Electronic Journal of Statistics
Vol. 15 (2021) 1085–1153
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1811

Oracle posterior contraction rates under hierarchical priors[*]

Qiyang Han

Department of Statistics,
Rutgers University
Piscataway, NJ 08854
e-mail: qh85@stat.rutgers.edu

Abstract: We offer a general Bayes theoretic framework to derive posterior contraction rates under a hierarchical prior design: the first-step prior serves to assess the model selection uncertainty, and the second-step prior quantifies the prior belief on the strength of the signals within the model chosen from the first step. In particular, we establish non-asymptotic oracle posterior contraction rates under (i) a local Gaussianity condition on the log likelihood ratio of the statistical experiment, (ii) a local entropy condition on the dimensionality of the models, and (iii) a sufficient mass condition on the second-step prior near the best approximating signal for each model. The first-step prior can be designed generically. The posterior distribution enjoys Gaussian tail behavior and therefore the resulting posterior mean also satisfies an oracle inequality, automatically serving as an adaptive point estimator in a frequentist sense. Model mis-specification is allowed in these oracle rates.

The local Gaussianity condition serves as a unified attempt of non-asymptotic Gaussian quantification of the experiments, and can be easily verified in various experiments considered in [GvdV07a] and beyond. The general results are applied in various problems including: (i) trace regression, (ii) shape-restricted isotonic/convex regression, (iii) high-dimensional partially linear regression, (iv) covariance matrix estimation in the sparse factor model, (v) detection of non-smooth polytopal image boundary, and (vi) intensity estimation in a Poisson point process model. These new results serve either as theoretical justification of practical prior proposals in the literature, or as an illustration of the generic construction scheme of a (nearly) minimax adaptive estimator for a complicated experiment.

MSC2020 subject classifications: Primary 62G20; secondary 62G05.
Keywords and phrases: Bayes nonparametrics, hierarchical priors, local Gaussianity, trace regression, shape-restricted regression, covariance matrix estimation, detection of image boundary, intensity estimation of a Poisson point process.

Received December 2019.

Contents

[*] Supported in part by NSF Grant DMS-1566514, DMS-1916221, and by Isaac Newton Institute for Mathematical Sciences, program *Statistical Scalability*, EPSRC Grant Number LNAG/036 RG91310.

1. Introduction

1.1. Overview

Suppose we observe $X^{(n)}$ from a statistical experiment $(\mathfrak{X}^{(n)}, \mathcal{A}^{(n)}, P_f^{(n)})$, where f belongs to a statistical model $\mathcal{F}$ and $\{P_f^{(n)}\}_{f\in\mathcal{F}}$ is dominated by a σ-finite measure μ. In many cases, instead of using a single 'big' model $\mathcal{F}$, a collection of suitably nested (sub-)models $\{\mathcal{F}_m\}_{m\in\mathcal{I}} \subset \mathcal{F}$ are available to statisticians. A hierarchical Bayesian approach assigns a first-step prior Λ_n assessing the uncertainty in which model to use, followed by a second-step prior $\Pi_{n,m}$ quantifying the prior belief in the strength of the signals within the specific chosen model $\mathcal{F}_m$ from the first step.

Such a hierarchical prior design is intrinsic in many proposals for different problems, including the canonical Gaussian white noise/regression and density estimation [AGR13, BG03, dJvZ10, GLvdV08, KRvdV10, LvdV07, RS17, Scr06], and the more recent sparse linear regression [CSHvdV15, CvdV12], trace regression [ACCR14], shape restricted regression [HD11, HH03], covariance matrix estimation [GZ15, PBPD14], etc. Despite many contraction rates available for different models (see e.g. [Cas14, CSHvdV15, CvdV12, GGvdV00, GvdV07a, GvdV17, HRSH15, Rou10, SW01, vdVvZ08, vdVvZ09] for some key contributions), a unified theoretical understanding towards the behavior of posterior distributions under the hierarchical prior design has been limited. [GLvdV08] focused on designing adaptive Bayes procedures with models primarily indexed by the smoothness level of function classes in the context of density estimation. Their conditions are complicated and seem not directly applicable to other settings. [dJvZ10] uses a specific location mixture prior for regression/density estimation/classification. [AGR13] considered a more general setting where the models are indexed by functions that admit a linear ℓ_2-basis structure (e.g. Sobolev/Besov type); see also [RS17]. [GvdVZ15] designed a prior specific to structured linear problems in the Gaussian regression model, with their main focus on high-dimensional (linear) and network problems. As such, all these results apriori require certain specific form of the prior, the model structure, or the statistical experiments.

The goal of this paper aims at giving a unified theoretical treatment of deriving posterior contraction rates under the common hierarchical prior design, without specifying particular forms for the prior, the model structure, or the experiments. More specifically, we aim at identifying common *structural assumptions* on the statistical experiments $(\mathfrak{X}^{(n)}, \mathcal{A}^{(n)}, P_f^{(n)})$, the collection of models $\{\mathcal{F}_m\}$ and the priors $\{\Lambda_n\}$ and $\{\Pi_{n,m}\}$ such that the posterior distribution both

(G1) contracts at an *oracle* rate with respect to some metric[1] d_n:

$$\inf_{m\in\mathcal{I}}\left(\inf_{g\in\mathcal{F}_m} d_n^2(f_0,g)+\mathrm{pen}(m)\right), \tag{1.1}$$

where $\mathrm{pen}(m)$[2] is related to the 'dimension' of $\mathcal{F}_m$, and

(G2) puts little mass on models that are substantially larger than the oracle one balancing the bias-variance tradeoff in (1.1).

The oracle formulation (1.1) follows the convention in the frequentist literature on model selection [BC91, YB98, BBM99, Mas07, Tsy14], and has several advantages: (i) (*minimaxity*) if the true signal f_0 can be well-approximated by the models $\{\mathcal{F}_m\}$, the contraction rate in (1.1) is usually (nearly) minimax optimal, (ii) (*adaptivity*) if f_0 lies in certain low-dimensional model $\mathcal{F}_m$, the contraction rate adapts to this unknown information, and (iii) (*mis-specification*) if the models $\mathcal{F}_m$ are mis-specified while $d_n^2(f_0,\cup_{m\in\mathcal{I}}\mathcal{F}_m)$ remains 'small', then the contraction rate should still be rescued by this relatively 'small' bias.

As the main abstract result of this paper (cf. Theorem 2.3), we show that our goals (G1)-(G2) can be accomplished under:

(i) (**Experiment**) a local Gaussianity condition on the log likelihood ratio for the statistical experiment with respect to d_n;

(ii) (**Models**) a dimensionality condition of the model $\mathcal{F}_m$ measured in terms of local entropy with respect to the metric d_n;

(iii) (**Priors**) exponential weighting for the first-step prior Λ_n, and sufficient mass of the second-step prior $\Pi_{n,m}$ near the 'best' approximating signal $f_{0,m}$ within the model $\mathcal{F}_m$ for the true signal f_0.

The local Gaussianity condition is rooted in the frequentist theory of the convergence rates of M-estimators (i.e. estimators maximizing certain likelihood) via the theory of Gaussian and empirical processes. In fact, the local Gaussianity serves as an essential ingredient for various (by-now standard) techniques, including the Gaussian concentration and the chaining with bracketing, that give a unification to the theory for, e.g. regression and density estimation [BM93, vdG00, vdVW96] (see Appendix E for more discussions).

From the Bayesian theoretic side, one important convention in studying posterior contraction rates in the literature has been the construction of appropriate tests with exponentially small type I and II errors with respect to certain metric, the Gaussian behavior of type II error being particularly crucial [GGvdV00, GvdV07a]. It is rather curious if the frequentist local Gaussianity can also be useful in the Bayes theory. Our formulation in (i) can be viewed as an attempt in this regard, and seems useful in that, local Gaussianity with respect to the intrinsic metric is a rather universal property in various statistical experiments including the ones considered in [GvdV07a] and beyond: Gaussian/Laplace/binary/Poisson regression, density estimation, Gaussian autoregression, Gaussian time series, covariance matrix estimation, image boundary

[1] The requirement of being a metric can be weakened.

[2] $\mathrm{pen}(m)$ may depend on n but we suppress this dependence for notational convenience.

detection, and support boundary recovery in a Poisson point process model, etc. Moreover, such local Gaussianity naturally entails the Gaussian tail behavior of the posterior distribution, thereby complementing a recent result of [HRSH15] who showed that such a Gaussian tail behavior cannot be uniformly improved under uniform posterior consistency.

Conditions (ii) and (iii) are familiar in Bayes nonparametrics literature. In particular, the first-step prior can be designed generically (cf. Proposition 2.2). Sufficient mass of the second-step prior $\Pi_{n,m}$ is a minimal condition in the sense that using $\Pi_{n,m}$ alone should lead to a (nearly) optimal posterior contraction rate on the model $\mathcal{F}_m$.

As an illustration of the scope of our general results in concrete applications, we justify the prior proposals in (i) [ACCR14, MA15] for the trace regression problem, and in (ii) [HD11, HH03] for the shape-restricted regression problems. Despite many theoretical results for Bayes high-dimensional models (cf. [BG14, CSHvdV15, CvdV12, GvdVZ15, GZ15, PBPD14]), it seems that the important low-rank trace regression problem has not yet been successfully addressed. Our result here fills in this gap. Furthermore, to the best knowledge of the author, the theoretical results concerning shape-restricted regression problems provide the first systematic approach that bridges the gap between Bayesian nonparametrics and shape-restricted nonparametric function estimation literature in the context of adaptive estimation[3].

Several other applications are considered, including: (iii) high-dimensional partially linear regression model, (iv) covariance matrix estimation in the sparse factor model, (v) detection of polytopal image boundary, and (vi) estimation of piecewise constant intensity in a Poisson point process model. These results serve as an illustration of the generic construction scheme of a (nearly) minimax adaptive estimator in multi-structured experiments, or in experiments that seem far from Gaussian. We also revisit some density estimation problems, in particular in the location mixture models. The purpose of this is to provide some guidance of how the local Gaussianity can be applied via appropriate localization of the parameter space, when such Gaussianity may fail to hold at a global scale.

During the preparation of this paper, we become aware of a very recent paper [YP17] who independently considered a similar problem. Both our approach and [YP17] shed light on the behavior of Bayes procedures under hierarchical priors, while differing in several important aspects (cf. Remark 2.6). Moreover, our work here applies to a wide range of applications that are not covered by [YP17].

1.2. Notation

Let $(\mathcal{F}, \|\cdot\|)$ be a subset of the normed space of real functions $f : \mathcal{X} \to \mathbb{R}$. Let $\mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|)$ be the ε-covering number; see page 83 of [vdVW96] for more details. For a real-valued measurable function f defined on $(\mathcal{X}, \mathcal{A}, P)$, $\|f\|_{L_p(P)} \equiv$

[3] Almost completed at the same time, [MRS20] considered a Bayes approach for univariate log-concave density estimation, where they derived contraction rates without addressing the adaptation issue.

$(P|f|^p)^{1/p}$ denotes the usual L_p-norm under P (where $p \geq 1$), and will be simplified as $\|f\|_p$ when there is no potential confusion. $\|f\|_\infty \equiv \|f\|_{L_\infty} \equiv \sup_{x\in\mathcal{X}}|f(x)|$ denotes the supremum norm.

For any $v \in \mathbb{R}^d$, we use $\|v\|_p$ to denote the usual Euclidean p-norm. For any $\varepsilon > 0$, denote $B_d(v,\varepsilon) \equiv \{u \in \mathbb{R}^d : \|u - v\|_2 \leq \varepsilon\}$ the Euclidean ball in $\mathbb{R}^d$ centered at v with radius ε.

C_x denotes a generic constant that depends only on x, whose numeric value may change from line to line. $a \lesssim_x b$ and $a \gtrsim_x b$ mean $a \leq C_x b$ and $a \geq C_x b$ respectively, and $a \asymp_x b$ means $a \lesssim_x b$ and $a \gtrsim_x b$. For $a, b \in \mathbb{R}$, $a \vee b := \max\{a,b\}$ and $a \wedge b := \min\{a,b\}$. $P_f^{(n)}T$ denotes the expectation of a random variable $T = T(X^{(n)})$ under the experiment $(\mathfrak{X}^{(n)}, \mathcal{A}^{(n)}, P_f^{(n)})$.

1.3. *Organization*

Section 2 is devoted to the general results on oracle posterior contraction rates. We work out a wide range of experiments and some concrete applications that fit into our general theory in Section 3. Detailed proofs are deferred to the Appendix.

2. General results

In the hierarchical prior design framework, we first put a prior Λ_n on the model index $\mathcal{I}$, followed by a prior $\Pi_{n,m}$ on the model $\mathcal{F}_m$ chosen from the first step. The overall prior is a probability measure on $\mathcal{F}$ given by $\Pi_n \equiv \sum_{m\in\mathcal{I}} \lambda_n(m)\Pi_{n,m}$. The posterior distribution is then a random measure on $\mathcal{F}$: for a measurable subset $B \subset \mathcal{F}$,

$$\Pi_n(B|X^{(n)}) = \int_B p_f^{(n)}(X^{(n)})\,\mathrm{d}\Pi_n(f) \Big/ \int p_f^{(n)}(X^{(n)})\,\mathrm{d}\Pi_n(f) \tag{2.1}$$

where $p_f^{(n)}(\cdot)$ denotes the probability density function of $P_f^{(n)}$ with respect to the dominating measure μ.

2.1. *Assumptions*

For some $v > 0, c \in [0,\infty)$ let

$$\psi_{v,c}(\lambda) = v\lambda^2 \cdot \mathbf{1}_{|\lambda|\leq 1/c} + \infty \cdot \mathbf{1}_{|\lambda|>1/c} \tag{2.2}$$

denote the local quadratic function.

Assumption A (**Experiment: Local Gaussianity condition**). There exist some constants $c_1 > 0$ and $\kappa = (\kappa_g, \kappa_\Gamma) \in (0,\infty) \times [0,\infty)$ such that for all $n \in \mathbb{N}, \lambda \in \mathbb{R}$, and $f_0, f_1 \in \mathcal{F}$,

$$P_{f_0}^{(n)} e^{\lambda\left(\log(p_{f_0}^{(n)}/p_{f_1}^{(n)}) - P_{f_0}^{(n)}\log(p_{f_0}^{(n)}/p_{f_1}^{(n)})\right)} \leq c_1 e^{\psi_{\kappa_g n d_n^2(f_0,f_1),\kappa_\Gamma}(\lambda)}.$$

Here $d_n : \mathcal{F} \times \mathcal{F} \to \mathbb{R}_{\geq 0}$ is a symmetric function satisfying

$$(2.3)\qquad \big(c_2 \cdot d_n^2(f_0, f_1) - d_0^2\big)_+ \leq n^{-1} P_{f_0}^{(n)} \log(p_{f_0}^{(n)}/p_{f_1}^{(n)}) \leq c_3 \cdot d_n^2(f_0, f_1) + d_0^2,$$

for some constants $c_2, c_3 > 0$ and $d_0 \geq 0$ (possibly depending on n).

In Assumption A, we require the log likelihood ratio to have local Gaussian behavior with respect to the intrinsic 'metric' d_n in the sense of (2.3). If κ_Γ can be chosen to be 0, then the log likelihood ratio exhibits global Gaussian behavior. In Section 3, many statistical experiments, beyond the apparent Gaussian ones, will be shown to satisfy this local Gaussianity condition in their respective intrinsic metrics. In some cases the local Gaussianity by itself may entail certain apriori compactness constraints on the parameter space, for instance boundedness requirements for the parameter space in binary/Poisson regression and density estimation. These constraints can be removed, in a technical way, by working with appropriately localized subsets of the parameter space on which the local Gaussianity holds. See Section 2.3 and Appendix F for more details and examples in this regard.

As already mentioned in the Introduction, this local Gaussianity point of view has its root in the unified treatment of deriving convergence rates of M-estimators—a formal connection to the theory of sieved MLE under local Gaussianity will be given in Appendix E.

A direct consequence of the local Gaussianity of the statistical experiment is the following.

Lemma 2.1. *Let Assumption A hold. For any $f_0, f_1 \in \mathcal{F}$ such that $d_n(f_0, f_1) \geq \sqrt{2/(c_2 \wedge c_3)} \cdot d_0$, there exists some test ϕ_n such that*

$$\sup_{f \in \mathcal{F}: d_n^2(f, f_1) \leq c_5 d_n^2(f_0, f_1)} \Big(P_{f_0}^{(n)} \phi_n + P_f^{(n)}(1 - \phi_n)\Big) \leq c_6 e^{-c_7 n d_n^2(f_0, f_1)}$$

where $c_5 \leq 1/4, c_6 \in [2, \infty)$ and $c_7 \in (0, 1)$ only depends on the constants in Assumption A.

Next we state the assumption on the complexity of the models $\{\mathcal{F}_m\}_{m \in \mathcal{I}}$. Let $\mathcal{I} = \mathbb{N}^q$ be a q-dimensional lattice with the natural order $(\mathcal{I}, \leq)$[4]. Here the dimension q is understood as *the number of different structures* in the models $\{\mathcal{F}_m\}_{m \in \mathcal{I}}$. For instance, in the trace regression problem (cf. Section 3.1.1), there is only one rank structure so $q = 1$; in the covariance matrix estimation problem in the sparse factor model (cf. Section 3.5.1), there are both rank and sparsity structures so $q = 2$. In the sequel we will not explicitly mention q unless otherwise specified. We require the models to be nested in the sense that $\mathcal{F}_m \subset \mathcal{F}_{m'}$ if and only if $m \leq m'$ [5].

Let $f_{0,m}$ denote the 'best' approximation of f_0 within the model $\mathcal{F}_m$ in the sense that $f_{0,m} \in \arg\inf_{g \in \mathcal{F}_m} d_n(f_0, g)$[6]. Our assumption on the model complexity below, at a heuristic level, says that $\mathcal{F}_m$ has dimension $n\delta_{n,m}^2$ measured

[4] For any $a, b \in \mathcal{I}$, $a \leq b$ iff $a_i \leq b_i$ for all $1 \leq i \leq q$. Similar definition applies to $<, \geq, >$.

[5] Nesting requirement is for simplicity; see Appendix F for examples of non-nesting models.

[6] We assume that $f_{0,m}$ is well-defined without loss of generality.

in a local entropy sense, for some $\delta_{n,m} > 0$. In typical cases, $\mathcal{F}_m$ has 'dimension' m, and $\delta^2_{n,m} \approx \frac{m}{n} \times$ poly-log is regarded as the contraction rate on $\mathcal{F}_m$ (up to logarithmic factors).

Assumption B (**Models: Local entropy condition**). Let $\{\delta_{n,m}\}_{m\in\mathcal{I}} \subset \mathbb{R}_{>0}$ be such that each $\delta_{n,m}$ depends on n, m only, and:

- For each $m \in \mathcal{I}$,

$$1 + \sup_{\varepsilon > \delta_{n,m}} \log \mathcal{N}\big(c_5\varepsilon, \{f \in \mathcal{F}_m : d_n(f,g) \leq 2\varepsilon\}, d_n\big) \leq (c_7/2) n\delta^2_{n,m} \tag{2.4}$$

holds for all $g \in \{f_{0,m'}\}_{m' \leq m}$.

- Furthermore there exist some constants $\mathfrak{c} \in [1,\infty), \gamma \in [1,\infty), \mathfrak{h}_0 \in [1,\infty]$ such that for any $m \in \mathcal{I}$, $\alpha \geq c_7/2$ and any $1 \leq h \leq \mathfrak{h}_0$,

$$\sum_{m' \geq hm} e^{-\alpha n \delta^2_{n,m'}} \leq 2e^{-\alpha n h \delta^2_{n,m}/\mathfrak{c}^2}, \quad \mathfrak{c}^{-2}\delta^2_{n,hm} \leq h^\gamma \delta^2_{n,m}. \tag{2.5}$$

Using $\delta_{n,m}$'s, the models can be divided into over-fitting or under-fitting ones according to whether $\delta^2_{n,m} \geq \inf_{g\in\mathcal{F}_m} d^2_n(f_0,g)$ or $\delta^2_{n,m} < \inf_{g\in\mathcal{F}_m} d^2_n(f_0,g)$.

Note that if we choose all models $\mathcal{F}_m = \mathcal{F}$, then (2.4) reduces to the local entropy condition in [GGvdV00, GvdV07a]. When $\mathcal{F}_m$ is finite-dimensional, typically we can check (2.4) for all $g \in \mathcal{F}_m$. Now we comment on (2.5). The left side of (2.5) essentially requires super linearity of the map $m \mapsto \delta^2_{n,m}$, while the right side of (2.5) controls the degree of this super linearity. As a leading example, (2.5) will be trivially satisfied with $\mathfrak{c} = \gamma = 1, \mathfrak{h}_0 = \infty$ when $n\delta^2_{n,m} = c \cdot m \log(en)$ for some absolute constant $c > 2/c_7$.

Finally we state assumptions on the priors.

Assumption C (**Priors: Mass condition**). For all m,

(P1) (First-step prior) There exists some $\mathfrak{h} \geq 1$ such that

$$\lambda_n(m) \geq e^{-2n\delta^2_{n,m}}/2, \quad \sum_{k > \mathfrak{h}m} \lambda_n(k) \leq 2e^{-n\delta^2_{n,m}}. \tag{2.6}$$

(P2) (Second-step prior)

$$\Pi_{n,m}\left(\{f \in \mathcal{F}_m : d^2_n(f, f_{0,m}) \leq \delta^2_{n,m}/c_3\}\right) \geq e^{-2n\delta^2_{n,m}}. \tag{2.7}$$

Condition (P1) can be verified by using the following generic prior Λ_n:

$$\lambda_n(m) \propto \exp(-2n\delta^2_{n,m}). \tag{2.8}$$

Proposition 2.2. *Suppose the first condition of (2.5) holds. Then (P1) in Assumption C holds for the prior (2.8) with $\mathfrak{h}_0 \geq \mathfrak{h} \geq 2\mathfrak{c}^2$.*

(2.8) will be the model selection (first-step) prior on the model index $\mathcal{I}$ in all examples in Section 3.

Condition (P2) is reminiscent of the classical prior mass condition considered in [GGvdV00, GvdV07a]. Since $\delta_{n,m}^2$ is understood as the 'posterior contraction rate' for the model $\mathcal{F}_m$, (P2) can also be viewed as a *solvability condition* imposed on each model. Note that (2.7) only requires a sufficient prior mass on a Kullback-Leibler ball near $f_{0,m}$, where [GGvdV00, GvdV07a] use more complicated metric balls induced by higher moments of the Kullback-Leibler divergence.

2.2. Main abstract results

We say an index set $\mathcal{M} \subset \mathcal{I}$ *rectangular* if and only if there exist some integers $1 \leq a_k \leq b_k \leq \infty (k = 1, \ldots, q)$ such that $\mathcal{M} = \prod_{k=1}^q \{a_k, \ldots, b_k\}$.

Theorem 2.3. *Suppose Assumptions A-C hold for some rectangular $\mathcal{M} \subset \mathcal{I}$ with $\mathfrak{h} \geq C_0 \mathfrak{c}^2$ and $\mathfrak{h}_0 \geq C_0', d_0^2 \leq \inf_{m \in \mathcal{M}} \varepsilon_{n,m}^2 / C_0'$, where $\varepsilon_{n,m}^2 \equiv \inf_{g \in \mathcal{F}_m} d_n^2(f_0, g) \vee \delta_{n,m}^2$. Suppose d_n satisfies the triangle inequality. Then:*

1. *For any $m \in \mathcal{M}$,*

$$P_{f_0}^{(n)} \Pi_n \big(f \in \mathcal{F} : d_n^2(f, f_0) > C_1 \varepsilon_{n,m}^2 \big| X^{(n)} \big) \leq C_2 e^{-n \varepsilon_{n,m}^2 / C_2}. \tag{2.9}$$

2. *For any $m \in \mathcal{M}$ such that $\delta_{n,m}^2 \geq \inf_{g \in \mathcal{F}_m} d_n^2(f_0, g)$*[7]*,*

$$P_{f_0}^{(n)} \Pi_n \big(f \notin \mathcal{F}_{C_3 m} | X^{(n)} \big) \leq C_2 e^{-n \varepsilon_{n,m}^2 / C_2}. \tag{2.10}$$

3. *Let $\hat{f}_n \equiv \Pi_n(f | X^{(n)})$ be the posterior mean. If $\mathfrak{h}_0 = \infty$ and $d_n(\cdot, \cdot)$ is convex in each of its arguments, then*

$$P_{f_0}^{(n)} d_n^2(\hat{f}_n, f_0) \leq C_4 \inf_{m \in \mathcal{M}} \varepsilon_{n,m}^2. \tag{2.11}$$

Here the constant C_0 depends on $\{c_i\}_{i=1}^3, \kappa$ and $C_0', \{C_i\}_{i=1}^4$ depend on the $\{c_i\}_{i=1}^3, \kappa, \mathfrak{c}, \mathfrak{h}$ and γ.

Remark 2.4. Some technical comments:

1. $f_{0,m}$ in Assumptions B and C may be taken other than the minimizer of $f \mapsto d_n^2(f_0, f)$ over $\mathcal{F}_m$. In this case, the conclusion of the above theorems is valid by using $\varepsilon_{n,m}^2 \equiv d_n^2(f_0, f_{0,m}) \vee \delta_{n,m}^2$.
2. The constants $\{C_i\}_{i=0}^4$ do not depend on $m \in \mathcal{M}$, so the conclusions in (1)-(2) hold simultaneously for all $m \in \mathcal{M}$.

Theorem 2.3 shows that the task of constructing Bayes procedures *adaptive* to a collection of models in the intrinsic metric of a given statistical experiment, can be essentially reduced to that of designing a suitable *non-adaptive* prior for each model, provided the model selection prior is chosen according to (P1).

[7] We use the convention that $\mathcal{F}_m \equiv \mathcal{F}_{m \wedge b}$ where $b = (b_1, \ldots, b_q)$ where $\mathcal{M} = \prod_{k=1}^q \{a_k, \ldots, b_k\}$.

Furthermore, the resulting posterior mean serves as an automatic adaptive point estimator in a frequentist sense. Besides being rate-adaptive to the collection of models, (2.10) shows that the posterior distribution does not spread too much mass on overly large models. Results of this type have been derived primarily in the Gaussian regression model (cf. [CSHvdV15, CvdV12, GvdVZ15]) and in density estimation [GLvdV08]; here our result shows that this is a general phenomenon for the hierarchical prior design.

As mentioned in the Introduction, previous results [AGR13, dJvZ10, GvdVZ15, GLvdV08, RS17] require certain specific form of the prior, model structure, or the experiments. Our Theorem 2.3 can thus be viewed as a generalization of these results without such apriori requirements under a hierarchical prior design. As will be clear from concrete applications in Section 3, another advantage of the formulation of Theorem 2.3 is that Assumptions B-C typically concern *finite-dimensional* models $\mathcal{F}_m$ so verification is easy and routine.

Note that f_0 is arbitrary and hence our oracle inequalities (2.9) and (2.11) account for model mis-specification errors. Previous work allowing model mis-specification includes [GvdVZ15] who mainly focuses on structured linear models in the Gaussian regression setting, and [KvdV06] who pursued generality at the cost of harder-to-check conditions.

Remark 2.5. We make some technical remarks.

1. The probability estimate in (2.9) is of Gaussian type and is therefore sharp (up to constants) in view of the lower bound result Theorem 2.1 in [HRSH15]. Such sharp estimates have been derived separately in the Hellinger metric [GGvdV00], or in individual settings, e.g. the sparse normal mean model [CvdV12], the sparse PCA model [GZ15], and the structured linear model [GvdVZ15], to name a few. The Gaussian estimate naturally implies good behavior of the posterior mean under bounded metrics (cf. page 507 of [GGvdV00]). In the leading case $\mathfrak{c} = \gamma = 1, \mathfrak{h}_0 = \infty$ in Assumption B, the posterior mean $\hat{f}_n$ satisfies an oracle inequality with a Gaussian tail[8].
2. (2.10) asserts that the posterior distribution does not concentrate on overly large models. It is also of significant interest to assert the converse in some models, i.e. the posterior distribution does not concentrate on overly small models under additional problem-specific conditions. We refer to the readers to [Bel17, CSHvdV15, RS16, YP17] and references therein for more details in this direction.
3. Assumption A implies, among other things, the existence of a good test (cf. Lemma 2.1). In this sense our approach here falls into the general testing approach adopted in [GGvdV00, GvdV07a]. Some alternative approaches for dealing with non-intrinsic metrics can be found in [Cas14, HRSH15, YG16].
4. The constants $\{C_i\}_{i=1}^4$ in Theorem 2.3 depend at most polynomially with respect to the constants involved in Assumption A. This will be useful

[8] This can be seen by a simple modification of the proof by calculating the moment generating function.

in handling models where the local Gaussianity only holds locally on the parameter space (cf. Appendix F).

5. If d_n does not satisfy the triangle inequality, then (2.9) and (2.10) in Theorem 2.3 hold if $f_0 \in \mathcal{F}_m$ for some m (i.e. the form of an exact oracle inequality may be lost at a general level).

Remark 2.6. We compare our results with Theorems 4 and 5 of [YP17]. Both their results and our Theorem 2.3 shed light on the general problem of Bayes model selection, while differing in several important aspects:

1. Theorem 4 of [YP17] targets at exact model selection consistency, under a set of additional 'separation' assumptions. Our Theorem 2.3 (2) requires no extra assumptions, and shows that the posterior distribution does not concentrate on overly large models. This is significant in non-parametric problems: the true signal typically need not belong to any specific model.
2. Theorem 5 of [YP17] contains a term involving the cardinality of the models, so their bound will be finite only if there are finitely many models. It remains open to see if this can be removed.

2.3. The localization (sieving) principle

Consider a sequence of models $\{\bar{\mathcal{F}}_n\}$, where $\bar{\mathcal{F}}_n$ is regarded as the localized model of $\mathcal{F}$ at sample size n. Note that any prior Π_n on $\mathcal{F}$ can be localized to a prior $\bar{\Pi}_n$ on $\bar{\mathcal{F}}_n$: for any $B \subset \bar{\mathcal{F}}_n$, define $\bar{\Pi}_n(B) \equiv \Pi_n(B \cap \bar{\mathcal{F}}_n)/\Pi_n(\bar{\mathcal{F}}_n)$. Now the quantity in Theorem 2.3 concerning posterior distribution can be decomposed by

$$\begin{aligned}(2.12)\quad & P_{f_0}^{(n)}\Pi_n\big(f \in \mathcal{F} : d_n^2(f, f_0) > C_1\varepsilon_{n,m}^2 \big| X^{(n)}\big) \\ & \le P_{f_0}^{(n)}\bar{\Pi}_n\big(f \in \bar{\mathcal{F}}_n : d_n^2(f, f_0) > C_1\varepsilon_{n,m}^2 \big| X^{(n)}\big) + P_{f_0}^{(n)}\Pi_n\big(f \notin \bar{\mathcal{F}}_n \big| X^{(n)}\big).\end{aligned}$$

In essence, (2.12) suggests that we can use the machinery of Assumptions A-C to the localized model $\mathcal{F}_n$ (typically by choosing the constants c_2, c_3, d_0 depending on n), as long as the residue term $P_{f_0}^{(n)}\Pi_n\big(f \notin \bar{\mathcal{F}}_n \big| X^{(n)}\big)$ is well-controlled. This typically reduces to a reasonable control of $\Pi_n(\mathcal{F} \setminus \bar{\mathcal{F}}_n)$ (cf. Lemma 1 of [GvdV07a], see also examples in Appendix F). The localization principle is under the name 'sieving' in [GGvdV00, GvdV07a].

2.4. Proof sketch

Here we sketch the main steps in the proof of our main abstract result Theorem 2.3. The details will be deferred to Appendix A. The proof can be roughly divided into two main steps.

(**Step 1**) We first solve a localized problem on the model $\mathcal{F}_m$ by 'projecting' the underlying probability measure from P_{f_0} to $P_{f_{0,m}}$. In particular, we establish exponential deviation inequality for the posterior contraction rate via the existence of tests guaranteed by Lemma 2.1:

$$P_{f_{0,m}}^{(n)}\Pi_n\big(f\in\mathcal{F}: d_n^2(f,f_{0,m}))>M\delta_{n,\tilde{m}}^2|X^{(n)}\big)\lesssim e^{-c_1 n\delta_{n,\tilde{m}}^2}, \tag{2.13}$$

where $\tilde{m}$ is the smallest index $\geq m$ such that $\delta_{n,\tilde{m}}^2 \gtrsim d_n^2(f_0,f_{0,m})$. This index may deviate from m substantially for small indices.
(**Step 2**) We argue that, the cost of the projection in Step 1 is essentially a multiplicative $\mathcal{O}\big(\exp(c_2 n\delta_{n,\tilde{m}}^2)\big)$ factor in the probability bound (2.13), cf. Lemma A.1, which is made possible by the local Gaussianity Assumption A. Then by choosing c_1 much larger than c_2 we obtain the conclusion by the definition of $\delta_{n,\tilde{m}}^2$ and the fact that $\delta_{n,\tilde{m}}^2\approx d_n^2(f_0,f_{0,m})\vee\delta_{n,m}^2$.

The existence of tests (Lemma 2.1) is used in Step 1. Step 2 is inspired by the work of [CGS15] in the context of frequentist least squares estimator over a polyhedral cone in the Gaussian regression setting, where the localized problem therein is estimation of signals on a low-dimensional face (where 'risk adaptation' happens). In the Bayesian context, [CSHvdV15, CvdV12] used a change of measure argument in the Gaussian regression setting for a different purpose. Our proof strategy can be viewed as an extension of these ideas beyond the (simple) Gaussian regression model.

3. Models and applications

In this section we work out a couple of specific statistical models that satisfy the local Gaussianity Assumption A to illustrate the scope of the general results in Section 2. Some of the examples come from [GvdV07a]; we identify the 'intrinsic' metric to use in these models. Some concrete applications are also given. The applications presented in this section serve as a demonstration of the scope of our general results in deriving new contraction rate results. More applications can be found in Appendix F to illustrate the localization principle (cf. Section 2.3) and aid calculations/formulation in complicated list of models.

3.1. Regression models

Suppose we want to estimate $\theta=(\theta_1,\ldots,\theta_n)$ in a given model $\Theta_n\subset\mathbb{R}^n$ in the following settings: for $1\leq i\leq n$,

1. (*Gaussian*) $X_i=\theta_i+\varepsilon_i$ where ε_i's are i.i.d. $\mathcal{N}(0,1)$ and $\Theta_n\subset\mathbb{R}^n$;
2. (*Laplace*) $X_i=\theta_i+\varepsilon_i$ where ε_i's are i.i.d. errors with density $x\mapsto\frac{1}{2}e^{-|x|}$, and $\Theta_n\subset[-M,M]^n$;
3. (*Binary*) $X_i\sim\mathrm{Bern}(\theta_i)$ are independent, where $\Theta_n\subset[\eta,1-\eta]^n$ for some $\eta>0$;
4. (*Poisson*) $X_i\sim_{\text{i.i.d.}}\mathrm{Poisson}(\theta_i)$ where $\Theta_n\subset[1/M,M]^n$ for some $M\geq 1$;

For any $\theta_0,\theta_1\in\Theta_n$, $\ell_n^2(\theta_0,\theta_1)\equiv n^{-1}\sum_{i=1}^n\big(\theta_{0,i}-\theta_{1,i}\big)^2$.

Lemma 3.1. *Assumption A holds for ℓ_n with*

1. *(Gaussian)* $c_1 = c_2 = c_3 = \kappa_g = 1$ *and* $\kappa_\Gamma = 0$*;*
2. *(Laplace)* $\kappa_\Gamma = 0$*,* κ_g *an absolute constant and constants* $\{c_i\}_{i=1}^3$ *depending on* M *only;*
3. *(Binary)* $\kappa_\Gamma = 0$ *and the constants* $\{c_i\}_{i=1}^3, \kappa_g$ *depend on* η *only;*
4. *(Poisson) constants* $\{c_i\}_{i=1}^3, \kappa$ *depending on* M *only.*

Corollary 3.2. *For Gaussian/Laplace/binary/Poisson regression models, let* $d_n \equiv \ell_n$*. If Assumptions B-C hold, then (2.9)-(2.11) hold.*

Using similar techniques we can derive analogous results for Gaussian regression with random design and white noise model. We omit the details.

