Electronic Journal of Statistics
Vol. 3 (2009) 1305–1321
ISSN: 1935-7524
DOI: 10.1214/09-EJS479

# Semiparametric minimax rates

**James Robins and Eric Tchetgen Tchetgen**

*Department of Biostatistics and Epidemiology*
*School of Public Health*
*Harvard University*
*e-mail:* robins@hsph.harvard.edu
*e-mail:* etchetge@hsph.harvard.edu

**Lingling Li**

*Department of Population Medicine*
*Harvard Medical School and Harvard Pilgrim Health Care*
*Boston, MA, 02215*
*e-mail:* lingling_li@post.harvard.edu

**Aad van der Vaart**

*Department of Mathematics*
*Vrije Universiteit Amsterdam*
*De Boelelaan 1081a*
*1081 HV Amsterdam*
*The Netherlands*
*e-mail:* aad@cs.vu.nl

**Abstract:** We consider the minimax rate of testing or estimation of nonlinear functionals defined on semiparametric models. Existing methods appear not capable of determining a lower bound on the minimax rate if the semiparametric model is indexed by several infinite-dimensional parameters. These methods test a single null distribution to a convex mixture of perturbed distributions. To cope with semiparametric functionals we extend these methods to comparing two convex mixtures. The first mixture is obtained by perturbing a first parameter of the model, and the second by perturbing in addition a second parameter. We obtain a lower bound on the affinity of the resulting pair of mixtures of product measures in terms of three parameters that measure the sizes and asymmetry of the perturbations. We apply the new result to two examples: the estimation of a mean response when response data are missing at random, and the estimation of an expected conditional covariance.

**AMS 2000 subject classifications:** Primary 62G05, 62G20, 62F25.
**Keywords and phrases:** Nonlinear functional, nonparametric estimation, Hellinger distance, missing data, Hellinger affinity, mixtures.

Received September 2009.

## 1. Introduction

Let $X_1, X_2, \ldots, X_n$ be a random sample from a density $p$ relative to a measure $\mu$ on a sample space $(\mathcal{X}, \mathcal{A})$. It is known that $p$ belongs to a collection $\mathcal{P}$ of densities, and we wish to estimate the value $\chi(p)$ of a functional $\chi: \mathcal{P} \to \mathbb{R}$. In

this setting the mimimax rate of estimation of $\chi(p)$ relative to squared error loss can be defined as the root of

$$\inf_{T_n} \sup_{p \in \mathcal{P}} \mathrm{E}_p |T_n - \chi(p)|^2,$$

where the infimum is taken over all estimators $T_n = T_n(X_1, \ldots, X_n)$. Determination of a minimax rate in a particular problem often consists of proving a "lower bound", showing that the mean square error of no estimator tends to zero faster than some rate $\varepsilon_n^2$, combined with the explicit construction of an estimator with mean square error $\varepsilon_n^2$.

The lower bound is often proved by a testing argument, which tries to separate two subsets of the set $\{P^n : p \in \mathcal{P}\}$ of possible distributions of the observation $(X_1, \ldots, X_n)$. Even though testing is a statistically easier problem than estimation under quadratic loss, the corresponding minimax rates are often of the same order. The testing argument can be formulated as follows. *If $P_n$ and $Q_n$ are in the convex hulls of the sets $\{P^n : p \in \mathcal{P}, \chi(p) \leq 0\}$ and $\{P^n : p \in \mathcal{P}, \chi(p) \geq \varepsilon_n\}$ and there exists no sequence of tests $\phi_n = \phi_n(X_1, \ldots, X_n)$ of $P_n$ versus $Q_n$ with both error probabilities $P_n\phi_n$ and $Q_n(1-\phi_n)$ tending to zero, then the minimax rate is not faster than a multiple of $\varepsilon_n$.* See [5], [4] page 47, or [1] Corollary 1. For easy reference we also present a readily applicable version of the result in the appendix.

Here existence of a sequence of tests with errors tending to zero (a *perfect sequence of tests*) is determined by the asymptotic separation of the sequences $P_n$ and $Q_n$ and can be described, for instance, in terms of the *Hellinger affinity*

$$\rho(P_n, Q_n) = \int \sqrt{dP_n}\sqrt{dQ_n}.$$

If $\rho(P_n, Q_n)$ is bounded away from zero as $n \to \infty$, then no perfect sequence of tests exists (see [5] or e.g. Section 14.5 in [10]).

One difficulty in applying this simple argument is that the relevant (approximately least favorable) two sequences of measures $P_n$ and $Q_n$ need not be product measures, but can be arbitrary convex combinations of product measures. In particular, it appears that for nonlinear functionals at least one of the two sequences must be a true mixture. This complicates the computation of the affinity $\rho(P_n, Q_n)$ considerably. Birgé and Massart [1] derived an elegant lower bound on the affinity when $P_n$ is a product measure and $Q_n$ a convex mixture of product measures, and used it to determine the testing rate for functionals of the type $\int f \circ p \, d\mu$, for a given smooth function $f : \mathbb{R} \to \mathbb{R}$, the function $f(x) = x^2$ being the crucial example. Other versions of this argument, which can also be conveniently framed using the chisquare-distance, can be found in e.g. [3] and [9].

In this paper we are interested in structured models $\mathcal{P}$ that are indexed by several subparameters and where the functional is defined in terms of the subparameters. It appears that testing a product versus a mixture is often not least favorable in this situation, but testing two mixtures is. Thus we extend the bound of [1] to the case that both $P_n$ and $Q_n$ are mixtures. In our examples

$P_n$ is equal to a convex mixture obtained by perturbing a first parameter of the model, and $Q_n$ is obtained by perturbing in addition a second parameter. We also refine the bound in other, less essential directions.

The main general result of the paper is stated in Section 2. In Sections 3 and 4 we use this result to obtain a (sharp) lower bound on the minimax rate in two examples of interest. The proof of the main result can be found in Section 5.

### 1.1. Notation

In our examples our a-priori models for the parameters are subsets of Hölder spaces. We define $\|\cdot\|_\alpha$ as the norm of the Hölder space $C^\alpha[0,1]^d$ of $\alpha$-smooth functions on $[0,1]^d$ (cf. e.g. Section 2.7.1 in [11]). The notation $a \lesssim b$ means $a \le Cb$ for a universal constant $C$, and $a \sim b$ means $a \lesssim b$ and $b \lesssim a$.

## 2. Main result

For $k \in \mathbb{N}$ let $\mathcal{X} = \cup_{j=1}^k \mathcal{X}_j$ be a measurable partition of the sample space. Given a vector $\lambda = (\lambda_1, \ldots, \lambda_k)$ in some product measurable space $\Lambda = \Lambda_1 \times \cdots \times \Lambda_k$ let $P_\lambda$ and $Q_\lambda$ be probability measures on $\mathcal{X}$ such that

(1) $P_\lambda(\mathcal{X}_j) = Q_\lambda(\mathcal{X}_j) = p_j$ for every $\lambda \in \Lambda$, for some probability vector $(p_1, \ldots, p_k)$.
(2) The restrictions of $P_\lambda$ and $Q_\lambda$ to $\mathcal{X}_j$ depend on the $j$th coordinate $\lambda_j$ of $\lambda = (\lambda_1, \ldots, \lambda_k)$ only.

