OPTIMAL ESTIMATION OF COMPONENTS IN STRUCTURED NONPARAMETRIC MODELS

MARTIN WAHL

Abstract. We consider the nonparametric random regression model
\[ Y = f_1(X_1) + f_2(X_2) + \epsilon \]
in which the function \( f_1 \) is the parameter of interest and the function \( f_2 \) is a nuisance parameter. We present a theory for estimating \( f_1 \) in settings where \( f_2(X_2) \) is more complex than \( f_1(X_1) \). The proposed estimation procedure combines two least squares criteria and can be written as an alternating projection procedure. We derive nonasymptotic risk bounds which reveal connections between the performance of our estimators of \( f_1 \) and the notions of minimal angles and Hilbert-Schmidt operators in the theory of Hilbert spaces. Under additional regularity conditions on the design densities, these bounds can be further improved. Our results establish general assumptions under which the estimators of \( f_1 \) have up to first order the same sharp upper bound as the corresponding estimators in the model \( Y = f_1(X_1) + \epsilon \). As an example we apply the results to an additive model where the number of covariates is large or the nuisance components are considerably less smooth than \( f_1 \).

1. Introduction

In this paper we consider the nonparametric random regression model
\[ Y = f_1(X_1) + f_2(X_2) + \epsilon, \]
where \( X_1 \) and \( X_2 \) are \( q_1 - \) and \( q - q_1 \)-dimensional random variables, respectively, and \( f_1 \) and \( f_2 \) are unknown regression functions. 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 part \( f_2(X_2) \) is more complex than the first part \( f_1(X_1) \). We focus on the case that both parts are nonparametric, i.e., on the case that the functions \( f_1 \) and \( f_2 \) are infinite-dimensional parameters. The estimation problem is similar to the one arising in semiparametric models in which 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. These models correspond to the case that \( q_1 = 1 \) and \( f_2 \) is an additive function of the various coordinates of \( X_2 \). One of the seminal theoretical papers is by Stone [29], who showed that each component can

2010 Mathematics Subject Classification. 62G08, 62G20, 62H05, 62H20.

Key words and phrases. Structured nonparametric models, projection on sumspaces, alternating projections, additive models, increasing number of covariates.
be estimated with the one-dimensional optimal rate of convergence. Since then many estimation procedures have been proposed, many of them consisting of several steps. In the papers of Linton [18] and Fan, Härdle, and Mammen [11] a preliminary undersmoothed estimator is updated in a second estimation step. In both papers it is shown that there exist estimators of single components which have the same asymptotic bias and variance as corresponding oracle estimators which know the other components.

Probably the most popular estimation procedures are the backfitting procedures which are empirical versions of the orthogonal projection on the subspace of additive functions in a Hilbert space setting (see, e.g., the book of Hastie and Tibshirani [12] and the references therein). This orthogonal projection was studied, e.g., by Breiman and Friedman [6]. 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 estimates leads to the backfitting procedures. Opsomer and Ruppert [23] and Opsomer [22] computed the asymptotic bias and variance of estimators which are based on the backfitting procedure in the case where the conditional expectation operators are replaced by the smoother matrices of the local polynomial regression. Mammen, Linton, and Nielsen [20] introduced the smooth backfitting procedure and showed that their estimators of single components achieve the same asymptotic bias and variance as oracle estimators which know the other components. 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 more than it is required in the Hilbert space setting (see [6]). In the work of Horowitz, Klemelä, and Mammen [13] a general two-step procedure was proposed in which a preliminary undersmoothed estimator is based on the smooth backfitting procedure of [20]. They also showed that there are estimators which are asymptotically efficient, i.e., achieve the asymptotic minimax risk, and that the corresponding constant is the same as in the case with only one component. In addition to the assumptions coming from the results in [20] they require a Lipschitz condition for all components.

The problem of estimating $f_1$ in cases where $f_2(X_2)$ is more complex than $f_1(X_1)$ is also considered in the paper by Efromovich [9] and the preprint by van de Geer [31]. In [9] 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 construction of the estimator is complicated and starts with a blockwise-shrinkage oracle estimator. The assumptions include smoothness and boundedness conditions on the full-dimensional density of $(X_1^T, X_2^T)^T$. Note that the model in [9] also allows the existence of an additional scale function in the error term depending on $X_1$ and $X_2$. In [31] a penalized least squares estimator is analyzed in cases where the function $f_2$ is less smooth than $f_1$. It is shown
that, under certain assumptions including smoothness and boundedness conditions on the design densities, the estimator achieves the optimal rate of convergence in both components, 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. Therefore, it seems to be an interesting problem to study for instance a backfitting procedure in the case that $q$ increases as $n$ increases.

Finally, we mention the recent work on high-dimensional additive models where the number of covariates is much larger than the number of observations. These models have been studied under certain sparsity constraints (see [21], [14], [17], [24], [7], and others). In the literature there are several closely related assumptions which make both the estimation of the whole model and the selection of the relevant covariates possible. These include restricted eigenvalue conditions and compatibility conditions.

In this paper we propose estimators of $f_1$ which are based on the composition of two least squares criteria. The first least squares criterion defines empirical versions of certain projections on sumspaces in a Hilbert space setting. We analyze these estimators of $f_1$ under the main assumption that there exists an upper bound on $\rho_0$, the cosine of the minimal angle between two closed subspaces of $L^2(\mathbb{P}^X_1)$ and $L^2(\mathbb{P}^X_2)$ which contain $f_1$ and $f_2$, respectively. This assumption allows for applying the theory of projections on sumspaces in Hilbert spaces (see, e.g., the book of Bickel, Klaassen, Ritov, and Wellner [4, Appendix 4]). We show that several geometric properties in the Hilbert space setting carry over to the finite sample setting with probability close to one. As a result we obtain that our estimation procedure can be written as a backfitting procedure (or more precisely, as an alternating projection procedure) which converges geometrically in terms of $\rho_0$ with probability close to one. In contrast to the aforementioned papers on the backfitting procedure, our approach is nonasymptotic and builds on recent concentration inequalities for structured random matrices (see, e.g., the work of Rauhut [25]). Under weak assumptions we prove several nonasymptotic risk bounds for our estimators of $f_1$. From these upper bounds we conclude that the estimators of $f_1$ achieve the same (nonasymptotic) optimal rate of convergence as the estimators in the model where $f_2$ is known if it is possible to choose a $d_2$-dimensional subspace $V_2$ of $L^2(\mathbb{P}^X_2)$ such that $\min_{g_2 \in V_2} \|f_2 - g_2\| = o(n^{-\alpha_1/(2\alpha_1+q_1)})$, while $\varphi^2 d_2 (\log n)/n \rightarrow 0$. Here, $\alpha_1$ is the degree of smoothness of $f_1$ and $\varphi$ is a real number (defined in Theorem 1 below) which depends on the degree of dependence between the coordinates of $(X_1^T, X_2^T)^T$. The real number $\varphi$ is allowed to increase as $n \rightarrow \infty$ meaning that we can apply these results to models where $q$ increases as $n \rightarrow \infty$.

As an example we apply these results to an additive model where $q$ is large generalizing several results obtained in [29], [20], and [13] in the case that $q$ is fixed, and we apply them to a model where $f_2$ is considerably less
smooth than \( f_1 \) relaxing the smoothness conditions required, e.g., in the results of [9] and [31]. Consider the latter and suppose for simplicity that \( q = 2 \) and \( q_1 = 1 \). Denote by \( \alpha_2 \) the degrees of smoothness of \( f_2 \). Then we find that the condition \( \alpha_2 > \alpha_1/(2\alpha_1 + 1) \) is sufficient. In semiparametric models one usually requires a global rate of convergence of order \( o(n^{-1/4}) \) to obtain the rate of convergence \( n^{-1/2} \) of the parametric component (see, e.g., the book of van de Geer [30, Chapter 11]). Since \( \alpha_1/(2\alpha_1 + 1) \) goes to 1/2 as \( \alpha_1 \to \infty \) our condition extends this result to the nonparametric case. Under additional smoothness conditions on the design densities we further relax this condition. For each fixed \( x_1 \) suppose that \( p(x_1, \cdot)/(p_1(x_1)p_2(\cdot)) \) has degree of smoothness \( \beta \) (\( p \), \( p_1 \), and \( p_2 \) are the corresponding densities, see Section 3). Then we find that the condition \( \alpha_2 + \beta > \alpha_1/(2\alpha_1 + 1) \) is sufficient. We also find generalizations of these conditions in models with \( q > 2 \) (for instance, the condition \((\alpha_2 + \beta)/(q - q_1) > \alpha_1/(2\alpha_1 + q_1) \) in the general case).

The paper is organized as follows. In Section 2 we present the assumptions on the model and state our main results in Theorems 1-4. In Section 3 we apply our results to an additive model. The proofs of our results are given in Sections 4-7.

2. Main Results

Let \((Y, X)\) be a pair of random variables such that

\[
Y = f(X) + \epsilon = f_1(X_1) + f_2(X_2) + \epsilon, \tag{2.1}
\]

where \( X_1 \) and \( X_2 \) take values in \( \mathbb{R}^{q_1} \) and \( \mathbb{R}^{q_2} \), respectively, \( q = q_1 + q_2 \), \( X = (X_1^T, X_2^T)^T \), and \( \epsilon \) satisfies \( \mathbb{E}[\epsilon|X] = 0 \) and \( \mathbb{E}[\epsilon^2|X] = \sigma^2 \). Concerning the unknown regression functions, we make the following assumption:

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

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

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.

Thus \( f_1 \) and \( f_2 \) are supposed to have finite 2-norm. Structural assumptions on \( f_2 \), as for instance additivity (see, e.g., Section 3), should be incorporated into the model by making assumptions on \( H_2 \). The spaces \( H_1 \) and \( H_2 \) are contained in \( L^2(\mathbb{P}^X) \) in a canonical way which is a Hilbert space with the inner product \( \langle g, h \rangle = \mathbb{E}[g(X)h(X)] \) and the corresponding norm \( \|g\| = \sqrt{\langle g, g \rangle} \). In order to state our main assumption, we give the following definition (see [15] 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 the inner product \( \langle \cdot, \cdot \rangle \). 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

\[
\rho_0 = \rho_0(\mathcal{H}_1, \mathcal{H}_2) = \sup\{\langle h_1, h_2 \rangle/(\|h_1\|\|h_2\|) \mid h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\}.
\]
Assumption 2. Suppose that the cosine of the minimal angle between $H_1$ and $H_2$ is strictly less than 1, i.e.,

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

We observe $n$ independent copies $(Y^i, X^i) = (Y^i, ((X^i_1)^T, (X^i_2)^T)^T)$, $i = 1, \ldots, n$, of $(Y, X)$. Based on this sample, we want to estimate the function $f_1$ by using least squares procedures. Let $V_1 \subseteq H_1$ and $V_2 \subseteq H_2$ be finite-dimensional subspaces, and let $V = V_1 + V_2$. The least squares estimator $\hat{f}_V$ on the model $V$ is given (not uniquely) by

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

By Assumption 2, we have $V_1 \cap V_2 = \{0\}$. Therefore, each $g \in V$ can be decomposed uniquely as $g = g_1 + g_2$ with $g_1 \in V_1$ and $g_2 \in V_2$. In particular, $\hat{f}_V = (\hat{f}_V)_1 + (\hat{f}_V)_2$ with $(\hat{f}_V)_1 \in V_1$ and $(\hat{f}_V)_2 \in V_2$. We define

$$\hat{f}_{1,V_1,V_2} = (\hat{f}_V)_1 \text{ if } ||(\hat{f}_V)_1||_\infty \leq k_n \text{ and } \hat{f}_{1,V_1,V_2} = 0 \text{ otherwise},$$

where $k_n$ is a real number to be chosen later (compare to the work of Baraud [3] Eq. (3)). By applying a second least squares criterion, we construct another estimator. Let $W_1 \subseteq V_1$ be a subspace. We define $\hat{f}_{1,W_1,V_1,V_2}$ equal to

$$\arg \min_{g_1 \in W_1} \frac{1}{n} \sum_{i=1}^{n} ((\hat{f}_V)_1(X^i_1) - g_1(X^i_1))^2$$

if (2.3) has $\infty$-norm bounded by $k_n$ and equal to the function 0 otherwise. Note that both estimators coincide if $V_1 = W_1$. We will see that the dependence of the bias term on $\rho_0$ is reduced if the second least squares criterion is applied.

Notation 1. Denote by $\Pi_V$ the orthogonal projection from $\mathbb{R}^n$ to the subspace $\{g(X^1), \ldots, g(X^n)^T | g \in V\}$. Analogously, we define $\Pi_{V_1}$, $\Pi_{V_2}$, and $\Pi_{W_1}$. Moreover, denote by $\Pi_V$ (resp. $\Pi_{V_1}$, $\Pi_{V_2}$, and $\Pi_{W_1}$) the orthogonal projection from $L^2(\mathbb{P}^X)$ to the subspace $V$ (resp. $V_1$, $V_2$, and $W_1$).

Both estimators have a considerably simplified representation if we restrict functions to vectors in $\mathbb{R}^n$, evaluated at the observations $X^1, \ldots, X^n$. For this let $Y = (Y^1, \ldots, Y^n)^T$. Then the estimators $\hat{f}_{1,V_1,V_2}$ and $\hat{f}_{1,W_1,V_1,V_2}$ (as vectors in $\mathbb{R}^n$) are equal to $(\hat{\Pi}_V Y)_1$ and $\hat{\Pi}_{W_1}(\hat{\Pi}_V Y)_1$, respectively. These expressions are well-defined if the subspaces $\{g_j(X^1_1), \ldots, g_j(X^n_1)^T | g_j \in V_j\}$, $j = 1, 2$, have intersection equals 0 which is satisfied with probability close to one (see Section 4 and Remark 3). Finally, one may object that these estimators are 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 it by $(1/n) \sum_{i=1}^{n} g_1(X^i_1) = 0$. In Appendix B we show how our results carry over to these modified estimators.
It is known that Assumption 2 implies that $H_1 + H_2$ is closed. Therefore, the orthogonal projection from $L^2(P^X)$ to $H_1 + H_2$ exists. Moreover, it can be computed by an iterative algorithm which converges geometrically in terms of $\rho_0$ (see [4, Theorem 2 in Appendix A.4] and the references therein). Obviously, the least squares criterion in (2) defines an empirical version of this orthogonal projection. In Section 4 we apply concentration of measure results to show that the norm $\| \cdot \|$ and the empirical norm $\| \cdot \|_n$ are close to each other, uniformly on $V$ and with probability close to one. If this holds, we obtain an upper bound for the cosine of the minimal angle between $V_1$ and $V_2$, computed with respect to $\langle \cdot, \cdot \rangle_n$, where $\langle g, h \rangle_n = (1/n) \sum_{i=1}^{n} g(X_i)h(X_i)$ and $\|g\|_n = \sqrt{\langle g, g \rangle_n}$. Since our approach is nonasymptotic, we provide explicit constants throughout. This upper bound will be the basis in the proof of Theorem 1. Furthermore, the previous results will imply that our estimators can be computed by an iterative algorithm which again converges geometrically in terms of $\rho_0$. This fact will play a key role in the proofs of the Theorems 2-4.

