Convergence et spike and Slab Bayesian posterior distributions in some high dimensional models

Romain Mismer

Convergence of Spike and Slab Bayesian posterior distributions in some high dimensional models.

Romain Mismer

Laboratoire de Probabilités, Statistique et Modélisation - UMR 8001
Université de Paris Diderot

Thèse soutenue publiquement le 12 juin 2019 pour l'obtention du grade de :
Docteur de l'Université Sorbonne Paris Cité

Sous la direction de : Ismaël Castillo *Professeur* (Sorbonne Université)

Rapportée par : Vincent Rivoirard
Aad van der Vaart

Jury composé de : Stéphane Boucheron *(Président du jury) Professeur* (Paris Diderot)
Ismaël Castillo *Professeur* (Sorbonne Université)
Vincent Rivoirard *Professeur* (CEREMADE, Université Paris Dauphine)
Aad van der vaart *Professeur* (Leiden University)
Pierre Alquier *Professeur* (ENSAE)
Julyan Arbel *Chargé de recherches* (INRIA Grenoble)
Cristina Butucea *Professeur* (ENSAE)

École Doctorale de Sciences Mathématiques
de Paris-Centre
Section Mathématiques appliquées

12 juin 2019

Remerciements

Lourde tâche que de remercier tous les gens qui ont pu m'accorder leur soutien ces dernières années, j'espère n'oublier personne. Et toi que j'aurais oublié, sache que je te remercie quand même.

La première personne que je souhaite remercier est bien sûr Ismaël, mon directeur de thèse, qui a toujours su faire preuve de gentillesse, de patience et a su être un soutien inconditionnel même dans mes moments les plus difficiles (et Dieu sait qu'il y en a eu). Sa passion pour les mathématiques et son extrême rigueur sont ce qui m'ont le plus marqué chez lui et j'aimerais toujours m'en inspirer à l'avenir, et il y a en réalité peu de mots pour dire toute l'admiration que j'ai pour lui, à la fois sur le plan scientifique et le plan humain.

Je remercie également les personnes qui ont accepté de faire partie de mon jury : Cristina Butucea, Julyan Arbel, Pierre Alquier, mes rapporteurs Aad van der Vaart et Vincent Rivoirard pour avoir eu la gentillesse de lire en détails ma thèse, et aussi tout particulièrement Stéphane Boucheron qui aura eu pour moi le rôle d'un second mentor au long de ma thèse.

Je souhaite aussi remercier mes collègues de Paris Diderot et particulièrement Ratha qui fût un peu une seconde mère pour moi et les autres membres de notre bureau Zakaria, Dominique et Inès que je remercie pour les bons moments passés ensemble, Ziaad avec qui on aura échangés quelques (nombreuses) heures d'enseignement et Yann, dont l'amitié remonte à bien plus loin que le début de ma thèse. Je n'oublie pas également Raphaël Lefevere et les élèves de MIASHS avec qui j'aurai passé de bons moments d'enseignement.

Merci aussi à tous les affamés de Jussieu pour tous ces bons moments passés au restaurant des personnels, appelé plus sobrement "la cantine", à savoir Eric, Alexandre, Nicolas, Nicolas, Rancy, Thibault, Michel, Paul, Carlo, Henri, Sarah, Olga, PA et d'autres bien sûr. Mes profondes amitiés aux membres du bureau 203 (où le bon mot du jour aura été chaque jour : "j'ai faim") qui devraient se reconnaître : Burrito pour toutes les distractions (mais aussi conseils et astuces scientifico-informatiques, je dois le rappeler quand même) dont tu as eu connaissance et que tu as bien sûr partagées, Léo dit Rasemoohret pour avoir été la garantie du bon goût dans ce bureau, un Zèbre Lambda

aussi (que je dédeste en fait), ainsi que l'homme sage du bureau dont aucune blague ne sera tombée à l'eau (de Co...) j'ai nommé.. Curve ? Tant que j'y suis merci aussi aux Missplayers pour leurs.. listes et leurs mythiques : Max dit 27, Clément, Charles, Hadrien ainsi qu'Arnaud.

J'ai une pensée aussi évidemment pour mes anciens camarades de Master/magistère à Orsay (et pour nos enseignants, en particulier Elisabeth Gassiat et Frédéric Paulin): Jeanne, Dimitri, Eugène, Younès, Florent, David et Caroline, à qui s'ajoutent aussi MF et Jérémy. Merci aussi à tous mes collocataires (anciens comme nouveaux) de m'avoir supporté (et réciproquement) : Thibault, Camille, François, Florine et sa maman Christine (qui aime beaucoup trop Disneyland)... et bien sûr plus particulièrement Arthur parce qu'après toutes ces années, on en a gros.

Côté Est, un merci pêle-mêle également à Aline, Toto, Laura, Nicolas, Malau, Zaza, Héléna et Charlène. A Chantal, Caroline et Florence.

A ma (nombreuse) famille, à Sylvie, Denis, Lulu, Sylvie, Pascale, à toutes mes tatans, mes tontons, mes cousines, mes cousins, et les mômes !

A Cécile.

Abstract

Title : Convergence of Spike and Slab Bayesian posterior distributions in some high dimensional models.

The first main focus is the sparse Gaussian sequence model. An Empirical Bayes approach is used on the Spike and Slab prior to derive minimax convergence of the posterior second moment for Cauchy Slabs and a suboptimality result for the Laplace Slab is proved. Next, with a special choice of Slab convergence with the sharp minimax constant is derived. The second main focus is the density estimation model using a special Pólya tree prior where the variables in the tree construction follow a Spike and Slab type distribution. Adaptive minimax convergence in the supremum norm of the posterior distribution as well as a nonparametric Bernstein-von Mises theorem are obtained.

Keywords: Bayesian nonparametrics, Spike and Slab prior, thresholding, Pólya tree, Bernstein-von Mises theorems.

Résumé

Titre : Convergence de lois a posteriori Spike and Slab bayésiennes dans des modèles de grande dimension.

On s'intéresse d'abord au modèle de suite gaussienne parcimonieuse. Une approche bayésienne empirique sur l'a priori Spike and Slab permet d'obtenir la convergence à vitesse minimax du moment d'ordre 2 a posteriori pour des Slabs Cauchy et on prouve un résultat de sous-optimalité pour un Slab Laplace. Un meilleur choix de Slab permet d'obtenir la constante exacte. Dans le modèle d'estimation de densité, un a priori arbre de Pólya tel que les variables de l'arbre ont une distribution de type Spike and Slab donne la convergence à vitesse minimax et adaptative pour la norme sup de la loi a posteriori et un théorème Bernstein-von Mises non paramétrique.

Mots-clé: Bayésien non paramétrique, a priori Spike and Slab, seuillage, arbre de Pólya, théorèmes Bernstein-von Mises.

Contents

Résumé détaillé

Ce document rassemble les travaux que j'ai effectués sous la direction d'Ismaël Castillo pendant la durée de ma thèse centrée sur l'utilisation dans un cadre bayésien de l'a priori Spike and Slab dans des modèles de dimension grande ou infinie, et des propriétés asymptotiques qui en découlent. Ce travail est divisé en 4 chapitres, un chapitre introductif et 3 chapitres qui font l'objet d'articles (un paru pour le deuxième chapitre, et deux à soumettre pour les suivants).

0.0.1 Analyse par bayésien empirique de lois a posteriori Spike and Slab.

On considère le modèle de suite gaussienne parcimonieuse, où l'on observe $X_1, \dots, X_n$ des variables aléatoires telles que pour tout $i \in \{1, \dots, n\}$

$$X_i = \theta_i + \varepsilon_i$$

avec le bruit ε tel que ses coordonnées ε_i suivent la loi normale standard (de densité notée ϕ) et $\theta \in \mathbb{R}^n$ le paramètre à estimer. On suppose que ce paramètre θ est parcimonieux, c'est-à-dire qu'il appartient à la classe $\ell_0[s_n]$ suivante :

$$\ell_0[s_n] = \{\theta \in \mathbb{R}^n, \#\{i; \theta_i \neq 0\} \leq s_n\}$$

avec $(s_n)_n$ une suite qui tend vers l'infini mais telle que $s_n/n \to 0$ quand $n \to \infty$. On considère la convergence de lois a posteriori bayésiennes de lois a priori Spike and Slab :

$$\Pi_\alpha = \prod_{i=1}^{n} (1-\alpha)\delta_0 + \alpha\Gamma,$$

où Γ est une loi à densité notée γ sur $\mathbb{R}$. La famille de lois Π_α permet de modéliser des vecteurs parcimonieux grâce au paramètre de parcimonie $\alpha \in]0; 1[$. Ce paramètre est

calibré par une approche bayésienne empirique : on le remplace par un estimateur $\hat{\alpha}$ construit en maximisant la vraisemblance marginale bayésienne empirique :

$$\prod_{i=1}^{n}((1-\alpha)\phi(X_i)+\alpha\phi*\gamma(X_i)).$$

Johnstone and Silverman (2004) ont montré que la médiane a posteriori avec plug-in de $\hat{\alpha}$ converge à vitesse optimale au sens minimax pour la perte quadratique sur la classe des vecteurs parcimonieux $\ell_0[s_n]$, dès que la loi Γ (dite Slab) a des queues de distribution au moins Laplace.

Dans ce travail, on considère la loi a posteriori plug-in complète $\Pi_{\hat{\alpha}}(\cdot|X)$. On s'intéresse principalement au moment d'ordre 2 a posteriori

$$\int \|\theta-\theta_0\|^2 d\Pi_{\hat{\alpha}}(\theta|X).$$

On montre que, sous certaines conditions sur Γ, le moment d'ordre 2 a posteriori converge lui aussi à vitesse minimax optimale pour la perte quadratique. De façon surprenante, ce n'est pas le cas pour Γ la loi Laplace : on montre qu'il est en effet nécessaire que Γ ait des queues polynomiales (plus lourdes que x^{-3}, par exemple Γ Cauchy) pour que le moment d'ordre 2 a posteriori converge à vitesse optimale. On montre que cette sous-optimalité pour un Slab Laplace n'est pas dûe au second moment puisqu'elle se traduit également sur la loi a posteriori entière.

Par ailleurs, on montre que des résultats similaires (à un facteur logarithmique près) sont vrais pour la classe de lois dite Spike and Slab LASSO récemment introduite par Ročková and George (2018) et Ročková (2018).

0.0.2 Constante exacte pour l'a posteriori Spike and Slab calibré par bayésien empirique.

Ce travail se situe dans le même cadre que le chapitre précédent et poursuit l'étude de la loi a posteriori plug-in complète $\Pi_{\hat{\alpha}}(\cdot|X)$. Les résultats d'optimalité évoqués ci-dessus le sont à constante près. Ainsi, pour $\hat{\theta}^{med}(X)$ la médiane a posteriori, Johnstone and Silverman (2004) montrent que pour n assez grand et pour une constante $C>0$ assez grande,

$$\sup_{\theta_0\in\ell_0[s_n]} E_{\theta_0}\left[\|\hat{\theta}^{med}(X)-\theta_0\|^2\right] \leq C s_n \log(\frac{n}{s_n})(1+o(1)).$$

Il est connu que la vitesse minimax pour ce problème est $2s_n\log(n/s_n))(1+o(1))$ quand $n\to\infty$. Il est possible de montrer que l'a posteriori Spike and Slab dans lequel on fait un

plug-in d'un paramètre α oracle fait atteindre la vitesse minimax avec constante exacte 2 au moment d'ordre 2, et ce même pour un Slab Laplace. On peut donc naturellement se demander si le second moment a posteriori avec plug-in du maximum de vraisemblance peut lui aussi converger à vitesse minimax, cette fois adaptative. On montre qu'en effet, pour un choix approprié de la loi slab Γ (celui-ci doit avoir des queues très lourdes, de l'ordre de $x^{-1}\log^{-2}(x)$), il est possible d'atteindre cette vitesse minimax exacte :

$$\sup_{\theta_0 \in \ell_0[s_n]} E_{\theta_0}\left[\int \|\theta-\theta_0\|^2 d\Pi_{\hat{\alpha}}(\theta|X)\right] \leq 2s_n \log(\frac{n}{s_n})(1+o(1)).$$

0.0.3 Estimation adaptative de densités par a priori arbres de Pólya Spike and Slab.

On se place désormais dans le modèle d'estimation de densité sur $[0;1]$, où l'on observe $X_1,\ldots,X_n$ des variables aléatoires indépendantes et identiquement distribuées de densité inconnue f. Le but de ce travail est d'étudier les propriétés d'une méthode bayésienne non-paramétrique reposant sur des lois a priori dites d'arbres de Pólya, pour l'estimation de f ainsi que l'inférence sur certains fonctionnelles de f.

Dans un travail récent, Castillo (2017b) a montré que les arbres de Pólya permettent notamment d'atteindre la vitesse optimale au sens minimax pour l'estimation de f en termes de la norme infinie avec la loi a posteriori, si les paramètres de l'arbre sont bien choisis en termes de la régularité $\beta \in]0,1]$ (au sens Hölder) de la densité f.

L'objet de ce chapitre est d'obtenir des versions adaptatives des résultats précédents. En effet, lorsque la régularité de f n'est pas connue, on montre qu'il est possible de modifier la construction d'origine de l'arbre de Pólya de façon à s'adapter automatiquement à la régularité inconnue. Pour cela, les lois Beta le long de l'arbre de la construction d'origine sont remplacées par des mélanges d'une Beta et d'une masse de Dirac en $1/2$. Pour ces arbres de Pólya Spike and Slab, on montre que la loi a posteriori converge à vitesse minimax optimale à constante près pour la norme infinie, ainsi qu'un théorème de Bernstein–von Mises non paramétrique dans un espace fonctionnel bien choisi. Du point de vue conceptuel, cette classe de lois a priori peut se voir comme un analogue des méthodes de seuillage par ondelettes, avec de plus une quantification de l'incertitude propre à l'utilisation de l'approche bayésienne. Un autre avantage conceptuel de l'approche est que, contrairement aux estimateurs par ondelettes de densités (qui ne sont pas nécessairement des densités), la loi a posteriori est ici automatiquement une densité.

Chapter 1

Introduction

1.1 General Frame : the non-parametric, frequentist Bayesian approach

1.1.1 The Bayesian approach

Take $(\mathcal{X}, \mathcal{A})$ a measurable space, where $\mathcal{A}$ is a σ-field over $\mathcal{X}$ and (Θ, d) a subset of a separable Banach space.
Consider a statistical experiment where one observes some data $X \in \mathcal{X}$, a random object whose law will be interpreted using a model, defined here as follows

$$\mathcal{P} = \{\mathbb{P}_\theta,\ \theta \in \Theta\}, \tag{1.1.1}$$

where the $\mathbb{P}_\theta$ are probability measures on $\mathcal{A}$.

The model depends on an unknown parameter θ, let us consider this parameter θ as a random variable too. Namely θ will follow the law Π, which is called the *a priori law* (or simply *prior*).

On the other hand, one views P_θ as the law of $X|\theta$. This gives us the following Bayesian diagram

$$\begin{array}{rcl} X|\theta & \sim & \mathbb{P}_\theta \\ \theta & \sim & \Pi. \end{array} \tag{1.1.2}$$

This defines the joint distribution of (θ, X), from which one can derive the law of $\theta|X$, which is called the *a posteriori law* (or simply *posterior*)

$$\theta|X \sim \Pi(\cdot|X). \tag{1.1.3}$$

We will henceforth assume that $\mathbb{P}_\theta$ and Π are absolutely continuous relatively to fixed σ-finite measures μ and ν. Denoting by f_θ and π their densities, the joint law (θ, X) has a density $h(\theta, x) = f_\theta(x)\pi(\theta)$ and X has a density $h(x) = \int_\Theta f_\theta(x)\pi(\theta)d\nu(\theta)$. Under standard measurability conditions, see pages 6-7 of Ghosal and van der Vaart (2017), Bayes' formula gives the following density for $\theta|X$

$$\Pi(\theta|X) = \frac{f_\theta(X)\pi(\theta)}{h(X)}\mathbb{1}_{h(X)>0} \tag{1.1.4}$$

In the classical approach, one generally builds a point estimator $\hat{\theta}(X) \in \Theta$. The Bayesian approach provides the user with an entire probability distribution which depends on our observations X and not just a point estimator. It also provides estimators which are "aspects" of the a posteriori law: if they exist, the mean of the a posteri law $\int \theta d\Pi(\theta|X)$, the posterior median, or the posterior mode(s) for instance. It can be used to find credible sets (which can turn out to be confidence sets), or to make tests $\mathcal{H}_0$ versus $\mathcal{H}_1$ using the quantities $\Pi(\mathcal{H}_0|X)$ and $\Pi(\mathcal{H}_1|X)$.

From now on we will assume that we have $n \in \mathbb{N}$ observations $X = X^{(n)} = (X_1, \ldots, X_n)$.

Score and Fisher Information in i.i.d. parametric models. A model $\mathcal{P}$ as above is said to be *differentiable in quadratic mean* (abbreviated *DQM*) at θ if there exists a vector l_θ (called the *score* at θ) of k functions such that, when $h \to 0$

$$\int \left(\sqrt{f_{\theta+h}} - \sqrt{f_\theta} - \frac{1}{2}h^T l_\theta \sqrt{f_\theta}\right)^2 d\mu = o(\|h\|^2) \tag{1.1.5}$$

The score is centered and has a variance I_θ which is called the *Fisher Information*. It is shown in van der Vaart (1998) that this also implies that the model is locally asymptotically normal (abbreviated LAN).

1.1.2 Frequentist Bayesian

We will follow the Frequentist approach by assuming that a true parameter θ_0 exists and has to be estimated

$$\exists \theta_0 \in \Theta \text{ such that } X \sim \mathbb{P}_{\theta_0} \tag{1.1.6}$$

The sequence $(\Pi(\cdot|X))_{n\in\mathbb{N}*}$ is said to be $\mathbb{P}_{\theta_0}$-consistent with respect to the distance d if, for every $\varepsilon > 0$ as $n \to \infty$

$$\Pi(d(\theta,\theta_0)\leq\varepsilon|X)\to 1 \text{ in } \mathbb{P}_{\theta_0}\text{-probability} \tag{1.1.7}$$

This result is equivalent to the sometimes more convenient version, denoting by $E_{\theta_0}=E_{\mathbb{P}_{\theta_0}}$ the expectation under $\mathbb{P}_{\theta_0}$

$$E_{\theta_0}[\Pi(d(\theta,\theta_0)\leq\varepsilon|X)]\to 1 \tag{1.1.8}$$

The sequence $(\Pi(\cdot|X))_{n\in\mathbb{N}*}$ will be strongly $\mathbb{P}_{\theta_0}$-consistent if the previous convergence is $\mathbb{P}_{\theta_0}$-almost surely.

Point Estimators. Let $\hat{\theta}$ be an estimator derived from the posterior (like the posterior mean $\bar{\theta}=\int\theta d\Pi(\theta|X)$ for example). One says that is $\hat{\theta}$ is consistent (uniformly in $\theta_0\in\Theta$) if, as $n\to\infty$

$$\sup_{\theta_0\in\Theta} E_{\theta_0}[d(\hat{\theta},\theta_0)]\to 0 \tag{1.1.9}$$

Minimax convergence rate. In terms of rate of convergence, one would like to build estimators converging to the true parameter 'as fast as possible' . To do so, one defines the minimax rate r_n^* over the set Θ of parameters with respect to the loss function (here a distance) d, as

$$r_n^*=\inf_{\hat{\theta}}\sup_{\theta\in\Theta} E_{\theta}[d(\theta,\hat{\theta})], \tag{1.1.10}$$

where the infimum is taken over all estimators of the parameter.
One says that $\hat{\theta}$ converges at minimax rate if there exists $N\in\mathbb{N}$ such that $\forall n\geq N$

$$\sup_{\theta_0\in\Theta} E_{\theta_0}[d(\hat{\theta},\theta_0)]\leq Cr_n^* \tag{1.1.11}$$

Actually the entire a posteriori law can converge at minimax rate (uniformly in $\theta_0\in\Theta$), namely if, as $n\to\infty$

$$\sup_{\theta_0\in\Theta} E_{\theta_0}[\Pi(d(\theta,\theta_0)\leq Cr_n^*|X)]\to 1 \tag{1.1.12}$$

Credible sets. A Credible set $\mathcal{C}=\mathcal{C}(X)$ of level $1-\gamma$ (with $\gamma\in(0,1)$) is defined as a set such that

$$\Pi(\mathcal{C}|X)=1-\gamma \tag{1.1.13}$$

One can define a credible set of level at least $1-\gamma$ by replacing the $=$ by a $\geq$ in the definition.

In general, one may want (this may not always be possible for complex models) the diameter of a credible set to be rate-optimal, in a minimax sense, as $n \to \infty$

$$\sup_{\theta_0 \in \Theta} E_{\theta_0}[Diam(\mathcal{C})] \asymp r_n^* \tag{1.1.14}$$

One would naturally ask if credible sets can be used as confidence sets, namely if

$$\liminf_{n\to\infty} \inf_{\theta_0 \in \Theta} \mathbb{P}_{\theta_0}(\theta_0 \in \mathcal{C}) \geq 1-\gamma \tag{1.1.15}$$

If $\Theta \subset \mathbb{R}^k$, it turns out that for quantile-type sets and i.i.d. data, one can positively answer that question using the following theorem

Theorem 1 (Bernstein-von Mises). Consider a model $\mathcal{P} = \{P_\theta^{\otimes n}, \theta \in \Theta\}$ such that $X_1, \dots, X_n | \theta \sim P_\theta^{\otimes n}$. Assume that the density π of the prior is positive and continuous at θ_0, the model $\mathcal{P}$ is DQM (see (1.1.5)) at the point θ_0 with an invertible Fisher Information I_{θ_0}. Assume also that for every $\varepsilon > 0$ there exists a sequence $(\phi_n)_n$ of tests such that $\lim_{n\to\infty} \mathbb{E}_{\theta_0}[\phi_n] = 0$ and $\lim_{n\to\infty} \sup_{\|\theta-\theta_0\| \geq \varepsilon} \mathbb{E}_\theta[\phi_n] = 0$. Then, as $n \to \infty$,

$$\|\mathcal{L}(\sqrt{n}(\theta-\theta_0)|X_1,\dots,X_n) - \mathcal{N}(\Delta_n(\theta_0), I_{\theta_0}^{-1})\|_{TV} = o_{\mathbb{P}_{\theta_0}}(1),$$

with $\Delta_n(\theta_0) = I_{\theta_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n l_{\theta_0}(X_i)$ and $\|\cdot\|_{TV}$ the total variation distance between two probability measures.

It can be checked that this theorem implies that, asymptotically, quantile-type credible sets built from the a posteriori law are confidence sets and have optimal diameter.

1.1.3 High and Infinite Dimension Models

Nonparametric prior distributions are harder to build and choose than in a parametric setting, as one has to define a distribution on a much larger space. Often the posterior distribution will still strongly depend on the choice of the prior distribution. One has to aim at posterior consistency at minimax rate, and it can be significantly harder than in parametric settings, where good consistency is often obtained as soon as the prior puts positive mass around the true parameter (in nonparametric setting, the precise amount of mass in vanishing neighbourhoods of the truth typically matters). The object one usually estimates in nonparametric Bayesian inference is a function or a density,

for instance through the analysis of an infinite sequence of its wavelet coefficients, and building a flexible enough prior (for instance to achieve adaptive results) will require some care. Tuning the involved parameters may also demand significantly more work than in parametric settings. Nonparametric and high-dimensional models include the Gaussian sequence model (which is the main focus of Chapters 2 and 3), the Gaussian White Noise model and the Density Estimation model (which is the main focus of Chapter 4).

Estimation. In i.i.d. settings, Ghosal, Ghosh and van der Vaart developped a general framework to derive posterior rates with respect to certain distances on the parameter space (later generalised in Ghosal and van der Vaart (2007) to non i.i.d. settings)

Theorem 2 (Ghosal et al. (2000)). Let $\Pi = \Pi_n$ be a sequence of a priori laws and assume that X are i.i.d. with density f_{θ_0}. Let ε_n be a sequence of positive reals such that $\varepsilon_n \to 0$ and $\sqrt{n}\varepsilon_n \to \infty$ as $n \to \infty$.
Assume the existence of some constants C and L such that

$$\Pi\left(\theta \in \Theta; -E_{\theta_0}[\log(\frac{f_\theta}{f_{\theta_0}}(X))] \leq \varepsilon_n^2, E_{\theta_0}[\log(\frac{f_\theta}{f_{\theta_0}}(X))^2] \leq \varepsilon_n^2\right) \geq e^{-Cn\varepsilon_n^2}$$

and

$$\Pi(\Theta \setminus \Theta_n) \leq Le^{-(C+4)n\varepsilon_n^2}$$

for a sequence $\Theta_n \subset \Theta$ such that there exist tests $\psi_n = \psi(X_1, \ldots, X_n)$ such that $\forall n \in \mathbb{N}$ and $M > 0$ large enough

$$E_{\theta_0}[\psi_n] \to 0 \text{ and } \sup_{\theta \in \Theta_n; d(f_\theta, f_{\theta_0}) \geq M\varepsilon_n} E_\theta[1 - \psi_n] \leq Le^{-(C+4)n\varepsilon_n^2}$$

Then $\Pi(d(f_\theta, f_{\theta_0}) > M\varepsilon_n | X) \to 0$ as $n \to \infty$ in $\mathbb{P}_{\theta_0}$-probability.

This result provides qualitative conditions such as the existence of tests (an entropy condition via ε-covering numbers of the Θ_n can also be used) for the minimax convergence of the a posteriori law. Directly using this theorem may be delicate to get more precise conditions on some a priori laws (such as the Spike and Slab introduced in the following section) or for some choices of metric. In the cases where no analog of this result have been proven one sometimes needs to use more direct reasonings on the posterior distribution.

Nonparametric Bernstein-von Mises and Uncertainty Quantification. A nonparametric Bernstein-von Mises result would take the following form :

$$\mathcal{L}(\sqrt{n}(\theta - T_n)|X) \to \mathcal{D} \tag{1.1.16}$$

where one has to ask several questions, whose answers may be unclear at first but certainly depend on the situation. Firstly, what is the limiting distribution $\mathcal{D}$? Secondly, what is the sense of the convergence in the result? Finally, which centering estimator T_n do we choose to get this convergence result?

Let us consider the Gaussian White Noise model as an example. For $f \in L^2([0,1])$, $t \in [0,1]$ and dW the standard Gaussian White Noise, the model is

$$dX^{(n)}(t) = f(t)dt + \frac{1}{\sqrt{n}}dW(t). \tag{1.1.17}$$

If one chooses a wavelet basis $\phi, (\psi_{lk})_{l\in\mathbb{N},0\leq k<2^l}$ (say the Haar basis to fix ideas), using the notation $f_{lk} = \langle f, \psi_{lk}\rangle = \int_0^1 f(t)\psi_{lk}(t)dt$, one can write (setting $\psi_{-1,-1/2} = \phi$ and letting $l \geq -1$ in what follows)

$$\int_0^1 \psi_{lk}(t)dX^{(n)}(t) = \int_0^1 \psi_{lk}(t)f(t)dt + \frac{1}{\sqrt{n}}\int_0^1 \psi_{lk}(t)dW(t)$$

that we can rewrite $\mathbb{X}_{lk} = f_{lk} + \frac{1}{\sqrt{n}}\varepsilon_{lk}$.
One now has

$$\mathbb{X}^{(n)} = f + \frac{1}{\sqrt{n}}\mathbb{W}$$

so, as $\sqrt{n}(\mathbb{X}^{(n)} - f) = \mathbb{W}$, one would naturally take the centering $T_n = \mathbb{X}^{(n)}$ in (1.1.16) and the limiting distribution $\mathcal{D} = \mathcal{L}(\mathbb{W}) := \mathcal{N}$ the law of white noise. We set $\tau : f \mapsto \sqrt{n}(f - \mathbb{X}^{(n)})$ and denote by Π_n the shifted posterior distribution $\Pi(\cdot|\mathbb{X}^{(n)}) \circ \tau$.
Recall also that, by definition of white noise, $\forall f, g \in L^2([0,1])$, one has $E[\mathbb{W}(f)\mathbb{W}(g)] = \langle f, g\rangle$.

To establish a nonparametric BVM result, one has to consider larger spaces (here larger than $L^2([0,1])$) as one needs a $1/\sqrt{n}$ rate that can only be achieved with weaker metrics. The impossibility to obtain a BVM result in L^2 has been shown by Cox (1993) and Freedman (1999). Consider, for $s > 0$ the Sobolev space H_2^{-s} defined as

$$H_2^{-s} = \{f; \|f\|_{s,2}^2 = \sum_{l\geq 0} 2^{-2ls}\sum_{k=0}^{2^l-1} |\langle \psi_{lk}, f\rangle|^2 < \infty\} \tag{1.1.18}$$

For every $s > 0$, $L^2 \subset H_2^{-s}$. Now, one builds a 'logarithmic' Sobolev space to be the 'smallest' containing $\mathbb{W}$, somewhat taking the limiting case $s = 1/2$. For that, one usually uses an 'admissible' sequence $\omega = (\omega_l)_{l\geq 0}$. Here we take, for $\delta > 1$, $\omega_l = l^{2\delta}$ and set

$$H(\omega) = \{f; \|f\|_\omega^2 = \sum_{l\geq 0} \frac{2^{-l}}{\omega_l}\sum_{k=0}^{2^l-1} |\langle \psi_{lk}, f\rangle|^2 < \infty\} \tag{1.1.19}$$

This set was built to ensure that $\mathbb{W}$ belongs to it, as for $\delta > 1/2$,

$$\begin{aligned} E[\|\mathbb{W}\|_\omega^2] &= \sum_{l\geq 0} \frac{2^{-l}}{\omega_l} \sum_{k=0}^{2^l-1} E[\varepsilon_{lk}^2] \\ &\leq \sum_{l\geq 0} \frac{2^{-l}}{\omega_l} 2^l \leq \sum_{l\geq 0} l^{-2\delta} < \infty \end{aligned}$$

We now have to state the convergence in (1.1.16). For that, we use the following metric.

Bounded Lipschitz metric. Let $(\mathcal{S}, d)$ be a metric space. The bounded Lipschitz metric $\beta_{\mathcal{S}}$ on probability measures of $\mathcal{S}$ is defined as follows, for any μ, ν probability measures of $\mathcal{S}$

$$\beta_{\mathcal{S}}(\mu, \nu) = \sup_{F; \|F\|_{BL}\leq 1} \left| \int_{\mathcal{S}} F(x)(d\mu(x) - d\nu(x)) \right|, \tag{1.1.20}$$

where $F : \mathcal{S} \to \mathbb{R}$ and

$$\|F\|_{BL} = \sup_{x\in\mathcal{S}} |F(x)| + \sup_{x\neq y} \frac{|F(x) - F(y)|}{d(x, y)}. \tag{1.1.21}$$

This metric metrizes the convergence in distribution: $\mu_n \to \mu$ in distribution as $n \to \infty$ if and only if $\beta_{\mathcal{S}}(\mu_n, \mu) \to 0$ as $n \to \infty$.

Bernstein-von Mises phenomenon. One will say that the model satisfies a Bernstein-von Mises phenomenon if, as $n \to \infty$

$$\beta_{H(\omega)}(\Pi_n, \mathcal{N}) \to 0 \text{ in } \mathbb{P}_{f_0}\text{-probability.} \tag{1.1.22}$$

Castillo and Nickl (2013) have shown that result for the Gaussian White Noise model and series priors in their Theorem 8.

In the Density Estimation model where the observations $X_1, \ldots, X_n$ are i.i.d. random variables of density f_0 assumed to be α-Hölder, one recenters the function with the help of a smoothed version of the empirical estimator $\frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ to get a convergence in the Bounded Lipschitz metric of a larger space $\mathcal{M}_0(\omega)$ to the law of the Gaussian White Bridge, see Section 1.3 for more details about the BVM phenomenon in density estimation.

Uncertainty Quantification. Bernstein-von Mises results are useful to build Confidence sets from Credible sets (recall the definitions (1.1.13), (1.1.14) and (1.1.15)), but in

nonparametric models this will not always work. Theorem 1 of Castillo and Nickl (2013) states that this works in the Gaussian White Noise model for fixed regularity, namely the credible set is built using the true regularity of the function which therefore assumed to be known. To get adaptive results, one often needs more conditions on the parameter to estimate, such as so-called polished-tails condition or self-similarity conditions. As seen in Szabó et al. (2015), one will often need to use a blow up factor to ensure that the credible sets are confidence sets. Ray (2017) derive adaptive confidence sets from credible sets for the Gaussian White Noise model under a self-similarity condition and Spike and Slab priors.

One can use the following approach to quantify uncertainty via inflated credible balls. Choose a consistent estimator $\hat{\theta}$ of the parameter θ_0 and let $r(X) = \int \|\theta - \hat{\theta}\|^2 d\Pi(\theta|X)$, which is the second posterior moment if one chooses $\hat{\theta} = \bar{\theta}$ the posterior mean. The credible ball is defined as

$$\mathcal{C}_L = \{\theta, \|\theta - \hat{\theta}\|^2 \leq M_\tau L r(X)\}$$

with $L \geq 1$ a blow-up factor. By Markov's inequality, one has $\Pi(\mathcal{C}_L|X) \geq 1 - \tau$ as long as $M_\tau \geq 1/\tau$. One needs now to prove that this credible set is a confidence set (1.1.15) and has an optimal diameter (1.1.14), which is the same as proving that the second posterior moment is consistent at minimax rate if $\hat{\theta} = \bar{\theta}$. This approach has been used for instance in Castillo and Szabo (2018).

1.1.4 Tuning the parameters

In the Bayesian approach, it frequently happens that the a priori put on θ also depends on a parameter. In this section, we will assume that $\theta \sim \Pi_\alpha$, with α an additionnal parameter, which is often called a hyperparameter. One problem that arises is how to choose a decent value for α. Usually, one uses one of the two following methods to handle this problem.

Hierarchical Bayes. The first natural method is to adopt an even more Bayesian approach and consider the paramater α random and put a prior $\tilde{\pi}$ on it. This results in the following Bayesian diagram

$$\begin{aligned} X|\theta, \alpha &\sim \mathbb{P}_\theta \\ \theta|\alpha &\sim \Pi_\alpha \\ \alpha &\sim \tilde{\pi} \end{aligned} \tag{1.1.23}$$

Even though one has to choose another a priori law, which may in turn depend on other parameters, the randomization it provides on α is often enough to correctly choose α in order to get optimal (or nearly optimal) rates in a majority of examples.

Empirical Bayes. Another natural idea is to choose α as $\hat{\alpha}$ the maximiser of the marginal likelihood of the α in the model, namely the likelihood integrated over the entire space of parameters Θ. Simply put, this is the marginal distribution of $\alpha|X$.

$$\hat{\alpha} = \arg\max_{\alpha} \int_{\Theta} \left(\prod_{k=1}^{n} f_{\theta}(X_i) \right) \pi_{\alpha}(\theta) d\theta \tag{1.1.24}$$

One now uses this quantity $\hat{\alpha}$ to form a prior by plugging $\hat{\alpha}$ in Π_{α}, resulting in the following diagram

$$\begin{array}{rcl} X|\theta & \sim & \mathbb{P}_{\theta} \\ \theta & \sim & \Pi_{\hat{\alpha}} \end{array} \tag{1.1.25}$$

These two methods are of prime interest in the following, especially the Empirical Bayes method.

1.2 Gaussian Sequence Model and Thresholding

1.2.1 Definition of the Model

We can write the Gaussian Sequence Model as follows, with X the observed vector of $\mathbb{R}^n$

$$X_i = \theta_{0,i} + \varepsilon_i, \quad i = 1, \dots n, \tag{1.2.1}$$

where $\varepsilon_1, \dots, \varepsilon_n$ are independent and identically distributed (iid) random variables following the $\mathcal{N}(0,1)$ law (whose density will be denoted ϕ), and the parameter $\theta_0 = (\theta_{0,1}, \dots, \theta_{0,n})$ belongs to the class $\ell_0[s_n]$ defined by

$$\ell_0[s_n] = \{\theta \in \mathbb{R}^n,\ |\{i \in \{1, \dots n\},\ \theta_i \neq 0\}| \leq s_n\},$$

for $0 \leq s_n \leq n$, where $|A|$ is the number of elements in the set A.
One commonly assumes that $s_n = o(n)$ when $n \to \infty$.
We denote by $\|\cdot\|$ the euclidean norm, $\|v\|^2 = \sum_{i=1}^{n} v_i^2$ for $v \in \mathbb{R}^n$.

We are interested in finding estimators of θ_0 that converge to θ_0 at the minimax rate of the class $\ell_0[s_n]$, which is, as proven in Donoho et al. (1992)

Theorem 3 (Donoho,Hoch,Johnstone,Stern,1992). Let r_n be the minimax rate for estimating θ in $\ell_0[s_n]$ with respect to $\|.\|$. Then,

$$r_n = r_{n,2}(\ell_0[s_n]) = \inf_{\hat{\theta}} \sup_{\theta \in \ell_0[s_n]} \frac{1}{n} \sum_{i=1}^{n} E_\theta(\hat{\theta}_i - \theta_i)^2 = \frac{2s_n}{n} \log(\frac{n}{s_n}) (1 + o(1))$$

when $n \to \infty$

For an estimator $\hat{\theta}$ of θ_0, it is then desirable that

$$\sup_{\theta_0 \in \ell_0[s_n]} \frac{1}{n} E_{\theta_0} \|\hat{\theta} - \theta_0\|_2^2 \leq C \frac{2s_n}{n} \log(\frac{n}{s_n}) (1 + o(1)), \tag{1.2.2}$$

where C is a positive constant, that we ideally would like to be equal to 1 (but this could represent a lot of additional work on its own).

In fact, we are mostly interested in more general results for the entire a posteriori law, namely

$$\sup_{\theta_0 \in \ell_0[s_n]} E_{\theta_0} \left[\Pi(\|\theta - \theta_0\|^2 > 2Cs_n \log(\frac{n}{s_n}) | X) \right] \to 0 \tag{1.2.3}$$

and for the posterior second moment

$$\sup_{\theta_0 \in \ell_0[s_n]} E_{\theta_0} \left[\int \|\theta - \theta_0\|_2^2 d\Pi(\theta|X) \right] \leq 2Cs_n \log(\frac{n}{s_n})(1 + o(1)) \tag{1.2.4}$$

The second moment will be a main focus in the following, as good results of convergence for the posterior second moment imply good results for the complete a posteriori law and lead to Uncertainty Quantification via inflated credible balls, as seen in 1.1.3. Also if ones has (1.2.4), the posterior mean (denoted by $\bar{\theta}$) will satisfy (1.2.2). Indeed, using the Jensen inequality, one has for every $\theta_0 \in \ell_0[s_n]$

$$\begin{aligned} \|\bar{\theta} - \theta_0\|_2^2 &= \| \int \theta d\Pi(\theta|X) - \theta_0 \|_2^2 \\ &\leq \int \|\theta - \theta_0\|_2^2 d\Pi(\theta|X) \end{aligned}$$

which leads to $\sup_{\theta_0 \in \ell_0[s_n]} \frac{1}{n} E_{\theta_0} \|\bar{\theta} - \theta_0\|_2^2 \leq C \frac{2s_n}{n} \log(\frac{n}{s_n}) (1 + o(1))$.

The natural way to handle the sparsity of the model and produce consistant estimators is to use thresholding.

Thresholding. The first idea is to estimate θ by keeping the observations larger than some threshold t_n, and set the remaining coordinates to zero, this is the hard thresholding estimator : $\hat{\theta}_i = X_i \mathbb{1}_{\{|X_i|>t_n\}}$ for $i \in \{1, \dots, n\}$.

One has then to choose the threshold t_n. The oracle choice, namely if the maximum number of nonzero coordinates of the true signal s_n is known, is $t_n = \sqrt{2\log(n/s_n)}$. It can be checked that $\hat{\theta}$ concentrates around the true signal θ_0 at minimax rate. As s_n is unknown, one can not choose this threshold. However, the choice $t_n = \sqrt{2\log(n)}$ provides a near-minimax rate, only missing the true minimax rate by a constant or a logarithmic factor. Such fixed thresholds are actually not flexible enough. Indeed, if one chooses a rather large t_n but the true signal happens to be too dense, too much observations will be set to 0, and if one chooses a rather small t_n but the true signal is too sparse, the estimator will keep too much observations. A good threshold should therefore adapt to the effective sparsity of the signal. Furthermore, one may also want the threshold to be stable to small changes of the data. We will see in what follows that a suitable (possibly empirical) choice of prior on θ leads to a thresholding estimator (the posterior median) which has a threshold with all these desirable properties.

Penalization and other frequentist methods. The hard thresholding estimator can in fact be viewed as an ℓ^0-penalized estimator, which was introduced in the context of model selection (see for instance Birgé and Massart (2001)). Another useful penalty is the ℓ^1-norm of θ, which leads to the LASSO estimator.

The LASSO estimator is defined as follows

$$\hat{\theta}_{LASSO} = \underset{\theta \in \mathbb{R}^n}{argmin}\{\sum_{i=1}^{n}(\theta_i - X_i)^2 + \lambda \sum_{i=1}^{n} |\theta_i|\}$$

where $\lambda \geq 0$ is the regularization parameter. The second term is called the ℓ^1 penalty and is what makes the LASSO work, as it allows the estimator to continuously shrink the coefficients. The larger λ the closer to 0 are the coefficients. The LASSO, which leads to a good prediction accuracy by providing to the user a bias-variance trade-off, has been largely studied over the years. Among many others, one can cite Tibshirani (1996), Bickel et al. (2009), Zhang (2005) and Zou (2006).

Several other frequentist methods have been developed, one can cite, among many other methods, estimates based on False Discovery Rate thresholds (see Abramovich et al. (2006)), who used the Benjamini and Hochberg threshold.

1.2.2 Bayesian approach and the Spike and Slab Prior

We will follow the approach introduced in 1.1, and view the parameter θ as a random variable following an a priori law that we now have to choose. The first natural law to think of may just be a product of Gaussian densities, for we know that this is a conjugate prior and that the a posteriori law will also be Gaussian. Let us try this prior and assume first that for every $i \in \{1, \dots, n\}$

$$\theta_i \sim \mathcal{N}(0, \sigma_i^2),$$

After some quick computing, one finds out that the a posteriori law is also a product of Gaussian densities with updated means and variances so that, for every $i \in \{1, \dots, n\}$

$$\theta_i | X_i \sim \mathcal{N}\left(\frac{\sigma_i^2}{1+\sigma_i^2} X_i, \frac{\sigma_i^2}{1+\sigma_i^2}\right)$$

Therefore the posterior mean estimator is $\hat{\theta}^{mean} = \frac{\sigma^2}{1+\sigma^2} X$.
Now if for example one assumes that, for every $i \in \{1, \dots, n\}$, $1/2 \leq \sigma_i^2 \leq 1$, let us consider $\theta_0 = 0 \in \ell_0[s_n]$ and look at its quadratic risk

$$E_{\theta_0}[\|\hat{\theta}^{moy} - \theta_0\|^2] = \sum_{i=1}^{n} \left(\frac{\sigma_i^2}{1+\sigma_i^2}\right)^2 E_{\theta_0}[X_i^2] \geq \frac{n}{9}$$

which is far from the minimax rate $2s_n \log(\frac{n}{s_n})$. This shows that this choice of prior does not properly take account of the sparsity of the model.
This choice also faces issues for large signals. Indeed if one assumes that for every $i \in \{1, \dots, n\}$, $\sigma_i^2 = 1$, the posterior mean becomes $\hat{\theta}^{moy} = \frac{X}{2}$, so if the real θ_0 has a first coordinate equal to 1000, the first coordinate of the estimator will be around 500. One sees that the estimator shrinks the signal too much to be useful, and it suggests that a density with heavier tails may also be useful.

A second idea is to use another continuous distribution and use a product of Laplace distributions instead. In this view we will now assume that for every $i \in \{1, \dots, n\}$

$$\theta_i \sim Lap(0, \lambda).$$

The posterior density can be written as a constant times $e^{-\frac{1}{2}\sum_{i=1}^{n}(\theta_i - X_i)^2 - \sum_{i=1}^{n}\lambda|\theta_i|}$ so the posterior mode $\hat{M}$ is

$$\begin{aligned}\hat{M} &= \underset{\theta\in\mathbb{R}^n}{argmax}\{e^{-\frac{1}{2}\sum_{i=1}^n(\theta_i-X_i)^2-\sum_{i=1}^n\lambda|\theta_i|}\}\\ &= \underset{\theta\in\mathbb{R}^n}{argmin}\{\sum_{i=1}^n(\theta_i-X_i)^2+2\lambda\sum_{i=1}^n|\theta_i|\}\\ &= \hat{\theta}_{LASSO}\end{aligned}$$

Therefore the posterior mode will show good consistency properties. But this represents only one aspect of the full a posteriori law, which has actually been shown in Theorem 7 in Castillo et al. (2015) to not contract at the same rate as its mode. Namely, for the standard choice $\lambda = \lambda_n = \sqrt{2\log n}$ the posterior distribution will not put any mass on balls around the true signal of radius $\sqrt{n}/\sqrt{2\log n}$. Thus this choice of prior is not very appropriate especially if one also aims at Uncertainty Quantification through the full posterior distribution.

Another idea that seems very natural and that will not use a continuous prior is to reflect the parcimonious nature of the model directly in the a priori law, which is done in the Spike and Slab prior.

The Spike and Slab Prior. Since the model is sparse, we already know that a certain number (in fact, most) of coordinates are equal to zero, the natural idea behind the Spike and Slab prior is to force some coordinates of θ to be equal to 0 and model the rest of the coordinates as an arbitrary signal (even possibly small).

$$\theta \sim \Pi_\alpha := \bigotimes_{i=1}^n (1-\alpha)\delta_0 + \alpha\Gamma \tag{1.2.5}$$

with δ_0 the Dirac in 0, Γ a probability law to be chosen which is absolutely continuous relatively to the Lebesgue measure and whose density will be noted γ, and $\alpha \in [0;1]$ a parameter to be chosen too.

Both because of their graphical representations, the part with the Dirac mass at 0 is called the Spike and the part with the density which is meant to have heavy tails is called the Slab. The closer α is to 0 the sparsier the model is, and one usually calls α the smoothing parameter.

Posterior Distribution. The a posteriori law is also a product. Indeed, writing $g = \phi * \gamma$

$$\Pi_\alpha(\theta|X) = \prod_{i=1}^n \frac{[(1-\alpha)\delta_0(\theta_i)+\alpha\gamma(\theta_i)]\phi(X_i-\theta_i)}{(1-\alpha)\phi(X_i)+\alpha g(X_i)}.$$

Thus we obtain

$$\Pi_\alpha(\theta|X) = \prod_{i=1}^{n} [(1 - a(X_i))\delta_0(\theta_i) + a(X_i)\psi_{X_i}(\theta_i)] \tag{1.2.6}$$

with

$$a(X_i) = a_\alpha(X_i) = \frac{\alpha g(X_i)}{(1-\alpha)\phi(X_i) + \alpha g(X_i)}$$

and the density

$$\psi_{X_i} = \frac{\phi(X_i - \cdot)\gamma(\cdot)}{g(X_i)}$$

Note that, for the moment, each $\theta_i|X$ only depends on the observation X_i and actually $\mathcal{L}(\theta_i|X) = \mathcal{L}(\theta_i|X_i)$.

Firstly, one has now to specify the choice of the parameters γ and α. If we first wish to choose the Slab density γ, one may want to use Gaussian densities.

Case where γ is $\mathcal{N}(0, \sigma^2)$. In that case, for every $i \in \{1, \dots, n\}$, ψ_{X_i} also is the density of a normal law, whose mean is $\frac{\sigma^2}{1+\sigma^2} X_i$ and whose variance is $\frac{\sigma^2}{1+\sigma^2}$.

Taking $\alpha = 2$ in Theorem 2.8 of Castillo and van der Vaart (2012) shows that if the true signal has coordinates that are too large, the posterior distribution will asymptotically not put any mass around the true signal. This shows that choosing γ Gaussian is not suitable. In fact, the hypotheses of the following properties used by Johnstone and Silverman also exclude the Gaussian case.

Hypotheses on the Slab. Following Johnstone and Silverman (2004), we would like the density γ to have heavy enough tails, that is why we will choose in the following a standard Laplace density instead of a standard normal law. Precisely, one assumes that

$$\sup_{u>0} |\frac{d}{du} \log \gamma(u)| = \Lambda < \infty \tag{1.2.7}$$

This gives us that, $\forall u > 0$, $\log \gamma(u) \geq \log \gamma(0) - \Lambda u$ and therefore, $\forall u > 0$, $\gamma(u) \geq \gamma(0) e^{-\Lambda|u|}$, which prevents us from choosing a gaussian γ.
One will furthermore assume that $u \to u^2\gamma(u)$ is bounded and that there exists $\kappa \in [1, 2]$ such that, when $y \to \infty$

$$\frac{1}{\gamma(y)} \int_y^\infty \gamma(u)du \asymp y^{\kappa-1} \tag{1.2.8}$$

Properties of the coordinate-wise posterior median. Under these hypotheses, the posterior median, denoted by $\hat{\theta}^{med}$, has the following properties

Proposition 1 (Johnstone, Silverman, 2004). The posterior median $\hat{\theta}^{med} = \hat{\theta}^{med}(x, \alpha)$ is an increasing function in x, antisymmetric and is a thresholding rule:

$$\forall x \geq 0 \ , \ 0 \leq \hat{\theta}^{med}(x, \alpha) \leq x.$$

Moreover it is a thresholding estimator: there exists $t = t(\alpha) > 0$ (see below for more on this threshold) such that

$$\hat{\theta}^{med}(x, \alpha) = 0 \Leftrightarrow |x| \leq t(\alpha). \tag{1.2.9}$$

It also has a bounded shrinkage property : there exists $b > 0$ such that, for $t(\alpha)$ as above

$$\forall x, \forall \alpha, \ |\hat{\theta}^{med}(x, \alpha) - x| \leq t(\alpha) + b. \tag{1.2.10}$$

Link between α and the threshold $t = t(\alpha)$. We will see t as a function of α defined as follows

$$t\colon \begin{cases} (0,1) \longrightarrow (0, +\infty) \\ \alpha \longrightarrow \text{The threshold of the posterior median obtained with prior } \Pi_\alpha \end{cases} \tag{1.2.11}$$

Using the following notation $\forall x \in \mathbb{R}$, $g_+(x) = \displaystyle\int_0^{+\infty} \phi(x-u)\gamma(u)du$ and $g_-(x) = \displaystyle\int_{-\infty}^{0} \phi(x-u)\gamma(u)du$ and recalling (1.2.6), we have

$$P(\theta > 0 | X = x) = \frac{\alpha g_+(x)}{(1-\alpha)\phi(x) + \alpha g(x)},$$

so that $2\alpha g_+(t) = (1-\alpha)\phi(t) + \alpha g(t)$ and therefore

$$\frac{1}{\alpha} = 1 + \frac{g_+(t) - g_-(t)}{\phi(t)} = 1 + 2\int_0^{+\infty} \sinh(tu) e^{-\frac{u^2}{2}} \gamma(u)du. \tag{1.2.12}$$

This gives us a threshold which is a continuous function of α, decreasing from $+\infty$ when α equals 0 to 0 when α equals 1.
One can further show that $t(\alpha)$ is of order $\sqrt{2\log(1/\alpha)}$ independently of the choice of the density γ (as long as γ satisfies (1.2.7) and (1.2.8)). One can refer back to Lemma 14 of Castillo and Roquain (2018) for an even finer result (as $\zeta^2(\alpha) - C < t^2(\alpha) < \zeta^2(\alpha)$

as shown in (52) and (53) of Johnstone and Silverman (2004)).
As an oracle choice of threshold is $\sqrt{2\log(n/s_n)}$, one sees that an oracle choice of α would be $\alpha^* = s_n/n$. Since s_n is still an unknown quantity, one may now ask how to properly choose α.

Choice of α. First note that if one chooses α constant in $(0,1)$, the results will not be more satisfying than the case $\alpha = 1$ which is a product of continuous densities. The intuition behind this is if one looks at the a priori law, the expected number of nonzero coordinates αn will be of order n, which is too high as we want it to be of order s_n. One therefore has to make α depend on n and have it tend to 0 with n to be able to handle with the sparsity of the model properly.

Taking this into account, one can expect that choosing α of order $1/n$ will provide better results. Indeed, for a Spike and Slab prior with $\alpha = 1/n$, it is shown in Mismer (2015) (available on author's webpage) that the posterior law concentrates itself around the true signal θ_0 at a near-minimax rate : $\frac{s_n}{n}\log n$ (this is the correct rate (3) only up to a logarithmic factor).

For better results (both in theory and practice), one may consider an automatic procedure to select α, namely to use a Hierarchical Bayes or an Empirical Bayes approach.

To be even more Bayesian, one can use the Hierarchical method and consider α itself as a random variable, and put a prior on it. As $\alpha \in (0,1)$, a natural prior is a Beta distribution. If $\alpha \sim \text{Beta}(a,b)$, the Beta distribution of parameters a, $b \in \mathbb{R}^{+*}$ which has as density $b(x) = x^{a-1}(1-x)^{b-1}\mathbb{1}_{[0,1]}(x)\Gamma(a+b)/(\Gamma(a)\Gamma(b))$, the expected number of nonzero coordinates for Π_α is $\frac{a}{a+b}n$. Ideally this number should be s_n but as seen before with the choice $\alpha = 1/n$ reasonable rates can be achieved if this expected number is smaller than s_n, which suggests that the quantity $\frac{a}{a+b}$ has to belong in $(c\frac{1}{n}, C\frac{s_n}{n})$. This suggests to take a small and b larger, and in this view one may take $\alpha \sim \beta(1, n+1)$, as in Castillo and van der Vaart (2012).
This choice leads to the minimax concentration of the corresponding posterior distribution, as can be seen in the paper of Castillo and van der Vaart (2012) (They actually derive a concentration result for a general prior in their Theorem 2.2, the Spike and Slab prior only being a special case treated in Example 2.2, more details on the general prior can be found in 1.2.3). Note that the rate of convergence has the right logarithm part $\log(n/s_n)$ even though we were not able to choose the first parameter equal to s_n as it is unknown (this would have led to an expected number of nonzero coordinates of order s_n). This shows that the Hierarchical Bayes approach provides more flexibility than, for instance, just taking $\alpha = 1/n$.