For $p_\lambda$ and $q_\lambda$ densities of the measures $P_\lambda$ and $Q_\lambda$ that are jointly measurable in the parameter $\lambda$ and the observation, and $\pi$ a probability measure on $\Lambda$, define $p = \int p_\lambda \, d\pi(\lambda)$ and $q = \int q_\lambda \, d\pi(\lambda)$, and set

$$a = \max_j \sup_\lambda \int_{\mathcal{X}_j} \frac{(p_\lambda - p)^2}{p_\lambda} \frac{d\mu}{p_j},$$
$$b = \max_j \sup_\lambda \int_{\mathcal{X}_j} \frac{(q_\lambda - p_\lambda)^2}{p_\lambda} \frac{d\mu}{p_j},$$
$$d = \max_j \sup_\lambda \int_{\mathcal{X}_j} \frac{(q - p)^2}{p_\lambda} \frac{d\mu}{p_j}.$$

**Theorem 2.1.** *If $np_j(1 \vee a \vee b) \le A$ for all $j$ and $\underline{B} \le p_\lambda \le \overline{B}$ for positive constants $A, \underline{B}, \overline{B}$, then there exists a constant $C$ that depends only on $A, \underline{B}, \overline{B}$ such that, for any product probability measure $\pi = \pi_1 \otimes \cdots \otimes \pi_k$,*

$$\rho\Big(\int P_\lambda^n \, d\pi(\lambda), \int Q_\lambda^n \, d\pi(\lambda)\Big) \ge 1 - Cn^2(\max_j p_j)(b^2 + ab) - Cnd.$$

The main difference with the bound obtained in [1] is that the affinity on the left side concerns two mixtures, rather than a product measure and a mixture. The measures $P_\lambda$ in the first mixture may be viewed as perturbations of the

measure with density $p$, with the parameter $a$ giving the size of these perturbations. The measures $Q_\lambda$ in the second mixture may be viewed as perturbations of the perturbations, of relative sizes measured by the parameter $b$. The parameters $a$ and $b$ enter into the bound on the right side in an asymmetric way. This is appropriate if they concern perturbations of different types, as in our examples of semiparametic models indexed by two functional parameters, which can be perturbed independently.

The parameter $d$ measures the size of the difference of the average perturbations, and should be small if the perturbations are inserted symmetrically. In our examples this parameter is identically zero.

For $a = 0$ and $d = 0$ the bound reduces to the one given in [1], apart from the fact that we consider general "priors" $\pi_j$ rather than measures supported on two points. We believe the latter generalization is not essential, but does make the result and its proof more transparent.

The theorem can be applied to proving a lower bound on a minimax rate by constructing perturbations such that the difference

$$\min_\lambda \chi(q_\lambda) - \max_\lambda \chi(p_\lambda)$$

of the functional of interest on the two types of perturbations is as large as possible, while the parameters $a, b, d$ are small enough to keep the right side of the theorem bounded away from zero. This is somewhat of an art, although for standard model classes the form of the perturbations seems to take a standard form. In the semiparametric case the main issue is where and how to insert these perturbations. We illustrate this in the next two sections on two examples.

The proof of the theorem is deferred to Section 5.

## 3. Estimating the mean response in missing data models

Suppose that a typical observation is distributed as $X = (YA, A, Z)$, for $Y$ and $A$ taking values in the two-point set $\{0, 1\}$ and conditionally independent given $Z$. We are interested in estimating the expected value $\mathrm{E}Y$ of $Y$.

This model is a canonical example of a study with missing response variable, which arises frequently in biostatistical studies. The value of a response variable of interest can often not be ascertained for some subset of the study population. Ignoring this fact would lead to a bias in the estimate of the response distribution. To avoid bias, covariate information is obtained for the complete population. The covariate is chosen so that it can explain why some responses are missing, or at least can explain causes that also influence the response. In the language of missing data models, the covariate should be such that, given the value of the covariate, a response is "missing at random".

This assumption can be described precisely for our model, as follows. The variable $Y$ is the response, and the variable $A$ indicates whether it is observed ($A = 1$, implying $AY = Y$) or not ($A = 0$, implying $AY = 0$). The indicator $A$ is always observed, and the "missing at random" assumption is made precise in the assumption that $Y$ and $A$ are conditionally independent given the covariate

$Z$. To make this true the covariate must contain the information on possible dependence between response and missingness. The conditional independence of $Y$ and $A$ can equivalently be described by saying that the conditional law of $Y$ given $(Z, A)$ is independent of the value of $A$. From this it is seen that $\mathrm{E}Y = \mathrm{EE}(YA|\, A = 1, Z)$. Thus the assumption of "missing at random" renders the parameter of interest $\mathrm{E}Y$ identifiable from the observed data.

While the introduction of the covariate is necessary to make the parameter of interest identifiable, it also comes at a price: the statistical model for the data $X = (YA, A, Z)$ will include the conditional laws of $Y$ and $A$ given $Z$, and the marginal law of $Z$. If these laws are only nonparametrically specified, then the resulting problem of estimating $\mathrm{E}Y$ is semiparametric, and may involve "smoothing". If the covariate $Z$ is high-dimensional, then this may lead to slow convergence rates. We shall assume that $Z$ is $d$-dimensional, and for definiteness assume that it takes its values in $\mathcal{Z} = [0, 1]^d$.

The model can be parameterized by the marginal density $f$ of $Z$ relative to Lebesgue measure measure $\nu$ on $\mathcal{Z}$, and the probabilities $b(z) = \mathrm{P}(Y = 1|\, Z = z)$ and $a(z)^{-1} = \mathrm{P}(A = 1|\, Z = z)$. Alternatively, the model can be parameterized by the function $g = f/a$, which is the conditional density of $Z$ given $A = 1$ up to the norming factor $\mathrm{P}(A = 1)$. Under this latter parametrization which we adopt henceforth, the density $p$ of an observation $X$ is described by the triple $(a, b, g)$ and the functional of interest $E\{E[Y|A = 1, Z]\}$ is expressed as

$$\chi(p) = \int abg\, d\nu.$$

The parameterization through $g$ rather than $f$ appears to correspond to an essential feature of the structure of the observational model. On the other hand, the parameterization of $\mathrm{P}(A = 1|\, Z = z)$ by the inverse of $a$ rather than this function itself is for convenience of notation, as our a-priori model, imposing smoothness of $a$, does not change if $a$ is replaced by $1/a$.

Define $\|\cdot\|_\alpha$ as the norm of the Hölder space $C^\alpha[0, 1]^d$ of $\alpha$-smooth functions on $[0, 1]^d$ (cf. e.g. Section 2.7.1 in [11]). For given positive constants $\alpha, \beta, \gamma$ and $m, M$, we consider the models

- $\mathcal{P}_1 = \{(a, b, g)$: $\|a\|_\alpha, \|b\|_\beta \le M, g = 1/2, m \le a^{-1}, b \le \underline{1} - M\}$.
- $\mathcal{P}_2 = \{(a, b, g)$: $\|a\|_\alpha, \|b\|_\beta, \|g\|_\gamma \le M, m \le a^{-1}, b \le \underline{1} - M, g \ge m\}$.

If $(\alpha+\beta)/2 \ge d/4$, then a $\sqrt{n}$-rate is attainable over $\mathcal{P}_2$ (see [6]), and a standard "two-point" proof can show that this rate cannot be improved. Here we are interested in the case $(\alpha + \beta)/2 < d/4$, when the rate becomes slower than $1/\sqrt{n}$. The paper [6] (or in part [7]) constructs an estimator that attains the rate

$$n^{-(2\alpha+2\beta)/(2\alpha+2\beta+d)}$$

uniformly over $\mathcal{P}_2$ if

$$\frac{\gamma}{2\gamma + d} > \Big(\frac{\alpha \vee \beta}{d}\Big)\Big(\frac{d - 2\alpha - 2\beta}{d + 2\alpha + 2\beta}\Big) := \gamma(\alpha, \beta). \tag{3.1}$$

We shall show that this result is optimal by showing that the minimax rate over the smaller model $\mathcal{P}_1$ is lower bounded by the same rate.