Remark 3.3. The boundedness assumption in Laplace/binary/Poisson models is imposed here for simplicity, and can be removed using the localization principle (cf. Section 2.3) for more concrete Θ_n's and priors. See Appendix F for an example.

Below we give three concrete applications in the Gaussian regression model $y_i = f_0(x_i) + \varepsilon_i (1 \leq i \leq n)$, where ε_i's are i.i.d. $\mathcal{N}(0,1)$. We slightly abuse ℓ_n to denote $\ell_n^2(f,g) \equiv n^{-1}\sum_{i=1}^n (f(x_i) - g(x_i))^2$.

3.1.1. Example: Trace regression

Consider fitting the Gaussian regression model $y_i = f_0(x_i) + \varepsilon_i (1 \leq i \leq n)$ by $\mathcal{F} \equiv \{f_A : A \in \mathbb{R}^{m_1 \times m_2}\}$ where $f_A(x) = \mathrm{tr}(x^\top A)$ for all $x \in \mathfrak{X} \equiv \mathbb{R}^{m_1 \times m_2}$. Let $\underline{m} \equiv m_1 \wedge m_2$ and $\bar{m} \equiv m_1 \vee m_2$. The index set is $\mathcal{I} = \mathcal{I}_1 \cup \mathcal{I}_2 \equiv \{1, \ldots, r_{\max}\} \cup \{r_{\max} + 1, \ldots\} = \mathbb{N}$ where $r_{\max} \leq \underline{m}$. For $r \in \mathcal{I}_1$, let $\mathcal{F}_r \equiv \{f_A : A \in \mathbb{R}^{m_1 \times m_2}, \mathrm{rank}(A) \leq r\}$, and for $r \in \mathcal{I}_2$, $\mathcal{F}_r \equiv \mathcal{F}_{r_{\max}}$[9].

Although various Bayesian methods have been proposed in the literature (cf. see [ACCR14] for a state-to-art summary), theoretical understanding has been limited. [MA15] derived an oracle inequality for an exponentially aggragated estimator for the matrix completion problem. Their result is purely frequentist. Below we consider a two step prior similar to [ACCR14, MA15], and derive the corresponding posterior contraction rates.

For a matrix $B = (b_{ij}) \in \mathbb{R}^{m_1 \times m_2}$ let $\|B\|_p$ denote its Schatten p-norm[10]. $p = 1$ and 2 correspond to the *nuclear norm* and the *Frobenius norm* respectively. To introduce the notion of RIP, let $\mathcal{X} : \mathbb{R}^{m_1 \times m_2} \to \mathbb{R}^n$ be the linear map defined via $A \mapsto (\mathrm{tr}(x_i^\top A))_{i=1}^n$.

Definition 3.4. The linear map $\mathcal{X} : \mathbb{R}^{m_1 \times m_2} \to \mathbb{R}^n$ is said to satisfy *RIP*$(r, \boldsymbol{\nu}_r)$ for some $1 \leq r \leq r_{\max}$ and some $\boldsymbol{\nu}_r = (\underline{\nu}_r, \bar{\nu}_r)$ with $0 < \underline{\nu}_r \leq \bar{\nu}_r < \infty$ iff $\underline{\nu}_r \leq \frac{\|\mathcal{X}(A)\|_2}{\sqrt{n}\|A\|_2} \leq \bar{\nu}_r$ holds for all matrices $A \in \mathbb{R}^{m_1 \times m_2}$ such that $\mathrm{rank}(A) \leq r$. For $r > r_{\max}$, $\mathcal{X}$ satisfies $\mathrm{RIP}(r, \boldsymbol{\nu}_r)$ iff $\mathcal{X}$ satisfies $\mathrm{RIP}(r_{\max}, \boldsymbol{\nu}_r)$. Furthermore, $\mathcal{X} : \mathbb{R}^{m_1 \times m_2} \to \mathbb{R}^n$ is said to satisfy *uniform RIP* $(\boldsymbol{\nu}; \mathcal{I})$ on an index set $\mathcal{I}$ iff $\mathcal{X}$ satisfies $\mathrm{RIP}(2r, \boldsymbol{\nu})$ for all $r \in \mathcal{I}$.

[9] This trick of defining models for high-dimensional experiments will also used in other applications in later subsections, but we will not explicitly state it again.

[10] That is, $\|B\|_p \equiv \left(\sum_{j=1}^{\underline{m}} \sigma_j(B)^p\right)^{1/p}$, where $\{\sigma_j(B)\}$ are the singular values of B.

RIP$(r, \boldsymbol{\nu}_r)$ is a variant of the RIP condition introduced in [CT05, CP11, RFP10] with scaling factors $\bar{\nu}_r = 1/(1-\delta_r)$ and $\underline{\nu}_r = 1/(1+\delta_r)$ for some $0 < \delta_r < 1$. This condition quantifies the degree in which the linear map $\mathcal{X}$ behaves like an isometry between $\mathbb{R}^{m_1\times m_2}$ and $\mathbb{R}^n$ in terms of the ℓ_2 metric. Below are two canonical examples.

Example 3.5 (Matrix completion)**.** Suppose that $x_i \in \mathbb{R}^{m_1\times m_2}$ takes value 1 at one position and 0 otherwise. Further assume that $\underline{A} \leq |A_0|_{ij} \leq \bar{A}$ for all $1 \leq i \leq m_1$ and $1 \leq j \leq m_2$[11]. Let $\Omega \equiv \Omega_{\mathcal{X}}$ denote the indices for which $\{x_i\}$'s take value 1. Then $\|\mathcal{X}(A)\|_2 = \|A\mathbf{1}_\Omega\|_2$. Easy calculations show that we can take $\boldsymbol{\nu} = (\bar{\nu}, \underline{\nu})$ defined by $\bar{\nu} = (\bar{A}\sqrt{m_1 m_2 \wedge n})/(\underline{A}\sqrt{n m_1 m_2})$, $\underline{\nu} = (\underline{A}\sqrt{m_1 m_2 \wedge n})/(\bar{A}\sqrt{n m_1 m_2})$ so that $\mathcal{X}$ is uniform RIP$(\boldsymbol{\nu}; \mathcal{I})$.

Example 3.6 (Gaussian measurement ensembles)**.** Suppose x_i's are i.i.d. random matrices whose entries are i.i.d. standard normal. Theorem 2.3 of [CP11] entails that $\mathcal{X}$ is uniform RIP$(\boldsymbol{\nu}; \mathcal{I})$ with $\bar{\nu} = 1+\delta$, $\underline{\nu} = 1-\delta$ for some $\delta \in (0,1)$, with probability at least $1 - C\exp(-cn)$[12], provided $n \gtrsim \bar{m} r_{\max}$.

Consider a prior Λ_n on $\mathcal{I}$ of form

$$\lambda_n(r) \propto \exp\big(-c \cdot (m_1+m_2) r \log \bar{m}\big), \tag{3.1}$$

where $c > 0$ is a constant to be specified later. Given the chosen index $r \in \mathcal{I}_1$, a prior on $\mathcal{F}_r$ is induced by a prior on all $m_1 \times m_2$ matrices of form $\sum_{i=1}^r u_i v_i^\top$ where $u_i \in \mathbb{R}^{m_1}$ and $v_i \in \mathbb{R}^{m_2}$. Here we use a product prior distribution G with Lebesgue density $(g_1 \otimes g_2)^{\otimes r}$ on $(\mathbb{R}^{m_1} \times \mathbb{R}^{m_2})^r$. For simplicity we use $g_i \equiv g^{\otimes m_i}$ for $i = 1, 2$ where g is symmetric about 0 and non-increasing on $(0,\infty)$[13]. Let $\tau^{\mathrm{tr}}_{r,g} \equiv \sup_{A_{0,r} \in \arg\min_{B:\mathrm{rank}(B)\leq r} \ell_n^2(f_B, f_0)} g\big(\sigma_{\max}(A_{0,r}) + 1\big)$ where $\sigma_{\max}$ denotes the largest singular value.

Theorem 3.7. *Fix $0 < \eta < 1/2$ and $r_{\max} \leq n$. Suppose that there exists some $\mathcal{M} \subset \mathcal{I}_1$ such that the linear map $\mathcal{X} : \mathbb{R}^{m_1\times m_2} \to \mathbb{R}^n$ satisfies uniform RIP$(\boldsymbol{\nu}; \mathcal{M})$, and that for all $r \in \mathcal{M}$, we have*

$$\tau^{\mathrm{tr}}_{r,g} \geq e^{-\log \bar{m}/(2\eta)}, \quad \bar{m} \geq 3 \vee \big(2\bar{\nu}(1 \vee \sigma_{\max}(A_{0,r}))n^2\big)^{2\eta}. \tag{3.2}$$

Then there exists some $c > 0$ in (3.1) depending on $\bar{\nu}/\underline{\nu}, \eta$ such that for any $r \in \mathcal{M}$,

$$P_{f_0}^{(n)} \Pi_n\big(A \in \mathbb{R}^{m_1\times m_2} : \ell_n^2(f_A, f_0) > C_1 (\varepsilon^{\mathrm{tr}}_{n,r})^2 \big| Y^{(n)}\big) \leq C_2 e^{-n(\varepsilon^{\mathrm{tr}}_{n,r})^2/C_2}. \tag{3.3}$$

Here $(\varepsilon^{\mathrm{tr}}_{n,r})^2 \equiv \max\{\inf_{B:\mathrm{rank}(B)\leq r} \ell_n^2(f_0, f_B), (m_1+m_2) r \log \bar{m}/n\}$, and the constants $C_i (i = 1, 2)$ depend on $\bar{\nu}/\underline{\nu}, \eta$.

[11] This assumption is usually satisfied in applications: in fact in the Netflix problem (which is the main motivating example for matrix completion), A_0 is the rating matrix with rows indexing the users and columns indexing movies, and we can simply take $\underline{A} = 1$ (one star) and $\bar{A} = 5$ (five stars).

[12] Note here we used the union bound to get a probability estimate $r_{\max}\exp(-cn) \lesssim \exp(-c'n)$ for some $c' < c$ under the assumption that $n \gtrsim \bar{m} r_{\max}$.

[13] We will always use such g in the prior design in the examples in this section.

By Theorem 5 of [RT11], the rate in (3.3) is minimax optimal up to a logarithmic factor. To the best knowledge of the author, the theorem above is the first result in the literature that addresses the posterior contraction rate in the context of trace regression in a fully Bayesian setup.

(3.2) may be verified in a case-by-case manner; or generically we can take $\mathcal{M} = \{r_0, r_0 + 1, \ldots\}$ if the model is well specified, at the cost of sacrificing the form of oracle inequalities (but still get nearly optimal posterior contraction rates) in (3.3). In particular, the first condition of (3.2) prevents the largest eigenvalue of $A_{0,r}$ from growing too fast. This is in similar spirit with Theorem 2.8 of [CvdV12], showing that the magnitude of the signals cannot be too large for light-tailed priors to work in the sparse normal mean model. The second condition of (3.2) is typically a mild technical condition: we only need to choose $\eta > 0$ small enough.

3.1.2. Example: Isotonic regression

Consider the isotonic regression model $Y_i = f_0(x_i) + \varepsilon_i$ by $\mathcal{F} \equiv \{f : [0,1] \to \mathbb{R} : f \text{ is non-decreasing}\}$. For simplicity the design points are assumed to be $x_i = i/(n+1)$ for all $1 \le i \le n$. Bayesian approaches for the isotonic regression model received considerable attention, cf. [HH03, SSW09, ND04, LD14, Sal14]. Let $\mathcal{F}_m \equiv \big\{f \in \mathcal{F}, f \text{ is piecewise constant with at most } m \text{ constant pieces}\big\}$. Consider the following prior Λ_n on $\mathcal{I} = \mathbb{N}$:

$$\lambda_n(m) \propto \exp\big(-c \cdot m \log(en)\big), \tag{3.4}$$

where $c > 0$ is a constant to be specified later. Let $g_m \equiv g^{\otimes m}$ where g is symmetric and non-increasing on $(0, \infty)$. Then $\bar{g}_m(\boldsymbol{\mu}) \equiv m! g_m \mathbf{1}_{\{\mu_1 \le \ldots \le \mu_m\}}(\boldsymbol{\mu})$ is a valid density on $\{\mu_1 \le \ldots \le \mu_m\}$. Given a chosen model $\mathcal{F}_m$ by the prior Λ_n, we randomly pick a set of change points $\{x_{i(k)}\}_{k=1}^m (i(1) < \ldots < i(m))$ and put a prior $\bar{g}_m$ on $\{f(x_{i(k)})\}$'s. [HH03] proposed a similar prior with Λ_n being uniform since they assumed the maximum number of change points is known *apriori*. Below we derive a theoretical result without assuming the knowledge of this. Let $\tau_{m,g}^{\mathrm{iso}} = \sup_{f_{0,m} \in \arg\min_{g \in \mathcal{F}_m} \ell_n^2(f_0, g)} g\big(\|f_{0,m}\|_\infty + 1\big)$[14].

Theorem 3.8. *Fix $0 < \eta < 1/2$. Suppose that*

$$\tau_{m,g}^{\mathrm{iso}} \ge e^{-\log(en)/(2\eta)}. \tag{3.5}$$

Then there exists some $c > 0$ in (3.4) depending on η such that

$$P_{f_0}^{(n)} \Pi_n\big(f \in \mathcal{F} : \ell_n^2(f, f_0) > C_1(\varepsilon_{n,m}^{\mathrm{iso}})^2 \big| Y^{(n)}\big) \le C_2 e^{-n(\varepsilon_{n,m}^{\mathrm{iso}})^2/C_2}. \tag{3.6}$$

Here $(\varepsilon_{n,m}^{\mathrm{iso}})^2 \equiv \max\{\inf_{g \in \mathcal{F}_m} \ell_n^2(f_0, g), m \log(en)/n\}$, and the constants $C_i (i = 1, 2)$ depend on η.

[14] The value of $f_{0,m}$ outside of $[1/(n+1), n/(n+1)]$ can be defined in a canonical way by extending $f_{0,m}(1/(n+1))$ and $f_{0,m}(n/(n+1))$ towards the endpoints.

(3.6) implies that if f_0 is piecewise constant, the posterior distribution contracts at nearly a parametric rate. For general isotonic signals $f_0 \in \mathcal{F}$ with $\|f_0\|_\infty < \infty$, by using Theorem 4.1 of [CGS15], we obtain a contraction rate on the order of $n^{-2/3}\log(en)$ in ℓ_n^2. (3.5) can be checked by the following.

Lemma 3.9. *If f_0 is square integrable, and the prior density g is heavy-tailed in the sense that there exists some $\alpha > 0$ such that $\liminf_{|x|\to\infty} x^\alpha g(x) > 0$. Then for any $\eta \in (0, 1/\alpha)$, (3.5) holds uniformly in all $m \in \mathbb{N}$ for n large enough depending on α and $\|f_0\|_{L_2([0,1])}$.*

3.1.3. Example: Convex regression

Consider fitting the Gaussian regression model $Y_i = f_0(x_i) + \varepsilon_i$ by $\mathcal{F}$, the class of convex functions on $\mathfrak{X} = [0,1]^d$. Let $\mathcal{F}_m \equiv \big\{f(x) = \max_{1\le i\le m}(a_i \cdot x + b_i) : a_i \in \mathbb{R}^d, b_i \in \mathbb{R}\big\}$ denote the class of piecewise affine convex functions with at most m pieces.

We will focus on the multivariate case since the univariate case can be easily derived using the techniques exploited in isotonic regression. A prior on each model $\mathcal{F}_m$ can be induced by a prior on the slopes and the intercepts $\{(a_i, b_i) \in \mathbb{R}^d \times \mathbb{R}\}_{i=1}^m$. We use a prior with density $\bigotimes_{i=1}^m g^{\otimes d} \otimes g$ on $(\mathbb{R}^d \times \mathbb{R})^m$ to induce a prior on $\mathcal{F}_m$. For any $f_{0,m} \in \arg\min_{g\in\mathcal{F}_m} \ell_n^2(f_0, g)$, it can be represented as $f_{0,m}(x) \equiv \max_{1\le i\le m}\big(a_i^{(m)} \cdot x + b_i^{(m)}\big)$. Let $\tau_{m,g}^{\mathrm{cvx}} \equiv \sup_{f_{0,m}\in\arg\min_{g\in\mathcal{F}_m}\ell_n^2(f_0,g)} \min_{1\le i\le m}\big\{g\big(\|a_i^{(m)}\|_\infty + 1\big), g\big(|b_i^{(m)}| + 1\big)\big\}$.

The prior Λ_n we will use on the index $\mathcal{I} = \mathbb{N}$ is given by

$$\lambda_n(m) \propto \exp\big(-c \cdot dm\log 3m \cdot \log n\big), \tag{3.7}$$

where $c > 0$ is a constant to be specified later. The first step prior used in [HD11] is a Poisson proposal, which slightly differs from (3.7) by a logarithmic factor. This would affect the contraction rate only by a logarithmic factor.

Theorem 3.10. *Fix $0 < \eta < 1/4$. Suppose that*

$$\tau_{m,g}^{\mathrm{cvx}} \ge e^{-\log n\cdot\log 3m/8\eta}, \tag{3.8}$$

and $n \ge d$. Then there exists some $c > 0$ in (3.7) depending on η such that

$$P_{f_0}^{(n)}\Pi_n\big(f \in \mathcal{F} : \ell_n^2(f, f_0) > C_1(\varepsilon_{n,m}^{\mathrm{cvx}})^2\big|Y^{(n)}\big) \le C_2 e^{-n(\varepsilon_{n,m}^{\mathrm{cvx}})^2/C_2}. \tag{3.9}$$

Here $(\varepsilon_{n,m}^{\mathrm{cvx}})^2 \equiv \max\{\inf_{g\in\mathcal{F}_m}\ell_n^2(f_0,g), d\log n \cdot m\log 3m/n\}$, and the constants $C_i (i = 1, 2)$ depend on η.

The above oracle inequality shows that the posterior contraction rate of [HD11] (Theorem 3.3 therein) is far from optimal. (3.8) can be satisfied by using heavy-tailed priors $g(\cdot)$ in the same spirit as Lemma 3.9—if f_0 is square integrable and the design points are regular enough (e.g. using regular grids on $[0,1]^d$). Explicit rates can be obtained using approximation techniques in

[HW16]. Using the same proof as Lemma 4.10 therein, if f_0 is Lipschitz, the contraction rate in ℓ_2^2 becomes the familiar one in the sense that $\inf_{m\in\mathbb{N}}(\varepsilon_{n,m}^{\mathrm{cvx}})^2 \lesssim \inf_{m\in\mathbb{N}}\max\{m^{-4/d}, \log n\cdot m\log 3m/n\} \asymp (\log^2 n/n)^{4/(d+4)}$.

Remark 3.11. For univariate convex regression, the term $\log(3m)$ in (3.7)-(3.9) can be removed. The logarithmic term is due to the fact that the *pseudo-dimension* of $\mathcal{F}_m$ scales as $m\log(3m)$ for $d\geq 2$, cf. Lemma C.9.

Remark 3.12. Using similar priors and proof techniques we can construct a (nearly) rate-optimal adaptive Bayes estimator for the support function regression problem for convex bodies [Gun12]. There the models $\mathcal{F}_m$ are support functions indexed by polytopes with m vertices, and a prior on $\mathcal{F}_m$ is induced by a prior on the location of the m vertices. The pseudo-dimension of $\mathcal{F}_m$ can be controlled using techniques developed in [Gun12]. Details are omitted.

3.1.4. Example: High-dimensional partially linear model

Consider fitting the Gaussian regression model $Y_i = f_0(x_i,z_i)+\varepsilon_i$ where $(x_i,z_i)\in\mathbb{R}^p\times[0,1]$, by a partially linear model $\mathcal{F}\equiv\{f_{\beta,u}(x,z) = x^\top\beta+u(z)\equiv h_\beta(x)+u(z):\beta\in\mathbb{R}^p, u\in\mathcal{U}\}$ where the dimension of the parametric part can diverge. We consider $\mathcal{U}$ to be the class of non-decreasing functions as an illustration (cf. Section 3.1.2). Consider models $\mathcal{F}_{(s,m)}\equiv\{f_{\beta,u}:\beta\in B_0(s), u\in\mathcal{U}_m\}$ where $\mathcal{U}_m$ denotes the class of piecewise constant non-decreasing functions with at most m constant pieces, and $B_0(s)\equiv\{v\in\mathbb{R}^p: |\mathrm{supp}(v)|\leq s\}$. In this example the model index $\mathcal{I}$ is a 2-dimensional lattice. Our goal here is to construct an estimator that satisfies an oracle inequality over the models $\{\mathcal{F}_{(s,m)}\}_{(s,m)\in\{1,\ldots,p\}\times\{1,\ldots,n\}}$. Consider the following model selection prior:

$$\lambda_n((s,m)) \propto \exp\big(-c\cdot(s\log(ep)\wedge \mathrm{rank}(X) + m\log(en))\big), \tag{3.10}$$

where $c>0$ is a constant to be specified later. Here $X\in\mathbb{R}^{n\times p}$ is the design matrix so that $X^\top X/n$ is normalized with diagonal elements taking value 1[15]. For a chosen model $\mathcal{F}_{(s,m)}$, consider the following prior $\Pi_{n,(s,m)}$: pick randomly a support $S\subset\{1,\ldots,p\}$ with $|S|=s$ and a set of change points $Q\equiv\{z_{i(k)}\}_{k=1}^m(i(1)<\ldots i(m))$, and then put a prior $g_{S,Q}$ on β_S and $u(z_{i(k)})$'s. For simplicity we use a product prior $g_{S,Q}\equiv g^{\otimes s}\otimes\bar{g}_m$ where $\bar{g}_m$ is a prior on $\{\mu_1\leq\ldots\leq\mu_m\}\subset\mathbb{R}^m$ constructed in Section 3.1.2. For any $f_{0,(s,m)}\in \inf_{g\in\mathcal{F}_{(s,m)}}\ell_n^2(f_0,g)$, write $f_{0,(s,m)}(x,z) = x^\top\beta_{0,s}+u_{0,m}(z)\equiv h_{0,s}(x)+u_{0,m}(z)$. Let $\tau_{m,g}\equiv \sup\limits_{f_{0,(s,m)}\in\inf_{g\in\mathcal{F}_{(s,m)}}\ell_n^2(f_0,g)} g(\lVert u_{0,m}\rVert_\infty+1)$.

Theorem 3.13. *Fix $0<\eta<1/4$. Suppose $p\geq n$ and $L>\big(\log(ep)/\mathrm{rank}(X)\big)\vee \inf\limits_{f_{0,(s,m)}\in\inf_{g\in\mathcal{F}_{(s,m)}}\ell_n^2(f_0,g)}\lVert\beta_{0,s}\rVert_\infty\vee\sigma_{\max}(X)$ for some $L>0$. Suppose that*

$$g(L+1) > e^{-\log(ep)/2\eta}, \quad \tau_{m,g}\geq e^{-\log(en)/(2\eta)}. \tag{3.11}$$

[15] This is a common assumption, cf. Section 6.1 of [BvdG11].

Then there exists some $c > 0$ in (3.10) depending on η, L such that

$$P_{f_0}^{(n)}\Pi_n\big(f \in \mathcal{F} : \ell_n^2(f, f_0) > C_1(\varepsilon_{n,(s,m)}^{\mathrm{hp}})^2 \big| Y^{(n)}\big) \leq C_2 e^{-n(\varepsilon_{n,(s,m)}^{\mathrm{hp}})^2/C_2}. \tag{3.12}$$

Here $(\varepsilon_{n,(s,m)}^{\mathrm{hp}})^2 \equiv \max\{\inf_{f_{\beta,u} \in \mathcal{F}_{(s,m)}} \ell_n^2(f_0, f_{\beta,u}), (s\log(ep) \wedge \mathrm{rank}(X) + m\log(en))/n\}$*, and the constants $C_i (i = 1, 2)$ depend on η, L.*

The condition $p \geq n$ can be replaced by $p \geq n^\delta$ for any $\delta > 0$ by changing the constants. $L > 0$ prevents p, $\|\beta_{0,s}\|_\infty$ and the maximal singular value of X from being too large. The second condition of (3.11) is the same as in (3.5) (so in particular can be checked using Lemma 3.9). When the model is well-specified in the sense that $f_0(x, z) = x^\top \beta_0 + u_0(z)$ for some $\beta_0 \in B_0(s_0)$ and $u_0 \in \mathcal{U}$, the oracle rate in (3.12) becomes

$$\frac{s_0 \log(ep) \wedge \mathrm{rank}(X)}{n} + \inf_{m\in\mathbb{N}} \bigg(\inf_{u\in\mathcal{U}_m} \ell_n^2(u_0, u) + \frac{m\log(en)}{n}\bigg). \tag{3.13}$$

The two terms in the rate (3.13) trades off two structures of the experiment: the sparsity of $h_\beta(x)$ and the smoothness level of $u(z)$. The resulting phase transition of the rate (3.13) in terms of these structures is in a sense similar to the results of [YLC19, YZ16]. It is also easy to derive some explicit rate results from (3.13). For instance, if $u_0 \in \mathcal{U}$ and $\|u_0\|_\infty < \infty$, then by using Theorem 4.1 of [CGS15], (3.13) reduces to $(s_0 \log(ep) \wedge \mathrm{rank}(X))/n + n^{-2/3}\log(en)$.

3.2. Density estimation

Suppose $X_1, \ldots, X_n$'s are i.i.d. samples from a density $f \in \mathcal{F}$ with respect to a measure ν on the sample space $(\mathfrak{X}, \mathcal{A})$. We consider the following form of $\mathcal{F}$: $f(x) = e^{g(x)}/\int_{\mathfrak{X}} e^g \,\mathrm{d}\nu$ for some $g \in \mathcal{G}$ for all $x \in \mathfrak{X}$. For any $f_0, f_1 \in \mathcal{F}$, $h^2(f_0, f_1) \equiv \frac{1}{2}\int_{\mathfrak{X}}(\sqrt{f_0} - \sqrt{f_1})^2 \,\mathrm{d}\nu$.

Lemma 3.14. *Suppose that $\mathcal{G}$ is uniformly bounded. Then Assumption A is satisfied for h with constants $\{c_i\}_{i=1}^3, \kappa$ depending on $\mathcal{G}$ only.*

Corollary 3.15. *For density estimation, let $d_n \equiv h$. If $\mathcal{G}$ is a class of uniformly bounded functions and Assumptions B-C hold, then (2.9)-(2.11) hold.*

Remark 3.16. Similar to the above remark, the uniform boundedness is included here for simplicity. See Appendix F for an example on location mixture model where this restriction is removed.

3.3. Gaussian autoregression

Suppose $X_0, X_1, \ldots, X_n$ is generated from $X_i = f(X_{i-1}) + \varepsilon_i$ for $1 \leq i \leq n$, where f belongs to a function class $\mathcal{F}$ with a uniform bound M, and ε_i's are i.i.d. $\mathcal{N}(0, 1)$. Then X_n is a Markov chain with transition density $p_f(y|x) = \phi(y - f(x))$ where ϕ is the normal density. By the arguments on page 209 of

[GvdV07a], this chain has a unique stationary distribution with density q_f with respect to the Lebesgue measure λ on $\mathbb{R}$. We assume that X_0 is generated from this stationary distribution under the true f. For any $f_0, f_1 \in \mathcal{F}$, $d^2_{r,M}(f_0, f_1) \equiv \int (f_0 - f_1)^2 r_M \,\mathrm{d}\lambda$ where $r_M(x) \equiv \frac{1}{2}\left(\phi(x-M) + \phi(x+M)\right)$.

Lemma 3.17. *Suppose that $\mathcal{F}$ is uniformly bounded by M. Then Assumption A is satisfied for $d_{r,M}$ with constants $\{c_i\}_{i=1}^3, \kappa$ depending on M only.*

Corollary 3.18. *For Gaussian autoregression model, if $\mathcal{F}$ is uniformly bounded by M, let $d_n \equiv d_{r,M}$. If Assumptions B-C hold, then (2.9)-(2.11) hold.*

[GvdV07a] (cf. Section 7.4) uses a weighted $L_s(s > 2)$ norm to check the local entropy condition, and an average Hellinger metric as the loss function. Our results here use the metric $d_{r,M}$ defined as a weighted L_2 norm.

3.4. Gaussian time series

Suppose $X_1, X_2, \ldots$ is a stationary Gaussian process with spectral density $f \in \mathcal{F}$ defined on $[-\pi, \pi]$. Then the covariance matrix of $X^{(n)} = (X_1, \ldots, X_n)$ is given by $(T_n(f))_{kl} \equiv \int_{-\pi}^{\pi} e^{\sqrt{-1}\lambda(k-l)} f(\lambda) \,\mathrm{d}\lambda$. We consider a special form of $\mathcal{F}$: $f \equiv f_g \equiv e^g$ for some $g \in \mathcal{G}$. For any $g_0, g_1 \in \mathcal{G}$, $D_n^2(g_0, g_1) \equiv n^{-1}\|T_n(f_{g_0}) - T_n(f_{g_1})\|_F^2$, where $\|\cdot\|_F$ denotes the matrix Frobenius norm.

Lemma 3.19. *Suppose that $\mathcal{G}$ is uniformly bounded. Then Assumption A is satisfied for D_n with constants $\{c_i\}_{i=1}^3, \kappa$ depending on $\mathcal{G}$ only.*

Corollary 3.20. *For the Gaussian time series model, if $\mathcal{G}$ is uniformly bounded, let $d_n \equiv D_n$. If Assumptions B-C hold, then (2.9)-(2.11) hold.*

D_n is bounded from above by the usual L_2 metric, and can be related to the L_2 metric from below (cf. Lemma B.3 of [GZ16]). Our result then shows that the metric to use in the entropy condition can be weakened to the L_2 norm rather than the much stronger L_∞ norm as in page 202 of [GvdV07a]. Such improvements are particularly important in, e.g. shape constrained models that are not totally bounded in L_∞ (cf. [GS13]). See also [CGR04, RCL12] for some related works in Bayesian spectral density estimation.

3.5. Covariance matrix estimation

Suppose $X_1, \ldots, X_n \in \mathbb{R}^p$ are i.i.d. observations from $\mathcal{N}_p(0, \Sigma)$ where $\Sigma \in \mathscr{S}_p(L)$, the set of $p \times p$ covariance matrices whose minimal and maximal eigenvalues are bounded by L^{-1} and L (where $L > 1$), respectively. For any $\Sigma_0, \Sigma_1 \in \mathscr{S}_p(L)$, $D_F^2(\Sigma_0, \Sigma_1) \equiv \|\Sigma_0 - \Sigma_1\|_F^2$.

Lemma 3.21. *Under the above setting, Assumption A holds for the metric D_F with constants $\{c_i\}_{i=1}^3, \kappa$ depending on L only.*

Corollary 3.22. *For covariance matrix estimation in $\mathscr{S}_p(L)$ for some $L < \infty$, let $d_n \equiv D_F$. If Assumptions B-C hold, then (2.9)-(2.11) hold.*

3.5.1. Example: Covariance matrix estimation in the sparse factor model

Suppose we observe i.i.d. $X_1,\ldots,X_n \in \mathbb{R}^p$ from $\mathcal{N}_p(0,\Sigma_0)$. The covariance matrix is modelled by the sparse factor model $\mathfrak{M} \equiv \cup_{(k,s)\in\mathbb{N}^2}\mathfrak{M}_{(k,s)}$ where $\mathfrak{M}_{(k,s)} \equiv \{\Sigma = \Lambda\Lambda^\top + I : \Lambda \in \mathscr{R}_{(k,s)}(L)\}$ with $\mathscr{R}_{(k,s)}(L) \equiv \{\Lambda \in \mathbb{R}^{p\times k}, \Lambda_{\cdot j} \in B_0(s), |\sigma_j(\Lambda)| \leq L^{1/2}, \forall 1 \leq j \leq k\}$. In this example, the model index $\mathcal{I}$ is a 2-dimensional lattice, and the sparsity structure depends on the rank structure. Consider the following model selection prior:

$$\lambda_n((k,s)) \propto \exp\left(-c\cdot ks\log(ep)\right), \tag{3.14}$$

where $c > 0$ is a constant to be specified later.

Theorem 3.23. *Let $p \geq n$. There exist some $c > 0$ in (3.14) and some sequence of sieve priors $\Pi_{n,(k,s)}$ on $\mathfrak{M}_{(k,s)}$ depending on L such that*

$$P^{(n)}_{\Sigma_0}\Pi_n\big(\Sigma\in\mathfrak{M} : \|\Sigma-\Sigma_0\|_F^2 > C_1(\varepsilon^{\mathrm{cov}}_{n,(k,s)})^2\big|X^{(n)}\big) \leq C_2 e^{-n(\varepsilon^{\mathrm{cov}}_{n,(k,s)})^2/C_2}.$$

Here $(\varepsilon^{\mathrm{cov}}_{n,(k,s)})^2 \equiv \max\{\inf_{\Sigma'\in\mathfrak{M}_{(s,k)}}\|\Sigma'-\Sigma_0\|_F^2, ks\log(ep)/n\}$, and the constants $C_i (i=1,2)$ depend on L.

Since spectral norm (non-intrinsic) is dominated by Frobenius norm (intrinsic), our result shows that if the model is well-specified (i.e. $\Sigma_0 \in \mathfrak{M}$), then we can construct an adaptive Bayes estimator with convergence rates in both norms no worse than $\sqrt{ks\log p/n}$. [PBPD14] considered the same sparse factor model, where they proved a strictly sub-optimal rate $\sqrt{k^3 s\log p\log n/n}$ in spectral norm under $ks \gtrsim \log p$. [GZ15] considered a closely related sparse PCA problem, where the convergence rate under spectral norm achieves the same rate as here (cf. Theorem 4.1 therein), while a factor of $\sqrt{k}$ is lost when using Frobenius norm as a loss function (cf. Remark 4.3 therein).

It should be mentioned that the sieve prior $\Pi_{n,(k,s)}$ is constructed using the metric entropy of $\mathfrak{M}_{(k,s)}$ and hence the resulting Bayes estimator and the posterior mean as a point estimator are purely theoretical. We use this example to illustrate (i) the construction scheme of a (nearly) optimal adaptive procedure for a multi-structured experiment based on the metric entropy of the underlying parameter space, and (ii) derivation of contraction rates in non-intrinsic metrics when these metrics can be related to the intrinsic metrics nicely.

It is also possible to use similar strategies as above in the closely related problem of estimating a sparse precision matrix (cf. [BG15]), but we refrain from repetitive details here.

3.6. Image boundary detection

Consider the setup in [LG17] as follows. Let $\{f(\cdot;\phi) : \phi \in \mathbb{R}^p\}$ be a class of densities dominated by a σ-finite measure μ and indexed by a p-dimensional parameter ϕ [16]. Suppose we observe $\{(X_i, Y_i) \in [0,1]^d \times \mathbb{R}\}_{i=1}^n$ according to the following law: X_i's are i.i.d. uniformly distributed on $[0,1]^d$, and there exists a closed region $\Gamma_0 \subset [0,1]^d$ such that $Y_i \sim f(\cdot;\xi_0)\mathbf{1}_{X_i\in\Gamma_0} + f(\cdot;\rho_0)\mathbf{1}_{X_i\in\Gamma_0^c}$. Here X_i can be understood as the location of i-th observation and Y_i the corresponding pixel intensity. Let $\theta = (\xi,\rho,\Gamma) \in \Theta$ be the parameter and define for any $\theta_i = (\xi_i,\rho_i,\Gamma_i)(i=0,1)$,

$$\begin{aligned} d_n^2(\theta_0,\theta_1) \equiv \|\xi_0-\xi_1\|_2^2\lambda(\Gamma_0\cap\Gamma_1) + \|\rho_0-\rho_1\|_2^2\lambda(\Gamma_0^c\cap\Gamma_1^c) \\ + \|\xi_0-\rho_1\|_2^2\lambda(\Gamma_0\cap\Gamma_1^c) + \|\rho_0-\xi_1\|_2^2\lambda(\Gamma_0^c\cap\Gamma_1). \end{aligned}$$

Here λ denotes the Lebesgue measure on $[0,1]^d$ and $\lambda(B) = \int_B \,\mathrm{d}\lambda$. Clearly d_n is symmetric, but may not satisfy the triangle inequality.