Choosing α by Empirical Bayes. We will now introduce the Empirical Bayes approach for our Spike and Slab prior, which will be the main focus for the results presented in this document for the Gaussian Sequence model. The idea of Johnstone and Silverman (2004) is to estimate α by maximising the marginal likelihood in α in the Bayesian model, which is the density of $\alpha \,|\, X$. The log-marginal likelihood in α can be written as

$$\ell(\alpha) = \ell_n(\alpha; X) = \sum_{i=1}^{n} \log((1-\alpha)\phi(X_i) + \alpha g(X_i)). \tag{1.2.13}$$

Let $\hat{\alpha}$ be defined as the maximiser of the log-marginal likelihood

$$\hat{\alpha} = \underset{\alpha \in \mathcal{A}_n}{\operatorname{argmax}}\ \ell_n(\alpha; X), \tag{1.2.14}$$

where the maximisation is restricted to $\mathcal{A}_n = [\alpha_n, 1]$, with α_n defined, with $t(\alpha)$ as in (1.2.11), by

$$t(\alpha_n) = \sqrt{2 \log n}.$$

The reason for this restriction is that one does not need to take α smaller than α_n, which would correspond to a choice of α 'more conservative' than hard-thresholding at threshold level $\sqrt{2 \log n}$.

The a priori law that will be therefore considered is the Spike and Slab where we have 'plugged' the value $\hat{\alpha}$:

$$\theta \sim \Pi_{\hat{\alpha}} := \bigotimes_{i=1}^{n} (1-\hat{\alpha})\delta_0 + \hat{\alpha}\Gamma \tag{1.2.15}$$

One will also denote the threshold of our new 'plug-in' posterior median, recalling (1.2.11),

$$\hat{t} = t(\hat{\alpha}) \tag{1.2.16}$$

The first result obtained with this approach is the following, which shows that some point estimators derived from the Empirical Bayes a posteriori law converge to the true signal at minimax rate as appears in Theorem 3

Theorem 1.2.1 (Johnstone, Silverman, 2004)**.** *Let $\hat{\mu}$ be a thresholding rule (see (1.2.9)) with threshold $\hat{t}$ and with the bounded shrinkage property (see (1.2.10)). For n large enough we have, for C a large enough constant, and provided $s_n \geq \log^2 n$*

$$\sup_{\theta \in \ell_0[s_n]} \frac{1}{n} E_\theta \|\hat{\mu} - \theta\|^2 \leq C r_n$$

With a slab γ verifying (1.2.7) and (1.2.8), the posterior median is a thresholding rule with the bounded shrinkage property. Johnstone and Silverman (2004) also prove that the result also holds for the posterior mean even though it only has the bounded shrinkage property. The parameter $\hat{\alpha}$ obtained by Empirical Bayes is computationally very tractable, and the authors developped the package `EBayesThresh` to compute the quantities involved in their results.

Looking at this theorem, one can now ask whether the entire a posteriori law will also concentrate around the true signal at minimax rate. The focus will be put on the posterior second moment, for which one would like to derive results of the form

$$\sup_{\theta_0 \in \ell_0[s_n]} E_{\theta_0} \int \|\theta - \theta_0\|_2^2 d\Pi_{\hat{\alpha}}(\theta \,|\, X) \leq C r_n \tag{1.2.17}$$

This topic was one of the main interests of this work and is further addressed in 1.2.5.

1.2.3 Other choices of a priori laws

There are of course many other choices of a priori laws in this sparse setting, and this section aims to introduce a few of them.

Spike and Slab LASSO(SSL). Ročková (2018) and Ročková and George (2018) used a slightly different prior, which will also be considered further in this document. The idea is to replace the Dirac mass by a probability distribution to make the whole prior absolutely continuous relatively to the Lebesgue measure

$$\theta \sim \Pi_\alpha := \bigotimes_{i=1}^{n} (1-\alpha)\Gamma_0 + \alpha\Gamma_1 \tag{1.2.18}$$

where the densities γ_0 and γ_1 are Laplace $((\lambda_i/2)\exp(-\lambda_i|x|))$ where parameters λ_0 and λ_1 serve very different purposes : the first is larger than the second, making the first density look like a Dirac (in a continuous way) and the second like a classical Slab.

The Spike and Slab LASSO prior can be interpreted as linking the Spike and Slab prior and the frequentist LASSO, as the Spike and Slab is obtained by letting $\lambda_0 \to \infty$ and the LASSO is obtained by setting $\lambda_0 = \lambda_1$ and considering the posterior mode. Note that in the SSL case, the posterior median is not a thresholding estimator anymore.

The modes of the a posteriori law are then well defined, and the mode is even unique as soon as $(\lambda_1 - \lambda_0)^2 \leq 4$. Ročková (2018) shows that the posterior global mode converges to the true signal at minimax rate with the oracle choice $\alpha = s_n/(s_n + n)$ (so this is not an adaptive result) as long as $\lambda_1 < e^{-2}$ and for the choice $\lambda_0 = n/s_n + 4$. However,

the author shows an adaptive result in a particular regime for the entire posterior law using a Hierarchical approach, setting $\alpha \sim \beta(1, 4n)$ and $\lambda_0 = (1-\alpha)/\alpha$. The additional assumption on the signal is that all the nonzero coordinates have to be greater than $(s_n/n)\log(n/s_n)$.

Horseshoe. The Horseshoe prior is a scale mixture of Gaussian distributions. It is the distribution (which has a density π) such that, $\forall i \in \{1, \dots, n\}$, with each λ_i following the standard half-Cauchy on the positive reals law (denoted by $C^+(0,1)$) and τ a global hyperparameter,

$$\begin{aligned} \theta_i|\lambda_i, \tau &\sim \mathcal{N}(0, \lambda_i^2\tau^2) \\ \lambda_i &\sim C^+(0,1) \end{aligned} \tag{1.2.19}$$

The name Horsehoe, as stated by Carvalho et al. (2010), comes from the fact that, with $\kappa_i = \frac{1}{1+\lambda_i}$,

$$\begin{aligned} E[\theta_i|X] &= \int_0^1 (1-\kappa_i)X_i\Pi(\kappa_i|X)d\kappa_i \\ &= (1 - E[\kappa_i|X])X_i \end{aligned}$$

The quantity $E[\kappa_i|X]$ can be seen as the a posteriori amount of shrinkage towards 0. Since the λ_i's are half-Cauchy, each shrinkage coefficient κ_i follows the $\beta(1/2, 1/2)$ law, which has the shape of a horseshoe.

Its density π satisfies the following inequality, proven by Carvalho et al. (2010)

$$\frac{1}{2\tau}\log(1 + \frac{4\tau^2}{\theta_i^2}) \lesssim \pi(\theta_i) \lesssim \frac{1}{\tau}\log(1 + \frac{2\tau^2}{\theta_i^2}), \theta_i \neq 0$$

The density π therefore has a pole at zero and Cauchy tails, which makes the Horseshoe and the Spike and Slab (with Cauchy Slab) strikingly similar, and the parameter τ seems to play the same role as the parameter α of the Spike and Slab.

van der Pas et al. (2017a) show adaptive near-minimax (without the $\log(n/s_n)$ part) rates of convergence for the posterior distribution for Empirical Bayes and Hierarchical Bayes under certain conditions. van der Pas et al. (2017b) give credible sets derived from the horseshoe posterior that can be used, asymptotically in n, as confidence sets.

A more general Prior. Castillo and van der Vaart (2012) use a more global a priori form, of which the Spike and Slab is just a particular case. The general prior is obtained through the following method

- Draw a dimension s using a law $\pi(s)$ on the set $\{0, 1, \dots, n\}$

- Draw a support $S \subset \{1, \dots, n\}$ uniformly on the sets of cardinal s : $\Pi(S) = \frac{\pi(s)}{\binom{n}{s}}$

- This leads to the following prior on θ

$$\theta \sim \sum_{S \subset \{1,\dots,n\}} \Pi(S) \left\{ \bigotimes_{i \in S} \Gamma \otimes \bigotimes_{i \notin S} \delta_0 \right\} \tag{1.2.20}$$

  with Γ probability distributions which are absolutely continuous relative to the Lebesgue measure and with density γ.

Note that in this general setting, the a posteriori does not shape as a product anymore, which can make the proofs harder.
One says that π has exponential decrease if there exist $C > 0$ and $D < 1$ such that

$$\pi(s) \leq D\pi(s-1) \tag{1.2.21}$$

for $s > Cs_n$. If this is satisfied with $C = 0$ then π is said to has strict exponential decrease.
One may now ask what to choose for the prior π on the dimension, here are some examples :

Binomial prior. If π is the binomial $\text{Bin}(n, \alpha)$, then the prior on θ is the Spike and Slab. This prior π has exponential decrease for $\alpha \lesssim \frac{s_n}{n}$.
Hierarchical approach using a Beta prior. As before, we want $s|\alpha$ to follow $Bin(n, \alpha)$. In this aim take $\alpha \sim \text{Beta}(\kappa, \lambda)$ and set

$$\pi(s) = \binom{n}{s} \frac{\beta(\kappa + s, \lambda + n - s)}{\text{Beta}(\kappa, \lambda)} \propto \frac{\Gamma(\kappa + s)\Gamma(\lambda + n - s)}{s!(n-s)!}$$

For $\kappa = 1$ and $\lambda = n+1$, we have $\pi(s) \propto \binom{2n-s}{n}$, which has strict exponential decrease with $D = \frac{1}{2}$. More generally and as seen before one can set $\kappa = 1$ and $\lambda = \kappa_1(n+1)$, which leads to $\pi(s) \propto \binom{(\kappa_1+1)n-s}{\kappa_1 n}$.
Complexity prior. This prior has the form $\pi(s) \propto e^{-as\log(\frac{bn}{s})}$. It shows to be quite fitting for the problem. Indeed, as $e^{s\log(\frac{n}{s})} \leq \binom{n}{s} \leq e^{s\log(\frac{en}{s})}$, it is inversely proportional to the number of models of size s and seems good to lessen the complexity of the problem. It has exponential decrease as soon as $b > 1 + e$.

One is not able to use Theorem 2 as our observations X_i are not i.i.d.. To get to their results, Castillo and van der Vaart (2012) first prove a result on the dimension :

Theorem 4 (Castillo, van der Vaart, 2012). If π has exponential decrease and γ is centered with a finite second moment, then there exists $M > 0$ such that $n \to \infty$:

$$\sup_{\theta_0 \in \ell_0[s_n]} E_{\theta_0}[\Pi(|S_\theta| > M s_n | X)] \to 0$$

This further leads to their main result

Theorem 5 (Castillo, van der Vaart, 2012). If π has exponential decrease and γ is centered with a finite second moment which can be written e^h with h such that $\forall x, y \in \mathbb{R}$, $|h(x) - h(y)| \lesssim 1 + |x - y|$, then, with r_n^* such that

$$r_n^{*2} \geq (s_n \log(\frac{n}{s_n})) \vee (\log(\frac{1}{\pi(s_n)}))$$

and $M > 0$ large enough, for $n \to \infty$

$$\sup_{\theta_0 \in \ell_0[s_n]} E_{\theta_0}[\Pi(\|\theta - \theta_0\|^2 > M r_n^{*2} | X)] \to 0$$

The hypothesis on γ is verified for Laplace, and the 3 priors on dimension seen before in the examples (so including the Spike and Slab) verify the hypotheses of the theorem. Moreover, considering the complexity prior, the authors showed that the posterior mean converge to the true signal at minimax rate and that convergence of the second posterior moment is obtained.

There are several other Bayesian methods in the Gaussian sequence setting, such as non-local priors (as in Johnson and Rossell (2010)), Gaussian mixture priors (see George and Foster (2000)), or adopting a fractional likelihood perspective (see Martin and Walker (2014)).

1.2.4 Exact constant

In the setting of the sparse sequence model, we say that the posterior distribution converges at minimax rate with exact constant (or converges at sharp minimax rate) with respect to the L^2-norm loss if

$$\sup_{\theta_0 \in \ell_0[s_n]} E_{\theta_0} \left[\int \|\theta - \theta_0\|_2^2 d\Pi(\theta | X) \right] \leq 2 s_n \log(\frac{n}{s_n})(1 + o(1)). \tag{1.2.22}$$

This is (1.2.4) with the constant $C = 1$, implying that this is a finer result. The definition immediately implies using Jensen's inequality that the posterior mean (denoted here by $\bar{\theta}$) converges at minimax rate with exact constant to the true signal

$$\sup_{\theta_0 \in \ell_0[s_n]} E_{\theta_0}\left[\|\bar{\theta} - \theta_0\|_2^2\right] \leq 2s_n \log(\frac{n}{s_n})(1 + o(1)). \tag{1.2.23}$$

Another application is that if one uses a randomised estimator $\tilde{\theta} = \tilde{\theta}(X, U)$ using the data X and uniform variables U on $[0, 1]$ to simulate from the a posteriori law, namely $\tilde{\theta}$ such that $\mathcal{L}(\tilde{\theta}(X, U)|X) = \Pi(\cdot|X)$; stating (1.2.22) is exactly stating the convergence to θ_0 at sharp minimax rate of $\tilde{\theta}$.
In the present setting, in order to effectively sample such a $\tilde{\theta}$ and have $\mathcal{L}(\tilde{\theta}(X, U)|X) = \Pi(\cdot|X)$, as the Spike and Slab a posteriori law is a product, one can take, denoting by $\mathcal{F}_{\theta_i|X}$ the cumulative distribution function of each $\theta_i|X_i$,

$$\tilde{\theta}(X, U) = (\mathcal{F}^{-1}_{\theta_1|X}(U_1), \ldots, \mathcal{F}^{-1}_{\theta_n|X}(U_n))$$

The convergence at minimax rate with exact constant for the Spike and Slab will require a specific choice of Slab, as will be seen in the following section.

1.2.5 Contributions using the Empirical Bayes method for the Spike and Slab prior

The following work, which is treated in more details in Chapters 2 and 3, was motivated by pursuing the work seen in 1.2.2 of Johnstone and Silverman (2004). Using the Empirical Bayes approach (1.2.14), they derived convergences at minimax rate for the posterior median and the posterior mean, as seen in Theorem 1.2.1, for suiting densities γ. Aiming at Uncertainty Quantification (which was later treated by Castillo and Szabo (2018) based on the present work), a natural question was to know if the second moment of the posterior law (1.2.15) behaved the same way. Namely, the form of the desired results is

$$\sup_{\theta_0 \in \ell_0[s_n]} E_{\theta_0} \int \|\theta - \theta_0\|_2^2 d\Pi_{\hat{\alpha}}(\theta \,|\, X) \leq C r_n \tag{1.2.24}$$

Suboptimality of the Laplace Slab. The first investigations were conducted with Γ taken as a standard Laplace distribution, and led to a quite surprising result. The posterior second moment for a Laplace Slab does not converge at minimax rate uniformly in $\theta \in \ell_0[s_n]$, even though the posterior median and mean do so (as was proved by Johnstone and Silverman (2004) and noted above)

Theorem 6. Let Π_α be the Spike and Slab prior distribution (1.2.5) with Slab distribution Γ equal to the Laplace distribution Lap(1). Let $\Pi_{\hat{\alpha}}[\cdot \,|\, X]$ be the corresponding plug-in posterior distribution given by (1.2.15), with $\hat{\alpha}$ chosen by the empirical Bayes procedure (1.2.14). There exist $D > 0$, $N_0 > 0$, and $c_0 > 0$ such that, for any $n \geq N_0$ and any s_n with $1 \leq s_n \leq c_0 n$, there exists $\theta_0 \in \ell_0[s_n]$ such that,

$$E_{\theta_0} \int \|\theta - \theta_0\|_2^2 d\Pi_{\hat{\alpha}}[\theta \,|\, X] \geq D s_n e^{\sqrt{\log(n/s_n)}}.$$

One can now ask whether this suboptimality result only comes from considering an integrated L^2–moment, instead of simply asking for a posterior convergence result in probability like (1.2.3). It is actually not the case, as the entire a posteriori law is also suboptimal for the Laplace Slab.

Theorem 7. Under the same notation as in Theorem 6, if Π_α is a Spike and Slab distribution with a slab Γ taken as the standard Laplace distribution, there exists $m > 0$ such that for any s_n with $s_n/n \to 0$ and $\log^2 n = O(s_n)$ as $n \to \infty$, there exists $\theta_0 \in \ell_0[s_n]$ such that, as $n \to \infty$,

$$E_{\theta_0} \Pi_{\hat{\alpha}} \left[\|\theta - \theta_0\|_2^2 \leq m s_n e^{\sqrt{2\log(n/s_n)}} \,|\, X \right] = o(1).$$

This is a stronger result than Theorem 6, but with an additional mild condition $s_n \gtrsim \log^2 n$. The fact that this result implies the preceding one follows from bounding from below the integral in the display of Theorem 6 by the integral restricted to the set where $\|\theta - \theta_0\|^2$ is larger than the target lower bound rate.

The intuition behind these two results is that the Empirical Bayes provides (for some specific signals) a parameter $\hat{\alpha}$ somewhat larger than the oracle parameter $\alpha^* = s_n/n$ (here $\hat{\alpha} \gtrsim \frac{s_n}{n} e^{\sqrt{\log(n/s_n)}}$).

One sees through this example that the behaviour of some aspects of an a posteriori law (such as the median or the mean) does not drive the behaviour of the complete a posteriori law.

One can also note that this is an example where the Empirical Bayes and the Hierarchical Bayes methods deliver different results, as the a posteriori law in the Hierarchical approach with a Laplace Slab does not show any suboptimality and converges at minimax rate, as seen in Theorem 5.

Optimal concentration for Cauchy Slab. The next direction was to use a standard Cauchy Slab instead of a standard Laplace, and this led to the following result, showing optimal concentration uniformly in $\theta \in \ell_0[s_n]$

Theorem 8. Let Π_α be the Spike and Slab prior distribution (1.2.5) with Slab distribution Γ equal to the standard Cauchy distribution. Let $\Pi_{\hat{\alpha}}[\cdot \,|\, X]$ be the corresponding plug-in posterior distribution given by (1.2.15), with $\hat{\alpha}$ chosen by the empirical Bayes procedure (1.2.14). There exist $C > 0$, $N_0 > 0$, and $c_0, c_1 > 0$ such that, for any $n \geq N_0$, for any s_n such that there exist constant c_0, c_1 such that $c_1 \log^2 n \leq s_n \leq c_0 n$,

$$\sup_{\theta_0 \in \ell_0[s_n]} E_{\theta_0} \int \|\theta - \theta_0\|_2^2 d\Pi_{\hat{\alpha}}(\theta \,|\, X) \leq C r_n.$$

Actually, any Slab density γ with tails of the order $x^{-1-\delta}$ with $\delta \in (0, 2)$ gives the same result. These densities are particularly suitable if one wants to consider d_q-distances instead of the d_2-distance (see Castillo and Szabo (2018)).

This result shows once again that heavy tails are crucial to make the Empirical Bayes method succeed and get minimax results.

Sharp minimax convergence. To go even further and get the exact constant 2 in the minimax rate, we use the special Slab density γ on $\mathbb{R}$ given by

$$\gamma(x) = \frac{1}{2}\Delta(1 + |x|), \qquad \Delta(u) = u^{-1}(1 + \log(u))^{-2} \text{ for } u > 0, \tag{1.2.25}$$

(The purpose of this new density is to have sufficiently heavy tails, heavier than Cauchy.) Apart from this specific tail property, γ still satisfies

$$\sup_{u>0} \left| \frac{d}{du} \log \gamma(u) \right| =: \Lambda < \infty.$$

but not (1.2.8). However, it satisfies a similar property, see Lemma 22 of Chapter 3

This choice leads to the following sharp result

Theorem 9. Let Π_α be the Spike and Slab prior distribution (1.2.5) with Slab density γ given by (1.2.25). Let $\Pi_{\hat{\alpha}}(\cdot \,|\, X)$ be the corresponding plug-in posterior distribution given by (1.2.15), with $\hat{\alpha}$ chosen by the empirical Bayes procedure (1.2.14). For any s_n such that there exist constants c_0, c_1 such that $c_1 \log^2 n \leq s_n \leq c_0 n$, for $n \to \infty$

$$\sup_{\theta_0 \in \ell_0[s_n]} E_{\theta_0} \int \|\theta - \theta_0\|_2^2 d\Pi_{\hat{\alpha}}(\theta \,|\, X) \leq 2 s_n \log(\frac{n}{s_n})(1 + o(1)).$$

An intuition for why it works is that one may decompose the L^2-norm in two parts, depending on whether the components of the true signal are different from zero or not. The nonzero signal part contributes for $2 s_n \log(\frac{n}{s_n})(1 + o(1))$. The other part is more

dependent on the choice of the Slab. This part indeed depends on $\hat{\alpha}$, which is too far from the oracle parameter $\alpha^* = s_n/n$ in the Laplace case, resulting in a zero signal contribution larger than the minimax rate. In the Cauchy case (actually also with tails $x^{-1-\delta}$, $\delta \in (0,2)$), the zero signal contribution appears to be exactly of the order of the minimax rate. With the special Slab (1.2.25), this contribution becomes lower than the minimax rate, finally resulting in $2s_n \log(\frac{n}{s_n})(1+o(1))$.

Results for the Spike and Slab LASSO prior (SSL). Deriving analog results for the SSL prior, which is the continuous counterpart to the Spike and Slab, was also of particular interest. The prior on θ is the following

$$\theta \sim \Pi_\alpha := \bigotimes_{i=1}^{n} (1-\alpha)\Gamma_0 + \alpha\Gamma_1, \tag{1.2.26}$$

with Γ_0 a Laplace distribution with parameter λ_0, but here we will not restrict the choice of Γ_1 to a Laplace.

As seen in section 1.2.3, it is convenient to let λ_0 depend on n, here we set, mostly for more convenience in the proofs (see Chapter 2)

$$\lambda_0 = 5n\sqrt{2\pi} \tag{1.2.27}$$

Let $\hat{\alpha}$ be defined as the maximiser of the log-marginal likelihood

$$\hat{\alpha} = \underset{\alpha \in \mathcal{A}_n}{\operatorname{argmax}}\ \ell_n(\alpha; X), \tag{1.2.28}$$

where the maximisation is restricted to $\mathcal{A}_n = [\alpha_n, 1]$, with α_n defined, in view of (1.2.11), by

$$t(\alpha_n) = \sqrt{2\log n}.$$

The a priori law that will be therefore considered is the Spike and Slab where we have 'plugged' the value $\hat{\alpha}$:

$$\theta \sim \Pi_{\hat{\alpha}} := \bigotimes_{i=1}^{n} (1-\hat{\alpha})\delta_0 + \hat{\alpha}\Gamma \tag{1.2.29}$$

Theorem 10. Let Π_α be the SSL prior distribution (1.2.26) with Cauchy slab and parameters λ_0 given by (1.2.27) and $\lambda_1 = 0.05$. Let $\Pi_{\hat{\alpha}}[\cdot \,|\, X]$ be the corresponding plug-in posterior distribution given by (1.2.29), with $\hat{\alpha}$ chosen by the Empirical Bayes procedure (1.2.28). There exist $C > 0$, $N_0 > 0$, for any $n \geq N_0$, for any s_n such that there exist

constant c_0, c_1 such that $c_1 \log^2 n \leq s_n \leq c_0 n$, then

$$\sup_{\theta_0 \in \ell_0[s_n]} E_{\theta_0} \int \|\theta - \theta_0\|^2 d\Pi_{\hat{\alpha}}(\theta \mid X) \leq C s_n \log n.$$

1.3 Density Estimation and Pólya Trees

1.3.1 Definition of the Model

We can write the Density Estimation Model as follows, with X the observed vector of $\mathbb{R}^n$

$$X_1, \ldots, X_n \text{ i.i.d. } \sim P \tag{1.3.1}$$

where P belongs to the model $\mathcal{P} = \{P; dP = f d\mu\}$ with μ the Lebesgue measure on $[0,1]$.
The goal here is to estimate the true density function f_0. We make the two following assumptions

$$f_0 \text{ is bounded away from 0 and } \infty \tag{1.3.2}$$

$$\exists \alpha \in (0,1] \text{ such that } f_0 \in \mathcal{C}^\alpha([0,1]) \tag{1.3.3}$$

where $\mathcal{C}^\alpha([0,1]) = \{f : [0,1] \to \mathbb{R}; \sup_{x \neq y \in [0,1]} \frac{|f(x) - f(y)|}{|x-y|^\alpha} < \infty\}$ is the set of α-Hölder functions of $[0,1]$. One would like to estimate f_0 in an adaptive way, namely a way that does not depend on the unknown parameter α.

Minimax rate. As proven by Ibragimov and Khas'minskii (1980), the minimax rate when estimating densities in $\mathcal{C}^\alpha([0,1])$ using the supremum norm as the loss function is

$$\varepsilon_{n,\alpha}^* = \left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+1}}. \tag{1.3.4}$$

Haar Basis. The Haar wavelet basis is $\{\phi, \psi_{lk}, 0 \leq k < 2^l, l \geq 0\}$, where $\phi = \mathbb{1}_{[0,1]}$ and, for $\psi = -\mathbb{1}_{(0,1/2]} + \mathbb{1}_{(1/2,1]}$,

$$\psi_{lk}(\cdot) = 2^{l/2}\psi(2^l \cdot - k), 0 \leq k < 2^l, l \geq 0 \tag{1.3.5}$$

As we focus on density functions on $[0,1]$, which are nonnegative functions g such that $\int_0^1 g\phi = \int_0^1 g = 1$, the first Haar-coefficient is always 1. That means that one only

needs to consider the basis functions ψ_{lk} and the Haar basis will simply be denoted as $\{\psi_{lk}\}$ in the following.

If a function g belongs to $\mathcal{C}^\alpha([0,1])$, with $\alpha \in (0,1]$, then the sequence of its Haar wavelet coefficients $\langle g, \psi_{lk}\rangle$ satisfies

$$\sup_{0\le k<2^l, l\ge 0} 2^{l(1/2+\alpha)}|\langle g, \psi_{lk}\rangle| < \infty. \tag{1.3.6}$$

Bayesian approach. To adopt a Bayesian perspective, one has to put a prior on P, therefore one has to build a probability distribution on probability distributions. A rather common choice would be a Dirichlet process, but as its draws are discrete almost surely it will not be suitable to estimate objects as smooth as a density. A more convenient distribution on distributions with densities is the Pólya tree, which is introduced in what follows. (Actually the Dirichlet process is a particular case of Pólya tree, but with the α_ε going to 0, see below)

1.3.2 The Pólya Tree Prior

Dyadic partitions. For any fixed indexes $l \ge 0$ and $0 \le k < 2^l$, the rational number $r = k2^{-l}$ can be written in a unique way as $\varepsilon(r) := \varepsilon_1(r)\dots\varepsilon_l(r)$, its finite expression of length l in base 1/2 (note that it can end with one or more 0). That is, $\varepsilon_i \in \{0,1\}$ and

$$k2^{-l} = \sum_{i=1}^{l} \varepsilon_i(r)2^{-i}.$$

Let $\mathcal{E} := \bigcup_{l\ge 0}\{0,1\}^l \cup \{\emptyset\}$ be the set of finite binary sequences. We write $|\varepsilon| = l$ if $\varepsilon \in \{0;1\}^l$ and $|\emptyset| = 0$.
Let us introduce a sequence of partitions $\mathcal{I} = \{(I_\varepsilon)_{|\varepsilon|=l}, l \ge 0\}$ of the unit interval. Set $I_\emptyset = (0,1]$ and, for any $\varepsilon \in \mathcal{E}$ such that $\varepsilon = \varepsilon(l;k)$ is the expression in base 1/2 of $k2^{-l}$, set

$$I_\varepsilon := \left(\frac{k}{2^l}, \frac{k+1}{2^l}\right] := I_k^l$$

For any $l \ge 0$, the collection of all such dyadic intervals is a partition of $(0,1]$.

The Pólya Tree Prior. The probability distribution P is said to follow a Pólya tree distribution on $\mathcal{I}$, denoted $PT(\mathcal{A})$ where $\mathcal{A} = \{\alpha_\varepsilon, \varepsilon \in \mathcal{E}\}$ is the set of parameters, if $\forall(\varepsilon,\varepsilon') \in \mathcal{E}^2$, there exists Y_ε random variables in $[0,1]$ verifying the following conditions

$$Y_{\varepsilon 0} \amalg Y_{\varepsilon' 0} \tag{1.3.7}$$

$$Y_{\varepsilon 0} \sim Beta(\alpha_{\varepsilon 0}, \alpha_{\varepsilon 1}) \tag{1.3.8}$$

$$Y_{\varepsilon 1} = 1 - Y_{\varepsilon 0} \tag{1.3.9}$$

$$P(I_\varepsilon) = \prod_{i=1}^{|\varepsilon|} Y_{\varepsilon_1 \dots \varepsilon_i} \tag{1.3.10}$$

One can then use a tree representation (see Figure 1.1) to visually compute $P(I_\varepsilon)$. One follows the path $\varepsilon_1, \varepsilon_1\varepsilon_2, \dots, \varepsilon_1\varepsilon_2 \dots \varepsilon_{|\varepsilon|-1}, \varepsilon$ alongside ε, resulting in a product of Beta variables with parameters depending on whether one goes left on the tree ($\varepsilon_j = 0$) or right ($\varepsilon_j = 1$)

$$P(I_\varepsilon) = \prod_{j=1, \varepsilon_j = 0}^{|\varepsilon|} Y_{\varepsilon_1, \dots, \varepsilon_{j-1} 0} \times \prod_{j=1, \varepsilon_j = 1}^{|\varepsilon|} (1 - Y_{\varepsilon_1, \dots, \varepsilon_{j-1} 0}) \tag{1.3.11}$$

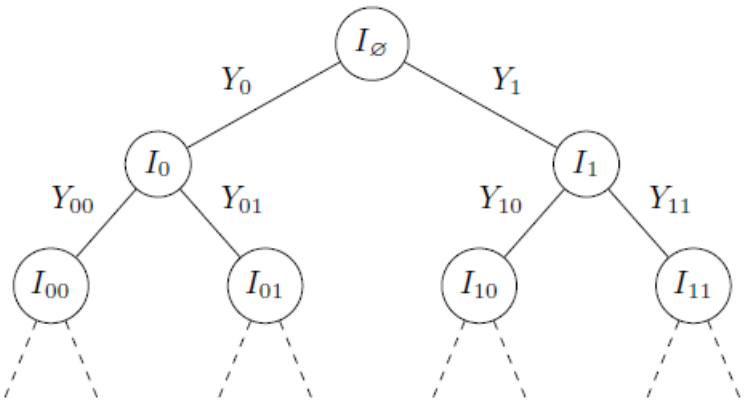

Fig. 1.1 Indexed binary tree with levels $l \leq 2$ represented. The nodes index the intervals I_ε. Edges are labelled with random variables Y_ε.

This defines a random probability distribution on the distributions of $[0, 1]$, so that the Pólya tree can be used as an a priori law on P in the Density Estimation Model. The Pólya tree prior has a conjugacy property, namely if one observes i.i.d. $X_1, \dots, X_n$ following a probability distribution P itself following a $PT(\mathcal{A})$ on $\mathcal{I}$, the a posteriori law of $P|X_1, \dots, X_n$ is also a Pólya tree $PT(\mathcal{A}^*)$, where $\mathcal{A}^* = \{\alpha^*_\varepsilon = \alpha_\varepsilon + \sum_{i=1}^{n} \mathbb{1}_{\{X_i \in I_\varepsilon\}}, \varepsilon \in \mathcal{E}\}$. A proof of this result can be found in the book of Ghosal and van der Vaart (2017).

The set of parameters $\mathcal{A}$ offers a large variety of choices, which leads to a large variety of different Pólya trees. However, the most common choice is to take the same parameters α_ε at each level. In the following, for any level $l \geq 1$, one takes

$$\forall \varepsilon \in \mathcal{E} \text{ such that } |\varepsilon| = l, \alpha_\varepsilon = a_l \tag{1.3.12}$$

In other words, one chooses in the following $\mathcal{A} = (a_l)_{l \geq 1}$ a sequence of positive numbers. Note that if one takes $a_l = 2^{-l}$, the corresponding Pólya tree is a Dirichlet process (see Ferguson (1973)).
As shown by Kraft (1964), if on the contrary one chooses a_l tending to ∞ as $l \to \infty$, more precisely if

$$\sum_{l=1}^{\infty} a_l^{-1} < \infty \tag{1.3.13}$$

the corresponding Pólya tree on the canonical dyadic partition on $[0, 1]$ is absolutely continuous relatively to the Lebesgue measure on $[0, 1]$. Therefore one will assume (1.3.13) in what follows.
For more details on Pólya trees, one can refer to Lavine (1992) or Mauldin et al. (1992).

One may note that, unlike other classical estimators such as kernel estimators (in case the kernel takes negative values) or wavelet density estimators, Pólya tree priors always sit on densities, so that the posterior is itself automatically a density. Furthermore, as we will see below, there is a natural way to equip the prior with a natural built-in choice of the regularity hyperparameter, which will allow for adaptive inference.

1.3.3 Contribution using a Hierarchical approach with the Spike and Slab prior

An analog of the Spike and Slab prior In the following, one defines the cutoff $L_{max} = \log_2(n)$ and L the largest integer such that

$$2^L L \leq n \tag{1.3.14}$$

Note that $L \leq L_{max}$ for every n.

Let $X^{(n)} = (X_1, \cdots, X_n)$ be i.i.d. from law P with density f.
Let Π be the prior on densities generated as follows. One keeps the Pólya tree random measure with respect to the canonical dyadic partition of $[0, 1]$ construction up to level L, replacing the Beta distributions by

$$\varepsilon \in \mathcal{E} \ , \ Y_{\varepsilon 0} \sim (1-\pi_{\varepsilon 0})\delta_{\frac{1}{2}} + \pi_{\varepsilon 0}\text{Beta}(\alpha_{\varepsilon 0}, \alpha_{\varepsilon 1}), \tag{1.3.15}$$

with parameters $\alpha_\varepsilon \in \mathbb{N}$ to be chosen and a real parameter π_ε (later to be taken of the form $2^{-\frac{l}{2}}e^{-Cl}$, where we wrote $l = |\varepsilon|$).

There are multiple probability distributions on Borelians of $[0,1]$ that coincide on dyadic intervals I_ε with $P(I_\varepsilon)$ resulting from the above construction. We consider the specific one that is absolutely continuous relatively to the Lebesgue measure on $[0,1]$ with a constant density on each I_ε, $|\varepsilon| = L+1$. So, both prior and posterior are histograms on dyadic intervals at depth L.

Definition. The prior distribution with parameters α_ε, π_ε, as above is called Spike and Slab Pólya tree and denoted $\Pi(\alpha_\varepsilon, \pi_\varepsilon)$.

This prior is based on an idea of Ghosal and van der Vaart, which is referred as Evenly Split Pólya tree in their book Ghosal and van der Vaart (2017). First note that the Haar coefficients f_{lk} of a density f can be expressed as

$$f_{lk} = \langle f, \psi_{lk} \rangle = 2^{\frac{l}{2}} P(I_\varepsilon)(1 - 2Y_{\varepsilon 0}) \tag{1.3.16}$$

The Spike and Slab Pólya tree can therefore be seen as a 'thresholding prior', as the thresholding takes place on the sequence of Haar coefficients of the function where $Y_{\varepsilon 0} = \frac{1}{2}$.

Using this Spike and Slab prior can be seen as taking a Hierarchical approach. The usual Pólya tree (PT) prior on densities (under (1.3.13)) leads to the following Bayesian diagram

$$\begin{aligned} X|f &\sim f \\ f &\sim PT((Y_{\varepsilon 0})) \text{ with } Y_{\varepsilon 0} \sim \text{Beta}(\alpha_{\varepsilon 0}, \alpha_{\varepsilon 1}), \end{aligned}$$

so the $Y_{\varepsilon 0}$ have fixed (Beta) distributions, whereas the Spike and Slab Pólya tree (SSPT) prior leads to the diagram

$$\begin{aligned} X|f &\sim f \\ f &\sim SSPT((Y_{\varepsilon 0})) \text{ with } Y_{\varepsilon 0} \sim (1-\pi_{\varepsilon 0})\delta_{\frac{1}{2}} + \pi_{\varepsilon 0}\ \text{Beta}(\alpha_{\varepsilon 0}, \alpha_{\varepsilon 1}) \end{aligned}$$

which can be seen as the following diagram, using a sequence $(\gamma_{\varepsilon 0})_\varepsilon$ of Bernoulli variables.

$$\begin{aligned} X|f &\sim f \\ f|(\gamma_{\varepsilon 0}) &\sim SSPT((Y_{\varepsilon 0})) \text{ with } Y_{\varepsilon 0} \sim (1-\gamma_{\varepsilon 0})\delta_{\frac{1}{2}} + \gamma_{\varepsilon 0} \text{ Beta}(\alpha_{\varepsilon 0}, \alpha_{\varepsilon 1}) \\ \gamma_{\varepsilon 0} &\sim \text{Be}(\pi_{\varepsilon 0}) \end{aligned}$$

So in this case the distributions followed by the $Y_{\varepsilon 0}$ are random, hence this approach can be viewed as hierarchical.

The a posteriori law. Proposition 6 of Chapter 4 states that the Spike and Slab type Pólya tree still satisfies conjugacy. Indeed, for every $\varepsilon \in \mathcal{E}$, the a posteriori law of $Y_{\varepsilon 0}$ knowing $X_1, \ldots, X_n$ is

$$Y_{\varepsilon 0}|X \sim (1-\tilde{\pi}_{\varepsilon 0})\delta_{\frac{1}{2}} + \tilde{\pi}_{\varepsilon 0} Beta(\alpha_{\varepsilon 0}(X), \alpha_{\varepsilon 1}(X)) \tag{1.3.17}$$

where the quantities $\tilde{\pi}_\varepsilon$, $T = T(\varepsilon, X)$ and $\alpha_\varepsilon(X)$ all depend on the observations.
Note that if $\pi_\varepsilon = 1$, meaning that the prior is also a product of Beta variables, we get that the posterior is a product of Beta variables too.

An adaptive concentration result. The following Theorem shows that the a posteriori law obtained with a Spike and Slab type Pólya tree prior concentrates around the true density f_0 at minimax rate for the supremum-norm loss.

Theorem 11. Let $f_0 \in \mathcal{C}^\alpha[0,1]$, for $\alpha \in (0,1]$ and suppose $\|\log f_0\|_\infty < \infty$. Let $X_1, \ldots, X_n$ be i.i.d. random variables on $[0,1]$ following P_{f_0}. Let Π be the prior on densities induced by a Spike and Slab Polya Tree prior $\Pi(\alpha_\varepsilon, \pi_\varepsilon)$ with the choices

$$\begin{aligned} \alpha_\varepsilon &= a \\ \pi_\varepsilon &= 2^{-\frac{l}{2}} e^{-\kappa l}, \quad l = |\varepsilon| \end{aligned}$$

for κ large enough constant and $a > 0$ constant. Then for any $M_n \to \infty$, in P_{f_0}-probability

$$\Pi\left[\|f - f_0\|_\infty \leq M_n \left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+1}} \mid X\right] \to 1$$

This theorem is an adaptive version of Theorem 1 of Castillo (2017b). There are few results so far in the literature in density estimation for the supremum-norm loss, among those are the results from Castillo (2014), Hoffmann et al. (2015) and Yoo and Ghosal (2016) for multivariate regression.

A Bernstein Von Mises result. To establish a nonparametric Bernstein Von Mises (BVM) result, one has first to find a space $\mathcal{M}_0$ large enough to have convergence at rate $\sqrt{n}$ of the posterior density to a Gaussian process. One can then derive results for some other space $\mathcal{F}$ using continuous mapping for continuous functionals $\psi : \mathcal{M}_0 \to \mathcal{F}$. A space that combines nicely with supremum norm structure was introduced by Castillo and Nickl (2014) and defined as follows, using an 'admissible' sequence $\omega = (\omega_l)_{l\geq 0}$ such that $\omega_l/\sqrt{l} \to \infty$ as $l \to \infty$

$$\mathcal{M}_0 = \mathcal{M}_0(\omega) = \left\{ x = (x_{lk})_{l,k} \; ; \; \lim_{l\to\infty} \max_{0\leq k<2^l} \frac{|x_{lk}|}{\omega_l} = 0 \right\} \tag{1.3.18}$$

Equipped with the norm $\|x\|_{\mathcal{M}_0} = \sup_{l\geq 0} \max_{0\leq k<2^l} \frac{|x_{lk}|}{\omega_l}$, this is a separable Banach space. In a slight abuse of notation, we will write $f \in \mathcal{M}_0$ if the sequence of its Haar wavelet coefficients belongs in that space : $(\langle f, \psi_{lk}\rangle)_{l,k} \in \mathcal{M}_0$.

P-white bridge process. For P a probability distribution in $[0,1]$, one defines the P-white bridge process, denoted by $\mathbb{G}_P$. This is the Gaussian process indexed by the Hilbert space $L^2(P) = \{f : [0,1] \to \mathbb{R}; \int_0^1 f^2 dP < \infty\}$ with covariance

$$E[\mathbb{G}_P(f)\mathbb{G}_P(g)] = \int_0^1 (f - \int_0^1 f dP)(g - \int_0^1 g dP) dP \tag{1.3.19}$$

We will denote by $\mathcal{N}$ the law of $\mathbb{G}_{P_0}$(with $P_0 = P_{f_0}$).

The main purpose of the admissible sequence ω is to ensure that $\mathbb{G}_P \in \mathcal{M}_0$. Intuitively, if one does not use these weights w_l, the maximum over 2^l Gaussian variables is of order $\sqrt{2\log(2^l)} = C\sqrt{l}$ and does not tend to 0 as $l \to \infty$, see Remark 1 of Castillo and Nickl (2014) for a precise proof of this result.

Recentering the distribution. To establish our BVM result, one also has to find a suitable way to center the posterior distribution. In this view, denote by P_n the empirical measure

$$P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}. \tag{1.3.20}$$

Let us also consider C_n, which is a smoothed version of P_n, defined by

$$\langle C_n, \psi_{lk}\rangle = \begin{cases} \langle P_n, \psi_{lk}\rangle \text{ if } l \leq L \\ 0 \text{ if } l > L, \end{cases} \tag{1.3.21}$$

where L is our original cutoff, defined by (4.1.5).

We finally introduce T_n, which depends on the true parameter α, defined by

$$\langle T_n, \psi_{lk} \rangle = \begin{cases} \langle P_n, \psi_{lk} \rangle \text{ if } l \leq L_n \\ 0 \text{ if } l > L_n, \end{cases} \tag{1.3.22}$$

where we defined L_n to be the integer such that

$$2^{L_n} = \lfloor c_0 \left(\frac{n}{\log n} \right)^{\frac{1}{1+2\alpha}} \rfloor \tag{1.3.23}$$

for a suitable constant $c_0 \in \mathbb{R}^{+*}$, whose precise value is made clear below.

Weak BVM result. We have the following Bernstein-von Mises phenomenon for f_0 in Hölder-type balls (standard Hölder balls are subsets of the following ones)

$$\mathcal{H}(\alpha, R) := \{ f = (f_{lk}) : |f_{lk}| \leq R 2^{-(\alpha+1/2)}, \forall l \geq 0, 0 \leq k < 2^l \}$$

Theorem 12. Let $\mathcal{N}$ be the law of $\mathbb{G}_{P_0}$. Let C_n be the centering defined in (4.2.6). Let $l_0(n)$ be an increasing and diverging sequence. We define the prior Π such that

$$\begin{aligned} Y_{\varepsilon 0} &\sim \text{Beta}(a, a) \text{ for } |\varepsilon| \leq l_0 \\ Y_{\varepsilon 0} &\sim (1 - \pi_{\varepsilon 0}) \delta_{\frac{1}{2}} + \pi_{\varepsilon 0} \text{Beta}(a, a) \text{ for } l_0 < |\varepsilon| \leq L \end{aligned}$$

where $\pi_\varepsilon = 2^{-\frac{l}{2}} e^{-\kappa |\varepsilon|}$ with κ a large enough constant. The posterior distribution then satisfies a weak BvM : for every $\alpha, R > 0$, recalling β_S from (1.1.20),

$$\sup_{f_0 \in \mathcal{H}(\alpha, R)} E_{f_0} \left[\beta_{\mathcal{M}_0(\omega)} (\Pi(\cdot | X) \circ \tau_{C_n}^{-1}, \mathcal{N}) \right] \to 0$$

as $n \to \infty$ and for any admissible sequence $\omega = (\omega_l)$ with $\omega_{l_0(n)} / \sqrt{\log(n)} \to \infty$.

The choice of recentering of the distribution is quite flexible, as it can be checked that the result also holds if one replaces C_n by the posterior mean $\bar{f}_n$ or by T_n which depends on α. Actually, the only required condition on where one cuts the empirical measure is to satisfy Theorem 1 of Castillo and Nickl (2014). One can see that the cutoff L is exactly the furthest one can go according to that theorem.

Using the methods of Castillo and Nickl (2014), this result leads to several applications, for instance derivation of BVM theorems for semiparametric functionals via the continuous mapping theorem and Donsker-type theorems, which do not appear here for the sake of

brievity. It may also lead to the construction of adaptive credible sets although it may require substantial additional work.

Chapter 2

Empirical Bayes analysis of spike and slab posterior distributions

2.1 Introduction

In the sparse normal means model, one observes a sequence $X = (X_1, \ldots, X_n)$

$$X_i = \theta_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.1.1}$$

with $\theta = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n$ and $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d. $\mathcal{N}(0,1)$. Given θ, the distribution of X is a product of Gaussians and is denoted by P_θ. Further, one assumes that the 'true' vector θ_0 belongs to

$$\ell_0[s_n] = \{\theta \in \mathbb{R}^n,\ \#\{i :\ \theta_i \neq 0\} \leq s_n\},$$

the set of vectors that have at most s_n nonzero coordinates, where $0 \leq s_n \leq n$. A typical *sparsity* assumption is that s_n is a sequence that may grow with n but is 'small' compared to n (e.g. in the asymptotics $n \to \infty$, one typically assumes $s_n/n = o(1)$ and $s_n \to \infty$). A natural problem is that of estimating θ with respect to the euclidean loss $\|\theta - \theta'\|^2 = \sum_{i=1}^n (\theta_i - \theta_i')^2$. A benchmark is given by the minimax rate for this loss over the class of sparse vectors $\ell_0[s_n]$. Denoting

$$r_n := 2s_n \log(n/s_n),$$

Donoho et al. (1992) show that the minimax rate equals $(1 + o(1))r_n$ as $n \to \infty$.

Taking a Bayesian approach, one of the simplest and arguably most natural classes of prior distributions in this setting is given by so-called *spike and slab* distributions,

$$\theta \sim \bigotimes_{i=1}^{n} (1-\alpha)\delta_0 + \alpha G,$$

where δ_0 denotes the Dirac mass at 0, the distribution G has density γ with respect to Lebesgue measure and α belongs to $[0,1]$. These priors were introduced and advocated in a number of papers, including Mitchell and Beauchamp (1988); George (2000); George and Foster (2000); Yuan and Lin (2005). One important point is the calibration of the tuning parameter α, which can be done in a number of ways, including: deterministic n-dependent choice, data-dependent choice based on a preliminary estimate $\hat{\alpha}$, fully Bayesian choice based on a prior distribution on α. Studying the behaviour of the posterior distributions in sparse settings is currently the object of a lot of activity. A brief (and by far not exhaustive) overview of recent works is given below. Given a prior distribution Π on θ, and interpreting P_θ as the law of X given θ, one forms the posterior distribution $\Pi[\cdot \,|\, X]$ which is the law of θ given X. The frequentist analysis of the posterior distribution consists in the study of the convergence of $\Pi[\cdot \,|\, X]$ in probability under P_{θ_0}, thus assuming that the data has actually been generated from some 'true' parameter θ_0.

In the present paper, we follow this path and are more particularly interested in obtaining a uniform bound on the posterior squared L^2-moment of the order of the optimal minimax rate, that is in proving, with C a large enough constant,

$$\sup_{\theta_0 \in \ell_0[s_n]} E_{\theta_0} \int \|\theta - \theta_0\|^2 d\Pi(\theta \,|\, X) \leq C r_n \tag{2.1.2}$$

for Π a prior distribution constructed using a spike and slab approach, whose prior parameters may be calibrated using the data, that is following an empirical Bayes method. This is of interest for at least three reasons

- this provides adaptive convergence rates for the entire posterior distribution, using a fully data-driven procedure. This is more than obtaining convergence of aspects of the posterior such that posterior mean or mode, and in fact may require different conditions on the prior, as we shall see below.

- the inequality (2.1.2) automatically implies convergence of several commonly used point estimators derived from the posterior $\Pi[\cdot \,|\, X]$: it implies convergence at rate Cr_n of the posterior mean $\int \theta d\Pi(\theta \,|\, X)$ (using Jensen's inequality, see e.g. Castillo

and van der Vaart (2012)), but also of the coordinatewise posterior median (see the supplement of Castillo and van der Vaart (2012) for details) and in fact of any fixed posterior coordinatewise quantile, for instance the quantile 1/4 of $\Pi[\cdot \,|\, X]$. It also implies, using Tchebychev's inequality, convergence of the posterior distribution at rate $M_n r_n$ for $\|\cdot\|^2$ as in (2.1.3) below with $M = M_n$, for any $M_n \to \infty$.

- knowing (2.1.2) is a first step towards results for *uncertainty quantification*, in particular for the study of certain *credible sets*. Indeed, (2.1.2) suggests a natural way to build such a set, that is $\mathcal{C} \subset \mathbb{R}^n$ with $\Pi[\mathcal{C} \,|\, X] \geq 1 - \alpha$ for a given $\alpha \in (0, 1)$. Namely, define $\mathcal{C} = \{\theta : \ \|\theta - \bar{\theta}\|^2 \leq r_X\}$, with $\bar{\theta}$ the posterior mean (or another suitable point estimate of θ) and r_X a large enough multiple of the $(1-\alpha)$–quantile of $\int \|\theta - \bar{\theta}\|^2 d\Pi(\theta \,|\, X)$.

The present work is the first of a series of papers where we study aspects of inference using spike and slab prior distributions. In particular, based on the present results, the behaviour of the previously mentioned credible sets is studied in the forthcoming paper Castillo and Szabo (2018).

Previous results on frequentist analysis of spike and slab type priors. In a seminal paper, Johnstone and Silverman Johnstone and Silverman (2004) considered estimation of θ using spike and slab priors combined with an empirical Bayes method for choosing α. They chose $\alpha = \hat{\alpha}$ based on a marginal maximum likelihood approach to be described in more details below. Denoting $\hat{\theta}$ the associated posterior median (or posterior mean), Johnstone and Silverman (2004) established that

$$\sup_{\theta_0 \in \ell_0[s_n]} E_{\theta_0} \|\hat{\theta} - \theta_0\|^2 \leq C r_n,$$

thereby proving minimaxity up to a constant of this estimator over $\ell_0[s_n]$. The estimator is adaptive, as the knowledge of s_n is not required in its construction.

In Castillo and van der Vaart (2012), convergence of the posterior distribution is studied in the case α is given a prior distribution. If $\alpha \sim \text{Beta}(1, n+1)$, Π is the corresponding hierarchical prior, and $\Pi[\cdot \,|\, X]$ the associated posterior distribution, it is established in Castillo and van der Vaart (2012) that for large enough M, as $n \to \infty$,

$$\sup_{\theta_0 \in \ell_0[s_n]} E_{\theta_0} \Pi[\|\theta - \theta_0\|^2 \leq M r_n \,|\, X] \to 1. \tag{2.1.3}$$

In Martin and Walker (2014), Martin and Walker use a fractional likelihood approach to construct a certain empirical Bayes spike and slab prior, where the idea is to reweight

the standard spike and slab prior by a power of the likelihood. They derive rate-optimal concentration results for the corresponding posterior distribution and posterior mean.

A related class of prior distributions recently put forward by Ročková Ročková (2018) and Ročková and George Ročková and George (2018), is given by

$$\theta \sim \bigotimes_{i=1}^{n} (1-\alpha)G_0 + \alpha G_1,$$

where both distributions G_0, G_1 have densities with respect to Lebesgue measure. The authors in particular consider the choices $G_0 = \text{Lap}(\lambda_0)$ and $G_1 = \text{Lap}(\lambda_1)$, where $\text{Lap}(\lambda)$ denotes the Laplace (double-exponential) distribution. Taking λ_0 large enough enables one to mimic the spike of the standard spike and slab prior, and the fact that both G_0, G_1 are continuous distributions offers some computational advantages, especially when working with the posterior mode. One can also note that the posterior mode when $\alpha = 1$ leads to the standard LASSO estimator. For this reason, the authors in Ročková (2018); Ročková and George (2018) call this prior the *spike and slab LASSO* prior. It is shown in Ročková (2018), Theorem 5.2 and corollaries, that a certain deterministic n-dependent choice of $\alpha, \lambda_0, \lambda_1$ (but independent on the unknown s_n) leads to posterior convergence at near-optimal rate $s_n \log n$, while putting a prior on α can yield (Ročková (2018), Theorem 5.4) the minimax rate for the posterior, if a certain condition on the strength of the true non-zero coefficiencents of θ_0 is verified.

