INAUGURAL - DISSERTATION
zur
Erlangung der Doktorwürde
der
Naturwissenschaftlich-Mathematischen Gesamtfakultät
der
Ruprecht-Karls-Universität
Heidelberg

vorgelegt von
Diplom-Mathematiker Martin Wahl
aus Altenkirchen

Tag der mündlichen Prüfung: .....

On the Mod-Gaussian Convergence of a Sum over Primes and Estimation of Components and Variable Selection in High-Dimensional Complex Models

Betreuer:
Prof. Dr. Enno Mammen
Prof. Dr. Ashkan Nikeghbali

Zusammenfassung

Diese Dissertation betrachtet zwei verschiedene Fragestellungen, eine aus der probabilistischen Zahlentheorie und eine aus der mathematischen Statistik.

In Kapitel 1 untersuchen wir die Verteilung der Werte des Logarithmus der Riemannschen Zeta-Funktion auf der kritischen Geraden. Wir beweisen Mod-Gaußsche Konvergenz für ein Dirichlet-Polynom, welches $\operatorname{Im}\log\zeta(1/2+it)$ approximiert. Dieses Dirichlet-Polynom ist lang genug um unter anderem einen neuen Beweis für Selbergs zentralen Grenzwertsatz mit explizitem Fehlerterm zu erhalten. Unter der Annahme der Riemannschen Hypothese und indem wir die Theorie der Riemannschen Zeta-Funktion anwenden, zeigen wir, dass sich diese Mod-Gaußsche Konvergenz auf die komplexe Zahlenebene erweitern lässt. Mit Hilfe dieser stärkeren Konvergenz und der Theorie Großer Abweichungen können wir beweisen, dass $\operatorname{Im}\log\zeta(1/2+it)$ auf der kritischen Geraden ein Prinzip großer Abweichungen erfüllt.

In Kapitel 2 betrachten wir ein nichtparametrisches Regressionsmodell $Y = f_1(X_1)+f_2(X_2)+\epsilon$ und beschäftigen uns mit dem Problem die Funktion f_1 zu schätzen. Den Term $f_2(X_2)$ sehen wir dabei als Störterm an, welcher wesentlich komplexer sein kann als $f_1(X_1)$. Unter minimalen Annahmen beweisen wir mehrere nichtasymptotische obere Schranken für das $L^2(\mathbb{P}^X)$-Risiko unserer Schätzer von f_1. Unsere Herangehensweise ist geometrisch und basiert auf Betrachtungen in Hilberträumen. Es zeigt sich, dass die Güte unserer Schätzer eng verknüpft ist mit geometrischen Größen aus der Theorie der Hilberträume, wie zum Beispiel den minimalen Winkeln und den Hilbert-Schmidt Normen. Mit Hilfe unserer Resultate lassen sich allgemeine Bedingungen aufstellen, unter denen unsere Schätzer von f_1 (in erster Ordnung) dieselbe scharfe obere Schranke besitzen wie die entsprechenden Schätzer von f_1 in dem Modell $Y = f_1(X_1) + \epsilon$. Als Anwendung betrachten wir unter anderem ein additives Modell, in dem die Anzahl der Kovariablen sehr groß oder die Glattheit der Störfunktionen sehr klein ist.

In Kapitel 3 und 4 betrachten wir das Problem der Variablenwahl in hochdimensionalen additiven Regressionsmodellen. Dabei interessieren wir uns für den Fall, dass die Komponenten in nichtparametrischen Funktionenklassen enthalten sind. Wir konstruieren ein Verfahren, in welchem die Normen der Projektionen der Daten auf verschiedene additive Unterräume

verglichen werden. Unsere Hauptresultate sind allgemeine Konzentrationsungleichungen, die zu Bedingungen führen, unter welchen konsistente Variablenwahl möglich ist. Unsere Herangehensweise ist wie in Kapitel 2 geometrisch und beruht auf Betrachtungen in Hilberträumen. Des Weiteren wenden wir Methoden aus der Theorie der Zufallsmatrizen an um den Übergang von der $L^2(\mathbb{P}^X)$-Norm zur empirischen Norm zu erreichen. Als Anwendung unserer Resulate stellen wir Bedingungen auf, unter denen eine einzelne Komponente mit derselben nichtasymptotischen optimalen Konvergenzrate geschätzt werden kann wie in dem Fall, dass die anderen Komponenten bekannt sind. Zuletzt betrachten wir auch das verwandte und einfachere Modell des additiven Signals in weißem Rauschen und leiten optimale Bedingungen für die Variablenwahl her.

Abstract

This thesis considers two problems, one in probabilistic number theory and another in mathematical statistics.

In Chapter 1, we study the distribution of values taken by the logarithm of the Riemann zeta-function on the critical line. We prove mod-Gaussian convergence for a Dirichlet polynomial which approximates $\operatorname{Im} \log \zeta(1/2+it)$. This Dirichlet polynomial is sufficiently long to deduce Selberg's central limit theorem with an explicit error term. Moreover, assuming the Riemann hypothesis, we apply the theory of the Riemann zeta-function to extend this mod-Gaussian convergence to the complex plane. From this and the theory of large deviations, we obtain that $\operatorname{Im} \log \zeta(1/2+it)$ satisfies a large deviation principle on the critical line. Results about the moments of the Riemann zeta-function follow.

In Chapter 2, we consider the nonparametric random regression model $Y = f_1(X_1) + f_2(X_2) + \epsilon$ and address the problem of estimating the function f_1. The term $f_2(X_2)$ is regarded as a nuisance term which can be considerably more complex than $f_1(X_1)$. Under minimal assumptions, we prove several nonasymptotic $L^2(\mathbb{P}^X)$-risk bounds for our estimators of f_1. Our approach is geometric and based on considerations in Hilbert spaces. It shows that the performance of our estimators is closely related to geometric quantities from the theory of Hilbert spaces, such as minimal angles and Hilbert-Schmidt norms. Our results establish general conditions under which the estimators of f_1 have up to first order the same sharp upper bound as the corresponding estimators of f_1 in the model $Y = f_1(X_1) + \epsilon$. As an example we apply the results to an additive model in which the number of components is very large or in which the nuisance components are considerably less smooth than f_1.

In Chapter 3 and 4, we consider the problem of variable selection in high-dimensional sparse additive models. We focus on the case that the components belong to nonparametric classes of functions. The proposed method consists of comparing the norms of the projections of the data onto various additive subspaces. Under minimal geometric assumptions, we prove concentration inequalities which lead to general conditions under which consistent variable selection is possible. Again, our approach is based on geometric considerations in Hilbert spaces. Moreover, we apply recent techniques from the theory of structured random matrices to accomplish the transition from the $L^2(\mathbb{P}^X)$-norm to the empirical norm. As an application,

we establish conditions under which a single component can be estimated with the rate of convergence corresponding to the situation in which the other components are known. Finally, we derive optimal conditions for variable selection in the related and more simple additive Gaussian white noise model.

Acknowledgments

First and foremost, I would like to thank both my advisors, Prof. Enno Mammen and Prof. Ashkan Nikeghbali, for introducing me to interesting topics in probability theory, number theory, and mathematical statistics and for always supporting me during the preparation of my thesis.

I also would like to thank Prof. Kowalski for several discussions on Chapter 1.

I am very grateful to Prof. Enno Mammen for enabling me to attend several interesting meetings in mathematical statistics, such as those in Oberwolfach and Luminy.

Several parts of my work have been proofread by friends and colleagues. Thank you for that.

Finally, I would like to thank my wife Katja and my family for their encouragement and support.

Contents

Zusammenfassung v

Abstract vii

Acknowledgments ix

Chapter 1. On the mod-Gaussian convergence of a sum over primes 1
1. Introduction 1
2. Moments of a sum over primes 4
3. Bessel functions 6
4. Mod-convergence of a sum over primes 7
5. Mod-convergence in the complex plane 9
6. Proof of Corollary 1 13
7. Proof of Corollary 2 and 3 16

Chapter 2. A theory of nonparametric regression in the presence of complex nuisance components 19
1. Introduction 19
2. The framework 21
2.1. The model 21
2.2. The main assumption 22
2.3. The estimation procedure 23
3. Main results 24
3.1. A first risk bound 24
3.2. A refined risk bound 25
3.3. Regularity conditions on the design densities 28
4. Applications 30
4.1. The two-dimensional case 30
4.2. The multidimensional case 32
4.3. The additive model with Sobolev smoothness 33
4.4. The additive model with Hölder smoothness 36
4.5. The additive model with smooth design densities 37
5. Proof of Theorem 3 and 4 38
5.1. The finite sample geometry 38
5.2. Analysis of the variance via von Neumann's theorem 40
5.3. End of the proof of Theorem 3 46
5.4. End of the proof of Theorem 4 49

6. Proof of Theorem 5 and 6 51
6.1. The bias term revisited 51
6.2. End of proof of Theorem 5 and 6 54

Chapter 3. Variable selection in high-dimensional additive models 57
1. Introduction 57
2. The main result 59
2.1. The variable selection problem 59
2.2. The main assumption 59
2.3. The selection criterion 61
2.4. The main result and some consequences 63
3. Structured random matrices and the event $\mathcal{E}_{\delta,q^*}$ 65
3.1. Independent covariates and the RIP 65
3.2. A general upper bound for $\mathbb{P}(\mathcal{E}^c_{\delta,q^*})$ 66
3.3. Conditions for variable selection 68
3.4. Estimation of single components 68
4. Outline of the proof of Theorem 10 69
4.1. The finite sample geometry 69
4.2. End of the proof of Theorem 10 70
5. Proofs 71
5.1. Proof of Remark 11 71
5.2. Proof of Lemma 12 71
5.3. Proof of Lemma 13 72
5.4. Proof of Lemma 14 72
5.5. Proof of Proposition 11 73
5.6. Proof of Equation (3.2) 75
5.7. Proof of Proposition 13 75
5.8. Proof of Lemma 16 76

Chapter 4. Optimal conditions for variable selection in additive Gaussian white noise models 79
1. Introduction 79
2. The main result 80
2.1. The Gaussian white noise framework 80
2.2. Model selection via penalization 80
2.3. A general variable selection theorem 82
3. Optimal Conditions 82
3.1. Sufficient conditions 82
3.2. Necessary conditions 84
4. Proofs 85
4.1. Proof of Theorem 15 85
4.2. Proof of Proposition 14 89

Appendix A. Appendix to Chapter 1 93
Selberg's result 93
Mean value estimates 94

Large deviation theory 94

Appendix B. Appendix to Chapter 2 97
Proof of Lemma 2 97
A feasible estimator 97
Proof of Lemma 4 and 5 98
Proof of Corollary 11 99
A remark on Theorem 9 100
Proof of Lemma 6 101
Proof of Lemma 7 102
An alternative proof of Corollary 12 103
Proof of (5.16) 104

Appendix C. Appendix to Chapter 4 105
Proof of Lemma 20 105

Bibliography 107

CHAPTER 1

On the mod-Gaussian convergence of a sum over primes

1. Introduction

In this chapter we study the distribution of values taken by $\log\zeta(1/2+it)$. A breakthrough was achieved by Selberg who showed that as t varies in $[T,2T]$, the distribution of $(\operatorname{Re}\log\zeta(1/2+it),\operatorname{Im}\log\zeta(1/2+it))$ is approximately Gaussian, with independent components each having expectation 0 and variance $(\log\log T)/2$. More precisely, he proved a central limit theorem which, by the Lévy continuity theorem, is equivalent to the statement that

$$\frac{1}{T}\int_T^{2T} e^{iu\frac{\operatorname{Re}\log\zeta(1/2+it)}{\sqrt{(\log\log T)/2}}+iv\frac{\operatorname{Im}\log\zeta(1/2+it)}{\sqrt{(\log\log T)/2}}}\,dt \to e^{-u^2/2-v^2/2}, \tag{1.1}$$

as $T\to\infty$, for all real numbers u and v. For the case of $\operatorname{Im}\log\zeta(1/2+it)$ see [**66**], [**67**], and also the work of Ghosh [**31**]. The general case is investigated for instance in the book of Joyner [**39**]. Some of Selberg's more recent results, for example about the rate of convergence, can be found in [**68**] and the thesis of Tsang [**73**]. Initially, Selberg obtained the asymptotics of the joint moments which lead to (1.1) by the method of moments. A more effective approach, applied in our analysis, too, is treated in the work of Bombieri and Hejhal [**14**]. A central limit theorem for the sum over primes $(1/\sqrt{(\log\log x)/2})\sum_{p\le x}p^{-1/2-iU_T}$, U_T being random variables uniformly distributed on $[T,2T]$, $\log x=\log T/(\log\log T)^{1/4}$, follows from the mean value theorem of Montgomery and Vaughan and the method of moments. To complete the proof (see [**14**, Lemma 3 and Corallary]), they showed that the L^1-norm of $\log\zeta(1/2+iU_T)-\sum_{p\le x}p^{-1/2-iU_T}$ is sufficiently small.

The convergence in (1.1) is also a consequence of a conjecture on the behaviour of the moments of the Riemann zeta-function on the critical line (see, e.g., the work of Keating and Snaith [**41**] and the references therein). It asserts that

$$e^{(z_1^2+z_2^2)(\log\log T)/4}\frac{1}{T}\int_T^{2T} e^{iz_1\operatorname{Re}\log\zeta(1/2+it)+iz_2\operatorname{Im}\log\zeta(1/2+it)}dt \\ \to \Phi_g(z_1,z_2)\Phi_a(z_1,z_2) \quad as\ T\to\infty \tag{1.2}$$

locally uniformly for $z_1,z_2\in\mathbb{C}$ with $\operatorname{Re}(iz_1)>-1$ and analytic functions Φ_g, Φ_a (see [**45**, Conjecture 9] and also [**33**, Conjecture 1]). This type of

convergence was introduced in [**38**] where it is called mod-Gaussian convergence.

A precise form of the function Φ_g was conjectured by Keating and Snaith and is based on calculations in the theory of random matrices (see [**41**], [**45**, formula (18)]). The arithmetic factor Φ_a can be explained, e.g., by computing the characteristic function of $\sum_{n\leq x}\Lambda(n)/(n^{1/2+iU_T}\log n)$ (see [**33**, Theorem 2], where x has to be $O((\log T)^{2-\epsilon})$) or of the corresponding stochastic model (replace $\{p^{iU_T}\}_{p\in\mathbb{P}}$ by an independent sequence of random variables uniformly distributed on the unit circle, see [**45**, Example 4]).

In this chapter we further investigate the distribution of the sum over primes $\sum_{p\leq x}p^{-1/2-it}$ as t varies in $[T,2T]$ and its consequences on the distribution of values of the Riemann zeta-function on the critical line. Here, we will restrict ourselves to the case of $\operatorname{Im}\log\zeta(1/2+it)$. Note that some of the arguments cannot be applied to the case of $\operatorname{Re}\log\zeta(1/2+it)$. It is our first aim to establish mod-Gaussian convergence if x fulfills certain conditions. Precisely, in Section 4 we prove the following:

THEOREM 1. *Let $x=e^{\log T/N}$ and N such that $N/\log\log T\to\infty$ and $x\to\infty$ as $T\to\infty$. Then*

$$e^{u^2(\log\log x+\gamma)/4}\frac{1}{T}\int_T^{2T}e^{iu\sum_{p\leq x}\frac{\sin(t\log p)}{\sqrt{p}}}dt\to\Phi(u)\quad\text{as } T\to\infty \tag{1.3}$$

locally uniformly for $u\in\mathbb{R}$. Here, γ denotes Euler's constant and Φ is the analytic function given by

$$\Phi(u)=\prod_{p\in\mathbb{P}}\Big(1-\frac{1}{p}\Big)^{-u^2/4}J_0\Big(\frac{u}{\sqrt{p}}\Big), \tag{1.4}$$

where J_0 denotes the zeroth Bessel function (see, e.g., Section 3).

One interesting point of the result seems to be the size of x. It can be chosen large enough to obtain Selberg's central limit theorem with Selberg's explicit error term (see [**68**, Theorem 2] and Appendix A). Moreover, we obtain the following improvement of (1.1):

COROLLARY 1. *Assume RH. For T sufficiently large, we have*

$$\frac{1}{T}\int_T^{2T}e^{iv\frac{\operatorname{Im}\log\zeta(1/2+it)}{\sqrt{(\log\log T)/2}}}dt=e^{-v^2/2}+v^2O\Big(\frac{\log\log\log T}{\log\log T}\Big)+O(1/\log T)$$

uniformly for $|v|\leq\sqrt{\log\log T/\log\log\log T}$.

In Section 5 we deal with the question if the convergence in Theorem 1 can be extended to the complex plane. Assuming the Riemann hypothesis, we prove such a result for a weighted sum over primes.

THEOREM 2. *Assume RH. Let $x=e^{\log T/N}$ and N such that $x\to\infty$ and $N/\log\log T\to\infty$ as $T\to\infty$. Furthermore, let f be the function*

$f(u) = (\pi u/2)\cot(\pi u/2)$ and $\gamma_f = -0.1080\ldots$ be the constant defined by $\prod_{p\le x}(1-f^2(\log p/\log x)/p) = (e^{-\gamma_f}/\log x)(1+o(1))$. Then

$$e^{z^2(\log\log x+\gamma_f)/4}\frac{1}{T}\int_T^{2T} e^{iz\sum_{p\le x}\frac{\sin(t\log p)}{\sqrt{p}}f\left(\frac{\log p}{\log x}\right)}dt \to \Phi(z) \quad as\ T\to\infty$$

locally uniformly for $z\in\mathbb{C}$, where Φ is given by (1.4).

More general sums are possible as well (see [**32**, Lemma 1] and [**14**, Lemma 1]). For the evaluation of γ_f see [**32**, proof of Lemma 6].

The crucial step from Theorem 1 to Theorem 2 is an estimate of the exponential moments of the above sum. For this purpose let $x\le T^2$ and $h\in\mathbb{R}$. Assuming the Riemann hypothesis, we then show that there exist constants C, C', and C'' such that

$$\frac{1}{T}\int_T^{2T} e^{h\sum_{n\le x}\frac{\Lambda(n)}{\log n}\frac{\sin(t\log n)}{\sqrt{n}}f\left(\frac{\log n}{\log x}\right)}dt \le C'' e^{C|h|\frac{\log T}{\log x}+C'h^2\log\log T}.$$

Note that this inequality, which is almost a subgaussian bound, is valid beyond the range which is contained in Theorem 1 and Theorem 2.

We turn to the applications of Theorem 2. As described above, Theorem 1 can be used to obtain results in connection with the central limit theorem. In addition, Theorem 2 yields large deviations results. Applying the Gärtner-Ellis theorem and Theorem 2, one obtains a large deviation principle (see [**22**, chapter 1.2] or Appendix A for the definition of the large deviation principle) from which we will deduce the following two Corollaries.

Corollary 2. *Assume RH. Let U_T be random variables uniformly distributed on $[T,2T]$. Then the family $(1/((\log\log T)/2))\operatorname{Im}\log\zeta(1/2+iU_T)$ satisfies the large deviation principle with the speed $1/((\log\log T)/2)$ and the rate function $I(h)=h^2/2$. For instance,*

$$\frac{1}{(\log\log T)/2}\log\Big(\frac{1}{T}\lambda(\{t\in[T,2T]:\operatorname{Im}\log\zeta(1/2+it)\ge h(\log\log T)/2\})\Big) \to -h^2/2 \quad as\ T\to\infty, \quad (1.5)$$

where $h>0$ and λ denotes the Lebesgue measure.

Corollary 3. *Assume RH. Let $h\in\mathbb{R}$. Then*

$$\frac{1}{(\log\log T)/2}\log\left(\frac{1}{T}\int_T^{2T} e^{h\operatorname{Im}\log\zeta(1/2+it)}dt\right)\to h^2/2 \quad as\ T\to\infty.$$

Related papers which also discuss large deviations results are the work of Radziwiłł [**56**], who extended the range of Selberg's central limit theorem for $\operatorname{Re}\log\zeta(1/2+it)$ and the work of Soundararajan [**69**], who proved large deviation bounds for $\operatorname{Re}\log\zeta(1/2+it)$. In fact, Soundararajan [**69**, Corollary A] completed the proof of Corollary 3 in the case of $\operatorname{Re}\log\zeta(1/2+it)$ by proving the upper bound. The result can be stated as follows. For all $\epsilon>0$ and all $h>0$ we have $(\log T)^{h^2-\epsilon}\ll_{h,\epsilon}\int_T^{2T}|\zeta(1/2+it)|^{2h}dt\ll_{h,\epsilon}$

$(\log T)^{h^2+\epsilon}$. Note that the proof of the upper bound also applies to the case of $\operatorname{Im}\log\zeta(1/2+it)$ and that we apply a slightly weaker upper bound in the proofs of Theorem 2 and Corollary 3.

Finally, I mention that Chapter 1 appeared in Math. Z., see [**79**].

NOTATION 1. For $y \geq 2$ and a function $g : [0,1] \to [0,1]$, we define

$$\Sigma_{g,y}(t) = \sum_{p\leq y} \frac{1}{p^{1/2+it}} g\Big(\frac{\log p}{\log y}\Big),$$
$$\Sigma^*_{g,y}(t) = \sum_{n\leq y} \frac{\Lambda(n)}{\log n} \frac{1}{n^{1/2+it}} g\Big(\frac{\log n}{\log y}\Big),$$

$r_{g,y}(t) = \log\zeta(1/2+it) - \Sigma_{g,y}(t)$, and $r^*_{g,y}(t) = \log\zeta(1/2+it) - \Sigma^*_{g,y}(t)$.

2. Moments of a sum over primes

Section 2 is devoted to some standard mean value calculations. In doing so, we will apply the following generalization of the mean value theorem of Montgomery and Vaughan contained in [**57**, Theorem 1.4.3] (see also [**73**, Lemma 3.1]). Let $a_1, \dots, a_M$ and $b_1, \dots, b_M$ be complex numbers, $M \geq 2$, and let $T > 0$. Then

$$\frac{1}{T}\int_T^{2T} \Big(\sum_{m\leq M} a_m m^{-it}\Big)\Big(\overline{\sum_{m\leq M} b_m m^{-it}}\Big) dt$$
$$= \sum_{m\leq M} a_m \bar{b}_m + \theta \frac{2D}{T} \sqrt{\sum_{m\leq M} m|a_m|^2} \sqrt{\sum_{m\leq M} m|b_m|^2}, \quad (2.1)$$

where θ depends on the various parameters but satisfies $|\theta| \leq 1$ and D is the universal constant in [**57**, Theorem 1.4.3].

PROPOSITION 1. *Let $x \geq 2$ and $T > 0$ be real numbers, k be a nonnegative integer, and $p_1, \dots, p_n$ be the prime numbers not exceeding x. Then*

$$\frac{1}{T}\int_T^{2T} \Big(\sum_{p\leq x} \frac{\sin(t\log p)}{\sqrt{p}}\Big)^{2k} dt$$
$$= \frac{1}{2^{2k}}\binom{2k}{k} \sum_{\lambda_1+\cdots+\lambda_n=k} \Big(\frac{k!}{\lambda_1!\cdots\lambda_n!}\Big)^2 p_1^{-\lambda_1}\cdots p_n^{-\lambda_n} + \theta\frac{2D}{T}\sqrt{n^{2k}(2k)!} \quad (2.2)$$

and $|(1/T)\int_T^{2T}(\sum_{p\leq x}\sin(t\log p)/\sqrt{p})^{2k+1}dt| \leq (2D/T)\sqrt{n^{2k+1}(2k+1)!}$ with $|\theta| \leq 1$ and D the constant in (2.1). Furthermore, the main term in (2.2) is bounded by $((2k)!/2^{2k}k!)(\sum_{p\leq x} 1/p)^k$.

PROOF. From $\sin(t\log p) = (p^{it} - p^{-it})/2i$, we obtain

$$\frac{1}{T}\int_T^{2T}\Big(\sum_{p\le x}\frac{\sin(t\log p)}{\sqrt{p}}\Big)^k dt$$
$$= \frac{1}{(2i)^k}\sum_{j=0}^{k}\binom{k}{j}\frac{(-1)^j}{T}\int_T^{2T}\Big(\sum_{p\le x}\frac{1}{p^{1/2+it}}\Big)^j\overline{\Big(\sum_{p\le x}\frac{1}{p^{1/2+it}}\Big)}^{k-j}dt. \quad (2.3)$$

For $j = 1,\dots,k$ the multinomial theorem yields

$$\Big(\sum_{p\le x}\frac{1}{p^{1/2+it}}\Big)^j = \sum_{\lambda_1+\cdots+\lambda_n=j}\frac{j!}{\lambda_1!\cdots\lambda_n!}(p_1^{-\lambda_1}\cdots p_n^{-\lambda_n})^{1/2+it}. \quad (2.4)$$

If we plug in (2.4) into (2.3) with k replaced by $2k$, we obtain from (2.1) that

$$\frac{1}{T}\int_T^{2T}\Big(\sum_{p\le x}\frac{\sin(t\log p)}{\sqrt{p}}\Big)^{2k}dt \quad (2.5)$$
$$= \frac{1}{2^{2k}}\binom{2k}{k}\sum_{\lambda_1+\cdots+\lambda_n=k}\Big(\frac{k!}{\lambda_1!\cdots\lambda_n!}\Big)^2 p_1^{-\lambda_1}\cdots p_n^{-\lambda_n}$$
$$+\frac{\theta 2D}{2^{2k}T}\sum_{j=0}^{2k}\binom{2k}{j}\sqrt{\sum_{\lambda_1+\cdots+\lambda_n=j}\Big(\frac{j!}{\lambda_1!\cdots\lambda_n!}\Big)^2}\sqrt{\sum_{\lambda_1+\cdots+\lambda_n=2k-j}\Big(\frac{(2k-j)!}{\lambda_1!\cdots\lambda_n!}\Big)^2}$$

with $|\theta| \le 1$. Applying $j!/(\lambda_1!\cdots\lambda_n!) \le j!$, $j = 0,\dots,2k$, we bound the absolute value of the remainder by

$$\frac{2D}{2^{2k}T}\sum_{j=0}^{2k}\binom{2k}{j}\sqrt{n^j j!\, n^{2k-j}(2k-j)!} \le \frac{2D}{T}\sqrt{n^{2k}(2k)!}. \quad (2.6)$$

The main term in (2.2) can be bounded similarly. As in (2.5) and (2.6), we also bound the $(2k+1)$th moment. Note that there is no main term in this case. This completes the proof. □

We want to compare these mean value estimates to some random variables expectations. Therefore, let $X_1, X_2, \dots$ be an i.i.d. sequence of random variables uniformly distributed on the unit circle and let $p_1,\dots,p_n$ be the primes not exceeding x. Then

$$\mathbb{E}\Big[\Big(\sum_{i=1}^{n}\frac{\operatorname{Im} X_i}{\sqrt{p_i}}\Big)^{2k}\Big] = \frac{1}{2^{2k}}\binom{2k}{k}\sum_{\lambda_1+\cdots+\lambda_n=k}\Big(\frac{k!}{\lambda_1!\cdots\lambda_n!}\Big)^2 p_1^{-\lambda_1}\cdots p_n^{-\lambda_n} \quad (2.7)$$

and $\mathbb{E}\left[(\sum_{i=1}^n \operatorname{Im} X_i/\sqrt{p_i})^{2k+1}\right] = 0$. To prove this, we replace $\sin(t\log p)$ by $\operatorname{Im} X_i$ and integration by expectation in (2.3) and (2.4) and then apply the formula $\mathbb{E}[X_1^{\lambda_1}\cdots X_n^{\lambda_n}X_1^{-\mu_1}\cdots X_n^{-\mu_n}] = 1$ if $\lambda_j = \mu_j$ for all $j = 1,\dots,n$ and $= 0$ else.

3. Bessel functions

The Bessel functions appear in the Fourier expansion of the function $e^{iz\sin\theta}$,

$$e^{iz\sin\theta} = \sum_{k=-\infty}^{\infty} J_k(z)e^{ik\theta}. \tag{3.1}$$

Explicitly the kth Bessel function $J_k(z)$ is given by

$$J_k(z) = \sum_{n=0}^{\infty} \frac{(-1)^n(z/2)^{k+2n}}{n!(k+n)!} \tag{3.2}$$

for $k \geq 0$ and given by the relation $J_k(z) = (-1)^k J_{-k}(z)$ for $k < 0$ (for these and more facts about Bessel functions see, e.g., the book of Andrews, Askey, and Roy [**1**]). This section is devoted to the following mod-Gaussian convergence result (compare to [**38**, Proposition 4.1]).

PROPOSITION 2. *Let $X_1, X_2, \dots$ be an i.i.d. sequence of random variables uniformly distributed on the unit circle and let $p_1, p_2, \dots$ be the increasing sequence of all primes. Then*

$$e^{z^2(\log\log x+\gamma)/4}\mathbb{E}\Big[e^{iz\sum_{j=1}^{\pi(x)}\frac{\operatorname{Im} X_j}{\sqrt{p_j}}}\Big] \to \Phi(z) \qquad \text{as } x \to \infty \tag{3.3}$$

locally uniformly for $z \in \mathbb{C}$. Here, γ denotes Euler's constant, $\pi(x)$ denotes the number of primes not exceeding x, and $\Phi(z)$ is given by (1.4).

PROOF. By (3.1), we have

$$\mathbb{E}\big[e^{iz\operatorname{Im} X_1}\big] = \frac{1}{2\pi}\int_0^{2\pi} e^{iz\sin\theta}d\theta = J_0(z). \tag{3.4}$$

Applying the independence of the X_j's, (3.4), and finally Merten's formula $\prod_{p\leq x}(1-1/p) = (e^{-\gamma}/\log x)(1+o(1))$, we obtain that the left hand side of (3.3) is equal to

$$e^{z^2(\log\log x+\gamma)/4}\prod_{p\leq x} J_0\Big(\frac{z}{\sqrt{p}}\Big) = (1+o(1))^{z^2/4}\prod_{p\leq x}\Big(1-\frac{1}{p}\Big)^{-z^2/4} J_0\Big(\frac{z}{\sqrt{p}}\Big).$$

It remains to show that the above product converges to $\Phi(z)$, locally uniformly for $z \in \mathbb{C}$. This follows from the fact that the product $\Phi(z)$ is normally convergent (see [**29**, Chapter IV.1, especially Remark IV.1.7]). This completes the proof. □

Consider the random variables $\operatorname{Im}\Sigma_{1,x}(-U_T)$, U_T being random variables uniformly distributed on $[T, 2T]$. As mentioned in the Introduction, one can use the method of moments to deduce that, as $x \to \infty$, $x = T^{o(1)}$, $(1/\sqrt{(\log\log x)/2})\operatorname{Im}\Sigma_{1,x}(-U_T)$ converges in distribution to a Gaussian random variable with expectation 0 and variance 1 (see [**14**, Proof of Theorem B]). We will generalize this result by considering the cumulants of $\operatorname{Im}\Sigma_{1,x}(-U_T)$.

If Y is a real random variable such that $\mathbb{E}[e^{zY}]$ exists and is finite for all $z \in \mathbb{C}$, $\mathbb{E}[e^{zY}]$ is an analytic function and there exists a neighbourhood of 0 where $\log \mathbb{E}[e^{zY}] = \sum_{m=1}^{\infty} \kappa_m(Y) z^m/m!$. The coefficients $\kappa_m(Y)$, $m \geq 1$, are called the cumulants of Y. Thus, $\kappa_m(Y)$ is equal to the mth derivative of $\log \mathbb{E}[e^{zY}]$ evaluated at 0.

COROLLARY 4. *Let $x = e^{\log T/N}$ and N such that $x \to \infty$ and $N \to \infty$ as $T \to \infty$ and let U_T be random variables uniformly distributed on $[T, 2T]$. Then, as $T \to \infty$, $\kappa_2(\operatorname{Im} \Sigma_{1,x}(-U_T)) - (\log\log x + \gamma)/2 \to c_2$ and for $m \neq 2$ $\kappa_m(\operatorname{Im} \Sigma_{1,x}(-U_T)) \to c_m$, where the c_m's are defined by the series expansion $\log \Phi(-iz) = \sum_{m=1}^{\infty} c_m z^m/m!$, for z in a neighbourhood of 0.*

PROOF. By the construction of Φ, there exists a real number $0 < r \leq 1$ such that for $|z| \leq r$, $\log \Phi(z) = \sum_{p \in \mathbb{P}}((-z^2/4)\log(1-1/p) + \log J_0(z/\sqrt{p}))$. Hence, by Merten's formula,

$$(-z^2/4)(\log\log x + \gamma) + \sum_{p \leq x} \log J_0\Big(\frac{-iz}{\sqrt{p}}\Big) \to \log \Phi(-iz) \quad as\ x \to \infty \tag{3.5}$$

uniformly for $|z| \leq r$, $z \in \mathbb{C}$. The uniform convergence implies (see [**29**, Theorem III.1.3]), that the mth derivative of the left hand side of (3.5) evaluated at 0 converges to c_m. Hence, under the assumptions of Proposition 2, the cumulants of $\sum_{j=1}^{\pi(x)} \operatorname{Im} X_j/\sqrt{p_j}$ satisfy the convergence described in Corollary 4, since $\mathbb{E}[\exp(z \sum_{j=1}^{\pi(x)} \operatorname{Im} X_j/\sqrt{p_j})] = \prod_{p \leq x} J_0(-iz/\sqrt{p})$. It remains to show that for $m \geq 1$

$$\kappa_m\left(\operatorname{Im} \Sigma_{1,x}(-U_T)\right) - \kappa_m\Bigg(\sum_{j=1}^{\pi(x)} \frac{\operatorname{Im} X_j}{\sqrt{p_j}}\Bigg) \to 0 \qquad as\ T \to \infty. \tag{3.6}$$

To prove this, we use the fact that the cumulants can be expressed in terms of the moments, namely $\kappa_m(Y) = \sum a_{\lambda_1,\ldots,\lambda_m} \mathbb{E}[Y^1]^{\lambda_1} \cdots \mathbb{E}[Y^m]^{\lambda_m}$, where the sum is over all positive integers such that $1\lambda_1 + 2\lambda_2 + \cdots + m\lambda_m = m$, $a_{\lambda_1,\ldots,\lambda_m}$ are integers, and Y is a random variable as above. If we plug in Proposition 1 and (2.7) into this formula, (3.6) follows from multiplying out since for $k \leq m$ and $x \geq 3$ the main terms in (2.2) are $O((\sum_{p \leq x} 1/p)^m) = O((\log\log x)^m)$ (see [**20**, (5) of chapter 7]), while for $k \leq m$ the remainders in (2.2) are $O(T^{(m/N)-1})$ which is $O(T^{-a})$ for some $0 < a < 1$ if T is sufficiently large. □

4. Mod-convergence of a sum over primes

By means of Proposition 1 and (2.7), we can apply the method of moments for fixed x and obtain the following convergence

$$\frac{1}{T}\int_T^{2T} e^{iu\sum_{p \leq x} \frac{\sin(t\log p)}{\sqrt{p}}}\, dt \to \prod_{p \leq x} J_0\Big(\frac{u}{\sqrt{p}}\Big) \quad as\ T \to \infty. \tag{4.1}$$

Another proof of (4.1) is contained in [**48**, Theorem 5.1]. The techniques used therein can be applied to get Theorem 1 and Theorem 2 for the choice $x = (\log T)^{2-\epsilon}$, $\epsilon > 0$ arbitrary. The improvement of Theorem 1 follows from Proposition 3 combined with Proposition 2.

PROPOSITION 3. *Let $c > 1$ be a constant. Define $x = e^{\log T/N}$ with $N = (c'ec^2/4)\log\log T$, where $c' > 1$ is allowed to depend on T but such that $x \to \infty$ as $T \to \infty$. For $T \geq 3$, sufficiently large such that $x \geq 2$ and $N \geq 1$, we have*

$$\frac{1}{T}\int_T^{2T} e^{iu\sum_{p\leq x}\frac{\sin(t\log p)}{\sqrt{p}}}dt = \prod_{p\leq x} J_0\Big(\frac{u}{\sqrt{p}}\Big) + O((1/c')^{N-1} + (2c^2/\log x)^N) \tag{4.2}$$

uniformly for $|u| \leq c$, $u \in \mathbb{R}$.

PROOF OF THEOREM 1. We apply Proposition 3 with x, N as in Theorem 1 and c an arbitrary constant with $c > 1$. Since $c' \to \infty$ in that case, the remainder in (4.2) is $o(\exp(-c^2(\log\log T)/4))$. If we multiply in (4.2) both sides by $\exp(u^2(\log\log x + \gamma)/4)$ and then apply Proposition 2, we obtain (1.3) uniformly for $|u| \leq c$. Since $c > 1$ is arbitrary, this completes the proof. □

PROOF OF PROPOSITION 3. Let $N' = \lfloor N \rfloor$. From the Taylor expansion $e^{iu} = \sum_{k\leq 2N'-1}(iu)^k/k! + \theta u^{2N'}/(2N')!$, $u \in \mathbb{R}$, with $|\theta| \leq 1$, we obtain

$$\begin{aligned}\frac{1}{T}\int_T^{2T} e^{iu\sum_{p\leq x}\frac{\sin(t\log p)}{\sqrt{p}}}dt = &\sum_{k\leq 2N'-1}\frac{(iu)^k}{k!}\frac{1}{T}\int_T^{2T}\Big(\sum_{p\leq x}\frac{\sin(t\log p)}{\sqrt{p}}\Big)^k dt \\ &+ \theta\frac{u^{2N'}}{(2N')!}\frac{1}{T}\int_T^{2T}\Big(\sum_{p\leq x}\frac{\sin(t\log p)}{\sqrt{p}}\Big)^{2N'} dt \end{aligned} \tag{4.3}$$

with $|\theta| \leq 1$. By Proposition 1, the remainder is

$$O\Big(\frac{c^{2N'}}{N'!}\frac{1}{2^{2N'}}\Big(\sum_{p\leq x}\frac{1}{p}\Big)^{N'} + \frac{(c^2\pi(x))^{N'}}{T}\Big).$$

Using the bound $(N')! \geq (N'/e)^{N'}$, elementary results in the theory of primes, namely the formulas $\sum_{p\leq x} 1/p = \log\log x + c_1 + O(1/\log x)$ and $\pi(x) \leq 2x/\log x$, and finally $N' = \lfloor N \rfloor$, this is

$$O\Big(\Big(\frac{ec^2\log\log T}{4N'}\Big)^{N'} + \frac{(c^2\pi(x))^N}{T}\Big) = O\Big(\Big(\frac{1}{c'}\Big)^{N-1} + \Big(\frac{2c^2}{\log x}\Big)^N\Big).$$

Now, let $X_1, X_2, \ldots$ be an i.i.d. sequence of random variables uniformly distributed on the unit circle. By Proposition 1 and (2.7), the moments in (4.3) are equal to those of the stochastic model plus a remainder which is bounded by $(2D/T)\sqrt{(\pi(x))^k k!}$. The resulting remainders in (4.3), $k \leq$

$2N'-1$, add up to $O((c^2\pi(x))^N/T) = O((2c^2/\log x)^N)$. Hence, (4.3) is equal to

$$\sum_{k\leq 2N'-1} \frac{(iu)^k}{k!}\mathbb{E}\Big[\Big(\sum_{j=1}^{\pi(x)} \frac{\operatorname{Im} X_j}{\sqrt{p_j}}\Big)^k\Big] + O((1/c')^{N-1} + (2c^2/\log x)^N).$$

Applying the above Taylor expansion again, we obtain

$$\begin{aligned}\prod_{p\leq x} J_0\Big(\frac{u}{\sqrt{p}}\Big) &= \mathbb{E}\Big[e^{iu\sum_{j=1}^{\pi(x)} \frac{\operatorname{Im} X_j}{\sqrt{p_j}}}\Big] \\ &= \sum_{k\leq 2N'-1} \frac{(iu)^k}{k!}\mathbb{E}\Big[\Big(\sum_{j=1}^{\pi(x)} \frac{\operatorname{Im} X_j}{\sqrt{p_j}}\Big)^k\Big] + \theta\frac{u^{2N'}}{(2N')!}\mathbb{E}\Big[\Big(\sum_{j=1}^{\pi(x)} \frac{\operatorname{Im} X_j}{\sqrt{p_j}}\Big)^{2N'}\Big]\end{aligned}$$

with $|\theta|\leq 1$. The remainder already appeared in (4.3) and is $O((1/c')^{N-1})$. This completes the proof. □

5. Mod-convergence in the complex plane

Section 5 is devoted to the proof of Theorem 2. Here, we will apply an explicit formula obtained by Goldston [**32**, Lemma 1] assuming RH. For $4\leq x\leq t^2$ and $t\neq\gamma$, we have

$$\begin{aligned}\operatorname{Im}\log\zeta(1/2+it) &= -\sum_{n\leq x}\frac{\Lambda(n)}{\log n}\frac{\sin(t\log n)}{\sqrt{n}} f\Big(\frac{\log n}{\log x}\Big) \\ &+ \sum_{\gamma}\sin((t-\gamma)\log x)\int_0^\infty \frac{u}{u^2+((t-\gamma)\log x)^2}\frac{du}{\sinh u} + O\Big(\frac{1}{t(\log x)^2}\Big),\end{aligned} \tag{5.1}$$

where $f(u) = (\pi u/2)\cot(\pi u/2)$. We will also apply the following estimate obtained by Soundararajan assuming RH. For every $h\in\mathbb{R}$ there exist constants $C', C''>0$ such that

$$\frac{1}{T}\int_T^{2T} e^{h\operatorname{Im}\log\zeta(1/2+it)}dt \leq C'' e^{C'h^2\log\log T}. \tag{5.2}$$

Soundararajan [**69**] proved (5.2) for $\operatorname{Re}\log\zeta(1/2+it)$. However, by using [**66**, Theorem 1] instead of [**69**, Proposition], his arguments apply to $\operatorname{Im}\log\zeta(1/2+it)$, too. We prove (compare to [**14**, Lemma 3 and Corallary]):

Proposition 4. *Assume RH. Let $4\leq x\leq T^2$, $f(u) = (\pi u/2)\cot(\pi u/2)$. For every $h\in\mathbb{R}$ there exist constants C, C', and C'' such that*

$$\frac{1}{T}\int_T^{2T} e^{h\sum_{n\leq x}\frac{\Lambda(n)}{\log n}\frac{\sin(t\log n)}{\sqrt{n}} f\left(\frac{\log n}{\log x}\right)}dt \leq C'' e^{C|h|\frac{\log T}{\log x}+C'h^2\log\log T}.$$

PROOF OF THEOREM 2. The proof mainly differs from the proof of Theorem 1 and Proposition 3 in its estimation of the remainder term. Nevertheless, we will repeat the main steps. We assume that $|z| \le c$, where $c > 1$ is an arbitrary constant and $z \in \mathbb{C}$, say $z = u - ih$ with $u, h \in \mathbb{R}$. Let $N' = \lfloor N/2 \rfloor$. From the Taylor expansion $e^{iz} = e^{h+iu} = \sum_{k \le 2N'-1} (iz)^k/k! + \theta e^h (|z|^{2N'}/(2N')!)$ with $|\theta| \le 1$, we obtain

$$\frac{1}{T}\int_T^{2T} e^{iz \operatorname{Im} \Sigma_{f,x}(-t)} dt = \sum_{k \le 2N'-1} \frac{(iz)^k}{k!} \frac{1}{T} \int_T^{2T} (\operatorname{Im} \Sigma_{f,x}(-t))^k \, dt + \theta \frac{c^{2N'}}{(2N')!} \frac{1}{T} \int_T^{2T} e^{h \operatorname{Im} \Sigma_{f,x}(-t)} (\operatorname{Im} \Sigma_{f,x}(-t))^{2N'} \, dt \quad (5.3)$$

with $|\theta| \le 1$. To continue as in the proof of Proposition 3, we use that under the assumptions of Proposition 1 we have

$$\frac{1}{T}\int_T^{2T} (\operatorname{Im} \Sigma_{f,x}(-t))^k dt = \mathbb{E}\left[\left(\sum_{j=1}^{\pi(x)} \frac{\operatorname{Im} X_j}{\sqrt{p_j}} f\left(\frac{\log p_j}{\log x} \right) \right)^k \right] + \theta \frac{2D}{T} \sqrt{(\pi(x))^k k!} \quad (5.4)$$

with $|\theta| \le 1$. Moreover, if we replace k by $2k$, the main term is bounded by $((2k)!/2^{2k}k!)(\sum_{p \le x} 1/p)^k$. These estimates follow as in the proof of Proposition 1 and (2.7). Applying the Cauchy-Schwarz inequality, the absolut value of the remainder in (5.3) can be bounded by

$$\frac{c^{2N'}}{(2N')!} \sqrt{\frac{1}{T}\int_T^{2T} (\operatorname{Im} \Sigma_{f,x}(-t))^{4N'} dt} \sqrt{\frac{1}{T}\int_T^{2T} e^{2h \operatorname{Im} \Sigma_{f,x}(-t)} dt}.$$