Lemma 3.24. *Suppose that $\{f(\cdot;\phi) : \phi \in \Theta \subset \mathbb{R}^p\}$ is any parametric class considered in Section 3.1 (i.e. Gaussian/Laplace/binary/Poisson models). Then Assumption A holds for d_n defined above with constants depending only through the specific parametric class.*

The following lemma relates d_n to the metric $\lambda(\cdot\Delta\cdot)$ of interest when two elements in Θ are close to each other in d_n.

Lemma 3.25. *Suppose that $\|\xi_0-\rho_0\|_2^2 = r_0^2 > 0$ and $\lambda(\Gamma_0^c\cap\Gamma_1^c) \geq \lambda_0^2 > 0$. If $d_n^2(\theta_0,\theta_1) \leq (\frac{\lambda_0^2}{4} \wedge \frac{\lambda(\Gamma_0)}{8})r_0^2$, then $\lambda(\Gamma_0\Delta\Gamma_1) \leq (8/r_0^2)\cdot d_n^2(\theta_0,\theta_1)$.*

Now we can state our main result in this section. Let $\Theta_1 \subset \ldots \subset \Theta_m \subset \ldots \subset \Theta$ be a sequence of nested models.

Corollary 3.26. *Suppose that $\{f(\cdot;\phi) : \phi \in \Theta \subset \mathbb{R}^p\}$ is any parametric class considered in Section 3.1, and that there exist some $m \in \mathbb{N}, \eta > 0$ such that $\theta_0 = (\xi_0,\rho_0,\Gamma_0) \in \Theta_m$ with $\Gamma_0 \subset [\eta,1-\eta]^d$ and $\xi_0 \neq \rho_0$, and $\Pi_n(\Gamma \subset [\eta,1-\eta]^d) = 1$. If Assumptions B-C hold for d_n described above with $\theta_{0,m}$ replaced by θ_0, then for n large enough (depending only on ξ_0,ρ_0,η), we have*

$$P_{\theta_0}^{(n)}\Pi_n\big(\Gamma : \lambda(\Gamma\Delta\Gamma_0) > C_1\delta_{n,m}^2 \big| \big(X^{(n)},Y^{(n)}\big)\big) \leq C_2 e^{-n\delta_{n,m}^2/C_2}.$$

Here the constants $\{C_i\}_{i=1}^2 > 0$ depend on ξ_0,ρ_0,η.

Our result can be used for smooth boundaries as studied in [LG17], but we will be mainly interested in non-smooth boundaries. Indeed, we will propose a hierarchical prior (cf. Section 3.6.1) so that the posterior distribution is nearly parametrically rate-adaptive to non-smooth polytopal regions Γ.

3.6.1. Example: Detection of polytopal image boundaries

For simplicity of presentation, we specify the binary model for $\{f(\cdot;\phi) : \phi \in [\eta,1-\eta]\}$, and consider $d = 2$. Suppose that $\theta_0 = (\xi_0,\rho_0,\Gamma_0)$ where $\Gamma_0 \subset$

[16] For instance, for the binary model considered in Section 3.1, we may take $p = 1$, $\phi \in [0,1]$ and $f(\cdot,\phi)$ to be the density of $\mathrm{Bern}(\phi)$ with respect to the counting measure on $\{0,1\}$.

$[\eta, 1-\eta]^2$ is a convex polytope. A natural nested sequence of models $\{\Theta_m\}_{m\in\mathbb{N}}$ is given by $\Theta_m \equiv \{(\xi,\rho,\Gamma): \xi \neq \rho, \Gamma \in \mathscr{C}_m\}$ where $\mathscr{C}_m$ contains all convex polytopes in $[\eta, 1-\eta]^2$ with at most m vertices. Consider the following model selection prior:

$$\lambda_n(m) \propto \exp\big(-c\cdot m\log(en)\big), \tag{3.15}$$

where $c>0$ is a constant to be specified later. A prior $\Pi_{n,m}$ on the model Θ_m can be induced by a product prior on (ξ,ρ,Γ). In particular, we put priors on ξ and ρ with densities g_ξ and g_ρ respectively, and a prior on Γ can be induced by taking the convex hull of randomly generated m points in $[\eta,1-\eta]^2$ with density $g_\Gamma^{\otimes m}$. For simplicity, we assume that g_ξ, g_ρ, g_Γ all follow the uniform distribution on $[\eta, 1-\eta]$.

Theorem 3.27. *In the above setting, if $\theta_0 \in \Theta_m$ with $\xi_0 \neq \rho_0$, then there exists some $c>0$ in (3.15) such that for n large enough,*

$$P_{\theta_0}^{(n)}\Pi_n\big(\Gamma : \lambda(\Gamma\Delta\Gamma_0) > C_1 m\log n/n \big| \big(X^{(n)}, Y^{(n)}\big)\big) \leq C_2 e^{-m\log n/C_2}.$$

Here the constants $\{C_i\}_{i=1}^2$ depend on ξ_0, ρ_0, η.

3.7. Intensity estimation in a Poisson point process model

Suppose we observe $\{(X_i,Y_i)\in[0,1]\times\mathbb{R}\}$ from a Poisson point process N defined on $[0,1]\times\mathbb{R}$ with intensity $\lambda(x,y)\equiv\lambda_f(x,y) = n\mathbf{1}_{f(x)\leq y}$. The goal is to recover the boundary $f:[0,1]\to\mathbb{R}$ of the support of the intensity λ [17].

Note that a dominating measure μ is not well-specified for all probability distributions $P_f^{(n)}$, and the likelihood ratio $\mathrm{d}P_{f_0}^{(n)}/\mathrm{d}P_{f_1}^{(n)}$ is well-defined only if $f_1 \leq f_0$. Indeed, [RSH17] showed (cf. Lemma 2.1 therein) that for $f_1 \leq f_0$, $\mathrm{d}P_{f_0}^{(n)}/\mathrm{d}P_{f_1}^{(n)} = e^{n\|f_0-f_1\|_1}\mathbf{1}_{\forall i: f_0(X_i)\leq Y_i}$, and therefore the Kullback-Leibler divergence is given by

$$\bar{L}_1(f_0,f_1) = \begin{cases} \|f_0-f_1\|_1, & f_1 \leq f_0; \\ \infty, & \text{otherwise.}\end{cases}$$

The technical problem here is that $\bar{L}_1$ is not symmetric—fortunately by a slight modification, our machinery can still be applied. To this end, suppose $\inf_{g\in\mathcal{F}_m}\bar{L}_1(f_0,g) < \infty$, and let $f_{0,m}\in\arg\min_{g\in\mathcal{F}_m}\bar{L}_1(f_0,g)$ (so that $f_{0,m}\leq f_0$), assumed to be well-defined.

Corollary 3.28. *For the support boundary recovery problem described above, let $d_n^2 \equiv \bar{L}_1$. If (i) Assumption B holds under entropy with left bracketing[18] and*

[17] This model can be regarded as a continuous analogue of the regression problem with irregular errors [MR13].

[18] For a generic function class $\mathcal{G}$ defined on $[0,1]$, the left bracketing number $\mathcal{N}_{[}(\varepsilon,\mathcal{G},\bar{L}_1)$ is the smallest number M of functions $g_1,\ldots,g_M$ such that for any $g\in\mathcal{G}$ there exists some $j\in\{1,\ldots,M\}$ with $g_j \leq g$ and $\int_0^1(g-g_j)\leq\varepsilon$. Note that in this definition g_j need not belong to $\mathcal{G}$.

the set in (2.4) restricted to $f \geq f_0$; (ii) Assumption C holds with the set in (P2) restricted to $f \geq f_{0,m}$, then (2.9)-(2.11) hold with the posterior distribution restricted to $f \geq f_0$.

In Section 3.7.1 we will use the above result to derive oracle contraction rates for estimating piecewise constant intensities.

It is also possible to consider the two-sided L_1 loss, at the expense of stronger conditions. Below is a result in this direction.

Corollary 3.29. *Suppose that for $m \in \mathcal{M}$, (i) $\log \mathcal{N}(\delta_{n,m}^2, \mathcal{F}_m, L_\infty) \leq C_1 n \delta_{n,m}^2$ and (ii) $\Pi_{n,m}(f \in \mathcal{F}_m : L_1(f, f_{0,m}) \leq C_2 \delta_{n,m}^2, f \leq f_{0,m}) \geq e^{-C_2 n \delta_{n,m}^2}$ hold for some $f_{0,m} \leq f_0$. Then using the prior (2.8), there exists some constant $C' > 0$ such that for any $m \in \mathcal{M}$,*

$$P_{f_0}^{(n)} \Pi_n \big(f \in \mathcal{F} : L_1(f, f_0) \geq C' \varepsilon_{n,m}^2 | (X^{(n)}, Y^{(n)})\big) \leq C' e^{-\varepsilon_{n,m}^2 / C'}.$$

Here $\varepsilon_{n,m}^2 \equiv \max\{L_1(f_0, f_{0,m}), \delta_{n,m}^2\}$.

3.7.1. Example: Estimating piecewise constant intensity in a Poisson point process model

Consider fitting the intensity λ_f in the Poisson point process model by the class of piecewise constant functions $\mathcal{F} \equiv \cup_{m=1}^\infty \mathcal{F}_m \equiv \{f : f = \sum_{j=1}^m a_j \mathbf{1}_{[t_{j-1}, t_j)}, 0 = t_0 < t_1 < \ldots < t_{m-1} < t_m = 1\}$. A prior on $\mathcal{F}_m$ can be induced by a prior $\Pi_{n,m}^t$ on $\{t_1 < \ldots < t_{m-1}\}$ followed by a prior $\Pi_{n,m}^a$ on $\{a_j\}_{j=1}^m$. More specifically, we choose $\Pi_{n,m}^t$ with density $\boldsymbol{t} = (t_1, \ldots, t_{m-1}) \mapsto (m-1)! \mathbf{1}_{t_1 < \ldots < t_{m-1}}(\boldsymbol{t})$, and $\Pi_{n,m}^a$ with product density $g_a^{\otimes m}$. As before, we assume that g_a is symmetric, non-increasing and satisfies the following: g_a has full support, and there exists some sequence $\{R_n\}$ with $\log R_n \lesssim \log n$, and a large enough absolute constant $C' > 0$ such that

$$\int_{|x| > R_n} g_a(x) \,\mathrm{d}x \leq n^{-C'}. \tag{3.16}$$

It is easily seen that this condition is very weak, and essentially does not require any tail condition on g_a. The reason for this to occur is that the information geometry of the model studied here does not change with the L_∞ size of the model—the impact of this only occurs through the complexity of the model by logarithmic factors.

Consider the following prior Λ_n on the model index $\mathcal{I} \equiv \mathbb{N}$:

$$\lambda_n(m) \propto \exp\big(-c \cdot m \log(en)\big), \tag{3.17}$$

where $c > 0$ is a constant to be specified later.

Theorem 3.30. *Suppose that $\|f_0\|_\infty < \infty$ and (3.16) holds for the prior density g_a. There exists some $c > 0$ in (3.17) such that for n large enough (depending only on f_0 and the prior g_a), with $(\varepsilon_{n,m}^{\mathrm{int}})^2 \equiv \max\{\inf_{g \in \mathcal{F}_m} \bar{L}_1(f_0, g), m \log(en)/n\}$,*

$$P_{f_0}^{(n)} \Pi_n \big(f \geq f_0 : \bar{L}_1(f, f_0) > C_1 (\varepsilon_{n,m}^{\mathrm{int}})^2 \big| N\big) \leq C_2 e^{-m \log(en)/C_2}.$$

Here the constants $C_i(i = 1, 2)$ are absolute.

Compared with Theorem 5.3 of [RSH17], our Theorem 3.30 works with a slightly weaker one-sided L_1 loss, but enjoys an exact form of an oracle posterior contraction rate. From here it is straightforward to derive rate result assuming Hölder smoothness on f_0 (as in [RSH17]). Note that here we do not require the technical condition $\log m \gtrsim \log n$ as in [RSH17], so our result here shows rate-adaptivity of the posterior distribution to intensities with fixed number of constant pieces.

Appendix A: Proofs for Section 2

A.1. Proof of Theorem 2.3: main steps

First we need a lemma allowing a change-of-measure argument.

Lemma A.1. *Let Assumption A hold. There exists some constant $c_4 \geq 1$ only depending on c_1, c_3 and κ such that for any random variable $U \in [0, 1]$, any $\delta_n \geq d_n(f_0, f_1)$ and any $j \in \mathbb{N}$,*

$$P_{f_0}^{(n)} U \leq c_4[P_{f_1}^{(n)} U \cdot e^{c_4 n j \delta_n^2} + e^{-c_4^{-1} n j \delta_n^2}].$$

The next propositions solve the posterior contraction problem for the 'local' model $\mathcal{F}_m$.

Proposition A.2. *Fix $m \in \mathcal{M}$ such that $\delta_{n,m}^2 \geq d_n^2(f_0, f_{0,m})$. Then there exists some constant $c_8 \geq 1$ (depending on the constants in Assumption A) such that for $j \geq 8\mathfrak{c}^2/c_7\mathfrak{h}$,*

$$P_{f_{0,m}}^{(n)} \Pi_n(f \in \mathcal{F} : d_n^2(f, f_{0,m}) > \mathfrak{c}^2(j\mathfrak{h})^\gamma \delta_{n,m}^2 \big| X^{(n)}) \leq c_8 e^{-nj\mathfrak{h}\delta_{n,m}^2/c_8\mathfrak{c}^2}. \tag{A.1}$$

Proposition A.3. *Fix $m \in \mathcal{M}$ such that $\delta_{n,m}^2 < d_n^2(f_0, f_{0,m})$. Let $\tilde{m} \equiv \tilde{m}(m) \equiv \inf\{m' \in \mathcal{M}, m' \geq m : \delta_{n,m'} \geq d_n(f_0, f_{0,m})\}$. Then for $j \geq 8\mathfrak{c}^2/c_7\mathfrak{h}$,*

$$P_{f_{0,m}}^{(n)} \Pi_n(f \in \mathcal{F} : d_n^2(f, f_{0,m}) > \mathfrak{c}^4(2j\mathfrak{h})^\gamma d_n^2(f_0, f_{0,m}) \big| X^{(n)}) \leq c_8 e^{-nj\mathfrak{h}\delta_{n,\tilde{m}}^2/c_8\mathfrak{c}^2}. \tag{A.2}$$

The proofs of these results will be detailed in later subsections.

Proof of Theorem 2.3: main steps. Instead of (2.9), we will prove a slightly stronger statement as follows: for any $j \geq 8\mathfrak{c}^2/c_7\mathfrak{h}$, and $\mathfrak{h} \geq 2c_4c_8\mathfrak{c}^2$,

$$P_{f_0}^{(n)} \Pi_n\big(f \in \mathcal{F} : d_n^2(f, f_0) > \mathfrak{c}_1 j^\gamma \varepsilon_{n,m}^2 \big| X^{(n)}\big) \leq \mathfrak{c}_2 e^{-jn\varepsilon_{n,m}^2/\mathfrak{c}_2}. \tag{A.3}$$

Here the constants $\mathfrak{c}_i(i = 1, 2)$ depends on the constants involved in Assumption A and $\mathfrak{c}, \mathfrak{h}$.

Proof of (A.3). First consider the overfitting case. By Proposition A.2 and Lemma A.1, we see that when $\delta_{n,m}^2 \geq d_n^2(f_0, f_{0,m})$ holds, for $j \geq 8\mathfrak{c}^2/c_7\mathfrak{h}$, it holds that

$$\begin{aligned}
&P_{f_0}^{(n)}\Pi_n\big(f \in \mathcal{F} : d_n^2(f, f_0) > 2d_n^2(f_0, f_{0,m}) + 2\mathfrak{c}^2(j\mathfrak{h})^\gamma \delta_{n,m}^2 \big| X^{(n)}\big)\\
&\leq P_{f_0}^{(n)}\Pi_n\big(f \in \mathcal{F} : d_n^2(f, f_{0,m}) > \mathfrak{c}^2(j\mathfrak{h})^\gamma \delta_{n,m}^2 \big| X^{(n)}\big)\\
&\leq c_4\big[P_{f_{0,m}}^{(n)}\Pi_n\big(f \in \mathcal{F} : d_n^2(f, f_{0,m}) > \mathfrak{c}^2(j\mathfrak{h})^\gamma \delta_{n,m}^2 \big| X^{(n)}\big)e^{c_4 nj\delta_{n,m}^2} + e^{-c_4^{-1}nj\delta_{n,m}^2}\big]\\
&\leq c_8 c_4 e^{-nj\delta_{n,m}^2\left(\frac{\mathfrak{h}}{c_8\mathfrak{c}^2} - c_4\right)} + c_4 e^{-c_4^{-1}nj\delta_{n,m}^2} \leq 2c_8c_4e^{-jn\delta_{n,m}^2 \min\{c_4, c_4^{-1}\}}.
\end{aligned}$$

Here in the second line we used the fact that $d_n^2(f, f_{0,m}) \geq d_n^2(f, f_0)/2 - d_n^2(f_0, f_{0,m})$.

Next consider the underfitting case: fix $m \in \mathcal{M}$ such that $\delta_{n,m}^2 < d_n^2(f_0, f_{0,m})$. Apply Proposition A.3 and Lemma A.1, and use similar arguments to see that for $j \geq 8\mathfrak{c}^2/c_7\mathfrak{h}$,

$$\begin{aligned}
&P_{f_0}^{(n)}\Pi_n\big(f \in \mathcal{F} : d_n^2(f, f_0) > \big[2\mathfrak{c}^4(2j\mathfrak{h})^\gamma + 2\big] d_n^2(f_0, f_{0,m}) \big| X^{(n)}\big)\\
&\leq c_4\big[P_{f_{0,m}}^{(n)}\Pi_n\big(f \in \mathcal{F} : d_n^2(f, f_{0,m}) > \mathfrak{c}^4(2j\mathfrak{h})^\gamma d_n^2(f_0, f_{0,m}) \big| X^{(n)}\big)e^{c_4 nj\delta_{n,\tilde{m}}^2}\\
&\qquad\qquad + e^{-c_4^{-1}jn\delta_{n,\tilde{m}}^2}\big]\\
&\leq 2c_8c_4e^{-nj\delta_{n,\tilde{m}}^2 \min\{c_4, c_4^{-1}\}}.
\end{aligned}$$

Here in the second line we used (i) $2d_n^2(f, f_{0,m}) \geq d_n^2(f, f_0) - 2d_n^2(f_0, f_{0,m})$, and (ii) $\delta_{n,\tilde{m}} \geq d_n(f_0, f_{0,m})$. The claim of (A.3) follows by combining the estimates.

Proof of (2.11). The proof is essentially integration of tail estimates by a peeling device. Let the event A_j be defined via

$$A_j := \{\mathfrak{c}_1 j^\gamma\big(d_n^2(f_0, f_{0,m}) + \delta_{n,m}^2\big) < d_n^2(f, f_0) \leq \mathfrak{c}_1(j+1)^\gamma\big(d_n^2(f_0, f_{0,m}) + \delta_{n,m}^2\big)\}.$$

Then,

$$\begin{aligned}
&P_{f_0}^{(n)} d_n^2(\hat{f}_n, f_0) = P_{f_0}^{(n)} d_n^2\left(\Pi_n(f|X^{(n)}), f_0\right) \leq P_{f_0}^{(n)}\Pi_n\left(d_n^2(f, f_0)|X^{(n)}\right)\\
&\leq C_{\mathfrak{c}_1,\mathfrak{c},c_7,\mathfrak{h},\gamma}\big(d_n^2(f_0, f_{0,m}) + \delta_{n,m}^2\big) + \sum_{j \geq 8\mathfrak{c}^2/c_7\mathfrak{h}} P_{f_0}^n\Pi_n\big(d_n^2(f, f_0)\mathbf{1}_{A_j}\big|X^{(n)}\big)\\
&\leq C_{\mathfrak{c}_1,\mathfrak{c},c_7,\mathfrak{h},\gamma}\big(d_n^2(f_0, f_{0,m}) + \delta_{n,m}^2\big) + \frac{2^{\gamma+1}\mathfrak{c}_1\mathfrak{c}_2}{n} \sum_{j \geq 8\mathfrak{c}^2/c_7\mathfrak{h}} j^\gamma n\varepsilon_{n,m}^2 e^{-jn\varepsilon_{n,m}^2/\mathfrak{c}_2}.
\end{aligned}$$

The inequality in the first line of the above display is due to Jensen's inequality applied with $d_n^2(\cdot, f_0)$ (the convexity follows since $f \mapsto d_n(f, f_0)$ is non-negatively convex, so is its square), followed by Cauchy-Schwarz inequality. The summation can be bounded up to a constant depending on $\gamma, \mathfrak{c}_1, \mathfrak{c}_2$ by

$$\sum_{j \geq 8\mathfrak{c}^2/c_7\mathfrak{h}} (jn\varepsilon_{n,m}^2)^\gamma e^{-jn\varepsilon_{n,m}^2/\mathfrak{c}_2}$$

$$\leq \sum_{j \geq 8\mathfrak{c}^2/c_7\mathfrak{h}} (jn\varepsilon_{n,m}^2)^\gamma e^{-jn\varepsilon_{n,m}^2/\mathfrak{c}_2}\big((j+1)n\varepsilon_{n,m}^2 - jn\varepsilon_{n,m}^2\big),$$

where the inequality follows since $n\varepsilon_{n,m}^2 \geq n\varepsilon_{n,1}^2 \geq 1$. This quantity can be bounded by a constant multiple of $\int_0^\infty x^\gamma e^{-x/\mathfrak{c}_2}\, \mathrm{d}x$ independent of m. Now the proof is complete by noting that $\delta_{n,m}^2$ majorizes $1/n$ up to a constant, and then taking infimum over $m \in \mathcal{M}$. □

A.2. Proofs of Propositions A.2 and A.3

We will need several lemmas before the proof of Propositions A.2 and A.3.

Lemma A.4. *Let Assumption A hold. Let $\mathcal{F}$ be a function class defined on the sample space $\mathfrak{X}$. Suppose that $N : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a non-increasing function such that for some $\varepsilon_0 \geq \sqrt{2/(c_2 \wedge c_3)} \cdot d_0$ and every $\varepsilon \geq \varepsilon_0$, the following entropy estimate holds:*

$$\mathcal{N}\left(c_5\varepsilon, \{f \in \mathcal{F} : \varepsilon < d_n(f, f_0) \leq 2\varepsilon\}, d_n\right) \leq N(\varepsilon).$$

Then for any $\varepsilon \geq \varepsilon_0$, there exists some test ϕ_n such that

$$P_{f_0}^{(n)}\phi_n \leq c_6 N(\varepsilon) e^{-c_7 n\varepsilon^2}/(1 - e^{-c_7 n\varepsilon^2}), \quad \sup_{f \in \mathcal{F}: d_n(f, f_0) \geq \varepsilon} P_f^{(n)}(1 - \phi_n) \leq c_6 e^{-c_7 n\varepsilon^2}.$$

The constants c_5, c_6, c_7 are taken from Lemma 2.1.

Lemma A.5. *Fix $\varepsilon > 0$. Let Assumption A holds for some d_0 such that $\varepsilon \geq \sqrt{2/(c_2 \wedge c_3)} \cdot d_0$. Suppose that Π is a probability measure on $\{f \in \mathcal{F} : d_n(f, f_0) \leq \varepsilon\}$. Then for every $C > 0$, there exists some $C' > 0$ depending on C, κ such that $P_{f_0}^{(n)}\big(\int p_f^{(n)}/p_{f_0}^{(n)}\, \mathrm{d}\Pi(f) \leq e^{-(C+c_3)n\varepsilon^2}\big) \leq c_1 e^{-C' n\varepsilon^2}$.*

The proof of these lemmas can be found in Appendix D.

Proof of Proposition A.2. Fix $m' \in \mathcal{M}$ with $m' \geq m$. Now we invoke Lemma A.4 with $\mathcal{F} \equiv \mathcal{F}_{m'}$, $f_0 \equiv f_{0,m} \in \mathcal{F}_m \subset \mathcal{F}_{m'}$ [since $m' \geq m$], $\varepsilon_0 \equiv \delta_{n,m'}$ and $\log N(\varepsilon) \equiv (c_7/2)n\delta_{n,m'}^2$ for $\varepsilon = \varepsilon_0$ to see that, there exists some test $\phi_{n,m'}$ such that

$$P_{f_{0,m}}^{(n)}\phi_{n,m'} \leq c_6 e^{\log N(\varepsilon) - c_7 n\delta_{n,m'}^2}/(1 - e^{-c_7 n\delta_{n,m'}^2}) \leq 2c_6 e^{-c_7 n\delta_{n,m'}^2/2}, \tag{A.4}$$

and that

$$\sup_{f \in \mathcal{F}_{m'}: d_n^2(f, f_{0,m}) \geq \delta_{n,m'}^2} P_f^{(n)}(1 - \phi_{n,m'}) \leq c_6 e^{-c_7 n\delta_{n,m'}^2}. \tag{A.5}$$

Note that here in (A.4) we used the fact that $n\delta_{n,m'}^2 \geq 2/c_7$ by definition of $\delta_{n,m'}$. Now for the fixed j, m as in the statement of the proposition, we let $\phi_n := \sup_{m' \in \mathcal{I}: m' \geq j\mathfrak{h}m} \phi_{n,m'}$ be a global test for big models. Then by (A.4),

$$P_{f_{0,m}}^{(n)}\phi_n \leq \sum_{m' \geq j\mathfrak{h}m} P_{f_{0,m}}^{(n)}\phi_{n,m'} \leq \sum_{m' \geq j\mathfrak{h}m} 2c_6 e^{-c_7 n\delta_{n,m'}^2/2} \leq 4c_6 e^{-(c_7/2\mathfrak{c}^2)nj\mathfrak{h}\delta_{n,m}^2}.$$

Here we used the left side of (2.5). This implies that for any random variable $U \in [0,1]$, we have

$$P^{(n)}_{f_{0,m}} U \cdot \phi_n \leq P^{(n)}_{f_{0,m}} \phi_n \leq 4c_6 e^{-(c_7/2\mathfrak{c}^2)nj\mathfrak{h}\delta^2_{n,m}}. \tag{A.6}$$

On the power side, with $m' = j\mathfrak{h}m$ applied to (A.5) we see that

$$\begin{aligned} &\sup_{\substack{f\in\mathcal{F}_{j\mathfrak{h}m}:\\ d_n^2(f,f_{0,m})\geq \mathfrak{c}^2(j\mathfrak{h})^\gamma \delta^2_{n,m}}} P^{(n)}_f(1-\phi_n) \leq \sup_{\substack{f\in\mathcal{F}_{j\mathfrak{h}m}:\\ d_n^2(f,f_{0,m})\geq \delta^2_{n,j\mathfrak{h}m}}} P^{(n)}_f(1-\phi_n) \\ &\quad\leq c_6 e^{-c_7 n\delta^2_{n,j\mathfrak{h}m}} \leq 2c_6 e^{-(c_7/\mathfrak{c}^2)nj\mathfrak{h}\delta^2_{n,m}}. \end{aligned} \tag{A.7}$$

The first inequality follows from the right side of (2.5) since $\mathfrak{c}^2(j\mathfrak{h})^\gamma\delta^2_{n,m} \geq \delta^2_{n,j\mathfrak{h}m}$, and the last inequality follows from the left side of (2.5). On the other hand, by applying Lemma A.5 with $C = c_3$ and $\varepsilon^2 \equiv c_7 j\mathfrak{h}\delta^2_{n,m}/8c_3\mathfrak{c}^2$, we see that there exists some event $\mathcal{E}_n$ such that

$$P^{(n)}_{f_{0,m}}(\mathcal{E}^c_n) \leq c_1 e^{-C' c_7 nj\mathfrak{h}\delta^2_{n,m}/8c_3\mathfrak{c}^2}$$

and it holds on the event $\mathcal{E}_n$ that

$$\begin{aligned} &\int p^{(n)}_f / p^{(n)}_{f_{0,m}}\ \mathrm{d}\Pi(f) \\ &\geq \lambda_n(m) \int_{\{f\in\mathcal{F}_m : d_n^2(f,f_{0,m})\leq c_7 j\mathfrak{h}\delta^2_{n,m}/8c_3\mathfrak{c}^2\}} p^{(n)}_f / p^{(n)}_{f_{0,m}}\ \mathrm{d}\Pi_{n,m}(f) \\ &\geq \lambda_n(m) e^{-\frac{c_7 nj\mathfrak{h}\delta^2_{n,m}}{4\mathfrak{c}^2}} \Pi_{n,m}\left(\{f\in\mathcal{F}_m : d_n^2(f,f_{0,m}) \leq c_7 j\mathfrak{h}\delta^2_{n,m}/8c_3\mathfrak{c}^2\}\right). \end{aligned} \tag{A.8}$$

Note that

(A.9)

$$\begin{aligned} &P^{(n)}_{f_{0,m}} \Pi_n\big(f\in\mathcal{F} : d_n^2(f,f_{0,m}) > \mathfrak{c}^2(j\mathfrak{h})^\gamma\delta^2_{n,m} \big| X^{(n)}\big)(1-\phi_n)\mathbf{1}_{\mathcal{E}_n} \\ &= P^{(n)}_{f_{0,m}}\left[\frac{\int_{f\in\mathcal{F}:d_n^2(f,f_{0,m})>\mathfrak{c}^2(j\mathfrak{h})^\gamma\delta^2_{n,m}} p^{(n)}_f/p^{(n)}_{f_{0,m}}\ \mathrm{d}\Pi_n(f)}{\int p^{(n)}_f/p^{(n)}_{f_{0,m}}\ \mathrm{d}\Pi_n(f)}(1-\phi_n)\mathbf{1}_{\mathcal{E}_n}\right] \\ &\leq \frac{e^{c_7 nj\mathfrak{h}\delta^2_{n,m}/4\mathfrak{c}^2}}{\lambda_n(m)\Pi_{n,m}(\{f\in\mathcal{F}_m : d_n^2(f,f_{0,m}) \leq c_7 j\mathfrak{h}\delta^2_{n,m}/8c_3\mathfrak{c}^2\})} \\ &\qquad \times P^{(n)}_{f_{0,m}}\left[\int_{f\in\mathcal{F}:d_n^2(f,f_{0,m})>\mathfrak{c}^2(j\mathfrak{h})^\gamma\delta^2_{n,m}} p^{(n)}_f/p^{(n)}_{f_{0,m}}\ \mathrm{d}\Pi_n(f)(1-\phi_n)\right] \\ &\equiv (I)\cdot(II) \end{aligned}$$

where the inequality follows from (A.8). On the other hand, the expectation term in the above display can be further calculated as follows:

$$(II) = \int_{f\in\mathcal{F}:d_n^2(f,f_{0,m})>\mathfrak{c}^2(j\mathfrak{h})^\gamma\delta^2_{n,m}} P^{(n)}_f(1-\phi_n)\ \mathrm{d}\Pi_n(f)$$

$$\begin{aligned}&\leq \sup_{f\in\mathcal{F}_{j\mathfrak{h}m}:d_n^2(f,f_{0,m})>\mathfrak{c}^2(j\mathfrak{h})^\gamma\delta_{n,m}^2} P_f^{(n)}(1-\phi_n)+\Pi_n\big(\mathcal{F}\setminus\mathcal{F}_{j\mathfrak{h}m}\big)\\&\leq 2c_6e^{-(c_7/\mathfrak{c}^2)nj\mathfrak{h}\delta_{n,m}^2}+4e^{-(1/\mathfrak{c}^2)nj\mathfrak{h}\delta_{n,m}^2}\leq 6c_6e^{-(c_7/\mathfrak{c}^2)nj\mathfrak{h}\delta_{n,m}^2}.\end{aligned}$$

The first term in the second inequality follows from (A.7) and the second term follows from (P1) in Assumption C along with the left side of (2.5). By (P1)-(P2) in Assumption C and $j\geq 8\mathfrak{c}^2/c_7\mathfrak{h}$,

$$\begin{aligned}&P_{f_{0,m}}^{(n)}\Pi_n\big(f\in\mathcal{F}:d_n^2(f,f_{0,m})>\mathfrak{c}^2(j\mathfrak{h})^\gamma\delta_{n,m}^2\big|X^{(n)}\big)(1-\phi_n)\mathbf{1}_{\mathcal{E}_n}\\&\leq Ce^{-(c_7/4\mathfrak{c}^2)nj\mathfrak{h}\delta_{n,m}^2}.\end{aligned}$$

We conclude (A.1) from (A.6), probability estimate on $\mathcal{E}_n^c$. □

Proof of Proposition A.3. The proof largely follows the same lines as that of Proposition A.2. See Appendix D for details. □

A.3. Completion of proof of Theorem 2.3

Proof of (2.10). For any $m\in\mathcal{M}$ such that $\delta_{n,m}^2\geq d_n^2(f_0,f_{0,m})$, following the similar reasoning in (A.9) with $j=8\mathfrak{c}^2/c_7\mathfrak{h}$,

$$\begin{aligned}&P_{f_{0,m}}^{(n)}\Pi_n\big(f\notin\mathcal{F}_{j\mathfrak{h}m}|X^{(n)}\big)\mathbf{1}_{\mathcal{E}_n}\\&\leq\frac{e^{c_7nj\mathfrak{h}\delta_{n,m}^2/4\mathfrak{c}^2}}{\lambda_n(m)\Pi_{n,m}\big(\big\{f\in\mathcal{F}_m:d_n^2(f,f_{0,m})\leq c_7j\mathfrak{h}\delta_{n,m}^2/8c_3\mathfrak{c}^2\big\}\big)}\cdot\Pi\big(\mathcal{F}\setminus\mathcal{F}_{j\mathfrak{h}m}\big)\\&\leq Ce^{-(c_7/4\mathfrak{c}^2)nj\mathfrak{h}\delta_{n,m}^2}.\end{aligned}$$

From here (2.10) can be established by controlling the probability estimate for $\mathcal{E}_n^c$ as in Proposition A.2, and a change of measure argument using Lemma A.1. □

A.4. Proof of Lemma 2.1

Proof of Lemma 2.1. Without loss of generality, we assume that $d_0=0$. Let $c>0$ be a constant to be specified later. Consider the test statistics $\phi_n\equiv\mathbf{1}\big(\log(p_{f_0}^{(n)}/p_{f_1}^{(n)})\leq -cnd_n^2(f_0,f_1)\big)$. We first consider type I error. Under the null hypothesis, we have for any $\lambda_1\in(0,1/\kappa_\Gamma)$,

$$\begin{aligned}P_{f_0}^{(n)}\phi_n&\leq P_{f_0}^{(n)}\big[\big(\log(p_{f_0}^{(n)}/p_{f_1}^{(n)})-P_{f_0}\log(p_{f_0}^{(n)}/p_{f_1}^{(n)})\big)\leq-(c+c_2)nd_n^2(f_0,f_1)\big]\\&\leq c_1e^{\psi_{\kappa_g nd_n^2(f_0,f_1),\kappa_\Gamma}(-\lambda_1)}\cdot e^{-\lambda_1(c+c_2)nd_n^2(f_0,f_1)}.\end{aligned}$$

Choosing $\lambda_1=\min\{1/(\kappa_\Gamma),(c+c_2)/(2\kappa_g)\}$ we get $P_{f_0}^{(n)}\phi_n\leq c_1e^{-C_1nd_n^2(f_0,f_1)}$ where $C_1=\lambda_1(c+c_2)/2$. Next we handle the type II error. To this end, for a constant $c'>c_3c_5$ to be specified later, consider the event $\mathcal{E}_n\equiv\mathbf{1}\big(\log(p_f^{(n)}/p_{f_1}^{(n)})<$

$c'nd_n^2(f_0,f_1)\big)$, where $f \in \mathcal{F}$ is such that $d_n^2(f,f_1) \leq c_5 d_n^2(f_0,f_1)$, and $\lambda_2 \in (0,1/\kappa_\Gamma)$,

$$\begin{aligned}
&P_f^{(n)}(\mathcal{E}_n^c)\\
&\leq P_f^{(n)}\big(\log(p_f^{(n)}/p_{f_1}^{(n)}) - P_f^{(n)}\log(p_f^{(n)}/p_{f_1}^{(n)}) > c'nd_n^2(f_0,f_1) - c_3nd_n^2(f,f_1)\big)\\
&\leq P_f^{(n)}\big(\log(p_f^{(n)}/p_{f_1}^{(n)}) - P_f^{(n)}\log(p_f^{(n)}/p_{f_1}^{(n)}) > (c'-c_3c_5)nd_n^2(f_0,f_1)\big)\\
&\leq e^{-\lambda_2(c'-c_3c_5)nd_n^2(f_0,f_1)}\cdot c_1 e^{\psi_{\kappa_g nd_n^2(f,f_1),\kappa_\Gamma}(\lambda_2)}.
\end{aligned}$$

By choosing $\lambda_2 = \min\{1/(\kappa_\Gamma),(c'-c_3c_5)/(2\kappa_g)\}$, we see that $P_f^{(n)}(\mathcal{E}_n^c) \leq c_1 e^{-C_2nd_n^2(f_0,f_1)}$ where $C_2 = \lambda_2(c'-c_3c_5)/2$. On the other hand, using the symmetry of $d_n(\cdot,\cdot)$ and for $0<c<c_2$, $\lambda_3 \in (0,1/\kappa_\Gamma)$,

$$\begin{aligned}
P_{f_1}^{(n)}\big(1-\phi_n\big) &= P_{f_1}^{(n)}\big(\log(p_{f_1}^{(n)}/p_{f_0}^{(n)}) < cnd_n^2(f_0,f_1)\big)\\
&= P_{f_1}^{(n)}\big(\log(p_{f_1}^{(n)}/p_{f_0}^{(n)}) - P_{f_1}^{(n)}\log(p_{f_1}^{(n)}/p_{f_0}^{(n)}) < -(c_2-c)nd_n^2(f_0,f_1)\big)\\
&\leq e^{-\lambda_3(c_2-c)nd_n^2(f_0,f_1)}\cdot c_1 e^{\psi_{\kappa_g nd_n^2(f_0,f_1),\kappa_\Gamma}(-\lambda_3)}.
\end{aligned}$$

Choosing $\lambda_3 = \min\{1/(\kappa_\Gamma),(c_2-c)/(2\kappa_g)\}$ we see that

$$P_{f_1}^{(n)}(1-\phi_n) \leq c_1 e^{-C_3nd_n^2(f_0,f_1)},$$

where $C_3 = \lambda_3(c_2-c)/2$. Hence it follows that

$$\begin{aligned}
P_f^{(n)}(1-\phi_n) &= P_{f_1}^{(n)}\big[\big(1-\phi_n\big)\cdot(p_f^{(n)}/p_{f_1}^{(n)})\big(\mathbf{1}_{\mathcal{E}_n}+\mathbf{1}_{\mathcal{E}_n^c}\big)\big]\\
&\leq e^{c'nd_n^2(f_0,f_1)}P_{f_1}^{(n)}\big(1-\phi_n\big) + c_1e^{-C_2nd_n^2(f_0,f_1)}\\
&\leq 2c_1e^{-\min\{(C_3-c'),C_2\}nd_n^2(f_0,f_1)}.
\end{aligned}$$

Now it suffices to choose c,c',c_5 such that $c' > c_3c_5$, $c<c_2$ and $c' < C_3$. To this end, we choose $c = c_2/2$, $c' \equiv \frac{C_3}{2} = \frac{\lambda_3(c_2-c)}{4} = \frac{\lambda_3 c_2}{8} = \frac{c_2}{8\kappa_\Gamma}\wedge\frac{c_2^2}{32\kappa_g}$, and $c_5 = \frac{c'}{2c_3}\wedge\frac{1}{4} = \frac{c_2}{16c_3\kappa_\Gamma}\wedge\frac{c_2^2}{64c_3\kappa_g}\wedge\frac{1}{4}$, completing the proof. □

A.5. Proof of Lemma A.1

We recall a standard fact.