It is our first aim to establish the following general nonasymptotic risk bound.

**Theorem 1.** Let $(Y^i, X^i)$, $i = 1, \ldots, n$, be $n$ independent copies of (2.1), and let Assumptions 1 and 2 hold. Let $V_1 \subseteq H_1$ be a $d_1$-dimensional subspace, and let $V_2 \subseteq H_2$ be a $d_2$-dimensional subspace. Let $V = V_1 + V_2$ and $d = d_1 + d_2$. Suppose that there exists a real number $\phi$ such that

$$\|g\|_\infty \leq \phi \sqrt{d \|g\|}$$

for all $g \in V$. Let $0 < \delta < 1$. Suppose that

$$\phi^2 d \leq \frac{n \delta^2}{\log n}.$$ 

Finally, let $\hat{f}_{1,V_1,V_2}$ be the estimator defined in (2.2). Then

$$E \left[ \|f_1 - \hat{f}_{1,V_1,V_2}\|^2 \right] \leq \frac{1 + \delta}{(1 - \delta)^2} \frac{1}{1 - \rho_0^2} \left( 2 \min_{g \in V} \|f - g\|^2 \right) + \frac{\sigma^2 d_1}{n} + R_n$$

with

$$R_n = \frac{2 \phi^2 d \|f_1\|^2 \|f\|^2 + \sigma^2}{(1 - \delta)(1 - \rho_0^2)k_2^2} + 2^{3/4} \left( \|f_1\| + k_n \right)^2 d \exp \left( - \frac{n \delta^2}{\phi^2 d} \kappa \right),$$

where $\kappa$ is the constant in Theorem 5.

**Remark 1.** If we apply this result to an asymptotic scenario, then for suitable chosen $k_n$ the remainder $R_n$ decreases faster than any power of $1/n$ if $\phi^2 d (\log n)/(n \delta^2) \to 0$ (compare this condition to [19, Eq. (1.6)]).

In Theorem 1 the bias term also depends on the function $f_2$. It can be made smaller by choosing a large space $V_2$ satisfying (2.4) and (2.5) without affecting the variance term. The assumptions in Theorem 1 are quit weak. Besides the (minimal) Assumptions 1 and 2, we only need some additional
conditions on the chosen statistical model \( V \). Firstly, we need an upper bound for the \( \infty \)-norm of functions of \( V \) in terms of their 2-norm. If \( q = 2 \), this is a weak additional assumption (see, e.g., (3.6)). In general \( \varphi \) depends on the amount of dependence between the coordinates of \( X \) (see Section 3). Secondly, we also need a weak upper bound on the dimension of \( V \).

To continue, we make a further assumption.

**Assumption 3.** Suppose that there exist measures \( \nu_1 \) and \( \nu_2 \) 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 densities of \( X_1 \) and \( X_2 \) with respect to the measures \( \nu_1 \) and \( \nu_2 \), respectively. Suppose that

\[
\|\Pi\|_{HS}^2 := \int \int \left( \frac{p(x_1, x_2)}{p_1(x_1)p_2(x_2)} \right)^2 p_1(x_1)p_2(x_2) d\nu_1(x_1)d\nu_2(x_2) < \infty. \tag{2.7}
\]

If \( \Pi : L^2(\mathbb{P}^{X_2}) \to L^2(\mathbb{P}^{X_1}) \), \( g_2 \mapsto \int g_2(x_2)(p(x_1, x_2)/p_1(x_1))d\nu_2(x_2) \), is the conditional expectation operator, then (2.7) implies that \( \Pi \) is Hilbert-Schmidt (see [2, Def. 3.4.1] and also [26, Theorem VI.23] for a converse statement). It is known that (2.7) and the fact \( \{g_1 \in L^2(\mathbb{P}^{X_1}) \mid \mathbb{E}[g_1(X_1)] = 0\} \cap L^2(\mathbb{P}^{X_2}) = \{0\} \) imply that \( \rho_0(\{g_1 \in L^2(\mathbb{P}^{X_1}) \mid \mathbb{E}[g_1(X_1)] = 0\}, L^2(\mathbb{P}^{X_2})) < 1 \), a strengthening of Assumption 2 (see [4, Proposition 2 in Appendix A.4]).

We prove:

**Theorem 2.** Let \((Y', X')\), \( i = 1, \ldots, n \), be \( n \) independent copies of (2.1), and let Assumptions 1, 2, and 3 hold. For \( j = 1, 2 \) let \( V_j \subseteq H_j \) be a \( d_j \)-dimensional subspace. Let \( V = V_1 + V_2 \) and \( d = d_1 + d_2 \). Let \( \varphi \) and \( 0 < \delta < 1 \) be real numbers such that (2.4) and (2.5) of Theorem 1 are satisfied. In addition, let \( W_1 \subseteq V_1 \) be a subspace, and let \( \hat{f}_{1,W_1,V_1,V_2} \) be the estimator defined in (2.3). Then

\[
\mathbb{E}\left[ \|f_1 - \hat{f}_{1,W_1,V_1,V_2}\|^2 \right] \leq \frac{(1 + \delta)^2}{1 - \delta} \left( 2 \min_{g_1 \in W_1} \|f_1 - g_1\|^2 + \frac{\sigma^2 \dim W_1}{n} \right) + 2 \left( 1 + \frac{1}{\delta} \right) \frac{1 + \delta}{(1 - \delta)^2} \frac{1}{1 - \rho_0^2} \min_{\varphi \in V} \|f - g\|^2 + R_{n2}
\]

with

\[
R_{n2} = \frac{(1 + \delta)^2}{(1 - \delta)^4} \frac{1}{1 - \rho_0^2} \left( \frac{\sigma^2 \|\Pi\|_{HS}^2}{n} + \frac{\varphi^2 d \sigma_1^2}{n} \right) + R_n, \tag{2.8}
\]

where \( R_n \) is given by (2.6) and \( \|\Pi\|_{HS}^2 \) is given in Assumption 3.

In Theorem 2, the first two terms on the right hand side are equal to the bias term and the variance term of the same estimator in the model \( Y = f_1(X_1) + \epsilon \). In particular, they no longer depend on \( \rho_0 \) as in Theorem 1. The third term contains the bias term of \( f \) in the model \( V \). It can be made smaller by choosing large spaces \( V_1, V_2 \) satisfying (2.4) and (2.5) without affecting the first two terms. Thus the theorem shows that, up to remainder terms, the estimators of \( f_1 \) have the same upper bound as the estimators in the model \( Y = f_1(X_1) + \epsilon \).
The following two theorems are devoted to a further analysis of the bias term. Under additional regularity conditions on the design densities, we show that the dependence of the bias term on the function \( f_2 \) decreases substantially. Theorem 3.3 is applicable if \( q \) is small, while Theorem 3.4 is designed for the case that \( q \) is large (see Section 3).

**Theorem 3.** Let \( (Y^i, X^i), i = 1, \ldots, n, \) be \( n \) independent copies of (2.1), and let Assumptions 1, 2, and 3 hold. For \( j = 1, 2 \) let \( V_j \subseteq H_j \) be a \( d_j \)-dimensional subspace. Let \( V = V_1 + V_2 \) and \( d = d_1 + d_2 \). Let \( \varphi \) and \( 0 < \delta < 1 \) be real numbers such that (2.4) and (2.5) of Theorem 1 are satisfied. Suppose that \( \|g_1\|_\infty \leq \varphi \sqrt{d_1} \|g_1\| \) for all \( g_1 \in V_1 \). Moreover, suppose that there exist real numbers \( \phi(V_2) \), \( \psi(V_2) \) and a function \( h_1 \in L^2(\mathbb{P}^{X_2}) \) such that \( \|f_2 - \Pi V_2 f_2\| \leq \phi(V_2) \) and

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

(2.9)

for all \( x_1 \). In addition, let \( W_1 \subseteq V_1 \) be a subspace, and let \( \hat{f}_{1,W_1,V_1,V_2} \) be the estimator defined in (2.3). Then

\[
\mathbb{E} \left[ \|f_1 - \hat{f}_{1,W_1,V_1,V_2}\|^2 \right] \leq \frac{(1 + \delta)^2}{1 - \delta} \left( 2 \min_{g_1 \in W_1} \|f_1 - g_1\|^2 + \frac{\sigma^2 \dim W_1}{n} \right) + 2 \left( 1 + \frac{1}{\delta} \right) \left( \frac{(1 + \delta)^4}{(1 - \delta)^5} \frac{1}{\rho_0^2} \right) \|h_1\|^2 (\psi(V_2)\phi(V_2))^2 + 8 \left( 1 + \frac{1}{\delta} \right) \left( \frac{1 + \delta}{1 - \delta} \right) \min_{g_1 \in V_1} \|f_1 - g_1\|^2 + R_{n3},
\]

where

\[
R_{n3} = 2 \left( 1 + \frac{1}{\delta} \right) \left( \frac{(1 + \delta)^4}{(1 - \delta)^5} \frac{1}{\rho_0^2} \right) \|h_1\|^2 (\psi(V_2)\phi(V_2))^2 \frac{\|f_2 - \Pi V_2 f_2\|_\infty^2}{n} + 2 \left( 1 + \frac{1}{\delta} \right) \left( \frac{(1 + \delta)^4}{(1 - \delta)^5} \frac{1}{\rho_0^2} \right) \frac{\varphi^2 d_1}{n} \|f_2 - \Pi V_2 f_2\|^2 + R_{n2}
\]

and \( R_{n2} \) is given by (2.5).

In Theorem 3.3, the term in the second line is of smaller order if \( f_2 \) and \( p(x_1, \cdot)/p_1(x_1)p_2(\cdot) \) are regular enough in the sense that \( (\psi(V_2)\phi(V_2))^2 \) is of smaller order than the first two terms. The first term in the third line contains the bias term of \( f_1 \) in the model \( V_1 \). It can easily be made smaller by choosing a large space \( V_1 \) satisfying (2.4) and (2.5) without affecting the previous terms. Thus the estimator of \( f_1 \) behaves well in cases where \( f_2 \) might be considerably less regular than \( f_1 \), i.e. in cases where Theorems 1 and 2 are not applicable.

**Theorem 4.** Let \( (Y^i, X^i), i = 1, \ldots, n, \) be \( n \) independent copies of (2.1), and let Assumptions 1, 2, and 3 hold. For \( j = 1, 2 \) let \( V_j \subseteq H_j \) be a \( d_j \)-dimensional subspace. Let \( V = V_1 + V_2 \) and \( d = d_1 + d_2 \). Let \( \varphi \) and \( 0 < \delta < 1 \) be real numbers such that (2.4) and (2.5) of Theorem 1 are satisfied.
For fixed \( x_1 \) let \( r \) be the orthogonal projection of \( p(x_1,\cdot)/(p_1(x_1)p_2(\cdot)) \) on \( H_2 \) (which by (2.7) is defined for \( \mathbb{P}^{X_1} \)-almost all \( x_1 \)). Suppose that there exist real numbers \( \phi(V_2),\psi(V_2) \) and a function \( h_1 \in L^2(\mathbb{P}^{X_1}) \) such that \( \|f_2 - \Pi V_2 f_2\| \leq \phi(V_2) \) and
\[
\|r(x_1,\cdot) - \Pi V_2 (r(x_1,\cdot))\|_{L^2(\mathbb{P}^{X_2})} \leq \psi(V_2)h_1(x_1)
\]
(2.10) for \( \mathbb{P}^{X_1} \)-almost all \( x_1 \). In addition, let \( W_1 \subseteq V_1 \) be a subspace, and let \( \hat{f}_{1,W_1,V_1, V_2} \) be the estimator defined in (2.3). Then
\[
\mathbb{E}\left[\|f_1 - \hat{f}_{1,W_1,V_1, V_2}\|^2\right] \leq \frac{(1+\delta)^2}{1-\delta} \left( \frac{2}{n} \min_{g_1 \in W_1} \|f_1 - g_1\|^2 + \frac{\sigma^2 \dim W_1}{n} \right) + 4 \left( 1 + \frac{1}{\delta} \right) \frac{1+\delta}{1-\delta} \frac{1}{(1-\rho_0^2)^2} \|h_1\|^2 (\psi(V_2)\phi(V_2))^2 \\
+ 4 \left( 1 + \frac{1}{\delta} \right) \frac{1+\delta}{1-\delta} \frac{1}{1-\rho_0^2} \min_{g_1 \in W_1} \|f_1 - g_1\|^2 + R_n4,
\]
where
\[
R_n4 = 2 \left( 1 + \frac{1}{\delta} \right) \frac{1+\delta}{(1-\delta)^2} \frac{1}{1-\rho_0^2} \frac{\varphi^2}{n} \|f - \Pi V f\|^2 + R_n2
\]
and \( R_n2 \) is given by (2.8).

