
[13] GEORGE, E. I. and XU, X. (2008). Predictive density estimation for multiple regression. *Econometric Theory* **24** 528–544. MR2391619 (2009a:62121)
[14] GHOSH, M., MERGEL, V. and DATTA, G. S. (2008). Estimation, prediction and the Stein phenomenon under divergence loss. *J. Multivariate Anal.* **99** 1941–1961. MR2466545 (2009m:62027)
[15] HARTIGAN, J. A. (1998). The maximum likelihood prior. *Ann. Statist.* **26** 2083–2103. MR1700222 (2000h:62030)
[16] ISHWARAN, H. and RAO, J. S. (2005a). Spike and slab variable selection: frequentist and Bayesian strategies. *Annals of Statistics* 730–773.
[17] ISHWARAN, H. and RAO, J. S. (2005b). Spike and slab gene selection for multigroup microarray data. *Journal of the American Statistical Association* **100** 764–780.
[18] JOHNSTONE, I. M. (1994). On minimax estimation of a sparse normal mean vector. *Ann. Statist.* **22** 271–289. MR1272083 (95g:62020)
[19] JOHNSTONE, I. M. (2013). Gaussian Estimation: Sequence and Wavelet Models. Version: 11 June, 2013. Available at "http://www-stat.stanford.edu/~imj".
[20] JOHNSTONE, I. M. and SILVERMAN, B. W. (2004). Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.* **32**. MR2089135 (2005h:62027)

[21] KOBAYASHI, K. and KOMAKI, F. (2008). Bayesian shrinkage prediction for the regression problem. *J. Multivariate Anal.* **99** 1888–1905. MR2466542 (2010a:62223)
[22] KOMAKI, F. (1996). On asymptotic properties of predictive distributions. *Biometrika* **83** 299–313. MR1439785 (98d:62048)
[23] KOMAKI, F. (2001). A shrinkage predictive distribution for multivariate normal observables. *Biometrika* **88** 859–864. MR1859415
[24] KOMAKI, F. (2006). Shrinkage priors for Bayesian prediction. *the Annals of Statistics* 808–819.
[25] KUBOKAWA, T., MARCHAND, É. and STRAWDERMAN, W. E. (2015). On predictive density estimation for location families under integrated squared error loss. *Journal of Multivariate Analysis* **142** 57–74.
[26] KUBOKAWA, T., MARCHAND, É., STRAWDERMAN, W. E. and TURCOTTE, J.-P. (2013). Minimaxity in predictive density estimation with parametric constraints. *Journal of Multivariate Analysis* **116** 382–397.
[27] KUBOKAWA, T., MARCHAND, É., STRAWDERMAN, W. E. et al. (2017). On predictive density estimation for location families under integrated absolute error loss. *Bernoulli* **23** 3197–3212.
[28] MALLOWS, C. (1978). Minimizing an Integral. *SIAM Review* **20** 183-183.
[29] MARUYAMA, Y., MATSUDA, T. and OHNISHI, T. (2019). Harmonic Bayesian prediction under alpha-divergence. *IEEE Transactions on Information Theory* **65** 5352–5366.
[30] MATSUDA, T. and KOMAKI, F. (2015). Singular value shrinkage priors for Bayesian prediction. *Biometrika* **102** 843–854.
[31] MITCHELL, T. J. and BEAUCHAMP, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association* **83** 1023–1032.
[32] MUKHERJEE, G. (2013). Sparsity and Shrinkage in Predictive Density Estimation, PhD thesis, Stanford University.
[33] MUKHERJEE, G. (2021). Minimax adaptive predictive density estimation for nonparametric regression. in preparation.
[34] MUKHERJEE, G. and JOHNSTONE, I. M. (2015). Exact minimax estimation of the predictive density in sparse Gaussian models. *Annals of Statistics* **43** 937.
[35] MUKHERJEE, G. and JOHNSTONE, I. (2020). Supplement to: On Minimax Optimality of Sparse Bayes Predictive Density Estimates.
[36] O'HARA, R. B., SILLANPÄÄ, M. J. et al. (2009). A review of Bayesian variable selection methods: what, how and which. *Bayesian analysis* **4** 85–117.
[37] PARK, T. and CASELLA, G. (2008). The bayesian lasso. *Journal of the American Statistical Association* **103** 681–686.
[38] ROCKOVÁ, V. (2017). Bayesian estimation of sparse signals with a continuous spike-and-slab prior. *Annals of Statistics*.
[39] ROČKOVÁ, V. and GEORGE, E. I. (2014). Negotiating multicollinearity with spike-and-slab priors. *Metron* **72** 217–229.
[40] ROČKOVÁ, V. and GEORGE, E. I. (2016). The Spike-and-Slab LASSO. *Journal of the American Statistical Association*.
[41] VAN ERVEN, T., SZABO, B. et al. (2020). Fast Exact Bayesian Inference for Sparse Signals in the Normal Sequence Model. *Bayesian Analysis*.
[42] XU, X. and LIANG, F. (2010). Asymptotic minimax risk of predictive density estimation for non-parametric regression. *Bernoulli* **16** 543–560. MR2668914 (2011g:62110)
[43] XU, X. and ZHOU, D. (2011). Empirical Bayes predictive densities for high-dimensional normal models. *J. Multivariate Analysis* **102** 1417-1428.
[44] YANO, K. and KOMAKI, F. (2017). Information Criteria For Prediction When The Distributions Of Current And Future Observations Differ. *Statistica Sinica* **27** 1205–1223.

Submitted to the Annals of Statistics

SUPPLEMENT TO "ON MINIMAX OPTIMALITY OF SPARSE BAYES PREDICTIVE DENSITY ESTIMATES"

BY GOURAB MUKHERJEE

AND

BY IAIN M. JOHNSTONE

University of Southern California and Stanford University

Here, we provide proofs of Lemmas 3.2, 4.1 and 7.1 as well as the proofs of all the inequalities used in Section 4 of the main paper. We also provide further insights on the risk properties of the predictive density estimates discussed in the main paper.

1. Further results on grid and bigrid priors.

Relative mass distributions of π_{G} and π_{B}. First we compare the normalizing constants c_{G} and c_{B} for ν_{G}^{+} and ν_{B}^{+} respectively. Let $\zeta = \eta^{v}$ and $\epsilon = \eta^{b^2 v} < \zeta$. We have $c_{\mathsf{G}}^{-1} = 1 + \zeta + \zeta^2 + \cdots$, while

$$c_{\mathsf{B}}^{-1} = 1 + \epsilon + \cdots + \epsilon^{K-1} + \epsilon^{K-1}(\zeta + \zeta^2 + \cdots),$$

so termwise comparison shows that $c_{\mathsf{B}} > c_{\mathsf{G}}$.