For $x \ge 2$, we have $|\operatorname{Im} \Sigma_{f,x}(-t) - \operatorname{Im} \Sigma^*_{f,x}(-t)| \le (\log\log x)/2 + O(1)$, say $\le CN$ for T sufficiently large. Hence, by (5.4) and Proposition 4, the above is

$$O\left(\frac{c^{2N'}}{(2N')!} \sqrt{\frac{(4N')!(\sum_{p \le x} 1/p)^{2N'}}{2^{4N'}(2N')!} + \frac{\sqrt{(4N')!(\pi(x))^{4N'}}}{T}} \sqrt{e^{4CcN + 4C'c^2N}} \right)$$

for T sufficiently large. Applying $(4N')!/(2^{4N'}(2N')!) \le (2N')!$, $\sqrt{(2N')!} \ge (2N'/e)^{N'}$, $\pi(x) \le 2x/\log x$, and $N' = \lfloor N/2 \rfloor$, there exists a constant $c'' > 0$ (depending on c, C, and C') such that this is

$$O\left(\left(\frac{c'' \sum_{p \le x} 1/p}{N'} \right)^{N'} + \left(\frac{c''}{\log x} \right)^{N/2} \right).$$

Since $N'/\sum_{p \le x} 1/p$ and $\log x$ go to infinity, this is $o(\exp(-c^2(\log\log x)/4))$. Hence, applying (5.4) to the other terms, (5.3) is equal to

$$\sum_{k \le 2N'-1} \frac{(iz)^k}{k!} \mathbb{E}\left[\left(\sum_{j=1}^{\pi(x)} \frac{\operatorname{Im} X_j}{\sqrt{p_j}} f\left(\frac{\log p_j}{\log x} \right) \right)^k \right] + o(e^{-c^2(\log\log T)/4})$$

uniformly for $|z| \leq c$. If we replace $\sin(t \log p)$ by $\operatorname{Im} X_i$ and integration by expectation in (5.3), we can bound the resulting remainder as above. In doing so, we apply (5.5) instead of Proposition 4. The result is that

$$\frac{1}{T}\int_T^{2T} e^{iz \operatorname{Im} \Sigma_{f,x}(-t)} dt = \mathbb{E}\left[e^{iz \sum_{j=1}^{\pi(x)} \frac{\operatorname{Im} X_j}{\sqrt{p_j}} f\left(\frac{\log p_j}{\log x}\right)}\right] + o(e^{-c^2(\log\log T)/4})$$

uniformly for $|z| \leq c$. If we multiply both sides by $\exp(z^2(\log\log x + \gamma_f)/4)$ and then apply the formula

$$e^{z^2(\log\log x+\gamma_f)/4}\mathbb{E}\left[e^{iz \sum_{j=1}^{\pi(x)} \frac{\operatorname{Im} X_j}{\sqrt{p_j}} f\left(\frac{\log p_j}{\log x}\right)}\right] \to \Phi(z) \qquad as\ x \to \infty \qquad (5.5)$$

locally uniformly for $z \in \mathbb{C}$, the statement of Theorem 2 follows by the same argument as in the proof of Theorem 1. (5.5) follows as in the proof of Proposition 2 by using the additional fact, that

$$\prod_{p \leq x} \left(1 - \frac{1}{p} f^2\left(\frac{\log p}{\log x}\right)\right)^{-z^2/4} J_0\left(\frac{z}{\sqrt{p}} f\left(\frac{\log p}{\log x}\right)\right) \to \Phi(z) \quad as\ x \to \infty$$

locally uniformly for $z \in \mathbb{C}$. We conclude by a brief argument why this holds. Split the product in $p \leq y$ and $y < p \leq x$. The product over $y < p \leq x$ converges locally uniformly to 1 if $y \to \infty$, while one can show that the product over $p \leq y$, say $y = \log x$, converges locally uniformly to $\phi(z)$, by using, e.g., $f^2(\log p/\log x) - 1 = O((\log\log x)/\log x)$ if $p \leq \log x$. This completes the proof. □

Proof of Proposition 4. From (5.1), (5.2) and the Cauchy-Schwarz inequality, we obtain

$$\frac{1}{T}\int_T^{2T} e^{h \sum_{n \leq x} \frac{\Lambda(n)}{\log n} \frac{\sin(t \log n)}{\sqrt{n}} f\left(\frac{\log n}{\log x}\right)} dt$$
$$\leq C''' e^{2C'h^2 \log\log T} \sqrt{\frac{1}{T}\int_T^{2T} e^{2h \sum_\gamma \sin((t-\gamma)\log x) \int_0^\infty \frac{u}{u^2+((t-\gamma)\log x)^2} \frac{du}{\sinh u}} dt}$$

where C''' is a constant. The absolute value of the sum over zeros is bounded by a constant times

$$\sum_{|(t-\gamma)\log x| \leq 1} 1 + \sum_{|(t-\gamma)\log x| > 1} \frac{1}{((t-\gamma)\log x)^2} \qquad (5.6)$$

and therefore it suffices to deal with the exponential moments of (5.6) with $h \geq 0$. Using the Cauchy-Schwarz inequality again, we obtain

$$\frac{1}{T}\int_T^{2T} e^{h \sum_{|(t-\gamma)\log x| \leq 1} 1 + h \sum_{|(t-\gamma)\log x| > 1} \frac{1}{((t-\gamma)\log x)^2}} dt$$
$$\leq \sqrt{\frac{1}{T}\int_T^{2T} e^{2h \sum_{|(t-\gamma)\log x| \leq 1} 1} dt} \sqrt{\frac{1}{T}\int_T^{2T} e^{2h \sum_{|(t-\gamma)\log x| > 1} \frac{1}{((t-\gamma)\log x)^2}} dt}.$$

We start with the first term, using the following fact on the number of zeros (see [**20**, (1) of Ch. 15])

$$N(t) = \frac{t}{2\pi}\log\frac{t}{2\pi} - \frac{t}{2\pi} + \frac{7}{8} + S(t) + O(1/t), \tag{5.7}$$

where $t \neq \gamma$ and $S(t) = (1/\pi)\operatorname{Im}\log\zeta(1/2+it)$. We compute (note that $h \geq 0$)

$$\begin{aligned}
&\frac{1}{T}\int_T^{2T} e^{h\sum_{|(t-\gamma)\log x|\leq 1} 1} dt \\
&\qquad = \frac{1}{T}\int_T^{2T} e^{h\left(N\left(t+\frac{1}{\log x}\right)-N\left(t-\frac{1}{\log x}\right)\right)} dt \\
&\qquad \leq \frac{1}{T}\int_T^{2T} e^{Ch\frac{\log t}{\log x}+h\left(S\left(t+\frac{1}{\log x}\right)-S\left(t-\frac{1}{\log x}\right)\right)} dt \\
&\qquad \leq e^{Ch\frac{\log T}{\log x}}\sqrt{\frac{1}{T}\int_T^{2T} e^{2hS\left(t+\frac{1}{\log x}\right)}dt}\sqrt{\frac{1}{T}\int_T^{2T} e^{-2hS\left(t-\frac{1}{\log x}\right)}dt} \\
&\qquad = O(e^{Ch\frac{\log T}{\log x}+4C'(h/\pi)^2\log\log T}).
\end{aligned} \tag{5.8}$$

In the last step we used (5.2). Next, we divide the sum over $|(t-\gamma)\log x| > 1$ into $|t-\gamma| \geq T$, $1 < |t-\gamma| < T$, and $1/\log x < |t-\gamma| \leq 1$.

For $t \in [T, 2T]$, we have

$$\sum_{|t-\gamma|\geq T} \frac{1}{((t-\gamma)\log x)^2} = O\Big(\sum_\gamma \frac{1}{\gamma^2(\log x)^2}\Big) = O\Big(\frac{1}{(\log x)^2}\Big).$$

The last step results from [**20**, (4) of Ch. 12]. For the second sum we use the fact that $N(t+1) - N(t) = O(1+\log^+|t|)$ (see [**20**, (2) of Ch. 15]). For $t \in [T, 2T]$, we obtain

$$\begin{aligned}
&\sum_{1<|t-\gamma|<T} \frac{1}{((t-\gamma)\log x)^2} \\
&\leq \sum_{k=1}^{\lceil T\rceil-1} \frac{N(t+k+1)-N(t+k)}{k^2(\log x)^2} + \sum_{k=1}^{\lceil T\rceil-1} \frac{N(t-k)-N(t-k-1)}{k^2(\log x)^2} \\
&= O\Big(\sum_{k=1}^{\lceil T\rceil-1} \frac{\log T}{k^2(\log x)^2}\Big) = O\Big(\frac{\log T}{(\log x)^2}\Big).
\end{aligned}$$

Next, we consider the sum over $1/\log x < \gamma - t \leq 1$. We have

$$\sum_{1/\log x<\gamma-t\leq 1} \frac{1}{((t-\gamma)\log x)^2} \leq \sum_{j=1}^{M} \frac{N\left(t+\frac{k_j}{\log x}\right) - N\left(t+\frac{k_{j-1}}{\log x}\right)}{k_{j-1}^2} \tag{5.9}$$

where $1 = k_0 < k_1 < \cdots < k_M$ with $k_{M-1} < \log x \le k_M$. By (5.7), this is bounded by, recall $t \in [T, 2T]$,

$$\sum_{j=1}^{M} \left(\frac{C(k_j - k_{j-1}) \log T}{(\log x) k_{j-1}^2} + \frac{S\left(t + \frac{k_j}{\log x}\right) - S\left(t + \frac{k_{j-1}}{\log x}\right)}{k_{j-1}^2} \right).$$

We choose $k_j = 2^{j/2}$ and bound the left hand side of (5.9) by

$$\sqrt{2} C \frac{\log T}{\log x} + \sum_{j=1}^{M} \frac{S\left(t + \frac{2^{j/2}}{\log x}\right) - S\left(t + \frac{2^{(j-1)/2}}{\log x}\right)}{2^{j-1}}.$$

It follows that

$$\frac{1}{T} \int_T^{2T} e^{h \sum_{1/\log x < \gamma - t \le 1} \frac{1}{((t-\gamma)\log x)^2}} dt$$
$$\le e^{\sqrt{2} C h \frac{\log T}{\log x}} \frac{1}{T} \int_T^{2T} e^{h \sum_{j=1}^{M} \frac{1}{2^{j-1}} \left(S\left(t + \frac{2^{j/2}}{\log x}\right) - S\left(t + \frac{2^{(j-1)/2}}{\log x}\right)\right)} dt.$$

Using $\mathbb{E}[e^{h \sum_{j=1}^{M} X_j / 2^j}] \le \prod_{j=1}^{M} (\mathbb{E}[e^{h X_j}])^{1/2^j}$, which follows from repeated application of the Cauchy-Schwarz inequality, this is

$$\le e^{\sqrt{2} C h \frac{\log T}{\log x}} \prod_{j=1}^{M} \left(\frac{1}{T} \int_T^{2T} e^{2h\left(S\left(t + \frac{2^{j/2}}{\log x}\right) - S\left(t + \frac{2^{(j-1)/2}}{\log x}\right)\right)} dt \right)^{1/2^j}.$$

Applying again the Cauchy-Schwarz inequality and then (5.2) (as in (5.8)), this is

$$O(e^{\sqrt{2} C h \frac{\log T}{\log x}} e^{16 C' (h/\pi)^2 \log\log T}).$$

The same bound is true for the sum over $1/\log x < t - \gamma \le 1$. The claim now follows from putting together all these estimates. □

6. Proof of Corollary 1

Let T, c, c', x, and N be as in Proposition 3, $T \ge 3$ sufficiently large such that $x \ge 2$ and $N \ge 2$. Assume further that $c' > 4$ is a constant such that the bound $(\log\log T)^{1/2} (c'/4)^{-N/2} = O(1/\log T)$ holds and that T is so big that the bound $(\log T)(2c^2/\log x)^{N/2} = O(1/\log T)$ holds, too. Then we show that

$$\frac{1}{T} \int_T^{2T} e^{iu \operatorname{Im} \log \zeta(1/2+it)} dt = \prod_{p \le x} J_0\Big(\frac{u}{\sqrt{p}}\Big)$$
$$- \sum_{\substack{p \le x \\ k \ge 3 \text{ odd}}} \frac{u}{k \sqrt{p}^k} J_k\Big(\frac{u}{\sqrt{p}}\Big) \prod_{\substack{q \le x \\ q \ne p}} J_0\Big(\frac{u}{\sqrt{q}}\Big) + u^2 O(\log\log\log T) + O(1/\log T) \tag{6.1}$$

uniformly for $|u| \le c$, $u \in \mathbb{R}$. One can deduce Corollary 1 from (6.1) as follows. Replace u by $v/\sqrt{(\log\log T)/2}$ with $|v| \le \sqrt{\log\log T / \log\log\log T}$ and

let T be sufficiently large. Then, by (3.2), the formula $\sum_{p\leq x} 1/p = \log\log x + c_1 + O(1/\log x)$, and $\log\log x/\log\log T = 1 + O(\log\log\log T/\log\log T)$, the first term on the right hand side of (6.1) is equal to

$$\exp\Big(\sum_{p\leq x}\log J_0(v/\sqrt{p(\log\log T)/2})\Big) = e^{-v^2/2}\Big(1+v^2O\Big(\frac{\log\log\log T}{\log\log T}\Big)\Big)$$

and, by using $|J_k(u)| \leq (|u|/2)^k/k!$ and $|J_0(u)| \leq 1$, $u\in\mathbb{R}$, the second term is $v^4O(1/(\log\log T)^2)$ which is smaller than $v^2O(\log\log\log T/\log\log T)$.

Hence, it remains to prove (6.1). From $\operatorname{Im}\log\zeta(1/2+it) = \operatorname{Im}\Sigma_{1,x}(t) + \operatorname{Im} r_{1,x}(t)$ and Taylor's theorem, we obtain

$$\frac{1}{T}\int_T^{2T} e^{iu\operatorname{Im}\log\zeta(1/2+it)}dt = \frac{1}{T}\int_T^{2T} e^{iu\operatorname{Im}\Sigma_{1,x}(t)}dt + iu\frac{1}{T}\int_T^{2T}\operatorname{Im} r_{1,x}(t)e^{iu\operatorname{Im}\Sigma_{1,x}(t)}dt + \theta\frac{u^2}{2}\frac{1}{T}\int_T^{2T}(\operatorname{Im} r_{1,x}(t))^2dt$$

with $|\theta|\leq 1$. By Proposition 3 and the above assumptions, the first term is equal to $\prod_{p\leq x} J_0(u/\sqrt{p}) + O(1/\log T)$ and by [**73**, Corollary of Theorem 5.1], the third term is $u^2O(\log\log\log T)$. It remains to consider the second term. We start showing that

$$\frac{1}{T}\int_T^{2T}\operatorname{Im}\log\zeta(1/2+it)e^{iu\operatorname{Im}\Sigma_{1,x}(t)}dt = \sum_{\substack{p\leq x\\ k\geq 1\ odd}}\frac{i}{k\sqrt{p}^k}J_k\Big(\frac{u}{\sqrt{p}}\Big)\prod_{\substack{q\leq x\\ q\neq p}}J_0\Big(\frac{u}{\sqrt{q}}\Big) + O(1/\log T) \quad (6.2)$$

uniformly for $|u|\leq c$. Let $N' = \lfloor N/2\rfloor$. From the Taylor expansion $e^{iu} = \sum_{k\leq 2N'-1}(iu)^k/k! + \theta u^{2N'}/(2N')!$, $u\in\mathbb{R}$, with $|\theta|\leq 1$, we obtain that the left hand side of (6.2) is equal to

$$\sum_{k\leq 2N'-1}\frac{(iu)^k}{k!}\frac{1}{T}\int_T^{2T}\operatorname{Im}\log\zeta(1/2+it)(\operatorname{Im}\Sigma_{1,x}(t))^k dt + \theta\frac{c^{2N'}}{(2N')!}\frac{1}{T}\int_T^{2T}|\operatorname{Im}\log\zeta(1/2+it)|(\operatorname{Im}\Sigma_{1,x}(t))^{2N'}dt \quad (6.3)$$

with $|\theta|\leq 1$. Applying the Cauchy-Schwarz inequality, the estimates in the proof of Proposition 3, and [**67**, Theorem 3], i.e. $(1/T)\int_T^{2T}(\operatorname{Im}\log\zeta(1/2+it))^2dt = (\log\log T)/2 + O(1)$, the remainder is $O((\log\log T)^{1/2}((c'/4)^{-N/2} + (2c^2/\log x)^{N/2}) = O(1/\log T)$. The remaining moments can be computed by using the following lemma which is a modification of [**66**, Lemma 5] and [**32**, equation (6.3)] and serves as a substitute for the mean value theorem of Montgomery and Vaughan in Section 2.

LEMMA 1. *Assume RH. Let $k, h \le T$ be two positive integers with $(k,h)=1$. Then*

$$\int_T^{2T} \log\zeta(1/2+it)\Big(\frac{k}{h}\Big)^{it}dt = \begin{cases} \dfrac{T\Lambda(k)}{\sqrt{k}\log k}+O(\sqrt{kh}\log T), & h=1 \\ O(\sqrt{kh}\log T), & h\neq 1, \end{cases}$$

$$\int_T^{2T} \operatorname{Im}\log\zeta(1/2+it)\Big(\frac{k}{h}\Big)^{it}dt = \begin{cases} \dfrac{-iT\Lambda(k)}{2\sqrt{k}\log k}+O(\sqrt{kh}\log T), & h=1 \\ \dfrac{iT\Lambda(h)}{2\sqrt{h}\log h}+O(\sqrt{kh}\log T), & k=1 \\ O(\sqrt{kh}\log T), & h,k\neq 1. \end{cases} \tag{6.4}$$

Denote by $p_1, p_2, \dots, p_n$ the prime numbers not exceeding x, and let $X_1, X_2, \dots$ be an i.i.d. sequence of random variables uniformly distributed on the unit circle. Furthermore, let $k, h \le T$ be positive integers with $k/h = p_1^{-k_1}\cdots p_n^{-k_n}$. Then (6.4) can be written as

$$\begin{aligned} &\frac{1}{T}\int_T^{2T} \operatorname{Im}\log\zeta(1/2+it)(p_1^{-k_1}\cdots p_n^{-k_n})^{it}dt \qquad (6.5)\\ =&\mathbb{E}\Big[-\sum_{j=1}^{n}\operatorname{Im}\log(1-X_j/\sqrt{p_j})X_1^{k_1}\cdots X_n^{k_n}\Big]+O\Big(\frac{1}{T}\sqrt{p_1^{|k_1|}\cdots p_n^{|k_n|}}\log T\Big). \end{aligned}$$

Expanding $(\operatorname{Im}\Sigma_{1,x}(t))^k$ as in (2.3) and (2.4), we deduce from (6.5) that

$$\begin{aligned} &\frac{1}{T}\int_T^{2T} \operatorname{Im}\log\zeta(1/2+it)(\operatorname{Im}\Sigma_{1,x}(t))^k dt \\ &\qquad = \mathbb{E}\Big[\Big(-\sum_{j=1}^{n}\operatorname{Im}\log(1-X_j/\sqrt{p_j})\Big)\Big(\sum_{j=1}^{n}\frac{\operatorname{Im}X_j}{\sqrt{p_j}}\Big)^k\Big] \\ &\quad + O\Big(\frac{\log T}{2^k T}\sum_{l=0}^{k}\binom{k}{l}\sum_{\lambda_1+\cdots+\lambda_n=l}\frac{l!}{\lambda_1!\cdots\lambda_n!}\sum_{\lambda_1+\cdots+\lambda_n=k-l}\frac{(k-l)!}{\lambda_1!\cdots\lambda_n!}\Big). \end{aligned}$$

The remainder is $O((\log T)n^k/T)$ and the resulting remainders in (6.3), $k \le 2N'-1$, add up to $O((\log T)(2c/\log x)^N) = O(1/\log T)$. Hence, (6.3) is

equal to

$$\sum_{k\le 2N'-1}\frac{(iu)^k}{k!}\mathbb{E}\Big[\Big(-\sum_{j=1}^n \operatorname{Im}\log(1-X_j/\sqrt{p_j})\Big)\Big(\sum_{j=1}^n\frac{\operatorname{Im}X_j}{\sqrt{p_j}}\Big)^k\Big]+O(1/\log T)$$
$$=\mathbb{E}\Big[\Big(-\sum_{j=1}^n \operatorname{Im}\log(1-X_j/\sqrt{p_j})\Big)e^{iu\sum_{j=1}^n\frac{\operatorname{Im}X_j}{\sqrt{p_j}}}\Big]$$
$$+\theta\frac{c^{2N'}}{(2N')!}\mathbb{E}\Big[\Big|-\sum_{j=1}^n \operatorname{Im}\log(1-X_j/\sqrt{p_j})\Big|\Big(\sum_{j=1}^n\frac{\operatorname{Im}X_j}{\sqrt{p_j}}\Big)^{2N'}\Big]+O(1/\log T). \tag{6.6}$$

The last equality follows from applying Taylor's theorem as in (6.3). If one treats the first remainder in the last row as the corresponding one in (6.3), using $\mathbb{E}[(\sum_{j=1}^{\pi(x)}\operatorname{Im}\log(1-X_j/\sqrt{p_j}))^2]=(\log\log x)/2+O(1)$ this time, one can show that it is also $O(1/\log T)$. By plugging in (3.1) and expanding the logarithm, we obtain that (6.6) is equal to

$$\sum_{\substack{p\le x\\ k\ge 1\ odd}}\frac{i}{k\sqrt{p}^k}J_k\Big(\frac{u}{\sqrt{p}}\Big)\prod_{\substack{q\le x\\ q\ne p}}J_0\Big(\frac{u}{\sqrt{q}}\Big)+O(1/\log T)$$

which completes the proof of (6.2). The last step in the proof of (6.1) is to show that

$$\frac{1}{T}\int_T^{2T}\operatorname{Im}\Sigma_{1,x}(t)e^{iu\operatorname{Im}\Sigma_{1,x}(t)}dt$$
$$=\mathbb{E}\Big[\Big(\sum_{j=1}^n\frac{\operatorname{Im}X_j}{\sqrt{p_j}}\Big)e^{iu\sum_{j=1}^n\frac{\operatorname{Im}X_j}{\sqrt{p_j}}}\Big]+O(1/\log T)$$
$$=\sum_{p\le x}\frac{i}{\sqrt{p}}J_1\Big(\frac{u}{\sqrt{p}}\Big)\prod_{\substack{q\le x\\ q\ne p}}J_0\Big(\frac{u}{\sqrt{q}}\Big)+O(1/\log T)$$

uniformly for $|u|\le c$. The first equality follows as above or as in the proof of Proposition 1, the second equality again by plugging in (3.1). This completes the proof. □

7. Proof of Corollary 2 and 3

Proof of Corollary 2. Let $x\ge 2$ be as in Theorem 2 with the additional property that $N/\log\log T=O(\log\log T)$. By Theorem 2 and the fact that $\log\log T/\log\log x\to 1$ in this case, we obtain for each $h\in\mathbb{R}$

$$\frac{1}{(\log\log T)/2}\log\Big(\frac{1}{T}\int_T^{2T}e^{h\operatorname{Im}\Sigma_{f,x}(t)}dt\Big)\to h^2/2\qquad as\ T\to\infty. \tag{7.1}$$

By Theorem 16, we obtain that the family $(1/((\log\log T)/2))\operatorname{Im}\Sigma_{f,x}(U_T)$ satisfies the large deviation principle with the speed $1/((\log\log T)/2)$ and

the rate function $I(h) = h^2/2$. Next, consider $\operatorname{Im} r_{f,x}(U_T)$. We will show that there exists a constant $C > 0$ (the constant in (7.4)) such that for each $\delta > 0$

$$(1/T)\lambda(\{t \in [T, 2T] : |\operatorname{Im} r_{f,x}(t)| \geq C\delta \log\log T\}) \\ \leq e^{-(1-o(1))(\delta \log\log T)\log(\delta \log\log T)}. \quad (7.2)$$

We postpone the proof of (7.2) to the end of this section. From (7.2) we deduce that for each $\delta > 0$

$$\frac{1}{(\log\log T)/2} \log\left(\frac{1}{T}\lambda(\{t \in [T, 2T] : |\operatorname{Im} r_{f,x}(t)| \geq \delta \log\log T\})\right) \\ \leq -2(\delta/C)(1 - o(1))(\log\log\log T + \log(\delta/C)).$$

As $T \to \infty$, the right hand side goes to $-\infty$. Hence, by Definition 3, the families $(1/((\log\log T)/2))\operatorname{Im}\log\zeta(1/2 + iU_T)$ and $(1/((\log\log T)/2))\Sigma_{f,x}(U_T)$ are exponentially equivalent. To obtain the statement of the theorem, we finally apply [**22**, Theorem 4.2.13], which states that if two families of random variables are exponentially equivalent, and one of them satisfies the large deviation principle with good rate function I, then the same large deviation principle holds for the other family.

It remains to show (7.2). Therefore, let $V = \delta \log\log T$ and decompose

$$\operatorname{Im} r_{f,x} = \operatorname{Im}(r^*_{g,T^{1/V}} + (\Sigma^*_{g,T^{1/V}} - \Sigma_{g,T^{1/V}}) + (\Sigma_{g,T^{1/V}} - \Sigma_{g,x}) + \Sigma_{g-f,x}).$$

If $|\operatorname{Im} r_{f,x}(t)| \geq CV$, there exists a summand on the right hand side whose absolute value is greater or equal to $CV/4$. Applying the union bound, we obtain

$$\begin{aligned}(1/T)\lambda(\{t \in [T, 2T] : |\operatorname{Im} r_{f,x}(t)| \geq CV\} \\ \leq (1/T)\lambda(\{t \in [T, 2T] : |\operatorname{Im} r^*_{g,T^{1/V}}(t)| \geq CV/4\}) \\ + (1/T)\lambda(\{t \in [T, 2T] : |\operatorname{Im}\Sigma^*_{g,T^{1/V}}(t) - \operatorname{Im}\Sigma_{g,T^{1/V}}(t)| \geq CV/4\}) \\ + (1/T)\lambda(\{t \in [T, 2T] : |\operatorname{Im}\Sigma_{g,T^{1/V}}(t) - \operatorname{Im}\Sigma_{g,x}(t)| \geq CV/4\}) \\ + (1/T)\lambda(\{t \in [T, 2T] : |\operatorname{Im}\Sigma_{g-f,x}(t)| \geq CV/4\}). \quad (7.3)\end{aligned}$$

If we choose Selberg's function $g(u) = e^{-2u}\min(1, 2(1-u))$, we can apply [**66**, Theorem 1], which says that, assuming RH, there exists constants $C, C' > 0$ such that for $2 \leq y \leq t^2$ and $t \geq 2$,

$$|\operatorname{Im} r^*_{g,y}(t)| \leq \left|\frac{C'}{\log y}\sum_{n \leq y}\frac{\Lambda(n)}{n^{1/2+it}} g\left(\frac{\log n}{\log y}\right)\right| + \frac{C}{16}\frac{\log t}{\log y}. \quad (7.4)$$

If we choose $y = T^{1/V}$ and $t \in [T, 2T]$, $T \geq 2$, we have $(C/16)(\log t/\log y) \leq CV/8$. For $T \geq 2$, sufficiently large such that $2 \leq T^{1/V} \leq T^2$, we obtain

$$\begin{aligned}&(1/T)\lambda(\{t \in [T, 2T] : |\operatorname{Im} r^*_{g,T^{1/V}}(t)| \geq CV/4\})\\ &\leq \frac{1}{T}\lambda\bigg(\bigg\{t \in [T, 2T] : \bigg|\frac{C'}{\log T^{1/V}} \sum_{n \leq T^{1/V}} \frac{\Lambda(n)}{n^{1/2+it}} g\bigg(\frac{\log n}{\log T^{1/V}}\bigg)\bigg| \geq CV/8\bigg\}\bigg).\end{aligned}$$

Now, we can apply Markov's inequality and (0.11) to bound the last term by

$$\bigg(\frac{8C'}{CV}\bigg)^{2\lfloor V \rfloor} 3^{2V}(2(AV)^V + O(1)^V) = e^{-(1-o(1))V \log V}.$$

Similarly, by using the other bounds in Appendix A, we can bound the three other terms in (7.3) by $\exp(-(1 - o(1))V \log V)$. Hence, (7.2) follows. This completes the proof. □

Proof of Corollary 3. The asserted formula is exactly content of Varadhan's integral lemma (see Theorem 17). The assumptions of the theorem are satisfied by Corollary 2 and Equation (5.2). □

CHAPTER 2

A theory of nonparametric regression in the presence of complex nuisance components

1. Introduction

In this chapter, we consider the nonparametric random regression model

$$Y = f_1(X_1) + f_2(X_2) + \epsilon. \tag{1.1}$$

We study the problem of estimating the function f_1, while the function f_2 is regarded as a nuisance parameter. We are interested in settings where the second term $f_2(X_2)$ is much more complex than the first term $f_1(X_1)$. A particular model of interest is the additive model

$$Y = f_1(X_1) + \sum_{j=1}^{q-1} f_{2j}(X_{2j}) + \epsilon \tag{1.2}$$

in which the nuisance components f_{2j} are considerably less smooth than f_1 or in which the number of components q is very large, for instance in the sense that q is allowed to increase with the sample size n. The estimation problem is similar to the one arising in semiparametric models where the aim is to estimate a finite-dimensional parameter in the presence of a (more complex) infinite-dimensional parameter.

Estimation in nonparametric additive models is a well-studied topic, especially when considering the problem of estimating all components in the case that q is fixed. One of the seminal theoretical papers is by Stone [**70**], who showed that each component can be estimated with the rate of convergence corresponding to the situation in which the other components are known. Since then, many estimation procedures have been proposed, many of them consisting of several steps. In the work by Linton [**50**] and Fan, Härdle, and Mammen [**26**], it is shown that there exist estimators of single components which have the same asymptotic bias and variance as the corresponding oracle estimators for which the other components are known.

Probably the most popular estimation procedures are the backfitting procedures, which are empirical versions of the orthogonal projection onto the subspace of additive functions in a Hilbert space setting (see, e.g., the book by Hastie and Tibshirani [**35**] and the references therein). This orthogonal projection was studied, e.g., by Breiman and Friedman [**15**] (see also the book by Bickel, Klaassen, Ritov, and Wellner [**8**, Appendix A.4]). They

showed that, under certain conditions including compactness of certain conditional expectation operators, it can be computed by an iterative procedure using only bivariate conditional expectation operators. Replacing these conditional expectation operators by empirical versions leads to the backfitting procedures. Opsomer and Ruppert [**55**] and Opsomer [**54**] computed the asymptotic bias and variance of estimators based on the backfitting procedure in the case where the conditional expectation operators are estimated using local polynomial regression. Mammen, Linton, and Nielsen [**51**] introduced the smooth backfitting procedure and showed that their estimators of single components achieve the same asymptotic bias and variance as oracle estimators for which the other components are known. Concerning the distribution of the covariates, they make some high-level assumptions which are satisfied under some boundedness conditions on the one- and two-dimensional densities. This is still more than is required in the Hilbert space setting (see [**15**]). In the work by Horowitz, Klemelä, and Mammen [**36**], a general two-step procedure was proposed in which a preliminary undersmoothed estimator is based on the smooth backfitting procedure of [**51**]. They also showed that there are estimators which are asymptotically efficient (i.e., achieve the asymptotic minimax risk) with the same constant as in the case with only one component. In addition to the assumptions coming from the results in [**51**], they require a Lipschitz condition for all components.

The problem of estimating f_1 in cases in which $f_2(X_2)$ is more complex than $f_1(X_1)$ is also considered in the work by Efromovich [**24**] and Muro and van de Geer [**75**]. In [**24**], an estimator of f_1 is constructed which is both adaptive to the unknown smoothness and asymptotically efficient with the same constant as in the case with only one component. The assumptions include smoothness and boundedness conditions on the full-dimensional density of (X_1, X_2). The construction of the estimator is involved and starts with a blockwise-shrinkage oracle estimator. In [**75**], a penalized least squares estimator is analyzed in cases where the function f_1 is smoother than the function f_2. Under certain assumptions including smoothness conditions on the design densities, it is shown that for both components, the estimator attains the rate of convergence corresponding to the situation in which the other component is known; i.e., no undersmoothing of the function f_2 is needed to estimate the function f_1.

The previously discussed literature on additive models focuses on the asymptotic behavior of estimators as the number of observations n goes to infinity in the case that q is fixed. Note that one of our purposes is to generalize several results to the case that q increases with n.

More recently, high-dimensional sparse additive models have been studied, e.g., in the work by Meier, van de Geer, and Bühlmann [**53**], Huang, Horowitz, and Wei [**37**], Koltchinskii and Yuan [**44**], Raskutti, Wainwright, and Yu [**58**], Suzuki and Sugiyama [**71**], and Dalalyan, Ingster, and Tsybakov [**19**]. These papers consider the case that the number of covariates q is much larger than the sample size n. The focus is on the problem

of estimating all components under sparsity constraints. In [**19**], e.g., the authors construct an estimator achieving optimal minimax rates of convergence. These rates of convergence depend on q and also on the smallest degree of smoothness of the f_{2j}. Hence, they may only lead to crude bounds for the rates of convergence of estimators of f_1. Let us mention that in this chapter, we do not consider a sparsity scenario. We are interested in cases in which the number of components q is very large, but smaller than n.

In this chapter, we consider model (1.1) in the case that the functions f_1 and f_2 belong to closed subspaces H_1 and H_2 of $\{g_1 \in L^2(\mathbb{P}^{X_1}) : \mathbb{E}[g_1(X_1)] = 0\}$ and $L^2(\mathbb{P}^{X_2})$, respectively. We propose an estimator of f_1 which is based on the composition of two least squares criteria. Our main contribution is to derive several nonasymptotic risk bounds which show that the performance of our estimators is closely related to geometric quantities of H_1 and H_2, such as minimal angles and Hilbert-Schmidt norms. These risk bounds lead to minimal conditions under which the function f_1 can be estimated (up to first order) just as well as in the model $Y = f_1(X_1) + \epsilon$. Our analysis is based on geometric considerations in Hilbert spaces, and relies on the theory of projections on sumspaces in Hilbert spaces (see, e.g., [**8**, Appendix A.4]). Moreover, we apply recent concentration inequalities for structured random matrices (see, e.g., the work by Rauhut [**59**]) in order to show that several geometric properties in the Hilbert space setting carry over to the finite sample setting with high probability. As a main example we apply our results to the additive model (1.2) which corresponds to the case that H_2 has an additive structure. Using our results, we establish new conditions on q and on the smoothness of the nuisance components under which our estimator of f_1 attains the same (nonasymptotic) optimal rate of convergence as the corresponding least squares estimator in the model $Y = f_1(X_1) + \epsilon$. We also address the question of when the corresponding constants coincide.

2. The framework

2.1. The model. Let (Y, X_1, X_2) be a triple of random variables satisfying (1.1), where X_1 and X_2 take values in some measurable spaces $(S_1, \mathcal{B}_1)$ and $(S_2, \mathcal{B}_2)$, respectively, ϵ is a real valued random variable such that $\mathbb{E}[\epsilon|X] = 0$ and $\mathbb{E}\left[\epsilon^2|X\right] = \sigma^2$, and the unknown regression functions satisfy the following assumption:

Assumption 1. *Suppose that $f_1 \in H_1$, where*

$$H_1 \subseteq \{g_1 \in L^2(\mathbb{P}^{X_1}) : \mathbb{E}[g_1(X_1)] = 0\}$$

is a closed subspace, and that $f_2 \in H_2$, where $H_2 \subseteq L^2(\mathbb{P}^{X_2})$ is a closed subspace.

Structural assumptions on f_1 and f_2 (see, e.g., Section 4 where we also consider the additive model) should be incorporated into the model by making assumptions on H_1 and H_2. From the above, we have that $X = (X_1, X_2)$ is a random variable taking values in $(S_1 \times S_2, \mathcal{B}_1 \otimes \mathcal{B}_2)$ (note

that in Section 4.3, we consider the example $S_1 = [0,1]$, $S_2 = [0,1]^{q-1}$, and $S_1 \times S_2 = [0,1]^q$, where all spaces are equipped with the Borel σ-algebra). Moreover, we have that the spaces $L^2(\mathbb{P}^{X_1})$ and $L^2(\mathbb{P}^{X_2})$ are (in a canonical way) subspaces of $L^2(\mathbb{P}^X)$, which implies that H_1 and H_2 are also closed subspaces of $L^2(\mathbb{P}^X)$. Finally, we denote by f the whole regression function given by $f = f_1 + f_2$. We assume that we observe n independent copies

$$(Y^1, X^1), \ldots, (Y^n, X^n)$$

of (Y, X), where $X^i = (X_1^i, X_2^i)$, $1 \leq i \leq n$. Based on this sample, we consider the problem of estimating the function f_1.

2.2. The main assumption. Our approach relies strongly on the fact that the space $L^2(\mathbb{P}^X)$ is a Hilbert space with the inner product $\langle g, h \rangle = \mathbb{E}[g(X)h(X)]$ and the corresponding norm $\|g\| = \sqrt{\langle g, g \rangle}$ (see, e.g., [**23**, Theorem 5.2.1]). In order to state our main assumption, we give the following general definition of a minimal angle in Hilbert spaces (see [**40**, Definition 1] and the references therein).

DEFINITION 1. Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two closed subspaces of a Hilbert space $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$. The minimal angle between $\mathcal{H}_1$ and $\mathcal{H}_2$ is the number $0 \leq \tau_0 \leq \pi/2$ whose cosine is given by

$$\rho_0 = \rho_0(\mathcal{H}_1, \mathcal{H}_2) = \sup \left\{ \frac{\langle h_1, h_2 \rangle}{\|h_1\| \|h_2\|} \;\middle|\; 0 \neq h_1 \in \mathcal{H}_1, 0 \neq h_2 \in \mathcal{H}_2 \right\}.$$

ASSUMPTION 2. *Suppose that the cosine of the minimal angle between H_1 and H_2 is strictly less than 1, i.e.,*

$$\rho_0(H_1, H_2) < 1.$$

The next lemma states two equivalent formulations of Assumption 2. Since we will also apply it to the finite sample setting in later sections, we again give a general statement.

LEMMA 2. *Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two closed subspaces of a Hilbert space $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$. Let $0 \leq \varrho < 1$ be a constant. Then the following assertions are equivalent:*

(i) *For all $0 \neq h_1 \in \mathcal{H}_1, 0 \neq h_2 \in \mathcal{H}_2$ we have*

$$\frac{|\langle h_1, h_2 \rangle|}{\|h_1\| \|h_2\|} \leq \varrho.$$

(ii) *For all $h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2$ we have*

$$\|h_1 + h_2\|^2 \geq (1 - \varrho)(\|h_1\|^2 + \|h_2\|^2).$$

(iii) *For all $h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2$ we have*

$$\|h_1 + h_2\|^2 \geq (1 - \varrho^2)\|h_1\|^2.$$

A proof of Lemma 2 is given in Appendix B.

2.3. The estimation procedure. Let $V_1 \subseteq H_1$ and $V_2 \subseteq H_2$ be d_1- and d_2-dimensional linear subspaces, respectively, and let $W_1 \subseteq V_1$ be a linear subspace. Let $V = V_1 + V_2$ and $d = d_1 + d_2$. By Assumption 2, we have $V_1 \cap V_2 = \{0\}$, which implies that d is equal to the dimension of V and that each $g \in V$ can be decomposed uniquely as $g = g_1 + g_2$ with $g_1 \in V_1$ and $g_2 \in V_2$. We will make only one assumption on V which relates the ∞-norm with the 2-norm, and which will be needed to apply concentration of measure inequalities (compare to, e.g., [**11**, Section 3.1.1] and [**5**, Section 1.1]).

ASSUMPTION 3. *Suppose that there is a real number $\varphi \geq 1$ such that*

$$\|g\|_\infty \leq \varphi\sqrt{d}\|g\| \tag{2.1}$$

for all $g \in V$.

REMARK 1. In view of Assumption 2, Equation (2.1) is satisfied if there are real numbers $\varphi_j \geq 1$ such that $\|g_j\|_\infty \leq \varphi_j\sqrt{d_j}\|g_j\|$ for all $g_j \in V_j$, $j = 1, 2$. Indeed, applying the Cauchy-Schwarz inequality and Lemma 2, we have

$$\|g_1 + g_2\|_\infty \leq \varphi_1\sqrt{d_1}\|g_1\| + \varphi_2\sqrt{d_2}\|g_2\| \leq \frac{\varphi_1 \vee \varphi_2}{\sqrt{1-\rho_0}}\sqrt{d_1+d_2}\|g_1+g_2\|.$$

The construction of our estimator is based on two least squares criteria. First, let $\hat{f}_V$ be the least squares estimator on the model V which is given (not uniquely) by

$$\hat{f}_V = \arg\min_{g\in V} \frac{1}{n}\sum_{i=1}^{n}(Y^i - g(X^i))^2. \tag{2.2}$$

By the definition of V, we have $\hat{f}_V = (\hat{f}_V)_1 + (\hat{f}_V)_2$ with $(\hat{f}_V)_1 \in V_1$ and $(\hat{f}_V)_2 \in V_2$. Next, by applying a second least squares criterion, we define the estimator $\hat{f}_1$ by

$$\hat{f}_1 = \arg\min_{g_1\in W_1} \frac{1}{n}\sum_{i=1}^{n}((\hat{f}_V)_1(X_1^i) - g_1(X_1^i))^2. \tag{2.3}$$

We will also consider the special case $W_1 = V_1$, in which we have $\hat{f}_1 = (\hat{f}_V)_1$. This means that the second least squares criterion can be dropped. However, we will see that choosing V_1 as a preliminary space of larger dimension leads to a smaller bias (it lowers the dependence on ρ_0). Finally, since we want to establish risk bounds, it is convenient to eliminate very large values. Therefore, we define our final estimator $\hat{f}_1^*$ by

$$\hat{f}_1^* = \hat{f}_1 \text{ if } \|\hat{f}_1\|_\infty \leq k_n \text{ and } \hat{f}_1^* \equiv 0 \text{ otherwise,} \tag{2.4}$$

where k_n is a real number to be chosen later (compare to the work by Baraud [**5**, Eq. (3)]). Finally, note that the estimator is not feasible since the distribution of X is not known and therefore the condition $\mathbb{E}[g_1(X_1)] = 0$ cannot be checked. However, one can replace this condition by $(1/n)\sum_{i=1}^n g_1(X_1^i) =$

0. In Appendix B, we show how our results carry over to these modified estimators.

In our analysis of $\hat{f}_1^*$, one important step is to carry over the geometric properties valid in the Hilbert space setting to the finite sample setting. For this, the following event

$$\mathcal{E}_\delta = \left\{(1-\delta)\|g\|^2 \leq \|g\|_n^2 \leq (1+\delta)\|g\|^2 \text{ for all } g \in V\right\},$$

$0 < \delta < 1$, will play the key role. Here, $\|\cdot\|_n$ denotes the empirical norm (see, e.g., Section 5.1). A first observation is that, under Assumptions 1 and 2, the estimator $\hat{f}_1^*$ is unique on the event $\mathcal{E}_\delta$. This can be seen as follows. If $\mathcal{E}_\delta$ holds, then $\|\cdot\|$ and $\|\cdot\|_n$ are equivalent norms on V, which in turn implies that each $g \in V$ is uniquely determined by $(g(X^1), \ldots, g(X^n))^T$. Hence, the solutions of the least squares criteria in (2.2) and (2.3) are unique (since the solutions are unique when restricted to vectors in $\mathbb{R}^n$ evaluated at the observations). Moreover, by Assumption 2, the decomposition $\hat{f}_V = (\hat{f}_V)_1 + (\hat{f}_V)_2$ is unique.

In addition, we also obtain a simple representation of our estimator. Therefore, let $\hat{\Pi}_V$ be the orthogonal projection from $\mathbb{R}^n$ to the subspace $\{(g(X^1), \ldots, g(X^n))^T | g \in V\}$, and let $\hat{\Pi}_{W_1}$ be defined analogously. If $\mathcal{E}_\delta$ holds, then we have

$$(\hat{f}_1(X_1^1), \ldots, \hat{f}_1(X_1^n))^T = \hat{\Pi}_{W_1}(\hat{\Pi}_V \mathbf{Y})_1,$$

where $\hat{\Pi}_V \mathbf{Y} = (\hat{\Pi}_V \mathbf{Y})_1 + (\hat{\Pi}_V \mathbf{Y})_2$ is the unique decomposition of the least squares estimator on the model V, considered as a vector in $\mathbb{R}^n$, with $(\hat{\Pi}_V \mathbf{Y})_j \in \{(g_j(X_j^1), \ldots, g_j(X_j^n))^T | g_j \in V_j\}$.