Lemma A.6. *If a random variable X satisfies $\mathbb{E}e^{\lambda X} \leq e^{\psi_{v,c}(\lambda)}$, then for $t>0$, $\mathbb{P}(X\geq t)\vee\mathbb{P}(X\leq -t) \leq e^{-\frac{t^2}{2(2v+ct)}}$.*

Proof. Noting that $\mathbb{E}e^{\lambda X} \leq e^{\psi_{v,c}(\lambda)} \leq e^{\frac{(2v)\lambda^2}{2(1-c|\lambda|)}}$. Then using arguments in page 29 of [BLM13] and Exercise 2.8 therein, we obtain the claim. □

Proof of Lemma A.1. For $c = 2c_3$, consider the event $\mathcal{E}_n \equiv \big\{\log(p_{f_0}^{(n)}/p_{f_1}^{(n)}) < cjn\delta_n^2\big\}$. By Lemma A.6, we have for some constant $C > 0$ depending on c_1, c_3 and κ,

$$\begin{aligned} P_{f_0}^{(n)}(\mathcal{E}_n^c) &\leq P_{f_0}^{(n)}\big(\log(p_{f_0}^{(n)}/p_{f_1}^{(n)}) - P_{f_0}^{(n)}\log(p_{f_0}^{(n)}/p_{f_1}^{(n)}) \geq cjn\delta_n^2 - c_3nd_n^2(f_0,f_1)\big) \\ &\overset{(*)}{\leq} P_{f_0}^{(n)}\big(\log(p_{f_0}^{(n)}/p_{f_1}^{(n)}) - P_{f_0}^{(n)}\log(p_{f_0}^{(n)}/p_{f_1}^{(n)}) \geq c_3jn\delta_n^2\big) \leq Ce^{-C^{-1}nj\delta_n^2}. \end{aligned}$$

Here in $(*)$ we used $d_n(f_0, f_1) \leq \delta_n$. Then

$$\begin{aligned} P_{f_0}^{(n)}U &= P_{f_0}^{(n)}U\mathbf{1}_{\mathcal{E}_n} + P_{f_0}^{(n)}U\mathbf{1}_{\mathcal{E}_n^c} \leq P_{f_1}^{(n)}\big[U(p_{f_0}^{(n)}/p_{f_1}^{(n)})\mathbf{1}_{\mathcal{E}_n}\big] + Ce^{-C^{-1}nj\delta_n^2} \\ &\leq P_{f_1}^{(n)}U \cdot e^{cnj\delta_n^2} + Ce^{-C^{-1}nj\delta_n^2}, \end{aligned}$$

completing the proof. □

A.6. Proof of Proposition 2.2

Proof of Proposition 2.2. Let $\Sigma_n = \sum_m e^{-2n\delta_{n,m}^2}$ be the total mass. Then

$$e^{-2n\delta_{n,1}^2} \leq \Sigma_n \leq 2e^{-2n\delta_{n,1}^2/\mathfrak{c}^2} \leq 2.$$

The first condition of (P1) is trivial. We only need to verify the second condition of (P1):

$$\sum_{k>\mathfrak{h}m}\lambda_n(k) = \Sigma_n^{-1}\sum_{k>\mathfrak{h}m}e^{-2n\delta_{n,k}^2} \leq e^{2n\delta_{n,1}^2}\cdot 2e^{-(2\mathfrak{h}/\mathfrak{c}^2)n\delta_{n,m}^2} \leq 2e^{-2n\delta_{n,m}^2},$$

where the first inequality follows from (2.5) and the second by the condition $\mathfrak{h} \geq 2\mathfrak{c}^2$. □

Appendix B: Proofs in Section 3 Part I: results for models

Proof of Lemma 3.1. Let $P_{\theta_0}^{(n)}$ denote the probability measure induced by the joint distribution of $(X_1, \ldots, X_n)$ when the underlying signal is θ_0.

First consider Gaussian regression case. Since

$$\begin{aligned} \log(p_{\theta_0}^{(n)}/p_{\theta_1}^{(n)})(X^{(n)}) &= \sum_{i=1}^n\Big[-\frac{1}{2}(X_i-\theta_{0,i})^2 + \frac{1}{2}(X_i-\theta_{1,i})^2\Big], \\ P_{\theta_0}^{(n)}\log(p_{\theta_0}^{(n)}/p_{\theta_1}^{(n)}) &= \frac{1}{2}n\ell_n^2(\theta_0,\theta_1). \end{aligned}$$

we have

$$\begin{aligned} &P_{\theta_0}^{(n)}e^{\lambda\big(\log(p_{\theta_0}^{(n)}/p_{\theta_1}^{(n)})(X^{(n)}) - P_{\theta_0}^{(n)}\log(p_{\theta_0}^{(n)}/p_{\theta_1}^{(n)})\big)} \\ &\leq Pe^{\sum_{i=1}^n\varepsilon_i\lambda\big(\theta_{0,i}-\theta_{1,i}\big)} \leq e^{\lambda^2n\ell_n^2(\theta_0,\theta_1)/2}. \end{aligned}$$

Secondly consider Laplace regression. Note

$$\begin{aligned}\log(p^{(n)}_{\theta_0}/p^{(n)}_{\theta_1})(X^{(n)}) &= \sum_{i=1}^n \big[\lvert X_i-\theta_{1,i}\rvert - \lvert X_i-\theta_{0,i}\rvert\big],\\ P^{(n)}_{\theta_0}\log(p^{(n)}_{\theta_0}/p^{(n)}_{\theta_1}) &= \sum_{i=1}^n \mathbb{E}\big[\lvert\varepsilon_i+\theta_{0,i}-\theta_{1,i}\rvert - \lvert\varepsilon_i\rvert\big].\end{aligned}$$

For any $v\in\mathbb{R}$, let $\varphi(v)\equiv\mathbb{E}\big(\lvert\varepsilon+v\rvert-\lvert\varepsilon\rvert\big)$. Clearly φ is twice differentiable with a strictly positive second derivative on compacta. Since $\lvert\theta_{0,i}-\theta_{1,i}\rvert\leq M$, this implies that there exists some $C_M>1$ such that $C_M^{-1}\lvert\theta_{0,i}-\theta_{1,i}\rvert^2\leq\varphi(\theta_{0,i}-\theta_{1,i})\leq C_M\lvert\theta_{0,i}-\theta_{1,i}\rvert^2$. Hence $P^{(n)}_{\theta_0}\log(p^{(n)}_{\theta_0}/p^{(n)}_{\theta_1})\asymp_M n\ell_n^2(\theta_0,\theta_1)$. To verify the local Gaussianity condition, note that

$$\lvert Z_i\rvert\equiv\lvert\big[\lvert\varepsilon_i+\theta_{0,i}-\theta_{1,i}\rvert-\lvert\varepsilon_i\rvert\big]-\mathbb{E}\big[\lvert\varepsilon_i+\theta_{0,i}-\theta_{1,i}\rvert-\lvert\varepsilon_i\rvert\big]\rvert\leq 2\lvert\theta_{0,i}-\theta_{1,i}\rvert,$$

so it follows from the Hoffmann-Jorgensen inequality (cf. Proposition A.1.6 of [vdVW96]) that

$$\begin{aligned}&\Big\lVert\log\log(p^{(n)}_{\theta_0}/p^{(n)}_{\theta_1})-P^{(n)}_{\theta_0}\log(p^{(n)}_{\theta_0}/p^{(n)}_{\theta_1})\Big\rVert_{\psi_2}\\ &=\Big\lVert\sum_{i=1}^n Z_i\Big\rVert_{\psi_2}\lesssim\Big\lVert\sum_{i=1}^n Z_i\Big\rVert_1+\bigg(\sum_{i=1}^n\lVert Z_i\rVert_{\psi_2}^2\bigg)^{1/2}\\ &\lesssim\bigg(\sum_{i=1}^n Z_i^2\bigg)^{1/2}+\bigg(\sum_{i=1}^n\lVert Z_i\rVert_{\psi_2}^2\bigg)^{1/2}\lesssim\big(n\ell_n^2(\theta_0,\theta_1)\big)^{1/2},\end{aligned}$$

where $\lVert\cdot\rVert_\psi$ denotes the usual Orlicz norm given by $\lVert X\rVert_\psi\equiv\inf\{C>0:\mathbb{E}\psi(\lvert X\rvert/C)\leq 1\}$, and $\psi_2(x)=e^{x^2}-1$. Hence

$$P^{(n)}_{\theta_0}e^{\lambda(\log(p^{(n)}_{\theta_0}/p^{(n)}_{\theta_1})-P^{(n)}_{\theta_0}\log(p^{(n)}_{\theta_0}/p^{(n)}_{\theta_1})}\leq e^{C\lambda^2 n\ell_n^2(\theta_0,\theta_1)},$$

where $C>0$ is an absolute constant.

Next consider binary regression. Note

$$\begin{aligned}\log(p^{(n)}_{\theta_0}/p^{(n)}_{\theta_1})(X^{(n)}) &= \sum_{i=1}^n X_i\log\frac{\theta_{0,i}}{\theta_{1,i}}+(1-X_i)\log\frac{1-\theta_{0,i}}{1-\theta_{1,i}},\\ P_{\theta_0^{(n)}}\log(p^{(n)}_{\theta_0}/p^{(n)}_{\theta_1}) &= \sum_{i=1}^n\theta_{0,i}\log\frac{\theta_{0,i}}{\theta_{1,i}}+(1-\theta_{0,i})\log\frac{1-\theta_{0,i}}{1-\theta_{1,i}}.\end{aligned}$$

Using the inequality $cx\leq\log(1+x)\leq x$ for all $-1<x\leq c'$ for some $c>0$ depending on $c'>-1$ only, we have shown $P_{\theta_0^{(n)}}\log(p^{(n)}_{\theta_0}/p^{(n)}_{\theta_1})\asymp n\ell_n^2(\theta_0,\theta_1)$ under the assumed condition that $\Theta_n\subset[\eta,1-\eta]^n$. Now we verify the local Gaussianity condition:

$$P^{(n)}_{\theta_0}e^{\lambda\left(\log(p^{(n)}_{\theta_0}/p^{(n)}_{\theta_1})-P_{\theta_0^{(n)}}\log(p^{(n)}_{\theta_0}/p^{(n)}_{\theta_1})\right)}=P^{(n)}_{\theta_0}e^{\lambda\sum_{i=1}^n(X_i-\theta_{0,i})t_i}\leq e^{\lambda^2\sum_{i=1}^n t_i^2/8}$$

where $t_i \equiv t_i(\theta_0, \theta_1) = \log\big(\frac{\theta_{0,i}}{1-\theta_{0,i}} \cdot \frac{1-\theta_{1,i}}{\theta_{1,i}}\big)$ and the last inequality follows from Hoeffding's inequality (cf. Section 2.6 of [BLM13]). The claim follows by noting that $t_i^2 = \big[\log\big(\frac{\theta_{0,i}-\theta_{1,i}}{(1-\theta_{0,i})\theta_{1,i}} + 1\big)\big]^2 \asymp (\theta_{0,i} - \theta_{1,i})^2$ by the assumed condition and the aforementioned inequality $\log(1 + x) \asymp x$ in a constrained range.

Finally consider Poisson regression. It is easy to see that

$$\log(p_{\theta_0}^{(n)}/p_{\theta_1}^{(n)})(X^{(n)}) = \sum_{i=1}^n X_i \log\frac{\theta_{0,i}}{\theta_{1,i}} + (\theta_{1,i} - \theta_{0,i}),$$

$$P_{\theta_0}^{(n)} \log(p_{\theta_0}^{(n)}/p_{\theta_1}^{(n)}) = \sum_{i=1}^n \theta_{0,i} \log\frac{\theta_{0,i}}{\theta_{1,i}} + (\theta_{1,i} - \theta_{0,i}).$$

Note that for any $1/M \leq p, q \leq M$,

$$p\log\frac{p}{q} - (p - q) = p\big(-\log\frac{q}{p} - 1 + \frac{q}{p}\big) \asymp p \cdot (\frac{q}{p} - 1)^2 \asymp (p - q)^2,$$

where in the middle we used the fact that $-\log x - 1 + x \asymp (x-1)^2$ for x bounded away from 0 and ∞. This shows that $P_{\theta_0}^{(n)} \log(p_{\theta_0}^{(n)}/p_{\theta_1}^{(n)}) \asymp n\ell_n^2(\theta_0, \theta_1)$. Hence

$$P_{\theta_0}^{(n)} e^{\lambda\big(\log(p_{\theta_0}^{(n)}/p_{\theta_1}^{(n)}) - P_{\theta_0^{(n)}} \log(p_{\theta_0}^{(n)}/p_{\theta_1}^{(n)})\big)}$$
$$\leq P_{\theta_0}^{(n)} e^{\lambda \sum_{i=1}^n (X_i - \theta_{0,i})t_i} \leq e^{\sum_{i=1}^n \theta_{0,i}(e^{\lambda t_i} - 1 - \lambda t_i)},$$

where $t_i = \log(\theta_{0,i}/\theta_{1,i})$. Now for any $|\lambda| \leq 1$, we have $e^{\lambda t_i} - 1 - \lambda t_i \asymp \lambda^2 t_i^2$. On the other hand, $\theta_{0,i} t_i^2 = \theta_{0,i}\left(\log(\theta_{0,i}/\theta_{1,i})\right)^2 \asymp (\theta_{0,i} - \theta_{1,i})^2$, completing the proof. □

Proof of Corollary 3.2. The claim follows from Lemma 3.1 and Theorem 2.3. □

Proof of Lemma 3.14. Since the log-likelihood ratio for $X_1, \ldots, X_n$ can be decomposed into sums of the log-likelihood ratio for single samples, and the log-likelihood ratio is uniformly bounded over $\mathcal{F}$ (since $\mathcal{G}$ is bounded), classical Bernstein inequality applies to see that for any couple (f_0, f_1), the local Gaussianity condition in Assumption A holds with $v = \kappa_g n \mathrm{Var}_{f_0}(\log f_0/f_1), c = \kappa_\Gamma$ where κ_g, κ_Γ depend only on $\mathcal{G}$. Hence we only need to verify that $\mathrm{Var}_{f_0}(\log f_0/f_1) \lesssim h^2(f_0, f_1)$ and $P_{f_0}(\log f_0/f_1) \asymp h^2(f_0, f_1)$. This can be seen by Lemma 8 of [GvdV07b] and the fact that Hellinger metric is dominated by the Kullback-Leiber divergence. □

Proof of Corollary 3.15. The claim follows from Lemma 3.14 and Theorem 2.3. □

Lemma B.1. *Let $Z \geq 0$ be a non-negative random variable bounded by $M > 0$. Then $\mathbb{E}\exp(Z) \leq \exp(e^M \mathbb{E}Z)$.*

Proof. Note that $\log \mathbb{E}\exp(Z) = \log(\mathbb{E}[\exp(Z)-1]+1) \leq \mathbb{E}[\exp(Z)-1] \leq e^M \mathbb{E}Z$, where the last inequality follows from Taylor expansion $e^x - 1 = \sum_{k=1}^n x^k/k! \leq x\sum_{k\geq 1} M^{k-1}/k! \leq xe^M$ for $x \geq 0$. □

Proof of Lemma 3.17. We omit explicit dependence of M on the notation $d_{r,M}$ and r_M in the proof. Let $P_{f_0}^{(n)}$ denote the probability measure induced by the joint distribution of $(X_0, \ldots, X_n)$ where X_0 is distributed according to the stationary density q_{f_0}. Easy computation shows that

$$\log(p_{f_0}^{(n)}/p_{f_1}^{(n)}) = \sum_{i=0}^{n-1}\bigg[\varepsilon_{i+1}(f_0(X_i) - f_1(X_i)) + \frac{1}{2}(f_0(X_i) - f_1(X_i))^2\bigg],$$

$$P_{f_0}^{(n)} \log(p_{f_0}^{(n)}/p_{f_1}^{(n)}) = \frac{n}{2}\int (f_0 - f_1)^2 q_{f_0}\ \mathrm{d}\lambda.$$

Here λ denotes the Lebesgue measure on $\mathbb{R}$. By the arguments on page 209 of [GvdV07a], we see that $r \lesssim q_{f_0} \lesssim r$. Hence we only need to verify the local Gaussianity condition. By Cauchy-Schwarz,

$$\text{(B.1)} \qquad \big[P_{f_0}^{(n)} e^{\lambda \log(p_{f_0}^{(n)}/p_{f_1}^{(n)})}\big]^2 \leq P_{f_0}^{(n)} e^{2\lambda\sum_{i=0}^{n-1}\varepsilon_{i+1}(f_0(X_i)-f_1(X_i))} \\ \times P_{f_0}^{(n)} e^{\lambda\sum_{i=0}^{n-1}(f_0(X_i)-f_1(X_i))^2} \equiv (I) \times (II).$$

The first term (I) can be handled by an inductive calculation. First note that for any $|\mu| \leq 2$ and $X_1 \in \mathbb{R}$,

$$\text{(B.2)} \quad P_{p(\cdot|X_1)} e^{\mu^2(f_0(X_2)-f_1(X_2)^2} \leq e^{e^{16M^2}\mu^2 P_{p(\cdot|X_1)}(f_0-f_1)(X_2)^2} \leq e^{C_M\mu^2 d_r^2(f_0,f_1)}$$

where the first inequality follows from Lemma B.1 and the second inequality follows from $r(\cdot) \lesssim p_f(\cdot|x) \lesssim r(\cdot)$ holds for all $x \in \mathbb{R}$ where the constant involved depends only on M. Let $S_n \equiv \sum_{i=0}^{n-1}\varepsilon_{i+1}(f_0(X_i)-f_1(X_i))$ and $\boldsymbol{\varepsilon}_n \equiv (\varepsilon_1, \ldots, \varepsilon_n)$. Then for $|\lambda| \leq 1$, let $\mu \equiv 2\lambda$,

$$\begin{aligned}
&P_{f_0}^{(n)} e^{2\lambda S_n} = P_{f_0}^{(n)} e^{\mu S_n} = \mathbb{E}_{X_0,\boldsymbol{\varepsilon}_{n-1}}\big[e^{\mu S_{n-1}}\mathbb{E}_{\varepsilon_n} e^{\mu\varepsilon_n(f_0(X_{n-1})-f_1(X_{n-1}))}\big] \\
&\leq \mathbb{E}_{X_0,\boldsymbol{\varepsilon}_{n-1}}\big[e^{\mu S_{n-1}} e^{\mu^2(f_0(X_{n-1})-f_1(X_{n-1}))^2/2}\big] \\
&\leq \mathbb{E}_{X_0,\boldsymbol{\varepsilon}_{n-2}}\big[e^{\mu S_{n-2}}\mathbb{E}_{\varepsilon_{n-1}} e^{\mu\varepsilon_{n-1}(f_0(X_{n-2})-f_1(X_{n-2}))+\mu^2(f_0(X_{n-1})-f_1(X_{n-1}))^2/2}\big] \\
&\leq \mathbb{E}_{X_0,\boldsymbol{\varepsilon}_{n-2}}\big[e^{\mu S_{n-2}}\big(\mathbb{E}_{\varepsilon_{n-1}} e^{2\mu\varepsilon_{n-1}(f_0(X_{n-2})-f_1(X_{n-2}))}\big)^{1/2} \\
&\qquad\qquad \times \big(\mathbb{E}_{p(\cdot|X_{n-2})} e^{\mu^2(f_0(X_{n-1})-f_1(X_{n-1}))^2}\big)^{1/2}\big] \\
&\leq \mathbb{E}_{X_0,\boldsymbol{\varepsilon}_{n-2}}\big[e^{\mu S_{n-2}} e^{\mu^2(f_0(X_{n-2})-f_1(X_{n-2}))^2}\big] \cdot e^{C_M\mu^2 d_r^2(f_0,f_1)/2}
\end{aligned}$$

where the last inequality follows from (B.2). Now we can iterate the above calculation to see that $(I) \leq e^{C_M\lambda^2 n d_r^2(f_0,f_1)}$. Next we consider (II). Since for any non-negative random variables $Z_1, \ldots, Z_n$, we have $\mathbb{E}\prod_{i=1}^n Z_i \leq \prod_{i=1}^n(\mathbb{E}Z_i^n)^{1/n}$. So

$$(II) \leq \prod_{i=1}^n (P_{f_0}^{(n)} e^{n\lambda(f_0(X_i)-f_1(X_i))^2})^{1/n} = P_{q_{f_0}} e^{n\lambda(f_0(X_0)-f_1(X_0))^2},$$

where the last inequality follows by stationarity. On the other hand, by Jensen's inequality,

$$e^{-\lambda P_{f_0}^{(n)} \log(p_{f_0}^{(n)}/p_{f_1}^{(n)})} \le e^{-\frac{\lambda n}{2} P_{q_{f_0}}(f_0-f_1)^2} \le P_{q_{f_0}} e^{-\lambda n (f_0(X_0)-f_1(X_0))^2/2}.$$

Collecting the above estimates, we see that for $|\lambda| \le 1$,

$$\begin{aligned} &P_{f_0^{(n)}} e^{\lambda \log(p_{f_0}^{(n)}/p_{f_1}^{(n)}) - P_{f_0^{(n)}} \log(p_{f_0}^{(n)}/p_{f_1}^{(n)})} \\ &\le \sqrt{(I)\cdot(II)} e^{-\lambda P_{f_0}^{(n)} \log(p_{f_0}^{(n)}/p_{f_1}^{(n)})} \le e^{C_M' \lambda^2 n d_r^2(f_0,f_1)}, \end{aligned}$$

completing the proof. □

Proof of Corollary 3.18. The claim follows from Lemma 3.17 and Theorem 2.3. □

Proof of Lemma 3.19. For any $g \in \mathcal{G}$, let $p_g^{(n)}$ denote the probability density function of a n-dimensional multivariate normal distribution with covariance matrix $\Sigma_g \equiv T_n(f_g)$, and $P_g^{(n)}$ the expectation taken with respect to the density $p_g^{(n)}$. Then for any $g_0, g_1 \in \mathcal{G}$,

$$\begin{aligned} \text{(B.3)} \quad \log \frac{p_{g_0}^{(n)}}{p_{g_1}^{(n)}}(X^{(n)}) &= -\frac{1}{2}(X^{(n)})^\top (\Sigma_{g_0}^{-1} - \Sigma_{g_1}^{-1}) X^{(n)} - \frac{1}{2} \log\det(\Sigma_{g_0}\Sigma_{g_1}^{-1}), \\ P_{g_0}^{(n)} \log \frac{p_{g_0}^{(n)}}{p_{g_1}^{(n)}} &= -\frac{1}{2} \operatorname{tr}(I - \Sigma_{g_0}\Sigma_{g_1}^{-1}) - \frac{1}{2} \log\det(\Sigma_{g_0}\Sigma_{g_1}^{-1}) \end{aligned}$$

where we used the fact that for a random vector X with covariance matrix Σ, $\mathbb{E} X^\top A X = \operatorname{tr}(\Sigma A)$. Let $G \equiv \Sigma_{g_0}^{-1/2} X^{(n)} \sim \mathcal{N}(0, I)$ under $P_{g_0}^{(n)}$, and $B \equiv I - \Sigma_{g_0}^{1/2} \Sigma_{g_1}^{-1} \Sigma_{g_0}^{1/2}$, then

$$\begin{aligned} Y_n &\equiv \log(p_{g_0}^{(n)}/p_{g_1}^{(n)})(X^{(n)}) - P_{g_0}^{(n)} \log(p_{g_0}^{(n)}/p_{g_1}^{(n)}) \\ &= -\frac{1}{2}[(X^{(n)})^\top (\Sigma_{g_0}^{-1} - \Sigma_{g_1}^{-1}) X^{(n)} - \operatorname{tr}(I - \Sigma_{g_0}\Sigma_{g_1}^{-1})] = -\frac{1}{2}[G^\top B G - \operatorname{tr}(B)]. \end{aligned}$$

Let $B = U^\top \Lambda U$ be the spectral decomposition of B where U is orthonormal and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix. Then we can further compute

$$-2Y_n =_d G^\top \Lambda G - \operatorname{tr}(\Lambda) = \sum_{i=1}^n \lambda_i (g_i^2 - 1),$$

where $g_1, \ldots, g_n$'s are i.i.d. standard normal. Note that for any $|t| < 1/2$,

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{t(x^2-1)} e^{-x^2/2} \, \mathrm{d}x = \frac{e^{-t}}{\sqrt{1-2t}} = e^{\frac{1}{2}(-\log(1-2t)-2t)} \le e^{t^2/(1-2t)},$$

where the inequality follows from

$$-\log(1-2t)-2t=\sum_{k\geq 2}\frac{1}{k}(2t)^k=4t^2\sum_{k\geq 0}\frac{1}{k+2}(2t)^k\leq\frac{2t^2}{1-2t}.$$

With $t=-\lambda\lambda_i/2$, we have that for any $|\lambda|<1/\max_i\lambda_i$,

$$\begin{aligned}\mathbb{E}e^{\lambda Y_n}&=\prod_{i=1}^{n}\mathbb{E}e^{-\lambda\cdot\lambda_i(g_i^2-1)/2}=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-\lambda\cdot\lambda_i(x^2-1)/2}e^{-x^2/2}\,\mathrm{d}x\\&\leq\prod_{i=1}^{n}e^{\frac{\lambda^2\lambda_i^2}{4+4\lambda\lambda_i}}\leq\exp\left(\frac{\lambda^2\sum_i\lambda_i^2}{4-4|\lambda|\max_i|\lambda_i|}\right).\end{aligned}$$

Denote $\|\cdot\|$ and $\|\cdot\|_F$ the matrix operator norm and Frobenius norm respectively. By the arguments on page 203 of [GvdV07a], we have $\|\Sigma_g\|\leq 2\pi\|e^g\|_\infty$ and $\|\Sigma_g^{-1}\|\leq(2\pi)^{-1}\|e^{-g}\|_\infty$. Since $\mathcal{G}$ is a class of uniformly bounded function classes, the spectrum of the covariance matrices Σ_g and their inverses running over g must be bounded. Hence

$$\max_i|\lambda_i|=\|B\|=\|(\Sigma_{g_1}-\Sigma_{g_0})\Sigma_{g_1}^{-1}\|\leq\|\Sigma_{g_1}-\Sigma_{g_0}\|\|\Sigma_{g_1}^{-1}\|\leq C_{\mathcal{G}}<\infty.$$

Next, note that

$$\begin{aligned}\left(\sum_i\lambda_i^2\right)^{1/2}&=(\mathrm{tr}(BB^\top))^{1/2}=\|B\|_F=\|(\Sigma_{g_1}-\Sigma_{g_0})\Sigma_{g_1}^{-1}\|_F\\&\leq\|\Sigma_{g_1}^{-1}\|\|\Sigma_{g_1}-\Sigma_{g_0}\|_F\leq C'_{\mathcal{G}}\sqrt{nD_n^2(g_0,g_1)},\end{aligned}$$

where in the first inequality we used $\|MN\|_F=\|NM\|_F$ for symmetric matrices M,N and the general rule $\|PQ\|_F\leq\|P\|\|Q\|_F$. Collecting the above estimates we see that Assumption A is satisfied for $v=\kappa_g nD_n^2(g_0,g_1)$ and $c=\kappa_\Gamma$ for constants κ_g,κ_Γ depending on $\mathcal{G}$ only.

Finally we relate $n^{-1}P_{g_0}^{(n)}\log(p_{g_0}^{(n)}/p_{g_1}^{(n)})$ and $D_n^2(g_0,g_1)$. First by (B.3), we have

$$\begin{aligned}&P_{g_0}^{(n)}\log(p_{g_0}^{(n)}/p_{g_1}^{(n)})=-\frac{1}{2}\mathrm{tr}(I-\Sigma_{g_0}\Sigma_{g_1}^{-1})-\frac{1}{2}\log\det(\Sigma_{g_0}\Sigma_{g_1}^{-1})\\&=\frac{1}{2}(\mathrm{tr}(\Sigma_{g_1}^{-1/2}(\Sigma_{g_0}-\Sigma_{g_1})\Sigma_{g_1}^{-1/2})-\log\det(I+\Sigma_{g_1}^{-1/2}(\Sigma_{g_0}-\Sigma_{g_1})\Sigma_{g_1}^{-1/2}))\\&\leq\frac{1}{4}\|I-\Sigma_{g_0}\Sigma_{g_1}^{-1}\|_F^2\leq\frac{1}{4}\|\Sigma_{g_1}-\Sigma_{g_0}\|_F^2\|\Sigma_{g_1}^{-1}\|^2\leq C''_{\mathcal{G}}nD_n^2(g_0,g_1).\end{aligned}$$

Here in the second line we used the fact that $\det(AB^{-1})=\det(I+B^{-1/2}(A-B)B^{-1/2})$, and in the third line we used the fact $-\log\det(I+A)+\mathrm{tr}(A)\leq\frac{1}{2}\mathrm{tr}(A^2)$ for any p.s.d. matrix A, due to the inequality $\log(1+x)-x\geq-\frac{1}{2}x^2$ for all $x\geq 0$. On the other hand, by using the reversed inequality $\log(1+x)-x\leq-cx^2$ for all $0\leq x\leq c'$ where c is a constant depending only on c', we can establish $P_{g_0}^{(n)}\log(p_{g_0}^{(n)}/p_{g_1}^{(n)})\geq C'''_{\mathcal{G}}nD_n^2(g_0,g_1)$, thereby completing the proof. □

Proof of Corollary 3.20. The claim follows from Lemma 3.19 and Theorem 2.3. □

Proof of Lemma 3.21. Note that

$$\begin{aligned}\log(p^{(n)}_{\Sigma_0}/p^{(n)}_{\Sigma_1})(X^{(n)}) &= -\sum_{i=1}^n \bigg[\frac{1}{2}X_i^\top(\Sigma_0^{-1}-\Sigma_1^{-1})X_i - \frac{1}{2}\log\det(\Sigma_0\Sigma_1^{-1})\bigg],\\ P^{(n)}_{\Sigma_0}\log(p^{(n)}_{\Sigma_0}/p^{(n)}_{\Sigma_1}) &= -\frac{n}{2}\operatorname{tr}(I-\Sigma_0\Sigma_1^{-1}) - \frac{n}{2}\log\det(\Sigma_0\Sigma_1^{-1}).\end{aligned}$$

The rest of the proof proceeds along the same line as in Lemma 3.19. □

Proof of Corollary 3.22. The claim follows from Lemma 3.21 and Theorem 2.3. □

Proof of Lemma 3.24. Note that

$$\begin{aligned}\log(p^{(n)}_{\theta_0}/p^{(n)}_{\theta_1})(X^{(n)},Y^{(n)}) = &\sum_{X_i\in\Gamma_0\cap\Gamma_1}\log\frac{f(Y_i;\xi_0)}{f(Y_i;\xi_1)} + \sum_{X_i\in\Gamma_0^c\cap\Gamma_1^c}\log\frac{f(Y_i;\rho_0)}{f(Y_i;\rho_1)}\\ &+\sum_{X_i\in\Gamma_0\cap\Gamma_1^c}\log\frac{f(Y_i;\xi_0)}{f(Y_i;\rho_1)} + \sum_{X_i\in\Gamma_0^c\cap\Gamma_1}\log\frac{f(Y_i;\rho_0)}{f(Y_i;\xi_1)}.\end{aligned}$$

Then we may verify Assumption A along the lines in the proof of Lemma 3.1, by considering each of the terms above by virtue of independence of X_i's. □

Proof of Lemma 3.25. Let $r>0$ be such that $d_n^2(\theta_0,\theta_1)=r^2$. By definition of d_n, we have $\|\rho_0-\rho_1\|_2^2\leq r^2/\lambda(\Gamma_0^c\cap\Gamma_1^c)\leq r^2/\lambda_0^2$. This implies that

$$\|\xi_0-\rho_1\|_2\geq\|\xi_0-\rho_0\|_2-\|\rho_0-\rho_1\|_2\geq r_0-\frac{r}{\lambda_0}\geq\frac{r_0}{2}$$

under the condition $r^2\leq\lambda_0^2r_0^2/4$. Hence $\lambda(\Gamma_0\cap\Gamma_1^c)\leq\frac{r^2}{\|\xi_0-\rho_1\|_2^2}\leq\frac{4r^2}{r_0^2}$, implying

$$\lambda(\Gamma_0\cap\Gamma_1)=\lambda(\Gamma_0)-\lambda(\Gamma_0\cap\Gamma_1^c)\geq\lambda(\Gamma_0)-\frac{4r^2}{r_0^2}\geq\frac{\lambda(\Gamma_0)}{2}$$

under the condition $r^2\leq\lambda(\Gamma_0)r_0^2/8$. This further implies that $\|\xi_0-\xi_1\|_2^2\leq\frac{r^2}{\lambda(\Gamma_0\cap\Gamma_1)}\leq\frac{2r^2}{\lambda(\Gamma_0)}$, whence

$$\|\rho_0-\xi_1\|_2\geq\|\rho_0-\xi_0\|_2-\|\xi_0-\xi_1\|_2\geq r_0-\sqrt{2/\lambda(\Gamma_0)}r\geq\frac{r_0}{2}$$

under the condition $r^2\leq\lambda(\Gamma_0)r_0^2/8$. Hence

$$\lambda(\Gamma_0^c\cap\Gamma_1)\leq\frac{r^2}{\|\rho_0-\xi_1\|_2^2}\leq\frac{4r^2}{r_0^2}.$$

The claim follows by noting that $\lambda(\Gamma_0\Delta\Gamma_1)=\lambda(\Gamma_0^c\cap\Gamma_1)+\lambda(\Gamma_0\cap\Gamma_1^c)$. □

Proof of Corollary 3.26. By Lemma 3.24 and Theorem 2.3, the claim of the corollary holds for d_n. Using Lemma 3.25, for n large, we may replace d_n with $\lambda(\cdot\Delta\cdot)$. □

Proof of Corollary 3.28. The main modification of the proof lies in part of Lemma 2.1. The modified Lemma 2.1 takes the following form: fix $f_0 \le f_1$, there exists some test ϕ_n such that

$$\sup_{f\ge f_1:\bar{L}_1(f,f_1)\le c_5^2\bar{L}_1(f_1,f_0)} \big(P_{f_0}^{(n)}\phi_n + P_f^{(n)}(1-\phi_n)\big) \le c_6 e^{-c_7 n\bar{L}_1(f_1,f_0)},$$

where $c_5 \le 1/4, c_6 \in [2,\infty), c_7 \in (0,1)$ are absolute constants.