3. Applications

In this section we consider the nonparametric additive regression model
\[
Y = f(X) + \epsilon = f_1(X_1) + \cdots + f_q(X_q) + \epsilon,
\]
(3.1)
where the random variables \( X_1,\ldots,X_q \) take values in \([0,1]\), the regression error satisfies \( \mathbb{E}[\epsilon|X] = 0 \) and \( \mathbb{E}[\epsilon^2|X] = \sigma^2 \), and \( X = (X_1,\ldots,X_q)^T \). Note that in this section we change our notation slightly which leads to an abuse of notation: we write \((X_2,\ldots,X_q)^T\) instead of \(X_2\) and \((x_2,\ldots,x_q)^T\) instead of \(x_2\). Moreover, we suppose that \( f_2(x_2) \) can be written as \( f_2(x_2) + \cdots + f_q(x_q) \), and later we take \( V_2 \) equal to \( V_2 + \cdots + V_q \). Let Assumptions \#1 and \#2 hold with \( H_1 = \{g_1 \in L^2(\mathbb{P}^{X_1})|\mathbb{E}[g_1(X_1)] = 0\} \) and \( H_2 = L^2(\mathbb{P}^{X_2}) + \cdots + L^2(\mathbb{P}^{X_q}) \), and let Assumption \#3 hold with \( \nu_1 \) and \( \nu_2 \) being the Lebesgue measures on \([0,1]\) and \([0,1]^{q-1}\), respectively. For \( j = 1,\ldots,q \) we assume that the density \( p_j \) of \( X_j \) with respect to the Lebesgue measure is bounded from below by \( c > 0 \), and that the function \( f_j \) is contained in the Hölder-space \( \mathcal{H}(\alpha_j,K_j) \) with \( \alpha_j > 0 \). Finally, we assume that there exists an \( \epsilon_q < 1 \) satisfying the following (compare to [17] Section 2.2)): for each \( h \in L^2(\mathbb{P}^{X_1}) + \cdots + L^2(\mathbb{P}^{X_q}) \) there exists a decomposition \( h = h_1 + \cdots + h_q \), \( h_j \in L^2(\mathbb{P}^{X_j}) \), \( j = 1,\ldots,q \), such that
\[
\|h\|^2 \geq (1-\epsilon_q)(\|h_1\|^2 + \cdots + \|h_q\|^2).
\]
(3.2)

Remark 2. Note that in models with an increasing number of covariates, \( 1-\epsilon_q \) might go to zero as \( n \to \infty \). If \( q = 2 \), then (3.2) holds by Assumption \#2 and Lemma \#2 below with \( \epsilon_q = \rho_0 \). If \( X_2,\ldots,X_q \) are independent, then
\((3.2)\) holds again with \(\epsilon_q = \rho_0\). In the general case \((3.2)\) is fulfilled for some \(\epsilon_q < 1\) if \(H(j) = L^2(\mathbb{P}^{X_j}) + \cdots + L^2(\mathbb{P}^{X_j})\) is closed for all \(j = 1, \ldots, q\). This can be seen as follows. By iteration (starting at \(j = q - 1\)) one can show that \((3.2)\) holds, e.g., for

\[1 - \epsilon_q = \prod_{j=1}^{q-1} \left(1 - \rho_0 \left(\left( H(j), \left( H(j) \cap L^2(\mathbb{P}^{X_{j+1}})\right)^c \cap L^2(\mathbb{P}^{X_{j+1}}) \right) \right) \right).\]

If \(H_1\) and \(H_2\) are two closed subspaces of a Hilbert space \(H\), then the assertions \(H_1 + H_2\) is closed and \(\rho_0(H_1,(H_1 \cap H_2)^c \cap H_2) < 1\) are equivalent. This follows from [16, Theorem 1a] and Lemma 2. Thus we have \(\epsilon_q < 1\).

In this section we want to apply Theorems 1111. First we construct \(V_j \subseteq L^2(\mathbb{P}^{X_j})\), \(j = 1, \ldots, q\), such that \(V = V_1 + \cdots + V_q\) satisfies (2.3) and (2.4). For \(j = 1, \ldots, q\) let \(P_{l_j,m_j}\) be the space of piecewise polynomials in the variable \(x_j\) on \([0, 1]\) of order \(l_j\) and with breakpoints \(0, 1/m_j, 2/m_j, \ldots, 1\), and let \(S_{l_j,m_j} \subseteq P_{l_j,m_j}\) be the spline space of the same order and with the same breakpoints (see [8, Chapter VII, VIII]). The spline spaces satisfy \(d_j := \dim S_{l_j,m_j} = m_j + l_j - 1\). If \(l_j \geq [\alpha_j] + 1\), then

\[
\inf_{g_j \in S_{l_j,m_j}} \| h_j - g_j \|_\infty \leq c_{\alpha_j} K_j d_j^{-\alpha_j} \tag{3.3}
\]

for all \(h_j \in \mathcal{H}(\alpha_j, K_j)\) and for all \(j = 2, \ldots, q\), where \(c_{\alpha_j}\) is a constant depending on \(\alpha_j\) and \(l_j\). Furthermore, if \(l_1 \geq [\alpha_1] + 1\), then also

\[
\inf_{g_1 \in S_{l_1,m_1} \cap H_1} \| h_1 - g_1 \|_\infty \leq c_{\alpha_1} K_1 d_1^{-\alpha_1} \tag{3.4}
\]

for all \(h_1 \in \mathcal{H}(\alpha_1, K_1) \cap H_1\), where \(c_{\alpha_1}\) is again a constant depending on \(\alpha_1\) and \(l_1\) (for these facts see [8, Chapter XII]).

We choose \(V_1 = S_{l_1,m_1} \cap H_1\) with \(l_1 \geq [\alpha_1] + 1\) and \(V_j = S_{l_j,m_j}\) with \(l_j \geq [\alpha_j] + 1\) for \(j = 2, \ldots, q\). We let \(V = V_1 + \cdots + V_q\). Before we choose the \(m_j\)s such that (2.5) is satisfied, we determine a constant \(\varphi\) satisfying (2.4). Since \(P_{l_j,m_j}\) is the orthogonal sum of \(m_j l_j\)-dimensional subspaces and since one can build an orthogonal basis by the shifted Legendre polynomials of degree \(\leq l_j\), one can show that (use the integral and boundedness properties of the Legendre polynomials; see, e.g., [33, Chapter XV])

\[
\|g_j\|_\infty^2 \leq l_j^2 m_j \int_0^1 (g_j(x_j))^2 dx_j
\]

for all \(g_j \in P_{l_j,m_j}\). Let \(l = \max_{j=1,\ldots,q} l_j\). Since the density \(p_j\) is bounded from below by \(c > 0\), we obtain

\[
\|g_j\|_\infty \leq \sqrt{\frac{l_j^2}{c} \sqrt{d_j}} \|g_j\| \tag{3.5}
\]
for all \( g_j \in V_j \). If \( q = 2 \), we conclude that

\[
\| g_1 + g_2 \|_\infty \leq \sqrt{\frac{l^2}{c} (\sqrt{d_1} \| g_1 \| + \sqrt{d_2} \| g_2 \|)} \\
\leq \sqrt{\frac{l^2}{c(1 - \rho_0)} d_1 + d_2 \| g_1 + g_2 \|}
\]

(3.6) for all \( g_1 \in V_1 \) and \( g_2 \in V_2 \), where the latter inequality follows from the Cauchy-Schwarz inequality, Assumption 2, and Lemma 2. Thus in the case \( q = 2 \), (2.4) is satisfied with \( \varphi^2 = l^2/(c(1 - \rho_0)) \). In the general case, (2.4) is satisfied with \( \varphi^2 = l^2/(c(1 - \epsilon_q)) \). This can be seen as in (3.6) by applying (3.2). We choose

\[
d_j = \left[ \frac{a_j n}{(\log n)^2 \varphi^2} \right] = \left[ \frac{c(1 - \epsilon_q)a_j n}{l^2 (\log n)^2} \right], \quad j = 1, \ldots, q,
\]

(3.7) with \( \sum_{j=1}^{q} a_j \leq 1 \). Then (2.5) is satisfied, where \( \delta \) is allowed to go to zero as \( n \to \infty \). We choose \( \delta \to 0 \) such that the lower order terms in Theorem 2 still remain of lower order. Finally, let \( W_1 = S_{l_1,m_{W_1}} \cap H_1 \) such that \( W_1 \subseteq V_1 \). Observing \( n \) independent copies of \((Y, X)\), we construct the estimator \( \hat{f}_{1,W_1,v_1,v_2} \) with the above choices. Applying Theorem 2, the bound \( \min_{g \in V} \| f - g \| = \| f - \Pi_V f \| \leq \sum_{j=1}^{q} \| f_j - \Pi_{V_j} f_j \| \), and (3.3), (3.4), we obtain:

**Corollary 1.** Let the assumptions at the beginning of this section hold. Then

\[
\mathbb{E} \left\| f_1 - \hat{f}_{1,W_1,v_1,v_2} \right\|^2 \leq \frac{(1 + \delta)^2}{1 - \delta} \left( 2c_{\alpha_1}^2 K_1^2 (\dim W_1)^{-2\alpha_1} + \frac{\sigma^2 \dim W_1}{n} \right) \\
+ 2 \frac{(1 + \delta)^2}{\delta(1 - \delta)^2} \frac{1}{1 - \rho_0} \left( \sum_{j=1}^{q} c_{\alpha_j} K_j \left( \frac{c(1 - \epsilon_q)a_j n}{l^2 (\log n)^2} \right)^{-\alpha_j} \right)^2 + R_{n2}.
\]

We choose

\[
\dim W_1 = \left\lfloor \left( \frac{2\alpha_1 2c_{\alpha_1}^2 K_1^2 n}{\sigma^2} \right)^{\frac{1}{2\alpha_1 + 1}} \right\rfloor.
\]

In order that \( W_1 \subseteq V_1 \) we need that \( m_1 \) is a multiple of \( m_{W_1} \), which can be achieved by choosing \( a_1 \) appropriately. If \( q \) is fixed, we obtain:

**Corollary 2.** Let the assumptions at the beginning of this section hold, and let \( q \) be fixed. Let \( \alpha_1 > 0 \) and \( \alpha_j > \alpha_1/(2\alpha_1 + 1) \), \( j = 2, \ldots, q \) (is satisfied if, e.g., \( \alpha_1 > 0 \) and \( \alpha_j > 1/2, j = 2, \ldots, q \)). Then

\[
\limsup_{n \to \infty} \mathbb{E} \left[ n^{2\alpha_1/(2\alpha_1 + 1)} \| f_1 - \hat{f}_{1,W_1,v_1,v_2} \|^2 \right] \leq \frac{2\alpha_1 + 1}{2\alpha_1} \frac{4\alpha_1}{\sigma^{2\alpha_1 + 1}} (4\alpha_1 c_{\alpha_1}^2 K_1^2)^{\frac{1}{2\alpha_1 + 1}}.
\]

(3.8)
As explained in the introduction, the condition on the $\alpha_j$'s generalizes a condition which is usually required in a semiparametric model. Next, we apply Theorem 3 in the case $q = 2$. We suppose that for each fixed $x_1$, we have

$$p(x_1, x_2) \in \mathcal{H}(\beta, h_1(x_1))$$

with $h_1 \in L^2(\mathbb{P}^{X_1})$. We make the same choices as above with $a_1 \leq 1/2$ and $a_2 = 1/2$. In addition, $l_2$ has to satisfy the bounds $l_2 \geq |\alpha_2| + 1$ and $l_2 \geq |\beta| + 1$. Then (2.9) is satisfied with this $h_1$ and $\psi(V_2) = c_\beta (\dim V_2)^{-\beta}$, where $c_\beta$ is a constant depending on $\beta$ and $l_2$. Thus

$$\|h_1\|\psi(V_2)\phi(V_2) \leq c_\alpha c_\beta \|h_1\|K_2 \left(\frac{c(1 - \rho_0)n}{2l^2(\log n)^2}\right)^{-\alpha_2 - \beta}$$

and we obtain:

**Corollary 3.** Let the assumptions at the beginning of this section hold with $q = 2$. Let (3.9) be satisfied, and let $\alpha_1 > 0$ and $\alpha_2 + \beta > \alpha_1/(2\alpha_1 + 1)$ (is satisfied if, e.g., $\alpha_1, \alpha_2 > 0$ and $\beta \geq 1/2$). Then (3.8) holds.

Finally, we apply Theorem 4. In the particular case that $X_2, \ldots, X_q$ are independent, one can show that (see (3.12))

$$r(x_1, x_2, \ldots, x_q) = \sum_{j=2}^q \frac{p_{jk}(x_1, x_j)}{p_j(x_1)p_j(x_j)} = (q - 2),$$

where $p_{jk}(x_j, x_k)$ denotes the joint density of $(X_j, X_k)$ and $p_j(x_j)$ denotes the density of $X_j$. Thus in this case condition (2.10) is much weaker than condition (2.9). In the general case we have a similar result:

**Lemma 1.** Let the assumptions at the beginning of this section hold. In addition, suppose that for each fixed $x_j$

$$\frac{p_{jk}(x_j, x_k)}{p_j(x_j)p_k(x_k)} \in \mathcal{H}(\beta, h_{jk}(x_j))$$

with $h_{jk} \in L^2(\mathbb{P}^{X_j})$, for all $j = 1, \ldots, q$, $k = 2, \ldots, q$. For $k = 2, \ldots, q$ let $l_j \geq |\beta| + 1$. Then (2.10) is satisfied with $\psi_1(V_2) = \sqrt{\sum_{k=2}^q d_k^{-2\beta}}$ and

$$h_1(x_1) = \sqrt{\sum_{k=1}^q c^2_{\beta} h^2_{1k}(x_1)}$$

$$+ \sqrt{\frac{c^2_{\beta}}{1 - c_q} \int \left(\frac{p(x_1, \ldots, x_q)}{p_1(x_1)p_2, \ldots, q(x_2, \ldots, x_q)}\right)^2 p_2, \ldots, q(x_2, \ldots, x_q) dx_2 \cdots dx_q},$$

where $p_2, \ldots, q$ denotes the density of $(X_2, \ldots, X_q)$ and $c' = \sum_{j,k=2,j\neq k} \|h_{jk}\|^2$. Note that $h_1 \in L^2(\mathbb{P}^{X_1})$ by Assumption 3.
Proof. Let \( x_1 \) be a fixed real number. By the projection theorem, the expression

\[
\int \left( \frac{p(x_1, \ldots, x_q)}{p_1(x_1)p_2(x_2, \ldots, x_q)} - \sum_{k=2}^{q} g_k(x_k) \right)^2 p_2, \ldots, q(x_2, \ldots, x_q) dx_2 \cdots dx_q,
\]

subject to the constraints \( g_k \in L^2(\mathbb{P}^{X_k}) \), \( k = 2, \ldots, q \), is minimized by \( r = r_2 + \cdots + r_q \). Suppose that \( r_2, \ldots, r_q \) are chosen such that (3.2) is satisfied. We have

\[
\mathbb{E} \left[ \frac{p(x_1, X_2, \ldots, X_q)}{p_1(x_1)p_2(x_2, \ldots, X_q)} \bigg| X_k = x_k \right] = \frac{p_{1k}(x_1, x_k)}{p_1(x_1)p_k(x_k)},
\]

(3.12)

\( k = 2, \ldots, q \). Thus the \( r_j \)'s satisfy the \( q - 1 \) equations

\[
 r_k(x_k) = \frac{p_{1k}(x_1, x_k)}{p_1(x_1)p_k(x_k)} - \sum_{j=2,j\neq k}^{q} \int r_j(x_j) \frac{p_{jk}(x_j, x_k)}{p_j(x_j)p_k(x_k)} p_j(x_j) dx_j,
\]

for \( \mathbb{P}^{X_k} \)-almost all \( x_k \), \( 2 \leq k \leq q \) (note that we omit the dependence of the \( r_j \)'s on \( x_1 \)). By (3.10) 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 \( \mathcal{H}(\beta, \sum_{j=2,j\neq k}^{q} \| r_j \|_{L^2(\mathbb{P}^{X_j})} \| h_{jk} \|) \), respectively. We conclude that

\[
\| r - \Pi V_2 + \cdots + V_q r \|_{L^2(\mathbb{P}^{X_2, \ldots, X_q})} \leq \sum_{k=2}^{q} \| r_k - \Pi V_k r_k \|_{L^2(\mathbb{P}^{X_k})}
\]

\[
\leq \sum_{k=2}^{q} c_{\beta} \left( h_{1k}(x_1) + \sum_{j=2,j\neq k}^{q} \| r_j \|_{L^2(\mathbb{P}^{X_j})} \| h_{jk} \| \right) \epsilon_k^{-\beta}.
\]

Applying the Cauchy-Schwarz inequality and (3.2), this is bounded by

\[
\leq \sum_{k=2}^{q} c_{\beta} \left( h_{1k}(x_1) + \frac{\| r \|^2_{L^2(\mathbb{P}^{X_2, \ldots, X_q})}}{1 - \epsilon_q} \sum_{j=2,j\neq k}^{q} \| h_{jk} \|^2 \right) \epsilon_k^{-\beta}.
\]

Applying the Cauchy-Schwarz inequality again and the fact that orthogonal projections lower the norm, we obtain (3.11). This completes the proof.

