BAYES MODEL SELECTION

QIYANG HAN

ABSTRACT. We offer a general Bayes theoretic framework to tackle the model selection problem under a two-step 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.

We establish non-asymptotic oracle posterior contraction rates under (i) a new Bernstein-inequality 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 resulting posterior mean also satisfies an oracle inequality, thus automatically serving as an adaptive point estimator in a frequentist sense. Model mis-specification is allowed in these oracle rates.

The new Bernstein-inequality condition not only eliminates the convention of constructing explicit tests with exponentially small type I and II errors, but also suggests the intrinsic metric to use in a given statistical experiment, both as a loss function and as an entropy measurement. This gives a unified reduction scheme for many experiments considered in [23] and beyond. As an illustration for the scope of our general results in concrete applications, we consider (i) trace regression, (ii) shape-restricted isotonic/convex regression, (iii) high-dimensional partially linear regression and (iv) covariance matrix estimation in the sparse factor 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 multi-structured experiment.

1. INTRODUCTION

1.1. Overview. Suppose we observe $X^{(n)}$ from a statistical experiment $(X^{(n)}, A^{(n)}, P^{(n)}_f)$, where $f$ belongs to a statistical model $F$ and $\{P^{(n)}_f\}_{f \in F}$ is dominated by a $\sigma$-finite measure $\mu$. Instead of using a single ‘big’ model $F$, a collection of (sub-)models $\{F_m\}_{m \in \mathcal{I}} \subset F$ are available to statisticians, and the art of model selection is to determine which one(s) to use.

Date: May 18, 2017.

2000 Mathematics Subject Classification. 60F17, 62E17.

Key words and phrases. Bayes nonparametrics, model selection, adaptive estimation, model mis-specification, Bernstein inequality.

Supported in part by NSF Grant DMS-1566514.
There are vast literatures on model selection from a frequentist point of view; we only refer the reader to [4, 36, 11, 44] as some representative pointers for various approaches of penalization, aggregation, etc. On the other hand, from a Bayes point of view, although posterior contraction rates have been derived for many different models (see e.g. [21, 43, 23, 16, 14, 15, 42, 45, 46, 28] for some key contributions), understanding towards general Bayes model selection procedures has been limited. [22] focused on designing adaptive Bayes procedures with models primarily indexed by the smoothness level of classical function classes in the context of density estimation. Their conditions are complicated and seem not directly applicable to other settings. [18] designed a prior specific to structured linear problems in the Gaussian regression model, with their main focus on high-dimensional (linear) and network problems. It seems non-trivial for their framework to handle other non-linear models.

Despite these limitations, [22, 18] give useful clues. One common feature in these papers is a two-step prior design, where the first-step prior \( \Lambda_n \) assesses the model selection uncertainty, followed by a second-step prior \( \Pi_{n,m} \) quantifying the prior belief in the strength of the signals within the specific chosen model \( F_m \) from the first step. Such a prior design is intrinsic in many proposals for different problems, e.g. [16, 15] for sparse linear regression, [2] for trace regression, [29, 27] for shape restricted regression, [19, 38] for problems related to covariance matrix estimation.

This is the starting point of this paper. We give a unified theoretical treatment to this two-step prior design by identifying common structural assumptions on the statistical experiments \((X^{(n)}, A^{(n)}, P^{(n)})\), the collection of models \( \{F_m\} \) and the priors \( \{\Lambda_n\} \) and \( \{\Pi_{n,m}\} \) such that the posterior distribution both

\[(G1) \text{ contracts at an oracle rate with respect to some metric } d_n; \]

\[
\inf_{m \in I} \left( \inf_{g \in F_m} d_n^2(f_0, g) + \text{pen}(m) \right),
\]

where \( \text{pen}(m) \)\(^1\) is related to the ‘dimension’ of \( F_m \), and

\[(G2) \text{ concentrates on the model } F_{m*}, \text{ where } m* \text{ is the ‘best’ model balancing the bias-variance tradeoff in (1.1)}.\]

The oracle formulation (1.1) follows the convention in the frequentist literature [36, 44], and has several advantages: (i) (minimaxity) if the true signal \( f_0 \) can be well-approximated by the models \( \{F_m\} \), the contraction rate in (1.1) is usually (nearly) minimax optimal, (ii) (adaptivity) if \( f_0 \) lies in certain low-dimensional model \( F_m \), the contraction rate adapts to this unknown information, and (iii) (mis-specification) if the models \( F_m \) are mis-specified while \( d_n^2(f_0, \cup_{m \in I} F_m) \) remains ‘small’, then the contraction rate should still be rescued by this relatively ‘small’ bias.

\(^1\)\( \text{pen}(m) \) may depend on \( n \) but we suppress this dependence for notational convenience.
As the main abstract result of this paper (cf. Theorem 1), we show that our goals (G1)-(G2) can be accomplished under:

(i) **(Experiment)** a Bernstein-inequality 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 $\Lambda_n$, 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 $f_0$.

One important ingredient in studying posterior contraction rates in Bayes nonparametrics literature has been the construction of appropriate tests with exponentially small type I and II errors with respect to certain metric [21, 23]. Such tests date back to the work of Le Cam [31, 32, 33] and Birgé [6, 7, 8], who brought out the special role of the Hellinger metric in which tests can be constructed generically. On the other hand, the testing framework [21, 23] requires the prior to spread sufficient mass near the Kullback-Leibler neighborhood of the true signal. The discrepancy of these two metrics can be rather delicate, particularly for non i.i.d. and complicated models, and it often remains unclear which metric is the natural one to use in these models. Moreover, it is usually a significant theoretical challenge to construct tests in complicated models, cf. [19, 38], to name a few.

Our Bernstein-inequality condition (i) closes these gaps by suggesting the usage of an ‘intrinsic metric’ that mimics the behavior of the Kullback-Leibler divergence in a given statistical experiment, in which a ‘good’ test can be constructed generically (cf. Lemma 1). Bernstein inequality is a fundamental tool in probability theory, and hence can be easily verified in many statistical experiments including various experiments considered in [23] and beyond: Gaussian/binary/poisson regression, density estimation, Gaussian autoregression, Gaussian time series and covariance matrix estimation problems. We identify the intrinsic metrics to use in these experiments. Furthermore, the Bernstein-inequality condition entails sharp exponential contraction of the posterior distribution near the ‘true’ signal, complementing a recent result of [28]. Results of this type typically do not follow directly from general principles in [21, 23], and have mainly been derived on a case-by-case basis, cf. [16, 19, 18]. As such, we provide a refinement of the seminal testing framework in [21, 23] so that the investigation of sharp posterior contraction rates in the intrinsic metric of an experiment essentially reduces to the study of prior design.

Conditions (ii) and (iii) are familiar in Bayes nonparametrics literature. In particular, the first-step prior can be designed generically (cf. Proposition 1). 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 $\mathcal{F}_m$. 
These conditions, albeit minimal, imply more than an optimal adaptive Bayes procedure in the sense of (G1)-(G2). In fact, we show that the posterior mean automatically serves as an adaptive point estimator in a frequentist sense. These results reveal, in a sense, that the task of constructing adaptive procedures with respect to the intrinsic metric in a given statistical experiment, in both frequentist and Bayes contexts, is not really harder than that of designing an optimal non-adaptive prior for each of the models.

A general theory would be less interesting without being able to address problems of different types. As an illustration of our general framework in concrete applications, we justify the prior proposals in (i) [2, 34] for the trace regression problem, and in (ii) [29, 27] for the shape-restricted regression problems. Despite many theoretical results for Bayes high-dimensional models (cf. [16, 15, 19, 18, 38, 3]), 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\footnote{Almost completed at the same time, [35] considered a Bayes approach for univariate log-concave density estimation, where they derived contraction rates without addressing the adaptation issue.}. We also consider adaptive Bayes procedures for the high-dimensional partially linear regression model and the covariance matrix estimation problem in the sparse factor model. These new results serve as an illustration of the generic construction scheme of a (nearly) minimax adaptive estimator in a complicated experiment with multiple structures. Some of these results improve the best known result in the literature.

During the preparation of this paper, we become aware of a very recent paper [48] who independently considered the Bayes model selection problem. Both our approach and [48] shed light on the general Bayes model selection problem, while differing in several important aspects (cf. Remark 2). Moreover, our work here applies to a wide range of applications that are not covered by [48].

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

1.3. Organization. Section 2 is devoted to the general model selection theory. We work out a wide range of experiments that fit into our general theory
in Section 3. Section 4 discusses various concrete applications as mentioned above. Most detailed proofs are deferred to Sections 5-6 and the Appendix.

2. General theory

In the two-step prior design framework, we first put a prior $\Lambda_n$ on the model index $\mathcal{I}$, followed by a prior $\Pi_{n,m}$ on the model $F_m$ chosen from the first step. The overall prior is a probability measure on $F$ 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 $F$: for a measurable subset $B \subset F$,

\[
\Pi_n(B | X^{(n)}) = \frac{\int_B p_f^{(n)}(X^{(n)}) \, d\Pi_n(f)}{\int p_f^{(n)}(X^{(n)}) \, d\Pi_n(f)}
\]

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

2.1. Assumptions. To state our assumption on the experiment, let

\[
\psi_{v,c}(\lambda) = v\lambda^2 \frac{2(1 - |c|\lambda)}{1 + |c|\lambda}, \quad 0 \leq |\lambda| < 1/c
\]

denote the ‘Bernstein’ function. This function plays a pivotal role in proving sub-gamma behavior of a given (complicated) random variable, cf. [9]. Here $v$ and $c$ are the $L_2$ size and $L_\infty$ size of the random variable controlling respectively the degree of its sub-Gaussian and sub-gamma behavior.

Assumption A (Experiment: Bernstein-inequality condition). There exist some absolute $c_1 > 0$ and $\kappa = (\kappa_g, \kappa_\Gamma) \in (0, \infty) \times [0, \infty)$ such that for all $n \in \mathbb{N}$, and $f_0, f_1 \in F$,

\[
P_{f_0}^{(n)} \exp \left[ \lambda \left( \log \frac{p_{f_0}^{(n)}}{p_{f_1}^{(n)}} - P_{f_0}^{(n)} \log \frac{p_{f_0}^{(n)}}{p_{f_1}^{(n)}} \right) \right] \leq c_1 \exp \left[ \psi_{\kappa_g d_n^2(f_0, f_1), \kappa_\Gamma}(\lambda) \right]
\]

holds for all $|\lambda| < 1/\kappa_\Gamma$. Here the metric $d_n : F \times F \to \mathbb{R}_{\geq 0}$ satisfies

\[
c_2 \cdot d_n^2(f_0, f_1) \leq \frac{1}{n} P_{f_0}^{(n)} \log \frac{p_{f_0}^{(n)}}{p_{f_1}^{(n)}} \leq c_3 \cdot d_n^2(f_0, f_1)
\]

for some absolute constants $c_2, c_3 > 0$.

In Assumption A, we require the log likelihood ratio to satisfy a Bernstein inequality. In particular, the log likelihood ratio has local Gaussian behavior. Conversely, if the log likelihood ratio behaves locally like Gaussian, then we can pick some $\kappa_\Gamma > 0$ so that the Bernstein inequality holds.

Lemma 1. Let Assumption A hold. Fix $f_0, f_1 \in F$, there exists some test $\phi_n$ such that

\[
\sup_{f \in F : d_n^2(f, f_0) \leq c_5 d_n^2(f_0, f_1)} \left( P_{f_0}^{(n)} \phi_n + P_f^{(n)} (1 - \phi_n) \right) \leq c_6 \exp(-c_7 n d_n^2(f_0, f_1))
\]
where \( c_5 \leq 1/4, c_6 \in [2, \infty), c_7 \in (0, 1) \) only depend on \( c_1, c_2, c_3, \kappa \).

This lemma suggests that under a Bernstein-inequality condition on the log likelihood ratio, tests exist automatically under the intrinsic metric \( d_n \) that mimics the behavior of the Kullback-Leibler divergence in the sense of (2.3). Several examples will be worked out in Section 3 to illustrate the choice of an intrinsic metric \( d_n \), including the discrete \( \ell_2 \) loss for regression models, a weighted \( L_2 \) metric for the Gaussian autoregression model, the Hellinger metric for density estimation, the Frobenius norm for covariance matrix estimation problem.

Next we state the assumption on the complexity of the models \( \{F_m\}_{m \in I} \). Let \( I = \mathbb{N}^q \) be a \( q \)-dimensional lattice with the natural order \((I, \leq)\).

The dimension \( q \) is understood as the number of different structures in the models \( \{F_m\}_{m \in I} \). In the sequel we will not explicitly mention \( q \) unless otherwise specified. We require the models to be nested in the sense that \( F_m \subset F_{m'} \) if \( m \leq m' \). Let \( f_{0,m} \) denote the ‘best’ approximation of \( f_0 \) within the model \( F_m \) in the sense that \( f_{0,m} \in \arg \inf_{g \in F_m} d_n(f_0, g) \).

**Assumption B (Models: Local entropy condition).** For each \( m \in I \),

\[
(2.4) \quad 1 + \sup_{\varepsilon > \delta_{n,m}} \log \mathcal{N}(c_5 \varepsilon, \{f \in F_m : d_n(f, g) \leq 2\varepsilon \}, d_n) \leq (c_7/2) n \delta_{n,m}^2
\]

holds for all \( g \in \{f_{0,m'}\}_{m' \leq m} \). Furthermore there exist absolute constants \( c \geq 1 \) and \( \gamma \geq 1 \) such that for any \( m \in I \), \( \alpha \geq c \gamma / 2 \) and any \( h \geq 1 \),

\[
(2.5) \quad \sum_{m' \geq h m} e^{-\alpha n \delta_{n,m'}^2} \leq 2 e^{-\alpha n h \delta_{n,m}^2 / c^2}, \quad c^{-2} \delta_{n,hm}^2 \leq h^\gamma \delta_{n,m}^2.
\]

Note that if we choose all models \( F_m = F \), then (2.4) reduces to the local entropy condition in [21, 23]. When \( F_m \) is finite-dimensional, typically we can check (2.4) for all \( g \in F_m \). Now we comment on (2.5). The left side of (2.5) essentially requires super linearity of the map \( m \mapsto \delta_{n,m}^2 \), 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 \( c = 1 \) and \( \gamma = 1 \) when \( n \delta_{n,m}^2 = cm \) for some absolute constant \( c > 2 / c_7 \).

Finally we state assumptions on the priors.

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

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

\[
(2.6) \quad \lambda_{n,m} \geq \frac{1}{2} \exp(-2n \delta_{n,m}^2), \quad \sum_{k \geq hm, k \in I} \lambda_{n,k} \leq 2 \exp(-n \delta_{n,m}^2).
\]

- (P2) (Second-step prior)

\[
(2.7) \quad \Pi_{n,m} \left\{ \{f \in F_m : d_n^2(f, f_{0,m}) \leq \delta_{n,m}^2 / c_3 \} \right\} \geq \exp(-2n \delta_{n,m}^2).
\]

\(^3\)For any \( a, b \in I \), \( a \leq b \) iff \( a_i \leq b_i \) for all \( 1 \leq i \leq q \). Similar definition applies to \( <, >, \geq, \leq \).

\(^4\)We assume that \( f_{0,m} \) is well-defined without loss of generality.
Condition (P1) can be verified by using the following generic prior \( \Lambda_n \):

\[
\lambda_{n,m} \propto \exp(-2n\delta_{n,m}^2).
\]

**Proposition 1.** Suppose the first condition of (2.5) holds. Then (P1) in Assumption C holds for the prior (2.8) with \( h \geq 2c^2 \).

(2.8) will be the model selection (first-step) prior on the model index \( \mathcal{I} \) in all examples in Section 4.

Condition (P2) is reminiscent of the classical prior mass condition considered in [21, 23] where \( \delta_{n,m}^2 \) is understood as the ‘posterior contraction rate’ for the model \( \mathcal{F}_m \). Hence (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 [21, 23] uses more complicated metric balls induced by higher moments of the Kullback-Leibler divergence.

### 2.2. Main results.

The following is the main abstract result of this paper.

**Theorem 1.** Suppose Assumptions A-C hold for some \( \mathcal{M} \subset \mathcal{I} \) with \( |\mathcal{M}| = \infty \), and \( h \geq C_0c^2 \). Let \( \varepsilon_{n,m}^2 = \inf_{g \in \mathcal{F}_m} d_n^2(f_0,g) \vee \delta_{n,m}^2 \).

1. For any \( m \in \mathcal{M} \),

\[
P_{f_0}^{(n)} \Pi_0(f \in \mathcal{F} : d_n^2(f, f_0) > C_1 \left( \inf_{g \in \mathcal{F}_m} d_n^2(f_0, g) + \delta_{n,m}^2 \right) | X^{(n)}) \leq C_2 \exp \left( -n \varepsilon_{n,m}^2 / C_2 \right).
\]

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

\[
P_{f_0}^{(n)} \Pi_0(f \notin \mathcal{F}_{C_m} | X^{(n)}) \leq C_2 \exp \left( -n \varepsilon_{n,m}^2 / C_2 \right).
\]

3. Let \( \hat{f}_n \equiv \Pi_0(f | X^{(n)}) \) be the posterior mean. Then

\[
P_{f_0}^{(n)} d_n^2(\hat{f}_n, f_0) \leq C_4 \inf_{m \in \mathcal{M}} \left( \inf_{g \in \mathcal{F}_m} d_n^2(f_0, g) + \delta_{n,m}^2 \right).
\]

Here the constant \( C_0 \) depends on \( \{c_i\}_{i=1}^{3} , \kappa \) and \( \{C_1\}_{i=1}^{4} \) depend on the \( \{c_i\}_{i=1}^{3} , \kappa, c, h \) and \( \gamma \).