In particular, the test ϕ_n is constructed in the 'same way' as in the proof of Lemma 2.1 with a modified way of writing:

$$\phi_n \equiv \mathbf{1}\big(\log(\mathrm{d}P_{f_1}^{(n)}/\mathrm{d}P_{f_0}^{(n)}) \ge cn\bar{L}_1(f_1,f_0)\big).$$

Now for type I error,

$$\begin{aligned} P_{f_0}^{(n)}\phi_n &= P_{f_0}^{(n)}(\forall i : f_1(X_i) \le Y_i) = P_{f_0}^{(n)}(N(\{(x,y) : y \le f_1(x)\}) = 0) \\ &= e^{-n\bar{L}_1(f_1,f_0)}. \end{aligned}$$

Here the last equality follows as

$$\begin{aligned} \int_{(x,y):y\le f_1(x)} \lambda_{f_0}(x,y)\ \mathrm{d}x\mathrm{d}y &= \int_0^1 \mathrm{d}x \int_{-\infty}^{f_1(x)} n\mathbf{1}_{f_0(x)\le y}\ \mathrm{d}y \\ &= n\int_0^1 \big(f_1(x) - f_0(x)\big)\ \mathrm{d}x. \end{aligned}$$

For type II error, note that as soon as $f \ge f_1$,

$$\begin{aligned} &P_f^{(n)}(1-\phi_n) = P_f^{(n)}\big(\log(\mathrm{d}P_{f_1}^{(n)}/\mathrm{d}P_{f_0}^{(n)}) < cn\bar{L}_1(f_1,f_0)\big) \\ &= P_f^{(n)}(\mathbf{1}_{\forall i:f_1(X_i)\le Y_i} < e^{-(1-c)n\bar{L}_1(f_1,f_0)}) = P_f^{(n)}(\exists i : f_1(X_i) > Y_i) = 0. \end{aligned}$$

This proves the modified version of Lemma 2.1 in the current setting. Then in the proof of Lemma A.4, the entropy condition needs to be replaced by the entropy with left bracketing, due to the reasoning towards the last display in the proof of Lemma A.4. Now in the proof of Proposition A.2, we apply Lemma A.4 with the set restricted to $f \ge f_0$. The set in the control of denominator in (A.8) can be restricted to $f \ge f_{0,m}$. The rest of the proofs carry over exactly so we omit the details. □

Proof of Corollary 3.29. The proof is a combination of the change of measure idea in the current paper combined with the results in [RSH17]. Let $m \in \mathcal{M}$ be such that $\delta_{n,m}^2 \ge L_1(f_0, f_{0,m})$. Note that condition (ii) entails that

$$\begin{aligned} &\Pi_n(f \in \mathcal{F} : L_1(f, f_{0,m}) \le C_2\delta_{n,m}^2, f \le f_{0,m}) \\ &\ge \lambda_n(m)\Pi_{n,m}(f \in \mathcal{F}_m : L_1(f, f_{0,m}) \le C_2\delta_{n,m}^2, f \le f_{0,m}) \ge e^{-C_2' n\delta_{n,m}^2} \end{aligned}$$

Then use Theorem 2.3 of [RSH17], we conclude that

$$P_{f_{0,m}}^{(n)} \Pi_n(f : L_1(f, f_{0,m}) \geq C_3 K \delta_{n,m}^2 | (X^{(n)}, Y^{(n)})) \leq C_3 e^{-nK\delta_{n,m}^2/C_3},$$

where $K > 0$ is a constant to be chosen later. Hence

$$\begin{aligned} & P_{f_0}^{(n)} \Pi_n(f : L_1(f, f_0) \geq L_1(f_0, f_{0,m}) + C_3 K \delta_{n,m}^2 | (X^{(n)}, Y^{(n)})) \\ & \leq P_{f_0}^{(n)} \Pi_n(f : L_1(f, f_{0,m}) \geq C_3 K \delta_{n,m}^2 | (X^{(n)}, Y^{(n)})) \\ & = P_{f_{0,m}}^{(n)} \Pi_n(f : L_1(f, f_{0,m}) \geq C_3 K \delta_{n,m}^2 | (X^{(n)}, Y^{(n)})) \big(\mathrm{d}P_{f_0}^{(n)} / \mathrm{d}P_{f_{0,m}}^{(n)}\big) \\ & \leq C_3 e^{-nK\delta_{n,m}^2/C_3 + nL_1(f_0, f_{0,m})} \leq C_3 e^{-n\delta_{n,m}^2}, \end{aligned}$$

by choosing $K = 2C_3$. We may similar consider $m \in \mathcal{M}$ such that $\delta_{n,m}^2 < L_1(f_0, f_{0,m})$. □

Appendix C: Proofs in Section 3 Part II: results for applications

C.1. Proof of Theorem 3.7

Lemma C.1. *Let $r \in \mathcal{I}$. Suppose that the linear map $\mathcal{X} : \mathbb{R}^{m_1 \times m_2} \to \mathbb{R}^n$ is uniform RIP($\boldsymbol{\nu}$;$\mathcal{I}$). Then for any $\varepsilon > 0$ and $A_0 \in \mathbb{R}^{m_1 \times m_2}$ such that* $\mathrm{rank}(A_0) \leq r$*, we have*

$$\log \mathcal{N}\big(c_5 \varepsilon, \{f_A \in \mathcal{F}_r : \ell_n(f_A, f_{A_0}) \leq 2\varepsilon\}, \ell_n\big) \leq 2(m_1 + m_2) r \cdot \log\big(18\bar{\nu}/c_5 \underline{\nu}\big).$$

We will need the following result.

Lemma C.2. *Let $S(r, B) = \{A \in \mathbb{R}^{m_1 \times m_2} : \mathrm{rank}(A) \leq r, \|A\|_2 \leq B\}$. Then $\mathcal{N}\big(\varepsilon, S(r, B), \|\cdot\|_2\big) \leq \big(\frac{9B}{\varepsilon}\big)^{(m_1+m_2-1)r}$.*

Proof of Lemma C.2. The case for $B = 1$ follows from Lemma 3.1 of [CP11] and the general case follows by a scaling argument. We omit the details. □

Proof of Lemma C.1. We only need to consider the case $r \leq r_{\max}$. First note that the entropy in question equals

$$\log \mathcal{N}\big(c_5 \sqrt{n} \varepsilon, \{\mathcal{X}(A - A_0) : \|\mathcal{X}(A - A_0)\|_2 \leq 2\sqrt{n}\varepsilon, \mathrm{rank}\, A \leq r\}, \|\cdot\|_2\big).$$

By uniform RIP($\boldsymbol{\nu}$;$\mathcal{I}$), the set to be covered is contained in

$$\{\mathcal{X}(A - A_0) : \|A - A_0\|_2 \leq 2\varepsilon/\underline{\nu}, \mathrm{rank}\, A \leq r\} \subset \mathcal{X}(S(2r, 2\varepsilon/\underline{\nu})).$$

On the other hand, again by uniform RIP($\boldsymbol{\nu}$;$\mathcal{I}$), a $c_5\varepsilon/\bar{\nu}$-cover of the set $S(2r, 2\varepsilon/\underline{\nu})$ under the Frobenius norm $\|\cdot\|_2$ induces a $c_5\sqrt{n}\varepsilon$-cover of $\mathcal{X}(S(2r, 2\varepsilon/\underline{\nu}))$ under the Euclidean $\|\cdot\|_2$ norm. This implies that the entropy can be further bounded from above by

$$\log \mathcal{N}\big(c_5 \varepsilon/\bar{\nu}, S(2r, 2\varepsilon/\underline{\nu}), \|\cdot\|_2\big) \leq 2(m_1 + m_2) r \cdot \log\big(18\bar{\nu}/c_5 \underline{\nu}\big),$$

where the last inequality follows from Lemma C.2. □

Now we take $\delta_{n,r}^2 = \big(\frac{4\log(18\bar{\nu}/c_5\underline{\nu})}{c_7} \vee \frac{1}{\eta}\big)\frac{\cdot(m_1+m_2)r\log\bar{m}}{n}$. Clearly $\delta_{n,r}^2$ satisfies (2.5) with $\mathfrak{c} = \gamma = 1, \mathfrak{h}_0 = \infty$.

Lemma C.3. *Suppose that $\mathcal{X} : \mathbb{R}^{m_1\times m_2} \to \mathbb{R}^n$ is uniform RIP($\boldsymbol{\nu}$;$\mathcal{I}$), and that (3.2) holds. Then (P2) in Assumption C holds.*

Proof of Lemma C.3. We only need to consider $r \le r_{\max}$. First note that

$$\begin{aligned}
\text{(C.1)}\quad & \Pi_{n,r}\left(\left\{f_A \in \mathcal{F}_r : \ell_n^2(f_A, f_{A_{0,r}}) \le \delta_{n,r}^2/c_3\right\}\right)\\
&= \Pi_G\left(\left\{A \in \mathbb{R}^{m_1\times m_2} : \|\mathcal{X}(A - A_{0,r})\|_2 \le \sqrt{n}\delta_{n,r}/\sqrt{c_3}, \mathrm{rank}(A) \le r\right\}\right)\\
&\ge \Pi_G\left(\left\{A \in \mathbb{R}^{m_1\times m_2} : \|A - A_{0,r}\|_2 \le \delta_{n,r}/\bar{\nu}\sqrt{c_3}, \mathrm{rank}(A) \le r\right\}\right).
\end{aligned}$$

Let $A_{0,r} \equiv \sum_{i=1}^r \sigma_i \bar{u}_i \bar{v}_i^\top$ be the spectral decomposition of $A_{0,r}$, and let $u_i \equiv \sqrt{\sigma_i}\bar{u}_i$ and $v_i \equiv \sqrt{\sigma_i}\bar{v}_i$. Then $A_{0,r} \equiv \sum_{i=1}^r u_i v_i^\top$. Now for $u_i^* \in B_{m_1}(u_i, \varepsilon)$ and $v_i^* \in B_{m_2}(v_i, \varepsilon)$, $i = 1, \ldots, r$, let $A^* \equiv \sum_{i=1}^r u_i^*(v_i^*)^\top$, then by noting that the Frobenius norm is sub-multiplicative and that $\|u_i\|_2 = \|v_i\|_2 = \sqrt{\sigma_i}$, we have for $\varepsilon \le 1$,

$$\begin{aligned}
\|A^* - A_{0,r}\|_2 &\le \sum_{i=1}^r \left(\|(u_i - u_i^*)v_i^\top\|_2 + \|u_i^*(v_i - v_i^*)^\top\|_2\right)\\
&\le \sum_{i=1}^r (\varepsilon\sqrt{\sigma_i} + (\sqrt{\sigma_i} + \varepsilon)\varepsilon) \le \rho_r\varepsilon,
\end{aligned}$$

where $\rho_r \equiv \sum_{i=1}^r (2\sqrt{\sigma_i} + 1)$. Now with $\bar{\varepsilon}_{n,r} \equiv \frac{\delta_{n,r}}{\bar{\nu}\sqrt{c_3}\rho_r} \wedge 1$ we see that (C.1) can be further bounded from below by

$$\begin{aligned}
&\Pi_G\big(\cap_{i=1}^r \{(u_i^*, v_i^*) : u_i^* \in B_{m_1}(u_i, \bar{\varepsilon}_{n,r}), v_i^* \in B_{m_2}(v_i, \bar{\varepsilon}_{n,r})\}\big)\\
&\ge (\tau_{r,g}^{\mathrm{tr}})^{(m_1+m_2)r} \prod_{i=1}^r \mathrm{vol}\left(B_{m_1}(u_i, \bar{\varepsilon}_{n,r})\right) \cdot \mathrm{vol}\left(B_{m_2}(v_i, \bar{\varepsilon}_{n,r})\right)\\
&\ge \left(\tau_{r,g}^{\mathrm{tr}} \cdot \bar{\varepsilon}_{n,r}\right)^{(m_1+m_2)r} v_{m_1}^r v_{m_2}^r \ge e^{-(m_1+m_2)r\cdot\left(\log\bar{m}/2 + \log\tau_{r,g}^{-1} + \log(\bar{\varepsilon}_{n,r}^{-1}\vee 1)\right)},
\end{aligned}$$

where $v_d = \mathrm{vol}(B_d(0,1))$, and $v_d \ge (1/\sqrt{d})^d$. The right side of the above display is bounded from below by $e^{-2n\delta_{n,r}^2}$, if we require

$$\max\left\{\log\tau_{r,g}^{-1}, \log(\bar{\varepsilon}_{n,r}^{-1} \vee 1)\right\} \le \log\bar{m}/(2\eta).$$

It is easy to calculate that

$$\begin{aligned}
(\bar{\varepsilon}_{n,r})^{-2} &\le \left(\bar{\nu}^2 c_3 \rho_r^2 \eta n\right) \vee 1 \le 8\eta\bar{\nu}^2 c_3 (1 \vee \sigma_{\max}(A_{0,r})) r_{\max}^2 n\\
&\le 4\bar{\nu}^2(1 \vee \sigma_{\max}(A_{0,r}))n^3 \le 4\bar{\nu}^2(1 \vee \sigma_{\max}(A_{0,r}))^2 n^4,
\end{aligned}$$

by using $r_{\max} \le n$ and $c_3 = 1$. Now the conclusion follows by noting that (3.2) implies the requirement. □

Proof of Theorem 3.7. The theorem follows by Corollary 3.2, Proposition 2.2 coupled with Lemmas C.1 and C.3. □

C.2. Proof of Theorem 3.8

Lemma C.4. *Let $n \geq 2$. Then for any $g \in \mathcal{F}_m$, $\log \mathcal{N}\big(c_5\varepsilon, \{f \in \mathcal{F}_m : \ell_n(f,g) \leq 2\varepsilon\}, \ell_n\big) \leq 2\log(6/c_5) \cdot m\log(en)$.*

Proof of Lemma C.4. Let $\mathscr{Q}_m$ denote all m-partitions of the design points $x_1, \ldots, x_n$. Then it is easy to see that $|\mathscr{Q}_m| = \binom{n}{m-1}$. For a given m-partition $Q \in \mathscr{Q}_m$, let $\mathcal{F}_{m,Q} \subset \mathcal{F}_m$ denote all monotonic non-decreasing functions that are constant on the partition Q. Then the entropy in question can be bounded by

$$\log\bigg[\binom{n}{m-1} \max_{Q\in\mathscr{Q}_m} \mathcal{N}\big(c_5\varepsilon, \{f \in \mathcal{F}_{m,Q} : \ell_n(f,g) \leq 2\varepsilon\}, \ell_n\big)\bigg].$$

On the other hand, for any fixed m-partition $Q \in \mathscr{Q}_m$, the entropy term above equals $\mathcal{N}\big(c_5\sqrt{n}\varepsilon, \{\boldsymbol{\gamma} \in \mathcal{P}_{n,m,Q} : \|\boldsymbol{\gamma} - \boldsymbol{g}\|_2 \leq 2\sqrt{n}\varepsilon\}, \|\cdot\|_2\big)$, where $\mathcal{P}_{n,m,Q} \equiv \{(f(x_1), \ldots, f(x_n)) : f \in \mathcal{F}_{m,Q}\}$. By Pythagoras theorem, the set involved in the entropy is included in $\{\boldsymbol{\gamma} \in \mathcal{P}_{n,m,Q} : \|\boldsymbol{\gamma} - \pi_{\mathcal{P}_{n,m,Q}}(\boldsymbol{g})\|_2 \leq 2\sqrt{n}\varepsilon\}$ where $\pi_{\mathcal{P}_{n,m,Q}}$ is the natural projection from $\mathbb{R}^n$ onto the subspace $\mathcal{P}_{n,m,Q}$. Clearly $\mathcal{P}_{n,m,Q}$ is contained in a linear subspace with dimension no more than m. Using entropy result for the finite-dimensional space [Problem 2.1.6 in [vdVW96], page 94 combined with the discussion in page 98 relating the packing number and covering number],

$$\log \mathcal{N}\big(c_5\varepsilon, \{f \in \mathcal{F}_{m,Q} : \ell_n(f, f_{0,m}) \leq 2\varepsilon\}, \ell_n\big) \leq \log\Big(\frac{3\cdot 2\sqrt{n}\varepsilon}{c_5\sqrt{n}\varepsilon}\Big)^m = m\log(6/c_5).$$

The claim follows by combining the estimates and $\log\binom{n}{m-1} \leq m\log(en)$. □

Hence we can take $\delta_{n,m}^2 \equiv \big(\frac{4\log(6/c_5)}{c_7} \vee \frac{1}{\eta}\big)\frac{m\log(en)}{n}$. It is clear that (2.5) is satisfied with $\mathfrak{c} = \gamma = 1, \mathfrak{h}_0 = \infty$.

Lemma C.5. *Suppose that (3.5) holds. Then (P2) in Assumption C holds.*

Proof of Lemma C.5. Let $Q_{0,m} = \{I_k\}_{k=1}^m$ be the associated m-partition of $\{x_1, \ldots, x_n\}$ of $f_{0,m} \in \mathcal{F}_m$ with the convention that $\{I_k\} \subset \{x_1, \ldots, x_n\}$ is ordered from smaller values to bigger ones. Then it is easy to see that $\boldsymbol{\mu}_{0,m} = (\mu_{0,1}, \ldots, \mu_{0,m}) \equiv \big(f_{0,m}(x_{i(1)}), \ldots, f_{0,m}(x_{i(m)})\big) \in \mathbb{R}^m$ is well-defined and $\mu_{0,1} \leq \ldots \leq \mu_{0,m}$. It is easy to see that any $f \in \mathcal{F}_{m,Q_{0,m}}$ satisfying the property that $\sup_{1\leq k\leq m}|f(x_{i(k)}) - \mu_{0,k}| \leq \delta_{n,m}/\sqrt{c_3}$ leads to the error estimate $\ell_n^2(f, f_{0,m}) \leq \delta_{n,m}^2/c_3$. Hence

$$\begin{aligned}
&\Pi_{n,m}(\{f \in \mathcal{F}_m : \ell_n^2(f, f_{0,m}) \leq \delta_{n,m}^2/c_3\}) \\
&\geq \binom{n}{m-1}^{-1} \Pi_{\bar{g}_m}(\{f \in \mathcal{F}_{m,Q_{0,m}} : \ell_n^2(f, f_{0,m}) \leq \delta_{n,m}^2/c_3\}) \\
&\geq \binom{n}{m-1}^{-1} \Pi_{\bar{g}_m}(\{\boldsymbol{\mu} \in \mathbb{R}^m : \boldsymbol{\mu} \equiv \big(\mu_{0,k} + \varepsilon_k\big)_{k=1}^m, 0 \leq \varepsilon_1 \leq \ldots \leq \varepsilon_m \leq \delta_{n,m}/\sqrt{c_3}\})
\end{aligned}$$

$$
\begin{aligned}
&\geq \binom{n}{m-1}^{-1} \cdot \inf_{\substack{\boldsymbol{\mu}\in\mathbb{R}^m:\boldsymbol{\mu}\equiv(\mu_{0,k}+\varepsilon_k)_{k=1}^m,\\ 0\leq\varepsilon_1\leq\ldots\leq\varepsilon_m\leq 1\wedge\delta_{n,m}/\sqrt{c_3}}} \bar{g}_m(\boldsymbol{\mu})(1\wedge\delta_{n,m}/\sqrt{c_3})^m \frac{1}{m!}\\
&\geq \binom{n}{m-1}^{-1} \cdot (\tau_{m,g}^{\mathrm{iso}})^m(1\wedge\delta_{n,m}/\sqrt{c_3})^m\\
&\geq e^{-m\log(en)-m\log\left((\tau_{m,g}^{\mathrm{iso}})^{-1}\vee 1\right)-m\log\left(\frac{\sqrt{c_3}}{\delta_{n,m}}\vee 1\right)}
\end{aligned}
$$

Here the first inequality in the last line follows from the definition of $\bar{g}_m$ and $\tau_{m,g}^{\mathrm{iso}}$. The claim follows by verifying (3.5) implies that the second and third term in the exponent above are both bounded by $\frac{1}{2\eta}\cdot m\log(en)$ [the third term does not contribute to the condition since $\sqrt{c_3}\delta_{n,m}^{-1}\leq n$ by noting $c_3=1$ in the Gaussian regression setting and definition of η]. □

Proof of Theorem 3.8. The theorem follows by Corollary 3.2, Proposition 2.2 coupled with Lemmas C.4 and C.5. □

We now prove Lemma 3.9. We need the following result.

Lemma C.6. *Let $\boldsymbol{f}_0 := (f_0(x_1),\ldots,f_0(x_n))\in\mathbb{R}^n$, and $\boldsymbol{f}_{0,m} := (f_{0,m}(x_1),\ldots,f_{0,m}(x_n))\in\mathbb{R}^n$ where $f_{0,m}\in\arg\min_{g\in\mathcal{F}_m}\ell_n^2(f_0,g)$. Suppose that $\|\boldsymbol{f}_0\|_2\leq L$, and that there exists some element $f\in\mathcal{F}_m$ such that $\boldsymbol{f}\equiv(f(x_1),\ldots,f(x_n))$ satisfies $\|\boldsymbol{f}\|_2\leq L$. Then $\|\boldsymbol{f}_{0,m}\|_2\leq 3L$.*

Proof of Lemma C.6. It can be seen that

$$\boldsymbol{f}_{0,m}\in\arg\min_{\boldsymbol{\gamma}\in\mathcal{P}_{n,m}}\mathcal{L}_{f_0}(\boldsymbol{\gamma})\equiv\arg\min_{\boldsymbol{\gamma}\in\mathcal{P}_{n,m}}\|\boldsymbol{f}_0-\boldsymbol{\gamma}\|_2,$$

where $\mathcal{P}_{n,m}\equiv\{(f(x_1),\ldots,f(x_n)):f\in\mathcal{F}_m\}$. For any $\boldsymbol{\gamma}\in\mathcal{P}_{n,m}$ such that $\|\boldsymbol{\gamma}\|_2\leq L$, the loss function satisfies $\mathcal{L}_{f_0}(\boldsymbol{\gamma})\leq 2L$ by triangle inequality. If $\|\boldsymbol{f}_{0,m}\|_2>3L$, then

$$\mathcal{L}_{f_0}(\boldsymbol{f}_{0,m})=\|\boldsymbol{f}_0-\boldsymbol{f}_{0,m}\|_2\geq\|\boldsymbol{f}_{0,m}\|_2-\|\boldsymbol{f}_0\|_2>3L-L=2L,$$

contradicting the definition of $\boldsymbol{f}_{0,m}$ as a minimizer of $\mathcal{L}_{f_0}(\cdot)$ over $\mathcal{P}_{m,n}$. This shows the claim. □

Proof of Lemma 3.9. Let $L=\int_0^1 f^2$. Note that $\|\boldsymbol{f}_0\|_2^2\leq 2n\int_0^1 f^2(x)\,\mathrm{d}x=2nL^2$. By Lemma C.6, we see that $\|\boldsymbol{f}_{0,m}\|_2\leq 3\sqrt{2n}L$ which entails that $\|f_{0,m}\|_\infty\leq 3\sqrt{2n}L$. Now the conclusion follows from $g(3\sqrt{2nL}+1)\geq(en)^{-1/(2\eta)}$ while the left side is at least on the order of $n^{-\alpha/2}$ as $n\to\infty$. □

C.3. Proof of Theorem 3.10

Checking the local entropy assumption B requires some additional work. The notion of *pseudo-dimension* will be useful in this regard. Following [Pol90] Section 4, a subset V of $\mathbb{R}^d$ is said to have *pseudo-dimension* t, denoted as $\mathrm{pdim}(V)=t$, if for every $x\in\mathbb{R}^{t+1}$ and indices $I=(i_1,\cdots,i_{t+1})\in\{1,\cdots,n\}^{t+1}$ with $i_\alpha\neq i_\beta$

for all $\alpha \neq \beta$, we can always find a sub-index set $J \subset I$ such that no $v \in V$ satisfies both $v_i > x_i$ for all $i \in J$ and $v_i < x_i$ for all $i \in I \setminus J$.

Lemma C.7. *Let $n \geq 2$. Suppose that $\mathrm{pdim}(\mathcal{P}_{n,m}) \leq D_m$ where $\mathcal{P}_{n,m} := \{\big(f(x_1), \ldots, f(x_n)\big) \in \mathbb{R}^n : f \in \mathcal{F}_m\}$. Then for all $g \in \mathcal{F}_m$,*

$$\log \mathcal{N}\big(c_5\varepsilon, \{f \in \mathcal{F}_m : \ell_n(f,g) \leq 2\varepsilon\}, \ell_n\big) \leq C \cdot D_m \log n$$

for some constant $C > 0$ depending on c_5.

To prove Lemma C.7, we need the following result, cf. Theorem B.2 [Gun12].

Lemma C.8. *Let V be a subset of $\mathbb{R}^n$ with $\sup_{v \in V} \|v\|_\infty \leq B$ and pseudo-dimension at most t. Then, for every $\varepsilon > 0$, we have*

$$\mathcal{N}(\varepsilon, A, \|\cdot\|_2) \leq \left(4 + \frac{2B\sqrt{n}}{\varepsilon}\right)^{\kappa t},$$

holds for some absolute constant $\kappa \geq 1$.

Proof of Lemma C.7. Note that the entropy in question can be bounded by $\log \mathcal{N}\big(c_5\varepsilon\sqrt{n}, \{\mathcal{P}_{n,m} - \boldsymbol{g}\} \cap B_n(0, 2\sqrt{n}\varepsilon), \|\cdot\|_2\big)$. Since translation does not change the pseudo-dimension of a set, $\mathcal{P}_{n,m} - \boldsymbol{g}$ has the same pseudo-dimension with that of $\mathcal{P}_{n,m}$, which is bounded from above by D_m by assumption. Further note that $\{\mathcal{P}_{n,m} - \boldsymbol{g}\} \cap B_n(0, 2\sqrt{n}\varepsilon)$ is uniformly bounded by $2\sqrt{n}\varepsilon$, hence an application of Lemma C.8 yields that the entropy can be further bounded as follows:

$$\log \mathcal{N}\big(c_5\varepsilon, \{f \in \mathcal{F}_m : \ell_n(f,g) \leq 2\varepsilon\}, \ell_n\big) \leq \kappa D_m \log\big(4 + 4n/c_5\big) \leq C \cdot D_m \log n$$

for some constant $C > 0$ depending on c_5 whenever $n \geq 2$. □

The pseudo-dimension of the class of piecewise affine functions $\mathcal{F}_m$ can be well controlled, as the following lemma shows.

Lemma C.9 (Lemma 4.9 in [HW16])**.** $\mathrm{pdim}(\mathcal{P}_{n,m}) \leq 6md \log 3m$.

As an immediate result of Lemmas C.7 and C.9, we can take for $n \geq 2$, $\delta_{n,m}^2 := (C \vee 1/\eta)d \cdot \frac{\log n}{n} \cdot m \log 3m$ for some $C \geq 2/c_7$ depending on c_5, c_7.

Lemma C.10. *Suppose that (3.8) holds and $n \geq d$. Then (P2) in Assumption C holds.*

Proof of Lemma C.10. We write $f_{0,m} \equiv \max_{1 \leq i \leq m}\big(a_i \cdot x + b_i\big)$ throughout the proof. We first claim that for any $a_i^* \in B_d(a_i, \delta_{n,m}/2\sqrt{c_3 d})$ and $b_i^* \in B_1(b_i, \delta_{n,m}/2\sqrt{c_3})$, let $g_m^*(x) := \max_{1 \leq i \leq m}(a_i^* \cdot x + b_i^*)$, then $\ell_\infty(g_m^*, f_{0,m}) \leq \delta_{n,m}/\sqrt{c_3}$. To see this, for any $x \in \mathfrak{X}$, there exists some index $i_x \in \{1, \ldots, m\}$ such that $g_m^*(x) = a_{i_x}^* \cdot x + b_{i_x}^*$. Hence

$$\begin{aligned} g_m^*(x) - f_{0,m}(x) &\leq \big(a_{i_x}^* - a_{i_x}\big) \cdot x + \big(b_{i_x}^* - b_{i_x}\big) \leq \|a_{i_x}^* - a_{i_x}\|_2 \|x\|_2 + |b_{i_x}^* - b_{i_x}| \\ &\leq \frac{\delta_{n,m}}{2\sqrt{c_3 d}} \cdot \sqrt{d} + \frac{\delta_{n,m}}{2\sqrt{c_3}} = \frac{\delta_{n,m}}{\sqrt{c_3}}. \end{aligned}$$

The reverse direction can be shown similarly, whence the claim follows by taking supremum over $x \in \mathfrak{X}$. This entails that

$$\begin{aligned}
&\Pi_{n,m}(\{f \in \mathcal{F}_m : \ell_n^2(f, f_{0,m}) \le \delta_{n,m}^2/c_3\})\\
&\ge \Pi_G(\cap_{i=1}^m \{(a_i^*, b_i^*) : a_i^* \in B_d(a_i, \delta_{n,m}/2\sqrt{c_3 d}), b_i^* \in B_1(b_i, \delta_{n,m}/2\sqrt{c_3})\})\\
&= \prod_{i=1}^m \Pi_{g^{\otimes d}}\big(B_d(a_i, \delta_{n,m}/2\sqrt{c_3 d})\big) \cdot \Pi_g\big(B_1(b_i, \delta_{n,m}/2\sqrt{c_3})\big)\\
&\ge \prod_{i=1}^m g(\|a_i\|_\infty + 1)^d \cdot g(|b_i| + 1) \cdot \bigg(\frac{\delta_{n,m}}{\sqrt{4c_3 d}} \wedge 1\bigg)^d v_d \bigg(\frac{\delta_{n,m}}{\sqrt{4c_3}} \wedge 1\bigg)\\
&\ge \exp\bigg(-2m(d+1)\log\big(\tau_{m,g}^{-1} \vee 1\big) - m(d+1)\log\big(\frac{\sqrt{4c_3 d}}{\delta_{n,m}} \vee 1\big) - \frac{1}{2} md\log d\bigg),
\end{aligned}$$

where $v_d \equiv \mathrm{vol}(B_d(0,1))$ and we used the fact that $v_d \ge (1/\sqrt{d})^d$. Now by requiring that $n \ge d$ and

$$\max\bigg\{2m(d+1)\log\big(\tau_{m,g}^{-1} \vee 1\big), m(d+1)\log\bigg(\frac{\sqrt{4c_3 d}}{\delta_{n,m}} \vee 1\bigg)\bigg\} \le \frac{d}{2\eta}\log n \cdot m\log 3m,$$

the claim follows by verifying (3.8) implies this requirement [since $\sqrt{4c_3 d}\,\delta_{n,m}^{-1} \le \sqrt{n}$, the second term is bounded by $md\log n$. The inequality follows by noting $\eta < 1/4$]. □

Lemma C.11. *For $n \ge 2$, (2.5) is satisfied for $\mathfrak{c} = 1, \gamma = 2, \mathfrak{h}_0 = \infty$.*

Proof. For fixed $n \ge 2$ and $\eta > 0$, write $n\delta_{n,m}^2 = c\log n(m\log 3m)$ throughout the proof, where $c \ge 2/c_7$. Then for any $\alpha \ge c_7/2$ and $h \ge 1$, since $\log(3m') \ge \log(3hm) \ge \log(3m)$ for any $m' \ge hm$, we have

$$\sum_{m' \ge hm} e^{-\alpha n\delta_{n,m'}^2} \le \sum_{m' \ge hm} e^{-\alpha cm'(\log n \cdot \log 3m)} = \frac{e^{-\alpha chm\log n\log 3m}}{1 - e^{-\alpha c\log n\log 3m}} \le 2e^{-\alpha hn\delta_{n,m}^2}.$$

For the second condition of (2.5), note that for $\gamma = 2$, in order to verify $\delta_{n,hm}^2 \le h^2\delta_{n,m}^2$, it suffices to have $hm\log(3hm) \le h^2 m\log(3m)$, equivalently $3hm \le (3m)^h$, and hence $3^{h-1} \ge h$ for all $h \ge 1$ suffices. This is valid and hence completing the proof. □

Proof of Theorem 3.10. This is a direct consequence of Corollary 3.2, Lemma C.10 and C.11, combined with Proposition 2.2. □

C.4. Proof of Theorem 3.13

Lemma C.12. *Let $n \ge 2$, then for any $g \in \mathcal{F}_{(s,m)}$,*

$$\begin{aligned}
&\log\mathcal{N}(c_5\varepsilon, \{f \in \mathcal{F}_{(s,m)} : \ell_n(f,g) \le 2\varepsilon\}, \ell_n)\\
&\le 2\log(6/c_5)\big(s\log(ep) \wedge \mathrm{rank}(X) + m\log(en)\big).
\end{aligned}$$

Proof. The proof borrows notation from the proof of Lemma C.4. Further let $\mathscr{S}_s$ denote all subsets of $\{1,\ldots,p\}$ with cardinality at most s. Then the entropy in the statement of the lemma can be further bounded by

$$\begin{aligned}
&\log\left[\binom{p}{s}\binom{n}{m-1}\max_{S\in\mathscr{S}_s,Q\in\mathscr{Q}_m}\mathcal{N}(c_5\varepsilon,\{f\in\mathcal{F}_{(s,m),(S,Q)}:\ell_n(f,g)\leq 2\varepsilon\},\ell_n)\right]\\
&\leq s\log(ep)+m\log(en)\\
&\qquad\qquad+\max_{S\in\mathscr{S}_s,Q\in\mathscr{Q}_m}\log\mathcal{N}(c_5\sqrt{n}\varepsilon,\{\boldsymbol{\gamma}\in\mathcal{P}_{n,(S,Q)}:\|\boldsymbol{\gamma}-\boldsymbol{g}\|_2\leq 2\sqrt{n}\varepsilon\},\|\cdot\|_2)
\end{aligned}$$

where $\mathcal{P}_{n,(S,Q)}\equiv\{(x_i^\top\beta+u(z_i))_{i=1}^n\in\mathbb{R}^n:\mathrm{supp}(\beta)=S,$
u is constant on the partitions of $Q\}$ is contained in a linear subspace of dimension no more than $s+m$. The entropy can also be bounded by

$$m\log(en)+\max_{Q\in\mathscr{Q}_m}\log\mathcal{N}(c_5\sqrt{n}\varepsilon,\{\boldsymbol{\gamma}\in\mathcal{P}_{n,(\{1,\ldots,p\},Q)}:\|\boldsymbol{\gamma}-\boldsymbol{g}\|_2\leq 2\sqrt{n}\varepsilon\},\|\cdot\|_2),$$

which is contained in a linear subspace of dimension no more than $\mathrm{rank}(X)+m$. Now using similar arguments as in Lemma C.4 proves the claim. □

Hence we can take $\delta^2_{n,(s,m)}\equiv c'\frac{s\log(ep)\wedge\mathrm{rank}(X)+m\log(en)}{n}$ for a large constant $c'>0$.