Other priors and related work. We briefly review other options to induce sparsity using a Bayesian approach. One option considered in Castillo and van der Vaart (2012) is first to draw a subset $S \subset \{1, \ldots, n\}$ at random and then to draw nonzero coordinates on this subset only. That is, sample first a dimension $k \in \{0, \ldots, n\}$ at random according to some prior π. Given k, sample S uniformly at random over subsets of size k and finally set

$$\begin{aligned} \theta_i &\sim G \quad i \in S \\ \theta_i &= 0 \quad i \notin S. \end{aligned}$$

Under the assumption that the prior π on k is of the form, referred to as the complexity prior,

$$\pi(k) = ce^{-ak\log(nb/k)}, \tag{2.1.4}$$

Castillo and van der Vaart (2012) show that under this prior, both (2.1.3) and (2.1.2) are satisfied. However, such a 'strong' prior on the dimension is not necessary at least for (2.1.3) to hold: it can be checked for instance, for π the prior on dimension induced

by the spike and slab prior on θ with $\alpha \sim \text{Beta}(1, n+1)$, that $\pi(s_n) \asymp \exp(-cs_n) \gg \exp(-cs_n \log(n/s_n))$. So in a sense the complexity prior 'penalises slightly more than necessary'.

Another popular way to induce sparsity is via the so-called *horseshoe* prior, which draws a θ from a continous distribution which is itself a mixture. As established in van der Pas et al. (2017a)–van der Pas et al. (2017b) the horseshoe yields the nearly-optimal rate $s_n \log n$ uniformly over the whole space $\ell_0[s_n]$, up again to the correct form of the logarithmic factor. In a different spirit but still without using Dirac masses at 0, the paper Jiang and Zhang (2009) shows that, remarkably, it is also possible to adopt an empirical Bayes approach on the entire unknown distribution function F of the vector θ, interpreting θ as sampled from a certain distribution, and the authors derive oracle results over $\ell^p, p > 0$, balls for the plug-in posterior mean (not including the case $p = 0$ though). We also note the interesting work van der Pas et al. (2016) that investigates necessary and sufficient conditions for sparse continuous priors to be rate-optimal. However the latter is for a fixed regularity parameter s_n, while the results decribed in Section 2.2 (in particularity the suboptimality phenomenon, but also upper-bounds using the empirical Bayes approach) are related to adaptation.

Using complexity–type priors on the number of non-zero coordinates, Belitser and co-authors Babenko and Belitser (2010)–Belitser and Nurushev (2015) consider Gaussian priors on non-zero coefficients, with a recentering of the posterior mean at the observation X_i– for those coordinates i that are selected– to adjust for overshrinkage. In Belitser and Nurushev (2015), oracle results for the corresponding posterior are derived, that in particular imply convergence at the minimax rate up to constant over $\ell_0[s_n]$, and the authors also derive results on uncertainty quantification by studying the frequentist coverage of credible sets using their procedure.

For further references on the topic, in particular about relationships between spike and slab priors and absolutely continuous counterparts such as the horseshoe or the spike and slab LASSO, we refer to the paper van der Pas et al. (2017b) and its discussion by several authors of the previously mentioned works.

Overview of results and outline. This paper obtains the following results.

1. For the spike and slab prior, in Section 2.2.2 we establish lower bound results that show that the popular Laplace slab yields suboptimal rates when the complete empirical Bayes posterior is considered.

2. In Sections 2.2.3 and 2.2.6, we establish rate-optimal results for the posterior squared L^2–moment for the usual spike and slab with a Cauchy slab, when the prior hyperparameter is chosen via a marginal maximum likelihood method.

3. In Section 2.2.4, the spike and slab LASSO prior is considered and we provide a near-optimal adaptive rate for the corresponding complete empirical Bayes posterior distribution.

Section 2.2 introduces the framework, notation, and the main results, ending with a brief simulation study in Section 2.2.5 and discussion. Section 2.3 gathers the proofs of the lower-bound results as well as upper-bounds on the spike and slab prior. Technical lemmas for the spike and slab prior can be found in Section 2.4, while Sections 2.5–2.6 contain the proof of the result for the spike and slab LASSO prior.

For real-valued functions f, g, we write $f \lesssim g$ if there exists a universal constant C such that $f(x) \leq Cg(x)$, and $f \gtrsim g$ is defined similarly. When x is a positive real number or an integer, we write $f(x) \asymp g(x)$ if there exists positive constants c, C, D such that for $x \geq D$, we have $cf(x) \leq g(x) \leq Cf(x)$. For reals a, b, one denotes $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$.

2.2 Framework and main results

2.2.1 Empirical Bayes estimation with spike and slab prior

In the setting of model (2.1.1), the spike and slab prior on θ with fixed parameter $\alpha \in [0, 1]$ is

$$\Pi_\alpha \sim \otimes_{i=1}^n (1 - \alpha)\delta_0 + \alpha G(\cdot), \tag{2.2.1}$$

where G is a given probability measure on $\mathbb{R}$. We consider the following choices

$$G = \begin{cases} & \text{Lap}(1) \\ or & \\ & \text{Cauchy}(1) \end{cases}$$

where $\text{Lap}(\lambda)$ denotes the Laplace (double exponential) distribution with parameter λ and $\text{Cauchy}(1)$ the standard Cauchy distribution. Different choices of parameters and prior distributions are possible (a brief discussion is included below) but for clarity of exposition we stick to these common distributions. In the sequel γ denotes the density of G with respect to Lebesgue measure.

By Bayes' formula the posterior distribution under (2.1.1) and (2.2.1) with fixed $\alpha \in [0, 1]$ is

$$\Pi_\alpha[\cdot \mid X] \sim \otimes_{i=1}^n (1 - a(X_i))\delta_0 + a(X_i)G_{X_i}(\cdot), \tag{2.2.2}$$

where, denoting by ϕ the standard normal density and $g(x) = \phi * G(x) = \int \phi(x-u)dG(u)$ the convolution of ϕ and G at point $x \in \mathbb{R}$, the posterior weight $a(X_i)$ is given by, for any i,

$$a(X_i) = a_\alpha(X_i) = \frac{\alpha g(X_i)}{(1-\alpha)\phi(X_i) + \alpha g(X_i)}. \tag{2.2.3}$$

The distribution G_{X_i} has density

$$\gamma_{X_i}(\cdot) := \frac{\phi(X_i - \cdot)\gamma(\cdot)}{g(X_i)} \tag{2.2.4}$$

with respect to Lebesgue measure on $\mathbb{R}$. The behaviour of the posterior distribution $\Pi_\alpha[\cdot \,|\, X]$ heavily depends on the choices of the smoothing parameters α and γ. It turns out that some aspects of this distribution are thresholding-type estimators, as established in Johnstone and Silverman (2004).

Posterior median and threshold $t(\alpha)$. The posterior median $\hat{\theta}_\alpha^{med}(X_i)$ of the ith coordinate has a thresholding property: there exists $t(\alpha) > 0$ such that $\hat{\theta}_\alpha^{med}(X_i) = 0$ if and only if $|X_i| \leq t(\alpha)$. A default choice can be $\alpha = 1/n$; one can check that this leads to a posterior median behaving similarly as a hard thresholding estimator with threshold $\sqrt{2\log n}$. One can significantly improve on this default choice by taking a well-chosen data-dependent α.

In order to choose α, in this paper we follow the empirical Bayes method proposed in Johnstone and Silverman (2004). The idea is to estimate α by maximising the marginal likelihood in α in the Bayesian model, which is the density of $\alpha \,|\, X$. The log-marginal likelihood in α can be written as

$$\ell(\alpha) = \ell_n(\alpha; X) = \sum_{i=1}^{n} \log((1-\alpha)\phi(X_i) + \alpha g(X_i)). \tag{2.2.5}$$

Let $\hat{\alpha}$ be defined as the maximiser of the log-marginal likelihood

$$\hat{\alpha} = \operatorname*{argmax}_{\alpha \in \mathcal{A}_n} \ell_n(\alpha; X), \tag{2.2.6}$$

where the maximisation is restricted to $\mathcal{A}_n = [\alpha_n, 1]$, with α_n defined by

$$t(\alpha_n) = \sqrt{2\log n}.$$

The reason for this restriction is that one does not need to take α smaller than α_n, which would correspond to a choice of α 'more conservative' than hard-thresholding at threshold level $\sqrt{2\log n}$.

In Johnstone and Silverman (2004), Johnstone and Silverman prove that the posterior median $\hat{\alpha}^{med}(X_i)$ has remarkable optimality properties, for many choices of the slab density γ. For γ with tails 'at least as heavy as' the Laplace distribution, then this point estimator converges at the minimax rate over $\ell_0[s_n]$. More precisely, it follows from Theorem 1 in Johnstone and Silverman (2004) that there exists constants C, c_0, c_1 such that if

$$c_1 \log^2 n \leq s_n \leq c_0 n, \tag{2.2.7}$$

then the posterior median $\hat{\theta}_{\hat{\alpha}}^{med} = (\hat{\theta}_{\hat{\alpha}}^{med}(X_i))_{1\leq i\leq n}$ is rate optimal

$$\sup_{\theta\in\ell_0[s_n]} E_\theta \|\hat{\theta}_{\hat{\alpha}}^{med} - \theta\|^2 \leq C s_n \log(n/s_n). \tag{2.2.8}$$

One can actually remove the lower bound in condition (2.2.7) – see Theorem 2 in Johnstone and Silverman (2004) – by a more complicated choice of $\hat{\alpha}$, for which $\hat{\alpha}$ in (2.2.6) is replaced by a smaller value if the empirical Bayes estimate is close to α_n given by $t(\alpha_n) = \sqrt{2\log n}$. In the present paper for simplicity of exposition we first work under the condition (2.2.7). In Section 2.2.6, we show that the lower bound part of the condition can be removed when working with the modified estimator as in Johnstone and Silverman (2004).

Plug-in posterior distribution. The posterior we consider in this paper is $\Pi_{\hat{\alpha}}[\cdot \,|\, X]$, that is the distribution given by (2.2.2), where α has been replaced by its empirical Bayes (EB) estimate $\hat{\alpha}$ given by (2.2.6). This posterior is called complete EB posterior in the sequel. The value $\hat{\alpha}$ is easily found numerically, as implemented in the R package `EbayesThresh`, see Johnstone and Silverman (2005). As noted in Johnstone and Silverman (2004), the posterior median $\hat{\alpha}^{med}(X_i)$ displays excellent behaviour in simulations. However, the entire posterior distribution $\Pi_{\hat{\alpha}}[\cdot \,|\, X]$ has not been studied so far. It turns out that the behaviour of the posterior median does not always reflect the behaviour of the complete posterior, as is seen in the next subsection.

2.2.2 Suboptimality of the Laplace slab for the complete EB posterior distribution

Theorem 13. Let Π_α be the spike and slab prior distribution (2.2.1) with slab distribution G equal to the Laplace distribution Lap(1). Let $\Pi_{\hat{\alpha}}[\cdot \,|\, X]$ be the corresponding plug-in posterior distribution given by (2.2.2), with $\hat{\alpha}$ chosen by the empirical Bayes procedure (2.2.6). There exist $D > 0$, $N_0 > 0$, and $c_0 > 0$ such that, for any $n \geq N_0$ and any s_n

with $1 \leq s_n \leq c_0 n$, there exists $\theta_0 \in \ell_0[s_n]$ such that,

$$E_{\theta_0} \int \|\theta - \theta_0\|^2 d\Pi_{\hat{\alpha}}[\theta \,|\, X] \geq D s_n e^{\sqrt{\log(n/s_n)}}.$$

Theorem 13 implies that taking a Laplace slab leads to a suboptimal convergence rate in terms of the posterior squared L^2–moment. This result is surprising at first, as we know by (2.2.8) that the posterior median converges at optimal rate r_n. The posterior mean also converges at rate r_n uniformly over $\ell_0[s_n]$, by Theorem 1 of Johnstone and Silverman (2004). So at first sight it would be quite natural to expect that so does the posterior second moment.

One can naturally ask whether the suboptimality result from Theorem 13 could come from considering an integrated L^2–moment, instead of simply asking for a posterior convergence result in probability, as is standard in the posterior rates literature following Ghosal et al. (2000). We now derive a stronger result than Theorem 13 under the mild condition $s_n \gtrsim \log^2 n$. The fact that the result is stronger follows from bounding from below the integral in the display of Theorem 13 by the integral restricted to the set where $\|\theta - \theta_0\|^2$ is larger than the target lower bound rate.

Theorem 14. Under the same notation as in Theorem 13, if Π_α is a spike and slab distribution with as slab G the Laplace distribution, there exists $m > 0$ such that for any s_n with $s_n/n \to 0$ and $\log^2 n = O(s_n)$ as $n \to \infty$, there exists $\theta_0 \in \ell_0[s_n]$ such that, as $n \to \infty$,

$$E_{\theta_0} \Pi_{\hat{\alpha}} \left[\|\theta - \theta_0\|^2 \leq m s_n e^{\sqrt{2\log(n/s_n)}} \,|\, X \right] = o(1).$$

Theorem 14, by providing a lower bound in the spirit of Castillo (2008), shows that the answer to the above question is negative, and for a Laplace slab, the plug-in posterior $\Pi_{\hat{\alpha}}[\cdot \,|\, X]$ does not converge at minimax rate uniformly over $\ell_0[s_n]$.

Note that the suboptimality occuring here does not result from an artificially constructed example (we work under exactly the same framework as Johnstone and Silverman (2004)) and that this has important (negative) consequences for construction of credible sets. Due to the rate-suboptimality of the EB Laplace-posterior, typical credible sets derived from it (such as, e.g., taking quantiles of a recentered posterior second moment) will inherit the suboptimality in terms of their diameter, and therefore will not be of optimal size. Fortunately, it is still possible to achieve optimal rates for certain spike and slab EB posteriors: the previous phenomenon indeed disappears if the tails of the slab in the prior distribution are heavy enough, as seen in the next subsection.

2.2.3 Optimal posterior convergence rate for the EB spike and Cauchy slab

The next result considers Cauchy tails, although other examples can be covered, as discussed below. In the sequel, we abbreviate by SAS prior a spike and slab prior with Cauchy slab.

Theorem 15. Let Π_α be the SAS prior distribution (2.2.1) with slab distribution G equal to the standard Cauchy distribution. Let $\Pi_{\hat\alpha}[\cdot \,|\, X]$ be the corresponding plug-in posterior distribution given by (2.2.2), with $\hat\alpha$ chosen by the empirical Bayes procedure (2.2.6). There exist $C > 0$, $N_0 > 0$, and $c_0, c_1 > 0$ such that, for any $n \geq N_0$, for any s_n such that (2.2.7) is satisfied for such c_0, c_1,

$$\sup_{\theta_0 \in \ell_0[s_n]} E_{\theta_0} \int \|\theta - \theta_0\|^2 d\Pi_{\hat\alpha}(\theta \,|\, X) \leq C r_n.$$

If one only assumes $s_n \leq c_0 n$ in (2.2.7), then the last statement holds with the bound Cr_n replaced by $Cr_n + C \log^3 n$.

Theorem 15 confirms that the empirical Bayes plug-in posterior, with $\hat\alpha$ chosen by marginal maximum likelihood, converges at optimal rate with precise logarithmic factor, at least under the mild condition (2.2.7), if tails of the slab distribution are heavy enough. Inspection of the proof of Theorem 15 reveals that any slab density γ with tails of the order $x^{-1-\delta}$ with $\delta \in (0, 2)$ gives the same result. Sensibility to the tails, in particular in view of posterior convergence in terms of d_q-distances, will be further investigated in Castillo and Szabo (2018).

We note that the horseshoe prior on θ considered in van der Pas et al. (2017a)–van der Pas et al. (2017b) also has Cauchy-like tails, which seems to confirm that for empirical Bayes–calibrated (product–type) sparse priors, heavy tails are important to ensure optimal or near-optimal behaviour, see also the discussion Castillo (2017a).

The lower bound in condition (2.2.7) is specific to the estimate $\hat\alpha$. Note that in the very sparse regime where $s_n \leq c_1 \log^2 n$, the rate is no more than $C \log^3 n$, thus missing the minimax rate by at most a logarithmic factor. This lower bound on s_n can be removed and the minimax rate obtained over the whole range of sparsities s_n if one modifies slightly $\hat\alpha$, where the estimator is changed if $\hat\alpha$ is too close to the lower boundary of the maximisation interval, see Section 2.2.6.

2.2.4 Posterior convergence for the EB spike and slab LASSO

Now consider the following prior on θ with fixed parameter $\alpha \in [0,1]$

$$\Pi_\alpha \sim \otimes_{i=1}^n (1-\alpha) G_0(\cdot) + \alpha G_1(\cdot), \tag{2.2.9}$$

where for $k = 0, 1$, G_k is given by

$$G_0 = \text{Lap}(\lambda_0), \quad G_1 = \begin{cases} & \text{Lap}(\lambda_1) \\ \text{or} & \\ & \text{Cauchy}(1/\lambda_1), \end{cases}$$

which leads to the spike and slab LASSO prior of Ročková and George (2018) in the case of a Laplace G_1, and to a heavy-tailed variant of the spike and slab LASSO if G_1 is Cauchy$(1/\lambda_1)$, that is if its density is $\gamma_1(x) = (\lambda_1/\pi)(1+\lambda_1^2 x^2)^{-1}$. In this setting γ_0, γ_1 denote the densities of G_0, G_1 with respect to Lebesgue measure. We call *SSL prior* a spike and slab LASSO prior with Cauchy slab.

By Bayes' formula the posterior distribution under (2.1.1) and (2.2.9) with fixed $\alpha \in [0,1]$ is

$$\Pi_\alpha[\cdot \,|\, X] \sim \otimes_{i=1}^n (1-a(X_i)) G_{0,X_i}(\cdot) + a(X_i) G_{1,X_i}(\cdot), \tag{2.2.10}$$

where $g_k(x) = \phi * G_k(x) = \int \phi(x-u) dG_k(u)$ is the convolution of ϕ and G_k at point $x \in \mathbb{R}$ for $k = 0, 1$, the posterior weight $a(X_i)$ is defined through the function $a(\cdot)$ given by

$$a(x) = a_\alpha(x) = \frac{\alpha g_1(x)}{(1-\alpha) g_0(x) + \alpha g_1(x)},$$

and if G_k has density γ_k with respect to Lebesgue measure, the distribution G_{k,X_i} has density

$$\gamma_{k,X_i}(\cdot) := \frac{\phi(X_i - \cdot)\gamma_k(\cdot)}{g_k(X_i)}.$$

In slight abuse of notation, we keep the same notation in the case of the SSL prior for quantities such as $a(x)$ or $\hat{\alpha}$ below, as it will always be clear from the context which prior we work with.

We consider the following specific choices for the constants λ_0, λ_1

$$\begin{cases} \lambda_0 &= L_0 n, & L_0 = 5\sqrt{2\pi}, \\ \lambda_1 &= L_1, & L_1 = 0.05. \end{cases} \tag{2.2.11}$$

The choice of the constants L_0, L_1 is mostly for technical convenience, and is similar to that of, e.g. Corollary 5.2 in Ročková (2018). Any other constant L_0 (resp. L_1) larger (resp. smaller) than the above value also works for the following result. The above numerical values may not be optimal.

Let $\hat{\alpha}$ be defined as the maximiser of the log-marginal likelihood,

$$\hat{\alpha} = \underset{\alpha \in [\mathcal{C} \log n/n, 1]}{\operatorname{argmax}} \ell_n(\alpha; X), \tag{2.2.12}$$

for $\mathcal{C} = \mathcal{C}_0(\gamma_0, \gamma_1)$ a large enough constant to be chosen below (this ensures that $\hat{\alpha}$ belongs to an interval on which we can verify that β is increasing, see (2.5.2)). This time we do not have access to the threshold t, since for the SSL prior the posterior median is not a threshold estimator, so here $\mathcal{C} \log n/n$ plays the role of an approximated version of α_n in (2.2.6).

Theorem 16. Let Π_α be the SSL prior distribution (2.2.9) with Cauchy slab and parameters (λ_0, λ_1) given by (2.2.11). Let $\Pi_{\hat{\alpha}}[\cdot \,|\, X]$ be the corresponding plug-in posterior distribution given by (2.2.10), with $\hat{\alpha}$ chosen by the empirical Bayes procedure (2.2.12). There exist $C > 0$, $N_0 > 0$, and $c_0, c_1 > 0$ such that, for any $n \geq N_0$, for any s_n such that (2.2.7) is satisfied for such c_0, c_1, then

$$\sup_{\theta_0 \in \ell_0[s_n]} E_{\theta_0} \int \|\theta - \theta_0\|^2 d\Pi_{\hat{\alpha}}(\theta \,|\, X) \leq C s_n \log n.$$

If one only assumes $s_n \leq c_0 n$ in (2.2.7), then the last bound holds with $C s_n \log n$ replaced by $C(s_n \log n + \log^3 n)$.

This result is an SSL version of Theorem 15. It shows that a spike and slab LASSO prior with heavy-tailed slab distribution and empirical Bayes choice of the weight parameter leads to a nearly optimal contraction rate for the entire posterior distribution. Hence it provides a theoretical guarantee of a fully data-driven procedure of calibration of the smoothing parameter in SSL priors.

2.2.5 A brief numerical study

Theorems 13–14 imply that the posterior distribution for the spike and slab prior and Laplace(1) slab does not converge at optimal rate and the discrepancy between the actual rate and the minimax rate for some 'bad' θ_0s is at least of order

$$R_n = \frac{\exp\left(\sqrt{2 \log(n/s_n)}\right)}{\log(n/s_n)},$$

up to a multiplicative constant factor, as both lower and upper bounds are up to a constant. Note that R_n grows more slowly than a polynomial in n/s_n, so the sub-optimality effect will typically be only visible for quite large values of n/s_n. For instance, if $n = 10^4$ and $s_n = 10$, one has $R_n \approx 6$, which is quite small given that an extra multiplicative constant is also involved.

For the present simulation study we took $n = 10^7$, $s_n = 10$, for which $R_n \approx 13.9$, and the non-zero values of θ_0 equal to $\{2\log(n/s_n)\}^{1/2}$, as the lower bound proof of Theorems 13–14 suggests. We computed $\hat{\alpha}$ using the package `EBayesThresh` of Johnstone and Silverman Johnstone and Silverman (2005) and computed $\int \|\theta - \theta_0\|_2^2 d\Pi_{\hat{\alpha}}(X)$ using its explicit expression, which can be obtained in closed form for a Laplace slab, with similar computations as in Johnstone and Silverman (2005), Section 6.3. We then took the empirical average over 100 repetitions to estimate the target expectation $R_2 := E_{\theta_0} \int \|\theta - \theta_0\|_2^2 d\Pi_{\hat{\alpha}}(X)$. We first took $\gamma = \text{Lap}(1)$ a standard Laplace slab and obtained $\hat{R}_2 \approx 1110$. For comparison, we computed the empirical quadratic risk $\hat{R}_{mean}$ for the posterior mean (approximating $E_{\theta_0}\|\hat{\theta}^{mean} - \theta_0\|^2$) and $\hat{R}_{median}$ the posterior median of the same posterior, obtaining $\hat{R}_{mean} \approx 158$ and $\hat{R}_{median} \approx 167$. So, in this case $\hat{R}_2$ is already 6 to 7 times larger than the risk of either mean or median.

To further illustrate the 'blow-up' in the rate for the posterior second moment R_2, we took a Laplace slab $\text{Lap}(a)$ with inverse-scale parameter a, for which the numerator in the definition of R_n becomes $\exp\{a\sqrt{2\log(n/s_n)}\}$ (let us also note that the multiplicative constant we refer to above also depends on a). The same simulation experiment as above was conducted, with the standard Laplace slab replaced by a $\text{Lap}(a)$ slab, for different values of a. The numerical results are presented in Table 2.1, which feature a noticeable increase in the second moment $\hat{R}_2$, while the risks of posterior mean and median stay around the same value, as expected.

a	0.5	1	1.5	2	2.5	3	3.5
Second moment	394	1110	2847	5716	8093	16530	34791
Median	173	167	169	174	185	209	219
Mean	157	158	166	172	182	224	336

Table 2.1 Empirical risks $\hat{R}_2, \hat{R}_{med}, \hat{R}_{mean}$ for Laplace slabs $\text{Lap}(a)$ and $a \in [0.5, 3.5]$

We also performed the same experiments for the quasi-Cauchy slab prior introduced in Johnstone and Silverman (2004)-Johnstone and Silverman (2005) (it is very close to the standard Cauchy slab – in particular it has the same Cauchy tails – but more convenient from the numerical perspective, see Johnstone and Silverman (2005), Section

6.4). We found $\hat{R}^{median} \approx 192$, $\hat{R}^{mean} \approx 191$ for the posterior mean and $\hat{R}_2 \approx 287$ for the posterior second moment. This time, as expected, the posterior second moment is not far from the two other risks.

2.2.6 Modified empirical Bayes estimator

For $n \geq 3$ and $A \geq 0$, let us set $t_n^2 = 2\log n - 5\log\log n$ and $t_A = \sqrt{2(1+A)\log n}$. For Π_α the SAS prior with a Cauchy slab, let as before $t(\alpha)$ be the posterior median threshold for fixed α. It is not hard to check that $t(\cdot)$ is continuous and strictly decreasing so has an inverse (see Johnstone and Silverman (2004), Section 5.3). In a similar fashion as in Johnstone and Silverman (2004), Section 4, let us introduce a modified empirical Bayes estimator as, for $A \geq 0$ and $\hat{t} := t(\hat{\alpha})$, $\alpha_A := t^{-1}(t_A)$,

$$\hat{\alpha}_A = \begin{cases} \hat{\alpha}, & \text{if } \hat{t} \leq t_n, \\ \alpha_A, & \text{if } \hat{t} > t_n. \end{cases} \tag{2.2.13}$$

Theorem 17. Let Π_α be the SAS prior distribution with slab distribution G equal to the standard Cauchy distribution. For a fixed $A > 0$, let $\Pi_{\hat{\alpha}_A}[\cdot \,|\, X]$ be the corresponding plug-in posterior distribution, with $\hat{\alpha}_A$ the *modified* estimator (2.2.13). There exist $C, c_0 > 0$, $N_0 > 0$, such that, for any $n \geq N_0$, for any s_n such that $s_n \leq c_0 n$,

$$\sup_{\theta_0 \in \ell_0[s_n]} E_{\theta_0} \int \|\theta - \theta_0\|^2 d\Pi_{\hat{\alpha}}(\theta \,|\, X) \leq C r_n.$$

Theorem 17 shows that the plug-in SAS posterior distribution using the modified estimator (2.2.13), $A > 0$, and a Cauchy slab attains the minimax rate of convergence r_n even in the very sparse regime $s_n \lesssim \log^2 n$, for which the unmodified estimate of Theorem 15 may lose a logarithmic factor.

2.2.7 Discussion

In this paper, we have developped a theory of empirical Bayes choice of the hyperparameter of spike and slab prior distributions. It extends the work of Johnstone and Silverman Johnstone and Silverman (2004) in that here the complete EB posterior distribution is considered. One important message is that such a generalisation preserves optimal convergence rates at the condition of taking slab distributions with heavy enough tails. If the tails of the slab are only moderate (e.g. Laplace), then the complete EB posterior rate may be suboptimal. This is in contrast with the hierarchical case considered in Castillo

and van der Vaart (2012), where a Laplace slab combined with a Beta distributed prior on α was shown to lead to an optimal posterior rate. On the one hand, the empirical Bayes method often leads to simpler or/and more easily tractable practical algorithms; on the other hand, we have illustrated here that the complete EB posterior may in some cases need slightly stronger conditions to conserve optimal theoretical guarantees. This phenomenon had not been pointed out so far in the literature, to the best of our knowledge.

We also note that Theorem 15 (or Theorem 17 if one allows for very sparse signals) enables one to recover the optimal form of the logarithmic factor $\log(n/s_n)$ in the minimax rate. This entails significant work, as one needs to control the empirical Bayes weight estimate $\hat{\alpha}$ both from above *and below*. This could work too in the SSL setting of Theorem 16, although this seems to need substantial extra technical work.

Looking at Theorems 13 and 14, it is natural to wonder why the Empirical Bayes approach fails for the Laplace slab where the full Bayes approach succeeds as seen in Castillo and van der Vaart (2012) Theorem 2.2. The reason why the hierarchical Bayes version works also for γ Laplace is the extra penalty in model size induced by the hierarchical prior on dimension. Indeed, in the full Bayes approach, the posterior distribution of α given X has density

$$f_{\alpha \,|\, X}(\alpha) \propto p(X \,|\, \alpha)\pi(\alpha),$$

where $p(X \,|\, \alpha)$ is the marginal density one maximises when considering the MMLE $\hat{\alpha}$. Hence adding a term $\log \pi(\alpha)$ for well-chosen π – for instance that arising from a Beta$(1, n+1)$ prior on α as considered in Castillo and van der Vaart (2012) – to the log-marginal likelihood one maximises forces $\hat{\alpha}$ to concentrate on smaller values. For instance, in the present setting, one could consider a penalised log-marginal maximum likelihood, which would force the estimate $\hat{\alpha}$ to concentrate on slightly smaller values, which would allow one to avoid the extra $e^{\sqrt{\log n/s_n}}$ term arising in Theorems 13–14.

The present work can also serve as a basis for constructing confidence regions using spike-and-slab posterior distributions. This question is considered in the forthcoming paper Castillo and Szabo (2018).

2.3 Proofs for the spike and slab prior

Let us briefly outline the ingredients of the proofs to follow. For Theorems 13 and 15, our goal is to bound the expected posterior risk $R_n(\theta_0) = E_{\theta_0} \int \|\theta - \theta_0\|^2 d\Pi_{\hat{\alpha}}(\theta \,|\, X)$. There

are three main tools. First, after introducing notation and basic bounds in Section 2.3.1, bounds on the posterior risk for fixed α are given in Section 2.3.2, as well as corresponding bounds for random α. Let us note that the corresponding upper bounds are different from those obtained on the quadratic risk for the posterior median in Johnstone and Silverman (2004) (and in fact, must be, in view of the negative result in Theorem 13). Second, inequalities on moments of the score function are stated in Section 2.3.3. As a third tool, we obtain deviation inequalities on the location of $\hat{\alpha}$ in Section 2.3.4. One of the bounds sharpens the corresponding bound from Johnstone and Silverman (2004) in case the signal belongs to the nearly-black class $\ell_0[s_n]$ which we assume here.

Proofs of Theorems 13 and 15 are given in Sections 2.3.5 and 2.3.6. For Theorem 15, we also needed to slightly complete the proof of one of the inequalities on thresholds stated in Johnstone and Silverman (2004), see Lemma 11. The proof of Theorem 14, which uses ideas from both previous proofs, is given in Section 2.3.7. Proofs of technical lemmas for the SAS prior are given in Section 2.4.

2.3.1 Notation and tools for the SAS prior

Expected posterior L^2–squared risk. For a fixed weight α, the posterior distribution of θ is given by (2.2.2). On each coordinate, the mixing weight $a(X_i)$ is given by (2.2.3) and the density of the non-zero component γ_{X_i} by (2.2.4). In the sequel we will obtain bounds on the following quantity, already for a given $\alpha \in [0,1]$,

$$\int \|\theta - \theta_0\|^2 d\Pi_\alpha(\theta \,|\, X) = \sum_{i=1}^n \int (\theta_i - \theta_{0,i})^2 d\Pi_\alpha(\theta_i \,|\, X_i).$$

To do so, we study $r_2(\alpha, \mu, x) := \int (u-\mu)^2 d\pi_\alpha(u \,|\, x)$, where $\pi_\alpha(\cdot \,|\, x) \sim (1 - a(x))\delta_0 + a(x)\gamma_x(\cdot)$. By definition

$$r_2(\alpha, \mu, x) = (1 - a(x))\mu^2 + a(x) \int (u-\mu)^2 \gamma_x(u) du.$$

This quantity is controlled by $a(x)$ and the term involving γ_x. From the definition of $a(x)$, bounding the denominator from below by one of its two components, and using $a(x) \leq 1$ yields, for any real x and $\alpha \in [0,1]$,

$$\alpha \frac{g}{g \vee \phi}(x) \leq a(x) \leq 1 \wedge \frac{\alpha}{1-\alpha} \frac{g}{\phi}(x). \tag{2.3.1}$$

The marginal likelihood in α. By definition, the empirical Bayes estimate $\hat{\alpha}$ in (2.2.6) maximises the logarithm of the marginal likelihood in α in (2.2.5). In case the maximum

is not taken at the boundary, $\hat{\alpha}$ is a zero of the derivative (score) of the previous likelihood. Its expression is $S(\alpha) = \sum_{i=1}^{n} \beta(X_i, \alpha)$, where following Johnstone and Silverman (2004) we set, for $0 \leq \alpha \leq 1$ and any real x,

$$\beta(x, \alpha) = \frac{\beta(x)}{1 + \alpha\beta(x)}, \qquad \beta(x) = \frac{g}{\phi}(x) - 1.$$

The study of $\hat{\alpha}$ below uses in a crucial way the first two moments of $\beta(X_i, \alpha)$, so we introduce the corresponding notation next. Let E_τ, for $\tau \in \mathbb{R}^n$, denote the expectation under $\theta_0 = \tau$. Define

$$\tilde{m}(\alpha) = -E_0\beta(X, \alpha) \tag{2.3.2}$$

and further denote

$$m_1(\tau, \alpha) = E_\tau[\beta(X, \alpha)] = \int_{-\infty}^{\infty} \beta(t, \alpha)\phi(t - \tau)dt.$$
$$m_2(\tau, \alpha) = E_\tau[\beta(X, \alpha)^2].$$

The thresholds $\zeta(\alpha)$, $\tilde{\tau}(\alpha)$ and $t(\alpha)$. Following Johnstone and Silverman (2004), we introduce several useful thresholds. From Lemma 1 in Johnstone and Silverman (2004), we know that g/ϕ, and therefore $\beta = g/\phi - 1$, is a strictly increasing function on $\mathbb{R}^+$. It is also continuous, so given α, a pseudo-threshold $\zeta = \zeta(\alpha)$ can be defined by

$$\beta(\zeta) = \frac{1}{\alpha}. \tag{2.3.3}$$

Further one can also define $\tau(\alpha)$ as the solution in x of

$$\Omega(x, \alpha) := \frac{a(x)}{1 - a(x)} = \frac{\alpha}{1 - \alpha}\frac{g}{\phi}(x) = 1.$$

Equivalently, $a(\tau(\alpha)) = 1/2$. Also, $\beta(\tau(\alpha)) = \alpha^{-1} - 2$ so $\tau(\alpha) \leq \zeta(\alpha)$. Define α_0 as $\tau(\alpha_0) = 1$ and set

$$\tilde{\tau}(\alpha) = \tau(\alpha \wedge \alpha_0). \tag{2.3.4}$$

Recall from Section 2.2 that $t(\alpha)$ is the threshold associated to the posterior median for given α. It is shown in Johnstone and Silverman (2004), Lemma 3, that $t(\alpha) \leq \zeta(\alpha)$. Finally, the following bound in terms of $\tau(\alpha)$, see Johnstone and Silverman (2004) p. 1623, is also useful for large x,

$$1 - a(x) \leq 1\!\!1_{|x| \leq \tilde{\tau}(\alpha)} + e^{-\frac{1}{2}(|x| - \tilde{\tau}(\alpha))^2} 1\!\!1_{|x| > \tilde{\tau}(\alpha)}. \tag{2.3.5}$$

2.3.2 Posterior risk bounds

Recall the notation $r_2(\alpha, \mu, x) = \int (u-\mu)^2 d\Pi_\alpha(u)$.

Lemma 1. Let γ be the Cauchy or Laplace density. For any x and $\alpha \in [0, 1/2]$,

$$r_2(\alpha, 0, x) \le C\Big[1 \wedge \frac{\alpha}{1-\alpha}\frac{g}{\phi}(x)\Big](1+x^2)$$
$$r_2(\alpha, \mu, x) \le (1-a(x))\mu^2 + Ca(x)((x-\mu)^2+1).$$

Let γ be the Cauchy density. For any real x and $\alpha \in [0, 1/2]$,

$$E_0 r_2(\alpha, 0, x) \le C\tau(\alpha)\alpha$$
$$E_\mu r_2(\alpha, \mu, x) \le C(1+\tilde{\tau}(\alpha)^2).$$

The following lower bound is used in the proof of Theorem 13.

Lemma 2. Let γ be the Laplace density. There exists $C_0 > 0$ such that, for $x \in \mathbb{R}$ and $\alpha \in [0, 1]$

$$r_2(\alpha, 0, x) \ge C_0\alpha.$$

We now turn to bounding $r_2(\hat{\alpha}, \mu, x)$. This is the quantity $r_2(\alpha, \mu, x)$, where α (which comes in via $a(x) = a_\alpha(x)$) is replaced by $\hat{\alpha}$. This is done with the help of the threshold $\tilde{\tau}(\alpha)$.

Lemma 3 (no signal or small signal). Let γ be the Cauchy density. Let α be a fixed non-random element of $(0, 1)$. Let $\hat{\alpha}$ be a random element of $[0, 1]$ that may depend on $x \sim \mathcal{N}(0, 1)$ and on other data. Then there exists $C_1 > 0$ such that

$$Er_2(\hat{\alpha}, 0, x) \le C_1\left[\alpha\tilde{\tau}(\alpha) + P(\hat{\alpha} > \alpha)^{1/2}\right].$$

There exists $C_2 > 0$ such that for any real μ, if $x \sim \mathcal{N}(\mu, 1)$,

$$Er_2(\hat{\alpha}, \mu, x) \le \mu^2 + C_2.$$

Lemma 4 (signal). Let γ be the Cauchy density. Let α be a fixed non-random element of $(0, 1)$. Let $\hat{\alpha}$ be a random element of $[0, 1]$ that may depend on $x \sim \mathcal{N}(\mu, 1)$ and on other data and such that $\tilde{\tau}(\hat{\alpha})^2 \le d\log(n)$ with probability 1 for some $d > 0$. Then there exists $C_2 > 0$ such that for all real μ,

$$Er_2(\hat{\alpha}, \mu, x) \le C_2\left[1 + \tilde{\tau}(\alpha)^2 + (1 + d\log n)P(\hat{\alpha} < \alpha)^{1/2}\right].$$

2.3.3 Moments of the score function

The next three lemmas are borrowed from Johnstone and Silverman (2004) and apply to any density γ such that $\log \gamma$ is Lipschitz on $\mathbb{R}$ and satisfies

$$\gamma(y)^{-1} \int_y^\infty \gamma(u)du \approx y^{\kappa-1}, \quad \text{as } y \to \infty. \tag{2.3.6}$$

Both Cauchy and Laplace densities satisfy (2.3.6), with $\kappa = 2$ and $\kappa = 1$ respectively, and their logarithm is Lipschitz.

Lemma 5. For $\kappa \in [1, 2]$ as in (2.3.6), as $\alpha \to 0$,

$$\tilde{m}(\alpha) \asymp \zeta^{\kappa-1} g(\zeta).$$

Also, the function $\alpha \to \tilde{m}(\alpha)$ is nonnegative and increasing in α.

Lemma 6. The function $\alpha \to m_1(\mu, \alpha)$ is decreasing in α. Also, $m_1(\zeta, \alpha) \sim 1/(2\alpha)$ as $\alpha \to 0$. For small enough α,

$$m_2(\mu, \alpha) \le C\alpha^{-1} m_1(\mu, \alpha), \quad \mu \ge 1.$$

Lemma 7. There exist a constant c_1 such that for any x and α,

$$|\beta(x, \alpha)| \le \frac{1}{\alpha \wedge c_1},$$

and constants c_2, c_3, c_4 such that for any α, and κ as in (2.3.6),

$$\begin{aligned} m_1(\mu, \alpha) &\le -\tilde{m}(\alpha) + c_2\zeta(\alpha)\mu^2, && |\mu| \le 1/\zeta(\alpha) \\ m_1(\mu, \alpha) &\le (\alpha \wedge c_3)^{-1} && \text{for all } \mu \end{aligned}$$

and

$$\begin{aligned} m_2(\mu, \alpha) &\le c_4 \frac{\tilde{m}(\alpha)}{\zeta(\alpha)^\kappa \alpha} && |\mu| \le 1/\zeta = 1/\zeta(\alpha) \\ m_2(\mu, \alpha) &\le (\alpha \wedge c_3)^{-2} && \text{for all } \mu. \end{aligned}$$

2.3.4 In-probability bounds for $\hat{\alpha}$

Lemma 9 below implies that, for any possible θ_0, the estimate $\hat{\alpha}$ is smaller than a certain α_1 with high probability. One can interpret this as saying that $\hat{\alpha}$ does not lead to too

much undersmoothing (i.e. too many nonzero coefficients). On the other hand, if there is enough signal in a certain sense, $\hat{\alpha}$ does not lead to too much oversmoothing (i.e. too many zero coefficients), see Lemma 10.

Although we generally follow the approach of Johnstone and Silverman (2004), there is one significant difference. One needs a fairly sharp bound on α_1 below. Using the definition from Johnstone and Silverman (2004) would lead to a loss in terms of logarithmic factors for the posterior L^2–squared moment. So we work with a somewhat different α_1, and shall thus provide a detailed proof of the corresponding Lemma 9. For the oversmoothing case, one can borrow the corresponding Lemma of Johnstone and Silverman (2004) as is.

Let $\alpha_1 = \alpha_1(d)$ be defined as the solution of the equation, with $\eta_n = s_n/n$,

$$d\alpha_1 \tilde{m}(\alpha_1) = \eta_n, \tag{2.3.7}$$

where d is a constant to be chosen (small enough for Lemma 9 to hold). A solution of (2.3.7) exists, as using Lemma 5, $\alpha \to \alpha\tilde{m}(\alpha)$ is increasing in α, and equals 0 at 0. Also, provided η_n is small enough, α_1 can be made smaller than any given arbitrary constant. The corresponding threshold ζ_1 is defined by $\beta(\zeta_1) = \alpha_1^{-1}$. From Lemma 5, we have $\tilde{m}(\alpha_1) \asymp \zeta g(\zeta_1)$ if γ is Cauchy and $\tilde{m}(\alpha_1) \asymp g(\zeta_1)$ if γ is Laplace.

Lemma 8. Let κ be the constant in (2.3.6). Let α_1 be defined by (2.3.7) for d a given constant and let ζ_1 be given by $\beta(\zeta_1) = \alpha_1^{-1}$. Then there exist real constants c_1, c_2 such that for large enough n,

$$\log(n/s_n) + c_1 \leq \frac{\zeta_1^2}{2} \leq \log(n/s_n) + \frac{\kappa - 1}{2} \log\log n + c_2,$$

with κ as in (2.3.6). Also, $\zeta_1^2 \sim 2\log(n/s_n)$ as n/s_n goes to ∞.

Lemma 9. Let α_1 be defined by (2.3.7) for d a given small enough constant and let ζ_1 be given by $\beta(\zeta_1) = \alpha_1^{-1}$. Suppose (2.2.7) holds. Then for some constant $C > 0$,

$$\sup_{\theta \in \ell_0[s_n]} P_\theta[\hat{\zeta} < \zeta_1] \leq \exp(-Cs_n).$$

For the oversmoothing case, one denotes the proportion of signals above a level τ by

$$\tilde{\pi}(\tau; \mu) = \frac{1}{n} \#\{i : |\mu_i| \geq \tau\}. \tag{2.3.8}$$

We also set, recalling that α_0 is defined via $\tau(\alpha_0) = 1$,

$$\alpha(\tau, \pi) = \sup\{\alpha \leq \alpha_0 : \ \pi m_1(\tau, \alpha) \geq 2\tilde{m}(\alpha)\}. \tag{2.3.9}$$

One defines $\zeta_{\tau,\pi}$ as the corresponding pseudo-threshold $\beta^{-1}(\alpha(\tau,\pi)^{-1})$.

Lemma 10 (Johnstone and Silverman (2004), Lemma 11). There exists C and π_0 such that if $\pi < \pi_0$, then for all $\tau \geq 1$,

$$\sup_{\theta:\ \tilde{\pi}(\tau;\theta)\geq\pi} P_\theta[\hat{\zeta} > \zeta_{\tau,\pi}] \leq \exp\{-Cn\phi(\zeta_{\tau,\pi})\}.$$

2.3.5 Proof of Theorem 13

Proof. Let α^* be defined as the solution in α of the equation,

$$\alpha\tilde{m}(\alpha) = \eta_n/4, \tag{2.3.10}$$

where $\eta_n = s_n/n$ (that is $\alpha^* = \alpha_1(d)$ with $d = 4$ in (2.3.7)). Let ζ^* be defined via $\beta(\zeta^*) = \alpha^*$.

Let θ_0 be the specific signal defined by, for α^*, ζ^* as in (2.3.10),

$$\theta_{0,i} = \begin{cases} \zeta^*, & 1 \leq i \leq s_n \\ 0, & s_n < i \leq n \end{cases}.$$

Using Lemma 5, one gets $\tilde{m}(\alpha^*) \asymp g(\zeta^*) \asymp \gamma(\zeta^*)$ as $\zeta^* \to \infty$. Lemma 8 implies $\zeta^{*2} \geq 2\log(1/\eta_n) + C$, for C a possibly negative constant. Combining this with the definition $\gamma(\zeta^*) = e^{-\zeta^*}/2$ leads to

$$\alpha^* \gtrsim \eta_n e^{\sqrt{\log(1/\eta_n)}}, \tag{2.3.11}$$

for c_0 in (2.2.7) small enough to have $2\log(1/\eta_n) + C \geq \log(1/\eta_n)$. We next prove that, for $\hat{\alpha}$ given by (2.2.6), for small enough $c > 0$,

$$P_{\theta_0}[\hat{\alpha} < \alpha^*] \leq e^{-cs_n}. \tag{2.3.12}$$

If $\alpha^* \leq \alpha_n$ the probability at stake is 0, as $\hat{\alpha}$ belongs to $[\alpha_n, 1]$ by definition. For $\alpha^* > \alpha_n$, we have $\{\hat{\alpha} < \alpha^*\} = \{S(\alpha^*) < 0\}$. With $A = \sum_{i=1}^n m_1(\mu_i, \alpha^*)$,

$$P_{\theta_0}[\hat{\alpha} < \alpha^*] = P_{\theta_0}[S(\alpha^*) < 0] = P_{\theta_0}\left[\sum_{i=1}^n \beta(\theta_{0,i} + Z_i, \alpha^*) - m_1(\theta_{0,i}, \alpha^*) < -A\right]$$

Setting $W_i = m_1(\theta_{0,i}, \alpha^*) - \beta(\theta_{0,i} + Z_i, \alpha^*)$, we have $|W_i| \leq 2C/\alpha^* =: M$ and W_i are independent. So by Bernstein's inequality,

$$P_{\theta_0}\left[\sum_{i=1}^n W_i > A\right] \leq \exp\left[-\frac{1}{2}\frac{A^2}{V + \frac{1}{3}MA}\right],$$

where V is an upper-bound for $\sum_{i=1}^n \mathrm{Var}(W_i)$. The term A equals

$$A = (n - s_n)(-\tilde{m}(\alpha^*)) + s_n m_1(\zeta^*, \alpha^*).$$

The function $\alpha \to \alpha\tilde{m}(\alpha)$ is increasing, as $\tilde{m}(\cdot)$ is (Lemma 5), so by its definition (2.3.10), α^* can be made smaller than any given positive constant, provided c_0 in (2.2.7) is small enough, ensuring $\eta_n = s_n/n$ is small enough. Using Lemma 6, $m_1(\zeta, \alpha) \sim 1/(2\alpha)$ as $\alpha \to 0$. So, using (2.3.10), one obtains, for small enough c_0,

$$A \geq \frac{s_n}{3\alpha^*} - \frac{s_n}{4\alpha^*} = \frac{s_n}{12\alpha^*}.$$

On the other hand, the last part of Lemma 7 implies

$$\begin{aligned} V &\leq \sum_{i \notin S_0} m_2(0, \alpha^*) + \sum_{i \in S_0} m_2(\zeta^*, \alpha^*) \\ &\leq C(n - s_n)\frac{\tilde{m}(\alpha^*)}{\zeta^*\alpha^*} + C\frac{s_n}{\alpha^{*2}}. \end{aligned}$$

Using the definition of α^*, one deduces $V \lesssim s_n/\alpha^{*2}$ and from this

$$\frac{V}{A^2} + \frac{MA}{3A^2} \lesssim \frac{1}{s_n},$$

which in turn implies (2.3.12), as then $P_{\theta_0}\left[\sum_{i=1}^n W_i > A\right] \leq \exp[-cs_n]$. Next one writes

$$\begin{aligned} \int \|\theta - \theta_0\|^2 d\Pi_{\hat{\alpha}}[\theta \,|\, X] &\geq \int \|\theta - \theta_0\|^2 d\Pi_{\hat{\alpha}}[\theta \,|\, X] \mathbb{1}_{\hat{\alpha} \geq \alpha^*} \\ &\geq \sum_{i \notin S_0} \int \theta_i^2 d\Pi_{\hat{\alpha}}(\theta \,|\, X) \mathbb{1}_{\hat{\alpha} \geq \alpha^*} \end{aligned}$$

Lemma 2 implies, for any possibly data-dependent weight α, that $\int \theta_i^2 d\Pi_\alpha(\theta \,|\, X) \gtrsim \alpha$, so

$$\int \|\theta - \theta_0\|^2 d\Pi_{\hat{\alpha}}[\theta \,|\, X] \geq (n - s_n)\hat{\alpha}\mathbb{1}_{\hat{\alpha} \geq \alpha^*} \geq (n - s_n)\alpha^*\mathbb{1}_{\hat{\alpha} \geq \alpha^*}.$$

As $(n - s_n)\alpha^* P_{\theta_0}[\hat{\alpha} \geq \alpha^*] \gtrsim Cn\alpha^*(1 - e^{-cs_n})$, an application of (2.3.11) concludes the proof. □

2.3.6 Proof of Theorem 15

Let us decompose the risk $R_n(\theta_0) = E_{\theta_0} \int \|\theta - \theta_0\|^2 d\Pi_{\hat{\alpha}}(\theta \,|\, X)$ according to whether coordinates of θ correspond to a 'small' or 'large' signal, the threshold being $\zeta_1 = \beta^{-1}(\alpha_1^{-1})$, with α_1 defined in (2.3.7). One can write

$$R_n(\theta_0) = \Big[\sum_{i:\ \theta_{0,i}=0} + \sum_{i:\ 0<|\theta_{0,i}|\leq\zeta_1} + \sum_{i:\ |\theta_{0,i}|>\zeta_1} \Big] E_{\theta_0} \int (\theta_i - \theta_{0,i})^2 d\Pi_{\hat{\alpha}}(\theta_i \,|\, X).$$

We next use the first part of Lemma 3 with $\alpha = \alpha_1$ and the second part of the Lemma to obtain, for any θ_0 in $\ell_0[s_n]$,

$$\begin{aligned}
&\Big[\sum_{i:\ \theta_{0,i}=0} + \sum_{i:\ 0<|\theta_{0,i}|\leq\zeta_1} \Big] E_{\theta_0} \int (\theta_i - \theta_{0,i})^2 d\Pi_{\hat{\alpha}}(\theta_i \,|\, X) \\
&\leq C_1 \sum_{i:\ \theta_{0,i}=0} [\alpha_1 \tau(\alpha_1) + P_{\theta_0}(\hat{\alpha} > \alpha_1)] + \sum_{i:\ 0<|\theta_{0,i}|\leq\zeta_1} (\theta_{0,i}^2 + C) \\
&\leq C_1 \Big[(n - s_n)\alpha_1 \tau(\alpha_1) + (n - s_n) e^{-c_1 \log^2 n} \Big] + (\zeta_1^2 + C) s_n,
\end{aligned}$$

where for the last inequality we use Lemma 9 and (2.2.7). From (2.3.7) one gets, with $\eta_n = s_n/n$,

$$n\alpha_1 \lesssim n\eta_n \zeta_1^{-1} g(\zeta_1)^{-1} \lesssim s_n \zeta_1.$$

Now using Lemma 8 and the fact that $\tau(\alpha_1) \leq \zeta_1$, one obtains that the contribution to the risk of the indices i with $|\theta_{0,i}| \leq \zeta_1$ is bounded by a constant times $s_n \log(n/s_n)$.

It remains to bound the part of the risk for indexes i with $|\theta_{0,i}| > \zeta_1$. To do so, one uses Lemma 4 with α chosen as $\alpha = \alpha_2 := \alpha(\zeta_1, \pi_1)$ and $\pi_1 = \tilde{\pi}(\zeta_1; \theta_0)$, following the definitions (2.3.8)–(2.3.9). One denotes by ζ_2 the pseudo-threshold associated to α_2. The following estimates are useful below

$$\zeta_1^2 < \zeta_2^2 \tag{2.3.13}$$

$$\pi_1 \zeta_2^2 \leq C\eta_n \log(1/\eta_n). \tag{2.3.14}$$

These are established in a similar way as in Johnstone and Silverman (2004), but with the updated definition of α_1, ζ_1 from (2.3.7), so we include the proof below for completeness.

One can now apply Lemma 4 with $\alpha = \alpha_2$,

$$\begin{aligned}
&\sum_{i:\ |\theta_{0,i}|>\zeta_1} E_{\theta_0} \int (\theta_i - \theta_{0,i})^2 d\Pi_{\hat\alpha}(\theta_i \,|\, X) \\
&\leq C_2 n\pi_1 \left[1 + \zeta_2^2 + (1 + d\log n) P_{\theta_0}(\hat\alpha < \alpha_2)^{1/2}\right] \\
&\leq C_2 n\pi_1 \left[1 + \zeta_2^2 + (1 + d\log n) P_{\theta_0}(\hat\zeta > \zeta_2)^{1/2}\right].
\end{aligned}$$

Let us verify that the term in brackets in the last display is bounded above by $C(1+\zeta_2^2)$. If $\zeta_2 > \log n$, this is immediate by bounding $P_{\theta_0}(\hat\zeta > \zeta_2)$ by 1. If $\zeta_2 \leq \log n$, Lemma 10 implies $P_{\theta_0}(\hat\zeta > \zeta_2) \leq \exp(-Cn\phi(\zeta_2)) \leq \exp(-C\sqrt{n})$, so this is also the case. Conclude that the last display is bounded above by $Cn\pi_1(1+\zeta_2^2) \leq C' n\pi_1 \zeta_2^2$. Using (2.3.14), this term is itself bounded by $Cs_n \log(n/s_n)$, which concludes the proof of the Theorem, given (2.3.13)–(2.3.14).