3. Main results

3.1. A first risk bound. In this section, we present a nonasymptotic risk bound in the case $W_1 = V_1$, which will be further improved (under additional assumptions) in later sections. We denote by Π_V (resp. Π_{V_1}, Π_{V_2}, and Π_{W_1}) the orthogonal projection from $L^2(\mathbb{P}^X)$ to the subspace V (resp. V_1, V_2, and W_1).

THEOREM 3. *Let Assumption 1, 2, and 3 be satisfied. Let $0 < \delta < 1$ be a real number. Let $W_1 = V_1$. Then*

$$\begin{aligned}&\mathbb{E}\left[\|f_1 - \hat{f}_1^*\|^2\right] \\ &\leq \frac{1+\delta}{(1-\delta)^3}\frac{1}{1-\rho_0^2}\left(\left(1 + \frac{\varphi^2 d}{n}\right)\|f - \Pi_V f\|^2 + \frac{\sigma^2 \dim V_1}{n}\right) + R_n\end{aligned}$$

with

$$R_n = \frac{2(1+\delta)\varphi^2 d\|f_1\|^2(\|f - \Pi_{V_2} f\|^2 + \sigma^2)}{(1-\delta)^2(1-\rho_0^2)k_n^2} + 2(\|f_1\| + k_n)^2 d \exp\left(-\kappa\frac{\delta^2 n}{\varphi^2 d}\right),$$

where κ is the universal constant in Theorem 9.

Before we discuss the two main terms, let us give conditions under which the remainder term R_n is small. Suppose that for some real number $c > 0$, we have

$$\varphi^2 d \leq \frac{c\delta^2 n}{\log n},$$

and let $k_n^2 = \|f_1\|^2 n^{\kappa/(2c)}$ (this is a theoretical choice of k_n leading to a simple upper bound for R_n, many other choices are possible, too). Then one can show that

$$R_n \leq \frac{12c(1+\delta)\delta^2}{(1-\delta)^2(1-\rho_0^2)} \left(\|f_1\|^2 + \|f_2 - \Pi_{V_2} f_2\|^2 + \sigma^2\right) n^{-\frac{\kappa}{2c}+1}.$$

Letting, e.g., $\delta = 1/\log n$ and $c = 1/\log n$, we obtain the following corollary of Theorem 3.

COROLLARY 5. *Let Assumption 1, 2, and 3 be satisfied. Suppose that*

$$\varphi^2 d \leq \frac{n}{(\log n)^4}. \tag{3.1}$$

Then there is a universal constant $C > 0$ such that

$$\begin{aligned}
&\mathbb{E}\left[\|f_1 - \hat{f}_1^*\|^2\right] \\
&\leq \frac{1}{1-\rho_0^2}\left(\|f_1 - \Pi_{V_1} f_1\|^2 + \frac{\sigma^2 \dim V_1}{n}\right)(1 + C/\log n) \\
&+ \frac{C}{1-\rho_0^2}\left((\log n)\|f_2 - \Pi_{V_2} f_2\|^2 + \|f_1\|^2 n^{-\frac{\kappa}{2}\log n + 1}\right).
\end{aligned}$$

The first two terms on the right hand side are (up to the factor $(1-\rho_0^2)^{-1}$) equal to the bias term and the variance term of the same estimator with $V_2 = 0$ in the model $Y = f_1(X_1) + \epsilon$. The third term is the approximation error of the function f_2 with respect to the space V_2. It decreases if V_2 is chosen larger. Moreover, the choice of V_2 does not effect any of the other terms, the only restriction is given by (3.1). The question arising now is as follows: Is it possible to choose a space V_2 subject to the constraint (3.1) such that $(1-\rho_0^2)^{-1}(\log n)\|f_2 - \Pi_{V_2} f_2\|^2$ is negligible with respect to the first two terms.

3.2. A refined risk bound. In this section, we improve Theorem 3 such that the factor $(1-\rho_0^2)^{-1}$ only appears in remainder terms. Since the refined upper bound for the variance term will also contain a Hilbert-Schmidt norm, we give the following general definition (see, e.g., [**81**]).

DEFINITION 2. Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be Hilbert spaces. A bounded linear operator $T : \mathcal{H}_1 \to \mathcal{H}_2$ is called Hilbert-Schmidt if for some orthonormal basis $\{\phi_{1\alpha}\}_{\alpha \in I}$ of $\mathcal{H}_1$,

$$\sum_{\alpha \in I} \|T\phi_{1\alpha}\|^2 < \infty. \tag{3.2}$$

This sum is independent of the choice of the orthonormal basis (see [**81**, Satz 3.18]). The square root of this sum is called the Hilbert-Schmidt norm of T, denoted by $\|T\|_{HS}$.

Let Π_{V_2} be the orthogonal projection from $L^2(\mathbb{P}^X)$ to V_2, and let $\Pi_{V_2}|_{W_1}$ be the restriction of Π_{V_2} to W_1. Then $\Pi_{V_2}|_{W_1}$ is a Hilbert-Schmidt operator, since W_1 is finite-dimensional. We prove:

THEOREM 4. *Let Assumption 1, 2, and 3 be satisfied. Let $0 < \delta < 1$ be a real number. Then*

$$\begin{aligned}
&\mathbb{E}\left[\|f_1 - \hat{f}_1^*\|^2\right] \\
&\leq \left(\|f_1 - \Pi_{W_1} f_1\|^2 + \frac{1}{1-\delta}\frac{\sigma^2 \dim W_1}{n}\right)\left(1 + \frac{1+\delta}{(1-\delta)^3}\frac{1}{1-\rho_0^2}\frac{2\varphi^2 d}{n}\right) \\
&+ \frac{1+\delta}{(1-\delta)^2}\frac{6}{1-\rho_0^2}\left(\|f_1 - \Pi_{V_1} f_1\|^2 + \|f_2 - \Pi_{V_2} f_2\|^2\right) \\
&+ \frac{1+\delta}{(1-\delta)^4}\frac{1}{1-\rho_0^2}\frac{\sigma^2 \|\Pi_{V_2}|_{W_1}\|_{HS}^2}{n} + R_n, \qquad (3.3)
\end{aligned}$$

where R_n is given in Theorem 3.

In order to state a corollary of Theorem 4 similar to Corollary 5, we have to discuss the quantity $\|\Pi_{V_2}|_{W_1}\|_{HS}^2$. If $\{\phi_{1k}\}_{1 \leq k \leq \dim W_1}$ is an orthonormal basis of W_1, then it can be bounded as follows:

$$\|\Pi_{V_2}|_{W_1}\|_{HS}^2 = \sum_{k=1}^{\dim W_1} \|\Pi_{V_2}\phi_{1k}\|^2 \leq \sum_{k=1}^{\dim W_1} \rho_0^2 \|\phi_{1k}\|^2 = \rho_0^2 \dim W_1, \qquad (3.4)$$

where the inequality can be shown as in (5.7). Using this bound, we get a variance term which coincides (up to first order) with the one in Theorem 3. However, (3.4) can be considerably improved under certain Hilbert-Schmidt Assumptions. In particular, we will derive upper bounds which are dimension free. The first assumption is as follows:

ASSUMPTION 4. *Suppose that there are measures ν_1 and ν_2 on $\mathcal{B}_1$ and $\mathcal{B}_2$, respectively, such that X has the density p with respect to the product measure $\nu_1 \otimes \nu_2$. Let p_1 and p_2 be the marginal densities of X_1 and X_2 with respect to the measures ν_1 and ν_2, respectively. Suppose that*

$$\begin{aligned}
\|K\|_{HS}^2 &= \int_{S_2}\int_{S_1}\left(\frac{p(x_1,x_2)}{p_1(x_1)p_2(x_2)}\right)^2 p_1(x_1)p_2(x_2)d\nu_1(x_1)d\nu_2(x_2) \\
&= \int_{S_2}\int_{S_1}\frac{(p(x_1,x_2))^2}{p_1(x_1)p_2(x_2)}d\nu_1(x_1)d\nu_2(x_2) < \infty.
\end{aligned}$$

If Assumption 4 is satisfied, then we can define the integral operator $K : L^2(\mathbb{P}^{X_1}) \to L^2(\mathbb{P}^{X_2})$ by

$$(Kg_1)(x_2) = \int_{S_1} g_1(x_1)\frac{p(x_1,x_2)}{p_1(x_1)p_2(x_2)}p_1(x_1)d\nu_1(x_1)$$

which is the orthogonal projection from $L^2(\mathbb{P}^X)$ to $L^2(\mathbb{P}^{X_2})$ restricted to $L^2(\mathbb{P}^{X_1})$. Applying [**81**, Satz 3.19], we obtain that K is a Hilbert-Schmidt operator with Hilbert-Schmidt norm $\|K\|_{HS}$. We conclude that

$$\|\Pi_{V_2}|_{W_1}\|_{HS} \leq \|K\|_{HS}.$$

Next, we present a more sophisticated upper bound, by using the spaces H_1 and H_2 instead of $L^2(\mathbb{P}^{X_1})$ and $L^2(\mathbb{P}^{X_2})$. Let Π_{H_2} be the orthogonal projection from $L^2(\mathbb{P}^X)$ to H_2, and let $\Pi_{H_2}|_{H_1}$ be the restriction of Π_{H_2} to H_1.

ASSUMPTION 5 (Weaker form of Assumption 4). *Suppose that $\Pi_{H_2}|_{H_1}$ is a Hilbert-Schmidt operator.*

If Assumption 5 is satisfied, then

$$\|\Pi_{V_2}|_{W_1}\|_{HS} \leq \|\Pi_{H_2}|_{H_1}\|_{HS}\,.$$

Letting now $\delta = 1/\log n$ and $c = 1/\log n$ as in Corollary 5, we obtain the following corollary of Theorem 4.

COROLLARY 6. *Let Assumption 1, 2, 3, and 4 be satisfied. Suppose that*

$$\varphi^2 d \leq \frac{n}{(\log n)^4}.$$

Then there is a universal constant $C > 0$ such that

$$\begin{aligned}
&\mathbb{E}\left[\|f_1 - \hat{f}_1^*\|^2\right] \\
&\leq \left(\|f_1 - \Pi_{W_1} f_1\|^2 + \frac{\sigma^2 \dim W_1}{n}\right)\left(1 + C'/\log n\right) \\
&+ C'\left(\|f_1 - \Pi_{V_1} f_1\|^2 + \|f_2 - \Pi_{V_2} f_2\|^2 + \frac{\sigma^2 \|K\|_{HS}^2}{n} + \frac{\|f_1\|^2}{n^{\frac{\kappa}{2}\log n - 1}}\right),
\end{aligned}$$

where $C' = C/(1-\rho_0^2)$. Moreover, if Assumption 5 holds instead of Assumption 4, then the above inequality holds if $\|K\|_{HS}^2$ is replaced by $\|\Pi_{H_2}|_{H_1}\|_{HS}^2$. Finally, if Assumption 5 and 4 are not satisfied, then the above inequality holds if $\|K\|_{HS}^2$ is replaced by $\rho_0^2 \dim W_1$.

Now the first two terms in the brackets on the right hand side are equal to the bias term and the variance term of the same estimator with $V_2 = 0$ in the model $Y = f_1(X_1) + \epsilon$. As in Corollary 5, we see that the choices of V_1 and V_2 do not effect any of the other terms, the only restriction is given by (3.1).

Finally, we give an alternative representation of the Hilbert-Schmidt norm $\|\Pi_{H_2}|_{H_1}\|_{HS}$ using the operator $\Pi_{H_1}\Pi_{H_2}\Pi_{H_1}$ (which we consider as a map from H_1 to H_1). To simplify the exposition, we suppose that H_1 is separable, which implies that each orthonormal basis of H_1 is countable (see, e.g., [**60**, Chapter II]). From Assumption 5, it follows that $\Pi_{H_1}\Pi_{H_2}\Pi_{H_1}$ is compact (see, e.g., [**49**, Chapter 30.8]). Since it is also symmetric and positive, the spectral theorem (see, e.g., [**49**, Theorem 3 in Chapter 28])

implies that there is an orthonormal basis for H_1 consisting of eigenvectors of $\Pi_{H_1}\Pi_{H_2}\Pi_{H_1}$. These all have non-negative eigenvalues. We arrange the positive eigenvalues of $\Pi_{H_1}\Pi_{H_2}\Pi_{H_1}$ in decreasing order: $\alpha_1 \geq \alpha_2 \cdots > 0$. We now have:

LEMMA 3. *Under the above assumptions, we have*

$$\rho_0^2 = \alpha_1$$

and

$$\|\Pi_{H_2}|_{H_1}\|_{HS}^2 = \operatorname{tr}\left(\Pi_{H_1}\Pi_{H_2}\Pi_{H_1}\right) = \sum_{k\geq 1}\alpha_k.$$

PROOF. We only prove the second equality. Let $\{\phi_{1k}\}_{k\geq 1}$ be an orthonormal basis for H_1 consisting of eigenvectors of $\Pi_{H_1}\Pi_{H_2}\Pi_{H_1}$. Then

$$\|\Pi_{H_2}|_{H_1}\|_{HS}^2 = \sum_{k\geq 1}\|\Pi_{H_2}\phi_{1k}\|^2 = \sum_{k\geq 1}\langle \Pi_{H_1}\Pi_{H_2}\Pi_{H_1}\phi_{1k}, \phi_{1k}\rangle^2 = \sum_{k\geq 1}\alpha_k.$$

□

EXAMPLE 1. Consider the case that $X = (X_1, X_2)$ is a bivariate Gaussian random variable such that $\mathbb{E}[X_1] = \mathbb{E}[X_2] = 0$, $\mathbb{E}[X_1^2] = \mathbb{E}[X_2^2] = 1$, and $\mathbb{E}\left[X_1X_2\right] = \rho$.

First, suppose that H_1 and H_2 are the spaces of linear centered functions, i.e., $H_1 = \{g_1 : g_1(x_1) = a\cdot x_1, a \in \mathbb{R}\}$ and $H_2 = \{g_2 : g_2(x_2) = a\cdot x_2, a \in \mathbb{R}\}$. Then it is easy to see that

$$\rho_0 = |\rho|$$

and

$$\|\Pi_{H_2}|_{H_1}\|_{HS}^2 = \rho^2.$$

Second, suppose that $H_1 = \{g_1 \in L^2(\mathbb{P}^{X_1}) : \mathbb{E}\left[g_1(X_1)\right] = 0\}$ and $H_2 = L^2(\mathbb{P}^{X_2})$. Then it follows from [**46**] that $\Pi_{H_1}\Pi_{H_2}\Pi_{H_1}$ has eigenvalues $\{\rho^2, \rho^4, \dots\}$. Hence, the above lemma implies that

$$\rho_0 = |\rho|$$

and

$$\|\Pi_{H_2}|_{H_1}\|_{HS}^2 = \sum_{k=1}^{\infty}\rho^{2k} = \frac{\rho^2}{1-\rho^2},$$

which is an improvement over (3.4) if $\dim W_1$ is large.

3.3. Regularity conditions on the design densities. In this section, we present two improvements of Theorem 4 which are possible under Assumption 4 and additional regularity conditions on the design densities. In particular, we show that the dependence of the bias term on the function f_2 can decrease considerably.

By Assumption 4 and Fubini's theorem, we have

$$\frac{p(x_1,\cdot)}{p_1(x_1)p_2(\cdot)} \in L^2(\mathbb{P}^{X_2}) \tag{3.5}$$

for $\mathbb{P}^{X_1}$-almost all x_1. Thus we can make the following assumption. Suppose that there is a real number $\psi(V_2)$ and a function $h_1 \in L^2(\mathbb{P}^{X_1})$ such that

$$\left\| (1 - \Pi_{V_2}) \frac{p(x_1, \cdot)}{p_1(x_1) p_2(\cdot)} \right\|_{L^2(\mathbb{P}^{X_2})} \leq h_1(x_1) \psi(V_2) \tag{3.6}$$

for $\mathbb{P}^{X_1}$-almost all x_1. In analogy, we let $\phi(V_2)$ be a real number such that $\|f_2 - \Pi_{V_2} f_2\| \leq \phi(V_2)$. We prove:

THEOREM 5. *Let Assumption 1, 2, 3, and 4 be satisfied. Let $0 < \delta < 1$ be a real number. Suppose that* (3.6) *is satisfied. Moreover, suppose that $\|g_1\|_\infty \leq \varphi \sqrt{d_1} \|g_1\|$ for all $g_1 \in V_1$, where φ is the constant from Assumption 3. Then* (3.3) *holds when*

$$\frac{1+\delta}{(1-\delta)^2} \frac{6}{1-\rho_0^2} \left(\|f_1 - \Pi_{V_1} f_1\|^2 + \|f_2 - \Pi_{V_2} f_2\|^2 \right)$$

is replaced by

$$\frac{(1+\delta)^2}{(1-\delta)^4} \frac{12}{1-\rho_0^2} \left(\|f_1 - \Pi_{V_1} f_1\|^2 + \frac{\|h_1\|^2 (\phi(V_2)\psi(V_2))^2}{1-\rho_0^2} + \frac{1}{n} \frac{\|h_1\|^2 \|f_2 - \Pi_{V_2} f_2\|_\infty^2 (\psi(V_2))^2}{1-\rho_0^2} + \frac{(\phi(V_2))^2}{1-\rho_0^2} \frac{\varphi^2 d_1}{n} \right).$$

Theorem 5 shows that the regularity conditions on $p/(p_1 p_2)$ and f_2 have similar effects, which can be seen from second term. In contrast to Theorems 3 and 4, Theorem 5 shows that the estimator $\hat{f}_1^*$ can also behave well when f_2 is considerably less regular than f_1. For instance, if we apply Theorem 5 to an asymptotic scenario, then, under suitable conditions on $\psi(V_2)$, the regularity conditions on f_2 can be (almost) reduced to $\phi(V_2) \to 0$ (see, e.g., Corollary 10).

For fixed x_1, let the function $r(x_1, \cdot)$ be the orthogonal projection of $p(x_1, \cdot)/(p_1(x_1)p_2(\cdot))$ on H_2. By (3.5), $r(x_1, \cdot)$ is defined for $\mathbb{P}^{X_1}$-almost all x_1. Thus we can consider the following weaker version of (3.6). Suppose that there exists a real number $\psi_\Pi(V_2)$ and a function $h_1 \in L^2(\mathbb{P}^{X_1})$ such that

$$\|(1 - \Pi_{V_2}) r(x_1, \cdot)\|_{L^2(\mathbb{P}^{X_2})} \leq h_1(x_1) \psi_\Pi(V_2) \tag{3.7}$$

for $\mathbb{P}^{X_1}$-almost all x_1. If (3.7) holds, then we obtain the following theorem. Note that, compared to Theorem 5, the last term is not always negligible.

THEOREM 6. *Let Assumption 1, 2, 3, and 4 be satisfied. Let $0 < \delta < 1$ be a real number. Suppose that* (3.7) *is satisfied. Then Theorem 4 also holds when the term*

$$\frac{1+\delta}{(1-\delta)^2} \frac{6}{1-\rho_0^2} \left(\|f_1 - \Pi_{V_1} f_1\|^2 + \|f_2 - \Pi_{V_2} f_2\|^2 \right)$$

is replaced by

$$\frac{1+\delta}{(1-\delta)^3}\frac{6}{1-\rho_0^2}\left(\|f_1-\Pi_{V_1}f_1\|^2\left(1+\frac{2\varphi^2 d}{n}\right)\right.$$
$$\left.+\frac{\|h_1\|^2(\phi(V_2)\psi_\Pi(V_2))^2}{1-\rho_0^2}+\frac{(\phi(V_2))^2}{1-\rho_0^2}\frac{2\varphi^2 d}{n}\right).$$

4. Applications

4.1. The two-dimensional case. In this section, we want to discuss Theorem 3 and 4 in the case that X_1 and X_2 take values in $\mathbb{R}$ and that Assumptions 1 and 2 are satisfied with

$$H_1=\{g_1\in L^2(\mathbb{P}^{X_1})|\mathbb{E}\,[g_1(X_1)]=0\}$$

and

$$H_2=L^2(\mathbb{P}^{X_2}).$$

The main remaining issue is to bound the approximation errors. This is possible if the f_j belong to certain nonparametric classes of functions and if the V_j are chosen appropriately. Here, we shall restrict our attention to (periodic) Sobolev smoothness and spaces of trigonometric polynomials. Note that we will also consider Hölder smoothness and spaces of piecewise polynomials in Section 4.4 and 4.5. Recall that the trigonometric basis is given by $\phi_0(x)=1$, $\phi_k(x)=\sqrt{2}\cos(2\pi kx)$ and $\phi_{-k}(x)=\sqrt{2}\sin(2\pi kx)$, $k\geq 1$, where $x\in[0,1]$.

Assumption 6. *Suppose that the X_j take values in $[0,1]$ and have densities p_{X_j} with respect to the Lebesgue measure on $[0,1]$, which satisfy $c\leq p_{X_j}\leq 1/c$ for some constant $c>0$. Moreover, suppose that the f_j belong to the Sobolev classes*

$$\tilde{W}_j(\alpha_j,K_j)=\left\{\sum_{k\in\mathbb{Z}}\theta_k\phi_k(x_j)\ :\ \sum_{k\in\mathbb{Z}}|k|^{2\alpha_j}\theta_k^2\leq K_j^2\right\},$$

where $\alpha_j>0$ and $K_j>0$ (see, e.g., [**74**, Definition 1.12]*).*

For $j=1,2$, let V_j be the intersection of H_j with the linear span of the ϕ_k (in the variable x_j) such that $|k|\leq m_j$, and let W_1 be the intersection of H_1 with the linear span of the ϕ_k (in the variable x_1) such that $|k|\leq m_{W_1}$. Note that we have $d_1=2m_1$ and $d_2=2m_2+1$. Using the definition of the $\tilde{W}_j(\alpha_j,K_j)$, we have for all $h_j\in\tilde{W}_j(\alpha_j,K_j)\cap H_j$,

$$\|h_j-\Pi_{V_j}h_j\|^2\leq(1/c)K_j^2(1+m_j)^{-2\alpha_j}$$

and thus

$$\|h_j-\Pi_{V_j}h_j\|^2\leq C_jK_j^2d_j^{-2\alpha_j},\tag{4.1}$$

where $C_j=2^{2\alpha_j}/c$. The same bound holds if V_1 and d_1 are replaced by W_1 and $\dim W_1$. Moreover, by applying the Cauchy-Schwarz inequality, we have $\|g_j\|_\infty\leq\sqrt{1/c}\sqrt{2m_j+1}\|g_j\|$ for all $g_j\in V_j$, which implies that

$\|g_j\|_\infty \sqrt{2/c}\sqrt{d_j}\|g_j\|$ for all $g_j \in V_j$. By Remark 1, this implies that Assumption 3 is satisfied with

$$\varphi^2 \leq \frac{2}{c(1-\rho_0)}. \tag{4.2}$$

We now choose d_1 and d_2 as the smallest possible integer satisfying

$$d_j \geq \frac{n}{4\varphi^2 \log^4 n}, \tag{4.3}$$

and we choose $\dim W_1$ (up to constant) equal to

$$\left(\frac{K_1^2 n}{\sigma^2}\right)^{\frac{1}{2\alpha_1+1}}.$$

If n is large enough, then these choices imply that (3.1) of Corollary 5 is satisfied. Applying (4.1) and (4.3), we obtain that

$$\|f_2 - \Pi_{V_2} f_2\|^2 \leq C_2 K_2^2 \left(\frac{n}{4\varphi^2 \log^4 n}\right)^{-2\alpha_2},$$

where the last expression is $o(n^{-(\alpha_1/(2\alpha_1+1))})$ if $\alpha_2 > \alpha_1/(2\alpha_1+1)$. Finally, note that $\|f_1\|^2 \leq \mathbb{E}[(f_1(X_1)-\theta_0)^2] \leq K_1^2/c$. From Corollary 6, we now obtain the following asymptotic result when the sample size n tends to infinity:

COROLLARY 7. *Let Assumption 2 and 6 be satisfied. Suppose that*

$$\alpha_2 > \alpha_1/(2\alpha_1+1). \tag{4.4}$$

Then

$$\limsup_{n\to\infty} \sup_{f_j \in \tilde{W}_j(\alpha_j, K_j)} \mathbb{E}\left[n^{\frac{2\alpha_1}{2\alpha_1+1}} \|f_1 - \hat{f}_1^*\|^2\right] \leq C\sigma^{\frac{4\alpha_1}{2\alpha_1+1}} K_1^{\frac{2}{2\alpha_1+1}}, \tag{4.5}$$

where C is a constant depending only on α_1, c, and ρ_0. If, in addition, Assumption 5 holds, then the dependence of C on ρ_0 disappears and we obtain the same constant as for the corresponding estimator with $V_2 = 0$ in the model $Y = f_1(X_1) + \epsilon$.

REMARK 2. Corollary 7 says that the estimator $\hat{f}_1^*$ attains the optimal rate of convergence in a minimax sense. Note that one can also take the supremum over random variables X such that c, $1/c$, and $1/(1-\rho_0^2)$ are bounded by a fixed constant.

REMARK 3. The assumptions of Corollary 7 are weaker than those needed in [**36**], where it is also assumed that the joint density of (X_1, X_2) is bounded away from zero and infinity and that the functions f_1 and f_2 are Lipschitz continuous. Note that (4.4) is always satisfied if, e.g., $\alpha_2 \geq 1/2$ and that the boundedness conditions imply Assumption 4 and also Lemma 2 (ii). Hence, the boundedness conditions imply Assumption 2 and 5.

REMARK 4. In semiparametric models, one often requires a global rate of convergence of order $o(n^{-1/4})$ in order to obtain the rate of convergence $n^{-1/2}$ for the parametric component (see, e.g., the book by van de Geer [**76**, Chapter 11]). Since $\alpha_1/(2\alpha_1+1)$ goes to $1/2$ as $\alpha_1 \to \infty$, condition (4.4) extends this result to the nonparametric case.

4.2. The multidimensional case. Now, we suppose that the X_1 and X_2 take values in $[0,1]^{q_1}$ and $[0,1]^{q_2}$, respectively, and that Assumptions 1 and 2 are again satisfied with

$$H_1 = \{g_1 \in L^2(\mathbb{P}^{X_1}) | \mathbb{E}\,[g_1(X_1)] = 0\}$$

and

$$H_2 = L^2(\mathbb{P}^{X_2}).$$

In this case, we consider the tensor product Fourier basis:

$$\phi_k(x_j) = \prod_{l=1}^{q_j} \phi_{k_l}(x_{jl}),$$

where $k \in \mathbb{Z}^{q_j}$ and $x_j \in [0,1]^{q_j}$. We define the following Sobolev class

$$\tilde{W}_j(\alpha_j, K_j) = \left\{ \sum_{k\in\mathbb{Z}^{q_j}}^{\infty} \theta_k \phi_k(x_j) \;:\; \sum_{k\in\mathbb{Z}^{q_j}}^{\infty} a_{jk}^2 \theta_k^2 \le K_j^2 \right\}$$

with $a_{jk} = \|k\|_\infty^{\alpha_j}$, where $\|k\|_\infty = \max_{l=1,\dots,q_j} |k_l|$, $\alpha_j > 0$, and $K_j > 0$ (an alternative choice would be, e.g., $a_{jk}^2 = \sum_{l=1}^{q_j} |k_l|^{\alpha_j}$).

ASSUMPTION 7. *Suppose that the X_j have densities p_{X_j} with respect to the Lebesgue measure on $[0,1]^{q_j}$, which satisfy $c \le p_{X_j} \le 1/c$ for some constant $c > 0$. Moreover, suppose that the f_j belong to the Sobolev classes $\tilde{W}_j(\alpha_j, K_j)$, where $\alpha_j > 0$ and $K_j > 0$.*

For $j = 1, 2$, let V_j be the intersection of H_j with the linear span of the ϕ_k (in the variable x_j) such that $\|k\|_\infty \le m_j$, and let W_1 be the intersection of H_1 with the linear span of the ϕ_k (in the variable x_1) such that $\|k\|_\infty \le m_{W_1}$. Note that we have $d_1 = (2m_1+1)^{q_1} - 1$ and $d_2 = (2m_2+1)^{q_2}$. Similarly as above, one can show that

$$\|h_j - \Pi_{V_j} h_j\|^2 \le C_j K_j^2 d_j^{-2\alpha_j/q_j}$$

for all $h_j \in \tilde{W}_j(\alpha_j, K_j) \cap H_j$, where $C_j = 2^{2\alpha_j}/c$, and that Assumption 3 holds with a φ satisfying again (4.2). We now choose d_1 and d_2 as in (4.3), and we choose $\dim W_1$ (up to constant) equal to

$$\left(\frac{K_1^2 n}{\sigma^2}\right)^{\frac{q_1}{2\alpha_1+q_1}}.$$

Similarly as above, we conclude:

COROLLARY 8. *Let Assumption 2 and 7 be satisfied. Suppose that*

$$\alpha_2/q_2 > \alpha_1/(2\alpha_1 + q_1).$$

Then

$$\limsup_{n\to\infty} \sup_{f_j \in \tilde{W}_j(\alpha_j, K_j)} \mathbb{E}\left[n^{\frac{2\alpha_1}{2\alpha_1+q_1}} \|f_1 - \hat{f}_1^*\|^2 \right] \leq C\sigma^{\frac{4\alpha_1}{2\alpha_1+q_1}} K_1^{\frac{2q_1}{2\alpha_1+q_1}},$$

where C is a constant depending only on α_1, c, and ρ_0. If, in addition, Assumption 5 holds, then the dependence of C on ρ_0 disappears and we obtain the same constant as for the corresponding estimator with $V_2 = 0$ in the model $Y = f_1(X_1) + \epsilon$.

4.3. The additive model with Sobolev smoothness. In this section, we want to discuss Theorem 3 and 4 in the case that the random variables X_1 and X_2 take values in $\mathbb{R}$ and $\mathbb{R}^{q-1}$, $q \geq 2$, respectively, and that Assumptions 1 and 2 are satisfied with

$$H_1 = \{g_1 \in L^2(\mathbb{P}^{X_1}) | \mathbb{E}[g_1(X_1)] = 0\}$$

and

$$H_2 = \sum_{j=1}^{q-1} L^2(\mathbb{P}^{X_{2j}}),$$

where the X_{2j}, $j = 1, \ldots, q-1$, are the components of X_2. If we define $H_{2j} = \{g_{2j} \in L^2(\mathbb{P}^{X_{2j}}) | \mathbb{E}[g_{2j}(X_{2j})] = 0\}$, if $j = 1, \ldots, q-2$, and $H_{2j} = L^2(\mathbb{P}^{X_{2j}})$, if $j = q-1$, then we can write $H_2 = \sum_{j=1}^{q-1} H_{2j}$.

ASSUMPTION 8. *Suppose that X_1 and the X_{2j} take values in $[0,1]$ and have densities p_{X_1} and $p_{X_{2j}}$ with respect to the Lebesgue measure on $[0,1]$, which satisfy $c \leq p_{X_1} \leq 1/c$ and $c \leq p_{X_{2j}} \leq 1/c$ for some constant $c > 0$.*

Moreover, suppose that $f_1 \in \tilde{W}_1(\alpha_1, K_1)$, where $\alpha_1 > 0$ and $K_1 > 0$, and that there is a decomposition $f_2 = \sum_{j=1}^{q-1} f_{2j}$ such that $f_{2j} \in \tilde{W}_{2j}(\alpha_2, K_2) \cap H_{2j}$, where $\alpha_2 > 0$ and $K_2 > 0$.

Now, let V_1 and W_1 be as in Section 4.1, and let $V_2 = \sum_{j=1}^{q-1} V_{2j}$, where V_{2j} is the intersection of H_{2j} with the linear span of the ϕ_k (in the variable x_{2j}) such that $|k| \leq m_2$. In order to show that Assumption 3 is satisfied, we we need the following additional assumption on H_2.

ASSUMPTION 9. *There is an $\epsilon_2 < 1$ such that for each $h_2 \in H_2$, there is a decomposition $h_2 = \sum_{j=1}^{q-1} h_{2j}$ with $h_{2j} \in H_{2j}$ such that*

$$\|h_2\|^2 \geq (1 - \epsilon_2) \sum_{j=1}^{q-1} \|h_{2j}\|^2. \tag{4.6}$$

Applying iteratively [**8**, Proposition 2.A in Appendix A.4], one can show that Assumption 9 is equivalent to the assertion that $\sum_{j\in J} H_{2j}$ is closed for all $J \subseteq \{1, \ldots, q-1\}$. In particular, Assumption 9 implies that H_2 is closed

meaning that Assumption 1 is included in Assumption 9. We mention that Assumption 9 can also be related to bounds on certain complementary angles (see [**8**, Definition 2 and Proposition 2.D in Appendix A.4]).

LEMMA 4. *Let Assumption 2, 8, and 9 be satisfied. Then Assumption 3 holds with a constant φ satisfying*

$$\varphi^2 \leq \frac{2}{c(1-\rho_0)(1-\epsilon_2)} \frac{d_1 + \sum_{j=1}^{q-1} d_{2j}}{d},$$

where $d_{2j} = \dim V_{2j}$. In particular, if the V_{2j} are linearly independent, then we have

$$\varphi^2 \leq \frac{2}{c(1-\rho_0)(1-\epsilon_2)}. \tag{4.7}$$

A proof of Lemma 4 is given in Appendix B. Thus (3.1) of Corollary 5 is satisfied if

$$\frac{2\left(d_1 + \sum_{j=1}^{q-1} d_{2j}\right)}{c(1-\rho_0)(1-\epsilon_2)} \leq \frac{n}{\log^4 n}.$$

We now choose $\dim W_1$ as in Section 4.1, and we choose d_1 and the d_{2j} equal to the smallest possible integers satisfying

$$d_1 \geq \frac{c(1-\rho_0)(1-\epsilon_2)n}{8\log^4 n}$$

and

$$d_{2j} \geq \frac{c(1-\rho_0)(1-\epsilon_2)n}{8(q-1)\log^4 n}. \tag{4.8}$$

If the right hand side of (4.8) is greater than or equal to 2, then (3.1) of Corollary 5 is satisfied. In order to bound the approximation error $\|f_2 - \Pi_{V_2} f_2\|$, we introduce ϵ_2' which is the smallest number such that

$$\|h_2\|^2 \leq (1+\epsilon_2') \sum_{j=1}^{q-1} \|h_{2j}\|^2 \tag{4.9}$$

for all $h_2 = \sum_{j=1}^{q-1} h_{2j}$ with $h_{2j} \in H_{2j}$. Note that

$$1 + \epsilon_2' \leq q - 1,$$

by the Cauchy-Schwarz inequality. Using the decomposition of f_2 in Assumption 8, the projection theorem, (4.9), and finally (4.1) and (4.8), we

have

$$\begin{aligned}\|f_2 - \Pi_{V_2} f_2\|^2 &\leq \left\| \sum_{j=1}^{q-1} (f_{2j} - \Pi_{V_{2j}} f_{2j}) \right\|^2 \\ &\leq (1+\epsilon_2') \sum_{j=1}^{q-1} \|f_{2j} - \Pi_{V_{2j}} f_{2j}\|^2 \\ &\leq (1+\epsilon_2') C_2 K_2^2 (q-1) \left(\frac{c(1-\rho_0)(1-\epsilon_2)n}{8(q-1)\log^4 n} \right)^{-2\alpha_2}.\end{aligned}$$

From Corollary 6, we now obtain:

THEOREM 7. *Let Assumption 2, 8, and 9 be satisfied. Suppose that the right hand side of* (4.8) *is greater than or equal to* 2 *and that* $\dim W_1 \leq d_1$. *Then*

$$\sup_{f_1 \in \tilde{W}_1(\alpha_1, K_1)} \sup_{f_{2j} \in \tilde{W}_{2j}(\alpha_2, K_2)} \mathbb{E}\left[\|f_1 - \hat{f}_1^*\|^2 \right] \leq C\sigma^{\frac{4\alpha_1}{2\alpha_1+1}} K_1^{\frac{2}{2\alpha_1+1}} n^{-\frac{2\alpha_1}{2\alpha_1+1}} + C'(1+\epsilon_2') q^{2\alpha_2+1} (\log n)^{8\alpha_2} ((1-\epsilon_2)n)^{-2\alpha_2}, \quad (4.10)$$

where C *is a constant depending only on* α_1, c *and* ρ_0, *and* C' *is a constant depending only on* α_2, c, ρ_0, K_2, *and* K_1.

REMARK 5. If the last expression in (4.10) is of smaller order, then Theorem 7 says that the estimator $\hat{f}_1^*$ attains the (nonasymptotic) optimal rate of convergence in a minimax sense. The last expression in (4.10) is of smaller order if, e.g.,

$$q^{2\alpha_2+1} (\log n)^{8\alpha_2+1} n^{-2\alpha_2} \leq C'' n^{-\frac{2\alpha_1}{2\alpha_1+1}}, \quad (4.11)$$

where C'' depends only on α_2, c, ρ_0, K_2, K_1, ϵ_2, and ϵ_2'. This result can be applied to an asymptotic scenario in which q and n tend to infinity such that (4.11) is satisfied.

Next, we apply Theorem 7 in the case that the sample size n tends to infinity, and q is a fixed constant.

COROLLARY 9. *Let Assumption 2, 8, and 9 be satisfied. Suppose that*

$$\alpha_2 > \alpha_1/(2\alpha_1+1).$$

Then

$$\limsup_{n\to\infty} \sup_{f_1 \in \tilde{W}_1(\alpha_1, K_1)} \sup_{f_{2j} \in \tilde{W}_{2j}(\alpha_2, K_2)} \mathbb{E}\left[n^{\frac{2\alpha_1}{2\alpha_1+1}} \|f_1 - \hat{f}_1^*\|^2 \right] \leq C\sigma^{\frac{4\alpha_1}{2\alpha_1+1}} K_1^{\frac{2}{2\alpha_1+1}},$$

where C *is a constant depending only on* α_1, c, *and* ρ_0. *If, in addition, Assumption 5 holds, then the dependence of* C *on* ρ_0 *disappears and we obtain the same constant as for the corresponding estimator with* $V_2 = 0$ *in the model* $Y = f_1(X_1) + \epsilon$.

REMARK 6. Again, the assumptions of Corollary 9 are weaker than those needed in [**36**] (compare to Remark 3). For instance, the condition that the one- and two-dimensional marginal densities of X_{2j} and $(X_{2j}, X_{2j'})$ are bounded away from zero and infinity implies that Assumption 9 is satisfied (see, e.g., [**8**, Proposition 2.C in Appendix A.4] for the case $q = 3$ and [**51**, Lemma 1] for the general case).

4.4. The additive model with Hölder smoothness. We continue the discussion of the additive model in Section 4.3, and consider briefly the case of Hölder smoothness and spaces of piecewise polynomials.

ASSUMPTION 10. *Suppose that X_1 and the X_{2j} take values in $[0,1]$ and have densities p_{X_1} and $p_{X_{2j}}$ with respect to the Lebesgue measure on $[0,1]$, which satisfy $p_{X_1} \geq c$ and $p_{X_{2j}} \geq c$ for some constant $c > 0$.*

*Moreover, suppose that the function f_1 is contained in the Hölder class $\mathcal{H}_1(\alpha_1, K_1)$ on $[0,1]$, where $\alpha_1 > 0$ and $K_1 \geq 0$ (see, e.g., [**74**, Definition 1.2]), and that there is a decomposition $f_2 = \sum_{j=1}^{q-1} f_{2j}$ such that $f_{2j} \in \mathcal{H}_{2j}(\alpha_2, K_2) \cap H_{2j}$, where $\alpha_2 > 0$ and $K_2 \geq 0$.*

Let V_1 be the intersection of H_1 with the space of regular piecewise polynomials (in the variable x_1) with integer-valued parameters $r_1 = \lfloor \alpha_1 \rfloor$ and m_1, where r_1 is the maximal degree of the polynomials and $\{0 < 1/m_1 < 2/m_1 < \cdots < 1\}$ generates the partition of $[0,1]$ into m_1 intervals (see, e.g., [**10**]), and let W_1 be the intersection of H_1 with the space of regular piecewise polynomials (in the variable x_1) with integer-valued parameters $r_1 = \lfloor \alpha_1 \rfloor$ and m_{W_1} (in order that $W_1 \subseteq V_1$ we need that m_1 is a multiple of m_{W_1}). Moreover, let $V_2 = \sum_{j=1}^{q-1} V_{2j}$, where V_{2j} is the intersection of H_{2j} with the space of regular piecewise polynomials (in the variable x_{2j}) with integer-valued parameters $r_2 = \lfloor \alpha_2 \rfloor$ and m_2. Note that alternatively, one could also consider spaces of splines with the same parameters (see, e.g., [**21**, Chapter VII, VIII]). We have $d_1 = (r_1 + 1)m_1 - 1$, $d_{2j} = (r_2 + 1)m_2 - 1$, if $j = 1, \ldots, q-1$, and $d_{2j} = (r_2 + 1)m_2$, if $j = q - 1$. Using Taylor's theorem, one can show that there are constants C_1 and C_2 depending only on α_1 and α_2, respectively, such that

$$\inf_{g_1 \in V_1} \|h_1 - g_1\|_\infty^2 \leq C_1 K_1^2 d_1^{-2\alpha_1} \tag{4.12}$$

for all $h_1 \in \mathcal{H}(\alpha_1, K_1) \cap H_1$ and

$$\inf_{g_{2j} \in V_{2j}} \|h_{2j} - g_{2j}\|_\infty^2 \leq C_2 K_2^2 d_{2j}^{-2\alpha_2} \tag{4.13}$$

for all $h_{2j} \in \mathcal{H}(\alpha_2, K_2) \cap H_{2j}$. Note that similar bounds also hold for spline spaces (see, e.g., [**21**, Chapter XII]). Concerning Assumption 3, we have a similar result as in the previous section:

LEMMA 5. *Let Assumptions 2 and 9, and 10 be satisfied. Let $r = r_1 \vee r_2$. Then we have*

$$\varphi^2 \leq \frac{2(r+1)}{c(1-\rho_0)(1-\epsilon_2)} \frac{d_1 + \sum_{j=1}^{q-1} d_{2j}}{d}.$$

In particular, if the V_{2j} are linearly independent, then we have

$$\varphi^2 \leq \frac{2(r+1)}{c(1-\rho_0)(1-\epsilon_2)}.$$

A proof of Lemma 5 is given in Appendix B. Choosing d_1, $\dim W_1$, and d_{2j}, $j = 1, \ldots, q-1$, similarly as in the previous section, we obtain the following analogue of Theorem 7:

THEOREM 8. *Let Assumption 2, 9, and 10 be satisfied. Suppose that the right hand side of* (4.8) *is greater than or equal to $r+1$ and that* $\dim W_1 \leq c(1-\rho_0)(1-\epsilon_2)n/(4(r+1)\log^4 n)$. *Then*

$$\begin{aligned}\mathbb{E}\left[\|f_1 - \hat{f}_1^*\|^2\right] &\leq C\sigma^{\frac{4\alpha_1}{2\alpha_1+1}} K_1^{\frac{2}{2\alpha_1+1}} n^{-\frac{2\alpha_1}{2\alpha_1+1}} \\ &\quad + C'(1+\epsilon_2')\varphi^{4\alpha_2} q^{2\alpha_2+1} (\log n)^{8\alpha_2+1} ((1-\epsilon_2)n)^{-2\alpha_2}\end{aligned}$$

where C is a constant depending only on α_1 and ρ_0, and C' is a constant depending only on α_2, c, r, ρ_0, K_2, and $\|f_1\|$.

4.5. The additive model with smooth design densities. We continue the example in Section 4.4 and discuss Condition (3.6) and (3.7) in Theorem 5 and 6, respectively. First, we apply Theorem 5 in the simple case $q = 2$. We suppose that Assumption 4 holds and that for each fixed x_1,

$$\frac{p_X(x_1, x_2)}{p_{X_1}(x_1) p_{X_2}(x_2)} \in \mathcal{H}(\beta, h_1(x_1)) \tag{4.14}$$

with $h_1 \in L^2(\mathbb{P}^{X_1})$. Let V_1 and W_1 as in the previous section, and let V_2 be the space of regular piecewise polynomials in the variable x_2 with parameters $r_2 = \lfloor \alpha_2 \rfloor \vee \lfloor \beta \rfloor$ and d_2 as in (4.3). Then (4.13) holds with a constant C_2 depending only on α_2 and r_2. Moreover, (3.6) is satisfied with h_1 from (4.14) and $\psi(V_2) = \sqrt{C_3} d_2^{-\beta}$, where C_3 is a constant depending only on β and r_2. Thus

$$\|h_1\| \psi(V_2) \phi(V_2) \leq \sqrt{C_2 C_3} K_2 \|h_1\| \left(\frac{n}{4\varphi^2 \log^4 n}\right)^{-\alpha_2 - \beta}$$

From Theorem 5, we obtain:

COROLLARY 10. *Let $q = 2$. Let Assumption 2, 10, and 4 be satisfied. Suppose that* (4.14) *holds, and that*

$$\alpha_2 + \beta > \alpha_1/(2\alpha_1 + 1). \tag{4.15}$$

Then

$$\limsup_{n\to\infty} \mathbb{E}\left[n^{\frac{2\alpha_1}{2\alpha_1+1}} \|f_1 - \hat{f}_1^*\|^2\right] \leq C\sigma^{\frac{4\alpha_1}{2\alpha_1+1}} K_1^{\frac{2}{2\alpha_1+1}}, \tag{4.16}$$

where C is a constant depending only on α_1.