If $0 < x_1 < x_2 < (1 + bK)\lambda$, then Riemann sum approximations give

$$\pi_{\mathsf{B}}[x_1, x_2] = c_{\mathsf{B}} \sum_{x_1 < bj\lambda < x_2} \eta^{(j-1)vb^2} \approx c_{\mathsf{B}} \int_{x_1/b\lambda}^{x_2/b\lambda} e^{-(x-1)b^2\lambda^2/2} dx$$

$$\pi_{\mathsf{G}}[x_1, x_2] = c_{\mathsf{G}} \sum_{x_1 < j\lambda < x_2} \eta^{(j-1)v} \quad \approx c_{\mathsf{G}} \int_{x_1/\lambda}^{x_2/\lambda} e^{-(x-1)\lambda^2/2} dx.$$

For $x > (1 + bK)\lambda$ the corresponding density ratio is the constant $\gamma = c_{\mathsf{B}}/c_{\mathsf{G}} > 1$. To summarize, approximately

$$\frac{d\pi_{\mathsf{B}}}{d\pi_{\mathsf{G}}}(x) \approx \begin{cases} \gamma\, b^{-1} \exp\left\{\frac{1}{2}(1-b)\lambda[x - (1+b)\lambda]\right\} & \text{if } (1+b)\lambda < x < (1+Kb)\lambda \\ \gamma & \text{if } x > (1+Kb)\lambda. \end{cases}$$

Proof of Lemma 3.2. We modify some of the methods used in Lemma 3.1 to incorporate now $v = r(1+r)^{-1} \in (0,1)$. With N_j defined as in (37) and (38), using (44) and arguing as around (64), we bound

$$\log N_{\theta,v}(Z) \le \log(1 + N_1) + \log(1 + \check{N}_1),$$

with $\check{N}_1 = \sum_{j \notin \{0,1\}} N_{j,1}$ and $N_{j,1} = N_j/N_1$. The desired control of the main term is easier here than in Lemma 3.1: for $l = 1$,

$$\mathbb{E}\log(1 + N_1) \le \log 2 + \mathbb{E}(\log N_1)_+$$

$$\leq \log 2 + \log(1-\eta)^{-1} + v^{-1/2}\lambda\,\mathbb{E}|Z| + 2^{-1}\lambda^2 n(1,\omega)$$
$$\leq 2^{-1}\lambda^2 n(1,\omega) + O(\lambda).$$

Turn now to the remainder term. Since $N_{-j} \stackrel{D}{=} N_j \exp(-2v^{-1/2}\mu_j\theta)$, we may argue as before to obtain the analog of (66):

$$\mathbb{E}\{\log(1+\check{N}_1)\} \leq 2\,\mathrm{Rem}(\lambda) + 1,$$

where, using Lemma 7.1 with $a_k = N_{k+1,1}$ and setting $\check{R}_k = N_{k+1}/N_k$,

$$\mathrm{Rem}(\lambda) = \mathbb{E}\log\Big(1 + \sum_{k=1}^{\infty} N_{k+1,1}\Big) \leq \log 2 + \mathbb{E}(\log \check{R}_1)_+ + \sum_{k=1}^{\infty} \mathbb{E}\,\check{R}_{k+1}.$$

Using definition (38) and then (41) and (31), we obtain

$$\check{R}_k = (\pi_{k+1}/\pi_k)\exp\{v^{-1/2}(\mu_{k+1}-\mu_k)(Z+v^{-1/2}\theta) - \tfrac{1}{2}v^{-1}(\mu_{k+1}^2-\mu_k^2)\}$$
$$= \exp\{\lambda v^{-1/2}\dot{\alpha}_k Z - \tfrac{1}{2}\lambda^2 v^{-1}\gamma_k(\omega,v)\},$$

where we put $\theta = \lambda(1+\omega)$ and used $\dot{\beta}_k = \dot{\alpha}_k^2$ and $\alpha_{k+1} = \dot{\alpha}_k + \alpha_k$ to write

$$\gamma_k(\omega,v) = -2\dot{\alpha}_k(1+\omega) + \dot{\alpha}_k^2 + 2\dot{\alpha}_k\alpha_k + v\dot{\alpha}_k^2$$
$$= \dot{\alpha}_k[(1+v)\dot{\alpha}_k + 2(\alpha_k - (1+\omega))].$$

Since $b \leq 1$, we necessarily have $K-1 = \lceil 2b^{-3/2}\rceil \geq 2$, and therefore $\alpha_1 = 1, \alpha_2 = 1+b$ and $\dot{\alpha}_1 = b$. Consequently, at $k=1$ we have

$$\gamma_1(\omega,v) = b[(1+v)b - 2\omega] \geq 0$$

for all ω satisfying $0 \leq \omega \leq \omega_1 = \dot{\alpha}_1(1+v)/2$. We arrive at $\log \check{R}_1 \leq \lambda v^{-1/2} b Z$, and therefore $\mathbb{E}(\log \check{R}_1)_+ \leq \lambda v^{-1/2} b \mathbb{E} Z_+ = O(\lambda)$.

For $k \geq 2$, we write $\log \mathbb{E}\check{R}_k = -\frac{1}{2}\lambda^2 v^{-1}(\gamma_k - \dot{\alpha}_k^2)$, with $\gamma_k = \gamma_k(\omega,v)$. Note that

$$\gamma_k - \dot{\alpha}_k^2 = \dot{\alpha}_k[v\dot{\alpha}_k + 2(\alpha_k - (1+\omega))].$$

Using $\alpha_2 \geq 1+\omega$ and then $\dot{\alpha}_k \geq b$ it follows that

$$\gamma_k - \dot{\alpha}_k^2 \geq \alpha_k[v\dot{\alpha}_k + 2(\alpha_k - \alpha_2)] \geq b^2(v + 2(k-2)) \geq 0.$$

This entails

$$\sum_{k=2}^{\infty} \mathbb{E}\check{R}_k \leq e^{-b^2\lambda^2/2} \sum_{k=0}^{\infty} \exp(-v^{-1}\lambda^2 b^2 k) = O(\lambda).$$

The last two paragraphs show that $\mathrm{Rem}(\lambda) = O(\lambda)$ and complete the proof.