We now check that (2.3.13)–(2.3.14) hold. We first compare α_1 and α_2. For small enough α, the bound on m_1 from Lemma 7 becomes $1/\alpha$, so that, using the definition (2.3.7) of α_1,

$$\frac{m_1(\zeta_1, \alpha_1)}{\tilde m(\alpha_1)} \leq \frac{1}{\alpha_1} \left(\frac{\eta_n}{d\alpha_1}\right)^{-1} \leq \frac{d}{\eta_n} \leq \frac{d}{\pi_1},$$

using the rough bound $\pi_1 \leq \eta_n$. Note that both functions $\tilde m(\cdot)^{-1}$ and $m_1(\zeta_1, \cdot)$ are decreasing via Lemmas 5–6, and so is their product on the interval where both functions are positive. As $d < 2$, by definition of α_2 this means $\alpha_2 < \alpha_1$ that is $\zeta_1 < \zeta_2$.

To prove (2.3.14), one compares ζ_2 first to a certain $\zeta_3 = \zeta(\alpha_3)$ defined by α_3 (largest) solution of

$$\bar\Phi(\zeta(\alpha_3) - \zeta_1) = \frac{8}{\pi_1} \alpha_3 \tilde m(\alpha_3),$$

with $\bar\Phi(x) = P[\mathcal{N}(0,1) > x]$. Using Lemma 11, which also gives the existence of ζ_3, one gets

$$\frac{m_1(\zeta_1, \alpha_3)}{\tilde m(\alpha_3)} \geq \frac{\frac{1}{4}\beta(\zeta_3)\bar\Phi(\zeta_3 - \zeta_1)}{\tilde m(\alpha_3)} = \frac{1}{4\alpha_3} \frac{8\alpha_3 \tilde m(\alpha_3)}{\pi_1 \tilde m(\alpha_3)} = \frac{2}{\pi_1}.$$

This shows, reasoning as above, that $\alpha_3 \leq \alpha_2$, that is $\zeta_2 \leq \zeta_3$. Following Johnstone and Silverman (2004), one distinguishes two cases to further bound ζ_3.

If $\zeta_3 > \zeta_1 + 1$, using $\zeta_2^2 \leq \zeta_3^2$ and $\tilde{m}(\alpha_3) \lesssim \zeta_3 g(\zeta_3)$,

$$\begin{aligned}\pi_1 \zeta_2^2 &\leq \zeta_3^2 \frac{8\alpha_3 \tilde{m}(\alpha_3)}{\bar{\Phi}(\zeta_3 - \zeta_1)} \lesssim \zeta_3^3 \frac{g(\zeta_3)}{\beta(\zeta_3)} \frac{\zeta_3 - \zeta_1}{\phi(\zeta_3 - \zeta_1)} \\ &\leq C\zeta_3^4 \frac{\phi(\zeta_3)}{\phi(\zeta_3 - \zeta_1)} = C\zeta_3^4 \phi(\zeta_1) e^{-(\zeta_3 - \zeta_1)\zeta_1} \\ &\leq C(\zeta_1 + 1)^4 e^{-\zeta_1} \phi(\zeta_1),\end{aligned}$$

where for the last inequality we have used that $x \to x^4 e^{-(x-\zeta_1)\zeta_1}$ is decreasing for $x \geq \zeta_1 + 1$. Lemma 8 now implies that $\phi(\zeta_1) \lesssim \eta_n$. As ζ_1 goes to ∞ with n/s_n, one gets $\pi_1 \zeta_2^2 \lesssim \eta_n$.

If $\zeta_1 \leq \zeta_3 \leq \zeta_1 + 1$, let $\zeta_4 = \zeta(\alpha_4)$ with α_4 solution in α of

$$\bar{\Phi}(1) = 8\alpha \tilde{m}(\alpha) \pi_1^{-1}.$$

By the definition of ζ_3, since $\bar{\Phi}(1) \leq \bar{\Phi}(\zeta_3 - \zeta_1)$, we have $8\alpha_4 \tilde{m}(\alpha_4) \leq 8\alpha_3 \tilde{m}(\alpha_3)$ so that $\alpha_4 \leq \alpha_3$. Using Lemma 5 as before,

$$\bar{\Phi}(1) \lesssim \frac{g(\zeta_4)}{\beta(\zeta_4)} \pi_1^{-1} \lesssim \phi(\zeta_4) \pi_1^{-1}.$$

Taking logarithms this leads to

$$\zeta_4^2 \leq C + 2\log(\pi_1^{-1}).$$

In particular, $\zeta_2^2 \leq 2\log(\pi_1^{-1}) + C$. As $x \to x\log(1/x)$ is increasing, one gets, using $\pi_1 \leq \eta_n$,

$$\pi_1 \zeta_2^2 \leq 2\eta_n \log(1/\eta_n) + C\eta_n,$$

which concludes the verification of (2.3.13)–(2.3.14) and the proof of Theorem 15.

In checking (2.3.14), one needs a lower bound on m_1. In Johnstone and Silverman (2004), the authors mention that it follows from their lower bound (82), Lemma 8. But this bound cannot hold uniformly for any smoothing parameter α (denoted by w in Johnstone and Silverman (2004)), as $m_1(0, w) = -\tilde{m}(w) < 0$ if $w \neq 0$. So, although the claimed inequality is correct, it does not seem to follow from (82). We state the inequality we use now, and prove it in Section 2.4.3.

Lemma 11. Let $\bar{\Phi}(t) = \int_t^\infty \phi(u)du$. For π_1, ζ_1 as above, a solution $0 < \alpha \leq \alpha_1$ to the equation

$$\bar{\Phi}(\zeta(\alpha) - \zeta_1) = 8\pi_1^{-1} \alpha \tilde{m}(\alpha). \tag{2.3.15}$$

exists. Let α_3 be the largest such solution. Then for c_0 in (2.2.7) small enough,

$$m_1(\zeta_1, \alpha_3) \geq \frac{1}{4}\beta(\zeta_3)\bar{\Phi}(\alpha_3 - \zeta_1). \tag{2.3.16}$$

2.3.7 Proof of Theorem 14

Let $\theta_0, \alpha^*, \zeta^*$ be defined as in the proof of Theorem 13. Below we show that the event $\mathcal{A} = \{\hat{\alpha} \in [\alpha^*, c\alpha^*]\}$, for c a large enough constant, has probability going to 1, faster than a polynomial in $1/n$. Recall from the proof of Theorem 13 that, if $\hat{\alpha} \geq \alpha^*$, so in particular on $\mathcal{A}$, we have $V_X \geq (n - s_n)\alpha^* \geq n\alpha^*/2 \geq C_1 s_n g(\zeta^*)^{-1}$. Denote

$$\begin{aligned} v_n &= m s_n g(\zeta^*)^{-1} \\ V_X &= \int \|\theta - \theta_0\|^2 d\Pi_{\hat{\alpha}}(\theta \,|\, X), \end{aligned}$$

where m is chosen small enough so that $v_n \leq V_X/2$ on $\mathcal{A}$. Then,

$$\begin{aligned} &\Pi_{\hat{\alpha}}\left[\|\theta - \theta_0\|^2 < v_n \,|\, X\right] \mathbb{1}_{\mathcal{A}} = \Pi_{\hat{\alpha}}\left[\|\theta - \theta_0\|^2 - V_X < v_n - V_X \,|\, X\right] \mathbb{1}_{\mathcal{A}} \\ &\leq \Pi_{\hat{\alpha}}\left[\|\theta - \theta_0\|^2 - V_X < -V_X/2 \,|\, X\right] \leq 4V_X^{-2} \int \{\|\theta - \theta_0\|^2 - V_X\}^2 d\Pi_{\hat{\alpha}}(\theta \,|\, X), \end{aligned}$$

where the second line follows from Markov's inequality. One now writes the L^2–norm in the previous display as sum over coordinates and one expands the square, while noting that given X the posterior $\Pi_{\hat{\alpha}}[\cdot \,|\, X]$ makes the coordinates of θ independent

$$\begin{aligned} &\int \{\|\theta - \theta_0\|^2 - V_X\}^2 d\Pi_{\hat{\alpha}}(\theta \,|\, X) \\ &= \int \sum_{i,j} \left[(\theta_i - \theta_{0,i})^2 - \int (\theta_i - \theta_{0,i})^2 d\Pi_{\hat{\alpha}}(\theta \,|\, X)\right] \\ &\qquad\qquad \times \left[(\theta_j - \theta_{0,j})^2 - \int (\theta_j - \theta_{0,j})^2 d\Pi_{\hat{\alpha}}(\theta \,|\, X)\right] d\Pi_{\hat{\alpha}}(\theta \,|\, X) \\ &= \sum_{i=1}^n \int \left[(\theta_i - \theta_{0,i})^2 - \int (\theta_i - \theta_{0,i})^2 d\Pi_{\hat{\alpha}}(\theta \,|\, X)\right]^2 d\Pi_{\hat{\alpha}}(\theta \,|\, X) \\ &\leq \sum_{i=1}^n \int (\theta_i - \theta_{0,i})^4 d\Pi_{\hat{\alpha}}(\theta \,|\, X). \end{aligned}$$

The last bound is the same as in the proof of the upper bound Theorem 15, except the fourth moment replaces the second moment. Denote $r_4(\alpha, \mu, x) = \int (u - \mu)^4 d\pi_\alpha(u \,|\, x)$,

then

$$r_4(\alpha, \mu, x) = (1 - a(x))\mu^4 + a(x) \int (u - \mu)^4 \gamma_x(u) du.$$

In a similar way as in the proof of Lemma 1, one obtains $\int (u-\mu)^4 \gamma_x(u) du \leq C(1+(x-\mu)^4)$. Next, noting that since now γ is Laplace so g has Laplace tails, $x \to (1 + x^4)g(x)$ is integrable, proceeding as in the proof of Lemma 1, one gets $E_0 r_4(\alpha, 0, x) \lesssim \alpha$ as well as $E_\mu r_4(\alpha, \mu, x) \lesssim 1 + \tilde{\tau}(\alpha)^4$, for any fixed α. Similarly as in Lemmas 3–4, one then derives the following random α bounds

$$E r_4(\hat{\alpha}, 0, x) \lesssim c\alpha^* + P(\hat{\alpha} > c\alpha^*)^{1/2}$$

and, for any μ,

$$E r_4(\hat{\alpha}, \mu, x) \lesssim 1 + \tau(\alpha^*)^4 + (1 + \log^2 n) P(\hat{\alpha} < \alpha^*)^{1/2}.$$

By using that the probabilities in the last displays go to 0 faster than $1/n$, which we show below, and gathering the bounds for all i,

$$E_{\theta_0} \sum_{i=1}^{n} \int (\theta_i - \theta_{0,i})^4 d\Pi_{\hat{\alpha}}(\theta \,|\, X) \lesssim s_n(1 + \tau(\alpha^*)^4) + n\alpha^*.$$

From this deduce that

$$\begin{aligned} E_{\theta_0} \Pi_{\hat{\alpha}} \left[\|\theta - \theta_0\|^2 < v_n \,|\, X \right] &\lesssim P[\mathcal{A}^c] + [s_n(1 + \tau(\alpha^*)^4) + n\alpha^*]/(s_n g(\zeta^*)^{-1})^2 \\ &\lesssim P[\mathcal{A}^c] + s_n^{-1}(1 + \tau(\alpha^*)^4) g(\zeta^*) + s_n^{-1} g(\zeta^*). \end{aligned}$$

The last bound goes to 0, as $\tau(\alpha^*) \leq \zeta_{\alpha^*} = \zeta^*$ and g has Laplace tails. To conclude the proof, we show that $P_{\theta_0}(\hat{\alpha} \in [\alpha^*, c\alpha^*])$ is small. From the proof of Theorem 13, one already has $P_{\theta_0}[\hat{\alpha} < \alpha^*] \leq \exp(-cs_n)$, which is a $o(1/n)$ using $s_n \gtrsim \log^2 n$. To obtain a bound on $P_{\theta_0}[\hat{\alpha} > c\alpha^*]$, one can now revert the inequalities in the reasoning leading to the Bernstein bound in the proof of Theorem 13. With $A = \sum_{i=1}^{n} m_1(\mu_i, \alpha)$, we have

$$\begin{aligned} P_{\theta_0}[\hat{\alpha} > c\alpha^*] &= P_{\theta_0}[S(c\alpha^*) > 0] \\ &= P_{\theta_0}\left[\sum_{i=1}^{n} \beta(\theta_{0,i} + Z_i, c\alpha^*) - m_1(\theta_{0,i}, c\alpha^*) > -A \right]. \end{aligned}$$

But here, $-A = (n - s_n)\tilde{m}(c\alpha^*) - s_n m_1(\zeta^*, c\alpha^*)$. As $\alpha \to \tilde{m}(\alpha)$ is increasing, $\tilde{m}(c\alpha^*) \geq \tilde{m}(\alpha^*)$. Now by Lemma 7,

$$m_1(\zeta^*, c\alpha^*) \leq (c\alpha^* \wedge c_3)^{-1} \leq \frac{1}{c\alpha^*},$$

provided $\alpha^* \leq c_3/c = c_3/16$, which is the case for η_n small enough. Since by definition $n\tilde{m}(\alpha^*) = s_n/(4\alpha^*)$, we have $-A \geq s_n/(8\alpha^*)$. From there one can carry over the same scheme of proof as for the previous Bernstein inequality, with now $\tilde{A} = -A$ and $\tilde{V}$ the variance proxy which is bounded by

$$\tilde{V} \leq (n - s_n)m_2(0, c\alpha^*) + s_n m_2(\zeta^*, c\alpha^*) \lesssim n\frac{\tilde{m}(c\alpha^*)}{\zeta_{c\alpha^*} c\alpha^*} + \frac{s_n}{(c\alpha^*)^2}.$$

Now $\tilde{m}(c\alpha^*) \lesssim Cg(\zeta_{c\alpha^*})$. Using bounds similar to those of Lemma 8, one can check that $C_1 + \zeta_{\alpha^*}^2 \leq \zeta_{c\alpha^*}^2 \leq C_2 + \zeta_{\alpha^*}^2$, which implies that $\tilde{m}(c\alpha^*)/\zeta_{c\alpha^*} \lesssim \tilde{m}(\alpha^*)/\zeta^* \lesssim \tilde{m}(\alpha^*)$. From this one deduces, with $\tilde{M} \leq C/s_n$,

$$\frac{\tilde{V}}{\tilde{A}^2} + \frac{\tilde{M}\tilde{A}}{3\tilde{A}^2} \lesssim \frac{C'}{s_n},$$

which by Bernstein's inequality implies $P_{\theta_0}[\hat{\alpha} > c\alpha^*] \leq \exp[-Cs_n]$, which completes the proof of Theorem 14.

2.4 Technical lemmas for the SAS prior

2.4.1 Proofs of posterior risk bounds: fixed α

Proof of Lemma 1. First one proves the first two bounds. To do so, we derive moment bounds on γ_x. Since $\gamma_x(\cdot)$ is a density function, we have for any x, $\int \gamma_x(u)du = 1$. This implies $(\log g)'(x) = \int (u - x)\gamma_x(u)du = \int u\gamma_x(u)du - x$. In Johnstone and Silverman (2004), the authors check, see p. 1623, that $\int u\gamma_x(u)du =: \tilde{m}_1(x)$ is a shrinkage rule, that is $0 \leq \tilde{m}_1(x) \leq x$ for $x \geq 0$, so by symmetry, for any real x,

$$\left| \int u\gamma_x(u)du \right| \leq |x|.$$

Decomposing $u^2 = (u-x)^2 + 2x(u-x) + x^2$ and noting that $\int (u-x)^2 \gamma_x(u)du = g''(x)/g(x) + 1$,

$$\int u^2 \gamma_x(u)du = \frac{g''}{g}(x) + 1 + 2x\frac{g'}{g}(x) + x^2.$$

Note that for γ Laplace or Cauchy, we have $|\gamma'| \le c_1\gamma$ and $|\gamma''| \le c_2\gamma$. This leads to

$$|g'(x)| = |\int \gamma'(x-u)\phi(u)du| \le c_1 \int \gamma(x-u)\phi(u)du = c_1 g(x)$$

and similarly $|g''| \le c_2 g$, so that $\int u^2\gamma_x(u)du \le C(1+x^2)$ which gives the first bound using (2.3.1). We note, *en passant*, that the one but last display also implies for any real x that

$$\int u^2 \gamma_x(u)du \ge 1 - c_2 - 2c_1|x| + x^2, \tag{2.4.1}$$

which implies that $\int u^2\gamma_x(u)du$ goes to ∞ with x. Also, for any real μ,

$$\int (u-\mu)^2 \gamma_x(u)du = (x-\mu)^2 + \frac{g''}{g}(x) + 1 + 2(x-\mu)\frac{g'}{g}(x).$$

Now using again $g'/g \le c_1$ and $g''/g \le c_2$ leads to

$$\int (u-\mu)^2 \gamma_x(u)du \le C(1 + (x-\mu)^2).$$

By using the expression of $r_2(\alpha, \mu, x)$, this yields the second bound of the lemma.

We now turn to the bounds in expectation. For a zero signal $\mu = 0$, one notes that $x = \tau(\alpha)$ is the value at which both terms in the minimum in the first inequality of the lemma are equal. So

$$E_0 r_2(\alpha, 0, x) \lesssim \int 1\!\!1_{|x|\le\tau(\alpha)} \frac{\alpha}{1-\alpha}\frac{g}{\phi}(x)\phi(x)(1+x^2)dx + \int 1\!\!1_{|x|>\tau(\alpha)}(1+x^2)\phi(x)dx.$$

For γ Cauchy, g has Cauchy tails and $x \to (1+x^2)g(x)$ is bounded, so one gets, with $\alpha \le 1/2$,

$$\begin{aligned} E_0 r_2(\alpha, 0, x) &\lesssim \alpha \int 1\!\!1_{|x|\le\tau(\alpha)}dx + \tau(\alpha)\phi(\tau(\alpha)) + \phi(\tau(\alpha))/\tau(\alpha) \\ &\lesssim \tau(\alpha)\alpha + \tau(\alpha)\phi(\tau(\alpha)) \lesssim \tau(\alpha)\alpha + \tau(\alpha)\alpha g(\tau(\alpha)) \lesssim \tau(\alpha)\alpha. \end{aligned}$$

Turning to the last bound of the lemma, we distinguish two cases. Set for the remaining of the proof $T := \tilde{\tau}(\alpha)$ for simplicity of notation. The first case is $|\mu| \leq 4T$, for which

$$E_\mu r_2(\alpha, \mu, x) \leq \mu^2 + C \leq C_1(1 + T^2).$$

The second case is $|\mu| > 4T$. We bound the expectation of each term in the second bound of the lemma (that for $r_2(\alpha, \mu, x)$) separately. First, $E[a(x)(1 + (x - \mu)^2)] \leq C$. It thus suffices to bound $\mu^2 E_\mu[1 - a(x)]$. To do so, one uses the bound (2.3.5) and starts by noting that, if $Z \sim \mathcal{N}(0, 1)$,

$$E[\mathbb{1}_{|Z+\mu|\leq T}] \leq P[|Z| \geq |\mu| - T] \leq P[|Z| \geq |\mu|/2].$$

This implies, with $\bar{\Phi}(u) = \int_u^\infty \phi(t)dt \leq \phi(u)/u$ for $u > 0$,

$$E_\mu[\mu^2 \mathbb{1}_{|x|\leq T}] \leq C_2|\mu|\phi(|\mu|) \leq C_3.$$

If $A = \{x,\ |x - \mu| \leq |\mu|/2\}$ and A^c denotes its complement,

$$\sqrt{2\pi} E_\mu[e^{-\frac{1}{2}(|x|-T)^2}] \leq \int_{A^c} e^{-\frac{1}{2}(x-\mu)^2} dx + \int_A e^{-\frac{1}{2}(|x|-T)^2} dx.$$

The first term in the last sum is bounded above by $2\bar{\Phi}(|\mu|/2)$. The second term, as $A \subset \{x,\ |x| \geq |\mu|/2\}$, is bounded above by $2\bar{\Phi}(|\mu|/4)$. This implies, in the case $|\mu| > 4T$, that

$$E_\mu r_2(\alpha, \mu, x) \leq C_4 + 4\mu^2 \bar{\Phi}(|\mu|/4) + 5 \leq C.$$

The last bound of the lemma follows by combining the previous bounds in the two cases. □

Proof of Lemma 2. From the expression of $r_2(\alpha, 0, x)$ it follows

$$\begin{aligned} r_2(\alpha, 0, x) &\geq a(x) \inf_{x\in\mathbb{R}} \int u^2 \gamma_x(u)du \geq \alpha \frac{g}{\phi \vee g}(x) \inf_{x\in\mathbb{R}} \int u^2 \gamma_x(u)du \\ &\geq \alpha \inf_{x\in\mathbb{R}} \frac{g}{\phi \vee g}(x) \inf_{x\in\mathbb{R}} \int u^2 \gamma_x(u)du \geq C_0\alpha, \end{aligned}$$

where $c_0 > 0$. Indeed, both functions whose infimum is taken in the last display are continuous in x, are strictly positive for any real x, and have respective limits 1 and $+\infty$ as $|x| \to \infty$, using (2.4.1), so these functions are bounded below on $\mathbb{R}$ by positive constants. □

2.4.2 Proofs of posterior risk bounds: random α

Proof of Lemma 3. Using the bound on $r_2(\alpha, 0, x)$ from Lemma 1,

$$\begin{aligned} r_2(\hat{\alpha}, 0, x) &= r_2(\hat{\alpha}, 0, x)\mathbb{1}_{\hat{\alpha}\leq\alpha} + r_2(\hat{\alpha}, 0, x)\mathbb{1}_{\hat{\alpha}>\alpha} \\ &\leq \left[\frac{\hat{\alpha}}{1-\hat{\alpha}}\frac{g}{\phi}(x) \wedge 1\right](1+x^2)\mathbb{1}_{\hat{\alpha}\leq\alpha} + C(1+x^2)\mathbb{1}_{\hat{\alpha}>\alpha} \\ &\leq \left[\frac{\alpha}{1-\alpha}\frac{g}{\phi}(x) \wedge 1\right](1+x^2)\mathbb{1}_{\hat{\alpha}\leq\alpha} + C(1+x^2)\mathbb{1}_{\hat{\alpha}>\alpha}. \end{aligned}$$

For the first term in the last display, one bounds the indicator from above by 1 and proceeds as in the proof of Lemma 1 to bound its expectation by $C\alpha\tilde{\tau}(\alpha)$. The first part of the lemma follows by noting that $E[(1+x^2)\mathbb{1}_{\hat{\alpha}>\alpha}]$ is bounded from above by $(2+2E_0[x^4])^{1/2}P(\hat{\alpha} > \alpha)^{1/2} \leq C_1 P(\hat{\alpha} > \alpha)^{1/2}$ by Cauchy-Schwarz inequality. The second part of the lemma follows from the fact that using Lemma 1, $r_2(\alpha, \mu, x) \leq (1-a(x))\mu^2 + Ca(x)((x-\mu)^2+1) \leq \mu^2 + C(x-\mu)^2 + C$ for any α.

□

Proof of Lemma 4. Combining (2.3.5) and the third bound of Lemma 1,

$$r_2(\hat{\alpha}, \mu, x) \leq \mu^2\left[\mathbb{1}_{|x|\leq\tilde{\tau}(\hat{\alpha})} + e^{-\frac{1}{2}(|x|-\tilde{\tau}(\hat{\alpha}))^2}\mathbb{1}_{|x|>\tilde{\tau}(\hat{\alpha})}\right] + C((x-\mu)^2+1).$$

Note that it is enough to bound the first term on the right hand side in the last display, as the last one is bounded by a constant under E_μ. Let us distinguish the two cases $\hat{\alpha} \geq \alpha$ and $\hat{\alpha} < \alpha$.

In the case $\hat{\alpha} \geq \alpha$, as $\tilde{\tau}(\alpha)$ is a decreasing function of α,

$$\begin{aligned} &\left[\mathbb{1}_{|x|\leq\tilde{\tau}(\hat{\alpha})} + e^{-\frac{1}{2}(|x|-\tilde{\tau}(\hat{\alpha}))^2}\mathbb{1}_{|x|>\tilde{\tau}(\hat{\alpha})}\right]\mathbb{1}_{\hat{\alpha}\geq\alpha} \\ &\leq \left[\mathbb{1}_{|x|\leq\tilde{\tau}(\hat{\alpha})} + \mathbb{1}_{\tilde{\tau}(\hat{\alpha})<|x|\leq\tilde{\tau}(\alpha)} + e^{-\frac{1}{2}(|x|-\tilde{\tau}(\hat{\alpha}))^2}\mathbb{1}_{|x|>\tilde{\tau}(\alpha)}\right]\mathbb{1}_{\hat{\alpha}\geq\alpha} \\ &\leq \mathbb{1}_{|x|\leq\tilde{\tau}(\alpha)} + e^{-\frac{1}{2}(|x|-\tilde{\tau}(\alpha))^2}\mathbb{1}_{|x|>\tilde{\tau}(\alpha)}, \end{aligned}$$

where we have used $e^{-\frac{1}{2}v^2} \leq 1$ for any v and that $e^{-\frac{1}{2}(u-c)^2} \leq e^{-\frac{1}{2}(u-d)^2}$ if $u > d \geq c$. As a consequence, one can borrow the fixed α bound obtained previously so that

$$E\left[r_2(\hat{\alpha}, \mu, x)1_{\hat{\alpha}\geq\alpha}\right] \leq 2E_\mu r_2(\alpha, \mu, x) \leq C\left[1+\tilde{\tau}(\alpha)^2\right].$$

In the case $\hat{\alpha} < \alpha$, setting $b_n = \sqrt{d \log n}$ and noting that $\tilde{\tau}(\hat{\alpha}) \leq b_n$ with probability 1 by assumption, proceeding as above, with b_n now replacing $\tilde{\tau}(\alpha)$, one can bound

$$\begin{aligned} &\mathbb{1}_{|x| \leq \tilde{\tau}(\hat{\alpha})} + e^{-\frac{1}{2}(|x| - \tilde{\tau}(\hat{\alpha}))^2} \mathbb{1}_{|x| > \tilde{\tau}(\hat{\alpha})} \\ &\leq \mathbb{1}_{|x| \leq b_n} + e^{-\frac{1}{2}(|x| - b_n)^2} \mathbb{1}_{|x| > b_n}. \end{aligned}$$

From this one deduces that

$$\begin{aligned} &E\left(\mu^2 \left[\mathbb{1}_{|x| \leq \tilde{\tau}(\hat{\alpha})} + e^{-\frac{1}{2}(|x| - \tilde{\tau}(\hat{\alpha}))^2} \mathbb{1}_{|x| > \tilde{\tau}(\hat{\alpha})}\right] \mathbb{1}_{\hat{\alpha} < \alpha}\right) \\ &\leq C \left(E_\mu \left[\mu^4 \mathbb{1}_{|x| \leq b_n} + \mu^4 e^{-(|x| - b_n)^2}\right]\right)^{1/2} P(\hat{\alpha} < \alpha)^{1/2}. \end{aligned}$$

Using similar bounds as in the fixed α case, one obtains

$$E_\mu \left[\mu^4 \mathbb{1}_{|x| \leq b_n} + \mu^4 e^{-(|x| - b_n)^2}\right] \leq C(1 + b_n^4).$$

Taking the square root and gathering the different bounds obtained concludes the proof. □

2.4.3 Proofs on pseudo-thresholds

Proof of Lemma 8. For small α, or equivalently large ζ, we have $(g/\phi)(\zeta) = \beta(\zeta) + 1 \asymp \beta(\zeta)$. Deduce that for large n, using $\eta_n = d\alpha_1 \tilde{m}(\alpha_1)$ and Lemma 5 on $\tilde{m}$,

$$\eta_n \asymp \alpha_1 \zeta_1^{\kappa - 1} \frac{g(\zeta_1)}{\beta(\zeta_1)} \beta(\zeta_1) \asymp \zeta_1^{\kappa - 1} \phi(\zeta_1) \asymp \zeta_1^{\kappa - 1} e^{-\zeta_1^2/2}.$$

From this deduce that

$$\left| \log c + (\kappa - 1) \log \zeta_1 - \frac{\zeta_1^2}{2} + \log(1/\eta_n) \right| \leq C.$$

In particular, using $\log \zeta \leq a + \zeta^2/4$ for some constant $a > 0$ large enough, one gets $\zeta_1^2 \leq 4(C + \log(1/\eta_n)) \leq 4(C + \log n)$. Inserting this back into the previous inequality leads to

$$\zeta_1^2/2 \leq \log(1/\eta_n) + C + (1/2)(\kappa - 1) \log \log n.$$

The lower bound is obtained by bounding $(\kappa - 1) \log(\zeta_1) \geq 0$, for small enough α_1. □

Proof of Lemma 9. Using (2.2.7), $\log(1/\eta_n) \leq \log(n) - 2 \log \log n$, and the bound on ζ from Lemma 8 gives $\zeta_1^2 \leq 2 \log n - \frac{3}{2} \log \log n$, so that $t(\alpha_1) \leq \zeta(\alpha_1) = \zeta_1 \leq \sqrt{2 \log n} =$

$t(\alpha_n)$. It follows that α_1 belongs to the interval $[\alpha_n, 1]$ over which the likelihood is maximised.

Then one notices that $\{\hat{\zeta} < \zeta_1\} = \{\hat{\alpha} > \alpha_1\} = \{S(\alpha_1) > 0\}$, regardless of the fact that the maximiser $\hat{\alpha}$ is attained in the interior or at the boundary of $[\alpha_n, 1]$. So

$$P_\theta[\hat{\zeta} < \zeta_1] = P_\theta[S(\alpha_1) > 0].$$

The score function equals $S(\alpha) = \sum_{i=1}^n \beta(X_i, \alpha)$, a sum of independent variables. By Bernstein's inequality, if W_i are centered independent variables with $|W_i| \leq M$ and $\sum_{i=1}^n \mathrm{Var}(W_i) \leq V$, then for any $A > 0$,

$$P\left[\sum_{i=1}^n W_i > A\right] \leq \exp\{-\frac{1}{2}A^2/(V + \frac{1}{3}MA)\}.$$

Set $W_i = \beta(X_i, \alpha_1) - m_1(\theta_{0,i}, \alpha_1)$ and $A = -\sum_{i=1}^n m_1(\theta_{0,i}, \alpha_1)$. Then one can take $M = c_3/\alpha_1$, using Lemma 7. One can bound $-A$ from above as follows, using the definition of α_1,

$$\begin{aligned} -A &\leq -\sum_{i \notin S_0} \tilde{m}(\alpha_1) + \sum_{i \in S_0} \frac{c}{\alpha_1} \leq -(n - s_n)\tilde{m}(\alpha_1) + cs_n/\alpha_1 \\ &\leq -n\tilde{m}(\alpha_1)/2 + cdn\tilde{m}(\alpha_1) \leq -n\tilde{m}(\alpha_1)/4, \end{aligned}$$

provided d is chosen small enough and, using again the definition of α_1,

$$\begin{aligned} V &\leq \sum_{i \notin S_0} m_2(0, \alpha_1) + \sum_{i \in S_0} m_2(\theta_{0,i}, \alpha_1) \leq \frac{C}{\alpha_1}\left[(n - s_n)\tilde{m}(\alpha_1)\zeta_1^{-\kappa} + cs_n/\alpha_1\right] \\ &\leq C\alpha_1^{-1}\left[n\tilde{m}(\alpha_1)\zeta_1^{-\kappa}/2 + cdn\tilde{m}(\alpha_1)\right] \leq C'dn\tilde{m}(\alpha_1)/\alpha_1, \end{aligned}$$

where one uses that ζ_1^{-1} is bounded. This leads to

$$\frac{V + \frac{1}{3}MA}{A^2} \leq \frac{C'd}{n\alpha_1\tilde{m}(\alpha_1)} + \frac{4c_3}{3n\alpha_1\tilde{m}(\alpha_1)} \leq \frac{c_5^{-1}}{n\alpha_1\tilde{m}(\alpha_1)}.$$

One concludes that $P[\hat{\alpha} > \alpha_1] \leq \exp\{-c_5 n\alpha_1\tilde{m}(\alpha_1)\} = \exp\{-Cs_n\}$ using (2.3.7). □

Proof of Lemma 10. It is the same proof as Lemma 11 of Johnstone and Silverman (2004), but one has actually to be careful as one needs a positive lower bound on $m_1(1, \alpha)$ (which cannot be true for every μ) to prove that $m_2(\mu, \alpha) \leq Cm_1(\mu, \alpha)/\alpha$. For more details, we refer to (Castillo and Szabo, 2018)'s proof of Lemma 18 or Lemma 24 and Lemma 25 of Chapter 3. □

Proof of Lemma 11. First we check the existence of a solution. Set $\zeta_\alpha = \zeta(\alpha)$ and $R_\alpha := \bar{\Phi}(\zeta_\alpha - \zeta_1)/(\alpha\tilde{m}(\alpha))$. For $\alpha \to 0$ we have $\zeta_\alpha - \zeta_1 \to \infty$ so by using $\bar{\Phi}(u) \asymp \phi(u)/u$ as $u \to \infty$ one gets, treating terms depending on ζ_1 as constants and using $\phi(\zeta_\alpha) \asymp \alpha g(\zeta_\alpha)$,

$$\bar{\Phi}(\zeta_\alpha - \zeta_1) \asymp \frac{\phi(\zeta_\alpha - \zeta_1)}{\zeta_\alpha - \zeta_1} \asymp \alpha g(\zeta_\alpha) e^{\zeta_\alpha \zeta_1}.$$

As $\tilde{m}(\alpha) \asymp \zeta_\alpha g(\zeta_\alpha)$, one gets $R_\alpha \asymp e^{\zeta_\alpha\zeta_1}/\zeta_\alpha \to \infty$ as $\alpha \to 0$. On the other hand, with $\pi_1 \leq s_n/n$ and $\alpha_1\tilde{m}(\alpha_1) = ds_n/n$,

$$R_{\alpha_1} = \frac{1}{2\alpha_1\tilde{m}(\alpha_1)} = \frac{dn}{2s_n} \leq \frac{8}{\pi_1}\frac{d}{16},$$

so that $R_{\alpha_1} < 8/\pi_1$ as $d < 2$. This shows that the equation at stake has at least one solution for α in the interval $(0, \alpha_1)$.

By definition of $m_1(\mu, \alpha)$, for any μ and α, and $\zeta = \zeta(\alpha)$,

$$\begin{aligned} m_1(\mu,\alpha) &= \int_{-\zeta}^{\zeta} \frac{\beta(x)}{1+\alpha\beta(x)}\phi(x-\mu)dx \; + \int_{|x|>\zeta} \frac{\beta(x)}{1+\alpha\beta(x)}\phi(x-\mu)dx \\ &= \qquad (A) \qquad\qquad + \qquad\qquad (B). \end{aligned}$$

By definition of ζ, the denominator in (B) is bounded from above by $2\alpha\beta(x)$ so

$$(B) \geq \frac{1}{2\alpha}\int_{|x|>\zeta} \phi(x-\mu)dx \geq \frac{1}{2}\beta(\zeta)\bar{\Phi}(\zeta-\mu).$$

One splits the integral (A) in two parts corresponding to $\beta(x) \geq 0$ and $\beta(x) < 0$. Let c be the real number such that $g/\phi(c) = 1$. By construction the part of the integral (A) with $c \leq |x| \leq \zeta$ is nonnegative, so, for $\alpha \leq |\beta(0)|^{-1}/2$,

$$\begin{aligned} (A) &\geq \int_{-c}^{c} \frac{\beta(x)}{1+\alpha\beta(x)}\phi(x-\mu)dx \\ &\geq -\int_{-c}^{c} \frac{|\beta(0)|}{1-\alpha|\beta(0)|}\phi(x-\mu)dx \\ &\geq -2|\beta(0)|\int_{-c}^{c}\phi(x-\mu)dx, \end{aligned}$$

where one uses the monotonicity of $y \to y/(1+\alpha y)$. For $\mu \geq c$, the integral $\int_{-c}^{c}\phi(x-\mu)dx$ is bounded above by $2\int_0^c \phi(x-\mu)dx \leq 2c\phi(\mu-c)$. To establish (2.3.16), it thus suffices

to show that

$$(i) := 4|\beta(0)|c\phi(\zeta_1 - c) \leq \frac{1}{4}\beta(\zeta_3)\bar{\Phi}(\zeta_3 - \zeta_1) =: (ii).$$

The right hand-side equals $2\tilde{m}(\alpha_3)/\pi_1$ by definition of ζ_3. Since γ is Cauchy, Lemma 5 gives $\tilde{m}(\alpha_3) \asymp \zeta_3 g(\zeta_3) \asymp \zeta_3^{-1}$. It is enough to show that $(\pi_1\zeta_3)^{-1}$ is larger than $C\phi(\zeta_1 - c)$, for suitably large $C > 0$.

Let us distinguish two cases. In the case $\zeta_3 \leq 2\zeta_1$, the previous claim is obtained, since ζ_1 goes to infinity with n/s_n by Lemma 8 and $\phi(\zeta_1 - c) = o(\zeta_1^{-1})$. In the case $\zeta_3 > 2\zeta_1$, we obtain an upper bound on ζ_3 by rewriting the equation defining it. For $t \geq 1$, one has $\bar{\Phi}(t) \geq C\phi(t)/t$. Since $\zeta_3 - \zeta_1 > \zeta_1$ in the present case, it follows from the equation defining ζ_3 that

$$C\frac{\phi(\zeta_3 - \zeta_1)}{\zeta_3 - \zeta_1} \leq 8\alpha_3\tilde{m}(\alpha_3)/\pi_1.$$

This can be rewritten using $\phi(\zeta_3 - \zeta_1) = \sqrt{2\pi}\phi(\zeta_3)\phi(\zeta_1)e^{\zeta_1\zeta_3}$, as well as $\phi(\zeta_3) = g(\zeta_3)\alpha_3/(1+\alpha_3) \gtrsim \alpha_3 g(\zeta_3)$ and $\tilde{m}(\alpha_3) \asymp \zeta_3 g(\zeta_3)$. This leads to

$$\frac{e^{\zeta_1\zeta_3}}{\zeta_3^2} \leq \frac{C}{\pi_1}e^{\zeta_1^2/2}.$$

By using $e^x/x^2 \geq Ce^{x/2}$ for $x \geq 1$ one obtains $\zeta_1^2 e^{\zeta_1\zeta_3/2} \leq e^{\zeta_1^2/2}C/\pi_1$, that is, using $\zeta_1^2 \geq 1$,

$$\pi_1\zeta_3 \leq \pi_1\zeta_1 + \frac{\pi_1 \log(C/\pi_1)}{\zeta_1} \leq \pi_1\zeta_1 + C \leq C'\zeta_1,$$

using that $u \to u\log(1/u)$ is bounded on $(0,1)$. So the previous claim is also obtained in this case, as $\phi(\zeta_1 - c)$ is small compared to $(C'\zeta_1)^{-1}$ for large ζ_1. □

2.4.4 Proof of the convergence rate for the modified estimator

Proof of Theorem 17. The proof is overall in the same spirit as that of Theorem 2 in Johnstone and Silverman (2004) and goes by distinguishing the two cases $s_n \geq \log^2 n$ and $s_n < \log^2 n$. The main difference is that here we work with the full posterior distribution, and the risk bounds require Lemmas 1–4, that bound the posterior risk in various settings, as well as a result, Lemma 13 below, in the same vein.

Also, we need to work with a modified version of ζ_1, to make sure that the probability in Lemma 9 goes to 0 fast enough. We note that this version of ζ_1 is the one used in Johnstone and Silverman (2004) for both their Theorems 1 and 2 (in our Theorem 15, such a modification is not needed and we worked with the simpler version there). To do

so, one replaces $\eta_n = s_n/n$ in the definition (2.3.7) of α_1 by

$$\tilde{\eta}_n = \max\left(\eta_n, \frac{\log^2 n}{n}\right).$$

To keep notation simple, we still denote the corresponding threshold by ζ_1. In the first part of the proof below, $\eta_n \geq \log^2(n)/n$, so this is the same version as in definition (2.3.7). In the second part of the proof, we have $\tilde{\eta}_n = \log^2 n/n$ and we now indicate the relevant properties of the corresponding modified threshold ζ_1. First, the statement of Lemma 8 becomes, with $\kappa = 2$ (as γ is Cauchy),

$$\log(1/\tilde{\eta}_n) + c_1 \leq \frac{\zeta_1^2}{2} \leq \log(1/\tilde{\eta}_n) + \frac{1}{2}\log\log n + c_2. \tag{2.4.2}$$

Second, we need below a bound on $P[\hat{\zeta} < \zeta_1]$ with the modified version of ζ_1 as above. It is not hard to check from the proof of Lemma 9 that this proof goes through with the new version of ζ_1 and η_n replaced by $\tilde{\eta}_n$. The only difference is with the term cs_n/α_1 which is bounded by $cn\tilde{\eta}_n/\alpha_1 = n\tilde{m}(\alpha_1)$, so that Bernstein's inequality gives

$$P[\hat{\zeta} < \zeta_1] \leq \exp\{-C'n\alpha_1\tilde{m}(\alpha_1)\} \leq \exp\{-Cn\tilde{\eta}_n\} \leq e^{-C\log^2 n}. \tag{2.4.3}$$

We are now ready for the proof of Theorem 17. First consider the case $s_n \geq \log^2 n$ and let us show that the risk of the empirical Bayes posterior $\Pi_{\hat{\alpha}_A}[\cdot \,|\, X]$ is not larger than that of the non-modified one. One decomposes

$$\begin{aligned}
&E_{\theta_0}\int \|\theta - \theta_0\|^2 d\Pi_{\hat{\alpha}_A}(\theta \,|\, X) \\
&= E_{\theta_0}\int \|\theta - \theta_0\|^2 d\Pi_{\hat{\alpha}}(\theta \,|\, X)1_{\hat{t}\leq t_n} + E_{\theta_0}\int \|\theta - \theta_0\|^2 d\Pi_{\hat{\alpha}_A}(\theta \,|\, X)1_{\hat{t}>t_n} \\
&\leq E_{\theta_0}\int \|\theta - \theta_0\|^2 d\Pi_{\hat{\alpha}}(\theta \,|\, X) + E_{\theta_0}\int \|\theta - \theta_0\|^2 d\Pi_{\alpha_A}(\theta \,|\, X)1_{\hat{t}>t_n} = (I) + (II).
\end{aligned}$$

The term (I) corresponds to the risk of the unmodified estimator, so is bounded as in Theorem 15. For (II), one splits it according to small and large signals $\theta_{0,i}$: $(II) = S + \tilde{S}$, with

$$S = \sum_{i:\ |\theta_{0,i}|\leq\zeta_1} E_{\theta_0}\int (\theta_i - \theta_{0,i})^2 d\Pi_{\alpha_A}(\theta_i \,|\, X)1_{\hat{t}>t_n},$$

and $\tilde{S} = (II) - S$. From Lemma 1, one knows that $r_2(\alpha_A, \mu, x) \leq \mu^2 + C(1 + (x-\mu)^2)$, while for $\mu = 0$, one can use the bound in expectation $E_0 r_2(\alpha, 0, x) \leq C\alpha\tau(\alpha)$, so that

$$S \leq \{ \sum_{i:\ |\theta_{0,i}|=0} + \sum_{i:\ 0<|\theta_{0,i}|\leq\zeta_1} \} E_{\theta_0} \int (\theta_i - \theta_{0,i})^2 d\Pi_{\alpha_A}(\theta_i \,|\, X) \leq Cn\alpha_A\tau(\alpha_A) + Cs_n\zeta_1^2.$$

We now use the definition of α_A to bound α_A and $\tau(\alpha_A)$. To bound $\tau(\alpha_A)$, note that for any $\alpha \in (0,1)$, by definition $a(\tau(\alpha)) = 1/2$, so for a signal of amplitude $\tau(\alpha)$, the posterior puts $1/2$ of its mass at zero, which means the posterior median is 0, implying $\tau(\alpha) \leq t(\alpha)$, so that $\tau(\alpha_A) \leq t_A$. Combining with the bound for α_A of Lemma 12,

$$n\alpha_A\tau(\alpha_A) \leq Cn^{-A}t_A^3.$$

For any fixed $A > 0$, this goes to 0 with n so it is a $o(s_n\zeta_1^2)$, while $s_n\zeta_1^2$ is bounded by $Cs_n \log(n/s_n)$ as follows from Lemma 8. Now to bound $\tilde{S}$, one adapts the last bound of Lemma 1 to accommodate for the indicator $1_{\hat{t}>t_n}$. This is done in Lemma 13 whose bound (2.4.5) implies $\tilde{S} \leq Cs_n t_A^2 P(\hat{t} > t_n)^{1/2}$. This bound coincides up to a universal constant with the corresponding bound (128) in Johnstone and Silverman (2004) (taken for $p = 0$, $\tilde{p} = 1$ and $q = 2$, which corresponds to our setting, i.e. working with ℓ_0 classes and quadratic risk). So the remaining bounds of Johnstone and Silverman (2004) for the case $s_n > c\log^2 n$ can be used directly (the distinction of the three cases as in Johnstone and Silverman (2004) p. 1646-1647 can be reproduced word by word, and is omitted for brevity), leading to $\tilde{S} \leq Cs_n \log(n/s_n)$.

Second, consider the case where $s_n \leq \log^2 n$. We note that for this regime of s_n, the inequalities (2.4.2) become, using that by definition $\tilde{\eta}_n = \log^2 n/n$,

$$\log n - 2\log\log n + c_1 \leq \frac{\zeta_1^2}{2} \leq \log n - \frac{3}{2}\log\log n + c_2. \tag{2.4.4}$$

Let us show that the risk of the plug-in posterior using the modified estimator is at most of the order of the minimax risk. For ζ_1 as above,

$$\begin{aligned}
&E_{\theta_0} \int \|\theta - \theta_0\|^2 d\Pi_{\hat{\alpha}_A}(\theta \,|\, X) \\
&= \Big[\sum_{i:\ \theta_{0,i}=0} + \sum_{i:\ 0<|\theta_{0,i}|\leq\zeta_1} + \sum_{i:\ |\theta_{0,i}|>\zeta_1} \Big] E_{\theta_0} \int (\theta_i - \theta_{0,i})^2 d\Pi_{\hat{\alpha}_A}(\theta_i \,|\, X) \\
&=: (i) + (ii) + (iii).
\end{aligned}$$

For the terms (i) and (ii), apply respectively each bound of Lemma 3 with $\alpha = \alpha_A$ to get $(ii) \leq Cs_n\left[\zeta_1^2 + 1\right] \leq C's_n\zeta_1^2 \lesssim s_n \log n$ using (2.4.2), which is bounded from above by $Cs_n \log(n/s_n)$ in the regime $s_n \leq \log^2 n$. Also,

$$(i) \leq Cn\left[\alpha_A\tilde{\tau}(\alpha_A) + P[\hat{\alpha}_A > \alpha_A]^{1/2}\right].$$

For large enough n, we have $\tilde{\tau}(\alpha_A) = \tau(\alpha_A)$ which is less than $t(\alpha_A) = t_A$ as noted above. Now α_A is bounded using Lemma 12, so that $n\alpha_A\tilde{\tau}(\alpha_A) \lesssim t_A(1+A)(\log n)n^{-A} = o(1)$ for $A > 0$.

We now bound the probability $P[\hat{\alpha}_A > \alpha_A]^{1/2}$. Recall the inequality $t(\alpha)^2 \geq \zeta(\alpha)^2 - C$ (see e.g. (53) in Johnstone and Silverman (2004)). Using (2.4.4), we have $\zeta_1^2 \geq 2\log n - 4\log\log n + 2c_1$ so, writing in slight abuse of notation $t(\zeta_1) = t(\alpha_1)$ seeing $t(\cdot)$ as a function of ζ_1 instead of α_1,

$$t(\zeta_1)^2 \geq t_n^2 + \log\log n - C + 2c_1$$

so that $t(\zeta_1) \geq t_n$ for n large enough. Deduce $\{\hat{\alpha}_A > \alpha_A\} = \{\hat{t} < t_n\} \subset \{\hat{t} < t(\zeta_1)\} = \{\hat{\zeta} < \zeta_1\}$. Using (2.4.3), we have $P[\hat{\zeta} < \zeta_1] \leq e^{-C\log^2 n}$, so that (i) goes to 0, and so is a $o(s_n \log(n/s_n))$.

Finally, for the term (iii) one uses Lemma 4 with $\alpha = \alpha_A$. Note $\{\hat{\alpha}_A < \alpha_A\} = \{t(\hat{\alpha}_A) > t_A\}$. But by definition note that $t(\hat{\alpha}_A)$ equals either t_A if $\hat{t} > t_n$ or $t(\hat{\alpha})$ if $\hat{t} = t(\hat{\alpha}) \leq t_n$, so that $t(\hat{\alpha}) \leq t_n$. As $t_n^2 < 2\log n < t_A^2$ for $A > 0$, conclude that in all cases $t(\hat{\alpha}_A) \leq t_A$ with probability one, so that $P[\hat{\alpha}_A < \alpha_A] = 0$. Thus

$$(iii) \leq \sum_{i:\ |\theta_{0,i}|>\zeta_1} E_{\theta_{0,i}} r_2(\hat{\alpha}_A, \theta_{0,i}, X_i) \leq C \sum_{i:\ |\theta_{0,i}|>\zeta_1} (1 + \tilde{\tau}(\alpha_A)^2 + 0) \leq Cs_n\tilde{\tau}(\alpha_A)^2,$$

which is no more than $2Cs_n(1+A)\log n \leq C's_n \log n$. As $s_n \leq c\log^2 n$, we have $\log n \lesssim \log(n/s_n)$ so $(iii) \leq Cs_n \log(n/s_n)$. Putting the previous bounds together, one gets $(i) + (ii) + (iii) \leq Cs_n \log(n/s_n)$, which concludes the proof. □

Lemma 12. For $A \geq 0$, with $t_A^2 = 2(1+A)\log n$ and $\alpha_A = t^{-1}(t_A)$, there exist $N_0 > 0$ and $C > 0$ both independent of A such that for $n \geq N_0$,

$$\alpha_A \leq C(1+A)(\log n)n^{-1-A}.$$

Proof. First recall the bound $t(\alpha) < \zeta(\alpha)$. Setting $\alpha = t^{-1}(u)$ in this inequality leads, using $\zeta(u) = \beta^{-1}(1/u)$, to $u < \beta^{-1}(1/t^{-1}(u))$. As β is increasing on $\mathbb{R}^+$, one has

$t^{-1}(u) < 1/\beta(u)$, so

$$\alpha_A < \frac{1}{\beta(t_A)} = \frac{g}{\phi - g}(t_A)\frac{\phi}{g}(t_A) \leq 2\frac{\phi}{g}(t_A) \leq Ct_A^2 e^{-t_A^2},$$

where we use that g has Cauchy tails. The result follows by using the expression of t_A. □

Lemma 13. For any real μ, for $B := \{\hat{t} > t_n\}$, and α_A, t_A as above,

$$E_\mu[r_2(\alpha_A, \mu, x)\mathbb{1}_B] \leq C(t_A^2 + 1)P(B)^{1/2}. \tag{2.4.5}$$

Proof. Similar to the proof of Lemma 1, one sets $T := \tau(\alpha_A)$ and distinguishes two cases: if $|\mu| \leq 4T$, Lemma 1 implies $r_2(\alpha_A, \mu, x) \leq \mu^2 + (1 + (x-\mu)^2)$, so using Cauchy-Schwarz inequality,

$$E_\mu[r_2(\alpha_A, \mu, x)\mathbb{1}_B] \leq CT^2P(B) + P(B) + E_\mu[(x-\mu)^4]^{1/2}P(B)^{1/2} \leq C(1+T^2)P(B)^{1/2}.$$

If $|\mu| > 4T$, one uses the bound on r_2 from Lemma 1 again keeping the dependence in $a(x)$. First,

$$E[a(x)\{1 + (x-\mu)^2\}\mathbb{1}_B] \leq E[\{1 + (x-\mu)^2\}^2]^{1/2}P(B)^{1/2} \leq CP(B)^{1/2}.$$

Let us now focus on $E_\mu[(1-a(x))\mu^2\mathbb{1}_B] \leq E_\mu[\{1_{|x|\leq T} + e^{-(|x|-T)^2/2}1_{|x|>T}\}\mathbb{1}_B]$. The first term, using $P_\mu[|x| < T] \leq \bar{\Phi}(|\mu|/2)$, is bounded by $\mu^2\Phi(|\mu|/2)^{1/2}P(B)^{1/2} \leq CP(B)^{1/2}$. The second term is bounded by $\mu^2\{E_\mu[e^{(-|x|-T)^2}]\}^{1/2}P(B)^{1/2}$. In the proof of Lemma 1, we showed that $E_\mu[e^{(-|x|-T)^2/2}]^{1/2}$ is bounded by a universal constant times $\bar{\Phi}(|\mu|/4)$. As $e^{-y^2} \leq e^{-y^2/2}$, the term at stake is bounded from above by $\mu^2\bar{\Phi}(|\mu|/4)P(B)^{1/2} \leq CP(B)^{1/2}$, which implies (2.4.5). □

2.5 Proof of Theorem 16: the SSL prior

Recall that we use the notation of the SAS case, keeping in mind that every instance of g is replaced by g_1 and (some of the) ϕs by g_0. Similarly, $\beta(x, \alpha)$, $\tilde{m}$, m_1 and m_2 are defined as in Section 2.3.1, but with $\beta(x) = g_1/g_0 - 1$.

The main steps of the proof generally follow those of Theorem 15, although technically there are quite a few differences. In the SSL case, we do not know whether the function $\beta = g_1/g_0 - 1$ is nondecreasing over the whole $\mathbb{R}^+$. Yet, we managed to show that β,

which is an even function, is nondecreasing on the interval

$$J_n = [2\lambda_1, \sqrt{2\log n}],$$

see Proposition 2 below. This allows us to define its inverse $\beta^{-1} = {\beta_{|J_n}}^{-1}$ on this interval. Further, we prove in Lemma 20 that β crosses the horizontal axis on the previous interval, is strictly negative on $[0, 2\lambda_1]$ and tends to ∞ when $x \to \infty$. As β is continuous, the graph of the function crosses any given horizontal line $y = c$, for any $c > 0$.