We state a corollary of Lemma 11 and Theorem 4 in the case that \( p \) is fixed. Because of the first term in \( R_{n4} \), we get a stronger condition on \( \beta \).

**Corollary 4.** Let the assumptions at the beginning of this section hold with \( q \) fixed. In addition, suppose that for each fixed \( x_j \)

\[
\frac{p_{jk}(x_j, x_k)}{p_j(x_j)p_k(x_k)} \in \mathcal{H}(\beta, h_{jk}(x_j))
\]

with \( h_{jk} \in L^2(\mathbb{P}^{X_j}) \), for all \( j = 1, \ldots, q \), \( k = 2, \ldots, q \). Let \( \alpha_1 > 0 \) and \( \alpha_j + \beta/(2\alpha_1 + 1) > \alpha_1/(2\alpha_1 + 1) \), \( j = 2, \ldots, q \). Then (3.8) holds.
4. Preliminaries

In the proofs of Theorems 1-4, we show in a first step that we can restrict our analysis to an event $\mathcal{E}_\delta$ with probability close to 1 on which $\| \cdot \|_n$ and $\| \cdot \|$ are close to each other, uniformly on $V$. This section is devoted to the construction of this event $\mathcal{E}_\delta$ using concentration of measure results. Moreover, if $\mathcal{E}_\delta$ holds, we upper bound the cosine of the minimal angle between $V_1$ and $V_2$ in $(V, \langle \cdot, \cdot \rangle_n)$ in terms of $\rho_0$. The main input is a combination of Talagrand’s inequality and Rudelson’s lemma (see [27, Theorem 1] and [28, Theorem 3.1]). We apply a version obtained by Rauhut [25, Theorem 7.3] which allows slightly better constants.

**Theorem 5** (Rauhut [25]). Let $X_1, \ldots, X_n$ be $n$ independent copies of a $q$-dimensional random variable $X$. Let $V \subseteq L^2(\mathbb{P}^X)$ be a $d$-dimensional subspace. Suppose that there exists a real number $\varphi$ such that
\begin{equation}
\| g \|_\infty^2 \leq \varphi^2 d \| g \|^2 \tag{4.1}
\end{equation}
for all $g \in V$. Let $0 < \delta < 1$. Then we have
\begin{equation}
(1 - \delta) \| g \|^2 \leq \| g \|_n^2 \leq (1 + \delta) \| g \|^2 \tag{4.2}
\end{equation}
with probability
\begin{equation}
\geq 1 - 2^{3/4} d \exp \left( -\frac{\kappa n \delta^2}{\varphi^2 d} \right),
\end{equation}
where $\kappa$ is a universal constant.

**Proof.** Let $b_1, \ldots, b_d$ be an orthonormal basis of $V$, and let
\[ B_n = (b_j, b_k) \] in $\mathbb{R}^d$.

Then [25, Theorem 7.3] yields
\begin{equation}
\mathbb{P}(\| B_n - I \|_{op} \leq \delta) \geq 1 - 2^{3/4} d \exp \left( -\frac{\kappa n \delta^2}{\varphi^2 d} \right) \tag{4.3}
\end{equation}
for $0 < \delta < 1$, where $\| B_n - I \|_{op} = \sup_{x \in \mathbb{R}^d, \| x \|_2 = 1} \| (B_n - I)x \|_2$ denotes the operator norm. Here, $\| \cdot \|_2$ denotes the Euclidean norm. Note that condition (4.1) is sufficient to apply [28, Theorem 3.1] and it is also sufficient to apply [25, Theorem 7.3] although condition (4.2) is slightly stronger (in fact in the proof of [25, Theorem 7.3] only [25, (7.5)] is used). A function $g \in V$ with $\| g \| = 1$ can be written uniquely as $g = \sum_{j=1}^d x_j b_j$ with $x \in \mathbb{R}^d$ and $\| x \|_2 = 1$. Then we have $\| g \|_n^2 = x^T B_n x$ and thus
\begin{equation}
\sup_{g \in V, \| g \| = 1} \| g \|_n^2 - \| g \|^2 = \sup_{x \in \mathbb{R}^d, \| x \|_2 = 1} | x^T (B_n - I)x | = \| B_n - I \|_{op}, \tag{4.4}
\end{equation}
where the latter equality follows from the spectral theorem. (4.3) and (4.4) yield that $\| g \|_n^2 - \| g \|^2 \leq \delta \| g \|^2$ for all $g \in V$ with probability greater or equal to $1 - 2^{3/4} d \exp(-\kappa n \delta^2/(\varphi^2 d))$. Applying the triangle inequality gives (4.2). This completes the proof. \(\square\)

Under Assumptions 1 and 2 we obtain:
Corollary 5. Let $X^1, \ldots, X^n$ be $n$ independent copies of $X$, where $X = (X^1, X^2)^T$ is a $q$-dimensional random variable as in (2.1). Let Assumptions 1 and 3 hold. For $j = 1, 2$ let $V_j \subseteq H_j$ be a $d_j$-dimensional subspace. Let $V = V_1 + V_2$ and $d = \dim V = d_1 + d_2$. Suppose that $V$ satisfies (4.1). Let $0 < \delta < 1$, and let $E_\delta$ be the event defined by (4.2). Then we have $P(E_\delta) \geq 1 - 2^{3/4} d \exp(-\kappa n \delta^2/(q^2 d))$. Furthermore, if the event $E_\delta$ holds, then

\[
\|g_1 + g_2\|_n^2 \geq \frac{(1-\delta)}{(1+\delta)}(1 - \rho_0)(\|g_1\|_n^2 + \|g_2\|_n^2) \quad \text{and} \quad (4.5)
\]

\[
\|g_1 + g_2\|_n^2 \geq \frac{(1-\delta)}{(1+\delta)}(1 - \rho_0^2)\|g_1\|_n^2 \quad (4.6)
\]

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

\[
\frac{|\langle g_1, g_2 \rangle_n|}{\|g_1\|_n \|g_2\|_n} \leq 1 - \frac{(1-\delta)}{(1+\delta)}(1 - \rho_0) \quad (4.7)
\]

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

Corollary 5 is a consequence of Theorem 5 and the following lemma:

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

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

\[
\|h_1 + h_2\|^2 \geq (1 - \rho)(\|h_1\|^2 + \|h_2\|^2)
\]

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

\[
\frac{|\langle h_1, h_2 \rangle|}{\|h_1\| \|h_2\|} \leq \rho
\]

Furthermore, both assertions imply $\|h_1 + h_2\|^2 \geq (1 - \rho^2)\|h_1\|^2$ for all $h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2$.

Proof of Lemma 2. Suppose (i) is true, and 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 (i) we have $2 - 2\langle h_1, h_2 \rangle \geq 2(1 - \rho)$ which gives (ii).

Conversely, suppose (ii) is true, and let $h_1 \in \mathcal{H}_1$ and $h_2 \in \mathcal{H}_2$. Then we have $\|h_1 + h_2\|^2 \geq \|h_1\|^2 - 2\rho\|h_1\|\|h_2\| + \|h_2\|^2$ and (i) follows from the inequality $2\|h_1\|\|h_2\| \leq \|h_1\|^2 + \|h_2\|^2$. Furthermore, the inequality $2\rho\|h_1\|\|h_2\| \leq \rho^2\|h_1\|^2 + \|h_2\|^2$ implies that $\|h_1 + h_2\|^2 \geq (1 - \rho^2)\|h_1\|^2$. This completes the proof. \qed
Proof of Corollary 5. Let \( g_1 \in V_1 \) and \( g_2 \in V_2 \), and let \( \mathcal{E}_\delta \) hold. Applying (4.2) and Lemma 2 combined with Assumption 2, we obtain

\[
\|g_1 + g_2\|^2_n \geq (1 - \delta)\|g_1 + g_2\|^2 \geq (1 - \delta)(1 - \rho_0)(\|g_1\|^2 + \|g_2\|^2)
\]

and similarly

\[
\|g_1 + g_2\|^2_n \geq (1 - \delta)\|g_1 + g_2\|^2 \geq (1 - \delta)(1 - \rho^2_0)\|g_1\|^2
\]

Finally, (4.7) follow from (4.5) and Lemma 2. This completes the proof. □

5. Proof of Theorem 1

5.1. The variance term. In this subsection we prove a first explicit upper bound for the variance term which does not depend on the dimension of \( V_2 \). In Section 6, we will further improve this result.

Proposition 1. Let the assumptions of Theorem 1 hold. If the event \( \mathcal{E}_\delta \) of Corollary 5 holds, then

\[
\mathbb{E} \left[ \| (\hat{\Pi}_V \epsilon)_1 \|_n^2 | X^1, \ldots, X^n \right] \leq \frac{1 + \delta}{1 - \delta} \frac{1}{1 - \rho_0^2} \frac{\sigma^2 d_1}{n},
\]

where \( \epsilon = (\epsilon^1, \ldots, \epsilon^n)^T \) and \( \hat{\Pi}_V \) is defined in Notation 7.

Remark 3. Suppose that the event \( \mathcal{E}_\delta \) holds. Then \( \| \cdot \| \) and \( \| \cdot \|_n \) are equivalent norms on \( V \). In particular, each \( g \in V \) is determined uniquely by \((g(X^1_1), \ldots, g(X^n_1))^T \). This and the fact that \( V_1 \cap V_2 = 0 \) imply that \( \hat{\Pi}_V \epsilon \) can be decomposed uniquely as \( \hat{\Pi}_V \epsilon = (\hat{\Pi}_V \epsilon)_1 + (\hat{\Pi}_V \epsilon)_2 \) with \( (\hat{\Pi}_V \epsilon)_j \in \{(g_j(X^1_1), \ldots, g_j(X^n_1))^T | g_j \in V_j \}, j = 1, 2 \) (since the latter subspaces have intersection equals 0). As a consequence, \( \hat{\Pi}_V \epsilon_1 \) is well-defined on \( \mathcal{E}_\delta \).

The proof is based on a result of Ehrenfeld [10] Lemma 2.1 which was already applied in [20]:

Lemma 3 (Ehrenfeld [10]). 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 \( d_1 \leq d \). Suppose that \( v^T F v \geq v_1^T F_1 v_1 \) for all \( v = (v_1^T, v_2^T)^T \) with \( v_1 \in \mathbb{R}^{d_1} \) and \( v_2 \in \mathbb{R}^{d-d_1} \). Then

\[
\begin{pmatrix} w_1 \\ 0 \end{pmatrix}^T F^{-1} \begin{pmatrix} w_1 \\ 0 \end{pmatrix} \leq w_1^T F_1^{-1} w_1
\]

for all \( w_1 \in \mathbb{R}^{d_1} \).
Proof of Proposition 4. Let \( \varphi_1, \ldots, \varphi_d \) be a basis of \( V_1 \) and \( \varphi_{d+1}, \ldots, \varphi_d \) be a basis of \( V_2 \). Let
\[
    Z_1 = (\varphi_j(X_i^1))_{1 \leq i \leq d_1, 1 \leq j \leq d_1} \in \mathbb{R}^{n \times d_1},
\]
\[
    Z_2 = (\varphi_j(X_i^2))_{1 \leq i \leq d_1, 1 \leq j \leq d} \in \mathbb{R}^{n \times (d-d_1)},
\]
and \( Z = (Z_1|Z_2) \in \mathbb{R}^{n \times d} \). Suppose that the event \( \mathcal{E}_\delta \) holds. Then by Theorem 5 (see (4.3)) \( Z^T Z \) is invertible. Thus the minimum of \( \|e - Z\beta\|_n^2 \) is taken at \( \beta = (Z^T Z)^{-1}Z^T e \) and we have \( \hat{\Pi}_V e = Z(Z^T Z)^{-1}Z^T e \). This implies that \( \hat{\Pi}_V 1 \) is equal to \( (Z_1|0)(Z^T Z)^{-1}Z^T e \). We conclude that
\[
    \mathbb{E} \left[ \|(\hat{\Pi}_V e)_1\|_n^2 | X_1, \ldots, X_n \right] = \text{tr}((Z_1|0)(Z^T Z)^{-1}(Z_1|0)^T) \frac{\sigma^2}{n}.
\]
On the other hand, (5.6) is equivalent to the inequality
\[
    \beta^T Z^T Z \beta \geq \frac{(1 + \delta)(1 - \rho_0^2)}{(1 - \delta)} \beta_1^T Z_1 \beta_1
\]
for all \( \beta = (\beta_1^T, \beta_2^T)^T \) with \( \beta_1 \in \mathbb{R}^{d_1} \) and \( \beta_2 \in \mathbb{R}^{d-d_1} \). Applying Lemma 3 with \( F = Z^T Z \) and \( F_1 = ((1 - \delta)/(1 + \delta))(1 - \rho_0^2)Z_1 \), we obtain
\[
    \text{tr}((Z_1|0)(Z^T Z)^{-1}(Z_1|0)^T) \leq \frac{1 + \delta}{1 - \delta} \frac{1}{1 - \rho_0^2} \text{tr}(Z_1 Z_1^T)^{-1}Z_1^T.
\]
Since \( \hat{\Pi}_V = Z_1(Z_1^T Z_1)^{-1}Z_1^T \), we have \( \text{tr}(Z_1(Z_1^T Z_1)^{-1}Z_1^T) = \dim V_1 = d_1 \). This completes the proof. \( \square \)