Proof of Lemma 7.1. We use induction. The bounds $1 + x + y < (1+x)(1+y/x)$ and $\log(1+x) < x$, valid for positive x, y, establish the case $n = 1$. For general n, let $a'_k = a_{k+1}/(1+a_1)$ for $k = 1, \ldots, n$. Then

$$\log\Big(1 + \sum_{k=1}^{n+1} a_k\Big) = \log(1+a_1) + \log\Big(1 + \sum_{k=1}^{n} a'_k\Big)$$
$$< \log(1+a_1) + \log(1+a'_1) + \sum_{k=1}^{n-1} \frac{a'_{k+1}}{a'_k}$$
$$< \log(1+a_1) + \frac{a_2}{a_1} + \sum_{k=2}^{n} \frac{a_{k+1}}{a_k},$$

where the inequalities use the cases $n-1$ and $n=1$ in turn.

2. Proofs of the lemmas and the inequalities used in Section 4.

Proof of Lemma 4.1. From (27), we find

$$\mathbb{E}N^S_{\theta,v}(Z) \le 1 + \frac{\eta}{1-\eta}\frac{1}{2l}\int_{-\infty}^{l} \exp\Big(\frac{\mu\theta}{v}\Big)\nu(d\mu) \le 1 + \frac{\eta}{(1-\eta)}\frac{e^w}{2w},$$

for $w = \theta l/v \ge 1$. From this and Jensen's inequality (26), we obtain

$$\mathbb{E}\log N^S_{\theta,v}(Z) \le w + \log\left\{e^{-w} + \frac{\eta}{(1-\eta)}\frac{1}{2w}\right\} \le w,$$

since the term in braces is bounded by $1/e + 1/2 \le 1$.

Proof of (59). Use $v \le 1$, then $0 \le \theta \le \ell$ and finally $\ell \ge 1$ to conclude

$$\Phi_{\ell,v} \ge \Phi_{\ell,1} \ge \Phi(-Z) - \Phi(-\ell - Z) \ge \Phi(-Z) - \Phi(-1-Z).$$

Now use symmetry of Z and then Jensen's inequality to get

$$\begin{aligned}\mathbb{E}\log\Phi_{\ell,v} \ge \mathbb{E}\log\int_Z^{Z+1}\phi(s)\,ds &\ge \mathbb{E}\int_Z^{Z+1}\log\phi(s)\,ds \\ &= \log\phi(0) - 2/3.\end{aligned}$$

Proof of (57). From the definition (58) with $v=1$ and using $\Phi_{\ell,1} \le 1$,

$$\log D^S_\theta(Z) \le \log 2 + [\log c(\eta) + (\theta+Z)^2/2]_+.$$

From the upper bound in (60), we have

$$\mathbb{E}\log D^S_\theta(Z) \le 2^{-1}\mathbb{E}[(\theta+Z)^2 - v^{-1}\lambda^2]_+ + O(1).$$

Now if $\theta < \tilde\lambda$, then

$$\begin{aligned}\mathbb{E}[(\theta+Z)^2 - v^{-1}\lambda^2]_+ &\le [\theta^2 - v^{-1}\lambda^2]_+ + 2|\theta|\,\mathbb{E}Z_+ + \mathbb{E}Z^2 \\ &\le [\tilde\lambda^2 - v^{-1}\lambda^2]_+ + 2\,\tilde\lambda\,\phi(0) + 1\end{aligned}$$

which suffices for the first bound.

If $\theta \ge \tilde\lambda$, we put $W = (\theta+Z)^2$ and $c = \lambda^2/v$ and apply the inequality $\mathbb{E}(W-c)_+ \le \mathbb{E}(W-c) + cP(W<c)$, valid for $W \ge 0$. For all $\theta \ge \tilde\lambda$, we have $\{|\theta+Z| \le \lambda/\sqrt{v}\} \subset \{Z < -\sqrt{2\log\lambda}\}$ and so

$$\begin{aligned}\mathbb{E}[(\theta+Z)^2 - v^{-1}\lambda^2]_+ &\le \mathbb{E}(\theta+Z)^2 - v^{-1}\lambda^2 + v^{-1}\lambda^2 P(Z > \sqrt{2\log\lambda}) \\ &\le \theta^2 - v^{-1}\lambda^2 + O(\lambda) \text{ as } \lambda \to \infty\,.\end{aligned}$$

where the last inequality uses the Mills ratio bound $P(Z>x) \le x^{-1}\phi(x)$.

3. Risk plots and further insights on the bi-grid priors.

3.1. *Risk Plots.* Figure 1 below shows the risk of the grid prior is well controlled below the desired limit when $r = 1$. Also, the plot reveals that the risk function exhibits periodicity for sufficiently large parametric values (in this case as $|\theta| \geq 2\lambda$) with a period of λ. In Figure 2 we have the risk plots of the grid and bi-grid priors for different values of r. The bi-grid comes in play for $r < r_0 \approx 0.309$. From the plots we see that for $|\theta| \in [\lambda, 2\lambda]$, the risk function for the grid prior is roughly decreasing in $|\theta|$ for large value of r but is bi-modal as r decreases towards r_0. At r_0 its two peaks are equal in height and as r decreases further the gap between the maximal risk of the grid prior and the bi-grid prior widens. In Figure 3, we exhibit a scenario where the pde based on the grid prior is no longer optimal and its risk is far dominated by the bi-grid prior based pde.

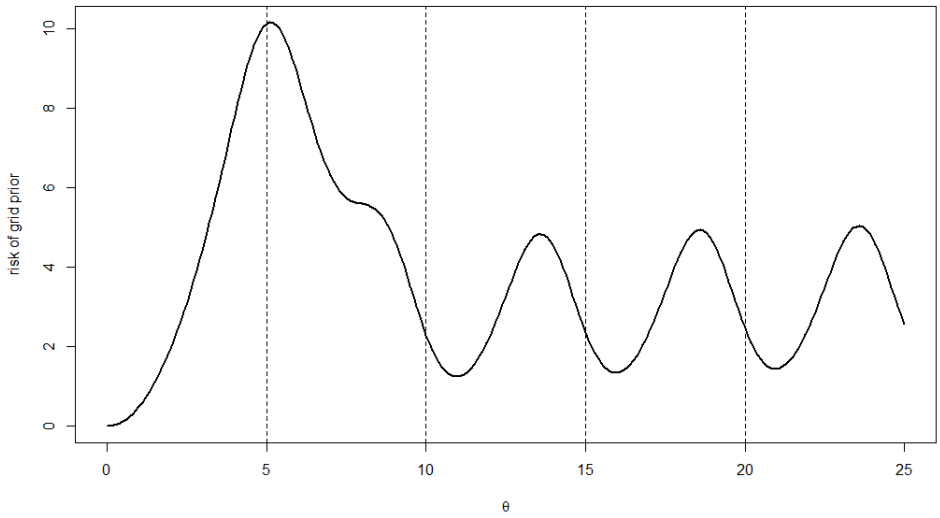

FIG 1. *Plot of the risk of $\hat{p}_G$ as the parameter θ varies over $\mathbb{R}^+$. Here, $\lambda = 5$, $r = 1$ and $\eta = \exp(-\lambda^2/2)$. The risk is well controlled below the asymptotic theory benchmark minimax value of $\lambda^2/(2r) = 12.5$. The vertical lines denote integer multiples of λ along the x-axis.*