The main message of Theorem 1 is 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 non-adaptive prior for each model. Furthermore, the resulting posterior mean serves as an automatic adaptive point estimator in a frequentist sense. In particular, if the non-adaptive priors we use on each model lead to (nearly) optimal posterior contraction rates on these models, adaptation happens automatically by designing a ‘correct’ model selection prior, e.g. (2.8).

Besides being rate-adaptive to the collection of models, (2.10) shows that the posterior distribution concentrates on the model \( \mathcal{F}_m \) that balances the bias and variance tradeoff in the oracle rates (2.9) and (2.11). Results of this type have been derived primarily in the Gaussian regression model (cf.
and in density estimation [22]; here our result shows that this
is a general phenomenon for the two-step prior design.

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 [18] who mainly focuses on structured linear mod-
els in the Gaussian regression setting, and [30] who pursued generality at
the cost of harder-to-check conditions. The condition $|\mathcal{M}| = \infty$ is as-
sumed purely for technical convenience. If we have finitely many models
$\{\mathcal{F}_1, \ldots, \mathcal{F}_{m'}\}$ at hand, then we can define $\mathcal{F}_m \equiv \mathcal{F}_{m'}$ for $m \geq m'$ so that
this condition is satisfied.

Remark 1. We make some technical remarks.

1. The probability estimate in (2.9) is sharp (up to constants) in view
of the lower bound result Theorem 2.1 in [28], thus closing a gap that
has not been attainable in a general setting by using [21, 23] directly.
Beyond of theoretical interest in its own right, the sharp estimate helps us to derive an oracle inequality for the posterior mean as an
important frequentist summary of the posterior distribution. Such
sharp estimates have been derived separately in different models, e.g.
the sparse normal mean model [16], the sparse PCA model [19], and
the structured linear model [18], to name a few.

2. Assumption A implies, among other things, the existence of a good
test (cf. Lemma 1). In this sense our approach here falls into the
general testing approach adopted in [21, 23]. The testing approach
has difficulties in handling non-intrinsic metrics, cf. [28]. Some
alternative approaches for dealing with non-intrinsic metrics can be
found in [14, 28, 49].

3. The constants $\{C_i\}_{i=1}^4$ in Theorem 1 depend at most polynomially
with respect to the constants involved in Assumption A. This al-
allows some flexibility in the choice of the constants therein. In fact,
Bernstein inequality in some dependent cases comes with logarithmic
factors in $n$, cf. [1, 37].

Remark 2. We compare our results with Theorems 4 and 5 of [48]. Both
their results and our Theorem 1 shed light on the general problem of Bayes
model selection, while differing in several important aspects:

1. Our Theorem 1 hinges on the new Bernstein-inequality condition,
while the results of [48] are based on the classical mechanism of [23]
which requires the construction of tests. Some merits of our approach
will be clear from Section 3 and (2) below along with Remark 1.

2. The probability estimate in [48] for the posterior distribution outside
a ball of radius at the targeted contraction rate is asymptotic in na-
ture, while our Theorem 1 provides non-asymptotic sharp estimates.

3. Theorem 4 of [48] targets at exact model selection consistency, under
a set of additional ‘separation’ assumptions. Our Theorem 1 (2) re-
quires no extra assumptions, and shows the concentration behavior
of the posterior distribution on the ‘best’ model that balances the bias-variance tradeoff. This is significant in non-parametric problems: the true signal typically need not belong to any specific model.

(4) Theorem 5 of \[48\] contains a term involving the cardinality of the models and hence the models need be \textit{apriori} finitely many for their bound to be finite. It remains open to see if this can be removed.

2.3. Proof sketch. Here we sketch the main steps in the proof of our main abstract result Theorem 1. The details will be deferred to Section 5. The proof can be roughly divided into two main steps.

**Step 1** We first solve a localized problem on the model \(F_m\) 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 1:

\[
(2.12) \quad P_{f_0,m}^{(n)} \left( f \in F : d_n^2(f, f_0,m) > M \delta_{n,\tilde{m}}^2 | X^{(n)} \right) \lesssim \exp \left( -c_1 n \delta_{n,\tilde{m}}^2 \right),
\]

where \(\tilde{m}\) is the smallest index \(\geq m\) such that \(\delta_{n,\tilde{m}}^2 > \ell_n^2(f_0, f_0,m)\). 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 \(O(\exp(c_2 n \delta_{n,\tilde{m}}^2))\) factor in the probability bound (2.12), cf. Lemma 8, which is made possible by the Bernstein-inequality Assumption A. Then by requiring \(c_1 \gg c_2\) we obtain the conclusion by the definition of \(\delta_{n,\tilde{m}}^2\) and the fact that \(\delta_{n,\tilde{m}}^2 \approx \ell_n^2(f_0, f_0,m) \lor \delta_{n,m}^2\).

The existence of tests (Lemma 1) is used in step 1. Step 2 is inspired by the work of \([17, 5]\) 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, \([16, 15]\) 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. Statistical experiments

In this section we work out a couple of specific statistical experiments that satisfy Bernstein-inequality Assumption A to illustrate the scope of the general theory in Section 2. Some of the examples come from \([23]\); we identify the ‘intrinsic’ metric to use in these examples. Since Bernstein inequality is a fundamental probabilistic tool, and has been derived in a wide range of complicated (dependent) settings (\([1, 37]\)), we expect many more experiments to be covered beyond the ones we present here.

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 regression models: for \(1 \leq i \leq n\),

1. **(Gaussian)** \(X_i = \theta_i + \varepsilon_i\) where \(\varepsilon_i\)'s are i.i.d. \(\mathcal{N}(0, 1)\) and \(\Theta_n \subset \mathbb{R}^n\);
2. **(Binary)** \(X_i \sim\text{i.i.d. Bern}(\theta_i)\) where \(\Theta_n \subset [\eta, 1 - \eta]^n\) for some \(\eta > 0\);
(3) (Poisson) \( X_i \sim_{i.i.d.} \text{Poisson}(\theta_i) \) where \( \Theta_n \subset [1/M, M]^n \) for some \( M \geq 1 \);

We will use the following metric: for any \( \theta_0, \theta_1 \in \Theta_n 
\)
\[ \ell_n^2(\theta_0, \theta_1) \equiv \frac{1}{n} \sum_{i=1}^n (\theta_{0,i} - \theta_{1,i})^2. \]

**Lemma 2.** Assumption A holds for \( \ell_n \) with

1. (Gaussian) \( c_1 = c_2 = c_3 = \kappa \) and \( \kappa \) depend on \( n \) only;
2. (Binary) \( \kappa = 0 \) and the constants \( \{c_i\}_{i=1}^3, \kappa \) depend on \( \eta \) only;
3. (Poisson) constants \( \{c_i\}_{i=1}^3, \kappa \) depending on \( M \) only.

**Theorem 2.** For Gaussian/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.

### 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 \( \nu \) on the sample space \((\mathcal{X}, \mathcal{A})\).

We consider the following form of \( \mathcal{F} \): \( f(x) = \int e^{g(x) \nu} \) for some \( g \in \mathcal{G} \) for all \( x \in \mathcal{X} \). A natural metric to use for density estimation is the Hellinger metric: for any \( f_0, f_1 \in \mathcal{F} 
\)
\[ h^2(f_0, f_1) \equiv \frac{1}{2} \int_\mathcal{X} (\sqrt{f_0} - \sqrt{f_1})^2 d\nu. \]

**Lemma 3.** 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.

**Theorem 3.** 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.

### 3.3. Gaussian autoregression.

Suppose \( X_0, X_1, \ldots, X_n \) is generated from \( X_i = f(X_{i-1}) + \epsilon_i \) for \( 1 \leq i \leq n \), where \( f \) belongs to a function class \( \mathcal{F} \) with a uniform bound \( M \), and \( \epsilon_i \)'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 \( \phi \) is the normal density.

By the arguments on page 209 of [23], this chain has a unique stationary distribution with density \( q_f \) with respect to the Lebesgue measure \( \lambda \) on \( \mathbb{R} \).

We assume that \( X_0 \) is generated from this stationary distribution under the true \( f \). Consider the following metric: for any \( f_0, f_1 \in \mathcal{F} 
\)
\[ d_{r,M}^2(f_0, f_1) \equiv \int (f_0 - f_1)^2 r_M d\lambda \]
where \( r_M(x) \equiv \frac{1}{2} (\phi(x-M) + \phi(x+M)) \).

**Lemma 4.** 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.

**Theorem 4.** 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.
Compared with results obtained in [23] (cf. Section 7.4), we identify the intrinsic metric $d_{r,M}$ (a weighted $L_2$ norm) for the Gaussian autoregression model, while [23] uses a weighted $L_s(s > 2)$ norm to check the local entropy condition, and an average Hellinger metric as the loss function.

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 $(T_n(f))_{kl} = \int_{-\pi}^{\pi} e^{\sqrt{-1} \lambda (k-l)} f(\lambda) \, d\lambda$. We consider a special form of $\mathcal{F}$: $f = f_g \equiv \exp(g)$ for some $g \in \mathcal{G}$. We will use the following metric: for any $g_0, g_1 \in \mathcal{G}$,

$$D_n^2(g_0, g_1) = \frac{1}{n} \| T_n(f_{g_0}) - T_n(f_{g_1}) \|^2_F,$$

where $\| \cdot \|_F$ denotes the matrix Frobenius norm.

Lemma 5. 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.

Theorem 5. 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.

The metric $D_n$ can always be 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 [20]). Our result then shows that the metric to use in the entropy condition can be weakened to the usual $L_2$ norm rather than the much stronger $L_\infty$ norm as in page 202 of [23].

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 $\Sigma \in \mathcal{S}_p(L)$, the set of $p \times p$ covariance matrices whose minimal and maximal eigenvalues are bounded by $L^{-1}$ and $L$, respectively. We will use the Frobenius norm: for any $\Sigma_0, \Sigma_1 \in \mathcal{S}_p(L)$,

$$D_F^2(\Sigma_0, \Sigma_1) \equiv \| \Sigma_0 - \Sigma_1 \|^2_F.$$

Lemma 6. Under the above setting, Assumption A holds for the metric $D_F$ with constants $\{c_i\}_{i=1}^3, \kappa$ depending on $L$ only.

Theorem 6. For covariance matrix estimation in $\mathcal{S}_p(L)$ for some $L < \infty$, let $d_n \equiv D_F$. If Assumptions B-C hold, then (2.9)-(2.11) hold.

4. Applications

In this section, we consider concrete applications. As we have seen in previous sections, construction of adaptive Bayes procedures in the intrinsic metric of an experiment essentially reduces to the design of non-adaptive priors, and hence we only consider the simplest setup for a particular structure. For instance, once we understand how to analyze the convex Gaussian regression problem, we can similarly consider convex binary/Poisson regression, convex density estimation, Gaussian autoregression with convex functions, Gaussian time series with convex spectral density problems in their
respective intrinsic metrics. Hence our emphasis in the examples will be focused on the analysis of different model structures. Models that can be handled using similar techniques will not be presented in detail (e.g. Remark 4).

We will only explicitly state the corresponding oracle inequalities in the form of (2.9) for each example to be considered below. The corresponding results for (2.10) and (2.11) are omitted.

4.1. 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 \( f_A(x) = \text{tr}(x^\top A) \) for all \( x \in \mathcal{X} \equiv \mathbb{R}^{m_1 \times m_2} \). Let \( m \equiv m_1 \wedge m_2 \) and \( \bar{m} \equiv m_1 \vee m_2 \). The index set is \( \mathcal{I} = \mathcal{I}_1 \cup \mathcal{I}_2 \equiv \{1, \ldots, r_{\max}\} \cup \{r_{\max} + 1, \ldots\} \) where \( r_{\max} \leq \bar{m} \). For \( r \in \mathcal{I}_1 \), let \( \mathcal{F}_r \equiv \{ f_A : A \in \mathbb{R}^{m_1 \times m_2}, \text{rank}(A) \leq r \} \), and for \( r \in \mathcal{I}_2 \), \( \mathcal{F}_r \equiv \mathcal{F}_{r_{\max}} \).

Although various Bayesian methods have been proposed in the literature (cf. see [2] for a state-to-art summary), theoretical understanding has been limited. [34] derived an oracle inequality for an exponentially aggregated estimator for the matrix completion problem. Their result is purely frequentist. Below we consider a two step prior similar to [34, 2], 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. \( p = 1 \) and 2 correspond to the nuclear norm and the Frobenius norm respectively. To introduce the notion of RIP, let \( \mathcal{X} : \mathbb{R}^{m_1 \times m_2} \rightarrow \mathbb{R}^n \) be the linear map defined via \( A \mapsto (\text{tr}(x_i^\top A))_{i=1}^n \).

**Definition 1.** The linear map \( \mathcal{X} : \mathbb{R}^{m_1 \times m_2} \rightarrow \mathbb{R}^n \) is said to satisfy RIP \((r, \nu_r)\) for some \( 1 \leq r \leq r_{\max} \) and some \( \nu_r = (\nu_r, \bar{\nu}_r) \) with \( 0 < \nu_r \leq \bar{\nu}_r < \infty \) iff
\[
\nu_r \leq \frac{\|\mathcal{X}(A)\|_2}{\sqrt{\text{rank}(A)}} \leq \bar{\nu}_r \text{ holds for all matrices } A \in \mathbb{R}^{m_1 \times m_2} \text{ such that } \text{rank}(A) \leq r.
\]

For \( r > r_{\max} \), \( \mathcal{X} \) satisfies RIP \((r, \nu_r)\) iff \( \mathcal{X} \) satisfies RIP \((r_{\max}, \nu_r)\). Furthermore, \( \mathcal{X} : \mathbb{R}^{m_1 \times m_2} \rightarrow \mathbb{R}^n \) is said to satisfy uniform RIP \((\nu; \mathcal{I})\) on an index set \( \mathcal{I} \) iff \( \mathcal{X} \) satisfies RIP \((2r, \nu)\) for all \( r \in \mathcal{I} \).

RIP \((r, \nu_r)\) is a variant of the RIP condition introduced in [13, 12, 40] with scaling factors \( \bar{\nu}_r = 1/(1 - \delta_r) \) and \( \nu_r = 1/(1 + \delta_r) \) for some \( 0 < \delta_r < 1 \).

**Example 1** (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 \( A \leq |A_0|_{ii} \leq \bar{A} \) for all \( 1 \leq i \leq m_1 \) and \( 1 \leq j \leq m_2 \). Let \( \Omega \equiv \Omega_X \) denote the indices for which \( \{x_i\}'s \) take value 1. Then \( \|\mathcal{X}(A)\|_2 = \|A1_\Omega\|_2 \). Easy calculations show that

---

5This 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.

6That is, \( \|B\|_p \equiv \left(\sum_{j=1}^{m_2} \sigma_j(B)^p\right)^{1/p} \), where \( \{\sigma_j(B)\} \) are the singular values of \( B \).

7This 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 \( \underline{\bar{A}} = 1 \) (one star) and \( \bar{A} = 5 \) (five stars).
we can take $\nu = (\bar{\nu}, \underline{\nu})$ defined by $\bar{\nu} = \frac{\sqrt{m_1 m_2}}{\sqrt{n m_1 m_2}} A^{\dagger} \frac{A}{A}$ so that $X$ is uniform RIP($\nu$; $\mathcal{I}$).

**Example 2** (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 [12] entails that $X$ is uniform RIP($\nu$; $\mathcal{I}$) with $\bar{\nu} = 1 + \delta, \underline{\nu} = 1 - \delta$ for some $\delta \in (0, 1)$, with probability at least $1 - C \exp(-cn)^8$, provided $n \geq \bar{m} r_{\text{max}}$.

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

$$
\lambda_{n,r} \propto \exp\left(-c^{tr}(m_1 + m_2)r \log \bar{m}\right).
$$

Given the chosen index $r \in \mathcal{I}$, a prior on $\mathcal{F}_r$ is induced by a prior on all $m_1 \times m_2$ matrices of form $\sum_{i=1}^r u_i v_i^\dagger$ 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 $(g_1 \otimes g_2)^\otimes$ on $(\mathbb{R}^{m_1} \times \mathbb{R}^{m_2})^\dagger$. 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 $(0, \infty)^9$. Let $A_{0,r} \in \arg\min_{B: \text{rank}(B) \leq r} \ell_n^2(f_B, f_0)$, and $\tau_r^{tr} \equiv g(\sigma_{\max}(A_{0,r}) + 1)$ where $\sigma_{\max}$ denotes the largest singular value.

**Theorem 7.** Fix $0 < \eta < 1/2$ and $r_{\text{max}} \leq n$. Suppose that there exists some $\mathcal{M} \subset \mathcal{I}$ such that the linear map $X : \mathbb{R}^{m_1 \times m_2} \to \mathbb{R}^n$ satisfies uniform RIP($\nu$; $\mathcal{M}$), and that for all $r \in \mathcal{M}$, we have

$$
\tau_{r,g}^{tr} \geq e^{-\log \bar{m}/2\eta}, \quad \bar{m} \geq 3 \vee (3\bar{\nu}(1 + \sigma_{\max}(A_{0,r}))n)^{2\eta}.
$$

Then there exists some $c^{tr}$ in (4.1) depending on $\bar{\nu}/\underline{\nu}, \eta$ such that for any $r \in \mathcal{M}$,

$$
P_{f_0}^{(n)} \Pi_n \left(A \in \mathbb{R}^{m_1 \times m_2} : \ell_n^2(f_A, f_0) > C_1^{tr} \left(\inf_{B: \text{rank}(B) \leq r} \ell_n^2(f_0, f_B) + \frac{(m_1 + m_2)r \log \bar{m}}{n}\right) | \mathcal{Y}(n) \right) \leq C_2^{tr} \exp\left(-n c_{n,r}^2 / C_2^{tr}\right).
$$

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

By Theorem 5 of [41], the rate in (4.3) 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.