In the case that $\alpha = \beta$ these results can be proved using the method of [1], but in general we need a construction as in Section 2 with $P_\lambda$ based on a perturbation of the coarsest parameter of the pair $(a, b)$ and $Q_\lambda$ constructed by perturbing in addition the smoothest of the two parameters. The rate $n^{-4\beta/(4\beta+d)}$ obtained if $\alpha = \beta$ is the same as the rate for estimating $\int p^2\,d\mu$ based on a sample from a $\beta$-smooth density $p$. In that sense the semiparametric structure changes the essence of the problem only if $\alpha \neq \beta$ (if (3.1) holds).

Because the left side of (3.1) is increasing in $\gamma$, this assumption requires that the (conditional) covariate density $g$ is smooth enough (relative to $a$ and $b$). We believe that the rate for estimating the functional may be slower if this condition fails. We have obtained some upper bounds in the situation of a very unsmooth covariate density, but not a closed theory, and do not address this situation in this paper.

**Theorem 3.1.** *If $(\alpha+\beta)/2 < d/4$, then the minimax rate over $\mathcal{P}_1$ for estimating $\int abg\,d\nu$ is at least $n^{-(2\alpha+2\beta)/(2\alpha+2\beta+d)}$.*

*Proof.* Let $H\colon \mathbb{R}^d \to \mathbb{R}$ be a $C^\infty$ function supported on the cube $[0, 1/2]^d$ with $\int H\,d\nu = 0$ and $\int H^2\,d\nu = 1$. Let $k$ be the integer closest to $n^{2d/(2\alpha+2\beta+d)}$ and let $\mathcal{Z}_1, \ldots, \mathcal{Z}_k$ be translates of the cube $k^{-1/d}[0, 1/2]^d$ that are disjoint and contained in $[0, 1]^d$. For $z_1, \ldots, z_k$ the bottom left corners of these cubes and $\lambda = (\lambda_1, \ldots, \lambda_k) \in \Lambda = \{-1, 1\}^k$, let

$$a_\lambda(z) = 2 + \left(\frac{1}{k}\right)^{\alpha/d} \sum_{i=1}^{k} \lambda_i H\big((z - z_i)k^{1/d}\big),$$

$$b_\lambda(z) = \frac{1}{2} + \left(\frac{1}{k}\right)^{\beta/d} \sum_{i=1}^{k} \lambda_i H\big((z - z_i)k^{1/d}\big).$$

These functions can be seen to be contained in $C^\alpha[0, 1]^d$ and $C^\beta[0, 1]^d$ with norms that are uniformly bounded in $k$. We choose a uniform prior $\pi$ on $\lambda$, so that $\lambda_1, \ldots, \lambda_k$ are i.i.d. Rademacher variables.

We partition the sample space $\{0, 1\} \times \{0, 1\} \times \mathcal{Z}$ into the sets $\mathcal{X}_j = \{0, 1\} \times \{0, 1\} \times \mathcal{Z}_j$ and the remaining set. An observation $X$ falls in such a set if and only if the corresponding covariate $Z$ falls in $\mathcal{Z}_j$. Because $Z$ has density $f_\lambda = ga_\lambda = \frac{1}{2}a_\lambda$ under model $\mathcal{P}_1$ and $\frac{1}{2}a_\lambda$ is a perturbation of the uniform density obtained by redistributing mass within each $\mathcal{Z}_j$ (as $\int H\,d\nu = 0$) we have that $P_\lambda(\mathcal{X}_j) = Q_\lambda(\mathcal{X}_j) = p_j$ is independent of $j$ and of the order $k^{-1}$.

The likelihood for the model $\mathcal{P}_1$, indexed by $(a, b, 1)$, can be written as

$$(a - 1)^{1-A}(Z)\Big(b^Y(Z)(1 - b)^{1-Y}(Z)\Big)^A.$$

Because $\int H\,d\nu = 0$ the values of the functional $\int abg\,d\nu$ at the parameter values $(a_\lambda, 1/2, 1/2)$ and $(2, b_\lambda, 1/2)$ are both equal to $1/2$, whereas the value at

$(a_\lambda, b_\lambda, 1/2)$ is equal to

$$\int a_\lambda b_\lambda \frac{d\nu}{2} = \frac{1}{2} + \Big(\frac{1}{k}\Big)^{\alpha/d+\beta/d} \int \left(\sum_{i=1}^{k} H\big((z-z_i)k^{1/d}\big)\right)^2 \frac{d\nu}{2} = \frac{1}{2} + \frac{1}{2}\Big(\frac{1}{k}\Big)^{\alpha/d+\beta/d}.$$

The minimax rate is not faster than $(1/k)^{\alpha/d+\beta/d}$ for $k = k_n$ such that the convex mixtures of the products of the perturbations do not separate completely as $n \to \infty$. We choose the mixtures differently in the cases $\alpha \le \beta$ and $\alpha \ge \beta$.

Assume $\alpha \le \beta$. We define $p_\lambda$ by the parameter $(a_\lambda, 1/2, 1/2)$ and $q_\lambda$ by the parameter $(a_\lambda, b_\lambda, 1/2)$. Because $\int a_\lambda \, d\pi(\lambda) = 2$ and $\int b_\lambda \, d\pi(\lambda) = 1/2$, we have

$$\begin{aligned} p(X) &:= \int p_\lambda(X) \, d\pi(\lambda) = \Big(b^Y(Z)(1-b)^{1-Y}(Z)\Big)^A, \\ (p_\lambda - p)(X) &= (1-A)(a_\lambda - 2)(Z), \\ (q_\lambda - p_\lambda)(X) &= A(b_\lambda - 1/2)^Y (1/2 - b_\lambda)^{1-Y}, \\ (q-p)(X) &:= \int (q_\lambda - p_\lambda)(X) \, d\pi(\lambda) = 0. \end{aligned}$$

Therefore, it follows that the number $d$ in Theorem 2.1 vanishes, while the numbers $a$ and $b$ are bounded above by

$$\max_j \int_{\mathcal{Z}_j} \frac{(a_\lambda - 2)^2}{p_\lambda} \frac{d\nu}{p_j}, \qquad \text{and} \qquad \max_j \int_{\mathcal{Z}_j} \frac{(b_\lambda - 1/2)^2}{p_\lambda} \frac{d\nu}{p_j},$$

respectively. Because $p_\lambda$ is bounded away from zero, these two expressions are equal to $k^{-2\alpha/d}$ and $k^{-2\beta/d}$ times a multiple of

$$\max_j \int_{\mathcal{Z}_j} \left(\sum_{i=1}^{k} \lambda_i H\big((z-z_i)k^{1/d}\big)\right)^2 \frac{d\nu}{p_j} \sim 1,$$

as $p_j \sim 1/k$. Theorem 2.1 shows that there exists a constant $C'$ such that

$$\rho\Big(\int P_\lambda^n \, d\pi(\lambda), \int Q_\lambda^n \, d\pi(\lambda)\Big) \ge 1 - C' n^2 \frac{1}{k}\Big(k^{-4\beta/d} + k^{-2\alpha/d} k^{-2\beta/d}\Big). \tag{3.2}$$

For $k \sim n^{2d/(2\alpha+2\beta+d)}$ the right side is bounded away from 0. Substitution of this number in the magnitude of separation $(1/k)^{\alpha/d+\beta/d}$ leads to the rate as claimed in the theorem.

Assume $\alpha \ge \beta$. We define $p_\lambda$ by the parameter $(2, b_\lambda, 1/2)$ and $q_\lambda$ by the parameter $(a_\lambda, b_\lambda, 1/2)$. The computations are very similar to the ones in the case $\alpha \le \beta$. □

## 4. Estimating an expected conditional covariance

Let a typical observation take the form $X = (Y, A, Z)$, where $Y$ and $A$ are dichotomous, with values in $\{0, 1\}$, and $Z$ takes its values in $\mathcal{Z} = [0, 1]^d$. Different

from the previous section we always observe $Y$; also we do not assume that $Y$ and $A$ are conditionally independent given $Z$. We are interested in estimating the expected conditional covariance $\mathrm{E}\,\mathrm{cov}(Y, A|\, Z)$.

