

hexagon:2610.00157v1

Nearly Minimax Rates for Functional Estimation Under Rough Random Design

P. M. Aronow

Nathan Kallus

Patrick Lopatto

AUTHORS' NOTE

This paper was written by the authors with the assistance of large language models, which were used to suggest mathematical arguments, draft and revise exposition, and write code for computational verification. We have taken care to supervise this writing to ensure that the proofs meet a minimum standard of readability and that previous literature is appropriately cited. However, the resulting document remains lacking in exposition and overall coherence. We are nonetheless releasing this manuscript on Hexagon to disseminate the result as quickly as possible, because we believe it may be of interest to the statistics community. We welcome suggestions concerning mathematical corrections, omitted citations, and attribution of ideas.

P.L. was partially supported by NSF grant DMS-2450004.

ABSTRACT

We establish nearly minimax bounds for missing-at-random means, treatment effects, and expected conditional covariances under rough random design. For two nuisance functions with average Hölder smoothness s in dimension d , the minimax root-mean-square error is $n^{-2s/d+o(1)}$ when $s < d/4$ and of order $n^{-1/2}$ when $s \geq d/4$. The first rate confirms the rough-design exponent suggested by higher-order influence function theory. In the generic model with an unknown bounded density weight, our upper and lower bounds differ by only polylogarithmic factors and identify a leading correction $e^{-\kappa\sqrt{\log n}}$ when $s < d/4$, with κ explicit in terms of s/d and the weight bounds. For the expected conditional covariance, the same exponent and the same constant κ were obtained independently and concurrently by S. Park (arXiv:2610.05006). For models with separately bounded density and propensity, we identify the same polynomial exponent and the explicit leading correction, with an $o(\sqrt{\log n})$ remainder in the logarithm of the risk.

Contents

1	Introduction	1
1.1	Background	1
1.2	Main results	2
1.3	Ideas of the proofs	3
1.4	Related work	5
1.5	Organization	10
2	Setting and main result	11
2.1	Functionals and examples	11
2.2	Model and nondegeneracy	12
2.3	Main theorem	14
3	Applications	16
3.1	Missing-data means and treatment effects	18
3.2	Conditional covariances and variances	19
3.3	Wald ratios and the overlap-weighted effect	20
4	Upper bounds	22
4.1	The reciprocal series	22
4.2	Projection increments and their polynomial approximations	26
4.3	Unbiased estimation and variance.	30
4.4	Choice of the degrees.	32
5	Lower bounds	37
5.1	Divergences and a two-point bound.	38
5.2	Phase priors and the Poisson comparison	38
5.3	Reduction to two estimates	43
6	Lattice priors	47
6.1	Lattice laws with compact Fourier support	47
6.2	Construction of the priors.	48
6.3	Proof of Lemma 6.2.	52
7	Embedding and proof of the lower bounds	58
7.1	Embedding the priors in the generic model	58
7.2	Localization	60

8	Conclusion	63
A	Reductions	64
A.1	Hard laws	64
A.2	Three reductions	65
B	Missing-data means and treatment effects	68
B.1	The MAR family and two maps.	68
B.2	A fixed-accuracy pilot	69
B.3	Proof of Corollary 3.2 for these targets	71
C	Conditional covariances and variances	73
D	Wald ratios and the overlap-weighted effect	75
D.1	Wald ratios	75
D.2	The overlap-weighted effect.	76
	References	79

Introduction

1.1 Background

Missing-data means, average treatment effects, and expected conditional covariances under random design are averages over the covariate distribution of functions of regressions given the covariates. Their estimation rates depend on what is known about both the regressions and the covariate distribution. In missing-response problems, Robins and Ritov [50] showed that attainable rates depend on what is known about the nuisance functions. We study *rough design*, in which the covariate density is unknown and subject only to fixed upper and lower bounds, while the regressions are assumed smooth. We show that below the root- n threshold, rough design reduces the minimax polynomial exponent from $4s/(4s+d)$ for a known or sufficiently smooth density to $2s/d$, where s is the average smoothness of the two nuisance functions and d is the covariate dimension. The threshold $s = d/4$ itself is unchanged.

For a missing-data mean, let α and β be the Hölder smoothness indices of the inverse observation probability and of the outcome regression. For the average product of two conditional means, which underlies the expected conditional covariance, let them be the indices of the two means. Let $s = (\alpha + \beta)/2$. Estimators of these parameters impose assumptions on a density weight g , which is the covariate density or its product with the observation probability, depending on the parameter. When g is known or belongs to a Hölder class of smoothness $\gamma > 0$ such that $\gamma/(2\gamma + d) > \{\max(\alpha, \beta)/d\}(d - 4s)/(d + 4s)$, higher-order influence function (HOIF) estimators attain root-mean-square error of order

$$n^{-1/2} \vee n^{-4s/(4s+d)}, \quad (1.1)$$

where $x \vee y = \max(x, y)$. This condition and the upper bound are from [52, Section 3.1], the corrected version of [53], which treats the missing-data mean there and indicates the extension to other doubly robust functionals in [52, Section 11.3]. Below the threshold, the rate (1.1) cannot be improved even when g is known, as shown in [49]; see [52, Section 3.1]. Rates of this form are classical for quadratic functionals of a density [5, 6]. For the average product of two conditional means, McGrath and Mukherjee [44, Corollaries 1 and 16] obtain this rate for a uniform or sufficiently smooth known density, and below $s = d/4$ their results also cover an unknown density that is sufficiently smooth.

The rough-design question arises in the HOIF literature [36, 48, 49, 52]. Immediately after their equation (1.3), Robins et al. [48, p. 338] conjecture that the rate given there is minimax, up to logarithmic factors, when the nuisance functions are equally smooth and the density smoothness γ is too small for their condition (1.2), which for equal indices is the condition stated before (1.1). They note that, as $\gamma \rightarrow 0$, this rate tends to $n^{-2s/d} \log n$, which agrees up to the logarithmic factor with the minimax rate $n^{-2s/d}$ for a constant conditional variance under an equally spaced fixed design [9, 71]. This limit suggests that the exponent under rough design is $2s/d$.

Robins et al. [52, Section 3.1] also state that heuristic arguments indicate a rate slower than (1.1) for smaller γ . Kennedy, Balakrishnan, and Wasserman [26, Section 4] describe the minimax problem with an unknown density as open. They suggest that the rate with an unknown covariate density might interpolate between the known-density rate and the fixed-design rate $n^{-2s/d}$, which would suggest agreement between rough random design and fixed design at the nonsmooth extreme. For a counterfactual mean, Liu, Mukherjee, and Robins [38, Section 4] describe an empirical HOIF estimator, based on [37], that attains $n^{-2s/d}$ up to logarithmic factors without complexity assumptions on the covariate density. McClean et al. [43, Section 4.2 and Theorem 1] construct an estimator of the expected conditional covariance with a convergence rate in probability of $n^{-2s/d}$, up to a polylogarithmic factor, when $s < d/4$, and they leave open whether this rate is minimax under rough design.

1.2 Main results

Our main result, Theorem 2.3, gives minimax bounds in a nonparametric model in which the weight g is unknown and satisfies only fixed bounds $0 < g_- \leq g \leq g_+ < \infty$ with $g_- < g_+$. It covers a class of doubly robust functionals [48, Section 3.1] that includes missing-data means and expected conditional covariances (Section 2.1). Under interior and nondegeneracy conditions stated in Section 2.2, the minimax root-mean-square error r_n satisfies

$$\log r_n = \begin{cases} -\frac{2s}{d} \log n - \frac{2}{d} \sqrt{2s(d-4s) \log n \log \frac{\sqrt{g_+} + \sqrt{g_-}}{\sqrt{g_+} - \sqrt{g_-}}} + O(\log \log n), & s < d/4, \\ -\frac{1}{2} \log n + O(1), & s \geq d/4. \end{cases} \quad (1.2)$$

Below $s = d/4$, (1.2) identifies the leading correction to the polynomial rate, a factor $e^{-\kappa \sqrt{\log n}}$ with the explicit constant κ of (2.9). In this regime the upper and lower bounds differ by a polylogarithmic factor, which is $(\log n)^{1/4}$ when both smoothness indices are at most one (Remark 2.4). The lower bounds hold even on the subclass in which the functions w, a, b of (2.2) stay within an arbitrary fixed distance of constant baseline values

(Section 2.2). In particular, they apply directly to models with additional overlap or range restrictions. They also persist if the density must be C^∞ , as long as the model imposes no fixed bound on its derivatives (Remark 2.5). Section 3 applies the theorem to the targets in Table 3.1.

By Theorem 2.3, the exponent $2s/d$ is minimax below the threshold, and our estimator attains it in root-mean-square error, uniformly over fixed Hölder classes, for every functional in the class of Theorem 2.3, whether or not the two smoothness indices are equal. For equal indices, $2s/d$ is the exponent suggested by the limit $\gamma \rightarrow 0$ of the rate that Robins et al. conjecture to be minimax. By (1.2), however, the minimax risk is smaller than $n^{-2s/d}$ by a factor $e^{-\kappa\sqrt{\log n} + O(\log \log n)}$, which tends to zero faster than every power of $1/\log n$. The limit $n^{-2s/d} \log n$ of their rate exceeds the minimax risk under rough design by more than any polylogarithmic factor. For the expected conditional covariance, Park [46] has obtained closely related bounds independently and concurrently, with the same threshold, polynomial exponent, and constant κ ; we compare the two works in Section 1.4.

1.3 Ideas of the proofs

We pair moment-matching priors with polynomial approximation, as in minimax estimation of nonsmooth functionals [8, 23, 33, 73, 74]. For an interval $I = [x_-, x_+]$ with $0 < x_- < x_+$, let $c_I = (x_+ + x_-)/2$ and $d_I = (x_+ - x_-)/2$. Then, for real t ,

$$\frac{1}{c_I + d_I \cos t} = \frac{1}{\sqrt{x_- x_+}} \left\{ 1 + 2 \sum_{k \geq 1} (-\rho_I)^k \cos(kt) \right\}, \quad \rho_I = \frac{\sqrt{x_+} - \sqrt{x_-}}{\sqrt{x_+} + \sqrt{x_-}} \quad (1.3)$$

(Lemma 4.1). Since $\cos(kt)$ is a polynomial of degree k in $c_I + d_I \cos t$, the partial sums of this series are polynomials that approximate the reciprocal on I , with error $O(\rho_I^m)$ at order m . For $I = [g_-, g_+]$, ρ_I is the ρ of (2.7).

Upper bound. Recall that $\psi(P) = \mathbb{E}_P W + \lambda \int abg$, with a, b, g as in (2.2). We project a and b in $L^2(g)$ onto piecewise polynomials on nested dyadic partitions. With K cells, the projection error of $\int abg$ is the inner product of the two projection errors, as for the truncation bias of HOIF estimators [48, Theorem 3.8], and it is of order $K^{-\theta}$, where $\theta = 2s/d$. By reciprocal approximation, we replace each projection increment by a polynomial in observable cell moments, with error $O(K^{-\theta}(m+1)^\nu \rho^m)$ at truncation order m , where ν is defined in (2.8) (Lemma 4.3). We estimate these polynomials from the n observations without bias by their lift, which replaces each monomial by a U-statistic [17, 19]; Lemma 4.4 expresses the covariance of the lifts exactly through centered derivatives. We then choose the degrees across resolutions by the variance bounds of Lemma 4.5, without an estimate of g . Polynomial approximations of bilinear forms in

the inverse of a Gram matrix, with U-statistic estimates of their terms, were used before in [10, 28] (Section 1.4).

Lower bound. We compare two priors by the method of two fuzzy hypotheses, and we follow the strategy of the lower bound of Aronow and Lopatto [3] for a constant conditional variance, with a random covariate density, a Poissonized sample, and a matching order M of order $\sqrt{\log n}$ (Section 1.4; see also Section 5). Here the matching comes from a random phase. On the bump supports of the two blocks of a pair, the density equals $r_0 \pm h_n \cos \Theta$, with a uniform phase Θ . The conditional law of the response is perturbed by u, v and is affine in (u, v) ; the perturbations are bumps of heights A_u, A_v divided by the density, with signs correlated through $\pm \cos(M\Theta)$ under the two priors, for an even integer M . In the generic model, they shift a and b by u/w and v/w , or both by $(u+v)/w$ when the response functions U and V of Section 2.1 coincide (Section 7.1). Each likelihood coefficient that differs between the priors contains an odd number of factors of each perturbation. The signs contribute $\cos(M\Theta)$, and apart from their signs the products pu and p_v do not depend on Θ . Conditional on the other phase parameters, j additional density factors form a trigonometric polynomial of degree at most j in Θ . We conclude that every such coefficient vanishes for $j < M$, including those with higher powers of the perturbations (Lemma 5.2). The target means separate by a multiple of $A_u A_v$ times the M th Fourier coefficient of the reciprocal density, which is of order ρ^M by (1.3). With block-constant phases, the case $J = 0$ of Lemma 6.2, we obtain $r_n \geq c a_n$ (Proposition 5.3), where a_n is defined in (2.9). In the lattice refinement, we let the phase vary across the R fine cells of each block. For $M \leq j < M + \sqrt{M}$, a coefficient then survives only where the binary digits of most of its points nearly agree, which gives the factor R^{1-M} ; a cruder bound suffices for larger j . We use a random gate to keep the Hölder norms bounded while the bump heights gain a factor $m \asymp M$ (Lemma 6.2), and the lower bound gains the factor $m^2 \asymp \log n$.

The common balance. Let $0 < \theta < 1/2$, $L = \log n$, and $\Delta = 1 - 2\theta$, and let τ, κ and ν be as in (2.7)–(2.9). In both bounds, let x be the logarithm of the number of cells per observation, up to a shift of order $\log L$. The degree in the upper bound and the frequency M in the lower bound are then about $\Delta L/x$. This keeps the variance at $n^{-2\theta}$, up to factors $e^{O(\sqrt{L})}$, in the upper bound, and it makes the coefficients with $j \geq M$ negligible in the lower bound. Relative to $n^{-\theta}$, the cells cost $e^{-\theta x}$ and the reciprocal approximation costs $\rho^{\Delta L/x}$. With $x_* = \sqrt{\tau \Delta L / \theta}$, the leading exponential cost satisfies

$$\theta x + \frac{\tau \Delta L}{x} = \kappa \sqrt{L} + \frac{\theta(x - x_*)^2}{x}.$$

In the lower bound, M is the nearest even integer to $M_* = \Delta L/x_*$. Since $\theta \Delta L/M + \tau M - \kappa \sqrt{L} = \tau(M - M_*)^2/M$, the rounding costs a factor $e^{O(1/M)}$. A balance between a loss exponential in the matching order and a loss exponential in $\log n$ divided by that order also underlies the lower bound of [3]. In the upper bound, the positive shifted values of

x on the capped resolutions are $O(\sqrt{L})$ and pairwise separated by at least $\log 2$, so only $O(L^{1/4})$ resolutions contribute comparable errors (Lemma 4.6). In both bounds, the shift of x contributes $L^{\theta/2}$. The lattice gain contributes a factor of order L to the lower bound, and the factor $(m+1)^\nu$, with truncation order $m = O(\sqrt{L})$, contributes $L^{\nu/2}$ to the upper bound. We obtain $a_n L$ and $a_n L^{\nu/2+1/4}$, which leaves the factor $L^{1/4}$ when $\nu = 2$.

1.4 Related work

Higher-order influence functions and rough design. The HOIF estimators of Robins et al. [48] are built from projection kernels, which they write through inverse Gram matrices [48, p. 357], and the truncation bias of the projected functional is the integral of the product of two projection residuals [48, Theorems 3.7–3.8]. When g is unknown, these estimators start from an initial estimate of g and add U-statistic terms of increasing order, which form a Neumann-type expansion of the inverse of a Gram matrix about its value under that estimate [48, Theorems 3.14–3.15]. Robins et al. [48, p. 381] conjecture that, for $s < d/4$, such estimators attain the rate $n^{-2s/d}$ in probability, up to a power of $\log n$. The conjecture assumes an initial estimate of g that is consistent in supremum norm at a rate $1/r(n)$, an order at least a multiple of $\log n / \log r(n)$, and a number of basis functions chosen according to the order. In their Remark 4.1, which concerns the mean of the square of a regression function, they observe that the empirical Gram matrix of more than n basis functions is singular. They give this as the reason why the minimax rate is slower than $n^{-4s/(4s+d)}$ without restrictions on the covariate density [48, Remark 4.1]. In their Remark 4.2, they refer to $n^{-2s/d}$, for $s < d/4$, as their conjectured minimax rate for the variance-weighted average treatment effect when the covariate density is not assumed smooth and the treatment effect is not assumed constant [48, Remark 4.2]. This effect is the overlap-weighted effect of Table 3.1, and we compare our result for it with their conjecture in Appendix D.2.

Our estimator has a form similar to the estimators in the conjecture on p. 381 of [48]: it is a sum of polynomials in cell moments, estimated without bias by U-statistics, with orders up to a multiple of $\log n$ at coarse resolutions. It differs from them in two respects. It approximates the inverses of the local Gram matrices of g by their Chebyshev series about the known interval $[g_-, g_+]$, and not about an initial estimate (Section 4.1). Further, its order decreases as the resolution increases. Keeping high orders only at coarse resolutions resembles the truncation of the higher-order kernels in [52, Section 9, (9.6)] (Section 4.4). Since our estimator has root-mean-square error $o(n^{-2s/d})$ uniformly over the class, it attains the rate of that conjecture, although it is not one of the estimators to which the conjecture refers. The empirical HOIF estimators of [37] avoid estimating g by using the inverse of a sample Gram matrix of $k = o(n)$ basis functions.

Polynomial approximation of a bilinear form in the inverse of a Gram matrix,

with each term estimated by a U-statistic, was used by Kong and Valiant [28, Section 2.2]. In high-dimensional linear regression with unknown covariance Σ , they estimate $\mathbb{E}[YX^\top]\Sigma^{-1}\mathbb{E}[XY]$ through unbiased estimates of $\mathbb{E}[YX^\top]\Sigma/\mathbb{E}[XY]$. Chen, Liu, and Mukherjee [10, Section 3.2] use Chebyshev polynomials in Σ for this purpose when the dimension exceeds the sample size. We apply this strategy to the local Gram matrices of g in a nonparametric model, approximate inverses and reciprocals of determinants jointly by a series about $[g_-, g_+]$, and choose the degrees across resolutions to obtain the rates of Theorem 2.3.

Estimators that do not use density smoothness. In dimension one, the difference-based estimator of Robins et al. [48, p. 380], which follows Wang et al. [71], has bias $O(n^{-2s})$ and variance $O(n^{-1})$ for the expected conditional covariance. Their bias bound combines the Hölder continuity of the two regressions with the theory of spacings and uses no smoothness of the density; it applies when both smoothness indices are at most one. With these bounds, the estimator attains the root-mean-square rate $n^{-1/2} \vee n^{-2s}$, that is, the exponent $\min\{2s, 1/2\}$ of Theorem 2.3 when $d = 1$ and both indices are at most one, with root- n risk at $s = 1/4$. As mentioned above, McClean et al. [43, Section 4.2 and Theorem 1] construct an estimator of the expected conditional covariance whose rate in probability is $n^{-2s/d}$, up to a polylogarithmic factor, when $s < d/4$. Their estimator applies in every dimension and requires only bounds on the covariate density and no knowledge of the nuisance smoothness. The lower bound of Corollary 3.2 on $\mathcal{P}_{\text{bil}}(\alpha, \beta)$, which also holds in probability, shows that the polynomial exponent of this rate is minimax (Section 3.2). Unlike the estimator of McClean et al., ours uses α, β , the bounds g_-, g_+ , and the remaining class parameters; it attains the exponent in root-mean-square error and the correction in (1.2) up to a polylogarithmic factor.

Above the threshold, root- n estimation without density smoothness was already known for some of these functionals. Double cross-fitting [45] and empirical HOIF estimators [37] attain semiparametric efficiency under their respective approximation and nuisance conditions, and the bounds of [48, p. 380] recalled above give root- n risk at $s = d/4$ in dimension one. McClean et al. [43, Section 4.2 and Theorem 1] obtain root- n consistency of their estimator of the expected conditional covariance when $s > d/4$, with only bounds on the covariate density. Our estimator has root- n risk for $s \geq d/4$ in every dimension, for all smoothness indices, and for every functional in the class of Theorem 2.3. In his setting, Park [46, Theorem 1] also obtains root- n risk for $s \geq d/4$, including $s = d/4$. Stabilized higher-order estimators of bilinear functionals without sample splitting were recently developed in [41], and higher-order estimators of treatment parameters defined implicitly by estimating equations in [76].

Structure-agnostic models. In the structure-agnostic models of Balakrishnan et al. [4], in which only convergence rates are imposed on black-box nuisance estimators, first-order estimators are minimax optimal for the expected conditional covariance and, when the

squared pilot-error radius is at least of order n^{-1} , for the two quadratic functionals studied there. Jin and Syrgkanis [24] prove the corresponding optimality for the average treatment effect, weighted average treatment effects, and the average treatment effect on the treated. Their subsequent generalization treats linear functionals of generalized regressions and includes a structure-agnostic lower bound for the expected conditional covariance even under a partially linear outcome restriction [25, Theorems 6.1–6.2 and 7.5]. Bonvini et al. [7] study classes that add smoothness constraints to structure-agnostic ones. The Hölder smoothness that we assume is what allows higher-order estimators to improve on first-order ones in the minimax rate. Minimax-rate optimality does not imply admissibility: in the structure-agnostic model, Liu, Mukherjee, and Robins [40, Theorems 0–2] show, under their other stated conditions and when the squared pilot-error radius is at least a constant multiple of $n^{-1/2}$, that second-order estimators asymptotically dominate first-order ones for two quadratic functionals. For the expected conditional covariance, neither estimator asymptotically dominates the other.