The threshold ζ in the SSL case. For every $\alpha \in (0, 1)$, one sets

$$\zeta = \zeta(\alpha) = \min\{s > 0,\ \beta(s) = 1/\alpha\}. \tag{2.5.1}$$

This is well defined by the property noted in the previous paragraph. Now one notes that $g_0 \leq 2\phi$ for $x \leq \lambda_0/2$, see Lemma 19, and that the function g_1/ϕ takes a value at $\sqrt{2\log n}$ not smaller than $Cn/\log n$, since $g_1 \lesssim \gamma_1$ has Cauchy tails. This implies the existence of a constant $\mathcal{C} > 1$ such that

$$\beta(\sqrt{2\log n}) \geq n/(\mathcal{C}\log n). \tag{2.5.2}$$

Now we claim that for any $\alpha \in (\mathcal{C}\log n/n, 1]$, we have the identity $\zeta(\alpha) = \beta^{-1}(\alpha^{-1})$. To see this, first note that for any $\alpha \in (\mathcal{C}\log n/n, 1]$, by (2.5.2) and $\beta(2\lambda_1) < 0$, we have $\alpha^{-1} \in \beta(J_n)$. This shows that $t = \beta^{-1}(\alpha^{-1})$ solves $\beta(t) = \alpha^{-1}$. Also, it is the smallest possible solution $t > 0$, as β takes negative values on $[0, 2\lambda_1]$, which establishes the identity.

The threshold ζ_1 in the SSL case. In the SSL case, the function $\alpha \to \tilde{m}(\alpha) = -E_0[\beta(X, \alpha)]$ is still nondecreasing, since for any real z, the map $\mathcal{M}_z : \alpha \to z/(1 + \alpha z)$ is nonincreasing and $\beta(X, \alpha) = \mathcal{M}_{\beta(X)}(\alpha)$. By Proposition 3, we also have that $\tilde{m}$ is positive for $\alpha \geq \mathcal{C}\log n/n$ and is of the order of a constant for $\alpha = 1$. So, the map $\alpha \to \alpha\tilde{m}(\alpha)$ is nondecreasing on $[\mathcal{C}\log n/n, 1]$, its value at $\mathcal{C}\log n/n$ is less than $C'\log n/n$, and its value at one is of the order of a constant. This shows, using $s_n \geq c_1 \log^2 n$ by (2.2.7), that the following equation has a unique solution $\alpha_1 \in (\mathcal{C}\log n/n, 1)$

$$\alpha_1\tilde{m}(\alpha_1) = ds_n/n, \tag{2.5.3}$$

with d a small enough constant to be chosen later (see the proof of Lemma 21). Thus we can set

$$\zeta_1 = \beta^{-1}(\alpha_1^{-1}),$$

and by the above arguments we have $\zeta_1 \in J_n$. So Proposition 3 gives $\alpha_1^{-1} \asymp \frac{n}{s_n}\zeta_1 g_1(\zeta_1) \asymp \frac{n}{s_n\zeta_1}$. Now we can follow the same proof as in Lemma 8, replacing up to constants instances of $g_0(\zeta_1)$ by $\phi(\zeta_1)$ thanks to Lemma 17 and (2.6.5) (as $\zeta_1 \le \sqrt{2\log n} < \lambda_0/2$), to obtain

$$\zeta_1^2 \lesssim C\log(n/s_n).$$

Defining $\tau(\alpha)$ and $\tilde\tau(\alpha)$. In the SSL case, we set

$$\Omega(x,\alpha) = \frac{\alpha}{1-\alpha}\frac{2g_1}{\phi}(x).$$

This definition is as in the SAS case except that g is replaced by $2g_1$. We still use the same notation for simplicity. As g_1 satisfies the same properties as g, one defines $\tau(\alpha)$ and $\tilde\tau(\alpha)$ similarly to the SAS case. More precisely, $\tau(\alpha)$ is the unique solution to the equation $\Omega(\tau(\alpha),\alpha) = 1$, whenever $\alpha \le \alpha^*$, where $\Omega(0,\alpha^*) = 1$. One sets $\tau(\alpha) = 0$ for $\alpha \ge \alpha^*$ and $\tilde\tau(\alpha) = \tau(\alpha \wedge \alpha_0)$ with $\tau(\alpha_0) = \lambda_1$ (this slightly differs from the SAS case).

As in the proof of Theorem 15, one can now decompose the risk $R_n(\theta_0) = E_{\theta_0}\int \|\theta - \theta_0\|^2 d\Pi_{\hat\alpha}(\theta \,|\, X)$ according to whether coordinates of θ correspond to a 'small' or 'large' signal, the threshold being ζ_1 that we define next. One can write

$$R_n(\theta_0) = \Big[\sum_{i:\ \theta_{0,i}=0} + \sum_{i:\ 0<|\theta_{0,i}|\le\zeta_1} + \sum_{i:\ |\theta_{0,i}|>\zeta_1}\Big] E_{\theta_0}\int (\theta_i - \theta_{0,i})^2 d\Pi_{\hat\alpha}(\theta_i \,|\, X).$$

We next use the first part of Lemma 16 with $\alpha = \alpha_1$ and the second part of the Lemma to obtain, for any θ_0 in $\ell_0[s_n]$,

$$\begin{aligned}
&\Big[\sum_{i:\ \theta_{0,i}=0} + \sum_{i:\ 0<|\theta_{0,i}|\le\zeta_1}\Big] E_{\theta_0}\int (\theta_i - \theta_{0,i})^2 d\Pi_{\hat\alpha}(\theta_i \,|\, X)\\
&\le C\sum_{i:\ \theta_{0,i}=0}\Big[\alpha_1\tilde\tau(\alpha_1) + P_{\theta_0}(\hat\alpha > \alpha_1) + \lambda_0^{-2}\Big] + \sum_{i:\ 0<|\theta_{0,i}|\le\zeta_1}(\theta_{0,i}^2 + C)\\
&\le C(n-s_n)\Big[\alpha_1\tilde\tau(\alpha_1) + e^{-C\log^2 n} + \lambda_0^{-2}\Big] + (\zeta_1^2 + C)s_n,
\end{aligned}$$

where for the last inequality we use Lemma 21. From (2.5.3) one gets

$$n\alpha_1 \lesssim s_n\zeta_1^{-1}g(\zeta_1)^{-1} \lesssim s_n\zeta_1.$$

Let us now check that $\tilde\tau(\alpha_1) \le \zeta_1$. First, $\beta(\zeta_1) = \alpha_1^{-1} > \alpha_1^{-1} - 1$. By definition of $\tau(\alpha_1)$, using $\phi \le 2g_0$ by Lemma 17,

$$\alpha_1^{-1} - 1 = 2(g_1/\phi)(\tau(\alpha_1)) \ge \beta(\tau(\alpha_1)) + 1.$$

This gives us that $\beta(\zeta_1) \geq \beta(\tau(\alpha_1)) + 1$ which implies the result as β is increasing here. Now with the previous bound on ζ_1 one obtains that the contribution to the risk of the indices i with $|\theta_{0,i}| \leq \zeta_1$ is bounded by a constant times $s_n \log(n/s_n)$.

It remains to bound the part of the risk for indexes i with $|\theta_{0,i}| > \zeta_1$. To do so, one uses the second part of Lemma 16 with α chosen as $\alpha_2' = \mathcal{C}(\log n/n)$, with $\mathcal{C}$ as in (2.5.2). By definition of $\hat{\alpha}$ in (2.2.12), the probability that $\hat{\alpha}$ is smaller than α_2' equals zero. Also, one has $\tilde{\tau}(\alpha_2')^2 \leq C \log n$. Indeed, setting $\zeta_2' = \beta^{-1}({\alpha_2'}^{-1})$, we have as before $\tau(\alpha_2') \leq \zeta_2' \leq \sqrt{2 \log n}$. This implies

$$\sum_{i:\ |\theta_{0,i}|>\zeta_1} E_{\theta_0} \int (\theta_i - \theta_{0,i})^2 d\Pi_{\hat{\alpha}}(\theta_i \,|\, X) \leq C s_n \log n,$$

which concludes the proof of Theorem 16.

2.6 Technical lemmas for the SSL prior

2.6.1 Fixed α bounds

As in the SAS case, we use the notation $r_2(\alpha, \mu, x) = \int (u-\mu)^2 d\pi_\alpha(u \,|\, x)$, where now $\pi_\alpha(\cdot \,|\, x)$ is the posterior on one coordinate (X_1, say) for fixed α in the SSL case, given $X_1 = x$.

Lemma 14. For a zero signal $\mu = 0$, we have for any x and $\alpha \in [0, 1/2]$,

$$\begin{aligned} r_2(\alpha, 0, x) &\leq C\Big[1 \wedge \frac{\alpha}{1-\alpha}\frac{g_1}{\phi}(x)\Big](1+x^2) + \int u^2 \gamma_{0,x}(u) du \\ E_0 r_2(\alpha, 0, x) &\leq C\tau(\alpha)\alpha + 4/\lambda_0^2. \end{aligned}$$

For an arbitrary signal $\mu \in \mathbb{R}$, we have that for any real x and $\alpha \in [0, 1/2]$,

$$\begin{aligned} r_2(\alpha, \mu, x) &\leq (1 - a(x)) \int (u-\mu)^2 \gamma_{0,x}(u) du + Ca(x)((x-\mu)^2 + 1) \\ E_\mu r_2(\alpha, \mu, x) &\leq C(1 + \tilde{\tau}(\alpha)^2). \end{aligned}$$

Proof. By definition, in the SSL case, $r_2(\alpha, 0, x) = (1-a(x)) \int u^2 \gamma_{0,x}(u) du + a(x) \int u^2 \gamma_{1,x}(u) du$. Similar to Lemma 1, we have $a(x) \int u^2 \gamma_{1,x}(u) du \leq C\Big[1 \wedge \frac{\alpha}{1-\alpha}\frac{g_1}{g_0}(x)\Big](1+x^2)$. The first bound now follows from the inequality $g_0 \geq \phi/2$ obtained in Lemma 17. For the bound

in expectation,

$$E_0\left[\int u^2\gamma_{0,x}(u)du\right] = \int\left(\int u^2\frac{\phi(x-u)\gamma_0(u)}{g_0(x)}du\right)\phi(x)dx$$
$$\leq 2\int u^2\int\phi(x-u)dx\gamma_0(u)du = 2\int u^2\gamma_0(u)du = 4/\lambda_0^2,$$

and one then proceeds as in Lemma 1 to obtain the bound for zero signal.

Now for a general signal μ, the bound for $r_2(\alpha,\mu,x)$ follows from the definition and the previous bound. For the bound in expectation, by symmetry one can assume $\mu \geq 0$. Also note that the term with the $a(x)$ factor is bounded in expectation by a constant, by using $a(x) \leq 1$. To handle the term with $1-a(x)$, we distinguish two cases. First, one assumes that $\mu \leq \lambda_0/2$. We have, using $(a+b)^2 \leq 2a^2+2b^2$,

$$(1-a(x))\int(u-\mu)^2\gamma_{0,x}(u)du \lesssim (1-a(x))\mu^2 + (1-a(x))\int u^2\phi(x-u)\frac{\gamma_0(u)}{g_0(x)}du.$$

For the first term we proceed as in Lemma 1, for the second using $g_0 \geq \phi/2$ from Lemma 17,

$$E_\mu\left[(1-a(x))\int u^2\phi(x-u)\frac{\gamma_0(u)}{g_0(x)}du\right] \leq 2\int u^2\gamma_0(u)\int\frac{\phi(x-u)\phi(x-\mu)}{\phi(x)}dxdu$$
$$\lesssim \int u^2\gamma_0(u)\int e^{-(x-(u+\mu))^2/2+u\mu}dxdu \lesssim \lambda_0\int u^2e^{-\lambda_0|u|+u\mu}du.$$

As $\mu \leq \lambda_0/2$, this is in turn bounded by a constant times $(\lambda_0)^{-2}$. Now in the case that $\mu > \lambda_0/2$, recall from the proof of Lemma 1 that for any real x,

$$\int(u-\mu)^2\gamma_{0,x}(u)du = (x-\mu)^2 + 1 + \frac{g_0''}{g_0}(x) + 2(x-\mu)\frac{g_0'}{g_0}(x). \tag{2.6.1}$$

The first two terms are, in expectation, bounded by a constant. Next one writes

$$E_\mu\left[(1-a(x))\frac{g_0''}{g_0}(x)\right] = \int(1-a(x))\frac{g_0''}{g_0}(x)\phi(x-\mu)dx$$

By Lemma 17, we have $|g_0''| = \lambda_0^2|g_0-\phi| \leq 1$. One splits the integral on the last display in two parts. For $|x| \leq \mu/4$, one uses that g_0'' is bounded together with the bound $g_0 \geq \phi/2$. For $|x| > \mu/4$, one uses $g_0''/g_0 = \lambda_0^2(g_0-\phi)/g_0 \leq \lambda_0^2$ together with $1-a(x) \leq (g_0/g_1)(x)/\alpha$,

which follows from the expression of $a(x)$. This leads to

$$E_\mu\left[(1-a(x))\frac{g_0''}{g_0}(x)\right] \leq \int_{|x|\leq\mu/4} e^{x\mu-\frac{\mu^2}{2}}dx + \frac{\lambda_0^2}{\alpha}\int_{|x|>\mu/4}\frac{g_0}{g_1}(x)\phi(x-\mu)dx.$$

The first term in the last expression is bounded. The second one is bounded by a constant given our choice of λ_0 by combining the following: $\alpha^{-1} \leq n$ from (2.2.12), $g_0 \lesssim \gamma_0$ for $\mu > \lambda_0/8$ from (2.6.6) and $g_1 \gtrsim \gamma_1$.

To conclude the proof, for the last term in (2.6.1), using (2.6.4), the bound on $1-a(x)$ from Lemma 15 below, and the fact that $x \mapsto x\phi(x)$ is bounded, $E_\mu\left[2(1-a(x))(x-\mu)\frac{g_0'}{g_0}(x)\right]$ is bounded by

$$\begin{aligned}
&2\int(1-a(x))|\frac{g_0'}{g_0}(x)||(x-\mu)\phi(x-\mu)|dx \lesssim \int(1-a(x))|x|dx\\
&\lesssim \int_{|x|\leq\tilde\tau(\alpha)}|x|dx + \int_{\tilde\tau(\alpha)\leq|x|\leq\frac{\lambda_0}{2}}|x|e^{-\frac{(|x|-\tilde\tau(\alpha))^2}{2}}dx + \int_{|x|\geq\frac{\lambda_0}{2}}|x|(1-a(x))dx\\
&\lesssim \tilde\tau(\alpha)^2 + 2(1-e^{-\frac{(\frac{\lambda_0}{2}-\tilde\tau(\alpha))^2}{2}}) + \tilde\tau(\alpha) + \int_{|x|\geq\frac{\lambda_0}{2}} n^3|x|\frac{\gamma_0}{\gamma_1}(x)dx \lesssim 1+\tilde\tau(\alpha)^2. \qquad \square
\end{aligned}$$

Lemma 15. For any $x \in [0, \lambda_0/2]$ and $\alpha \in [0,1]$,

$$1-a(x) \leq 1\!\!1_{|x|\leq\tilde\tau(\alpha)} + 4e^{-\frac{1}{2}(|x|-\tilde\tau(\alpha))^2}1\!\!1_{|x|>\tilde\tau(\alpha)}.$$

Proof. One first notes that $1-a(x) \leq 4\Omega(x,\alpha)^{-1}$ for $x \leq \lambda_0/2$, using the fact that for such x, $g_0(x) \leq 2\phi(x)$ as found in Lemma 19. The following inequalities hold for $\tilde\tau(\alpha) \leq x \leq \lambda_0/2$, using $\tilde\tau(\alpha) \geq \lambda_1$ by definition and that $|(\log g_1)'| \leq \lambda_1$ as seen in (2.6.3),

$$\begin{aligned}
\Omega(x,\alpha) &= \Omega(\tilde\tau(\alpha),\alpha)\exp\left(\int_{\tilde\tau(\alpha)}^x((\log g_1)'(u)-(\log\phi)'(u))du\right)\\
&\geq \exp\left(\int_{\tilde\tau(\alpha)}^x(u-\lambda_1)du\right) \geq \exp\left(\int_{\tilde\tau(\alpha)}^x(u-\tilde\tau(\alpha))du\right) = e^{\frac{(x-\tilde\tau(\alpha))^2}{2}}. \qquad \square
\end{aligned}$$

2.6.2 Random α bounds

Lemma 16. Let α be a fixed non-random element of $(0,1)$. Let $\hat\alpha$ be a random element of $[0,1]$ that may depend on $x \sim \mathcal{N}(0,1)$ and on other data. Then there exists $C_1 > 0$ such that

$$Er_2(\hat\alpha, 0, x) \leq C_1\left[\alpha\tilde\tau(\alpha) + P(\hat\alpha > \alpha)^{1/2}\right] + \frac{4}{\lambda_0^2}.$$

There exists $C_2 > 0$ such that for any real μ, if $x \sim \mathcal{N}(\mu, 1)$,

$$Er_2(\hat{\alpha}, \mu, x) \leq \mu^2 + C_2.$$

Suppose now that $\tilde{\tau}(\hat{\alpha})^2 \leq d \log(n)$ with probability 1 for some $d > 0$, and that $x \sim \mathcal{N}(\mu, 1)$. Then there exists $C_2 > 0$ such that for all real μ,

$$Er_2(\hat{\alpha}, \mu, x) \leq C_2 \left[1 + \tilde{\tau}(\alpha)^2 + (1 + d \log n) P(\hat{\alpha} < \alpha)^{1/2}\right].$$

Proof of Lemma 16. For the first two inequalities, the proof is the same as in the SAS case in Lemma 3, the only difference being the presence of the term $4/\lambda_0^2$ coming from Lemma 14 for the first inequality. For the third inequality , it follows from Lemma 14 that

$$r_2(\hat{\alpha}, \mu, x) \leq (1 - a_{\hat{\alpha}}(x)) \int (u - \mu)^2 \gamma_{0,x}(u) du + C[(x - \mu)^2 + 1].$$

In expectation the last term is constant. For the first term, with Lemma 15,

$$1 - a_{\hat{\alpha}}(x) \leq \mathbb{1}_{|x| \leq \tilde{\tau}(\hat{\alpha})} + 4 e^{-\frac{1}{2}(|x| - \tilde{\tau}(\hat{\alpha}))^2} \mathbb{1}_{\frac{\lambda_0}{2} \geq |x| > \tilde{\tau}(\hat{\alpha})} + \mathbb{1}_{|x| \geq \frac{\lambda_0}{2}} n \frac{g_0}{g_1}(x),$$

where the last estimate uses the bound $\alpha \geq 1/n$. As in Lemma 4, let us distinguish the two cases $\hat{\alpha} \geq \alpha$ and $\hat{\alpha} < \alpha$. In the case $\hat{\alpha} \geq \alpha$, as $\tilde{\tau}(\alpha)$ is a decreasing function of α,

$$\begin{aligned}
&\left[\mathbb{1}_{|x| \leq \tilde{\tau}(\hat{\alpha})} + 4 e^{-\frac{1}{2}(|x| - \tilde{\tau}(\hat{\alpha}))^2} \mathbb{1}_{\frac{\lambda_0}{2} \geq |x| > \tilde{\tau}(\hat{\alpha})}\right] \mathbb{1}_{\hat{\alpha} \geq \alpha} \\
&\lesssim \left[\mathbb{1}_{|x| \leq \tilde{\tau}(\hat{\alpha})} + \mathbb{1}_{\tilde{\tau}(\hat{\alpha}) < |x| \leq \tilde{\tau}(\alpha)} + e^{-\frac{1}{2}(|x| - \tilde{\tau}(\hat{\alpha}))^2} \mathbb{1}_{\frac{\lambda_0}{2} \geq |x| > \tilde{\tau}(\alpha)}\right] \mathbb{1}_{\hat{\alpha} \geq \alpha} \\
&\lesssim \mathbb{1}_{|x| \leq \tilde{\tau}(\alpha)} + e^{-\frac{1}{2}(|x| - \tilde{\tau}(\alpha))^2} \mathbb{1}_{\frac{\lambda_0}{2} \geq |x| > \tilde{\tau}(\alpha)},
\end{aligned}$$

where we have used $e^{-\frac{1}{2}v^2} \leq 1$ for any v and that $e^{-\frac{1}{2}(u-c)^2} \leq e^{-\frac{1}{2}(u-d)^2}$ if $u > d \geq c$.

For the third term, we have to control $E_\mu \left[\mathbb{1}_{|x| \geq \frac{\lambda_0}{2}} n \frac{g_0}{g_1}(x) \int (u - \mu)^2 \gamma_{0,x}(u) du\right]$. To do so, one uses (2.6.1). In expectation, the term in factor of $(x - \mu)^2 + 1$ is bounded by a constant. Using (2.6.6) and the fact that $g_0''/g_0 \leq \lambda_0^2$, the term in factor g_0''/g_0 is bounded by

$$\begin{aligned}
&\lambda_0^2 n \int_{|x| \geq \frac{\lambda_0}{2}} \frac{g_0}{g_1}(x) \phi(x - \mu) dx \lesssim n^3 \int_{|x| \geq \frac{\lambda_0}{2}} \frac{\gamma_0}{\gamma_1}(x) dx \\
&\lesssim n^4 \int_{|x| \geq \frac{\lambda_0}{2}} x^2 e^{-\lambda_0 |x|} dx \lesssim n^4 e^{-Cn^2}.
\end{aligned}$$

Finally, using (2.6.4) and the fact that $x \mapsto x\phi(x)$ is bounded, one obtains

$$E_\mu\left[\mathbb{1}_{|x|\geq\frac{\lambda_0}{2}} n\frac{g_0}{g_1}(x)(x-\mu)\frac{g_0'}{g_0}(x)\right] \leq \int_{|x|\geq\frac{\lambda_0}{2}} n\frac{g_0}{g_1}(x)|x||(x-\mu)\phi(x-\mu)|dx$$
$$\lesssim \int_{|x|\geq\frac{\lambda_0}{2}} n\frac{g_0}{g_1}(x)|x|dx.$$

As a consequence, one can borrow the fixed α bound obtained previously so that

$$E\left[r_2(\hat{\alpha},\mu,x)1_{\hat{\alpha}\geq\alpha}\right] \lesssim E_\mu r_2(\alpha,\mu,x) \lesssim \left[1+\tilde{\tau}(\alpha)^2\right].$$

In the case $\hat{\alpha}<\alpha$, setting $b_n=\sqrt{d\log n}$ and noting that $\tilde{\tau}(\hat{\alpha})\leq b_n$ with probability 1 by assumption, proceeding as above, with b_n now replacing $\tilde{\tau}(\alpha)$, one can bound

$$\mathbb{1}_{|x|\leq\tilde{\tau}(\hat{\alpha})} + 4e^{-\frac{1}{2}(|x|-\tilde{\tau}(\hat{\alpha}))^2}\mathbb{1}_{\frac{\lambda_0}{2}\geq|x|>\tilde{\tau}(\hat{\alpha})} + \mathbb{1}_{|x|\geq\frac{\lambda_0}{2}} n\frac{g_0}{g_1}(x)$$
$$\lesssim \mathbb{1}_{|x|\leq b_n} + e^{-\frac{1}{2}(|x|-b_n)^2}\mathbb{1}_{\frac{\lambda_0}{2}\geq|x|>b_n} + \mathbb{1}_{|x|\geq\frac{\lambda_0}{2}} n\frac{g_0}{g_1}(x).$$

From this one deduces that $E\left[(1-a_{\hat{\alpha}}(x))\int(u-\mu)^2\gamma_{0,x}(u)du\right]$ is bounded from above by a constant times

$$\left(E_\mu\left[\left(\int(u-\mu)^2\gamma_{0,x}(u)du\right)^2\left[\mathbb{1}_{|x|\leq b_n}+e^{-(|x|-b_n)^2}\right]\right]\right)^{1/2} P(\hat{\alpha}<\alpha)^{1/2}.$$

Using the same bounds but squared as in the fixed α case, one obtains that the expectation in the last display is bounded from above by $C(1+b_n^4)$. Taking the square root and gathering the different obtained bounds concludes the proof. □

2.6.3 Properties of the functions g_0 and β for the SSL prior

Recall the notation ϕ, γ_0, g_0 from Section 2.2. For any real x, we also write $\psi(x) = \int_x^\infty e^{-u^2/2}du$. Our key result on β is the following.

Proposition 2. $\beta = \frac{g_1}{g_0} - 1$ is strictly increasing on $[2\lambda_1; \sqrt{2\log n}]$.

We next state and prove some Lemmas used in the proof of Proposition 2 below.

Lemma 17. The convolution $g_0 = \phi * \gamma_0$ satisfies $g_0'' = \lambda_0^2(g_0-\phi)$ as well as

$$\frac{1}{g_0} \leq \frac{2}{\phi} \qquad \text{and} \qquad |g_0-\phi| \leq \frac{1}{\lambda_0^2}.$$

Proof. The first identity follows by differentiation. One computes $g_0(x)$ by separating the integral in a positive and negative part to get, for any real x,

$$g_0(x) = \frac{\lambda_0 e^{\frac{\lambda_0^2}{2}}}{2\sqrt{2\pi}} \left[e^{\lambda_0 x}\psi(\lambda_0 + x) + e^{-\lambda_0 x}\psi(\lambda_0 - x) \right]. \tag{2.6.2}$$

Now combining the standard inequality $(1 - x^{-2})e^{-x^2/2} \leq x\psi(x) \leq e^{-x^2/2}$, for $x > 0$, with the expression of $g_0(0)$ obtained from (2.6.2), we get $\frac{1}{2} \leq \frac{g_0}{\phi}(0) \leq 1$ for large enough n. By Johnstone and Silverman (2004), Lemma 1, the function g_0/ϕ is increasing, which implies the first inequality of the lemma.

The approximation property of ϕ by g_0 is obtained by a Taylor expansion. For any $x, u \in \mathbb{R}$, there exists c between x and $x - u$ such that $\phi(x - u) - \phi(x) = ux\phi(x) + u^2(c^2 - 1)\phi(c)/2$, so that

$$2(g_0(x) - \phi(x)) = \int (2ux\phi(x) + u^2(c^2 - 1)\phi(c))\gamma_0(u)du = \int u^2(c^2 - 1)\phi(c)\gamma_0(u)du,$$

whose absolute value is bounded by $\int u^2|c^2-1|\phi(c)\gamma_0(u)du$. This is less than $\int u^2\gamma_0(u)du = \lambda_0^{-2}$. □

Lemma 18. Let $L_0 = 5\sqrt{2\pi}$. Then for all $x \in [0; \sqrt{2\log(\lambda_0/L_0)}]$,

$$(\log g_0)'(x) \leq -x/2.$$

Proof. Let $g_{o+}(x) = \int_0^\infty \phi(v + x)\gamma_0(v)dv$ and $g_{o-}(x) = \int_{-\infty}^0 \phi(v + x)\gamma_0(v)dv$. First we check that for any x in the prescribed interval, we have

$$\lambda_0(g_{o+} - g_{o-})(x) \leq -x(\phi(x) - 2/\lambda_0) \leq 0.$$

For any real x, using the inequality $e^v \geq 1 + v$,

$$\begin{aligned} g_{o-}(x) &= \int_0^\infty \phi(x - u)\gamma_0(u)du = \int_0^\infty \phi(x + u)e^{2xu}\gamma_0(u)du \\ &\geq \int_0^\infty \phi(x + u)(1 + 2xu)\gamma_0(u)du \\ &\geq g_{o+}(x) + \lambda_0 x \int_0^\infty u\phi(x + u)e^{-\lambda_0 u}du. \end{aligned}$$

Setting $\Delta(x) = \int_0^\infty u\phi(x+u)e^{-\lambda_0 u}du$, one can write

$$\begin{aligned}\Delta(x) &= \int_0^\infty u(\phi(x+u)-\phi(x))e^{-\lambda_0 u}du + \phi(x)\int_0^\infty ue^{-\lambda_0 u}du \\ &= \int_0^\infty u(\phi(x+u)-\phi(x))e^{-\lambda_0 u}du + \phi(x)/\lambda_0^2.\end{aligned}$$

As ϕ is 1–Lipshitz, one can bound from below $\phi(x+u)-\phi(x) \geq -u$, which leads to, for any $x \geq 0$,

$$\Delta(x) \geq -\int_0^\infty u^2 e^{-\lambda_0 u}du + \phi(x)/\lambda_0^2 \geq -2/\lambda_0^3 + \phi(x)/\lambda_0^2.$$

This leads to inequality on $g_{o+} - g_{o-}$ above, using that x belongs to the prescribed interval to get the nonpositivity. From this one deduces

$$g_0'(x) = \lambda_0(g_{o+}-g_{o-})(x) \leq -x(\phi(x) - 2/\lambda_0).$$

This now implies

$$\frac{g_0'}{g_0}(x) \leq -x\frac{\phi(x)-2\lambda_0^{-1}}{\phi(x)+\lambda_0^{-2}}$$

On the prescribed interval $\phi(x) \geq 5/\lambda_0$, so using that $t \to (t-a)/(t+b)$ is increasing,

$$\frac{g_0'}{g_0}(x) \leq -x\frac{5\lambda_0^{-1}-2\lambda_0^{-1}}{5\lambda_0^{-1}+\lambda_0^{-2}} = -\frac{3x}{5+\lambda_0^{-1}} \leq -\frac{x}{2},$$

for large enough n, which concludes the proof. □

Proof of Proposition 2. We will firstly note that if G_1 has a Cauchy$(1/\lambda_1)$ law,

$$|(\log g_1)'(x)| \leq \lambda_1. \tag{2.6.3}$$

Indeed, for any real x, recalling that $\gamma_1(x) = (\lambda_1/\pi)(1+\lambda_1^2x^2)^{-1}$, one sees that $\gamma_1'(x)/\gamma_1(x) = (-2\lambda_1^2 x)/(1+2\lambda_1^2x^2)$ and $|\gamma_1'(x)/\gamma_1(x)| \leq 2\sqrt{2}\lambda_1/3$. This implies (2.6.3), as

$$\begin{aligned}|(\log g_1)'(x)| &= |\int \phi(x-u)\gamma_1'(u)du|/g_1(x) \\ &\leq \frac{2\sqrt{2}}{3}\lambda_1 \int \phi(x-u)\gamma_1(u)du/g_1(x) \leq \frac{2\sqrt{2}}{3}\lambda_1 \leq \lambda_1.\end{aligned}$$

Let $(x,y) \in [2\lambda_1; \lambda_0/4]^2$ with $x \leq y$. Using Lemma 18 one can find $c \in [x;y]$ with $\log(g_0(x)/g_0(y)) = (x-y)(\log g_0)'(c) \geq (x-y)(-c/2) \geq (y-x)x/2$. On the other

hand, by (2.6.3) one deduces that for some $c \in [x; y]$, we have $\log(g_1(x)/g_1(y)) = (x-y)(\log g_1)'(c) \leq (y-x)\lambda_1$. Thus for any x, y as before,

$$\frac{g_1(x)}{g_1(y)} \leq e^{(y-x)\lambda_1} \qquad \text{and} \qquad e^{(y-x)\frac{x}{2}} \leq \frac{g_0(x)}{g_0(y)}.$$

As $x \geq 2\lambda_1$ by assumption, this leads to the announced inequality. □

Lemma 19. For n large enough, recalling that λ_0 depends on n, we have

$$(\log g_0)'(x) \geq -x \qquad \text{for any } x > 0, \tag{2.6.4}$$

$$g_0(x) \leq 2\phi(x) \qquad \text{for any } 0 \leq x \leq \lambda_0/2, \tag{2.6.5}$$

$$g_0(x) \lesssim \gamma_0(x) \qquad \text{for any } x \geq \lambda_0/8. \tag{2.6.6}$$

Proof. For any real x, we set $\mu_{0,1}(x) = \int u \frac{\phi(x-u)\gamma_0(u)}{g_0(x)} du$, the expectation of $\gamma_{0,x}$. A direct computation shows, for $x > 0$ that $(\log g_0)'(x) = -x + \mu_{0,1}(x)$. But

$$\begin{aligned}
\mu_{0,1}(x) &= \int_0^\infty u \frac{\lambda_0 \phi(x-u) e^{-\lambda_0 u}}{2g_0(x)} du + \int_{-\infty}^0 u \frac{\lambda_0 \phi(x-u) e^{\lambda_0 u}}{2g_0(x)} du \\
&= \int_0^\infty u \frac{\lambda_0 e^{-\lambda_0 u}}{2g_0(x)} (\phi(x-u) - \phi(x+u)) du \\
&= \int_0^\infty u \frac{\lambda_0 e^{-\lambda_0 u}}{2g_0(x)} \phi(x+u)(e^{2xu} - 1) du \geq 0,
\end{aligned}$$

which leads to (2.6.4). For the second point, we first prove the identity, for $x > 0$,

$$\begin{aligned}
g_0(x) = \frac{e^{\lambda_0^2/2}}{\sqrt{2\pi}} \psi(\lambda_0)\gamma_0(x) + \phi(x)\frac{\lambda_0}{2}\Big(& e^{(\lambda_0 - x)^2/2}(\psi(\lambda_0 - x) - \psi(\lambda_0)) \\
& + e^{(\lambda_0+x)^2/2}\psi(\lambda_0 + x)\Big).
\end{aligned}$$

Indeed, $g_0(x) = \int_0^\infty \phi(u)(\gamma_0(x+u)+\gamma_0(x-u))du = \gamma_0(x)\int_0^\infty \phi(u)e^{-\lambda_0 u}du + \int_0^\infty \phi(u)\gamma_0(x-u)du$, for $x > 0$. The first term equals $e^{\lambda_0^2/2}\psi(\lambda_0)\gamma_0(x)/\sqrt{2\pi}$. The second one equals

$$\begin{aligned}
&\int_{-x}^\infty \phi(x+v)\gamma_0(v)dv = \phi(x)\int_{-x}^\infty e^{-\frac{v^2}{2}-vx}\gamma_0(v)dv \\
&= \phi(x)\frac{\lambda_0}{2}\left(\int_{-x}^0 e^{-\frac{v^2}{2}-vx+\lambda_0 v}dv + \int_0^\infty e^{-\frac{v^2}{2}-vx-\lambda_0 v}dv\right) \\
&= \phi(x)\frac{\lambda_0}{2}\left(\int_0^x e^{-\frac{v^2}{2}+vx-\lambda_0 v}dv + e^{\frac{(x+\lambda_0)^2}{2}}\int_0^\infty e^{-\frac{(v+x+\lambda_0)^2}{2}}dv\right) \\
&= \phi(x)\frac{\lambda_0}{2}\left(e^{\frac{(\lambda_0-x)^2}{2}}\int_{\lambda_0-x}^{\lambda_0} e^{-\frac{u^2}{2}}du + e^{\frac{(x+\lambda_0)^2}{2}}\psi(x+\lambda_0)\right)
\end{aligned}$$

which gives the announced identity. If $x \le \lambda_0/2$, using the inequality $y\psi(y) \le e^{-y^2/2}$ for $y > 0$, we have

$$g_0(x) \le \lambda_0^{-1}\gamma_0(x)/\sqrt{2\pi} + \phi(x)(\lambda_0/2)\left[(\lambda_0 - x)^{-1} + (\lambda_0 + x)^{-1}\right].$$

This leads, using $\gamma_0(x)/\lambda_0 \le e^{-\lambda_0^2/2}$ for $x \le \lambda_0/2$, to $g_0(x) \le \phi(x)(1/2+1+1/2) = 2\phi(x)$.

For the third point, if $x \ge \lambda_0/8$, the first term is bounded as follows:

$$\begin{aligned}
\lambda_0 e^{\lambda_0^2/2}e^{\lambda_0 x}\psi(\lambda_0+x) &\le \lambda_0 e^{\lambda_0^2/2}e^{\lambda_0 x}e^{-\lambda_0^2/2-x^2/2-\lambda_0 x}(\lambda_0+x)^{-1} \\
&\le \lambda_0(\lambda_0+x)^{-1}e^{-x^2/2} \le \lambda_0(9\lambda_0/8)^{-1}e^{-x^2/2}.
\end{aligned}$$

Now $\psi(\lambda_0 - x) \le e^{-\lambda_0^2/2-x^2/2+\lambda_0 x}(\lambda_0 - x)^{-1} \le 4e^{-\lambda_0^2/2-x^2/2+\lambda_0 x}\lambda_0^{-1}$ if $\lambda_0/8 \le x \le 3\lambda_0/4$, which leads to $g_0(x) \lesssim \phi(x)$. If $x \ge 3\lambda_0/4$ one bounds the second term by $\lambda_0 e^{\lambda_0^2/2-\lambda_0 x} \le \lambda_0 e^{2\lambda_0 x/3-\lambda_0 x} \le \lambda_0 e^{-\lambda_0 x/3}$, so that, for $x \ge \lambda_0/8$,

$$g_0(x) \lesssim \gamma_0(x). \qquad \square$$

The next lemma is useful to control β outside $[2\lambda_1, \sqrt{2\log n}]$.

Lemma 20. Set $\lambda_1 = 0.05$. For n large enough, for some $C > 0$, we have

$$\begin{aligned}
&(g_1/g_0)(2\lambda_1) < 0.25, \\
&\beta(x) < 0 && \text{for all } x \in [0, 2\lambda_1], \\
&\beta(x) \gtrsim n/\log n, && \text{for all } \sqrt{2\log n} \le x \le \lambda_0/2, \\
&\beta(x) \gtrsim e^{Cn^2}\gamma_1(n)/n && \text{for all } x \ge \lambda_0/8.
\end{aligned}$$

Proof. 1) We have $\frac{g_1}{g_0}(2\lambda_1) \leq \frac{\lambda_1\sqrt{2\pi}}{\lambda_0 \int e^{-(u-2\lambda_1)^2/2} e^{-\lambda_0|u|} du}$. For the denominator, we have

$$\begin{aligned}
\int e^{-(u-2\lambda_1)^2/2} e^{-\lambda_0|u|} du &\geq \int_0^\infty e^{-(u-2\lambda_1)^2/2-\lambda_0 u} du \\
&\geq e^{\lambda_0^2/2-2\lambda_1\lambda_0} \int_0^\infty e^{-(u-(2\lambda_1-\lambda_0))^2/2} du \\
&\geq e^{-2\lambda_1^2}\psi(\lambda_0-2\lambda_1)/(\lambda_0-2\lambda_1) \\
&\geq e^{-2\lambda_1^2}(\lambda_0-2\lambda_1)^{-1}(1-(\lambda_0-2\lambda_1)^{-2}) \\
&\geq 0.99 e^{-2\lambda_1^2}(\lambda_0-2\lambda_1)^{-1} \text{for } n \text{ large enough}
\end{aligned}$$

This implies $(g_1/g_0)(2\lambda_1) < 0.25$ for $\lambda_1 = 0.05$.

2) Let $x \in [0, 2\lambda_1]$, using Lemma 17, we have $\beta \leq 2g_1/\phi - 1$. As the last function is increasing as we know from the SAS case, we have $\beta(x) \leq 2(g_1/\phi)(2\lambda_1) - 1$. With (2.6.5) we end up with $\beta(x) \leq 4(g_1/\phi)(2\lambda_1) - 1$, which is strictly negative by the first point.

3) Let $x \in [\sqrt{2\log n}, \lambda_0/2]$. With (2.6.5), we have $\beta(x) \geq (g_1/2\phi)(x) - 1 \geq (g_1/2\phi)(\sqrt{2\log n}) - 1$, and as $g_1 \gtrsim \gamma_1$, we end up with $\beta(x) \gtrsim n/\log n$.

4) For $x \geq \lambda_0/8$, via (2.6.6) we have $\beta(x) + 1 \geq (\gamma_1/\gamma_0)(x) \geq (\gamma_1/\gamma_0)(\lambda_0/8)$ which gives the result.

□

2.6.4 Bounds on moments of the score function

Recall that, for all $k \geq 1$, $\mu \in \mathbb{R}$ and $\alpha \in [0,1]$, $m_k(\mu, \alpha) = E[\beta(Z+\mu)^k]$ where $Z \sim \mathcal{N}(0,1)$, and $\tilde{m}(\alpha) = -m_1(0,\alpha) = -2\int_0^\infty \beta(z,\alpha)\phi(z)dz$.

Proposition 3. With κ as in (2.3.6), there exist constants D_1 and D_2 such that for $\alpha \in (\mathcal{C}\log n/n, 1]$, $D_1\zeta^{\kappa-1}g_1(\zeta) \leq \tilde{m}(\alpha) \leq D_2\zeta^{\kappa-1}g_1(\zeta)$. Also, $c \leq \tilde{m}(1) \leq C$ with c, C independent of n.

Proof. Recall that for $\alpha \in (\mathcal{C}\log n/n, 1]$, we have $\zeta = \beta^{-1}(\alpha^{-1})$ and $\zeta \leq \sqrt{2\log n}$.

$$\begin{aligned}
\tilde{m}(\alpha) &= -2\int_0^\infty \frac{\beta(z)}{1+\alpha\beta(z)}\phi(z)dz \\
&= -2\int_0^\infty \beta(z)\phi(z)dz + 2\int_0^\infty \frac{\alpha\beta^2(z)}{1+\alpha\beta(z)}\phi(z)dz \\
&= -2\int_0^\infty \beta(z)\phi(z)dz + 2\int_0^\zeta \frac{\alpha\beta^2(z)}{1+\alpha\beta(z)}\phi(z)dz \\
&\quad +2\int_\zeta^\infty \frac{\alpha\beta^2(z)}{1+\alpha\beta(z)}\phi(z)dz \\
&:= A + B + C
\end{aligned}$$

- For the first term, with K a positive constant one can write

$$\begin{aligned} A &= 2\int_0^\infty (\phi - \frac{g_1}{g_0}\phi) = 2\int_0^\infty (\phi - \frac{g_1}{g_0}(\phi - g_0 + g_0)) \\ &= 2\int_0^\infty (\phi - g_1) + 2\int_0^\infty \frac{g_1(g_0-\phi)}{g_0} \\ &= 0 + 2\int_0^{K\zeta} \frac{g_1(g_0-\phi)}{g_0} + 2\int_{K\zeta}^\infty \frac{g_1(g_0-\phi)}{g_0} \\ &:= (i) + (ii). \end{aligned}$$

  Using the fact that g_1/ϕ is increasing, we have

$$\begin{aligned} |(i)| &\leq 2\lambda_0^{-2}\int_0^{K\zeta} g_1/g_0 \leq 4\lambda_0^{-2}\int_0^{K\zeta} g_1/\phi \\ &\leq 4K\zeta g_1(K\zeta)\lambda_0^{-2}/\phi(K\zeta) \lesssim Kn^{K^2-2}\zeta g_1(K\zeta) \end{aligned}$$

  Taking $K = 6/5$, we end up with $|(i)| \lesssim \zeta n^{-2/5} g_1(6\zeta/5)$ and this term is strictly dominated by $\zeta^{\kappa-1}g_1(\zeta)$. By Lemma 17, and the fact that $g_1 \asymp \gamma_1$, we have :

$$\begin{aligned} |(ii)| &\leq 2\int_{K\zeta}^\infty g_1(1+\phi/g_0) \leq 6\int_{K\zeta}^\infty g_1 \\ &\lesssim (6\zeta/5)^{\kappa-1} g_1(6\zeta/5) \text{ using (2.3.6)} \end{aligned}$$

  This term too is dominated by $\zeta^{\kappa-1}g_1(\zeta)$.

- For the second term, we use the fact that on $(0,\zeta)$, $\alpha|\beta| < 1$, so $1+b_0 \leq 1+\alpha\beta \leq 2$, where $b_0 = g_1(2\lambda_1)/2\phi(0) - 1$ does not depend on n, so that

$$B \asymp \int_0^\zeta \alpha\beta^2(z)\phi(z)dz$$

  We will now use the fact that, with $h := g_1^2/\phi$, $\int_0^\zeta h(z)dz \leq 16h(\zeta)/\zeta$. This is a direct corollary of lemma 4 in (Johnstone and Silverman, 2004). We have, also using (2.6.5):

$$\begin{aligned} \int_0^\zeta \beta^2(z)\phi(z)dz &\lesssim \int_0^\zeta (g_1^2/g_0^2)\phi \lesssim \int_0^\zeta g_1^2/\phi \\ &\lesssim g_1^2(\zeta)/(\zeta\phi(\zeta)) \lesssim \beta(\zeta)g_1(\zeta)/\zeta \lesssim g_1(\zeta)(\alpha\zeta)^{-1} \end{aligned}$$

hence $B \lesssim g_1(\zeta)\zeta^{-1}$, dominated by $\zeta^{\kappa-1}g_1(\zeta)$.

- For the last term, we first use the fact that $\alpha\beta(z) < 1 + \alpha\beta(z)$, so that $C \lesssim \int_\zeta^\infty \beta(z)\phi(z)dz$.

$$\begin{aligned} C &\lesssim \int_\zeta^\infty g_1\phi/g_0 \lesssim \int_\zeta^\infty g_1(z)dz \text{ using Lemma 17} \\ &\asymp \zeta^{\kappa-1}g_1(\zeta) \text{ using (2.3.6)} \end{aligned}$$

For an upper bound we write

$$C = 2\int_\zeta^{\lambda_0/2} \frac{\alpha\beta^2(z)}{1+\alpha\beta(z)}\phi(z)dz + 2\int_{\lambda_0/2}^\infty \frac{\alpha\beta^2(z)}{1+\alpha\beta(z)}\phi(z)dz =: (i) + (ii).$$

For the first term, using (2.6.5), we have for every $z \in [\zeta, \lambda_0/2]$, $\beta(z) \geq \frac{g_1}{2\phi}(z) - 1 \geq \frac{g_1}{4\phi}(z)$ and $\alpha\frac{g_1}{4\phi}(z) \gtrsim \alpha\frac{n}{\log n} \gtrsim 1$, so that

$$\begin{aligned} (i) &\geq 2\int_\zeta^{\lambda_0/2} \frac{\alpha(g_1^2/16\phi^2)(z)}{1+\alpha(g_1/4\phi)(z)}\phi(z)dz \\ &\gtrsim \int_\zeta^{\lambda_0/2} g_1(z)dz \gtrsim \zeta^{\kappa-1}g_1(\zeta) \end{aligned}$$

For the second term, we have

$$\begin{aligned} (ii) &\lesssim \int_{\lambda_0/2}^\infty \beta(z)\phi(z)dz \\ &\lesssim \int_{\lambda_0/2}^\infty g_1(z)dz \lesssim \lambda_0^{\kappa-1}g_1(\lambda_0) \lesssim \lambda_0^{-1}. \end{aligned}$$

Putting the bounds together finally leads to $\tilde{m}(\alpha) \asymp g_1(\zeta)\zeta^{\kappa-1}$.

To prove $\tilde{m}(1) \leq \phi(0)/g_1(2\lambda_1)$, write $\tilde{m}(1) = -2\int_0^\infty \phi + 2\int_0^\infty \phi/(1+\beta)$.

Now $\int_0^\infty \phi/(1+\beta) = \int_0^{2\lambda_1} \phi/(1+\beta) + \int_{2\lambda_1}^{\lambda_0/2} \phi/(1+\beta) + \int_{\lambda_0/2}^{+\infty} \phi/(1+\beta)$.

Using that on $[0, 2\lambda_1]$, $1 + \beta \geq 1 + b_0 = g_1(2\lambda_1)/2\phi(0)$ and (2.6.5) and (2.6.6), we have

$$\begin{aligned}\int_0^\infty \phi/(1+\beta) &\leq \int_0^{2\lambda_1} \phi/(1+b_0) + \int_{2\lambda_1}^{\lambda_0/2} \phi^2/g_1 + \int_{\lambda_0/2}^\infty \gamma_0\phi/g_1 \\ &\leq \int_0^{2\lambda_1} \phi/(1+b_0) + \int_{2\lambda_1}^\infty \phi^2/g_1 + \int_0^\infty \phi/g_1 \leq C.\end{aligned}$$

For the lower bound, recall that $\tilde{m}(1) = -2\int_0^\infty \phi + 2\int_0^\infty \phi/(1+\beta)$ and use Lemma 17 to write $2\int_0^\infty \phi/(1+\beta) \geq \int_0^\infty \phi^2/g_1$ which does not depend on n. □

Proposition 4. Let $\alpha \in [\mathcal{C}\log n/n, 1]$.

1) For small enough α, we have $m_2(0,\alpha) \lesssim \tilde{m}(\alpha)(\alpha\zeta^\kappa)^{-1}$

2) For $k = 1$ or 2, for all μ and all α small enough, $m_k(\mu,\alpha) \leq (\alpha \wedge |B_0|/(1+B_0))^{-k}$ with $B_0 = g_1(0)/2\phi(0) - 1$.

Proof. 1) Let $\alpha \in [0;1]$, we have

$$\begin{aligned}m_2(0,\alpha) &= 2\int_0^\infty \frac{\beta^2(z)}{(1+\alpha\beta(z))^2}\phi(z)dz \\ &= 2\int_0^\zeta \frac{\beta^2(z)}{(1+\alpha\beta(z))^2}\phi(z)dz + 2\int_\zeta^\infty \frac{\beta^2(z)}{(1+\alpha\beta(z))^2}\phi(z)dz\end{aligned}$$

For the first term, as in Proposition 3, and using Proposition 17, we have

$$\int_0^\zeta \frac{\beta^2(z)}{(1+\alpha\beta(z))^2}\phi(z)dz \lesssim \int_0^\zeta \beta^2(z)\phi(z)dz \lesssim g_1(\zeta)(\alpha\zeta)^{-1}$$

For the last term, by the fact that β is increasing on $[\zeta, \sqrt{2\log n}]$, (2.6.5) and (2.6.6) we have that $\beta > 0$ on $[\zeta, \infty]$ so that

$$\int_\zeta^\infty \frac{\beta^2(z)}{(1+\alpha\beta(z))^2}\phi(z)dz \lesssim 1/\alpha^2 \int_\zeta^\infty \phi(z)dz \lesssim \beta^2(\zeta)\phi(\zeta)/\zeta \lesssim \beta(\zeta)g_1(\zeta)/\zeta$$

hence $m_2(0,\alpha) \lesssim \frac{g_1(\zeta)}{\alpha\zeta}$. Yet $\tilde{m}(\alpha) \asymp \zeta^{\kappa-1}g_1(\zeta)$ when $\alpha \to 0$, which yields the first point.

2) Recall the definition $m_k(\mu,\alpha) = \int \left(\frac{\beta(t)}{1+\alpha\beta(t)}\right)^k \phi(t-\mu)dt$. If $\beta(t) \geq 0$, $\left|\frac{\beta(t)}{1+\alpha\beta(t)}\right| \leq 1/\alpha$. Otherwise we have $|t| < \lambda_0/2$ so using (2.6.5) for the numerator leads to $\beta(t) \geq g_1(0)/2\phi(0) - 1 = B_0$ and for the denominator $|1+\alpha\beta(t)| = 1+\alpha\beta(t) \geq 1+\beta(t) \geq 1+B_0$. □

2.6.5 In-probability bounds

Lemma 21. We take $\alpha = \alpha_1$ and $\zeta = \zeta_1$ as defined by (2.5.3). There exists $C > 0$ such that

$$\sup_{\theta \in \ell_0(s_n)} P_\theta(\hat{\zeta} < \zeta) \leq \exp(-C(\log n)^2).$$

Proof. First note that, almost surely, $\hat{\alpha}^{-1} \geq 1 > \beta(2\lambda_1)$ with the help of the first point of Lemma 20, so $\hat{\zeta} = \beta^{-1}(\hat{\alpha}^{-1}) > 2\lambda_1$. Since β is increasing on $(2\lambda_1, \sqrt{2\log n})$ and $\zeta \leq \sqrt{2\log n}$, we have $\{\hat{\zeta} < \zeta\} = \{\hat{\alpha} > \alpha\}$, so $P(\hat{\zeta} < \zeta) = P(\hat{\alpha} > \alpha) = P(\hat{\alpha} > \alpha \cap S(\alpha) > 0) + P(\hat{\alpha} > \alpha \cap S(\alpha) \leq 0)$.

Let us now focus on the event $\{\hat{\alpha} > \alpha\} \cap \{S(\alpha) \leq 0\}$. If $S(\alpha) \leq 0$, since S is decreasing, $S < 0$ on $]\alpha, \hat{\alpha}]$. So the likelihood l is decreasing on $]\alpha, \hat{\alpha}[$. It implies that there exists $\alpha' \in]\alpha, \hat{\alpha}[$ such that $l(\alpha') > l(\hat{\alpha})$. But this contradicts the maximality of $\hat{\alpha}$. Therefore $\{\hat{\alpha} > \alpha\} \cap \{S(\alpha) \leq 0\} = \emptyset$. Hence $P(\hat{\zeta} < \zeta) = P(\hat{\alpha} > \alpha \cap S(\alpha) > 0) \leq P(S(\alpha) > 0)$.

The score function $S(\alpha) = \sum_{i=1}^n \beta(\theta_i + Z_i, \alpha)$ is a sum of independent random variables, each bounded by α^{-1}. We have $P(S(\alpha) > 0) = P(\sum_{i=1}^n W_i > A)$, with $A = -\sum_{i=1}^n m_1(\theta_i, \alpha)$ and $W_i = \beta(\theta_i + Z_i, \alpha) - m_1(\theta_i, \alpha)$ centered variables, bounded by $M = (1 + c)/\alpha$ using the second point of Proposition 4. Setting $V = \sum_{i=1}^n var(W_i)$, Bernstein's inequality gives

$$P(S(\alpha) > 0) \leq \exp\left(\frac{-A^2}{2(V + \frac{MA}{3})}\right).$$

Moreover, proceeding as in Lemma 9 in the SAS case, we have $-A \lesssim -n\tilde{m}(\alpha)$ and $V \lesssim n\frac{\tilde{m}(\alpha)}{\alpha}$, so $\left(\frac{A^2}{2(V+\frac{MA}{3})}\right)^{-1} = \frac{V}{A^2} + \frac{M}{3A} \leq \frac{C}{\alpha n\tilde{m}(\alpha)} + \frac{C'}{\alpha n\tilde{m}(\alpha)} \lesssim (\alpha n\tilde{m}(\alpha))^{-1}$ therefore $\frac{A^2}{2(V+\frac{MA}{3})} \gtrsim \alpha n\tilde{m}(\alpha) \gtrsim s_n \gtrsim (\log n)^2$ and finally

$$P(S(\alpha) > 0) \leq \exp(-C(\log n)^2). \quad \square$$

Chapter 3

Sharp asymptotic minimaxity of spike and slab empirical Bayes procedures

3.1 Introduction

3.1.1 Model

Consider the sequence model (1.2)

$$X_i = \theta_i + \varepsilon_i, \quad i = 1, \dots, n,$$

with $\theta = (\theta_1, \dots, \theta_n) \in \mathbb{R}^n$ and $\varepsilon_1, \dots, \varepsilon_n$ i.i.d. $\mathcal{N}(0,1)$.