This functional often arises in the biostatistical and epidemiological literature on estimation of the effect of a binary treatment, in the following manner. The most common model used in observational studies to analyze the causal effect of a treatment $A$ on a continuous response $Y$ in the presence of a vector $Z$ of continuous pretreatment confounding variables is the semiparametric regression model

$$\mathrm{E}(Y|\, A, Z) = \xi A + \nu(Z), \tag{4.1}$$

where $\xi$ is an unknown parameter and $\nu$ is an unknown function. Specifically, it is shown in [8] that this model arises whenever we assume (i) no unmeasured confounders (i.e. ignorability of treatment $A$ within levels of $Z$) and (ii) a constant additive effect of treatment $A$ on the mean of $Y$. A nonzero estimate of the parameter $\xi$ in this model indicates a causal effect of the treatment $A$.

In situations where the assumption of additivity is not expected to hold, one may more generally consider the the *treatment effect function*

$$c(z) = \mathrm{E}(Y|\, A = 1, Z = z) - \mathrm{E}(Y|\, A = 0, Z = z),$$

which measures the difference in effect of the treatments as a function of covariate value. Under model (4.1) this reduces to the constant $\xi$. More generally, a (weighted) average of this function could be used as a summary measure of the causal effect of the treatment. A convenient average (see [2] for further discussion) is the *variance-weighted average treatment effect*, given by

$$\tau := \frac{\mathrm{E}\,\mathrm{var}(A|\, Z)c(Z)}{\mathrm{E}\,\mathrm{var}(A|\, Z)} = \frac{\mathrm{E}\,\mathrm{cov}(Y, A|\, Z)}{\mathrm{E}\,\mathrm{var}(A|\, Z)}. \tag{4.2}$$

Under model (4.1) this again reduces to the constant $\xi$. (The equality follows by simple algebra, for instance starting from the identity $\mathrm{cov}(Y, A|\, Z) = \big(\mathrm{E}(Y|\, A = 1, Z) - \mathrm{E}(Y|\, Z)\big)\mathrm{E}(A|\, Z)$ and the corresponding identity for $\mathrm{cov}(Y, 1 - A)$.)

A convenient method of inference for the parameter $\tau$ is as follows. For any $t \in \mathbb{R}$, define $Y(t) = Y - tA$ and the corresponding functional $\psi(t) = \mathrm{E}\,\mathrm{cov}\big(Y(t), A|\, Z\big)$. It is easily checked that the parameter $\tau$ given in (4.2) is the unique solution to the equation $\psi(t) = 0$. Thus inference on $\tau$ can be obtained by "inverting" the inference on $\psi(t)$, and we can aim on the estimation of the functional $\psi(t)$ at a fixed value of $t$. For simplicity we restrict ourselves to the estimation of $\psi(0)$, which is the expected conditional covariance $\mathrm{E}\,\mathrm{cov}(Y, A|\, Z)$.

The expected conditional covariance functional can be decomposed as $\mathrm{E}YA - \mathrm{E}\mathrm{E}(Y|\, Z)\mathrm{E}(A|\, Z)$, where the first part $\mathrm{E}YA$ is estimable at $1/\sqrt{n}$-rate by the empirical mean of the variables $Y_i A_i$. Therefore, we shall concentrate on obtaining a lower bound for estimating the functional $\mathrm{E}\mathrm{E}(Y|\, Z)\mathrm{E}(A|\, Z)$.

Define functions $a$, and $b$ by $a(z) = \mathrm{P}(A = 1|Z = z)$, and $b(z) = \mathrm{P}(Y = 1|Z = z)$, furthermore, recall that $c(z) = \mathrm{P}(Y = 1|\, A = 1, Z = z) - \mathrm{P}(Y = 1|A = 0, Z = z)$. In view of the dichotomous nature of $A$, we then have

$$\mathrm{P}(Y = 1|A, Z) = c(Z)\big(A - a(Z)\big) + b(Z).$$

Therefore we can parameterize the model by the quadruple $(a, b, c, f)$, for $f$ the marginal density of $Z$. In terms of this parameterization the functional of interest is

$$\chi(p) = \int abf \, d\nu.$$

The parameter $c$, which vanishes if $Y$ and $A$ are conditionally independent given $Z$, does not enter into the functional, and appears not to be important for the problem otherwise either.

For given positive constants $\alpha, \beta, \gamma$ and $m, M$, we consider the models

- $\mathcal{P}_1 = \{(a, b, c, f)\colon \|a\|_\alpha, \|b\|_\beta, \|c\|_{\alpha\wedge\beta} \le M, f = 1, m \le a, b \le 1 - m\}$,
- $\mathcal{P}_2 = \{(a, b, c, f)\colon \|a\|_\alpha, \|b\|_\beta, \|f\|_\gamma \le M, m \le a, b \le 1 - m, f \ge m\}$.

We are mainly interested in the case $(\alpha+\beta)/2 < d/4$, when the rate of estimation of $\chi(p)$ is slower than $1/\sqrt{n}$. The paper [6] constructs an estimator that attains the rate

$$n^{-(2\alpha+2\beta)/(2\alpha+2\beta+d)}$$

uniformly over $\mathcal{P}_2$ if equation (3.1) holds. (The same rate as in Section 3.) We shall show that this rate is optimal by showing that the minimax rate over the smaller model $\mathcal{P}_1$ is not faster.

**Theorem 4.1.** *If $(\alpha+\beta)/2 < d/4$, then the minimax rate over $\mathcal{P}_1$ for estimating $\int abf \, d\nu$ is at least $n^{-(2\alpha+2\beta)/(2\alpha+2\beta+d)}$.*

*Proof.* We use the same partition of the sample space, the same function $H$, and the same priors as in the proof of Theorem 3.1. The likelihood of the model $\mathcal{P}_1$, indexed by the parameter $(a, b, c, 1)$, is given by

$$\begin{aligned} a(Z)^A(1-a)(Z)^{1-A}\big(c(1-a)+b\big)(Z)^{YA}\big(1-c(1-a)-b\big)(Z)^{(1-Y)A} \\ \times(-ca+b)(Z)^{Y(1-A)}(1+ca-b)(Z)^{(1-Y)(1-A)}. \end{aligned}$$

The perturbations are defined differently in the cases $\alpha < \beta$ and $\alpha \ge \beta$.

Assume $\alpha < \beta$. Set

$$\begin{aligned} a_\lambda(z) &= \frac{1}{2} + \Big(\frac{1}{k}\Big)^{\alpha/d}\sum_{i=1}^{k}\lambda_i H\big((z - z_i)k^{1/d}\big), \\ b_\lambda(z) &= \frac{1}{2} + \Big(\frac{1}{k}\Big)^{\beta/d}\sum_{i=1}^{k}\lambda_i H\big((z - z_i)k^{1/d}\big), \\ c_\lambda(z) &= \frac{1/2 - b_\lambda(z)}{1 - a_\lambda(z)}. \end{aligned}$$

At the parameter values $(a_\lambda, 1/2, 0, 1)$ the functional takes the value $\int a_\lambda b_\lambda 1 \, d\nu = 1/4$ and the likelihood is $p_\lambda(X) = \frac{1}{2}a_\lambda(Z)^A(1 - a_\lambda)(Z)^{1-A}$, whereas at the parameter values $(a_\lambda, b_\lambda, c_\lambda, 1)$ the functional attains the value $1/4 + (1/k)^{\alpha/d+\beta/d}$ and the likelihood is given by $q_\lambda(X) = (a_\lambda/2)(Z)^A(b_\lambda -$

$a_\lambda/2)(Z)^{Y(1-A)}(1-b_\lambda-a_\lambda/2)(Z)^{(1-Y)(1-A)}$. We conclude that

$$\begin{aligned} p(X) &:= \int p_\lambda(X)\,d\pi(\lambda) = 1/4, \\ (p_\lambda - p)(X) &= \tfrac{1}{2}(a_\lambda - 1/2)(Z)^A(1/2-a_\lambda)(Z)^{1-A}, \\ (q_\lambda - p_\lambda)(X) &= (1-A)(b_\lambda - 1/2)(Z)^Y(1/2-b_\lambda)(Z)^{1-Y}, \\ (q-p)(X) &:= \int (q_\lambda - p_\lambda)(X)\,d\pi(\lambda) = 0. \end{aligned}$$

Therefore, it follows that the number $d$ in Theorem 2.1 vanishes, while the numbers $a$ and $b$ are bounded above by a multiple of

$$\max_j \int_{\mathcal{Z}_j} \frac{h^2}{p_\lambda}\frac{d\nu}{p_j}$$

for $h$ equal to the functions $a_\lambda - 1/2$ and $b_\lambda - 1/2$, respectively. As in the proof of Theorem 3.1 these two expressions are of the orders $k^{-2\alpha/d}$ and $k^{-2\beta/d}$, respectively.