Moment matching and polynomial approximation. Recall that, as in minimax estimation of nonsmooth functionals [8, 23, 33, 73, 74], we pair moment-matching priors with polynomial approximation, with the reciprocal in place of the nonsmooth function. To estimate the support size of a discrete distribution, Wu and Yang [74] use best polynomial approximation of the reciprocal on an interval bounded away from zero. Our priors use the framework of the lower bounds of Robins et al. [49, Theorem 2.1 and Sections 3–4] for these functionals: independent, randomly signed smooth bumps on the cells of a partition, with cell masses that do not depend on the prior parameters, and finite response spaces. Our lower bound follows the strategy of the lower bound of Aronow and Lopatto [3] for a constant conditional variance under rough design. There, the covariate density is random under both priors, approximate moment matching makes the local laws of the data under the two priors close simultaneously for every number $k \leq M$ of observations in a local region, the sample is Poissonized, and $M \asymp \sqrt{\log n}$ gives a factor $e^{-C\sqrt{\log n}}$ in the lower bound, for some constant C . Relative to [3], what is new here is the random phase of the density, under which the two priors differ only at frequency M , and the separation through the M th Fourier coefficient of the reciprocal density. This identifies the constant κ in (1.2), which the lower bound of [46] also identifies for the expected conditional covariance, and the lattice phases of Section 6 gain a further factor $\log n$. We compare the two constructions in more detail at the start of Section 5.

Variance estimation. In the semiparametric model $\text{Var}(Y | X) = \sigma^2$, the variance can be estimated faster under a random design without density smoothness than under an equally spaced fixed design [9, 71] when $0 < s \leq 1$ and $s < d/4$. This was sketched in [48, Section 4] for $s < \min\{1, d/4\}$ and proved in [3, 59]. For $s > 1$ and $d > 4s$, Aronow and Lopatto [3] identify the minimax exponent $2(s+1)/(d+4)$, and Dobriban et al. [14] give a two-scale estimator that attains it and whose analysis uses only upper and lower

bounds on the design density. We discuss these results in Section 3.2; they do not apply to the nonparametric models of this paper.

Discrete covariates. In a discrete analogue, Zeng et al. [75] give nearly matching minimax bounds for the average treatment effect with high-dimensional discrete covariates. Their lower bound [75, Section 3.2] uses priors on the covariate probabilities, propensities, and regressions of the categories that match moments up to order $\log n$, through the duality between moment matching and best polynomial approximation. They also show that the expected conditional covariance and a variance-weighted average treatment effect can be estimated faster when the covariate distribution is known, but that this knowledge may not help for the average treatment effect [75, Section 5].

Concurrent work. Park [46] studies the expected conditional covariance $\mathbb{E}[\text{Cov}(A, Y \mid X)]$ for responses bounded by one in absolute value, covariates in $[0, 1]^d$ with a density between known bounds $0 < g_- < 1 < g_+$, and conditional means $\mathbb{E}[A \mid X]$ and $\mathbb{E}[Y \mid X]$ in Hölder classes of orders α and β . With $U = A$ and $V = Y$, these are the response and density conditions of $\mathcal{P}_{\text{bil}}(\alpha, \beta)$ in Section 3.2, but the Hölder norms differ: Park requires extensions to a fixed open neighborhood of $[0, 1]^d$ and sums the supremum norms of the derivatives, whereas we take their maximum on the closed cube. The present paper, in an unpublished version dated 2 October 2026, and [46], submitted to arXiv on 4 October 2026, were developed independently and concurrently. Park states his bounds for the mean squared error. His exponent $(\alpha + \beta)/d$ and his constant $\log\{(\sqrt{g_+} + \sqrt{g_-})/(\sqrt{g_+} - \sqrt{g_-})\}$ are our θ and τ , and his dimension of the polynomials of degree at most $\lceil t \rceil - 1$ in d variables is our q_t . In this notation, for $0 < \theta < 1/2$, Park [46, Theorem 1] bounds the minimax mean squared error below and above by constant multiples of

$$n^{-2\theta}(\log n)^\theta e^{-4\sqrt{\tau\theta(1-2\theta)}\log n} \quad \text{and} \quad n^{-2\theta}(\log n)^{\theta+q_\alpha+q_\beta+1/2} e^{-4\sqrt{\tau\theta(1-2\theta)}\log n},$$

and he obtains the order n^{-1} when $\theta \geq 1/2$, including $\theta = 1/2$. Park presents these bounds as a resolution of the conjecture in [48, Remark 4.2]. That remark concerns the variance-weighted average treatment effect, a ratio whose numerator is the expected conditional covariance [48, Example 1c]; the conjecture after their equation (1.3), which we recall in Section 1.1, is a different passage. Since $2\sqrt{\tau\theta(1-2\theta)}$ is the constant κ of (2.9), the threshold, the polynomial exponent, and the constant in the stretched-exponential factor agree with those of Theorem 2.3. In root-mean-square error, Park's upper bound is $C a_n (\log n)^{v/2+1/4}$, which is our upper bound, and his lower bound is $c a_n$, which is the bound that our priors with block-constant phases give (Remark 2.4). Our lower bound $c a_n \log n$ is larger by the factor $\log n$ that the lattice refinement of Section 6 gains, and we deduce that it also holds on Park's class. Since his class is contained in $\mathcal{P}_{\text{bil}}(\alpha, \beta)$ for a suitable radius, our upper bound holds on it as well. The hard laws that we use for this target (Appendix C) have Rademacher responses with conditional means u and v ,

and C^∞ densities with values in $[g_-, g_+]$. The functions u, v are C^∞ and vanish outside a fixed cube in the interior of $[0, 1]^d$. Their extensions by zero satisfy Park's Hölder conditions once the constant ϵ of Proposition 7.1 is small enough, so these laws lie in his class, and Proposition 7.1 gives the lower bound $c a_n \log n$ there. The unresolved factor in root-mean-square error is $(\log n)^{v/2-3/4}$ here and $(\log n)^{v/2+1/4}$ in [46], and neither work determines the power of $\log n$ in the minimax rate.

The two proofs share their main ingredients. Both use polynomial approximation of the reciprocal on $[g_-, g_+]$, for the bias of the estimator and for the separation in the lower bound. Both upper bounds write a projection error as a telescoping sum of increments over nested partitions, remove the local Taylor polynomials of the two regressions from each increment, approximate inverses of local Gram matrices by matrix polynomials, estimate the resulting polynomials without bias by products over distinct observations, and lower the degree as the partition is refined. Park refines cubes into 2^d children and subtracts his corrections from a double cross-fit local polynomial estimator; the result is clipped to $[-1, 1]$. We bisect one coordinate at a time and approximate $\int abg$ directly by polynomials in cell moments, without regression fits. Both lower bounds randomize the density while keeping the masses of fixed regions constant, which are cubes in [46] and pairs of blocks here, and both choose the priors so that regions holding few observations do not distinguish them. The main text of [46] does not describe the lower-bound construction in more detail.

Park also shows that a double cross-fit local polynomial estimator whose degree depends only on d attains the exponent without knowledge of α and β [46, Corollary 1]. Further, his paper states the lower bound also for conditional covariances that are smooth and bounded away from zero and, when $\alpha = \beta$, for the average conditional variance of a binary response. It treats the normalized covariance $\mathbb{E}[\text{Cov}(A, Y \mid X)] / \mathbb{E}[\text{Var}(Y \mid X)]$, and it extends the bounds to responses that satisfy moment conditions in place of boundedness. By a data-driven selection, Park attains the known-density rate at each fixed density that is piecewise constant on a dyadic partition. He also reports an estimator that attains his upper bound up to a larger power of $\log n$ when (α, β) is unknown but lies in a known compact subset of $\{\alpha + \beta < d/2\}$. The paper also contains simulations and an analysis of diamond price data. The present paper treats the general class of Theorem 2.3 and the applications in Table 3.1, of which the expected conditional variance and the overlap-weighted effect have counterparts in [46], refines the lower bound by the factor $\log n$, and gives lower bounds on the local classes (2.6) and for smooth densities (Remark 2.5). The proofs of [46] are in a supplement that is not part of the arXiv posting, and we could not locate the supplement online as of 7 October 2026. Our comparison is based on the main text of arXiv:2610.05006v1.

1.5 Organization

Sections 2 and 3 state the model, Theorem 2.3, and its applications. Section 4 proves the upper bounds. Section 5 gives the two-point lemma for the root- n lower bound and reduces the bound below the threshold to two estimates on priors constructed in Section 6 and embedded in the model in Section 7.1. Section 7.2 completes the proof. Section 8 discusses open problems, and the appendices prove the applications.

Setting and main result

2.1 Functionals and examples

We observe n independent copies of $O = (X, Z)$ with unknown law P . Here $X \in [0, 1]^d$ is a covariate with unknown Lebesgue density p and Z is a response in a fixed measurable space. We let $D = D(Z)$, $U = U(Z)$, $V = V(Z)$, and $W = W(Z)$ be known bounded functions such that $0 \leq D \leq M_0$ and $|U|, |V| \leq M_0$. We fix a known $\lambda \neq 0$ and assume that $\mathbb{E}_P[D \mid X]$ is bounded away from zero. The parameter of interest is

$$\psi(P) = \mathbb{E}_P W + \lambda \mathbb{E}_P \left[\frac{\mathbb{E}_P[U \mid X] \mathbb{E}_P[V \mid X]}{\mathbb{E}_P[D \mid X]} \right]. \quad (2.1)$$

Examples 2.1 and 2.2 give the choices of (D, U, V, W, λ) for missing-data means, average treatment effects, and conditional covariances and variances. We write

$$w = \mathbb{E}_P[D \mid X], \quad a = \frac{\mathbb{E}_P[U \mid X]}{w}, \quad b = \frac{\mathbb{E}_P[V \mid X]}{w}, \quad g = wp. \quad (2.2)$$

Then $\psi(P) = \mathbb{E}_P W + \lambda \int_{[0,1]^d} abg$. We impose Hölder restrictions on a and b and upper and lower bounds on g .

The functional ψ belongs to the doubly robust class of Robins et al. [48, Section 3.1] and to the mixed-bias class of Rotnitzky, Smucler, and Robins [55]. For bounded candidate functions $\tilde{a}, \tilde{b}$, the moment $\mathbb{E}_P[W + \lambda\{U\tilde{b}(X) + V\tilde{a}(X) - D\tilde{a}(X)\tilde{b}(X)\}]$ differs from $\psi(P)$ by $-\lambda \int (\tilde{a} - a)(\tilde{b} - b)g$. This bias vanishes when either candidate function is correct and otherwise depends on the product of their errors. For missing-data means, the augmented inverse-probability-weighted estimating function is a special case of the construction of Robins, Rotnitzky, and Zhao [51, equation (9), p. 849, and Section 8]. The score of the average treatment effect in double/debiased machine learning has the same product-bias property [12, Section 9.4, equations (9.4.4)–(9.4.5)]. The upper bound uses the same product structure for projection errors.

Example 2.1 (Missing-data means and average treatment effects). We consider a full-data distribution $\bar{P}$ on (X, R, Y) such that R is binary and $Y \in [0, 1]$. Let P be the law of the coarsened observation $O = (X, (R, RY))$. If we take $(D, U, V) = (R, 1, RY)$, $W = 0$, and $\lambda = 1$, then we have

$$w = P(R = 1 \mid X), \quad a = 1/w, \quad b = \mathbb{E}_P[Y \mid R = 1, X], \quad g = wp, \quad \psi(P) = \mathbb{E}_P[b(X)].$$

Under the missing-at-random (MAR) condition $R \perp Y \mid X$ [56], this equals the full-data mean $\mathbb{E}_P Y$. For treatment observations (X, A, Y) such that A is binary, we write $b_j = \mathbb{E}_P[Y \mid A = j, X]$ for $j = 0, 1$. If we take $R = \mathbf{1}_{\{A=j\}}$, then we have $\psi(P) = \mathbb{E}_P[b_j(X)]$. Under consistency, conditional exchangeability, and positivity [18], the difference $\mathbb{E}_P[b_1(X) - b_0(X)]$ is the average treatment effect (ATE).

Example 2.2 (Conditional covariance and variance). We take $D = 1$ with $|U|, |V| \leq 1$, and we write $m_U = \mathbb{E}_P[U \mid X]$ and $m_V = \mathbb{E}_P[V \mid X]$. Then $w = 1$ and $g = p$. If we take $W = UV$ and $\lambda = -1$, then we have

$$\psi(P) = \mathbb{E}_P[UV] - \mathbb{E}_P[m_U(X)m_V(X)] = \mathbb{E}_P[\text{Cov}_P(U, V \mid X)].$$

If we take $U = V = Y$ and $f = \mathbb{E}_P[Y \mid X]$, then we have

$$\psi(P) = \mathbb{E}_P[Y^2] - \mathbb{E}_P[f(X)^2] = \mathbb{E}_P[\text{Var}_P(Y \mid X)].$$

This is the smallest mean squared error achievable by predicting Y from X . Taking $W = 0$ and $\lambda = 1$ instead gives $\psi(P) = \mathbb{E}_P[f(X)^2]$. Subtracting $(\mathbb{E}_P Y)^2$, we obtain the explained variance $\text{Var}_P(f(X))$. When $\text{Var}_P(Y) > 0$, the ratio $\text{Var}_P(f(X)) / \text{Var}_P(Y)$ is the population R^2 . In a randomized trial with $P(A = 1 \mid X) = 1/2$ and outcome $Y_{\text{obs}} \in [0, 1]$, we take $Y = (2A - 1)(2Y_{\text{obs}} - 1)$. Then f is the conditional average treatment effect (CATE), so that $\text{Var}_P(f(X))$ measures the heterogeneity of treatment effects [34].

The response in the trial of Example 2.2 is obtained by centering the outcome at $1/2$ in the modified-outcome construction discussed by Tian et al. [63, Section 2.1], who attribute it to Signorovitch's thesis. Its conditional mean under balanced randomization is the CATE.

Functionals of the form (2.1) also include some instrumental-variable parameters, and we obtain further treatment-effect parameters from them through ratios and differences (Corollary 3.2).

2.2 Model and nondegeneracy

For $t > 0$, we write $k = \lceil t \rceil - 1$ and $\vartheta = t - k \in (0, 1]$. For any $f : [0, 1]^d \rightarrow \mathbb{R}$ with continuous partial derivatives of order at most k , we define the Hölder norm

$$\|f\|_{C^t([0,1]^d)} = \max_{|\mathbf{j}| \leq k} \|D^{\mathbf{j}} f\|_{\infty} + \max_{|\mathbf{j}|=k} \sup_{x \neq y \in [0,1]^d} \frac{|D^{\mathbf{j}} f(x) - D^{\mathbf{j}} f(y)|}{\|x - y\|_2^{\vartheta}}.$$

Here $D^{\mathbf{j}}$ is the partial derivative with multi-index $\mathbf{j}$, of total order $|\mathbf{j}|$. We set $\|f\|_{C^t([0,1]^d)} = +\infty$ if these derivatives do not exist or are not continuous, and we define the Hölder ball

by $\mathcal{H}^t(H) = \{f : \|f\|_{C^t([0,1]^d)} \leq H\}$. In particular, for integer t the functions in this ball have Lipschitz partial derivatives of order $t - 1$.

We fix $d \geq 1$, $\alpha, \beta, H, \delta > 0$, and $0 < g_- < g_+ < \infty$. Let $\mathcal{P}(\alpha, \beta)$ be the class of all laws on the stated observation space such that

$$w \geq \delta, \quad a \in \mathcal{H}^\alpha(H), \quad b \in \mathcal{H}^\beta(H), \quad g_- \leq g \leq g_+ \quad \text{almost everywhere.} \quad (2.3)$$

These are the only restrictions on the conditional response law and the covariate density; we impose no smoothness bound on g or semiparametric shape restrictions. Since $\delta \leq w \leq M_0$, the covariate density satisfies $g_-/M_0 \leq p \leq g_+/\delta$. The functions that define the target, the smoothness indices, the radii, and the bounds are all fixed as n grows.

For a real-valued parameter T on a class $\mathcal{P}$ of observation laws, we define the minimax root-mean-square error (RMSE) by

$$r_n(T, \mathcal{P}) = \inf_{\hat{T}} \sup_{P \in \mathcal{P}} \{\mathbb{E}_P[(\hat{T} - T(P))^2]\}^{1/2}.$$

The infimum is over all measurable estimators based on the n observations, including randomized ones. For the parameter (2.1) on $\mathcal{P}(\alpha, \beta)$, we write $r_n = r_n(\psi, \mathcal{P}(\alpha, \beta))$. The constants in $\lesssim$ and $\asymp$ never depend on n . In statements about $\mathcal{P}(\alpha, \beta)$, they may depend on $d, \alpha, \beta, H, \delta, g_-, g_+, M_0$, on the response functions D, U, V, W , and on λ . In lower bounds, they may also depend on the baseline law π_0 and the radius r that we introduce below; in a lemma, they may also depend on the lemma's fixed inputs.

For the lower bound, we need a law at which every constraint in (2.3) holds strictly and perturbations of the conditional law of Z within the model move a and b . For example, if $U = D$, then $\psi(P) = \mathbb{E}_P[W + \lambda V]$ is an observable mean, which can be estimated at rate $n^{-1/2}$. We let π_0 be a probability law for Z with finite support, and we write

$$w_0 = \mathbb{E}_{\pi_0} D, \quad a_0 = \frac{\mathbb{E}_{\pi_0} U}{w_0}, \quad b_0 = \frac{\mathbb{E}_{\pi_0} V}{w_0}. \quad (2.4)$$

If we take X uniform on $[0, 1]^d$ and independent of $Z \sim \pi_0$, we obtain a joint law P_0 under which $w = g = w_0$, $a = a_0$, and $b = b_0$ are constant. For the lower bound, we assume the interior and nondegeneracy condition

$$\begin{aligned} \max\{\delta, g_-\} < w_0 < g_+, \quad |a_0| < H, \quad |b_0| < H, \\ \text{and either } \quad \text{Cov}_{\pi_0} \begin{pmatrix} U - a_0 D \\ V - b_0 D \end{pmatrix} \text{ is positive definite} \\ \text{or } \quad U = V, \quad \alpha = \beta, \quad \text{Var}_{\pi_0}(U - a_0 D) > 0. \end{aligned} \quad (2.5)$$

The first line states that P_0 satisfies (2.3) strictly, since a constant has Hölder norm equal to its absolute value. The residuals $U - a_0 D$ and $V - b_0 D$ have mean zero under

π_0 ; a positive definite covariance allows us to shift their means, and hence a and b , independently. The second alternative covers the square of one conditional mean, for which this covariance is singular. For $U = D$, the first residual vanishes and neither alternative holds.

For $r > 0$, we define the neighborhood

$$\mathcal{P}_r(\pi_0) = \{P \in \mathcal{P}(\alpha, \beta) : |w - w_0| \leq r, |a - a_0| \leq r, |b - b_0| \leq r \text{ almost everywhere on } [0, 1]^d\}. \quad (2.6)$$

Under (2.5), this class contains P_0 , and a lower bound on it holds on every larger class. For example, suppose that (2.3) is supplemented by bounds $w_- \leq w \leq w_+$, $a_- \leq a \leq a_+$, and $b_- \leq b \leq b_+$ with $w_- < w_0 < w_+$, $a_- < a_0 < a_+$, and $b_- < b_0 < b_+$, such as an overlap condition for missing-data means. If r is smaller than the six differences $w_0 - w_-$, $w_+ - w_0$, $a_0 - a_-$, $a_+ - a_0$, $b_0 - b_-$, $b_+ - b_0$, then the resulting class contains $\mathcal{P}_r(\pi_0)$ and is contained in $\mathcal{P}(\alpha, \beta)$.

2.3 Main theorem

The rates depend on the smoothness and on the bounds g_-, g_+ through

$$s = \frac{\alpha + \beta}{2}, \quad \theta = \frac{2s}{d}, \quad \rho = \frac{\sqrt{g_+} - \sqrt{g_-}}{\sqrt{g_+} + \sqrt{g_-}}, \quad \tau = -\log \rho. \quad (2.7)$$

The number ρ is the geometric rate of the reciprocal series of Lemma 4.1 on $[g_-, g_+]$, and τ determines the correction to the polynomial rate. Since $\rho = (\sqrt{g_+/g_-} - 1)/(\sqrt{g_+/g_-} + 1)$, it is also the rate in the classical convergence bound for the conjugate gradient method applied to a linear system with a symmetric positive definite matrix of condition number g_+/g_- [65, Lecture 38]. The upper bound also depends on

$$q_t = \binom{d + \lceil t \rceil - 1}{d} \quad (t > 0), \quad v = q_\alpha + q_\beta. \quad (2.8)$$

The dimension q_t counts the coefficients of a polynomial of degree $\lceil t \rceil - 1$ in d variables. We have $q_t = 1$ if and only if $0 < t \leq 1$, so that $v = 2$ if and only if $0 < \alpha, \beta \leq 1$. For $0 < \theta < 1/2$ and $\tau > 0$, we define

$$\kappa = 2\sqrt{\theta(1 - 2\theta)\tau}, \quad a_n = a_n(\theta, \tau) = n^{-\theta}(\log n)^{\theta/2}e^{-\kappa\sqrt{\log n}}. \quad (2.9)$$

We call $\theta = 1/2$, that is, $s = d/4$, the threshold.