Lemma C.13. *(2.5) holds with $\mathfrak{c},\gamma$ depending on $\mathfrak{h}_0\in[1,\infty)$ and L.*

Proof. For the first condition of (2.5), note that for any $h\in[1,\mathfrak{h}_0]$ and $\alpha\geq c_7/2$, choose $c'>0$ such that $\alpha c'\geq 2L\vee 2$, it follows that

$$\begin{aligned}
&\sum_{(s',m')\geq(hs,hm)}e^{-\alpha n\delta^2_{n,(s',m')}}=\sum_{s'\geq hs}e^{-\alpha c'(s\log(ep)\wedge\mathrm{rank}(X))}\sum_{m'\geq hm}e^{-\alpha c' m\log(en)}\\
&\leq(1-e^{-\alpha c'})^{-1}e^{-(\alpha c'/2\mathfrak{h}_0)h(s\log(ep)\wedge\mathrm{rank}(X)+m\log(en))}\leq 2e^{-\alpha nh\delta^2_{n,(s,m)}/\mathfrak{c}^2}.
\end{aligned}$$

The inequality in the middle for the previous display follows as

$$\begin{aligned}
&\sum_{s'\geq hs}e^{-\alpha c'(s\log(ep)\wedge\mathrm{rank}(X))}\leq e^{-\alpha c'(hs\log(ep)\wedge\mathrm{rank}(X))+\log p}\\
&\leq e^{-\min\{\alpha c' hs\log(ep)-\log p,\alpha c'\,\mathrm{rank}(X)-\log p\}}\leq e^{-(\alpha c'/2)(hs\log(ep)\wedge\mathrm{rank}(X))}\\
&\leq e^{-(\alpha c'/2\mathfrak{h}_0)h(s\log(ep)\wedge\mathrm{rank}(X))}.
\end{aligned}$$

The second condition of (2.5) is easy to verify. □

Lemma C.14. *Suppose (3.11) holds. Then (P2) in Assumption C holds.*

Proof. Let $\delta^2_{n,s}\equiv c'(s\log(ep)\wedge\mathrm{rank}(X))/n$ and $\delta^2_{n,m}\equiv c' m\log(en)/n$. Let $\tau_{s,g}\equiv\sup_{f_{0,(s,m)}}g(\|\beta_{0,s}\|_\infty+1)$.

First consider $s\log(ep)\leq\mathrm{rank}(X)$. Using notation in Lemma C.12,

$$\Pi_{n,(s,m)}(\{f\in\mathcal{F}_{(s,m)}:\ell^2_n(f,f_{0,(s,m)})\leq\delta^2_{n,(s,m)}/c_3\})$$

$$\geq \binom{p}{s}^{-1}\binom{n}{m-1}^{-1}\Pi_{g^{\otimes s}\otimes\bar{g}_m}(\{f\in\mathcal{F}_{(s,m),(S_0,Q_0)}:\ell_n^2(f,f_{0,(s,m)})\leq\delta_{n,(s,m)}^2/c_3\})$$

where $f_{0,(s,m)}\in\mathcal{F}_{(s,m),(S_0,Q_0)}$. To bound the prior mass of the above display from below, it suffices to bound the product of the following two terms:

$$\begin{aligned}\text{(C.2)}\qquad &\pi_s\equiv\Pi_{g^{\otimes s}}(\{\beta\in B_0(s):\beta_{S_0^c}=0,\ell_n^2(h_\beta,h_{\beta_{0,s}})\leq\delta_{n,s}^2/2c_3\}),\\ &\pi_m\equiv\Pi_{\bar{g}_m}(\{u\in\mathcal{U}_{m,Q_0}:\ell_n^2(u,u_{0,m})\leq\delta_{n,m}^2/2c_3\}).\end{aligned}$$

The first term equals

$$\begin{aligned}&\Pi_{g^{\otimes s}}(\{\beta\in B_0(s):\beta_{S_0^c}=0,\|X\beta-X\beta_{0,s}\|_2\leq\sqrt{n}\delta_{n,s}/\sqrt{2c_3}\})\\ &\geq\Pi_{g^{\otimes s}}\left(\left\{\beta\in B_0(s):\beta_{S_0^c}=0,\|\beta-\beta_{0,s}\|_2\leq\frac{1}{\sigma_\Sigma}\cdot\frac{\delta_{n,s}}{\sqrt{2c_3}}\right\}\right).\end{aligned}$$

Here the inequality follows by noting $\|X\beta-X\beta_{0,s}\|_2^2\leq n(\beta-\beta_{0,s})^\top\Sigma(\beta-\beta_{0,s})\leq n\sigma_\Sigma^2\|\beta-\beta_{0,s}\|_2^2$, where σ_Σ denotes the largest singular value of $X^\top X/n$. Note that $\sigma_\Sigma\leq\sqrt{p}$ since the trace for $X^\top X/n$ is p and the trace of a p.s.d. matrix dominates the largest eigenvalue. The set above is supported on $\mathbb{R}^p_{S_0}$ and hence can be further bounded from below by $\tau_{s,g}^s\big(\frac{1}{\sigma_\Sigma}\cdot\frac{\delta_{n,s}}{\sqrt{2c_3}}\wedge 1\big)^s v_s$ where $v_s=\mathrm{vol}(B_s(0,1))$. Hence

$$\pi_s\geq(\tau_{s,g}\wedge 1)^s\left(\frac{1}{\sigma_\Sigma}\cdot\frac{\delta_{n,s}}{\sqrt{2c_3}}\wedge 1\right)^s v_s\geq e^{-\frac{1}{2}s\log s-s\log\left(\tau_{s,g}^{-1}\vee 1\right)-\frac{s}{2}\log\left(\frac{2c_3\sigma_\Sigma^2}{\delta_{n,s}^2}\vee 1\right)},$$

where in the last inequality we used that $v_s\geq(1/\sqrt{s})^s$. By repeating the arguments in the proof of Lemma C.5, we have

$$\pi_m\geq e^{-m\log\left(\tau_{m,g}^{-1}\vee 1\right)-\frac{m}{2}\log\left(\frac{2c_3}{\delta_{n,m}^2}\vee 1\right)}.$$

Combining above estimates,

$$\begin{aligned}&\Pi_{n,(s,m)}(\{f\in\mathcal{F}_{(s,m)}:\ell_n^2(f,f_{0,(s,m)})\leq\delta_{n,(s,m)}^2/c_3\})\\ &\geq e^{-2s\log(ep)-m\log(en)-s\log\left(\tau_{s,g}^{-1}\vee 1\right)-m\log\left(\tau_{m,g}^{-1}\vee 1\right)}\\ &\quad\times e^{-\frac{s}{2}\log\left(\frac{2c_3\sigma_\Sigma^2}{\delta_{n,s}^2}\vee 1\right)-\frac{m}{2}\log\left(\frac{2c_3}{\delta_{n,m}^2}\vee 1\right)}.\end{aligned}$$

The right side is bounded from below by $e^{-2n\delta_{n,(s,m)}^2}$, if we require both

$$\begin{aligned}&\min\left\{e^{-s\log(\tau_{s,g}^{-1}\vee 1)},e^{-s\log\left(\frac{\sqrt{2c_3}\sigma_\Sigma}{\delta_{n,s}}\vee 1\right)}\right\}\geq e^{-\frac{1}{2\eta}s\log(ep)},\\ &\min\left\{e^{-m\log(\tau_{m,g}^{-1}\vee 1)},e^{-m\log\left(\frac{\sqrt{2c_3}}{\delta_{n,m}}\vee 1\right)}\right\}\geq e^{-\frac{1}{2\eta}m\log(en)}.\end{aligned}$$

The first terms in the above two lines can be verified by (3.11). The other terms in the above two lines do not contribute by noting that $2c_3/\delta_{n,m}^2\leq\frac{2c_3c_7}{4\log(6/c_5)}n\leq$

$(1/2)n \leq en$ since $c_3 = 1$ (in Gaussian regression model) and $c_7 \in (0,1)$, while $2c_3\sigma_\Sigma^2/\delta_{n,s}^2 \leq \sigma_\Sigma^2 n \leq pn \leq p^2$ and $\eta < 1/4$.

Next for $s\log(ep) > \mathrm{rank}(X)$, we may proceed with

$$\begin{aligned}&\Pi_{n,(s,m)}(\{f \in \mathcal{F}_{(s,m)} : \ell_n^2(f, f_{0,(s,m)}) \leq \delta_{n,(s,m)}^2/c_3\})\\ &\geq \binom{n}{m-1}^{-1} \Pi\Big(\{f \in \cup_{|S|=s}\mathcal{F}_{(s,m),(S,Q_0)} : \ell_n^2(f, f_{0,(s,m)}) \leq \delta_{n,(s,m)}^2/c_3\}\Big).\end{aligned}$$

To bound the prior mass of the above display from below, it suffices to bound from below the product of π_m and

$$\tilde{\pi}_s \equiv \Pi\Big(\{\beta \in B_0(s) : \|X(\beta - \beta_{0,s})\| \leq \sqrt{n}\delta_{n,s}/\sqrt{2c_3}\}\Big). \tag{C.3}$$

Let $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{p\times p}$ give rise to the SVD of X: $X = U\Lambda V \equiv U\mathrm{diag}(\sigma_1, \ldots, \sigma_{\mathrm{rank}(X)}, 0)V$ where $\sigma_1 \geq \ldots \geq \sigma_{\mathrm{rank}(X)} > 0$ are non-trivial singular values of X. It follows by writing $V = (v_1^\top \cdots v_p^\top)^\top$ that

$$\begin{aligned}\tilde{\pi}_s &\geq \Pi\big(\beta : \|\Lambda V(\beta - \beta_{0,s})\| \leq \sqrt{n}\delta_{n,s}/\sqrt{2c_3}\big)\\ &= \sum_{|S|=s} \binom{p}{s}^{-1} \Pi\Bigg(\beta : \beta_{S^c} = 0, \sum_{j=1}^{\mathrm{rank}(X)} \sigma_j^2 (v_j^\top(\beta - \beta_{0,s}))^2 \leq n\delta_{n,s}^2/2c_3\Bigg)\\ &\geq \sum_{|S|=s} \binom{p}{s}^{-1} \Pi\big(\beta : \beta_{S^c} = 0, \|\beta - \beta_{0,s}\|_2^2 \leq c'/(2c_3\sigma_1^2)\big).\end{aligned}$$

By choosing $c' > 2c_3\sigma_1^2(\|\beta_{0,s}\|_\infty + 1)^2$, the RHS of the previous display can be bounded from below by $g(1)$, as desired. π_m can be handled similarly as in the case $s\log(ep) \leq \mathrm{rank}(X)$. □

Proof of Theorem 3.13. The claim of the theorem follows by Corollary 3.2, Proposition 2.2 and Lemmas C.12-C.14. □

C.5. Proof of Theorem 3.23

Lemma C.15. *For any $\Sigma_0 \in \mathfrak{M}_{(k,s)}$, the following entropy estimate holds:*

$$\begin{aligned}&\log \mathcal{N}\left(c_5\varepsilon, \{\Sigma \in \mathfrak{M}_{(k,s)} : \|\Sigma - \Sigma_0\|_F \leq C_L\varepsilon\}, \|\cdot\|_F\right)\\ &\leq ks\log(ep/s) + ks\log(6\sqrt{kL}/c_5\varepsilon).\end{aligned}$$

Proof. The set involved in the entropy is equivalent to

$$\left\{\Lambda \in \mathscr{R}_{(k,s)}(L) : \|\Lambda\Lambda^\top - \Lambda_0\Lambda_0^\top\|_F \leq C_L\varepsilon, \|\cdot\|_F\right\}. \tag{C.4}$$

We claim that $\sup_{\Lambda \in \mathscr{R}_{(k,s)}} \|\Lambda\Lambda^\top\|_F \leq \sqrt{kL}$. To see this, let $\Lambda \equiv P\Xi Q^\top$ be the singular value decomposition of Λ, where $P \in \mathbb{R}^{p\times p}, Q \in \mathbb{R}^{k\times k}$ are unitary matrices and $\Xi \in \mathbb{R}^{p\times k}$ is a diagonal matrix. Then $\|\Lambda\Lambda^\top\|_F^2 = \|\Xi\Xi^\top\|_F^2 \leq kL$, proving the claim. Combined with (C.4) and Euclidean embedding, we see that the entropy in question can be bounded as follows:

$$\begin{aligned}&\log\mathscr{N}\Big(c_5\varepsilon,\{v\in B_0(ks;pk):\|v\|_2\leq 2\sqrt{kL}\},\|\cdot\|_2\Big)\\&\leq\log\left[\binom{pk}{ks}\left(\frac{6\sqrt{kL}}{c_5\varepsilon}\right)^{ks}\right]\leq ks\log(ep/s)+ks\log(6\sqrt{kL}/c_5\varepsilon),\end{aligned}$$

where $B_0(s;pk)\equiv\{v\in\mathbb{R}^{pk}:|\mathrm{supp}(v)|\leq s\}$. □

Proof of Theorem 3.23. Take $\delta^2_{n,(k,s)}=KC'ks\log(C'p)/n$ for some $C'\geq e$ depending on c_5,c_7,L and some absolute constant $K\geq 1$. Apparently (2.5) holds with $\mathfrak{c}=1,\gamma=1,\mathfrak{h}_0=\infty$. The prior $\Pi_{n,(k,s)}$ on $\mathfrak{M}_{(k,s)}$ will be the uniform distribution on a minimal $\sqrt{C'ks\log(C'p)/c_3n}$ covering-ball of the set $\{\Sigma\in\mathfrak{M}_{(k,s)}\}$ under the Frobenius norm $\|\cdot\|_F$. The above lemma entails that the cardinality for such a cover is no more than $e^{C''ks\log(C''p)}$ for another constant $C''\geq e$ depending on c_3,c_5,c_7,L. Hence we have that

$$\Pi_{n,(k,s)}(\{\Sigma\in\mathfrak{M}_{(k,s)}:\|\Sigma-\Sigma_{0,(k,s)}\|_F\leq\delta^2_{n,(k,s)}/c_3\})\geq e^{-C''ks\log(C''p)},$$

which can be bounded from below by $e^{-2n\delta^2_{n,(k,s)}}$ by choosing K large enough. The claim of Theorem 3.23 now follows from these considerations along with Corollary 3.22, Proposition 2.2. □

C.6. Proof of Theorem 3.27

Lemma C.16. *For $\theta_0\in\Theta_m$, we have $\log\mathscr{N}\big(c_5\varepsilon,\{\theta\in\Theta_m,d_n(\theta,\theta_0)\leq 2\varepsilon\},d_n\big)\leq 4m\log\big(\frac{C_\eta m}{c_5^4\varepsilon^4}\big)$.*

Proof. We first claim that for $\varepsilon\leq 1$,

$$\log\mathscr{N}\big(\varepsilon,\{\Gamma\in\mathscr{C}_m\},\lambda(\cdot\Delta\cdot)\big)\leq m\log\left(\frac{9em}{\varepsilon^2}\right).$$

To see this, fix $\delta>0$ to be chosen later, and partition $[0,1]^2$ into small squares with side length δ. Let $\mathscr{D}_\delta$ be the set of all polytopes in $[0,1]^2$ with its at most m vertices all located on the grid points of these small squares. Apparently $|\Delta_\delta|\leq\binom{(1+1/\delta)^2}{m}$. Then for each $\Gamma\in\mathscr{C}_m$, let $\Gamma_\delta\in\mathscr{D}_\delta$ be such that $\Gamma_\delta\supset\Gamma$ and that for every vertex v of Γ, there exists a vertex v_δ of Γ_δ so that both v and v_δ are in the same small square, with distance at most $\sqrt{2}\delta$. Then the points on the boundary of Γ_δ is within distance $\sqrt{2}\delta$ to Γ, and therefore $\lambda(\Gamma_\delta\Delta\Gamma)\leq\sqrt{2}(\sqrt{2}\delta)m=2\delta m$ (the estimate can be done in a conservative way by collapsing the set of vertices in Γ that corresponding to the same vertex in Γ_δ into one vertex). Now let $\varepsilon=2\delta m$ yields the claim.

Since

$$d_n^2(\theta_0,\theta_1)\leq C_1^2\big(|\xi_0-\xi_1|^2+|\rho_0-\rho_1|^2+\lambda(\Gamma_0\Delta\Gamma_1)\big)$$

for some constant $C_1^2 > 0$ depending only through η, it follows that

$$\begin{aligned}
&\log\mathcal{N}\big(c_5\varepsilon, \{\theta \in \Theta_m, d_n(\theta,\theta_0) \le 2\varepsilon\}, d_n\big) \le 2\log\mathcal{N}\big(c_5\varepsilon/(\sqrt{3}C_1), [\eta, 1-\eta], |\cdot|\big)\\
&\qquad\qquad + \log\mathcal{N}\big(c_5^2\varepsilon^2/(3C_1^2), \{\Gamma \in \mathscr{C}_m, \Gamma \subset [\eta, 1-\eta]^m\}, \lambda(\cdot\Delta\cdot)\big)\\
&\le 2\log\left(\frac{\sqrt{3}C_1}{c_5\varepsilon}\right) + m\log\left(\frac{81C_1^4 em}{c_5^4\varepsilon^4}\right) \le 4m\log\left(\frac{C_\eta m}{c_5^4\varepsilon^4}\right),
\end{aligned}$$

as desired. □

Now we take $\delta_{n,m}^2 \equiv C_\eta' \frac{m\log n}{n}$ for some large constant $C_\eta' > 0$.

Lemma C.17. *For $\theta_0 \in \Theta_m$, (P2) is satisfied for n large enough depending on θ_0.*

Proof. Let $\{v_i(\Gamma)\}_{i=1}^m$ be the vertices of $\Gamma \in \mathscr{C}_m$. Using again

$$d_n^2(\theta_0, \theta) \le C_1^2\big(|\xi_0 - \xi|^2 + |\rho_0 - \rho|^2 + \lambda(\Gamma_0\Delta\Gamma)\big),$$

and that for n large enough depending on Γ_0, for any $v_i \notin \Gamma_0$ such that $\|v_i - v_i(\Gamma_0)\|_2 \le \delta_{n,m}^2/(3\sqrt{2}mC_1^2c_3)$, $\Gamma \equiv \mathrm{conv}(\{v_i\})$ has vertices exactly given by $\{v_i\}$, and $\lambda(\Gamma\Delta\Gamma_0) \le \sqrt{2}\cdot\big(\delta_{n,m}^2/(3\sqrt{2}mC_1^2c_3)\big)m = \delta_{n,m}^2/(3C_1^2c_3)$, we have

$$\begin{aligned}
&\Pi_{n,m}\big(\{\theta \in \Theta_m : d_n^2(\theta,\theta_0) \le \delta_{n,m}^2/c_3\}\big)\\
&\ge \Pi_\xi\big(|\xi - \xi_0|^2 \le \delta_{n,m}^2/(3C_1^2c_3)\big)\cdot\Pi_\rho\big(|\rho - \rho_0|^2 \le \delta_{n,m}^2/(3C_1^2c_3)\big)\\
&\qquad\times\Pi_\Gamma\big(\|v_i(\Gamma) - v_i(\Gamma_0)\|_2 \le \delta_{n,m}^2/(3\sqrt{2}mC_1^2c_3, v_i(\Gamma) \notin \Gamma_0)\big)\\
&\gtrsim_\eta \delta_{n,m}^2\big(\delta_{n,m}^2/m^{3/2}\big)^m \ge \exp(-2n\delta_{n,m}^2),
\end{aligned}$$

as long as $C_\eta' > 0$ is large enough. □

Proof of Theorem 3.27. The claim follows by Corollary 3.26, Proposition 2.2 coupled with Lemmas C.16 and C.17. □

C.7. Proof of Theorem 3.30

Lemma C.18. *For any $g \in \mathcal{F}_m$ such that $g \le f_0$, and any $R \ge \|f_0\|_\infty \vee 1$,*

$$\log\mathcal{N}_{[}\big(c_5\varepsilon^2, \{f \in \mathcal{F}_m, R \ge f \ge f_0 : \bar{L}_1(f,g) \le 4\varepsilon^2\}, \bar{L}_1\big) \le 2m\log\left(\frac{8emR^2}{c_5\varepsilon^2}\right).$$

Proof of Lemma C.18. Note that the local entropy with left bracketing in question can be bounded by its global counterpart $\mathcal{N}_{[}\big(c_5\varepsilon^2, \{f \in \mathcal{F}_m, |f| \le R\}, \bar{L}_1\big)$.

Let $m \ge 2$. Fix $\varepsilon > 0$, let $\delta^2 = c_5\varepsilon^2/(2Rm+1)$. Without loss of generality, we assume that $1/\delta^2 \in \mathbb{N}$, and we partition the interval $[0,1)$ into $\cup_{j=1}^{1/\delta^2} I_{j,\delta} \equiv \cup_{j=1}^{1/\delta^2}[(j-1)\delta^2, j\delta^2)$. For any $f \in \mathcal{F}_m$, let $f \equiv \sum_{j=1}^m a_j\mathbf{1}_{[t_{j-1},t_j)}$ for some $0 = t_0 < t_1 < \ldots < t_{m-1} < t_m = 1$. Then $\{t_1, \ldots, t_{m-1}\}$ must be contained in

$m-1$ intervals amongst $\{I_{j,\delta}\}_{j=1}^{1/\delta^2}$, namely, $\{\bar{I}_{k,\delta;f}\}_{k=1}^{m-1}$. Furthermore, $[0,1]\setminus\cup_{k=1}^{m-1}\bar{I}_{k,\delta;f}$ contains at most m intervals. Now define $\bar{f}$ as follows:

$$\bar{f}\equiv\sum_{k=1}^{m-1}(-R)\cdot\mathbf{1}_{\bar{I}_{k,\delta;f}}+\left\lfloor\frac{f}{\delta^2}\right\rfloor\delta^2\cdot\mathbf{1}_{[0,1]\setminus\cup_{k=1}^{m-1}\bar{I}_{k,\delta;f}}.$$

Clearly $\bar{f}\le f$, and

$$\int_0^1\big(f(x)-\bar{f}(x)\big)\,\mathrm{d}x\le 2Rm\delta^2+\int_{[0,1]\setminus\cup_{k=1}^{m-1}\bar{I}_{k,\delta;f}}\delta^2\,\mathrm{d}x\le c_5\varepsilon^2.$$

On the other hand, there are at most $\binom{1/\delta^2}{m-1}\cdot\big(\frac{2R}{\delta^2}\big)^m$ many choices of $\bar{f}$, and hence

$$\begin{aligned}&\log\mathcal{N}_{[}\big(c_5\varepsilon^2,\{f\in\mathcal{F}_m,|f|\le R\},\bar{L}_1\big)\le\log\left[\binom{1/\delta^2}{m-1}\cdot\big(\frac{2R}{\delta^2}\big)^m\right]\\&\le(m-1)\log\left(\frac{e(2Rm+1)}{c_5\varepsilon^2(m-1)}\right)+m\log\left(\frac{2R(2Rm+1)}{c_5\varepsilon^2}\right)\le 2m\log\left(\frac{8emR^2}{c_5\varepsilon^2}\right).\end{aligned}$$

For $m=1$, it is clear the above bound holds so the proof is complete. □

Hence we can take $\delta_{n,m}^2\equiv\big(\frac{4}{c_7}\vee 2\big)\frac{m}{n}\log\big(8enR^2/c_5\big)$. Clearly (2.5) is satisfied with $\mathfrak{c}=\gamma=1,\mathfrak{h}_0=\infty$.

Lemma C.19. *Suppose that g_a has full support. For n large enough depending on f_0 and the prior g_a, (P2) in Assumption C restricted to $\{f\ge f_{0,m},\|f\|_\infty\le\|f_0\|_\infty+1\}$ holds.*

Proof. Let $f_{0,m}\equiv\sum_{j=1}^m a_j^*\mathbf{1}_{[t_{j-1}^*,t_j^*)}$ for some $t^*=(t_1^*,\ldots,t_{m-1}^*)$ with $0=t_0^*<t_1^*<\ldots<t_{m-1}^*<t_m^*=1$. Without loss of generality, we may assume that $\min\{t_j^*-t_{j-1}^*:j\}>1/(2n\|f_0\|_\infty)$ (otherwise we may merge such short intervals to construct a surrogate $\tilde{f}_{0,m}$, and the total difference between $\tilde{f}_{0,m}$ and $f_{0,m}$ in L_1 metric by doing this does not exceed m/n so that there is no effect in the final oracle inequality). Let $u_j^*\equiv 2\cdot\mathbf{1}_{a_{j+1}^*<a_j^*}-1$. For any $t=(t_1,\ldots,t_{m-1})$ such that $t_j=t_j^*+u_j^*\delta$ with $\delta<1/(4n\|f_0\|_\infty+1)$, and any $a=(a_1,\ldots,a_m)$ such that $a_j\ge a_j^*$ and $\max_j|a_j-a_j^*|\le 1/(4n)$, let $f\equiv\sum_{j=1}^m a_j\mathbf{1}_{[t_{j-1},t_j)}\ge f_{0,m}$. Then

$$\begin{aligned}&\int_0^1\big(f(x)-f_{0,m}(x)\big)\,\mathrm{d}x\\&\le\frac{1}{4n}+m\cdot\frac{1}{4n\|f_0\|_\infty+1}\cdot\left(\|f_0\|_\infty+\frac{1}{4n}\right)\le\frac{m}{n}\le\delta_{n,m}^2/c_3\end{aligned}$$

by the definition of $\delta_{n,m}^2$ and the fact that $c_3=1$. This implies that with $\tau_{g_a}^{\mathrm{int}}\equiv g_a(\|f_0\|_\infty+1)$ (it is easy to see $\|f_{0,m}\|_\infty\le\|f_0\|_\infty$),

$$\Pi_{n,m}\big(\{f\in\mathcal{F}_m:f\ge f_{0,m},\bar{L}_1(f,f_{0,m})\le\delta_{n,m}^2/c_3\}\big)$$

$$\geq (4n\|f_0\|_\infty + 1)^{-(m-1)} \bigg(1 \wedge \frac{\tau_{g_a}^{\mathrm{int}}}{4n}\bigg)^m \geq e^{-m\big(\log(4n\|f_0\|_\infty+1)+\log\big(1\vee \frac{4n}{\tau_{g_a}^{\mathrm{int}}}\big)\big)}$$

Since $2n\delta_{n,m}^2 \geq 4m\log(32en)$, it suffices to require that $\log(4n\|f_0\|_\infty + 1) \vee \log\big(1 \vee \frac{4n}{\tau_{g_a}^{\mathrm{int}}}\big) \leq 2\log(32en)$, which is satisfied for n large. □

Proof of Theorem 3.30. Let $R_n \to \infty$ be a sequence such that $\log R_n \lesssim \log n$. We omit the superscript in the constants in the proof. Let $\bar{\mathcal{F}}_n \equiv \{f : [0,1] \to \mathbb{R} : |f| \leq R_n, f \in \mathcal{F}_m\}$ be the localized subset of $\mathcal{F}$. By the decomposition (2.12), the probability in question can be bounded by

(C.5)
$$P_{f_0}^{(n)}\bar{\Pi}_n\big(f \geq f_0, f \in \bar{\mathcal{F}}_n : \bar{L}_1(f, f_0) > C_2\big(\inf_{g\in\mathcal{F}_m} \bar{L}_1(f_0, g) + \frac{m\log(R_n^2 n)}{n}\big)\big|N\big) \\ + P_{f_0}^{(n)}\Pi_n\big(f \notin \bar{\mathcal{F}}_n\big|N\big).$$

We first handle the first term in (C.5). Now Corollary 3.28 combined with Lemma C.18 and C.19 yields that for n large enough

$$P_{f_0}^{(n)}\bar{\Pi}_n\bigg(f \geq f_0, f \in \bar{\mathcal{F}}_n : \bar{L}_1(f, f_0) > C_2\big(\inf_{g\in\mathcal{F}_m\cap\bar{\mathcal{F}}_n} \bar{L}_1(f_0, g) + m\log(R_n^2 n)/n\big)\big|N\bigg) \\ \leq C_3 e^{-n\varepsilon_{n,m}^2/C_3},$$

where $\varepsilon_{n,m}^2 \equiv \max\{\inf_{g\in\mathcal{F}_m\cap\bar{\mathcal{F}}_n} \bar{L}_1(f_0, g), m\log(R_n^2 n)/n\}$. Here $C_2, C_3 > 0$ are absolute constants that do not depend on R_n. Note that in applying (modified) Lemma C.19 we (implicitly) used the fact that the induced localized prior mass satisfies the following:

$$\bar{\Pi}_{n,m}\big(\{f \in \mathcal{F}_m \cap \bar{\mathcal{F}}_n : f \geq f_{0,m}, \bar{L}_1(f, f_{0,m}) \leq \delta_{n,m}^2/c_3\}\big) \\ \geq \Pi_{n,m}\big(\{f \in \mathcal{F}_m \cap \bar{\mathcal{F}}_n : f \geq f_{0,m}, \bar{L}_1(f, f_{0,m}) \leq \delta_{n,m}^2/c_3\}\big).$$

Next we handle the second term in (C.5). Applying Lemma A.5 to the localized model with

$$\bar{\varepsilon}_{n,m}^2 \equiv \inf_{g\in\mathcal{F}_m\cap\bar{\mathcal{F}}_n} \bar{L}_1(f_0, g) + C_4 m\log(en)/n = \bar{L}_1(f_0, f_{0,m}) + C_4 m\log(en)/n$$

for $C_4 > 0$ large enough and n large enough, we see that on an event $\mathcal{E}_n$ with $P_{f_0}^{(n)}$ probability at least $1 - e^{-C_5 n\bar{\varepsilon}_{n,m}^2}$, it holds that

$$\int_{\bar{\mathcal{F}}_n} p_f^{(n)}/p_{f_0}^{(n)}\, \mathrm{d}\bar{\Pi}_n(f) \geq \lambda_n(m) \int_{f\in\mathcal{F}_m\cap\bar{\mathcal{F}}_n : \bar{L}_1(f,f_0)\leq\bar{\varepsilon}_{n,m}^2} p_f^{(n)}/p_{f_0}^{(n)}\, \mathrm{d}\bar{\Pi}_{n,m}(f) \\ \gtrsim e^{-C_6 m\log(en)} \times \bar{\Pi}_{n,m}\big(\{f \in \mathcal{F}_m \cap \bar{\mathcal{F}}_n, f \geq f_{0,m} : \bar{L}_1(f, f_{0,m}) \leq C_4\frac{m}{n}\log(en)\}\big) \\ \gtrsim e^{-C_7 m\log(en)}\big(\Pi_n(\bar{\mathcal{F}}_n)\big)^{-1}$$

where the last inequality holds for n large enough, and follows essentially from the same argument used in the proof of Lemma C.19. Now we have

$$\begin{aligned}
&P_{f_0}^{(n)}\Pi_n\big(f\notin\bar{\mathcal{F}}_n|N\big)\leq P_{f_0}^{(n)}\Pi_n\big(f\notin\bar{\mathcal{F}}_n|N\big)\mathbf{1}_{\mathcal{E}_n}+P_{f_0}^{(n)}(\mathcal{E}_n^c)\\
&\leq P_{f_0^{(n)}}\big[\frac{\int_{f\notin\bar{\mathcal{F}}_n}p_f^{(n)}/p_{f_0}^{(n)}\ \mathrm{d}\Pi_n(f)}{\int_{\bar{\mathcal{F}}_n}p_f^{(n)}/p_{f_0}^{(n)}\ \mathrm{d}\Pi_n(f)}\mathbf{1}_{\mathcal{E}_n}\big]+P_{f_0}^{(n)}(\mathcal{E}_n^c)\\
&\leq\frac{1}{\Pi_n(\bar{\mathcal{F}}_n)}\cdot P_{f_0^{(n)}}\big[\frac{\int_{f\notin\bar{\mathcal{F}}_n}p_f^{(n)}/p_{f_0}^{(n)}\ \mathrm{d}\Pi_n(f)}{\int_{\bar{\mathcal{F}}_n}p_f^{(n)}/p_{f_0}^{(n)}\ \mathrm{d}\bar{\Pi}_n(f)}\mathbf{1}_{\mathcal{E}_n}\big]+P_{f_0}^{(n)}(\mathcal{E}_n^c)\\
&\lesssim e^{C_7 m\log(en)}\cdot\Pi_n(\mathcal{F}\setminus\bar{\mathcal{F}}_n)+e^{-C_5 n\bar{\varepsilon}_{n,m}^2}
\end{aligned}$$

Furthermore we have,

$$\begin{aligned}
\Pi_n(\mathcal{F}\setminus\bar{\mathcal{F}}_n)&\leq\sum_{k>m}\lambda_n(k)\Big(\int_{|x|>R_n}g(x)\ \mathrm{d}x\Big)^k\\
&\lesssim\sum_{k>m}e^{-C_6(k-1)\log(en)-k\log(\int_{|x|>R_n}g(x)\ \mathrm{d}x)^{-1}}\lesssim e^{-2C_7 m\log(en)},
\end{aligned}$$

where the last inequality follows as $\log(\int_{|x|>R_n}g(x)\ \mathrm{d}x)^{-1}\geq C'\log(en)$ holds for a large enough constant $C'>0$. Combining the above estimates concludes the proof. □

Appendix D: Proofs of auxiliary lemmas in Appendix A

Proof of Lemma A.4. Without loss of generality we assume $d_0=0$. Let $\mathcal{F}_j:=\{f\in\mathcal{F}:j\varepsilon<d_n(f,f_0)\leq 2j\varepsilon\}$ and $\mathcal{G}_j\subset\mathcal{F}_j$ be the collection of functions that form a minimal $c_5 j\varepsilon$ covering set of $\mathcal{F}_j$ under the metric d_n. Then by assumption $|\mathcal{G}_j|\leq N(j\varepsilon)$. Furthermore, for each $g\in\mathcal{G}_j$, it follows by Lemma 2.1 that there exists some test $\omega_{n,j,g}$ such that

$$\sup_{f\in\mathcal{F}:d_n(f,g)\leq c_5 d_n(g,f_0)}\big[P_{f_0}^{(n)}\omega_{n,j,g}+P_f^{(n)}(1-\omega_{n,j,g})\big]\leq c_6 e^{-c_7 n d_n^2(g,f_0)}.$$

Recall that $g\in\mathcal{G}_j\subset\mathcal{F}_j$, then $d_n(g,f_0)>j\varepsilon$. Hence the indexing set above contains $\{f\in\mathcal{F}:d_n(f,g)\leq c_5 j\varepsilon\}$. Now we see that

$$P_{f_0}^{(n)}\omega_{n,j,g}\leq c_6 e^{-c_7 n j^2\varepsilon^2},\qquad\sup_{f\in\mathcal{F}:d_n(f,g)\leq c_5 j\varepsilon}P_f^{(n)}(1-\omega_{n,j,g})\leq c_6 e^{-c_7 n j^2\varepsilon^2}.$$

Consider the global test $\phi_n:=\sup_{j\geq1}\max_{g\in\mathcal{G}_j}\omega_{n,j,g}$, then

$$\begin{aligned}
P_{f_0}^{(n)}\phi_n&\leq P_{f_0}^{(n)}\sum_{j\geq1}\sum_{g\in\mathcal{G}_j}\omega_{n,j,g}\leq c_6\sum_{j\geq1}N(j\varepsilon)e^{-c_7 n j^2\varepsilon^2}\\
&\leq c_6 N(\varepsilon)\sum_{j\geq1}e^{-c_7 n j^2\varepsilon^2}\leq c_6 N(\varepsilon)e^{-c_7 n\varepsilon^2}\cdot\big(1-e^{-c_7 n\varepsilon^2}\big)^{-1}.
\end{aligned}$$

On the other hand, for any $f \in \mathcal{F}$ such that $d_n(f, f_0) \geq \varepsilon$, there exists some $j^* \geq 1$ and some $g_{j^*} \in \mathcal{G}_{j^*}$ such that $d_n(f, g_{j^*}) \leq j^* c_5 \varepsilon$. Hence

$$P_f^{(n)}(1-\phi_n) \leq P_f^{(n)}(1-\omega_{n,j^*,g_{j^*}}) \leq c_6 e^{-c_7 n (j^*)^2 \varepsilon^2} \leq c_6 e^{-c_7 n \varepsilon^2}.$$

The right hand side is independent of individual $f \in \mathcal{F}$ such that $d_n(f, f_0) \geq \varepsilon$ and hence the claim follows. □