Theorem 2.1 shows again that (3.2) holds, and for $k \sim n^{2d/(2\alpha+2\beta+d)}$ the two mixtures have affinity bounded away from zero. For this choice of $k$ the separation of the functional is the minimax rate $n^{-(2\alpha+2\beta)/(2\alpha+2\beta+d)}$.

Assume that $\alpha \geq \beta$. Even though the model and the functional of interest is symmetric in $Y$ and $A$, the parameterization of the model is not, and therefore this case needs to be treated separately. Define the functions $a_\lambda$ and $b_\lambda$ as above, and set

$$c_\lambda = \frac{(1/2-a_\lambda)b_\lambda}{(1-a_\lambda)a_\lambda}.$$

At the parameter values $(1/2, b_\lambda, 0, 1)$ the functional takes the value $1/4$ with corresponding likelihood $p_\lambda(X) = \frac{1}{2}b_\lambda(Z)^Y(1-b_\lambda)(Z)^{1-Y}$, whereas at the parameter values $(a_\lambda, b_\lambda, c_\lambda, 1)$ the functional attains the value $1/4 + k^{-\alpha/d-\beta/d}$ and the likelihood is given by $q_\lambda(X) = (b_\lambda/2)(Z)^Y(a_\lambda - b_\lambda/2)(Z)^{(1-Y)A}(1 - a_\lambda - b_\lambda/2)(Z)^{(1-Y)(1-A)}$. It follows that

$$\begin{aligned} p(X) &:= \int p_\lambda(X)\,d\pi(\lambda) = 1/4, \\ (p_\lambda - p)(X) &= \tfrac{1}{2}(b_\lambda - 1/2)(Z)^Y(1/2-b_\lambda)(Z)^{1-Y}, \\ (q_\lambda - p_\lambda)(X) &= (1-Y)(a_\lambda - 1/2)(Z)^A(1/2-a_\lambda)(Z)^{1-A}, \\ (q-p)(X) &:= \int (q_\lambda - p_\lambda)(X)\,d\pi(\lambda) = 0. \end{aligned}$$

These are the same equations as in the case that $\alpha < \beta$, except that $Y$ and $A$ (and $a_\lambda$ and $b_\lambda$) are permuted. Therefore, the proof can be finished as before. □

## 5. Proof of main result

The proof of Theorem 2.1 is based on two lemmas. The first lemma factorizes the affinity between two mixtures of product measures into (conditional) affinities

of certain products of restrictions to the partitioning sets. The latter are next lower bounded using the second lemma. The reduction to the partioning sets is useful, because it reduces the $n$-fold products to lower order products for which the second lemma is accurate.

Define probability measures $P_{j,\lambda_j}$ and $Q_{j,\lambda_j}$ on $\mathcal{X}_j$ by

$$dP_{j,\lambda_j} = \frac{1_{\mathcal{X}_j}\, dP_\lambda}{p_j}, \qquad dQ_{j,\lambda_j} = \frac{1_{\mathcal{X}_j}\, dQ_\lambda}{p_j}. \tag{5.1}$$

**Lemma 5.1.** *For any product probability measure $\pi = \pi_1 \otimes \cdots \otimes \pi_k$ on $\Lambda$ and every $n \in \mathbb{N}$,*

$$\rho\Big(\int P_\lambda^n\, d\pi(\lambda), \int Q_\lambda^n\, d\pi(\lambda)\Big) = \mathrm{E} \prod_{j=1}^k \rho_j(N_j),$$

*where $(N_1, \ldots, N_k)$ is multinomially distributed on $n$ trials with success probability vector $(p_1, \ldots, p_k)$ and $\rho_j\colon \{0, \ldots, n\} \to [0,1]$ is defined by $\rho_j(0) = 1$ and*

$$\rho_j(m) = \rho\Big(\int P_{j,\lambda_j}^m\, d\pi_j(\lambda_j), \int Q_{j,\lambda_j}^m\, d\pi_j(\lambda_j)\Big), \qquad m \geq 1.$$

*Proof.* Set $\bar{P}_n := \int P_\lambda^n\, d\pi(\lambda)$ and consider this as the distribution of the vector $(X_1, \ldots, X_n)$. Then, for $p_\lambda$ and $q_\lambda$ densities of $P_\lambda$ and $Q_\lambda$ relative to some dominating measure, the left side of the lemma can be written as

$$\rho\Big(\int P_\lambda^n\, d\pi(\lambda), \int Q_\lambda^n\, d\pi(\lambda)\Big) = \mathrm{E}_{\bar{P}_n} \sqrt{\frac{\int \prod_{j=1}^k \prod_{i:X_i \in \mathcal{X}_j} q_\lambda(X_i)\, d\pi(\lambda)}{\int \prod_{j=1}^k \prod_{i:X_i \in \mathcal{X}_j} p_\lambda(X_i)\, d\pi(\lambda)}}.$$

Because by assumption on each partitioning set $\mathcal{X}_j$ the measures $Q_\lambda$ and $P_\lambda$ depend on $\lambda_j$ only, the expressions $\prod_{i:X_i \in \mathcal{X}_j} q_\lambda(X_i)$ and $\prod_{i:X_i \in \mathcal{X}_j} p_\lambda(X_i)$ depend on $\lambda$ only through $\lambda_j$. In fact, within the quotient on the right side of the preceding display, they can be replaced by $\prod_{i:X_i \in \mathcal{X}_j} q_{j,\lambda_j}(X_i)$ and $\prod_{i:X_i \in \mathcal{X}_j} p_{j,\lambda_j}(X_i)$ for $q_{j,\lambda_j}$ and $p_{j,\lambda_j}$ densities of the measures $Q_{j,\lambda_j}$ and $P_{j,\lambda_j}$. Because $\pi$ is a product measure, we can next use Fubini's theorem and rewrite the resulting expression as

$$\mathrm{E}_{\bar{P}_n} \sqrt{\frac{\prod_{j=1}^k \int \prod_{i:X_i \in \mathcal{X}_j} q_{j,\lambda_j}(X_i)\, d\pi_j(\lambda_j)}{\prod_{j=1}^k \int \prod_{i:X_i \in \mathcal{X}_j} p_{j,\lambda_j}(X_i)\, d\pi_j(\lambda_j)}}.$$

Here the two products over $j$ can be pulled out of the square root and replaced by a single product preceding it. A product over an empty set (if there is no $X_i \in \mathcal{X}_j$) is interpreted as 1.

Define variables $I_1, \ldots, I_n$ that indicate the partitioning sets that contain the observations: $I_i = j$ if $X_i \in \mathcal{X}_j$ for every $i$ and $j$, and let $N_j = (\#1 \leq i \leq n\colon I_i = j)$ be the number of $X_i$ falling in $\mathcal{X}_j$.