Theorem 2.3 (Minimax bounds). *Consider the parameter (2.1) on the class (2.3).*

(a) There are $C > 0$ and an integer $n_0 \geq 3$, depending only on $d, \alpha, \beta, H, \delta, g_-, g_+, M_0, |\lambda|$, and $\|W\|_\infty$, such that, if $\mathcal{P}(\alpha, \beta)$ is nonempty, then for every integer $n \geq n_0$

$$r_n \leq Cn^{-1/2} \quad \text{if } \theta \geq 1/2, \quad r_n \leq C a_n (\log n)^{v/2+1/4} \quad \text{if } 0 < \theta < 1/2. \quad (2.10)$$

(b) Suppose that a law π_0 for Z with finite support satisfies (2.5), and fix $r > 0$. There are $c > 0$ and an integer $n_0 \geq 3$, depending only on $d, \alpha, \beta, H, \delta, g_-, g_+, M_0$, the response functions $D, U, V, W, \lambda, \pi_0$, and r , such that the following holds for every class $\mathcal{P}'$ with $\mathcal{P}_r(\pi_0) \subseteq \mathcal{P}' \subseteq \mathcal{P}(\alpha, \beta)$ and every integer $n \geq n_0$. If $\theta \geq 1/2$, then $r_n(\psi, \mathcal{P}') \geq cn^{-1/2}$. If $0 < \theta < 1/2$, then $r_n(\psi, \mathcal{P}') \geq c a_n \log n$ and

$$\inf_{\hat{\psi}} \sup_{P \in \mathcal{P}'} \mathbb{P}_P \{ |\hat{\psi} - \psi(P)| \geq 2c a_n \log n \} \geq \frac{1}{4}. \quad (2.11)$$

In particular, if such a π_0 exists, then for all $n \geq n_0$, with n_0 the larger of the two thresholds,

$$cn^{-1/2} \leq r_n \leq Cn^{-1/2} \quad (\theta \geq 1/2), \quad (2.12)$$

$$c a_n \log n \leq r_n \leq C a_n (\log n)^{v/2+1/4} \quad (0 < \theta < 1/2), \quad (2.13)$$

and the same bounds hold for $r_n(\psi, \mathcal{P}')$ for every class $\mathcal{P}'$ as in (b).

Remark 2.4 (Logarithmic form and the remaining gap). Taking logarithms in (2.13), we obtain

$$\log r_n = -\theta \log n - 2\sqrt{\theta(1-2\theta)\tau} \log n + O(\log \log n), \quad (2.14)$$

which is (1.2). The priors of Lemma 6.2 with $J = 0$, whose phases are constant on each block, give $r_n \geq c a_n$ by the proof of Proposition 7.1, and with the upper bound this implies (2.14). The lattice phases, with $J \asymp \log \log n$, gain the factor $\log n$, so that $r_n/a_n \rightarrow \infty$, and a_n is not the minimax rate up to fixed constants. The bounds leave the factor $(\log n)^{v/2-3/4}$ unresolved, which is $(\log n)^{1/4}$ when $0 < \alpha, \beta \leq 1$. The constants may grow as $\theta \uparrow 1/2$ or as the residual covariance in (2.5) approaches singularity.

Remark 2.5 (Smooth and nearly uniform densities). Every density in the lower-bound constructions of this paper is C^∞ , but its derivatives are not bounded uniformly in n . All lower bounds therefore persist if we require that p be C^∞ , provided that no uniform bound on its derivatives is imposed. When $D = 1$, as in Example 2.2, we have $w = 1$ and $g = p$. Applying Theorem 2.3 with $g_\pm = 1 \pm \omega$ and $\delta < 1$, we find that, under the remaining conditions of (2.5), even the bounds $1 - \omega \leq p \leq 1 + \omega$, for any fixed $0 < \omega < 1$, give the exponent $2s/d$ below the threshold. The same holds, with $p_\pm = 1 \pm \omega$, for the rows of Table 3.1 with the interval $[p_-, p_+]$. The width of the interval changes only the correction in (2.14).

Applications

Corollary 3.2 and Table 3.1 collect the applications of Theorem 2.3, with proofs in Appendices A to D.

For an interval $I = [i_-, i_+]$ with $0 < i_- < i_+ < \infty$, we write ρ_I and τ_I for the values of ρ and τ in (2.7) computed from $[g_-, g_+] = I$. For $\theta, \tau > 0$ and $\nu \geq 2$, we define

$$\underline{a}_n(\theta, \tau) = a_n(\theta, \tau) \log n, \quad \bar{a}_n(\theta, \tau, \nu) = a_n(\theta, \tau) (\log n)^{\nu/2+1/4} \quad (3.1)$$

if $\theta < 1/2$, with $a_n(\theta, \tau)$ as in (2.9), and $\underline{a}_n(\theta, \tau) = \bar{a}_n(\theta, \tau, \nu) = n^{-1/2}$ if $\theta \geq 1/2$.

Definition 3.1. Let T be a real parameter on a class $\mathcal{P}$ of observation laws, let $\theta > 0$, $\nu \geq 2$, and let I be an interval as above. We say that $(T, \mathcal{P})$ satisfies the *lower bracket* (θ, I) if there are $c > 0$ and an integer $n_0 \geq 3$ such that, for every integer $n \geq n_0$, we have $r_n(T, \mathcal{P}) \geq c \underline{a}_n(\theta, \tau_I)$ and, if $\theta < 1/2$,

$$\inf_{\hat{T}} \sup_{P \in \mathcal{P}} \mathbb{P}_P\{|\hat{T} - T(P)| \geq 2c \underline{a}_n(\theta, \tau_I)\} \geq \frac{1}{4}. \quad (3.2)$$

We say that $(T, \mathcal{P})$ satisfies the *upper bracket* (θ, I, ν) if there are $C > 0$ and $n_0 \geq 3$ such that $r_n(T, \mathcal{P}) \leq C \bar{a}_n(\theta, \tau_I, \nu)$ for every integer $n \geq n_0$. It satisfies the *bracket* (θ, I, ν) if it satisfies both.

The targets and classes in Table 3.1 are defined below. Most of these classes bound the covariate density by

$$0 < p_- < 1 < p_+ < \infty, \quad p_- \leq p \leq p_+. \quad (3.3)$$

We write $\tau_p = \tau_{[p_-, p_+]}$, we set $\kappa_p(\theta) = 2\sqrt{\theta(1-2\theta)\tau_p}$ for $0 < \theta < 1/2$, and we write $I_\zeta = [(1-\zeta)p_-, (1+\zeta)p_+]$ for $\zeta \in (0, 1)$.

Corollary 3.2 (Applications). Fix $d \in \mathbb{N}$ and the parameters of the classes in Table 3.1. Each pair $(T, \mathcal{P})$ of a target and a class in the table satisfies the bracket $(2s/d, I, \nu)$ of its row, with the following qualification and addition.

- (a) On $\mathcal{M}^{\text{sep}}$ and $\mathcal{T}^{\text{sep}}$, the lower bracket holds with $I = [p_-, p_+]$, and the upper bracket holds with I_ζ in place of I , for every fixed $\zeta \in (0, 1)$. In particular, if $s < d/4$, then

$$\log r_n(T, \mathcal{P}) = -\frac{2s}{d} \log n - \kappa_p(2s/d) \sqrt{\log n} + o(\sqrt{\log n}). \quad (3.4)$$

(b) Let f_0 be a fixed known measurable predictor with $B_0 = \|f_0\|_\infty < \infty$. On $\mathcal{P}_{\text{quad}}(s)$, the excess loss $\mathbb{E}_P[(f(X) - f_0(X))^2]$ satisfies the bracket $(2s/d, [p_-, p_+], 2q_s)$ of $\bar{\sigma}^2$.

The constants c, C, n_0 depend only on the fixed quantities that define T and $\mathcal{P}$, and also on ζ in (a); in (b), they depend on f_0 only through B_0 .

Table 3.1. The applications in Corollary 3.2. Here β_* is $\min(\beta_0, \beta_1)$ for the ATE, β_0 for the ATT, and β_1 for the ATU. The last column names the source of the lower bound: Theorem 2.3(b), or the families of laws and the Markov kernels constructed in Appendices B to D.

Target	Class	s	I	ν	Lower bound from
MAR mean	$\mathcal{M}_\eta$	$\frac{\alpha+\beta}{2}$	$[\eta, \eta^{-1}]$	$q_\alpha + q_\beta$	Theorem 2.3(b)
ATE, ATT, ATU	$\mathcal{T}_\eta$	$\frac{\alpha+\beta_*}{2}$	$[\eta, \eta^{-1}]$	$q_\alpha + q_{\beta_*}$	MAR family, kernel
MAR mean	$\mathcal{M}^{\text{sep}}$	$\frac{\alpha+\beta}{2}$	$[p_-, p_+]$	$q_\alpha + q_\beta$	MAR family
ATE, ATT, ATU	$\mathcal{T}^{\text{sep}}$	$\frac{\alpha+\beta_*}{2}$	$[p_-, p_+]$	$q_\alpha + q_{\beta_*}$	MAR family, kernel
$\mathbb{E}[\text{Cov}(U, V \mid X)],$ $\mathbb{E}[m_U m_V]$	$\mathcal{P}_{\text{bil}}(\alpha, \beta)$	$\frac{\alpha+\beta}{2}$	$[p_-, p_+]$	$q_\alpha + q_\beta$	Theorem 2.3(b)
$\mathbb{E}[\text{Var}(Y \mid X)], \mathbb{E}[f^2]$	$\mathcal{P}_{\text{quad}}(s)$	s	$[p_-, p_+]$	$2q_s$	Theorem 2.3(b)
$\text{Var}(f(X))$	$\mathcal{P}_{\text{quad}}(s)$	s	$[p_-, p_+]$	$2q_s$	Rademacher family
R^2	$\mathcal{P}_{\text{quad}}(s; v_{\min})$	s	$[p_-, p_+]$	$2q_s$	Rademacher family
$\mathbb{E}[f^2], \text{Var}(f(X))$ in a trial	$\mathcal{P}_{\text{trial}}(s)$	s	$[p_-, p_+]$	$2q_s$	kernel from $\mathcal{P}_{\text{quad}}(s)$
Wald ratio	$\mathcal{I}_{\eta, c_W}$	$\frac{\alpha+\beta}{2}$	$[\eta, \eta^{-1}]$	$q_\alpha + q_\beta$	ATE, kernel
Average conditional Wald ratio	$\mathcal{J}_\eta$	$\frac{\alpha+\beta}{2}$	$[\eta, \eta^{-1}]$	$q_\alpha + q_\beta$	MAR mean, kernel
Overlap-weighted effect	$\mathcal{P}_{\text{OW}}(\alpha, \beta)$	$\min\{\frac{\alpha+\beta}{2}, \alpha\}$	$[p_-, p_+]$	$q_\alpha + q_{\min(\alpha, \beta)}$	two families

In every row, $r_n(T, \mathcal{P}) = n^{-\min\{1/2, 2s/d\} + o(1)}$, and $\nu = 2$ when all smoothness indices relevant to the row are at most one. When $s < d/4$, the lower bound rules out, even up to factors $n^{o(1)}$, the exponent $4s/(4s + d)$ that higher-order influence function estimators attain when the density weight g is known or sufficiently smooth [52, Section 3.1 and Corollary 9.1]. Every density in the lower-bound constructions is C^∞ , so that, as in Remark 2.5, all these lower bounds persist if p must be C^∞ with no uniform bound on its derivatives.

3.1 Missing-data means and treatment effects

We use the missing-data observations $O = (X, (R, RY))$ of Example 2.1, with $(D, U, V, W, \lambda) = (R, 1, RY, 0, 1)$, so that $w = P(R = 1 \mid X)$, $a = 1/w$, $b = \mathbb{E}_P[Y \mid R = 1, X]$, and the target is the MAR mean $\mathbb{E}_P[b(X)]$. We fix $0 < \eta < 1/4$ and $H > 2$, and we let $\mathcal{M}_\eta(\alpha, \beta, H)$ consist of all laws of these observations such that $Y \in [0, 1]$ and

$$1/w \in \mathcal{H}^\alpha(H), \quad b \in \mathcal{H}^\beta(H), \quad \eta \leq w \leq 1 - \eta, \quad \eta \leq wp \leq \eta^{-1}.$$

This is the class $\mathcal{P}(\alpha, \beta)$ with $M_0 = 1$, $\delta = \eta$, and $[g_-, g_+] = [\eta, \eta^{-1}]$, so that $\rho = (1 - \eta)/(1 + \eta)$, together with the overlap condition $w \leq 1 - \eta$.

For treatment effects, we observe $O = (X, A, Y)$ with $A \in \{0, 1\}$ and $Y \in [0, 1]$, and we write

$$\begin{aligned} w(x) &= P(A = 1 \mid X = x), & w_1 &= w, & w_0 &= 1 - w, \\ a_j &= 1/w_j, & g_j &= w_j p, & b_j(x) &= \mathbb{E}_P[Y \mid A = j, X = x] \quad (j = 0, 1). \end{aligned}$$

Here the subscripts index the treatment arms; w_0 and a_0 are not the baseline constants of Section 2. We let $\mathcal{T}_\eta(\alpha, \beta_0, \beta_1, H)$ consist of all laws of O such that

$$\eta \leq w \leq 1 - \eta, \quad a_0, a_1 \in \mathcal{H}^\alpha(H), \quad \eta \leq g_0, g_1 \leq \eta^{-1}, \quad b_j \in \mathcal{H}^{\beta_j}(H) \quad (j = 0, 1). \quad (3.5)$$

These restrictions imply $2\eta \leq p \leq 2\eta^{-1}$. We write $\mu_j = \mathbb{E}_P[b_j(X)]$ and consider

$$\begin{aligned} T_{\text{ATE}} &= \mu_1 - \mu_0, \\ T_{\text{ATT}} &= \mathbb{E}_P[b_1(X) - b_0(X) \mid A = 1], \quad T_{\text{ATU}} = \mathbb{E}_P[b_1(X) - b_0(X) \mid A = 0]. \end{aligned} \quad (3.6)$$

Under consistency, conditional exchangeability, and positivity [18], these are the average effects in the population, among treated units, and among untreated units. Their effective smoothness indices are

$$s_{\text{ATE}} = \frac{\alpha + \min(\beta_0, \beta_1)}{2}, \quad s_{\text{ATT}} = \frac{\alpha + \beta_0}{2}, \quad s_{\text{ATU}} = \frac{\alpha + \beta_1}{2}. \quad (3.7)$$

To bound the difference from below, we vary one arm regression and keep the other constant (Appendix B).

If we bound the covariate density and the propensity separately, the correction is determined by the density interval rather than by the range of wp . We fix $H_0 > 1$ and $0 < \epsilon < 1/2$. We let $\mathcal{M}^{\text{sep}}$ consist of the laws of the missing-data observations such that $Y \in [0, 1]$, p satisfies (3.3), $w \in \mathcal{H}^\alpha(H_0)$, $\epsilon \leq w \leq 1 - \epsilon$, and $b \in \mathcal{H}^\beta(H_0)$. We let $\mathcal{T}^{\text{sep}}$ consist of the laws of (X, A, Y) such that $Y \in [0, 1]$, p satisfies (3.3), $w = P(A = 1 \mid X)$ satisfies the same conditions, and $b_j \in \mathcal{H}^{\beta_j}(H_0)$ for $j = 0, 1$. Here α is the smoothness of

w itself, equivalently of $1/w$ up to the Hölder radius, since $\epsilon \leq w \leq 1 - \epsilon$. In all four classes, the conditional law of the response is otherwise unrestricted, and no smoothness of p is assumed.

For the upper bounds on $\mathcal{M}^{\text{sep}}$ and $\mathcal{T}^{\text{sep}}$, we divide R and RY by a smooth pilot estimate $\hat{w}$ of w computed from half of the sample. This preserves the target and replaces the weight wp by $wp/\hat{w}$, which lies within a factor $1 \pm \zeta$ of p except on an event of exponentially small probability (Appendix B.2). For μ_j , we apply this construction with $R = \mathbf{1}_{\{A=j\}}$ and observation probability w_j as in Example 2.1; the arm-0 pilot estimates $1 - w$. The correction computed from I_ζ tends to $\kappa_p(2s/d)$ as $\zeta \rightarrow 0$, which gives (3.4).

3.2 Conditional covariances and variances

We take $D = 1$ as in Example 2.2, so that $g = p$, and we fix $H_0 > 1$ and the density bounds (3.3). Let $\mathcal{P}_{\text{bil}}(\alpha, \beta)$ consist of all laws of (X, U, V) that satisfy (3.3), $|U|, |V| \leq 1$, $m_U \in \mathcal{H}^\alpha(H_0)$, and $m_V \in \mathcal{H}^\beta(H_0)$, where $m_U = \mathbb{E}_P[U | X]$ and $m_V = \mathbb{E}_P[V | X]$. For $U = V = Y$, let $\mathcal{P}_{\text{quad}}(s)$ consist of all laws of (X, Y) that satisfy (3.3), $|Y| \leq 1$, and $f \in \mathcal{H}^s(H_0)$, where $f = \mathbb{E}_P[Y | X]$. We write

$$Q_2 = \mathbb{E}_P[f(X)^2], \quad \bar{\sigma}^2 = \mathbb{E}_P[\text{Var}_P(Y | X)], \quad E_* = \text{Var}_P(f(X)) = Q_2 - (\mathbb{E}_P Y)^2.$$

For fixed $v_{\min} \in (0, 1)$, we let $\mathcal{P}_{\text{quad}}(s; v_{\min})$ be the subclass of $\mathcal{P}_{\text{quad}}(s)$ on which $\text{Var}_P(Y) \geq v_{\min}$, and on it we define the population coefficient of determination $R_*^2 = E_* / \text{Var}_P(Y)$. In a randomized trial we observe (X, A, Y_{obs}) , and we let $\mathcal{P}_{\text{trial}}(s)$ consist of the laws such that $A \in \{0, 1\}$, $Y_{\text{obs}} \in [0, 1]$, $P(A = 1 | X) = 1/2$, p satisfies (3.3), and the conditional average treatment effect (CATE) $f = \mathbb{E}_P[Y_{\text{obs}} | A = 1, X] - \mathbb{E}_P[Y_{\text{obs}} | A = 0, X]$ lies in $\mathcal{H}^s(H_0)$. All these classes leave the conditional law of the response otherwise unrestricted.

For E_* and R_*^2 , the upper bounds combine the estimator of Q_2 with sample means of Y and Y^2 , and the lower bounds use hard laws on which $\text{Var}_P(Y) \rightarrow 1$. In Gaussian linear regression, the same relations through observable moments imply that the signal strength, the noise variance, and the proportion of explained variance are equally hard to estimate, up to a parametric loss [69, Section 1.1]. In the trial, the ATE $\mathbb{E}_P[f(X)]$ is an observable mean, whereas $\text{Var}_P(f(X))$ is as hard to estimate as E_* ; only f needs to be smooth. Confidence intervals for this variance in experiments, which remain valid when it is zero, are constructed in [57]. In Corollary 3.2(b), the loss $\mathbb{E}_P[(Y - f_0(X))^2]$ is an observable mean, but the excess loss is as hard to estimate as $\bar{\sigma}^2$.

For the expected conditional covariance, the probability bound (3.2) on $\mathcal{P}_{\text{bil}}(\alpha, \beta)$ shows that the polynomial exponent in McClean et al.'s [43] upper bound in probability is minimax optimal. Evans and Jones [15, Section 1(a) and Section 2] studied nearest-neighbor estimators of residual moments and covariance under Lipschitz regression,

regularity conditions on the design density, and additive errors independent of the covariates; Williamson et al. [72, Section 2.3, Example 1, and Section 3, Theorems 1–2] studied R^2 inference under convergence conditions on estimated regressions. Variance estimation has also been studied in the semiparametric model $\text{Var}(Y | X) = \sigma^2$ [48, 59]. There, when $0 < s \leq 1$ and $s < d/4$, a rough random design permits faster estimation than an equally spaced fixed design [9, 71]. This was sketched in [48, Section 4, pp. 382–383] for $s < \min\{1, d/4\}$, proved in [59] for errors independent of X , and proved in [3, Theorem 1.3] for a model in which the law of the error may depend on X . For $s > 1$, Aronow and Lopatto [3] resolved the uniform minimax formulation of an open question of Robins, recorded by Richardson and Rotnitzky [47, Section 6]. For $s > 1$ and $d > 4s$, the two-scale estimator of Dobriban et al. [14, Section 3.3, Proposition 3.4] has root-mean-square error $O(n^{-2(s+1)/(d+4)})$, the rate of the upper bound in [3, Theorem 1.1]. The proof for this estimator uses only upper and lower bounds on the covariate density [14, Appendix A.5]. Corollary 3.2 concerns the fully nonparametric model, to which these faster rates do not apply.