Finally, we apply Theorem 4. In the particular case that the X_{2j}, $j = 1, \ldots, q-1$, are independent, one can show that (see (0.15))

$$r(x_1, x_2) = \sum_{j=1}^{q-1} \frac{p_{X_1, X_{2j}}(x_1, x_{2j})}{p_{X_1}(x_1) p_{X_{2j}}(x_{2j})} - (q-2),$$

where $p_{X_1, X_{2j}}$ denotes the joint density of (X_1, X_{2j}). In this particular case, condition (3.7) is thus much weaker than condition (3.6): we only need smoothness conditions on several kernels which involve only one- and two-dimensional design densities. In the general case, we have a similar result. Suppose that for each fixed x_1

$$\frac{p_{X_1, X_{2k}}(x_1, x_{2k})}{p_{X_1}(x_1) p_{X_{2k}}(x_{2k})} \in \mathcal{H}(\beta, h_{1k}(x_1)) \tag{4.17}$$

with $h_{1k} \in L^2(\mathbb{P}^{X_1})$, for all $k = 1, \ldots, q-1$. Moreover, suppose that for each fixed x_{2j}

$$\frac{p_{X_{2j}, X_{2k}}(x_{2j}, x_{2k})}{p_{X_{2j}}(x_{2j}) p_{X_{2k}}(x_{2k})} \in \mathcal{H}(\beta, h'_{jk}(x_{2j})) \tag{4.18}$$

with $h'_{jk} \in L^2(\mathbb{P}^{X_{2j}})$, for all $j, k = 1, \ldots, q-1$. Then we have:

COROLLARY 11. *Let $q > 2$. Let Assumption 2, 9, 10, and 4 be satisfied. Suppose that* (4.17) *and* (4.18) *are satisfied. Moreover, suppose that $\alpha_2 + \beta/(2\alpha_1 + 1) > \alpha_1/(2\alpha_1 + 1)$. Then* (4.16) *holds, where C is a constant depending only on α_1.*

A proof of Corollary 11 is given in Appendix B. Note that in Corollary 11, the smoothness condition is stronger than the smoothness condition given in (4.15).

5. Proof of Theorem 3 and 4

5.1. The finite sample geometry. In this section, we present an empirical version of Assumption 2, which holds on the event

$$\mathcal{E}_\delta = \left\{ (1-\delta)\|g\|^2 \leq \|g\|_n^2 \leq (1+\delta)\|g\|^2 \ \text{ for all } g \in V \right\},$$

where $0 < \delta < 1$ is the constant from Theorem 3. Moreover, using concentration of measure inequalities for structured random matrices, we lower bound the probability that the event $\mathcal{E}_\delta$ occurs.

In order to state our first result, we introduce the empirical inner product $\langle \cdot, \cdot \rangle_n$ and the corresponding empirical norm $\| \cdot \|_n$ which are given by

$$\langle g, h \rangle_n = \frac{1}{n} \sum_{i=1}^{n} g(X^i) h(X^i)$$

and $\|g\|_n^2 = \langle g, g \rangle_n$ for $g, h \in L^2(\mathbb{P}^X)$.

PROPOSITION 5. *Let Assumptions 1 and 2 hold. If $\mathcal{E}_\delta$ holds, then we have*

$$\|g_1+g_2\|_n^2 \geq \frac{(1-\delta)}{(1+\delta)}(1-\rho_0)(\|g_1\|_n^2+\|g_2\|_n^2) \quad \textit{and} \tag{5.1}$$

$$\|g_1+g_2\|_n^2 \geq \frac{(1-\delta)}{(1+\delta)}(1-\rho_0^2)\|g_1\|_n^2 \tag{5.2}$$

for all $g_1 \in V_1$, $g_2 \in V_2$, and also

$$\frac{|\langle g_1, g_2\rangle_n|}{\|g_1\|_n\|g_2\|_n} \leq 1-\frac{(1-\delta)}{(1+\delta)}(1-\rho_0) \tag{5.3}$$

for all $0 \neq g_1 \in V_1$, $0 \neq g_2 \in V_2$.

PROOF. Let $g_1 \in V_1$ and $g_2 \in V_2$. By the definition of $\mathcal{E}_\delta$ and Lemma 2 combined with Assumption 2, we have

$$\begin{aligned}\|g_1+g_2\|_n^2 \geq (1-\delta)\|g_1+g_2\|^2 &\geq (1-\delta)(1-\rho_0)(\|g_1\|^2+\|g_2\|^2)\\ &\geq \frac{(1-\delta)}{(1+\delta)}(1-\rho_0)(\|g_1\|_n^2+\|g_2\|_n^2)\end{aligned}$$

and similarly

$$\begin{aligned}\|g_1+g_2\|_n^2 \geq (1-\delta)\|g_1+g_2\|^2 &\geq (1-\delta)(1-\rho_0^2)\|g_1\|^2\\ &\geq \frac{(1-\delta)}{(1+\delta)}(1-\rho_0^2)\|g_1\|_n^2.\end{aligned}$$

This gives (5.1) and (5.2). (5.3) follows from (5.1) and Lemma 2. This completes the proof. □

The following result follows from Rauhut [**59**, Theorem 7.3] (see also [**63**, Theorem 3.1]). It can also be obtained from a combination of Talagrand's inequality and Rudelson's lemma (see [**62**, Theorem 1]). The details are also given in Appendix B.

THEOREM 9. *Let Assumption 3 hold. Then we have*

$$\mathbb{P}\left(\mathcal{E}_\delta\right) \geq 1-2^{3/4}d\exp\left(-\kappa\frac{n\delta^2}{\varphi^2 d}\right),$$

where κ is a universal constant.

PROOF. Let $b_1,\ldots,b_d$ be an orthonormal basis of V. By Assumption 3 and [**11**, Lemma 1], we have

$$\Big\|\sum_{j=1}^d b_j^2\Big\|_\infty \leq \varphi^2 d. \tag{5.4}$$

Now let

$$B_n = (\langle b_j, b_k\rangle_n)_{1\leq j,k\leq d}.$$

Then [**59**, Theorem 7.3] (in the case $s = N = d$) yields for $0 < \delta < 1$,

$$\mathbb{P}(\|B_n - I\|_{\mathrm{op}} \le \delta) \ge 1 - 2^{3/4} d \exp\left(-\kappa \frac{n\delta^2}{\varphi^2 d}\right), \tag{5.5}$$

where $\kappa > 0$ is a universal constant. Here, $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm (see Lemma 7 (iii) for the definition). Note that we can apply [**59**, Theorem 7.3] since in the proof the condition [**59**, (4.2)] is only used in the form [**59**, (7.5)], which is satisfied by (5.4). A similar result follows from [**63**, Theorem 3.1].

Now, a function $g \in V$ with $\|g\| \le 1$ can be written uniquely as $g = \sum_{j=1}^d x_j b_j$ with $x \in \mathbb{R}^d$ and $\|x\|_2 \le 1$. Using this and $\|g\|_n^2 = x^T B_n x$, we obtain

$$\sup_{g \in V, \|g\| \le 1} |\|g\|_n^2 - \|g\|^2| = \sup_{x \in \mathbb{R}^d, \|x\|_2 \le 1} |x^T (B_n - I) x| = \|B_n - I\|_{\mathrm{op}}, \tag{5.6}$$

where the latter equality follows from the spectral theorem. Moreover, we have that $\mathcal{E}_\delta$ holds if and only if

$$\sup_{g \in V, \|g\| \le 1} |\|g\|_n^2 - \|g\|^2| \le \delta.$$

Applying this, (5.5), and (5.6), we complete the proof. □

5.2. Analysis of the variance via von Neumann's theorem. The basic theorem in the theory of projections on sumspaces is due to von Neumann [**78**]. We state the following version dealing only with the first component, which is a consequence of [**3**, (15) on page 378] (see also [**8**, Theorem 2.C in Appendix A.4]).

LEMMA 6. *Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two closed subspaces of a Hilbert space $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$. Suppose that $\rho_0(\mathcal{H}_1, \mathcal{H}_2) < 1$. Let Π, Π_1, Π_2 be the orthogonal projections on $\mathcal{H}_1 + \mathcal{H}_2$, $\mathcal{H}_1$, $\mathcal{H}_2$, respectively. Let $h \in \mathcal{H}$, and let $(\Pi h)_1 \in \mathcal{H}_1$, $(\Pi h)_2 \in \mathcal{H}_2$ be the unique elements such that $\Pi h = (\Pi h)_1 + (\Pi h)_2$. Then*

$$\Big\| (\Pi h)_1 - \big(\Pi_1 - \sum_{j=1}^k (\Pi_1 \Pi_2)^j (1 - \Pi_1)\big) h \Big\| \to 0$$

as k goes to infinity.

REMARK 7. If we set $h_1^{(1)} = \Pi_1 h$ and proceed iteratively by setting $h_2^{(m)} = \Pi_2(h - h_1^{(m)})$ and $h_1^{(m+1)} = \Pi_1(h - h_2^{(m)})$, $m \ge 1$, then Lemma 6 can be rewritten as $\|(\Pi h)_1 - h_1^{(m)}\| \to 0$. This procedure is often called "backfitting".

For completeness, a proof of Lemma 6 is given in Appendix B. In this section, we apply Lemma 6 to the finite sample setting, using results from the previous section. Recall that $\hat{\Pi}_V$ is the orthogonal projection from $\mathbb{R}^n$ to the subspace $\{(g(X^1), \ldots, g(X^n))^T | g \in V\}$, and that $\hat{\Pi}_{V_1}$, $\hat{\Pi}_{V_2}$, and $\hat{\Pi}_{W_1}$

are defined analogously (replace V by V_1, V_2, and W_1, respectively). We first prove:

PROPOSITION 6. *Let Assumption 1 and 2 be satisfied. Let*

$$\rho_{0,\delta} = 1 - \frac{(1-\delta)}{(1+\delta)}(1-\rho_0).$$

If $\mathcal{E}_\delta$ holds, then we have

$$\mathbb{E}\left[\|\hat{\Pi}_{W_1}(\hat{\Pi}_V \boldsymbol{\epsilon})_1\|_n^2 | X^1, \ldots, X^n\right] \leq \frac{\sigma^2 \dim W_1}{n} + \frac{1}{1-\rho_{0,\delta}^2}\frac{\sigma^2 \operatorname{tr}(\hat{\Pi}_{W_1}\hat{\Pi}_{V_2})}{n}.$$

In the proof, we will need the following result:

LEMMA 7. *Let $A \in \mathbb{R}^{k_1 \times k_2}$ and $B \in \mathbb{R}^{k_2 \times k_1}$. Then*

(i) $\operatorname{tr}(AB) = \operatorname{tr}(BA)$.
(ii) $|\operatorname{tr}(AB)| \leq \sqrt{\operatorname{tr}(AA^T)\operatorname{tr}(BB^T)}$.
(iii) *Let $k_1 = k_2$ and B be symmetric and positive semi-definite. Then*

$$|\operatorname{tr}(AB)| \leq \|A\|_{\mathrm{op}} \operatorname{tr}(B),$$

where $\|A\|_{\mathrm{op}} = \sup_{\|x\|_2=1} \|Ax\|_2$ denotes the operator norm. Here, $\|\cdot\|_2$ denotes the Euclidean norm.

For completeness, a proof of Lemma 7 is given in Appendix B.

PROOF OF PROPOSITION 6. Throughout the proof, we suppose that $\mathcal{E}_\delta$ holds. Furthermore, we consider V as a subset of $\mathbb{R}^n$. This is no restriction, since (on $\mathcal{E}_\delta$) each element $g \in V$ is uniquely determined by $(g(X^1), \ldots, g(X^n))^T$. From (5.3) and Lemma 6 applied to $(\mathcal{H}, \langle\cdot,\cdot\rangle) = (V, \langle\cdot,\cdot\rangle_n)$ and $\mathcal{H}_j = V_j$, $j = 1, 2$, we have

$$\Big\|(\hat{\Pi}_V \boldsymbol{\epsilon})_1 - \big(\hat{\Pi}_{V_1} - \sum_{j=1}^{k}(\hat{\Pi}_{V_1}\hat{\Pi}_{V_2})^j(1-\hat{\Pi}_{V_1})\big)\boldsymbol{\epsilon}\Big\|_n \to 0$$

as k goes to infinity. From (5.3), we have

$$\|\hat{\Pi}_{V_1} g_2\|_n \leq \rho_{0,\delta}\|g_2\|_n \tag{5.7}$$

for all $g_2 \in V_2$, which follows from

$$\|\hat{\Pi}_{V_1} g_2\|_n^2 = \langle \hat{\Pi}_{V_1} g_2, \hat{\Pi}_{V_1} g_2\rangle_n = \langle \hat{\Pi}_{V_1} g_2, g_2\rangle_n \leq \rho_{0,\delta}\|\hat{\Pi}_{V_1} g_2\|_n \|g_2\|_n.$$

Similarly, we have

$$\|\hat{\Pi}_{V_2} g_1\|_n \leq \rho_{0,\delta}\|g_1\|_n \tag{5.8}$$

for all $g_1 \in V_1$. This gives the improved convergence result

$$\begin{aligned}
&\big\|(\hat{\Pi}_V \boldsymbol{\epsilon})_1 - \big(\hat{\Pi}_{V_1} - \sum_{j=1}^{k} (\hat{\Pi}_{V_1}\hat{\Pi}_{V_2})^j (1-\hat{\Pi}_{V_1})\big)\boldsymbol{\epsilon}\big\|_n \\
&\leq \sum_{j=k+1}^{\infty} \|(\hat{\Pi}_{V_1}\hat{\Pi}_{V_2})^j (1-\hat{\Pi}_{V_1})\boldsymbol{\epsilon}\|_n \\
&\leq \sum_{j=k+1}^{\infty} \rho_{0,\delta}^{2j-1} \|\boldsymbol{\epsilon}\|_n \\
&= \frac{\rho_{0,\delta}^{2k+1}}{1-\rho_{0,\delta}^2} \|\boldsymbol{\epsilon}\|_n,
\end{aligned}$$

and also

$$\begin{aligned}
&\big\|\hat{\Pi}_{W_1}(\hat{\Pi}_V \boldsymbol{\epsilon})_1 - \hat{\Pi}_{W_1}\big(\hat{\Pi}_{V_1} - \sum_{j=1}^{k} (\hat{\Pi}_{V_1}\hat{\Pi}_{V_2})^j (1-\hat{\Pi}_{V_1})\big)\boldsymbol{\epsilon}\big\|_n \\
&\leq \frac{\rho_{0,\delta}^{2k+1}}{1-\rho_{0,\delta}^2} \|\boldsymbol{\epsilon}\|_n. \qquad (5.9)
\end{aligned}$$

Applying (5.9) and the bound $(x+y)^2 \leq (1+\epsilon)x^2 + (1+1/\epsilon)y^2$, $\epsilon > 0$, and then taking expectation, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\|\hat{\Pi}_{W_1}(\hat{\Pi}_V \boldsymbol{\epsilon})_1\|_n^2 \mid X^1, \ldots, X^n\right] \\
&\leq (1+\epsilon)\mathbb{E}\left[\big\|\hat{\Pi}_{W_1}\big(\hat{\Pi}_{V_1} - \sum_{j=1}^{k} (\hat{\Pi}_{V_1}\hat{\Pi}_{V_2})^j (1-\hat{\Pi}_{V_1})\big)\boldsymbol{\epsilon}\big\|_n^2 \mid X^1, \ldots, X^n\right] \\
&+ (1+1/\epsilon)\,\frac{\rho_{0,\delta}^{4k+2}}{(1-\rho_{0,\delta}^2)^2}\sigma^2. \qquad (5.10)
\end{aligned}$$

Since $\mathbb{E}\left[\|A\boldsymbol{\epsilon}\|_n^2\right] = \sigma^2 \operatorname{tr}(AA^T)/n$ for all $A \in \mathbb{R}^{n\times n}$, it remains to bound the trace of

$$\hat{\Pi}_{W_1}\big(\hat{\Pi}_{V_1} - \sum_{j=1}^{k} (\hat{\Pi}_{V_1}\hat{\Pi}_{V_2})^j (1-\hat{\Pi}_{V_1})\big)\Big(\hat{\Pi}_{V_1} - \sum_{j=1}^{k} (\hat{\Pi}_{V_1}\hat{\Pi}_{V_2})^j (1-\hat{\Pi}_{V_1})\Big)^T \hat{\Pi}_{W_1}. \qquad (5.11)$$

Using

$$\begin{aligned}
&\hat{\Pi}_{V_1} - \sum_{j=1}^{k} (\hat{\Pi}_{V_1}\hat{\Pi}_{V_2})^j (1-\hat{\Pi}_{V_1}) \\
&= \sum_{j=0}^{k-1} \hat{\Pi}_{V_1}(\hat{\Pi}_{V_2}\hat{\Pi}_{V_1})^j (1-\hat{\Pi}_{V_2}) + (\hat{\Pi}_{V_1}\hat{\Pi}_{V_2})^k \hat{\Pi}_{V_1},
\end{aligned}$$

(5.11) is equal to

$$\hat{\Pi}_{W_1}\sum_{j=0}^{k}(\hat{\Pi}_{V_1}\hat{\Pi}_{V_2})^j\hat{\Pi}_{V_1}\Big(\hat{\Pi}_{V_1}-(1-\hat{\Pi}_{V_1})\sum_{j=1}^{k}(\hat{\Pi}_{V_2}\hat{\Pi}_{V_1})^j\Big)\hat{\Pi}_{W_1}$$
$$-\hat{\Pi}_{W_1}\sum_{j=1}^{k}(\hat{\Pi}_{V_1}\hat{\Pi}_{V_2})^j\Big((1-\hat{\Pi}_{V_2})\sum_{j=0}^{k-1}(\hat{\Pi}_{V_1}\hat{\Pi}_{V_2})^j\hat{\Pi}_{V_1}+(\hat{\Pi}_{V_1}\hat{\Pi}_{V_2})^k\hat{\Pi}_{V_1}\Big)\hat{\Pi}_{W_1}$$

and, since $\hat{\Pi}_{V_1}(1-\hat{\Pi}_{V_1})=0$ and $\hat{\Pi}_{V_2}(1-\hat{\Pi}_{V_2})=0$, this is equal to

$$\hat{\Pi}_{W_1}\sum_{j=0}^{k}(\hat{\Pi}_{V_1}\hat{\Pi}_{V_2})^j\hat{\Pi}_{W_1}-\hat{\Pi}_{W_1}\sum_{j=1}^{k}(\hat{\Pi}_{V_1}\hat{\Pi}_{V_2})^{j+k}\hat{\Pi}_{W_1}.$$

By Lemma 7 (i) and the identities $\hat{\Pi}_{W_1}\hat{\Pi}_{V_1}=\hat{\Pi}_{W_1}$ and $\hat{\Pi}_{V_2}\hat{\Pi}_{V_2}=\hat{\Pi}_{V_2}$, we have for $j=1,\dots,2k$,

$$\begin{aligned}\operatorname{tr}(\hat{\Pi}_{W_1}(\hat{\Pi}_{V_1}\hat{\Pi}_{V_2})^j\hat{\Pi}_{W_1}) &= \operatorname{tr}((\hat{\Pi}_{V_2}\hat{\Pi}_{V_1})^{j-1}\hat{\Pi}_{V_2}\hat{\Pi}_{W_1}\hat{\Pi}_{V_2})\\ &= \operatorname{tr}((\hat{\Pi}_{V_2}\hat{\Pi}_{V_1}\hat{\Pi}_{V_2})^{j-1}\hat{\Pi}_{V_2}\hat{\Pi}_{W_1}\hat{\Pi}_{V_2}).\end{aligned}$$

Thus the trace of (5.11) is bounded by

$$\dim W_1+\sum_{j=1}^{2k}|\operatorname{tr}((\hat{\Pi}_{V_2}\hat{\Pi}_{V_1}\hat{\Pi}_{V_2})^{j-1}\hat{\Pi}_{V_2}\hat{\Pi}_{W_1}\hat{\Pi}_{V_2})|. \tag{5.12}$$

Applying Lemma 7 (iii), this can be bounded by

$$\dim W_1+\sum_{j=0}^{2k-1}\|\hat{\Pi}_{V_2}\hat{\Pi}_{V_1}\hat{\Pi}_{V_2}\|_{\mathrm{op}}^j\operatorname{tr}(\hat{\Pi}_{V_2}\hat{\Pi}_{W_1}\hat{\Pi}_{V_2}).$$

By (5.7) and (5.8), we have

$$\|\hat{\Pi}_{V_2}\hat{\Pi}_{V_1}\hat{\Pi}_{V_2}\|_{\mathrm{op}}\le\rho_{0,\delta}^2. \tag{5.13}$$

Moreover, we have $\operatorname{tr}(\hat{\Pi}_{V_2}\hat{\Pi}_{W_1}\hat{\Pi}_{V_2})=\operatorname{tr}(\hat{\Pi}_{W_1}\hat{\Pi}_{V_2})$. Thus we obtain that (5.12) is bounded by

$$\dim W_1+\frac{1}{1-\rho_{0,\delta}^2}\operatorname{tr}(\hat{\Pi}_{W_1}\hat{\Pi}_{V_2}). \tag{5.14}$$

From (5.10)-(5.14), we conclude that

$$\begin{aligned}&\mathbb{E}\left[\|\hat{\Pi}_{W_1}(\hat{\Pi}_V\boldsymbol{\epsilon})_1\|_n^2|X^1,\dots,X^n\right]\\ &\le(1+\epsilon)\left(\frac{\sigma^2\dim W_1}{n}+\frac{1}{1-\rho_{0,\delta}^2}\frac{\sigma^2\operatorname{tr}(\hat{\Pi}_{W_1}\hat{\Pi}_{V_2})}{n}\right)\\ &+(1+1/\epsilon)\frac{\rho_{0,\delta}^{4k+2}}{(1-\rho_{0,\delta}^2)^2}\sigma^2.\end{aligned}$$

Now, send k off to infinity first, and then let ϵ go to zero. This completes the proof. □

From Proposition 6, we obtain a first upper bound for the variance term, which does not depend on the dimension of V_2. Note that in Proposition 7 and Corollary 13, we show that this upper bound can be further refined.

COROLLARY 12. *Let Assumption 1 and 2 be satisfied. If $\mathcal{E}_\delta$ holds, then we have*

$$\mathbb{E}\left[\|\hat{\Pi}_{W_1}(\hat{\Pi}_V \boldsymbol{\epsilon})_1\|_n^2 | X^1, \dots, X^n\right] \leq \frac{1+\delta}{1-\delta} \frac{1}{1-\rho_0^2} \frac{\sigma^2 \dim W_1}{n}.$$

PROOF. By Lemma 7 (i) and (ii), we have

$$\operatorname{tr}(\hat{\Pi}_{W_1}\hat{\Pi}_{V_2}) = \operatorname{tr}(\hat{\Pi}_{W_1}\hat{\Pi}_{V_2}\hat{\Pi}_{W_1}\hat{\Pi}_{W_1}).$$

Applying Lemma 7 (iii) and (5.13), we obtain on $\mathcal{E}_\delta$,

$$\operatorname{tr}(\hat{\Pi}_{W_1}\hat{\Pi}_{V_2}) \leq \|\hat{\Pi}_{W_1}\hat{\Pi}_{V_2}\hat{\Pi}_{W_1}\|_{\mathrm{op}} \operatorname{tr}(\hat{\Pi}_{W_1}) \leq \rho_{0,\delta}^2 \dim W_1.$$

Thus, if $\mathcal{E}_\delta$ holds, Proposition 6 yields

$$\mathbb{E}\left[\|\hat{\Pi}_{W_1}(\hat{\Pi}_V \boldsymbol{\epsilon})_1\|_n^2 | X^1, \dots, X^n\right] \leq \frac{1}{1-\rho_{0,\delta}^2} \frac{\sigma^2 \dim W_1}{n}.$$

Since $(1-c(1-\varrho))^2 \leq 1-c(1-\varrho^2)$ for $\varrho \in [0,1]$ and a constant $0 \leq c \leq 1$ (both functions are equal to 1 at the right endpoint $\varrho = 1$ and the derivative of the left hand side is greater or equal than the derivative of the right hand side for all $\varrho \in [0,1]$), we obtain (set $c = (1-\delta)/(1+\delta)$ and $\varrho = \rho_0$)

$$\frac{1}{1-\rho_{0,\delta}^2} \leq \frac{1+\delta}{1-\delta} \frac{1}{1-\rho_0^2}. \tag{5.15}$$

This completes the proof. □

REMARK 8. An alternative proof of Corollary 12 is given in Appendix B.

PROPOSITION 7. *Let Assumption 1, 2, and 3 be satisfied. Then*

$$\frac{1}{n}\mathbb{E}\left[1_{\mathcal{E}_\delta} \operatorname{tr}(\hat{\Pi}_{W_1}\hat{\Pi}_{V_2})\right] \leq \frac{1}{(1-\delta)^2}\left(\frac{\|\Pi_{V_2}|_{W_1}\|_{HS}^2}{n} + \frac{\dim W_1}{n}\frac{\varphi^2 d}{n}\right).$$

PROOF. Let $\{\phi_{1j}\}_{1 \leq j \leq \dim W_1}$ be an orthonormal basis of W_1, and let $\{\phi_{2j}\}_{1 \leq j \leq d_2}$ be an orthonormal basis of V_2. Let

$$Z_1 = (\phi_{1j}(X_1^i))_{1 \leq i \leq n, 1 \leq j \leq \dim W_1}$$

and

$$Z_2 = (\phi_{2j}(X_2^i))_{1 \leq i \leq n, 1 \leq j \leq d_2},$$

Now, suppose that $\mathcal{E}_\delta$ holds. Then we have $\hat{\Pi}_{W_1} = Z_1(Z_1^TZ_1)^{-1}Z_1^T$ and $\hat{\Pi}_{V_2} = Z_2(Z_2^TZ_2)^{-1}Z_2^T$. Thus

$$\operatorname{tr}(\hat{\Pi}_{W_1}\hat{\Pi}_{V_2}) = \operatorname{tr}\left(\left(\frac{1}{n}Z_1^TZ_1\right)^{-1}\frac{1}{n}Z_1^TZ_2\left(\frac{1}{n}Z_2^TZ_2\right)^{-1}\frac{1}{n}Z_2^TZ_1\right),$$

where we applied Lemma 7 (i). By Theorem 9, we have $\|(1/n)Z_j^TZ_j - I\|_{\mathrm{op}} \leq \delta$ and thus

$$\left(\frac{1}{n}Z_j^TZ_j\right)^{-1} = \sum_{k\geq 0}\left(I - \frac{1}{n}Z_j^TZ_j\right)^k$$

for $j = 1, 2$. We conclude that

$$\begin{aligned}
&\mathbb{E}\left[1_{\mathcal{E}_\delta}\operatorname{tr}(\hat{\Pi}_{W_1}\hat{\Pi}_{V_2})\right] \\
&\leq \sum_{k,l=0}^{\infty}\mathbb{E}\left[1_{\mathcal{E}_\delta}\left|\operatorname{tr}\left(\left(I - \frac{1}{n}Z_1^TZ_1\right)^k\frac{1}{n}Z_1^TZ_2\left(I - \frac{1}{n}Z_2^TZ_2\right)^l\frac{1}{n}Z_2^TZ_1\right)\right|\right] \\
&\leq \sum_{k,l=0}^{\infty}\mathbb{E}\left[1_{\mathcal{E}_\delta}\sqrt{\operatorname{tr}\left(\left(I - \frac{1}{n}Z_1^TZ_1\right)^{2k}\frac{1}{n}Z_1^TZ_2\left(\frac{1}{n}Z_1^TZ_2\right)^T\right)}\right. \\
&\qquad\qquad \left.\cdot\sqrt{\operatorname{tr}\left(\left(I - \frac{1}{n}Z_2^TZ_2\right)^{2k}\frac{1}{n}Z_2^TZ_1\left(\frac{1}{n}Z_2^TZ_1\right)^T\right)}\right] \\
&\leq \sum_{k,l=0}^{\infty}\delta^{k+l}\mathbb{E}\left[\operatorname{tr}\left(\frac{1}{n}Z_1^TZ_2\left(\frac{1}{n}Z_1^TZ_2\right)^T\right)\right],
\end{aligned}$$

where we applied Lemma 7 (i) and (ii) in the second inequality and Lemma 7 (i) and (iii) and the bound $\|(1/n)Z_j^TZ_j - I\|_{\mathrm{op}} \leq \delta$ in the third inequality. Now

$$\begin{aligned}
&\mathbb{E}\left[\operatorname{tr}\left(\frac{1}{n}Z_1^TZ_2\left(\frac{1}{n}Z_1^TZ_2\right)^T\right)\right] \\
&= \sum_{j=1}^{\dim W_1}\sum_{k=1}^{d_2}\left(\left(\mathbb{E}\left[\phi_{1j}(X_1)\phi_{2k}(X_2)\right]\right)^2 + \frac{1}{n}\operatorname{Var}(\phi_{1j}(X_1)\phi_{2k}(X_2))\right) \\
&\leq \sum_{j=1}^{\dim W_1}\sum_{k=1}^{d_2}\langle\phi_{1j},\phi_{2k}\rangle^2 + \dim W_1\frac{\varphi^2 d}{n},
\end{aligned}$$

where we applied (5.4). Finally, we use

$$\|\Pi_{V_2}|_{W_1}\|_{HS}^2 = \sum_{j=1}^{\dim W_1}\|\Pi_{V_2}\phi_{1j}\|^2 = \sum_{j=1}^{\dim W_1}\sum_{k=1}^{d_2}\langle\phi_{1j},\phi_{2k}\rangle^2$$

This completes the proof. □

REMARK 9. By considering $\Pi_{W_1}\Pi_{V_2}\Pi_{W_1}$ as a map from $W_1 \subseteq L^2(\mathbb{P}^{X_1})$ to itself, we have $\|\Pi_{V_2}|_{W_1}\|_{HS}^2 = \operatorname{tr}(\Pi_{W_1}\Pi_{V_2}\Pi_{W_1})$.

Combining Proposition 6 and 7, we obtain the following improvement of Corollary 12:

COROLLARY 13. *Let Assumption 1, 2, and 3 be satisfied. Then*

$$\begin{aligned}&\mathbb{E}\left[1_{\mathcal{E}_\delta}\|\hat{\Pi}_{W_1}(\hat{\Pi}_V\boldsymbol{\epsilon})_1\|_n^2\right]\\&\leq \frac{\sigma^2\dim W_1}{n} + \frac{(1+\delta)}{(1-\delta)^3}\frac{1}{1-\rho_0^2}\left(\frac{\sigma^2\,\|\Pi_{V_2}|_{W_1}\|_{HS}^2}{n} + \frac{\sigma^2\dim W_1}{n}\frac{\varphi^2 d}{n}\right).\end{aligned}$$

5.3. End of the proof of Theorem 3. Applying the arguments of [**5**], we obtain

$$E\left[\|f_1 - \hat{f}_1^*\|^2\right] \leq \mathbb{E}\left[1_{\mathcal{E}_\delta}\|f_1 - (\hat{f}_V)_1\|^2\right] + R_n. \tag{5.16}$$

The details can be found in Appendix B. By the projection theorem (see, e.g., [**60**, Theorem II.3]), we have

$$\|f_1 - (\hat{f}_V)_1\|^2 = \|f_1 - \Pi_{V_1}f_1\|^2 + \|\Pi_{V_1}f_1 - (\hat{f}_V)_1\|^2. \tag{5.17}$$

By the definition of $\mathcal{E}_\delta$ and the definition of $\hat{f}_V$, we have

$$\begin{aligned}&\mathbb{E}\left[1_{\mathcal{E}_\delta}\|\Pi_{V_1}f_1 - (\hat{f}_V)_1\|^2\right]\\&\leq \frac{1}{(1-\delta)}\mathbb{E}\left[1_{\mathcal{E}_\delta}\|\Pi_{V_1}f_1 - (\hat{f}_V)_1\|_n^2\right]\\&= \frac{1}{(1-\delta)}\mathbb{E}\left[1_{\mathcal{E}_\delta}\|\Pi_{V_1}f_1 - (\hat{\Pi}_V\mathbf{Y})_1\|_n^2\right]\\&= \frac{1}{(1-\delta)}\mathbb{E}\left[1_{\mathcal{E}_\delta}\mathbb{E}\left[\|\Pi_{V_1}f_1 - (\hat{\Pi}_V\mathbf{Y})_1\|_n^2|X^1,\dots,X^n\right]\right].\end{aligned} \tag{5.18}$$

Recall from Section 2.3 that on $\mathcal{E}_\delta$ each $g \in V$ is determined uniquely by $(g(X^1),\dots,g(X^n))^T$, which implies that on $\mathcal{E}_\delta$ we don't have to distinguish between those objects. For instance, if $\mathcal{E}_\delta$ holds, then $\hat{\Pi}_V$ and $\hat{\Pi}_{V_1}$ can also be seen as maps from $L^2(\mathbb{P}^X)$ to V and V_1, respectively (by letting $\hat{\Pi}_V h$ (resp. $\hat{\Pi}_{V_1}h$) equal to $\hat{\Pi}_V(h(X^1),\dots,h(X^n))^T$ (resp. $\hat{\Pi}_{V_1}(h(X^1),\dots,h(X^n))^T$) for $h \in L^2(\mathbb{P}^X)$). Moreover, if $\mathcal{E}_\delta$ holds, then we have $\mathbb{E}[(\hat{\Pi}_V\boldsymbol{\epsilon})_1|X^1,\dots,X^n] = 0$ and $(\hat{\Pi}_V\mathbf{Y})_1 = (\hat{\Pi}_V f)_1 + (\hat{\Pi}_V\boldsymbol{\epsilon})_1$ (the former follows for instance from (5.9)). Thus (5.18) is equal to

$$\frac{1}{(1-\delta)}\mathbb{E}\left[1_{\mathcal{E}_\delta}\left(\|\Pi_{V_1}f_1 - (\hat{\Pi}_V f)_1\|_n^2 + \mathbb{E}\left[\|(\hat{\Pi}_V\boldsymbol{\epsilon})_1\|_n^2|X_1,\dots,X_n\right]\right)\right]. \tag{5.19}$$

From (5.17)-(5.19) and the definition of $\mathcal{E}_\delta$, we obtain

$$\begin{aligned}
&\mathbb{E}\left[1_{\mathcal{E}_\delta}\|f_1-(\hat{f}_V)_1\|^2\right]\\
&\quad\le\frac{(1+\delta)}{(1-\delta)}\mathbb{E}\left[1_{\mathcal{E}_\delta}\left(\|f_1-\Pi_{V_1}f_1\|^2+\|\Pi_{V_1}f_1-(\hat{\Pi}_V f)_1\|^2\right)\right]\\
&\quad+\frac{1}{(1-\delta)}\mathbb{E}\left[1_{\mathcal{E}_\delta}\mathbb{E}\left[\|(\hat{\Pi}_V\boldsymbol{\epsilon})_1\|_n^2|X^1,\dots,X^n\right]\right]. \qquad (5.20)
\end{aligned}$$

Applying the projection theorem and Corollary 12, this is bounded by

$$\frac{(1+\delta)}{(1-\delta)}\mathbb{E}\left[1_{\mathcal{E}_\delta}\|f_1-(\hat{\Pi}_V f)_1\|^2\right]+\frac{(1+\delta)}{(1-\delta)^2}\frac{1}{1-\rho_0^2}\frac{\sigma^2\dim V_1}{n}.$$

By Assumption 2 and Lemma 2, we have

$$\|f_1-(\hat{\Pi}_V f)_1\|^2\le\frac{1}{1-\rho_0^2}\|f-\hat{\Pi}_V f\|^2.$$

The projection theorem implies that

$$\|f-\hat{\Pi}_V f\|^2=\|f-\Pi_V f\|^2+\|\Pi_V f-\hat{\Pi}_V f\|^2.$$

Now, the following proposition completes the proof.

LEMMA 8. *Let Assumption 3 be satisfied. Then*

$$\mathbb{E}\left[1_{\mathcal{E}_\delta}\|(\Pi_V-\hat{\Pi}_V)f\|^2\right]\le\frac{1}{(1-\delta)^2}\frac{\varphi^2 d}{n}\|f-\Pi_V f\|^2.$$

REMARK 10. Instead of applying Lemma 8, one can also apply the easier and weaker bound

$$\|\Pi_V f-\hat{\Pi}_V f\|^2\le\frac{1}{1-\delta}\|\Pi_V f-\hat{\Pi}_V f\|_n^2\le\frac{1}{1-\delta}\|f-\Pi_V f\|_n^2,$$

which follows from the definition of $\mathcal{E}_\delta$ and the projection theorem.

PROOF OF LEMMA 8. Throughout the proof we suppose that the event $\mathcal{E}_\delta$ holds. Let $b_1,\dots,b_d$ be an orthonormal basis of V. Let

$$B_n=(\langle b_j,b_k\rangle_n)_{1\le j,k\le d},$$

and let

$$x=(\langle b_1,f\rangle,\dots,\langle b_d,f\rangle)^T\quad\text{and}\quad x_n=(\langle b_1,f\rangle_n,\dots,\langle b_d,f\rangle_n)^T.$$

Then we have

$$\Pi_V f=\sum_{j=1}^d x_j b_j\quad\text{and}\quad\hat{\Pi}_V f=\sum_{j=1}^d(B_n^{-1}x_n)_j b_j$$

and thus

$$\|(\Pi_V-\hat{\Pi}_V)f\|^2=\|B_n^{-1}x_n-x\|_2^2. \qquad (5.21)$$

Since $\mathcal{E}_\delta$ holds, we have $\|B_n - I\|_{\text{op}} \leq \delta$ (see the proof of Theorem 9). This implies that

$$B_n^{-1} = \sum_{k\geq 0}(I - B_n)^k.$$

Thus

$$\begin{aligned} B_n^{-1}x_n - x &= \sum_{k\geq 0}(I - B_n)^k x_n - x \\ &= \sum_{k\geq 0}(I - B_n)^k(x_n - x) + \sum_{k\geq 1}(I - B_n)^k x. \end{aligned}$$

Applying the bounds $\|B_n - I\|_{\text{op}} \leq \delta$ and $(x+y)^2 \leq (1+\epsilon)x^2 + (1+1/\epsilon)y^2$, $\epsilon > 0$ arbitrary, we obtain

$$\begin{aligned} \|B_n^{-1}x_n - x\|_2^2 &\leq \frac{1}{(1-\delta)^2}(\|x_n - x\|_2 + \|(B_n - I)x\|_2)^2 \\ &\leq \frac{1}{(1-\delta)^2}((1+\epsilon)\|x_n - x\|_2^2 + (1+1/\epsilon)\|(B_n - I)x\|_2^2). \end{aligned} \tag{5.22}$$

Now we have

$$\begin{aligned} \mathbb{E}\left[\|x_n - x\|_2^2\right] &= \mathbb{E}\Big[\sum_{j=1}^d(\langle b_j, f\rangle_n - \langle b_j, f\rangle)^2\Big] \\ &= \frac{1}{n}\sum_{j=1}^d \text{Var}(b_j(X)f(X)) \leq \frac{\varphi^2 d}{n}\|f\|^2, \end{aligned} \tag{5.23}$$

where we applied (5.4). The jth coordinate of $(B_n - I)x$ is is equal to

$$\langle b_j, \sum_{k=1}^d b_k x_k\rangle_n - \langle b_j, f\rangle = \langle b_j, \Pi_V f\rangle_n - \langle b_j, \Pi_V f\rangle \tag{5.24}$$

and thus

$$\begin{aligned} \mathbb{E}\left[\|(B_n - I)x\|_2^2\right] &= \mathbb{E}\Big[\sum_{j=1}^d(\langle b_j, \Pi_V f\rangle_n - \langle b_j, \Pi_V f\rangle)^2\Big] \\ &= \frac{1}{n}\sum_{j=1}^{d_1} \text{Var}(b_j(X)\Pi_V f(X)) \leq \frac{\varphi^2 d}{n}\|\Pi_V f\|^2. \end{aligned} \tag{5.25}$$

Applying (5.21)-(5.25), we conclude that

$$\mathbb{E}\left[1_{\mathcal{E}_\delta}\|(\Pi_V - \hat{\Pi}_V)f\|^2\right] \leq \frac{1}{(1-\delta)^2}\frac{\varphi^2 d}{n}((1+\epsilon)\|f\|^2 + (1+1/\epsilon)\|\Pi_V f\|^2). \tag{5.26}$$

Finally, since Π_V and $\hat{\Pi}_V$ fix elements in V, we obtain $(\Pi_V - \hat{\Pi}_V)\Pi_V f = 0$ and thus $(\Pi_V - \hat{\Pi}_V)f = (\Pi_V - \hat{\Pi}_V)(1 - \Pi_V)f$. Combining this with (5.26), we see that

$$\mathbb{E}\left[1_{\mathcal{E}_\delta}\|(\Pi_V - \hat{\Pi}_V)f\|^2\right] \leq \frac{1}{(1-\delta)^2}\frac{\varphi^2 d}{n}(1+\epsilon)\|f - \Pi_V f\|^2.$$

Since $\epsilon > 0$ is arbitrary, this completes the proof. □

5.4. End of the proof of Theorem 4. Compared with the previous section, we modify our analysis of the bias term, which is based the following two lemmas. Moreover, we replace Corollary 13 by Corollary 12.

LEMMA 9. *Let Assumption 1 and 2 be satisfied. If $\mathcal{E}_\delta$ holds, then we have*

$$\|\hat{\Pi}_{V_1}h_1 - (\hat{\Pi}_V h_1)_1\|_n^2 \leq \frac{(1+\delta)}{(1-\delta)}\frac{1}{(1-\rho_0^2)}\|h_1 - \hat{\Pi}_{V_1}h_1\|_n^2$$

for all $h_1 \in H_1$ and

$$\|(\hat{\Pi}_V h_2)_1\|_n^2 \leq \frac{(1+\delta)}{(1-\delta)}\frac{1}{(1-\rho_0^2)}\|h_2 - \hat{\Pi}_{V_2}h_2\|_n^2$$

for all $h_2 \in H_2$.