The measure $\bar{P}_n$ arises as the distribution of $(X_1, \ldots, X_n)$ if this vector is generated in two steps. First $\lambda$ is chosen from $\pi$ and next given this $\lambda$ the variables $X_1, \ldots, X_n$ are generated independently from $P_\lambda$. Then given $\lambda$ the vector

$(N_1, \ldots, N_k)$ is multinomially distributed on $n$ trials and probability vector $\big(P_\lambda(\mathcal{X}_1), \ldots, P_\lambda(\mathcal{X}_k)\big)$. Because the latter vector is independent of $\lambda$ and equal to $(p_1, \ldots, p_k)$ by assumption, the vector $(N_1, \ldots, N_k)$ is stochastically independent of $\lambda$ and hence also unconditionally, under $\bar{P}_n$, multinomially distributed with parameters $n$ and $(p_1, \ldots, p_k)$. Similarly, given $\lambda$ the variables $I_1, \ldots, I_n$ are independent and the event $I_i = j$ has probability $P_\lambda(\mathcal{X}_j)$, which is independent of $\lambda$ by assumption. It follows that the random elements $(I_1, \ldots, I_n)$ and $\lambda$ are stochastically independent under $\bar{P}_n$.

The conditional distribution of $X_1, \ldots, X_n$ given $\lambda$ and $I_1, \ldots, I_n$ can be described as: for each partitioning set $\mathcal{X}_j$ generate $N_j$ variables independently from $P_\lambda$ restricted and renormalized to $\mathcal{X}_j$, i.e. from the measure $P_{j,\lambda_j}$; do so independently across the partitioning sets; and attach correct labels $\{1, \ldots, n\}$ consistent with $I_1, \ldots, I_n$ to the $n$ realizations obtained. The conditional distribution under $\bar{P}_n$ of $X_1, \ldots, X_n$ given $I_n$ is the mixture of this distribution relative to the conditional distribution of $\lambda$ given $(I_1, \ldots, I_n)$, which was seen to be the unconditional distribution, $\pi$. Thus we obtain a sample from the conditional distribution under $\bar{P}_n$ of $(X_1, \ldots, X_n)$ given $(I_1, \ldots, I_n)$ by generating for each partitioning set $\mathcal{X}_j$ a set of $N_j$ variables from the measure $\int P_{j,\lambda_j}^{N_j}\, d\pi_j(\lambda_j)$, independently across the partitioning sets, and next attaching labels consistent with $I_1, \ldots, I_n$.

Now rewrite the right side of the last display by conditioning on $I_1, \ldots, I_n$ as

$$\mathrm{E}_{\bar{P}_n} \mathrm{E}_{\bar{P}_n} \left[ \prod_{j=1}^{k} \sqrt{\frac{\int \prod_{i:I_i=j} q_{j,\lambda_j}(X_i)\, d\pi_j(\lambda_j)}{\int \prod_{i:I_i=j} p_{j,\lambda_j}(X_i)\, d\pi_j(\lambda_j)}} \Big| I_1, \ldots, I_n \right].$$

The product over $j$ can be pulled out of the conditional expectation by the conditional independence across the partitioning sets. The resulting expression can be seen to be of the form as claimed in the lemma. □

The second lemma does not use the partitioning structure, but is valid for mixtures of products of arbitrary measures on a measurable space. For $\lambda$ in a measurable space $\Lambda$ let $P_\lambda$ and $Q_\lambda$ be probability measures on a given sample space $(\mathcal{X}, \mathcal{A})$, with densities $p_\lambda$ and $q_\lambda$ relative to a given dominating measure $\mu$, which are jointly measurable. For a given (arbitrary) probability density $p$ define functions $\ell_\lambda = q_\lambda - p_\lambda$ and $\kappa_\lambda = p_\lambda - p$, and set

$$\begin{aligned}
a &= \sup_{\lambda \in \Lambda} \int \frac{\kappa_\lambda^2}{p_\lambda}\, d\mu, \\
b &= \sup_{\lambda \in \Lambda} \int \frac{\ell_\lambda^2}{p_\lambda}\, d\mu, \\
c &= \sup_{\lambda \in \Lambda} \int \frac{p^2}{p_\lambda}\, d\mu, \\
d &= \sup_{\lambda \in \Lambda} \int \frac{\big(\int \ell_\mu\, d\pi(\mu)\big)^2}{p_\lambda}\, d\mu.
\end{aligned}$$

**Lemma 5.2.** *For any probability measure $\pi$ on $\Lambda$ and every $n \in \mathbb{N}$,*

$$\rho\Big(\int P_\lambda^n \, d\pi(\lambda), \int Q_\lambda^n \, d\pi(\lambda)\Big) \geq 1 - \sum_{r=2}^{n} \binom{n}{r} b^r - 2n^2 \sum_{r=1}^{n-1} \binom{n-1}{r} a^r b - 2n^2 c^{n-1} d.$$

*Proof.* Consider the measure $\bar{P}_n = \int P_\lambda^n \, d\pi(\lambda)$, which has density $\bar{p}_n(\vec{x}_n) = \int \prod_{i=1}^n p_\lambda(x_i) \, d\pi(\lambda)$ relative to $\mu^n$, as the distribution of $(X_1, \ldots, X_n)$. Using the inequality $\mathrm{E}\sqrt{1+Y} \geq 1 - \mathrm{E}Y^2/2$, valid for any random variable $Y$ with $1 + Y \geq 0$ and $\mathrm{E}Y = 0$ (see [1], Lemma 1), we see that

$$\begin{aligned}
&\rho\Big(\int P_\lambda^n \, d\pi(\lambda), \int Q_\lambda^n \, d\pi(\lambda)\Big) \\
&\quad = \mathrm{E}_{\bar{P}_n} \sqrt{1 + \frac{\int \big[\prod_{i=1}^n q_\lambda(X_i) - \prod_{i=1}^n p_\lambda(X_i)\big] \, d\pi(\lambda)}{\bar{p}_n(X_1, \ldots, X_n)}} \\
&\quad \geq 1 - \frac{1}{2} \mathrm{E}_{\bar{P}_n} \frac{\int \big[\prod_{i=1}^n q_\lambda(X_i) - \prod_{i=1}^n p_\lambda(X_i)\big] \, d\pi(\lambda)^2}{\bar{p}_n(X_1, \ldots, X_n)^2}.
\end{aligned} \tag{5.2}$$

It suffices to upper bound the expected value on the right side. To this end we expand the difference $\prod_{i=1}^n q_\lambda(X_i) - \prod_{i=1}^n p_\lambda(X_i)$ as $\sum_{|I| \geq 1} \prod_{i \in I^c} p_\lambda(X_i) \times \prod_{i \in I} \ell_\lambda(X_i)$, where the sum ranges over all nonempty subsets $I \subset \{1, \ldots, n\}$. We split this sum in two parts, consisting of the terms indexed by subsets of size 1 and the subsets that contain at least 2 elements, and separate the square of the sum of these two parts by the inequality $(A+B)^2 \leq 2A^2 + 2B^2$.