3.3 Wald ratios and the overlap-weighted effect

We observe (X, Z, A, Y) with a binary instrument Z , a binary treatment A , and $Y \in [0, 1]$. We write $w_z(x) = P(Z = z | X = x)$, $m_{Y,z}(x) = \mathbb{E}_P[Y | Z = z, X = x]$, $m_{A,z}(x) = \mathbb{E}_P[A | Z = z, X = x]$, $\Delta_Y = m_{Y,1} - m_{Y,0}$, and $\Delta_A = m_{A,1} - m_{A,0}$. When $\mathbb{E}_P[\Delta_A(X)] > 0$, the *Wald ratio* is $T_W = \mathbb{E}_P[\Delta_Y(X)] / \mathbb{E}_P[\Delta_A(X)]$. If $P(Z = 1 | X) = 1/2$, $A \leq Z$, and $q = P(A = 1 | Z = 1, X)$ is bounded away from zero, then $\Delta_A = q$, and the *average conditional Wald ratio* is $T_{CW} = \mathbb{E}_P[b(X)]$ with $b = \Delta_Y/q$. We fix $\alpha, \beta > 0$, $H > 2$, $0 < \eta < 1/4$, and $0 < c_W \leq 1$. We let $\mathcal{I}_{\eta, c_W}(\alpha, \beta, H)$ consist of the laws such that, for $z = 0, 1$,

$$\begin{aligned} \eta \leq w_z \leq 1 - \eta, \quad 1/w_z \in \mathcal{H}^\alpha(H), \quad \eta \leq w_z p \leq \eta^{-1}, \\ m_{Y,z}, m_{A,z} \in \mathcal{H}^\beta(H), \quad \mathbb{E}_P[\Delta_A(X)] \geq c_W, \end{aligned}$$

and we let $\mathcal{J}_\eta(\alpha, \beta, H)$ consist of the laws such that $P(Z = 1 | X) = 1/2$, $A \leq Z$, and

$$1/q \in \mathcal{H}^\alpha(H), \quad b \in \mathcal{H}^\beta(H), \quad \eta \leq q \leq 1 - \eta, \quad \eta \leq qp \leq \eta^{-1}.$$

For observations (X, A, Y) with a binary treatment, $Y \in [0, 1]$, propensity $w = P(A = 1 | X)$, and regression $m = \mathbb{E}_P[Y | X]$, the *overlap-weighted effect* is

$$\gamma_{OW} = \frac{\mathbb{E}_P[\text{Cov}_P(A, Y | X)]}{\mathbb{E}_P[\text{Var}_P(A | X)]} = \frac{\mathbb{E}_P[w(1-w)(b_1 - b_0)(X)]}{\mathbb{E}_P[w(1-w)(X)]}, \quad (3.8)$$

with $b_a = \mathbb{E}_P[Y | A = a, X]$. It is the coefficient of A in the least-squares projection of Y onto functions $\gamma A + h(X)$. The first expression in (3.8) is the population version of the

partialing-out construction of Robinson [54] for the partially linear model, which we do not assume to hold. We let $\mathcal{P}_{\text{OW}}(\alpha, \beta)$ consist of the laws that satisfy (3.3), $w \in \mathcal{H}^\alpha(H_0)$, $m \in \mathcal{H}^\beta(H_0)$, and $\epsilon \leq w \leq 1 - \epsilon$, with $H_0 > 1$ and $0 < \epsilon < 1/2$. In all three classes, the conditional law of the response is otherwise unrestricted. Appendix D recalls the causal interpretations of the three targets. The numerator of γ_{OW} has effective smoothness $(\alpha + \beta)/2$; its denominator depends only on the propensity and has effective smoothness α . The ratio has $s_{\text{OW}} = \min\{(\alpha + \beta)/2, \alpha\}$, as in Table 3.1.

Upper bounds

We prove Theorem 2.3(a) by approximating $\int abg$ by polynomials in expectations of known functions of one observation. These have unbiased estimators, so we never estimate g . Our estimator is a sum of U-statistics of growing order, like the HOIF estimators of [48, 52], which correct an initial estimate by U-statistics of increasing order. Our estimator uses no initial estimate of g (Section 1.4). For linear models, Kong and Valiant [28] developed a method that approximates a quadratic form in the inverse of an unknown covariance matrix by polynomials, whose terms are estimated without bias by averages over distinct observations. Chen, Liu, and Mukherjee [10, Section 3.2] use Chebyshev polynomials of the reciprocal in this way. We apply this idea cell by cell, to the local Gram matrices of g . We use reciprocal series to approximate projection increments over nested partitions, bound the variance of their unbiased estimators, and choose the resolution and polynomial degrees.

We first reduce to $W = 0$ and $\lambda = 1$, so that $\psi = \int abg$. For an estimator $\hat{T}$ of $T(P) = \int abg$, by the triangle inequality, we have

$$\left\| \lambda \hat{T} + \frac{1}{n} \sum_{i=1}^n W(O_i) - \psi(P) \right\|_{L^2(P^n)} \leq |\lambda| \|\hat{T} - T(P)\|_{L^2(P^n)} + \|W\|_{\infty} n^{-1/2},$$

and the last term is at most a fixed multiple of both upper rates. We may also assume that $\alpha \leq \beta$, since exchanging (a, U, α) with (b, V, β) changes neither $\int abg$ nor the class $\mathcal{P}(\alpha, \beta)$, and the rates are symmetric in (α, β) . We set $\ell = \lceil \beta \rceil - 1$, the cellwise polynomial degree of the estimator, and $s_0 = \min(\alpha, \beta)$. For $h \leq 1$, h^{s_0} bounds both h^{α} and h^{β} . Recall from (2.7) that $\theta = (\alpha + \beta)/d$ and $\tau = -\log \rho$, and that ν is given by (2.8).

In this section, constants depend only on $d, \alpha, \beta, H, \delta, g_-, g_+, M_0$, and, for the general target, on $|\lambda|$ and $\|W\|_{\infty}$. The estimator uses these known parameters and the values of D, U, V, W at the observations. The arguments use only the bounds $0 \leq D \leq M_0$, $|U|, |V| \leq M_0$, and $\|W\|_{\infty}$, with no property of the response space or baseline law π_0 .

4.1 The reciprocal series

Reciprocals enter both bounds. In the upper bound, we approximate products of reciprocals of cell moments of g by polynomials in those moments. In the lower bound, the separation between the target means of two priors is a Fourier coefficient of the reciprocal

of an oscillating density. In Lemma 4.1, we record the Chebyshev series of $1/x$ on a positive interval I , with geometric rate ρ_I ; the parameter ρ_I^{-1} is that of the Bernstein ellipse through the pole at $x = 0$, as in the classical theory of Chebyshev approximation of analytic functions [64, Chapter 8]. The Chebyshev coefficients of $1/(z - x)$ are recorded explicitly in [42, Lemma 1, equation (10), p. 598]. For $I = [g_-, g_+]$, or any positive multiple of it, ρ_I equals the ρ of (2.7).

Lemma 4.1 (Reciprocal expansion). *Fix an interval $I = [x_-, x_+]$ with $0 < x_- < x_+$, and set*

$$c_I = \frac{x_+ + x_-}{2}, \quad d_I = \frac{x_+ - x_-}{2}, \quad \bar{g}_I = \sqrt{x_- x_+}, \quad \rho_I = \frac{\sqrt{x_+} - \sqrt{x_-}}{\sqrt{x_+} + \sqrt{x_-}}.$$

Let T_k be the Chebyshev polynomial determined by $T_k(\cos t) = \cos(kt)$. For $x \in I$ and $|z| < \rho_I^{-1}$, we define

$$\mathcal{K}_I(z; x) = 1 + 2 \sum_{k \geq 1} (-\rho_I)^k T_k((x - c_I)/d_I) z^k.$$

Then

$$\mathcal{K}_I(z; x) = \frac{1 - \rho_I^2 z^2}{1 + 2\rho_I z(x - c_I)/d_I + \rho_I^2 z^2}. \quad (4.1)$$

In particular, $\mathcal{K}_I(1; x) = \bar{g}_I/x$; equivalently, writing $x = c_I + d_I \cos t$,

$$\frac{1}{c_I + d_I \cos t} = \frac{1}{\bar{g}_I} \left\{ 1 + 2 \sum_{k \geq 1} (-\rho_I)^k \cos(kt) \right\}. \quad (4.2)$$

Proof. We write $x = c_I + d_I \cos t$ with t real and $q = -\rho_I z$, so that $T_k((x - c_I)/d_I) = \cos(kt)$ and $|q| < 1$. Summing two absolutely convergent geometric series, we obtain the Poisson-kernel identity [61, Chapter 2]

$$1 + \sum_{k \geq 1} q^k (e^{ikt} + e^{-ikt}) = \frac{1}{1 - qe^{it}} + \frac{1}{1 - qe^{-it}} - 1 = \frac{1 - q^2}{1 - 2q \cos t + q^2},$$

which is (4.1). At $z = 1$, the identities $1 + \rho_I^2 = 2\rho_I c_I/d_I$ and $(1 - \rho_I^2)d_I/(2\rho_I) = \bar{g}_I$ give (4.2). $\square$

For $I = [g_-, g_+]$, the rate $\rho_I = \rho$ determines the constant κ of the upper bound through $\tau = -\log \rho$. The Neumann series $1/x = c_I^{-1} \sum_{k \geq 0} (1 - x/c_I)^k$, the most direct expansion of a reciprocal, converges on I only at the rate $d_I/c_I = 2\rho_I/(1 + \rho_I^2)$, which exceeds ρ_I . If we replaced (4.2) by the Neumann series in the argument of this section, we would obtain the correction with $-\log\{2\rho/(1 + \rho^2)\} < \tau$ in place of τ , and the upper bound would no longer match the lower bound. No other choice of polynomials improves on ρ_I : by a classical result of Chebyshev, the error of the best uniform approximation of $1/x$ on I by polynomials of degree m is a constant multiple of ρ_I^m [29, Theorem 2.1 and Corollary 2.2].

The HOIF estimators of Robins et al. with an estimated density expand the inverse of a Gram matrix about its value at an initial estimate [48, Theorems 3.14–3.15]. At resolutions with order n cells, each cell contains only order-one observations on average, so the normalized cell moments of g need not be uniformly consistently estimable under the stated assumptions, and we expand about the known interval $[g_-, g_+]$ instead.

In the upper bound we use these series only on the interval $[g_-, g_+]$, but with matrix arguments. We set $c_g = (g_- + g_+)/2$, $d_g = (g_+ - g_-)/2$, and $\bar{g} = \sqrt{g_- g_+}$, the values of $c_I, d_I, \bar{g}_I$ for this interval, for which $\rho_I = \rho$. From now on I denotes an identity matrix, and we regard scalars as 1×1 matrices. For a complex square matrix M , we define $A_M = (M - c_g I)/d_g$ and

$$\begin{aligned} Q(z; M) &= I + 2\rho z A_M + \rho^2 z^2 I, \\ \mathcal{K}(z; M) &= (1 - \rho^2 z^2) Q(z; M)^{-1} = I + 2 \sum_{k \geq 1} (-\rho)^k T_k(A_M) z^k. \end{aligned} \quad (4.3)$$

The last identity holds as a power series at $z = 0$: by the recurrence $T_{k+1}(A) = 2AT_k(A) - T_{k-1}(A)$, with $T_0(A) = I$ and $T_1(A) = A$, the product of $Q(z; M)$ and the series on the right is $(1 - \rho^2 z^2)I$, and $Q(0; M) = I$. The coefficient of z^k is a polynomial of degree at most k in the entries of M . For real symmetric M with spectrum in $[g_-, g_+]$, A_M has spectrum in $[-1, 1]$, so that $\|T_k(A_M)\|_{\text{op}} \leq 1$. By orthogonal diagonalization and Lemma 4.1, we obtain absolute convergence of (4.3) for $|z| < \rho^{-1}$, with $\mathcal{K}(1; M) = \bar{g} M^{-1}$.

We expand all reciprocal factors in the same variable z and truncate their combined index to keep the degree at m . The next lemma bounds the approximation error and, for Lemma 4.5, the derivatives uniformly in the degree. For a function f on $\mathbb{C}^p$, a point t_0 , and directions $v_1, \dots, v_r \in \mathbb{R}^p$, we write $D^r f(t_0)[v_1, \dots, v_r]$ for the derivative $\partial_{s_1} \cdots \partial_{s_r} f(t_0 + s_1 v_1 + \cdots + s_r v_r)$ at $s = 0$.

Lemma 4.2 (Reciprocal truncations and derivative bounds). *(a) Fix integers $N \geq 0$, $p_0 \geq 1$, and $p_1, \dots, p_N \geq 0$, and set $v' = p_1 + \cdots + p_N$. For real symmetric matrices $M_0, \dots, M_N$ of dimensions $p_0, \dots, p_N$ and every integer $m \geq 0$, we define G_k and S_m by the expansion in powers of z*

$$\mathcal{K}(z; M_0) \prod_{i=1}^N \det \mathcal{K}(z; M_i) = \sum_{k \geq 0} G_k z^k, \quad S_m(M_0, \dots, M_N) = \bar{g}^{-(v'+1)} \sum_{k=0}^m G_k, \quad (4.4)$$

where a determinant of dimension zero equals one. Then S_m is a matrix polynomial of total degree at most m in the entries of $M_0, \dots, M_N$. If the spectra of $M_0, \dots, M_N$ lie in $[g_-, g_+]$, then

$$\begin{aligned} \left\| \frac{M_0^{-1}}{\det M_1 \cdots \det M_N} - S_m(M_0, \dots, M_N) \right\|_{\text{op}} &\leq C(m+1)^{v'} \rho^m, \\ C &= 2^{v'+1} \bar{g}^{-(v'+1)} \sum_{k \geq 0} (k+1)^{v'} \rho^k. \end{aligned} \quad (4.5)$$

- (b) Let $\mathcal{V} \subset \mathbb{R}^p$ be bounded, and let $M_1(t), \dots, M_N(t)$ be fixed affine maps from $\mathbb{C}^p$ into complex square matrices that are real symmetric with spectra in $[g_-, g_+]$ for $t \in \mathcal{V}$. Let $\mathcal{S}(z; t) = \sum_{k \geq 0} \mathcal{S}_k(t) z^k$ be formed by fixed finite sums and products of entries of the kernels $\mathcal{K}(z; M_i(t))$, $1 \leq i \leq N$, and of fixed polynomials in t . Then there are $\varepsilon \in (0, 1]$ and C such that $|\sum_{k=0}^m \mathcal{S}_k(t)| \leq C$ for every $m \geq 0$ and every $t \in \mathbb{C}^p$ with $\|t - t_0\|_\infty < \varepsilon$ for some $t_0 \in \mathcal{V}$. The constants depend only on $g_\pm$, the maps M_i , the expression defining $\mathcal{S}$, and the bound on $\mathcal{V}$.
- (c) Let f be a polynomial on $\mathbb{C}^p$, let $t_0 \in \mathbb{R}^p$ and $\varepsilon > 0$, and suppose that $|f(t)| \leq A$ whenever $\|t - t_0\|_\infty < \varepsilon$. Then, for all $r \geq 1$ and $v_1, \dots, v_r \in \mathbb{R}^p$,

$$|D^r f(t_0)[v_1, \dots, v_r]| \leq A r! (e/\varepsilon)^r \prod_{i=1}^r \|v_i\|_\infty.$$

Proof. (a) The coefficient of z^k in $\mathcal{K}(z; M)$ has degree at most k in the entries of M . Since a determinant is a sum of products of entries, the coefficient of z^k in $\det \mathcal{K}(z; M_i)$ also has total degree at most k . The same holds for G_k , and S_m has degree at most m . For spectra in $[g_-, g_+]$, diagonalizing M_i , we have $\det \mathcal{K}(z; M_i) = \prod_{l=1}^{p_i} \mathcal{K}(z; \lambda_{il})$ for $|z| < \rho^{-1}$, where the λ_{il} are the eigenvalues of M_i . The left side of (4.4) is therefore a product of $v' + 1$ absolutely convergent series, and in each of them the coefficient of z^k has norm at most $2\rho^k$. There are $\binom{k+v'}{v'} \leq (k+1)^{v'}$ ways to distribute an index k among $v' + 1$ factors, so that $\|G_k\|_{\text{op}} \leq 2^{v'+1} (k+1)^{v'} \rho^k$. In particular, $\sum_k G_k$ converges absolutely, and since $\mathcal{K}(1; M) = \bar{g}M^{-1}$ and $\det(\bar{g}M_i^{-1}) = \bar{g}^{p_i} / \det M_i$, its sum is $\bar{g}^{v'+1} M_0^{-1} / (\det M_1 \cdots \det M_N)$. Writing $k = m + j$ with $j \geq 1$ and using $m + j + 1 \leq (m+1)(j+1)$, we bound the left side of (4.5) by

$$\bar{g}^{-(v'+1)} \sum_{k > m} \|G_k\|_{\text{op}} \leq 2^{v'+1} \bar{g}^{-(v'+1)} (m+1)^{v'} \rho^m \sum_{j \geq 1} (j+1)^{v'} \rho^j,$$

which is at most the right side of (4.5).

(b) Let $r_* = \rho^{-1/2}$ and $b = (1 - \sqrt{\rho})^2$. If M is real symmetric with spectrum in $[g_-, g_+]$, then $Q(z; M)$ is diagonal in an orthonormal eigenbasis of M , with diagonal entries $(1 + \rho z e^{i\phi})(1 + \rho z e^{-i\phi})$ for real ϕ , and these have absolute value at least $(1 - \rho r_*)^2 = b$ for $|z| \leq r_*$. Thus $\|Q(z; M)^{-1}\|_{\text{op}} \leq b^{-1}$ for $|z| \leq r_*$. Let $\Lambda \geq 1$ be such that $\|M_i(t) - M_i(t_0)\|_{\text{op}} \leq \Lambda \|t - t_0\|_\infty$ for all i and all complex t, t_0 , and let $\varepsilon = \min\{1, d_g b / (4\Lambda)\}$. If $t_0 \in \mathcal{V}$ and $\|t - t_0\|_\infty < \varepsilon$, then

$$\|Q(z; M_i(t)) - Q(z; M_i(t_0))\|_{\text{op}} \leq 2\rho r_* \Lambda \varepsilon / d_g \leq b/2, \quad |z| \leq r_*,$$

and the Neumann series shows that $\|Q(z; M_i(t))^{-1}\|_{\text{op}} \leq 2/b$ and $\|\mathcal{K}(z; M_i(t))\|_{\text{op}} \leq 2(1 + \rho)/b$ there. Since the inequality $\|t - t_0\|_\infty < \varepsilon$ is strict, the kernels are holomorphic on a neighborhood of $|z| \leq r_*$. The fixed polynomials are bounded on the bounded set of such t , and we conclude that $\mathcal{S}(\cdot; t)$ is holomorphic and bounded by a uniform

A on $|z| \leq r_*$. Its Taylor coefficients are $\mathcal{S}_k(t)$ by (4.3); by Cauchy's inequality, we have $|\mathcal{S}_k(t)| \leq Ar_*^{-k}$. Summing over $k \leq m$, we obtain the bound $A/(1 - \sqrt{\rho})$.

(c) By multilinearity we may assume that $\|v_i\|_\infty = 1$ for each i . The polynomial $\varphi(w_1, \dots, w_r) = f(t_0 + \sum_i w_i v_i)$ on $\mathbb{C}^r$ is bounded by A when $\max_i |w_i| < \varepsilon/r$, and

$$D^r f(t_0)[v_1, \dots, v_r] = \partial_{w_1} \cdots \partial_{w_r} \varphi(0).$$

By Cauchy's formula on the torus $|w_1| = \cdots = |w_r| = \varepsilon'$ with $\varepsilon' < \varepsilon/r$, this derivative is at most $A\varepsilon'^{-r}$ in absolute value. Letting $\varepsilon' \uparrow \varepsilon/r$ and using $r^r \leq e^r r!$, we obtain the claim. $\square$

4.2 Projection increments and their polynomial approximations

We bisect $[0, 1]^d$ successively in coordinates $1, \dots, d$, and we repeat this cycle. At level j , the partition $\mathcal{C}_j$ has 2^j equal-volume rectangles, each of aspect ratio at most two, and each cell has two children at the next level. Let $\mathbb{V}_j$ be the space of functions whose restriction to each cell of $\mathcal{C}_j$ is a polynomial of total degree at most ℓ , and let Π_j be the orthogonal projection onto $\mathbb{V}_j$ in $L^2(g)$. These spaces are nested. In dimension one, they are the spaces of piecewise polynomials on dyadic intervals that underlie the multiwavelet bases of Alpert [1, Section 1.1]. Haar projections of this kind were combined with U-statistics to estimate integral functionals of a density in [27]. We define $\psi_j = \langle \Pi_j a, \Pi_j b \rangle_g$, where $\langle f_1, f_2 \rangle_g = \int f_1 f_2 g$. By orthogonality,

$$\psi - \psi_j = \langle a - \Pi_j a, b - \Pi_j b \rangle_g, \quad |\psi - \psi_j| \leq C2^{-j\theta}. \quad (4.6)$$

Indeed, the cellwise Taylor polynomials of degrees $\lceil \alpha \rceil - 1$ and $\lceil \beta \rceil - 1$ belong to $\mathbb{V}_j$ and have supremum errors $C2^{-j\alpha/d}$ and $C2^{-j\beta/d}$. The orthogonal projection minimizes the $L^2(g)$ error, and the Cauchy–Schwarz inequality bounds the inner product by g_+ times the product of the two supremum errors. We note that the identity in (4.6) has the form of the truncation bias of the HOIF estimators, which is the expectation of a product of two projection residuals [48, Theorems 3.7–3.8].

The decomposition $\psi_J = \psi_0 + \sum_{j=1}^J (\psi_j - \psi_{j-1})$ allows us to use different approximation orders at different resolutions. Each increment is a sum over parent cells of terms in integrals of g , ag , and bg against polynomials, since the projections act cell by cell.