PROOF OF LEMMA 9. By (5.2) and the fact that projections lower the norm, we have

$$\begin{aligned}\|\hat{\Pi}_{V_1}h_1 - (\hat{\Pi}_V h_1)_1\|_n^2 &= \|(\hat{\Pi}_V(h_1 - \hat{\Pi}_{V_1}h_1))_1\|_n^2 \\ &\leq \frac{(1+\delta)}{(1-\delta)}\frac{1}{(1-\rho_0^2)}\|h_1 - \hat{\Pi}_{V_1}h_1\|_n^2.\end{aligned}$$

This gives the first inequality. Since $\hat{\Pi}_V\hat{\Pi}_{V_2}h_2 = \hat{\Pi}_{V_2}h_2$ and $(\hat{\Pi}_{V_2}h_2)_1 = 0$, we have $(\hat{\Pi}_V h_2)_1 = (\hat{\Pi}_V(h_2 - \hat{\Pi}_{V_2}h_2))_1$. Using the previous arguments, we conclude that

$$\|(\hat{\Pi}_V h_2)_1\|_n^2 = \|(\hat{\Pi}_V(h_2 - \hat{\Pi}_{V_2}h_2))_1\|_n^2 \leq \frac{(1+\delta)}{(1-\delta)}\frac{1}{(1-\rho_0^2)}\|h_2 - \hat{\Pi}_{V_2}h_2\|_n^2.$$

This completes the proof. □

LEMMA 10. *Let Assumption 1 and 2 be satisfied. Then we have*

$$\begin{aligned}&\mathbb{E}\left[1_{\mathcal{E}_\delta}\|f_1 - (\hat{\Pi}_V f)_1\|_n^2\right] \\ &\leq \frac{(1+\delta)}{(1-\delta)}\frac{3}{(1-\rho_0^2)}\left(\|f_1 - \Pi_{V_1}f_1\|^2 + \|f_2 - \Pi_{V_2}f_2\|^2\right).\end{aligned}$$

PROOF OF LEMMA 10. By the projection theorem, we have

$$\|f_1 - (\hat{\Pi}_V f)_1\|_n^2 = \|f_1 - \hat{\Pi}_{V_1}f_1\|_n^2 + \|\hat{\Pi}_{V_1}f_1 - (\hat{\Pi}_V f_1)_1 - (\hat{\Pi}_V f_2)_1\|_n^2.$$

Applying this, the bound $\|x+y\|^2 \leq 2\|x\|^2 + 2\|y\|^2$, and Lemma 10, we obtain

$$\begin{aligned}&\mathbb{E}\left[1_{\mathcal{E}_\delta}\|f_1 - (\hat{\Pi}_V f)_1\|_n^2\right]\\&\leq \frac{(1+\delta)}{(1-\delta)}\frac{3}{(1-\rho_0^2)}\left(\mathbb{E}\left[\|f_1 - \hat{\Pi}_{V_1} f_1\|_n^2 + \|f_2 - \hat{\Pi}_{V_2} f_2\|_n^2\right]\right).\end{aligned}$$

By the projection theorem, we have for $j = 1, 2$,

$$\|f_j - \hat{\Pi}_{V_j} f_j\|_n^2 \leq \|f_j - \Pi_{V_j} f_j\|_n^2.$$

Moreover, taking expectation, we get for $j = 1, 2$,

$$\mathbb{E}\left[\|f_j - \Pi_{V_j} f_j\|_n^2\right] = \|f_j - \Pi_{V_j} f_j\|^2.$$

This completes the proof. □

Now, we begin with the proof of Theorem 4. Repeating the steps (5.16)-(5.19) in the proof of Theorem 3, we have

$$\begin{aligned}&\mathbb{E}\left[\|f_1 - \hat{f}_1^*\|^2\right]\\&\leq \|f_1 - \Pi_{W_1} f_1\|^2 + \frac{1}{(1-\delta)}\mathbb{E}\left[1_{\mathcal{E}_\delta}\|\Pi_{W_1} f_1 - \hat{\Pi}_{W_1}(\hat{\Pi}_V f)_1\|_n^2\right]\\&+ \frac{1}{(1-\delta)}\mathbb{E}\left[1_{\mathcal{E}_\delta}\mathbb{E}\left[\|\hat{\Pi}_{W_1}(\hat{\Pi}_V \boldsymbol{\epsilon})_1\|_n^2 | X^1, \ldots, X^n\right]\right] + R_n.\end{aligned}$$

Applying the bound $\|x+y\|^2 \leq 2\|x\|^2 + 2\|y\|^2$ and the fact that projections lower the norm, we obtain

$$\begin{aligned}&\|\Pi_{W_1} f_1 - \hat{\Pi}_{W_1}(\hat{\Pi}_V f)_1\|_n^2\\&\leq 2\|\Pi_{W_1} f_1 - \hat{\Pi}_{W_1} f_1\|_n^2 + 2\|\hat{\Pi}_{W_1} f_1 - \hat{\Pi}_{W_1}(\hat{\Pi}_V f)_1\|_n^2\\&\leq 2\|\Pi_{W_1} f_1 - \hat{\Pi}_{W_1} f_1\|_n^2 + 2\|f_1 - (\hat{\Pi}_V f)_1\|_n^2.\end{aligned}$$

Thus

$$\begin{aligned}&\mathbb{E}\left[\|f_1 - \hat{f}_1^*\|^2\right]\\&\leq \|f_1 - \Pi_{W_1} f_1\|^2 + \frac{2}{(1-\delta)}\mathbb{E}\left[1_{\mathcal{E}_\delta}\|\Pi_{W_1} f_1 - \hat{\Pi}_{W_1} f_1\|_n^2\right]\\&+ \frac{1}{(1-\delta)}\mathbb{E}\left[1_{\mathcal{E}_\delta}\mathbb{E}\left[\|\hat{\Pi}_{W_1}(\hat{\Pi}_V \boldsymbol{\epsilon})_1\|_n^2 | X^1, \ldots, X^n\right]\right]\\&+ \frac{2}{(1-\delta)}\mathbb{E}\left[1_{\mathcal{E}_\delta}\|f_1 - (\hat{\Pi}_V f)_1\|_n^2\right] + R_n. \qquad (5.27)\end{aligned}$$

Similarly as in Lemma 8, we have

$$\mathbb{E}\left[1_{\mathcal{E}_\delta}\|\Pi_{W_1} f_1 - \hat{\Pi}_{W_1} f_1\|^2\right] \leq \frac{1}{(1-\delta)^2}\frac{\varphi^2 d}{n}\|f_1 - \Pi_{W_1} f_1\|^2. \qquad (5.28)$$

Inserting (5.28), Lemma 10, and Corollary 13 into (5.27), we complete the proof. □

6. Proof of Theorem 5 and 6

6.1. The bias term revisited. In this subsection, we prove two results which lead to improvements of the bias term considered in Lemma 10. This improvement is possible under the additional regularity conditions on the design densities, namely (3.6) or (3.7). We first prove:

PROPOSITION 8. *Let Assumption 1, 2, and 4 be satisfied. Let r, $\phi(V_2)$, $\psi_\Pi(V_2)$, and h_1 be as in Theorem 6. Then*

$$\|(\Pi_V f_2)_1\| \le \frac{1}{1-\rho_0^2}\|h_1\|\psi_\Pi(V_2)\phi(V_2). \tag{6.1}$$

PROOF. Let $\phi_1, \dots, \phi_{d_1}$ be an orthonormal basis of V_1. We have

$$\begin{aligned}
&\|\Pi_{V_1}(f_2 - \Pi_{V_2} f_2)\|^2 \\
&= \sum_{j=1}^{d_1} \Big(\int_{S_1\times S_2} \phi_j(x_1)(f_2(x_2) - \Pi_{V_2} f_2(x_2)) p(x_1,x_2) d(\nu_1\otimes\nu_2)(x_1,x_2)\Big)^2 \\
&= \sum_{j=1}^{d_1} \Big(\int_{S_1} \phi_j(x_1) \int_{S_2} ((1-\Pi_{V_2})f_2(x_2)) \frac{p(x_1,x_2)}{p_1(x_1)p_2(x_2)} d\mathbb{P}^{X_2}(x_2) d\mathbb{P}^{X_1}(x_1)\Big)^2 \\
&= \sum_{j=1}^{d_1} \Big(\int_{S_1} \Big\langle (1-\Pi_{V_2})f_2, \frac{p(x_1,\cdot)}{p_1(x_1)p_2(\cdot)} \Big\rangle_{L^2(\mathbb{P}^{X_2})} \phi_j(x_1) d\mathbb{P}^{X_1}(x_1)\Big)^2,
\end{aligned}$$

where we already applied Fubini's theorem and Assumption 5 in the second equality. Since orthogonal projections are idempotent and self-adjoint, the above is equal to

$$\begin{aligned}
&= \sum_{j=1}^{d_1} \Big(\int_{S_1} \langle (1-\Pi_{V_2})f_2, (1-\Pi_{V_2})(r(x_1,\cdot))\rangle_{L^2(\mathbb{P}^{X_2})} \phi_j(x_1) d\mathbb{P}^{X_1}(x_1)\Big)^2 \\
&\le \int_{S_1} \langle (1-\Pi_{V_2})f_2, (1-\Pi_{V_2})(r(x_1,\cdot))\rangle^2_{L^2(\mathbb{P}^{X_2})} d\mathbb{P}^{X_1}(x_1) \\
&\le \|h_1\|^2 (\psi_\Pi(V_2)\phi(V_2))^2,
\end{aligned}$$

where we applied Bessel's inequality in the first inequality and the Cauchy-Schwarz inequality and (3.7) in the second inequality. Thus we have shown that

$$\|\Pi_{V_1}(f_2 - \Pi_{V_2} f_2)\| \le \|h_1\|\psi_\Pi(V_2)\phi(V_2). \tag{6.2}$$

Now, by Lemma 6, we have

$$\Big\|(\Pi_V h)_1 - \big(\Pi_{V_1} - \sum_{j=1}^{k} (\Pi_{V_1}\Pi_{V_2})^j (1-\Pi_{V_1})\big)h\Big\| \to 0, \tag{6.3}$$

as $k\to\infty$, for all $h \in L^2(\mathbb{P}^X)$. By Assumption 2, we have

$$\|\Pi_{V_1}\Pi_{V_2} h\| \le \rho_0 \|\Pi_{V_2} h\| \quad \text{and} \quad \|\Pi_{V_2}\Pi_{V_1} h\| \le \rho_0 \|\Pi_{V_1} h\|$$

for all $h \in L^2(\mathbb{P}^X)$, which follows as in the proof of (5.7). Applying this and (6.2), we obtain

$$\begin{aligned}
&\big\|\big(\Pi_{V_1} - \sum_{j=1}^{k}(\Pi_{V_1}\Pi_{V_2})^j(1-\Pi_{V_1})\big)(f_2 - \Pi_{V_2}f_2)\big\| \\
&= \big\|\sum_{j=0}^{k}(\Pi_{V_1}\Pi_{V_2})^j\Pi_{V_1}(f_2 - \Pi_{V_2}f_2)\big\| \\
&\leq \sum_{j=0}^{k}\rho_0^{2j}\|\Pi_{V_1}(f_2 - \Pi_{V_2}f_2)\| \leq \frac{1}{1-\rho_0^2}\|h_1\|\psi_\Pi(V_2)\phi(V_2). \qquad (6.4)
\end{aligned}$$

Since $\Pi_V\Pi_{V_2}f_2 = \Pi_{V_2}f_2$ and $(\Pi_{V_2}f_2)_1 = 0$, we have $(\Pi_V f_2)_1 = (\Pi_V(f_2 - \Pi_{V_2}f_2))_1$. Applying this, (6.3), and (6.4), we conclude that

$$\|(\Pi_V f_2)_1\| \leq \frac{1}{1-\rho_0^2}\|h_1\|\psi_\Pi(V_2)\phi(V_2).$$

This completes the proof. □

PROPOSITION 9. *Let Assumption 1, 2, 3, and 4 be satisfied. Let $\phi(V_2)$, $\psi(V_2)$, and h_1 be as in Theorem 5. Suppose that $\|g_1\|_\infty \leq \varphi\sqrt{d_1}\|g_1\|$ for all $g_1 \in V_1$. Then*

$$\begin{aligned}
\mathbb{E}\left[1_{\mathcal{E}_\delta}\|(\hat{\Pi}_V f_2)_1\|_n^2\right] \leq \frac{(1+\delta)^2}{(1-\delta)^3}\frac{2}{(1-\rho_0^2)^2}\Bigg(&\|h_1\|^2(\psi(V_2)\phi(V_2))^2 \\
&+ \frac{1}{n}\|h_1\|^2\|(1-\Pi_{V_2})f_2\|_\infty^2\psi^2(V_2) + \phi^2(V_2)\frac{\varphi^2 d_1}{n}\Bigg). \qquad (6.5)
\end{aligned}$$

PROOF. The proof is similar to the proof of Proposition 8. In the following, suppose that the event $\mathcal{E}_\delta$ holds. Then $\hat{\Pi}_V$ is a well-defined map from $L^2(\mathbb{P}^X)$ to V. Let $\phi_1, \ldots, \phi_{d_1}$ be an orthonormal basis of V_1. By repeating the arguments at the beginning of the proof of Lemma 8, we obtain

$$\|\hat{\Pi}_{V_1}(f_2 - \hat{\Pi}_{V_2}f_2)\|_n^2 \leq \frac{1}{1-\delta}\sum_{j=1}^{d_1}\langle\phi_j, (1-\hat{\Pi}_{V_2})f_2\rangle_n^2.$$

Thus

$$\begin{aligned}
&\mathbb{E}\left[1_{\mathcal{E}_\delta}\|\hat{\Pi}_{V_1}(f_2 - \hat{\Pi}_{V_2}f_2)\|_n^2\right] \\
&\leq \frac{2}{(1-\delta)}\mathbb{E}\left[\sum_{j=1}^{d_1}\langle\phi_{j,\Pi_2}, (1-\hat{\Pi}_{V_2})f_2\rangle_n^2\right] \qquad (6.6) \\
&+ \frac{2}{(1-\delta)}\mathbb{E}\left[\sum_{j=1}^{d_1}\langle\phi_j - \phi_{j,\Pi_2}, (1-\hat{\Pi}_{V_2})f_2\rangle_n^2\right], \qquad (6.7)
\end{aligned}$$

where

$$\phi_{j,\Pi_2}(x_2) = \int \phi_j(x_1) \frac{p(x_1,x_2)}{p_1(x_1)p_2(x_2)} p_1(x_1) d\nu_1(x_1)$$

is the conditional expectation of $\phi_j(X_1)$ given $X_2 = x_2$ (for $\mathbb{P}^{X_2}$-almost all x_2, by Assumption 4). The expectation in (6.6) is equal to

$$\mathbb{E}\left[\sum_{j=1}^{d_1}\left(\int \left\langle \frac{p(x_1,\cdot)}{p_1(x_1)p_2(\cdot)}, (1-\hat{\Pi}_{V_2})f_2 \right\rangle_n \phi_j(x_1)p_1(x_1)d\nu_1(x_1)\right)^2\right]$$
$$\leq \int \mathbb{E}\left[\left\langle \frac{p(x_1,\cdot)}{p_1(x_1)p_2(\cdot)}, (1-\hat{\Pi}_{V_2})f_2 \right\rangle_n^2\right] p_1(x_1)d\nu_1(x_1),$$

where we applied Bessel's inequality and Fubini's theorem in the last inequality. Applying the fact that orthogonal projections are idempotent and self-adjoint and then the Cauchy-Schwarz inequality, this is

$$\leq \int \mathbb{E}\left[\left\|(1-\hat{\Pi}_{V_2})\frac{p(x_1,\cdot)}{p_1(x_1)p_2(\cdot)}\right\|_n^2 \left\|(1-\hat{\Pi}_{V_2})f_2\right\|_n^2\right] p_1(x_1)d\nu_1(x_1).$$

Applying the projection theorem and then (3.6), this is

$$\leq \int \mathbb{E}\left[\left\|(1-\Pi_{V_2})\frac{p(x_1,\cdot)}{p_1(x_1)p_2(\cdot)}\right\|_n^2 \|(1-\Pi_{V_2})f_2\|_n^2\right] p_1(x_1)d\nu_1(x_1)$$
$$\leq \frac{n-1}{n}\|h_1\|^2(\psi(V_2)\phi(V_2))^2 + \frac{1}{n}\|h_1\|^2\|(1-\Pi_{V_2})f_2\|_\infty^2(\psi(V_2))^2.$$

Now we turn to the expectation in (6.7). We have

$$\mathbb{E}\left[\sum_{j=1}^{d_1}\langle \phi_j - \phi_{j,\Pi_2}, (1-\hat{\Pi}_{V_2})f_2\rangle_n^2\right] \leq \frac{\varphi^2 d_1}{n}\mathbb{E}\left[\|(1-\hat{\Pi}_{V_2})f_2\|_n^2\right]$$
$$\leq \frac{\varphi^2 d_1}{n}\|(1-\Pi_{V_2})f_2\|^2.$$

To prove the first inequality, first note that the $((1-\hat{\Pi}_{V_2})f_2)(X_2^i)$ depend only on $X_2^1, \ldots, X_2^n$ and we have

$$\mathbb{E}\left[(\phi_j - \phi_{j,\Pi_2})(X^i)|X_2^1, \ldots, X_2^n, X_1^{i'}\right] = \mathbb{E}\left[(\phi_j - \phi_{j,\Pi_2})(X^i)|X_2^i\right] = 0$$

for $i \neq i'$. This implies that the nondiagonal terms vanish. Next, apply the inequalities $\mathbb{E}[(\phi_j - \phi_{j,\Pi_2})^2(X)|X_2] \leq \mathbb{E}[\phi_j^2(X_1)|X_2]$ and

$$\left\|\sum_{j=1}^{d_1}\phi_j^2\right\|_\infty \leq \varphi^2 d_1,$$

which follows from the bound $\|g_1\|_\infty \le \varphi\sqrt{d_1}\|g_1\|$, for all $g_1 \in V_1$, and [**11**, Lemma 1]. Thus we have shown that

$$\mathbb{E}\left[1_{\mathcal{E}_\delta}\|\hat{\Pi}_{V_1}(f_2 - \hat{\Pi}_{V_2}f_2)\|_n^2\right] \le \frac{2}{(1-\delta)}\Big(\|h_1\|^2(\psi(V_2)\phi(V_2))^2 + \frac{1}{n}\|h_1\|^2\|(1-\Pi_{V_2})f_2\|_\infty^2(\psi(V_2))^2 + \phi^2(V_2)\frac{\varphi^2 d_1}{n}\Big). \quad (6.8)$$

The remaining arguments are as in the proof of Proposition 8. From (5.3) and Lemma 6 we have

$$\big\|(\hat{\Pi}_V h)_1 - \big(\hat{\Pi}_{V_1} - \sum_{j=1}^{k}(\hat{\Pi}_{V_1}\hat{\Pi}_{V_2})^j(1-\hat{\Pi}_{V_1})\big)h\big\|_n \to 0 \quad (6.9)$$

as $k \to \infty$, for all $h \in L^2(\mathbb{P}^X)$. Applying (5.7) and (5.8) as in (6.4), we obtain

$$\big\|\big(\hat{\Pi}_{V_1} - \sum_{j=1}^{k}(\hat{\Pi}_{V_1}\hat{\Pi}_{V_2})^j(1-\hat{\Pi}_{V_1})\big)(f_2 - \hat{\Pi}_{V_2}f_2)\big\|_n \le \frac{1}{1-\rho_{0,\delta}^2}\|\hat{\Pi}_{V_1}(f_2 - \hat{\Pi}_{V_2}f_2)\|_n. \quad (6.10)$$

From (6.9) and (6.10), we conclude that

$$\|(\hat{\Pi}_V f_2)_1\|_n^2 = \|(\hat{\Pi}_V(f_2 - \hat{\Pi}_{V_2}f_2))_1\|_n^2 \le \frac{1}{(1-\rho_{0,\delta}^2)^2}\|\hat{\Pi}_{V_1}(f_2 - \hat{\Pi}_{V_2}f_2)\|_n^2.$$

Combining this with (6.8) and (5.15) gives (6.5). This completes the proof. □

6.2. End of proof of Theorem 5 and 6. The only place where we modify the proof of Theorem 4 is the analysis of the term

$$\frac{2}{(1-\delta)}\mathbb{E}\left[1_{\mathcal{E}_\delta}\|f_1 - (\hat{\Pi}_V f)_1\|_n^2\right], \quad (6.11)$$

which we bounded by

$$\frac{1+\delta}{(1-\delta)^2}\frac{6}{1-\rho_0^2}\left(\|f_1 - \Pi_{V_1}f_1\|^2 + \|f_2 - \Pi_{V_2}f_2\|^2\right), \quad (6.12)$$

by using Lemma 10. We show that, under the additional Assumptions (3.6) or (3.7), one can replace the upper bound (6.12) by the ones given in Theorem 5 and 6, respectively. To achieve this, we replace Lemma 10 by Proposition 8 and 9.

In order to proof Theorem 5 we decompose

$$f_1 - (\hat{\Pi}_V f)_1 = f_1 - (\Pi_V f_1)_1 + (\Pi_V f_1)_1 - (\hat{\Pi}_V f_1)_1 - (\hat{\Pi}_V f_2)_1.$$

Thus

$$\begin{aligned}\mathbb{E}\left[1_{\mathcal{E}_\delta}\|(f_1-(\hat{\Pi}_V f_1)_1\|^2\right] &\leq 3\mathbb{E}\left[\|f_1-(\Pi_V f_1)_1\|_n^2\right]\\ &\quad+3\mathbb{E}\left[1_{\mathcal{E}_\delta}\|(\Pi_V f_1)_1-(\hat{\Pi}_V f_1)_1\|_n^2\right]\\ &\quad+3\mathbb{E}\left[1_{\mathcal{E}_\delta}\|(\hat{\Pi}_V f_2)_1\|_n^2\right].\end{aligned}$$

The third term on the right-hand side is part of Proposition 9. Using Lemma 2 and the projection theorem, the first term can be bounded by

$$\begin{aligned}\mathbb{E}\left[\|f_1-(\Pi_V f_1)_1\|_n^2\right] &= \|f_1-(\Pi_V f_1)_1\|^2\\ &\leq \frac{1}{(1-\rho_0^2)}\|f_1-\Pi_V f_1\|^2\\ &\leq \frac{1}{(1-\rho_0^2)}\|f_1-\Pi_{V_1} f_1\|^2. \qquad (6.13)\end{aligned}$$

Applying Proposition 5 and the projection theorem, the second term can be bounded by

$$\begin{aligned}&\mathbb{E}\left[1_{\mathcal{E}_\delta}\|(\Pi_V f_1)_1-(\hat{\Pi}_V f_1)_1\|_n^2\right]\\ &\leq \frac{(1+\delta)}{(1-\delta)}\frac{1}{(1-\rho_0^2)}\mathbb{E}\left[\|\Pi_V f_1-\hat{\Pi}_V f_1\|_n^2\right]\\ &\leq \frac{(1+\delta)}{(1-\delta)}\frac{1}{(1-\rho_0^2)}\mathbb{E}\left[\|f_1-\Pi_V f_1\|_n^2\right]\\ &\leq \frac{(1+\delta)}{(1-\delta)}\frac{1}{(1-\rho_0^2)}\|f_1-\Pi_{V_1} f_1\|^2.\end{aligned}$$

This completes the proof of Theorem 5. In order to proof Theorem 6 we decompose

$$f_1-(\hat{\Pi}_V f)_1 = f_1-(\Pi_V f_1)_1-(\Pi_V f_2)_1+(\Pi_V f)_1-(\hat{\Pi}_V f)_1.$$

Thus

$$\begin{aligned}\mathbb{E}\left[1_{\mathcal{E}_\delta}\|(f_1-(\hat{\Pi}_V f_1)_1\|^2\right] &\leq 3\mathbb{E}\left[\|f_1-(\Pi_V f_1)_1\|_n^2\right]\\ &\quad+3\mathbb{E}\left[\|(\Pi_V f_2)_1\|_n^2\right]\\ &\quad+3\mathbb{E}\left[1_{\mathcal{E}_\delta}\|(\Pi_V f)_1-(\hat{\Pi}_V f)_1\|_n^2\right].\end{aligned}$$

The first term on the right-hand side is bounded in (6.13), the second one in Proposition 8. Using the definition of $\mathcal{E}_\delta$ and Lemma 2, we obtain

$$\begin{aligned}\mathbb{E}\left[1_{\mathcal{E}_\delta}\|(\Pi_V f)_1-(\hat{\Pi}_V f)_1\|_n^2\right] &\leq (1+\delta)\mathbb{E}\left[1_{\mathcal{E}_\delta}\|(\Pi_V f)_1-(\hat{\Pi}_V f)_1\|^2\right]\\ &\leq (1+\delta)\frac{1}{(1-\rho_0^2)}\mathbb{E}\left[1_{\mathcal{E}_\delta}\|\Pi_V f-\hat{\Pi}_V f\|^2\right].\end{aligned}$$

By Proposition 8, this can be bounded by

$$\frac{(1+\delta)}{(1-\delta)^2}\frac{2}{(1-\rho_0^2)}\frac{\varphi^2 d}{n}\left(\|f_1-\Pi_{V_1}f_1\|^2+\|f_2-\Pi_{V_2}f_2\|^2\right).$$

This completes the proof. □

CHAPTER 3

Variable selection in high-dimensional additive models

1. Introduction

In this chapter, we consider the two related problems of variable selection and component estimation in high-dimensional nonparametric additive models in which the number of covariates is much larger than the number of observations. We study these models under the assumption that most components are equal to zero.

High-dimensional linear models have been investigated intensively in the literature. A great deal of attention has been given to the Lasso (see, e.g., the book by Bühlmann and van de Geer [**7**] and the references therein). The Lasso is based on l_1-penalization and can be used for both estimation and variable selection. There is also a huge literature on estimation and variable selection via l_0-penalization. These procedures can be found, e.g., in the book by Massart [**52**], where a general approach to model selection via penalization is developed (see also the work by Barron, Birgé, and Massart [**6**] and the references therein). Finally, there is a third approach which is based on exponentially weighted aggregation (see, e.g., the work by Rigollet and Tsybakov [**61**] and Arias-Castro and Lounici [**2**] and the references therein).

More recently, high-dimensional additive models have been studied, e.g., in the work by Meier, van de Geer, and Bühlmann [**53**], Huang, Horowitz, and Wei [**37**], Koltchinskii and Yuan [**44**], Raskutti, Wainwright, and Yu [**58**], Gayraud and Ingster [**30**], Suzuki and Sugiyama [**71**], and Dalalyan, Ingster, and Tsybakov [**19**]. One approach generalizes the (group) Lasso and combines sparsity penalties with smoothness penalties or constraints (see [**53, 37, 44, 58, 71**]). As in the case of the Lasso, these procedures can be used for both estimation and variable selection (see [**53, 37**]). Another approach based on exponential aggregation is developed in the work by Dalalyan, Ingster, and Tsybakov [**19**]. They considered the problem of estimation in a more general model which they called the compound model and which includes the additive model as a special case. In a Gaussian white noise setting, they showed that their estimator achieves non-asymptotic minimax rates of convergence.

Comminges and Dalalyan [**18**] considered the problem of variable selection in a high-dimensional Gaussian white noise model and established

tight conditions which make the estimation of the relevant variables possible. They also extended their method to a high-dimensional random regression model, but they assumed that the joint density of all covariates is known. Similar results were obtained earlier by Wainwright [**80**] for high-dimensional linear models with Gaussian measurement matrices.

Several results in the theory of high-dimensional statistical inference are initiated by achievements in the theory of compressive sensing (see, e.g., the introductory book chapters by Fornasier and Rauhut [**28**] and Rauhut [**59**] and the references therein). A popular method is the l_1-minimization which enables sparse recovery if the measurement matrix satisfies, for instance, a restricted isometry property (RIP). It is known that several random matrices satisfy the RIP with probability close to one, important examples being the Gaussian random matrices and the so-called structured random matrices (see, e.g., the work by Candès and Tao [**17**], Baraniuk, Davenport, DeVore, and Walkin [**4**], and Rauhut [**59**]). These results were generalized to high-dimensional linear models by Candès and Tao [**16**] (see also the work by Bickel, Ritov, and Tsybakov [**9**] and the book by Koltchinskii [**43**, Chapters 7 and 8]).

In this chapter, we study a method for variable selection which consists of comparing the norms of the projections of the data onto various finite-dimensional additive subspaces. Given an upper bound q^* for the number of nonzero components, the procedure selects the subset of cardinality smaller than or equal to q^* which best explains the data in the finite sample setting. The basis of this procedure is a selection criterion in the population setting which works well under the essential assumption that the minimal angles between various disjoint additive subspaces are bounded away from zero. Applying this assumption and tools from the theory of structured random matrices, we derive a strong uniform concentration property of the empirical norm around the $L^2(\mathbb{P}^X)$-norm, which, in the special case of independent covariates, can be rewritten as a restricted (block)-isometry property. This property enables us to carry over the geometry in the population setting to the finite sample setting, and thus leads to an analysis of our procedure. Our results are of theoretical interest. Under minimal geometric assumptions, we prove upper bounds for the probability that our procedure misses relevant variables. These concentration inequalities lead to conditions making consistent estimation of the relevant variables possible. In the case of the linear model with random measurement and also in settings considered in the theory of compressive sensing, these conditions coincide with what can be usually found in the literature (see, e.g., [**80, 59**]). In the general case of the nonparametric additive model, we find conditions which are, to the best of our knowledge, new. As an application of our variable selection procedure, we consider the problem of estimating single components. Combining the results of this chapter with those obtained in Chapter 2, we establish conditions under which a single component can be estimated with

the rate of convergence corresponding to the situation in which the other components are known.

2. The main result

2.1. The variable selection problem. Let (Y, X) be a pair of random variables such that $X = (X_1, \dots, X_q)^T$ and

$$Y = \sum_{j=1}^{q} f_j(X_j) + \epsilon, \tag{2.1}$$

where the X_j are real-valued random variables, the f_j are unknown functions which are contained in $L^2(\mathbb{P}^{X_j})$, and ϵ is a Gaussian random variable with expectation 0 and variance σ^2 which is independent of X. Moreover, we suppose that f_j satisfies $\mathbb{E}[f_j(X_j)] = 0$ for $j = 1, \dots, q-1$. We denote by f the whole regression function given by $f(x) = \sum_{j=1}^{q} f_j(x_j)$. We assume that we observe n independent copies $(Y^1, X^1), \dots, (Y^n, X^n)$ of (Y, X), i.e.,

$$Y^i = \sum_{j=1}^{q} f_j(X_j^i) + \epsilon^i, \quad i = 1, \dots, n. \tag{2.2}$$

The number of covariates q can be much larger than the number of observations n, but we assume that the number of non-zero components is smaller than n. Thus we consider a high-dimensional sparse additive model. We define $J_0 = \{j \in \{1, \dots, q\} : \|f_j\| > 0\}$, meaning that we have $f(x) = \sum_{j \in J_0} f_j(x_j)$. Moreover, we denote by s the cardinality of J_0, i.e., $s = |J_0|$. The set J_0 is supposed to be unknown, but we assume that we are given an integer q^* such that $|J_0| \leq q^*$. We aim at selecting a subset of cardinality smaller than or equal to q^* which contains J_0.

2.2. The main assumption. Without any further assumption, the components are not necessarily uniquely determined. In this section, we give an assumption which implies uniqueness and furthermore makes the variable selection task accessible. We define $H_q = L^2(\mathbb{P}^{X_q})$ and

$$H_j = \left\{h_j \in L^2(\mathbb{P}^{X_j}) | \mathbb{E}\left[h_j(X_j)\right] = 0\right\}$$

for $j = 1, \dots, q-1$. Note that $f_j \in H_j$. The spaces H_j are all canonically contained in $L^2(\mathbb{P}^X)$ which is a Hilbert space with the inner product $\langle g, h \rangle = \mathbb{E}[g(X)h(X)]$ and the corresponding norm $\|g\| = \sqrt{\langle g, g \rangle}$. Moreover, for $J \subseteq \{1, \dots, q\}$, we define

$$H_J = \sum_{j \in J} H_j$$

(with the convention that $H_J = 0$ if $J = \emptyset$).

Assumption 11. *There exists a constant $0 \leq \rho < 1$ such that for all subsets $J_1, J_2 \subseteq \{1, \dots, q\}$ satisfying $J_1 \cap J_2 = \emptyset$ and $|J_1|, |J_2| \leq q^*$, we have*

$$\langle h_{J_1}, h_{J_2} \rangle \leq \rho \|h_{J_1}\| \|h_{J_2}\| \tag{2.3}$$

for all $h_{J_1} \in H_{J_1}$, $h_{J_2} \in H_{J_2}$.

It follows from the fact that the spaces H_j are closed combined with Assumption 11 and [**42**, Theorem 1a] (applied inductively) that all spaces H_J with $J \subseteq \{1, \dots, q\}$ and $|J| \leq 2q^*$ are closed. The real number

$$\rho_0(H_{J_1}, H_{J_2}) = \sup \left\{ \frac{\langle h_{J_1}, h_{J_2} \rangle}{\|h_{J_1}\| \|h_{J_2}\|} \middle| 0 \neq h_{J_1} \in H_{J_1}, 0 \neq h_{J_2} \in H_{J_2} \right\}$$

is the cosine of the minimal angle between H_{J_1} and H_{J_2} (see, e.g., [**40**, Definition 1]). Letting $\rho_{q^*} = \max \rho_0(H_{J_1}, H_{J_2})$, where the maximum is taken over all subsets $J_1, J_2 \subseteq \{1, \dots, q\}$ satisfying $J_1 \cap J_2 = \emptyset$ and $|J_1|, |J_2| \leq q^*$, then Assumption 11 says that $\rho_{q^*} < 1$. By a simple argument which is given in Section 5.1, one can show that Assumption 11 can be written as follows:

REMARK 11 (Equivalent form of Assumption 11). For all subsets $J_1, J_2 \subseteq \{1, \dots, q\}$ satisfying $J_1 \cap J_2 = \emptyset$ and $|J_1|, |J_2| \leq q^*$, we have

$$\|h_{J_1} + h_{J_2}\|^2 \geq (1 - \rho_{q^*}^2) \|h_{J_1}\|^2 \tag{2.4}$$

for all $h_{J_1} \in H_{J_1}$, $h_{J_2} \in H_{J_2}$.

Remark 11 shows that Assumption 11 is essential for variable selection: if (2.4) does not hold, then it is possible that f is arbitrary close to a sparse additive function which is based on a completely different set of variables. From (2.4) and the definition of J_0, we obtain:

LEMMA 11. *Let Assumption 11 be satisfied. Then*

$$\kappa := \min_{\emptyset \neq J \subseteq J_0} \Big\| \sum_{j \in J} f_j \Big\|^2 > 0.$$

For $J \subseteq \{1, \dots, q\}$ let Π_{H_J} be the orthogonal projection from $L^2(\mathbb{P}^X)$ to H_J. In the following we abbreviate Π_{H_J} as Π_J. Since projections lower the norm, the set J_0 maximizes the quantity $\|\Pi_J f\|^2$. If Assumption 11 holds, the following Lemma shows that $\|\Pi_{J_0} f\|^2 - \|\Pi_J f\|^2$ is strictly positive for all subsets $J \subseteq \{1, \dots, q\}$ with $|J| \leq q^*$ and $J_0 \setminus J \neq \emptyset$. This means that a subset $J \subseteq \{1, \dots, q\}$ with $|J| \leq q^*$ which maximizes $\|\Pi_J f\|^2$ always contains J_0 (and is equal to J_0 in the special case when $|J_0| = q^*$). These observations will be the theoretical basis for our selection criterion in the finite sample setting.

PROPOSITION 10. *Let Assumption 11 be satisfied, and let* $J \subseteq \{1, \dots, q\}$ *be a subset such that* $|J| \leq q^*$ *and* $J_0 \setminus J \neq \emptyset$. *Then*

$$\|\Pi_{J_0} f\|^2 - \|\Pi_J f\|^2 = \|f - \Pi_J f\|^2 \geq (1 - \rho_{q^*}^2) \kappa_l,$$

where $l = |J_0 \setminus J|$ *and*

$$\kappa_l := \min_{J' \subseteq J_0, |J'| = l} \Big\| \sum_{j \in J'} f_j \Big\|^2.$$

PROOF. The equality follows from $\Pi_{J_0} f = f$ and the projection theorem. We turn to the proof of the inequality. We have $f = \sum_{j \in J_0 \cap J} f_j + \sum_{j \in J_0 \setminus J} f_j =: f_{J_0 \cap J} + f_{J_0 \setminus J}$. Hence

$$\Pi_J f = f_{J_0 \cap J} + \Pi_J f_{J_0 \setminus J}$$

and

$$f - \Pi_J f = f_{J_0 \setminus J} - \Pi_J f_{J_0 \setminus J}.$$

We have $f_{J_0 \setminus J} \in H_{J_0 \setminus J}$, $\Pi_J f_{J_0 \setminus J} \in H_J$, and $l = |J_0 \setminus J| \geq 1$. Thus (2.4) and the definition of κ_l yield

$$\|f - \Pi_J f\|^2 = \left\|f_{J_0 \setminus J} - \Pi_J f_{J_0 \setminus J}\right\|^2 \geq (1 - \rho_{q^*}^2)\|f_{J_0 \setminus J}\|^2 \geq (1 - \rho_{q^*}^2)\kappa_l.$$

This completes the proof. □

Finally, we show that ρ_{q^*} can be related to a quantity which is known in the literature on sparse additive models (see, e.g., [**44**]).

LEMMA 12. *Let ϵ_{2q^*} be the smallest number such that*

$$\Big\| \sum_{j \in J} f_j \Big\|^2 \geq (1 - \epsilon_{2q^*}) \Big(\sum_{j \in J} \|f_j\|^2 \Big) \tag{2.5}$$

for all $J \subseteq \{1, \ldots, q\}$ with $|J| \leq 2q^*$ and all $\sum_{j \in J} f_j \in H_J$. Then we have $\rho_{q^*} < 1$ if and only if $\epsilon_{2q^*} < 1$.

A proof of this lemma is given in Section 5.2.

2.3. The selection criterion. In this section, we construct the selection criterion. For $j = 1, \ldots, q$, let $V_j \subseteq H_j$ be finite-dimensional linear subspaces. For $J \subseteq \{1, \ldots, q\}$, let

$$V_J = \sum_{j \in J} V_j$$

and $d_J = \dim V_J$. Moreover, for $l = 1, \ldots, q$, let $d_l = \max_{|J|=l} d_J$.

In order to proceed, we introduce some further notation. Let $\|\cdot\|_n$ be the empirical norm which is defined by

$$\|h\|_n^2 = \frac{1}{n} \sum_{i=1}^n h^2(X^i)$$

for $h \in L^2(\mathbb{P}^X)$, and which is defined by $\|\cdot\|_n^2 = (1/n)\|\cdot\|_2^2$ if applied to vectors in $\mathbb{R}^n$. Here, $\|\cdot\|_2$ denotes the usual Euclidean norm. Let $\hat{\Pi}_J$ be the orthogonal projection from $\mathbb{R}^n$ to the subspace $\{(g_J(X^1), \ldots, g_J(X^n))^T | g_J \in V_J\}$. If $h \in L^2(\mathbb{P}^X)$, then we abbreviate $\hat{\Pi}_J(h(X^1), \ldots, h(X^n))^T$ as $\hat{\Pi}_J h$. Finally, let $\mathbf{Y} = (Y^1, \ldots, Y^n)^T$ and $\boldsymbol{\epsilon} = (\epsilon^1, \ldots, \epsilon^n)^T$. Motivated by Proposition 10, we define an estimator $\hat{J}_0$ of J_0 as follows:

$$\hat{J}_0 = \operatorname*{arg\,max}_{J \subseteq \{1, \ldots, q\}, |J| \leq q^*} \left(\left\|\hat{\Pi}_J \mathbf{Y}\right\|_n^2 - \sigma^2 d_J / n \right). \tag{2.6}$$

Conditioning on $X^1, \ldots, X^n$, the random variable $(n/\sigma^2)\|\hat{\Pi}_J \boldsymbol{\epsilon}\|_n^2$ has a chi-square distribution with $\operatorname{rank}(\hat{\Pi}_J) \leq d_J$ degrees of freedom and the last term is supposed to cancel its expectation. The last term can also be seen as a penalty term. In fact, the criterion in (2.6) can be written as a penalized least squares criterion (see, e.g., [**52**]).

The success of the criterion depends on a suitable choice of the V_j, which in turn depends on the regularity conditions of the f_j. For instance, if the f_j belong to some known finite-dimensional linear subspaces of H_j, then we let the V_j be equal to these spaces. In the following, we consider the nonparametric case. Without loss of generality, we shall restrict our attention to (periodic) Sobolev smoothness and spaces of trigonometric polynomials. A similar treatment is possible, e.g., for Hölder smoothness and spaces of piecewise polynomials or spaces of splines. Recall that the trigonometric basis is given by $\phi_1(x) = 1$ and $\phi_{2k}(x) = \sqrt{2}\cos(2\pi kx)$ and $\phi_{2k+1}(x) = \sqrt{2}\sin(2\pi kx)$, $k \geq 1$, where $x \in [0,1]$.

ASSUMPTION 12. *Suppose that the X_j take values in $[0,1]$ and have densities p_j with respect to the Lebesgue measure on $[0,1]$, which satisfy $c \leq p_j \leq 1/c$ for some constant $c > 0$. Moreover, suppose that the f_j belong to the Sobolev classes*

$$\tilde{W}_j(\alpha_j, K_j) = \left\{ \sum_{k=1}^{\infty} \theta_k \phi_k(x_j) \;:\; \sum_{k=1}^{\infty} (2\pi k)^{2\alpha_j} (\theta_{2k}^2 + \theta_{2k+1}^2) \leq K_j^2 \right\},$$

where $\alpha_j > 1/2$ and $K_j > 0$ (see, e.g., [**74**, Definition 1.12]*).*

For $j = 1, \ldots, q$, let V_j be the intersection of H_j with the linear span of $\phi_1, \ldots, \phi_{m_j}$ (in the variable x_j). The choice of the m_j will depend on the following approximation properties.

LEMMA 13. *Let Assumption 12 be satisfied. Then there exists a constant $C_j > 0$ depending only on α_j and c (given explicitly in the proof) such that*

$$\|h_j - \Pi_{V_j} h_j\|^2 \leq C_j K_j^2 m_j^{-2\alpha_j} \quad \text{and}$$
$$\|h_j - \Pi_{V_j} h_j\|_\infty^2 \leq C_j K_j^2 m_j^{1-2\alpha_j}$$

for all $h_j \in \tilde{W}_j(\alpha_j, K_j) \cap H_j$, where Π_{V_j} is the orthogonal projection from $L^2(\mathbb{P}^X)$ to V_j.

For completeness, a proof of this lemma is given in Section 5.3. We suppose that for $j = 1, \ldots, q$,

$$m_j \geq \left(\frac{C_j K_j^2 q^* (1 + \epsilon'_{q^*})}{c'(1 - \rho_{q^*}^2)\kappa} \right)^{1/2\alpha_j}, \tag{2.7}$$

where $0 < c' < 1$ is a small constant satisfying (4.3) and ϵ'_{q^*} is a positive real number such that

$$\Big\| \sum_{j \in J} f_j \Big\|^2 \leq (1 + \epsilon'_{q^*}) \Big(\sum_{j \in J} \|f_j\|^2 \Big) \tag{2.8}$$

for all $J \subseteq \{1, \dots, q\}$ with $|J| \leq q^*$ and all $\sum_{j \in J} f_j \in H_J$. Note that, by the Cauchy-Schwarz inequality, we can always choose $1 + \epsilon'_{q^*} = q^*$. The m_j are chosen such that the following upper bound holds

$$\left\| f - \sum_{j \in J_0} \Pi_{V_j} f_j \right\|^2 \leq (1 + \epsilon'_{q^*}) \sum_{j \in J_0} \|f_j - \Pi_{V_j} f_j\|^2 \leq c'(1 - \rho^2_{q^*})\kappa, \qquad (2.9)$$

where we used (2.8) and Lemma 13. Applying Bennett's inequality and Lemma 13, one can show that a similar bound holds with high probability when the $L^2(\mathbb{P}^X)$-norm is replaced by the empirical norm $\|\cdot\|_n$. The result is as follows:

LEMMA 14. *Let Assumptions 11 and 12 be satisfied. Suppose that* (2.7) *is satisfied for* $j = 1, \dots, q$. *Let the event* $\mathcal{A}$ *be given by*

$$\mathcal{A} = \left\{ \left\| f - \sum_{j \in J_0} \Pi_{V_j} f_j \right\|_n^2 \leq 2c'(1 - \rho^2_{q^*})\kappa \right\}.$$

Then

$$\mathbb{P}(\mathcal{A}^c) \leq \exp\left(-\frac{3}{16} \frac{n}{d_{q^*}} \right). \qquad (2.10)$$

A proof of Lemma 14 is given in Section 5.4

2.4. The main result and some consequences. In this section, we present our first main theorem and derive several consequences. These results will be further developed in Section 3, where the final results can be found.

For $J \subseteq \{1, \dots, q\}$ and $0 < \delta < 1$ (e.g. $\delta = 1/2$), we define the events

$$\mathcal{E}_{\delta, J} = \left\{ (1 - \delta)\|g_J\|^2 \leq \|g_J\|_n^2 \leq (1 + \delta)\|g_J\|^2 \text{ for all } g_J \in V_J \right\}.$$

Moreover, we define

$$\mathcal{E}_{\delta, q^*} = \bigcap_{J \subseteq \{1, \dots, q\}, |J| \leq q^*} \mathcal{E}_{\delta, J \cup J_0}.$$

We prove:

THEOREM 10. *Let Assumptions 11 and 12 be satisfied. Let* $0 < \delta < 1$. *Suppose that* (2.7) *is satisfied for* $j = 1, \dots, q$. *Then there is a constant* $c_1 > 0$ *depending only on* δ *(given explicitly in the proof) such that*

$$\mathbb{P}\left(J_0 \subseteq \hat{J}_0 \right) \geq 1 - \mathbb{P}\left(\mathcal{E}^c_{\delta, q^*} \right) - \exp\left(-\frac{3}{16} \frac{n}{d_{q^*}} \right)$$
$$- \sum_{l=1}^{s} \sum_{m=0}^{q^* - (s-l)} \binom{s}{l} \binom{q-s}{m} 4 \exp\left(-c_1 \frac{n^2 (1 - \rho^2_{q^*})^2 \kappa_l^2}{\sigma^4 d_{q^* - s + l} + \sigma^2 n (1 - \rho^2_{q^*}) \kappa_l} \right). \qquad (2.11)$$

Recall, that the d_l *are given by* $d_l = \max_{|J| = l} d_J$.