Suppose as before that the 'true' vector θ_0 belongs to

$$\ell_0[s_n] = \{\theta \in \mathbb{R}^n,\ \#\{i :\ \theta_i \neq 0\} \leq s_n\}. \tag{3.1.1}$$

Recall that the minimax rate over $\ell_0[s_n]$ for the Euclidean norm (not renormalized by n) is $2s_n \log(\frac{n}{s_n})(1 + o(1))$.

3.1.2 Posterior convergence at sharp minimax rate

One defines a notion of posterior convergence at the sharp minimax rate, or convergence at the minimax rate with exact constant, with respect to the L^2-norm loss, as follows

$$\sup_{\theta_0 \in \ell_0[s_n]} E_{\theta_0}\left[\int \|\theta - \theta_0\|_2^2 d\Pi(\theta|X)\right] \leq 2s_n \log(\frac{n}{s_n})(1+o(1)) \tag{3.1.2}$$

If (3.1.2) holds, then at least two estimators (one of these randomised) converge at the minimax rate with exact constant in the usual sense. First, using the Jensen inequality, it implies that the posterior mean (denoted here by $\bar{\theta}$) converges at minimax rate with exact constant to the true signal

$$\sup_{\theta_0 \in \ell_0[s_n]} E_{\theta_0}\left[\|\bar{\theta} - \theta_0\|_2^2\right] \leq 2s_n \log(\frac{n}{s_n})(1+o(1)) \tag{3.1.3}$$

Second, let us consider a draw from the posterior distribution. More formally, it is a $\tilde{\theta} = \tilde{\theta}(X, U)$, using the data X and uniform variables U on $[0,1]$, such that $\mathcal{L}(\tilde{\theta}(X,U)|X) = \Pi(\cdot|X)$, stating (3.1.2) is exactly stating the convergence to θ_0 at minimax rate with exact constant of $\tilde{\theta}$.
To construct such a $\tilde{\theta}$ in practice in the setting of the sequence model with a Spike and Slab prior, as the a posteriori law is a product, one can take, denoting by $\mathcal{F}_{\theta_i|X}$ the cumulative distribution function of each $\theta_i|X_i$,

$$\tilde{\theta}(X,U) = (\mathcal{F}^{-1}_{\theta_1|X}(U_1), \ldots, \mathcal{F}^{-1}_{\theta_n|X}(U_n))$$

3.1.3 Spike and Slab prior

3.1.3.1 Prior

The spike and slab prior with smoothing parameter α is given by

$$\Pi_\alpha \sim \bigotimes_{i=1}^{n} (1-\alpha)\delta_0 + \alpha G(\cdot), \tag{3.1.4}$$

where δ_0 denotes the Dirac mass at 0 and G is a given probability measure of density γ, often taken to be the Laplace distribution or some heavy tailed distribution.

3.1.3.2 Posterior

The posterior distribution under (3.1.1)-(3.1.4) is

$$\Pi_\alpha[\cdot \mid X] \sim \bigotimes_{i=1}^{n}(1 - a_\alpha(X_i))\delta_0 + a_\alpha(X_i)\gamma_{X_i}(\cdot), \tag{3.1.5}$$

where we have set, denoting ϕ the standard normal density and $\phi * G(x) = \int \phi(x-u)dG(u)$ the convolution of ϕ and G,

$$\begin{aligned} g(X_i) &= (\phi * G)(X_i), \\ \gamma_{X_i}(\cdot) &= \frac{\phi(X_i - \cdot)\gamma(\cdot)}{g(X_i)}, \\ a_\alpha(X_i) &= \frac{\alpha g(X_i)}{(1-\alpha)\phi(X_i) + \alpha g(X_i)} \end{aligned}$$

3.1.3.3 A special Slab density γ

Consider the unimodal symmetric density γ on $\mathbb{R}$ given by

$$\gamma(x) = \frac{1}{2}\Delta(1 + |x|), \qquad \Delta(u) = u^{-1}(1 + \log(u))^{-2}, \tag{3.1.6}$$

The purpose of this new density is to have sufficiently heavy tails, heavier than Cauchy. Apart from this specific tail property, γ still satisfies

$$\sup_{u>0}\left|\frac{d}{du}\log\gamma(u)\right| =: \Lambda < \infty. \tag{3.1.7}$$

Let us denote by $\mathfrak{g} = \phi * \gamma$ the convolution of the heavy-tailed γ given by (3.1.6) and the noise density ϕ. Basic properties of $\mathfrak{g}$ are gathered in Lemma 22, while Lemma 23 provides bounds on corresponding moments of the score function.

3.1.4 Useful Thresholds

Let us recall the following useful threshold properties already used in Chapter 2. As noted in Castillo and Szabo (2018), the properties established by Johnstone and Silverman (2004) are extended without difficulties to slabs with heavier tails than Cauchy.

Posterior median and threshold $t(\alpha)$. The a posteriori median $\hat{\theta}_\alpha^{med}(X) = \left(\hat{\theta}_i^{med}(X_i)\right)_{i\in\{1,\dots,n\}}$ has the following thresholding property: there exists $t(\alpha) > 0$, depending on the smoothing parameter α of (3.1.4) such that $\hat{\theta}_i^{med}(X_i) = 0$ if and only if $|X_i| \le t(\alpha)$. A default

choice can be $\alpha = 1/n$, which leads to a posterior median behaving similarly as a hard thresholding estimator with threshold $\sqrt{2\log n}$. One can significantly improve on this choice by taking a well-chosen data-dependent α, as will be seen in 3.1.5.

The thresholds $\zeta(\alpha)$, $\tau(\alpha)$ and $\tilde{\tau}(\alpha)$. Following Johnstone and Silverman (2004), we introduce several useful thresholds. From Lemma 1 in Johnstone and Silverman (2004), we know that $\mathfrak{g}/\phi$, and therefore $\mathfrak{B} = \mathfrak{g}/\phi - 1$, is a strictly increasing function on $\mathbb{R}^+$. It is also continuous, so given α, a pseudo-threshold $\zeta = \zeta(\alpha)$ can be defined by

$$\mathfrak{B}(\zeta) = \frac{1}{\alpha}. \tag{3.1.8}$$

It is shown in Johnstone and Silverman (2004), Lemma 3, that

$$t(\alpha) \leq \zeta(\alpha). \tag{3.1.9}$$

Further one can also define $\tau(\alpha)$ as the solution in x of

$$\Omega(x, \alpha) := \frac{a(x)}{1 - a(x)} = \frac{\alpha}{1 - \alpha}\frac{g}{\phi}(x) = 1.$$

Equivalently, $a(\tau(\alpha)) = 1/2$. Also, $\mathfrak{B}(\tau(\alpha)) = \alpha^{-1} - 2$ so $\tau(\alpha) \leq \zeta(\alpha)$. Define α_0 as $\tau(\alpha_0) = 1$ and set

$$\tilde{\tau}(\alpha) = \tau(\alpha \wedge \alpha_0). \tag{3.1.10}$$

In the sequel, one can always take α small enough, so it will be silently understood that $\alpha \leq \alpha_0$ so that $\tilde{\tau}(\alpha) = \tau(\alpha)$.

These thresholds are useful to understand the behaviour of the a posteriori law, in particular to have bounds on the updated smoothing parameter $a_\alpha(X)$.

3.1.5 Empirical Bayes choice of α

The log-marginal likelihood in α can be written as

$$\ell(\alpha) = \ell_n(\alpha; X) = \sum_{i=1}^{n} \log((1 - \alpha)\phi(X_i) + \alpha g(X_i)). \tag{3.1.11}$$

Let $\hat{\alpha}$ be defined as the maximiser of the log-marginal likelihood

$$\hat{\alpha} = \underset{\alpha \in \mathcal{A}_n}{\operatorname{argmax}}\ \ell_n(\alpha; X), \tag{3.1.12}$$

where the maximisation is restricted to $\mathcal{A}_n = [\alpha_n, 1]$, with α_n defined, in view of (1.2.11), by

$$t(\alpha_n) = \sqrt{2 \log n}.$$

The reason for this restriction is that one does not need to take α smaller than α_n, which would correspond to a choice of α 'more conservative' than hard-thresholding at threshold level $\sqrt{2 \log n}$.

The a priori law that will be therefore considered is the Spike and Slab where we have 'plugged' the value $\hat{\alpha}$:

$$\theta \sim \Pi_{\hat{\alpha}} := \bigotimes_{i=1}^{n} (1 - \hat{\alpha})\delta_0 + \hat{\alpha}\Gamma \tag{3.1.13}$$

One will also denote the threshold of our new 'plug-in' posterior median, in view of 3.1.4,

$$\hat{t} = t(\hat{\alpha}) \tag{3.1.14}$$

the threshold of the posterior median corresponding to the Spike and Slab prior with plugged-in parameter $\hat{\alpha}$.

3.2 Main result

Our specific choice of Slab density leads the following result, assuming that s_n satisfies the mild condition that there exist constants c_0, c_1 such that

$$c_1 \log^2 n \leq s_n \leq c_0 n. \tag{3.2.1}$$

As in Chapter 2, we note that it is under this condition that the rate in Theorem 1 of Johnstone and Silverman (2004) is optimal.

Theorem 18. Let Π_α be the Spike and Slab prior distribution (3.1.4) with Slab density γ given by (3.1.6). Let $\Pi_{\hat{\alpha}}[\cdot \,|\, X]$ be the corresponding plug-in posterior distribution given by (3.1.13), with $\hat{\alpha}$ chosen by the empirical Bayes procedure (3.1.12). For any s_n verifying (3.2.1) , for $n \to \infty$

$$\sup_{\theta_0 \in \ell_0[s_n]} E_{\theta_0} \int \|\theta - \theta_0\|_2^2 d\Pi_{\hat{\alpha}}(\theta \,|\, X) \leq 2 s_n \log(\frac{n}{s_n})(1 + o(1)).$$

Theorem 18 states that the second moment of the a posteriori law whose smoothing parameter has been chosen with Empirical Bayes and whose Slab is the specific density

(3.1.6) converges to 0 at sharp minimax rate, i.e. where one obtains the exact constant 2. We note that similarly to Chapter 2 or to Theorem 2 of Johnstone and Silverman (2004), one could also consider a modified estimator for $\hat{\alpha}$. As in Chapter 2, we believe that working with this modified estimator $\hat{\alpha}$ should enable one to remove condition (3.2.1) and get the sharp minimax rate also in the regime $s_n \lesssim \log^2 n$.

3.2.1 Why it works

Let us first consider the case where α is a fixed constant in $(0,1)$ to get an intuition on why it is possible at all to obtain this sharp minimax rate result already in case the regularity parameter s_n is given to us. In the quantity

$$\int \|\theta - \theta_0\|^2 d\Pi_\alpha(\theta \,|\, X) = \sum_{i=1}^{n} \int (\theta_i - \theta_{0,i})^2 d\Pi_\alpha(\theta_i \,|\, X_i),$$

let us distinguish two parts: the coordinates of θ_0 that are just equal to zero on one hand, and the nonzero coordinates on the other hand.
For $\alpha \in (0,1)$, $\mu \in \mathbb{R}$ and $x \in \mathbb{R}$, we use the following notation

$$r_2(\alpha, \mu, x) = (1 - a(x))\mu^2 + a(x) \int (u - \mu)^2 \gamma_x(u) du.$$

so that

$$\int \|\theta - \theta_0\|^2 d\Pi_\alpha(\theta \,|\, X) = \sum_{i=1}^{n} r_2(\alpha, \theta_{0,i}, X_i).$$

For the part where the signal is zero, we have the following result, for α small enough

$$E_0 r_2(\alpha, 0, x) \lesssim \alpha \tau(\alpha)^2 \{1 + \log(1 + \tau(\alpha))\}^{-2} \tag{3.2.2}$$

This result immediately follows from Lemma 27, as $\tau(\alpha)$ goes to infinity as $\alpha \to 0$.

For the nonzero coordinates, the following bound directly follows from Lemma 28

$$E_\mu[r_2(\alpha, \mu, x)] \leq \tilde{\tau}(\alpha)^2 (1 + o(1)) \tag{3.2.3}$$

Expected posterior squared norm at fixed α. Putting together the previous bounds leads to, with S_0 the support of θ_0,

$$\begin{aligned} E_{\theta_0} \int \|\theta - \theta_0\|^2 d\Pi_\alpha(\theta \,|\, X) &= \sum_{i=1}^n E_{\theta_0} \int (\theta_i - \theta_{0,i})^2 d\Pi_\alpha(\theta_i \,|\, X_i) \\ &\leq \sum_{i \notin S_0} C\alpha\tau(\alpha)^2 / \log^2(\tau(\alpha)) + \sum_{i \in S_0} \tilde{\tau}(\alpha)^2(1 + o(1)) \\ &\leq Cn\alpha\tau(\alpha)^2 / \log^2(\tau(\alpha)) + s_n\tilde{\tau}(\alpha)^2(1 + o(1)) \end{aligned}$$

which leads to the following result for fixed α

Proposition 5. For a fixed $\alpha \in [s_n/n; \log(\log n/s_n)s_n/n]$, the Spike and Slab prior with parameter α and γ as in (3.1.6) yields the exact constant for the posterior squared norm, that is, for $n \to \infty$

$$\sup_{\theta_0 \in \ell_0[s_n]} E_{\theta_0} \int \|\theta - \theta_0\|^2 d\Pi_\alpha(\theta \,|\, X) \leq 2s_n \log(n/s_n)(1 + o(1))$$

Note that for α as in Proposition 5 (and as soon as $\alpha \geq (s_n/n)^\eta$ for $\eta > 0$ for instance) the nonzero part of the signal always contribute for $2s_n \log(n/s_n)$. The part of zero signal is more dependent on the choice of the Slab. When α is data-driven, this part may interfere with the nonzero part. In the Laplace case, one can check that $\hat{\alpha}$ is too far from the oracle parameter $\alpha^* = s_n/n$, resulting in a zero signal contribution larger than the minimax rate. In the Cauchy case, the zero signal contribution becomes exactly of the order of the minimax rate. With the special Slab (3.1.6), we shall prove that for data-driven α this contribution becomes lower than the minimax rate, finally resulting in $2s_n \log(\frac{n}{s_n})(1 + o(1))$.

3.3 Proofs

3.3.1 Thresholds and Useful Bounds

The following bounds are borrowed from Johnstone and Silverman (2004) (again, they extend without difficulty to the heavy-tailed γ)

Bounds on $a_\alpha(x)$. For any real x and $\alpha \in [0, 1]$,

$$\alpha \frac{\mathfrak{g}}{\mathfrak{g} \vee \phi}(x) \leq a_\alpha(x) \leq 1 \wedge \frac{\alpha}{1 - \alpha} \frac{\mathfrak{g}}{\phi}(x). \tag{3.3.1}$$

The following bound in terms of $\tau(\alpha)$, see Johnstone and Silverman (2004) p. 1623 (one has $\tilde{\tau}(\alpha) = \tau(\alpha)$ for any $\alpha \leq \alpha_0$), is useful for large x,

$$1 - a_\alpha(x) \leq 1\!\!1_{|x| \leq \tau(\alpha)} + e^{-\frac{1}{2}(|x| - \tau(\alpha))^2} 1\!\!1_{|x| > \tau(\alpha)}. \tag{3.3.2}$$

3.3.2 Properties of $\mathfrak{g}$ and moments of the score function

While qualitative properties of $\mathfrak{g}$ and of the score function (such as $\mathfrak{g}/\phi$ or $\tilde{m}$ are increasing functions) do not change with the present heavy-tailed choice of γ, some of the equivalents of $\mathfrak{g}$ and moments of the score function change, in a way that we describe now.

Lemma 22. For γ defined by (3.1.6) and $\mathfrak{g} = \phi * \gamma$, as $x \to \infty$,

$$\mathfrak{g}(x) \asymp \gamma(x)$$
$$\mathfrak{g}(x)^{-1} \int_x^\infty \mathfrak{g}(u)du \asymp x \log x.$$

Also, $\mathfrak{g}/\phi$ is strictly increasing from $(\mathfrak{g}/\phi)(0) < 1$ to $+\infty$ as $x \to \infty$.

The monotonicity property in Lemma 22 enables one to define a pseudo-threshold, still denoted ζ, from the function $\mathfrak{B} = (\mathfrak{g}/\phi) - 1$ as $\zeta(\alpha) = \mathfrak{B}^{-1}(\alpha^{-1})$.

The posterior median, by the same proof as Lemma 2 in Johnstone and Silverman (2004), is a threshold rule: there exists $t(\alpha) > 0$ such that the posterior median on coordinate i is 0 if and only if $|X_i| \leq t(\alpha)$. Also, by the same proof as in Johnstone and Silverman (2004), one has $t(\alpha)^2 < \zeta(\alpha)^2$ and $\phi(t) < C\phi(\zeta)$.

Now turning to the moments of the score function, let us denote

$$\mathfrak{B}(x, \alpha) = \frac{\mathfrak{B}(x)}{1 + \alpha\mathfrak{B}(x)},$$

and similarly as for $\tilde{m}, m_1, m_2$, let us set

$$\tilde{\mathfrak{m}}(\alpha) = -E_0\mathfrak{B}(X, \alpha), \quad \mathfrak{m}_1(\mu, \alpha) = E_\mu\mathfrak{B}(X, \alpha), \quad \mathfrak{m}_2(\mu, \alpha) = E_\mu\mathfrak{B}(X, \alpha)^2.$$

Lemma 23. The function $\alpha \to \tilde{\mathfrak{m}}(\alpha)$ is nonnegative and increasing in α. As $\alpha \to 0$,

$$\tilde{\mathfrak{m}}(\alpha) \asymp (\log \zeta)^{-1}. \tag{3.3.3}$$

$$\mathfrak{m}_1(\mu,\alpha) \leq \begin{cases} -\tilde{\mathfrak{m}}(\alpha) + C\left(\frac{\zeta(\alpha)}{\log\zeta(\alpha)}\right)^2\mu^2, & \text{for } |\mu| < 1/\zeta(\alpha), \\ C\frac{\phi(\zeta/2)}{\alpha}, & \text{for } |\mu| < \zeta(\alpha)/2 \\ (\alpha\wedge c)^{-1}, & \text{for all } \mu. \end{cases}$$

$$\mathfrak{m}_2(\mu,\alpha) \leq \begin{cases} \frac{C}{\log\zeta(\alpha)}\frac{1}{\zeta(\alpha)^2}\frac{\tilde{\mathfrak{m}}(\alpha)}{\alpha}, & \text{for } |\mu| < 1/\zeta(\alpha), \\ \frac{C}{\zeta}\frac{\phi(\zeta/2)}{\alpha^2}, & \text{for } |\mu| < \zeta(\alpha)/2 \\ (\alpha\wedge c)^{-1}, & \text{for all } \mu. \end{cases}$$

Proof of Lemma 22. The derivative $(\log\gamma)'(u)$ is bounded in absolute value for $|u| \geq 1$ by a universal constant, and γ is unimodal and symmetric, so $\mathfrak{g} \asymp \gamma$ and the monotonicity of $\mathfrak{g}/\phi$ follow from the proof of Lemma 1 in Johnstone and Silverman (2004). The second estimate is immediate using the first one and the fact that $\int_y^\infty (u\log^2 u)^{-1}du = (\log y)^{-1}$. □

Proof of Lemma 23. Using that $\int \mathfrak{B}\phi = 0$, one can rewrite, as in Johnstone and Silverman (2004), $\tilde{\mathfrak{m}}$ as

$$\tilde{\mathfrak{m}}(\alpha) = 2\int_0^\infty \frac{\alpha\mathfrak{B}(z)^2}{1+\alpha\mathfrak{B}(z)}\phi(z)dz$$

from which it follows that, separating into $z \leq \zeta$ anf $z > \zeta$,

$$\tilde{\mathfrak{m}}(\alpha) \asymp \int_0^\zeta \alpha\mathfrak{B}(z)^2\phi(z)dz + \int_\zeta^\infty \mathfrak{B}(z)\phi(z)dz.$$

The first term is dealt with using the estimate of Corollary 1 of Johnstone and Silverman (2004), valid for any γ log-Lipshitz on $\mathbb{R}$ which is the case here, which leads to, for ζ large enough, or equivalently α small enough,

$$\alpha\int_0^\zeta \mathfrak{B}(z)^2\phi(z)dz \leq \alpha\frac{C}{\zeta}\frac{\mathfrak{g}(\zeta)^2}{\phi(\zeta)} \lesssim C\frac{\mathfrak{g}(\zeta)}{\zeta}.$$

For the second term, noting that $\mathfrak{B}\phi \sim \mathfrak{g}$ and using Lemma 22 for the tail bound on (minus) the primitive of $\mathfrak{g}$, one gets that this term is asymptotic to $\mathfrak{g}(\zeta)\zeta\log\zeta \asymp \log^{-1}\zeta$ and always dominates the first term. This proves the claim on $\tilde{\mathfrak{m}}$.

Now turning to $\mathfrak{m}_1$ and $\mathfrak{m}_2$, note that the global bounds directly follow from the fact that $|\mathfrak{B}(x,\alpha)| \leq C \vee \alpha^{-1}$. The intermediate bounds are derived as in Johnstone and Silverman (2004), since the proofs involve only the log-Lipschitz property of γ.

For small signals $|\mu| \leq 1/\zeta$ and the first moment, one proceeds as in Johnstone and Silverman (2004) by (Taylor-) expanding the function $\mu \to \mathfrak{m}_1(\mu,\alpha)$ at the order 2 around $\mu = 0$. The first derivative in μ is 0, since the function is symmetric. The following bound on the second derivative is as in Johnstone and Silverman (2004): on $[-\zeta,\zeta]$ one bounds $\phi''(u)$ by $C(1+u^2)\phi(u)$ and uses $\phi(z-\mu) \leq C\phi(z)$ thanks to the fact that $|\mu| \leq \zeta^{-1}$,

$$\begin{aligned}\left|\frac{\partial^2}{\partial\mu^2}\mathfrak{m}_1(\mu,\alpha)\right| &\leq \int_{-\infty}^{\infty} |\mathfrak{B}(z,\alpha)\phi''(z-\mu)|dz \\ &\leq \int_{-\zeta}^{\zeta} |\mathfrak{B}(z)|(1+z^2)\phi(z)dz + \frac{2}{\alpha}\int_{|z|>\zeta} \phi''(z-\mu)dz \\ &\leq \int_{-\zeta}^{\zeta} \mathfrak{g}(z)(1+z^2)dz + C\mathfrak{B}(\zeta)\phi(\zeta) = (i)+(ii).\end{aligned}$$

The term (ii) is bounded by a constant times $\zeta\mathfrak{g}(\zeta) \leq C(\log\zeta)^{-2}$. The integral defining (i) can be separated in $|z| \leq 2$, for which it is bounded by a constant, and $|z| > 2$, part on which one integrates by part to obtain

$$\int_2^{\zeta} \mathfrak{g}(z)z^2dz \leq C\int_2^{\zeta} \frac{z}{\log^2 z}dz = \frac{\zeta^2}{2\log^2\zeta} - \frac{4}{2\log^2 2} - \int_2^{\zeta}\frac{z}{2\log^3 z}dz \leq \frac{\zeta^2}{2\log^2\zeta}.$$

One concludes that the term (i) dominates in the expression of the second derivative, and the bound for $\mathfrak{m}_1$ follows by a Taylor expansion.

The bound for the second moment is obtained by separating again $|z| \leq \zeta$ and $|z| > \zeta$ to obtain

$$\begin{aligned}\mathfrak{m}_2(\mu,\alpha) &\leq C\int_{-\zeta}^{\zeta} \mathfrak{B}(z)^2\phi(z-\mu)dz + \frac{1}{\alpha^2}\int_{|z|>\zeta}\phi(z-\mu)dz \\ &\leq C\int_0^{\zeta} \mathfrak{B}(z)^2\phi(z)dz + \frac{2}{\alpha^2}\frac{\phi(z-\mu)}{z-\mu} \\ &\leq C\frac{\mathfrak{g}(\zeta)}{\zeta}\frac{1}{\alpha} + C\frac{\mathfrak{g}(\zeta)}{\zeta}\frac{1}{\alpha}.\end{aligned}$$

By using the estimate $\mathfrak{g}(\zeta) \asymp \tilde{\mathfrak{m}}(\alpha)/(\zeta\log\zeta)$ which follows from the estimate on $\tilde{\mathfrak{m}}$, the bound on $\mathfrak{m}_2$ for small μ follows. □

Let us now state a simplified version of Lemma 26 of Castillo and Roquain (2018). Note that the authors introduce a quantity $T_\mu(\alpha)$ not appearing here since $T_\mu(\alpha) \geq 1$ which will be sufficient in what follows. Another proof of Lemma 24 can be obtained using a slightly different approach, see Lemma 18 of Castillo and Szabo (2018).

Lemma 24. Let $\bar{\Phi}(t) = \int_t^\infty \phi(u)du$. There exist $M_0 > 0$ and $a_0 \in (0,1)$ such that $\forall \mu \geq M_0$ and $\forall \alpha \leq a_0$

$$\mathfrak{m}_1(\mu,\alpha) \geq \frac{1}{4}\mathfrak{B}(\zeta)\bar{\Phi}(\zeta-\mu)$$

.

Proof. We follow the proof of Lemma 26 in Castillo and Roquain (2018) that stays valid for our special Slab (3.1.6) as it only needs $\mathfrak{g}$ to be decreasing and $(\log \mathfrak{g})'$ to be bounded which is stated in (3.1.7). In the page 59 of their paper, one has the following inequalities

$$\forall \alpha \leq a_0, \forall M_0 \leq \mu \leq \zeta - 1,\ \mathfrak{m}_1(\mu,\alpha) \geq \frac{1}{2}\mathfrak{B}(\zeta)\bar{\Phi}(\zeta-\mu) + C\mathfrak{B}(\zeta)\frac{\phi(\zeta-\mu)}{\mu}$$

and

$$\forall \alpha \leq a_0, \forall \mu \geq \zeta - 1,\ \mathfrak{m}_1(\mu,\alpha) \geq \frac{1}{4}\mathfrak{B}(\zeta)\bar{\Phi}(\zeta-\mu)$$

which leads to the result. □

We will also need the following result

Lemma 25. Let M_0 and a_0 be the constants appearing in Lemma 24. For every $\mu \geq M_0$ and $\alpha \leq a_0$,

$$\mathfrak{m}_2(\mu,\alpha) \lesssim \frac{\mathfrak{m}_1(\mu,\alpha)}{\alpha}.$$

Proof. We have $E[|\mathfrak{B}(\mu+Z,\alpha)|] = \mathfrak{m}_1(\mu,\alpha) + E[|\mathfrak{B}(\mu+Z,\alpha)| - \mathfrak{B}(\mu+Z,\alpha)]$ with $Z \sim \mathcal{N}(0,1)$.
We first use Lemma 24 to show that, for $\mu \geq M_0$ and $\alpha \leq a_0$, $\mathfrak{m}_1(\mu,\alpha) \gtrsim 1$.

For $M_0 \leq \mu < \zeta$, we have, for small enough α

$$\begin{aligned}
\mathfrak{m}_1(\mu,\alpha) &\gtrsim \mathfrak{B}(\zeta)\bar{\Phi}(\zeta-\mu) \gtrsim \frac{\mathfrak{g}(\zeta)}{\phi(\zeta)}\left(\frac{1}{\zeta}-\frac{1}{\zeta^3}\right)\phi(\zeta-\mu) \\
&\gtrsim \mathfrak{g}(\zeta)\frac{\zeta^2-1}{\zeta^3}e^{-\frac{\mu^2}{2}+\mu\zeta} \gtrsim \frac{\mathfrak{g}(\zeta)}{\zeta^3}e^{-\frac{\mu\zeta}{2}+\mu\zeta} \\
&\gtrsim \frac{1}{\zeta^4\log^2(\zeta)}e^{\frac{\mu\zeta}{2}} \gtrsim 1
\end{aligned}$$

For $\mu > \zeta$, we have, for small enough α

$$\begin{aligned}\mathfrak{m}_1(\mu,\alpha) &\gtrsim \mathfrak{B}(\zeta)\bar{\Phi}(\zeta-\mu)\\ &\gtrsim \frac{\mathfrak{g}(\zeta)}{\phi(\zeta)}\left(\frac{1}{2}+\int_0^{\mu-\zeta}\phi(t)dt\right)\\ &\gtrsim \frac{1}{2}\frac{\mathfrak{g}(\zeta)}{\phi(\zeta)} \gtrsim 1\end{aligned}$$

Using the fact that, as noted in (88) of Johnstone and Silverman (2004), $|\mathfrak{B}(x,\alpha)| \leq c = \mathfrak{B}(0)/(1+\mathfrak{B}(0))$ if $\mathfrak{B}(x) < 0$ and $|\mathfrak{B}(x,\alpha)| \leq 1/\alpha$ if $\mathfrak{B}(x) \geq 0$, we have, for $\mu \geq M_0$ and $\alpha \leq a_0$, as $\mathfrak{m}_1(\mu,\alpha) \gtrsim 1$,

$$E[|\mathfrak{B}(\mu+Z,\alpha)|] \leq \mathfrak{m}_1(\mu,\alpha) + 2c \lesssim \mathfrak{m}_1(\mu,\alpha).$$

We also have, for α small enough, $|\mathfrak{B}(\mu+Z,\alpha)| \leq \frac{1}{\alpha}$.
Hence $\mathfrak{m}_2(\theta,\alpha) \leq E[|\mathfrak{B}(\mu+Z,\alpha)|^2] \leq \frac{1}{\alpha}E[|\mathfrak{B}(\mu+Z,\alpha)|] \lesssim \frac{\mathfrak{m}_1(\mu,\alpha)}{\alpha}$ for α small enough. □

3.3.3 Bounds for posterior moments and fixed α

Here we study $r_2(\alpha,\mu,x) := \int(u-\mu)^2 d\pi_\alpha(u\,|\,x)$, where for x real we denote in slight abuse of notation

$$\pi_\alpha(\cdot\,|\,x) \sim (1-a_\alpha(x))\delta_0 + a_\alpha(x)\gamma_x(\cdot).$$

For any real μ and $\alpha \in [0,1]$, by definition

$$r_2(\alpha,\mu,x) = (1-a_\alpha(x))\mu^2 + a_\alpha(x)\int(u-\mu)^2\gamma_x(u)du. \tag{3.3.4}$$

We first need to study the integral $\int(u-\mu)^2\gamma_x(u)du$. The following Lemma is a general result on densities that one could check for our special Slab γ but also for Laplace, Cauchy or other classical choices of Slab densities.

Lemma 26. Let γ be any density on $\mathbb{R}$ such that there exist positive constants c_1 and c_2 such that $|(\log\gamma)'| \leq c_1$ and $|(\log\gamma)''| \leq c_2$

$$\int(u-\mu)^2\gamma_x(u)du = (x-\mu)^2 + \frac{g''}{g}(x) + 1 + 2(x-\mu)\frac{g'}{g}(x). \tag{3.3.5}$$

which leads to the following upper bound

$$\int (u-\mu)^2 \gamma_x(u)du \leq (x-\mu)^2 + c_2 + 2c_1|x-\mu| \leq 2(x-\mu)^2 + c_3. \tag{3.3.6}$$

Proof. Since $\gamma_x(\cdot)$ is a density function, for any x,

$$\int \gamma_x(u)du = 1.$$

Noting that for any x, $\int u\gamma_x(u)du = x + (\log g)'(x)$ and that this quantity is in absolute value less than $|x|$ *(check, cf before)*, one obtains

$$|\int u\gamma_x(u)du| \leq |x|.$$

Decomposing $u^2 = (u-x)^2 + 2x(u-x) + x^2$ and noting that $\int (u-x)^2\gamma_x(u)du = g''(x)/g(x) + 1$,

$$\int u^2\gamma_x(u)du = \frac{g''}{g}(x) + 1 + 2x\frac{g'}{g}(x) + x^2.$$

Using that $|\gamma'| \leq c_1\gamma$ and $|\gamma''| \leq c_2\gamma$, this leads to

$$|g'(x)| \leq \int |\gamma'(x-u)|\phi(u)du \leq c_1 \int \gamma(x-u)\phi(u)du = c_1 g(x)$$

and similarly $|g''| \leq c_2 g$, so that $\int u^2\gamma_x(u)du \leq C(1+x^2)$.

Similarly, for any real μ,

$$\int (u-\mu)^2\gamma_x(u)du = (x-\mu)^2 + \frac{g''}{g}(x) + 1 + 2(x-\mu)\frac{g'}{g}(x). \tag{3.3.7}$$

□

Bounds for zero signal.

Lemma 27. Let γ be as in (3.1.6) and let $r_2(\alpha, \mu, x)$ be as in (3.3.4).

$$E_0 r_2(\alpha, 0, x) \lesssim \alpha\tau(\alpha)^2\{1 + \log(1+\tau(\alpha))\}^{-2}$$

Proof. Suppose for now that $x \sim \mathcal{N}(0,1)$. One has

$$r_2(\alpha, 0, x) = a_\alpha(x) \int u^2 \gamma_x(u) du.$$

Using the simple bound (3.3.1) for $a_\alpha(x)$, this implies for all x

$$r_2(\alpha, 0, x) \lesssim \left[1 \wedge \frac{\alpha}{1-\alpha} \frac{g}{\phi}(x)\right] (1 + x^2). \tag{3.3.8}$$

By taking the expectation and noticing that $\tau(\alpha)$ is the number that makes both sides of the infimum of the last display equal, one obtains

$$E_0 r_2(\alpha, 0, x) \lesssim \int \mathbb{1}_{|x|>\tau(\alpha)}(1 + x^2)\phi(x)dx + \int \mathbb{1}_{|x|\leq\tau(\alpha)} \frac{\alpha}{1-\alpha} \frac{g}{\phi}(x)\phi(x)(1 + x^2)dx.$$

As $\int_M^\infty x^2\phi(x)dx \leq CM\phi(M)$, the first term of the last display is bounded by

$$\phi(\tau(\alpha))/\tau(\alpha) + \tau(\alpha)\phi(\tau(\alpha)).$$

For γ as in (3.1.6), as $\mathfrak{g} \asymp \gamma$ by Lemma 22, the second term of the last display is bounded by a constant times

$$\begin{aligned} \alpha \int_{-\tau(\alpha)}^{\tau(\alpha)} \gamma(x)(1 + x^2)dx &= \alpha \int_0^{\tau(\alpha)} \frac{1 + x^2}{(1 + x)(1 + \log(1 + x))^2} dx \\ &\leq \alpha \int_1^{1+\tau(\alpha)} \frac{y^2}{y(1 + \log y)^2} dy, \end{aligned}$$

setting $y = 1 + x$ and using that $1 + x^2 \leq y^2$ for $y \geq 1$. An integration by parts gives

$$\int_1^{1+\tau(\alpha)} \frac{y}{(1 + \log y)^2} dy = \left[\frac{y^2}{2(1 + \log y)^2}\right]_1^{1+\tau(\alpha)} + \int_1^{1+\tau(\alpha)} \frac{y}{(1 + \log y)^3} dy.$$

The map $y \to y/(1 + \log y)^3$ is increasing for $y \geq e^2$. This implies, for $C > 0$ a universal constant,

$$\int_1^{1+\tau(\alpha)} \frac{y}{(1 + \log y)^3} dy \leq C + \tau(\alpha) \frac{1 + \tau(\alpha)}{\{1 + \log(1 + \tau(\alpha))\}^3}.$$

□

Bounds for nonzero signal. The definition of $r_2(\alpha, \mu, x)$ and the moment bound (3.3.6) lead to

$$r_2(\alpha, \mu, x) \leq (1 - a_\alpha(x))\mu^2 + a_\alpha(x)(2(x-\mu)^2 + C). \tag{3.3.9}$$

The weight $1 - a_\alpha(x)$ is bounded using (3.3.2), so that

$$r_2(\alpha, \mu, x) \leq \mu^2 \left[\mathbb{1}_{|x| \leq \tau(\alpha)} + e^{-\frac{1}{2}(|x|-\tau(\alpha))^2} \mathbb{1}_{|x|>\tau(\alpha)} \right] + 2(x-\mu)^2 + C. \tag{3.3.10}$$

Let us prove the following lemma

Lemma 28. Let a_0 be the solution of the equation $\tau(\alpha) = 20$. There exist $C_1, C_2 > 0$ such that for any $0 < \alpha \leq a_0$ (and alpha small enough for $\tilde{\tau} = \tau$) and any real μ,

$$E_\mu r_2(\alpha, \mu, x) \leq (\mu^2 + C_1)\mathbb{1}_{|\mu| \leq \tau(\alpha) + \sqrt{\tau(\alpha)}} + C_2 \mathbb{1}_{|\mu| > \tau(\alpha) + \sqrt{\tau(\alpha)}}.$$

In particular, there exists $C_3 > 0$ such that for any $0 < \alpha \leq a_0$ and any real μ,

$$E_\mu r_2(\alpha, \mu, x) \leq \tau(\alpha)^2 \left[1 + \tau(\alpha)^{-1/2} \right] \mathbb{1}_{|\mu| \leq \tau(\alpha) + \sqrt{\tau(\alpha)}} + C_3.$$

Proof. Let us set as shorthand notation for the proof

$$T := \tau(\alpha), \qquad \epsilon = T^{-1/2}, \qquad \delta = \frac{\epsilon}{2(1+\epsilon)}.$$

The condition on α implies $\epsilon < 1/4$. Let us distinguish two cases. First, if $|\mu| \leq (1+\epsilon)T$, one simply bounds the first term on the right hand side of (3.3.9) by μ^2 and the expectation of the second one by a constant, which leads to the first term displayed in the upper-bound of the lemma.

In the case that $|\mu| > (1+\epsilon)T$, one uses (3.3.10) to get

$$E_\mu r_2(\alpha, \mu, x) \leq \mu^2 \left(P_\mu[|x| \leq T] + E_\mu[e^{-\frac{1}{2}(|x|-T)^2}] \right) + C.$$

Under the present assumption on μ, note that $|\mu| - T = 2\delta|\mu| + (1-2\delta)|\mu| - T > 2\delta|\mu|$. The triangle inequality implies

$$P_\mu[|x| \leq T] \leq P_\mu[|\varepsilon| \geq |\mu| - T] = 2\bar{\Phi}\left(2\delta|\mu|\right).$$

Let us consider the interval $A = [\mu - \delta|\mu|, \mu + \delta|\mu|]$. One can split

$$E_\mu[e^{-\frac{(|x|-T)^2}{2}}] = \int e^{-\frac{(|x|-T)^2}{2}} e^{-\frac{(x-\mu)^2}{2}} \frac{dx}{\sqrt{2\pi}} \\ \leq \int_{A^c} e^{-\frac{(x-\mu)^2}{2}} \frac{dx}{\sqrt{2\pi}} + \int_A e^{-\frac{(|x|-T)^2}{2}} \frac{dx}{\sqrt{2\pi}}.$$

By definition of A, the first term of the last bound is $2\bar{\Phi}(\delta|\mu|)$. Moreover, on A, we have $|x| \geq (1-\delta)|\mu|$, so $|x| - T \geq (1 - \delta - \frac{1}{1+\epsilon})|\mu| = \delta|\mu|$. This leads to

$$\int_A e^{-\frac{(|x|-T)^2}{2}} \frac{dx}{\sqrt{2\pi}} \leq 2\delta|\mu| e^{-\delta^2\mu^2/2} \leq Ce^{-\delta^2\mu^2/4},$$

where one uses $xe^{-x^2/2} \leq Ce^{-x^2/4}$ for $x \geq 0$. Putting the previous bounds together implies

$$E_\mu r_2(\alpha, \mu, x) \leq C\mu^2 \left[\bar{\Phi}(\delta|\mu|) + \phi(\delta\mu/2)\right] + C \leq C\mu^2\phi(\delta\mu/2) + C \\ \leq C\delta^{-2}\phi(\delta\mu/4) + C \leq CTe^{-dT^{1/2}} + C,$$

where we have used that $ve^{-v} \leq 2e^{-v/2}$ for $v \geq 0$ and where $d = 1/256$. The last bound in the previous display is bounded by a universal constant, which leads to the second term displayed in the upper-bound of the lemma. □

This finally leads to the bound, for $\alpha = o(1)$ and any μ

$$E_\mu[r_2(\alpha, \mu, x)] \leq \tau(\alpha)^2(1 + o(1)) \tag{3.3.11}$$

3.3.4 Risk bound for fixed α: proof of Proposition 5

Proof of Proposition 5. Using the facts that $\tau^2(\alpha) \leq \zeta^2(\alpha) \leq 2\log(1/\alpha) + o(\log(1/\alpha))$ (use for example Lemma 14 of Castillo and Roquain (2018)) and that, for α small enough, there exist $c_1 > 1$ and $c_2 > c_1$ two constants such that $2c_1 \log(1/\alpha) \leq \tau(\alpha) \leq 2c_2 \log(1/\alpha)$, we have, for α small enough

$$\begin{aligned} E_{\theta_0} \int \|\theta - \theta_0\|^2 d\Pi_\alpha(\theta \,|\, X) &\leq Cn\alpha\tau(\alpha)^2/\log^2(\tau(\alpha)) + s_n\tau(\alpha)^2(1 + o(1)) \\ &\leq 2c_1 Cn\alpha \log(1/\alpha)/\log^2(2c_2\log(1/\alpha)) + 2s_n \log(1/\alpha)(1 + o(1)) \\ &\leq 2c_1 Cn\alpha \log(1/\alpha)/\log^2(2c_2\log(1/\alpha)) + 2s_n \log(n/s_n)(1 + o(1)) \end{aligned}$$

The first term, as it involves an increasing function of α (as here α is small enough), is bounded by $2c_1Cs_n \log\left((n/s_n)\log(\log n/s_n)\right)/\log^2(2c_2\log(n/s_n\log(\log n/s_n)))$, which is bounded by

$$2c_1Cs_n \log\left((n/s_n)\log(\log n/s_n)\right)/\log^2(\log(n^{c_2})) = o(s_n\log(n/s_n))$$

which completes the proof. □

3.3.5 Random α bounds

For convenience we work with the 'threshold' $\tau(\alpha)$ (as we take $\alpha \leq \alpha_0$ such that $\tilde{\tau}(\alpha) = \tau(\alpha)$), although other choices $t(\alpha), \zeta(\alpha)$ should be essentially equivalent.

Lemma 29 (no signal or small signal). Let α be a fixed non-random element of $(0,1)$. Let $\hat{\alpha}$ be a random element of $[0,1]$ (chosen small enough so that $\tilde{\tau}(\alpha) = \tau(\alpha)$) that may depend on $x \sim \mathcal{N}(0,1)$ and on other data. Then there exists $C_1 > 0$ such that

$$Er_2(\hat{\alpha}, 0, x) \leq C_1\left[\alpha\tau(\alpha)^2(1+\log(1+\tau(\alpha))^{-2} + P(\hat{\alpha} > \alpha)^{1/2}\right].$$

There exists $C_2 > 0$ such that for any real μ, if $x \sim \mathcal{N}(\mu, 1)$,

$$Er_2(\hat{\alpha}, \mu, x) \leq \mu^2 + C_2.$$

Proof. Using the bound (3.3.8) on $r_2(\alpha, 0, x)$,

$$\begin{aligned} r_2(\hat{\alpha}, 0, x) &= r_2(\hat{\alpha}, 0, x)\mathbb{1}_{\hat{\alpha}\leq\alpha} + r_2(\hat{\alpha}, 0, x)\mathbb{1}_{\hat{\alpha}>\alpha} \\ &\lesssim \left[\frac{\hat{\alpha}}{1-\hat{\alpha}}\frac{g}{\phi}(x) \wedge 1\right](1+x^2)\mathbb{1}_{\hat{\alpha}\leq\alpha} + (1+x^2)\mathbb{1}_{\hat{\alpha}>\alpha} \\ &\lesssim \left[\frac{\alpha}{1-\alpha}\frac{g}{\phi}(x) \wedge 1\right](1+x^2)\mathbb{1}_{\hat{\alpha}\leq\alpha} + (1+x^2)\mathbb{1}_{\hat{\alpha}>\alpha}. \end{aligned}$$

For the first term in the last display, one bounds the indicator from above by 1 and proceeds as in the proof of (3.2.2) to bound its expectation by $C\tau(\alpha)^2(1+\log(1+\tau(\alpha))^{-2}$. The first part of the lemma follows by noting that $E[(1+x^2)\mathbb{1}_{\hat{\alpha}>\alpha}]$ is bounded from above by $(2+2E_0[x^4])^{1/2}P(\hat{\alpha}>\alpha)^{1/2} \leq C_1P(\hat{\alpha}>\alpha)^{1/2}$ by Cauchy-Schwarz inequality. The second part of the lemma follows from the fact that using (3.3.9), $r_2(\alpha,\mu,x) \leq (1-a_\alpha(x))\mu^2 + Ca_\alpha(x)((x-\mu)^2+1) \leq \mu^2 + C(x-\mu)^2 + C$ for any α.

□

Lemma 30 (signal). Let α be a fixed non-random element of $(0,1)$ (chosen small enough so that $\tilde{\tau}(\alpha) = \tau(\alpha)$). Let $\hat{\alpha}$ be a random element of $[0,1]$ that may depend on $x \sim \mathcal{N}(\mu, 1)$ and on other data and such that $\tau(\hat{\alpha})^2 \leq d \log(n)$ with probability 1 for some $d > 0$. Then there exists $C_2 > 0$ such that for all real μ,

$$Er_2(\hat{\alpha}, \mu, x) \leq \tau(\alpha)^2(1 + o(1)) + C_2(1 + d \log n)P(\hat{\alpha} < \alpha)^{1/2}.$$

Proof. Combining (3.3.2) and (3.3.9),

$$r_2(\hat{\alpha}, \mu, x) \leq \mu^2 \left[\mathbb{1}_{|x| \leq \tau(\hat{\alpha})} + e^{-\frac{1}{2}(|x| - \tau(\hat{\alpha}))^2} \mathbb{1}_{|x| > \tau(\hat{\alpha})} \right] + C((x - \mu)^2 + 1).$$

Note that it is enough to bound the first term on the right hand side in the last display, as the last one is bounded by a constant under E_μ. Let us distinguish the two cases $\hat{\alpha} \geq \alpha$ and $\hat{\alpha} < \alpha$.

In the case $\hat{\alpha} \geq \alpha$, as $\tau(\alpha)$ is a decreasing function of α,

$$\begin{aligned}
&\left[\mathbb{1}_{|x| \leq \tau(\hat{\alpha})} + e^{-\frac{1}{2}(|x| - \tau(\hat{\alpha}))^2} \mathbb{1}_{|x| > \tau(\hat{\alpha})} \right] \mathbb{1}_{\hat{\alpha} \geq \alpha} \\
&\leq \left[\mathbb{1}_{|x| \leq \tau(\hat{\alpha})} + \mathbb{1}_{\tau(\hat{\alpha}) < |x| \leq \tau(\alpha)} + e^{-\frac{1}{2}(|x| - \tau(\hat{\alpha}))^2} \mathbb{1}_{|x| > \tau(\alpha)} \right] \mathbb{1}_{\hat{\alpha} \geq \alpha} \\
&\leq \mathbb{1}_{|x| \leq \tau(\alpha)} + e^{-\frac{1}{2}(|x| - \tau(\alpha))^2} \mathbb{1}_{|x| > \tau(\alpha)},
\end{aligned}$$

where we have used $e^{-\frac{1}{2}v^2} \leq 1$ for any v and that $e^{-\frac{1}{2}(u-c)^2} \leq e^{-\frac{1}{2}(u-d)^2}$ if $u > d \geq c$. As a consequence, one can borrow the fixed α bound (3.3.11) obtained previously so that

$$E\left[r_2(\hat{\alpha}, \mu, x) 1_{\hat{\alpha} \geq \alpha}\right] \leq E_\mu \left[\mu^2 (\mathbb{1}_{|x| \leq \tau(\alpha)} + e^{-\frac{1}{2}(|x| - \tau(\alpha))^2} \mathbb{1}_{|x| > \tau(\alpha)}) + C((x - \mu)^2 + 1) \right] \leq \tau(\alpha)^2 (1 + o(1)).$$

In the case $\hat{\alpha} < \alpha$, setting $b_n = \sqrt{d \log n}$ and noting that $\tau(\hat{\alpha}) \leq b_n$ with probability 1 by assumption, proceeding as above, with b_n now replacing $\tilde{\tau}(\alpha)$, one can bound

$$\begin{aligned}
&\mathbb{1}_{|x| \leq \tau(\hat{\alpha})} + e^{-\frac{1}{2}(|x| - \tau(\hat{\alpha}))^2} \mathbb{1}_{|x| > \tau(\hat{\alpha})} \\
&\leq \mathbb{1}_{|x| \leq b_n} + e^{-\frac{1}{2}(|x| - b_n)^2} \mathbb{1}_{|x| > b_n}.
\end{aligned}$$

From this one deduces that

$$\begin{aligned}
&E\left(\mu^2 \left[\mathbb{1}_{|x| \leq \tau(\hat{\alpha})} + e^{-\frac{1}{2}(|x| - \tau(\hat{\alpha}))^2} \mathbb{1}_{|x| > \tau(\hat{\alpha})} \right] \mathbb{1}_{\hat{\alpha} < \alpha} \right) \\
&\leq C \left(E_\mu \left[\mu^4 \mathbb{1}_{|x| \leq b_n} + \mu^4 e^{-(|x| - b_n)^2} \right] \right)^{1/2} P(\hat{\alpha} < \alpha)^{1/2}.
\end{aligned}$$

Using similar bounds as in the fixed α case, one obtains

$$E_\mu \left[\mu^4 \mathbb{1}_{|x|\le b_n} + \mu^4 e^{-(|x|-b_n)^2} \right] \le C(1 + b_n^4).$$

Taking the square root and gathering the different bounds obtained concludes the proof. □

3.3.6 Undersmoothing

Let α_1 be defined as the solution in α of the equation,

$$d\alpha\tilde{\mathfrak{m}}(\alpha) = \tilde{\eta}_n, \tag{3.3.12}$$

where d is a constant to be chosen small enough, see below, and

$$\tilde{\eta}_n = (s_n \vee \log^2 n)/n.$$

Note that under (3.2.1), we have $\tilde{\eta} = s_n/n$. As in Johnstone and Silverman (2004) we note that $\alpha \to \alpha\tilde{\mathfrak{m}}(\alpha)$ is increasing in α for α small enough, and equals 0 at 0. So a solution of (3.3.12) exists. Also, provided η_n is small enough, it can be made smaller than any given arbitrary constant.
Let ζ_1 be defined via $\mathfrak{B}(\zeta_1) = 1/\alpha_1$ for α_1 as in (3.3.12). As stated in (3.3.3), $\tilde{\mathfrak{m}}(\alpha) \asymp (\log \zeta)^{-1}$, as $\alpha = o(1)$.
Note that in Johnstone and Silverman (2004), ζ_1 is defined as the solution in ζ of $\alpha\tilde{m}(\alpha) = \tilde{\eta}_n\zeta^\kappa$. The following result is an adaptation of Lemma 10 in Johnstone and Silverman (2004) to accommodate this different choice. As already noted in Chapter 2, it seems choosing $\alpha = \alpha_1$ as in (3.3.12) is necessary to obtain a sharp posterior integrated squared rate.

Lemma 31. There exist universal constants C and η_0 such that if $\eta \le \eta_0$ and $n/\log^2(n) \ge \eta_0^{-1}$, then

$$\sup_{\theta \in \ell_0[\eta]} P_\theta[\hat{\zeta} < \zeta_1] \le \exp(-Cn\tilde{\eta}_n).$$

Proof. Using $\log(1/\tilde{\eta}_n) \le \log(n) - 2\log\log n$, the bound on ζ from Lemma 32 gives that $\zeta_1^2 \le 2\log n - \frac{3}{2}\log\log n$, so that $t(\alpha_1) \le \zeta(\alpha_1) = \zeta_1 \le \sqrt{2\log n} = t(\alpha_n)$, so α_1 belongs to the interval $[\alpha_n, 1]$ over which the likelihood is maximised. For the rest of the proof let us denote $\alpha = \alpha_1$.