Lemma 4.3 (Polynomials for projection increments). *Fix a level $j \geq 1$, and set $K = 2^{j-1}$ and $h = K^{-1/d}$. For each parent cell $C \in \mathcal{C}_{j-1}$ there are a known observable vector $Z_C(O)$, its moment vector $\mathbf{m}_C = \mathbb{E}_P Z_C(O)$, and a function F_C such that*

$$\psi_j - \psi_{j-1} = K^{-1} \sum_{C \in \mathcal{C}_{j-1}} F_C(\mathbf{m}_C), \quad |F_C(\mathbf{m}_C)| \leq C_0 h^{\alpha+\beta},$$

and

$$\|Z_C(O)\|_2 \leq C_0 K 1_C(X), \quad P(X \in C) \leq C_0/K. \quad (4.7)$$

For every integer $m \geq 0$, there is a polynomial $F_{C,m}$ in these moments of total degree at most $m + \nu + 2$ such that

$$|F_C(\mathbf{m}_C) - F_{C,m}(\mathbf{m}_C)| \leq C_0 h^{\alpha+\beta} (m+1)^\nu \rho^m. \quad (4.8)$$

At every population moment vector, that is, at $\mathbf{m}_C = \mathbf{m}_C(P)$ for every $P \in \mathcal{P}(\alpha, \beta)$, it satisfies

$$|F_{C,m}(\mathbf{t})| \leq C_0 \quad \text{for complex } \mathbf{t} \text{ with } \|\mathbf{t} - \mathbf{m}_C\|_\infty < 1/C_0, \quad (4.9)$$

$$\|\nabla F_{C,m}(\mathbf{m}_C)\|_2 \leq C_0 h^{s_0}. \quad (4.10)$$

The functions Z_C and the coefficients of $F_{C,m}$ depend only on the class parameters, the cell, and m . The dimension of Z_C and the constant $C_0 \geq 1$ do not depend on j , m , C , or P . For the base term there are a known observable vector $Z_0(O)$ and, for every $m \geq 0$, a polynomial $F_{0,m}$ of degree at most $m + 2$ in $\mathbf{m}_0 = \mathbb{E}_P Z_0(O)$ with $|\psi_0 - F_{0,m}(\mathbf{m}_0)| \leq C_0 \rho^m$, for which (4.7), (4.9), and (4.10) hold with $K = h = 1$.

Proof. Bases. For an increment, let $C \in \mathcal{C}_{j-1}$ be a parent cell, of volume $1/K$ and diameter at most $2\sqrt{d}h$. Its *child space* consists of the functions on C whose restriction to each child is a polynomial of total degree at most ℓ . For $\eta = a, b$, its η -*parent space* consists of the polynomials on C of total degree at most $\lceil \alpha \rceil - 1$ for $\eta = a$ and ℓ for $\eta = b$, and has dimension $p_a = q_\alpha$ and $p_b = q_\beta$, respectively. Since $\alpha \leq \beta$, the a -parent space is contained in the b -parent space, which is contained in the child space. For the base term, we let $C = [0, 1]^d$ and $K = h = 1$, we let the child space consist of the polynomials of total degree at most ℓ , and we let both parent spaces be $\{0\}$, so that $p_a = p_b = 0$.

Let z_C be a basis of the child space that is orthonormal for $K dx$ on C , and for $\eta = a, b$, let $L_\eta^\top z_C$ be an orthonormal basis of the η -parent space for $K dx$ on C . Then L_η has p_η columns, $L_\eta^\top L_\eta = I$, and the column space of L_a is contained in that of L_b . Each parent cell is the image of one of the d reference rectangles $[0, \frac{1}{2}]^r \times [0, 1]^{d-r}$, $0 \leq r < d$, under a dilation followed by a translation, and it is split in coordinate $r + 1$. We fix the bases on the reference rectangles and on $[0, 1]^d$, and we transport them by these affine maps. The maps preserve polynomial degrees and the normalized Lebesgue measures on the cell and on its children. In particular, the functions in z_C are bounded uniformly over levels and cells, and L_a, L_b take finitely many values.

Moments. Suppressing the cell index, we define

$$\begin{aligned} \Omega &= K \mathbb{E}_P [1_C(X) D z_C(X) z_C(X)^\top], & \Gamma_\eta &= L_\eta^\top \Omega L_\eta, \\ \mathbf{u} &= K \mathbb{E}_P [1_C(X) U z_C(X)], & \mathbf{v} &= K \mathbb{E}_P [1_C(X) V z_C(X)]. \end{aligned} \quad (4.11)$$

They equal $K \int_C g z_C z_C^\top$, $K \int_C a g z_C$, and $K \int_C b g z_C$. Since z_C is orthonormal for $K dx$ on C and $g_- \leq g \leq g_+$, we have $g_- I \preceq \Omega \preceq g_+ I$, and since $L_\eta^\top L_\eta = I$, the same bounds hold for Γ_η if $p_\eta \geq 1$. By Bessel's inequality, for every bounded function f on C ,

$$\left\| K \int_C f g z_C \right\|_2 \leq \left(K \int_C f^2 g^2 \right)^{1/2} \leq g_+ \sup_C |f|. \quad (4.12)$$

With $f = a, b$, this gives $\|u\|_2, \|v\|_2 \leq H g_+$. Let m_C collect the upper-triangular entries of Ω and the entries of u and v . Then $m_C = \mathbb{E}_P Z_C(O)$, where $Z_C(O)$ collects the corresponding entries of $K 1_C(X) D z_C(X) z_C(X)^\top$, $K 1_C(X) U z_C(X)$, and $K 1_C(X) V z_C(X)$. Since the bases and the responses are bounded, $\|Z_C(O)\|_2 \leq C_0 K 1_C(X)$, and $p = g/w \leq g_+/\delta$ gives $P(X \in C) \leq C_0/K$.

The increment. We set $B_\eta = L_\eta \Gamma_\eta^{-1} L_\eta^\top$ and $N_\eta = \Omega^{-1} - B_\eta$, with $B_\eta = 0$ and $\det \Gamma_\eta = 1$ if the η -parent space is $\{0\}$. For the inner product $\langle f_1, f_2 \rangle_{g,C} = K \int_C f_1 f_2 g$, the Gram matrices of the bases z_C and $L_\eta^\top z_C$ are Ω and Γ_η , and the inner products of z_C with a and b are u and v . By the normal equations, the projections of a onto the child space and onto the η -parent space therefore have the coefficient vectors $\Omega^{-1}u$ and $B_\eta u$ in the basis z_C . Thus $N_\eta u$ represents $E_\eta a$, where E_η projects orthogonally onto the complement of the η -parent space in the child space for $\langle \cdot, \cdot \rangle_{g,C}$, and similarly for b with v . Since the ranges of Π_j and Π_{j-1} on C are the child space and the b -parent space, K times the contribution of C to $\psi_j - \psi_{j-1}$ is

$$F_C = u^\top \Omega^{-1} v - u^\top B_\eta v = u^\top N_\eta v.$$

For the base term, the same formula gives $F_C = u^\top \Omega^{-1} v = \psi_0$. As in the projection kernels of the HOIF estimators [48, p. 357], we represent a projection through the inverse of a Gram matrix; here the Gram matrices are local, one for each parent cell.

Two small polynomial factors. Both terms of F_C are of order one. For the inner product $x^\top \Omega y$ on coefficient vectors, $B_\eta \Omega$ is the orthogonal projection onto the column space of L_η , and $N_\eta \Omega = I - B_\eta \Omega$ is the complementary projection. Since the column space of L_a is contained in that of L_b , we have $(N_a \Omega)(N_b \Omega) = N_b \Omega$, that is, $N_a \Omega N_b = N_b$. We define

$$R_\eta = (\det \Gamma_\eta) I - \Omega L_\eta \text{adj}(\Gamma_\eta) L_\eta^\top = (\det \Gamma_\eta) \Omega N_\eta,$$

where $\text{adj}(\Gamma_\eta) = (\det \Gamma_\eta) \Gamma_\eta^{-1}$ is the adjugate matrix, with $R_\eta = I$ if the η -parent space is $\{0\}$, and

$$r_a = R_a u, \quad r_b = R_b v, \quad S = \frac{\Omega^{-1}}{\det \Gamma_a \det \Gamma_b}. \quad (4.13)$$

Since N_a and Ω are symmetric, $R_a^\top S R_b = N_a \Omega \Omega^{-1} \Omega N_b = N_b$, and therefore $F_C = r_a^\top S r_b$. Since $\det \Gamma_\eta$ and the entries of $\text{adj}(\Gamma_\eta)$ are polynomials of degrees p_η and $p_\eta - 1$ in the entries of Ω , the vectors r_a and r_b are polynomials in m_C of degrees at most $p_a + 1$ and $p_b + 1$.

To see that r_a is small, let $\tilde{a}$ be the Taylor polynomial of a of degree $\lceil \alpha \rceil - 1$ at a point of C , so that $|a - \tilde{a}| \leq Ch^\alpha$ on C ; for the base term, we let $\tilde{a} = 0$, and then $|a - \tilde{a}| \leq H$. Since $\tilde{a}$ lies in the a -parent space, $E_a \tilde{a} = 0$. We write $\|x\|_\Omega^2 = x^\top \Omega x$. Since $\Omega^2 \preceq g_+ \Omega$, we have $\|\Omega x\|_2 \leq g_+^{1/2} \|x\|_\Omega$, and since $\det \Gamma_a \leq g_+^{p_a}$ and E_a is a contraction,

$$\|r_a\|_2 \leq g_+^{p_a+1/2} \|N_a u\|_\Omega = g_+^{p_a+1/2} \|E_a(a - \tilde{a})\|_{g,C} \leq g_+^{p_a+1} \sup_C |a - \tilde{a}| \leq Ch^\alpha,$$

where $\|f\|_{g,C}^2 = \langle f, f \rangle_{g,C}$. In the same way, with β and the b -parent space, $\|r_b\|_2 \leq Ch^\beta$. Since $\|S\|_{\text{op}} \leq g_-^{-(p_a+p_b+1)}$, we conclude that $|F_C| \leq Ch^{\alpha+\beta}$. Since we use the a -parent space, S contains $1/\det \Gamma_a$ in place of a second copy of $1/\det \Gamma_b$, and this keeps the power of $m+1$ below at $v = q_\alpha + q_\beta$.

Polynomial approximation. We apply Lemma 4.2(a) with $N = 2$ and $(M_0, M_1, M_2) = (\Omega, \Gamma_a, \Gamma_b)$, so that $v' = p_a + p_b$, and we define

$$F_{C,m}(m_C) = r_a^\top S_m(\Omega, L_a^\top \Omega L_a, L_b^\top \Omega L_b) r_b. \quad (4.14)$$

The substitutions $\Gamma_\eta = L_\eta^\top \Omega L_\eta$ are linear, and the total degree of $F_{C,m}$ is therefore at most $m + p_a + p_b + 2$. Since $F_C - F_{C,m} = r_a^\top (S - S_m) r_b$, the bounds on r_a, r_b and (4.5) imply that $|F_C - F_{C,m}| \leq Ch^{\alpha+\beta} (m+1)^{p_a+p_b} \rho^m$. For an increment, $p_a + p_b = v$, which proves (4.8); for the base term, $p_a = p_b = 0$, which gives the degree $m+2$ and the error $C\rho^m$.

The complex bound and the gradient. For each of the finitely many choices of (L_a, L_b) , and for the base term, let $\mathcal{V} = \{t : g_- I \preceq \Omega \preceq g_+ I, \|u\|_2, \|v\|_2 \leq Hg_+\}$, where Ω, u, v are read off from t as from m_C . This bounded set contains $m_C(P)$ for every level, cell, and $P \in \mathcal{P}(\alpha, \beta)$, and the affine maps $\Omega, L_a^\top \Omega L_a$, and $L_b^\top \Omega L_b$ are real symmetric with spectra in $[g_-, g_+]$ on it. By (4.4), $F_{C,m}$ and the entries of S_m are the truncations at order m in z of

$$\bar{g}^{-(v'+1)} r_a^\top \mathcal{K}(z; \Omega) r_b \det \mathcal{K}(z; L_a^\top \Omega L_a) \det \mathcal{K}(z; L_b^\top \Omega L_b)$$

and of the entries of the corresponding matrix, in which each determinant is a fixed sum of products of kernel entries, or one, and the entries of r_a, r_b are fixed polynomials in t . By finitely many applications of Lemma 4.2(b), there exist $\varepsilon \in (0, 1]$ and C , uniform over levels and cells, such that $F_{C,m}$ and the entries of S_m are bounded by C on the complex ε -neighborhood of $\mathcal{V}$ for every m . This proves (4.9). By Lemma 4.2(c) with $r = 1$, the entries of S_m have bounded gradients on $\mathcal{V}$, and so do the fixed polynomials r_a, r_b . By the product rule, at $m_C(P)$ we therefore have

$$\|\nabla F_{C,m}\|_2 \leq C(\|r_a\|_2 + \|r_b\|_2 + \|r_a\|_2 \|r_b\|_2) \leq Ch^{s_0},$$

which is (4.10). Let $C_0 \geq 1$ be at least as large as all constants above and the inverses of the radii of the complex neighborhoods. $\square$

For $0 < \alpha, \beta \leq 1$, each projection is a g -weighted cell mean. Let C_1, C_2 be the children of C , and let $x_i = 2K \int_{C_i} g$, $u_i = 2K \int_{C_i} ag$, $v_i = 2K \int_{C_i} bg$, and $\bar{x} = (x_1 + x_2)/2$. With $z_C = (\sqrt{2}1_{C_1}, \sqrt{2}1_{C_2})^\top$ and $L_a = L_b = (1, 1)^\top / \sqrt{2}$, we have $\Omega = \text{diag}(x_1, x_2)$, $u = (u_1, u_2)^\top / \sqrt{2}$, $\Gamma_a = \Gamma_b = \bar{x}$, and $r_a = A_C(1, -1)^\top / (2\sqrt{2})$, so that

$$F_C = \frac{A_C B_C}{4x_1 x_2 \bar{x}}, \quad A_C = x_2 u_1 - x_1 u_2, \quad B_C = x_2 v_1 - x_1 v_2. \quad (4.15)$$

The contrasts $A_C = x_1 x_2 (u_1/x_1 - u_2/x_2)$ and B_C multiply the two normalized masses by the differences of the g -weighted child means of a and b , respectively.

4.3 Unbiased estimation and variance

Let $O_1, \dots, O_n$ be the observations, let Z be a bounded measurable map from the observation space to $\mathbb{R}^p$, and let $m = \mathbb{E}_P Z(O)$. A real polynomial F on $\mathbb{R}^p$ of degree at most R satisfies $F(m) = \sum_{q=0}^R D^q F(0)[m, \dots, m]/q!$. For $0 \leq q \leq n$, we set $(n)_q = n(n-1) \cdots (n-q+1)$, with $(n)_0 = 1$. If $R \leq n$, we define the *lift* of F by

$$\hat{F} = \sum_{q=0}^R \frac{1}{q! (n)_{q_1, \dots, i_q \leq n}} \sum^{\neq} D^q F(0)[Z(O_{i_1}), \dots, Z(O_{i_q})], \quad (4.16)$$

where $\sum^{\neq}$ runs over ordered q -tuples of pairwise distinct indices in $\{1, \dots, n\}$, and the term $q=0$ is $F(0)$. The lift replaces each monomial by a U-statistic [17, 19], is linear in F , and uses only its coefficients, Z , and the observations.

Lemma 4.4 (Mean and covariance of the lift). *Let F_1, F_2 be real polynomials on $\mathbb{R}^p$ of degree at most n , and let $Y_i = Z(O'_i) - m$, where $O'_1, O'_2, \dots$ are i.i.d. with law P . Then $\mathbb{E} \hat{F}_1 = F_1(m)$ and*

$$\text{Cov}(\hat{F}_1, \hat{F}_2) = \sum_{r \geq 1} \frac{1}{r! (n)_r} \mathbb{E} \left[D^r F_1(m)[Y_1, \dots, Y_r] D^r F_2(m)[Y_1, \dots, Y_r] \right],$$

where the sum stops at the smaller degree.

Proof. Let $Y_i = Z(O_i) - m$, and let F be F_1 or F_2 . We substitute $Z(O_i) = m + Y_i$ in every argument in (4.16) and expand by multilinearity. An ordered r -tuple $(k_1, \dots, k_r)$ of pairwise distinct indices then arises from $\binom{q}{r} (n-r)_{q-r}$ terms of order q , and by the symmetry of $D^q F(0)$, each of them equals $D^q F(0)[Y_{k_1}, \dots, Y_{k_r}, m, \dots, m]$. Since $\binom{q}{r} (n-r)_{q-r} / \{q! (n)_q\} = 1 / \{r! (q-r)! (n)_r\}$, and since $D^r F(m)[y_1, \dots, y_r] = \sum_{q \geq r} D^q F(0)[y_1, \dots, y_r, m, \dots, m] / (q-r)!$ by Taylor's formula, we obtain

$$\hat{F} = F(m) + \sum_{r=1}^{\deg F} \frac{1}{r! (n)_{k_1, \dots, k_r \leq n}} \sum^{\neq} D^r F(m)[Y_{k_1}, \dots, Y_{k_r}].$$

The Y_i are independent with mean zero, and each kernel is linear in each of its arguments. If an index appears in one of two tuples but not in the other, then integrating first over that observation shows that the product of the two corresponding terms has mean zero. In particular, every term with $r \geq 1$ has mean zero, so that $\mathbb{E}\hat{F}_1 = F_1(\mathbf{m})$. The surviving pairs have the same index set and length r , and there are $(n)_r r!$ such pairs. By symmetry, each contributes $\mathbb{E}[D^r F_1(\mathbf{m})[Y_1, \dots, Y_r] D^r F_2(\mathbf{m})[Y_1, \dots, Y_r]] / \{r!(n)_r\}^2$, which gives the covariance identity. $\square$

The expansion in this proof is the Hoeffding decomposition of the U-statistics in (4.16) into degenerate U-statistics; see [67, Sections 11.4 and 12.3] and [58, Section 5.1.5, Lemma A]. Hoeffding [19, Sections 5–6] derived exact variance and covariance formulas by grouping pairs of index sets according to their overlap. A covariance identity of the same form as that of Lemma 4.4, for linear combinations of complete U-statistics estimating trace polynomials, appears in [60, Theorem 6.2].

For a multilinear form H , the kernel $H[Y_1, \dots, Y_r]$ is obtained from $H[Z(O'_1), \dots, Z(O'_r)]$ by centering each argument, that is, by applying $\prod_{l=1}^r (I - \mathcal{E}_l)$, where $\mathcal{E}_l$ integrates the l th observation against P . These commuting orthogonal projections in $L^2(P^{\otimes r})$ have an orthogonal projection as their product, which implies that

$$\mathbb{E}|H[Y_1, \dots, Y_r]|^2 \leq \mathbb{E}|H[Z(O'_1), \dots, Z(O'_r)]|^2. \quad (4.17)$$

This allows us to bound the variance using the cell support of Z ; the centered variables $Z_C(O) - \mathbf{m}_C$ themselves are not supported in the cell C . The centered kernel $H[Y_1, \dots, Y_r]$ is the component of highest order in the Hoeffding decomposition of $H[Z(O'_1), \dots, Z(O'_r)]$ [67, Section 11.4], and (4.17) is the corresponding contraction property of that decomposition.

Lemma 4.5 (Variance of a polynomial in cell moments). *Fix integers $n, K, R \geq 1$ with $2R \leq n$, a number $h \in (0, 1]$, and a partition of $[0, 1]^d$ into K cells. For every cell C , let Z_C be a bounded observable vector with $\mathbf{m}_C = \mathbb{E}_P Z_C(O)$, and let F_C be a real polynomial of degree at most R . Let $\hat{T}$ be the lift of $T = K^{-1} \sum_C F_C(\mathbf{m}_C)$, regarded as a polynomial in the moment vector $(\mathbf{m}_C)_C$ with $Z = (Z_C)_C$. Suppose that, for some $C_0 \geq 1$, $s_0 > 0$, and every cell,*

$$\|Z_C(O)\|_2 \leq C_0 K 1_C(X), \quad P(X \in C) \leq C_0/K,$$

that $|F_C(\mathbf{t})| \leq C_0$ for complex $\mathbf{t}$ with $\|\mathbf{t} - \mathbf{m}_C\|_\infty < 1/C_0$, and that $\|\nabla F_C(\mathbf{m}_C)\|_2 \leq C_0 h^{s_0}$. Then, with $C_v = 2e^2 C_0^7$,

$$\text{Var } \hat{T} \leq \frac{C_v h^{2s_0}}{n} + \frac{1}{n} \sum_{r=2}^R C_v^r r! \left(\frac{K}{n}\right)^{r-1}. \quad (4.18)$$

If moreover $C_v R K / n \leq 1/2$, then

$$\text{Var } \hat{T} \leq \frac{C_v h^{2s_0}}{n} + \frac{4C_v^2 K}{n^2}. \quad (4.19)$$

Proof. We write $\mathbf{m} = (\mathbf{m}_C)_C$ and $H_r = D^r T(\mathbf{m})[Z(O'_1), \dots, Z(O'_r)]$. By Lemma 4.4 and (4.17), the term of order r in the variance of $\hat{T}$ is at most $\mathbb{E}|H_r|^2 / \{r!(n)_r\}$, and the terms with $r > R$ vanish. For observations $o_l = (x_l, \zeta_l)$, we have

$$D^r T(\mathbf{m})[Z(o_1), \dots, Z(o_r)] = K^{-1} \sum_C D^r F_C(\mathbf{m}_C)[Z_C(o_1), \dots, Z_C(o_r)].$$

Since the cells are disjoint and Z_C is supported on C , the summand for C vanishes unless $x_1, \dots, x_r \in C$, and at most one summand is nonzero. Thus

$$|D^r T(\mathbf{m})[Z(o_1), \dots, Z(o_r)]|^2 = K^{-2} \sum_C |D^r F_C(\mathbf{m}_C)[Z_C(o_1), \dots, Z_C(o_r)]|^2.$$

For $r \geq 2$, by Lemma 4.2(c) with $\varepsilon = 1/C_0$ and $A = C_0$, together with $\|Z_C\|_\infty \leq \|Z_C\|_2 \leq C_0 K$, the summand for C is at most $C_0^2 (r!)^2 (eC_0)^{2r} (C_0 K)^{2r}$ on the event that $X'_1, \dots, X'_r \in C$, which has probability at most $(C_0/K)^r$. Summing over the K cells, we obtain

$$\mathbb{E}|H_r|^2 \leq C_0^2 (r!)^2 (e^2 C_0^5)^r K^{r-1} \leq (r!)^2 (e^2 C_0^7)^r K^{r-1}.$$

For $r = 1$, by the gradient bound, we have $|\nabla F_C(\mathbf{m}_C) \cdot Z_C(o)|^2 \leq C_0^4 h^{2s_0} K^2 1_C(x)$, and since the cells partition $[0, 1]^d$, we obtain $\mathbb{E}|H_1|^2 \leq C_0^4 h^{2s_0} \leq C_v h^{2s_0}$. Since $2R \leq n$, we have $(n)_r \geq (n/2)^r$ for $r \leq R$, and these bounds imply (4.18). In (4.18), the ratio of the $(r+1)$ st to the r th term is $C_v(r+1)K/n \leq C_v RK/n$ for $r+1 \leq R$. If this is at most $1/2$, the sum over $r \geq 2$ is at most twice its first term $2C_v^2 K/n^2$, which proves (4.19). $\square$

For a fixed order r , variances of order $n^{-1} \max\{1, (K/n)^{r-1}\}$, with K the number of basis functions, were obtained for the terms of the HOIF estimators in [48, Theorems 3.20–3.21]. In (4.18), we keep the dependence on the order explicit, including the factorial, since our orders grow with n .