(4.2) 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

---

8Note here we used the union bound to get a probability estimate $r_{\text{max}} \exp(-cn) \lesssim \exp(-c'n)$ for some $c' < c$ under the assumption that $n \geq \bar{m} r_{\text{max}}$.

9We will always use such $g$ in the prior design in the examples in this section.
the form of oracle inequalities (but still get nearly optimal posterior contraction rates) in (4.3). In particular, the first condition of (4.2) prevents the largest eigenvalue of $A_{0,r}$ from growing too fast. This is in similar spirit with Theorem 2.8 of [16], 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 (4.2) is typically a mild technical condition: we only need to choose $\eta > 0$ small enough.

4.2. Isotonic regression. Consider fitting the Gaussian regression model $Y_i = f_0(x_i) + \varepsilon_i$ by $F \equiv \{ f : [0, 1] \to \mathbb{R} : f \text{ is non-decreasing} \}$. For simplicity the design points are assumed to be $x_i = i/(n + 1)$ for all $1 \leq i \leq n$. Let $F_m \equiv \{ f \in F : f \text{ is piecewise constant with at most } m \text{ constant pieces} \}$. Consider the following prior $\Lambda_n$ on $I = \mathbb{N}$:

\[
\lambda_{n,m} \propto \exp \left( -c_{\text{iso}} m \log(en) \right).
\]

Let $g_m \equiv g^{\otimes m}$ where $g$ is symmetric and non-increasing on $(0, \infty)$. Then $g_m(\mu) \equiv m! g_{m1} \{ \mu_1 \leq \ldots \leq \mu_m \}$ is a valid density on $\{ \mu_1 \leq \ldots \leq \mu_m \}$. Given a chosen model $F_m$ by the prior $\Lambda_n$, we randomly pick a set of change points $\{ x(i(k)) \}_{k=1}^{m} (i(1) < \ldots < i(m))$ and put a prior $g_m$ on $\{ f(x(i(k))) \}$'s. [29] proposed a similar prior with $\Lambda_n$ 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 $f_{0,m} \in \arg\min_{g \in F_m} \ell_\ast_n(f_0, g)$, and $\tau_{\text{iso}} = g(\|f_{0,m}\|_\infty + 1)^{10}$.

**Theorem 8.** Fix $0 < \eta < 1/2$. Suppose that

\[
\tau_{\text{iso}} \geq e^{-\log(en)/2\eta}.
\]

Then there exists some $c_{\text{iso}}$ in (4.4) depending on $\eta$ such that

\[
P_{f_0}^{(n)} \Pi_n \left( f \in F : \ell_\ast_n(f, f_0) > C_{1\text{iso}} \left( \inf_{g \in F_m} \ell_\ast_n(f_0, g) + \frac{m \log(en)}{n} \right) \right) \leq C_{2\text{iso}} \exp \left( -n(\varepsilon_{\text{iso}})^2 / C_{2\text{iso}} \right).
\]

Here $(\varepsilon_{\text{iso}})^2 \equiv \max \left\{ \inf_{g \in F_m} \ell_\ast_n(f_0, g), \frac{m \log(en)}{n} \right\}$, and the constants $C_i\text{iso}$ depend on $\eta$.

(4.6) implies that if $f_0$ is piecewise constant, the posterior distribution contracts at nearly a parametric rate. (4.5) can be checked by the following.

**Lemma 7.** 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 $\eta \in (0, 1/\alpha)$, (4.5) holds uniformly in all $m \in \mathbb{N}$ for $n$ large enough depending on $\alpha$ and $\|f_0\|_{L_2([0,1])}$.

\(^{10}\)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.
4.3. 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 \( \mathbb{X} = [0,1]^d \). Let \( \mathcal{F}_m \equiv \{ f(x) = \max_{1 \leq i \leq m}(a_i \cdot x + b_i) : a_i \in \mathbb{R}^d, b_i \in \mathbb{R} \} \) 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 intercepts \( \{ (a_i,b_i) \in \mathbb{R}^d \times \mathbb{R} \}_{i=1}^m \). We use a prior with density \( \bigotimes_{i=1}^m g \circ d \otimes g \) on \( (\mathbb{R}^d \times \mathbb{R})^m \) to induce a prior on \( \mathcal{F}_m \). Let \( f_{0,m} \in \arg \min_{g \in \mathcal{F}_m} \ell_n^2(f_0,g) \) admit the representation given by 
\[ f_{0,m}(x) \equiv \max_{1 \leq i \leq m}(a_i^{(m)} \cdot x + b_i^{(m)}) \] . Further let 
\[ \tau_{m,g} \equiv \min_{1 \leq i \leq m} \left\{ g \left( \| a_i^{(m)} \|_{\infty} + 1 \right), g \left( |b_i^{(m)}| + 1 \right) \right\} \]

The prior \( \Lambda_n \) we will use on the index \( \mathcal{I} = \mathbb{N} \) is given by
\[ (4.7) \quad \lambda_{n,m} \propto \exp \left( -c_{\text{cvx}} \log m \cdot \log n \right). \]

The first step prior used in [27] is a Poisson proposal, which slightly differs from (4.7) by a logarithmic factor. This would affect the contraction rate only by a logarithmic factor.

**Theorem 9.** Fix \( 0 < \eta < 1/4 \). Suppose that 
\[ (4.8) \quad \tau_{m,g} \geq e^{-\log n \cdot \log 3 m / 8 \eta}, \]
and \( n \geq d \). Then there exists some \( \tau_{\text{cvx}} \) in (4.7) depending on \( \eta \) such that 
\[ (4.9) \quad P_{f_0}^{(n)} \Pi_n \left( f \in \mathcal{F} : \ell_n^2(f,f_0) > C_{1,m}^{\text{cvx}} \left( \inf_{g \in \mathcal{F}_m} \ell_n^2(f_0,g) + \frac{d \log n \cdot m \log 3 m}{n} \right) \right) \leq C_2^{\text{cvx}} \exp \left( -n \left( \tau_{m,n}^{\text{cvx}} \right)^2 / C_2^{\text{cvx}} \right). \]

Here \( (\tau_{m,n}^{\text{cvx}})^2 \equiv \max \left\{ \inf g, \ell_n^2(f_0,g), \frac{d \log n \cdot m \log 3 m}{n} \right\} \), and the constants \( C_i^{\text{cvx}} (i = 1,2) \) depend on \( \eta \).

Our oracle inequality shows that the posterior contraction rate of [27] (Theorem 3.3 therein) is far from optimal. (4.8) can be satisfied by using heavy-tailed priors \( g(\cdot) \) in the same spirit as Lemma 7–if \( f_0 \) is square integrable and the design points are regular enough (e.g. using regular grids on \([0,1]^d\)). Moreover, explicit rate results can be obtained using approximation techniques in [26] (cf. Lemma 4.10 therein). We omit detailed derivations.

**Remark 3.** For univariate convex regression, the term \( \log(3m) \) in (4.7)-(4.9) 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 22.

**Remark 4.** 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 [25]. 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 [25]. Details are omitted.
4.4. 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 $\mathcal{F} \equiv \{ f_{\beta,u}(x,z) = x^\top \beta + u(z) : \beta \in \mathbb{R}^p, u \in \mathcal{U} \}$ 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 4.2). 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 $B_0(s) \equiv \{ v \in \mathbb{R}^p : |\text{supp}(v)| \leq s \}$. In this example the model index $I$ is a 2-dimensional lattice. Our goal here is to construct an estimator that satisfies an oracle inequality over the models $\{ \mathcal{F}_{(s,m)} \}_{(s,m) \in \mathbb{N}^2}$. Consider the following model selection prior:

$$\lambda_{n,(s,m)} \propto \exp \left( -e^{hp}(s \log(ep) + m \log(en)) \right).$$

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 $Q \equiv \{ z_{(i)} \}_{i=1}^m$ with $t(i(1)) < \ldots < t(i(m))$, and then put a prior $g_{s,Q} \equiv g_{s,s} \otimes g_m$ where $g_m$ is a prior on $\{ \mu_1 \leq \ldots \leq \mu_m \} \subset \mathbb{R}^m$ constructed in Section 4.2. Let $f_{0,(s,m)} \in \inf_{g \in \mathcal{F}_{(s,m)}} \ell_2^n(f_0, g)$, and 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 $\tau_{s,g} \equiv g(\|\beta_{0,s}\|_{\infty} + 1)$ and $\tau_{m,g} \equiv g(\|u_{0,m}\|_{\infty} + 1)$. Let $X \in \mathbb{R}^{n \times p}$ be the design matrix so that $X^\top X/n$ is normalized with diagonal elements taking value $1^{11}$.

**Theorem 10.** Fix $0 < \eta < 1/4$ and $p \geq n$. Suppose that

$$\tau_{s,g} \geq e^{-\log(ep)/2\eta}, \quad \tau_{m,g} \geq e^{-\log(en)/2\eta}.$$  

Then there exists some $e^{hp}$ in (4.10) depending on $\eta$ such that

$$P_{f_0} \Pi_n \left( f \in \mathcal{F} : \ell_2^n(f_0, f) > C_{1}^{hp} \left( \inf_{f_{\beta,u} \in \mathcal{F}_{(s,m)}} \ell_2^n(f_0, f_{\beta,u}) + \frac{s \log(ep) + m \log(en)}{n} \right) \right) \leq C_{2}^{hp} \exp \left( -e^{hp}(s \log(ep) + m \log(en)) \right).$$

Here $(\varepsilon_{n,(s,m)})^2 \equiv \max \{ \inf_{f_{\beta,u} \in \mathcal{F}_{(s,m)}} \ell_2^n(f_0, f_{\beta,u}), \frac{s \log(ep) + m \log(en)}{n} \}$, and the constants $C_{i}^{hp}$ $(i = 1, 2)$ depend on $\eta$.

The first condition of (4.11) requires that the magnitude of $\|\beta_{0,s}\|_{\infty}$ does not grow too fast; see also comments following Theorem 7. The second condition of (4.11) is the same as in (4.5). 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 (4.12) becomes

$$\frac{s_0 \log(ep)}{n} + \inf_{m \in \mathbb{N}} \left( \inf_{u \in \mathcal{U}_m} \ell_2^n(u_0, u) + \frac{m \log(en)}{n} \right).$$

---

11 This is a common assumption, cf. Section 6.1 of [10].
The two terms in the rate (4.13) trades off two structures of the experiment: the sparsity of \( h_p(x) \) and the smoothness level of \( u(z) \). The resulting phase transition of the rate (4.13) in terms of these structures is in a sense similar to the results of \([51, 50]\). It is not hard to see that (4.13) cannot be improved in general. Hence our Bayes estimator automatically serves as a theoretically (nearly) optimal adaptive estimator for the high-dimensional partially linear regression model.

### 4.5. 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 \( \mathfrak{M} \equiv \bigcup_{(k,s) \in \mathbb{N}^2} \mathfrak{M}_{(k,s)} \) where \( \mathfrak{M}_{(k,s)} \equiv \{ \Sigma = \Lambda \Lambda^\top + I : \Lambda \in \mathfrak{F}_{(k,s)}(L) \} \) with \( \mathfrak{F}_{(k,s)}(L) \equiv \{ \Lambda \in \mathbb{R}^{p \times k} : \Lambda \in B_0(s), |\epsilon_j(\Lambda)| \leq L^{1/2}, \forall 1 \leq j \leq k \} \). 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:

\[
\lambda_{n,(k,s)} \propto \exp(-C^{\text{cov}} k s \log(ep)).
\]  

**Theorem 11.** Let \( p \geq n \). There exist some \( C^{\text{cov}} \) in (4.14) and some sequence of sieve priors \( \Pi_{n,(k,s)} \) on \( \mathfrak{M}_{(k,s)} \) depending on \( L \) such that

\[
P_{\Sigma_0} \Pi_n \left( \Sigma \in \mathfrak{M} : \| \Sigma - \Sigma_0 \|_F > C^{\text{cov}}_1 \left( \inf_{\Sigma' \in \mathfrak{M}_{(k,s)}} \| \Sigma' - \Sigma_0 \|_F + \frac{ks \log(ep)}{n} \right) \right) \leq C^{\text{cov}}_2 \exp \left( -n \left( \frac{C^{\text{cov}}_{n,(k,s)}}{2} \right)^2 / C^{\text{cov}}_2 \right).
\]

Here \((C^{\text{cov}}_{n,(k,s)})^2 \equiv \max \left\{ \inf_{\Sigma' \in \mathfrak{M}_{(k,s)}} \| \Sigma' - \Sigma_0 \|_F^2, \frac{ks \log(ep)}{n} \right\} \), and the constants \( C^{\text{cov}}_i (i = 1, 2) \) depend on \( L \).

Since spectral norm (non-intrinsic) is dominated by Frobenius norm (intrinsic), our result shows that if the model is well-specified in the sense that \( \Sigma_0 \in \mathfrak{M} \), then we can construct an adaptive Bayes estimator with convergence rates in both norms no worse than \( \sqrt{ks \log p/n} \). [38] 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 \( ks \gtrsim \log p \). [19] 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.
5. Proofs for the main results

5.1. Proof of Theorem 1: main steps. First we need a lemma allowing a change-of-measure argument.

Lemma 8. Let Assumption A hold. There exists some constant $c_4 \geq 1$ only depending on $c_1, c_3$ and $\kappa$ such that for any random variable $U \in [0, 1]$, any $\delta_n \geq d_n(f_0, f_1)$ and any $j \in \mathbb{N}$,

$$P_{f_0}^{(n)}U \leq c_4 \left[ P_{f_1}^{(n)}U \cdot e^{c_4nj\delta_n^2} + e^{-c_4^{-1}nj\delta_n^2} \right].$$

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

Proposition 2. Fix $m \in \mathcal{M}$ such that $\delta_{n,m}^2 \geq d_n^2(f_0, f_m)$. Then there exists some constant $c_8 \geq 1$ (depending on the constants in Assumption A) such that for $j \geq 8\epsilon^2/c_7b$,

$$P_{f_0,m}^{(n)}\left( f \in \mathcal{F} : d_n^2(f, f_{0,m}) > c^2(jb)^\gamma \delta_{n,m}^2 | X^{(n)} \right) \leq c_8e^{-nj\delta_{n,m}^2/c_8\epsilon^2}.$$

Proposition 3. Fix $m \in \mathcal{M}$ such that $\delta_{n,m}^2 \leq d_n^2(f_0, f_m)$. 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 $j \geq 8\epsilon^2/c_7b$,

$$P_{f_0,m}^{(n)}\left( f \in \mathcal{F} : d_n^2(f, f_{0,m}) > c^4(2jb)^\gamma \delta_n^2(f_0, f_{0,m}) | X^{(n)} \right) \leq c_8e^{-nj\delta_{n,m}^2/c_8\epsilon^2}.$$

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

Proof of Theorem 1: main steps. Instead of (2.9), we will prove a slightly stronger statement as follows: for any $j \geq 8\epsilon^2/c_7b$, and $b \geq 2c_4c_8\epsilon^2$,

$$P_{f_0}^{(n)}\left( f \in \mathcal{F} : d_n^2(f, f_0) > c_1(j)^\gamma \left( \inf_{g \in \mathcal{F}_m} d_n^2(f_0, g) + \delta_{n,m}^2 \right) | X^{(n)} \right) \leq c_2e^{-j}\epsilon^{2, n\delta_{n,m}^2/c_2}.$$

Here the constants $c_i (i = 1, 2)$ depends on the constants involved in Assumption A and $c, b$.

Proof of (5.3).

First consider the overfitting case. By Proposition 2 and Lemma 8, we see that when $\delta_{n,m}^2 \geq d_n^2(f_0, f_{0,m})$ holds, for $j \geq 8\epsilon^2/c_7b$,

$$P_{f_0}^{(n)}\left( f \in \mathcal{F} : d_n^2(f, f_0) > 2d_n^2(f_0, f_{0,m}) + 2c^2(jb)^\gamma \delta_{n,m}^2 | X^{(n)} \right)$$

$$\leq P_{f_0}^{(n)}\left( f \in \mathcal{F} : d_n^2(f, f_0) > c^2(jb)^\gamma \delta_{n,m}^2 | X^{(n)} \right)$$

$$\leq c_4 \left[ P_{f_0,m}^{(n)}\left( f \in \mathcal{F} : d_n^2(f, f_{0,m}) > c^2(jb)^\gamma \delta_{n,m}^2 | X^{(n)} \right) e^{c_4nj\delta_n^2} + e^{-c_4^{-1}nj\delta_n^2} \right]$$

$$\leq c_8c_4e^{-nj\delta_{n,m}^2}\left( \frac{b}{\epsilon^2} - c_4 \right) + c_4e^{-c_4^{-1}nj\delta_n^2} \leq 2c_8c_4e^{-jn\delta_{n,m}^2}\min\{c_4, c_4^{-1}\}.$$
Here in the second line we used the fact that \( d_n^2(f, f_{0,m}) \geq d_n^2(f, f_0)/2 - d_n^2(f_0, f_{0,m}) \). (5.4) completes the estimate for overfitting \( m \in \mathcal{M} \).