REMARK 12. Theorem 10 also holds in the parametric case, i.e., if $f_j \in V_j$ for $j \in J$. In this case, only Assumption 11 has to be satisfied, Assumption 12 and the condition (2.7) disappear. Moreover, in (2.11) the term $\exp(-3n/(16d_{q^*}))$ can be dropped.

The bound (2.11) yields the following simpler one

$$\mathbb{P}\left(J_0 \subseteq \hat{J}_0\right) \geq 1 - \mathbb{P}\left(\mathcal{E}^c_{\delta,q^*}\right) - \exp\left(-\frac{3}{16}\frac{n}{d_{q^*}}\right) - \left(\frac{eq}{q^*}\right)^{q^*} 4\exp\left(-c_1 \frac{n^2(1-\rho^2_{q^*})^2\kappa^2}{\sigma^4 d_{q^*} + \sigma^2 n(1-\rho^2_{q^*})\kappa}\right). \tag{2.12}$$

This can be seen as follows. First, we successively apply the bounds $\kappa_l \geq \kappa$ and $d_{q^*-s+l} \leq d_{q^*}$. Then, we use the following combinatorial result (for a proof see, e.g., [**52**, Proposition 2.5])

$$\sum_{j=0}^{q^*} \binom{q}{j} \leq \left(\frac{eq}{q^*}\right)^{q^*}. \tag{2.13}$$

From (2.12), we conclude:

COROLLARY 14. *Suppose that the assumptions of Theorem 10 hold. Then for each constant $c_2 > 0$, there is a constant $c_3 > 0$ (depending only on c_1 and c_2) such that*

$$\mathbb{P}\left(J_0 \subseteq \hat{J}_0\right) \geq 1 - \mathbb{P}\left(\mathcal{E}^c_{\delta,q^*}\right) - q^{-c_2},$$

provided that

$$\max\left\{\frac{\sigma^2\sqrt{q^* d_{q^*}\log(eq/q^*)}}{(1-\rho^2_{q^*})\kappa}, \frac{\sigma^2 q^*\log(eq/q^*)}{(1-\rho^2_{q^*})\kappa}, d_{q^*}\log q\right\} \leq c_3 n.$$

REMARK 13. Corollary 14 also holds if the assumptions of Remark 12 are satisfied. In this case, the term $d_{q^*}\log q$ can be dropped.

Next, we present another analysis of (2.11) in the case that $q^* = s$. Then $J_0 \subseteq \hat{J}_0$ if and only if $J_0 = \hat{J}_0$. Thus, we can rewrite (2.11) as

$$\mathbb{P}\left(J_0 \neq \hat{J}_0\right) \leq \mathbb{P}\left(\mathcal{E}^c_{\delta,q^*}\right) + \exp\left(-\frac{3}{16}\frac{n}{d_{q^*}}\right) + \sum_{l=1}^{s}\sum_{m=0}^{l}\binom{s}{l}\binom{q-s}{m} 4\exp\left(-c_1\frac{n^2(1-\rho^2_s)^2\kappa_l^2}{\sigma^4 d_l + \sigma^2 n(1-\rho^2_s)\kappa_l}\right). \tag{2.14}$$

Applying $\kappa_l \geq (1-\epsilon_s)l\kappa_1$, $d_l \leq l d_1$, and

$$\sum_{m=0}^{l}\binom{s}{l}\binom{q-s}{m} \leq q^{2l},$$

the last expression in (2.14) can be bounded by

$$\sum_{l=1}^{s} 4q^{2l} \exp\left(-c_1 \frac{l(n(1-\rho_s^2)(1-\epsilon_s)\kappa_1)^2}{\sigma^4 d_1 + \sigma^2 n(1-\rho_s^2)(1-\epsilon_s)\kappa_1}\right).$$

We obtain:

COROLLARY 15. *Suppose that the assumptions of Theorem 10 hold. Moreover, suppose that $q^* = s$. Then for each constant $c_2 > 0$, there is a constant $c_3 > 0$ (depending only on c_1 and c_2) such that*

$$\mathbb{P}\left(J_0 \neq \hat{J}_0\right) \leq \mathbb{P}\left(\mathcal{E}_{\delta,q^*}^c\right) + q^{-c_2}, \tag{2.15}$$

provided that

$$\max\left\{\frac{\sigma^2\sqrt{d_1 \log q}}{(1-\rho_s^2)(1-\epsilon_s)\kappa_1}, \frac{\sigma^2 \log q}{(1-\rho_s^2)(1-\epsilon_s)\kappa_1}, d_s \log q\right\} \leq c_3 n.$$

REMARK 14. Corollary 15 also holds if the assumptions of Remark 12 are satisfied. In this case, the term $d_s \log q$ can be dropped. In the special case that the covariates are also independent, we have $\rho_s = \epsilon_s = 0$ and the conditions become

$$\max\left\{\frac{\sigma^2\sqrt{d_1 \log q}}{\kappa_1}, \frac{\sigma^2 \log q}{\kappa_1}\right\} \leq c_3 n.$$

Note that in Chapter 4, we show that this conditions are also necessary.

Finally, we mention that in the case $q^* = s$, the conditions in Corollary 14 and 15 are both consequences of a more general condition. One can show that for each $c_2 > 0$, there is a $c_3 > 0$ such that (2.15) holds, provided that for $l = 1, \ldots, s$,

$$\max\left\{\frac{\sigma^2\sqrt{l d_l \log(eq/l)}}{(1-\rho_s^2)\kappa_l}, \frac{\sigma^2 l \log(eq/l)}{(1-\rho_s^2)\kappa_l}, d_s \log q\right\} \leq c_3 n. \tag{2.16}$$

Note that Corollary 14 follows from the bounds $\kappa_l \geq \kappa$ and the fact that $l \log(eq/l)$ is increasing in l for $1 \leq l \leq q$, and Corollary 15 follows (up to the constant e in the logarithm) from the bounds $\kappa_l \geq (1-\epsilon_s) l \kappa_1$ and $d_l \leq l d_1$ and the fact that $\log(eq/l)$ is decreasing in l.

3. Structured random matrices and the event $\mathcal{E}_{\delta,q^*}$

3.1. Independent covariates and the RIP. In this subsection, we suppose that $X_1, \ldots, X_n$ are independent, which implies that the spaces $V_1, \ldots, V_q$ are orthogonal in $L^2(\mathbb{P}^X)$. In this particular case, we rewrite the event $\mathcal{E}_{\delta,q^*}$ as a restricted (block)-isometry property. This allows us to apply known concentration inequalities.

For $j = 1, \ldots, q$, let $\{\phi_{jk}\}_{1 \leq k \leq \dim V_j}$ be an orthonormal basis of V_j. Then we define the $n \times \dim V_j$-matrix

$$A_j = \frac{1}{\sqrt{n}} \left(\phi_{jk}(X_j^i)\right)_{1 \leq i \leq n, 1 \leq k \leq \dim V_j}$$

and for $J \subseteq \{1, \dots, q\}$, we define the $n \times d_J$-matrix $A_J = (A_j)_{j \in J}$ (we abbreviate $A_{\{1,\dots,q\}}$ as A). With these definitions, it is easy to see that $\mathcal{E}_{\delta,J}$ is the event such that

$$(1-\delta)\|z_J\|_2^2 \leq \|A_J z_J\|_2^2 \leq (1+\delta)\|z_J\|_2^2$$

for all $z_J \in \mathbb{R}^{d_J}$. Here, we have used that the spaces $V_1, \dots, V_q$ are orthogonal. Thus, if we define

$$\delta_{q^*} = \max_{J \subseteq \{1,\dots,q\}, |J| \leq q^*} \|A_{J \cup J_0}^T A_{J \cup J_0} - I\|_{\mathrm{op}},$$

then we have

$$\mathcal{E}_{\delta,q^*} = \{\delta_{q^*} \leq \delta\}.$$

The constant δ_{q^*} is bounded by the restricted isometry constant of order d_{2q^*} of the matrix A (see [**59**, Definition 2.4]). Note that the restricted isometry constant plays a prominent role in the theory of sparse recovery. Moreover, there exist many concentration inequalities for the restricted isometry constant in many ensembles of random matrices. We give two examples.

EXAMPLE 2. Consider the model $Y = \sum_{j=1}^q X_j \beta_j + \epsilon$, where the X_j are independent centered Gaussian random variables and the β_j are real numbers. Then A is a Gaussian random matrix (the entries are independent Gaussian random variables, each with expectation zero and variance $1/n$), and [**4**, Theorem 5.2] implies that there exist constants $c_3, c_4 > 0$ depending only on δ such that $\mathbb{P}(\delta_{q^*} \leq \delta) \geq 1 - 2\exp(-c_4 n)$, provided that $q^* \log(q/q^*) \leq c_3 n$. Combining this with Corollary 15, we obtain (in the case $q^* = s$) that $\mathbb{P}(J_0 \neq \hat{J}_0) \leq 2\exp(-c_4 n) + q^{-c_2}$, provided that

$$\max\left\{ s \log(q/s), \frac{\sigma^2 \log q}{\kappa_1} \right\} \leq c_3 n.$$

These conditions are also known to be necessary (see, e.g., [**59**, Section 2.6] for the setting without noise and [**80**, Theorem 2] for the noisy setting).

EXAMPLE 3. Consider the nonparametric case where $X_1, \dots, X_n$ are independent and uniformly distributed on $[0,1]$. Then the trigonometric bases of the V_j are also orthonormal bases and we can apply [**59**, Theorem 8.4] (recall that the constant δ_{q^*} is bounded by the restricted isometry constant of order d_{2q^*}) which says that there are constants $c_3, c_4 > 0$ such that for $\delta \leq 1/2$, $\mathbb{P}(\delta_{q^*} > \delta) \leq \exp(-c_4 n \delta^2 / d_{2q^*})$, provided that $d_{2q^*} \log^2(100 d_{2q^*}) \log(4d_q) \log(10n) \leq c_3 n \delta^2$.

3.2. A general upper bound for $\mathbb{P}(\mathcal{E}_{\delta,q^*}^c)$. In this section, we give a general upper bound for the probability that the event $\mathcal{E}_{\delta,q^*}^c$ occurs. This upper bound is a generalization of [**64**, Theorem 3.3] and [**59**, Theorem 8.1 and 8.4]. The derivation will consist in two steps. The first step is the following generalization of Theorem 3.6 by Rudelson and Vershynin [**64**].

PROPOSITION 11. *Let Assumptions 11 and 12 be satisfied. Then there is a universal constant $C_1 > 0$ such that*

$$\mathbb{E}\left[\sup_{g\in V_J, |J|\leq 2q^*, \|g\|\leq 1} \left|\|g\|_n^2 - \|g\|^2\right|\right] \leq C_1\sqrt{\frac{d_{2q*}}{c(1-\epsilon_{2q^*})n}}\log^2(d_q \vee n), \quad (3.1)$$

provided that the last expression is smaller than 1.

A proof of Proposition 11 is given in Section 5.5. The second step is an application of Talagrand's inequality (see [**72**]). Here, we state a version of Talagrand's inequality presented in [**11**, Corollary 2]:

THEOREM 11 (Talagrand's inequality). *Let $X^1, \ldots, X^n$ be n independent and identically distributed random variables taking values in some measurable space $(S, \mathcal{B})$. Let $\mathcal{G}$ be a countable family of real-valued measurable functions on $(S, \mathcal{B})$ that are uniformly bounded by some constant b. Let $Z = \sup_{g\in\mathcal{G}} \left|\frac{1}{n}\sum_{i=1}^n g(X^i) - \mathbb{E}\left[g(X^i)\right]\right|$ and $v = \sup_{g\in\mathcal{G}} \mathbb{E}\left[g^2(X^1)\right]$. Then for every positive number λ,*

$$\mathbb{P}\left(Z \geq 2\mathbb{E}\left[Z\right] + \lambda\right) \leq 3\exp\left(-n\kappa\left(\frac{\lambda^2}{v} \wedge \frac{\lambda}{b}\right)\right),$$

where κ is a universal constant.

We want to apply Talagrand's inequality to the family $\mathcal{G} = \{g^2 : g \in V_J, |J| \leq 2q^*, \|g\| \leq 1\}$. This family is not countable, but the value of Z does not change if we restrict the supremum to a countable and dense subset (note that the V_J are finite-dimensional spaces). For $J \subseteq \{1, \ldots, q\}$, let

$$\varphi_J = \frac{1}{\sqrt{d_J}} \sup_{0\neq g\in V_J} \frac{\|g\|_\infty}{\|g\|}.$$

Moreover, let $\varphi_{2q^*} = \max_{|J|\leq 2q^*} \varphi_J$. Under Assumptions 11 and 12, we have

$$\varphi_{2q^*}^2 \leq \frac{2}{c(1-\epsilon_{2q^*})}, \quad (3.2)$$

the details are given in Section 5.6. Therefore, for all $g^2 \in \mathcal{G}$, we have

$$\|g\|_\infty^2 \leq \varphi_{2q^*}^2 d_{2q^*}\|g\|^2 \leq \frac{2d_{2q^*}}{c(1-\epsilon_{2q^*})}.$$

Using this and the bound $\mathbb{E}\left[g^4(X^1)\right] \leq \|g\|_\infty^2\|g\|^2$, we conclude that $b, v \leq 2d_{2q^*}/(c(1-\epsilon_{2q^*}))$. Now, suppose that the last expression in (3.1) is smaller than $\delta/4$, $0 < \delta < 1$. Then Theorem 11, applied with $\lambda = \delta/2$, yields

$$\begin{aligned}\mathbb{P}\left(\mathcal{E}_{\delta,q^*}^c\right) &\leq \mathbb{P}\left(\sup_{g\in V_J, |J|\leq 2q^*, \|g\|\leq 1} \left|\|g\|_n^2 - \|g\|^2\right| > \delta\right) \\ &\leq 3\exp\left(-n\kappa\frac{c(1-\epsilon_{2q^*})\delta^2}{8d_{2q^*}}\right).\end{aligned}$$

We have shown:

THEOREM 12. *Let Assumptions 11 and 12 be satisfied. Let* $c_4 = c\kappa/8$ *and* $c_3 = \sqrt{c}/(4C_1)$*, where* C_1 *and* κ *are the constants in Proposition 11 and Talagrand's inequality, respectively. Let* $\delta \in (0,1)$*. Suppose that*

$$\sqrt{\frac{d_{2q^*}}{(1-\epsilon_{2q^*})n}}\log^2(d_q \vee n) \leq c_3\delta.$$

Then

$$\mathbb{P}\left(\mathcal{E}^c_{\delta,q^*}\right) \leq 3\exp\left(-c_4\frac{(1-\epsilon_{2q^*})n\delta^2}{d_{2q^*}}\right).$$

3.3. Conditions for variable selection. In this section, we combine Corollary 15 with Theorem 12. Therefore, suppose that the assumptions of Theorem 10 hold. To simplify the exposition, we will treat the quantities $\alpha = \min_j \alpha_j$, $K = \max_j K_j$, and c from Assumption 12 and the geometric quantities ρ_s, ϵ_{2s}, and ϵ'_s as constants. Moreover, we assume that $q^* = s$ and that $q \geq n$. Recall from (2.7) that in this case it suffices to choose the m_j of size constant times $(s/\kappa)^{1/(2\alpha)}$. By the inequalities $\kappa_l \geq l(1-\epsilon_l)\kappa_1$, we have that κ is bounded from below by a constant times κ_1, which in turn implies that the m_j can be chosen of size constant times $(s/\kappa_1)^{1/(2\alpha)}$. Inserting this into Corollary 15 and Theorem 12 (let, e.g., $\delta = 1/2$), we obtain:

THEOREM 13. *Make the above assumptions. Then for each constant* $c_2 > 0$*, there are constants* $c_3 > 0$ *and* $c_5 > 0$ *such that*

$$\mathbb{P}\left(J_0 \neq \hat{J}_0\right) \leq q^{-c_2} + q^{-c_5\log^3 q},$$

provided that

$$\max\left\{\frac{\sigma^2 s^{1/(4\alpha)}\sqrt{\log q}}{\kappa_1^{(4\alpha+1)/(4\alpha)}}, \frac{\sigma^2\log q}{\kappa_1}, \frac{s^{(2\alpha+1)/(2\alpha)}\log^4 q}{\kappa_1^{1/(2\alpha)}}\right\} \leq c_3 n. \tag{3.3}$$

REMARK 15. In Chapter 4, we will show that the condition

$$\max\left\{\frac{\sigma^2\sqrt{\log q}}{\kappa_1^{(4\alpha+1)/(4\alpha)}}, \frac{\sigma^2\log q}{\kappa_1}\right\} \leq c_3 n.$$

is optimal in an additive Gaussian white noise model. Obviously, this condition is weaker than (3.3). In (3.3), we have the additional factor $s^{1/(4\alpha)}$ in the first term, and we have an additional term coming from the event $\mathcal{E}_{\delta,q^*}$ (note that this event disappears in the Gaussian white noise framework).

3.4. Estimation of single components. The proposed selection criterion can be seen as a method to reduce the dimension of the model. We start with n independent observations of a sparse additive model with q covariates and an unknown subset J_0 of indices corresponding to the non-zero components, and we end up with a subset $\hat{J}_0$ such that $|\hat{J}_0| \leq q^*$ and

$J_0 \subseteq \hat{J}_0$ with high probability. More precisely, if $\{J_0 \subseteq \hat{J}_0\}$ holds, then we have successfully reduced the model (2.1) to

$$Y = \sum_{j \in \hat{J}_0} f_j(X_j) + \epsilon. \tag{3.4}$$

We now consider the problem of estimating a single component f_j of the model (2.1) with $j \in J_0$. We may assume without loss of generality that $j = 1$. To simplify the exposition, we make the same assumptions as in the previous Section 3.3. We split the sample into two parts. More precisely, we assume that we observe an even number of independent copies $(Y^1, X^1), \ldots, (Y^{2n}, X^{2n})$ of (Y, X). The estimator $\hat{J}_0$ of J_0 is constructed as in Section 2.3 using the sample $(Y^1, X^1), \ldots, (Y^n, X^n)$, and the estimator $\hat{f}_1$ of f_1 is constructed as in Section 4.3 of Chapter 2 using $\hat{J}_0$ and the sample $(Y^{n+1}, X^{n+1}), \ldots, (Y^{2n}, X^{2n})$. We have

$$\mathbb{E}\left[\|f_1 - \hat{f}_1^*\|^2\right] \leq \mathbb{E}\left[1_{\{\hat{J}_0 = J_0\}} \|f_1 - \hat{f}_1^*\|^2\right] + (\|f_1\| + k_n)^2 \mathbb{P}\left(J_0 \neq \hat{J}_0\right)$$

meaning that we can apply Theorem 7 to the first term (note that Assumption 11 implies Assumptions 1 and 2) and Theorem 13 to the second term.

THEOREM 14. *Make the same assumptions as in Theorem 13. Then there are constants $c_3 > 0$, $C > 0$ such that*

$$\mathbb{E}\left[\|f_1 - \hat{f}_1^*\|^2\right] \leq C n^{\frac{-2\alpha_1}{2\alpha_1 + 1}},$$

provided that (3.3) *is satisfied and that*

$$s^{(2\alpha+1)/(2\alpha)} n^{\frac{2\alpha_1}{2\alpha(2\alpha_1+1)}} \log^4 n \leq c_3 n.$$

4. Outline of the proof of Theorem 10

4.1. The finite sample geometry. In this section, we present empirical versions of Assumption 11 and Proposition 10. Throughout this section, let $0 < \delta < 1$ be the constant in Theorem 10. Recall that in Section 2.4, we defined the events

$$\mathcal{E}_{\delta,J} = \left\{(1-\delta)\|g_J\|^2 \leq \|g_J\|_n^2 \leq (1+\delta)\|g_J\|^2 \text{ for all } g_J \in V_J\right\}$$

for $J \subseteq \{1, \ldots, q\}$. Written in the equivalent form of Remark 11, we have:

LEMMA 15. *Let Assumption 11 be satisfied. Let $J_1, J_2 \subseteq \{1, \ldots, q\}$ be two subsets such that $J_1 \cap J_2 = \emptyset$ and $|J_1|, |J_2| \leq q^*$. If $\mathcal{E}_{\delta, J_1 \cup J_2}$ holds, then we have*

$$\|g_{J_1} + g_{J_2}\|_n^2 \geq \frac{(1-\delta)}{(1+\delta)}(1 - \rho_{q^*}^2) \|g_{J_1}\|_n^2 \tag{4.1}$$

for all $g_{J_1} \in V_{J_1}$, $g_{J_2} \in V_{J_2}$.

PROOF. Under the assumptions of Lemma 15, we have

$$\begin{aligned}\|g_{J_1} + g_{J_2}\|_n^2 &\geq (1-\delta)\|g_{J_1} + g_{J_2}\|^2 \\ &\geq (1-\delta)(1-\rho_{q^*}^2)\|g_{J_1}\|^2 \\ &\geq \frac{(1-\delta)}{(1+\delta)}(1-\rho_{q^*}^2)\|g_{J_1}\|_n^2.\end{aligned}$$

This completes the proof. □

Applying (4.1) as in the proof of Proposition 10, we obtain:

PROPOSITION 12. *Let Assumption 11 be satisfied. Let $J \subseteq \{1,\dots,q\}$ be a subset such that $|J| \leq q^*$ and $J_0 \setminus J \neq \emptyset$. Let $v = \sum_{j\in J_0} v_j$ with $v_j \in V_j$ for $j \in J_0$. If $\mathcal{E}_{\delta,J\cup J_0}$ holds, then we have*

$$\|\hat{\Pi}_{J_0} v\|_n^2 - \|\hat{\Pi}_J v\|_n^2 = \|v - \hat{\Pi}_J v\|_n^2 \geq \frac{(1-\delta)}{(1+\delta)}(1-\rho_{q^*}^2)\Big\|\sum_{j\in J_0\setminus J} v_j\Big\|_n^2.$$

By decomposing f as $v + f - v$ with $v = \sum_{j\in J_0} \Pi_{V_j} f_j$, we can apply Proposition 12 to v and Lemma 14 to $f - v$. The result is the following empirical version of Proposition 10.

PROPOSITION 13. *Let Assumption 11 and Assumption 12 be satisfied. Suppose that (2.7) is satisfied for $j = 1,\dots,q$. Let $J \subseteq \{1,\dots,q\}$ be a subset such that $|J| \leq q^*$ and $J_0 \setminus J \neq \emptyset$. Let $l = |J_0 \setminus J|$. If $\mathcal{E}_{\delta,J\cup J_0} \cap \mathcal{A}$ holds, then we have*

$$\|\hat{\Pi}_{J_0} f\|_n^2 - \|\hat{\Pi}_J f\|_n^2 \geq \frac{1}{2}\frac{(1-\delta)^2}{(1+\delta)}(1-\rho_{q*}^2)\kappa_l, \tag{4.2}$$

provided that

$$(2/3)(1-\sqrt{c'})^2 - 8\frac{(1+\delta)}{(1-\delta)^2}c' \geq 1/2. \tag{4.3}$$

A proof of Proposition 13 is given in Section 5.7. In the absence of noise, Proposition 12 and 13 already prove Theorem 10. In fact, if the event $\mathcal{E}_{\delta,q^*} \cap \mathcal{A}$ holds, then (2.6) selects a subset $\hat{J}_0 \subseteq \{1,\dots,q\}$ with $|\hat{J}_0| \leq q^*$ and $J_0 \subseteq \hat{J}_0$. Proposition 13 applies to the nonparametric setting, while Proposition 12 applies if the components f_j satisfy $f_j \in V_j$, the latter being a commonly used setting in the theory of compressive sensing (see, e.g., [**28**] and the references therein).

4.2. End of the proof of Theorem 10. We have

$$\begin{aligned}\mathbb{P}\left(J_0 \not\subseteq \hat{J}_0\right) &= \mathbb{P}\left(J_0 \setminus \hat{J}_0 \neq \emptyset\right) \\ &\leq \mathbb{P}\left(\exists J \subseteq \{1,\dots,q\}, |J| \leq q^* \text{ with } J_0 \setminus J \neq \emptyset\right. \\ &\qquad \left.\text{and } \|\hat{\Pi}_J \mathbf{Y}\|_n^2 - d_J/n \geq \|\hat{\Pi}_{J_0}\mathbf{Y}\|_n^2 - d_{J_0}/n\right).\end{aligned}$$

Applying the union bound, we obtain

$$\mathbb{P}\left(J_0 \nsubseteq \hat{J}_0\right) \leq \mathbb{P}\left(\mathcal{E}^c_{\delta,q^*}\right) + \mathbb{P}\left(\mathcal{A}^c\right) + \sum_{\substack{J\subseteq\{1,\dots,q\} \\ |J|\leq q^*, J_0\setminus J\neq\emptyset}} \mathbb{P}\left(\mathcal{E}_{\delta,J\cup J_0} \cap \mathcal{A} \cap \|\hat{\Pi}_J \mathbf{Y}\|_n^2 - d_J/n \geq \|\hat{\Pi}_{J_0}\mathbf{Y}\|_n^2 - d_{J_0}/n\right),$$

where $\mathcal{A}$ is the event defined in Proposition 13. We have:

LEMMA 16. *Let Assumptions 11 and 12 be satisfied. Suppose that* (2.7) *is satisfied for $j = 1, \dots, q$. Let $J \subseteq \{1, \dots, q\}$ be a subset such that $|J| \leq q^*$ and $J_0 \setminus J \neq \emptyset$. Let $l = |J_0 \setminus J|$. Then there is a constant c_1 depending only on δ (given explicitly in the proof) such that*

$$\mathbb{P}\left(\mathcal{E}_{\delta,J\cup J_0} \cap \mathcal{A} \cap \|\hat{\Pi}_J \mathbf{Y}\|_n^2 - \sigma^2 d_J/n \geq \|\hat{\Pi}_{J_0}\mathbf{Y}\|_n^2 - \sigma^2 d_{J_0}/n\right) \leq 4\exp\left(-c_1 \frac{n^2(1-\rho_{q^*}^2)^2\kappa_l^2}{\sigma^4 d_{q^*-s+l} + \sigma^2 n(1-\rho_{q^*}^2)\kappa_l}\right).$$

A proof of Lemma 16 is given in Section 5.8. Thus

$$\mathbb{P}\left(J_0 \nsubseteq \hat{J}_0\right) \leq \mathbb{P}\left(\mathcal{E}^c_{\delta,q^*}\right) + \mathbb{P}\left(\mathcal{A}^c\right) + \sum_{l=1}^{s}\sum_{m=0}^{q^*-(s-l)} \binom{s}{l}\binom{q-s}{m} 4\exp\left(-c_1 \frac{n^2(1-\rho_{q^*}^2)^2\kappa_l^2}{\sigma^4 d_{q^*-s+l} + \sigma^2 n(1-\rho_{q^*}^2)\kappa_l}\right).$$

Now apply Lemma 14. This completes the proof. □

5. Proofs

5.1. Proof of Remark 11. Suppose that (2.3) holds, and let $h_{J_1} \in H_{J_1}$ and $h_{J_2} \in H_{J_2}$. Then $\|h_{J_1} + h_{J_2}\|^2 \geq \|h_{J_1}\|^2 - 2\rho_{q^*}\|h_{J_1}\|\|h_{J_2}\| + \|h_{J_2}\|^2$ and (2.4) follows from the inequality $2\rho_{q^*}\|h_{J_1}\|\|h_{J_2}\| \leq \rho_{q^*}^2\|h_{J_1}\|^2 + \|h_{J_2}\|^2$.

Conversely, suppose that (2.4) holds, and let $h_{J_1} \in H_{J_1}$ and $h_{J_2} \in H_{J_2}$. We may assume without loss of generality that $h_{J_2} \neq 0$ and that $\|h_{J_2}\| = 1$. Then $\|h_{J_1}\|^2 - \langle h_{J_1}, h_{J_2}\rangle^2 = \|h_{J_1} - \langle h_{J_1}, h_{J_2}\rangle h_{J_2}\|^2 \geq (1-\rho_{q^*}^2)\|h_{J_1}\|^2$ which gives (2.3). This completes the proof. □

5.2. Proof of Lemma 12. Let $J_1, J_2 \subseteq \{1, \dots, q\}$ be two subsets satisfying $J_1 \cap J_2 = \emptyset$ and $|J_1|, |J_2| \leq q^*$. Applying (2.5) and (2.8), we see that

$$\|f_{J_1} + f_{J_2}\|^2 \geq \frac{1-\epsilon_{2q^*}}{1+\epsilon'_{q^*}}\left(\|f_{J_1}\|^2 + \|f_{J_2}\|^2\right)$$

for all $f_{J_1} \in H_{J_1}$, $f_{J_2} \in H_{J_2}$. Thus Remark 11 gives the "if" part. Conversely, applying (2.3) iteratively, one gets for instance

$$1 - \epsilon_{2q^*} \geq (1-\rho_{q^*}^2)^{\log_2 q^*+1}$$

which gives the "only if" part. □

5.3. Proof of Lemma 13. Let $\sum_{k=1}^{\infty}\theta_k\phi_k \in \tilde{W}_j(\alpha_j, K_j)$. Then there is a constant c_{α_j} depending only on α_j such that (see, e.g., [**74**, Proof of Lemma 1.8 and Theorem 1.9])

$$\sum_{k>m_j}\theta_j^2 \le c_{\alpha_j}K_j^2 m_j^{-2\alpha_j} \tag{5.1}$$

and

$$\left(\sum_{k>m_j}|\theta_j|\right)^2 \le c_{\alpha_j}K_j^2 m_j^{1-2\alpha_j}. \tag{5.2}$$

We now define $U_j = V_j + \mathbb{R}$, where $\mathbb{R}$ denotes the constant functions. Since V_j and $\mathbb{R}$ are orthogonal for $j = 1, \ldots, q-1$, and since $V_q = U_q$, we have $\Pi_{V_j}h_j = \Pi_{U_j}h_j$ for $h_j \in H_j$ and $j = 1, \ldots, q$. Now, let $f_j \in H_j \cap \tilde{W}_j(\alpha_j, K_j)$. Then $f_j(x_j) = \sum_{k=1}^{\infty}\theta_k\phi_k(x_j)$. Let $q_j(x_j) = \sum_{j=1}^{m_j}\theta_k\phi_k(x_j)$. Then $q_j - \mathbb{E}[q_j(X_j)] \in V_j$ and thus $q_j \in U_j$. We conclude that

$$\|f_j - \Pi_{V_j}f_j\|^2 = \|f_j - \Pi_{U_j}f_j\|^2 \le \|f_j - q_j\|^2 \le (1/c)\sum_{k>m_j}\theta_j^2$$

and similarly that

$$\begin{aligned}
\|f_j - \Pi_{V_j}f_j\|_\infty^2 &\le 2\|f_j - q_j\|_\infty^2 + 2\|q_j - \Pi_{V_j}f_j\|_\infty^2 \\
&\le 4\left(\sum_{k>m_j}|\theta_j|\right)^2 + 2(1/c)m_j\|q_j - \Pi_{U_j}f_j\|^2 \\
&\le 4\left(\sum_{k>m_j}|\theta_j|\right)^2 + 2(1/c)m_j\|q_j - f_j\|^2 \\
&\le 4\left(\sum_{k>m_j}|\theta_j|\right)^2 + 2(1/c)^2 m_j\sum_{k>m_j}\theta_j^2
\end{aligned} \tag{5.3}$$

Using (5.1)-(5.3), we obtain Lemma 13. This completes the proof. □

5.4. Proof of Lemma 14. Let $v = \sum_{j\in J_0}v_j$ with $v_j = \Pi_{V_j}f_j$ for $j \in J_0$. By (2.9), we have

$$\|f - v\|^2 \le c'(1 - \rho_{q^*}^2)\kappa.$$

Moreover, by Lemma 13 and the Cauchy-Schwarz inequality, we also have

$$\|f - v\|_\infty^2 \le q^*\sum_{j\in J_0}C_jK_j^2 m_j^{1-2\alpha_j} \le \frac{2d_{q^*}c(1-\rho_{q^*}^2)\kappa}{(1+\epsilon'_{q^*})}. \tag{5.4}$$

Thus, letting $x = c'(1-\rho_{q^*}^2)\kappa$, Bennett's inequality (see, e.g., [**52**, Comment after Proposition 2.8]) yields

$$\begin{aligned}\mathbb{P}\left(\|f-v\|_n^2 > 2x\right) &\le \mathbb{P}\left(\|f-v\|_n^2 - \|f-v\|^2 > x\right)\\ &\le \exp\left(-\frac{nx^2}{2\|(f-v)^2\|^2 + (2/3)\|f-v\|_\infty^2 x}\right)\\ &\le \exp\left(-\frac{3nx}{8\|f-v\|_\infty^2}\right).\end{aligned}$$

Using this and (5.4), we obtain (2.10). This completes the proof □

5.5. Proof of Proposition 11. The proof is taken from [**64**, proof of Theorem 3.6] (see also [**59**, proof of Theorem 8.1]). However, we have to modify several details. For $j = 1, \dots, q-1$, the spaces V_j are spanned by the functions $\psi_{jk}(x_j) = \phi_k(x_j) - \mathbb{E}[\phi_k(X_j)]$, $2 \le k \le m_j$, and the space V_q is spanned by $\psi_{qk}(x_q) = \phi_k(x_q)$, $1 \le k \le m_q$. Thus each function in $\sum_{j=1}^q V_j$, can be written as $g_\alpha = \sum_{j,k} \alpha_{jk}\psi_{jk}$, for some $\alpha = (\alpha_1^T, \dots, \alpha_q^T)^T \in \mathbb{R}^{d_q}$. Letting

$$T = \left\{\alpha \in \mathbb{R}^{d_q} : g_\alpha \in V_J, |J| \le 2q^*, \|g_\alpha\| \le 1\right\},$$

we have to show that there is a constant $C_1 > 0$ such that

$$E := \mathbb{E}\left[\sup_{\alpha\in T}\left|\|g_\alpha\|_n^2 - \|g_\alpha\|^2\right|\right] \le C_1\sqrt{\frac{d_{2q*}}{c(1-\epsilon_{2q^*})n}}\log^2(d_q \vee n),$$

provided that the last expression is smaller than 1. Using the symmetrization lemma (see, e.g., [**77**, Lemma 2.3.1]), we obtain

$$E \le 2\mathbb{E}\left[\sup_{\alpha\in T}\frac{1}{n}\sum_{i=1}^n \delta^i g_\alpha^2(X^i)\right],$$

where $\delta^1, \dots, \delta^n$ are independent Rademacher random variables. Applying [**77**, Corollary 2.2.8], we have for a universal constant C_2,

$$E_1 := \mathbb{E}\left[\sup_{\alpha\in T}\frac{1}{n}\sum_{i=1}^n \delta^i g_\alpha^2(X^i)\middle| X^1, \dots, X^n\right] \le C_2\int_0^\infty \sqrt{\log N(T,d,u)}du,$$

where $N(T,d,u)$ denotes the minimal number of balls of radius u in the semimetric d needed to cover T and d is the given by

$$d(\alpha,\beta) = \left(\frac{1}{n^2}\sum_{i=1}^n \left(g_\alpha^2(X^i) - g_\beta^2(X^i)\right)^2\right)^{1/2}.$$

Now,

$$d(\alpha,\beta) \leq \left(\frac{1}{n^2}\sum_{i=1}^{n}\left(g_\alpha(X^i)+g_\beta(X^i)\right)^2\right)^{1/2} \max_{i=1,\dots,n}\left|g_\alpha(X^i)-g_\beta(X^i)\right|$$
$$\leq \frac{2}{\sqrt{n}}\sup_{\alpha\in T}\|g_\alpha\|_n \max_{i=1,\dots,n}\left|g_\alpha(X^i)-g_\beta(X^i)\right|.$$

Applying a linear change of variables, we obtain

$$E_1 \leq \sup_{\alpha\in T}\|g_\alpha\|_n \frac{2C_2}{\sqrt{n}}\int_0^\infty \sqrt{\log N(T,\|\cdot\|_X,u)}du,$$

where the seminorm $\|\cdot\|_X$ is given by

$$\|\alpha\|_X = \max_{i=1,\dots,n}\left|g_\alpha(X^i)\right| = \max_{i=1,\dots,n}\left|\langle\alpha,x_i\rangle\right|.$$

Here, the x_i are the vectors of the basis functions evaluated at X^i. Note that the x_i are uniformly bounded by $K=2\sqrt{2}$ and that the last expression coincides with the definition of $\|\cdot\|_X$ in [**64**]. Now, if $\alpha\in T$, then

$$\|\alpha\|_0 = |\{j:\alpha_j\neq 0\}| \leq d_{2q^*}$$

and

$$\|\alpha\|_2 \leq \frac{1}{\sqrt{c(1-\epsilon_{2q^*})}}.$$

The first inequality follows from the definition, the second one from Assumption 12, (2.5), and $\|g_\alpha\|\leq 1$. Thus

$$T \subseteq \frac{1}{\sqrt{c(1-\epsilon_{2q^*})}} D_2^{d_{2q^*},d_q},$$

where

$$D_2^{d_{2q^*},d_q} = \left\{\alpha\in\mathbb{R}^{d_q} : \|\alpha\|_0\leq d_{2q^*}, \|\alpha\|_2\leq 1\right\}.$$

Applying again a linear change of variables, we obtain

$$E_1 \leq$$
$$\sup_{\alpha\in T}\|g_\alpha\|_n C_2\sqrt{\frac{d_{2q^*}}{c(1-\epsilon_{2q^*})n}}\int_0^\infty \log^{1/2} N\left(\frac{1}{\sqrt{d_{2q^*}}}D_2^{d_{2q^*},d_q},\|\cdot\|_X,u\right)du.$$

The above integral is the same as in [**64**, (3.7)] and can be bounded by $C_3\log(d_{2q^*})\sqrt{\log n}\sqrt{\log d_q}\leq C_3\log^2(d_q\vee n)$ (here, we use that the x_i are uniformly bounded by $2\sqrt{2}$). We conclude that

$$E_1 \leq C(q^*,q,n)\sup_{\alpha\in T}\|g_\alpha\|_n,$$

where

$$C(q^*,q,n) = C_2C_3\sqrt{\frac{d_{2q^*}}{c(1-\epsilon_{2q^*})n}}\log^2(d_q\vee n)$$

Using this and the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned} E &\leq C(q^*, q, n) \left(\mathbb{E} \left[\sup_{\alpha \in T} \|g_\alpha\|_n^2 \right] \right)^{1/2} \\ &\leq C(q^*, q, n) \left(E + 1 \right)^{1/2}. \end{aligned}$$

If

$$C(q^*, q, n) \leq 1,$$

then we get

$$E \leq 2C(q^*, q, n).$$

This completes the proof. □

5.6. Proof of Equation (3.2)**.** By the Cauchy-Schwarz inequality, we have for $g = \sum_{k=1}^m \theta_k \phi_k$,

$$\|g\|_\infty^2 \leq m \sum_{k=1}^m \theta_k^2 = m \int_0^1 g^2(x) dx.$$

This implies that

$$\|g_j\|_\infty^2 \leq (2/c) \dim V_j \, \|g_j\|^2 \tag{5.5}$$

for all $g_j \in V_j$. Now, let $J \subseteq \{1, \ldots, q\}$ be a subset with $|J| \leq 2q^*$. Applying (5.5), the Cauchy-Schwarz inequality, and Lemma 12, we obtain

$$\begin{aligned} \|g_J\|_\infty \leq \sum_{j \in J} \|g_j\|_\infty &\leq \sqrt{2/c} \sqrt{\sum_{j \in J} \dim V_j} \sqrt{\sum_{j \in J} \|g_j\|^2} \\ &\leq \sqrt{\frac{2}{c(1 - \epsilon_{2q^*})}} \sqrt{\dim V_J} \, \|g_J\| \end{aligned}$$

for all $g_J = \sum_{j \in J} g_j \in V_J$. This completes the proof. □

5.7. Proof of Proposition 13. Let $v = \sum_{j \in J_0} v_j$ with $v_j = \Pi_{V_j} f_j$ for $j \in J_0$. We have

$$\begin{aligned} &\|\hat{\Pi}_{J_0} f\|_n^2 - \|\hat{\Pi}_J f\|_n^2 \\ &= \|\hat{\Pi}_{J_0} v\|_n^2 + 2\langle \hat{\Pi}_{J_0} v, \hat{\Pi}_{J_0}(f - v) \rangle_n + \|\hat{\Pi}_{J_0}(f - v)\|_n^2 \\ &- \|\hat{\Pi}_J v\|_n^2 - 2\langle \hat{\Pi}_J v, \hat{\Pi}_J (f - v) \rangle_n - \|\hat{\Pi}_J (f - v)\|_n^2 \\ &\geq \|\hat{\Pi}_{J_0} v\|_n^2 - \|\hat{\Pi}_J v\|_n^2 + 2\langle \hat{\Pi}_{J_0} v - \hat{\Pi}_J v, f - v \rangle_n - \|f - v\|_n^2, \end{aligned} \tag{5.6}$$

where the inequality holds since orthogonal projections are self-adjoint and lower the norm. Since $\hat{\Pi}_{J_0} v = v$ and $\|v - \hat{\Pi}_J v\|_n^2 = \|v\|_n^2 - \|\hat{\Pi}_J v\|_n^2$, we get

$$\begin{aligned} &2\langle \hat{\Pi}_{J_0} v - \hat{\Pi}_J v, f - v \rangle_n \\ &\leq (1/3) \|\hat{\Pi}_{J_0} v - \hat{\Pi}_J v\|_n^2 + 3\|f - v\|_n^2 \\ &= (1/3) \left(\|\hat{\Pi}_{J_0} v\|_n^2 - \|\hat{\Pi}_J v\|_n^2 \right) + 3\|f - v\|_n^2, \end{aligned} \tag{5.7}$$

where we also applied the bound $2xy \le 3x^2 + (1/3)y^2$. Combining (5.6) and (5.7), we conclude that

$$\|\hat{\Pi}_{J_0} f\|_n^2 - \|\hat{\Pi}_J f\|_n^2 \ge (2/3)\left(\|\hat{\Pi}_{J_0} v\|_n^2 - \|\hat{\Pi}_J v\|_n^2\right) - 4\|f - v\|_n^2. \tag{5.8}$$

If $\mathcal{E}_{\delta, J\cup J_0}$ holds, then Proposition 12 says that

$$\left\|\hat{\Pi}_{J_0} v\right\|_n^2 - \left\|\hat{\Pi}_J v\right\|_n^2 \ge \frac{(1-\delta)}{(1+\delta)}(1-\rho_{q^*}^2)\|v_{J_0\setminus J}\|_n^2,$$

where $v_{J_0\setminus J} = \sum_{j\in J_0\setminus J} v_j$. If $\mathcal{E}_{\delta, J_0}$ holds, then

$$\|v_{J_0\setminus J}\|_n^2 \ge (1-\delta)\|v_{J_0\setminus J}\|^2 \ge (1-\delta)\left(\|f_{J_0\setminus J}\| - \|f_{J_0\setminus J} - v_{J_0\setminus J}\|\right)^2,$$

where $f_{J_0\setminus J} = \sum_{j\in J_0\setminus J} f_j$. As in (2.9), we have

$$\|f_{J_0\setminus J} - v_{J_0\setminus J}\|^2 \le c'(1-\rho_{q*}^2)\kappa \le c'\kappa_l.$$

Thus

$$\|v_{J_0\setminus J}\|_n^2 \ge (1-\delta)(1-\sqrt{c'})^2\kappa_l$$

If $\mathcal{E}_{\delta, J\cup J_0}$ holds, then we obtain

$$\left\|\hat{\Pi}_{J_0} v\right\|_n^2 - \left\|\hat{\Pi}_J v\right\|_n^2 \ge (1-\sqrt{c'})^2\frac{(1-\delta)^2}{(1+\delta)}(1-\rho_{q^*}^2)\kappa_l. \tag{5.9}$$

If $\mathcal{E}_{\delta, J\cup J_0} \cap \mathcal{A}$ holds, then we conclude from (5.8) and (5.9) that

$$\|\hat{\Pi}_{J_0} f\|_n^2 - \|\hat{\Pi}_J f\|_n^2 \ge \frac{1}{2}\frac{(1-\delta)^2}{(1+\delta)}(1-\rho_{q*}^2)\kappa_l,$$

provided that (4.3) is satisfied. This completes the proof. □

5.8. Proof of Lemma 16. We have

$$\|\hat{\Pi}_J \mathbf{Y}\|_n^2 - \sigma^2 d_J/n \ge \|\hat{\Pi}_{J_0}\mathbf{Y}\|_n^2 - \sigma^2 d_{J_0}/n$$

if and only if

$$\begin{aligned}&\|\hat{\Pi}_J\boldsymbol{\epsilon}\|_n^2 - \|\hat{\Pi}_{J_0}\boldsymbol{\epsilon}\|_n^2 - \sigma^2 d_J/n + \sigma^2 d_{J_0}/n + 2\langle(\hat{\Pi}_J - \hat{\Pi}_{J_0})f, \boldsymbol{\epsilon}\rangle_n \\ &\ge \|\hat{\Pi}_{J_0} f\|_n^2 - \|\hat{\Pi}_J f\|_n^2.\end{aligned}$$

If $\mathcal{E}_{\delta, J\cup J_0} \cap \mathcal{A}$ holds, then (5.8), (4.3), and (4.2) yield

$$\|\hat{\Pi}_{J_0} f\|_n^2 - \|\hat{\Pi}_J f\|_n^2 \ge \frac{1}{2}\frac{(1-\delta)^2}{(1+\delta)}(1-\rho_{q^*}^2)\kappa_l$$

and also

$$\|\hat{\Pi}_{J_0} f\|_n^2 - \|\hat{\Pi}_J f\|_n^2 \ge \frac{1}{2(1-\sqrt{c'})^2}\|v - \hat{\Pi}_J v\|_n^2 \ge \frac{1}{2}\|v - \hat{\Pi}_J v\|_n^2.$$

Recall that the random variables $\epsilon^1, \ldots, \epsilon^n$ are independent and Gaussian, each with expectation 0 and variance σ^2. Moreover, they are independent of $X^1, \ldots, X^n$. One can show that, conditioned on $X^1, \ldots, X^n$ and if $\mathcal{E}_{\delta, J\cup J_0}$ holds, we have

$$\|\hat{\Pi}_J\boldsymbol{\epsilon}\|_n^2 - \|\hat{\Pi}_{J_0}\boldsymbol{\epsilon}\|_n^2 \stackrel{d}{=} (\sigma^2/n)\chi^2(d_{J\setminus J_0}) - (\sigma^2/n)\chi^2(d_{J_0\setminus J}),$$

where $\stackrel{d}{=}$ denotes equality in distribution, and where $\chi^2(d_{J\setminus J_0})$ and $\chi^2(d_{J_0\setminus J})$ are chi-square distributed random variables with $d_{J\setminus J_0}$ and $d_{J_0\setminus J}$ degrees of freedom, respectively. Applying all these arguments and the union bound, we conclude that

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{E}_{\delta,J\cup J_0}\cap\mathcal{A}\cap\|\hat{\Pi}_J\mathbf{Y}\|_n^2-\sigma^2 d_J/n\geq\|\hat{\Pi}_{J_0}\mathbf{Y}\|_n^2-\sigma^2 d_{J_0}/n\right)\\
&\quad\leq\mathbb{P}\left(\frac{\sigma^2}{n}\left(\chi^2(d_{J\setminus J_0})-d_{J\setminus J_0}\right)\geq\frac{1}{8}\frac{(1-\delta)^2}{(1+\delta)}(1-\rho_{q^*}^2)\kappa_l\right)\\
&\quad+\mathbb{P}\left(\frac{\sigma^2}{n}\left(\chi^2(d_{J_0\setminus J})-d_{J_0\setminus J}\right)\leq-\frac{1}{8}\frac{(1-\delta)^2}{(1+\delta)}(1-\rho_{q^*}^2)\kappa_l\right)\\
&\quad+\mathbb{P}\left(\mathcal{E}_{\delta,J\cup J_0}\cap\mathcal{A}\cap 2\langle(\hat{\Pi}_J-\hat{\Pi}_{J_0})f,\boldsymbol{\epsilon}\rangle_n\geq\frac{1}{4}\|v-\hat{\Pi}_J v\|_n^2\right).
\end{aligned}$$

The first and the second term can be bounded by standard concentration inequalities for chi-square distributions.