4.4 Choice of the degrees

For a level $j \geq 1$ and an order $m \geq 0$, we set $K_j = 2^{j-1}$, the number of parent cells, and $h_j = K_j^{-1/d}$, and we let $T_{j,m}(P) = K_j^{-1} \sum_{C \in \mathcal{C}_{j-1}} F_{C,m}(\mathbf{m}_C)$ be the order- m polynomial approximation of $\psi_j - \psi_{j-1}$. For the base term, we let $K_0 = h_0 = 1$ and $T_{0,m}(P) = F_{0,m}(\mathbf{m}_0)$. In particular, $K_0 = K_1 = 1$. We write $\hat{T}_{j,m}$ for the lift of $T_{j,m}$ and $C_v = 2e^2 C_0^7$, with C_0 from Lemma 4.3. By that lemma, if $R = m + v + 2$ satisfies $2R \leq n$, then the hypotheses of Lemma 4.5 hold for $T_{j,m}$ with $K = K_j$, $h = h_j$, and this R , including the base term. Since $r! \leq R^{r-1}$ for $2 \leq r \leq R$, the bound (4.18) then implies

$$\text{Var } \hat{T}_{j,m} \leq \frac{C_v}{n} \left(h_j^{2s_0} + \sum_{r=1}^{R-1} y^r \right), \quad y = \frac{C_v R K_j}{n}. \quad (4.20)$$

For a terminal level J and orders $m_0, \dots, m_J$ with $m_j + \nu + 2 \leq n/2$, the estimator of $T(P) = \int abg$ is $\hat{T} = \sum_{j=0}^J \hat{T}_{j,m_j}$, the lift of $\sum_{j=0}^J T_{j,m_j}$. Its mean is $\sum_{j=0}^J T_{j,m_j}(P)$ by Lemma 4.4. By the telescoping decomposition of ψ_J , (4.6), (4.8), the base-term bound, and $h_j^{\alpha+\beta} = K_j^{-\theta}$, we obtain

$$|\mathbb{E}\hat{T} - T(P)| \leq C2^{-J\theta} + C_0 \sum_{j=0}^J f_j(m_j), \quad f_j(m) = K_j^{-\theta} (m+1)^\nu \rho^m. \quad (4.21)$$

By the L^2 triangle inequality, the standard deviation of $\hat{T}$ is at most the sum of the standard deviations of the $\hat{T}_{j,m_j}$.

We choose the order on each level as the smaller of two values. The first, a bias rule, makes the approximation errors decrease geometrically from the terminal level toward the coarse levels. The second, a variance rule, is needed only below the threshold. We keep high orders only at coarse resolutions, and in this respect our estimator resembles that of Robins et al. [52, Section 9, (9.6)]. Their kernel of each order three or higher is a sum over the contiguous pairs of projection kernels in its product; in each summand, that pair is truncated below a hyperbola in the indices of the basis functions, and all other kernels are truncated at n basis functions.

On a level with K parent cells and degree R , the bound (4.20) is at most about $n^{-2\theta}$ if $y^R \leq n^{1-2\theta}$. If $x = \log y > 0$, this limits the degree to about $(1 - 2\theta) \log n / x$. Since $K = ne^x / (C_\nu R)$, the level error $K^{-\theta} \rho^m$ at this degree is about $n^{-\theta} (C_\nu R)^\theta \exp\{-\theta x - \tau(1 - 2\theta) \log n / x\}$. The minimum of $\theta x + \tau(1 - 2\theta) \log n / x$ over $x > 0$ is $\kappa \sqrt{\log n}$, and the relevant levels have R of order $\sqrt{\log n}$. The three factors of $(\log n)^{\theta/2+\nu/2+1/4}$ in $a_n(\log n)^{\nu/2+1/4} / (n^{-\theta} e^{-\kappa \sqrt{\log n}})$ come from $(C_\nu R)^\theta$, which reflects the factorials in the variance, from the factor $(m+1)^\nu$ in (4.8), and from the sum over levels. The next lemma bounds this sum. It is the only source of the factor $(\log n)^{1/4}$ in the upper bound.

Lemma 4.6 (Window sum). *Let $\theta, \tau, B, \bar{X} > 0$, and let $S \subset (0, \bar{X}]$ be a finite set whose points are pairwise at distance at least $\log 2$. Then*

$$\sum_{x \in S} \exp\left(-\theta x - \frac{\tau B}{x}\right) \leq \left(2 + \frac{(\pi \bar{X} / \theta)^{1/2}}{\log 2}\right) e^{-2\sqrt{\theta \tau B}}.$$

Proof. Let $\xi = \sqrt{\tau B / \theta}$. Completing the square, we obtain

$$\theta x + \frac{\tau B}{x} = 2\sqrt{\theta \tau B} + \frac{\theta(x - \xi)^2}{x} \geq 2\sqrt{\theta \tau B} + \frac{\theta(x - \xi)^2}{\bar{X}}, \quad 0 < x \leq \bar{X}.$$

We list the points of S in $[\xi, \infty)$ in increasing order and those in $(0, \xi)$ in decreasing order. The k th point of either list, counted from $k = 0$, satisfies $|x - \xi| \geq k \log 2$. With $c = \theta(\log 2)^2 / \bar{X}$, the sum is therefore at most $2e^{-2\sqrt{\theta \tau B}} \sum_{k \geq 0} e^{-ck^2}$, and $\sum_{k \geq 0} e^{-ck^2} \leq 1 + \int_0^\infty e^{-cu^2} du = 1 + \frac{1}{2}(\pi/c)^{1/2}$. $\square$

The construction below also gives estimators satisfying, for every integer $n \geq 3$,

$$\sup_{P \in \mathcal{P}(\alpha, \beta)} \|\hat{\psi}_n - \psi(P)\|_{L^2(P^n)} \leq C\{n^{-1/2} + n^{-\theta}\}. \quad (4.22)$$

Moreover, fix $0 \leq M_W < \infty$ and a compact set $\mathcal{G} \subset \{(x, y) : 0 < x < y < \infty\}$. With the remaining class parameters fixed, the bounds hold with one constant C for all $(g_-, g_+) \in \mathcal{G}$ and all admissible observables with $\|W\|_\infty \leq M_W$, using each pair's own ρ, τ, κ and a_n where applicable. The tuning constants can be chosen common. The same assertions hold when the observables depend on X .

Proof of Theorem 2.3(a). We write $L = \log n$, and we let $\bar{r}_n = n^{-1/2}$ if $\theta \geq 1/2$ and $\bar{r}_n = a_n L^{\nu/2+1/4} = n^{-\theta} L^{\theta/2+\nu/2+1/4} e^{-\kappa\sqrt{L}}$ if $0 < \theta < 1/2$. In both cases $n^{-1/2} \leq C\bar{r}_n$; in the second, $n^{-1/2}/\bar{r}_n \leq e^{-(1/2-\theta)L+\kappa\sqrt{L}}$ is bounded. By the reduction at the start of this section, it therefore suffices to construct estimators $\hat{T}$ of $T(P) = \int abg$ such that $\mathbb{E}(\hat{T} - T(P))^2 \leq C\bar{r}_n^2$ for every $P \in \mathcal{P}(\alpha, \beta)$. Below, $n_1 \geq 3$ is an integer, depending only on the parameters listed in Theorem 2.3(a), that is large enough for every condition that we impose for $n \geq n_1$. For $3 \leq n < n_1$ we use the estimator 0, whose error $|T(P)| \leq H^2 g_+$ is at most a constant multiple of $\bar{r}_n$. Let $n \geq n_1$.

The two rules. We fix $A > 0$ such that $\tau A \geq \theta \log 2 + 1$. We write $K_* = 2^J$ for the terminal resolution and $l_j = J - j + 1$ for $0 \leq j \leq J$. Then $K_j = K_* 2^{-l_j}$ for $j \geq 1$ and $K_0 = 1 = 2K_* 2^{-l_0}$, so that $K_* 2^{-l_j} \leq K_j \leq 2K_* 2^{-l_j}$ for every j . The *bias rule* is the order $\lceil Al_j \rceil$. Since $(\lceil Al_j \rceil + 1)^\nu \leq (Al_j + 2)^\nu$ and $\rho^{\lceil Al_j \rceil} \leq e^{-\tau Al_j}$, we have

$$\sum_{j=0}^J f_j(\lceil Al_j \rceil) \leq K_*^{-\theta} \sum_{l \geq 1} (Al + 2)^\nu e^{-l} = C_A K_*^{-\theta}, \quad (4.23)$$

because the l_j are distinct positive integers and $\tau A - \theta \log 2 \geq 1$. In each case below, the order at level j is $m_j = \min\{\lceil Al_j \rceil, M_j\}$ for a *variance rule* $M_j \in \{0, 1, 2, \dots\} \cup \{\infty\}$, and we call level j *capped* if $M_j < \lceil Al_j \rceil$. Then (4.21), (4.23), and $2^{-J\theta} = K_*^{-\theta}$ imply that

$$|\mathbb{E}\hat{T} - T(P)| \leq CK_*^{-\theta} + C \sum_{j \text{ capped}} f_j(M_j). \quad (4.24)$$

The degree $R_j = m_j + \nu + 2$ satisfies $R_j \leq Al_j + \nu + 3$. In both cases below $J \leq CL$, so that $R_j \leq CL \leq n/2$, and (4.20) applies to level j with $y = y_j = C_\nu R_j K_j / n$.

The case $\theta \geq 1/2$. Let $c_K \in (0, 1]$ satisfy $c_K \leq \{4C_\nu \sup_{l \geq 1} (Al + \nu + 3)2^{-l}\}^{-1}$, let $J = \lfloor \log_2(c_K n) \rfloor \geq 0$, and let $M_j = \infty$ for all j . Since $K_* \leq c_K n$, we have $C_\nu R_j K_j / n \leq 2C_\nu c_K (Al_j + \nu + 3)2^{-l_j} \leq 1/2$, and (4.19) gives $\text{Var} \hat{T}_{j, m_j} \leq C_\nu h_j^{2s_0} / n + 4C_\nu^2 K_j / n^2$. Since $\sum_{j \geq 0} h_j^{s_0} \leq 1 + \sum_{i \geq 0} 2^{-is_0/d}$ and $\sum_{j=0}^J K_j^{1/2} \leq CK_*^{1/2} \leq Cn^{1/2}$, the standard deviation of $\hat{T}$ is at most $Cn^{-1/2}$. Since no level is capped, (4.24) implies that the bias is at most $CK_*^{-\theta} \leq C(c_K n/2)^{-\theta} \leq Cn^{-1/2}$.

The case $0 < \theta < 1/2$. Let $\Delta = 1 - 2\theta$, so that $\kappa = 2\sqrt{\theta\Delta\tau}$. We take

$$J = \left\lceil \log_2 n + \frac{\kappa\sqrt{L}}{\theta \log 2} \right\rceil, \quad \text{so that} \quad K_*^{-\theta} \leq n^{-\theta} e^{-\kappa\sqrt{L}} \leq \bar{r}_n, \quad \log \frac{K_*}{n} \leq \frac{\kappa}{\theta} \sqrt{L} + \log 2.$$

We call level j *high* if $K_j > n/L^2$, and *low* otherwise. The levels 0 and 1 are low. On a high level, $2^{l_j} = K_*/K_j < K_*L^2/n$, so that

$$R_j \leq \bar{R} = A \log_2(K_*L^2/n) + \nu + 3 \leq C\sqrt{L},$$

and there are at most $\log_2(K_*L^2/n) \leq C\sqrt{L}$ high levels. We set $c_L = \log(C_v\bar{R})$, so that $0 < c_L \leq \frac{1}{2} \log L + C$, and

$$B = \Delta L - 2\kappa\sqrt{L}, \quad x_j = \log \frac{K_j}{n} + c_L, \quad \bar{X} = \log \frac{K_*}{n} + c_L \leq C\sqrt{L}.$$

We define the variance rule by

$$M_j = \lfloor B/x_j \rfloor - \nu - 2 \quad \text{if } x_j > 0, \quad M_j = \infty \quad \text{if } x_j \leq 0.$$

If $x_j > 0$, then $K_j > n/(C_v\bar{R}) \geq n/L^2$, so that level j is high, and $0 < x_j \leq \bar{X}$. Since $B \geq \Delta L/2$ and $B/\bar{X} \geq \nu + 3$, we then have $M_j \geq 1$.

Bias. By (4.24) and $K_*^{-\theta} \leq \bar{r}_n$, it remains to bound $f_j(M_j)$ on the capped levels. Since a capped level has $M_j < \infty$, it is high and has $x_j > 0$, and we have $M_j + 1 \leq \lceil Al_j \rceil \leq \bar{R}$, so that $(M_j + 1)^\nu \leq CL^{\nu/2}$. Moreover $\rho^{M_j} \leq e^{\tau(\nu+3)} e^{-\tau B/x_j}$ and $K_j^{-\theta} = n^{-\theta} e^{\theta c_L} e^{-\theta x_j} \leq Cn^{-\theta} L^{\theta/2} e^{-\theta x_j}$, since $e^{\theta c_L} = (C_v\bar{R})^\theta \leq CL^{\theta/2}$. Combining these bounds, we find

$$f_j(M_j) \leq Cn^{-\theta} L^{\theta/2+\nu/2} \exp\left(-\theta x_j - \frac{\tau B}{x_j}\right).$$

The points x_j of distinct capped levels lie in $(0, \bar{X}]$ and differ by nonzero multiples of $\log 2$. By Lemma 4.6 and $\bar{X} \leq C\sqrt{L}$, the sum of these bounds is at most $Cn^{-\theta} L^{\theta/2+\nu/2+1/4} e^{-2\sqrt{\theta\tau B}}$. Finally,

$$\kappa\sqrt{L} - 2\sqrt{\theta\tau B} = \frac{2\sqrt{\theta\tau}(\Delta L - B)}{\sqrt{\Delta L} + \sqrt{B}} \leq \frac{4\kappa\sqrt{\theta\tau}}{\sqrt{\Delta}} = 8\theta\tau,$$

so that the bias is at most $C\bar{r}_n$.

Variance. Since $h_j \leq 1$, (4.20) implies that $\text{Var } \hat{T}_{j,m_j} \leq C_v R_j n^{-1} \max\{1, y_j\}^{R_j-1}$. On a low level, we have $R_j \leq CL$ and $K_j \leq n/L^2$, so that $y_j \leq CL^{-1} \leq 1$, and the variance is at most CL/n . There are at most $J + 1 \leq CL$ low levels, and their standard deviations sum to at most $CL^{3/2} n^{-1/2}$. On a high level, we have $R_j \leq \bar{R}$, so that $y_j \leq C_v \bar{R} K_j/n = e^{x_j}$. If $x_j \leq 0$, then $y_j \leq 1$. If $x_j > 0$, then $R_j \leq M_j + \nu + 2 \leq B/x_j$, and therefore $y_j^{R_j-1} \leq e^{R_j x_j} \leq e^B$.

Since $e^B \geq 1$, in both cases the variance is at most $C_v \bar{R} n^{-1} e^B = C_v \bar{R} n^{-2\theta} e^{-2\kappa\sqrt{L}}$, and the standard deviations of the at most $C\sqrt{L}$ high levels sum to at most $CL^{3/4} n^{-\theta} e^{-\kappa\sqrt{L}}$. Since $\nu \geq 2$, this is at most $C\bar{r}_n$, and so is $CL^{3/2} n^{-1/2}$, since $L^{3/2} e^{-\Delta L/2 + \kappa\sqrt{L}}$ is bounded.

With λ and the observable mean restored as at the start of the section, this proves Theorem 2.3(a), and (4.22) follows because $L^{\theta/2 + \nu/2 + 1/4} e^{-\kappa\sqrt{L}}$ is bounded below the threshold. All constants depend only on the parameters listed in Theorem 2.3(a). The estimator is an explicit function of the covariates X_i and the values of D, U, V, W at the observations, and it involves no baseline law.

It remains to prove the uniformity assertions. For compact $\mathcal{G}$, g_- and d_g are bounded below, g_+ is bounded above, and ρ lies in a compact subinterval of $(0, 1)$. The approximation constants, the radii of the complex neighborhoods, and the polynomial bounds in Lemmas 4.2 and 4.3 can therefore be chosen uniformly, which gives common C_0 and C_v . We take A with $A \inf_{\mathcal{G}} \tau \geq \theta \log 2 + 1$, and then we choose c_K uniformly. Below the threshold, $\Delta > 0$ is fixed, and τ and κ have positive lower and finite upper bounds on $\mathcal{G}$, so that $\bar{R} \leq C\sqrt{L}$ and $\bar{X} \leq C\sqrt{L}$ uniformly. Then every large-sample condition, rate comparison, and finite-sample enlargement above holds beyond one common n_1 and with one constant C . The reduction at the start of the section costs at most $M_W n^{-1/2}$. Dependence of the observables on X changes none of the moment identities or observable bounds; equivalently, we may regard (X, Z) as the response. $\square$

Lower bounds

We prove Theorem 2.3(b) in this section and the next two. The root- n bound follows from a two-point argument (Section 5.1). Below the threshold, we use two fuzzy hypotheses [66, Section 2.7.4]: the prior mixtures are close in total variation, while their target means differ by more than the prior fluctuations. In a finite-response model, we bound the Hellinger distance between Poissonized mixtures (Lemma 5.2) and reduce the lower bound to coefficient and separation estimates (Proposition 5.3). Section 6 constructs priors giving $c a_n$ with block-constant phases and $c a_n \log n$ with lattice phases. Section 7.1 embeds them in the generic model; Section 7.2 establishes localization and completes the proof.

For the construction below the threshold, we follow the strategy of the lower bound of Aronow and Lopatto [3, Section 1.2] for estimating a constant conditional variance under rough design. There, the covariate density is random under both priors, approximate moment matching makes the local laws of the data under the two priors close simultaneously for every number $k \leq M$ of observations in the same local region, and the sample is Poissonized, with a reweighting of the priors that cancels the exponential likelihood factor arising from the random total mass of the perturbed density. A matching order $M \asymp \sqrt{\log n}$, balanced against regions that hold about $e^{-C(\log n)/M}$ observations on average, gives a factor $e^{-C\sqrt{\log n}}$. Here, too, a random density hides the difference between the priors from every configuration with fewer than about M observations in a pair of blocks, and we Poissonize the sample and take $M \asymp \sqrt{\log n}$ to balance the two losses. Relative to [3], the hiding and the separation take a new form. Our priors share the law of a random phase Θ of the density and differ only through sign correlations at frequency M in this phase. The two blocks of each pair oscillate with opposite signs, so that the mass of each pair is fixed and no reweighting is needed. The target sees the difference between the priors through the M th Fourier coefficient of the reciprocal density, which identifies the constant κ . The lattice phases of Section 6 gain a further factor $\log n$.

5.1 Divergences and a two-point bound

For probability laws $\mathbb{P}, \mathbb{Q}$, we write

$$\text{TV}(\mathbb{P}, \mathbb{Q}) = \sup_E |\mathbb{P}(E) - \mathbb{Q}(E)|, \quad d_H^2(\mathbb{P}, \mathbb{Q}) = \int (\sqrt{d\mathbb{P}} - \sqrt{d\mathbb{Q}})^2.$$

Any test between $\mathbb{P}$ and $\mathbb{Q}$ has error probabilities that sum to at least $1 - \text{TV}(\mathbb{P}, \mathbb{Q})$. We use the inequality $\text{TV} \leq d_H$ and the fact that the affinity $1 - d_H^2/2$ is multiplicative over product laws [66, Section 2.4 and Lemma 2.3]. Since $\prod_i (1 - x_i) \geq 1 - \sum_i x_i$ for $x_i \in [0, 1]$, the squared Hellinger distance between two product laws is at most the sum of the squared Hellinger distances between their factors. The next lemma applies the two-point method [66, Sections 2.3 and 2.4.2] along a one-parameter path.