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 3 and Lemma 8, and use arguments similar to (5.4) to see that for \( j \geq 8c^2/c_7h \),

\[
(5.5)
\]

\[
P_{f_0}^{(n)}(f \in \mathcal{F} : d_n^2(f, f_0) > [2\mathfrak{c}^4(2j\mathfrak{d})^\gamma + 2] d_n^2(f_0, f_{0,m})|X^{(n)})
\]

\[
\leq c_4 \left[ P_{f_0,m}^{(n)}(f \in \mathcal{F} : d_n^2(f, f_{0,m}) > \mathfrak{c}^4(2j\mathfrak{d})^\gamma d_n^2(f_0, f_{0,m})|X^{(n)}) e^{c_4 nj\delta_{n,m}^2} + e^{-c_4 nj\delta_{n,m}^2} \right] \leq 2c_8c_4 e^{-n\delta_{n,m}^2 \min\{c_4, c_4^{-1}\}}.
\]

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) \( \delta_{n,m} \geq d_n(f_0, f_{0,m}) \). The claim of (5.3) follows by combining (5.4) and (5.5).

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

\[
A_j := \{c_1 j^\gamma (d_n^2(f_0, f_{0,m}) + \delta_{n,m}^2) < d_n^2(f, f_0) \leq c_1 (j + 1)^\gamma (d_n^2(f_0, f_{0,m}) + \delta_{n,m}^2) \}.
\]

Then,

\[
P_{f_0}^{(n)}d_n^2(\hat{f}_n, f_0) = P_{f_0}^{(n)}d_n^2 \left( \Pi_n(f|X^{(n)}), f_0 \right) \leq P_{f_0}^{(n)}\Pi_n \left( d_n^2(f, f_0)|X^{(n)} \right)
\]

\[
\leq C_{c_1, c_7, h, \gamma} \left( d_n^2(f_0, f_{0,m}) + \delta_{n,m}^2 \right) + \sum_{j \geq 8c^2/c_7h} P_{f_0}^{(n)}(d_n^2(f, f_0)1_{A_j}|X^{(n)})
\]

\[
\leq C_{c_1, c_7, h, \gamma} \left( d_n^2(f_0, f_{0,m}) + \delta_{n,m}^2 \right) + \frac{2^{\gamma+1}c_1c_2}{n} \sum_{j \geq 8c^2/c_7h} j^{\gamma}n\varepsilon_{n,m}^2 e^{-j\varepsilon_{n,m}^2/c_2}.
\]

The inequality in the first line of the above display is due to Jensen’s inequality applied with \( d_n(\cdot, f_0) \), followed by Cauchy-Schwarz inequality. The summation can be bounded up to a constant depending on \( \gamma, c_1, c_2 \) by

\[
\sum_{j \geq 8c^2/c_7h} j^{\gamma}n\varepsilon_{n,m}^2 e^{-j\varepsilon_{n,m}^2/c_2}
\]

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 \( \int_0^{\infty} x^{\gamma}e^{-x/c_2} dx \) 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 \( m \in \mathcal{M} \).

**5.2. Proofs of Propositions 2 and 3.** We will need several lemmas before the proof of Propositions 2 and 3.

**Lemma 9.** Let Assumption A hold. Let \( \mathcal{F} \) be a function class defined on the sample space \( \mathcal{X} \). Suppose that \( N : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0} \) is a non-increasing function
such that for some \( \varepsilon_0 \geq 0 \) and every \( \varepsilon \geq \varepsilon_0 \), it holds that

\[
N(\varepsilon, \{f \in \mathcal{F} : \varepsilon < d_n(f, f_0) \leq 2\varepsilon\}, d_n) \leq N(\varepsilon).
\]

Then for any \( \varepsilon \geq \varepsilon_0 \), there exists some test \( \phi_n \) such that

\[
P_{f_0}^{(n)} \phi_n \leq \frac{c_6 N(\varepsilon)e^{-c_7n\varepsilon^2}}{(1 - e^{-c_7n\varepsilon^2})^+}, \quad \sup_{f \in \mathcal{F}, d_n(f, f_0) \geq \varepsilon} P_f^{(n)}(1 - \phi_n) \leq c_6 e^{-c_7n\varepsilon^2}.
\]

The constants \( c_5, c_6, c_7 \) are taken from Lemma 1.

**Lemma 10.** Let Assumption A hold. Suppose that \( \Pi \) is a probability measure on \( \{f \in \mathcal{F} : d_n(f, f_0) \leq \varepsilon\} \). Then for every \( C > 0 \), there exists some \( C'_0 > 0 \) depending on \( C, \kappa \) such that

\[
P_{f_0}^{(n)} \left( \int \frac{P_f^{(n)}}{P_{f_0}^{(n)}} \text{d}\Pi(f) \leq e^{-\left(C + c_3\right)n\varepsilon^2} \right) \leq c_1 e^{-C'n\varepsilon^2}.
\]

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

**Proof of Proposition 2.** Fix \( m' \in \mathcal{I} \) with \( m' \geq m \). Now we invoke Lemma 9 with \( \mathcal{F} \equiv \mathcal{F}_{m'}, f_0 \equiv f_{0,m} \in \mathcal{F}_{m} \subset \mathcal{F}_{m'} \) [since \( m' \geq m \)], \( \varepsilon_0 \equiv \varepsilon_{n,m'} \) 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

\[
P_{f_0}^{(n)} \phi_{n,m'} \leq \frac{c_6 e^{\log N(\varepsilon) - c_7 n \delta_{n,m'}^2}}{1 - e^{-c_7 n \delta_{n,m'}^2}} \leq 2c_6 e^{-c_7 n \delta_{n,m'}^2 / 2},
\]

and that

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

Note that here in (5.7) 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 \( \phi_n := \sup_{m' \geq j'h_m} \phi_{n,m'} \) be a global test for big models. Then by (5.7),

\[
P_{f_0}^{(n)} \phi_n \leq \sum_{m' \geq j'h_m} P_{f_0}^{(n)} \phi_{n,m'} \leq \sum_{m' \geq j'h_m} 2c_6 e^{-c_7 n \delta_{n,m'}^2 / 2} \leq 4c_6 e^{-(c_7/2)c^2 n j'h_m^2}.
\]

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

\[
P_{f_0}^{(n)} U \cdot \phi_n \leq P_{f_0}^{(n)} \phi_n \leq 4c_6 e^{-(c_7/2)c^2 n j'h_m^2}.
\]

On the power side, with \( m' = j'h_m \) applied to (5.8) we see that

\[
\sup_{f \in \mathcal{F}_{j'h_m} : d_n^2(f, f_{0,m}) \geq \gamma^2(j) \delta_{n,m}^2} P_f^{(n)}(1 - \phi_n) \leq \sup_{f \in \mathcal{F}_{j'h_m} : d_n^2(f, f_{0,m}) \geq \delta_{n,j'h_m}^2} P_f^{(n)}(1 - \phi_n) \leq c_6 e^{-c_7 n \delta_{n,j'h_m}^2} \leq 2c_6 e^{-(c_7/2)c^2 n j'h_m^2}.
\]
The first inequality follows from the right side of (2.5) since $c^2(jb)^\gamma \delta_{n,m}^2 \geq \delta_{n,jb,m}^2$, and the last inequality follows from the left side of (2.5). On the other hand, by applying Lemma 10 with $C = c_3$ and $\varepsilon^2 \equiv \frac{c_7 jbh_2^2}{8c_3 \varepsilon^2 c^2}$, we see that there exists some event $\mathcal{E}_n$ such that $P_{f_{0,m}}(\mathcal{E}_n^c) \leq c_1 e^{-C'C_jbh_2^2 \delta_{n,m}^2/8c_3 \varepsilon^2}$ and it holds on the event $\mathcal{E}_n$ that

\begin{equation}
\int \frac{p_f(n)}{p_{f_{0,m}}(n)} \, d\Pi(f) \geq \lambda_{n,m} \int f \in \mathcal{F} : d_n(f, f_{0,m}) > c^2(jb)^\gamma \delta_{n,m}^2 \bigl| X(n) \bigr) (1 - \phi_n) \, d\mathcal{E}_n(n)
\end{equation}

Note that

\begin{equation}
P_{f_{0,m}}(n) \Pi_n \left( f \in \mathcal{F} : d_n(f, f_{0,m}) > c^2(jb)^\gamma \delta_{n,m}^2 \bigl| X(n) \bigr) (1 - \phi_n) \, d\mathcal{E}_n(n) \right)
\end{equation}

where the inequality follows from (5.11). On the other hand, the expectation term in the above display can be further calculated as follows:

\begin{equation}
(II) = \int f \in \mathcal{F} : d_n(f, f_{0,m}) > c^2(jb)^\gamma \delta_{n,m}^2 \bigl| \, d\Pi(f) (1 - \phi_n)
\end{equation}

The first term in the third line follows from (5.10) 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 $j \geq 8c^2/c_7b$, 

\begin{equation}
P_{f_{0,m}}(n) \Pi_n \left( f \in \mathcal{F} : d_n(f, f_{0,m}) > c^2(jb)^\gamma \delta_{n,m}^2 \bigl| X(n) \bigr) (1 - \phi_n) \, d\mathcal{E}_n(n) \right) \leq C e^{-(c_7/c^2)jbh_2^2 \delta_{n,m}^2}
\end{equation}

Hence we conclude (5.1) from (5.9), probability estimate on $\mathcal{E}_n^c$, and (5.14).
Choosing where \( C \) to this end, for a constant \( 5.4 \).

Using Lemma 5.3.

Completion of proof of Theorem 1. Proposition Proof of Proposition 22 Q. HAN

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 (5.12) with \( j = 8c^2/\tau h \),

\[
P_{f_{0,m}}(f \notin \mathcal{F}_{jhm}|X^{(n)}) \leq \frac{e^{c\tau njh\delta_{n,m}^2/4c^2} \cdot \Pi(\mathcal{F} \setminus \mathcal{F}_{jhm})}{\lambda_{n,m} \Pi_{n,m}(\{f \in \mathcal{F}_m: d_n^2(f, f_{0,m}) \leq c\tau jh\delta_{n,m}^2/8c^2\})} \leq C e^{-(c/2\lambda^2)njh\delta_{n,m}^2}.
\]

From here (2.10) can be established by controlling the probability estimate for \( \mathcal{E}_n^c \) as in Proposition 2, and a change of measure argument as in (5.4) using Lemma 8.

5.4. Proof of Lemma 1.

Proof of Lemma 1. Let \( c > 0 \) be a constant to be specified later. Consider the test statistics \( \phi_n \equiv 1\left(\log \frac{p_{f_0}^{(n)}}{p_f^{(n)}} \leq -cdn_n^2(f_0, f_1)\right) \). We first consider type I error. Under the null hypothesis, we have for any \( \lambda_1 \in (0,1/\kappa_\Gamma) \),

\[
P_{f_0}(\phi_n) \leq P_{f_0}^{(n)} \left[ \left(\log \frac{p_{f_0}^{(n)}}{p_f^{(n)}} - P_{f_0} \log \frac{p_{f_0}^{(n)}}{p_f^{(n)}}\right) \leq -(c + c_2)nd_n^2(f_0, f_1) \right] \leq c_1 \exp \left(\psi_{\kappa_\gamma nd_n^2(f_0,f_1),\kappa_\Gamma}(\lambda_1)\right) \exp \left(-(\lambda_1(c + c_2)nd_n^2(f_0, f_1))\right).
\]

Choosing \( \lambda_1 > 0 \) small enough (depending on \( \kappa, c_2, c \)) we get

\[
P_{f_0}(\phi_n) \leq C_1 \exp(-C_2nd_n^2(f_0, f_1))
\]

where \( C_1, C_2 > 0 \) depend on \( c_1, c_2, c, \kappa \). Next we handle the type II error. To this end, for a constant \( c' > c_3c_5 \) to be specified later, consider the event \( \mathcal{E}_n \equiv 1\left(\log \frac{p_{f_0}^{(n)}}{p_f^{(n)}} \leq c'nd_n^2(f_0, f_1)\right) \), where \( f \in \mathcal{F} \) is such that \( d_n^2(f, f_1) \leq c_5d_n^2(f_0, f_1) \), and \( \lambda_2 \in (0, 1/\kappa_\Gamma) \),

\[
P_f(\mathcal{E}_n^c) \leq P_f^{(n)} \left( \log \frac{p_f^{(n)}}{p_{f_1}^{(n)}} > c'nd_n^2(f_0, f_1) - c_3nd_n^2(f_0, f_1) \right) \leq P_f^{(n)} \left( \log \frac{p_f^{(n)}}{p_{f_1}^{(n)}} > (c' - c_3c_5)nd_n^2(f_0, f_1) \right) \leq e^{-\lambda_2(c' - c_3c_5)nd_n^2(f_0,f_1)c_1 \exp(\psi_{\kappa_\gamma nd_n^2(f_0,f_1),\kappa_\Gamma}(\lambda_2))}.
\]
By choosing $\lambda_2 > 0$ small enough depending on $c_3, c_5, c', \kappa$, we see that

$$P_f^{(n)}(\mathcal{E}_n^c) \leq C_3 \exp(-C_4 n d_n^2(f_0, f_1))$$

for some constants $C_3, C_4$ depending only on $c_1, c_3, c_5, c', \kappa$ (in particular, does not depend on $f$). On the other hand,

$$P_f^{(n)}(1 - \phi_n) = P_f^{(n)} \left( \log \frac{p_f^{(n)}}{p_{f_0}^{(n)}} + \log \frac{p_f^{(n)}}{p_{f_0}^{(n)}} < c n d_n^2(f_0, f_1) \right) (1_{\mathcal{E}_n} + 1_{\mathcal{E}_n^c})$$

$$\leq P_f^{(n)} \left( \log \frac{p_f^{(n)}}{p_{f_0}^{(n)}} < (c + c')n d_n^2(f_0, f_1) \right) + P_f^{(n)}(\mathcal{E}_n^c) \equiv (*) + P_f^{(n)}(\mathcal{E}_n^c).$$

Since $P_f^{(n)} \log \frac{p_f^{(n)}}{p_{f_0}^{(n)}} \geq c_2 d_n^2(f, f_0) \geq c_2(1 - \sqrt{c_3})^2 d_n^2(f_0, f_1)$, we continue our computation: for $c, c'$ such that $c + c' \leq c_2(1 - \sqrt{c_3})^2$ and $\lambda_3 \in (0, 1/2)$,

$$(*) \leq P_f^{(n)} \left( \log \frac{p_f^{(n)}}{p_{f_0}^{(n)}} - \log \frac{p_f^{(n)}}{p_{f_0}^{(n)}} < -[c_2(1 - \sqrt{c_3})^2 - (c + c')] n d_n^2(f_0, f_1) \right)$$

$$\leq e^{-\lambda_3 [c_2(1 - \sqrt{c_3})^2 - (c + c')] n d_n^2(f_0, f_1)} \leq C_5 \exp(-C_6 n d_n^2(f_0, f_1))$$

where $C_5, C_6$ depending on $c_1, c_2, c_3, c', \kappa, C_3, C_4$. Now we need to choose $c, c', c_5$ such that $c > 0, c' > c_3 c_5, c + c' < c_2(1 - \sqrt{c_3})^2$. This can be done by choosing $c_5 \leq \min\{1/4, c_2/16c_3\}$ and $c' = c = 2c_3c_5$. \hfill $\square$

5.5. **Proof of Lemma 8.** We recall a standard fact.

**Lemma 11.** If a random variable $X$ satisfies $\mathbb{E} \exp(\lambda X) \leq \exp(\psi_{\nu, c}(\lambda))$, then for $t > 0$, $\mathbb{P}(X \geq t) \vee \mathbb{P}(X \leq -t) \leq \exp \left(-\frac{t^2}{2(\nu + c)}\right)$.

**Proof of Lemma 8.** For $c = 2c_3$, consider the event $\mathcal{E}_n \equiv \left\{ \log \frac{p^{(n)}_{f_0}}{p^{(n)}_{f_1}} < c j n \delta_n^2 \right\}$.