If $n = 1$, then there are no subsets with at least two elements and the second part is empty. Otherwise the sum over subsets with at least two elements contributes two times

$$\begin{aligned}
&\int \frac{\int \sum_{|I| \geq 2} \prod_{i \in I^c} p_\lambda(x_i) \prod_{i \in I} \ell_\lambda(x_i) \, d\pi(\lambda)^2}{\int \prod_i p_\lambda(x_i) \, d\pi(\lambda)} \, d\mu^n(\vec{x}_n) \\
&\quad \leq \int \int \Big( \sum_{|I| \geq 2} \prod_{i \in I^c} \sqrt{p_\lambda}(x_i) \prod_{i \in I} \frac{\ell_\lambda}{\sqrt{p_\lambda}}(x_i) \Big)^2 \, d\pi(\lambda) \, d\mu^n(\vec{x}_n) \\
&\quad = \sum_{|I| \geq 2} \int \int \prod_{i \in I^c} p_\lambda(x_i) \prod_{i \in I} \frac{\ell_\lambda^2}{p_\lambda}(x_i) \, d\pi(\lambda) \, d\mu^n(\vec{x}_n).
\end{aligned}$$

Here to derive the first inequality we use the inequality $(\mathrm{E}U)^2/\mathrm{E}V \leq \mathrm{E}(U^2/V)$, valid for any random variables $U$ and $V \geq 0$, which can be derived from Cauchy-Schwarz' or Jensen's inequality. The last step follows by writing the square of the sum as a double sum and noting that all off-diagonal terms vanish, as they contain at least one "loose" $\ell_\lambda$ and $\int \ell_\lambda \, d\mu = 0$. The order of integration in the right side can be exchanged, and next the integral relative to $\mu^n$ can be factorized, where the integrals $\int p_\lambda \, d\mu$ are equal to 1. This yields the contribution $2 \sum_{|I| \geq 2} b^{|I|}$ to the bound on the expectation in (5.2).

The sum over sets with exactly one element contributes two times

$$\int \frac{\int \sum_{j=1}^n \prod_{i \neq j} p_\lambda(x_i) \ell_\lambda(x_j) \, d\pi(\lambda)^2}{\int \prod_i p_\lambda(x_i) \, d\pi(\lambda)} \, d\mu^n(\vec{x}_n). \tag{5.3}$$

Here we expand

$$\prod_{i\neq j} p_\lambda(x_i) - \prod_{i\neq j} p(x_i) = \prod_{i\neq j} p_\lambda(x_i) - \prod_{i\neq j}(p_\lambda - \kappa_\lambda)(x_i)$$
$$= -\sum_{|I|\geq 1, j\notin I} \prod_{i\in I^c} p_\lambda(x_i) \prod_{i\in I}(-\kappa_\lambda)(x_i),$$

where the sum is over all nonempty subsets $I \subset \{1,\ldots,n\}$ that do not contain $j$. Replacement of $\prod_{i\neq j} p_\lambda(x_i)$ by $\prod_{i\neq j} p(x_i)$ changes (5.3) into

$$\int \frac{\int \sum_{j=1}^n \prod_{i\neq j} p(x_i)\ell_\lambda(x_j)\,d\pi(\lambda)^2}{\int \prod_i p_\lambda(x_i)\,d\pi(\lambda)}\,d\mu^n(\vec{x}_n)$$
$$\leq n\sum_{j=1}^n \int \frac{\prod_{i\neq j} p^2(x_i)\int \ell_\lambda(x_j)\,d\pi(\lambda)^2}{\int \prod_i p_\lambda(x_i)\,d\pi(\lambda)}\,d\mu^n(\vec{x}_n)$$
$$\leq n\sum_{j=1}^n \int\int \prod_{i\neq j} \frac{p^2}{p_\mu}(x_i)\frac{\int \ell_\lambda\,d\pi(\lambda)^2}{p_\mu}(x_j)\,d\pi(\mu)\,d\mu^n(\vec{x}_n).$$

In the last step we use that $1/EV \leq E(1/V)$ for any positive random variable $V$. The integral with respect to $\mu^n$ in the right side can be factorized, and the expression bounded by $n^2c^{n-1}d$. Four times this must be added to the bound on the expectation in (5.2).

Finally the remainder after substituting $\prod_{i\neq j} p(x_i)$ for $\prod_{i\neq j} p_\lambda(x_i)$ in (5.3) contributes

$$\int \frac{\int \sum_{j=1}^n \sum_{|I|\geq 1, j\notin I} \prod_{i\in I^c} p_\lambda(x_i) \prod_{i\in I}(-\kappa_\lambda)(x_i)\ell_\lambda(x_j)\,d\pi(\lambda)^2}{\int \prod_i p_\lambda(x_i)\,d\pi(\lambda)}\,d\mu^n(\vec{x}_n)$$
$$\leq \int\int \left(\sum_{j=1}^n \sum_{|I|\geq 1, j\notin I} \prod_{i\in I^c} \sqrt{p_\lambda}(x_i) \prod_{i\in I} \frac{-\kappa_\lambda}{\sqrt{p_\lambda}}(x_i)\frac{\ell_\lambda}{\sqrt{p_\lambda}}(x_j)\right)^2 d\pi(\lambda)\,d\mu^n(\vec{x}_n)$$
$$\leq n\sum_{j=1}^n \int\int \left(\sum_{|I|\geq 1, j\notin I} \prod_{i\in I^c} \sqrt{p_\lambda}(x_i) \prod_{i\in I} \frac{-\kappa_\lambda}{\sqrt{p_\lambda}}(x_i)\right)^2 \frac{\ell_\lambda^2}{p_\lambda}(x_j)\,d\pi(\lambda)\,d\mu^n(\vec{x}_n)$$
$$= n\sum_{j=1}^n \sum_{|I|\geq 1, j\notin I} \int\int \prod_{i\in I^c} p_\lambda(x_i) \prod_{i\in I} \frac{\kappa_\lambda^2}{p_\lambda}(x_i)\frac{\ell_\lambda^2}{p_\lambda}(x_j)\,d\pi(\lambda)\,d\mu^n(\vec{x}_n).$$

In the last step we use that $\int \kappa_\lambda\,d\mu = 0$ to reduce the square sum to the sum over the squares of its terms. We exchange the order of integration and factorize the integral with respect to $\mu^n$ to bound the far right side by $n^2\sum_{|I|\geq 1, j\notin I} a^{|I|}b$. □

We are ready for the proof of Theorem 2.1.

The numbers $a$, $b$ and $d$ in Theorem 2.1 are the maxima over $j$ of the numbers $a$, $b$ and $d$ defined in Lemma 5.2, but with the measures $P_\lambda$ and $Q_\lambda$ replaced

there by the measures $P_{j,\lambda_j}$ and $Q_{j,\lambda_j}$ given in (5.1). Define a number $c$ similarly as

$$c = \max_j \sup_\lambda \int_{\mathcal{X}_j} \frac{p^2}{p_\lambda} \frac{d\mu}{p_j}.$$

Under the assumptions of the theorem $c$ is bounded above by $\overline{B}/\underline{B}$.

Define $(N_1, \ldots, N_k)$ and the functions $\rho_j$ as in the statement of Lemma 5.1. By Lemma 5.2

$$\begin{aligned}\prod_{j=1}^k \rho_j(N_j) &\geq \prod_{j=1}^k \left(1 - \sum_{r=2}^{N_j} \binom{N_j}{r} b^r - 2N_j^2 \sum_{r=1}^{N_j-1} \binom{N_j-1}{r} a^r b - 2N_j^2 c^{N_j-1} d\right) \\ &\geq 1 - \sum_{j=1}^k \left(\sum_{r=2}^{N_j} \binom{N_j}{r} b^r + 2N_j^2 \sum_{r=1}^{N_j-1} \binom{N_j-1}{r} a^r b + 2N_j^2 c^{N_j-1} d\right),\end{aligned}$$

provided every of the $k$ terms in the product in the middle is nonnegative, where in the second step we use that $\prod_{j=1}^k (1-a_j) \geq 1 - \sum_{j=1}^k a_j$ for any numbers $a_1, \ldots, a_k$ in $[0,1]$. If one or more of the terms are negative, then these inequalities may be false, but then the far right side is negative and hence still is a lower bound for the far left side. Hence in all cases

$$\prod_{j=1}^k \rho_j(N_j) \geq 1 - \sum_{j=1}^k \left(\sum_{r=2}^{N_j} \binom{N_j}{r} b^r + 2N_j^2 \sum_{r=1}^{N_j-1} \binom{N_j-1}{r} a^r b + 2N_j^2 c^{N_j-1} d\right).$$