5.2. End of the proof of Theorem 4. As defined in Notation 1 let \( \Pi_V \) and \( \Pi_{V_1} \) be the orthogonal projections from \( L^2(\mathbb{P}^X) \) to \( V \) and \( V_1 \), respectively. Applying the arguments of \( 3 \), we obtain
\[
    \mathbb{E} \left[ \|f_1 - \hat{f}_1_{V_1V_2}\|_n^2 \right] \leq \mathbb{E} \left[ 1_{\mathcal{E}_\delta} \|f_1 - (\hat{f}_V)_1\|_n^2 \right] + R_n. \tag{5.2}
\]

The details can be found in Appendix A. By the projection theorem (see, e.g., [26] Theorem II.3),
\[
    \|f_1 - (\hat{f}_V)_1\|_n^2 = \|f_1 - \Pi_{V_1} f_1\|_n^2 + \|\Pi_{V_1} f_1 - (\hat{f}_V)_1\|_n^2. \tag{5.3}
\]

By (4.2) and the definition of \( \hat{f}_V \), we have
\[
    \mathbb{E} \left[ 1_{\mathcal{E}_\delta} \|\Pi_{V_1} f_1 - (\hat{f}_V)_1\|_n^2 \right]
\]
\[
    \leq \frac{1}{1 - \delta} \mathbb{E} \left[ 1_{\mathcal{E}_\delta} \|\Pi_{V_1} f_1 - (\hat{f}_V)_1\|_n^2 \right]
\]
\[
    = \frac{1}{1 - \delta} \mathbb{E} \left[ 1_{\mathcal{E}_\delta} \|\Pi_{V_1} f_1 - (\hat{\Pi}_V \mathbf{Y})_1\|_n^2 \right]
\]
\[
    = \frac{1}{1 - \delta} \mathbb{E} \left[ 1_{\mathcal{E}_\delta} \left[ \|\Pi_{V_1} f_1 - (\hat{\Pi}_V \mathbf{Y})_1\|_n^2 | X_1, \ldots, X_n \right] \right]. \tag{5.4}
\]

If \( \mathcal{E}_\delta \) holds, each \( g \in V \) is determined uniquely by \( (g(X^1), \ldots, g(X^n))^T \) (see Remark 3). This explains why we don’t have to distinguish between those objects. Furthermore, we define \( \hat{\Pi}_V h \) as \( \hat{\Pi}_V(h(X^1), \ldots, h(X^n))^T \) for
all \( h \in L^2(\mathbb{P}^X) \). Therefore, if \( E_\delta \) holds, \( \hat{\Pi}_V \) can also be seen as a map from \( L^2(\mathbb{P}^X) \) to \( V \). From now on suppose that the event \( E_\delta \) holds. Then by the proof of Proposition 1, we have (5.3) is equal to

\[
\frac{1}{1 - \delta} E \left[ 1_{E_\delta} \left( \| \Pi_V f_1 - (\hat{\Pi}_V f_1) \|_n^2 \right) + E \left[ \| (\hat{\Pi}_V \epsilon)_1 \|_n^2 | X_1, \ldots, X_n \right] \right].
\]  

(5.5)

From (5.3) - (5.5) and (4.2), we obtain

\[
E \left[ 1_{E_\delta} \left( \| f_1 - (\hat{f}_V)_1 \|_n^2 \right) \right] \leq \frac{(1 + \delta)}{(1 - \delta)} E \left[ 1_{E_\delta} (\| f_1 - \Pi_V f_1 \|_n^2 + \| \Pi_V f_1 - (\hat{\Pi}_V f)_1 \|_n^2) \right] + \frac{(1 + \delta)}{(1 - \delta)} \frac{1}{\sigma^2 d_1} \frac{1}{n}.
\]

By Assumption 2 and Lemma 2, we have

\[
\| f_1 - (\hat{\Pi}_V f)_1 \|_n^2 \leq \frac{1}{1 - \rho_0} \| f - \hat{\Pi}_V f \|^2.
\]

The projection theorem implies that

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

By (4.2) (since \( E_\delta \) holds) and the projection theorem, we have

\[
\| \Pi_V f - \hat{\Pi}_V f \|^2 \leq \frac{1}{1 - \delta} \| \Pi_V f - \hat{\Pi}_V f \|^2 \leq \frac{1}{1 - \delta} \| f - \Pi_V f \|^2.
\]

Taking expectation, we obtain

\[
E \left[ 1_{E_\delta} (\| f_1 - (\hat{\Pi}_V f)_1 \|_n^2) \right] \leq E \left[ (\| f - \Pi_V f \|_n^2) \right] = \| f - \Pi_V f \|^2.
\]

We conclude that

\[
E \left[ 1_{E_\delta} (\| f_1 - (\hat{\Pi}_V f)_1 \|_n^2) \right] \leq \frac{1}{1 - \rho_0} \frac{2}{1 - \delta} \| f - \Pi_V f \|^2.
\]

This completes the proof. \( \square \)

6. Proof of Theorem 2

6.1. The variance term revisited. This subsection is devoted to a further analysis of the variance term of the estimator \( \hat{f}_{1,W_1, V_1, V_2} \) under the additional Assumption 3. In doing so, we apply the theory of projections on sumspaces in Hilbert spaces. We show that, up to terms of smaller order, the variance is equal to the variance of the estimator in the model \( Y = f_1(X_1) + \epsilon \). We begin with the following improvement of Proposition 1.
Proposition 2. Let the assumptions of Theorem 2 hold. If the event $E_\delta$ of Corollary 5 holds, then

$$E \left[ \| \tilde{\Pi}_{W_1} (\hat{\Pi}_V \epsilon)_1 \|_n^2 | X_1, \ldots, X_n \right]$$

$$\leq (1 + \delta) \left( \frac{\sigma^2 \dim W_1}{n} + \frac{1 + \delta}{1 - \rho_0^2} \frac{\sigma^2 \text{tr}(\hat{\Pi}_{W_1} \hat{\Pi}_{12})}{n} \right).$$

Remark 4. Applying the bound $\text{tr}(\hat{\Pi}_{W_1} \hat{\Pi}_{V_1}) \leq \text{tr}(\hat{\Pi}_{W_1}) = \dim W_1$, we obtain Proposition 1 with a slightly larger constant. However, in Proposition 3 we show that $\text{tr}(\hat{\Pi}_{W_1} \hat{\Pi}_{V_2})$ is of smaller order if Assumption 3 holds.

The basic theorem in the theory of projections on sumspaces is due to von Neumann [32]. In the proof of Proposition 2, we apply the following version which is a consequence of [1, (15) on page 378] (see also [4, Theorem 2.C]). Since we also apply it in the population setting in Proposition 4 we give a general statement.

Lemma 4. Let $H_1$ and $H_2$ be two closed subspaces of a Hilbert space $H$. Suppose that $\rho_0(H_1, H_2) < 1$. Let $\Pi$, $\Pi_1$, $\Pi_2$ be the orthogonal projections on $H_1 + H_2$, $H_1$, $H_2$, respectively. Let $h \in H$, and let $h^*_1 \in H_1$, $h^*_2 \in H_2$ be the unique elements satisfying $\Pi h = h^*_1 + h^*_2$. Then

$$\| h^*_1 - (\Pi_1 - \sum_{j=1}^k (\Pi_1 \Pi_2)^j (1 - \Pi_1)) h \| \rightarrow 0$$

as $k$ goes to infinity.

We will also need the following results:

Lemma 5. Let $A \in \mathbb{R}^{k_1 \times k_2}$ and $B \in \mathbb{R}^{k_2 \times k_1}$. Then

(i) $\text{tr}(AB) = \text{tr}(BA)$.

(ii) $| \text{tr}(AB) | \leq \sqrt{\text{tr}(AA^T) \text{tr}(BB^T)}$.

(iii) Let $k_1 = k_2$ and $B$ be symmetric and positive semi-definite. Then

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

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

Proof of Lemma 5. 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

$$B = V^T \text{diag}(\lambda_1(B), \ldots, \lambda_{k_1}(B)) V. \quad (6.1)$$

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 (6.1),
part (i) of this Lemma, the fact that the $\lambda_j(B)$’s are nonnegative, and the previous argument, we obtain

$$|\text{tr}(AB)| = \left| \sum_{j=1}^{k_1} (VAV^T)_{jj} \lambda_j(B) \right| \leq \max_{j=1,\ldots,k_1} |(VAV^T)_{jj}| \text{tr}(B) \leq \|A\|_{\text{op}} \text{tr}(B).$$

This completes the proof. \[\square\]

**Proof of Proposition 2.** 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 $\|\cdot\|_n$ is a norm on $V$ and thus each element $g \in V$ is determined uniquely by $(g(X^1), \ldots, g(X^n))^T$ (since $\mathcal{E}_\delta$ holds, see Remark 3). From (4.7) and Lemma 4 we have

$$\|\hat{\Pi}V\epsilon_1 - (\hat{\Pi}V_1 - \sum_{j=1}^{k_1} (\hat{\Pi}V_1 \hat{\Pi}V_2)^j(1 - \hat{\Pi}V_1))\epsilon_1\|_n \to 0$$

as $k$ goes to infinity. Now we introduce the constant

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

From (6.7) we have $\|\hat{\Pi}V_1 g_2\|_n \leq \rho_{0,\delta}\|g_2\|_n$ for all $g_2 \in V_2$, which follows from $\|\hat{\Pi}V_1 g_2\|^2_n = \langle \hat{\Pi}V_1 g_2, \hat{\Pi}V_1 g_2 \rangle_n = \langle \hat{\Pi}V_1 g_2, g_2 \rangle_n \leq \rho_{0,\delta}\|\hat{\Pi}V_1 g_2\| \|g_2\|_n$. Similarly, we have $\|\hat{\Pi}V_2 g_1\|_n \leq \rho_{0,\delta}\|g_1\|_n$ for all $g_1 \in V_1$. We obtain $\|\hat{\Pi}V_1 \hat{\Pi}V_2\|_{\text{op}} \leq \rho_{0,\delta}$ and

$$\|\hat{\Pi}V_1 \hat{\Pi}V_2 \hat{\Pi}V_1\|_{\text{op}} \leq \rho_{0,\delta}^2.$$ 

(6.2)

Note that $\|A\|_{\text{op}} = \sup_{\|x\|_2=1} \|Ax\|_2 = \sup_{\|x\|_n=1} \|Ax\|_n$ for all $A \in \mathbb{R}^{n \times n}$. This gives the improved convergence result

$$\|\hat{\Pi}V\epsilon_1 - (\hat{\Pi}V_1 - \sum_{j=1}^{k_1} (\hat{\Pi}V_1 \hat{\Pi}V_2)^j(1 - \hat{\Pi}V_1))\epsilon_1\|_n \leq \sum_{j=k+1}^{\infty} \|\hat{\Pi}V_1 \hat{\Pi}V_2 \hat{\Pi}V_1\|_{\text{op}}^{j-1} \|\hat{\Pi}V_1 \hat{\Pi}V_2 (1 - \hat{\Pi}V_1)\epsilon_1\|_n \leq \sum_{j=k+1}^{\infty} \rho_{0,\delta}^{2(j-1)} \rho_{0,\delta}\|\epsilon_1\|_n \leq \frac{\rho_{0,\delta}^{2k+1}}{1 - \rho_{0,\delta}}\|\epsilon_1\|_n,$$
where we applied \((\hat{\Pi}_1 \hat{\Pi}_2)^j = (\hat{\Pi}_1 \hat{\Pi}_2 \hat{\Pi}_1)^j \hat{\Pi}_1 \hat{\Pi}_2, j \geq 1\), and thus we also have

\[
\|\hat{\Pi}_W (\Pi_V \epsilon)_1 - \hat{\Pi}_W (\Pi_V - \sum_{j=1}^{k} (\hat{\Pi}_V \hat{\Pi}_V)_{\hat{j}} (1 - \hat{\Pi}_V)) \epsilon\| \leq \frac{\rho_{0,\delta}^{2k+1}}{1 - \rho_{0,\delta}^2} \|\epsilon\|_n. \tag{6.3}
\]

Applying \((6.3)\) and the bound \((x + y)^2 \leq (1 + \delta)x^2 + (1 + 1/\delta)y^2\) and taking expectation, we obtain

\[
\mathbb{E} \left[\left\|\hat{\Pi}_W (\Pi_V \epsilon)_1 \right\|_n^2 \mid X^1, \ldots, X^n\right] \leq (1 + \delta)\mathbb{E} \left[\left\|\hat{\Pi}_W (\Pi_V - \sum_{j=1}^{k} (\hat{\Pi}_V \hat{\Pi}_V)_{\hat{j}} (1 - \hat{\Pi}_V)) \epsilon\|_n^2 \mid X^1, \ldots, X^n\right]\right] + \left(1 + \frac{1}{\delta}\right) \frac{\rho_{0,\delta}^{2k+2}}{(1 - \rho_{0,\delta}^2)^2} \sigma^2. \tag{6.4}
\]

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

\[
\hat{\Pi}_W (\hat{\Pi}_V - \sum_{j=1}^{k} (\hat{\Pi}_V \hat{\Pi}_V)_{\hat{j}} (1 - \hat{\Pi}_V)) \left(\hat{\Pi}_W (\hat{\Pi}_V - \sum_{j=1}^{k} (\hat{\Pi}_V \hat{\Pi}_V)_{\hat{j}} (1 - \hat{\Pi}_V))\right)^T. \tag{6.5}
\]