By Lemma 11, we have for some constant $C > 0$ depending on $c_1, c_3$ and $\kappa$,

$$P^{(n)}_{f_0}(\mathcal{E}_n^c) \leq P^{(n)}_{f_0} \left( \log \frac{p^{(n)}_{f_0}}{p^{(n)}_{f_1}} - \log \frac{p^{(n)}_{f_0}}{p^{(n)}_{f_1}} \geq c j n \delta_n^2 - c_3 j n \delta_n^2(f_0, f_1) \right)$$

$$\leq P^{(n)}_{f_0} \left( \log \frac{p^{(n)}_{f_0}}{p^{(n)}_{f_1}} - \log \frac{p^{(n)}_{f_0}}{p^{(n)}_{f_1}} \geq c j n \delta_n^2 \right) \quad (\text{since } d_n(f_0, f_1) \leq \delta_n)$$

$$\leq C \exp \left(-\frac{n^2 j^2 \delta_n^4}{C(nj \delta_n^2 + 1)}\right) \leq C \exp \left(-C^{-1} n j \delta_n^2\right).$$
We remind the reader that the constant $C$ may not be the same in the above series of inequalities, and hence the last inequality follows by noting that (i) if $nj\delta_n^2 \geq 1$, then we replace the denominator of the second last line by 2, (ii) if $nj\delta_n^2 < 1$, then we increase $C \geq 1$. Then

\[
P_{f_0^{(n)}}U = P_{f_0^{(n)}}U_1e_n + P_{f_0^{(n)}}U_1\xi_n \leq P_{f_1^{(n)}} \left[U_1\frac{P_{f_0^{(n)}}}{P_{f_1^{(n)}}}1e_n\right] + Ce^{-C^{-1}nj\delta_n^2} \\
\leq P_{f_1^{(n)}}U \cdot e^{cnj\delta_n^2} + Ce^{-C^{-1}nj\delta_n^2},
\]

completing the proof. \hfill \square

5.6. Proof of Proposition 1.

Proof of Proposition 1. Let $\Sigma_n = \sum_{m\in I} e^{-2nd_{n,m}}$ be the total mass. Then $e^{-2nd_{n,1}} \leq \Sigma_n \leq 2e^{-2nd_{n,1}/c^2} \leq 2$. The first condition of (P1) is trivial. We only need to verify the second condition of (P1): $\sum_{k>bm} \lambda_{n,k} = \Sigma_n^{-1} \sum_{k>bm} e^{-2nd_{n,k}} \leq e^{2nd_{n,1}} \cdot 2e^{-2(n/c^2)n\delta_n^2,m} \leq 2e^{-2nd_{n,m}}$, where the first inequality follows from (2.5) and the second by the condition $h \geq 2c^2$. \hfill \square

6. Proofs for Applications

The proofs of the theorems in Section 4 follow the similar route by verifying (i) the local entropy condition in Assumption B, (ii) the summability condition in (2.5) and (iii) the sufficient mass condition (P2) in Assumption C. We remind the reader that we use (2.8) in all examples as the first-step (model selection) prior. We only prove Theorems 7 and 11 in this section. The proofs for Theorems 8-10 are deferred to Appendix B.

6.1. Proof of Theorem 7.

Lemma 12. Let $r \in I$. Suppose that the linear map $\mathcal{X} : \mathbb{R}^{m_1 \times m_2} \to \mathbb{R}^n$ is uniform RIP($\nu, I$). Then for any $\varepsilon > 0$ and $A_0 \in \mathbb{R}^{m_1 \times m_2}$ such that rank($A_0$) $\leq r$, we have

$$\log N(c_5\varepsilon, \{f_A \in F_r: \ell_n(f_A, f_{A_0}) \leq 2\varepsilon, \ell_n\} \leq 2(m_1 + m_2)r \cdot \log \left(18\nu/c_5\nu\right).$$

We will need the following result.

Lemma 13. Let $S(r, B) = \{A \in \mathbb{R}^{m_1 \times m_2} : \text{rank}(A) \leq r, \|A\|_2 \leq B\}$. Then $N(\varepsilon, S(r, B), \|\cdot\|_2) \leq \left(\frac{2B}{\varepsilon}\right)^{(m_1 + m_2 - 1)r}$.

Proof of Lemma 13. The case for $B = 1$ follows from Lemma 3.1 of [12] and the general case follows by a scaling argument. We omit the details. \hfill \square

Proof of Lemma 12. We only need to consider the case $r \leq r_{\text{max}}$. First note that the entropy in question equals

\[
(6.1) \quad \log N(c_5\sqrt{n}\varepsilon, \{\mathcal{X}(A) : \|\mathcal{X}(A - A_0)\|_2 \leq 2\sqrt{n}\varepsilon, \text{rank } A \leq r\}, \|\cdot\|_2).
\]
By uniform RIP($\nu; I$), the set to be covered in the above display is contained in $\{X(A): \|A - A_0\|_2 \leq 2\varepsilon/\nu, \text{rank } A \leq r\} \subset X(S(2r, 2\varepsilon/\nu))$. On the other hand, again by uniform RIP($\nu; I$), a $c_3\varepsilon/\nu$-cover of the set $S(2r, 2\varepsilon/\nu)$ under the Frobenius norm $\|\cdot\|_2$ induces a $c_5\sqrt{n}\varepsilon$-cover of $X(S(2r, 2\varepsilon/\nu))$ under the Euclidean $\|\cdot\|_2$ norm. This implies that (6.1) can be further bounded from above by

$$\log N(c_5\varepsilon/\nu, S(2r, 2\varepsilon/\nu), \|\cdot\|_2) \leq 2(m_1 + m_2)r \cdot \log (18\nu/c_5\nu),$$

where the last inequality follows from Lemma 13.

Now we take $\delta_{n,r}^2 = \left(\frac{4(\log(18\varepsilon/c_5\nu))}{c_7} \vee \frac{1}{7}\right) \cdot (m_1 + m_2)r \log \frac{m}{n}$. Clearly $\delta_{n,r}^2$ satisfies (2.5) with $c \equiv 1$ and $\gamma = 1$.

**Lemma 14.** Suppose that $X : \mathbb{R}^{m_1 \times m_2} \rightarrow \mathbb{R}^n$ is uniform RIP($\nu; I$), and that (4.2) holds. Then (P2) in Assumption C holds.

**Proof of Lemma 14.** We only need to consider $r \leq r_{\text{max}}$. First note that

$$(6.2) \quad \Pi_{r,n}\left(\{f_A \in F_r: \|f_A, f_{A_0,r}\|_2 \leq \delta_{n,r}/c_3\}\right) = \Pi_G \left(\{A \in \mathbb{R}^{m_1 \times m_2}: \|X(A - A_{0,r})\|_2 \leq \sqrt{n}\delta_{n,r}/c_3, \text{rank}(A) \leq r\}\right) \geq \Pi_G \left(\{A \in \mathbb{R}^{m_1 \times m_2}: \|A - A_{0,r}\|_2 \leq \delta_{n,r}/\nu c_3, \text{rank}(A) \leq r\}\right).$$

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

$$\|A_{i,j} - A_{0,r}\|_2 \leq \sum_{i=1}^r \left(\|u_i - u_{i,j}\|_2 \|v_i\|_2 + \|v_i - v_{i,j}\|_2 \|u_i\|_2\right) \leq \sum_{i=1}^r (\varepsilon \sqrt{\sigma_i} + (\sqrt{\sigma_i} + \varepsilon)\varepsilon) \leq \rho_r \varepsilon$$

where $\rho_r \equiv \sum_{i=1}^r (2\sqrt{\sigma_i} + 1)$. Now with $\varepsilon_{n,r} \equiv \frac{\delta_{n,r}}{\sqrt{c_3 \rho_r}} \wedge 1$ we see that (6.2) can be further bounded from below as follows:

$$(6.2) \geq \Pi_G \left(\bigcap_{i=1}^r \{(u_{i,j}, v_{i,j}) : u_{i,j} \in B_{m_1}(u_i, \varepsilon_{n,r}), v_{i,j} \in B_{m_2}(v_i, \varepsilon_{n,r})\}\right) \geq (\tau_{r,g}^m) (m_1 + m_2) \prod_{i=1}^r \text{vol}(B_{m_1}(u_i, \varepsilon_{n,r})) \cdot \text{vol}(B_{m_2}(v_i, \varepsilon_{n,r})) \geq (\tau_{r,g}^m \cdot \varepsilon_{n,r}) (m_1 + m_2) v_d r v_m \geq e^{-2n\delta_{n,r}^2},$$

where $v_d = \text{vol}(B_d(0,1))$, and $v_d \geq (1/\sqrt{d})^d$. Hence in order that the right side of the above display can be bounded from below by $e^{-2n\delta_{n,r}^2}$, it suffices
to require that
\[(6.3) \quad \max \{\log \tau_{r,g}, \log(\varepsilon^{-1}_{n,r} \lor 1)\} \leq \frac{\log \bar{m}}{2\eta}.
\]

It is easy to calculate that \(\varepsilon^{-2}_{n,r} \leq 9\bar{m}^2 c_3 (1 \lor \sigma_{\max}(A_{0,r})^2 r_{\max} n).\) Now the conclusion follows by noting that (4.2) implies (6.3) since \(r_{\max} \leq n\) and \(c_3 = 1.\)

**Proof of Theorem 7.** The theorem follows by Theorems 1 and 2, Proposition 1 coupled with Lemmas 12 and 14. \(\square\)

### 6.2. Proof of Theorem 11.

**Lemma 15.** For any \(\Sigma_0 \in \mathcal{M}_{(k,s)},\)

\[
\log \mathcal{N}(c_5 \varepsilon, \{\Sigma \in \mathcal{M}_{(k,s)} : \|\Sigma - \Sigma_0\|_F \leq C L \varepsilon\}, \|\cdot\|_F) \\
\leq ks \log(ep/s) + ks \log(6\sqrt{kL}/c_5 \varepsilon).
\]

**Proof.** The set involved in the entropy is equivalent to

\[(6.4) \quad \{\Lambda \in \mathcal{B}_{(k,s)}(L) : \|\Lambda \Lambda^\top - \Lambda_0 \Lambda_0^\top\|_F \leq C L \varepsilon, \|\cdot\|_F\}.
\]

We claim that \(\sup_{\Lambda \in \mathcal{B}_{(k,s)}(L)} \|\Lambda \Lambda^\top\|_F \leq \sqrt{kL}.\) To see this, let \(\Lambda \equiv P \Xi Q^\top\) be the singular value decomposition of \(\Lambda,\) 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 \(\|\Lambda \Lambda^\top\|_F = \|\Xi \Xi^\top\|_F \leq kL,\) proving the claim. Combined with (6.4) and Euclidean embedding, we see that the entropy in question can be bounded by

\[
\log \mathcal{N}(c_5 \varepsilon, \{v \in B_0(k,s; pk) : \|v\|_2 \leq 2\sqrt{kL}\}, \|\cdot\|_2) \\
\leq \log \left(\frac{pk}{ks} \left(\frac{6\sqrt{kL}}{c_5 \varepsilon}\right)^{ks}\right) \\
\leq ks \log(ep/s) + ks \log(6\sqrt{kL}/c_5 \varepsilon),
\]

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

**Proof of Theorem 11.** Take \(\delta_{n,(k,s)}^2 = KC''^n s \log(C'p)\) for some \(C' \geq c\) depending on \(c_5, c_7, L\) and some absolute constant \(K \geq 1.\) Apparently (2.5) holds with \(c = 1, \gamma = 1.\) The prior \(\Pi_{n,(k,s)}\) on \(\mathcal{M}_{(k,s)}\) will be the uniform distribution on a minimal \(\sqrt{C'' s \log(C'p)}\) covering-ball of the set \(\{\Sigma \in \mathcal{M}_{(k,s)}\}\) under the Frobenius norm \(\|\cdot\|_F.\) The above lemma entails that the cardinality for such a cover is no more than \(\exp(C''^n s \log(C''p))\) for another constant \(C'' \geq c\) depending on \(c_3, c_5, c_7, L.\) Hence

\[
\Pi_{n,(k,s)} \left(\left\{\Sigma \in \mathcal{M}_{(k,s)} : \|\Sigma - \Sigma_{0,(k,s)}\|_F \leq \delta_{n,(k,s)}^2 / c_3\right\}\right) \geq \exp(-C''^n s \log(C''p)),
\]

which can be bounded from below by \(\exp(-2n\delta_{n,(k,s)}^2)\) by choosing \(K\) large enough. The claim of Theorem 11 now follows from these considerations along with Theorems 1 and 6, Proposition 1. \(\square\)
Appendix A. Proof of lemmas in Section 3

Proof of Lemma 2. 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 $\theta_0$.

First consider Gaussian regression case. It is easy to calculate that

$$\log \frac{P_{\theta_0}^{(n)}}{P_{\theta_i}^{(n)}}(X^{(n)}) = \sum_{i=1}^n \left[ -\frac{1}{2}(X_i - \theta_{0,i})^2 + \frac{1}{2}(X_i - \theta_{1,i})^2 \right],$$

$$P_{\theta_0}^{(n)} \log \frac{P_{\theta_0}^{(n)}}{P_{\theta_i}^{(n)}}(X^{(n)}) = \frac{1}{2} n \ell_n^2(\theta_0, \theta_1).$$

Then

$$P_{\theta_0}^{(n)} \exp \left[ \lambda \left( \log \frac{P_{\theta_0}^{(n)}}{P_{\theta_i}^{(n)}}(X^{(n)}) - P_{\theta_0}^{(n)} \log \frac{P_{\theta_0}^{(n)}}{P_{\theta_i}^{(n)}}(X^{(n)}) \right) \right]$$

$$\leq P \exp \left( \sum_{i=1}^n \varepsilon_i \lambda (\theta_{0,i} - \theta_{1,i}) \right) \leq \exp \left( \lambda^2 n \ell_n^2(\theta_0, \theta_1)/2 \right).$$

Next consider binary regression. Easy calculation shows that

$$\log \frac{P_{\theta_0}^{(n)}}{P_{\theta_i}^{(n)}} = \sum_{i=1}^n X_i \log \frac{\theta_{0,i}}{\theta_{1,i}} + (1 - X_i) \log \frac{1 - \theta_{0,i}}{1 - \theta_{1,i}},$$

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

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 $P_{\theta_0}^{(n)} \log \frac{P_{\theta_0}^{(n)}}{P_{\theta_i}^{(n)}} \asymp n \ell_n^2(\theta_0, \theta_1)$ under the assumed condition that $\Theta_n \subset [\eta, 1 - \eta]^n$. Now we verify the Bernstein condition:

$$P_{\theta_0}^{(n)} \exp \left[ \lambda \left( \log \frac{P_{\theta_0}^{(n)}}{P_{\theta_i}^{(n)}} - P_{\theta_0}^{(n)} \log \frac{P_{\theta_0}^{(n)}}{P_{\theta_i}^{(n)}} \right) \right]$$

$$= P_{\theta_0}^{(n)} \exp \left( \lambda \sum_{i=1}^n (X_i - \theta_{0,i}) t_i \right) \leq \exp \left( \lambda^2 \sum_{i=1}^n t_i^2 / 8 \right)$$

where $t_i \equiv t_i(\theta_0, \theta_1) = \log \left( \frac{\theta_{0,i}}{1 - \theta_{0,i}} \frac{1 - \theta_{1,i}}{\theta_{1,i}} \right)$ and the last inequality follows from Hoeffding’s inequality (cf. Section 2.6 of [9]). The claim follows by noting that $t_i^2 = \left[ \log \left( \frac{\theta_{0,i} - \theta_{1,i}}{1 - \theta_{0,i} + \theta_{1,i}} + 1 \right) \right]^2 \asymp (\theta_{0,i} - \theta_{1,i})^2$ 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
\[
\log \frac{p_{\theta_0}^{(n)}}{p_{\theta_1}^{(n)}} = \sum_{i=1}^{n} X_i \log \frac{\theta_{0,i}}{\theta_{1,i}} + (\theta_{1,i} - \theta_{0,i}),
\]
where
\[
P_{\theta_0}^{(n)} \log \frac{p_{\theta_0}^{(n)}}{p_{\theta_1}^{(n)}} = \sum_{i=1}^{n} \theta_{0,i} \log \frac{\theta_{0,i}}{\theta_{1,i}} + (\theta_{1,i} - \theta_{0,i}).
\]

Note that for any \(1/M \leq p, q \leq M\),
\[
p \log \frac{p}{q} - (p-q) = p \left( -\log \frac{q}{p} - 1 + \frac{q}{p} \right) \asymp p \cdot \left( \frac{1}{p} - 1 \right)^2 \asymp (p-q)^2,
\]
where in the middle we used the fact that \(-\log x - 1 + x \asymp (x-1)^2\) for \(x\) bounded away from 0 and \(\infty\). This shows that \(P_{\theta_0}^{(n)} \log \frac{p_{\theta_0}^{(n)}}{p_{\theta_1}^{(n)}} \asymp n \ell^{(2)}(\theta_0, \theta_1)\).

Next we verify the Bernstein condition:
\[
P_{\theta_0}^{(n)} \exp \left[ \lambda \left( \log \frac{p_{\theta_0}^{(n)}}{p_{\theta_1}^{(n)}} - P_{\theta_0}^{(n)} \log \frac{p_{\theta_0}^{(n)}}{p_{\theta_1}^{(n)}} \right) \right]
\leq P_{\theta_0}^{(n)} \exp \left[ \lambda \sum_{i=1}^{n} (X_i - \theta_{0,i})t_i \right] \leq \exp \left[ \sum_{i=1}^{n} \theta_{0,i} \left( e^{\lambda t_i} - 1 - \lambda t_i \right) \right]
\]
where \(t_i = \log(\theta_{0,i}/\theta_{1,i})\). Now for any \(|\lambda| \leq 1\), we have \(e^{\lambda t_i} - 1 - \lambda t_i \asymp \lambda^2 t_i^2 \leq \lambda^2 t_i^2/(1-|\lambda|)\). On the other hand, \(\theta_{0,i}t_i^2 = \theta_{0,i} (\log(\theta_{0,i}/\theta_{1,i}))^2 \asymp (\theta_{0,i} - \theta_{1,i})^2\), completing the proof. \(\square\)

**Proof of Lemma 3.** 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 Bernstein condition in Assumption A holds with \(v = \kappa_g n \text{Var}_{f_0}(\log f_0/f_1), c = \kappa_\Gamma\) where \(\kappa_g, \kappa_\Gamma\) depend only on \(\mathcal{G}\). Hence we only need to verify that
\[
\text{Var}_{f_0}(\log f_0/f_1) \lesssim h^2(f_0, f_1), \quad P_{f_0}(\log f_0/f_1) \asymp h^2(f_0, f_1).
\]
This can be seen by Lemma 8 of [24] and the fact that Hellinger metric is dominated by the Kullback-Leibler divergence. \(\square\)

**Lemma 16.** Let \(Z\) be a random variable bounded by \(M > 0\). Then \(\mathbb{E} \exp(Z) \leq \exp \left( e^M \mathbb{E} Z \right)\).