Then one notices that $\{\hat{\zeta} < \zeta_1\} = \{\hat{\alpha} > \alpha_1\} = \{S(\alpha_1) > 0\}$ as well as $\{\hat{\zeta} > \zeta_1\} = \{S(\alpha_1) < 0\}$: the sign of S at any particular w determines on which side of $\hat{\alpha}$ the given

α lies. So,

$$P_\theta[\hat{\zeta} < \zeta_1] = P_\theta[S(\alpha_1) > 0].$$

The score function equals $S(\alpha) = \sum_{i=1}^n \beta(X_i, \alpha)$, a sum of independent variables. By Bernstein's inequality, if W_i are centered independent variables with $|W_i| \leq M$ and $\sum_{i=1}^n \mathrm{Var}(W_i) \leq V$, then for any $A > 0$,

$$P\left[\sum_{i=1}^n W_i > A\right] \leq \exp\{-\frac{1}{2}A^2/(V + \frac{1}{3}MA)\}.$$

Set $W_i = \mathfrak{B}(X_i, \alpha) - \mathfrak{m}_1(\theta_{0,i}, \alpha)$ and $A = -\sum_{i=1}^n \mathfrak{m}_1(\theta_{0,i}, \alpha)$. Then one can take $M = c_3/\alpha$, using Lemma 7. One can bound $-A$ from above as follows

$$\begin{aligned}
-A &\leq -\sum_{i \notin S_0} \tilde{\mathfrak{m}}(\alpha) + \sum_{i \in S_0} \frac{c}{\alpha} \\
&\leq -(n - s_n)\tilde{\mathfrak{m}}(\alpha) + cs_n/\alpha \\
&\leq -n\tilde{\mathfrak{m}}(\alpha)/2 + cdn\tilde{\mathfrak{m}}(\alpha) \\
&\leq -n\tilde{\mathfrak{m}}(\alpha)/4,
\end{aligned}$$

and

$$\begin{aligned}
V(\alpha) &\leq \sum_{i \notin S_0} \mathfrak{m}_2(0, \alpha) + \sum_{i \in S_0} \mathfrak{m}_2(\theta_{0,i}, \alpha) \\
&\leq c_4(n - s_n)\frac{\tilde{\mathfrak{m}}(\alpha)}{\zeta^\kappa \alpha} + cs_n/\alpha^2 \\
&\leq \frac{C}{\alpha}\left[(n - s_n)\tilde{\mathfrak{m}}(\alpha)\zeta^{-\kappa} + cs_n/\alpha\right] \\
&\leq C\alpha^{-1}\left[n\tilde{\mathfrak{m}}(\alpha)\zeta^{-\kappa}/2 + cdn\tilde{\mathfrak{m}}(\alpha)\right] \\
&\leq C'dn\tilde{\mathfrak{m}}(\alpha)/\alpha, \qquad (3.3.13)
\end{aligned}$$

where one uses that ζ^{-1} is bounded *[in fact, goes to* 0 *if* $\eta_0 = o(1)$*]*. This leads to

$$\frac{V + \frac{1}{3}MA}{A^2} \leq \frac{C'd}{n\alpha\tilde{\mathfrak{m}}(\alpha)} + \frac{4c_3}{3n\alpha\tilde{\mathfrak{m}}(\alpha)} \leq \frac{c_5^{-1}}{n\alpha\tilde{\mathfrak{m}}(\alpha)}.$$

Deduce that

$$P\left[S(\alpha) > 0\right] \leq \exp\{-c_5 n\alpha\tilde{\mathfrak{m}}(\alpha)\}.$$

□

Lemma 32 (Basic bounds on $\zeta(\alpha), \tau(\alpha)$ and $t(\alpha)$). Let $\alpha = \alpha_1$ be defined by (3.3.12) for d a given constant and $\tilde{\eta}_n$ small enough, and let $\zeta(\alpha)$ be given by $\mathfrak{B}(\zeta(\alpha)) = \alpha^{-1}$. Then for some constants c_1, c_2,

$$\log(1/\tilde{\eta}_n) + c_1 \leq \frac{\zeta(\alpha)^2}{2} \leq \log(1/\tilde{\eta}_n) + \frac{1}{2}\log(1 + \log(1/\tilde{\eta}_n)) + c_2.$$

The same upper and lower bound hold (with possible different constants c_1 and c_2) for $\tau(\alpha)^2$ and $t(\alpha)$. In particular, $\zeta(\alpha)^2 \sim \tau(\alpha)^2 \sim t(\alpha)^2 \sim 2\log(1/\tilde{\eta}_n)$ as $\tilde{\eta}_n \to 0$.

Proof. But for small $\tilde{\eta}_n$, we have α small, or equivalently ζ large, so that $(\mathfrak{g}/\mathfrak{B})(\zeta) \sim \phi(\zeta)$. Now from the definition (3.3.12) of α combined with (3.3.3), one has

$$\tilde{\eta}_n \asymp d\alpha\zeta\log(\zeta)\frac{\mathfrak{g}(\zeta)}{\mathfrak{B}(\zeta)}\mathfrak{B}(\zeta) \asymp \zeta\log(\zeta)\phi(\zeta) \asymp \zeta\log(\zeta)e^{-\zeta^2/2}.$$

From this deduce that

$$|\log c + \log\zeta + \log(\log(\zeta)) - \frac{\zeta^2}{2} + \log(1/\tilde{\eta}_n)| \leq C.$$

$$|\log c + C'\log\zeta - \frac{\zeta^2}{2} + \log(1/\tilde{\eta}_n)| \leq C.$$

In particular, using $\log\zeta \leq a + \zeta^2/4$ for some constant $a > 0$ large enough, one gets $\zeta^2 \leq 4(C + \log(1/\tilde{\eta}_n))$. Inserting this back into the previous inequality leads to

$$\zeta^2/2 \leq \log(1/\tilde{\eta}_n) + C + \frac{1}{2}\log(1 + \log(1/\tilde{\eta}_n)).$$

To prove that the same statement holds for $\tau(\alpha)$ and $t(\alpha)$ note that following from the definition of $\tau(\alpha)$ and $\zeta(\alpha)$ we have $\zeta(\alpha/2) \leq \tau(\alpha) \leq \zeta(\alpha)$ and from page 1622 of Johnstone and Silverman (2004) we have that $\zeta(\alpha)^2 - c \leq t(\alpha)^2 \leq \zeta(\alpha)^2$.

□

3.3.7 Oversmoothing

Following Johnstone and Silverman (2004), Section 8.3, let us define

$$\tilde{\pi}(\tau;\mu) = \frac{1}{n}\#\{i : |\mu_i| \geq \tau\}. \tag{3.3.14}$$

We also set, recalling that α_0 is defined via $\tau(\alpha_0) = 1$,

$$\alpha(\tau,\pi) = \sup\{\alpha \le \alpha_0 :\ \pi\mathfrak{m}_1(\tau,\alpha) \ge 2\tilde{\mathfrak{m}}(\alpha)\}. \tag{3.3.15}$$

One also defines $\zeta_{\tau,\pi}$ as the corresponding pseudo-threshold $\mathfrak{B}^{-1}(\alpha(\tau,\pi)^{-1})$.

Lemma 33. There exists C and π_0 such that if $\pi < \pi_0$, then for all $\tau \ge 1$,

$$\sup_{\theta:\ \tilde{\pi}(\tau;\theta)\ge\pi} P_\theta[\hat{\zeta} > \zeta_{\tau,\pi}] \le \exp\{-Cn\phi(\zeta_{\tau,\pi})\}.$$

Proof. This is the same proof as in Johnstone and Silverman (2004), Lemma 11, where we use Lemma 23 to bound $\mathfrak{m}_2(0,\alpha)$ and Lemma 25 for $\mathfrak{m}_2(\tau,\alpha)$. □

3.3.8 Proof of Theorem 18

Let us decompose the risk $R_n(\theta_0) = E_{\theta_0}\int \|\theta-\theta_0\|^2 d\Pi_{\hat{\alpha}}(\theta\,|\,X)$ according to whether coordinates of θ correspond to a 'small' or 'large' signal, the threshold being $\zeta_1 = \beta^{-1}(\alpha_1^{-1})$, with α_1 defined in (3.3.12). One can write

$$R_n(\theta_0) = \Big[\sum_{i:\ \theta_{0,i}=0} + \sum_{i:\ 0<|\theta_{0,i}|\le\zeta_1} + \sum_{i:\ |\theta_{0,i}|>\zeta_1}\Big] E_{\theta_0}\int (\theta_i-\theta_{0,i})^2 d\Pi_{\hat{\alpha}}(\theta_i\,|\,X).$$

We next use the first part of Lemma 29 with $\alpha = \alpha_1$ to obtain, for any θ_0 in $\ell_0[s_n]$,

$$\begin{aligned}(I) &:= \sum_{i:\ \theta_{0,i}=0} E_{\theta_0}\int (\theta_i-\theta_{0,i})^2 d\Pi_{\hat{\alpha}}(\theta_i\,|\,X)\\ &\le C_1 \sum_{i:\ \theta_{0,i}=0}\Big[\alpha_1\tau(\alpha_1)^2(1+\log(1+\tau(\alpha_1))^{-2} + P_{\theta_0}(\hat{\alpha}>\alpha_1)\Big]\\ &\le C_1\Big[(n-s_n)\alpha_1\tau(\alpha_1)^2(1+\log(1+\tau(\alpha_1))^{-2} + (n-s_n)e^{-c_1\log^2 n}\Big],\end{aligned}$$

where for the last inequality we use Lemma 31 and (3.2.1). From (3.3.12) one gets, with $\eta_n = s_n/n$,

$$n\alpha_1 \lesssim n\eta_n\zeta_1^{-1}g(\zeta_1)^{-1} \lesssim s_n\log(\zeta_1).$$

Therefore, as in Proposition 5, the term (I) is a $o(s_n\log(n/s_n))$

For the 'intermediate' signal part, using the second part of the Lemma 29, we have

$$\begin{aligned}
&\sum_{i:\ 0<|\theta_{0,i}|\le\zeta_1} E_{\theta_0}\int(\theta_i-\theta_{0,i})^2 d\Pi_{\hat\alpha}(\theta_i\,|\,X)\\
&\le \sum_{i:\ 0<|\theta_{0,i}|\le\zeta_1}(\theta_{0,i}^2+C)\\
&\le (\zeta_1^2+C)\#\{i:\ 0<|\theta_{0,i}|\le\zeta_1\},
\end{aligned}$$

Now using Lemma 32 and the fact that $\tau(\alpha_1)\le\zeta_1$, one obtains that the contribution to the risk of the indices i with $0<|\theta_{0,i}|\le\zeta_1$ is bounded by

$$2\log(n/s_n)(1+o(1))\#\{i:\ 0<|\theta_{0,i}|\le\zeta_1\} \tag{3.3.16}$$

It remains to bound the part of the risk for indexes i with $|\theta_{0,i}|>\zeta_1$. To do so, one uses Lemma 30 with α chosen as $\alpha=\alpha_2:=\alpha(\zeta_1,\pi_1)$ and $\pi_1=\tilde\pi(\zeta_1;\theta_0)$ the proportion of components of the true signal above ζ_1, following the definitions (3.3.14)–(3.3.15). Also, $\tilde\tau(\alpha_2)=\tau(\alpha_2)$. One denotes by ζ_2 the pseudo-threshold associated to α_2.

Let us first compare α_1 and α_2. For small enough α, the bound on $\mathfrak{m}_1$ from Lemma 23 becomes $1/\alpha$, so that, using the definition (3.3.12) of α_1,

$$\frac{\mathfrak{m}_1(\zeta_1,\alpha_1)}{\tilde{\mathfrak{m}}(\alpha_1)}\le\frac{1}{\alpha_1}\left(\frac{\eta_n}{d\alpha_1}\right)^{-1}\le\frac{d}{\eta_n}\le\frac{d}{\pi_1},$$

using the rough bound $\pi_1\le\eta_n$. Note that both functions $\tilde{\mathfrak{m}}(\cdot)^{-1}$ and $\mathfrak{m}_1(\zeta_1,\cdot)$ are decreasing via Lemma 23, and so is their product on the interval where both functions are positive. As $d<2$, by definition of α_2 this means $\alpha_2<\alpha_1$ that is $\zeta_1<\zeta_2$.

One can now apply Lemma 30 with $\alpha=\alpha_2$ and use the fact that $\tau(\alpha_2)\le\zeta_2$

$$\begin{aligned}
&\sum_{i:\ |\theta_{0,i}|>\zeta_1} E_{\theta_0}\int(\theta_i-\theta_{0,i})^2 d\Pi_{\hat\alpha}(\theta_i\,|\,X)\\
&\le n\pi_1\left[\tau(\alpha_2)^2(1+o(1))+C_2(1+d\log n)P_{\theta_0}(\hat\alpha<\alpha_2)^{1/2}\right]\\
&\le n\pi_1\left[\zeta_2^2(1+o(1))+C_2(1+d\log n)P_{\theta_0}(\hat\zeta>\zeta_2)^{1/2}\right].
\end{aligned}$$

Let us verify that this term in the last display is bounded above by $n\pi_1\zeta_2^2(1+o(1))$. If $\zeta_2>\log n$, this is immediate by bounding $P_{\theta_0}(\hat\zeta>\zeta_2)$ by 1. If $\zeta_2\le\log n$, Lemma 33 implies $P_{\theta_0}(\hat\zeta>\zeta_2)\le\exp(-Cn\phi(\zeta_2))\le\exp(-C\sqrt{n})$, so this is also the case.

One now compares ζ_2 first to a certain $\zeta_3 = \zeta(\alpha_3)$ defined by α_3 (largest) solution of

$$\bar{\Phi}(\zeta(\alpha_3) - \zeta_1) = \frac{8}{\pi_1}\alpha_3 \tilde{\mathfrak{m}}(\alpha_3),$$

with $\bar{\Phi}(x) = P[\mathcal{N}(0,1) > x]$. Using Lemma 34, which also gives the existence of ζ_3, one gets

$$\frac{\mathfrak{m}_1(\zeta_1, \alpha_3)}{\tilde{\mathfrak{m}}(\alpha_3)} \geq \frac{\frac{1}{4}\mathfrak{B}(\zeta_3)\bar{\Phi}(\zeta_3 - \zeta_1)}{\tilde{\mathfrak{m}}(\alpha_3)} = \frac{1}{4\alpha_3}\frac{8\alpha_3\tilde{\mathfrak{m}}(\alpha_3)}{\pi_1\tilde{\mathfrak{m}}(\alpha_3)} = \frac{2}{\pi_1}.$$

This shows, reasoning as above, that $\alpha_3 \leq \alpha_2$, that is $\zeta_2 \leq \zeta_3$. Following Johnstone and Silverman (2004), one distinguishes two cases to further bound ζ_3.

If $\zeta_3 > \zeta_1 + 1$, using $\zeta_2^2 \leq \zeta_3^2$ and $\tilde{\mathfrak{m}}(\alpha_3) \lesssim \zeta_3 \mathfrak{g}(\zeta_3)$,

$$\begin{aligned}
\pi_1\zeta_2^2 &\leq \zeta_3^2 \frac{8\alpha_3\tilde{\mathfrak{m}}(\alpha_3)}{\bar{\Phi}(\zeta_3 - \zeta_1)} \lesssim \zeta_3^3 \frac{\mathfrak{g}(\zeta_3)}{\mathfrak{B}(\zeta_3)} \frac{\zeta_3 - \zeta_1}{\phi(\zeta_3 - \zeta_1)} \\
&\leq C\zeta_3^4 \frac{\phi(\zeta_3)}{\phi(\zeta_3 - \zeta_1)} = C\zeta_3^4 \phi(\zeta_1) e^{-(\zeta_3 - \zeta_1)\zeta_1} \\
&\leq C(\zeta_1 + 1)^4 e^{-\zeta_1}\phi(\zeta_1),
\end{aligned}$$

where for the last inequality we have used that $x \to x^4 e^{-(x-\zeta_1)\zeta_1}$ is decreasing for $x \geq \zeta_1 + 1$. Lemma 32 now implies that $\phi(\zeta_1) \lesssim \eta_n$. As ζ_1 goes to ∞ with n/s_n, one gets $\pi_1\zeta_2^2 \lesssim \eta_n$.

In this case, gathering the three different bounds leads us to

$$\begin{aligned}
R_n(\theta_0) &\leq o(s_n \log(n/s_n)) + 2\log(n/s_n)(1 + o(1))\#\{i : \ 0 < |\theta_{0,i}| \leq \zeta_1\} + s_n \\
&\leq o(s_n \log(n/s_n)) + 2s_n \log(n/s_n)(1 + o(1)) + s_n \\
&\leq 2s_n \log(n/s_n)(1 + o(1))
\end{aligned}$$

If $\zeta_1 \leq \zeta_3 \leq \zeta_1 + 1$, let $\zeta_4 = \zeta(\alpha_4)$ with α_4 solution in α of

$$\bar{\Phi}(1) = 8\alpha\tilde{\mathfrak{m}}(\alpha)\pi_1^{-1}.$$

By the definition of ζ_3, since $\bar{\Phi}(1) \leq \bar{\Phi}(\zeta_3 - \zeta_1)$, we have $8\alpha_4\tilde{\mathfrak{m}}(\alpha_4) \leq 8\alpha_3\tilde{\mathfrak{m}}(\alpha_3)$ so that $\alpha_4 \leq \alpha_3$, that is also $\zeta_3 \leq \zeta_4$. Using (3.3.3) as before,

$$\bar{\Phi}(1) \lesssim \frac{\mathfrak{g}(\zeta_4)}{\mathfrak{B}(\zeta_4)}\pi_1^{-1} \lesssim \phi(\zeta_4)\pi_1^{-1}.$$

Taking logarithms this leads to

$$\zeta_4^2 \leq C + 2\log(\pi_1^{-1}).$$

In particular, $\zeta_2^2 \leq 2\log(\pi_1^{-1}) + C$. It follows that

$$\begin{aligned} n\pi_1\zeta_2^2 &\leq 2n\pi_1\log(\pi_1^{-1}) + Cn\pi_1 = 2n\pi_1\log(n/s_n) + 2n\pi_1\log(s_n/(n\pi_1)) + Cn\pi_1 \\ &\leq 2n\pi_1\log(n/s_n) + 2s_n\log(n/s_n)\left[\tfrac{n\pi_1}{s_n}\log(s_n/(n\pi_1))(\log(n/s_n))^{-1} + C\tfrac{n\pi_1}{s_n}(\log(n/s_n))^{-1}\right] \\ &\leq 2n\pi_1\log(n/s_n) + 2s_n\log(n/s_n)\left[C'(\log(n/s_n))^{-1}\right], \end{aligned}$$

Where the last line uses that $n\pi_1 \leq s_n$ and that $u \in [1;+\infty] \mapsto \log(u)/u$ is bounded.

Now gathering the three different bounds and using the fact that $\#\{i:\ 0 < |\theta_{0,i}| \leq \zeta_1\} + n\pi_1 = \#\{i:\ 0 < |\theta_{0,i}| \leq \zeta_1\} + \#\{i:\ |\theta_{0,i}| \geq \zeta_1\} \leq s_n$ leads us to

$$\begin{aligned} R_n(\theta_0) &\leq o(s_n\log(n/s_n)) + 2\log(n/s_n)\#\{i:\ 0 < |\theta_{0,i}| \leq \zeta_1\} + 2n\pi_1\log(n/s_n) \\ &\leq 2s_n\log(n/s_n)(1+o(1)) \end{aligned}$$

which concludes the proof of Theorem 18.

Lemma 34. Let $\bar{\Phi}(t) = \int_t^\infty \phi(u)du$. For π_1, ζ_1 as above, a solution $0 < \alpha \leq \alpha_1$ to the equation

$$\bar{\Phi}(\zeta(\alpha) - \zeta_1) = 8\pi_1^{-1}\alpha\tilde{\mathfrak{m}}(\alpha). \tag{3.3.17}$$

exists. Let α_3 be the largest such solution. Then for c_0 in (3.2.1) small enough,

$$\mathfrak{m}_1(\zeta_1, \alpha_3) \geq \frac{1}{4}\mathfrak{B}(\zeta_3)\bar{\Phi}(\alpha_3 - \zeta_1). \tag{3.3.18}$$

Proof. First we check the existence of a solution. Set $\zeta_\alpha = \zeta(\alpha)$ and $R_\alpha := \bar{\Phi}(\zeta_\alpha - \zeta_1)/(\alpha\tilde{\mathfrak{m}}(\alpha))$. For $\alpha \to 0$ we have $\zeta_\alpha - \zeta_1 \to \infty$ so by using $\bar{\Phi}(u) \asymp \phi(u)/u$ as $u \to \infty$ one gets, treating terms depending on ζ_1 as constants and using $\phi(\zeta_\alpha) \asymp \alpha\mathfrak{g}(\zeta_\alpha)$,

$$\bar{\Phi}(\zeta_\alpha - \zeta_1) \asymp \frac{\phi(\zeta_\alpha - \zeta_1)}{\zeta_\alpha - \zeta_1} \asymp \alpha\mathfrak{g}(\zeta_\alpha)e^{\zeta_\alpha\zeta_1}.$$

As $\tilde{\mathfrak{m}}(\alpha) \asymp \zeta_\alpha\mathfrak{g}(\zeta_\alpha)$, one gets $R_\alpha \asymp e^{\zeta_\alpha\zeta_1}/\zeta_\alpha \to \infty$ as $\alpha \to 0$. On the other hand, with $\pi_1 \leq s_n/n$ and $\alpha_1\tilde{\mathfrak{m}}(\alpha_1) = ds_n/n$,

$$R_{\alpha_1} = \frac{1}{2\alpha_1\tilde{\mathfrak{m}}(\alpha_1)} = \frac{dn}{2s_n} \leq \frac{8}{\pi_1}\frac{d}{16},$$

so that $R_{\alpha_1} < 8/\pi_1$ as $d < 2$. This shows that the equation at stake has at least one solution for α in the interval $(0, \alpha_1)$.

Finally (3.3.18) is a direct consequence of Lemma 24. □

Chapter 4

Adaptive Pólya trees on densities using a Spike and Slab type prior

4.1 Introduction

The paper Castillo (2017b) showed that, for well chosen parameters, Pólya trees are able to model smooth functions and to induce posterior distributions with optimal convergence rates in the minimax sense for a range of Hölder regularities and also derived a Bernstein-von Mises theorem as well as a Donsker-type theorem, but the chosen parameters depend on the Hölder regularity of the true density. Here, as in Castillo (2017b), we will follow a multiscale approach to obtaining adaptive rates and limiting shape results, introduced in Castillo and Nickl (2013), Castillo and Nickl (2014), Castillo (2014) with connections to semiparametric functionals Castillo and Rousseau (2015).

4.1.1 Definition of a Pólya tree

Here we recall the construction of a standard Pólya tree.

First let us introduce some notation relative to dyadic partitions. For any fixed indexes $l \geq 0$ and $0 \leq k < 2^l$, the rational number $r = k2^{-l}$ can be written in a unique way as $\varepsilon(r) := \varepsilon_1(r) \dots \varepsilon_l(r)$, its finite expression of length l in base $1/2$ (note that it can end with one or more 0). That is, $\varepsilon_i \in \{0, 1\}$ and

$$k2^{-l} = \sum_{i=1}^{l} \varepsilon_i(r) 2^{-i}.$$

Let $\mathcal{E} := \bigcup_{l\geq 0}\{0,1\}^l \cup \{\emptyset\}$ be the set of finite binary sequences. We write $|\varepsilon| = l$ if $\varepsilon \in \{0;1\}^l$ and $|\emptyset| = 0$. We also use the notation $\varepsilon' = \varepsilon_1\varepsilon_2\ldots\varepsilon_{l-1}(1-\varepsilon_l)$.
Let us introduce a sequence of partitions $\mathcal{I} = \{(I_\varepsilon)_{|\varepsilon|=l}, l \geq 0\}$ of the unit interval. Here we will consider regular partitions, as defined below. This is mostly for simplicity of presentation, and other partitions, based for instance on quantiles of a given distribution, could be considered as well. Set $I_\emptyset = (0,1]$ and, for any $\varepsilon \in \mathcal{E}$ such that $\varepsilon = \varepsilon(l;k)$ is the expression in base $1/2$ of $k2^{-l}$, set

$$I_\varepsilon := \left(\frac{k}{2^l}, \frac{k+1}{2^l}\right] := I_k^l$$

For any $l \geq 0$, the collection of all such dyadic intervals is a partition of $(0,1]$. A random probability measure P follows a Pólya tree distribution $PT(\mathcal{A})$ with parameters $\mathcal{A} = \{\alpha_\varepsilon; \varepsilon \in \mathcal{E}\}$ on the sequence of partitions $\mathcal{I}$ if there exist random variables $0 \leq Y_\varepsilon \leq 1$ such that,
1. the variables $Y_{\varepsilon 0}$ for $\varepsilon \in \mathcal{E}$ are mutually independent and $Y_{\varepsilon 0}$ follows a $Beta(\alpha_{\varepsilon 0}, \alpha_{\varepsilon 1})$ distribution.
2. for any $\varepsilon \in \mathcal{E}$, we have $Y_{\varepsilon 1} = 1 - Y_{\varepsilon 0}$
3. for any $l \geq 0$ and $\varepsilon = \varepsilon_1\ldots\varepsilon_l \in \{0;1\}^l$, we have

$$P(I_\varepsilon) = \prod_{j=1}^{l} Y_{\varepsilon_1\ldots\varepsilon_j} \tag{4.1.1}$$

.

This construction can be visualised using a tree representation, see Figure 4.1 : to compute the random mass that P assigns to the subset I_ε of $[0,1]$, one follows a dyadic tree along the expression of ε : $\varepsilon_1; \varepsilon_1\varepsilon_2, \ldots, \varepsilon_1\varepsilon_2\ldots\varepsilon_l = \varepsilon$. The mass $P(I_\varepsilon)$ is a product of Beta variables whose parameters depend on whether one goes 'left' ($\varepsilon_j = 0$) or 'right' ($\varepsilon_j = 1$) along the tree :

$$P(I_\varepsilon) = \prod_{j=1,\varepsilon_j=0}^{l} Y_{\varepsilon_1,\ldots,\varepsilon_{j-1}0} \times \prod_{j=1,\varepsilon_j=1}^{l} (1 - Y_{\varepsilon_1,\ldots,\varepsilon_{j-1}0}) \tag{4.1.2}$$

This construction uniquely defines a random probability distribution on distributions on $[0,1]$. For more details we refer to Ferguson (1974) and Lavine (1992).
The corresponding object, the class of Pólya tree distributions, is quite flexible : different behaviours of the sequence of parameters $(\alpha_\varepsilon)_{\varepsilon\in\mathcal{E}}$ give a Pólya tree with different properties.

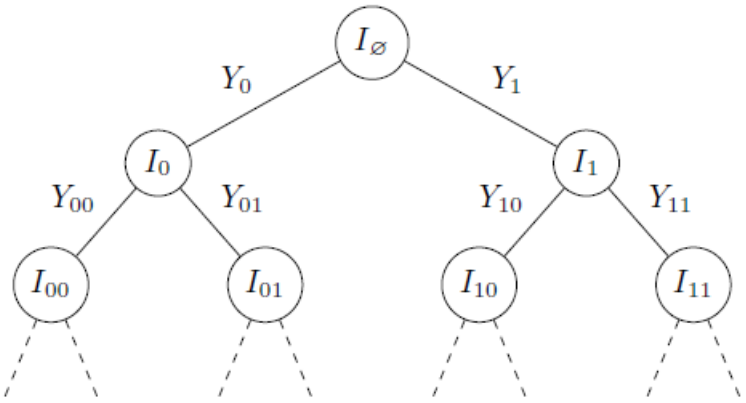

Fig. 4.1 Indexed binary tree with levels $l \leq 2$ represented. The nodes index the intervals I_ε. Edges are labelled with random variables Y_ε.

A standard assumption is that the parameters α_ε only depend on the depth $|\varepsilon|$, so that

$$\forall \varepsilon \in \mathcal{E}, \alpha_\varepsilon = a_l$$

for any $l \geq 1$ and a sequence $(a_l)_{l \geq 1}$ of positive numbers, which will be assumed henceforth.

Paths along the tree. A given $\varepsilon = \varepsilon_1, \ldots, \varepsilon_l \in \mathcal{E}$ gives rise to a path $\varepsilon_1 \to \varepsilon_1\varepsilon_2 \to \varepsilon_1\varepsilon_2 \ldots \varepsilon_l$. We denote $I_\varepsilon^{[i]} := I_{\varepsilon_1 \ldots \varepsilon_i}$, for any i in $\{1, \ldots, l\}$. Similarly, denote

$$Y_\varepsilon^{[i]} = Y_{\varepsilon_1 \ldots \varepsilon_i}$$

Conversely, any pair (l, k) with $l \geq 0$ and $k \in \{0, \ldots 2^l - 1\}$ is associated with a unique $\varepsilon = \varepsilon(l, k)$, the expression of length l in base $1/2$ of $k2^l$.

4.1.2 Function spaces and wavelets

We briefly introduce some standard notation appearing in the statements below.
Haar basis. The Haar wavelet basis is $\{\phi, \psi_{lk}, 0 \leq k < 2^l, l \geq 0\}$, where $\phi = \mathbb{1}_{[0,1]}$ and, for $\psi = -\mathbb{1}_{(0,1/2]} + \mathbb{1}_{(1/2,1]}$,

$$\psi_{lk}(\cdot) = 2^{l/2}\psi(2^l \cdot -k), 0 \leq k < 2^l, l \geq 0.$$

In this paper our interest is in density functions, that is nonnegative functions g with $\int_0^1 g\phi = \int_0^1 g = 1$, so that their first Haar-coefficient is always 1. So, we will only need to consider the basis functions ψ_{lk} and simply write informally (ψ_{lk}) for the Haar basis.

Function classes. Let $L^2 = L^2[0,1]$ denote the space of square-integrable functions on $[0,1]$ relative to Lebesgue measure equipped with the $\|\cdot\|_2$-norm. For $f, g \in L^2$, denote $\langle f, g\rangle := \langle f, g\rangle_2 = \int_0^1 fg$.
Let $L^\infty = L^\infty[0,1]$ denote the space of all measurable functions on $[0,1]$ that are bounded up to a set of Lebesgue measure 0, equipped with the (essential) supremum norm $\|\cdot\|_\infty$. The class $\mathcal{C}^\alpha[0,1]$, $\alpha \in (0,1]$, of Hölder functions on the interval $[0,1]$ is the set of functions g on $[0,1]$ such that $\sup_{x\neq y\in[0,1]} |g(x) - g(y)|/|x-y|^\alpha$ is finite. Let us recall that if a function g belongs to $\mathcal{C}^\alpha$, $\alpha \in (0,1]$, then the sequence of its Haar-wavelet coefficients $\langle g, \psi_{lk}\rangle$ satisfies

$$\sup_{0\leq k<2^l, l\geq 0} 2^{l(1/2+\alpha)} |\langle g, \psi_{lk}\rangle| < \infty. \tag{4.1.3}$$

For a given $\alpha > 0$, and $n \geq 1$, define

$$\varepsilon^*_{n,\alpha} = \left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+1}}. \tag{4.1.4}$$

This is the minimax rate for estimating a density function in a ball of α-Hölder functions, when the supremum norm is considered as a loss, see Ibragimov and Khas'minskii (1980) and Khas'minskii (1979).

4.1.3 Spike and Slab prior distributions 'truncated' at a certain level L.

In the following, one defines the cutoff $L_{max} = \log_2(n)$ and L the largest integer such that

$$2^L L \leq n \tag{4.1.5}$$

Note that $L \leq L_{max}$ for every n.

Let $X^{(n)} = (X_1, \cdots, X_n)$ be i.i.d. from law P with density f.
Let Π be the prior on densities generated as follows. One keeps the Pólya tree random

measure with respect to the canonical dyadic partition of $[0,1]$ construction up to level L, replacing the Beta distributions by

$$\varepsilon \in \mathcal{E} \ , \ Y_{\varepsilon 0} \sim (1-\pi_{\varepsilon 0})\delta_{\frac{1}{2}} + \pi_{\varepsilon 0}\text{Beta}(\alpha_{\varepsilon 0}, \alpha_{\varepsilon 1}), \tag{4.1.6}$$

with parameters $\alpha_\varepsilon \in \mathbb{N}$ to be chosen and a real parameter π_ε (later to be taken of the form $2^{-\frac{l}{2}}e^{-Cl}$, where we wrote $l = |\varepsilon|$).

There are multiple probability distributions on Borelians of $[0,1]$ that coincide on dyadic intervals I_ε with $P(I_\varepsilon)$ resulting from the above construction. We consider the specific one that is absolutely continuous relatively to the Lebesgue measure on $[0,1]$ with a constant density on each I_ε, $|\varepsilon| = L+1$. So, both prior and posterior are histograms on dyadic intervals at depth L.

Definition. The prior distribution with parameters α_ε, π_ε, as above is called Spike and Slab Pólya tree and denoted $\Pi(\alpha_\varepsilon, \pi_\varepsilon)$.

This prior is based on an idea of Ghosal and van der Vaart, which is referred as Evenly Split Pólya tree in their book Ghosal and van der Vaart (2017). First note that the Haar coefficients f_{lk} of a density f can be expressed as

$$f_{lk} = \langle f, \psi_{lk}\rangle = 2^{\frac{l}{2}}P(I_\varepsilon)(1-2Y_{\varepsilon 0}) \tag{4.1.7}$$

The Spike and Slab Pólya tree can therefore be seen as a 'thresholding prior', as the thresholding takes place on the sequence of Haar coefficients of the function where $Y_{\varepsilon 0} = \frac{1}{2}$.

Using this Spike and Slab prior can be seen as taking a Hierarchical approach. The usual Pólya tree (PT) prior on densities (under (1.3.13)) leads to the following Bayesian diagram

$$\begin{aligned} X|f &\sim f \\ f &\sim PT((Y_{\varepsilon 0})) \text{ with } Y_{\varepsilon 0} \sim \text{Beta}(\alpha_{\varepsilon 0}, \alpha_{\varepsilon 1}), \end{aligned}$$

so the $Y_{\varepsilon 0}$ have fixed (Beta) distributions, whereas the Spike and Slab Pólya tree (SSPT) prior leads to the diagram

$$\begin{aligned} X|f &\sim f \\ f &\sim SSPT((Y_{\varepsilon 0})) \text{ with } Y_{\varepsilon 0} \sim (1-\pi_{\varepsilon 0})\delta_{\frac{1}{2}} + \pi_{\varepsilon 0}\ \text{Beta}(\alpha_{\varepsilon 0}, \alpha_{\varepsilon 1}) \end{aligned}$$

which can be seen as the following diagram, using a sequence $(\gamma_{\varepsilon 0})_\varepsilon$ of Bernoulli variables.

$$\begin{aligned} X|f &\sim f \\ f|(\gamma_{\varepsilon 0}) &\sim SSPT((Y_{\varepsilon 0})) \text{ with } Y_{\varepsilon 0} \sim (1-\gamma_{\varepsilon 0})\delta_{\frac{1}{2}} + \gamma_{\varepsilon 0} \text{ Beta}(\alpha_{\varepsilon 0}, \alpha_{\varepsilon 1}) \\ \gamma_{\varepsilon 0} &\sim \text{Be}(\pi_{\varepsilon 0}) \end{aligned}$$

So in this case the distributions followed by the $Y_{\varepsilon 0}$ are random, hence this approach can be viewed as hierarchical.

The a posteriori law. In the following, we will write

$$N_X(I_\varepsilon) = \sum_{i=1}^{n} \mathbb{1}_{X_i \in I_\varepsilon} \text{ , } p_{\alpha_{\varepsilon 0}} = \frac{(\alpha_{\varepsilon 0} + \alpha_{\varepsilon 1} - 1)!}{(\alpha_{\varepsilon 0} - 1)!(\alpha_{\varepsilon 1} - 1)!} \tag{4.1.8}$$

and

$$p_X = \frac{(N_X(I_\varepsilon) + \alpha_{\varepsilon 0} + \alpha_{\varepsilon 1} - 1)!}{(N_X(I_{\varepsilon 0}) + \alpha_{\varepsilon 0} - 1)!(N_X(I_{\varepsilon 1}) + \alpha_{\varepsilon 1} - 1)!} \tag{4.1.9}$$

The following proposition shows that the Spike and Slab Pólya tree prior, as does the classical Pólya tree prior, is conjugate in the Density Estimation model.

Proposition 6. For every $\varepsilon \in \mathcal{E}$ with $|\varepsilon| \leq L$, the a posteriori law of $Y_{\varepsilon 0}$ knowing $X_1, \ldots, X_n$ is

$$Y_{\varepsilon 0}|X \sim (1-\tilde{\pi}_{\varepsilon 0})\delta_{\frac{1}{2}} + \tilde{\pi}_{\varepsilon 0}\text{Beta}(\alpha_{\varepsilon 0}(X), \alpha_{\varepsilon 1}(X))$$

where $\tilde{\pi}_{\varepsilon 0}$ is defined via a quantity T as follows

$$\begin{aligned} \tilde{\pi}_{\varepsilon 0} =& \frac{\pi_{\varepsilon 0} T}{(1-\pi_{\varepsilon 0}) + \pi_{\varepsilon 0} T}, \\ T =& T(\varepsilon, X) = 2^{N_X(I_\varepsilon)} \frac{p_{\alpha_{\varepsilon 0}}}{p_X} \end{aligned}$$

and where

$$\alpha_\varepsilon(X) = N_X(I_\varepsilon) + \alpha_\varepsilon.$$

Proof. We have

$$\Pi(Y|X) \propto f(X_1, \ldots, X_n) \prod_{|\varepsilon|=0}^{L-1} \left((1-\pi_{\varepsilon 0})\delta_{\frac{1}{2}}(Y_{\varepsilon 0}) + \pi_{\varepsilon 0} p_{\alpha_{\varepsilon 0}} Y_{\varepsilon 0}^{\alpha_{\varepsilon 0}-1}(1-Y_{\varepsilon 0})^{\alpha_{\varepsilon 1}-1} \right)$$

with $f(X_1, \ldots, X_n) = \prod_{i=1}^{n} \prod_{|\varepsilon|=L} (2^L P(I_\varepsilon))^{\mathbb{1}_{X_i \in I_\varepsilon}} = (2^L)^n \prod_{|\varepsilon|=L} P(I_\varepsilon)^{N_X(I_\varepsilon)}$.

Noticing that $\prod_{|\varepsilon|=0}^{L-1} 2^{N_X(I_\varepsilon)} = 2^{nL}$, we have

$$\begin{aligned} f(X_1,\dots,X_n) &= (2^L)^n \prod_{|\varepsilon|=0}^{L-1} Y_{\varepsilon 0}^{N_X(I_{\varepsilon 0})}(1-Y_{\varepsilon 0})^{N_X(I_{\varepsilon 1})} \\ &= \prod_{|\varepsilon|=0}^{L-1} 2^{N_X(I_\varepsilon)} Y_{\varepsilon 0}^{N_X(I_{\varepsilon 0})}(1-Y_{\varepsilon 0})^{N_X(I_{\varepsilon 1})}. \end{aligned}$$

This gives us that

$$\begin{aligned} \Pi(Y|X) &= \tfrac{1}{A}\prod_{|\varepsilon|=0}^{L-1}\Big((1-\pi_{\varepsilon 0})2^{N_X(I_\varepsilon)}Y_{\varepsilon 0}^{N_X(I_{\varepsilon 0})}(1-Y_{\varepsilon 0})^{N_X(I_{\varepsilon 1})}\delta_{\frac12}(Y_{\varepsilon 0}) + \\ &\quad 2^{N_X(I_\varepsilon)}\pi_{\varepsilon 0}p_{\alpha_{\varepsilon 0}}Y_{\varepsilon 0}^{N_X(I_{\varepsilon 0})+\alpha_{\varepsilon 0}-1}(1-Y_{\varepsilon 0})^{N_X(I_{\varepsilon 1})+\alpha_{\varepsilon 1}-1}\Big) \\ &= \tfrac{1}{A}\prod_{|\varepsilon|=0}^{L-1}\Big((1-\pi_{\varepsilon 0})\delta_{\frac12}(Y_{\varepsilon 0}) + 2^{N_X(I_\varepsilon)}\pi_{\varepsilon 0}p_{\alpha_{\varepsilon 0}}Y_{\varepsilon 0}^{N_X(I_{\varepsilon 0})+\alpha_{\varepsilon 0}-1}(1-Y_{\varepsilon 0})^{N_X(I_{\varepsilon 1})+\alpha_{\varepsilon 1}-1}\Big) \end{aligned}$$

with $A = \prod_{|\varepsilon|=0}^{L-1}\left((1-\pi_{\varepsilon 0}) + 2^{N_X(I_\varepsilon)}\pi_{\varepsilon 0}\frac{p_{\alpha_{\varepsilon 0}}}{p_X}\right)$, which concludes the proof.

□

Note that if $\pi_\varepsilon = 1$, meaning that the prior is also a product of Beta variables, one recovers the standard conjugacy for the (truncated at L) usual Pólya tree.

We will henceforth use the following notations

1. Tilded notation, posterior distribution. We denote by $\tilde{P}$ a distribution sampled from the posterior distribution and by $\tilde{Y}$ the corresponding variables Y in (4.1.1). In particular, the variable $\tilde{Y}_{\varepsilon 0}$ is distributed following the marginal a posteriori law

$$\tilde{Y}_{\varepsilon 0} \sim (1-\tilde{\pi}_{\varepsilon 0})\delta_{\frac12} + \tilde{\pi}_{\varepsilon 0} Beta(\alpha_{\varepsilon 0}(X), \alpha_{\varepsilon 1}(X)).$$

2. Bar notation, posterior mean. Let $\bar{f} = \int f d\Pi(f|X)$ denote the posterior mean density and $\bar{P}$ the corresponding probability measure. We use the notation $\bar{Y}$ for the variables defining $\bar{P}$ via (4.1.1).

4.2 Main results

By definition, we take as prior as above the realisation of the Spike and Slab Pólya tree P that is absolutely continuous with respect to Lebesgue's measure with density equal to a histogram and histogram heights equal to $P(I_\varepsilon)$. The posterior is, by Proposition 6, again a Spike and Slab Pólya tree with density w.r.t. Lebesgue equal to a histogram and histogram heights equal to $\tilde{P}(I_\varepsilon)$. In particular, it induces a posterior on densities that we consider in the main results below.

4.2.1 An adaptive concentration result

The following Theorem shows that the a posteriori law obtained with a Spike and Slab type Pólya tree prior concentrates around the true density f_0 at minimax rate for the supremum-norm loss.

Theorem 19. Let $f_0 \in \mathcal{C}^\alpha[0,1]$, for $\alpha \in (0,1]$ and suppose $\|\log f_0\|_\infty < \infty$. Let $X_1, \dots, X_n$ be i.i.d. random variables on $[0,1]$ following P_{f_0}. Let Π be the prior on densities induced by a Spike and Slab Polya Tree prior $\Pi(\alpha_\varepsilon, \pi_\varepsilon)$ with the choices

$$\begin{aligned} \alpha_\varepsilon &= a \\ \pi_\varepsilon &= 2^{-\frac{l}{2}} e^{-\kappa l}, \quad l = |\varepsilon| \end{aligned}$$

for κ large enough constant and $a > 0$ constant. Then for any $M_n \to \infty$, in P_{f_0}-probability

$$\Pi\left[\|f - f_0\|_\infty \leq M_n \left(\frac{\log n}{n} \right)^{\frac{\alpha}{2\alpha+1}} \mid X \right] \to 1$$

This theorem is an adaptive version of Theorem 1 of Castillo (2017b). There are few results so far in the literature in density estimation for the supremum-norm loss, among those are the results from Castillo (2014), Hoffmann et al. (2015) and Yoo and Ghosal (2016) for multivariate regression.

4.2.2 A Bernstein Von Mises result

To establish a nonparametric Bernstein Von Mises (BVM) result, one has first to find a space $\mathcal{M}_0$ large enough to have convergence at rate $\sqrt{n}$ of the posterior density to a Gaussian process. One can then derive results for some other space $\mathcal{F}$ using continuous mapping for continuous functionals $\psi : \mathcal{M}_0 \to \mathcal{F}$. A space that combines nicely with

supremum norm structure was introduced by Castillo and Nickl (2014) and defined as follows, using an 'admissible' sequence $\omega = (\omega_l)_{l\geq 0}$ such that $\omega_l/\sqrt{l} \to \infty$ as $l \to \infty$

$$\mathcal{M}_0 = \mathcal{M}_0(\omega) = \left\{ x = (x_{lk})_{l,k} \; ; \; \lim_{l\to\infty} \max_{0\leq k<2^l} \frac{|x_{lk}|}{\omega_l} = 0 \right\} \tag{4.2.1}$$

Equipped with the norm $\|x\|_{\mathcal{M}_0} = \sup_{l\geq 0} \max_{0\leq k<2^l} \frac{|x_{lk}|}{\omega_l}$, this is a separable Banach space. In a slight abuse of notation, we will write $f \in \mathcal{M}_0$ if the sequence of its Haar wavelet coefficients belongs in that space : $(\langle f, \psi_{lk}\rangle)_{l,k} \in \mathcal{M}_0$.

P-white bridge process. For P a probability distribution in $[0,1]$, one defines the P-white bridge process, denoted by $\mathbb{G}_P$. This is the Gaussian process indexed by the Hilbert space $L^2(P) = \{f : [0,1] \to \mathbb{R}; \int_0^1 f^2 dP < \infty\}$ with covariance

$$E[\mathbb{G}_P(f)\mathbb{G}_P(g)] = \int_0^1 (f - \int_0^1 f dP)(g - \int_0^1 g dP) dP \tag{4.2.2}$$

We will denote by $\mathcal{N}$ the law of $\mathbb{G}_{P_0}$(with $P_0 = P_{f_0}$).

The main purpose of the admissible sequence ω is to ensure that $\mathbb{G}_P \in \mathcal{M}_0$. Intuitively, if one does not use these weights w_l, the maximum over 2^l Gaussian variables is of order $\sqrt{2\log(2^l)} = C\sqrt{l}$ and does not tend to 0 as $l \to \infty$, see Remark 1 of Castillo and Nickl (2014) for a precise proof of this result.

Bounded Lipschitz metric. Let $(\mathcal{S}, d)$ be a metric space. The bounded Lipschitz metric $\beta_{\mathcal{S}}$ on probability measures of $\mathcal{S}$ is defined as, for any μ, ν probability measures of $\mathcal{S}$,

$$\beta_{\mathcal{S}}(\mu,\nu) = \sup_{F;\|F\|_{BL}\leq 1} \left| \int_{\mathcal{S}} F(x)(d\mu(x) - d\nu(x)) \right|, \tag{4.2.3}$$

where $F : \mathcal{S} \to \mathbb{R}$ and

$$\|F\|_{BL} = \sup_{x\in\mathcal{S}} |F(x)| + \sup_{x\neq y} \frac{|F(x) - F(y)|}{d(x,y)}. \tag{4.2.4}$$

This metric metrizes the convergence in distribution: $\mu_n \to \mu$ in distribution as $n \to \infty$ if and only if $\beta_{\mathcal{S}}(\mu_n, \mu) \to 0$ as $n \to \infty$.

Recentering the distribution. To establish our BVM result, one also has to find a suitable way to center the posterior distribution. In this view, denote by P_n the empirical

measure

$$P_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}. \tag{4.2.5}$$

Let us also consider C_n, which is a smoothed version of P_n, defined by

$$\langle C_n, \psi_{lk}\rangle = \begin{cases} \langle P_n, \psi_{lk}\rangle \text{ if } l \leq L \\ 0 \text{ if } l > L, \end{cases} \tag{4.2.6}$$

where L is our original cutoff, defined by (4.1.5).
We finally introduce T_n, which depends on the true parameter α, defined by

$$\langle T_n, \psi_{lk}\rangle = \begin{cases} \langle P_n, \psi_{lk}\rangle \text{ if } l \leq L_n \\ 0 \text{ if } l > L_n, \end{cases} \tag{4.2.7}$$

where we defined L_n to be the integer such that

$$2^{L_n} = \lfloor c_0 \left(\frac{n}{\log n}\right)^{\frac{1}{1+2\alpha}} \rfloor \tag{4.2.8}$$

for a suitable constant $c_0 \in \mathbb{R}^{+*}$, whose precise value is made clear below.

Weak BVM result. We have the following Bernstein-von Mises phenomenon for f_0 in Hölder balls (standard Hölder balls are subsets of the following ones)

$$\mathcal{H}(\alpha, R) := \{f = (f_{lk}) : |f_{lk}| \leq R2^{-(\alpha+1/2)}, \forall l \geq 0, 0 \leq k < 2^l\}$$

Theorem 20. Let $\mathcal{N}$ be the law of $\mathbb{G}_{P_0}$. Let C_n be the centering defined in (4.2.6). Let $l_0(n)$ be an increasing and diverging sequence. We define the prior Π such that

$$\begin{aligned} &Y_{\varepsilon 0} \sim \text{Beta}(a, a) \text{ for } |\varepsilon| \leq l_0 \\ &Y_{\varepsilon 0} \sim (1 - \pi_{\varepsilon 0})\delta_{\frac{1}{2}} + \pi_{\varepsilon 0}\text{Beta}(a, a) \text{ for } l_0 < |\varepsilon| \leq L \end{aligned}$$

where $\pi_\varepsilon = 2^{-\frac{l}{2}}e^{-\kappa|\varepsilon|}$ with κ a large enough constant. The posterior distribution then satisfies a weak BvM : for every $\alpha, R > 0$

$$\sup_{f_0 \in \mathcal{H}(\alpha,R)} E_{f_0}\left[\beta_{\mathcal{M}_0(\omega)}(\Pi(\cdot|X) \circ \tau_{C_n}^{-1}, \mathcal{N})\right] \to 0$$

as $n \to \infty$ and for any admissible sequence $\omega = (\omega_l)$ with $\omega_{l_0(n)}/\sqrt{\log(n)} \to \infty$.

The choice of recentering of the distribution is quite flexible, as it can be checked that the result also holds if one replaces C_n by the posterior mean $\bar{f}_n$ or by T_n which depends on α. Actually, the only required condition on where one cuts the empirical measure is to satisfy Theorem 1 of Castillo and Nickl (2014). One can see that the cutoff L is exactly the furthest one can go according to that theorem.

Using the methods of Castillo and Nickl (2014), this result leads to several applications, for instance derivation of BVM theorems for semiparametric functionals via the continuous mapping theorem and Donsker-type theorems, which do not appear here for the sake of brievity. It may also lead to the construction of adaptive credible sets although it may require substantial additional work.

4.3 Proofs

4.3.1 Preliminaries and notation

The following notations will be used throughout the proofs.

For a given distribution P with distribution function F and density f on $[0,1]$, denote $P(B) = F(B) = \int_B f$, for any measurable subset B of $[0,1]$. In particular under the "true" distribution, we denote $P_0(B) = F_0(B) = \int_B f_0$. We will also denote by p_ε the quantity $P(I_\varepsilon)$.

In the sequel C denotes a universal constant whose value only depends on other fixed quantities of the problem.

For a function f in L^2, and L_n an integer, denote by f^{L_n} the L^2-projection of f onto the linear span of all elements of the basis $\{\psi_{lk}\}$ up to level $l = L_n$. Also, denote $f^{L_n^c}$ the projection of f onto the orthocomplement $Vect\{\psi_{lk}, l > L_n\}$. In the proofs, we shall use the decomposition $f = f^{L_n} + f^{L_n^c}$, which holds in L^2 and L^∞ under prior and posterior as f is truncated at level L so has a finite Haar expansion under prior and posterior.

Recall the definition of L_n from (4.2.8).

We will write, for $i \in \{1, \dots, l\}$,

$$y_i = \frac{F_0(I_\varepsilon^{[i]})}{F_0(I_\varepsilon^{[i-1]})} \text{ and } \Delta_i = \sqrt{C_0 \frac{L_n 2^i}{n}} \tag{4.3.1}$$

.

For any integer l, set

$$\Lambda_n(l)^2 := (l + L_n)\frac{n}{2^l} \tag{4.3.2}$$

Define $\mathcal{B}$ an event on the dataspace on which, simultaneously for the countable family of indexes $l \geq 1, 0 \leq k < 2^l$, for M large enough to be chosen,

$$M^{-1}|N_X(I_k^l) - nF_0(I_k^l)| \leq \Lambda_n(l) \vee (l + L_n) \tag{4.3.3}$$

We recall Lemma 4 in Castillo (2017b) which ensures that one can restrict to the event $\mathcal{B}$ in the following. We note that, as here the levels $l > L$ are truncated in the prior and posterior, one actually only needs the control (4.3.2) for $l \leq L$.

Lemma 35. Let $X_1, \ldots, X_n$ be i.i.d. of density f_0 on $[0, 1]$, with f_0 bounded away from 0 and infinity. Then for M large enough there exists $B > \log(2)$ such that for every positive integer n

$$P_{f_0}^n(\mathcal{B}^c) \lesssim e^{-BL_n}$$

4.3.2 Proof of Theorem 19

Let us recall the definition of L_n in (4.2.8).

Since we have $f - f_0 = (f^{L_n} - \bar{f}^{L_n}) + (\bar{f}^{L_n} - f_0^{L_n}) + (f^{L_n^c}) - (f_0^{L_n^c})$, the proof will be split in four parts, where we study each of these terms separately.

4.3.2.1 Term $f_0^{L_n^c}$

$$\|f_0^{L_n^c}\|_\infty \leq \sum_{l=L_n+1}^{\infty} \left(\{ \max_{0 \leq k < 2^l} |f_{0,lk}| \} \left\| \sum_{k=0}^{2^l-1} |\psi_{lk}| \right\|_\infty \right) \lesssim \sum_{l > L_n} 2^{-l\alpha} \lesssim \varepsilon_{n,\alpha}^*.$$

4.3.2.2 Term $\bar{f}^{L_n} - f_0^{L_n}$

One uses the key identities

$$\bar{f}_{lk} = 2^{\frac{l}{2}} \bar{p}_\varepsilon (1 - 2\bar{Y}_{\varepsilon 0}) \text{ and } f_{0,lk} = 2^{\frac{l}{2}} p_{0,\varepsilon}(1 - 2y_{\varepsilon 0})$$

By Lemma 37 and Lemma 38 below, we have on the event $\mathcal{B}$

$$|\bar{f}_{lk} - f_{0,lk}| = \left| f_{0,lk}\left(\frac{\bar{p}_\varepsilon}{p_{0,\varepsilon}} - 1\right) + 2^{1+\frac{l}{2}} \bar{p}_\varepsilon (y_{\varepsilon 0} - \bar{Y}_{\varepsilon 0}) \right| \lesssim |f_{0,lk}| \left(\frac{2^l}{n} + \sqrt{\frac{L_n 2^l}{n}} \right) + \sqrt{\frac{L_n}{n}} \tag{4.3.4}$$

This gives us that, for $\varepsilon^*_{n,\alpha}$ as in (4.1.4),

$$\begin{aligned}\|\bar{f}^{L_n} - f_0^{L_n}\|_\infty &\lesssim \sum_{l=0}^{L_n} 2^{\frac{l}{2}} \max_{0\le k<2^l} |\bar{f}_{lk} - f_{0,lk}| \\ &\lesssim \frac{1}{n}\sum_{l=0}^{L_n} 2^{l(\frac{1}{2}-\alpha)} + \sqrt{\frac{L_n}{n}}\sum_{l=0}^{L_n} 2^{-l\alpha} + \sqrt{\frac{L_n}{n}2^{L_n}} \\ &\lesssim \varepsilon^*_{n,\alpha},\end{aligned}$$

where we have used $|f_{0,lk}| \lesssim 2^{-l(\frac{1}{2}+\alpha)}$.