To understand the differences in the risk properties between the grid and bi-grid prior as r varies, we now concentrate on the behavior of the risk components theoretically analyzed in Section 3. Following the proofs of theorems 1.1 and 1.2, we concentrate on the dominant component of the risk $\sigma(l, \omega)$ as defined in (47) where, $\theta = \lambda(\alpha_l + \omega)$. Figure 4 plots $\sigma(l, \omega)$ for $\hat{p}_G$ and $\hat{p}_B$ as l, ω vary. Note that the representation of θ is different for π_G and π_B as the α_l and β_l are different. As such, for $\hat{p}_G$, ω always varies over $[0, 1]$ where as for the bi-grid prior ω only varies over $[0, b]$ for some initial l denominations. In figure 4, we have 3 plots. The top plot shows the dominant risk for $\hat{p}_G$ when $r = 1$. The dominant risk here is always contained below the minimax value (shown in dotted horizontal line). Here, $d^+(l, \omega)$ starts being positive roughly at $l = 1$ and $\omega = 0.5$ onwards; and so, the plot of $\sigma(l, \omega)$ for $\hat{p}_G$ has 3 piecewise quadratics for $\theta \in [\lambda, 2\lambda]$ but 2 piecewise quadratics for the following intervals. The next plot shows the dominant risk for $\hat{p}_G$ when $r = 0.25$. In this case, $d^+(l, \omega)$ starts being positive roughly at $l = 2$ and $\omega = 0.5$ onwards; and so, the plot of $\sigma(l, \omega)$ for $\hat{p}_G$ has 3 piecewise quadratics for θ in the interval $[2\lambda, 3\lambda]$. Here, the dominant risk is not always contained below the minimax value. The bottom plot shows the dominant risk for $\hat{p}_B$ when $r = 0.25$ and it is always contained below the minimax value. Note, that we have used

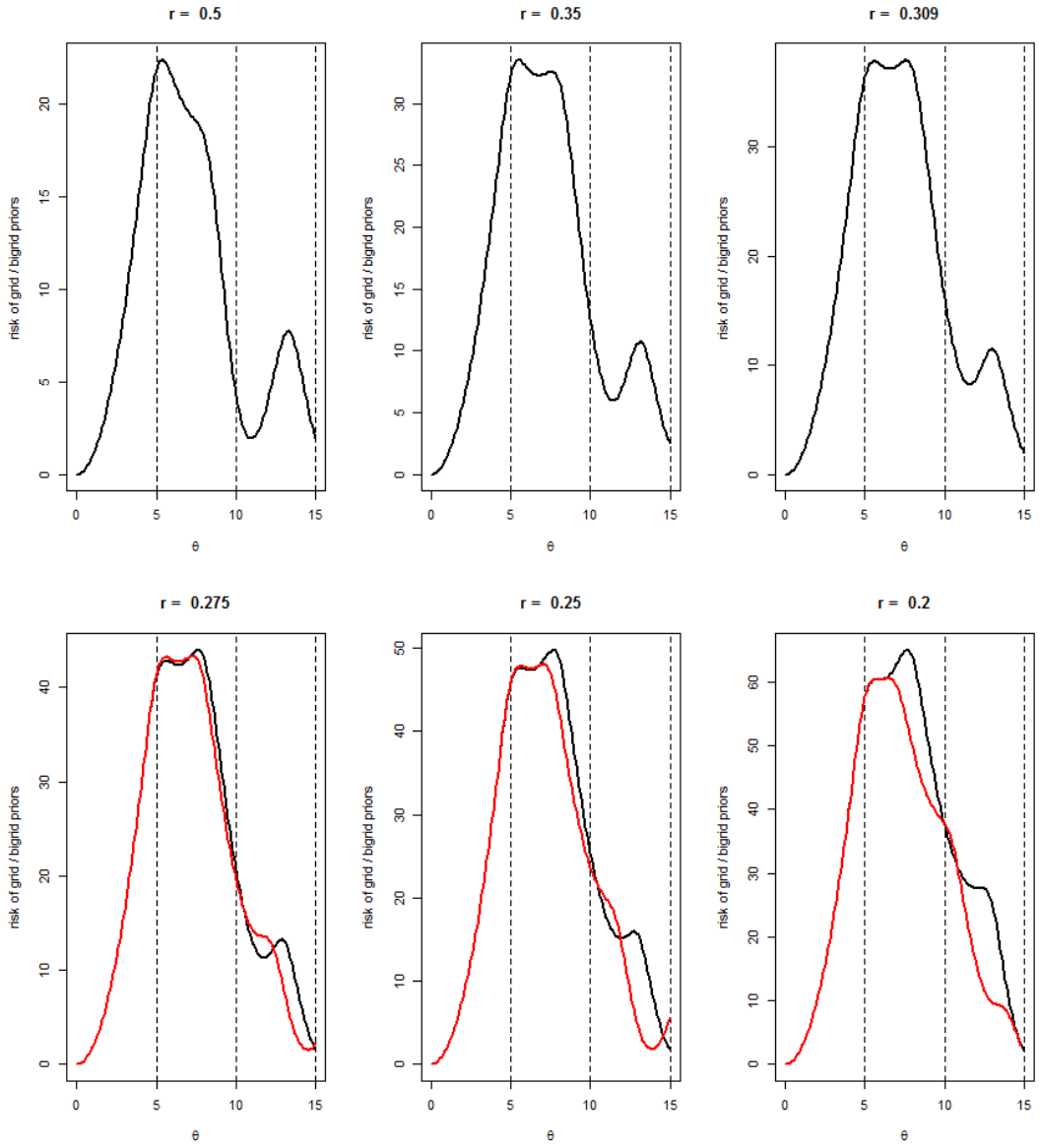

FIG 2. *Plot of the risk of $\hat{p}_G$ (in black) and $\hat{p}_B$ (in red for $r < r_0 = (\sqrt{5}-1)/4 \approx 0.309$) as the parameter θ varies over $\mathbb{R}^+$. Here, $\lambda = 5$ and $\eta = \exp(-\lambda^2/(2v))$. The vertical lines denote integer multiples of λ along the x-axis.*

different colors to display the dominant risk for different values of l and have used vertical lines to partition the θ values belonging to different values of l. These partitions are same for the top two plots as they both involve $\hat{p}_G$ but is different for $\hat{p}_B$ in the bottom plot.