Using \(\hat{\Pi}_V - \sum_{j=1}^{k} (\hat{\Pi}_V \hat{\Pi}_V)_{\hat{j}} (1 - \hat{\Pi}_V) = \sum_{j=0}^{k-1} \hat{\Pi}_V (\hat{\Pi}_V \hat{\Pi}_V)_{\hat{j}} (1 - \hat{\Pi}_V) + (\hat{\Pi}_V \hat{\Pi}_V)^{k} \hat{\Pi}_V\), \(6.5\) is equal to

\[
\hat{\Pi}_W \sum_{j=0}^{k} (\hat{\Pi}_V \hat{\Pi}_V)_{\hat{j}} \hat{\Pi}_V \left( (\hat{\Pi}_V - (1 - \hat{\Pi}_V)) \sum_{j=1}^{k} (\hat{\Pi}_V \hat{\Pi}_V)_{\hat{j}} \hat{\Pi}_W \right) - \hat{\Pi}_W \sum_{j=1}^{k} (\hat{\Pi}_V \hat{\Pi}_V)_{\hat{j}} \left( ((1 - \hat{\Pi}_V) \sum_{j=0}^{k-1} (\hat{\Pi}_V \hat{\Pi}_V)_{\hat{j}} \hat{\Pi}_V + (\hat{\Pi}_V \hat{\Pi}_V)^{k} \hat{\Pi}_V) \hat{\Pi}_W \right)
\]

and, since \(\hat{\Pi}_V (1 - \hat{\Pi}_V) = 0\) and \(\hat{\Pi}_V (1 - \hat{\Pi}_V) = 0\), this is equal to

\[
\hat{\Pi}_W \sum_{j=0}^{k} (\hat{\Pi}_V \hat{\Pi}_V)_{\hat{j}} \hat{\Pi}_W - \hat{\Pi}_W \sum_{j=1}^{k} (\hat{\Pi}_V \hat{\Pi}_V)_{\hat{j}} \hat{\Pi}_W = \hat{\Pi}_W + \hat{\Pi}_W \sum_{j=1}^{k} (\hat{\Pi}_V \hat{\Pi}_V \hat{\Pi}_V)_{\hat{j}} \hat{\Pi}_W - \hat{\Pi}_W \sum_{j=k+1}^{2k} (\hat{\Pi}_V \hat{\Pi}_V \hat{\Pi}_V)_{\hat{j}} \hat{\Pi}_W.
\]

Thus the trace of \(6.5\) is bounded by

\[
\dim W_1 + \sum_{j=1}^{2k} \left| \text{tr}(\hat{\Pi}_W (\hat{\Pi}_V \hat{\Pi}_V \hat{\Pi}_V)_{\hat{j}} \hat{\Pi}_W) \right|.
\]
Applying Lemma 5 (i) and (iii), this can be bounded by
\[
\dim W_1 + \frac{2k-1}{\delta_0^2} \cdot \nabla_{\hat{\Pi}V_1^2} |\text{tr}(\hat{\Pi}V_1^2 \hat{\Pi}W_1)|. \tag{6.6}
\]

We have \(\text{tr}(\hat{\Pi}V_1^2 \hat{\Pi}W_1) = \text{tr}(\hat{\Pi}W_1^2 \hat{\Pi}V_1) = \text{tr}(\hat{\Pi}W_1 \hat{\Pi}V_1)\) and this is a non-negative real number, since also \(\text{tr}(\hat{\Pi}W_1 \hat{\Pi}V_1) = \text{tr}(\hat{\Pi}W_1 \hat{\Pi}V_2 \hat{\Pi}W_1)\). Applying (6.2), we obtain that (6.6) is bounded by
\[
\dim W_1 + \frac{1}{\frac{1 - \rho_0^2}{\delta_0^2}} \cdot \text{tr}(\hat{\Pi}W_1^2 \hat{\Pi}V_1^2). \tag{6.7}
\]

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 \(\varrho = \rho_0\))
\[
\frac{1}{1 - \rho_0^2} \leq \frac{1 + \delta}{1 - \delta} \cdot \frac{1}{1 - \rho_0^2}. \tag{6.8}
\]

From (6.4)-(6.8), we conclude that
\[
\mathbb{E} \left[ \|\hat{\Pi}W_1^2 (\hat{\Pi}V_1^2)\|_n^2 |X^1, \ldots, X^n \right] \\
\leq (1 + \delta) \left( \frac{\sigma^2 \dim W_1}{n} + \frac{1 + \delta}{1 - \delta} \cdot \frac{1}{1 - \rho_0^2} \cdot \frac{\sigma^2 \text{tr}(\hat{\Pi}W_1^2 \hat{\Pi}V_1)}{n} \right) \\
+ \left( 1 + \frac{1}{\delta} \right) \cdot \frac{\rho_0^{4k+2}}{(1 - \rho_0^2)\delta^2} \cdot \sigma^2.
\]

Now, let \(k\) go to infinity. This completes the proof. \(\square\)

**Proposition 3.** Let the assumptions of Theorem 2 hold. Let \(E_\delta\) be the event of Corollary 5. Then

\[
\frac{1}{n} \mathbb{E} \left[ 1_{E_\delta} \text{tr}(\hat{\Pi}W_1^2 \hat{\Pi}V_1^2) \right] \leq \frac{1}{(1 - \delta)^2} \left( \frac{\|\Pi\|^2_{HS}}{n} + \frac{\dim W_1 \varphi^2 d}{n} \right),
\]

where \(\|\Pi\|^2_{HS}\) is defined in Assumption 3.

**Proof.** Let \(d_{W_1} = \dim W_1\). Let \(\varphi_1, \ldots, \varphi_{d_{W_1}}\) be an orthonormal basis of \(W_1\) and \(\varphi_{d_{W_1}+1}, \ldots, \varphi_{d_{W_1}+d_2}\) be an orthonormal basis of \(V_2\). Now, define the matrices \(Z_1, Z_2, \) and \(Z\) similarly as in the proof of Proposition 1 but with the above functions instead. Suppose that the event \(E_\delta\) holds. Then we have \(\hat{\Pi}W_1 = Z_1(Z_1^T Z_1)^{-1} Z_1^T, \hat{\Pi}V_2 = Z_2(Z_2^T Z_2)^{-1} Z_2^T\) and thus
\[
\text{tr}(\hat{\Pi}W_1 \hat{\Pi}V_2) = \text{tr} \left( \left( \frac{1}{n} Z_1^T Z_1 \right)^{-1} \frac{1}{n} Z_1^T Z_2 \left( \frac{1}{n} Z_2^T Z_2 \right)^{-1} \frac{1}{n} Z_2^T Z_1 \right),
\]
where we applied Lemma 5 (i). By Theorem 5 we have \( \| (1/n) Z_j^T Z_j - I \|_{\text{op}} \leq \delta \) and thus \( (1/n) Z_j^T Z_j \) for \( j = 1, 2 \). We conclude that

\[
\mathbb{E} \left[ 1_{\mathcal{E}_3} \text{tr}(\hat{\Pi}_{W_1} \hat{\Pi}_{V_2}) \right] \\
\leq \sum_{k,l=0}^{\infty} \mathbb{E} \left[ 1_{\mathcal{E}_3} \right] \text{tr} \left( \left( I - \frac{1}{n} Z_1^T Z_1 \right)^k \frac{1}{n} Z_1^T Z_2 \left( I - \frac{1}{n} Z_2^T Z_2 \right)^l \frac{1}{n} Z_2^T Z_1 \right) \\
\leq \sum_{k,l=0}^{\infty} \mathbb{E} \left[ 1_{\mathcal{E}_3} \right] \text{tr} \left( \left( I - \frac{1}{n} Z_1^T Z_1 \right)^{2k} \frac{1}{n} Z_1^T Z_2 \left( \frac{1}{n} Z_1^T Z_2 \right)^l \left( I - \frac{1}{n} Z_2^T Z_2 \right)^{2l} \frac{1}{n} Z_2^T Z_1 \right) \\
\leq \sum_{k,l=0}^{\infty} \delta^{k+l} \mathbb{E} \left[ 1_{\mathcal{E}_3} \right] \text{tr} \left( \frac{1}{n} Z_1^T Z_2 \left( \frac{1}{n} Z_1^T Z_2 \right)^T \right)
\]

where we applied Lemma 5 (i) and (ii) in the second inequality and Lemma 5 (i) and (iii) and the bound \( \| (1/n) Z_j^T Z_j - I \|_{\text{op}} \leq \delta \) in the third inequality. Now

\[
\mathbb{E} \left[ \text{tr} \left( \frac{1}{n} Z_1^T Z_2 \left( \frac{1}{n} Z_1^T Z_2 \right)^T \right) \right] \\
= \sum_{j=1}^{d_{W_1}} \sum_{k=d_{W_1}+1}^{d_{W_1}+d_2} \left( (\mathbb{E} [\mathcal{F}_j(X_1)\mathcal{F}_k(X_2)])^2 + \frac{1}{n} \text{Var}(\mathcal{F}_j(X_1)\mathcal{F}_k(X_2)) \right) \\
\leq \| \mathcal{F} \|_{\text{HS}}^2 + d_{W_1} \frac{\phi^2 d}{n},
\]

where the first part of the inequality follows from Bessel’s inequality and the second part from [2,4] and [5, Lemma 1]. This completes the proof. \( \square \)

6.2. End of the proof of Theorem 2. Repeating the steps (5.2)-(5.5) in the proof of Theorem 1 we obtain

\[
\mathbb{E} \left[ \| f_1 - \hat{f}_{1,W_1,V_1,V_2} \|^2 \right] \\
\leq \| f_1 - \Pi_{W_1} f_1 \|^2 + \frac{1}{1-\delta} \mathbb{E} \left[ 1_{\mathcal{E}_3} \| \Pi_{W_1} f_1 - \hat{\Pi}_{W_1} (\hat{\Pi}_V f_1) \|_n^2 \right] \\
+ \frac{1}{1-\delta} \mathbb{E} \left[ 1_{\mathcal{E}_3} \mathbb{E} \left[ \| \hat{\Pi}_{W_1} (\hat{\Pi}_V \epsilon)_1 \|_n^2 \| X^1, \ldots, X^n \right] \right] + R_n. \quad (6.9)
\]

By Proposition 2 and 3 the third term on the right-hand side is bounded by

\[
\frac{1 + \delta \sigma^2 \dim W_1}{1-\delta} \frac{\sqrt{\dim W_1}}{n} + \frac{(1 + \delta)^2}{(1-\delta)^4} \frac{1}{1-\rho_0^2} \left( \frac{\sigma^2 \| \mathcal{F} \|_{\text{HS}}^2 \dim W_1}{n} + \frac{\sigma^2 \dim W_1 \phi^2 d}{n} \right)
\]
Thus it remains to deal with the second term. Suppose that the event $\mathcal{E}_\delta$ holds. By the projection theorem and \((2.2)\), we have
\[
\|\Pi W_1 f_1 - \hat{\Pi} W_1 (\hat{\Pi} V f)_1\|^2_n \leq \|\Pi W_1 f_1 - (\hat{\Pi} V f)_1\|^2_n \\
\leq (1 + \delta)\|\Pi W_1 f_1 - (\hat{\Pi} V f)_1\|^2.
\] (6.10)
Applying the bound $(x + y)^2 \leq (1 + \delta)x^2 + (1 + 1/\delta)y^2$, we obtain
\[
\|\Pi W_1 f_1 - (\hat{\Pi} V f)_1\|^2 \leq (1 + \delta)\|f_1 - \Pi W_1 f_1\|^2 + (1 + 1/\delta)\|f_1 - (\hat{\Pi} V f)_1\|^2.
\] (6.11)
At the end of the proof of Theorem 1, we already showed that
\[
\text{applying the bound } (x + y)^2 \leq (1 + \delta)x^2 + (1 + 1/\delta)y^2, \text{ we have}
\]
\[
\|f_1 - (\hat{\Pi} V f)_1\|^2 \leq \frac{2}{1 - \delta} \frac{1}{1 - \rho_0^2} \|f - \Pi V f\|^2.
\] (6.12)
Inserting (6.10)-(6.12) into (6.9), completes the proof. \qed

7. Proof of Theorem 3 and 4

7.1. The bias term. In this subsection we improve the upper bounds for the bias term given in the proofs of Theorem 1 and 2. This is possible under Assumption 3 and the additional regularity conditions on the design densities, namely (2.9) or (2.10).

**Proposition 4.** Let $X = (X_1^T, X_2^T)^T$ be a $q$-dimensional random variable as in (2.1). Let Assumptions 1, 2 and 3 hold. For $j = 1, 2$ let $V_j \subseteq H_j$ be a $d_j$-dimensional subspace. Let $V = V_1 + V_2$. For fixed $x_1$ let $r(x_1, \cdot)$ be the orthogonal projection of $p(x_1, \cdot)/(p_1(x_1)p_2(\cdot))$ on $H_2$ (defined for $\mathbb{P}^{X_1}$-almost all $x_1$, by (2.7)). Suppose that there exist real numbers $\phi(V_2)$, $\psi(V_2)$ and a function $h_1 \in L^2(\mathbb{P}^{X_1})$ such that $\|f_2 - \Pi V_2 f_2\| \leq \phi(V_2)$ and
\[
\|r(x_1, \cdot) - \Pi V_2 (r(x_1, \cdot))\|_{L^2(\mathbb{P}^{X_2})} \leq h_1(x_1)\psi(V_2)
\] (7.1)
for $\mathbb{P}^{X_1}$-almost all $x_1$. Then
\[
\|\Pi V_2 f_2\|_1 \leq \frac{1}{1 - \rho_0^2} \|h_1\|\psi(V_2)\phi(V_2).
\] (7.2)

**Proof.** Let $\varphi_1, \ldots, \varphi_{d_1}$ be an orthonormal basis of $V_1$. We have
\[
\|\Pi V_1 (f_2 - \Pi V_2 f_2)\|^2 = \sum_{j=1}^{d_1} \left( \int \int \varphi_j(x_1)(f_2(x_2) - \Pi V_2 f_2(x_2))p(x_1, x_2)d(\nu_1 \otimes \nu_2)(x_1, x_2) \right)^2
\]
\[
= \sum_{j=1}^{d_1} \left( \int \varphi_j(x_1) \left( \int \frac{p(x_1, x_2)}{p_1(x_1)p_2(x_2)}d\mathbb{P}^{X_2}(x_2) \right)d\mathbb{P}^{X_1}(x_1) \right)^2
\]
\[
= \sum_{j=1}^{d_1} \left( \int \left( (1 - \Pi V_2) f_2, \frac{p(x_1, \cdot)}{p_1(x_1)p_2(\cdot)} \right)_{L^2(\mathbb{P}^{X_2})} \varphi_j(x_1)d\mathbb{P}^{X_1}(x_1) \right)^2,
\]
where we already applied Assumption 3 in the second equality. Since orthogonal projections are idempotent and self-adjoint, the above is equal to

\[
\sum_{j=1}^{d_1} \left( \int (1 - \Pi V_2) f_2, (1 - \Pi V_2) r(x_1, \cdot) \right)_{L^2(\mathbb{P}X_2)}^2 \leq \int (1 - \Pi V_2) f_2, (1 - \Pi V_2) r(x_1, \cdot) \right)_{L^2(\mathbb{P}X_2)}^2 \leq \|h_1\|_2^2 (\psi \Pi V_2 \phi(V_2))^2,
\]

where we applied Bessel’s inequality in the first inequality and the Cauchy-Schwarz inequality and (7.1) in the second inequality. Thus we have shown that

\[
\|\Pi V_1 (f_2 - \Pi V_2 f_2)\| \leq \|h_1\|_2 \psi \Pi V_2 \phi(V_2). \tag{7.3}
\]