By Lemma 5.1 the expectation of the left side is a lower bound on the left side of the theorem. The expected values on the binomial variables $N_j$ in the right side can be evaluated explicitly, using the identities, for $N$ a binomial variable with parameters $n$ and $p$,

$$\begin{aligned}\mathrm{E} \sum_{r=2}^N \binom{N}{r} b^r &= \mathrm{E}\big((1+b)^N - 1 - Nb\big) = (1+bp)^n - 1 - npb, \\ \mathrm{E} N^2 c^{N-1} &= np(cp + 1 - p)^{n-2}(cnp + 1 - p), \\ \mathrm{E} N^2 \sum_{r=1}^{N-1} \binom{N-1}{r} a^r &= \mathrm{E} N^2 \big((1+a)^{N-1} - 1\big) \\ &= np(1+ap)^{n-2}(1 + nap + np - p) - np(1-p) - n^2p^2.\end{aligned}$$

Under the assumption that $np(1 \vee a \vee b \vee c) \lesssim 1$, the right sides of these expressions can be seen (by Taylor expansions with remainder) to be bounded by multiples of $(npb)^2$, $np$ and $(np)^2 a$, respectively. We substitute these bounds into the expectation of the second last display, and use the equality $\sum_j p_j = 1$ to complete the proof.

**Remark 5.1.** *If* $\min p_j \sim \max_j p_j \sim 1/n^{1+\varepsilon}$ *for some* $\varepsilon > 0$, *which arises for equiprobable partitions in* $k \sim n^{1+\varepsilon}$ *sets, then there exists a number* $n_0$ *such that*

$\mathrm{P}(\max_j N_j > n_0) \to 0$. *(Indeed, the probability is bounded by* $k(n \max_j p_j)^{n_0+1}$*.) Under this slightly stronger assumption the computations need only address* $N_j \leq n_0$ *and hence can be simplified.*

## 6. Appendix: Minimax rates and testing hulls

Let $\mathcal{P}$ be a set of probability densities $p$ on a measurable space $(\mathcal{X}, \mathcal{A})$ with corresponding distributions $P$, and let $\chi: \mathcal{P} \to \mathbb{R}$ besome functional. Let $P^n$ be the $n$-fold product measure of $P$, and define

$$\mathcal{P}_{n,\leq 0} = \{P^n : p \in \mathcal{P}, \chi(p) \leq 0\}, \qquad \mathcal{P}_{n,\geq \varepsilon} = \{P^n : p \in \mathcal{P}, \chi(P) \geq \varepsilon\}.$$

The convex hulls of these sets are the sets of measures $\sum_{i=1}^k \lambda_i P_i^n$ for $k \in \mathbb{N}$, $\lambda_1, \ldots, \lambda_k \geq 0$ with $\sum_{i=1}^k \lambda_i = 1$, and $P_1^n, \ldots, P_k^n$ ranging over the set under consideration.

We are given $n$ i.i.d. observations $X_1, \ldots, X_n$ distributed according to one of the densities $p \in \mathcal{P}$. An *estimator* is a measurable function $T_n: (\mathcal{X}^n, \mathcal{A}^n) \to \mathbb{R}$, and a *test* $\phi_n$ is an estimator that takes its values in the interval $[0, 1]$.

**Proposition 6.1.** *Suppose that for some* $\varepsilon_n \to 0$ *there exist measures* $P_n$ *and* $Q_n$ *in the convex hulls of* $\mathcal{P}_{n,\leq 0}$ *and* $\mathcal{P}_{n,\geq \varepsilon_n}$*, respectively, such that*

$$\liminf_{n\to\infty} \bigl(P_n\phi_n + Q_n(1-\phi_n)\bigr) > 0,$$

*for any sequence of tests* $\phi_n$*. Then,*

$$\liminf_{n\to\infty} \frac{1}{\varepsilon_n^2} \inf_{T_n} \sup_{p\in\mathcal{P}} \mathrm{E}_p |T_n - \chi(p)|^2 > 0.$$

*Proof.* The assertion is equivalent to the statement that for every estimator sequence $T_n$ the liminf of $\varepsilon_n^{-2} \sup_{p\in\mathcal{P}} \mathrm{E}_p|T_n - \chi(p)|^2$ is positive. We shall in fact show that

$$\liminf_{n\to\infty} \sup_{p\in P} \mathrm{P}_p\bigl(|T_n - \chi(p)| > \varepsilon_n/2\bigr) > 0.$$

The assertion then follows, because $\varepsilon^{-2}\mathrm{E}Y^2 \geq \mathrm{P}\bigl(|Y| \geq \varepsilon\bigr)$, for any random variable $Y$ and every $\varepsilon > 0$.

To prove the assertion in the preceding display suppose that $T_n$ were a sequence of estimators for which the right side of the display is 0. We can then define tests by $\phi_n = 1_{T_n \geq \varepsilon_n/2}$.

If $\chi(p) \leq 0$, then $T_n \geq \varepsilon_n/2$ implies that $|T_n - \chi(p)| \geq \varepsilon_n/2$, and hence $P^n\phi_n \leq \mathrm{P}_p\bigl(|T_n - \chi(p)| \geq \varepsilon_n/2\bigr)$ for every $P^n \in \mathcal{P}_{n,\leq 0}$. It follows that $P_n\phi_n \leq \sup_p \mathrm{P}_p\bigl(|T_n - \chi(p)| \geq \varepsilon_n/2\bigr) \to 0$.

Similarly, if $\chi(p) \geq \varepsilon_n$, then $T_n \leq \varepsilon_n/2$ implies that $|T_n - \chi(p)| \geq \varepsilon_n/2$, and hence $P^n(1-\phi_n) \leq \mathrm{P}_p\bigl(|T_n - \chi(p)| \geq \varepsilon_n/2\bigr)$ for every $P^n \in \mathcal{P}_{n,\geq\varepsilon_n}$. It follows that $Q_n(1-\phi_n) \to 0$. We have arrived at a contradiction, because both error probabilities of $\phi_n$ tends to zero. □

## References


[1] Birgé, L. and Massart, P. (1995). Estimation of integral functionals of a density. *Ann. Statist.* **23**, 1, 11–29. MR1331653 (96c:62065)

[2] Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika 96*, 187–199.

[3] Ingster, Y. I. and Suslina, I. A. (2003). *Nonparametric goodness-of-fit testing under Gaussian models*. Lecture Notes in Statistics, Vol. **169**. Springer-Verlag, New York. MR1991446 (2005k:62003)

[4] Le Cam, L. (1986). *Asymptotic methods in statistical decision theory*. Springer Series in Statistics. Springer-Verlag, New York. MR856411 (88a:62004)

[5] LeCam, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist. 1*, 38–53. MR0334381 (48 #12700)

[6] Robins, J., Li, L., Tchetgen, E., and van der Vaart, A. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman*. Inst. Math. Stat. Collect., Vol. **2**. Inst. Math. Statist., Beachwood, OH, 335–421. MR2459958

[7] Robins, J., Li, L., Tchetgen, E., and van der Vaart, A. (2009). Quadratic semiparametric von mises calculus. *Metrika 69*, 227–247.

[8] Robins, J. M., Mark, S. D., and Newey, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics* **48**, 2, 479–495. MR1173493 (93e:62260)

[9] Tsybakov, A. B. (2004). *Introduction à l'estimation non-paramétrique*. Mathématiques & Applications (Berlin) [Mathematics & Applications], Vol. **41**. Springer-Verlag, Berlin. MR2013911 (2005a:62007)

[10] van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics, Vol. **3**. Cambridge University Press, Cambridge. MR1652247 (2000c:62003)

[11] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics. MR1385671 (97g:60035)