**Proof.** Note that
\[
\log \mathbb{E} \exp(Z) = \log \left( \mathbb{E} [\exp(Z) - 1] + 1 \right) \leq \mathbb{E} [\exp(Z) - 1] \leq e^M \mathbb{E} Z.
\]
where the last inequality follows from Taylor expansion \(e^x - 1 = \sum_{k=1}^{n} x^k/k! \leq x \sum_{k \geq 1} M^{k-1}/k! \leq xe^M\). \(\square\)
Proof of Lemma 4. We omit explicit dependence of $M$ on the notation $d_{r,M}$ and $r_M$ in the proof. Let $P_f^{(n)}$ 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
\[
\log \frac{P_f^{(n)}}{P_f^{(n)}} = \sum_{i=0}^{n-1} \left[ \varepsilon_{i+1} (f_0(X_i) - f_1(X_i)) + \frac{1}{2} \varepsilon_{i+1} (f_0(X_i) - f_1(X_i))^2 \right],
\]
\[
P_f^{(n)} \log \frac{P_f^{(n)}}{P_f^{(n)}} = \frac{n}{2} \int (f_0 - f_1)^2 q_{f_0} \, d\lambda.
\]
Here $\lambda$ denotes the Lebesgue measure on $\mathbb{R}$. By the arguments on page 209 of [23], we see that $r \lesssim q_{f_0} \lesssim r$. Hence we only need to verify the Bernstein condition. By Cauchy-Schwarz inequality,
\[
(A.1) \quad \left[ P_f^{(n)} \exp \left( \lambda \log \frac{P_f^{(n)}}{P_f^{(n)}} \right) \right]^2 \leq P_f^{(n)} \exp \left( 2\lambda \sum_{i=0}^{n-1} \varepsilon_{i+1} (f_0(X_i) - f_1(X_i)) \right) \times P_f^{(n)} \exp \left( \lambda \sum_{i=0}^{n-1} (f_0(X_i) - f_1(X_i))^2 \right) \equiv (I) \times (II).
\]
The first term $(I)$ can be handled by an inductive calculation. First note that for any $|\mu| \leq 2$ and $X_1 \in \mathbb{R}$,
\[
(A.2) \quad P_{p(\cdot|x_1)} \mu_2^2(f_0(\cdot|x_2) - f_1(\cdot|x_2))^2 \leq e^{16M^2 \mu^2} P_{p(\cdot|x_1)}(f_0 - f_1)(\cdot|x_2)^2 \leq e^{CM \mu^2} d_{p,f_0,f_1}^2
\]
where the first inequality follows from Lemma 16 and the second inequality follows from $r(\cdot) \lesssim q_{f_0} \lesssim r(\cdot)$ 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 $\varepsilon_n \equiv (\varepsilon_1, \ldots, \varepsilon_n)$. Then for $|\lambda| \leq 1$, let $\mu \equiv 2\lambda$,
\[
P_{p(\cdot|x_1)}^{(n)} e^{2\lambda S_n} = P_{p(\cdot|x_1)}^{(n)} e^{\mu S_n}
\]
\[
\leq E_{X_0,\varepsilon_{n-1}} \left[ e^{\mu S_{n-1}} e^{\varepsilon_{n}} e^{\mu \varepsilon_{n}} (f_0(X_{n-1}) - f_1(X_{n-1})) \right]
\leq E_{X_0,\varepsilon_{n-1}} \left[ e^{\mu S_{n-1}} e^{\mu^2 (f_0(X_{n-1}) - f_1(X_{n-1}))^2 / 2} \right]
\leq E_{X_0,\varepsilon_{n-2}} \left[ e^{\mu S_{n-2}} e^{\varepsilon_{n-1} \mu (f_0(X_{n-2}) - f_1(X_{n-2}) + \mu^2 (f_0(X_{n-1}) - f_1(X_{n-1}))^2 / 2} \right]
\leq E_{X_0,\varepsilon_{n-2}} \left[ e^{\mu S_{n-2}} \left( E_{\varepsilon_{n-1}} e^{2\mu \varepsilon_{n-2} (f_0(X_{n-2}) - f_1(X_{n-2}))} \right)^{1/2} \right]
\leq E_{X_0,\varepsilon_{n-2}} \left[ e^{\mu S_{n-2}} \left( E_{\varepsilon_{n-1}} e^{2\mu \varepsilon_{n-2} (f_0(X_{n-2}) - f_1(X_{n-2}))} \right)^{1/2} \right] \times \left( E_{p(\cdot|x_{n-2})} e^{\mu^2 (f_0(X_{n-1}) - f_1(X_{n-1}))^2 / 2} \right)^{1/2}
\leq E_{X_0,\varepsilon_{n-2}} \left[ e^{\mu S_{n-2}} e^{\mu^2 (f_0(X_{n-2}) - f_1(X_{n-2}))^2 / 2} \right] \cdot e^{CM \mu^2} d_{p,f_0,f_1}^2 / 2,
where the last inequality follows from (A.2). Now we can iterate the above calculation to see that
\[(A.3) \quad (I) \leq \exp(C_M \lambda^2 n d_f^2(f_0, f_1)).\]

Next we consider (II). Since for any non-negative random variables $Z_1, \ldots, Z_n$, we have $\mathbb{E} \prod_{i=1}^n Z_i \leq \prod_{i=1}^n (\mathbb{E}Z_i)^{1/n}$. It follows that
\[(A.4) \quad (II) \leq \prod_{i=1}^n \left( P_{f_0}^{(n)} \exp \left( n \lambda (f_0(X_i) - f_1(X_i))^2 \right) \right)^{1/n} = P_{f_0} \exp(n \lambda (f_0(X_0) - f_1(X_0))^2)
\]
where the last inequality follows by stationarity. On the other hand, by Jensen’s inequality,
\[(A.5) \quad \exp \left( -\lambda P_{f_0}^{(n)} \log \frac{P_{f_0}^{(n)}}{P_{f_1}^{(n)}} \right) \leq \exp \left( -\frac{\lambda n}{2} P_{f_0} (f_0 - f_1)^2 \right) \leq P_{f_0} \exp(-\lambda n (f_0(X_0) - f_1(X_0))^2/2).
\]
Collecting (A.1) and (A.3)-(A.5), we see that for $|\lambda| \leq 1$,
\[
P_{f_0} \exp \left( \lambda \log \frac{P_{f_0}^{(n)}}{P_{f_1}^{(n)}} - P_{f_0}^{(n)} \log \frac{P_{f_0}^{(n)}}{P_{f_1}^{(n)}} \right) \leq \sqrt{(I) \cdot (II)} \exp \left( -\lambda P_{f_0}^{(n)} \log \frac{P_{f_0}^{(n)}}{P_{f_1}^{(n)}} \right) \leq \exp(C_M \lambda^2 n d_f^2(f_0, f_1)),
\]
completing the proof.  \(\square\)

**Proof of Lemma 5.** 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 $g_0, g_1 \in \mathcal{G}$,
\[
\log \frac{p_{g_0}^{(n)}}{p_{g_1}^{(n)}}(X(n)) = -\frac{1}{2}(X(n))^\top (\Sigma_{g_0}^{-1} - \Sigma_{g_1}^{-1}) X(n) - \frac{1}{2} \log \det(\Sigma_{g_0} \Sigma_{g_1}^{-1}),
\]
(A.6)
\[
P_{g_0} \log \frac{p_{g_0}^{(n)}}{p_{g_1}^{(n)}} = -\frac{1}{2} \text{tr}(I - \Sigma_{g_0} \Sigma_{g_1}^{-1}) - \frac{1}{2} \log \det(\Sigma_{g_0} \Sigma_{g_1}^{-1})
\]
where we used the fact that for a random vector $X$ with covariance matrix $\Sigma$, $\mathbb{E}X^\top AX = \text{tr}(\Sigma A)$. 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
\[
Y_n \equiv \log \frac{p_{g_0}^{(n)}}{p_{g_1}^{(n)}}(X(n)) - P_{g_0}^{(n)} \log \frac{p_{g_0}^{(n)}}{p_{g_1}^{(n)}}
\]
\[
= -\frac{1}{2} \left[ (X(n))^\top (\Sigma_{g_0}^{-1} - \Sigma_{g_1}^{-1}) X(n) - \text{tr}(I - \Sigma_{g_0} \Sigma_{g_1}^{-1}) \right] = -\frac{1}{2} \left[ G^\top B G - \text{tr}(B) \right].
\]
Let $B = U^T \Lambda U$ be the spectral decomposition of $B$ where $U$ is orthonormal and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix. Then we can further compute

$$
-2Y_n = d^2 G^T \Lambda G - \text{tr}(\Lambda) = \sum_{i=1}^n \lambda_i (g_i^2 - 1)
$$

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

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

where the inequality follows from $-\log(1-2t) - 2t = \sum_{k \geq 2} \frac{1}{k} (2t)^k = 4t^2 \sum_{k \geq 0} \frac{1}{2^k} (2t)^k < 2t^2$. Hence apply the above display with $t = -\lambda_i / 2$, we have that for any $|\lambda| < 1/\max_i \lambda_i$,

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

Denote $\|\cdot\|$ and $\|\cdot\|_F$ the matrix operator norm and Frobenius norm respectively. By the arguments on page 203 of [23], 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 $\Sigma_g$ and their inverses running over $g$ must be bounded. Hence

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

Next, note that

$$
\left( \sum_i \lambda_i^2 \right)^{1/2} = (\text{tr}(BB^T))^{1/2} = \|B\|_F = \|(\Sigma_{g_1} - \Sigma_{g_0})\Sigma_{g_1}^{-1}\|_F \leq \|\Sigma_{g_1}^{-1}\| \|\Sigma_{g_1} - \Sigma_{g_0}\|_F \leq C_g \sqrt{nD_n^2(g_0, g_1)}
$$

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 (A.7)-(A.9) we see that Assumption A is satisfied for $v = \kappa_g nD_n^2(g_0, g_1)$ and $c = \kappa_G$ for some constants $\kappa_g, \kappa_G$ depending on $\mathcal{G}$ only.
Finally we establish equivalence of \( \frac{1}{n} P_{g_0}^{(n)} \log \frac{p_{g_0}^{(n)}}{p_{g_1}^{(n)}} \) and \( D_n^2(g_0, g_1) \). First by (A.6), we have

\[
P_{g_0}^{(n)} \log \frac{p_{g_0}^{(n)}}{p_{g_1}^{(n)}} = -\frac{1}{2} \text{tr}(I - \Sigma_{g_0}^{-1} \Sigma_{g_1}^{-1}) - \frac{1}{2} \log \det(\Sigma_{g_0}^{-1} \Sigma_{g_1}^{-1}) \\
= \frac{1}{2} \left( \text{tr} \left( \Sigma_{g_1}^{-1/2} (\Sigma_{g_0} - \Sigma_{g_1}) \Sigma_{g_1}^{-1/2} \right) - \log \det \left( I + \Sigma_{g_1}^{-1/2} (\Sigma_{g_0} - \Sigma_{g_1}) \Sigma_{g_1}^{-1/2} \right) \right) \\
\leq \frac{1}{4} \|I - \Sigma_{g_0}^{-1} \Sigma_{g_1}^{-1}\|_F^2 \\
\leq \frac{1}{4} \|\Sigma_{g_1}^{-1}\|_F^{-2} \leq C g'' n D_n^2(g_0, g_1).
\]

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 \( -\log \det(I + A) + \operatorname{tr}(A) \leq \frac{1}{2} \operatorname{tr}(A^2) \) 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 \( P_{g_0}^{(n)} \log \frac{p_{g_0}^{(n)}}{p_{g_1}^{(n)}} \geq C g'' n D_n^2(g_0, g_1) \), thereby completing the proof. \( \square \)

Proof of Lemma 6. Note that

\[
\log \frac{p_{\Sigma_0}^{(n)}}{p_{\Sigma_1}^{(n)}}(X^{(n)}) = -\sum_{i=1}^{n} \left[ \frac{1}{2} X_i^\top (\Sigma_0^{-1} - \Sigma_1^{-1}) X_i - \frac{1}{2} \log \det(\Sigma_0^{-1}) \right], \\
P_{\Sigma_0}^{(n)} \log \frac{p_{\Sigma_0}^{(n)}}{p_{\Sigma_1}^{(n)}} = -\frac{n}{2} \operatorname{tr}(I - \Sigma_0^{-1}) - \frac{n}{2} \log \det(\Sigma_0^{-1}).
\]

The rest of the proof proceeds along the same line as in Lemma 5. \( \square \)

APPENDIX B. PROOF OF REMAINING THEOREMS IN SECTION 4

B.1. PROOF OF THEOREM 8.

Lemma 17. Let \( n \geq 2 \). Then for any \( g \in \mathcal{F}_m \),

\[
\log \mathcal{N}(c_5 \varepsilon, \{ f \in \mathcal{F}_m : \ell_n(f, g) \leq 2\varepsilon \}, \ell_n) \leq 2 \log(6/c_5) \cdot m \log(en).
\]

Proof of Lemma 17. Let \( \mathcal{Q}_m \) denote all \( m \)-partitions of the design points \( x_1, \ldots, x_n \). Then it is easy to see that \( |\mathcal{Q}_m| = \binom{n}{m-1} \). For a given \( m \)-partition \( Q \in \mathcal{Q}_m \), 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

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

On the other hand, for any fixed \( m \)-partition \( Q \in \mathcal{Q}_m \), the entropy term above equals \( \mathcal{N}(c_5 \sqrt{n} \varepsilon, \{ \gamma \in \mathcal{P}_{n,m,Q} : \|\gamma - g\|_2 \leq 2\sqrt{n} \varepsilon, \|\cdot\|_2 \} \), 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 \( \{ \gamma \in P_{n,m,Q} : \| \gamma - \pi_{P_{n,m,Q}}(g) \|_2 \leq 2\sqrt{\varepsilon} \} \) where \( \pi_{P_{n,m,Q}} \) is the natural projection from \( \mathbb{R}^n \) onto the subspace \( P_{n,m,Q} \). Clearly \( P_{n,m,Q} \) 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 [47], page 94 combined with the discussion in page 98 relating the packing number and covering number],

(B.2) 

\[
\log N(c_5\varepsilon, \{ f \in F_{m,Q} : \ell_n(f, f_{0,m}) \leq 2\varepsilon \}, \ell_n) \leq \log \left( \frac{3 \cdot 2\sqrt{\varepsilon}}{c_5\sqrt{\varepsilon}} \right)^m = m \log(6/c_5).
\]

The claim follows by (B.1)-(B.2), and \( \log (\frac{n}{m-1}) \leq m \log(en) \).

Hence we can take \( \delta_{n,m}^2 \equiv \left( \frac{4\log(6/c_3)}{c_7} \vee \frac{1}{\eta} \right)^m \frac{m \log(en)}{n} \). It is clear that (2.5) is satisfied with \( \varepsilon \equiv 1 \) and \( \gamma = 1 \).

**Lemma 18.** Suppose that (4.5) holds. Then (P2) in Assumption C holds.

**Proof of Lemma 18.** Let \( Q_{0,m} = \{ I_k \}_{k=1}^m \) be the associated \( m \)-partition of \( \{ x_1, \ldots, x_n \} \) of \( f_{0,m} \in F_m \) 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 \( \mu_{0,m} = (\mu_0, \ldots, \mu_{0,m}) \equiv (f_{0,m}(x_{i(1)}), \ldots, f_{0,m}(x_{i(m)})) \in \mathbb{R}^m \) is well-defined and \( \mu_{0,1} \leq \ldots \leq \mu_{0,m} \). It is easy to see that any \( f \in F_{m,Q_{0,m}} \) 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

(B.3) 

\[
\Pi_{n,m} (\{ f \in F_m : \ell_n^2(f, f_{0,m}) \leq \delta_{n,m}^2/c_3 \}) \geq \left( \frac{n}{m-1} \right)^{-1} \Pi_{\tilde{g}_m} (\{ f \in F_{m,Q_{0,m}} : \ell_n^2(f, f_{0,m}) \leq \delta_{n,m}^2/c_3 \}) \
\geq \left( \frac{n}{m-1} \right)^{-1} \Pi_{\tilde{g}_m} (\{ \mu \in \mathbb{R}^m : \mu \equiv (\mu_{0,k} + \varepsilon_k)_{k=1}^m, 0 \leq \varepsilon_1 \leq \ldots \leq \varepsilon_m \leq \delta_{n,m}/\sqrt{c_3} \}) \
\geq \left( \frac{n}{m-1} \right)^{-1} \inf_{\mu \in \mathbb{R}^m : \mu \equiv (\mu_{0,k} + \varepsilon_k)_{k=1}^m, 0 \leq \varepsilon_1 \leq \ldots \leq \varepsilon_m \leq \delta_{n,m}/\sqrt{c_3}} \tilde{g}_m(\mu) (1 \wedge \delta_{n,m}/\sqrt{c_3})^m \frac{1}{m!} \
\geq \left( \frac{n}{m-1} \right)^{-1} \tau_{m,g}^{iso}(1 \wedge \delta_{n,m}/\sqrt{c_3})^m \geq m \log(en)^{-1} - m \log(m) \left( \frac{\tau_{m,g}^{iso}}{\tau_{n,m}^{iso}} \right)^{-1} \vee 1.
\]

Here the first inequality in the last line follows from the definition of \( \tilde{g}_m \) and \( \tau_{m,g}^{iso} \). The claim follows by verifying (4.5) implies that the second and third term in the exponent above are both bounded by \( \frac{1}{2m} \cdot m \log(en) \) [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 \( \eta \)].