Now, by Lemma 4 we have

\[
\| (\Pi V h)_1 - (\Pi V_1 - \sum_{j=1}^k (\Pi V_1 \Pi V_2)^j (1 - \Pi V_1)) h_1 \| \to 0, \tag{7.4}
\]

as \( k \to \infty \), for all \( h \in L^2(\mathbb{P}^X) \). By Assumption 2 we have \( \|\Pi V_1 \Pi V_2 h\| \leq \rho_0 \|\Pi V_1 h\| \) and \( \|\Pi V_1 \Pi V_2 h\| \leq \rho_0 \|\Pi V_1 h\| \) for all \( h \in L^2(\mathbb{P}^X) \), which follows as in the proof of (6.2). Applying this and (7.3), we obtain

\[
\| (\Pi V_1 - \sum_{j=1}^k (\Pi V_1 \Pi V_2)^j (1 - \Pi V_1)) (f_2 - \Pi V_2 f_2) \|
\]

\[
= \left\| \sum_{j=0}^k (\Pi V_1 \Pi V_2)^j \Pi V_1 (f_2 - \Pi V_2 f_2) \right\|
\]

\[
\leq \sum_{j=0}^k \rho_0^{2j} \|\Pi V_1 (f_2 - \Pi V_2 f_2)\| \leq \frac{1}{1 - \rho_0^2} \|h_1\|_2 \psi \Pi V_2 \phi(V_2). \tag{7.5}
\]

Since \( \Pi V_1 \Pi V_2 f_2 = \Pi V_2 f_2 \) 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, (7.4), and (7.5), we conclude that

\[
\|(\Pi V f_2)_1\| = \|(\Pi V (f_2 - \Pi V_2 f_2))_1\| \leq \frac{1}{1 - \rho_0^2} \|h_1\|_2 \psi \Pi V_2 \phi(V_2). \tag{7.6}
\]

This completes the proof. \( \square \)

**Proposition 5.** Let \( X^1, \ldots, X^n \) be \( n \) independent copies of a \( q \)-dimensional random variable \( X \). Let \( V \subseteq L^2(\mathbb{P}^X) \) be \( d \)-dimensional subspace. Suppose that there exists a real number \( \varphi \) such that

\[
\|g\|_{\infty}^2 \leq \varphi^2 d \|g\|^2 \tag{7.7}
\]
for all \( g \in V \). Let \( 0 < \delta < 1 \), and let \( \mathcal{E}_\delta \) be the event defined by (4.2). Let \( f \in L^2(\mathbb{P}^X) \). Then
\[
\mathbb{E} \left[ 1_{\mathcal{E}_\delta} \| (\Pi_V - \hat{\Pi}_V) f \|_2^2 \right] \leq \left( \frac{1}{1 - \delta} \right)^2 \frac{\varphi^2 d}{n} \| f - \Pi_V f \|_2^2. \tag{7.8}
\]

**Proof.** Let \( b_1, \ldots, b_d \) be an orthonormal basis of \( V \) and let \( B_n = (\langle b_j, b_k \rangle)_{1 \leq j, k \leq d} \).

Throughout the proof suppose that the event \( \mathcal{E}_\delta \) holds. Then \( \hat{\Pi}_V \) is a well-defined map from \( L^2(\mathbb{P}^X) \) to \( V \) (see the discussion after (5.4)). For \( f \in L^2(\mathbb{P}^X) \) we have
\[
\Pi_V f = \begin{pmatrix} \langle b_1, f \rangle \\ \vdots \\ \langle b_d, f \rangle \end{pmatrix}^T \begin{pmatrix} b_1 \\ \vdots \\ b_d \end{pmatrix} \quad \text{and} \quad \hat{\Pi}_V f = \begin{pmatrix} \langle b_1, f \rangle \\ \vdots \\ \langle b_d, f \rangle \end{pmatrix}^T B_n^{-1} \begin{pmatrix} b_1 \\ \vdots \\ b_d \end{pmatrix}.
\]

Let \( x = (\langle b_1, f \rangle, \ldots, \langle b_d, f \rangle)^T \) and \( x_n = (\langle b_1, f \rangle_n, \ldots, \langle b_d, f \rangle_n)^T \). Then
\[
\| (\Pi_V - \hat{\Pi}_V) f \|_2^2 = \| B_n^{-1} x_n - x \|_2^2. \tag{7.9}
\]

Since \( \mathcal{E}_\delta \) holds, we have \( \| B_n - I \|_{\text{op}} \leq \delta \) by (4.3). This implies that
\[
B_n^{-1} - I = \sum_{k \geq 1} (I - B_n)^k.
\]

Thus
\[
B_n^{-1} x_n - x = x_n - x + \sum_{k \geq 1} (I - B_n)^k x_n
\]
\[
= \sum_{k \geq 0} (I - B_n)^k (x_n - x) + \sum_{k \geq 1} (I - B_n)^k x.
\]

Applying the bound \( \| B_n - I \|_{\text{op}} \leq \delta \), we obtain
\[
\| B_n^{-1} x_n - x \|_2^2 \leq \left( \frac{1}{1 - \delta} \right)^2 \left( \| x_n - x \|_2 + \| (B_n - I)x \|_2 \right)^2
\]
\[
\leq \left( \frac{1}{1 - \delta} \right)^2 \left( (1 + \epsilon) \| x_n - x \|_2^2 + (1 + 1/\epsilon) \| (B_n - I)x \|_2^2 \right), \tag{7.10}
\]

where we also applied the bound \((x + y)^2 \leq (1 + \epsilon) x^2 + (1 + 1/\epsilon) y^2\), for \( \epsilon > 0 \) arbitrary. Now we have
\[
\mathbb{E} \left[ \| x_n - x \|_2^2 \right] = \mathbb{E} \left[ \sum_{j=1}^d (\langle b_j, f \rangle_n - \langle b_j, f \rangle)^2 \right]
\]
\[
= \frac{1}{n} \sum_{j=1}^d \text{Var}(b_j(X)f(X)) \leq \frac{\varphi^2 d}{n} \| f \|_2^2, \tag{7.11}
\]
where we applied the bound \( \| \sum_{j=1}^d b_j^2 \|_\infty \leq \varphi^2 d \), which follows from (7.7) and [5 Lemma 1]. The \( j \)th coordinate of \((B_n - I)x\) is is equal to

\[
(b_j, \sum_{k=1}^d b_k x_k)_n - (b_j, f) = (b_j, \Pi_V f)_n - (b_j, \Pi_V f)
\]  

(7.12)

and thus

\[
\mathbb{E} \left[ \| (B_n - I)x \|_2^2 \right] = \mathbb{E} \left[ \sum_{j=1}^d (\langle b_j, \Pi_V f \rangle_n - \langle b_j, \Pi_V f \rangle)^2 \right]
\]

\[
= \frac{1}{n} \sum_{j=1}^{d_2} \text{Var}(b_j(X)\Pi_V f(X)) \leq \frac{\varphi^2 d}{n} \| \Pi_V f \|^2.
\]

(7.13)

Applying (7.9)–(7.13), we conclude that

\[
\mathbb{E} \left[ 1_{E_\delta} \| (\Pi_V - \hat{\Pi})f \|_2^2 \right] \leq \left( \frac{1}{1 - \delta} \right)^2 \frac{\varphi^2 d}{n} (1 + \epsilon)\| f \|^2 + (1 + 1/\epsilon)\| \Pi_V f \|^2.
\]

(7.14)

Finally, since \( \Pi_V \) and \( \hat{\Pi} \) fix elements in \( V \), we obtain \((\Pi_V - \hat{\Pi})f = 0 \) and thus \((\Pi_V - \hat{\Pi})f = (\Pi_V - \hat{\Pi})(1 - \Pi_V)f \). Combining this with (7.14), we see that

\[
\mathbb{E} \left[ 1_{E_\delta} \| (\Pi_V - \hat{\Pi})f \|_2^2 \right] \leq \left( \frac{1}{1 - \delta} \right)^2 \frac{\varphi^2 d}{n} (1 + \epsilon)\| f - \Pi_V f \|^2.
\]

(7.15)

Since \( \epsilon \) is arbitrary, this completes the proof. \( \square \)

**Proposition 6.** Let the assumptions of Theorem 3 hold, and let \( E_\delta \) be the event constructed in Corollary 5. Suppose that there exist real numbers \( \phi(V_2), \psi_\Pi(V_2) \) and a function \( h_1 \in L^2(\mathbb{P}^X_1) \) such that \( \| f_2 - \Pi_{V_2} f_2 \| \leq \phi(V_2) \) and

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

(7.16)

for all \( x_1 \). Then

\[
\mathbb{E} \left[ 1_{E_\delta} \| (\hat{\Pi}_V f_2) \|_n^2 \right] \leq \left( 1 + \delta \right)^3 \frac{1}{(1 - \delta)^3} \left( \frac{\varphi^2 d}{\delta n} \right)^2 \left( \| h_1 \|^2 (\psi(V_2)\phi(V_2))^2 \right)
\]

\[
+ \frac{1}{n} \left( h_1 \| \psi(V_2) \|_2^2 \| (1 - \Pi_{V_2}) f_2 \|_\infty^2 + \frac{\varphi^2 d_1^2}{\delta n} \| (1 - \Pi_{V_2}) f_2 \|_n^2 \right).
\]

(7.17)

**Proof.** The proof is similar to the proof of Proposition 4. Let \( \varphi_1, \ldots, \varphi_{d_1} \) be an orthonormal basis of \( V_1 \). Throughout the proof suppose that the event \( E_\delta \) holds. Then \( \hat{\Pi}_V \) is a well-defined map from \( L^2(\mathbb{P}^X) \) to \( V \) (see the discussion after (7.4)). By repeating the arguments at the beginning of the proof of Proposition 5 we obtain

\[
\| \hat{\Pi}_V (f_2 - \hat{\Pi}_V f_2) \|_n^2 \leq \frac{1}{1 - \delta} \sum_{j=1}^{d_1} \langle \varphi_j, (1 - \hat{\Pi}_V) f_2 \rangle_n^2
\]

and thus
Thus
\[
\mathbb{E} \left[ 1_{\xi_i} \| \hat{\Pi}_{V_1} (f_2 - \hat{\Pi}_{V_2} f_2) \|^2_n \right]
\leq \frac{1 + \delta}{1 - \delta} \mathbb{E} \left[ \sum_{j=1}^{d_1} (\varphi_{j, n_2}, (1 - \hat{\Pi}_{V_2}) f_2)^2_n \right]^{\frac{2}{1 - \delta}} (7.18)
\]
\[
+ \frac{1 + \delta}{\delta(1 - \delta)} \mathbb{E} \left[ \sum_{j=1}^{d_1} (\varphi_{j} - \varphi_{j, n_2}, (1 - \hat{\Pi}_{V_2}) f_2)^2_n \right], \quad (7.19)
\]
where
\[
\varphi_{j, n_2}(x_2) = \int \varphi_j(x_1) \frac{p(x_1, x_2)}{p_n(x_1)p_2(x_2)} p_1(x_1) dx_1
\]
is the conditional expectation of $\varphi_j(X_1)$ given $X_2 = x_2$ (for $\mathbb{P}^{X_2}$-almost all $x_2$, by Assumption [3]). The expectation in (7.18) is equal to
\[
\mathbb{E} \left[ \sum_{j=1}^{d_1} \left( \int \left( \frac{p(x_1, \cdot)}{p_n(x_1)p_2(\cdot)} \right)^2 (1 - \hat{\Pi}_{V_2}) f_2^2_n \right) \varphi_j(x_1)p_1(x_1) dx_1 \right]^{\frac{1}{2}}
\leq \int \mathbb{E} \left[ \left( \frac{p(x_1, \cdot)}{p_n(x_1)p_2(\cdot)} \right)^2 (1 - \hat{\Pi}_{V_2}) f_2^2_n \right] p_1(x_1) dx_1,
\]
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
\[
\leq \int \mathbb{E} \left[ \left( (1 - \hat{\Pi}_{V_2}) \frac{p(x_1, \cdot)}{p_n(x_1)p_2(\cdot)} \right)^2 \right] p_1(x_1) dx_1.
\]
Applying the projection theorem and then (7.16), this is
\[
\leq \int \mathbb{E} \left[ \left( (1 - \Pi_{V_2}) \frac{p(x_1, \cdot)}{p_n(x_1)p_2(\cdot)} \right)^2 \right] p_1(x_1) dx_1
\leq \frac{n - 1}{n} \| h_1 \|_2^2 (\psi(V_2)\phi(V_2))^2 + \frac{1}{n} \| (1 - \Pi_{V_2}) f_2 \|^2_\infty \| h_1 \|_2^2 (\psi(V_2))^2.
\]
Now we turn to the expectation in (7.19). We have
\[
\mathbb{E} \left[ \sum_{j=1}^{d_1} (\varphi_j - \varphi_{j, n_2}, (1 - \hat{\Pi}_{V_2}) f_2)^2_n \right]
\leq \frac{\varphi^2 d_1}{n} \mathbb{E} \left[ \| (1 - \hat{\Pi}_{V_2}) f_2 \|^2_n \right]
\leq \frac{\varphi^2 d_1}{n} \| (1 - \Pi_{V_2}) f_2 \|^2.
To prove the first inequality, first note that the \((1 - \hat{\Pi} V_2) f_2)(X_2)\)s only depend on \(X_1^1, \ldots, X_2^n\) and we have
\[
\mathbb{E} \left[ (\varphi_j - \varphi_j, \Pi_2)(X_1^1, \ldots, X_2^n, X_1) \right] = \mathbb{E} \left[ (\varphi_j - \varphi_j, \Pi_2)(X_1^1) \right] = 0
\]
for \(i \neq i'\). This implies that the nondiagonal terms vanish. Next, apply the inequalities \(\mathbb{E}[(\varphi_j - \varphi_j, \Pi_2)^2(X)|X_2] \leq \mathbb{E}[(\varphi_j^2(X_1)|X_2] \) and \(\sum_{j=1}^{d_1} \varphi_j^2 \leq \varphi^2 \) (which follows from the assumptions of Theorem 3 and Lemma 1). Thus we have shown that
\[
\mathbb{E} \left[ 1_{\epsilon_n} \|\hat{\Pi} V_1 (f_2 - \hat{\Pi} V_2)\|^2 \right] \leq \frac{1 + \delta}{1 - \delta} \left( \|h_1\|^2 (\psi(V_2) \phi(V_2))^2 + \frac{1}{n} \|h_1\|^2 (1 - \Pi V_2) f_2 \|_2^2 + \frac{\varphi^2 d_1}{\delta n} \| (1 - \Pi V_2) f_2 \|^2 \right). \tag{7.20}
\]
The remaining arguments are as in the proof of Proposition 4. From (4.7) and Lemma 4 we have
\[
\|(\hat{\Pi} V h_1) - (\hat{\Pi} V_1 - \sum_{j=1}^{k} (\hat{\Pi} V_1 \hat{\Pi} V_2) j (1 - \hat{\Pi} V_1)) h_1\|_n \to 0 \tag{7.21}
\]
as \(k \to \infty\), for all \(h \in L^2(\mathbb{P}^X)\). In the proof of Proposition 2 we showed that \(\|\hat{\Pi} V_1 \hat{\Pi} V_2 h\| \leq \rho_0 \|\hat{\Pi} V_1 h\|\) and \(\|\hat{\Pi} V_2 \hat{\Pi} V_1 h\| \leq \rho_0 \|\hat{\Pi} V_1 h\|\) for all \(h \in L^2(\mathbb{P}^X)\). As in (7.21) this implies
\[
\|(\hat{\Pi} V_1 - \sum_{j=1}^{k} (\hat{\Pi} V_1 \hat{\Pi} V_2) j (1 - \hat{\Pi} V_1)) (f_2 - \hat{\Pi} V_2 f_2)\|_n \leq \frac{1}{1 - \rho_0^2} \|\hat{\Pi} V_1 (f_2 - \hat{\Pi} V_2 f_2)\|_n. \tag{7.22}
\]
From (7.21) and (7.22) we conclude that
\[
\|(\hat{\Pi} V f_2)\|_n^2 = \|(\hat{\Pi} V(f_2 - \hat{\Pi} V_2 f_2))\|_n^2 \leq \frac{1}{(1 - \rho_0^2)^2} \|\hat{\Pi} V_1 (f_2 - \hat{\Pi} V_2 f_2)\|_n^2.
\]
Combining this with (7.20) and (6.8) gives (7.17). This completes the proof.
\(\square\)