Figure 5 displays the risk plots for Spike and Slab prior based pdes $\hat{p}_S[l]$. For three different choices of l the risk plots were tracked the interval $[0, t]$. It was seen that the risk is controlled below the benchmark value of $\lambda^2/(2r) = 4.5$ when $l = t$. In that case, the risk function initially increases at a quadratic rate and peaks near λ and thereafter decreases at a rapid rate.

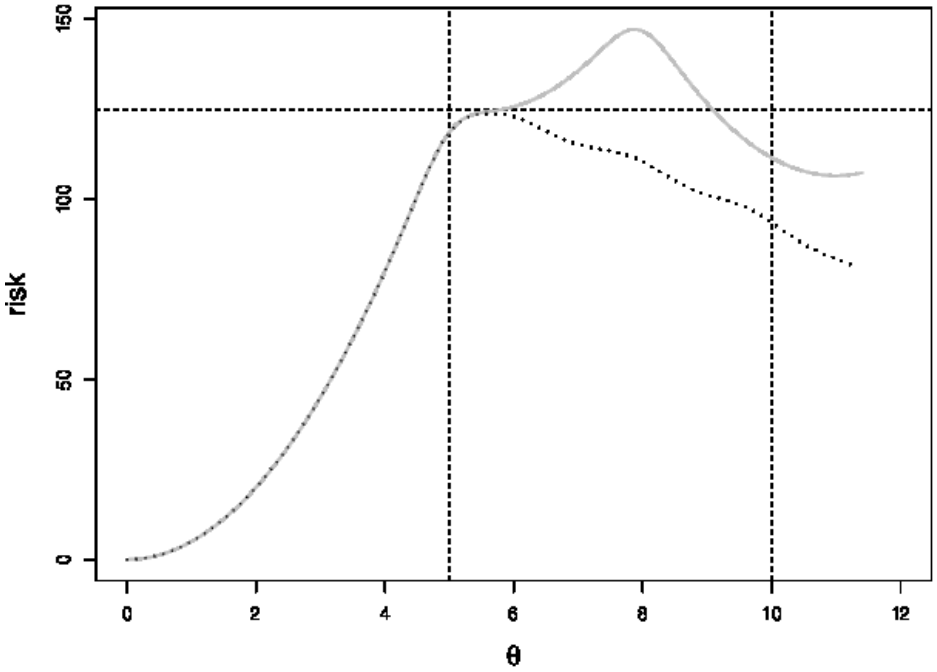

FIG 3. *In black dotted line we have the plot of $\hat{p}_B$ and in gray we have plot of $\hat{p}_G$. The vertical lines mark λ and 2λ while the horizontal line represents the bench mark $\lambda^2/(2r)$. Here, $\lambda = 5$, $r = 0.1$ and $\eta = \exp(-\lambda^2/(2v))$. The maximal risk of $\hat{p}_G$ is 1.18 times the bench mark value. The risk of $\hat{p}_B$ is controlled below the bench mark value.*

3.2. *Effect of sparsity level on the structure of the Bi-grid prior.* In the univariate Bi-grid prior, the spacings between the support points increases linearly with increase in $\lambda_n^2 = -2v\log(s_n/n)$. So, as sparsity increases, i.e., s_n/n decreases, the spacings increases. The probability assignments corresponding to the non-origin points decrease exponentially with decrease in s_n/n. However, note that b in the definition of the Bi-Grid prior depends entirely on $r = v_y/v_x$ not on the sparsity levels. So, the cardinality of the inner zone will be unchanged due to change in the sparsity levels and is still given by:

$$K = 1 + \lceil 2b^{-3/2} \rceil.$$

To see the qualitative difference, we provide the representation of the Bi-grid prior for two different sparsity levels: (a) $s_n/n = \eta_n \to 0$ (b) $s_n/n = \alpha\,\eta_n$ where α is a positive constant. Noting, that the Bi-grid prior is a product prior, we present the representation of the prior for each co-ordinate. For case (a), the univariate Bi-grid prior is $\pi_{\mathsf{B}}^{(a)} = (1-\eta_n)\delta_0 + 2^{-1}\eta_n(\nu_B^+ + \nu_B^-)$ and for case (b), $\pi_{\mathsf{B}}^{(b)} = (1-\alpha\,\eta_n)\delta_0 + 2^{-1}\alpha\,\eta_n(\tilde{\nu}_B^+ + \tilde{\nu}_B^-)$ where,

$$\nu_B^+ = c\sum_{k=1}^{\infty} a_k \delta_{\nu_k} \text{ and } \tilde{\nu}_B^+ = \tilde{c}\sum_{k=1}^{\infty} \tilde{a}_k \delta_{\tilde{\nu}_k}$$

are related as:

$$\log(\tilde{a}_k/a_k) = \begin{cases} (k-1)vb^2\log\alpha & \text{for } k = 1,\ldots,K \\ \{(K-1)b^2 + (k-K)\}\,v\,\log\alpha & \text{for } k > K\ . \end{cases}$$