Proof of Lemma A.5. WLOG we assume $d_0 = 0$. By Jensen's inequality, the probability in question is bounded by

$$\begin{aligned}
&P_{f_0}^{(n)}\bigg\{ \int \big(\log(p_{f_0}^{(n)}/p_f^{(n)}) - P_{f_0}^{(n)} \log(p_{f_0}^{(n)}/p_f^{(n)})\big)\ \mathrm{d}\Pi(f) \\
&\qquad\qquad\qquad \geq (C+c_3)n\varepsilon^2 - c_3 n \int d_n^2(f_0, f)\ \mathrm{d}\Pi(f)\bigg\} \\
&\leq P_{f_0}^{(n)}\bigg[\int \big(\log(p_{f_0}^{(n)}/p_f^{(n)}) - P_{f_0}^{(n)} \log(p_{f_0}^{(n)}/p_f^{(n)})\big)\ \mathrm{d}\Pi(f) \geq Cn\varepsilon^2 \bigg] \\
&\leq e^{-C\lambda n\varepsilon^2} \cdot c_1 P_{f_0}^{(n)} e^{\lambda \int \big(\log(p_{f_0}^{(n)}/p_f^{(n)}) - P_{f_0}^{(n)} \log(p_{f_0}^{(n)}/p_f^{(n)})\big)\ \mathrm{d}\Pi(f)} \\
&\leq P_{f_0}^{(n)} \int e^{\lambda\big(\log(p_{f_0}^{(n)}/p_f^{(n)}) - P_{f_0}^{(n)} \log(p_{f_0}^{(n)}/p_f^{(n)})\big)}\ \mathrm{d}\Pi(f) \leq \int e^{\psi_{\kappa_g n d_n^2(f_0,f),\kappa_\Gamma}(\lambda)} \mathrm{d}\Pi(f),
\end{aligned}$$

where the last inequality follows from Fubini's theorem and Assumption A. Now the condition on the prior Π entails that

$$P_{f_0}^{(n)}\bigg(\int (p_f^{(n)}/p_{f_0}^{(n)})\ \mathrm{d}\Pi(f) \leq e^{-(C+c_3)n\varepsilon^2} \bigg) \leq c_1 e^{-C\lambda n\varepsilon^2 + \psi_{\kappa_g n\varepsilon^2, \kappa_\Gamma}(\lambda)}.$$

The claim follows by choosing $\lambda > 0$ small enough depending on C, κ. □

Proof of Proposition A.3. By definition we have $\delta_{n,\tilde{m}} \geq d_n(f_0, f_{0,m})$ and $\delta_{n,\tilde{m}-1} < d_n(f_0, f_{0,m})$. In this case, the global test can be constructed via

$$\tilde{\phi}_n := \sup_{m' \in \mathcal{I}, m' \geq j\mathfrak{h}\tilde{m}} \phi_{n,m'}.$$

Then analogous to (A.6) and (A.7), for any random variable $U \in [0,1]$, we have exponential testability:

$$\begin{aligned}
P_{f_{0,m}}^{(n)} U \cdot \tilde{\phi}_n &\leq 4c_6 e^{-(c_7/2\mathfrak{c}^2) nj\mathfrak{h}\delta_{n,\tilde{m}}^2}, \\
\sup_{f \in \mathcal{F}_{j\mathfrak{h}\tilde{m}} : d_n^2(f, f_{0,m}) \geq \mathfrak{c}^2 (j\mathfrak{h})^\gamma \delta_{n,\tilde{m}}^2} P_f^{(n)}(1-\tilde{\phi}_n) &\leq 2c_6 e^{-(c_7/\mathfrak{c}^2) nj\mathfrak{h}\delta_{n,\tilde{m}}^2}.
\end{aligned}$$

Similar to (A.8), there exists an event $\tilde{\mathcal{E}}_n$ with

$$P_{f_{0,m}}^{(n)}(\tilde{\mathcal{E}}_n^c) \leq c_1 e^{-C' c_7 nj\mathfrak{h}\delta_{n,\tilde{m}}^2 / 8c_3 \mathfrak{c}^2}$$

and on the event $\tilde{\mathcal{E}}_n$,

$$\begin{aligned}&\int \prod_{i=1}^n \frac{p_f}{p_{f_{0,m}}}\,\mathrm{d}\Pi_n(f)\\ &\geq \lambda_n(m)e^{-c_7 nj\mathfrak{h}\delta_{n,\tilde{m}}^2/4\mathfrak{c}^2}\Pi_{n,m}(\{f\in\mathcal{F}_m : d_n^2(f,f_{0,m})\leq c_7 j\mathfrak{h}\delta_{n,\tilde{m}}^2/8c_3\mathfrak{c}^2\}).\end{aligned}$$

Repeating as in (A.9),

$$\begin{aligned}&P_{f_{0,m}}^{(n)}\Pi_n\big(f\in\mathcal{F}: d_n^2(f,f_{0,m}) > \mathfrak{c}^4(2j\mathfrak{h})^\gamma d_n^2(f_0,f_{0,m})\big|X^{(n)}\big)(1-\tilde{\phi}_n)\mathbf{1}_{\tilde{\mathcal{E}}_n}\\ &\leq \frac{e^{c_7 nj\mathfrak{h}\delta_{n,\tilde{m}}^2/4\mathfrak{c}^2}}{\lambda_n(m)\Pi_{n,m}(\{f\in\mathcal{F}_m : d_n^2(f,f_{0,m})\leq c_7 j\mathfrak{h}\delta_{n,\tilde{m}}^2/8c_3\mathfrak{c}^2\})}\\ &\qquad\qquad \times\int_{f\in\mathcal{F}:d_n^2(f,f_{0,m})>\mathfrak{c}^4(2j\mathfrak{h})^\gamma d_n^2(f_0,f_{0,m})} P_f^{(n)}(1-\tilde{\phi}_n)\,\mathrm{d}\Pi_n(f)\\ &\leq (\cdots)\times\bigg(\sup_{f\in\mathcal{F}_{j\mathfrak{h}\tilde{m}}:d_n^2(f,f_{0,m})\geq\mathfrak{c}^2(j\mathfrak{h})^\gamma\delta_{n,\tilde{m}}^2} P_f^{(n)}(1-\tilde{\phi}_n)+\Pi_n\big(\mathcal{F}\setminus\mathcal{F}_{j\mathfrak{h}\tilde{m}}\big)\bigg)\\ &\leq Ce^{-(c_7/4\mathfrak{c}^2)nj\mathfrak{h}\delta_{n,\tilde{m}}^2}.\end{aligned}$$

Here the third line is valid since $\mathfrak{c}^4(2j\mathfrak{h})^\gamma d_n^2(f_0,f_{0,m}) > \mathfrak{c}^4(2j\mathfrak{h})^\gamma\delta_{n,\tilde{m}-1}^2 \geq \mathfrak{c}^2(j\mathfrak{h})^\gamma\delta_{n,\tilde{m}}^2$ by the right side of (2.5), which entails $\delta_{n,\tilde{m}}^2\leq\mathfrak{c}^2 2^\gamma\delta_{n,\tilde{m}-1}^2$. The fourth line uses exponential testability and assumption (P1), together with the fact that $\delta_{n,\tilde{m}}\geq\delta_{n,m}$. (A.2) follows from exponential testability, probability estimate for $\mathcal{E}_n^c$. □

Appendix E: Some formal connections with frequentist theory for *M*-estimators

In this section, we establish some formal structural similarities between the Bayes theory developed in this paper under the local Gaussianity condition Assumption A, and the frequentist theory for M-estimators.

Let us consider the simplest setup where only one big model $\mathcal{F}$ is available, and we consider the sieved MLE $\hat{f}_n$ for illustration of the Gaussian concentration technique. To this end, let $\delta_n > 0$ be determined by the entropy condition

$$\log\mathcal{N}(\delta_n,\mathcal{F},d_n)\leq\kappa\cdot n\delta_n^2, \tag{E.1}$$

where $\kappa>0$ is a small enough constant depending on the constants in Assumption A. The sieved MLE $\hat{f}_n$ is defined by $\hat{f}_n\equiv\arg\max_{f\in\mathcal{F}_{\delta_n}}\log p_f^{(n)}(X^{(n)})$, where $\mathcal{F}_{\delta_n}$ is a minimal δ_n-net of $\mathcal{F}$ under d_n.

Proposition E.1. *Suppose the local Gaussianity condition Assumption A and the entropy condition (E.1) hold. Then the sieved MLE defined above satisfies* $P_{f_0}^{(n)}\big(d_n^2(\hat{f}_n,f_0)>\delta_n^2\big)\leq\exp(-\kappa' n\delta_n^2)$, *where* $\kappa'>0$ *is a constant depending on the constants in Assumption A.*

The entropy condition (E.1) used for the sieved MLE is of global type since the construction of the net $\mathcal{F}_{\delta_n}$ does not allow information on f_0. Results of this type in the context of Gaussian regression and density estimation have long been known in the literature; we only refer the readers to [vdVW96, vdG00]. Our result here seems to yield some new results for other locally Gaussian experiments considered in Section 3.

The structural similarity of Theorem 2.3 (when only one model is used) and Proposition E.1 is obvious: both assertions hold under the same local Gaussianity structure of the experiment and the entropy condition, and the posterior distribution in Theorem 2.3 and the sieved MLE in Proposition E.1 both enjoy Gaussian tail behavior. Furthermore, the proofs for both results use (one-sided) Gaussian concentration in an essential way.

Proof of Proposition E.1. Let $S_j \equiv \{f \in \mathcal{F}_{\delta_n} : 2^{j-1}\delta_n \leq d_n(f, f_0) \leq 2^j \delta_n\}$. If $\hat{f}_n \in S_j$, then since $\log p^{(n)}_{f_0}/p^{(n)}_{\hat{f}_n} \leq 0$, it follows that

$$\max_{f\in S_j}\big(P^{(n)}_{f_0}\log(p^{(n)}_{f_0}/p^{(n)}_f) - \log(p^{(n)}_{f_0}/p^{(n)}_f)\big) \geq P^{(n)}_{f_0}\log(p^{(n)}_{f_0}/p^{(n)}_{\hat{f}_n}) \geq c_2 2^{2j-2} n\delta_n^2.$$

This implies that

$$\begin{aligned}
&P^{(n)}_{f_0}\big(d_n(\hat{f}_n, f_0) > \delta_n\big)\\
&\leq \sum_{j=1}^\infty P^{(n)}_{f_0}\bigg(\max_{f\in S_j}\big(P^{(n)}_{f_0}\log(p^{(n)}_{f_0}/p^{(n)}_f) - \log(p^{(n)}_{f_0}/p^{(n)}_f)\big) \geq c_2 2^{2j-2} n\delta_n^2\bigg)\\
&\leq \sum_{j=1}^\infty \sum_{f\in S_j} P^{(n)}_{f_0}\bigg(P^{(n)}_{f_0}\log(p^{(n)}_{f_0}/p^{(n)}_f) - \log(p^{(n)}_{f_0}/p^{(n)}_f) \geq c_2 2^{2j-2} n\delta_n^2\bigg)\\
&\leq \sum_{j=1}^\infty N(\delta_n) e^{-C_1 2^{2j} n\delta_n^2} \leq e^{-C_2 n\delta_n^2 + \log N(\delta_n)} \leq e^{-C_3 n\delta_n^2},
\end{aligned}$$

as desired. □

Appendix F: More examples

This section contains addition examples, including (i) regression models without boundedness restrictions, (ii) density estimation in location mixtures, (iii) estimation of piecewise constant signals in the Gaussian autoregression model and (iv) subset selection for sparse approximation of regression functions. The main purpose of (i) and (ii) is to demonstrate how the localization principle (cf. Section 2.3) can be applied in situations where local Gaussianity may fail over the entire parameter space, but still essentially holds on suitably localized subsets of the parameter space. The purpose of (iii) is to perform some explicit calculations without losing additional logarithmic factors, when the parameter space is non-compact. The purpose of (iv) is to demonstrate how to adapt the machinery in the paper to complicated model structures that are non-nested.

F.1. Removing boundedness restrictions in Section 3.1

The boundedness assumption in many examples in Section 3.1 is imposed for simplicity. Below we will remove the boundedness restriction in the binary regression model as a proof of concept.

Let $n \geq 3$. Consider fitting $X_i \sim_{\text{i.i.d.}} \text{Bern}(\theta_i)$ by piecewise constant model $\Theta \equiv \{\theta \in [0,1]^n\} = \cup_{m=1}^n \Theta_m$, where $\Theta_m \equiv \{\theta \in \Theta$ has at most m constant pieces$\}$. The model selection prior Λ_n on m is chosen as

$$\lambda_n(m) \propto \exp(-c^{\text{bin}} m \log(en)). \tag{F.1}$$

For the selected model Θ_m, we use the prior $\Pi_{n,m}$ which first randomly selects $m-1$ change points from $\{2, \ldots, n-1\}$, and then assigns a product prior with density $g^{\otimes(m-1)}$ where g is a density on $[0,1]$.

Proposition F.1. *Suppose $\theta_0 \in \Theta_m$ and $\theta_0 \in [\eta, 1-\eta]^n$ for some $\eta > 0$. If g is such that $\int_{x \in [0,t] \cup [1-t,1]} g(x)\, \mathrm{d}x \leq e^{-1/t^C}$ for some large constant $C > 0$ and $t > 0$ small. Then there exists $C' > 0$ (depending on η and the prior) such that $P_{\theta_0}^{(n)} \Pi_n\big(\theta \in \Theta : \|\theta - \theta_0\|_2^2 > C' m \log^{C'} n/n\big) \to 0$.*

The boundedness restrictions in other Laplace/Poisson models can be removed in a completely similar fashion so we omit these digressions.

Proof. Let $\delta_{n,m}^2 \equiv cm \log^c n/n$ for some large constant $c > 0$. Let the localized parameter spaces be defined by $\bar{\Theta}_n \equiv \{\theta \in \Theta : w_n \leq \theta_1, \ldots, \theta_n \leq 1 - w_n\}$, where $w_n \equiv 1/\log n$. By the decomposition (2.12),

$$\begin{aligned} &P_{\theta_0}^{(n)} \Pi_n\big(\theta \in \Theta : \|\theta - \theta_0\|_2^2 > \delta_{n,m}^2 \big| X^{(n)}\big) \\ &\quad \leq P_{\theta_0}^{(n)} \bar{\Pi}_n\big(\theta \in \bar{\Theta}_n : \|\theta - \theta_0\|_2^2 > \delta_{n,m}^2 \big| X^{(n)}\big) + P_{\theta_0}^{(n)} \Pi_n(\theta \notin \bar{\Theta}_n | X^{(n)}). \end{aligned} \tag{F.2}$$

For the first term in (F.2), we use Theorem 2.3. By the proof of Lemma 3.1, for any $\theta_0, \theta_1 \in \bar{\Theta}_n$,

$$w_n^2 \log(1/w_n) \|\theta_0 - \theta_1\|_2^2 \lesssim n^{-1} P_{\theta_0}^{(n)} \log(p_{\theta_0}^{(n)}/p_{\theta_1}^{(n)}) \lesssim (w_n \log(1/w_n))^{-1} \|\theta_0 - \theta_1\|_2^2.$$

Similarly we may verify the local Gaussianity condition with constants $\kappa = (\kappa_g, \kappa_\Gamma)$ depending polynomially on w_n. So Assumption A is verified by choosing $\{c_i\}$ and κ (or its inverse) on the order of $\mathcal{O}(w_n^{C_1})$ for some $C_1 > 0$. Assumption B can be verified immediately using the similar arguments as in Lemma C.4. Assumption C follows by similar (and simpler) arguments in Lemma C.5 and the fact that $\bar{\Pi}_{n,m}(A) \geq \Pi_{n,m}(A)$ for any A. Hence, the first term on the RHS of (F.2) is bounded by

$$\exp\big(C_2 \log(1/\omega_n) - n\delta_{n,m}^2 \omega_n^{C_2}\big),$$

which is $o(1)$ by our choice of w_n and $c > 0$ large enough.

We handle the second term on the right hand side of (F.2) below. By applying Lemma A.5 to the localized model with $\varepsilon^2 \equiv \delta^2_{n,m}$, we see that on an event $\mathcal{E}_n$ with $P^{(n)}_{\theta_0}$ probability at least $1 - e^{-m \log^C n \cdot w_n^{C_3}} = 1 - o(1)$,

$$\begin{aligned}
&\int_{\bar{\Theta}_n} p^{(n)}_\theta / p^{(n)}_{\theta_0}\ \mathrm{d}\bar{\Pi}_n(\theta) \geq \lambda_n(m) \int_{\theta \in \bar{\Theta}_n : \|\theta - \theta_0\|_2^2 \leq \delta^2_{n,m}/c_3} p^{(n)}_\theta / p^{(n)}_{\theta_0}\ \mathrm{d}\bar{\Pi}_{n,m}(\theta) \\
&\gtrsim e^{-m \log^c n \cdot w_n^{C_4}} \cdot \bar{\Pi}_{n,m}\big(\{\theta \in \Theta_m \cap \bar{\Theta}_n : \|\theta - \theta_0\|_2^2 \leq \delta^2_{n,m}/c_3\}\big) \\
&\gtrsim e^{-m \log^c n \cdot w_n^{C_4} - m \log n / C_5} \big(\Pi_n(\bar{\Theta}_n)\big)^{-1} \gtrsim e^{-m \log^{C_6} n / C_6} \big(\Pi_n(\bar{\Theta}_n)\big)^{-1}
\end{aligned}$$

by choosing $c > 0$ large enough. Now we have that

$$\begin{aligned}
&P^{(n)}_{\theta_0} \Pi_n\big(\theta \notin \bar{\Theta}_n | X^{(n)}\big) \leq P^{(n)}_{\theta_0} \Pi_n\big(\theta \notin \bar{\Theta}_n | X^{(n)}\big) \mathbf{1}_{\mathcal{E}_n} + P^{(n)}_{\theta_0}(\mathcal{E}_n^c) \\
&= P_{\theta_0^{(n)}} \left[\frac{\int_{\theta \notin \bar{\Theta}_n} p^{(n)}_\theta / p^{(n)}_{\theta_0}\ \mathrm{d}\Pi_n(\theta)}{\int_\Theta p^{(n)}_\theta / p^{(n)}_{\theta_0}\ \mathrm{d}\Pi_n(\theta)} \mathbf{1}_{\mathcal{E}_n} \right] + P^{(n)}_{\theta_0}(\mathcal{E}_n^c) \\
&\leq \frac{1}{\Pi_n(\bar{\Theta}_n)} \cdot P_{\theta_0^{(n)}} \left[\frac{\int_{\theta \notin \bar{\Theta}_n} p^{(n)}_\theta / p^{(n)}_{\theta_0}\ \mathrm{d}\Pi_n(\theta)}{\int_{\bar{\Theta}_n} p^{(n)}_\theta / p^{(n)}_{\theta_0}\ \mathrm{d}\bar{\Pi}_n(\theta)} \mathbf{1}_{\mathcal{E}_n} \right] + P^{(n)}_{\theta_0}(\mathcal{E}_n^c) \\
&\lesssim e^{m \log^{C_6} n / C_6} \cdot \Pi_n(\Theta \setminus \bar{\Theta}_n) + o(1),
\end{aligned}$$

where in the last inequality we used a previous inequality and Fubini's theorem. On the other hand,

$$\begin{aligned}
\Pi_n(\Theta \setminus \bar{\Theta}_n) &\leq \sum_{k \geq 1} \lambda_n(k) k \int_{x \in [0, w_n] \cup [1 - w_n, 1]} g(x)\ \mathrm{d}x \\
&\lesssim \int_{x \in [0, w_n] \cup [1 - w_n, 1]} g(x)\ \mathrm{d}x \leq e^{-\log^{C_7} n / C_7}
\end{aligned}$$

for some large $C_7 > 0$ by the assumption on g. □

F.2. Density estimation in location mixtures

Consider estimation of a density f_0 on $\mathbb{R}$ from the class of location mixtures $\cup_{m=1}^\infty \mathcal{F}_m$ where $\mathcal{F}_m$ consists densities of the type

$$h(x; m, \mu, w, \sigma) \equiv \sum_{j=1}^m w_j \psi_\sigma(x - \mu_j),$$

where $\sigma > 0$, $w_j \geq 0$, $\sum_{j=1}^m w_j = 1$, $\mu_j \in \mathbb{R}$ and $\psi_\sigma(x) \equiv e^{-x^2/2\sigma^2} / \sqrt{2\pi\sigma^2}$. This problem has received considerable attention, see e.g. [GvdV01, Rou10, KRvdV10, Scr16, DRRS18] and references therein for some Bayesian developments. The model selection prior Λ_n on m is chosen as

$$\lambda_n(m) \propto \exp(-c^{\mathrm{mix}} m \log(en)). \tag{F.3}$$

A prior $\Pi_{n,m}$ on the model $\mathcal{F}_m$ is naturally induced by a product prior $\Pi_w \otimes \Pi_\mu \otimes \Pi_\sigma$. For simplicity, we assume that Π_w has the standard Dirichlet distribution, Π_μ, Π_σ have Lebesgue density $g_\mu^{\otimes m}, g_\sigma$ with the following properties: g_μ has full support on $\mathbb{R}$ such that $-\log g_\mu(x) \asymp \log(x)$ as $x \to \infty$, and $-\log g_\sigma(x) \asymp \log(1/x)$ as $x \to 0$ and $-\log g_\sigma(x) \asymp \log(x)$ as $x \to \infty$.

Proposition F.2. *Suppose that $f_0 \in \mathcal{F}_m$, and the priors are specified as above. Then there exist $C > 0, \gamma > 0$ depending only on the priors such that $P_{f_0}^{(n)} \Pi_n\big(f \in \mathcal{F} : h^2(f, f_0) > Cm \log^\gamma n/n \big| X^{(n)}\big) \to 0$.*

Proposition F.2 says that the posterior distribution under such hierarchical priors adapts to the finite mixtures at a nearly parametric rate. Although this result does not seem to be explicitly spelled out in the literature, we believe that it can also be derived along the lines, e.g. [KRvdV10]. Indeed, [KRvdV10] proved adaptive behavior of the posterior contraction rates with respect to the local smoothness of the density, under similar hierarchical priors. It is clear from the above proposition that adaptation to the smoothness of the density can be accomplished once the quantity $\inf_{f \in \mathcal{F}_m} h^2(f, f_0)$ can be shown to be adaptive to the smoothness of f_0. This has been the main focus of [KRvdV10] (in Kullback-Leibler divergence). The main purpose here, instead of repeating along the lines of [KRvdV10], rests in demonstrating how the localization principle can be used in the mixture model.

It can also be seen immediately from the proof that the Gaussian kernel can be replaced by any kernel of form considered in [KRvdV10].

Proof of Proposition F.2. Let $\bar{\mathcal{F}}_n \equiv \{h(\cdot; m, \mu, w, \sigma) : \mu \in [-b_n, b_n]^m, \sigma \in [\underline{\sigma}_n, \bar{\sigma}_n]\}$ where $b_n \asymp (\log n)^{\gamma_1}, \bar{\sigma}_n \asymp \underline{\sigma}_n^{-1} \asymp (\log n)^{\gamma_2}$ for a sufficiently large $\gamma_1 > \gamma_2$. For any $f \in \bar{\mathcal{F}}_n$, define $\tilde{f} \equiv f\mathbf{1}_{[-2b_n, 2b_n]} + f_0 \mathbf{1}_{\mathbb{R}\setminus[-2b_n, 2b_n]}$, and $f^* = \tilde{f}/\int \tilde{f}$. Note that

$$\int_{\mathbb{R}} \tilde{f}(x)\ \mathrm{d}x = 1 - \int_{\mathbb{R}\setminus[-2b_n, 2b_n]} (f + f_0)(x)\ \mathrm{d}x = 1 + \mathcal{O}(e^{-b_n^2/(2\bar{\sigma}_n^2)}),$$

since

$$\begin{aligned}\int_{\mathbb{R}\setminus[-2b_n, 2b_n]} f(x)\ \mathrm{d}x &\lesssim \Big(\sum_j w_j\Big) \int_{2b_n}^\infty \psi_\sigma(x - b_n)\ \mathrm{d}x \\ &\lesssim \int_{b_n/\bar{\sigma}_n}^\infty e^{-x^2/2}\ \mathrm{d}x \lesssim e^{-b_n^2/(2\bar{\sigma}_n^2)},\end{aligned}$$

and for n large

$$\int_{\mathbb{R}\setminus[-2b_n, 2b_n]} f_0(x)\ \mathrm{d}x \lesssim e^{-b_n^2/(2\bar{\sigma}_n^2)}.$$

Now define $\bar{\mathcal{F}}_n^*$ to be the set containing all f^* defined as above from some $f \in \bar{\mathcal{F}}_n$. Note that for any $f \in \bar{\mathcal{F}}_n$, we have that

$$h^2(f, f_0) \lesssim h^2(f^*, f_0) + h^2(f^*, f) \lesssim h^2(f^*, f_0) + \mathcal{O}(e^{-b_n^2/(2\bar{\sigma}_n^2)}).$$

Then for a large enough constant $C > 0$, by the decomposition (2.12), we have for n large,

$$
\begin{aligned}
&P_{f_0}^{(n)}\Pi_n\bigg(f \in \mathcal{F} : h^2(f, f_0) > C\frac{m \log^\gamma n}{n}\bigg| X^{(n)}\bigg) \\
&\le P_{f_0}^{(n)}\bar{\Pi}_n\bigg(f \in \bar{\mathcal{F}}_n : h^2(f^*, f_0) > C_1\frac{m \log^\gamma n}{n}\bigg| X^{(n)}\bigg) + P_{f_0}^{(n)}\Pi_n\big(f \notin \bar{\mathcal{F}}_n \big| X^{(n)}\big),
\end{aligned}
$$

which can be bounded by

$$
\begin{aligned}
\text{(F.4)} \qquad & P_{f_0}^{(n)}\bar{\Pi}_n^*\big(f^* \in \bar{\mathcal{F}}_n^* : h^2(f^*, f_0) > C_1 m \log^\gamma n/n \big| X^{(n)}\big) \\
&\quad + P_{f_0}^{(n)}\Pi_n^*\big(f^* \notin \bar{\mathcal{F}}_n^* \big| X^{(n)}\big) + 1/n
\end{aligned}
$$

where $\Pi_n^*, \bar{\Pi}_n^*$ are the natural induced priors from $\Pi_n, \bar{\Pi}_n$. The last inequality follows by noting that

$$
P_{f_0}^{(n)}\big(\max_{1\le i\le n} |X_i| > 2b_n\big) \lesssim n e^{-b_n^2/(2\bar{\sigma}_n^2)} \le 1/(2n)
$$

for $\gamma_1 \gg \gamma_2$.

We handle the first term on the right hand side of (F.4). To this end, we first verify the local Gaussianity condition Assumption A. Clearly for any $f_0^*, f_1^* \in \bar{\mathcal{F}}_n^*$,

$$
\sup_{x\in\mathbb{R}} \left|\frac{f_0^*(x)}{f_1^*(x)}\right| \le \sup_{x\in[-2b_n, 2b_n]} \left|\frac{f_0(x)}{f_1(x)}\right| \cdot \frac{1 + \mathcal{O}(e^{-b_n^2/(2\bar{\sigma}_n^2)})}{1 - \mathcal{O}(e^{-b_n^2/(2\bar{\sigma}_n^2)})} \lesssim \frac{\bar{\sigma}_n}{\underline{\sigma}_n} e^{(3b_n/\underline{\sigma}_n)^2}.
$$

By Lemma 8 of [GvdV07b],

$$
\begin{aligned}
h^2(f_0^*, f_1^*) &\le \frac{1}{n} P_{f_0^*}^{(n)} \log\big(P_{f_0^*}^{(n)}/P_{f_1^*}^{(n)}\big) \lesssim h^2(f_0^*, f_1^*)\big(1 + \log\|f_0^*/f_1^*\|_\infty\big) \\
&\lesssim h^2(f_0^*, f_1^*)\big(1 + (b_n/\underline{\sigma}_n)^2 + \log(\bar{\sigma}_n/\underline{\sigma}_n)\big), \\
\mathrm{Var}_{f_0^*}\big(\log(f_0^*/f_1^*)\big) &\lesssim h^2(f_0^*, f_1^*)\big(1 + \log\|f_0^*/f_1^*\|_\infty\big)^2 \\
&\lesssim h^2(f_0^*, f_1^*)\big(1 + (b_n/\underline{\sigma}_n)^2 + \log(\bar{\sigma}_n/\underline{\sigma}_n)\big)^2.
\end{aligned}
$$

By the classical Bernstein inequality, the local Gaussianity condition on $\bar{\mathcal{F}}_n^*$ holds with $c_1 = c_2 = 1$, $c_3 \asymp \kappa_\Gamma \asymp \big(1 + (b_n/\underline{\sigma}_n)^2 + \log(\bar{\sigma}_n/\underline{\sigma}_n)\big)$ and $\kappa_g \asymp \big(1 + (b_n/\underline{\sigma}_n)^2 + \log(\bar{\sigma}_n/\underline{\sigma}_n)\big)^2$.

Next we verify Assumption B. Let $\delta_{n,m}^2 \equiv C'\frac{m}{n}\log n$ for some large constant $C' > 0$. Since the Hellinger distance is bounded by the square root of total variational distance, we have

$$
\log \mathcal{N}(c_5\varepsilon, \bar{\mathcal{F}}_n^* \cap \mathcal{F}_m, h) \le \log \mathcal{N}(c_5^2\varepsilon^2, \bar{\mathcal{F}}_n^* \cap \mathcal{F}_m, d_{\mathrm{TV}}).
$$

By Lemma 3 of [KRvdV10], for any $f_0^*, f_1^* \in \bar{\mathcal{F}}_n^* \cap \mathcal{F}_m$ that are defined through $f_i = h(\cdot; m, \mu^i, w^j, \sigma^j)(j = 0, 1)$,

$$
d_{\mathrm{TV}}(f_0^*, f_1^*)
$$

$$
\begin{aligned}
&\lesssim \mathcal{O}(e^{-b_n^2/(2\bar{\sigma}_n^2)}) + \|w^0 - w^1\|_1 + \|\psi\|_\infty \sum_{i=1}^m \frac{w_i^0 \wedge w_i^1}{\sigma^0 \wedge \sigma^1} |\mu_i^0 - \mu_i^1| + \frac{|\sigma^0 - \sigma^1|}{\sigma^0 \wedge \sigma^1} \\
&\leq C_2 \big(e^{-b_n^2/(2\bar{\sigma}_n^2)} + \|w^0 - w^1\|_1 + \underline{\sigma}_n^{-1} \|\mu^0 - \mu^1\|_1 + \underline{\sigma}_n^{-1} |\sigma^0 - \sigma^1| \big).
\end{aligned}
$$

Here $C_2 > 0$ is an absolute constant. Now for any $1 \geq \varepsilon^2 \geq 4C_2 e^{-b_n^2/(2\bar{\sigma}_n^2)}/c_5^2$, with Δ^m denoting the unit simplex in $\mathbb{R}^m$ and using Lemma 5 of [KRvdV10], we have

$$
\begin{aligned}
&\log \mathcal{N}(c_5^2 \varepsilon^2, \bar{\mathcal{F}}_n^* \cap \mathcal{F}_m, d_{\mathrm{TV}}) \\
&\leq \log \mathcal{N}\bigg(\frac{c_5^2 \varepsilon^2}{4C_2}, \Delta^m, \|\cdot\|_1 \bigg) + \log \mathcal{N}\bigg(\frac{c_5^2 \varepsilon^2 \underline{\sigma}_n}{4C_2}, [-b_n, b_n]^m, \|\cdot\|_1 \bigg) \\
&\qquad\qquad + \log \mathcal{N}\bigg(\frac{c_5^2 \varepsilon^2 \underline{\sigma}_n}{4C_2}, [\underline{\sigma}_n, \bar{\sigma}_n], |\cdot| \bigg) \\
&\leq m \log \bigg(\frac{20 C_2}{c_5^2 \varepsilon^2} \bigg) + \log \bigg(\frac{m! (b_n + 1)^m (4C_2)^m}{(c_5^2 \varepsilon^2 \underline{\sigma}_n)^m} \bigg) + \log \bigg(\frac{4C_2 (\bar{\sigma}_n - \underline{\sigma}_n)}{c_5^2 \varepsilon^2 \underline{\sigma}_n} \bigg).
\end{aligned}
$$

Using that $c_5 \asymp (c_3 \kappa_\Gamma)^{-1} \wedge (c_3 \kappa_g)^{-1}$ and $\log(m!) \lesssim m \log m$, we have

$$
\begin{aligned}
\log \mathcal{N}(c_5 \varepsilon^2, \bar{\mathcal{F}}_n^* \cap \mathcal{F}_m, d_{\mathrm{TV}}) &\lesssim m \bigg(\log m + \log \bigg(\frac{C_3 b_n \vee (\bar{\sigma}_n - \underline{\sigma}_n)}{c_5^2 \varepsilon^2 \underline{\sigma}_n} \bigg) \bigg) \\
&\lesssim m \log n \leq (c_7/2) n \delta_{n,m}^2.
\end{aligned}
$$

It is easy to check that ε^2 hits the boundary $\delta_{n,m}^2$ by choosing $\gamma > 0$ large enough.

We continue to verify Assumption C. As before, it suffices to control from below the quantity $\Pi_{n,m}\big(\{f \in \bar{\mathcal{F}}_n \cap \mathcal{F}_m, h^2(f, f_0) \leq \delta_{n,m}^2/(2c_3)\}\big)$. Again by Lemma 3 of [KRvdV10], for any $f_1, f_2 \in \bar{\mathcal{F}}_n \cap \mathcal{F}_m$ with $f_i = h(\cdot; m, \mu^i, w^j, \sigma^j)(j = 1, 2)$, we have

$$
h^2(f_1, f_2) \lesssim d_{\mathrm{TV}}(f_1, f_2) \leq C_4 \big(\|w^1 - w^2\|_1 + \underline{\sigma}_n^{-1} \|\mu^1 - \mu^2\|_1 + \underline{\sigma}_n^{-1} |\sigma^1 - \sigma^2| \big).
$$

In view of Lemma 6 of [KRvdV10], the above display implies

$$
\begin{aligned}
&\Pi_{n,m}\big(\{f \in \bar{\mathcal{F}}_n \cap \mathcal{F}_m, h^2(f, f_0) \leq \delta_{n,m}^2/(2c_3)\}\big) \\
&\geq \Pi_w\big(\Delta_m(w^0, \delta_{n,m}^2/(6C_4 c_3))\big) \\
&\qquad\qquad \times \prod_{j=1}^m \Pi_\mu \bigg(|\mu_j - \mu_j^0| \leq \frac{\delta_{n,m}^2 \underline{\sigma}_n}{6C_4 m c_3} \bigg) \Pi_\sigma \bigg(|\sigma - \sigma^0| \leq \frac{\delta_{n,m}^2 \underline{\sigma}_n}{6C_4 c_3} \bigg) \\
&\gtrsim e^{-C_5 m \log n} \geq e^{-2n\delta_{n,m}^2}
\end{aligned}
$$

Now apply Theorem 2.3, we see that the first term on the right hand side of (F.4) can be bounded by $\exp(-C \log^{\gamma'} n)$ for some $\gamma' > 0$ if $\gamma > 0$ is chosen large enough.