7.2. End of proof of Theorem 3 and 4. In the proof of Theorem 2 we showed that
\[
\mathbb{E} \left[ \|f_1 - \hat{f}_1, W_1, V_1, V_2\|^2 \right] \leq \|f_1 - \Pi W_1 f_1\|^2 + \frac{1 + \delta \sigma^2 \dim W_1}{1 - \delta} n + R_{n2}
\]
\[
+ \frac{1 + \delta}{1 - \delta} \mathbb{E} \left[ 1_{\epsilon_n} \|\Pi W_1 f_1 - (\hat{\Pi} V f_1)\|^2 \right]. \tag{7.23}
\]
Instead of applying (6.11) and (6.12) to the last term in (7.23), we present an alternative analysis which is possible under the additional assumptions of Theorem 4. We decompose
\[
\Pi W_1 f_1 - (\hat{\Pi} V f)_1 = \Pi W_1 f_1 - f_1 + f_1 - (\Pi V f_1)_1 - (\Pi V f_2)_1 + (\Pi V f)_1 - (\hat{\Pi} V f)_1.
\]
Thus
\[\mathbb{E} \left[ \|\epsilon_{\delta} \| \Pi_{w_1} f_1 - (\hat{\Pi}_V f)_1 \|_2^2 \right] \leq (1 + \delta) \|\Pi_{w_1} f_1 - f_1 \|^2 + 4(1 + 1/\delta) \left( \|f_1 - (\Pi_V f_1)_1\|^2 + \|\Pi_V f_2\|^2 \right) + 2(1 + 1/\delta) \mathbb{E} \left[ \epsilon_{\delta} \|\Pi_V f_1 - (\hat{\Pi}_V f)_1 \|^2 \right].\]

Applying this, Lemma 2, and Proposition 4 and 5, we obtain
\[\mathbb{E} \left[ \|\epsilon_{\delta} \| \Pi_{w_1} f_1 - (\hat{\Pi}_V f)_1 \|_2^2 \right] \leq (1 + \delta) \|\Pi_{w_1} f_1 - f_1 \|^2 + 4 \left( \frac{1}{1 - \rho_0^2} \|f_1 - \Pi_V f_1\|^2 + \frac{1}{(1 - \rho_0^2)^2} \|h_1\|^2 \right) + 2 \left( \frac{1}{1 - \rho_0^2} \frac{\varphi^2 d}{n} \|f - \Pi_V f\|^2 \right).
\]

Inserting this into (7.23) gives Theorem 4. In order to prove Theorem 3 we decompose
\[\Pi_{w_1} f_1 - (\hat{\Pi}_V f)_1 = \Pi_{w_1} f_1 - f_1 + (\Pi_V f_1)_1 + (\Pi_V f_1)_1 - (\hat{\Pi}_V f)_1.\]

Now we proceed as above, using Proposition 6 and the bound
\[\mathbb{E} \left[ \|\epsilon_{\delta} \| \Pi_V f_1 - (\hat{\Pi}_V f)_1 \|_2^2 \right] \leq \frac{1}{1 - \rho_0^2} \mathbb{E} \left[ \epsilon_{\delta} \|\Pi_V f_1 - \Pi_V f\|^2 \right] \leq \frac{1}{1 - \rho_0^2} \frac{1}{1 - \delta} \mathbb{E} \left[ \epsilon_{\delta} \|f_1 - \Pi_V f_1\|^2 \right] \leq \frac{1}{1 - \rho_0^2} \frac{1}{1 - \delta} \|f_1 - F_V f_1\|^2 \leq \frac{1}{1 - \rho_0^2} \frac{1}{1 - \delta} \|f_1 - \Pi_{w_1} f_1\|^2.
\]

This completes the proof. \(\square\)

**APPENDIX A. PROOF OF (5.2)**

In this appendix we prove (5.2). As mentioned in the proof of Theorem 11, the main arguments are taken from [3, page 139 and 140]. We define the event \(A = \{\|(\hat{f}_V)_1\|_\infty \leq k_n\}\). Then
\[\mathbb{E} \left[ \|f_1 - \hat{f}_{1,V_1,V_2}\|^2 \right] = \mathbb{E} \left[ (\epsilon_{\delta} \|A + \epsilon_{\delta} \|A^c + \epsilon_{\delta}^c) \|f_1 - \hat{f}_{1,V_1,V_2}\|^2 \right] \leq \mathbb{E} \left[ \epsilon_{\delta} \|f_1 - (\hat{f}_V)_1\|^2 \right] + \mathbb{E} \left[ \epsilon_{\delta} \|f_1\|^2 \right] + \mathbb{E} \left[ \epsilon_{\delta}^c \|f_1 + k_n\|^2 \right].
\]

Thus it remains to consider the last two terms on the right hand side. By Corollary 5, the second of these two terms 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 first one. If $\mathcal{E}_{\delta}$ holds, we have
\[ \|\hat{f}_{V,1}\|_{2,\infty}^2 \leq \varphi^2 d \|\hat{f}_{V,1}\|^2 \leq \frac{\varphi^2 d}{(1-\delta)(1-\rho_0^2)} \|\hat{f}_V\|_{n,1}^2, \]
where we used (2.4), Assumption 2 and Lemma 2 and (4.2). Using \( \|\hat{f}_V\|_n = \|\hat{\Pi}_V Y\|_n \leq \|\hat{\Pi}_V f\|_n + \|\hat{\Pi}_V \epsilon\|_n \leq \|f\|_n + \|\epsilon\|_n \) and Markov’s inequality, we conclude that
\[ \mathbb{P}(\mathcal{E}_{\delta} \cap \mathcal{A}^c) \leq \mathbb{P} \left( \frac{\varphi^2 d}{(1-\delta)(1-\rho_0^2)} (\|f\|_n + \|\epsilon\|_n)^2 > k_n^2 \right) \]
\[ \leq \frac{2\varphi^2 d (\|f\|^2 + \sigma^2)}{(1-\delta)(1-\rho_0^2) k_n^2}. \]
This completes the proof. \( \blacksquare \)

**Appendix B. A feasible estimator**

In this appendix we show that estimators which are based on the condition 
\((1/n) \sum_{i=1}^n g_1(X_i) = 0\) have (up to a small constant and a term of smaller order) the same upper bound for the risk as estimators based on the condition 
\( \mathbb{E} [g_1(X_1)] = 0 \). Suppose 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 \( V'_1 = \{ g_1 \in U_1 | (1/n) \sum_{i=1}^n g_1(X_i) = 0 \} \) and \( V_1 = \{ g_1 \in U_1 | \mathbb{E} [g_1(X_1)] = 0 \} \). Then our theory only applies to the estimator \( \hat{f}_{V_1,V_2} \). However, since \( V_2 \) contains all constants, we have \( V_1 + V_2 = V'_1 + V_2 \) and therefore
\[ (\hat{f}_{V'_1,V_2})_1 = (\hat{f}_{V_1,V_2})_1 - \frac{1}{n} \sum_{i=1}^n (\hat{f}_{V_1,V_2})_1(X_i^1), \]
i.e., both estimators differ only by a constant. Now, by the bound 
\((x+y)^2 \leq (1+\delta)x^2 + (1+1/\delta)y^2\), we have
\[ \mathbb{E} \left[ 1_{\mathcal{E}_{\delta}} \left( \frac{1}{n} \sum_{i=1}^n (\hat{f}_{V_1,V_2})_1(X_i^1) \right)^2 \right] \]
\[ \leq (1+\delta) \mathbb{E} \left[ 1_{\mathcal{E}_{\delta}} \left( \frac{1}{n} \sum_{i=1}^n (\hat{f}_{V_1,V_2})_1(X_i^1) - f_1(X_i^1) \right)^2 \right] \]
\[ + \left( 1 + \frac{1}{\delta} \right) \mathbb{E} \left[ 1_{\mathcal{E}_{\delta}} \left( \frac{1}{n} \sum_{i=1}^n f_1(X_i^1) \right)^2 \right]. \]
Applying the bound \((1/n) \sum_{i=1}^n x_i^2 \leq (1/n) \sum_{i=1}^n x_i^2 \) and the facts that 
\( \mathbb{E} [f_1(X_i^1)] = 0 \) and that the \( X_i^1 \)'s are independent, this is bounded by
\[ (1+\delta) \mathbb{E} \left[ 1_{\mathcal{E}_{\delta}} \| (\hat{f}_{V_1,V_2})_1 - f_1 \|_n^2 \right] + \left( 1 + \frac{1}{\delta} \right) \frac{\|f_1\|_n^2}{n}. \]
This term can be analyzed as in the proof of Theorem 11.
Finally, I sincerely would like to thank Prof. Enno Mammen for his support and guidance during the preparation of this paper.

REFERENCES

[1] N. Aronszajn. Theory of reproducing kernels. Trans. Amer. Math. Soc., 68:337–404, 1950.
[2] A. V. Balakrishnan. Applied functional analysis. Springer, New York, 2nd edition, 1981.
[3] Y. Baraud. Model selection for regression on a random design. ESAIM Probab. Statist., 6:127–146, 2002.
[4] 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.
[5] L. Birgé and P. Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. Bernoulli, 4:329–375, 1998.
[6] L. Breiman and J. H. Friedman. Estimating optimal transformations for multiple regression and correlation. J. Amer. Statist. Assoc., 80:580–598, 1985.
[7] A. Dalalyan, Y. Ingster, and A. B. Tsybakov. Statistical inference in compound functional models. Probab. Theory Relat. Fields, 2013. doi:10.1007/s00440-013-0487-y.
[8] C. de Boor. A practical guide to splines. Springer, New York, revised edition, 2001.
[9] S. Efromovich. Nonparametric regression with the scale depending on auxiliary variable. Ann. Statist., 41:1542–1568, 2013.
[10] 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.
[11] J. Fan, W. Härdle, and E. Mammen. Direct estimation of low-dimensional components in additive models. Ann. Statist., 26:943–971, 1998.
[12] T. J. Hastie and R. J. Tibshirani. Generalized additive models. Chapman and Hall, London, 1990.
[13] J. Horowitz, J. Klemeli, and E. Mammen. Optimal estimation in additive regression models. Bernoulli, 12:271–298, 2006.
[14] J. Huang, J. L. Horowitz, and F. Wei. Variable selection in nonparametric additive models. Ann. Statist., 38:2282–2313, 2010.
[15] S. Kayalar and H. L. Weinert. Error bounds for the method of alternating projections. Math. Control Signals Systems, 1:43–59, 1988.
[16] H. Kober. A theorem on banach spaces. Compositio Math., 7:135–140, 1939.
[17] V. Koltchinskii and M. Yuan. Sparsity in multiple kernel learning. Ann. Statist., 38:3660–3695, 2010.
[18] O. B. Linton. Efficient estimation of additive nonparametric regression models. Biometrika, 84:469–473, 1997.
[19] E. Mammen. Asymptotics with increasing dimension for robust regression with applications to the bootstrap. Ann. Statist., 17:382–400, 1989.
[20] 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.
[21] L. Meier, S. van de Geer, and P. Bühlmann. High-dimensional additive modeling. Ann. Statist., 37:3779–3821, 2009.
[22] J. D. Opsomer. Asymptotic properties of backfitting estimators. J. Multivariate Anal., 73:166–179, 2000.
[23] J. D. Opsomer and D. Ruppert. Fitting a bivariate additive model by local polynomial regression. *Ann. Statist.*, 25:186–211, 1997.

[24] 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.

[25] 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.

[26] M. Reed and B. Simon. *Methods of modern mathematical physics. I. Functional analysis*. Academic Press, Inc., New York, 2nd edition, 1980.

[27] M. Rudelson. Random vectors in the isotropic position. *J. Funct. Anal.*, 164:60–72, 1999.

[28] M. Rudelson and R. Vershynin. Sampling from large matrices: an approach through geometric functional analysis. *J. ACM*, 54:19 pp, 2007.

[29] C. J. Stone. Additive regression and other nonparametric models. *Ann. Statist.*, 13:689–705, 1985.

[30] S. A. van de Geer. *Applications of empirical process theory*. Cambridge University Press, Cambridge, 2000.

[31] S. A. van de Geer. Penalized least squares for an additive model. preprint, 2012.

[32] J. von Neumann. *Functional Operators. II. The Geometry of Orthogonal Spaces*. Princeton University Press, Princeton, 1950.

[33] E. T. Whittaker and G. N. Watson. *A course of modern analysis*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, reprint of the fourth (1927) edition, 1996.

DEPARTMENT OF ECONOMICS, UNIVERSITY OF MANNHEIM, L7, 3-5, 68131 MANNHEIM, GERMANY

E-mail address: mawahl@mail.uni-mannheim.de