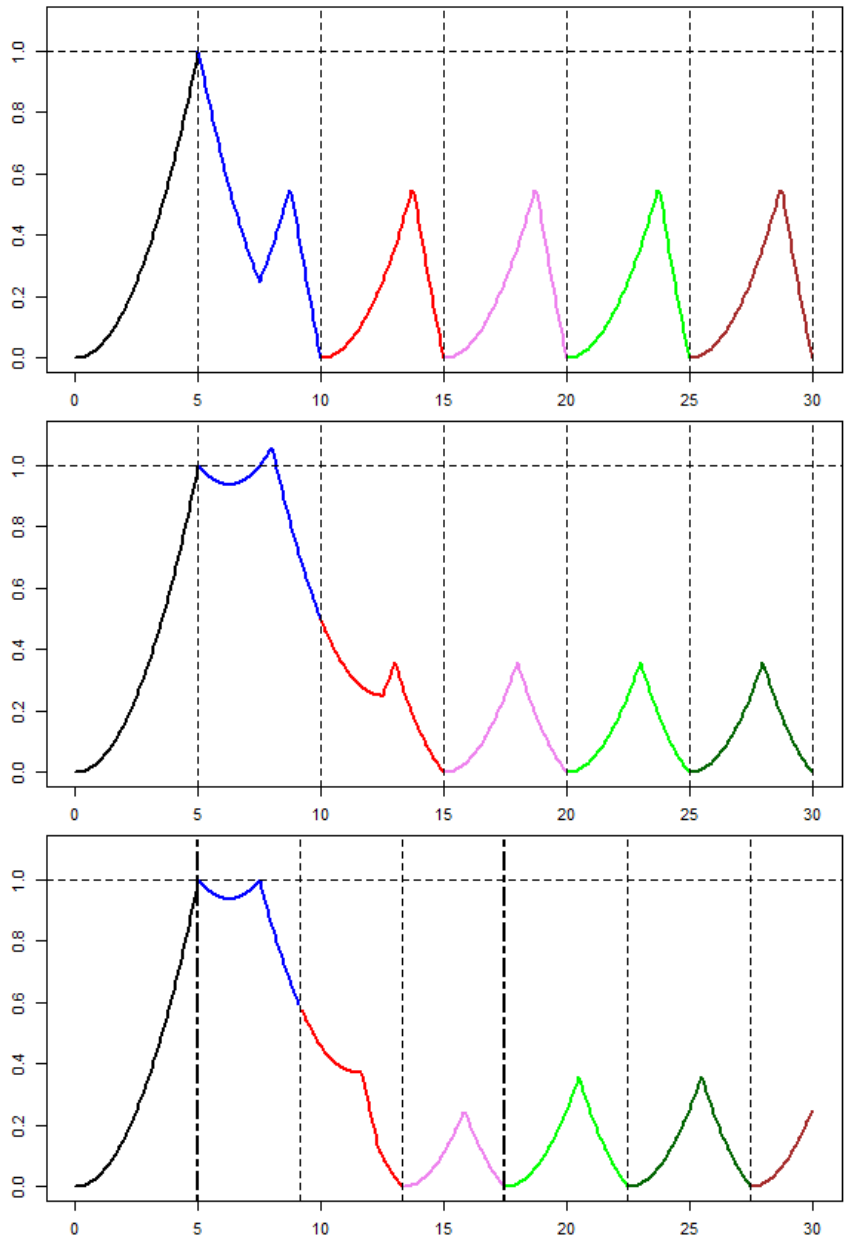

FIG 4. *For $r = 1$, the top plot displays the dominant risk $\sigma(l,\omega)$ of $\hat{p}_G$ in the y-axis as $|\theta| = \lambda(\alpha_l + \omega)$ varies in the x-axis. The next two plots show the dominant risk of $\hat{p}_G$ and $\hat{p}_B$ respectively for $r = 0.25$. For each plot, the benchmark minimax value (adjusted, as in (47)) of 1 is represented by the dotted horizontal line. The vertical lines partition $|\theta|$ into intervals corresponding to the different values l (based on the co-ordinate system for θ used in Section 3 of main paper; the dashed vertical lines denotes the boundaries for differential spacings) and different colors are used to display the dominant risk in these partitions. Here, $\lambda = 5$ for all the 3 plots.*

and $\log(\tilde{c}/c) \to 1$ as $s_n/n \to 0$. The support points satisfy the following relation with $\tilde{\Delta}_k := (\tilde{\nu}_{k+1} - \tilde{\nu}_k)$ and $\Delta_k := (\nu_{k+1} - \nu_k)$:

$$\tilde{\Delta}_k^2 - \Delta_k^2 = \begin{cases} -2vb^2 \log \alpha & \text{for } k = 1, \ldots, K-1 \\ -2v \log \alpha & \text{for } k = 0 \text{ or } k > K \end{cases}$$

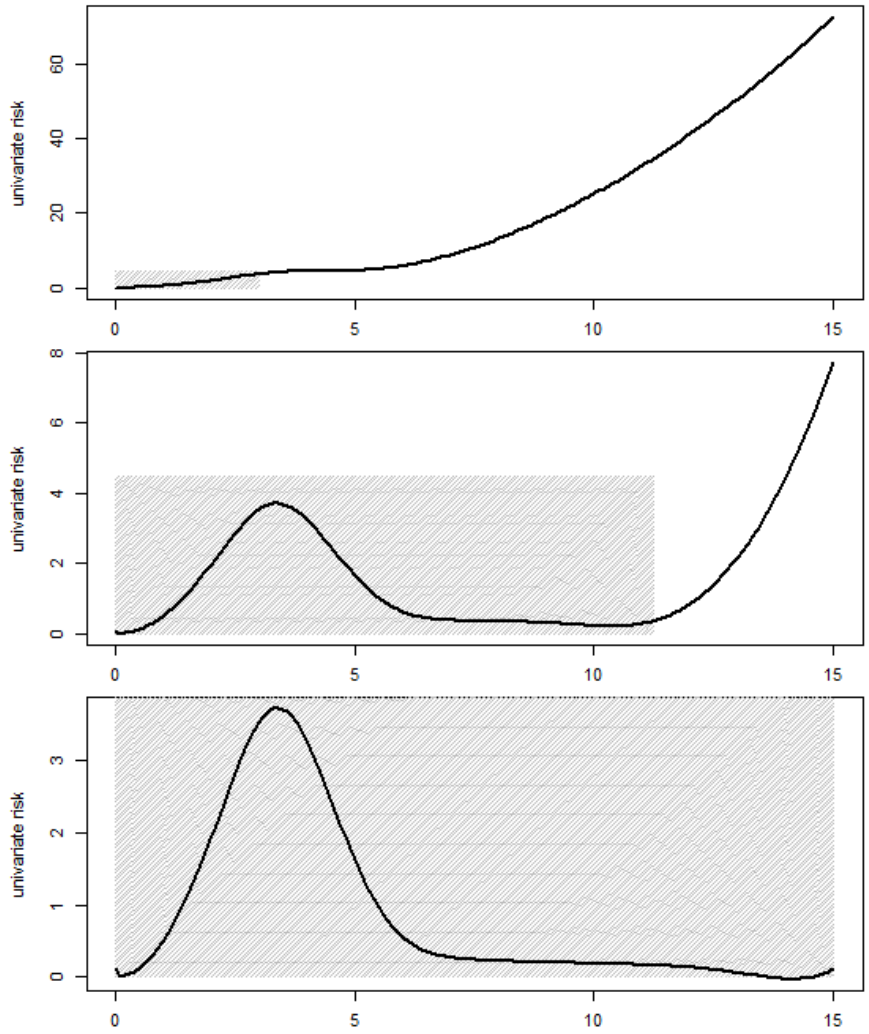

FIG 5. *From top to bottom we have the plots of the risk of* $\hat{p}_S[\ell]$ *as* $|\theta|$ *varies over the interval* $[0, t]$ *for* $\ell = \lambda, 0.75t, t$ *respectively. Here,* $r = 1$*,* $\lambda = 3$ *and* $t = 5\lambda$*. The shaded rectangle in each plot has base on the support of the prior* $[0, \ell]$ *and has height equaling the minimax benchmark value of* $\lambda^2/(2r) = 4.5$*.*