Lemma 5.1 (A one-dimensional testing bound). *Suppose a class $\mathcal{P}$ contains probability laws P_t for t in an open interval J , with*

$$\frac{dP_t}{dP_*} = f_0 + tf_1, \quad f_0 + tf_1 \geq c_f > 0, \quad |f_1| \leq C_f \quad P_*\text{-almost surely},$$

where P_* is a probability law and c_f, C_f are independent of $t \in J$. If $t \mapsto T(P_t)$ is differentiable at an interior point t_0 and its derivative there is nonzero, then $r_n(T, \mathcal{P}) \geq cn^{-1/2}$ for all sufficiently large n . The constant $c > 0$ is independent of n .

Proof. Let $f_t = f_0 + tf_1$. Since $(\sqrt{f_t} - \sqrt{f_s})^2 = (t - s)^2 f_1^2 / (\sqrt{f_t} + \sqrt{f_s})^2$ and $f_t, f_s \geq c_f$, we have $d_H^2(P_t, P_s) \leq C_f^2 (t - s)^2 / (4c_f)$. Let $t_{\pm} = t_0 \pm a/\sqrt{n}$, where $a > 0$ is fixed and small enough that $C_f a / \sqrt{c_f} \leq 1/2$; for all large n , these points lie in J . By the subadditivity of d_H^2 over products, we have $d_H^2(P_{t_+}^{\otimes n}, P_{t_-}^{\otimes n}) \leq C_f^2 a^2 / c_f \leq 1/4$, so that the total variation distance between the product laws is at most $1/2$. Since the derivative of the target at t_0 is nonzero, the separation $D_n = |T(P_{t_+}) - T(P_{t_-})|$ is at least $c'n^{-1/2}$ for a constant $c' > 0$ and all sufficiently large n . We threshold any estimator at the midpoint of the two target values. The two error probabilities of the resulting test sum to at least $1/2$. At one of the laws, the estimator therefore has probability at least $1/4$ of an estimation error of at least $D_n/2$, and its RMSE is at least $D_n/4$. $\square$

5.2 Phase priors and the Poisson comparison

Throughout the lower-bound constructions, we fix $d \geq 1$, $\alpha, \beta > 0$ and $0 < r_- < 1 < r_+ < \infty$, and we write

$$\begin{aligned} a_* &= \alpha/d, & b_* &= \beta/d, & \theta &= a_* + b_* < \frac{1}{2}, \\ \rho &= \frac{\sqrt{r_+} - \sqrt{r_-}}{\sqrt{r_+} + \sqrt{r_-}}, & \tau &= -\log \rho, & \delta_n &= (\log n)^{-2}. \end{aligned}$$

Here r_-, r_+ are the endpoints of the density range, and ρ, τ are as in (2.7) for the interval $[r_-, r_+]$. The constructions produce densities with the margin $r_- + \delta_n \leq p \leq r_+ - \delta_n$. This margin is small enough that shrinking the density range by $O(\delta_n)$ changes ρ^M only by a factor $1 + o(1)$ when $M = O(\sqrt{\log n})$. It is large enough to absorb the change of $\mathbb{E}_P[D \mid X]$ caused by the perturbations (Section 7.2). Let $\mathcal{E}$ be a finite response space, with conditional probabilities, for $z \in \mathcal{E}$,

$$Q_z(u, v) = q_{z0} + uq_{zu} + vq_{zv}, \quad q_{z0} > 0, \quad \sum_z q_{z0} = 1, \quad \sum_z q_{zu} = \sum_z q_{zv} = 0. \quad (5.1)$$

Here $u = u(x)$ and $v = v(x)$ are the perturbations. We call $s_u(z) = q_{zu}/q_{z0}$ and $s_v(z) = q_{zv}/q_{z0}$ the score functions of u and v ; they may coincide. Let F be C^4 near $(0, 0)$ with $F_{uv}(0, 0) \neq 0$, and assume that the target is a parameter that equals $T(P) = \int pF(u, v)$ on the laws that we construct. We state the lower bounds in terms of $a_n = a_n(\theta, \tau)$ from (2.9), with these θ and τ .

We next describe the priors. We fix a cube $\mathcal{Q}$ in the interior of $[0, 1]^d$, of volume v_0 , and we partition it into B congruent subcubes, the *blocks*, where B is the d th power of an even integer. We group the blocks into $B/2$ pairs of adjacent blocks, and we label the two blocks of each pair by $\ell \in \{1, 2\}$. In each block we use rescaled coordinates in $\mathcal{U} = [0, 1]^d$, and we identify a pair with $\mathcal{U} \times \{1, 2\}$, which carries the product of Lebesgue measure and counting measure. Let $M \geq 2$ be an integer, $C_0 > 0$, and $A_u, A_v > 0$. We call two laws, the $+$ and $-$ priors, of random functions (p, u, v) on $[0, 1]^d$ *admissible with frequency M* if the following hold under both of them.

- (A1) The parameters of distinct pairs are independent.
- (A2) The function p is a smooth probability density with values in $[r_-, r_+]$, it equals a fixed constant p_0 outside $\mathcal{Q}$, and its mass on a pair is the same for every pair and in every realization.
- (A3) We have $u = v = 0$ outside $\mathcal{Q}$, $|u| \leq C_0 A_u$, and $|v| \leq C_0 A_v$.
- (A4) The parameters of a pair are $(\Theta, H, \sigma_u, \sigma_v)$, where (Θ, H) has the same law under both priors, Θ is uniform on $[0, 2\pi)$ and independent of H , and $\sigma_u, \sigma_v \in \{-1, 1\}$. There are functions p_c, p_1, p_2, h_u, h_v on the pair, which depend on H but neither on Θ nor on the prior, such that, in block coordinates,

$$p = p_c + p_1 \cos \Theta + p_2 \sin \Theta, \quad pu = A_u \sigma_u h_u, \quad pv = A_v \sigma_v h_v.$$

- (A5) Conditional on (Θ, H) , the signs satisfy

$$P_{\pm}\{\sigma_u = s, \sigma_v = t \mid \Theta, H\} = \frac{1}{4}\{1 \pm st \cos(M\Theta)\} \quad (s, t \in \{-1, 1\}).$$

We write $E_{\pm}$ for expectation under the $\pm$ prior, and E for expectation over the parameters (Θ, H) of one pair. Given a realization, the law of one observation is that of $O = (X, Z)$,

where X has density p and $P(Z = z \mid X = x) = Q_z(u(x), v(x))$. In [49, Theorem 2.1 and Sections 3–4], Robins et al. used priors with independent parameters on disjoint cells whose masses do not depend on the parameters, with a finite response space, to prove lower bounds for the missing-data mean and the expected conditional covariance. There the perturbations are bumps with independent random signs, and in the missing-data case the covariate density changes with the parameters without changing the mass of any cell.

By (A4), the unsigned products pu and $p v$ do not depend on Θ ; by (A5), only the sign correlation differs between the priors. Summing over the four sign values, for integers $k_u, k_v \geq 0$ and every bounded function f of (Θ, H) , we obtain

$$(E_+ - E_-)[\sigma_u^{k_u} \sigma_v^{k_v} f(\Theta, H)] = 2 \mathbf{1}_{\{k_u, k_v \text{ odd}\}} E[\cos(M\Theta) f(\Theta, H)]. \quad (5.2)$$

In particular, each prior is invariant under $(\sigma_u, \sigma_v) \mapsto (-\sigma_u, -\sigma_v)$, which maps (p, u, v) to $(p, -u, -v)$, and the $-$ prior is the image of the $+$ prior under $\sigma_v \mapsto -\sigma_v$, which maps (p, u, v) to $(p, u, -v)$.

As in [3, 23, 73], we Poissonize the sample: we compare the priors through the marked Poisson process on $[0, 1]^d \times \mathcal{E}$ whose intensity, with respect to dx times counting measure, is $2np(x)Q_z(u(x), v(x))$. Given the parameters, its restrictions to distinct pairs are independent. By (A2) and since $\int p = 1$, the mean value of p over a pair in block coordinates is $\bar{p} = \{1 - p_0(1 - v_0)\}/v_0$ in every realization. Let $\Lambda = 2nv_0/B$. As reference law on a pair we use the marked Poisson process with intensity measure $\mu(dx, z) = \Lambda \bar{p} q_{z0} dx$ on $\mathcal{Y} = (\mathcal{U} \times \{1, 2\}) \times \mathcal{E}$. In block coordinates, the process on the pair has intensity $\Lambda p(x)Q_z(u(x), v(x))$, whose density with respect to μ is $1 + \varphi$, where

$$\varphi(x, z) = e_0(x) + e_u(x)s_u(z) + e_v(x)s_v(z), \quad e_0 = \frac{p}{\bar{p}} - 1, \quad e_u = \frac{pu}{\bar{p}}, \quad e_v = \frac{pv}{\bar{p}}. \quad (5.3)$$

Since $\sum_z q_{zu} = \sum_z q_{zv} = 0$ and the pair has mass $2\bar{p}$ in block coordinates, we have $\int \varphi d\mu = 0$ in every realization.

The next lemma is in the spirit of the affinity bound of Robins et al. [49, Theorem 2.1] for two mixtures of product laws under a product prior over the cells of a partition. We use the Poisson process in place of the sample, and we expand in the factors that the points carry.

Lemma 5.2 (Comparison of the Poisson mixtures). *Let two priors be admissible with frequency M . Assume $B \geq n$, $A_u, A_v \leq 1$, and $\sup_{x,z} (|u(x)q_{zu}| + |v(x)q_{zv}|)/q_{z0} \leq 1/2$. Let $\mathbb{P}_{B,\pm}^{\text{Poi}}$ be the mixtures, under the $\pm$ prior, of the laws of the marked Poisson process restricted to $\mathcal{Q}$. For $j \geq 0$, define on $(\mathcal{U} \times \{1, 2\})^{j+2}$, in the block coordinates of a given pair,*

$$\Gamma_j(x, y, \zeta_1, \dots, \zeta_j) = (E_+ - E_-) \left[(pu)(x)(pv)(y) \prod_{i=1}^j \{p(\zeta_i) - \bar{p}\} \right],$$

and let $\bar{\Gamma}_j^2$ be the maximum over the pairs of $\|\Gamma_j\|_2^2 = \|\Gamma_j\|_{L^2((\mathcal{U} \times \{1,2\})^{j+2})}^2$. Then $\Gamma_j = 0$ for $j < M$, and

$$\begin{aligned} d_{\text{H}}^2(\mathbb{P}_{B,+}^{\text{Poi}}, \mathbb{P}_{B,-}^{\text{Poi}}) &\leq CB\Lambda^2 \sum_{j \geq M} \frac{(C\Lambda)^j}{j!} \bar{\Gamma}_j^2 \\ &\quad + CB\Lambda^4 A_u^2 A_v^2 (A_u^2 + A_v^2)^2 \frac{(C\Lambda)^M}{M!}. \end{aligned} \quad (5.4)$$

Here $C \geq 1$ depends only on $r_{\pm}$, v_0 , C_0 , and the probabilities q_{z0}, q_{zu}, q_{zv} . In particular, it is independent of n , B , M , A_u , and A_v .

Proof. By (A1), each mixture is a product of mixtures over the pairs, and the squared Hellinger distance is at most the sum of the distances for the pairs. We fix a pair. Recall that a Poisson process with finite intensity measure consists of a Poisson number of independent points [30, Propositions 3.2 and 3.5]. Conditioning on the number of points and using $\int \varphi \, d\mu = 0$, we find that the likelihood ratio of the process on the pair with respect to the reference process, given the parameters, is $\prod_{y \in \eta} \{1 + \varphi(y)\}$, where η is the configuration of marked points. The mixture likelihood ratios are therefore $f_{\pm}(\eta) = E_{\pm} \prod_{y \in \eta} \{1 + \varphi(y)\}$. By the amplitude assumption and $p/\bar{p} \geq r_-/r_+$, we have $1 + \varphi \geq c_{\varphi}$, where $c_{\varphi} = r_-/(2r_+)$, so that $f_{\pm} \geq c_{\varphi}^N$ for a configuration with N points. Since $(\sqrt{f} - \sqrt{g})^2 = (f - g)^2 / (\sqrt{f} + \sqrt{g})^2$, we obtain

$$d_{\text{H}}^2 \leq \frac{1}{4} E_{\mathbb{R}} [c_{\varphi}^{-N} (f_+ - f_-)^2], \quad (5.5)$$

where $E_{\mathbb{R}}$ is the expectation under the reference process.

We compute the right side exactly. Let $\nu = \Pi_+ - \Pi_-$ be the difference of the laws of the parameters of the pair under the two priors, and let φ_{ω} be the point factor at the parameters ω . Conditioning on the number of points gives $E_{\mathbb{R}} \prod_{y \in \eta} h(y) = \exp\{\int (h - 1) \, d\mu\}$ for bounded h ; for $0 \leq h \leq 1$, this is the formula for the probability generating functional of a Poisson process [30, Section 3.3 and Exercise 3.6]. With $h = c_{\varphi}^{-1}(1 + \varphi_{\omega})(1 + \varphi_{\omega'})$ and $\int \varphi_{\omega} \, d\mu = \int \varphi_{\omega'} \, d\mu = 0$, this gives

$$E_{\mathbb{R}} [c_{\varphi}^{-N} (f_+ - f_-)^2] = e^{|\mu|(1/c_{\varphi}-1)} \iint \exp\{c_{\varphi}^{-1} \langle \varphi_{\omega}, \varphi_{\omega'} \rangle_{L^2(\mu)}\} \, d\nu(\omega) \, d\nu(\omega').$$

We expand the exponential and use $\langle \varphi_{\omega}, \varphi_{\omega'} \rangle_{L^2(\mu)}^k = \langle \varphi_{\omega}^{\otimes k}, \varphi_{\omega'}^{\otimes k} \rangle_{L^2(\mu^k)}$. Since ν has total mass zero, the term $k = 0$ vanishes, and we obtain the exact identity

$$E_{\mathbb{R}} [c_{\varphi}^{-N} (f_+ - f_-)^2] = e^{|\mu|(1/c_{\varphi}-1)} \sum_{k \geq 1} \frac{c_{\varphi}^{-k}}{k!} \|G_k\|_{L^2(\mu^k)}^2, \quad G_k = (E_+ - E_-) \varphi^{\otimes k}. \quad (5.6)$$

The interchanges are justified because the point factors are bounded and ν has finite total variation. The prefactor is bounded, since $|\mu| = 2\Lambda\bar{p} \leq 4v_0r_+$ when $B \geq n$. We note

that c_φ^{-N} is, up to the factor $e^{|\mu|(1/c_\varphi-1)}$, the density of the Poisson process with intensity measure μ/c_φ relative to the reference process, and that the kernels of $\prod_{y \in \eta} \{1 + \varphi \omega(y)\}$ under that process are the tensor powers $\varphi_\omega^{\otimes k}$. Thus (5.6) is also a special case of the Fock-space representation of Poisson functionals [31, Theorem 1.1].

For a labeling $\omega \in \{0, u, v\}^k$, let k_u and k_v be the numbers of labels u and v , let $j = k - k_u - k_v$, and let $\Gamma_\omega(x_1, \dots, x_k) = (E_+ - E_-) \prod_{i=1}^k e_{\omega_i}(x_i)$. Then $G_k = \sum_\omega s_\omega \Gamma_\omega$, where $s_\omega(z_1, \dots, z_k)$ is the product of $s_{\omega_i}(z_i)$ over the i with $\omega_i \neq 0$. Let $\bar{s} \geq 1$ bound $|s_u|$ and $|s_v|$. By the Cauchy–Schwarz inequality over the 3^k labelings, and since $\sum_z q_{z0} = 1$,

$$\|G_k\|_{L^2(\mu^k)}^2 \leq (3\bar{s}^2 \Lambda \bar{p})^k \sum_\omega \|\Gamma_\omega\|_{L^2((\mathcal{U} \times \{1,2\})^k)}^2. \quad (5.7)$$

Only few labelings contribute. By (A4), we have $e_u = A_u \sigma_u h_u / \bar{p}$ and $e_v = A_v \sigma_v h_v / \bar{p}$, and all other factors are functions of (Θ, H) . By (5.2), $\Gamma_\omega = 0$ unless k_u and k_v are both odd, and then Γ_ω is $2A_u^{k_u} A_v^{k_v} \bar{p}^{-k}$ times the expectation over (Θ, H) of the product of $\cos(M\Theta)$, of the $k_u + k_v$ factors $h_u(x_i)$ and $h_v(x_i)$, and of the j factors $p(x_i) - \bar{p}$. Conditional on H , the factors h_u, h_v do not depend on Θ , and the product of the j density factors is a trigonometric polynomial of degree at most j in Θ . Since Θ is uniform and independent of H , we conclude that $\Gamma_\omega = 0$ when $j < M$, whatever the value of $k_u + k_v$.

If $k_u = k_v = 1$, then Γ_ω equals $\bar{p}^{-k} \Gamma_j$ with permuted arguments; in particular, $\Gamma_j = 0$ for $j < M$. There are $k(k-1)$ such labelings, and $k(k-1)/k! = 1/j!$. Since $\bar{p} \geq r_-$, (5.5)–(5.7) imply that their contribution to the distance for the pair is at most $C \sum_{j \geq M} (C_2 \Lambda)^{j+2} \|\Gamma_j\|_2^2 / j!$, where $C_2 = 3\bar{s}^2 r_+ / (c_\varphi r_-^2)$. Summing over the $B/2$ pairs gives the first term of (5.4).

For the remaining labelings, $q = k_u + k_v \geq 4$ and $j \geq M$. Since $|e_0| \leq r_+ / r_-$, $|e_u| \leq C_0 r_+ A_u / r_-$, $|e_v| \leq C_0 r_+ A_v / r_-$, and $(\mathcal{U} \times \{1,2\})^k$ has measure 2^k , we have $\|\Gamma_\omega\|_2^2 \leq 4C_3^k A_u^{2k_u} A_v^{2k_v}$ with $C_3 = 2 \max(1, C_0)^2 (r_+ / r_-)^2$. Let $\mathcal{A} = A_u^2 + A_v^2 \leq 2$. Since $\binom{q}{k_u} \leq q(q-1) \binom{q-2}{k_u-1}$ for $1 \leq k_u \leq q-1$, we have

$$\sum_{k_u=1}^{q-1} \binom{q}{k_u} A_u^{2k_u} A_v^{2(q-k_u)} \leq q(q-1) A_u^2 A_v^2 \mathcal{A}^{q-2}.$$

There are $\binom{j+q}{q}$ choices of the points labeled u or v , and $\binom{j+q}{q} / (j+q)! = 1/(j!q!)$. With $z = C_2 C_3 r_-^2 \Lambda$, these labelings therefore contribute at most

$$C A_u^2 A_v^2 \left(\sum_{j \geq M} \frac{z^j}{j!} \right) \left(\sum_{q \geq 4} \frac{z^q \mathcal{A}^{q-2}}{(q-2)!} \right) \leq C A_u^2 A_v^2 \mathcal{A}^2 \frac{z^{M+4}}{M!} e^{z(1+\mathcal{A})}$$

to the distance for the pair, where we used $(M+i)! \geq M! i!$ and $(i+2)! \geq i!$. Since $\Lambda \leq 2v_0$, the exponential factor is bounded, and summing over the pairs gives the second term of (5.4). $\square$

5.3 Reduction to two estimates

The priors of Section 6 divide each block into $R \geq 1$ fine cells and multiply the amplitudes $(BR)^{-a_*}$ and $(BR)^{-b_*}$ by $m \geq 1$; block-constant phases have $R = m = 1$. In (5.8), R^{1-M} records overlap after integration over the fine cells, $R^{(j-M)\sqrt{M}}$ allows a cost $R^{\sqrt{M}}$ per additional density factor, and $e^{C_E M}$ collects the remaining losses.

Proposition 5.3 (Reduction to two estimates). *Under the standing assumptions of Section 5.2, let $L = \log n$ and $\Delta_\theta = 1 - 2\theta$, let M be the nearest even integer to $\sqrt{\theta\Delta_\theta L/\tau}$, and let ρ_n be the parameter ρ of (2.7) computed for $[r_- + \delta_n, r_+ - \delta_n]$. Let $R = R_n \geq 1$ and $m = m_n \geq 1$ satisfy*

$$m \leq c_* M, \quad m \leq \min(R^{a_*}, R^{b_*}), \quad \log R = O(\log M)$$

for a constant $c_* > 0$, and fix $\epsilon_u, \epsilon_v \in (0, 1]$. Suppose that there are constants $C_0, C_\Gamma \geq 1$, $C_E \geq 0$, and $c_2 > 0$ such that the following holds for all large n and every $B \geq n$ that is the d th power of an even integer. There are two priors that are admissible with frequency M , with the constant C_0 and the amplitudes

$$A_u = \epsilon_u (BR)^{-a_*} m, \quad A_v = \epsilon_v (BR)^{-b_*} m,$$

such that, for every pair and every $j \geq 0$,

$$\|\Gamma_j\|_2^2 \leq C_\Gamma^{j+1} A_u^2 A_v^2 e^{C_E M} R^{1-M+(j-M)\sqrt{M}} \mathbf{1}_{j \geq M}, \quad (5.8)$$

and

$$\left| (E_+ - E_-) \int puv \right| \geq c_2 A_u A_v \rho_n^M. \quad (5.9)$$

Then, for all large n , there is such a B with $B \geq n$, for which $A_u \leq \epsilon_u B^{-a_*}$ and $A_v \leq \epsilon_v B^{-b_*}$, with the following property. Let $\mathcal{P}$ be any class that contains the law of one observation for every realization of either prior with this B , and let T be a parameter on $\mathcal{P}$ with $T(P) = \int pF(u, v)$ on these laws. Then there is $c_1 > 0$ such that, for all large n ,

$$\inf_{\hat{T}} \sup_{P \in \mathcal{P}} P\{|\hat{T} - T(P)| \geq c_1 m^2 a_n\} \geq \frac{3}{8}. \quad (5.10)$$

The infimum includes randomized estimators. Writing B_n for this choice of B , we also have $B_n / \{n(\log n)^q\} \rightarrow \infty$ for every fixed $q > 0$. The constant c_1 depends only on the fixed quantities, on $\epsilon_u \epsilon_v$, F , the response law, and the constants of the hypotheses.