Next we handle the second term on the RHS of (F.4). By applying Lemma A.5 to the localized model with $\varepsilon^2 \equiv \delta_{n,m}^2$ and using the same arguments as before, on an event $\mathcal{E}_n$ with $P_{f_0}^{(n)}$ probability at least $1 - e^{-C_6 m \log^{\gamma''} n}$,

$$\int_{\bar{\mathcal{F}}_n^*} p_{f^*}^{(n)}/p_{f_0}^{(n)} \, \mathrm{d}\bar{\Pi}_n^*(f^*) \geq \lambda_n(m) \int_{f \in \bar{\mathcal{F}}_n \cap \mathcal{F}_m, h^2(f,f_0) \leq \delta_{n,m}^2/c_3} p_{f^*}^{(n)}/p_{\theta_0}^{(n)} \, \mathrm{d}\bar{\Pi}_{n,m}^*(f^*)$$
$$\gtrsim e^{-C_7 m \log^{C_8} n} \cdot (\Pi_n^*(\bar{\mathcal{F}}_n))^{-1} \Pi_{n,m}^*\big(\{f \in \bar{\mathcal{F}}_n \cap \mathcal{F}_m, h^2(f,f_0) \leq \delta_{n,m}^2/c_3\}\big)$$
$$\gtrsim e^{-m \log^{C_9} n} (\Pi_n^*(\bar{\mathcal{F}}_n))^{-1}.$$

Similar as above, we have

$$P_{f_0}^{(n)} \Pi_n^*\big(f^* \notin \bar{\mathcal{F}}_n^* \big| X^{(n)}\big)$$
$$\leq \frac{1}{\Pi_n(\bar{\mathcal{F}}_n^*)} \cdot P_{f_0^{(n)}} \left[\frac{\int_{f^* \notin \bar{\mathcal{F}}_n^*} p_{f^*}^{(n)}/p_{f_0}^{(n)} \, \mathrm{d}\Pi_n^*(f^*)}{\int_{\bar{\mathcal{F}}_n^*} p_{f^*}^{(n)}/p_{f_0}^{(n)} \, \mathrm{d}\bar{\Pi}_n^*(f^*)} \mathbf{1}_{\mathcal{E}_n} \right] + P_{f_0}^{(n)}(\mathcal{E}_n^c)$$
$$\lesssim e^{m \log^{C_9} n} \cdot \Pi_n(\mathcal{F} \setminus \bar{\mathcal{F}}_n) + e^{-C_6 m \log^{\gamma''} n}.$$

Furthermore, for γ_1, γ_2 large enough,

$$\Pi_n(\mathcal{F} \setminus \bar{\mathcal{F}}_n) \leq \Pi_\sigma\big(\sigma \notin [\underline{\sigma}_n, \bar{\sigma}_n]\big) + \sum_{m=1}^\infty \lambda_n(m) \Pi_\mu\bigg(\max_{1 \leq j \leq m} |\mu_j| > b_n\bigg)$$
$$\lesssim e^{-\log^{(C_9+1)} n} + \sum_{m=1}^\infty e^{-C_{10}(m-1)\log n} m \bigg(\int_{\mathbb{R} \setminus [-b_n, b_n]} g_\mu(x)\bigg) \lesssim e^{-\log^{C_{11}} n}.$$

Hence $P_{f_0}^{(n)} \Pi_n^*\big(f^* \notin \bar{\mathcal{F}}_n^* \big| X^{(n)}\big) = o(1)$ and the proof is complete. □

F.3. *Estimation of piecewise constant signals in the Gaussian autoregression model*

Consider fitting the Gaussian autoregression model (cf. Section 3.3) by the class of piecewise constant functions $\mathcal{F} \equiv \cup_{m=1}^\infty \mathcal{F}_m \equiv \{f : f = \sum_{j=1}^m a_j \mathbf{1}_{[t_{j-1}, t_j)}, -\infty = t_0 < t_1 < \ldots < t_{m-1} < t_m = \infty, |a_j| \leq M\}$. Consider the following model selection prior Λ_n on the model index $\mathcal{I} \equiv \mathbb{N}$:

$$\lambda_n(m) \propto \exp\big(-c \cdot m \log(en)\big), \tag{F.5}$$

where $c > 0$ is a constant to be specified later. Similar to the development in Section 3.7.1, we choose the prior $\Pi_{n,m}^t$ on $(t_1, \ldots, t_{m-1})$ with density $\boldsymbol{t} = (t_1, \ldots, t_{m-1}) \mapsto (m-1)! g_t^{\otimes(m-1)} \mathbf{1}_{t_1 < \ldots < t_{m-1}}(\boldsymbol{t})$, and the prior $\Pi_{n,m}^a$ on $(a_1, \ldots, a_m)$ with a product density $g_a^{\otimes m}$. Here we assume that g_t is symmetric and non-increasing on $[0, \infty)$, and g_a is uniform on $[-M, M]$ for simplicity. The difference in the Gaussian autoregression example, compared with the results in Section 3.7.1, is that the metric $d_{r,M}$ is defined on the entire real line $\mathbb{R}$. As commented on page 210 of [GvdV07a], "*....The logarithmic factor in the convergence rate appears to be a consequence of the fact that the regression functions are defined on the full real line...*". Below we perform some explicit computation to address this non-compact issue, with a particular goal of avoiding additional logarithmic factors (compared with the results in Section 3.7.1) in the contraction rates.

Proposition F.3. *Suppose that $M > \|f_0\|_\infty$, and that the prior density g_t satisfies $\limsup_{x\to\infty} \frac{1}{x^2}\log(1\vee \frac{1}{g_t(x)}) < \infty$. Then there exists some $c > 0$ in (F.5) such that*

$$P_{f_0}^{(n)}\Pi_n\big(f\in\mathcal{F}: d_{r,M}^2(f,f_0) > C_1(\varepsilon_{n,m}^{\mathrm{aut}})^2\big|X^{(n)}\big) \leq C_2 e^{-n(\varepsilon_{n,m}^{\mathrm{aut}})^2/C_2}.$$

Here $(\varepsilon_{n,m}^{\mathrm{aut}})^2 \equiv \max\{\inf_{g\in\mathcal{F}_m} d_{r,M}^2(f_0,g), m\log n/n\}$, and the constants $C_i(i=1,2)$ depend on M.

Note that the condition on g_t is quite mild: it essentially requires that the tail of g_t is not lighter than Gaussian.

Lemma F.4. *For any $g\in\mathcal{F}_m$, and $\varepsilon\in(0,1/e)$, $\log\mathcal{N}\big(c_5\varepsilon,\{f\in\mathcal{F}_m, d_{r,M}(f,g)\leq 2\varepsilon\}, d_{r,M}\big) \lesssim m\log\Big(\frac{C_M\log(1/\varepsilon)}{\varepsilon^4}\Big)$.*

Proof. We only need to consider global entropy $\mathcal{N}\big(c_5\varepsilon,\{f\in\mathcal{F}_m\},d_{r,M}\big)$.

Let $m\geq 2$. Fix $\varepsilon\in(0,1/e)$, let $R_\varepsilon = \lceil M+\sqrt{2\log(24M^2)+4\log(1/(c_5\varepsilon))}\rceil$ and $\delta^2 = \frac{c_5^2\varepsilon^2}{4(2M^2+1)R_\varepsilon}$. We partition the interval $[-R_\varepsilon,R_\varepsilon]$ into small intervals $\{I_{j,\delta}\}_{j=1}^{N_\delta}$ of length $R_\varepsilon\delta^2/m$ (ignoring the rounding issue here). For any $f\in\mathcal{F}_m$, let $f\equiv\sum_{j=1}^m a_j\mathbf{1}_{[t_{j-1},t_j)}$ for some $-\infty=t_0<t_1<\ldots<t_{m-1}<t_m=\infty$. Then $\{t_1,\ldots,t_{m-1}\}\cap[-R_\varepsilon,R_\varepsilon]$ must be contained in at most $m-1$ intervals amongst $\{I_{j,\delta}\}_{j=1}^{N_\delta}$, namely, $\{\bar{I}_{k,\delta;f}\}_{k=1}^{m_f}$. Furthermore, $[-R_\varepsilon,R_\varepsilon]\setminus\cup_{k=1}^{m_f}\bar{I}_{k,\delta;f}$ contains at most m intervals. Now define $\bar{f}$ as follows:

$$\bar{f}\equiv M\cdot\mathbf{1}_{\mathbb{R}\setminus[-R_\varepsilon,R_\varepsilon]}+\sum_{k=1}^{m_f}M\cdot\mathbf{1}_{\bar{I}_{k,\delta;f}}+\left\lfloor\frac{f}{\delta}\right\rfloor\delta\cdot\mathbf{1}_{[-R_\varepsilon,R_\varepsilon]\setminus\cup_{k=1}^{m_f}\bar{I}_{k,\delta;f}}.$$

Then using the well-known fact that $\int_t^\infty\phi(x)\,\mathrm{d}x\leq e^{-t^2/2}/(\sqrt{2\pi}t)(t>0)$, we have

$$\begin{aligned}&\int_{\mathbb{R}}\big(f(x)-\bar{f}(x)\big)^2 r_M(x)\,\mathrm{d}x\\&\leq 4M^2\int_{\mathbb{R}\setminus[-R_\varepsilon,R_\varepsilon]} r_M(x)\,\mathrm{d}x+4M^2m(R_\varepsilon\delta^2/m)+\int_{[-R_\varepsilon,R_\varepsilon]\setminus\cup_{k=1}^{m_f}\bar{I}_{k,\delta;f}}\delta^2\,\mathrm{d}x\\&\leq 8M^2\int_{R_\varepsilon}^\infty\phi(x-M)\,\mathrm{d}x+(4M^2+2)R_\varepsilon\delta^2\leq c_5^2\varepsilon^2\end{aligned}$$

by our choice of R_ε and δ. On the other hand, there are at most $\binom{2m/\delta^2}{m-1}\cdot\big(\frac{2M}{\delta}\big)^m$

many choices of $\bar{f}$, and hence

$$\log \mathcal{N}\big(c_5\varepsilon, \{f \in \mathcal{F}_m\}, d_{r,M}\big) \leq \log \left[\binom{2m/\delta^2}{m-1} \cdot \big(\frac{2M}{\delta}\big)^m\right] \lesssim m \log \left(\frac{C_M \log(1/\varepsilon)}{\varepsilon^4}\right).$$

For $m = 1$, we define $\bar{f} \equiv M \cdot \mathbf{1}_{\mathbb{R}\setminus[-R_\varepsilon, R_\varepsilon]} + \left\lfloor \frac{f}{\delta} \right\rfloor \delta \cdot \mathbf{1}_{[-R_\varepsilon, R_\varepsilon]}$, and repeat the above calculation to see that the entropy bound holds. □

Hence we can take $\delta_{n,m}^2 = Cm \log n/n$ for some large constant $C > 0$.

Lemma F.5. *Let $M > \|f_0\|_\infty$. Suppose g_t is such that $\limsup_{x\to\infty} \frac{1}{x^2} \log(1 \vee \frac{1}{g_t(x)}) < \infty$. For n large enough depending on $\|f_0\|_\infty$ and M, (P2) in Assumption C holds.*

Proof. The proof uses similar ideas as that of Lemma C.19. Let $f_{0,m} \equiv \sum_{j=1}^m a_j^* \mathbf{1}_{[t_{j-1}^*, t_j^*)}$ for some $t^* = (t_1^*, \ldots, t_{m-1}^*)$ with $-\infty = t_0^* < t_1^* < \ldots < t_{m-1}^* < t_m^* = \infty$. Without loss of generality, we may assume that $\min\{t_j^* - t_{j-1}^* : 2 \leq j \leq m-1\} > 1/(4nM^2)$ (otherwise we may merge such short intervals to construct a surrogate $\tilde{f}_{0,m}$, and the total difference between $\tilde{f}_{0,m}$ and $f_{0,m}$ in squared L_2 metric by doing this does not exceed m/n so that there is no effect in the final oracle inequality). For any $t = (t_1, \ldots, t_{m-1})$ such that $|t_j - t_j^*| < 1/(8nM^2)$ where $|t_j^*| \leq L_n$ with L_n specified later on, and any $a = (a_1, \ldots, a_m)$ such that $\max_j |a_j - a_j^*| \leq 1/\sqrt{n}$, let $f \equiv \sum_{j=1}^m a_j \mathbf{1}_{[t_{j-1}, t_j)}$. Then, $\|f\|_\infty \leq M$ for n large enough depending only through $\|f_0\|_\infty$ and M. Now with $L_n \equiv M + \sqrt{2\log(8M^2) + 2\log n}$,

$$\begin{aligned}
&\int_{-\infty}^{\infty} \big(f(x) - f_{0,m}(x)\big)^2 r_M(x)\,\mathrm{d}x \\
&\leq 8M^2 \int_{L_n}^{\infty} \phi(x-M)\,\mathrm{d}x + \int_{-L_n}^{L_n} \big(f(x) - f_{0,m}(x)\big)^2 r_M(x)\,\mathrm{d}x \\
&\leq \frac{1}{n} + \big(\frac{1}{n} + 4M^2 \cdot m \cdot \frac{1}{8nM^2}\big) \leq \frac{3m}{n} \leq \delta_{n,m}^2/c_3
\end{aligned}$$

by choosing the constant $C = C_M > 0$ in the definition of $\delta_{n,m}^2$ large enough. This implies that

$$\begin{aligned}
&\Pi_{n,m}\big(\{f \in \mathcal{F}_m : d_{r,M}^2(f, f_{0,m}) \leq \delta_{n,m}^2/c_3\}\big) \\
&\geq g_t(L_n)^{m-1}(16nM^2)^{-(m-1)}\left(\frac{2}{\sqrt{n}}\right)^m \\
&\geq e^{-m\left(\log g_t(L_n)^{-1} + \log(16nM^2) + \log\left(\sqrt{n}/2\right)\right)} \geq e^{-2n\delta_{n,m}^2}
\end{aligned}$$

by the assumption on g_t and again choosing the constant $C = C_M > 0$ in the definition of $\delta_{n,m}^2$ large enough. □

Proof of Proposition F.3. Proposition F.3 follows from Corollary 3.18 combined with Lemmas F.4 and F.5. □

F.4. Subset selection for sparse approximation of functions

Consider Gaussian regression with random design $Y_i = f_0(X_i) + \varepsilon_i (1 \leq i \leq n)$. We assume that X_i's are i.i.d. uniformly distributed on $[0,1]$ and are independent of ε_i's for simplicity of discussion. Let $\{\phi_k\}_{k=1}^\infty$ be an orthonormal basis of $L_2([0,1])$. Let $\boldsymbol{N} \equiv \{N_1, N_2, \ldots\} \subset \mathbb{N}$. For any $\boldsymbol{\gamma} \equiv (\gamma_0, \gamma_1, \ldots)$, let the $\boldsymbol{\gamma}$-sparse approximation space $\mathcal{S}(\boldsymbol{\gamma}, \boldsymbol{N}) \equiv \{f \in L_2([0,1]) : \min_{\ell_j \leq N_j, 1 \leq j \leq k} \min_{(a_{\ell_1}, \ldots, a_{\ell_k})} \|f - \sum_{j=1}^k a_{\ell_j} \phi_{\ell_j}\|_{L_2([0,1])} \leq \gamma_k, k = 0, 1, \ldots\}$. For any $\boldsymbol{\gamma}$ and $k \in \mathbb{N}$, let $\boldsymbol{\gamma}^{(k)} \equiv (\gamma_0, \gamma_1, \ldots, \gamma_{k-1}, 0, 0, \ldots)$, and $\mathcal{F}_k \equiv \mathcal{S}(\boldsymbol{\gamma}^{(k)}, \boldsymbol{N})$. Clearly $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots$. We use the model selection prior:

$$\lambda_n(k) \propto \exp\big(- c \cdot k \log(en) \big). \tag{F.6}$$

For each model $\mathcal{F}_k$, we use the prior $\Pi_{n,k}$ which first picks randomly a subset $I \subset \{1, \ldots, N_k\}$ with cardinality k, then puts a product prior $g^{\otimes |I|}$ on the coefficients $(a_j)_{j \in I}$. We assume for simplicity that g is symmetric and non-decreasing on $[0, \infty)$. Note that in Section 6.3 of [Yan99], the model index corresponds to (k, I) in our notation.

Proposition F.6. *Let $f_0 \in \mathcal{S}(\boldsymbol{\gamma}, \boldsymbol{N})$ be such that $\sup_k |\int f_0 \phi_k| < \infty$. Suppose the priors are specified as above and g satisfies $g\big(\sup_k |\int f_0 \phi_k| + 1\big) > 0$. Then if $\log N_k \lesssim \log k$, with $\varepsilon_{n,k}^2 \equiv \gamma_k + k \log(N_k \vee n)/n$, for n large,*

$$P_{f_0}^{(n)} \Pi_n \big(f : L_2^2(f, f_0) > C_1 \varepsilon_{n,k}^2 \big| X^{(n)}, Y^{(n)} \big) \leq C_2 e^{-n \varepsilon_{n,k}^2 / C_2}.$$

The constants $C_i (i = 1, 2)$ do not depend on k.

Proof. We only sketch the proof. For the entropy condition, we claim that for any $g \in \mathcal{F}_k$,

$$\log \mathcal{N}(c_5 \varepsilon, \{f \in \mathcal{F}_k : L_2(f, g) \leq 2\varepsilon\}, L_2) \leq C_{c_5} k \log(e N_k).$$

To see this, the entropy can be bounded by

$$\log \left[\binom{N_k}{k} \max_{I \subset \{1, \ldots, N_k\}, |I| = k} \mathcal{N}(c_5 \varepsilon, \{f \in \mathcal{F}_{k,I} : L_2(f, g) \leq 2\varepsilon\}, L_2) \right],$$

where $\mathcal{F}_{k,I} \equiv \{f = \sum_{\ell_j \in I} a_{\ell_j} \phi_{\ell_j}\}$. Now we may use the standard entropy bound for Euclidean balls to conclude. The sufficient mass condition can be checked along similar lines as many examples before, by using $f_{0,k} \in \mathcal{F}_k$ as the best linear approximation amongst $\{\sum_{\ell_j \in I} a_{\ell_j} \phi_{\ell_j} : I \subset \{1, \ldots, N_k\}, |I| = k\}$, and $\delta_{n,k}^2 \equiv C k \log(N_k \vee n)/n$ for a large enough constant $C > 0$. □

It is straightforward from here to compute a more concrete contraction rate by specifying concrete orders of $\boldsymbol{\gamma}, \boldsymbol{N}$. Details are omitted.

The above proposition holds for a pre-specified $\boldsymbol{N}$. Let us now consider 'adaptation' problem with respect to $\boldsymbol{N}$. We will consider this in the framework of Corollary 2 of [Yan99]. Let $\boldsymbol{N}^{(1)} \equiv (N_1^{(1)}, N_2^{(1)}, \ldots)$ and $\boldsymbol{N}^{(2)} \equiv (N_1^{(2)}, N_2^{(2)}, \ldots)$

where $N_k^{(2)} \geq N_k^{(1)}$ and $\log N_k^{(i)} \lesssim \log k$. In this case, we may formulate formally two models: $\tilde{\mathcal{F}}_1 \equiv \mathcal{S}(\boldsymbol{\gamma}, \boldsymbol{N}^{(1)})$ and $\tilde{\mathcal{F}}_2 \equiv \mathcal{S}(\boldsymbol{\gamma}, \boldsymbol{N}^{(1)}) \cup \mathcal{S}(\boldsymbol{\gamma}, \boldsymbol{N}^{(2)})$, and we put a uniform prior on the index $\{1, 2\}$. The prior $\tilde{\Pi}_i$ on $\tilde{\mathcal{F}}_i$ is given by $\sum_k \lambda_n(k)\Pi_{n,k}(\boldsymbol{N}^{(i)})$ as specified above in (F.6) and satisfies the conditions in the proceeding proposition (so the prior on $\tilde{\mathcal{F}}_2$ only charges mass on $\mathcal{S}(\boldsymbol{\gamma}, \boldsymbol{N}^{(2)})$). Let $\gamma_k = k^{-\alpha}$ for $\alpha > 0$.

Proposition F.7. *Consider the above setup. Let $f_0 \in L_2([0,1])$ be such that $\sup_k |\int f_0 \phi_k| < \infty$. Then with $\varepsilon_{n,\alpha}^2 \equiv (\log n/n)^{2\alpha/(2\alpha+1)}$, for $f \in \mathcal{S}(\boldsymbol{\gamma}, \boldsymbol{N}^{(1)}) \cup \mathcal{S}(\boldsymbol{\gamma}, \boldsymbol{N}^{(2)})$,*

$$P_{f_0}^{(n)} \Pi_n\big(f : L_2^2(f, f_0) > C_1 \varepsilon_{n,\alpha}^2 \big| X^{(n)}, Y^{(n)}\big) \leq C_2 e^{-n\varepsilon_{n,\alpha}^2/C_2}$$

holds for n large enough.

Proof. Let $f_{0,i}$ be the best linear approximation amongst $\{\sum_{\ell_j \in I} a_{\ell_j} \phi_{\ell_j} : I \subset \{1, \ldots, N_{k_n}^{(i)}\}, |I| = k_n\}$, so $\|f_0 - f_{0,i}\|_{L_2([0,1])}^2 \leq \gamma_{k_n}^2 (i = 1, 2)$, where $k_n = (n/\log n)^{1/(1+2\alpha)}$. In particular, write $f_{0,i} = \sum_{\ell_j^{(i)} \in I^{(i)}} a_{\ell_j^{(i)}} \phi_{\ell_j^{(i)}}$. Using the result on page 1586 of [YB99], $\log \mathcal{N}(c_5\varepsilon, \tilde{\mathcal{F}}_i, L_2) \lesssim \varepsilon^{-1/\alpha} \log(1/\varepsilon)$. So we may take $\delta_{n,i}^2 \equiv C(n/\log n)^{-2\alpha/(2\alpha+1)}$ for $i = 1, 2$ and a large constant $C > 0$. To verify the sufficient mass condition, note that

$$\begin{aligned}
&\tilde{\Pi}_i(\{f \in \tilde{\mathcal{F}}_i : L_2^2(f, f_{0,i}) \leq \delta_{n,i}^2/c_3\}) \\
&\geq \lambda_n(k_n) \cdot \binom{N_{k_n}}{k_n}^{-1} \big(\delta_{n,i}/\sqrt{c_3}\big)^{k_n} g\big(\sup_k \Big|\int f_0 \phi_k\Big| + 1\big)^{k_n} \geq e^{-2n\delta_{n,k_n}^2}
\end{aligned}$$

by choosing $C > 0$ large enough. □

The proposition shows that under the specified prior, it is indeed possible to achieve adaptive rate over $\mathcal{S}(\boldsymbol{\gamma}, \boldsymbol{N}^{(1)}) \cup \mathcal{S}(\boldsymbol{\gamma}, \boldsymbol{N}^{(2)})$. It is straightforward to extend this result to multiple lists of models so we omit the details.

Acknowledgments

The author is indebted to Chao Gao, Johannes Schmidt-Hieber, and the anonymous referees for numerous comments and suggestions that lead to substantial improvements of an earlier version of the paper. The author would also like to thank Jon Wellner for constant support and continuous encouragement as this work developed.

References

[ACCR14] Pierre Alquier, Vincent Cottet, Nicolas Chopin, and Judith Rousseau, *Bayesian matrix completion: prior specification*, arXiv preprint 1406.1440 (2014).

[AGR13] Julyan Arbel, Ghislaine Gayraud, and Judith Rousseau, *Bayesian optimal adaptive estimation using a sieve prior*, Scand. J. Stat. **40** (2013), no. 3, 549–570. MR3091697

[BBM99] Andrew Barron, Lucien Birgé, and Pascal Massart, *Risk bounds for model selection via penalization*, Probab. Theory Related Fields **113** (1999), no. 3, 301–413. MR1679028 (2000k:62049)

[BC91] Andrew R. Barron and Thomas M. Cover, *Minimum complexity density estimation*, IEEE Trans. Inform. Theory **37** (1991), no. 4, 1034–1054. MR1111806

[Bel17] Eduard Belitser, *On coverage and local radial rates of credible sets*, Ann. Statist. **45** (2017), no. 3, 1124–1151. MR3662450

[BG03] Eduard Belitser and Subhashis Ghosal, *Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution*, Ann. Statist. **31** (2003), no. 2, 536–559, Dedicated to the memory of Herbert E. Robbins. MR1983541

[BG14] Sayantan Banerjee and Subhashis Ghosal, *Posterior convergence rates for estimating large precision matrices using graphical models*, Electron. J. Stat. **8** (2014), no. 2, 2111–2137. MR3273620

[BG15] Sayantan Banerjee and Subhashis Ghosal, *Bayesian structure learning in graphical models*, J. Multivariate Anal. **136** (2015), 147–162. MR3321485

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart, *Concentration inequalities: A nonasymptotic theory of independence*, Oxford University Press, Oxford, 2013. MR3185193

[BM93] Lucien Birgé and Pascal Massart, *Rates of convergence for minimum contrast estimators*, Probab. Theory Related Fields **97** (1993), no. 1-2, 113–150. MR1240719 (94m:62095)

[BvdG11] Peter Bühlmann and Sara van de Geer, *Statistics for high-dimensional data*, Springer Series in Statistics, Springer, Heidelberg, 2011, Methods, theory and applications. MR2807761

[Cas14] Ismaël Castillo, *On Bayesian supremum norm contraction rates*, Ann. Statist. **42** (2014), no. 5, 2058–2091. MR3262477

[CGR04] Nidhan Choudhuri, Subhashis Ghosal, and Anindya Roy, *Bayesian estimation of the spectral density of a time series*, J. Amer. Statist. Assoc. **99** (2004), no. 468, 1050–1059. MR2109494

[CGS15] Sabyasachi Chatterjee, Adityanand Guntuboyina, and Bodhisattva Sen, *On risk bounds in isotonic and other shape restricted regression problems*, Ann. Statist. **43** (2015), no. 4, 1774–1800. MR3357878

[CP11] Emmanuel J. Candès and Yaniv Plan, *Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements*, IEEE Trans. Inform. Theory **57** (2011), no. 4, 2342–2359. MR2809094

[CSHvdV15] Ismaël Castillo, Johannes Schmidt-Hieber, and Aad van der Vaart, *Bayesian linear regression with sparse priors*, Ann. Statist. **43** (2015), no. 5, 1986–2018. MR3375874

[CT05] Emmanuel J. Candès and Terence Tao, *Decoding by linear programming*, IEEE Trans. Inform. Theory **51** (2005), no. 12, 4203–4215. MR2243152

[CvdV12] Ismaël Castillo and Aad van der Vaart, *Needles and straw in a haystack: posterior concentration for possibly sparse sequences*, Ann. Statist. **40** (2012), no. 4, 2069–2101. MR3059077

[dJvZ10] R. de Jonge and J. H. van Zanten, *Adaptive nonparametric Bayesian inference using location-scale mixture priors*, Ann. Statist. **38** (2010), no. 6, 3300–3320. MR2766853

[DRRS18] Sophie Donnet, Vincent Rivoirard, Judith Rousseau, and Catia Scricciolo, *Posterior concentration rates for empirical Bayes procedures with applications to Dirichlet process mixtures*, Bernoulli **24** (2018), no. 1, 231–256. MR3706755

[GGvdV00] Subhashis Ghosal, Jayanta K. Ghosh, and Aad van der Vaart, *Convergence rates of posterior distributions*, Ann. Statist. **28** (2000), no. 2, 500–531. MR1790007

[GLvdV08] Subhashis Ghosal, Jüri Lember, and Aad van der Vaart, *Nonparametric Bayesian model selection and averaging*, Electron. J. Stat. **2** (2008), 63–89. MR2386086

[GS13] Adityanand Guntuboyina and Bodhisattva Sen, *Covering numbers for convex functions*, IEEE Trans. Inform. Theory **59** (2013), no. 4, 1957–1965. MR3043776

[Gun12] Adityanand Guntuboyina, *Optimal rates of convergence for convex set estimation from support functions*, Ann. Statist. **40** (2012), no. 1, 385–411. MR3014311

[GvdV01] Subhashis Ghosal and Aad W. van der Vaart, *Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities*, Ann. Statist. **29** (2001), no. 5, 1233–1263. MR1873329

[GvdV07a] Subhashis Ghosal and Aad van der Vaart, *Convergence rates of posterior distributions for non-i.i.d. observations*, Ann. Statist. **35** (2007), no. 1, 192–223. MR2332274

[GvdV07b] Subhashis Ghosal and Aad van der Vaart, *Posterior convergence rates of Dirichlet mixtures at smooth densities*, Ann. Statist. **35** (2007), no. 2, 697–723. MR2336864

[GvdV17] Subhashis Ghosal and Aad van der Vaart, *Fundamentals of nonparametric Bayesian inference*, Cambridge Series in Statistical and Probabilistic Mathematics, vol. 44, Cambridge University Press, Cambridge, 2017. MR3587782

[GvdVZ15] Chao Gao, Aad van der Vaart, and Harrison H Zhou, *A general framework for bayes structured linear models*, arXiv preprint 1506.02174 (2015).

[GZ15] Chao Gao and Harrison H. Zhou, *Rate-optimal posterior contraction for sparse PCA*, Ann. Statist. **43** (2015), no. 2, 785–818. MR3325710

[GZ16] Chao Gao and Harrison H. Zhou, *Rate exact Bayesian adaptation* *with modified block priors*, Ann. Statist. **44** (2016), no. 1, 318–345. MR3449770

[HD11] Lauren A Hannah and David B Dunson, *Bayesian nonparametric multivariate convex regression*, arXiv preprint 1109.0322 (2011).

[HH03] CC Holmes and NA Heard, *Generalized monotonic regression using random change points*, Statistics in Medicine **22** (2003), no. 4, 623–638.

[HRSH15] Marc Hoffmann, Judith Rousseau, and Johannes Schmidt-Hieber, *On adaptive posterior concentration rates*, Ann. Statist. **43** (2015), no. 5, 2259–2295. MR3396985

[HW16] Qiyang Han and Jon A. Wellner, *Multivariate convex regression: global risk bounds and adaptation*, arXiv preprint 1601.06844 (2016).

[KRvdV10] Willem Kruijer, Judith Rousseau, and Aad van der Vaart, *Adaptive Bayesian density estimation with location-scale mixtures*, Electron. J. Stat. **4** (2010), 1225–1257. MR2735885

[KvdV06] B. J. K. Kleijn and Aad van der Vaart, *Misspecification in infinite-dimensional Bayesian statistics*, Ann. Statist. **34** (2006), no. 2, 837–877. MR2283395

[LD14] Lizhen Lin and David B. Dunson, *Bayesian monotone regression using Gaussian process projection*, Biometrika **101** (2014), no. 2, 303–317. MR3215349

[LG17] Meng Li and Subhashis Ghosal, *Bayesian detection of image boundaries*, Ann. Statist. **45** (2017), no. 5, 2190–2217. MR3718166

[LvdV07] Jüri Lember and Aad van der Vaart, *On universal Bayesian adaptation*, Statist. Decisions **25** (2007), no. 2, 127–152. MR2388859

[MA15] The Tien Mai and Pierre Alquier, *A Bayesian approach for noisy matrix completion: optimal rate under general sampling distribution*, Electron. J. Stat. **9** (2015), no. 1, 823–841. MR3331862

[Mas07] Pascal Massart, *Concentration inequalities and model selection*, Lecture Notes in Mathematics, vol. 1896, Springer, Berlin, 2007, Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard. MR2319879 (2010a:62008)

[MR13] Alexander Meister and Markus Reiß, *Asymptotic equivalence for nonparametric regression with non-regular errors*, Probab. Theory Related Fields **155** (2013), no. 1-2, 201–229. MR3010397

[MRS20] Ester Mariucci, Kolyan Ray, and Botond Szabó, *A Bayesian nonparametric approach to log-concave density estimation*, Bernoulli **26** (2020), no. 2, 1070–1097. MR4058361

[ND04] Brian Neelon and David B. Dunson, *Bayesian isotonic regression and trend analysis*, Biometrics **60** (2004), no. 2, 398–406. MR2066274

[PBPD14] Debdeep Pati, Anirban Bhattacharya, Natesh S. Pillai, and David Dunson, *Posterior contraction in sparse Bayesian factor models for massive covariance matrices*, Ann. Statist. **42** (2014), no. 3, 1102–1130. MR3210997

[Pol90] David Pollard, *Empirical processes: theory and applications*, NSF-CBMS Regional Conference Series in Probability and Statistics, 2, Institute of Mathematical Statistics, Hayward, CA; American Statistical Association, Alexandria, VA, 1990. MR1089429 (93e:60046)

[RCL12] Judith Rousseau, Nicolas Chopin, and Brunero Liseo, *Bayesian nonparametric estimation of the spectral density of a long or intermediate memory Gaussian process*, Ann. Statist. **40** (2012), no. 2, 964–995. MR2985940

[RFP10] Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo, *Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization*, SIAM Rev. **52** (2010), no. 3, 471–501. MR2680543

[Rou10] Judith Rousseau, *Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density*, Ann. Statist. **38** (2010), no. 1, 146–180. MR2589319

[RS16] Judith Rousseau and Botond Szabo, *Asymptotic frequentist coverage properties of bayesian credible sets for sieve priors in general settings*, arXiv preprint 1609.05067 (2016).

[RS17] Judith Rousseau and Botond Szabo, *Asymptotic behaviour of the empirical Bayes posteriors associated to maximum marginal likelihood estimator*, Ann. Statist. **45** (2017), no. 2, 833–865. MR3650402

[RSH17] Markus Reiss and Johannes Schmidt-Hieber, *Nonparametric bayesian analysis for support boundary recovery*, arXiv preprint 1703.08358 (2017).

[RT11] Angelika Rohde and Alexandre B. Tsybakov, *Estimation of high-dimensional low-rank matrices*, Ann. Statist. **39** (2011), no. 2, 887–930. MR2816342

[Sal14] Jean-Bernard Salomond, *Adaptive Bayes test for monotonicity*, The contribution of young researchers to Bayesian statistics, Springer Proc. Math. Stat., vol. 63, Springer, Cham, 2014, pp. 29–33. MR3133254

[Scr06] Catia Scricciolo, *Convergence rates for Bayesian density estimation of infinite-dimensional exponential families*, Ann. Statist. **34** (2006), no. 6, 2897–2920. MR2329472

[Scr16] Catia Scricciolo, *Rates for Bayesian estimation of location-scale mixtures of super-smooth densities*, Topics in theoretical and applied statistics, Stud. Theor. Appl. Stat. Sel. Papers Stat. Soc., Springer, Cham, 2016, pp. 49–57. MR3838069

[SSW09] Thomas S. Shively, Thomas W. Sager, and Stephen G. Walker, *A Bayesian approach to non-parametric monotone function estimation*, J. R. Stat. Soc. Ser. B Stat. Methodol. **71** (2009), no. 1, 159–175. MR2655528

[SW01] Xiaotong Shen and Larry Wasserman, *Rates of convergence of* *posterior distributions*, Ann. Statist. **29** (2001), no. 3, 687–714. MR1865337

[Tsy14] Alexandre B Tsybakov, *Aggregation and minimax optimality in high-dimensional estimation*, Proceedings of the International Congress of Mathematicians, 2014, pp. 225–246.

[vdG00] Sara van de Geer, *Applications of Empirical Process Theory*, Cambridge Series in Statistical and Probabilistic Mathematics, vol. 6, Cambridge University Press, Cambridge, 2000. MR1739079 (2001h:62002)

[vdVvZ08] Aad van der Vaart and J. H. van Zanten, *Rates of contraction of posterior distributions based on Gaussian process priors*, Ann. Statist. **36** (2008), no. 3, 1435–1463. MR2418663

[vdVvZ09] Aad van der Vaart and J. H. van Zanten, *Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth*, Ann. Statist. **37** (2009), no. 5B, 2655–2675. MR2541442

[vdVW96] Aad van der Vaart and Jon A. Wellner, *Weak Convergence and Empirical Processes*, Springer Series in Statistics, Springer-Verlag, New York, 1996. MR1385671 (97g:60035)

[Yan99] Yuhong Yang, *Model selection for nonparametric regression*, Statist. Sinica **9** (1999), no. 2, 475–499. MR1707850

[YB98] Yuhong Yang and Andrew R. Barron, *An asymptotic property of model selection criteria*, IEEE Trans. Inform. Theory **44** (1998), no. 1, 95–116. MR1486651

[YB99] Yuhong Yang and Andrew Barron, *Information-theoretic determination of minimax rates of convergence*, Ann. Statist. **27** (1999), no. 5, 1564–1599. MR1742500

[YG16] William Weimin Yoo and Subhashis Ghosal, *Supremum norm posterior contraction and credible sets for nonparametric multivariate regression*, Ann. Statist. **44** (2016), no. 3, 1069–1102. MR3485954

[YLC19] Zhuqing Yu, Michael Levine, and Guang Cheng, *Minimax optimal estimation in partially linear additive models under high dimension*, Bernoulli **25** (2019), no. 2, 1289–1325. MR3920373

[YP17] Yun Yang and Debdeep Pati, *Bayesian model selection consistency and oracle inequality with intractable marginal likelihood*, arXiv preprint 1701.00311 (2017).

[YZ16] Ming Yuan and Ding-Xuan Zhou, *Minimax optimal rates of estimation in high dimensional additive models*, Ann. Statist. **44** (2016), no. 6, 2564–2593. MR3576554