Note that $\alpha < 1 \Leftrightarrow \tilde{a}_k < a_k$ and $\tilde{\nu}_k > \nu_k$. Thus, greater sparsity corresponds to more widely spaced support points, each with reduced probability mass.

4. Additional numerical results.

4.1. *Table 1.* Table 1 shows the locations of the maxima of the pdes discussed in Section 5.

4.2. *Sensitivity of predictive performance .* We conduct simulations to assess sensitivity under small misspecifications of the sparsity level. To do this, we compare predictive performance of the prdes when sparsity levels are not known but instead are estimated from

Sparsity	r	Asymp	H-Plugin	C-Thresh	Grid	Bi-Grid	SS
0.1	1	1.52	2.46	2.50	2.47	2.47	2.41
	0.5	1.24	2.41	2.56	2.16	2.16	2.10
	0.25	0.96	2.37	2.62	1.83	1.83	1.78
	0.1	0.65	2.33	2.70	1.41	1.35	1.31
0.001	1	2.63	3.39	3.00	3.12	3.12	3.12
	0.5	2.15	3.39	2.91	2.84	2.84	2.75
	0.25	1.66	3.39	2.28	2.38	2.28	2.29
	0.1	1.12	3.39	1.71	1.72	1.69	1.70
1E-10	1	4.80	5.93	0.00	4.90	4.90	5.06
	0.5	3.92	5.93	4.41	4.35	4.35	4.36
	0.25	3.03	5.93	3.45	4.39	3.90	3.57
	0.1	2.05	5.92	2.52	3.05	2.54	2.57

TABLE 1

Numerical evaluation of the location of maxima of the risk plots over $[-\ell, \ell]$ for the different univariate predictive densities as the degree of sparsity (η) and predictive difficulty r varies. Here, we have chosen $\ell = 5\lambda$, where λ is defined in (11). In 'Asymp' column we report the theoretically obtained first order asymptotic maxima.

datasets with simulated variation. The departures of these estimated sparsity levels from the 'ground truth' provide the misspecifications whose effect we assess.

Estimated sparsities: We use the EbayesThresh package [2] to generate data and estimate sparsity levels. A product sparse prior whose marginals are a mixture of an atom of probability at zero and the laplace prior with scale $a = 0.5$ is considered. The mixing weight is estimated by maximizing the log-likelihood. Three different estimates $\hat{\eta}_{\mathsf{med}}, \hat{\eta}_{\mathsf{bf}}, \hat{\eta}_{\mathsf{mix}}$ of the sparsity level are then used:
(a) For point estimation [1] suggests using a threshold $\hat{\lambda}_{\mathsf{med}}$ based on the posterior median of the above prior. For known sparsity level η_n the ideal threshold in point estimation is $\lambda_{\mathsf{Ideal}} = \sqrt{2 \log \eta_n^{-1}}$. So, we use $\hat{\eta}_{\mathsf{med}} = \exp(-\hat{\lambda}^2_{\mathsf{med}}/2)$ as an estimate of the unknown sparsity level.
(b) Derive the threshold choice $\hat{\lambda}_{\mathsf{bf}}$ based on the Bayes factor as in [2] and use $\hat{\eta}_{\mathsf{bf}} = \exp(-\hat{\lambda}^2_{\mathsf{bf}}/2)$ as an estimate of the unknown sparsity level.
(c) Use the probability corresponding to the non-origin support points in the estimated mixture prior as an estimate for the ℓ_0 sparsity of the parametric space: $\hat{\eta}_{\mathsf{mix}} = (2 + \beta(\hat{\lambda}_{\mathsf{bf}}))^{-1}$ where $\beta(w) = g(w)/\phi(w) - 1$ where g is the marginal density of the mixture prior.

Performance comparison. We study the KL risk of the prdes described in Section 5. We consider prde in the following three regimes each of which had $n = 10000$:
(a) $\eta_n = 0.1$. Thus $\boldsymbol{\theta}$ has 1000 non-zero values.
(b) $\eta_n = 0.01$. Thus, $\boldsymbol{\theta}$ has 100 non-zero values.
(c) $\eta_n = 0.001$. Thus, $\boldsymbol{\theta}$ has 10 non-zero values.
The noise variance was set to 1 in both cases.

We write $\hat{p}_{\mathsf{E}}[\eta]$ as shorthand for the estimators $\hat{p}_{\mathsf{E}}(y|x)$ for $\mathsf{E} \in \{\mathsf{H}, \mathsf{T}, \mathsf{G}, \mathsf{B}, \mathsf{S}\}$ and with the sparsity η now shown explicitly. When the sparsity is estimated as described above, we write $\hat{p}_{\mathsf{E}}[\hat{\eta}_{\mathsf{e}}]$ for $\mathsf{e} \in \{\mathsf{med}, \mathsf{bf}, \mathsf{mix}\}$.