LEMMA 17. *Let d be a positive integer. Then, for all $x\geq 0$, we have*

$$\mathbb{P}\left(\chi^2(d)-d\geq x\right)\leq\exp\left(-\frac{x^2}{2(2d+2x)}\right)$$

and

$$\mathbb{P}\left(\chi^2(d)-d\leq -x\right)\leq\exp\left(-\frac{x^2}{4d}\right).$$

For a proof of this lemma see [**47**, Lemma 1] and [**11**, Lemma 8]. Since $|J_0\setminus J|=l$ and $|J\setminus J_0|\leq q^*-s+l$, we have $d_{J_0\setminus J},d_{J\setminus J_0}\leq d_{q^*-s+l}$. Applying this and Lemma 17, we obtain

$$\begin{aligned}
&\mathbb{P}\left(\frac{\sigma^2}{n}\left(\chi^2(d_{J\setminus J_0})-d_{J\setminus J_0}\right)\geq\frac{1}{8}\frac{(1-\delta)^2}{(1+\delta)}(1-\rho_{q^*}^2)\kappa_l\right)\\
&\quad+\mathbb{P}\left(\frac{\sigma^2}{n}\left(\chi^2(d_{J_0\setminus J})-d_{J_0\setminus J}\right)\leq-\frac{1}{8}\frac{(1-\delta)^2}{(1+\delta)}(1-\rho_{q^*}^2)\kappa_l\right)\\
&\quad\leq 2\exp\left(-\frac{1}{32}\frac{c_\delta^2 n^2(1-\rho_{q^*}^2)^2\kappa_l^2}{8\sigma^4 d_{q^*-s+l}+c_\delta\sigma^2 n(1-\rho_{q^*}^2)\kappa_l}\right),
\end{aligned}\tag{5.10}$$

where $c_\delta=(1-\delta)^2/(1+\delta)$. Thus it remains the third term. It can be bounded by

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{E}_{\delta,J\cup J_0}\cap\langle\hat{\Pi}_J v-v,\boldsymbol{\epsilon}\rangle_n\geq\frac{1}{16}\|v-\hat{\Pi}_J v\|_n^2\right)\\
+&\mathbb{P}\left(\mathcal{E}_{\delta,J\cup J_0}\cap\mathcal{A}\cap\langle(\hat{\Pi}_J-\hat{\Pi}_{J_0})(f-v),\boldsymbol{\epsilon}\rangle_n\geq\frac{1}{16}\|v-\hat{\Pi}_J v\|_n^2\right).
\end{aligned}\tag{5.11}$$

These terms can be bounded by standard concentration inequalities for Gaussian random variables. Applying (5.9), we obtain

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{E}_{\delta,J\cup J_0} \cap \langle \hat{\Pi}_J v - v, \boldsymbol{\epsilon}\rangle_n \geq \frac{1}{16}\|v - \hat{\Pi}_J v\|_n^2\right) \\
&\leq \mathbb{E}\left[1_{\mathcal{E}_{\delta,J\cup J_0}} \exp\left(-\frac{n}{2^9}\frac{\|v - \hat{\Pi}_J v\|_n^2}{\sigma^2}\right)\right] \\
&\leq \exp\left(-\frac{c_\delta}{2^{10}}\frac{n(1-\rho_{q^*}^2)\kappa_l}{\sigma^2}\right),
\end{aligned}$$

which bounds the first term in (5.11). If $\mathcal{A}$ holds, then

$$\|(\hat{\Pi}_J - \hat{\Pi}_{J_0})(f - v)\|_n^2 \leq 4\|f - v\|_n^2 \leq 8c'(1-\rho_{q^*}^2)\kappa \leq 8c'(1-\rho_{q^*}^2)\kappa_l.$$

Applying this and (5.9), we obtain

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{E}_{\delta,J\cup J_0} \cap \mathcal{A} \cap \langle (\hat{\Pi}_J - \hat{\Pi}_{J_0})(f - v), \boldsymbol{\epsilon}\rangle_n \geq \frac{1}{16}\|v - \hat{\Pi}_J v\|_n^2\right) \\
&\leq \mathbb{P}\left(\mathcal{E}_{\delta,J\cup J_0} \cap \mathcal{A} \cap \langle (\hat{\Pi}_J - \hat{\Pi}_{J_0})(f - v), \boldsymbol{\epsilon}\rangle_n \geq \frac{1}{32}\frac{(1-\delta)^2}{(1+\delta)}(1-\rho_{q^*}^2)\kappa_l\right) \\
&\leq \exp\left(-\frac{c_\delta^2}{2^{14}c'}\frac{n(1-\rho_{q^*}^2)\kappa_l}{\sigma^2}\right)
\end{aligned}$$

which bounds the second term in (5.11). This completes the proof. □

CHAPTER 4

Optimal conditions for variable selection in additive Gaussian white noise models

1. Introduction

In this chapter, we study the problem of variable selection in high-dimensional Gaussian white noise models. We suppose that the regression function has an additive form and that its additive components belong to nonparametric classes of functions. The aim is to derive optimal conditions under which consistent variable selection is possible.

In order to obtain sufficient conditions, we analyze a penalized least squares criterion. In contrast to Chapter 3, we do not assume that an upper bound q^* for the number of non-zero components is known. This forces us to introduce an additional penalty term. Our main result is a general exponential bound for the probability that our procedure recovers the support, i.e., the set of indices corresponding to the non-zero components. In the case that the components belong to Sobolev classes, we establish conditions making consistent estimation of the support possible. Finally, we prove minimax lower bounds showing that these conditions are also optimal.

High-dimensional additive models have been recently studied in a series of papers by Meier, van de Geer, and Bühlmann [**53**], Huang, Horowitz, and Wei [**37**], Koltchinskii and Yuan [**44**], Raskutti, Wainwright, and Yu [**58**], Gayraud and Ingster [**30**], Suzuki and Sugiyama [**71**], and Dalalyan, Ingster, and Tsybakov [**19**]. Most of these papers focus on the problem of estimation. The latter, e.g., constructs an estimator achieving optimal minimax rates of convergence. In this chapter, we analyze a penalized least squares criterion similar to those developed in the work by Birgé and Massart [**12, 13**] (see also [**10, 6**]). Since we are dealing with nonparametric components and since we focus on the problem of variable selection instead of estimation, we introduce a slightly different penalty term. This chapter can also be seen as a complement to Chapter 3. By switching to the Gaussian white noise framework, the constructions, statements, and proofs become simpler and more transparent. In addition, we also derive minimax lower bounds. The proofs of these lower bounds use standard tools, such as Fano's lemma and the method of several fuzzy hypothesis (see, e.g., the book by Tsybakov [**74**]), and are based on proofs of lower bounds obtained previously in the linear model (see the work by Wainwright [**80**] and Arias-Castro and Lounici

[**2**]) and in the single atom model (see the work by Comminges and Dalalyan [**18**]).

2. The main result

2.1. The Gaussian white noise framework. We suppose that we observe a process $(Y_\epsilon(h) : h \in L^2([0,1]^q))$ such that

$$Y_\epsilon(h) = \langle f, h \rangle + \epsilon W(h),$$

where f is an unknown function in $L^2([0,1]^q)$, $\epsilon > 0$ is a known parameter, and $(W(h) : h \in L^2([0,1]^q))$ is a centered Gaussian process with covariance given by

$$\mathbb{E}\left[W(h)W(h')\right] = \left\langle h, h' \right\rangle \tag{2.1}$$

(see, e.g., the book by Massart [**52**, Chapter 3.5] and the references therein). Here, $\langle \cdot, \cdot \rangle$ denotes the inner product on the Hilbert space $L^2([0,1]^q)$. Moreover, we suppose that f has the form

$$f(x_1, \ldots, x_q) = \sum_{j=1}^{q} f_j(x_j),$$

where the f_j satisfy $\int_0^1 f_j(x_j)dx_j = 0$. We denote by J_0 the set of non-zero functions in this sum, i.e.,

$$J_0 = \{j \in \{1, \ldots, q\} : \|f_j\| > 0\},$$

where $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$. This implies that f can also be written as

$$f(x_1, \ldots, x_q) = \sum_{j \in J_0} f_j(x_j).$$

In this chapter, we consider the problem of estimating J_0. We are interested in settings where q is very large, for instance in the sense that $q\epsilon^2$ is still very large. Note that ϵ^2 plays the same role as σ^2/n in the regression framework.

2.2. Model selection via penalization. We begin with introducing some notation. For $j = 1, \ldots, q$, let $L_j^2([0,1]) \subseteq L^2([0,1]^q)$ be those functions which depend only on the variable x_j, and let

$$H_j = \left\{ h_j \in L_j^2([0,1]) : \int_0^1 h_j(x_j)dx_j = 0 \right\}.$$

For $j = 1, \ldots, q$, let $V_j \subseteq H_j$ be finite-dimensional linear subspaces, and for $J \subseteq \{1, \ldots, q\}$, let

$$V_J = \sum_{j \in J} V_j$$

and $d_J = \dim V_J$. Moreover, we define $d_l = \max_{|J|=l} d_J$, $l = 1, \ldots, q$. For simplicity, we suppose that the V_j all have the same dimension. This implies that $d_J = |J|d_1$ and also that $d_l = ld_1$. Note that the spaces V_J will be the models used in the selection criterion below. In this and in the next section, we let these spaces unspecified. We only assume that the

f_j can be approximated well (see Theorem 15 below) by some function in V_j. Concrete examples can be found in Section 3.1. We denote by Π_{V_J} the orthogonal projection from $L^2([0,1]^q)$ to V_J. Recall that Π_{V_J} satisfies $\langle \Pi_{V_J} h, \Pi_{V_J} h' \rangle = \langle \Pi_{V_J} h, h' \rangle$ for all $h, h' \in L^2([0,1]^q)$, which is one of its key properties. If $\{\phi_{jk}\}_{1 \le k \le d_1}$ are orthonormal bases of the V_j, then Π_{V_J} is given by

$$\Pi_{V_J} h = \sum_{j \in J} \sum_{k=1}^{d_1} \langle h, \phi_{jk} \rangle \, \phi_{jk}$$

for all $h \in L^2([0,1]^q)$. This representation shows that we can also apply Π_{V_J} to W and Y by letting

$$\Pi_{V_J} W = \sum_{j \in J} \sum_{k=1}^{d_1} W(\phi_{jk}) \phi_{jk} \tag{2.2}$$

and

$$\Pi_{V_J} Y = \Pi_{V_J} f + \epsilon \Pi_{V_J} W.$$

Note that we have

$$\langle \Pi_{V_J} W, \Pi_{V_J} h \rangle = W(\Pi_{V_J} h)$$

for all $h \in L^2([0,1]^q)$, which can be compared to the above key property.

We now turn to the construction of the variable selection criterion. We will consider the following penalized least squares procedure (see, e.g., [**10, 6, 12, 13**], or the book by Massart [**52**])

$$\hat{J}_0 = \underset{J \subseteq \{1,\dots,q\}}{\arg\max} \left(\|\Pi_{V_J} Y\|^2 - \epsilon^2 d_J - \epsilon^2 p(J) \right) \tag{2.3}$$

in the special case that

$$p(J) = p(|J|) = 2\sqrt{c|J| d_{|J|} \log q} + 2c|J| \log q,$$

where $c > 1$ is a constant. Letting $k = |J|$, we have

$$p(k) = 2k \left(\sqrt{c d_1 \log q} + c \log q \right). \tag{2.4}$$

The choice of the penalty function has the following two explanations. The first term $\epsilon^2 d_J$ is equal to the expectation of $\|\Pi_{V_J} \epsilon W\|^2$ and is there for centering. The second term $\epsilon^2 p(|J|)$ is chosen such that the procedure will not choose large values of $k = |J|$. Its particular form is based on the following exponential inequality for chi-square random variables

$$\mathbb{P}\left(\chi^2(d_k) - d_k > p(k) \right) \le \exp(-ck \log q), \tag{2.5}$$

where $\chi^2(d_k)$ denotes a chi-square random variable with d_k degrees of freedom (see Lemma 18). We mention that in the special case $d_1 = 1$, the whole penalty is equal to

$$\epsilon^2 k (1 + 2\sqrt{c \log q} + 2c \log q),$$

which has the form considered in [**13**] (see, e.g., [**13**, Equation (26)]). Finally, note that if $\epsilon > 0$ (which we assume in this chapter), then the solution of

(2.3) is almost surely unique, since the probability that the expression inside the argmax takes the same value for two different sets is equal to 0. Here, one uses that the difference of two independent non central chi-square random variables is absolute continuous with respect to the Lebesgue measure.

2.3. A general variable selection theorem. In this section, we state our main theorem, which roughly speaking asserts that variable selection is possible if $\|\Pi_{V_j} f_j\|^2$ is greater than a constant (strictly larger than 1) times $\epsilon^2 p(1)$, for $j = 1, \dots, q$. Using the quantity

$$\kappa_1 = \min_{j \in J_0} \|f_j\|^2,$$

we will derive this condition from two more concrete ones. Note that applications can be found in the next section.

THEOREM 15. *Let $c > 8$. Suppose that the following two assumptions hold:*

(i) *For $j = 1, \dots, q$, we have*

$$\|f_j - \Pi_{V_j} f_j\|^2 \leq \kappa_1/2.$$

(ii) *We have*

$$\epsilon^2 p(1) \leq \kappa_1/8.$$

Then

$$\mathbb{P}\left(\hat{J}_0 \neq J_0\right) \leq 9q^{-(c/4-2)}.$$

3. Optimal Conditions

3.1. Sufficient conditions. In this section, we want to apply Theorem 15 in a parametric and in a nonparametric setting. The parametric one is as follows. Suppose that the f_j belong to known finite-dimensional linear subspaces V_j of H_j, and that these spaces are also chosen in the selection criterion (2.3). In this case, Theorem 15 has the following corollary. Note that, since $\Pi_{V_j} f_j = f_j$, the approximation condition (i) disappears.

COROLLARY 16. *Suppose that the f_j satisfy $f_j \in V_j$. Let $c > 8$. Moreover, suppose that the following two conditions hold:*

(i)

$$32\epsilon^2 c \log q \leq \kappa_1,$$

(ii)

$$32\epsilon^2 \sqrt{c d_1 \log q} \leq \kappa_1.$$

Then

$$\mathbb{P}\left(\hat{J}_0 \neq J_0\right) \leq 9q^{-(c/4-2)}.$$

Next, we consider the nonparametric case. Now, the success of the criterion depends on a suitable choice of the V_j, which in turn depends on the regularity assumptions on the f_j. In the following, we shall restrict our attention to the case that the coefficients of the f_j with respect to some orthonormal basis belong to some Sobolev ellipsoid. For $j = 1, \ldots, q$, let $\{\phi_{jk}\}_{k\geq 1}$ be orthonormal bases of the H_j. Moreover, let $\Sigma_j(\alpha, K)$ be the Sobolev class of functions defined by

$$\Sigma_j(\alpha, K) = \left\{ h_j \in L_j^2([0,1]) : \sum_{k=1}^{\infty} k^{2\alpha} \langle h_j, \phi_{jk} \rangle^2 \leq K \right\}, \tag{3.1}$$

where $\alpha > 0$ and $K > 0$ are real numbers (see, e.g., the book by Tsybakov [**74**]). Specifically, we can consider trigonometric bases (by omitting the constant function). We now make the following assumption:

ASSUMPTION 13. *There are $\alpha > 0$ and $K > 0$ such that that $f_j \in H_j \cap \Sigma_j(\alpha, K)$ for $j = 1, \ldots, q$.*

For $j = 1, \ldots, q$, let V_j be the linear span of the first d_1 basis functions $\phi_{j1}, \ldots, \phi_{jd_1}$. Using (3.1), we have

$$\|h_j - \Pi_{V_j} h_j\|^2 \leq K(1 + d_1)^{-2\alpha}$$

for each $h_j \in \Sigma_j(\alpha, K)$. In particular, if

$$1 + d_1 \geq \left(\frac{2K}{\kappa_1} \right)^{1/(2\alpha)}, \tag{3.2}$$

then we have

$$\|h_j - \Pi_{V_j} h_j\|^2 \leq \kappa_1/2 \tag{3.3}$$

for each $h_j \in \Sigma_j(\alpha, K)$, meaning that Assumption (i) of Theorem 15 holds. From Theorem 15, we now obtain:

COROLLARY 17. *Let Assumption 13 be satisfied. Let the V_j be chosen as above, where d_1 is the smallest integer such that* (3.2) *is satisfied. Moreover, suppose that the following two conditions hold:*

(i)
$$32\epsilon^2 c \log q \leq \kappa_1,$$

(ii)
$$32\epsilon^2 \sqrt{c \log q} \leq \kappa_1^{(4\alpha+1)/4\alpha} (2K)^{-1/(4\alpha)}.$$

Then

$$\mathbb{P}\left(\hat{J}_0 \neq J_0 \right) \leq 9q^{-(c/4-2)}.$$

PROOF. Recall that Assumption (i) of Theorem 15 is satisfied by the choice of d_1 and (3.3). Thus it remains to verify that Assumption (ii) of Theorem 15 is also satisfied. By the definition of d_1, we have

$$d_1 \leq \left(\frac{2K}{\kappa_1} \right)^{1/(2\alpha)}.$$

Using this and the Assumptions of the corollary, we conclude that

$$p(1) = 2\epsilon^2 \left(\sqrt{cd_1 \log q} + c \log q \right) \leq \kappa_1/16 + \kappa_1/16 = \kappa/8,$$

which is Assumption (ii) of Theorem 15. This completes the proof. □

3.2. Necessary conditions. In this section, we complete Corollaries 16 and 17 by showing that the conditions are (up to constants) also necessary, thus optimal.

We define

$$\Sigma_j(\alpha, K, \kappa_1) = \left\{ f_j \in \Sigma_j(\alpha, K) : \|f_j\|^2 \geq \kappa_1 \right\}$$

and

$$\Sigma(\alpha, K, \kappa_1, s) = \left\{ f = \sum_{j \in J} f_j : f_j \in \Sigma_j(\alpha, K, \kappa_1), |J| = s \right\}.$$

We prove:

PROPOSITION 14. *Let $s \neq 0$, $s \leq q-3$, and $q \geq 5$. If*

$$\epsilon^2 \max\left\{ \log(q-s+1), \log(s+1) \right\} \geq 4\kappa_1$$

or if

$$\epsilon^2 \sqrt{\log(q-s+1)} \geq 8\kappa_1^{(4\alpha+1)/4\alpha} K^{-1/(4\alpha)},$$

then we have

$$\inf_{\hat{J}} \sup_{f \in \Sigma(\alpha, K, \kappa_1, s)} \mathbb{P}_f \left(\hat{J} \neq J_f \right) \geq 1/4.$$

REMARK 16. Applying the bound $\max\{\log(q-s+1), \log(s+1)\} \geq \log(q/2)$, the first condition can also be replaced by $\epsilon^2 \log(q/2) \geq 4\kappa_1$.

The proof of this proposition is given in Section 4.2, and uses standard techniques, namely Fano's lemma and the method of several fuzzy hypothesis (see, e.g., [**74**]).

A similar result also holds in the setting considered in Corollary 16. In this case, we define

$$V_j(\kappa_1) = \left\{ g_j \in V_j : \|g_j\|^2 \geq \kappa_1 \right\}$$

and

$$V(\kappa_1, s) = \left\{ f = \sum_{j \in J} f_j : f_j \in V_j(\kappa_1), |J| = s \right\}.$$

Then we have (see Remark 17):

PROPOSITION 15. *Let $s \neq 0$, $s \leq q-3$, and $q \geq 5$. If*

$$\epsilon^2 \max\left\{ \log(q-s+1), \log(s+1) \right\} \geq 4\kappa_1$$

or if

$$\epsilon^2 \sqrt{d_1 \log(q-s+1)} \geq 8\kappa_1,$$

then we have

$$\inf_{\hat{J}} \sup_{f \in V(\kappa_1, s)} \mathbb{P}_f \left(\hat{J} \neq J_f \right) \geq 1/4.$$

4. Proofs

4.1. Proof of Theorem 15. Before we begin with the proof of Theorem 15, we present some known exponential inequalities for Gaussian and chi-square random variables which will be used in the proof. Moreover, we derive a consequence of Assumption (i) and (ii) of Theorem 15.

From (2.1), we have that the random variables $W(h)$, $h \in L^2([0,1]^q)$, are Gaussian, with expectation 0 and variance $\|h\|^2$. Hence,

$$\mathbb{P}\left(W(h) \geq x\right) \leq \exp\left(-\frac{x^2}{2\|h\|^2}\right) \tag{4.1}$$

(see, e.g., [**52**, Chapter 2]). Moreover, from (2.2), we have

$$\|\Pi_{V_J} W\|^2 = \sum_{j \in J} \sum_{k=1}^{d_1} (W(\phi_{jk}))^2,$$

which is the sum of the squares of d_J independent standard Gaussian random variables, thus a chi-square random variable with d_J degrees of freedom. In the proof, we will make use of the following lemma.

LEMMA 18. *Let d be a positive integer, and let $\chi^2(d)$ be a chi-square random variable with d degrees of freedom. Then, for all $x \geq 0$, we have*

$$\mathbb{P}\left(\chi^2(d) - d \geq 2\sqrt{dx} + 2x\right) \leq \exp\left(-x\right)$$

and

$$\mathbb{P}\left(\chi^2(d) - d \leq -2\sqrt{dx}\right) \leq \exp\left(-x\right).$$

Moreover, let d' be another positive integer, and let $\chi^2(d')$ be a chi-square random variable independent of $\chi^2(d)$. Then, for all $x \geq 0$, we have

$$\mathbb{P}\left(\chi^2(d) - d - (\chi^2(d') - d') \geq 2\sqrt{(d+d')x} + 2x\right) \leq \exp\left(-x\right).$$

The first two inequalities of Lemma 18 follow from [**47**, Lemma 1]. The third one follows by the same arguments. Let $Z = \chi^2(d) - d - (\chi^2(d') - d')$. Then, for $0 < u < 1/2$,

$$\log\left(\mathbb{E}\left[\exp(uZ)\right]\right) \leq \frac{du^2}{1-2u} + d'u^2 \leq \frac{(d+d')u^2}{1-2u}.$$

Now, proceed as in [**47**] or as in the proof of Proposition 2.9 in [**52**].

In the proof of Theorem 15, we will apply Assumption (i) and (ii) of Theorem 15 in the following form:

LEMMA 19. *Let $J' \subseteq J_0$ be a subset. Then Assumption (i) and (ii) of Theorem 15 imply that*

$$\|\Pi_{V_{J'}} f\|^2 \geq 4\epsilon^2 p(|J'|) \tag{4.2}$$

PROOF. Applying (i), we obtain

$$\begin{aligned}\|\Pi_{V_{J'}} f\|^2 &= \sum_{j\in J'} \|\Pi_{V_j} f_j\|^2 \\ &= \sum_{j\in J'} \left(\|f_j\|^2 - \|f_j - \Pi_{V_j} f_j\|^2\right) \\ &\geq |J'|\frac{\kappa_1}{2}.\end{aligned}$$

Combining this with (ii), we get

$$4\epsilon^2 p(|J'|) = 4\epsilon^2 |J'| p(1) \leq |J'|\frac{\kappa_1}{2} \leq \|\Pi_{V_{J'}} f\|^2.$$

This completes the proof. □

PROOF OF THEOREM 15. Let

$$P_{J,J_0} = \mathbb{P}\left(\|\Pi_{V_J} Y\|^2 - \epsilon^2 d_J - \epsilon^2 p(J) > \|\Pi_{V_{J_0}} Y\|^2 - \epsilon^2 d_{J_0} - \epsilon^2 p(J_0)\right).$$

We have

$$\|\Pi_{V_J} Y\|^2 - \epsilon^2 d_J - \epsilon^2 p(J) > \|\Pi_{V_{J_0}} Y\|^2 - \epsilon^2 d_{J_0} - \epsilon^2 p(J_0) \tag{4.3}$$

if and only if

$$\begin{aligned}\epsilon^2(\|\Pi_{V_{J\setminus J_0}} W\|^2 - d_{J\setminus J_0}) - \epsilon^2(\|\Pi_{V_{J_0\setminus J}} W\|^2 - d_{J_0\setminus J}) + 2\epsilon W(\Pi_{V_{J_0\setminus J}} f) \\ > \|\Pi_{V_{J_0\setminus J}} f\|^2 + \epsilon^2(p(J) - p(J_0)).\end{aligned}$$

Let $l = |J_0 \setminus J|$ and $m = |J \setminus J_0|$. Then (4.3) is equivalent to

$$\begin{aligned}\epsilon^2(\|\Pi_{V_{J\setminus J_0}} W\|^2 - d_m) - \epsilon^2(\|\Pi_{V_{J_0\setminus J}} W\|^2 - d_l) + 2\epsilon W(\Pi_{V_{J_0\setminus J}} f) \\ > \|\Pi_{V_{J_0\setminus J}} f\|^2 + \epsilon^2(p(m) - p(l)).\end{aligned} \tag{4.4}$$

By the union bound, we have

$$\begin{aligned}\mathbb{P}(\hat{J}_0 \neq J_0) &\leq \sum_{J\neq J_0} P_{J,J_0} \\ &= \sum_{|J|<|J_0|} P_{J,J_0} + \sum_{|J|>|J_0|, J_0\subset J} P_{J,J_0} + \sum_{|J|\geq|J_0|, J_0\not\subseteq J} P_{J,J_0} \\ &=: S_1 + S_2 + S_3.\end{aligned}$$

Step 1. We first consider S_1, i.e., the sum over subsets J satisfying $|J| < |J_0|$. For such a set, we have $l > m$. By (4.2), we have

$$\|\Pi_{V_{J_0\setminus J}} f\|^2/2 - \epsilon^2 p(l) + \epsilon^2 p(m) \geq \epsilon^2 p(l) + \epsilon^2 p(m) = \epsilon^2 p(l+m). \tag{4.5}$$

Using (4.4), (4.5), and the union bound, we obtain

$$\begin{aligned}
&P_{J,J_0} \\
&\le \mathbb{P}\left(\epsilon^2(\|\Pi_{V_{J\setminus J_0}}W\|^2 - d_m) - \epsilon^2(\|\Pi_{V_{J_0\setminus J}}W\|^2 - d_l) > \epsilon^2 p(l+m)\right) \\
&+ \mathbb{P}\left(2\epsilon W(\Pi_{V_{J_0\setminus J}}f) > \|\Pi_{V_{J_0\setminus J}}f\|^2/2\right).
\end{aligned}$$

Thus Lemma 18 (see also (2.5)) and (4.1) imply

$$P_{J,J_0} \le \exp\left(-c(l+m)\log q\right) + \exp\left(-\frac{\|\Pi_{V_{J_0\setminus J}}f\|^2}{32\epsilon^2}\right).$$

Using the bound $\|\Pi_{V_{J_0\setminus J}}f\|^2 \ge 4\epsilon^2 p(l) \ge 8\epsilon^2 cl\log q$, we get

$$P_{J,J_0} \le q^{-c(l+m)} + q^{-(c/4)l} \le 2q^{-(c/4)l}.$$

We conclude that

$$\begin{aligned}
S_1 &\le \sum_{l=1}^{s}\sum_{m=0}^{l-1}\binom{s}{l}\binom{q-s}{m}2q^{-(c/4)l} \\
&\le \sum_{l=1}^{s}\sum_{m=0}^{l}\binom{s}{l}\binom{q-s}{m}2q^{-(c/4)l} \\
&\le \sum_{l=1}^{s}\frac{s^l}{l!}\left(\frac{e(q-s)}{l}\right)^l 2q^{-(c/4)l} \\
&\le \sum_{l=1}^{s}\frac{2}{l!}q^{-(c/4-2)l} \\
&\le 2(e-1)q^{-(c/4-2)},
\end{aligned}$$

where we used the following combinatorial result (for a proof see, e.g., [**52**, Proposition 2.5])

$$\sum_{m=0}^{l}\binom{q}{m} \le \left(\frac{eq}{l}\right)^l \tag{4.6}$$

and the inequality $s(q-s)e \le q^2 e/4 \le q^2$.

Step 2. Second, we consider S_2, i.e., the sum over subsets J satisfying $J_0 \subset J$ and $|J| > |J_0|$. For such a set, we have $m = |J \setminus J_0| \ge 1$ and $l = |J_0 \setminus J| = 0$. Thus (4.4) and Lemma 18 (see also (2.5)) yield

$$P_{J,J_0} \le \mathbb{P}\left(\|\Pi_{V_{J\setminus J_0}}W\|^2 - d_m > p(m)\right) \le \exp\left(-cm\log q\right)$$

We conclude that

$$\begin{aligned} S_2 &\le \sum_{m=1}^{q-s} \binom{q-s}{m} q^{-cm} \\ &\le \sum_{m=1}^{q-s} \frac{1}{m!} q^{-(c-1)m} \\ &\le (e-1) q^{-(c-1)}. \end{aligned}$$

Step 3. Finally, we consider S_3, i.e., the sum over subsets J satisfying $J_0 \nsubseteq J$ and $|J| \ge |J_0|$. For such a set, we have $m \ge l \ge 1$. Similarly as in Step 1, we have

$$\|\Pi_{V_{J_0 \setminus J}} f\|^2/2 - \epsilon^2 p(l) + \epsilon^2 p(m)/2 \ge \epsilon^2 p(l+m)/2.$$

Using this, (4.4), and the union bound, we obtain

$$\begin{aligned} &P_{J,J_0} \\ &\le \mathbb{P}\left(\epsilon^2(\|\Pi_{V_{J\setminus J_0}} W\|^2 - d_m) - \epsilon^2(\|\Pi_{V_{J_0\setminus J}} W\|^2 - d_l) > \epsilon^2 p(l+m)/2\right) \\ &+ \mathbb{P}\left(2\epsilon W(\Pi_{V_{J_0\setminus J}} f) > \left(\|\Pi_{V_{J_0\setminus J}} f\|^2 + \epsilon^2 p(m)\right)/2\right). \end{aligned}$$

Thus Lemma 18 and (4.1) imply

$$\begin{aligned} &P_{J,J_0} \le \\ &\exp\left(-(c/4)(l+m)\log q\right) + \exp\left(-\frac{\left(\epsilon^2 p(m) + \|\Pi_{V_{J_0\setminus J}} f\|^2\right)^2}{32\epsilon^2 \|\Pi_{V_{J_0\setminus J}} f\|^2}\right). \end{aligned}$$

The last expression can be bounded by

$$\exp\left(-\frac{1}{32}\inf_{y\ge 0}(yp(m) + 1/y)^2\right) \le \exp\left(-\frac{1}{8}p(m)\right).$$

Using $p(m) \ge 2cm \log q$, we get

$$P_{J,J_0} \le q^{-(c/4)(l+m)} + q^{-(c/4)m} \le 2q^{-(c/4)m}.$$

By proceeding as in Step 1, we conclude that

$$\begin{aligned} S_1 &\le \sum_{m=1}^{q-s} \sum_{l=1}^{m} \binom{s}{l}\binom{q-s}{m} 2q^{-(c/4)m} \\ &\le \sum_{m=1}^{q-s} \frac{(q-s)^m}{m!} \left(\frac{es}{m}\right)^m 2q^{-(c/4)m} \\ &\le \sum_{m=1}^{q-s} \frac{2}{m!} q^{-(c/4-2)m}. \\ &\le 2(e-1)q^{-(c/4-2)} \end{aligned}$$

Finally, from Steps 1-3, we conclude that

$$\mathbb{P}(\hat{J}_0 \neq J_0) \leq (2(e-1) + (e-1) + 2(e-1))\, q^{-(c/4-2)} \leq 9q^{-2(c/8-1)}.$$

This completes the proof. □

4.2. Proof of Proposition 14. The first part of Proposition 14 follows from:

PROPOSITION 16. *Let $s \neq 0$, $s \neq q$, and $q \geq 3$. Then*

$$\inf_{\hat{J}} \sup_{f \in \Sigma(\alpha,K,\kappa_1,s)} \mathbb{P}_f\left(\hat{J} \neq J_f\right) \geq 1 - \frac{\kappa_1/\epsilon^2 + \log 2}{\max\{\log(q-s+1), \log(s+1)\}}.$$

PROOF. The proof is based on Fano's Lemma (see, e.g., [**74**] and also [**80**, proof of Theorem 2]). We have

$$\begin{aligned} &\inf_{\hat{J}} \sup_{f \in \Sigma} \mathbb{P}_f\left(\hat{J} \neq J_f\right) \\ &\geq \inf_{\hat{J}} \frac{1}{M} \sum_{l=1}^{M} \mathbb{P}_{f_l}\left(\hat{J} \neq J_{f_l}\right), \end{aligned}$$

where $f_l \in \Sigma = \Sigma(\alpha, K, \kappa_1, s)$ have different support sets J_{f_l}. By Fano's Lemma, the right hand side is bounded from below by

$$1 - \frac{\frac{1}{M^2}\sum_{l,m=1}^{M} K(\mathbb{P}_{f_l}, \mathbb{P}_{f_m}) + \log 2}{\log M}, \tag{4.7}$$

provided that $M \geq 3$. Let $M = \max\{q-s+1, s+1\}$, which greater than or equal to 3, since $q \geq 3$. If $M = s+1$, then we choose

$$J_l = \{1, \ldots, s+1\} \setminus \{l\},$$

for $l = 1, \ldots, M$. On the other hand, if $M = q-s+1$, then we choose

$$J_l = \{1, 2, \ldots, s-1\} \cup \{s-1+l\},$$

for $l = 1, \ldots, M$. The sets are chosen such that for $l \neq m$, we have

$$|J_l \setminus J_m| = |J_m \setminus J_l| = 1.$$

Moreover, let $h_j \in \Sigma_j(\alpha, K, \kappa_1)$ be elements such that $\|h_j\|^2 = \kappa_1$, for $j = 1, \ldots, q$. We choose

$$f_l = \sum_{j \in J_l} h_j,$$

for $l = 1, \ldots, M$. By [**52**, Lemma 4.6], we have

$$K(\mathbb{P}_{f_l}, \mathbb{P}_{f_m}) = \frac{1}{2\epsilon^2}\|f_l - f_m\|^2 = \frac{\kappa_1}{\epsilon^2},$$

for $l \neq m$. We conclude that

$$\inf_{\hat{J}} \sup_{f \in \Sigma} \mathbb{P}_f \left(\hat{J} \neq J_f \right) \geq 1 - \frac{\kappa_1}{\epsilon^2 \log M} - \frac{\log 2}{\log M}.$$

This completes the proof. □

The second part of Proposition 14 follows from:

PROPOSITION 17. *Let $s \neq 0$ and $s \leq q-2$. We have*

$$\inf_{\hat{J}} \sup_{f \in \Sigma(\alpha, K, \kappa_1, s)} \mathbb{P}_f \left(\hat{J} \neq J_f \right) \geq 1 - \frac{2\kappa_1^{\frac{4\alpha+1}{2\alpha}} K^{-\frac{1}{2\alpha}}/\epsilon^4 + \log 2}{\log(q-s+1)}.$$

PROOF. The proof is based on with Fano's Lemma combined with the method of several fuzzy hypothesis (see, e.g., [**74**, Section 2.7.4 and 2.7.5]). Again the proof is standard (compare to [**74**, Chapter 2.7.5] and [**18**, Theorem 2]). We start with the inequality

$$\inf_{\hat{J}} \sup_{f \in \Sigma} \mathbb{P}_f \left(\hat{J} \neq J_f \right) \geq \inf_{\hat{J}} \frac{1}{M} \sum_{l=0}^{M} \int_{\Sigma} \mathbb{P}_f \left(\hat{J} \neq J_l \right) \mu_l(df),$$

where the J_l are different support sets and where μ_l are probability measures on Σ each supported on a finite set of functions having support J_l. By Fano's Lemma, the right hand side is bounded from below by

$$1 - \frac{\frac{1}{M} \sum_{l=1}^{M} K(\mathbf{P}_l, \mathbf{Q}) + \log 2}{\log M}, \tag{4.8}$$

provided that $M \geq 3$, where

$$\mathbf{P}_l(\cdot) = \int_{\Sigma} \mathbb{P}_f(\cdot) \mu_l(df)$$

and $\mathbf{Q}$ is an arbitrary probability measure to be chosen later (for this version of Fano's lemma, see [**34**, Eq. (1.3)]). Let $M = q - s + 1$, and let

$$J_l = \{1, 2, \ldots, s-1\} \cup \{s - 1 + l\},$$

for $l = 1, \ldots, M$. Moreover, let μ_l be the uniform measure on the set

$$\mathcal{S}_l = \left\{ f = \sum_{j \in J_l} f_j : f_j = A \sum_{k=1}^{d} \omega_{jk} \phi_{jk}, \omega_{jk} \in \{-1, 1\} \right\},$$

where

$$A = \sqrt{\frac{\kappa_1}{d}}$$

and d is the greatest integer satisfying

$$d \leq \left(\frac{K}{\kappa_1}\right)^{\frac{1}{2\alpha}}.$$

Thus, μ_l is supported on 2^{sd} functions which all have the support J_l. Moreover, for such a function $f = \sum_{j \in J_l} f_j$, we have

$$\|f_j\|^2 = \frac{\kappa_1}{d} \sum_{k=1}^{d} \|\phi_{jk}\|^2 = \kappa_1$$

and

$$\sum_{k=1}^{\infty} k^{2\alpha} \langle f_j, \phi_{jk} \rangle^2 = \frac{\kappa_1}{d} \sum_{k=1}^{d} k^{2\alpha} \leq \kappa_1 d^{2\alpha} \leq K$$

which implies that $f \in \Sigma(\alpha, K, \kappa_1, s)$, as required. Next, we choose the probability measure $\mathbf{Q}$ as follows:

$$\mathbf{Q}(\cdot) = \int_{\Sigma} \mathbb{P}_f(\cdot) \mu(df),$$

where μ is the uniform measure on the set

$$\mathcal{S} = \left\{ f = \sum_{j=1}^{s-1} f_j : f_j = A \sum_{k=1}^{d} \omega_{jk} \phi_{jk}, \omega_{jk} \in \{-1, 1\} \right\},$$

with A and d as above. We now use:

LEMMA 20. *We have*

$$K(\mathbf{P}_l, \mathbf{Q}) \leq \frac{dA^4}{\epsilon^4}.$$

A proof of Lemma 20 is given in Appendix C. Inserting the choices of A and d, we get

$$K(\mathbf{P}_l, \mathbf{Q}) \leq \frac{2\kappa_1^{\frac{4\alpha+1}{2\alpha}} K^{-\frac{1}{2\alpha}}}{\epsilon^4}.$$

We conclude that

$$\inf_{\hat{J}} \sup_{f \in \Sigma} \mathbb{P}_f \left(\hat{J} \neq J_f \right) \geq 1 - \frac{2\kappa_1^{\frac{4\alpha+1}{2\alpha}} K^{-\frac{1}{2\alpha}}}{\epsilon^4 \log M} - \frac{\log 2}{\log M}.$$

□

REMARK 17. The proof of Proposition 15 follows essentially the same lines as that of Proposition 14. The only difference is that, in the proof of Proposition 17, we choose $d = d_1$ and we let the ϕ_{jk} be orthonormal bases of the V_j.