4.3.2.3 Term $f^{L_n} - \bar{f}^{L_n}$

Consider the event

$$\mathcal{A} = \{\forall i \le L_n, |\varepsilon| = l \le L_n, |\bar{Y}_\varepsilon^{[i]} - \tilde{Y}_\varepsilon^{[i]}| \le r_\varepsilon^{[i]}\} \text{ with } r_\varepsilon^{[i]} = M\sqrt{\frac{L_n}{nF_0(I_\varepsilon^{[i]})}}. \tag{4.3.5}$$

Note that, by Lemma 39,

$$\Pi(\mathcal{A}^c|X) \lesssim e^{-CL_n}\sum_{l=0}^{L_n} 2^l \lesssim 2^{L_n}e^{-CL_n},$$

which tends to 0 when $n \to \infty$.

Now using similar arguments as in the proof of Theorem 1 of Castillo (2017b) where the bound for the Pólya tree given in Lemma 2 is replaced by the bound for the Spike and Slab Pólya tree given in Lemma 38, we have, on $\mathcal{A}$ and $\mathcal{B}$,

$$|\frac{\tilde{p}_\varepsilon}{\bar{p}_\varepsilon} - 1| \lesssim \sum_{i=0}^{l-1} r_\varepsilon^{[i]} \lesssim \sqrt{\frac{L_n 2^l}{n}} \tag{4.3.6}$$

Using the fact that $\tilde{Y}_{\varepsilon 0}$ is a Beta variable and is therefore bounded by 1, we have

$$\begin{aligned}|f_{lk} - \bar{f}_{lk}| &= 2^{\frac{l}{2}}|(\tilde{p}_\varepsilon - \bar{p}_\varepsilon) + 2\bar{p}_\varepsilon(\bar{Y}_{\varepsilon 0} - \tilde{Y}_{\varepsilon 0}) + 2\tilde{Y}_{\varepsilon 0}(\bar{p}_\varepsilon - \tilde{p}_\varepsilon)| \\ &\le 2^{\frac{l}{2}}\left(|\tilde{p}_\varepsilon - \bar{p}_\varepsilon| + 2|\bar{p}_\varepsilon(\bar{Y}_{\varepsilon 0} - \tilde{Y}_{\varepsilon 0})| + 2|\tilde{p}_\varepsilon - \bar{p}_\varepsilon|\right) \\ &\le 2^{\frac{l}{2}}\bar{p}_\varepsilon\left(3|\tfrac{\tilde{p}_\varepsilon}{\bar{p}_\varepsilon} - 1| + |\bar{Y}_{\varepsilon 0} - \tilde{Y}_{\varepsilon 0}|\right)\end{aligned}$$

Using the fact that $\bar{p}_\varepsilon \lesssim 2^{-l}$, $|\bar{Y}_{\varepsilon 0} - \tilde{Y}_{\varepsilon 0}| \lesssim \sqrt{\frac{L_n 2^l}{n}}$ on $\mathcal{A}$ and (4.3.6), we obtain that, on $\mathcal{A}$ and $\mathcal{B}$,

$$|f_{lk} - \bar{f}_{lk}| \lesssim \sqrt{\frac{L_n}{n}}. \tag{4.3.7}$$

Which leads, on $\mathcal{A}$ and $\mathcal{B}$, to $\|f^{L_n} - \bar{f}^{L_n}\|_\infty \lesssim \sqrt{\frac{L_n 2^{L_n}}{n}} \lesssim \varepsilon^*_{n,\alpha}$.

4.3.2.4 Term $f^{L_n^c}$

We have, denoting by E_X the expectation under the posterior distribution,

$$E_X[\|f^{L_n^c}\|_\infty] \leq \sum_{l=L_n+1}^{\log_2(n)} 2^{\frac{l}{2}} E_X[\max_{0\leq k<2^l} |f_{lk}|] \leq \sum_{l=L_n+1}^{\log_2(n)} 2^{\frac{l}{2}} \left(\sum_{k=0}^{2^l-1} E_X[|f_{lk}|] \right)$$

We have $E_X[|f_{lk}|] = 2^{\frac{l}{2}} E_X[\tilde{p}_\varepsilon] E_X[|1 - 2\tilde{Y}_{\varepsilon 0}|]$.

On one hand, we have

$$E_X[|1 - 2\tilde{Y}_{\varepsilon 0}|] = (1 - \tilde{\pi}_\varepsilon) \int (1-2u)\delta_{\frac{1}{2}}(u) + \tilde{\pi}_\varepsilon E[|1-2Z|] = \tilde{\pi}_\varepsilon E[|1-2Z|],$$

with Z drawn from a Beta$(\alpha_{\varepsilon 0}, \alpha_{\varepsilon 1})$. This gives us that, using Lemma 40

$$E_X[|1 - 2\tilde{Y}_{\varepsilon 0}|] \leq \tilde{\pi}_\varepsilon \lesssim \frac{e^{-C_2 l}}{\sqrt{n}}.$$

On the other hand, we have $E_X[\tilde{p}_\varepsilon] = \prod_{i=0}^{l-1} Q_{X,\varepsilon}(i,0)$ with

$$Q_{X,\varepsilon}(i,0) = (1 - \tilde{\pi}_\varepsilon^{[i+1]}) \frac{1}{2} + \tilde{\pi}_\varepsilon^{[i+1]} \frac{l + N_X(I_\varepsilon^{[i+1]})}{2l + N_X(I_\varepsilon^{[i]})}$$

using the notations $I_\varepsilon^{[i]} = I_{\varepsilon_1 \varepsilon_2 \ldots \varepsilon_i}$ and $\tilde{\pi}_\varepsilon^{[i]} = \tilde{\pi}_{\varepsilon_1 \varepsilon_2 \ldots \varepsilon_i}$.

Let us distinguish two regimes, when $i \leq L_n$ and when $i > L_n$. For the former, we have that, writing as before for $i \in \{1, \ldots, l\}$,

$$y_i = \frac{F_0(I_\varepsilon^{[i]})}{F_0(I_\varepsilon^{[i-1]})} \text{ and } \Delta_i = \sqrt{C_0 \frac{L_n 2^i}{n}}.$$

$$\begin{aligned} (1 - \tilde{\pi}_\varepsilon^{[i+1]}) \frac{1}{2} &\leq (1 - \tilde{\pi}_\varepsilon^{[i+1]})(y_{i+1} + |\tfrac{1}{2} - y_{i+1}|) \\ &\leq (1 - \tilde{\pi}_\varepsilon^{[i+1]}) y_{i+1} + \frac{1}{n} + \Delta_i. \end{aligned}$$

This gives us that there exists $C_1, C_1' > 0$ such that :

$$\begin{aligned} Q_{X,\varepsilon}(i,0) &\leq y_{i+1}\left(\frac{\frac{1}{n}+\Delta_i}{y_{i+1}} + (1-\tilde{\pi}_\varepsilon^{[i+1]}) + \tilde{\pi}_\varepsilon^{[i+1]}\left(1 + C_1\frac{(l+1)2^i+\sqrt{in}2^{\frac{i}{2}}}{n}\right)\right) \\ &\leq \frac{F_0(I_\varepsilon^{[i+1]})}{F_0(I_\varepsilon^{[i]})}\left(1 + C_1'\frac{(l+1)2^i+\sqrt{in}2^{\frac{i}{2}}}{n}\right) \end{aligned}$$

This implies that

$$\begin{aligned} \prod_{i=0}^{L_n} Q_{X,\varepsilon}(i,0) &\lesssim F_0(I_\varepsilon^{[L_n+1]})\prod_{i=0}^{L_n}\left(1 + C_2\frac{l2^i+\sqrt{in}2^{\frac{i}{2}}}{n}\right) \\ &\lesssim \frac{1}{2^{L_n}}\exp\left(C_4\sum_{i=0}^{L_n}\frac{l2^i+\sqrt{in}2^{\frac{i}{2}}}{n}\right) \\ &\lesssim \frac{1}{2^{L_n}}\exp\left(C_5\frac{L_n2^{L_n}+\sqrt{L_nn}2^{\frac{L_n}{2}}}{n}\right) \\ &\lesssim \frac{1}{2^{L_n}}. \end{aligned}$$

When $i > L_n$, we have, $Q_{X,\varepsilon}(i,0) \leq \frac{1}{2}(1+\frac{2}{\sqrt{n}})$, therefore :

$$\begin{aligned} \prod_{i=L_n+1}^{l} Q_{X,\varepsilon}(i,0) &\lesssim \frac{2^{L_n}}{2^l}\prod_{i=L_n+1}^{l}(1+\frac{2}{\sqrt{n}}) \\ &\lesssim \frac{2^{L_n}}{2^l}(1+\frac{2}{\sqrt{n}})^{l-L_n} \end{aligned}$$

We finally have $E_X[\tilde{p}_\varepsilon] \leq \frac{1}{2^l}(1+\frac{2}{\sqrt{n}})^{l-L_n}$

This leads to $E_X[\|f^{L_n^c}\|_\infty] \leq \frac{1}{\sqrt{n}}\sum_{l=L_n+1}^{\log_2(n)} 2^l e^{-C_2 l}(1+\frac{2}{\sqrt{n}})^{l-L_n}$,

therefore

$$\begin{aligned} E_X[\|f^{L_n^c}\|_\infty] &\leq \frac{e^{(L_n+1)(\log 2-C_2)}}{\sqrt{n}}(1+\tfrac{2}{\sqrt{n}})\frac{1-\left[(1+\frac{2}{\sqrt{n}})e^{\log 2-C_2}\right]^{\log_2(n)-L_n}}{1-(1+\frac{2}{\sqrt{n}})e^{\log 2-C_2}} \\ &\lesssim \frac{e^{(L_n+1)(\log 2-C_2)}}{\sqrt{n}} \end{aligned}$$

which tends to 0 faster than any power of n provided C_2 is chosen large enough.

4.3.2.5 Conclusion

Gathering the different bounds, one obtains that $E_{f_0}\left[\Pi\left(\|f-f_0\|_\infty \geq M_n\left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+1}} \mid X\right)\right]$ is bounded by

$$\frac{E_{f_0}\left[E_X\left(\|f-f_0\|_\infty\right)\right]}{M_n\varepsilon_{n,\alpha}^*} \lesssim \frac{1}{M_n} + \frac{2^{L_n}e^{-CL_n}}{M_n\varepsilon_{n,\alpha}^*} + \frac{e^{(L_n+1)(\log 2-C_2)}}{M_n\varepsilon_{n,\alpha}^*\sqrt{n}}$$

This tends to 0 when $n \to \infty$, which concludes the proof oh Theorem 19.

4.3.3 Proof of Theorem 20

In what follows, one will denote, for $l \geq 0$, by π_l the projection onto the finite-dimensional subspace V_l of L^2 defined by

$$V_l = span\{\psi_{l'k} : 0 \leq l' \leq l, 0 \leq k < 2^{l'}\}.$$

One will also similarly denote by $\pi_{>l}$ the projection onto $span\{\psi_{l'k} : l' > l, 0 \leq k < 2^{l'}\}$.

The proof uses a similar approach as Ray (2017), but the argument has to be adapted to the density estimation model and to the specific Spike and Slab procedure considered here.

Let us introduce the sets

$$\mathcal{J}_n(\gamma) = \left\{(l,k); |f_{0,lk}| > \gamma\sqrt{\log n/n}\right\} \tag{4.3.8}$$

for every $\gamma \in \mathbb{R}^{+*}$. Note that, recalling (4.2.8) and for $f_0 \in \mathcal{C}^\alpha([0,1])$, $(l,k) \in \mathcal{J}_n(\gamma)$ implies $l \leq L_n$. We will also denote by S the support of f:

$$S = \{(l,k); f_{lk} \neq 0\} \tag{4.3.9}$$

One will firstly need the following tightness result.

A tightness result

Theorem 21. Under the assumptions of Theorem 20, for every $\eta > 0$, $R > 0$ and $\alpha \in (0,1)$, there exist $M > 0$ and $n_0 \in \mathbb{N}$ such that for every $n \geq n_0$

$$\sup_{f_0 \in \mathcal{H}(\alpha,R)} E_{f_0}\left[\Pi(\|f-f_0\|_{\mathcal{M}_0} \geq M/\sqrt{n}|X)\right] < \eta$$

Proof. Fix $\eta > 0$. Consider the event

$$A_n = \{S^c \cap \mathcal{J}_n(\bar{\gamma}) = \emptyset\} \cap \{S \cap \{l > L_n\} \neq \emptyset\} \cap \{\max_{(l,k):l \leq L_n} |f_{0,lk} - f_{lk}| \leq \bar{\gamma}\sqrt{\log(n)/n}\}$$

By Lemma 41, there exist $\bar{\gamma} > 0$ such that for every $\alpha \in (0,1)$ there exists $B > 0$ such that, for every $f_0 \in \mathcal{H}(\alpha, R)$, $E_{f_0}[\Pi(A_n^c|X)] \lesssim n^{-B}$.

Now with $D > 0$ to be chosen, $E_{f_0}[\Pi(\|f - f_0\|_{\mathcal{M}_0} \geq M/\sqrt{n}|X)]$ is bounded above by $T_1 + T_2 + T_3$, where

$$\begin{aligned} T_1 &= E_{f_0}[\Pi(\{\|f - f_0\|_{\mathcal{M}_0} \geq M/\sqrt{n}\} \cap \{\|\pi_{l_0}(f - f_0)\|_{\mathcal{M}_0} \leq D/\sqrt{n}\} \cap A_n|X)] \\ T_2 &= E_{f_0}[\Pi(\{\|f - f_0\|_{\mathcal{M}_0} \geq M/\sqrt{n}\} \cap \{\|\pi_{l_0}(f - f_0)\|_{\mathcal{M}_0} > D/\sqrt{n}\} \cap A_n|X)] \\ T_3 &= E_{f_0}[\Pi(A_n^c|X) \end{aligned}$$

The last term is a $o(1)$. The first term is bounded by

$$E_{f_0}\left[\Pi(\{\|\pi_{>l_0}(f - f_0)\|_{\mathcal{M}_0} \geq (M - D)/\sqrt{n}\} \cap A_n|X)\right].$$

We now proceed as in Hoffmann et al. (2015). As $f_0 \in \mathcal{H}(\alpha, R)$, there exists $J_n(\alpha)$ with $2^{J_n(\alpha)} \lesssim (n/\log n)^{1/(2\alpha+1)}$ such that $\mathcal{J}_n(\underline{\gamma}) \subset \{(l,k) : l \leq J_n(\alpha), 0 \leq k < 2^l\}$ and

$$\sup_{f_0 \in \mathcal{H}(\alpha,R)} \sup_{l > J_n(\alpha)} \omega_l^{-1} \max_k |f_{0,lk}| \leq \frac{R2^{-J_n(\alpha)(\alpha+1/2)}}{\sqrt{J_n(\alpha)}} \leq C(\alpha, R)/\sqrt{n}$$

It now remains to bound the part with the frequencies $l_0 < l \leq J_n(\alpha)$. On A_n, we have

$$\sup_{l_0 < l \leq J_n(\alpha)} \omega_l^{-1} \max_k |f_{0,lk} - f_{lk}| \leq \frac{\bar{\gamma}}{\omega_{l_0}} \sqrt{\frac{\log n}{n}} \leq \bar{\gamma}/(c\sqrt{n})$$

since by hypothesis $\omega_{l_0} \geq c\sqrt{\log n}$. This gives us that, on A_n, $\|\pi_{>l_0}(f - f_0)\|_{\mathcal{M}_0} = O(n^{-1/2})$ for every $f_0 \in \mathcal{H}(\alpha, R)$. We therefore choose $M = M(\eta)$ to make the term T_1 smaller than $\eta/2$.

The term T_2 is bounded by $E_{f_0}[\Pi(\sqrt{n}\|\pi_{l_0}(f - f_0)\|_{\mathcal{M}_0} > D|X)]$, which is bounded, by Markov's inequality, by $\frac{\sqrt{n}}{D} E_{f_0}\left[E^{\Pi}(\|\pi_{l_0}(f - f_0)\|_{\mathcal{M}_0}|X)\right]$. We have

$$\frac{\sqrt{n}}{D}E_{f_0}\left[E^{\Pi}(\|\pi_{l_0}(f-f_0)\|_{\mathcal{M}_0}|X)\right] = \frac{\sqrt{n}}{D}E_{f_0}\left[E^{\Pi}(\sup_{l\leq l_0}\frac{1}{\omega_l}\max_k|f_{l,k}-f_{0,l,k}|\ |X)\right]$$
$$\leq \frac{\sqrt{n}}{D}\left(E_{f_0}\left[E^{\Pi}(\sup_{l\leq l_0}\frac{1}{\omega_l}\max_k|\langle f-T_n,\psi_{l,k}\rangle|\ |X)\right] + E_{f_0}\left[E^{\Pi}(\sup_{l\leq l_0}\frac{1}{\omega_l}\max_k|\langle f_0-T_n,\psi_{l,k}\rangle|\ |)\right]\right)$$

The first expectation is bounded by $C/\sqrt{n}$ as in Castillo (2017b) (see Lemma 8 and the proof of tightness starting page 2091). Indeed, when $l \leq l_0$ our prior is only a Beta, just as in Castillo (2017b), except that the parameters $a_l \equiv a$ of the Beta are constant in our case and do not decay to 0, but this decline to 0 is irrelevant to frequencies $l \leq l_0$. The second term can be bounded, following the approach of Castillo and Nickl (2014) in their first theorem, using that $(\omega_l)_l$ is admissible and with $\kappa > 0$ large enough, by

$$\frac{1}{D}\left[\sup_{l\leq l_0}\frac{\sqrt{l}}{\omega_l}\right]E_{f_0}\left[\sup_{l\leq l_0}\frac{1}{\sqrt{l}}\max_k|\sqrt{n}\langle f_0-T_n,\psi_{lk}\rangle|\right]$$
$$\lesssim \frac{\kappa}{D}+\frac{1}{D}\int_\kappa^\infty P_{f_0}\left(\sup_{l\leq l_0}\frac{1}{\sqrt{l}}\max_k|\sqrt{n}\langle f_0-T_n,\psi_{lk}\rangle|>u\right)du$$
$$\lesssim \frac{\kappa}{D}+\frac{1}{D}\sum_{l\leq l_0,k}\int_\kappa^\infty P_{f_0}\left(|\sqrt{n}\langle f_0-T_n,\psi_{lk}\rangle|>\sqrt{l}u\right)du$$
$$\lesssim \frac{\kappa}{D}+\frac{1}{D}\sum_{l\leq l_0}2^l\int_\kappa^\infty e^{-Clu}du \lesssim \frac{\kappa}{D}+\frac{1}{D}\sum_{l\leq l_0}e^{-C'\kappa l}\lesssim\frac{1}{D}$$

where the third inequality follows from an application of Bernstein's inequality.

This finally gives us that by taking $D = D(\eta)$ large enough the second term can be made smaller than $\eta/2$, which concludes the proof of Theorem 21. □

Proof of Theorem 20

Fix $\eta > 0$ and denote $\tilde{\Pi}_n = \Pi(\cdot|X)\circ\tau_{C_n}^{-1}$. By the triangle inequality, uniformly over the relevant class of functions, for fixed $l > 0$, we have

$$\beta_{\mathcal{M}_0}(\tilde{\Pi}_n,\mathcal{N}) \leq \beta_{\mathcal{M}_0}(\tilde{\Pi}_n,\tilde{\Pi}_n\circ\pi_l^{-1})+\beta_{\mathcal{M}_0}(\tilde{\Pi}_n\circ\pi_l^{-1},\mathcal{N}\circ\pi_l^{-1})+\beta_{\mathcal{M}_0}(\mathcal{N}\circ\pi_l^{-1},\mathcal{N}) \quad (4.3.10)$$

Let us now look more precisely at the first term of (4.3.10). Take a function $F : \mathcal{M}_0 \to \mathbb{R}$ such that $\|F\|_{BL} \leq 1$, F_n a random variable following $\tilde{\Pi}_n$ and $(\bar{\omega}_{l'})$ an admissible sequence such that $\bar{\omega}_{l'}/\omega_{l'} \to 0$ as $l' \to \infty$. Let us also consider the events

$$D = \{\|f\|_{\mathcal{M}_0} \leq M\} \text{ and } D_n = \{\|f - C_n\|_{\mathcal{M}_0} \leq M/\sqrt{n}\}$$

where M is large enough to have $E_{f_0}\left[\Pi(\|f-f_0\|_{\mathcal{M}_0} \geq M/\sqrt{n}|X)\right] < \eta/9$ as in Theorem 21. One has

$$\begin{aligned}
\left|\int_{\mathcal{M}_0} F d\tilde{\Pi}_n - \int_{\mathcal{M}_0} F d\tilde{\Pi}_n \circ \pi_l^{-1}\right| \leq & E^{\tilde{\Pi}_n}\left[|F(F_n) - F(\pi_l(F_n))||X\right] \\
\leq & E^{\tilde{\Pi}_n}\left[|F(F_n) - F(\pi_l(F_n))|(\mathbb{1}_D + \mathbb{1}_{D^c})|X\right] \\
\leq & E^{\tilde{\Pi}_n}\left[\|F_n - \pi_l(F_n)\|_{\mathcal{M}_0}\|F\|_{BL}\mathbb{1}_D|X\right] + 2\tilde{\Pi}_n(D^c|X) \\
\leq & E\left[\sup_{l'>l}\frac{1}{\omega_{l'}}\max_{0\leq k<2^{l'}}|\sqrt{n}\langle f - C_n, \psi_{l'k}\rangle|\mathbb{1}_{D_n}|X\right] \\
& + 2\Pi(\|f - C_n\|_{\mathcal{M}_0} \geq M/\sqrt{n}|X) \\
& \leq \left(\sup_{l'>l}\frac{\bar{\omega}_{l'}}{\omega_{l'}}\right) E\left[\sup_{l'>l}\frac{1}{\bar{\omega}_{l'}}\max_{0\leq k<2^{l'}}|\sqrt{n}\langle f - C_n, \psi_{l'k}\rangle|\mathbb{1}_{D_n}|X\right] \\
& + 2\Pi(\|(f - f_0) + (f_0 - C_n)\|_{\mathcal{M}_0} \geq M/\sqrt{n}|X) \\
& \leq \left(\sup_{l'>l}\frac{\bar{\omega}_{l'}}{\omega_{l'}}\right) M + 2\Pi(\|f - f_0\|_{\mathcal{M}_0} \geq M/(2\sqrt{n})|X) \\
& + 2\Pi(\|f_0 - C_n\|_{\mathcal{M}_0} \geq M/(2\sqrt{n})|X)
\end{aligned}$$

The first term can be made smaller than $\eta/9$ by taking l large enough. Using Theorem 21, the expectation of the second term can be made smaller than $\eta/9$ by taking n large enough. The last term can be handled as in Theorem 1 of Castillo and Nickl (2014) (as here j_n in that statement corresponds to our cutoff L) and be made smaller than $\eta/9$ by taking n large enough. Besides, one can note that the result holds when replacing C_n by T_n as in that case j_n would correspond to L_n which satisfies the required condition of Theorem 1 of Castillo and Nickl (2014).

This gives us that the first term of (4.3.10) is smaller than $\eta/3$. A similar result holds for the last term (see the proof of Theorem 1 of Castillo and Nickl (2014)). For the middle term, note that $l_0(n) \geq l$ for n large enough. For such n, the projected prior onto the first l coordinates is a product of Beta variables and we are exactly in the setting of Castillo (2017b), except that the parameters $a_l \equiv a$ of the Beta are constant in our case and do not decay to 0. Since l is fixed, the fact that the parameters of our Beta do not depend on l does not change the outcome. Therefore, following the proof of the convergence of the finite-dimensional projections from page 2089 to page 2091 of Castillo (2017b), the middle term can be made smaller than $\eta/3$, which concludes the proof.

4.3.4 Technical Lemmas

Recall the notation, for $i \in \{1, \dots, l\}$,

$$y_i = \frac{F_0(I_\varepsilon^{[i]})}{F_0(I_\varepsilon^{[i-1]})} \text{ and } \Delta_i = \sqrt{C_0 \frac{L_n 2^i}{n}}. \tag{4.3.11}$$

Lemma 36. For $l \leq L_n$, on the event $\mathcal{B}$, for $\varepsilon \in \mathcal{E}$ with $|\varepsilon| = l$, there exist some nonnegative real constants C and C' such that :

$$1 - \tilde{\pi}_\varepsilon^{[i]} \leq (1 - \tilde{\pi}_\varepsilon^{[i]}) \mathbb{1}_{|y_i - \frac{1}{2}| \leq \Delta_i} + C \frac{2^{-\frac{i}{2}} \sqrt{n}}{\pi_\varepsilon^{[i]}} e^{-C' n F_0(I_\varepsilon^{[i-1]}) \Delta_i^2} \mathbb{1}_{|y_i - \frac{1}{2}| > \Delta_i}.$$

In particular

$$1 - \tilde{\pi}_\varepsilon^{[i]} \lesssim \mathbb{1}_{|y_i - \frac{1}{2}| \leq \Delta_i} + \frac{1}{n} \mathbb{1}_{|y_i - \frac{1}{2}| > \Delta_i}.$$

Proof. Let us write

$$s = s_X = N_X(I_\varepsilon^{[i-1]}) + 2a - 2 \text{ and } q = q_X = N_X(I_\varepsilon^{[i]}) + a - 1 \tag{4.3.12}$$

so that we can rewrite $\frac{N_X(I_\varepsilon^{[i-1]}) + 2a - 1}{p_X} = \frac{q!(s-q)!}{s!}$. Using the fact that $\sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n} \leq n! \leq \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n+\frac{1}{12n}}$ for any integer n, this gives us that

$$\frac{N_X(I_\varepsilon^{[i-1]}) + 2a - 1}{p_X} \geq \sqrt{2\pi \frac{q(s-q)}{s}} \frac{(\frac{q}{e})^q (\frac{s-q}{e})^{s-q}}{(\frac{s}{e})^s} e^{-\frac{1}{12s}}.$$

We have, denoting by $B(a)$ the Bernoulli distribution of parameter a and $KL(P, Q)$ the Kullback-Leibler divergence between distributions P and Q, that

$$2^s \frac{(\frac{q}{e})^q (\frac{s-q}{e})^{s-q}}{(\frac{s}{e})^s} = e^{s\left(\frac{q}{s}\log(\frac{2q}{s}) + (1-\frac{q}{s})\log(2(1-\frac{q}{s}))\right)} = e^{sKL(B(\frac{q}{s})||B(\frac{1}{2}))}.$$

We also know that $KL(B(\frac{q}{s})||B(\frac{1}{2})) \geq \frac{1}{4}\|B(\frac{q}{s}) - B(\frac{1}{2})\|_{L_1}^2 = \frac{1}{4}(2|\frac{q}{s} - \frac{1}{2}|)^2$.

Recalling T from Proposition 6, this leads to

$$T \gtrsim \frac{1}{\sqrt{s+1}} \sqrt{\frac{q(s-q)}{s(s+1)}} e^{s|\frac{q}{s} - \frac{1}{2}|^2 - \frac{1}{12s}}$$

On $\mathcal{B}$, we can write $N_X(I_\varepsilon^{[i]}) = nF_0(I_\varepsilon^{[i]}) + \delta_{i,\varepsilon}$ with $|\delta_{i,\varepsilon}| \lesssim \sqrt{\frac{nL_n}{2^i}}$.

We have $|\frac{q}{s} - \frac{1}{2}| = \left| y_i - \frac{1}{2} + y_i \frac{\frac{\delta_{i,\varepsilon}}{nF_0(I_\varepsilon^{[i]})} - \frac{\delta_{i-1,\varepsilon}}{nF_0(I_\varepsilon^{[i-1]})}}{1+\frac{\delta_{i-1,\varepsilon}}{nF_0(I_\varepsilon^{[i-1]})}} \right| \geq |y_i - \frac{1}{2}| - y_i \frac{|\frac{\delta_{i,\varepsilon}}{nF_0(I_\varepsilon^{[i]})} - \frac{\delta_{i-1,\varepsilon}}{nF_0(I_\varepsilon^{[i-1]})}|}{1+\frac{\delta_{i-1,\varepsilon}}{nF_0(I_\varepsilon^{[i-1]})}}$.

Since $\frac{\delta_{i,\varepsilon}}{nF_0(I_\varepsilon^{[i]})}$ tends to 0 when $n \to \infty$, we have $|\frac{q}{s} - \frac{1}{2}| \gtrsim |y_i - \frac{1}{2}|$.

We now have $T \gtrsim \frac{e^{CnF_0(I_\varepsilon^{[i-1]})|y_i - \frac{1}{2}|^2}}{2^{-\frac{i}{2}}\sqrt{n}}$, which concludes the proof, since $1 - \tilde{\pi}_\varepsilon^{[i]} = \frac{1-\pi_\varepsilon^{[i]}}{(1-\pi_\varepsilon^{[i]})+\pi_\varepsilon^{[i]}T}$. □

Lemma 37. For $l \leq L_n$, $\varepsilon \in \mathcal{E}$ such that $|\varepsilon| = l$, on the event $\mathcal{B}$, we have

$$\left| \frac{\bar{p}_\varepsilon}{p_{0,\varepsilon}} - 1 \right| \leq C \left(\sum_{i=1}^{l} \frac{2^i}{n} + \sqrt{\frac{L_n 2^l}{n}} \right).$$

Proof. We have $p_{0,\varepsilon} = \prod_{i=1}^{l} y_i$ and $\bar{p}_\varepsilon = \prod_{i=1}^{l} w_i$ with $w_i = \frac{1}{2}(1 - \tilde{\pi}_\varepsilon^{[i]}) + \tilde{\pi}_\varepsilon^{[i]} \frac{N_X(I_\varepsilon^{[i]})+l}{N_X(I_\varepsilon^{[i-1]})+2l}$.
We can write

$$\left| \frac{w_i}{y_i} - 1 \right| = (1 - \tilde{\pi}_\varepsilon^{[i]}) \left| \frac{1}{2y_i} - 1 \right| + \tilde{\pi}_\varepsilon \left| \frac{N_X(I_\varepsilon^{[i]}) + a}{y_i(N_X(I_\varepsilon^{[i-1]}) + 2a)} - 1 \right|.$$

Using Lemma 42, the second term is bounded by a constant times $(\frac{2^i}{n} + \sqrt{\frac{L_n 2^i}{n}})$. The first term is bounded, by Lemma 36, by $\Delta_i + \frac{2^{-\frac{i}{2}}\sqrt{n}}{\pi_\varepsilon} e^{-CnF_0(I_\varepsilon^{[i-1]})\Delta_n^2} \lesssim \sqrt{\frac{L_n 2^i}{n}} + \frac{1}{n}$. One uses Lemma 43 to conclude the proof.

□

Lemma 38. For $l \leq L_n$,on the event $\mathcal{B}$, $\varepsilon \in \mathcal{E}$ with $|\varepsilon| = l$,

$$\left| \bar{Y}_{\varepsilon 0} - \frac{F_0(I_{\varepsilon 0})}{F_0(I_\varepsilon)} \right| \lesssim \frac{2^{\frac{l}{2}}}{n} (2^l |f_{0,lk}| + \sqrt{nL_n})$$

Proof. We have, with $y_{\varepsilon 0} = \frac{F_0(I_{\varepsilon 0})}{F_0(I_\varepsilon)}$,

$$\left| \frac{\bar{Y}_{\varepsilon 0}}{y_{\varepsilon 0}} - 1 \right| \leq (1 - \tilde{\pi}_{\varepsilon 0}) \left| \frac{1}{2y_{\varepsilon 0}} - 1 \right| + \tilde{\pi}_{\varepsilon 0} \left| \frac{N_X(I_{\varepsilon 0}) + a}{y_{\varepsilon 0}(N_X(I_\varepsilon) + 2a)} - 1 \right|$$

The second term is bounded using Lemma 44 by a constant times $(\frac{2^{2l+\frac{l}{2}}}{n} + \sqrt{\frac{L_n 2^l}{n}})$. The first term is bounded by $\sqrt{\frac{L_n 2^l}{n}} + \frac{1}{n}$, which concludes the proof. □

Lemma 39. For $\varepsilon \in \mathcal{E}$ with $|\varepsilon| = i \leq L_n$, let us write

$$B = \frac{M\sqrt{L_n}}{\sqrt{nF_0(I_\varepsilon^i)}} + \frac{4}{nF_0(I_\varepsilon^i)}.$$

Then, for $M > 0$ large enough, on the event $\mathcal{B}$, there exists $C > 2\log(2)$ such that

$$\Pi(|Y_\varepsilon - \bar{Y}_\varepsilon| > B|X) \lesssim e^{-CL_n}.$$

Proof. We will write Z a random variable drawn from a $\text{Beta}(a_\varepsilon(X), a_{\varepsilon'}(X))$, whose density will be noted as b.

Note first that the set $\{|u-\bar{Y}_\varepsilon| > B\}$ can be written as $\{|(1-\tilde{\pi}_\varepsilon)(u-\frac{1}{2})+\tilde{\pi}_\varepsilon(u-E[Z])| > B\}$, so that

$$\begin{aligned}
\Pi(|Y_\varepsilon - \bar{Y}_\varepsilon| > B|X) &= (1-\tilde{\pi}_\varepsilon)\int_{|u-\bar{Y}_\varepsilon|>B} \delta_{\frac{1}{2}}(u) + \tilde{\pi}_\varepsilon \int_{|u-\bar{Y}_\varepsilon|>B} b(u)du \\
&= (1-\tilde{\pi}_\varepsilon)\mathbb{1}_{\{\tilde{\pi}_\varepsilon|\frac{1}{2}-E[Z]|>B\}} + \tilde{\pi}_\varepsilon \int_{|u-\bar{Y}_\varepsilon|>B} b(u)du \\
&= (I) + (II)
\end{aligned}$$

For the first term, due to Lemma 36 we have

$$(I) \leq (1-\tilde{\pi}_\varepsilon)\mathbb{1}_{\{\tilde{\pi}_\varepsilon|\frac{1}{2}-E[Z]|>B, |y_i-\frac{1}{2}|\leq\Delta_i\}} + C\frac{2^{-\frac{i}{2}}\sqrt{n}}{\pi_\varepsilon} e^{-C'nF_0(I_\varepsilon^{[i-1]})\Delta_i^2}\mathbb{1}_{|y_i-\frac{1}{2}|>\Delta_i}$$

The first term is 0 so $(I) \lesssim \frac{2^{-\frac{i}{2}}\sqrt{n}}{\pi_\varepsilon} e^{-C'nF_0(I_\varepsilon^{[i-1]})\Delta_i^2} \lesssim e^{-CL_n}$ given our choice of Δ_i.

For the second term, we first write that, for Z as above,

$$(II)\mathbb{1}_{\{|y_i-\frac{1}{2}|>\Delta_i\}} \leq \tilde{\pi}_\varepsilon \mathbb{1}_{\{|y_i-\frac{1}{2}|>\Delta_i\}} \left(P\left((1-\tilde{\pi}_\varepsilon)|Z-\frac{1}{2}| > \frac{B}{2}\right) + P\left(\tilde{\pi}_\varepsilon|Z-E[Z]| > \frac{B}{2}\right)\right).$$

The first probability is 0 and the second one is bounded by $e^{-\frac{M^2L_n}{16}}$ using Lemma 45.

We also have, using Lemma 45

$$(II)\mathbb{1}_{\{|y_i-\frac{1}{2}|\leq\Delta_i\}} \leq \mathbb{1}_{\{|y_i-\frac{1}{2}|\leq\Delta_i\}} P\left(|Z-E[Z]| \geq C'\sqrt{\frac{L_n 2^i}{n}}\right) \lesssim e^{-\frac{M^2 L_n}{16}}.$$

□

Lemma 40. There exists $C_2 > 0$ such that for any $l > L_n$, $\varepsilon \in \mathcal{E}$ with $|\varepsilon| = l$, on $\mathcal{B}$,

$$\tilde{\pi}_\varepsilon \lesssim \frac{e^{-C_2 l}}{\sqrt{n}}.$$

Proof. As in Lemma 36, we will write $s = N_X(I_\varepsilon^{[i-1]}) + 2a - 2$ and $q = N_X(I_\varepsilon^{[i]}) + a - 1$, and $\delta = N_X(I_\varepsilon^{[i]}) - N_X(I_{\varepsilon'}^{[i]})$. We can note first that $q = \frac{s+\delta}{2}$. Let $M_l = M\left(\sqrt{\frac{nl}{2^l}} \vee l\right)$ be the constant appearing in (4.3.3) when $l > L_n$.

We have $|\delta| \leq n|F_0(I_\varepsilon) - F_0(I_{\varepsilon'})| + M_l \lesssim n2^{-l(\alpha+1)} + M_l \lesssim M_l$ because f_0 is α-Hölder.

As in Lemma 36,

$$\begin{aligned} T &\lesssim \frac{1}{\sqrt{s+1}}\sqrt{\frac{q(s-q)}{s(s+1)}} e^{s\left(\frac{q}{s}\log(\frac{2q}{s})+(1-\frac{q}{s})\log(2(1-\frac{q}{s}))\right)} \\ &\lesssim \frac{1}{\sqrt{s+1}} e^{s\left(\frac{1}{2}(1+\frac{\delta}{s})\log(1+\frac{\delta}{s})+\frac{1}{2}(1-\frac{\delta}{s})\log(1-\frac{\delta}{s})\right)} \lesssim \frac{1}{\sqrt{s+1}} e^{\frac{\delta^2}{2s}} \end{aligned}$$

As $l \leq L$ with L defined as in (4.1.5), we have

$$s \gtrsim \frac{n}{2^l} + \left(\sqrt{\frac{l+L_n}{2^l}n} \vee (l+L_n)\right).$$

which leads, as $l \leq L$, to

$$\frac{\delta^2}{2s} \lesssim \frac{2^l}{n}\left(\frac{nl}{2^l} \vee l^2\right)$$

This gives us that $T \lesssim \frac{2^{\frac{l}{2}}}{\sqrt{n}} e^{C_1 l}$. Choosing $\kappa = C_1 + C_2$ in the definition of π_l leads us to $\tilde{\pi}_\varepsilon \lesssim \pi_l T \lesssim \frac{e^{-C_2 l}}{\sqrt{n}}$. □

Lemma 41. Let $\mathcal{J}_n(\gamma)$ be defined in (4.3.8). There exist $\bar{\gamma} > 0$ such that for every $\alpha \in (0,1)$, there exists $B > 0$ such that, for every $f_0 \in \mathcal{H}(\alpha, R)$,

$$1) E_{f_0}\left[\Pi(S^c \cap \mathcal{J}_n(\bar{\gamma}) \neq \emptyset | X)\right] \lesssim \frac{n^{-2\alpha/(1+2\alpha)}}{(\log n)^{1/(2\alpha+1)}}$$
$$2) E_{f_0}\left[\Pi(S \cap \{l > L_n\} \neq \emptyset | X)\right] \lesssim 1/\sqrt{n}$$
$$3) E_{f_0}\left[\Pi(\max_{(l,k):l\leq L_n} |f_{0,lk} - f_{lk}| > \bar{\gamma}\sqrt{\log(n)/n} | X)\right] \lesssim 1/n^B$$

Proof. 1) We have, on $\mathcal{B}$, using Lemma 36

$$\begin{aligned}
\Pi(S^c \cap \mathcal{J}_n(\bar{\gamma}) \neq \emptyset | X) &\leq \sum_{(l,k)\in\mathcal{J}_n(\bar{\gamma})} \Pi(f_{lk} = 0 | X) \\
&\leq \sum_{(l,k)\in\mathcal{J}_n(\bar{\gamma})} (1 - \tilde{\pi}_{lk}) \\
&\leq \sum_{(l,k)\in\mathcal{J}_n(\bar{\gamma})} \left(\mathbb{1}_{|y_l - \frac{1}{2}| \leq \Delta_l} + \frac{1}{n} \mathbb{1}_{|y_l - \frac{1}{2}| > \Delta_l} \right)
\end{aligned}$$

The first term is in fact 0, as $|y_l - \frac{1}{2}| = |f_{0,l,k}| \frac{2^{-l/2}}{2F_0(I_k^l)} \gtrsim 2^{l/2}\sqrt{\log(n)/n} > \Delta_l$ for $(l,k) \in \mathcal{J}_n(\bar{\gamma})$ with $\bar{\gamma}$ large enough. This finally gives us that

$$\Pi(S^c \cap \mathcal{J}_n(\bar{\gamma}) \neq \emptyset | X) \lesssim \sum_{l \leq L_n} 2^l / n$$

2) We have, on $\mathcal{B}$, using Lemma 40

$$\begin{aligned}
\Pi(S \cap \{l > L_n\} \neq \emptyset | X) &\leq \sum_{(l,k):l>L_n} \Pi(f_{lk} \neq 0 | X) \\
&\leq \sum_{(l,k):l>L_n} \tilde{\pi}_{lk} \\
&\lesssim \sum_{(l,k):l>L_n} \frac{e^{-C_2 l}}{\sqrt{n}} \\
&\lesssim (2e^{-C_2})^{L_n} / \sqrt{n}
\end{aligned}$$

Choosing $C_2 > \log(2)$ then leads to the result.
3) A union bound gives us

$$\Pi(\max_{(l,k):l\leq L_n}|f_{0,lk}-f_{lk}|>\bar{\gamma}\sqrt{\frac{\log(n)}{n}}|X)$$
$$\leq\sum_{(l,k):l\leq L_n}\Pi(|f_{0,lk}-f_{lk}|>\bar{\gamma}\sqrt{\frac{\log(n)}{n}}|X)$$

Looking at the expectation under f_0 of each term, and in view of recentering by the posterior mean $\bar{f}$, one writes

$$E_{f_0}\left[\Pi(|f_{0,lk}-f_{lk}|>\bar{\gamma}\sqrt{\frac{\log(n)}{n}}|X)\right]\leq P^n_{f_0}\left(|f_{0,lk}-\bar{f}_{lk}|>\frac{\bar{\gamma}}{2}\sqrt{\frac{\log(n)}{n}}\right)$$
$$+E_{f_0}\left[\Pi(|f_{0,lk}-f_{lk}|>\bar{\gamma}\sqrt{\frac{\log(n)}{n}}|X)\mathbb{1}_{\{|f_{0,lk}-\bar{f}_{lk}|\leq\frac{\bar{\gamma}}{2}\sqrt{\frac{\log(n)}{n}}\}}\right] \tag{4.3.13}$$

On the event $\mathcal{B}$, using (4.3.4), we have that

$$|f_{0,lk}-\bar{f}_{lk}|\lesssim 2^{-l(\alpha+1/2)}\left(\frac{2^l}{n}+\sqrt{\frac{L_n2^l}{n}}\right)+\sqrt{\frac{L_n}{n}}\lesssim\frac{1}{n^{\frac{6\alpha+1}{4\alpha+2}}\log(n)^{\frac{1-2\alpha}{4\alpha+2}}}+\sqrt{\frac{\log((n/\log n)^{1/(2\alpha+1)})}{n}}$$

if $\alpha<1/2$. If $\alpha\geq 1/2$, we have

$$|f_{0,lk}-\bar{f}_{lk}|\lesssim 2^{-l(\alpha+1/2)}\left(\frac{2^l}{n}+\sqrt{\frac{L_n2^l}{n}}\right)+\sqrt{\frac{L_n}{n}}\lesssim\frac{1}{n}+\sqrt{\frac{\log((n/\log n)^{1/(2\alpha+1)})}{n}}.$$

This means that, for n large enough, the event in the probability in (4.3.13) is a subset of the event $\mathcal{B}^c$. Using Lemma 35, the probability in the last display is therefore bounded by a constant times e^{-BL_n}.
The second term of (4.3.13) is bounded by

$$E_{f_0}\left[\Pi(|\bar{f}_{lk}-f_{lk}|>\frac{\bar{\gamma}}{2}\sqrt{\frac{\log(n)}{n}}|X)\right].$$

As in (4.3.7), $|f_{lk}-\bar{f}_{lk}|\lesssim\sqrt{\frac{L_n}{n}}$ on $\mathcal{A}$ and $\mathcal{B}$, so the second term is bounded by $P^n_{f_0}(\mathcal{B}^c)+\Pi(\mathcal{A}^c|X)\lesssim e^{-CL_n}$. This finally leads to

$$E_{f_0}\left[\Pi(\max_{(l,k)\in\mathcal{J}_n(\underline{\gamma})}|f_{0,lk}-f_{lk}|>\bar{\gamma}\sqrt{\frac{\log(n)}{n}}|X)\right]\lesssim\sum_{(l,k):l\leq L_n}e^{-CL_n}\lesssim 2^{-L_n}$$

provided C is chosen greater than $2\log(2)$, which concludes the proof.

□

The following Lemmas are borrowed from Castillo (2017b).

Lemma 42. On the event $\mathcal{B}$,

$$\left|\frac{N_X(I_\varepsilon^{[i]})+a}{y_i(N_X(I_\varepsilon^{[i-1]})+2a)}-1\right| \lesssim \frac{2^i}{n}+\sqrt{\frac{L_n 2^i}{n}}$$

Proof. This is the quantity $\left|\frac{w_i}{y_i}-1\right|$ in Lemma 1 of Castillo (2017b). □

Lemma 43 (Lemma 3 of Castillo (2017b)). Let $\{y_i\}_{1\le i\le L}$, $\{w_i\}_{1\le i\le L}$ be two sequences of positive real numbers such that there are constants c_1, c_2 with

$$\max_{1\le i\le L}|\frac{w_i}{y_i}-1|\le c_1<1,\ \sum_{i=1}^{L}|\frac{w_i}{y_i}-1|\le c_2<\infty$$

Then there exists c_3 depending on c_1, c_2 only such that

$$\prod_{i=1}^{L}|\frac{w_i}{y_i}-1|\le c_3\sum_{i=1}^{L}|\frac{w_i}{y_i}-1|.$$

Lemma 44. On the event $\mathcal{B}$

$$\left|\frac{N_X(I_{\varepsilon 0})+a}{y_{\varepsilon 0}(N_X(I_\varepsilon)+2a)}-1\right| \lesssim (\frac{2^{2l+\frac{l}{2}}}{n}+\sqrt{\frac{L_n 2^l}{n}})$$

Proof. This is the quantity $\left|\frac{\bar{Y}_{\varepsilon 0}}{y_{\varepsilon 0}}-1\right|$ in Lemma 2 of (Castillo, 2017b). □

Lemma 45 (Lemma 6 of Castillo (2017b)). Let ϕ,ψ belong to $(0,\infty)$. Let Z follow a $Beta(\phi,\psi)$ distribution. Suppose, for some reals c_0, c_1,

$$0<c_0\le\phi/(\phi+\psi)\le c1<1 \text{ and } \phi\wedge\psi>8$$

Then there exists $D>0$ depending on c_0, c_1 only such that for any $x>0$,

$$P\left[|Z-E[Z]|>\frac{x}{\sqrt{\phi+\psi}}+\frac{2}{\phi+\psi}\right]\le De^{-\frac{x^2}{4}}.$$

References

Abramovich, F., Benjamini, Y., Donoho, D. L., and Johnstone, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.*, 34(2):584–653.

Babenko, A. and Belitser, E. (2010). Oracle convergence rate of posterior under projection prior and Bayesian model selection. *Math. Methods Statist.*, 19(3):219–245.

Belitser, E. and Nurushev, N. (2015). Needles and straw in a haystack: robust empirical Bayes confidence for possibly sparse sequences. *ArXiv e-print 1511.01803*.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.*, 37(4):1705–1732.

Birgé, L. and Massart, P. (2001). Gaussian model selection. *Journal of the European Mathematical Society*, 3(3):203–268.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Castillo, I. (2008). Lower bounds for posterior rates with Gaussian process priors. *Electronic Journal of Statistics*, 2:1281–1299.

Castillo, I. (2014). On bayesian supremum norm contraction rates. *Ann. Statist.*, 42(5):2058–2091.

Castillo, I. (2017a). Discussion of: Uncertainty quantification with the horseshoe. *Bayesian Analysis*, 12(4):1250–1253.

Castillo, I. (2017b). Polya tree posterior distributions on densities. *Ann. Inst. H. Poincaré Probab. Statist.*, 53(4):2074–2102.

Castillo, I. and Nickl, R. (2013). Nonparametric bernstein–von mises theorems in gaussian white noise. *Ann. Statist.*, 41(4):1999–2028.

Castillo, I. and Nickl, R. (2014). On the bernstein von mises phenomenon for nonparametric bayes procedures. *Ann. Statist.*, 42(5):1941–1969.

Castillo, I. and Roquain, E. (2018). On spike and slab empirical Bayes multiple testing. *arXiv e-prints*, page arXiv:1808.09748.

Castillo, I. and Rousseau, J. (2015). A bernstein–von mises theorem for smooth functionals in semiparametric models. *Ann. Statist.*, 43(6):2353–2383.

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.*, 43(5):1986–2018.

Castillo, I. and Szabo, B. (2018). Spike and slab empirical Bayes sparse credible sets. *arXiv e-prints*, page arXiv:1808.07721.

Castillo, I. and van der Vaart, A. W. (2012). Needles and straw in a haystack: posterior concentration for possibly sparse sequences. *Ann. Statist.*, 40(4):2069–2101.

Cox, D. D. (1993). An analysis of bayesian inference for nonparametric regression. *Ann. Statist.*, 21(2):903–923.

Donoho, D. L., Johnstone, I. M., Hoch, J. C., and Stern, A. S. (1992). Maximum entropy and the nearly black object. *J. Roy. Statist. Soc. Ser. B*, 54(1):41–81. With discussion and a reply by the authors.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1(2):209–230.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics*, 2(4):615–629.

Freedman, D. (1999). Wald lecture: On the bernstein-von mises theorem with infinite-dimensional parameters. *Ann. Statist.*, 27(4):1119–1141.

George, E. I. (2000). The variable selection problem. *J. Amer. Statist. Assoc.*, 95(452):1304–1308.

George, E. I. and Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika*, 87(4):731–747.

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531.

Ghosal, S. and van der Vaart, A. (2007). Convergence rates of posterior distributions for noniid observations. *Ann. Statist.*, 35(1):192–223.

Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Hoffmann, M., Rousseau, J., and Schmidt-Hieber, J. (2015). On adaptive posterior concentration rates. *Ann. Statist.*, 43(5):2259–2295.

Ibragimov, I. A. and Khas'minskii, R. Z. (1980). An estimate of the density of a distribution. *Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI). 98:61-85, 161-162, 166.*

Jiang, W. and Zhang, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *Ann. Statist.*, 37(4):1647–1684.

Johnson, V. E. and Rossell, D. (2010). On the use of non-local prior densities in bayesian hvoothesis tests hypothesis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 72(2):143–170.

Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.*, 32(4):1594–1649.

Johnstone, I. M. and Silverman, B. W. (2005). EbayesThresh: R Programs for Empirical Bayes Thresholding. *Journal of Statistical Software*, 12(8).

Khas'minskii, R. Z. (1979). A lower bound on the risks of non-parametric estimates of densities in the uniform metric. *Theory of Probability & Its Applications*, 23(4):794–798.

Kraft, C. H. (1964). A class of distribution function processes which have derivatives. *Journal of Applied Probability*, 1(2):385–388.

Lavine, M. (1992). Some aspects of polya tree distributions for statistical modelling. *The Annals of Statistics*, 20(3):1222–1235.

Martin, R. and Walker, S. G. (2014). Asymptotically minimax empirical bayes estimation of a sparse normal mean vector. *Electron. J. Statist.*, 8(2):2188–2206.

Mauldin, R. D., Sudderth, W. D., and Williams, S. C. (1992). Polya trees and random distributions. *Ann. Statist.*, 20(3):1203–1221.

Mismer, R. (2015). Master's thesis.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.*, 83(404):1023–1036. With comments by James Berger and C. L. Mallows and with a reply by the authors.

Ray, K. (2017). Adaptive bernstein von mises theorems in gaussian white noise. *Ann. Statist.*, 45(6):2511–2536.

Ročková, V. (2018). Bayesian estimation of sparse signals with a continuous spike-and-slab prior. *Ann. Statist.*, 46(1):401–437.

Ročková, V. and George, E. I. (2018). The spike-and-slab LASSO. *J. Amer. Statist. Assoc.*, 113(521):431–444.

Szabó, B., van der Vaart, A. W., and van Zanten, J. H. (2015). Rejoinder to discussions of "frequentist coverage of adaptive nonparametric bayesian credible sets". *Ann. Statist.*, 43(4):1463–1470.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

van der Pas, S., Szabó, B., and van der Vaart, A. (2017a). Adaptive posterior contraction rates for the horseshoe. *Electron. J. Stat.*, 11(2):3196–3225.

van der Pas, S., Szabó, B., and van der Vaart, A. (2017b). Uncertainty quantification for the horseshoe (with discussion). *Bayesian Anal.*, 12(4):1221–1274. With a rejoinder by the authors.

van der Pas, S. L., Salomond, J.-B., and Schmidt-Hieber, J. (2016). Conditions for posterior contraction in the sparse normal means problem. *Electron. J. Stat.*, 10(1):976–1000.

van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

Yoo, W. W. and Ghosal, S. (2016). Supremum norm posterior contraction and credible sets for nonparametric multivariate regression. *Ann. Statist.*, 44(3):1069–1102.

Yuan, M. and Lin, Y. (2005). Efficient empirical Bayes variable selection and estimation in linear models. *J. Amer. Statist. Assoc.*, 100(472):1215–1225.

Zhang, C.-H. (2005). General empirical bayes wavelet methods and exactly adaptive minimax estimation. *Ann. Statist.*, 33(1):54–100.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.