**Proof of Theorem 8.** The theorem follows by Theorems 1 and 2, Proposition 1 coupled with Lemmas 17 and 18.

We now prove Lemma 7. We need the following result.
Lemma 19. Let \( f_0 := (f_0(x_1), \ldots, f_0(x_n)) \in \mathbb{R}^n \), and \( f_{0,m} := (f_{0,m}(x_1), \ldots, f_{0,m}(x_n)) \in \mathbb{R}^n \) where \( f_{0,m} \in \arg\min_{g \in F_m} f_0^2(f_0, g) \). Suppose that \( \|f_0\|_2 \leq L \), and that there exists some element \( f \in F_m \) such that \( f \equiv (f(x_1), \ldots, f(x_n)) \) satisfies \( \|f\|_2 \leq L \). Then \( \|f_{0,m}\|_2 \leq 3L \).

Proof of Lemma 19. It can be seen that \( f_{0,m} \in \arg\min_{\gamma \in \mathcal{P}_{n,m}} \mathcal{L}_f(\gamma) \equiv \arg\min_{\gamma \in \mathcal{P}_{n,m}} \|f_0 - \gamma\|_2 \) where \( \mathcal{P}_{n,m} \equiv \{(f(x_1), \ldots, f(x_n)) : f \in F_m\} \). For any \( \gamma \in \mathcal{P}_{n,m} \) such that \( \|\gamma\|_2 \leq L \), the loss function satisfies \( \mathcal{L}_f(\gamma) \leq 2L \) by triangle inequality. If \( \|f_{0,m}\|_2 > 3L \), then \( \mathcal{L}_f(f_{0,m}) = \|f_0 - f_{0,m}\|_2 \geq \|f_{0,m}\|_2 - \|f_0\|_2 > 3L - L = 2L \), contradicting the definition of \( f_{0,m} \) as a minimizer of \( \mathcal{L}_f(\cdot) \) over \( \mathcal{P}_{n,m} \). This shows the claim. \( \square \)

Proof of Lemma 7. Let \( L = \int_0^1 f^2 \). Note that \( \|f_0\|_2^2 = 2n \int_0^1 f^2(x) \, dx = 2nL^2 \). By Lemma 19, we see that \( \|f_{0,m}\|_2 \leq 3\sqrt{2nL} \), which entails that \( \|f_{0,m}\|_\infty \leq 3\sqrt{2nL} \). 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 \( n \to \infty \). \( \square \)

B.2. Proof of Theorem 9. Checking the local entropy assumption B requires some additional work. The notion of pseudo-dimension will be useful in this regard. Following [39] Section 4, a subset \( V \) of \( \mathbb{R}^d \) is said to have pseudo-dimension \( t \), denoted as pdim(\( V \)) = \( t \), if for every \( x \in \mathbb{R}^{t+1} \) and indices \( I = (i_1, \ldots, i_{t+1}) \in \{1, \ldots, n\}^{t+1} \) 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 20. Let \( n \geq 2 \). Suppose that pdim(\( \mathcal{P}_{n,m} \)) \( \leq D_m \) where \( \mathcal{P}_{n,m} := \{(f(x_1), \ldots, f(x_n)) \in \mathbb{R}^n : f \in F_m\} \). Then for all \( g \in F_m \),

\[
\log \mathcal{N}(c_5 \varepsilon, \{f \in F_m : \ell_n(f, g) \leq 2\varepsilon\}, \ell_n) \leq C \cdot D_m \log n
\]

for some constant \( C > 0 \) depending on \( c_5 \).

To prove Lemma 20, we need the following result, cf. Theorem B.2 [25].

Lemma 21. 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 \( \mathcal{N}(\varepsilon, A, \|\cdot\|_2) \leq \left(4 + \frac{2B\sqrt{n}}{\varepsilon}\right)^t \), holds for some absolute constant \( \kappa \geq 1 \).

Proof of Lemma 20. Note that the entropy in question can be bounded by

\[
\log \mathcal{N}(c_5 \varepsilon \sqrt{n}, \{\mathcal{P}_{n,m} - g\} \cap B_n(0, 2\sqrt{n}\varepsilon), \|\cdot\|_2).
\]

Since translation does not change the pseudo-dimension of a set, \( \mathcal{P}_{n,m} - g \) 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 \( \{\mathcal{P}_{n,m} - g\} \cap B_n(0, 2\sqrt{n}\varepsilon) \) is uniformly bounded by \( 2\sqrt{n}\varepsilon \), hence an application of Lemma 21 yields that the above display can be further bounded as follows:

\[
\log \mathcal{N}(c_5 \varepsilon, \{f \in F_m : \ell_n(f, g) \leq 2\varepsilon\}, \ell_n) \leq \kappa D_m \log \left(4 + 4n/c_5\right) \leq C \cdot D_m \log n
\]

for some constant \( C > 0 \) depending on \( c_5 \) whenever \( n \geq 2 \). \( \square \)
The pseudo-dimension of the class of piecewise affine functions $\mathcal{F}_m$ can be well controlled, as the following lemma shows.

**Lemma 22** (Lemma 4.9 in [26]), \( \text{pdim}(\mathcal{P}_{n,m}) \leq 6md \log 3m. \)

As an immediate result of Lemmas 20 and 22, we can take for \( n \geq 2, \delta_{n,m}^2 := (C \lor 1/\eta) d \cdot \frac{\log n}{n} \cdot m \log 3m \) for some \( C \geq 2/c_\gamma \) depending on \( c_3, c_\gamma. \)

**Lemma 23.** Suppose that (4.8) holds and \( n \geq d. \) Then (P2) in Assumption C holds.

**Proof of Lemma 23.** We write \( f_{0,m} \equiv \max_{1 \leq i \leq m} (a_i \cdot x + b_i) \) throughout the proof. We first claim that for any \( a^*_i \in B_d(a_i, \delta_{n,m}/2\sqrt{c_3d}) \) 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 \( x \in \mathcal{X}, \) there exists some index \( i_x \in \{1, \ldots, m\} \) such that \( g^*_m(x) = a^*_i \cdot x + b^*_i \). Hence

\[
g^*_m(x) - f_{0,m}(x) = (a^*_i - a_{i_x}) \cdot x + (b^*_i - b_{i_x}) \leq \|a^*_i - a_{i_x}\|_2 \|x\|_2 + |b^*_i - b_{i_x}|
\]

\[
\leq \frac{\delta_{n,m}}{2\sqrt{c_3d}} \cdot \sqrt{d} + \frac{\delta_{n,m}}{2\sqrt{c_3}} = \frac{\delta_{n,m}}{\sqrt{c_3}}.
\]

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

(B.4)

\[
\Pi_{n,m} \left( \{ f \in \mathcal{F}_m : \ell_n^2(f, f_{0,m}) \leq \delta_{n,m}/c_3 \} \right)
\]

\[
\geq \Pi_G \left( \cap_{i=1}^m \left\{ (a^*_i, b^*_i) : a^*_i \in B_d(a_i, \delta_{n,m}/2\sqrt{c_3d}), b^*_i \in B_1(b_i, \delta_{n,m}/2\sqrt{c_3}) \right\} \right)
\]

\[
= \prod_{i=1}^m \Pi_{g^{ad}}(B_d(a_i, \delta_{n,m}/2\sqrt{c_3d})) \cdot \Pi_g(B_1(b_i, \delta_{n,m}/2\sqrt{c_3}))
\]

\[
\geq \prod_{i=1}^m g(||a_i||_{\infty} + 1)^d \cdot g(|b_i| + 1) \cdot \left( \frac{\delta_{n,m}}{4c_3d} \land 1 \right)^d v_d \left( \frac{\delta_{n,m}}{\sqrt{4c_3}} \land 1 \right)
\]

\[
\geq \exp \left( -2m(d + 1) \log (\tau_{m,g}^{-1} \lor 1) - m(d + 1) \log \left( \frac{\sqrt{4c_3d}}{\delta_{n,m}} \lor 1 \right) - \frac{1}{2} md \log d \right)
\]

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

(B.5)

\[
\text{max} \left\{ 2m(d + 1) \log (\tau_{m,g}^{-1} \lor 1), m(d + 1) \log \left( \frac{\sqrt{4c_3d}}{\delta_{n,m}} \lor 1 \right) \right\} \leq \frac{d}{2\eta} \log n \cdot m \log 3m,
\]

the claim follows by verifying (4.8) implies (B.5) since \( \sqrt{4c_3d}\delta_{n,m}^{-1} \leq \sqrt{n}, \) the second term is bounded by \( md \log n. \) The inequality follows by noting \( \eta < 1/4. \)

**Lemma 24.** For \( n \geq 2, \) (2.5) is satisfied for \( c = 1 \) and \( \gamma = 2. \)
Proof. For fixed $n \geq 2$ and $\eta > 0$, write $n\delta_{n,m}^2 = c \log n (m \log 3m)$ throughout the proof, where $c \geq 2/c_7$. Then for any $\alpha \geq c_7/2$ and $h \geq 1$, since $\log(3m') \geq \log(3hm) \geq \log(3m)$ for any $m' \geq hm$, we have

$$\sum_{m' \geq hm} e^{-\alpha\delta_{n,m'}^2} \leq \sum_{m' \geq hm} e^{-\alpha cm'(\log n \log 3m)} = \frac{e^{-\alpha cm \log n \log 3m}}{1 - e^{-\alpha c \log n \log 3m}} \leq 2e^{-\alpha hm\delta_{n,m}^2}. $$

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

Proof of Theorem 9. This is a direct consequence of Theorems 1 and 2, Lemma 23 and 24, combined with Proposition 1. □

B.3. Proof of Theorem 10.

Lemma 25. Let $n \geq 2$, then for any $g \in F_{(s,m)}$,

$$\log \mathcal{N}(c_5\varepsilon, \{f \in F_{(s,m)} : \ell_n(f, g) \leq 2\varepsilon\}, \ell_n) \leq 2 \log(6/c_5)(s \log(ep) + m \log(en)).$$

Proof. The proof borrows notation from the proof of Lemma 17. Further let $\mathcal{S}_s$ denote all subsets of $\{1, \ldots, p\}$ with cardinality at most $s$. Then the entropy in question can be bounded by

$$\log \left[ \binom{p}{s} \binom{n}{m-1} \max_{S \in \mathcal{S}_s, Q \in \mathcal{Q}_m} \mathcal{N}(c_5\varepsilon, \{f \in F_{(s,m),(S,Q)} : \ell_n(f, g) \leq 2\varepsilon\}, \ell_n) \right] \leq s \log(ep) + m \log(en)$$

$$+ \max_{S \in \mathcal{S}_s, Q \in \mathcal{Q}_m} \log \mathcal{N}(c_5\sqrt{n}\varepsilon, \{\gamma \in \mathcal{P}_{n,(S,Q)} : \|\gamma - g\|_2 \leq 2\sqrt{n}\varepsilon, \|\cdot\|_2\})$$

where $\mathcal{P}_{n,(S,Q)} \equiv \{(x_i^T\beta + u(z_i))_{i=1}^n \in \mathbb{R}^n : \supp(\beta) = S, u \text{ is constant on the partitions of } Q\}$ is contained in a linear subspace of dimension no more than $s + m$. Now similar arguments as in Lemma 17 shows that the entropy term in the above display can be bounded by $(s + m) \log(6/c_5)$, proving the claim.

Hence we can take $\delta_{n,(s,m)}^2 \equiv \left( \frac{4 \log(6/c_5)}{c_7} \wedge \frac{1}{\eta} \right) \frac{s \log(ep) + m \log(en)}{n}$.

Lemma 26. (2.5) holds with $\epsilon = 1$ and $\gamma = 1$.

Proof. For the first condition of (2.5), note that for any $h \geq 1$, with $c' \equiv \frac{4 \log(6/c_5)}{c_7} \wedge \frac{1}{\eta}$ in the proof, for any $\alpha \geq c_7/2$, $\alpha c' \geq 2 \log(6/c_5) \geq 2$ since $c_5 \leq 1/4$,

$$\sum_{(s',m') \geq (hs,hm)} e^{-\alpha\delta_{n,(s',m')}^2} \leq \sum_{s' \geq hs} e^{-\alpha c's \log(ep)} \sum_{m' \geq hm} e^{-\alpha c'm \log(en)} \leq (1 - e^{-\alpha c'})^{-2} e^{-\alpha \delta_{(hs,hm)}^2.}$$

The second condition of (2.5) is easy to verify by our choices of $\epsilon, \gamma$. □

Lemma 27. Suppose (4.11) holds. Then (P2) in Assumption C holds.
Proof. Using notation in Lemma 25, (B.6)
\[
\Pi_{n,(s,m)} \left( \{ f \in \mathcal{F}_{(s,m)} : \ell_n^2(f, f_{0,(s,m)}) \leq \delta_{n,(s,m)}/c_3 \} \right)
\geq \left( \frac{p}{s} \right)^{-1} \left( \frac{n}{m-1} \right)^{-1} \Pi_{g^\otimes s \otimes \bar{g}^m} \left( \{ f \in \mathcal{F}_{(s,m),(S_0,Q_0)} : \ell_n^2(f, f_{0,(s,m)}) \leq \delta_{n,(s,m)}/c_3 \} \right)
\]
where \( f_{0,m} \in \mathcal{F}_{(s,m),(S_0,Q_0)}. \) Let \( c' \equiv \frac{4\log(6/c_3)}{n} \) throughout the proof, and \( \delta_{n,s}^2 \equiv c's \log(ep)/n \) and \( \delta_{n,m}^2 \equiv c'm \log(en)/n. \) To bound the prior mass of the above display from below, it suffices to bound the product of the following two terms:
(B.7)
\[
\pi_s \equiv \Pi_{g^\otimes s} \left( \{ \beta \in B_0(s) : \beta_{S_0}^c = 0, \ell_n^2(h_{\beta}, h_{\beta_0,s}) \leq \delta_{n,s}^2/2c_3 \} \right),
\]
\[
\pi_m \equiv \Pi_{\bar{g}^m} \left( \{ u \in \mathcal{U}_{m,Q_0} : \ell_n^2(u, u_{0,m}) \leq \delta_{n,m}^2/2c_3 \} \right).
\]
The first term equals
(B.8)
\[
\Pi_{g^\otimes s} \left( \{ \beta \in B_0(s) : \beta_{S_0}^c = 0, \| X \beta - X \beta_{0,s} \|_2 \leq \sqrt{n} \delta_{n,s}/\sqrt{2c_3} \} \right)
\]
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 \( \sigma_{\Sigma} \) 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 in the last line of (B.8) is supported on \( \mathbb{R}_{S_0}^p \) and hence can be further bounded from below by \( \tau_{s,g} \left( \frac{1}{\sigma_{\Sigma}} \cdot \frac{\delta_{n,s}}{\sqrt{2c_3}} \land 1 \right)^s v_s \) where \( v_s = \text{vol}(B_s(0,1)). \) Hence
(B.9)
\[
\pi_s \geq (\tau_{s,g} \land 1)^s \left( \frac{1}{\sigma_{\Sigma}} \cdot \frac{\delta_{n,s}}{\sqrt{2c_3}} \land 1 \right)^s v_s
\]
\[
\geq \exp \left( - \frac{1}{2} s \log s - s \log (\tau_{s,g} \land 1) - \frac{s}{2} \log \left( \frac{2c_3 \sigma_{\Sigma}^2}{\delta_{n,s}^2} \land 1 \right) \right),
\]
where in the last inequality we used that \( v_s \geq (1/\sqrt{3})^s. \) By repeating the arguments in the proof of Lemma 18, we have
(B.10)
\[
\pi_m \geq \exp \left( - m \log (\tau_{m,g}^{-1} \land 1) - m \log \left( \frac{2c_3}{\delta_{n,m}^2} \land 1 \right) \right).
\]
Combining (B.6), (B.7), (B.9) and (B.10), we see that
(B.11)
\[
\Pi_{n,(s,m)} \left( \{ f \in \mathcal{F}_{(s,m)} : \ell_n^2(f, f_{0,(s,m)}) \leq \delta_{n,(s,m)}/c_3 \} \right)
\geq \exp \left( -2s \log(ep) - m \log(en) - s \log (\tau_{s,g}^{-1} \land 1) - m \log (\tau_{m,g}^{-1} \land 1) \right)
\]
\[
\times \exp \left( - \frac{s}{2} \log \left( \frac{2c_3 \sigma_{\Sigma}^2}{\delta_{n,s}^2} \land 1 \right) - \frac{m}{2} \log \left( \frac{2c_3}{\delta_{n,m}^2} \land 1 \right) \right).
\]
In order that the right side of the above display bounded from below by $\exp(-2n\delta^2_{n,(s,m)})$, we only need to require that
\[
\begin{align*}
\min \left\{ e^{-s \log(\tau_{-1}^1/v1)}, e^{-s \log\left(\frac{\sqrt{\Sigma_n}}{\delta_{n,m}}\right) v1} \right\} & \geq e^{-\frac{1}{2n} s \log(ep)}, \\
\min \left\{ e^{-m \log(\tau_{-1}^1/v1)}, e^{-m \log\left(\frac{\sqrt{\Sigma_n}}{\delta_{n,m}}\right) v1} \right\} & \geq e^{-\frac{1}{2n} m \log(en)}.
\end{align*}
\]

The first terms in the above two lines lead to (4.11). The other terms in the above two lines do not contribute by noting that $2c_3/\delta^2_{n,m} \leq \frac{2c_3\epsilon^7}{8 \log(6/c_3)} n \leq (1/2)n \leq en$ since $c_3 = 1$ (in Gaussian regression model) and $c_7 \in (0,1)$, while $2c_3\sigma^2_{\Sigma}/\delta^2_{n,s} \leq \sigma^2_{\Sigma} n \leq pn \leq p^2$ and $\eta < 1/4$.