APPENDIX A

Appendix to Chapter 1

Selberg's result

In this appendix we briefly discuss Selberg's result about the rate of convergence in the central limit theorem of $\operatorname{Im}\log\zeta(1/2+it)$ (see [**68**, Theorem 2] and [**73**, Theorem 6.2]). From Theorem 1 we deduce:

LEMMA 21. *Let $x = e^{\log T/N}$ and N such that $N/\log\log T \to \infty$ and $x \to \infty$ as $T \to \infty$. Suppose further that $N/\log\log T = O(\log\log T)$. Then*

$$\sup_{a<b}\Bigg(\frac{1}{T}\lambda\Big(\Big\{t\in[T,2T]: \frac{1}{\sqrt{(\log\log x+\gamma)/2}}\sum_{p\le x}\frac{\sin(t\log p)}{\sqrt{p}}\in[a,b]\Big\}\Big) - \int_a^b e^{-t^2/2}\frac{dt}{\sqrt{2\pi}}\Bigg) = O(1/\sqrt{\log\log T}). \quad (0.9)$$

PROOF. We denote by $\Phi_n(u)$ the left hand side of (1.3). Using [**27**, XVI.3, formula 3.13] we can bound the left hand side of (0.9) by

$$\frac{2}{\pi}\int_{-c\sqrt{\log\log x}}^{c\sqrt{\log\log x}} e^{-u^2/2}|(\Phi_n(u/\sqrt{(\log\log x+\gamma)/2})-1)/u|du + O\Big(\frac{1}{c\sqrt{\log\log x}}\Big). \quad (0.10)$$

An inspection of the proof of Proposition 2 combined with (4.2) shows that $\Phi_n(u) = \Phi(u)(1+O(1/\log x)) + O(1/\log T)$, $|u|\le c$. If we choose $c>0$ such that $\Phi(u)$ has no zeros for $|u|\le c$, we obtain $\Phi_n(u) = \Phi(u)(1+O(1/\log x))$, $|u|\le c$. On the other hand, we have $\Phi(u/\sqrt{(\log\log x+\gamma)/2}) = 1+O(u^2/\log\log x)$, $|u|\le c\sqrt{\log\log x}$. Plugging in these estimates gives that (0.10) is $O(1/\sqrt{\log\log x})$. From $N/\log\log T = O(\log\log T)$ we conclude that $\log\log T/\log\log x \to 1$ and this completes the proof. □

This lemma combined with the bound (see [**73**, Lemma 6.2])

$$|\{t\in[T,2T]: |r_{1,x}(t)|\ge c'\log\log\log T\}| = O(1/\sqrt{\log\log T}),$$

where $c' > 0$ is a constant, yields Selberg's result

$$\sup_{a<b}\Bigg(\frac{1}{T}\lambda\Big(\Big\{t\in[T,2T]:\frac{\operatorname{Im}\log\zeta(1/2+it)}{\sqrt{(\log\log T)/2}}\in[a,b]\Big\}\Big) - \int_a^b e^{-t^2/2}\frac{dt}{\sqrt{2\pi}}\Bigg) = O\Big(\frac{\log\log\log T}{\sqrt{\log\log T}}\Big).$$

Mean value estimates

For completeness we present some standard mean value estimates which we applied in the proof of Corollary 2 (see [**66**, Lemma 3] and [**69**, Lemma 3]). For this purpose let x and y be positive real numbers, a_p and b_p be complex numbers with $|a_p| \le 1$ and $|b_p| \le \log p/\log x$, and k be a nonnegative integer. By repeating the arguments in the proof of Proposition 1, we obtain

$$\frac{1}{T}\int_T^{2T}\Big|\sum_{p\le x}\frac{a_p}{p^{1+2it}}\Big|^{2k}dt \le k!\Big(\sum_{p\le x}\frac{1}{p^2}\Big)^k + 2Dk!(\pi(x))^k/T,$$

$$\frac{1}{T}\int_T^{2T}\Big|\sum_{y<p\le x}\frac{a_p}{p^{1/2+it}}\Big|^{2k}dt \le k!\Big(\sum_{y<p\le x}\frac{1}{p}\Big)^k + 2Dk!(\pi(x)-\pi(y))^k/T$$

$$\frac{1}{T}\int_T^{2T}\Big|\sum_{p\le x}\frac{b_p}{p^{1/2+it}}\Big|^{2k}dt \le k!\frac{1}{(\log x)^k}\Big(\sum_{p\le x}\frac{\log p}{p}\Big)^k + 2Dk!(\pi(x))^k/T.$$

If $x \le T^{1/k}$, the first and the third term are bounded by $(Ak)^k$ and the second by $(k(\log\log x - \log\log y + A))^k$, $A > 0$ some constant.

For example, we obtain for a function $|g(u)| \le 1$

$$\begin{aligned}\frac{1}{T}\int_T^{2T}&\Big|\frac{1}{\log T^{1/V}}\sum_{n\le T^{1/V}}\frac{\Lambda(n)}{n^{1/2+it}}g\Big(\frac{\log n}{\log T^{1/V}}\Big)\Big|^{2\lfloor V\rfloor}dt\\ &= \frac{1}{T}\int_T^{2T}\Big|\sum_{p\le T^{1/V}}\frac{b_p}{p^{1/2+it}} + \sum_{p^2\le T^{1/V}}\frac{a_p}{p^{1+2it}} + O(1)\Big|^{2\lfloor V\rfloor}dt\\ &\le 3^{2V}((AV)^V + (AV)^V + O(1)^V). \qquad (0.11)\end{aligned}$$

Large deviation theory

In this appendix we give the definition of the large deviation principle and state two important results which we used in the proofs of Corollary 2 and 3 (see [**22**]).

A function $I : \mathbb{R} \to [0,\infty]$ is called a rate function (resp. good rate function), if for all $\alpha \in [0,\infty)$, the sets $\{x : I(x) \le \alpha\}$ are closed (resp. compact). A family $\{Z_\epsilon\}$ of real-valued random variables satisfies the large deviation principle with the speed ϵ and the rate function I, if

(a) For any closed set $F \subseteq \mathbb{R}$

$$\limsup_{\epsilon \to 0} \epsilon \log \mathbb{P}(Z_\epsilon \in F) \leq - \inf_{x \in F} I(x).$$

(b) For any open set $G \subseteq \mathbb{R}$

$$\liminf_{\epsilon \to 0} \epsilon \log \mathbb{P}(Z_\epsilon \in G) \geq - \inf_{x \in G} I(x).$$

THEOREM 16 (Gärtner-Ellis, see Theorem 2.3.6 or 4.5.20 in [**22**]). *Suppose that for each* $\lambda \in \mathbb{R}$

$$\Lambda(\lambda) := \lim_{\epsilon \to 0} \epsilon \log \mathbb{E}\left[e^{\lambda Z_\epsilon/\epsilon}\right]$$

exists and that Λ *is differentiable. Then the family* $\{Z_\epsilon\}$ *satisfies the large deviation principle with the good rate function* $I(x) = \sup_{\lambda \in \mathbb{R}}(\lambda x - \Lambda(\lambda))$.

THEOREM 17 (Varadhan, see Theorem 4.3.1 in [**22**]). *Suppose that* $\{Z_\epsilon\}$ *satisfies the large deviation principle with a good rate function* I *and let* $h \in \mathbb{R}$. *Assume further that for some* $\gamma > 1$

$$\limsup_{\epsilon \to 0} \epsilon \log \mathbb{E}\left[e^{\gamma h Z_\epsilon/\epsilon}\right] < \infty. \tag{0.12}$$

Then

$$\lim_{\epsilon \to 0} \epsilon \log \mathbb{E}\left[e^{h Z_\epsilon/\epsilon}\right] = \sup_{x \in \mathbb{R}}(xh - I(x)).$$

DEFINITION 3 (see Definition 4.2.10 in [**22**]). Let $\{Z_\epsilon\}$ and $\{\tilde{Z}_\epsilon\}$ be two families of real-valued random variables, defined on the same probability space. Then $\{Z_\epsilon\}$ and $\{\tilde{Z}_\epsilon\}$ are called exponentially equivalent if for each $\delta > 0$,

$$\limsup_{\epsilon \to 0} \epsilon \log \mathbb{P}(|Z_\epsilon - \tilde{Z}_\epsilon| > \delta) = -\infty. \tag{0.13}$$

APPENDIX B

Appendix to Chapter 2

Proof of Lemma 2

We first show how (i) implies (ii) and (iii). Let $h_1 \in \mathcal{H}_1$ and $h_2 \in \mathcal{H}_2$. Then by (i) we have $\|h_1 + h_2\|^2 \geq \|h_1\|^2 - 2\varrho\|h_1\|\|h_2\| + \|h_2\|^2$ and (ii) follows from the inequality $2\|h_1\|\|h_2\| \leq \|h_1\|^2 + \|h_2\|^2$, while (iii) follows from $2\varrho\|h_1\|\|h_2\| \leq \varrho^2\|h_1\|^2 + \|h_2\|^2$.

Next, we show how (ii) implies (i). Let $0 \neq h_1 \in \mathcal{H}_1$ and $0 \neq h_2 \in \mathcal{H}_2$. We may assume without loss of generality that $\|h_1\| = \|h_2\| = 1$ and that $\langle h_1, h_2\rangle \geq 0$. Then by (ii) we have $2 - 2\langle h_1, h_2\rangle = \|h_1 - h_2\|^2 \geq 2(1 - \varrho)$ which gives (i).

Finally, suppose that (iii) is true. Let $0 \neq h_1 \in \mathcal{H}_1$ and $0 \neq h_2 \in \mathcal{H}_2$. Again, we may assume that $\|h_1\| = \|h_2\| = 1$. Then by (iii) we have $1 - \langle h_1, h_2\rangle^2 = \|h_1 - \langle h_1, h_2\rangle h_2\|^2 \geq 1 - \varrho^2$ which gives (i). This completes the proof. □

A feasible estimator

In this appendix, we show that estimators which are based on the condition $(1/n)\sum_{i=1}^n g_1(X_1^i) = 0$ have (up to a constant and a term of smaller order) the same risk bound as our estimators based on the condition $\mathbb{E}[g_1(X_1)] = 0$. We only sketch the main arguments in the case $W_1 = V_1$. Suppose that we choose $U_1 \subset L^2(\mathbb{P}^{X_1})$ and $V_2 \subset L^2(\mathbb{P}^{X_2})$, where V_2 contains all constant functions. Let $V_1' = \{g_1 \in U_1 | (1/n)\sum_{i=1}^n g_1(X_1^i) = 0\}$ and $V_1 = \{g_1 \in U_1 | \mathbb{E}[g_1(X_1)] = 0\}$. Since V_2 contains all constants, we have $V = V_1 + V_2 = V_1' + V_2$. This implies that the first components of $\hat{f}_{V_1+V_2}$ and $\hat{f}_{V_1'+V_2}$ in V_1 and V_1', respectively, differ only by the constant

$$\frac{1}{n}\sum_{i=1}^n (\hat{f}_V)_1(X_1^i).$$

The risk of this constant can be bounded as follows. By the bound $(x+y)^2 \leq 2x^2 + 2y^2$, we have

$$\begin{aligned}
&\mathbb{E}\left[1_{\mathcal{E}_\delta}\left(\frac{1}{n}\sum_{i=1}^n (\hat{f}_V)_1(X_1^i)\right)^2\right] \\
&\leq 2\mathbb{E}\left[1_{\mathcal{E}_\delta}\left(\frac{1}{n}\sum_{i=1}^n (\hat{f}_V)_1(X_1^i) - f_1(X_1^i)\right)^2\right] \\
&+ 2\mathbb{E}\left[1_{\mathcal{E}_\delta}\left(\frac{1}{n}\sum_{i=1}^n f_1(X_1^i)\right)^2\right].
\end{aligned}$$

Applying the Cauchy-Schwarz inequality and the fact that the $f_1(X_1^i)$ are independent and centered, this can be bounded by

$$2\mathbb{E}\left[1_{\mathcal{E}_\delta}\|(\hat{f}_V)_1 - f_1\|_n^2\right] + \frac{2\|f_1\|^2}{n}.$$

Now apply Lemma 10 to the first term.

Proof of Lemma 4 and 5

First, we prove Lemma 4. In Section 4.1, we have shown that

$$\|g_1\|_\infty^2 \leq \varphi_1^2 d_1 \|g_1\|^2$$

and

$$\|g_{j2}\|_\infty^2 \leq \varphi_2^2 d_{j2} \|g_{j2}\|^2$$

for all $g_1 \in V_1$, $g_{j2} \in V_{j2}$, $1 \leq j \leq q-1$, with $\varphi_1^2 = 2/c$ and $\varphi_2^2 = 2/c$. Now let $g = g_1 + g_2 \in V$. Suppose that $g_2 = \sum_{j=1}^{q-1} g_{2j}$ is the decomposition satisfying (4.6). Applying the above bounds, the Cauchy-Schwarz inequality, and Assumption 9, we obtain

$$\|g_2\|_\infty \leq \sum_{j=1}^{q-1} \varphi_2 \sqrt{d_{2j}} \|g_{2j}\| \leq \frac{\varphi_2}{\sqrt{1-\epsilon_2}} \sqrt{\sum_{j=1}^{q-1} d_{2j}} \|g_2\|.$$

Applying again the Cauchy-Schwarz inequality and then Assumption 2 and Lemma 2, we conclude that

$$\begin{aligned}
\|g_1 + g_2\|_\infty &\leq \varphi_1 \sqrt{d_1}\|g_1\| + \frac{\varphi_2}{\sqrt{1-\epsilon_2}} \sqrt{\sum_{j=1}^{q-1} d_{2j}}\|g_2\| \\
&\leq \sqrt{\frac{\varphi_1 \vee \varphi_2}{(1-\epsilon_2)(1-\rho_0)}} \sqrt{d_1 + \sum_{j=1}^{q-1} d_{2j}}\|g_1 + g_2\|. \qquad (0.14)
\end{aligned}$$

This completes the proof of Lemma 4. The proof of Lemma 5 is similar. In [**10**], it is shown that

$$\|g_1\|_\infty^2 \leq (r_1+1)^2 m_1 \int_0^1 g_1^2(x_1)dx_1$$

and

$$\|g_{j2}\|_\infty^2 \leq (r_2+1)^2 m_2 \int_0^1 g_{j2}^2(x_{j2})dx_{j2}$$

for all $g_1 \in V_1$, $g_{j2} \in V_{j2}$, $1 \leq j \leq q-1$. This implies that

$$\|g_1\|_\infty^2 \leq \varphi_1^2 d_1 \|g_1\|^2$$

and

$$\|g_{j2}\|_\infty^2 \leq \varphi_2^2 d_{j2} \|g_{j2}\|^2$$

with $\varphi_1^2 = 2(r_1+1)/c$ and $\varphi_2^2 = 2(r_2+1)/c$. Now proceed as above. This completes the proof. □

Proof of Corollary 11

LEMMA 22. *Let Assumption 4 and 9 be satisfied. Suppose that* (4.17) *and* (4.18) *are satisfied. Then* (3.7) *is satisfied with*

$$\psi_\Pi(V_2) = \sqrt{C_3}\sqrt{\sum_{k=1}^{q-1} d_{2k}^{-2\beta}}$$

and

$$h_1(x_1) = \sqrt{\sum_{k=1}^{q} h_{1k}^2(x_1)} + \sqrt{\frac{c'}{1-\epsilon_2}\int \left(\frac{p_X(x_1,x_2)}{p_{X_1}(x_1)p_{X_2}(x_2)}\right)^2 p_{X_2}(x_2)dx_2},$$

where $c' = \sum_{j,k=1, j\neq k}^{q-1} \|h'_{jk}\|^2$. *Note that* $h_1 \in L^2(\mathbb{P}^{X_1})$, *by Assumption 4.*

PROOF. Let x_1 be fixed (such that $p_X(x_1,\cdot)/(p_{X_1}(x_1)p_{X_2}(\cdot)) \in L^2(\mathbb{P}^{X_2})$, which is satisfied for $\mathbb{P}^{X_1}$-almost all x_1, by Assumption 4). By the projection theorem, the expression

$$\int \left(\frac{p_X(x_1,x_2)}{p_{X_1}(x_1)p_{X_2}(x_2)} - g(x_2)\right)^2 p_2(x_2)dx_2,$$

subject to the constraints $g \in H_2$, is minimized by r. Suppose that $r = \sum_{k=1}^{q-1} r_k$ is the decomposition such that (4.6) is satisfied (note that we omit the dependence of r and the r_k on x_1). For $k = 1, \ldots, q-1$, we have

$$\mathbb{E}\left[\frac{p_X(x_1,X_2)}{p_{X_1}(x_1)p_{X_2}(X_2)}\middle| X_{2k} = x_{2k}\right] = \frac{p_{X_1,X_{2k}}(x_1,x_{2k})}{p_{X_1}(x_1)p_{X_{2k}}(x_{2k})}. \tag{0.15}$$

Thus the r_k satisfy the $q-1$ equations

$$r_k(x_{2k}) = \frac{p_{X_1,X_{2k}}(x_1,x_{2k})}{p_{X_1}(x_1)p_{X_{2k}}(x_{2k})} - \sum_{j=1,j\neq k}^{q-1} \int r_j(x_{2j}) \frac{p_{X_{2j},X_{2k}}(x_{2j},x_{2k})}{p_{X_{2j}}(x_{2j})p_{X_{2k}}(x_{2k})} p_{X_{2j}}(x_{2j}) dx_{2j},$$

for $\mathbb{P}^{X_{2k}}$-almost all x_{2k}, $1 \leq k \leq q-1$ (note again that we omit the dependence of the r_k on x_1). By (4.17), (4.18), and the Cauchy-Schwarz inequality, the first and the second term on the right hand side are contained in $\mathcal{H}(\beta, h_{1k}(x_1))$ and $\mathcal{H}(\beta, \sum_{j=1,j\neq k}^{q-1} \|r_j\|_{L^2(\mathbb{P}^{X_{2j}})} \|h'_{jk}\|)$, respectively. We conclude that

$$\begin{aligned}
&\|r - \Pi_{V_2} r\|_{L^2(\mathbb{P}^{X_2})} \\
&\leq \sum_{k=1}^{q-1} \|r_k - \Pi_{V_{2k}} r_k\|_{L^2(\mathbb{P}^{X_{2k}})} \\
&\leq \sum_{k=1}^{q-1} C_3 \left(h_{1k}(x_1) + \sum_{j=1,j\neq k}^{q-1} \|r_j\|_{L^2(\mathbb{P}^{X_{2j}})} \|h'_{jk}\| \right) d_{2k}^{-\beta}.
\end{aligned}$$

Applying the Cauchy-Schwarz inequality and Assumption 9, this is bounded by

$$\leq \sum_{k=1}^{q-1} C_3 d_{2k}^{-\beta} \left(h_{1k}(x_1) + \sqrt{\frac{\|r\|^2_{L^2(\mathbb{P}^{X_2})}}{1-\epsilon_2} \sum_{j=1,j\neq k}^{q-1} \|h'_{jk}\|^2} \right).$$

Applying the Cauchy-Schwarz inequality again and the fact that orthogonal projections lower the norm, we obtain the claimed $\psi_\Pi(V_2)$ and $h_1(x_1)$. This completes the proof. □

A remark on Theorem 9

In this appendix, we show how to derive Theorem 9 from Rudelson's lemma combined with Talagrand's inequality. We use the same notation as in the proof of Theorem 9. Then Rudelson's lemma [**62**, Theorem 1] says that there is a universal constant C_1 such that

$$\mathbb{E}\left[\|B_n - I\|_{\mathrm{op}} \right] \leq C_1 \sqrt{\frac{\varphi^2 d \log d}{n}}, \tag{0.16}$$

provided that the last expression is smaller than 1. Here, we also used (5.4). From (5.6), Theorem 11 (applied with $\lambda = \delta/2$, note that $v, b \leq \varphi^2 d$), and (0.16), we now conclude:

THEOREM 18. *Let $\delta \in (0,1)$. Suppose that*

$$4C_1 \sqrt{\frac{\varphi^2 d \log d}{n}} \leq \delta, \tag{0.17}$$

where C_1 is the constant (0.16). *Then*

$$\mathbb{P}\left(\sup_{g\in V,\|g\|\leq 1} |\|g\|_n^2 - \|g\|^2| > \delta\right) \leq 3\exp\left(-\frac{\kappa}{4}\frac{n\delta^2}{\varphi^2 d}\right),$$

where κ is the constant from Talagrand's inequality.

Proof of Lemma 6

In this appendix, we prove the following stronger convergence in norm result

$$\big\|(\Pi h)_1 - \big(\Pi_1 - \sum_{j=1}^{k}(\Pi_1\Pi_2)^j(1-\Pi_1)\big)h\big\| \leq \frac{\rho_0^{2k+1}}{1-\rho_0^2}\|h\|.$$

First, note that

$$\|\Pi_1 h_2\| \leq \rho_0\|h_2\| \text{ and } \|\Pi_2 h_1\| \leq \rho_0\|h_1\| \tag{0.18}$$

for all $h_2 \in \mathcal{H}_2$, $h_1 \in \mathcal{H}_1$. The first inequality follows from

$$\|\Pi_1 h_2\|^2 = \langle \Pi_1 h_2, \Pi_1 h_2\rangle = \langle \Pi_1 h_2, h_2\rangle \leq \rho_0\|\Pi_1 h_2\|\|h_2\|,$$

the second one can be shown analogously. Now, define the alternating sums

$$\begin{aligned} L_1^{(k)} &= \Pi_1 - \Pi_1\Pi_2 + \Pi_1\Pi_2\Pi_1 - \cdots + \Pi_1(\Pi_2\Pi_1)^k \\ &= \Pi_1 - \sum_{j=1}^{k}(\Pi_1\Pi_2)^j(1-\Pi_1) \\ &= \sum_{j=0}^{k-1}\Pi_1(\Pi_2\Pi_1)^j(1-\Pi_2) + \Pi_1(\Pi_2\Pi_1)^k \end{aligned} \tag{0.19}$$

and

$$\begin{aligned} L_2^{(k)} &= \Pi_2 - \Pi_2\Pi_1 + \cdots - (\Pi_2\Pi_1)^k \\ &= \Pi_2 - \sum_{j=1}^{k-1}(\Pi_2\Pi_1)^j(1-\Pi_2) - (\Pi_2\Pi_1)^k \\ &= \sum_{j=0}^{k-1}\Pi_2(\Pi_1\Pi_2)^j(1-\Pi_1). \end{aligned} \tag{0.20}$$

Applying (0.18), we obtain

$$\begin{aligned}\|L_1^{(l)} - L_1^{(k)}\|_{\text{op}} &\leq \sum_{j=k+1}^{l} \|(\Pi_1\Pi_2)^j(1-\Pi_1)\|_{\text{op}} \\ &\leq \sum_{j=k+1}^{l} \rho_0^{2j-1} \\ &\leq \frac{\rho_0^{2k+1}}{1-\rho_0^2}.\end{aligned}$$

Similarly, we obtain

$$\|L_2^{(l)} - L_2^{(k)}\|_{\text{op}} \leq \frac{\rho_0^{2k}}{1-\rho_0^2}.$$

We conclude that $\{L_j^{(k)}\}$ is a Cauchy sequence in the Banach space of all bounded linear mappings (see, e.g., [**65**, Theorem 4.1]). Hence $L_j^{(k)}$ converges to a bounded linear mapping L_j and we have

$$\|L_1 - L_1^{(k)}\|_{\text{op}} \leq \frac{\rho_0^{2k+1}}{1-\rho_0^2}.$$

Moreover, $L^{(n)} = L_1^{(n)} + L_2^{(n)}$ converges to the linear map $L = L_1 + L_2$. Since L_j takes values in $\mathcal{H}_j$, it remains to show that $L = \Pi$. Since $L^{(n)}$ is self-adjoint, the limit L is also self-adjoint. Applying (0.19), (0.20), and (0.18), we have

$$L_1h_1 = h_1,\ L_2h_2 = h_2, \text{ and } L_1h_2 = L_2h_1 = 0 \tag{0.21}$$

for all $h_1 \in \mathcal{H}_1$, $h_2 \in \mathcal{H}_2$. This implies that L is idempotent, i.e. satisfies $L^2 = L$. It follows from [**8**, Proposition 2] that L is an orthogonal projection. But (0.21) also implies that the range of L is equal to $\mathcal{H}_1 + \mathcal{H}_2$. This gives $L = \Pi$. This completes the proof. □

Proof of Lemma 7

We only proof (iii), since (i) and (ii) are standard. By the spectral theorem, there exists an orthogonal matrix V and nonnegative real numbers $\lambda_1(B), \ldots, \lambda_{k_1}(B)$ such that

$$B = V^T \operatorname{diag}(\lambda_1(B), \ldots, \lambda_{k_1}(B))V. \tag{0.22}$$

Now, by the Cauchy-Schwarz inequality, each entry of a matrix is bounded by the operator norm of that matrix. In particular, we have $|(VAV^T)_{jk}| \leq \|VAV^T\|_{\text{op}} = \|A\|_{\text{op}}$ for all j, k, since V is orthogonal. Applying (0.22), part (i) of this Lemma, the fact that the $\lambda_j(B)$ are nonnegative, and the previous

argument, we obtain

$$|\operatorname{tr}(AB)| = \Big|\sum_{j=1}^{k_1}(VAV^T)_{jj}\lambda_j(B)\Big| \leq \max_{j=1,\dots,k_1}|(VAV^T)_{jj}|\operatorname{tr}(B) \leq \|A\|_{\mathrm{op}}\operatorname{tr}(B).$$

This completes the proof. □

An alternative proof of Corollary 12

Let $\{\phi_{1j}\}_{1\leq j\leq d_1}$ be a basis of V_1 and let $\{\phi_{2j}\}_{1\leq j\leq d_2}$ be a basis of V_2. Let

$$Z_1 = (\phi_{1j}(X_1^i))_{1\leq i\leq n, 1\leq j\leq d_1} \in \mathbb{R}^{n\times d_1}$$

and

$$Z_2 = (\phi_{2j}(X_2^i))_{1\leq i\leq n, 1\leq j\leq d_2} \in \mathbb{R}^{n\times d_2}.$$

Moreover, let

$$Z = (Z_1|Z_2) \in \mathbb{R}^{n\times d}.$$

In the following, suppose that the $\mathcal{E}_\delta$ holds. Then Theorem 9 implies that Z^TZ is invertible (see, e.g., (5.5)). Thus the minimum of $\|\boldsymbol{\epsilon} - Z\beta\|_n^2$ is taken at $\beta = (Z^TZ)^{-1}Z^T\boldsymbol{\epsilon}$ and we have $\hat{\Pi}_V\boldsymbol{\epsilon} = Z(Z^TZ)^{-1}Z^T\boldsymbol{\epsilon}$. This implies that $(\hat{\Pi}_V\boldsymbol{\epsilon})_1 = (Z_1|0)(Z^TZ)^{-1}Z^T\boldsymbol{\epsilon}$. We conclude that

$$\mathbb{E}\left[\|(\hat{\Pi}_V\boldsymbol{\epsilon})_1\|_n^2|X^1,\dots,X^n\right] = \operatorname{tr}\left((Z_1|0)(Z^TZ)^{-1}(Z_1|0)^T\right)\frac{\sigma^2}{n}.$$

On the other hand, (5.2) is equivalent to the inequality

$$\beta^TZ^TZ\beta \geq \frac{(1-\delta)}{(1+\delta)}(1-\rho_0^2)\beta_1^TZ_1^TZ_1\beta_1 \tag{0.23}$$

for all $\beta = (\beta_1^T, \beta_2^T)^T$ with $\beta_1 \in \mathbb{R}^{d_1}$ and $\beta_2 \in \mathbb{R}^{d_2}$. Now, we apply the following lemma. Note that a proof of a more general result can be found in [**25**, Lemma 2.1] and that this result was also applied in [**70**].

LEMMA 23. *Let $F \in \mathbb{R}^{d\times d}$ and $F_1 \in \mathbb{R}^{d_1\times d_1}$ be two symmetric and positive definite matrices with $d_2 = d - d_1 \geq 0$. Suppose that $v^TFv \geq v_1^TF_1v_1$ for all $v = (v_1^T, v_2^T)^T$ with $v_1 \in \mathbb{R}^{d_1}$ and $v_2 \in \mathbb{R}^{d_2}$. Then, for all $w_1 \in \mathbb{R}^{d_1}$,*

$$\begin{pmatrix} w_1 \\ 0 \end{pmatrix}^T F^{-1}\begin{pmatrix} w_1 \\ 0 \end{pmatrix} \leq w_1^TF_1^{-1}w_1. \tag{0.24}$$

Applying (0.23) and this Lemma with

$$F = Z^TZ \text{ and } F_1 = \frac{(1-\delta)}{(1+\delta)}(1-\rho_0^2)Z_1^TZ_1,$$

we obtain

$$\operatorname{tr}((Z_1|0)(Z^TZ)^{-1}(Z_1|0)^T) \leq \frac{(1+\delta)}{(1-\delta)}\frac{1}{(1-\rho_0^2)}\operatorname{tr}(Z_1(Z_1^TZ_1)^{-1}Z_1^T).$$

Since $\hat{\Pi}_{V_1} = Z_1(Z_1^T Z_1)^{-1} Z_1^T$, we have $\mathrm{tr}(Z_1(Z_1^T Z_1)^{-1} Z_1^T) = \dim V_1 = d_1$. This completes the proof. □

Proof of (5.16)

In this appendix, we prove (5.16). As mentioned in the proof of Theorem 3, the main arguments are taken from [**5**, page 139 and 140]. We define the event $\mathcal{A} = \{\|\hat{f}_1\|_\infty \leq k_n\}$. Then

$$\begin{aligned}\mathbb{E}\left[\|f_1 - \hat{f}_1^*\|^2\right] &= \mathbb{E}\left[(1_{\mathcal{E}_\delta} 1_{\mathcal{A}} + 1_{\mathcal{E}_\delta} 1_{\mathcal{A}^c} + 1_{\mathcal{E}_\delta^c})\|f_1 - \hat{f}_1^*\|^2\right] \\ &\leq \mathbb{E}\left[1_{\mathcal{E}_\delta}\|f_1 - \hat{f}_1\|^2\right] + \mathbb{E}\left[1_{\mathcal{E}_\delta} 1_{\mathcal{A}^c}\|f_1\|^2\right] + \mathbb{E}\left[1_{\mathcal{E}_\delta^c}(\|f_1\| + k_n)^2\right].\end{aligned}$$

Thus it remains to consider the last two expressions. By Theorem 9, the last one is bounded by

$$2^{3/4}(\|f_1\| + k_n)^2 d \exp\left(-\kappa \frac{n\delta^2}{\varphi^2 d}\right).$$

Consider the other one. By Assumption 3, we have $\|\hat{f}_1\|_\infty^2 \leq \varphi^2 d \|\hat{f}_1\|^2$. If $\mathcal{E}_\delta$ holds, then

$$\|\hat{f}_1\|_\infty^2 \leq \frac{\varphi^2 d}{(1-\delta)} \|\hat{\Pi}_{W_1}(\hat{\Pi}_V \mathbf{Y})_1\|_n^2 \leq \frac{\varphi^2 d}{(1-\delta)} \|(\hat{\Pi}_V \mathbf{Y})_1\|_n^2,$$

where we applied the definition of $\mathcal{E}_\delta$ and the fact that projections lower the norm. By Proposition 5, the last expression is bounded by

$$\frac{(1+\delta)\varphi^2 d}{(1-\delta)^2(1-\rho_0^2)} \|\hat{\Pi}_V \mathbf{Y} - g_2\|_n^2,$$

for $g_2 \in V_2$ arbitrary. Using $\|\hat{\Pi}_V \mathbf{Y} - g_2\|_n \leq \|\hat{\Pi}_V(f - g_2)\|_n + \|\hat{\Pi}_V \boldsymbol{\epsilon}\|_n \leq \|f - g_2\|_n + \|\boldsymbol{\epsilon}\|_n$ and Markov's inequality, we conclude that

$$\begin{aligned}\mathbb{P}(\mathcal{E}_\delta \cap \mathcal{A}^c) &\leq \mathbb{P}\left(\frac{(1+\delta)\varphi^2 d}{(1-\delta)^2(1-\rho_0^2)}(\|f - g_2\|_n + \|\boldsymbol{\epsilon}\|_n)^2 > k_n^2\right) \\ &\leq \frac{2(1+\delta)\varphi^2 d(\|f - g_2\|^2 + \sigma^2)}{(1-\delta)^2(1-\rho_0^2)k_n^2}.\end{aligned}$$

Letting $g_2 = \Pi_{V_2} f$, this completes the proof. □

APPENDIX C

Appendix to Chapter 4

Proof of Lemma 20

The proof of Lemma 20 is similar to the proof of Lemma 10 in [**18**] (see also [**74**, Chapter 2.7.5]). For completeness, we repeat the arguments. First, we compute $d\mathbf{P}_l/d\mathbb{P}_0(y)$. By [**52**, Lemma 4.6], we have

$$\begin{aligned}\frac{d\mathbb{P}_{f_l}}{d\mathbb{P}_0}(y) &= \exp\left[\epsilon^{-2}\left(y(f_l) - \frac{\|f_l\|^2}{2}\right)\right] \\ &= \exp\left[-\frac{A^2 sd}{2\epsilon^2}\right]\exp\left[A\sum_{j\in J_l}\sum_{k=1}^{d}\omega_{jk}y(\phi_{jk})\right].\end{aligned}$$

which implies that

$$\begin{aligned}&\frac{d\mathbf{P}_l}{d\mathbb{P}_0}(y) = \frac{1}{2^{sd}}\sum_{f_l\in\mathcal{S}_l}\frac{d\mathbb{P}_f}{d\mathbb{P}_0}(y) \\ &= \exp\left[-\frac{A^2 sd}{2\epsilon^2}\right]\prod_{j\in J_l}\prod_{k=1}^{d}\left\{\frac{\exp\left[Ay(\phi_{jk})/\epsilon^2\right] + \exp\left[-Ay(\phi_{jk})/\epsilon^2\right]}{2}\right\}.\end{aligned}$$

Setting

$$E_{jk}(y) = \frac{\exp\left[Ay(\phi_{jk})/\epsilon^2 - A^2/(2\epsilon^2)\right] + \exp\left[-Ay(\phi_{jk})/\epsilon^2 - A^2/(2\epsilon^2)\right]}{2},$$

we obtain

$$\frac{d\mathbf{P}_l}{d\mathbb{P}_0}(y) = \prod_{j\in J_l}\prod_{k=1}^{d}E_{jk}(y).$$

Similarly, we have

$$\frac{d\mathbf{Q}}{d\mathbb{P}_0}(y) = \prod_{j=1}^{s-1}\prod_{k=1}^{d}E_{jk}(y).$$

We have that the $E_{jk}(\epsilon W)$ are independent, and that $\mathbb{E}\left[E_{jk}(\epsilon W)\right] = 1$ and

$$\mathbb{E}\left[E_{jk}^2(\epsilon W)\right] = \frac{\exp\left[A^2/\epsilon^2\right] - \exp\left[-A^2/\epsilon^2\right]}{2}.$$

Using these facts, one can compute the χ^2 divergence between $\mathbf{P}_l$ and $\mathbf{Q}$

$$\begin{aligned}\chi^2(\mathbf{P}_l, \mathbf{Q}) + 1 &= \mathbb{E}\left[\left(\frac{d\mathbf{P}_l}{d\mathbb{P}_0}(\epsilon W)\right)^2 / \frac{d\mathbf{Q}}{d\mathbb{P}_0}(\epsilon W)\right] \\ &= \left(\frac{\exp\left[A^2/\epsilon^2\right] - \exp\left[-A^2/\epsilon^2\right]}{2}\right)^d.\end{aligned}$$

Applying the bound $(e^x - e^{-x})/2 \leq e^{x^2}$, we obtain

$$\chi^2(\mathbf{P}_l, \mathbf{Q}) + 1 \leq \exp\left[\frac{dA^4}{\epsilon^4}\right]$$

Now apply $K(\mathbf{P}_l, \mathbf{Q}) \leq \log(1+\chi^2(\mathbf{P}_l, \mathbf{Q}))$ (see, e.g., [**74**, Lemma 2.7]). This completes the proof. □

Bibliography

[1] G. E. Andrews, R. Askey, and R. Roy. *Special Functions. Encyclopedia of Mathematics and its Applications 71*. Cambridge University Press, Cambridge, 1999.

[2] E. Arias-Castro and K. Lounici. Estimation and variable selection with exponential weights. *Electron. J. Stat.*, 8:328–354, 2014.

[3] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

[4] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constr. Approx.*, 28:253–263, 2008.

[5] Y. Baraud. Model selection for regression on a random design. *ESAIM Probab. Statist.*, 6:127–146, 2002.

[6] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113:301–413, 1999.

[7] P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data. Methods, Theory and Applications*. Springer, Heidelberg, 2011.

[8] P. J. Bickel, C. A. J. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and adaptive estimation for semiparametric models*. Springer, New York, 1998. Reprint of the 1993 original.

[9] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.*, 37:1705–1732, 2009.

[10] L. Birgé and P. Massart. From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York, 1997.

[11] L. Birgé and P. Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4:329–375, 1998.

[12] L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc.*, 3:203–268, 2001.

[13] L. Birgé and P. Massart. Minimal penalties for gaussian model selection. *Probab. Theory Related Fields*, 138:33–73, 2007.

[14] E. Bombieri and D. A. Hejhal. On the distribution of zeros of linear combinations of Euler products. *Duke Math. J.*, 80:821–862, 1995.

[15] L. Breiman and J. H. Friedman. Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.*, 80:580–598, 1985.

[16] E. Candès and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.*, 35:2313–2351, 2007.

[17] E. J. Candès and T. Tao. Near-optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inform. Theory*, 52:5406–5425, 2006.

[18] L. Comminges and A. S. Dalalyan. Tight conditions for consistency of variable selection in the context of high dimensionality. *Ann. Statist.*, 40:2667–2696, 2012.

[19] A. Dalalyan, Y. Ingster, and A. B. Tsybakov. Statistical inference in compound functional models. *Probab. Theory Related Fields*, 158:513–532, 2014.

[20] H. Davenport. *Multiplicative Number Theory*. Springer, New York, 3rd edition, 2000.

[21] C. De Boor. *A practical guide to splines*. Springer, New York, revised edition, 2001.

[22] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Springer, Berlin, corrected reprint of the 2nd edition, 2010.

[23] R. M. Dudley. *Real analysis and probability*. Cambridge University Press, Cambridge, 2002. Revised reprint of the 1989 original.

[24] S. Efromovich. Nonparametric regression with the scale depending on auxiliary variable. *Ann. Statist.*, 41:1542–1568, 2013.

[25] S. Ehrenfeld. Complete class theorems in experimental design. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I*, pages 57–67. University of California Press, Berkeley and Los Angeles, 1956.

[26] J. Fan, W. Härdle, and E. Mammen. Direct estimation of low-dimensional components in additive models. *Ann. Statist.*, 26:943–971, 1998.

[27] W. Feller. *An Introduction to Probability Theory and its Applications, vol. 2*. Wiley, New York, 2nd edition, 1971.

[28] M. Fornasier and H. Rauhut. Compressive sensing. In *Handbook of Mathematical Methods in Imaging*, pages 187–228. Springer, 2011.

[29] E. Freitag and R. Busam. *Complex Analysis*. Springer, Berlin Heidelberg, 2005.

[30] G. Gayraud and Y. Ingster. Detection of sparse additive functions. *Electron. J. Stat.*, 6:1409–1448, 2012.

[31] A. Ghosh. On the Riemann zeta-function–mean value theorems and the distribution of $|S(T)|$. *J. Number Theory*, 17:93–102, 1983.

[32] D. A. Goldston. On the function $S(T)$ in the theory of the Riemann zeta-function. *J. Number Theory*, 27:149–177, 1987.

[33] S. M. Gonek, C. P. Hughes, and J. P. Keating. A hybrid Euler-Hadamard product for the Riemann zeta function. *Duke Math. J.*, 136:507–549, 2007.

[34] A. A. Gushchin. On Fano's lemma and similar inequalities for the minimax risk. *Theory Probab. Math. Statist.*, 67:29–41, 2003.

[35] T. J. Hastie and R. J. Tibshirani. *Generalized additive models*. Chapman and Hall, London, 1990.

[36] J. Horowitz, J. Klemelä, and E. Mammen. Optimal estimation in additive regression models. *Bernoulli*, 12:271–298, 2006.

[37] J. Huang, J. L. Horowitz, and F. Wei. Variable selection in nonparametric additive models. *Ann. Statist.*, 38:2282–2313, 2010.

[38] J. Jacod, E. Kowalski, and A. Nikeghbali. Mod-Gaussian convergence: new limit theorems in probability and number theory. *Forum Math.*, 23:835–873, 2011.

[39] D. Joyner. *Distribution theorems of L-functions*. Longman Scientific, Harlow; Wiley, New York, 1986.

[40] S. Kayalar and H. L. Weinert. Error bounds for the method of alternating projections. *Math. Control Signals Systems*, 1:43–59, 1988.

[41] J. P. Keating and N. C. Snaith. Random matrix theory and $\zeta(1/2+it)$. *Comm. Math. Phys.*, 214:57–89, 2000.

[42] H. Kober. A theorem on banach spaces. *Compositio Math.*, 7:135–140, 1939.

[43] V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer, Heidelberg, 2011.

[44] V. Koltchinskii and M. Yuan. Sparsity in multiple kernel learning. *Ann. Statist.*, 38:3660–3695, 2010.

[45] E. Kowalski and A. Nikeghbali. Mod-Gaussian convergence and the value distribution of $\zeta(1/2 + it)$ and related quantities. *J. Lond. Math. Soc. (2)*, 86:291–319, 2012.

[46] H. O. Lancaster. Some properties of the bivariate normal distribution considered in the form of a contingency table. *Biometrika*, 44:289–292, 1957.

[47] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28:1302–1338, 2000.

[48] A. Laurinčikas. *Limit Theorems for the Riemann Zeta-Function*. Kluwer, Dordrecht, 1996.

[49] P. D. Lax. *Functional analysis*. John Wiley and Sons, Inc. New York, 2002.

[50] O. B. Linton. Efficient estimation of additive nonparametric regression models. *Biometrika*, 84:469–473, 1997.

[51] E. Mammen, O. Linton, and J. Nielsen. The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. Statist.*, 27:1443–1490, 1999.

[52] P. Massart. *Concentration Inequalities and Model Selection*. Springer, Berlin, 2007.

[53] L. Meier, S. van de Geer, and P. Bühlmann. High-dimensional additive modeling. *Ann. Statist.*, 37:3779–3821, 2009.

[54] J. D. Opsomer. Asymptotic properties of backfitting estimators. *J. Multivariate Anal.*, 73:166–179, 2000.

[55] J. D. Opsomer and D. Ruppert. Fitting a bivariate additive model by local polynomial regression. *Ann. Statist.*, 25:186–211, 1997.

[56] M. Radziwiłł. Large deviations in Selberg's central limit theorem. http://arxiv.org/abs/1108.5092, 2011.

[57] K. Ramachandra. *On the Mean-Value and Omega-Theorems for the Riemann Zeta-Function. Tata Institute of Fundamental Research Lectures on Mathematics and Physics, 85*. Published for the Tata Institute of Fundamental Research, Bombay, Springer, 1995.

[58] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *J. Mach. Learn. Res.*, 13:389–427, 2012.

[59] H. Rauhut. Compressive sensing and structured random matrices. In *Theoretical Foundations and Numerical Methods for Sparse Recovery*, Radon Ser. Comput. Appl. Math., 9, pages 1–92. Walter de Gruyter, Berlin, 2010.

[60] M. Reed and B. Simon. *Methods of modern mathematical physics. I. Functional analysis*. Academic Press, Inc., New York, 2nd edition, 1980.

[61] P. Rigollet and A. Tsybakov. Exponential screening and optimal rates of sparse estimation. *Ann. Statist.*, 39:731–771, 2011.

[62] M. Rudelson. Random vectors in the isotropic position. *J. Funct. Anal.*, 164:60–72, 1999.

[63] M. Rudelson and R. Vershynin. Sampling from large matrices: an approach through geometric functional analysis. *J. ACM*, 54:19 pp, 2007.

[64] M. Rudelson and R. Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Comm. Pure Appl. Math.*, 61:1025–1045, 2008.

[65] W. Rudin. *Functional analysis*. McGraw-Hill, Inc., New York, 2nd edition, 1991.

[66] A. Selberg. On the remainder in the formula for $N(T)$, the number of zeros of $\zeta(s)$ in the strip $0 < t < T$. *Avh. Norske Vid. Akad. Oslo. I.*, 1944:27 pp, 1944.

[67] A. Selberg. Contributions to the theory of the Riemann zeta-function. *Arch. Math. Naturvid.*, 48:89–155, 1946.

[68] A. Selberg. Old and new conjectures and results about a class of Dirichlet series. In *Proceedings of the Amalfi Conference on Analytic Number Theory (Maiori 1989)*, pages 367–385. Univ. Salerno, Salerno, 1992.

[69] K. Soundararajan. Moments of the Riemann zeta function. *Ann. of Math.*, 170:981–993, 2009.

[70] C. J. Stone. Additive regression and other nonparametric models. *Ann. Statist.*, 13:689–705, 1985.

[71] T. Suzuki and M. Sugiyama. Fast learning rate of multiple kernel learning: trade-off between sparsity and smoothness. *Ann. Statist.*, 41:1381–1405, 2013.

[72] M. Talagrand. New concentration inequalities in product spaces. *Invent. Math.*, 126:505–563, 1996.

[73] K. M. Tsang. *The Distribution of the values of the Riemann zeta-function*. PhD thesis, Princeton University, 1984.

[74] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, 2009.

[75] S. Van de Geer and A. Muro. The additive model with different smoothness for the components. http://arxiv.org/abs/1405.6584, 2014.

[76] S. A. Van de Geer. *Applications of empirical process theory*. Cambridge University Press, Cambridge, 2000.

[77] A. W. Van der Vaart and J. A. Wellner. *Weak convergence and empirical processes. With applications to statistics*. Springer, New York, 1996.

[78] J. Von Neumann. *Functional Operators. II. The Geometry of Orthogonal Spaces*. Princeton University Press, Princeton, 1950.

[79] M. Wahl. On the mod-Gaussian convergence of a sum over primes. *Math. Z.*, 276:635–654, 2014.

[80] M. J. Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inform. Theory*, 55:5728–5741, 2009.

[81] J. Weidmann. *Lineare Operatoren in Hilberträumen. Teil 1. Grundlagen*. B. G. Teubner, Stuttgart, 2000.