When the sparsity $\eta_n = s_n/n$ is estimated, we evaluate the maximal average KL risk: $\rho_{\mathsf{M}}(\hat{p}) := (1 - \eta_n)\rho(0, \hat{p}) + \eta_n \sup_\theta \rho(\theta, \hat{p})$ for $\hat{p} = \hat{p}_{\mathsf{E}}[\hat{\eta}_{\mathsf{e}}]$. When $\eta_n = s_n/n$ is known, the maximum risks of $\hat{p}_{\mathsf{E}}[\eta_n]$ were computed in Table 1 of the main paper (where they are reported relative to the theoretical minimax risk $n^{-1} R_N(\Theta[s_n])$). In tables 2, 3 and 4, corre-

TABLE 2

Numerical evaluation of the performance of the different prdes *under unknown sparsity level. As r varies, the maximal KL risk for 3 different sparsity estimation methods are reported as ratios to the maximal KL risk of these prdes under known sparsity. Here, $\eta_n = 0.100$ and $n = 10000$.*

Sparsity	r	H-Plugin	C-Thresh	Grid	Bi-Grid	SS
$\hat{\eta}_{\mathsf{med}}$	1	1.139	0.940	1.106	1.106	1.135
	0.5	1.182	0.957	1.128	1.128	1.125
	0.25	1.213	0.936	1.088	1.112	1.131
	0.1	1.240	0.886	1.114	1.158	1.098
$\hat{\eta}_{\mathsf{bf}}$	1	1.127	0.942	1.104	1.104	1.118
	0.5	1.167	0.958	1.121	1.121	1.111
	0.25	1.196	0.943	1.083	1.115	1.121
	0.1	1.221	0.887	1.104	1.145	1.097
$\hat{\eta}_{\mathsf{mix}}$	1	1.008	1.207	0.885	0.885	0.899
	0.5	0.897	1.158	0.909	0.909	0.910
	0.25	0.824	1.198	0.880	0.890	0.934
	0.1	0.766	1.264	0.896	0.906	0.939

TABLE 3

Numerical evaluation of the performance of the different prdes *under unknown sparsity level. As r varies, the maximal KL risk for 3 different sparsity estimation methods are reported as ratios to the maximal KL risk of these prdes under known sparsity. Here, $\eta_n = 0.010$ and $n = 10000$.*

Sparsity	r	H-Plugin	C-Thresh	Grid	Bi-Grid	SS
$\hat{\eta}_{\mathsf{med}}$	1	1.346	0.949	1.237	1.237	1.251
	0.5	1.355	0.987	1.280	1.280	1.249
	0.25	1.363	0.967	1.249	1.259	1.229
	0.1	1.371	0.989	1.240	1.217	1.198
$\hat{\eta}_{\mathsf{bf}}$	1	1.345	0.941	1.231	1.231	1.251
	0.5	1.354	0.982	1.269	1.269	1.244
	0.25	1.362	0.966	1.258	1.260	1.230
	0.1	1.370	0.985	1.239	1.222	1.188
$\hat{\eta}_{\mathsf{mix}}$	1	0.911	1.111	0.947	0.947	0.940
	0.5	0.905	1.102	0.957	0.957	0.947
	0.25	0.901	1.117	0.944	0.961	0.963
	0.1	0.897	1.110	0.957	0.960	0.990

sponding to sparsities $\eta_n = 0.1, 0.01$ and 0.001, we report the ratios

$$\frac{\rho_{\mathsf{M}}(\hat{p}_{\mathsf{E}}[\hat{\eta}_{\mathsf{e}}])}{\rho_{\mathsf{M}}(\hat{p}_{\mathsf{E}}[\eta_n])}.$$

for $\mathsf{e} \in \{\mathsf{med}, \mathsf{bf}, \mathsf{mix}\}$ and for $\mathsf{E} \in \{\mathsf{H}, \mathsf{T}, \mathsf{G}, \mathsf{B}, \mathsf{S}\}$. These ratios reflect the change in the maximum risk of the respective prdes when sparsity level is estimated compared to when it was known.

We see that $\hat{p}_{\mathsf{B}}$ and $\hat{p}_{\mathsf{G}}$ perform similarly under misspecification of sparsity levels and across all the studied regimes: their maximum risks are not too far from their respective maximum risks under known sparsity. While the other prdes also do not have much inflated maximum risk, recall that some of them have much higher comparative worst-case risk under known sparsity. Thus, their maximum risks under misspecification are also much higher than that of the grid and bi-grid priors. This is evident if tables 2-4 here are read along with table

TABLE 4

Numerical evaluation of the performance of the different prdes *under unknown sparsity level. As r varies, the maximal KL risk for 3 different sparsity estimation methods are reported as ratios to the maximal KL risk of these prdes under known sparsity. Here, $\eta_n = 0.001$ and $n = 10000$.*

Sparsity	**r**	**H-Plugin**	**C-Thresh**	**Grid**	**Bi-Grid**	**SS**
$\hat{\eta}_{\mathsf{med}}$	1	1.359	0.978	1.267	1.267	1.286
	0.5	1.363	1.054	1.285	1.285	1.266
	0.25	1.367	0.987	1.295	1.284	1.246
	0.1	1.371	1.017	1.271	1.242	1.202
$\hat{\eta}_{\mathsf{bf}}$	1	1.359	0.986	1.263	1.263	1.284
	0.5	1.363	1.055	1.283	1.283	1.274
	0.25	1.367	0.991	1.296	1.280	1.247
	0.1	1.371	1.019	1.275	1.240	1.200
$\hat{\eta}_{\mathsf{mix}}$	1	0.972	1.013	0.974	0.974	0.989
	0.5	0.972	1.082	0.990	0.990	0.989
	0.25	0.972	1.013	0.988	0.990	0.988
	0.1	0.971	1.045	0.993	0.990	0.983

1 of the main paper. It may seem curious that $\hat{p}_B$ and $\hat{p}_G$ outperform versions with known η_n for method $\mathsf{e} = \mathsf{mix}$: we interpret this a "finite-η" effect, especially as it diminshes as $\eta \to 0$.

In table 5, we report the average of the estimation error $(\hat{\eta}_{\mathsf{e}} - \eta_n)/\eta_n$. Methods med and bf show comparable negative bias while mix has positive bias. For med and bf, all prdes except C-Thresh had higher maximum risks while their maximal risk is lower when the mix estimate of sparsity levels is used. C-Thresh has comparatively higher risk at the origin than the other prdes and so, unlike others it peformed better when sparsity levels were underestimated than overestimated.

TABLE 5

Relative estimation errors of the sparsity level averaged over replications

η_n	med	bf	mix
0.1	-0.4770	-0.4504	1.0642
0.01	-0.7985	-0.7981	0.6248
0.001	-0.9000	-0.9000	0.2028

In summary, the estimated sparsity levels $\hat{\eta}_{\mathrm{e}}$ can show significant departures both above and below the targets η_n, thus providing reasonable 'perturbations', but the deterioration of the maximum risks of the grid and bi-grid priors (among others) remains well controlled.

REFERENCES

[1] JOHNSTONE, I. M. and SILVERMAN, B. W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Annals of Statistics* 1594–1649.

[2] JOHNSTONE, I. M. and SILVERMAN, B. W. (2005). EbayesThresh: R and S-Plus programs for Empirical Bayes thresholding. *Journal of Statistical Software*.