**Proof of Theorem 10.** The claim of the theorem follows by Theorems 1 and 2, Proposition 1 and Lemmas 25-27. □

APPENDIX C. PROOF OF AUXILIARY LEMMAS IN SECTION 5

**Proof of Lemma 9.** Let $F_j := \{ f \in F : j \varepsilon < d_n(f, f_0) \leq 2j \varepsilon \}$ and $G_j \subset F_j$ be the collection of functions that form a minimal $c_5 j \varepsilon$ covering set of $F_j$ under the metric $d_n$. Then by assumption $|G_j| \leq N(j \varepsilon)$. Furthermore, for each $g \in G_j$, it follows by Lemma 1 that there exists some test $\omega_{n,j,g}$ such that
\[
(C.1) \quad \sup_{f \in F : d_n(f, g) \leq c_5 j \varepsilon} P^{(n)}_{f_0}(\omega_{n,j,g} + P^{(n)}_f(1 - \omega_{n,j,g})) \leq c_6 e^{-c_7 n d^2_{n}(f_0,g)}.
\]

Recall that $g \in G_j \subset F_j$, then $d_n(f_0, g) \geq j \varepsilon$. Hence the indexing set above contains $\{ f \in F : d_n(f, g) \leq c_5 j \varepsilon \}$. Now we see that
\[
P^{(n)}_{f_0}(\omega_{n,j,g}) \leq c_6 e^{-c_7 n j^2 \varepsilon^2}, \quad \sup_{f \in F : d_n(f, g) \leq c_5 j \varepsilon} P^{(n)}_f(1 - \omega_{n,j,g}) \leq c_6 e^{-c_7 n j^2 \varepsilon^2}.
\]

Consider the global test $\phi_n := \sup_{j \geq 1} \max_{g \in G_j} \omega_{n,j,g}$, then
\[
P^{(n)}_{f_0}(\phi_n) \leq P^{(n)}_{f_0} \left( \sum_{j \geq 1} \sum_{g \in G_j} \omega_{n,j,g} \leq c_6 \sum_{j \geq 1} N(j \varepsilon) e^{-c_7 n j^2 \varepsilon^2} \right.
\leq c_6 N(\varepsilon) \sum_{j \geq 1} e^{-c_7 n j^2 \varepsilon^2} \leq c_6 N(\varepsilon) e^{-c_7 n \varepsilon^2} \cdot (1 - e^{-c_7 n \varepsilon^2})^{-1}.
\]

On the other hand, for any $f \in F$ such that $d_n(f, f_0) \geq \varepsilon$, there exists some $j^* \geq 1$ and some $g_{j^*} \in G_{j^*}$ such that $d_n(f, g_{j^*}) \leq j^* c_5 \varepsilon$. Hence
\[
P^{(n)}_f(1 - \phi_n) \leq P^{(n)}_f \left( 1 - \omega_{n,j^*, g_{j^*}} \right) \leq c_6 e^{-c_7 n (j^*)^2 \varepsilon^2} \leq c_6 e^{-c_7 n \varepsilon^2}.
\]

The right hand side of the above display is independent of individual $f \in F$ such that $d_n(f, f_0) \geq \varepsilon$ and hence the claim follows. □
Proof of Lemma 10. By Jensen’s inequality, the left side of (5.6) is bounded by
\[
P_{f_0}^{(n)} \left\{ \int \left( \log \frac{p_{f_0}^{(n)}}{p_f^{(n)}} - P_{f_0}^{(n)} \log \frac{p_{f_0}^{(n)}}{p_f^{(n)}} \right) d\Pi(f) \right\} \geq (C + c_3)n\varepsilon^2 - c_3n \int d_n^2(f_0, f) d\Pi(f) \}
\[
\leq P_{f_0}^{(n)} \left[ \int \left( \log \frac{p_{f_0}^{(n)}}{p_f^{(n)}} - P_{f_0}^{(n)} \log \frac{p_{f_0}^{(n)}}{p_f^{(n)}} \right) d\Pi(f) \right] \geq Cn\varepsilon^2 \]
\[
\leq \exp(-C\lambda n\varepsilon^2) \cdot c_1 P_{f_0}^{(n)} \exp \left( \lambda \int \left( \log \frac{p_{f_0}^{(n)}}{p_f^{(n)}} - P_{f_0}^{(n)} \log \frac{p_{f_0}^{(n)}}{p_f^{(n)}} \right) d\Pi(f) \right) .
\]

Using Jensen’s inequality, the last term in the right side of the above display can be further bounded by
\[
P_{f_0}^{(n)} \int \exp \left[ \lambda \left( \log \frac{p_{f_0}^{(n)}}{p_f^{(n)}} - P_{f_0}^{(n)} \log \frac{p_{f_0}^{(n)}}{p_f^{(n)}} \right) \right] d\Pi(f) \leq \int e^{\psi_{\kappa^2 n^2, \kappa} \lambda} d\Pi(f)
\]
where the last inequality follows from Fubini’s theorem and Assumption A. Now the condition on the prior \( \Pi \) entails that
\[
P_{f_0}^{(n)} \left( \int \frac{p_f^{(n)}}{p_{f_0}^{(n)}} d\Pi(f) \right) \leq e^{-(C+c_3)n\varepsilon^2} \leq c_1 \exp \left( -C\lambda n\varepsilon^2 + \psi_{\kappa^2 n^2, \kappa} \right) .
\]
The claim follows by choosing \( \lambda > 0 \) small enough depending on \( C, \kappa \). \( \square \)

Proof of Proposition 3. We may assume without loss of generality that \( d_n(f_0, f_{0,m}) < \infty \) so that \( \tilde{m} \) is well-defined since \( |\mathcal{M}| = \infty \) and (2.5). By definition we have \( \delta_{n,m} \geq d_n(f_0, f_{0,m}) \) and \( \delta_{n,m-1} < d_n(f_0, f_{0,m}) \). In this case, the global test can be constructed via \( \tilde{\phi}_n := \sup_{n' \in \mathcal{M}^2} \phi_{n,m'} \). Then analogously to (5.9) and (5.10), for any random variable \( U \in [0,1] \),
\[
P_{f_0,m}^{(n)} U \cdot \tilde{\phi}_n \leq 4c_6e^{-(c_7/2^2)\varepsilon^2}\delta^2_{n,m},
\]
\[
\sup_{f \in \mathcal{F}, h : d_n(f, f_{0,m}) \geq (\varepsilon/2)^2 \delta^2_{n,m}} P_f^{(n)} (1 - \tilde{\phi}_n) \leq 2c_6e^{-(c_7/\varepsilon^2)\varepsilon^2}\delta^2_{n,m}.
\]

Similar to (5.11), there exists an event \( \tilde{\mathcal{E}}_n \) with \( P_{f_0,m}^{(n)} (\tilde{\mathcal{E}}_n) \leq c_1 e^{-c_7^2 \varepsilon^2 \delta^2_{n,m}/8c_3 \varepsilon^2} \) and the following is true on the event \( \tilde{\mathcal{E}}_n \):
\[
\int \prod_{i=1}^n \frac{p_f}{p_{f_0,m}} d\Pi(f) \geq \lambda_{n,m} e^{-c_7 \varepsilon^2 \delta^2_{n,m}/8c_3 \varepsilon^2} \Pi_{n,m} \left( \{ f \in \mathcal{F}_m : d_n(f, f_{0,m}) \leq c_7 \varepsilon^2 \delta^2_{n,m}/8c_3 \varepsilon^2 \} \right).
\]
Repeating the reasoning in (5.12), (5.13) and (5.14) we see that (C.4)

\[ P_{f_{0,m}}^{(n)} \Pi_n \left( f \in \mathcal{F} : d_n^2(f, f_{0,m}) > c^4(2 jh)^\gamma d_n^2(f_{0,m}) \right) X^{(n)} \left( 1 - \tilde{\phi}_n \right) 1_{\tilde{\xi}_n} \]

\[ \leq e^{c_7 n^2 jh \delta_{n,m}^2 / 4c^2} \lambda_{n,m} \Pi_{n,m} \left( \left\{ f \in \mathcal{F}_m : d_n^2(f, f_{0,m}) \leq c_7 jh \delta_{n,m}^2 / 8c_3 \right\} \right) \]

\[ \times \int_{f \in \mathcal{F} : d_n^2(f, f_{0,m}) > c^4(2 jh)^\gamma d_n^2(f_{0,m})} P_f^{(n)}(1 - \tilde{\phi}_n) d\Pi(f) \]

\[ \leq (\cdots) \times \sup_{f \in \mathcal{F}_{jmh_n}} P_f^{(n)}(1 - \tilde{\phi}_n) + \Pi (\mathcal{F} \setminus \mathcal{F}_{jmh_n}) \]

\[ \leq Ce^{- (c_7/4c^2) njh \delta_{n,m}^2} \].

Here the third line is valid since \( c^4(2 jh)^\gamma d_n^2(f_{0,m}) > c^4(2 jh)^\gamma \delta_{n,m}^2 \) by the right side of (2.5), which entails \( \delta_{n,m}^2 \leq c^2 \gamma \delta_{n,m}^2 \). The fourth line uses (C.2) and assumption (P1), together with the fact that \( \delta_{n,m} \geq \delta_{n,m} \). (5.2) follows from (C.2), probability estimate for \( \mathcal{E}_n^c \) and (C.4).

**Acknowledgements**

The author is indebted to Chao Gao for his numerous suggestions that lead to a substantially improved version of the paper. He thanks Johannes Schmidt-Hieber for very helpful comments on 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**

[1] R. a. Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.*, 13:no. 34, 1000–1034, 2008.
[2] P. Alquier, V. Cottet, N. Chopin, and J. Rousseau. Bayesian matrix completion: prior specification. *arXiv preprint arXiv:1406.1440*, 2014.
[3] S. Banerjee and S. Ghosal. Posterior convergence rates for estimating large precision matrices using graphical models. *Electron. J. Stat.*, 8(2):2111–2137, 2014.
[4] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.
[5] P. C. Bellec. Sharp oracle inequalities for least squares estimators in shape restricted regression. *arXiv preprint arXiv:1510.08029*, 2015.
[6] L. Birgé. Approximation dans les espaces métriques et théorie de l’estimation. *Z. Wahrsch. Verw. Gebiete*, 65(2):181–237, 1983.
[7] L. Birgé. Robust testing for independent nonidentically distributed variables and Markov chains. In *Specifying statistical models (Louvain-la-Neuve, 1981)*, volume 16 of *Lect. Notes Stat.*, pages 134–162. Springer, New York, 1983.
[8] L. Birgé. Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Ann. Inst. H. Poincaré Probab. Statist.*, 42(3):273–325, 2006.
[9] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

[10] P. Bühlmann and S. van de Geer. *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg, 2011. Methods, theory and applications.

[11] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.*, 35(4):1674–1697, 2007.

[12] E. J. Candès and Y. Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Trans. Inform. Theory*, 57(4):2342–2359, 2011.

[13] E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Trans. Inform. Theory*, 51(12):4203–4215, 2005.

[14] I. Castillo. On Bayesian supremum norm contraction rates. *Ann. Statist.*, 42(5):2058–2091, 2014.

[15] I. Castillo, J. Schmidt-Hieber, and A. van der Vaart. Bayesian linear regression with sparse priors. *Ann. Statist.*, 43(5):1986–2018, 2015.

[16] I. Castillo and A. van der Vaart. Needles and straw in a haystack: posterior concentration for possibly sparse sequences. *Ann. Statist.*, 40(4):2069–2101, 2012.

[17] S. Chatterjee, A. Guntuboyina, and B. Sen. On risk bounds in isotonic and other shape restricted regression problems. *Ann. Statist.*, 43(4):1774–1800, 2015.

[18] C. Gao, A. W. van der Vaart, and H. H. Zhou. A general framework for bayes structured linear models. *arXiv preprint arXiv:1506.02174*, 2015.

[19] C. Gao and H. H. Zhou. Rate-optimal posterior contraction for sparse PCA. *Ann. Statist.*, 43(2):785–818, 2015.

[20] C. Gao and H. H. Zhou. Rate exact Bayesian adaptation with modified block priors. *Ann. Statist.*, 44(2):318–345, 2016.

[21] S. Ghosal, J. K. Ghosh, and A. W. van der Vaart. Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531, 2000.

[22] S. Ghosal, J. Lember, and A. van der Vaart. Nonparametric Bayesian model selection and averaging. *Electron. J. Stat.*, 2:63–89, 2008.

[23] S. Ghosal and A. van der Vaart. Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist.*, 35(1):192–223, 2007.

[24] S. Ghosal and A. van der Vaart. Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.*, 35(2):697–723, 2007.

[25] A. Guntuboyina. Optimal rates of convergence for convex set estimation from support functions. *Ann. Statist.*, 40(1):385–411, 2012.

[26] Q. Han and J. A. Wellner. Multivariate convex regression: global risk bounds and adaptation. *arXiv preprint arXiv:1601.06844*, 2016.

[27] L. A. Hannah and D. B. Dunson. Bayesian nonparametric multivariate convex regression. *arXiv preprint arXiv:1109.0322*, 2011.

[28] M. Hoffmann, J. Rousseau, and J. Schmidt-Hieber. On adaptive posterior concentration rates. *Ann. Statist.*, 43(5):2259–2295, 2015.

[29] C. Holmes and N. Heard. Generalized monotonic regression using random change points. *Statistics in Medicine*, 22(4):623–638, 2003.

[30] B. J. K. Kleijn and A. W. van der Vaart. Misspecification in infinite-dimensional Bayesian statistics. *Ann. Statist.*, 34(2):837–877, 2006.

[31] L. Le Cam. Convergence of estimates under dimensionality restrictions. *Ann. Statist.*, 1:38–53, 1973.

[32] L. Le Cam. On local and global properties in the theory of asymptotic normality of experiments. pages 13–54, 1975.

[33] L. Le Cam. *Asymptotic methods in statistical decision theory*. Springer Series in Statistics. Springer-Verlag, New York, 1986.
[34] T. T. Mai and P. Alquier. A Bayesian approach for noisy matrix completion: optimal rate under general sampling distribution. *Electron. J. Stat.*, 9(1):823–841, 2015.

[35] E. Mariucci, K. Ray, and B. Szabo. A bayesian nonparametric approach to log-concave density estimation. *arXiv preprint arXiv:1703.09531*, 2017.

[36] P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. 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.

[37] F. Merlevède, M. Peligrad, and E. Rio. Bernstein inequality and moderate deviations under strong mixing conditions. In *High dimensional probability V: the Luminy volume*, volume 5 of *Inst. Math. Stat. (IMS) Collect.*, pages 273–292. Inst. Math. Statist., Beachwood, OH, 2009.

[38] D. Pati, A. Bhattacharya, N. S. Pillai, and D. Dunson. Posterior contraction in sparse Bayesian factor models for massive covariance matrices. *Ann. Statist.*, 42(3):1102–1130, 2014.

[39] D. 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.

[40] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, 52(3):471–501, 2010.

[41] A. Rohde and A. B. Tsybakov. Estimation of high-dimensional low-rank matrices. *Ann. Statist.*, 39(2):887–930, 2011.

[42] J. Rousseau. Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density. *Ann. Statist.*, 38(1):146–180, 2010.

[43] X. Shen and L. Wasserman. Rates of convergence of posterior distributions. *Ann. Statist.*, 29(3):687–714, 2001.

[44] A. B. Tsybakov. Aggregation and minimax optimality in high-dimensional estimation. In *Proceedings of the International Congress of Mathematicians*, pages 225–246, 2014.

[45] A. W. van der Vaart and J. H. van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.*, 36(3):1435–1463, 2008.

[46] A. W. van der Vaart and J. H. van Zanten. Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.*, 37(5B):2655–2675, 2009.

[47] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996.

[48] Y. Yang and D. Pati. Bayesian model selection consistency and oracle inequality with intractable marginal likelihood. *arXiv preprint arXiv:1701.00311*, 2017.

[49] W. W. Yoo and S. Ghosal. Supremum norm posterior contraction and credible sets for nonparametric multivariate regression. *Ann. Statist.*, 44(3):1069–1102, 2016.

[50] Z. Yu, M. Levine, and G. Cheng. Minimax optimal estimation in high dimensional semiparametric models. *arXiv preprint arXiv:1612.05906*, 2016.

[51] M. Yuan and D.-X. Zhou. Minimax optimal rates of estimation in high dimensional additive models: Universal phase transition. *arXiv preprint arXiv:1503.02817*, 2015.

(Q. Han) DEPARTMENT OF STATISTICS, BOX 354322, UNIVERSITY OF WASHINGTON, SEATTLE, WA 98195-4322, USA.

E-mail address: royhan@uw.edu