Proof. Scales. Let $C \geq 1$ be the constant of Lemma 5.2 for the constant C_0 , let $C_1 = 2v_0 C C_\Gamma$, and fix $c_0 > 1 + \log C_1 + C_E$. We define

$$t_n = \frac{\Delta_\theta L}{M} - \log M + c_0,$$

we let B be the d th power of the smallest even integer k with $k^d R \geq ne^{t_n}$, and we set $x = BR/n$. Since M is within one of $M_* = \sqrt{\theta \Delta_\theta L / \tau}$ and $\Delta_\theta L = (\tau/\theta)M_*^2$, we have $\Delta_\theta L/M = (\tau/\theta)M + O(1)$, so that $t_n = (\tau/\theta)M + O(\log M)$. As $\log R = O(\log M)$, the number $k_* = (ne^{t_n}/R)^{1/d}$ tends to infinity, and $k_* \leq k < k_* + 2$ yields $1 \leq xe^{-t_n} \leq (1 + 2/k_*)^d \rightarrow 1$. In particular, $x \geq e^{t_n}$, and

$$\log x = t_n + o(1), \quad \log(B/n) = \frac{\tau}{\theta}M + O(\log M), \quad \log \Lambda = -\frac{\tau}{\theta}M + O(\log M), \quad (5.11)$$

where we used $\log(B/n) = \log x - \log R$ and $\Lambda = 2v_0 n/B$. In particular, $B \geq n$ for all large n , and since $M \asymp \sqrt{L}$, we have $B/\{n(\log n)^q\} \rightarrow \infty$ for every fixed $q > 0$. Since $m \leq R^{a_*}$, we have $A_u \leq \epsilon_u B^{-a_*} \leq n^{-a_*}$, and similarly $A_v \leq \epsilon_v B^{-b_*} \leq n^{-b_*}$. Since $|u| \leq C_0 A_u$ and $|v| \leq C_0 A_v$, the amplitude assumptions of Lemma 5.2 hold for all large n .

Indistinguishability. Let $y = CC_\Gamma \Lambda = C_1 n/B$. Since $\log y = -(\tau/\theta)M + O(\log M)$ by (5.11), and $\sqrt{M} \log R = O(\sqrt{M} \log M) = o(M)$, we have $yR^{\sqrt{M}} \leq 1/2$ for all large n . By (5.8), the first term of (5.4) is at most

$$CC_\Gamma B \Lambda^2 A_u^2 A_v^2 e^{C_E M} R^{1-M} \sum_{j \geq M} \frac{y^j R^{(j-M)\sqrt{M}}}{j!} \leq 2CC_\Gamma \mathcal{I}_n, \quad \mathcal{I}_n = B \Lambda^2 A_u^2 A_v^2 R^{1-M} e^{C_E M} \frac{y^M}{M!},$$

where we wrote $j = M + i$ and used $M!/(M+i)! \leq 1$ and $yR^{\sqrt{M}} \leq 1/2$. Since $B \Lambda^2 = 4v_0^2 n R/x$, $(BR)^{-2\theta} = (nx)^{-2\theta}$, and $R^{1-M} y^M = R(C_1/x)^M$, we have

$$\mathcal{I}_n = 4v_0^2 (\epsilon_u \epsilon_v)^2 n^{\Delta_\theta} R^2 m^4 e^{C_E M} \frac{C_1^M}{M! x^{M+1+2\theta}}.$$

Using $x \geq e^{t_n}$, $Mt_n = \Delta_\theta L - M \log M + c_0 M$, and $\log M! \geq M \log M - M$, we obtain

$$\log \mathcal{I}_n \leq \log(4v_0^2) + 2 \log R + 4 \log m - (1 + 2\theta)t_n - (c_0 - 1 - \log C_1 - C_E)M.$$

The terms $2 \log R$ and $4 \log m$ are $O(\log M)$, since $m \leq c_* M$, and $t_n > 0$ for all large n . By the choice of c_0 , we conclude that $\mathcal{I}_n \rightarrow 0$. Let $s_* = \min(a_*, b_*)$. The second term of (5.4), divided by $\mathcal{I}_n$, equals

$$C \Lambda^2 (A_u^2 + A_v^2)^2 R^{M-1} e^{-C_E M} C_\Gamma^{-M} \leq C n^{-4s_*} e^{O(M \log M)},$$

since $\Lambda \leq 2v_0$, $A_u^2 + A_v^2 \leq 2n^{-2s_*}$, and $\log R = O(\log M)$. This tends to zero, since $M \log M = o(L)$. It follows that $d_H(\mathbb{P}_{B,+}^{\text{Poi}}, \mathbb{P}_{B,-}^{\text{Poi}}) \rightarrow 0$.

Separation. Let $D_n = |E_+ T - E_- T|$, and let $F^{\text{oo}}(u, v) = \frac{1}{4}\{F(u, v) - F(-u, v) - F(u, -v) + F(-u, -v)\}$ be the part of F that is odd in each argument. By the two symmetries that follow from (A5), we have

$$E_+ T - E_- T = E_+ \int p\{F(u, v) - F(u, -v)\} = 2E_+ \int pF^{\text{oo}}(u, v),$$

and in the same way $(E_+ - E_-) \int puv = 2E_+ \int puv$. Moreover, $4F^{\text{oo}}(u, v) = \int_{-u}^u \int_{-v}^v F_{uv}(s, t) dt ds$ for small u, v . The linear Taylor terms of F_{uv} at $(0, 0)$ integrate to zero over this symmetric rectangle, and since $F \in C^4$ the remainder is at most $C(s^2 + t^2)$. This gives

$$F^{\text{oo}}(u, v) = F_{uv}(0, 0)uv + O\{|uv|(u^2 + v^2)\}.$$

Since $|u| \leq C_0 A_u$, $|v| \leq C_0 A_v$, and $\int p = 1$, it follows from (5.9) that

$$D_n \geq |F_{uv}(0, 0)|c_2 A_u A_v \rho_n^M - C A_u A_v (A_u^2 + A_v^2).$$

Since $A_u^2 + A_v^2 \leq 2n^{-2s_*}$, while $\rho_n^{-M} = e^{O(\sqrt{L})}$, the second term is at most half the first for all large n . Since $\log \rho_n - \log \rho = O(\delta_n)$ and $M\delta_n \rightarrow 0$, we also have $\rho_n^M \geq \rho^M/2$ for all large n . With $f(z) = \theta \Delta_\theta L/z + \tau z$, we have

$$f(M) - f(M_*) = \frac{\tau(M - M_*)^2}{M} = O(M^{-1}), \quad f(M_*) = 2\sqrt{\theta \Delta_\theta \tau L} = \kappa \sqrt{L},$$

with κ as in (2.9). By (5.11), we have $\log(BR) = L + t_n + o(1)$, and $\theta t_n + \tau M = f(M) - \theta \log M + \theta c_0$. Then

$$\log(A_u A_v \rho^M) = \log(\epsilon_u \epsilon_v m^2) - \theta L - f(M) + \theta \log M - \theta c_0 + o(1) = \log(m^2 a_n) + O(1),$$

where we used $\theta \log M = (\theta/2) \log L + O(1)$. Thus $D_n \geq c_3 m^2 a_n$ for a constant $c_3 > 0$ and all large n .

Testing. The target T is a constant plus a sum of independent contributions of the pairs, each bounded by C/B , since F is bounded near the origin. Then $\text{Var}_\pm T \leq C/B$, and since $B \geq n$, $m \geq 1$, and $\theta < 1/2$, we have $BD_n^2 \geq c_3^2 n a_n^2 = c_3^2 n^{1-2\theta} L^\theta e^{-2\kappa \sqrt{L}} \rightarrow \infty$. Outside $\mathcal{Q}$ the Poisson process has the same law in every realization, and it is independent of the process on $\mathcal{Q}$. Given the parameters, the number N of points has law $\text{Poisson}(2n)$, and given N the points are independent observations from the law P . We order the points randomly, keep the first n when $N \geq n$, and return a fixed sample otherwise. This defines a kernel that does not depend on the parameters. Its output mixtures are $(1 - \pi_n)Q_\pm^{(n)} + \pi_n \delta_*$, where $Q_\pm^{(n)}$ are the mixtures of the laws of n observations, δ_* is the law of the fixed sample, and $\pi_n = P\{\text{Poisson}(2n) < n\} \leq e^{-(1-\log 2)n}$. Since a kernel does not increase total variation and $\text{TV} \leq d_H$, we obtain $\text{TV}(Q_+^{(n)}, Q_-^{(n)}) \leq d_H(\mathbb{P}_{B,+}^{\text{Poi}}, \mathbb{P}_{B,-}^{\text{Poi}})/(1 - \pi_n) = o(1)$.

Given an estimator $\hat{T}$, we decide for the prior whose mean $E_\pm T$ is closer to $\hat{T}$. Under the $\pm$ prior, the decision is correct when $|T - E_\pm T| < D_n/4$ and $|\hat{T} - T(P)| < D_n/4$, and by Chebyshev's inequality the first event fails with prior probability at most $16 \text{Var}_\pm T / D_n^2$. Since the two testing errors sum to at least $1 - \text{TV}(Q_+^{(n)}, Q_-^{(n)})$, we obtain

$$\sup_{P \in \mathcal{P}} P\{|\hat{T} - T(P)| \geq D_n/4\} \geq \frac{1}{2} \left\{ 1 - \text{TV}(Q_+^{(n)}, Q_-^{(n)}) - \frac{16(\text{Var}_+ T + \text{Var}_- T)}{D_n^2} \right\} = \frac{1}{2} - o(1).$$

Since $D_n/4 \geq c_1 m^2 a_n$ with $c_1 = c_3/4$, this proves (5.10) for all large n . This argument is the total-variation version of the method of two fuzzy hypotheses [66, Theorem 2.15], and we control the prior concentration by Chebyshev's inequality. Since an independent random seed appended to the observations does not change the total variation distance, the same argument applies to randomized estimators. $\square$

Lattice priors

We construct priors satisfying Proposition 5.3 (Lemma 6.2). In each pair, the density on the bump supports is $r_0 \pm h_n \cos(\Theta + \phi(x))$, where Θ is uniform and ϕ is a random block function. The perturbations are bumps divided by the density, multiplied by a gate $S \in [0, 1]$, with sign correlation $\pm \cos(M\Theta)$. The phase ϕ sums independent coefficients $\xi_{a,q}$ times smoothed binary digits over levels $a = 1, \dots, J$ and coordinates $q = 1, \dots, d$. For $J = 0$, we have $\phi = 0$ and $S = 1$, and the bound is $c a_n$; for $J \geq 1$, three properties gain the factor $\log n$.

- (i) Outside a set of small measure, $\phi \in (2\pi/M)\mathbb{Z}$, so the M th Fourier coefficient of $1/p$ in Θ retains the separation factor ρ_n^M of block-constant phases.
- (ii) The gated characteristic functions of $\xi_{a,q}$ vanish outside $\bigcup_{l \in \mathbb{Z}} [lM - K_a, lM + K_a]$. For a likelihood term with one factor each of u, v and j density factors, averaging over Θ gives zero for $j < M$. For $M \leq j < M + \sqrt{M}$, the band restriction implies the digit mismatch bounds (6.21), whose integrals yield $R^{1-M} e^{C_E M}$, where R counts fine cells. For larger j , a crude bound suffices.
- (iii) The gate controls weighted coefficient powers: S times the change of ϕ across a fine cell is $O(1/m)$, with $m \asymp M$. Dividing a bump by the density costs a factor $1/m$ of a fine-scale derivative, allowing amplitudes $m(BR)^{-a_*}$ and $m(BR)^{-b_*}$.

6.1 Lattice laws with compact Fourier support

Throughout this section we fix the integer $Q = \lceil \max(\alpha, \beta) \rceil + 2$, and we define

$$\Pi(y) = c_Q^{-1} \left(\frac{\sin y}{y} \right)^{2Q}, \quad c_Q = \int_{\mathbb{R}} \left(\frac{\sin y}{y} \right)^{2Q} dy,$$

where $\sin y/y$ takes its continuous value one at $y = 0$. For $Y \sim \Pi$ we write $\mu_2 = EY^2$, which is finite because $\Pi(y) \leq c_Q^{-1} \min(1, |y|^{-2Q})$, and we write $\hat{f}(w) = \int_{\mathbb{R}} f(x) e^{iwx} dx$.

Lemma 6.1 (Gated lattice laws). *Let $\eta > 0$ and $\lambda \geq 1$, let $M \geq 2$ be an integer, and let $h_M = 2\pi/M$, $\pi_\eta(x) = \eta^{-1} \Pi(x/\eta)$, and $K = 6Q/\eta$. We define*

$$G(x) = \left(\frac{\sin(x/(\lambda\eta))}{x/(\lambda\eta)} \right)^{2Q} \quad (x \in \mathbb{R}), \quad G(0) = 1.$$

If $4K \leq M$, then the numbers $h_M \pi_\eta(x)$, $x \in h_M \mathbb{Z}$, sum to one. If ξ has the law $P\{\xi = x\} = h_M \pi_\eta(x)$ on $h_M \mathbb{Z}$, then

$$0 \leq G \leq 1, \quad 1 - EG(\xi) \leq \frac{Q\mu_2}{3\lambda^2}, \quad G(x)|x|^r \leq (\lambda\eta)^r \quad (x \in \mathbb{R}, 0 \leq r \leq 2Q - 2). \quad (6.1)$$

Moreover, the function $\widehat{\Xi}(w) = E[G(\xi)^2 e^{i w \xi}]$ of $w \in \mathbb{R}$ satisfies $|\widehat{\Xi}| \leq 1$, and $\widehat{\Xi}(w) = 0$ unless $|w - lM| \leq K$ for some $l \in \mathbb{Z}$.

Proof. The identity $\sin y/y = \frac{1}{2} \int_{-1}^1 e^{i\omega y} d\omega$ shows that $(\sin y/y)^{2Q}$ is the Fourier transform of the $2Q$ -fold convolution of $\frac{1}{2} \mathbf{1}_{[-1,1]}$, which is supported in $[-2Q, 2Q]$. By Fourier inversion, $\widehat{\pi}_\eta$ and $\widehat{G}$ are supported in $[-2Q/\eta, 2Q/\eta]$ and $[-2Q/(\lambda\eta), 2Q/(\lambda\eta)]$, respectively. Since multiplication convolves Fourier transforms and adds their supports, and since $\lambda \geq 1$, $\widehat{G^k \pi_\eta}$ is supported in $[-K, K]$ for $k = 0, 1, 2$.

Let $f = G^k \pi_\eta$ with $k \in \{0, 1, 2\}$, and let $w \in \mathbb{R}$. Since $0 \leq f(x) \leq C_\eta(1 + |x|)^{-2Q}$, the h_M -periodic function $y \mapsto \sum_{x \in h_M \mathbb{Z}} f(y+x) e^{i w (y+x)}$ is continuous, and its Fourier coefficients are $h_M^{-1} \widehat{f}(w - lM)$, $l \in \mathbb{Z}$. Since only finitely many of them are nonzero, the function is a trigonometric polynomial and equals its Fourier series everywhere. At $y = 0$ we obtain an instance of the Poisson summation formula [61, Chapter 5, Section 3],

$$h_M \sum_{x \in h_M \mathbb{Z}} f(x) e^{i w x} = \sum_{l \in \mathbb{Z}} \widehat{f}(w + lM). \quad (6.2)$$

At $w = 0$ only the term $l = 0$ remains, since $K < M$. With $k = 0$, the numbers $h_M \pi_\eta(x)$ sum to $\widehat{\pi}_\eta(0) = 1$, and with $k = 1$ we obtain $EG(\xi) = \int G \pi_\eta$. With $k = 2$, (6.2) gives $\widehat{\Xi}(w) = \sum_l \widehat{f}(w + lM)$. Since $\widehat{f}$ vanishes outside $[-K, K]$ and $2K < M$, at most one term is nonzero, and it vanishes unless $|w + lM| \leq K$; moreover $|\widehat{\Xi}| \leq EG(\xi)^2 \leq 1$.

The bound $|\sin z/z| \leq \min(1, |z|^{-1})$ implies that $0 \leq G \leq 1$ and $G(x)|x|^r \leq (\lambda\eta)^r$ for $0 \leq r \leq 2Q$. For independent U, V uniform on $[-1, 1]$, we have $(\sin z/z)^2 = E \cos(z(U+V)) \geq 1 - z^2 E(U+V)^2/2 = 1 - z^2/3$. Then $1 - (\sin z/z)^{2Q} \leq Q\{1 - (\sin z/z)^2\} \leq Qz^2/3$. Since $\int G \pi_\eta = E(\sin(Y/\lambda)/(Y/\lambda))^{2Q}$ for $Y \sim \Pi$, we obtain $1 - EG(\xi) \leq Q\mu_2/(3\lambda^2)$. $\square$

Lattice laws obtained in this way, by sampling a probability density whose Fourier transform has compact support, are constructed in [68, Section V, Proposition 5 and Corollary 6]. There, the sampled masses are normalized by the same formula, the characteristic function is the sum of the translated copies of the Fourier transform, and Example 1 uses powers of $\sin y/y$.

6.2 Construction of the priors

We use the cube $\mathcal{Q}$ of volume v_0 , its blocks and pairs, and the block coordinates $x \in \mathcal{U} = [0, 1]^d$ of Section 5.2. We fix smooth functions $\iota_{\text{in}}, \iota_{\text{out}} : \mathcal{U} \rightarrow [0, 1]$ with compact supports

in the interior of $\mathcal{U}$, such that ι_{in} is not identically zero and $\iota_{\text{out}} = 1$ on the support of ι_{in} . We extend them by zero to $\mathbb{R}^d$, and we write $\bar{\iota} = \int_{\mathcal{U}} \iota_{\text{out}}$ and $I_{\text{in}} = \int_{\mathcal{U}} \iota_{\text{in}}^2 > 0$. We set

$$r_0 = \frac{r_- + r_+}{2}, \quad h_n = \frac{r_+ - r_-}{2} - \delta_n, \quad p_0 = \frac{1 - v_0 \bar{\iota} r_0}{1 - v_0 \bar{\iota}},$$

and we choose v_0 so small that p_0 lies in a fixed interior subinterval of (r_-, r_+) . Throughout this section, n is large enough that

$$0 < \delta_n \leq \frac{1}{2} \min\{r_0 - r_-, p_0 - r_-, r_+ - p_0\}.$$

We write $s_0 = \min(\alpha, \beta)$ and $\alpha_0 = \min(s_0, 1)/2$; we use only $0 < \alpha_0 < 1$ and $\alpha_0 < s_0$. Let $J \geq 0$ be an integer, and let $R = 2^{Jd}$ and $m = 2^{Js_0}$. The fine cells of a block are its R dyadic subcubes of side 2^{-J} . Sums over an empty set of levels are zero, and products over it are one.

Digit selectors. For $y \in \mathbb{R}$ and $a \geq 1$ we write $e_a(y) = \lfloor 2^a y \rfloor \bmod 2$. If x is uniform on $\mathcal{U}$, the digits $e_a(x_q)$, $1 \leq a \leq J$, $1 \leq q \leq d$, are independent and uniform on $\{0, 1\}$, and the string of these dJ digits determines the fine cell of x . To smooth the digits, we fix a nondecreasing C^∞ function Y with $Y = 0$ on $(-\infty, -1]$, $Y = 1$ on $[1, \infty)$, and $Y(0) = 1/2$. For $0 < \gamma \leq 1/4$ and $y \in \mathbb{R}$ we let z be a nearest integer to y , and we define

$$\omega_\gamma(y) = Y((-1)^{z+1}(y - z)/\gamma).$$

At distance at least γ from the integers, $\omega_\gamma(y)$ equals $\lfloor y \rfloor \bmod 2$; in particular it is locally constant near the half-integers, where the two choices of z agree. This implies that ω_γ is C^∞ , with values in $[0, 1]$ and $|\omega_\gamma^{(N)}| \leq C_N \gamma^{-N}$ for $N \geq 1$. If $\lfloor y \rfloor$ is odd, then either $z = \lfloor y \rfloor$ and $y - z \geq 0$, or $z = \lfloor y \rfloor + 1$ is even and $z - y > 0$; in both cases $\omega_\gamma(y) \geq Y(0) = 1/2$. In the same way $\omega_\gamma(y) \leq 1/2$ if $\lfloor y \rfloor$ is even. Therefore

$$|\omega_\gamma(y) - e| \geq \frac{1}{2} \quad \text{if } e \in \{0, 1\} \text{ and } e \neq \lfloor y \rfloor \bmod 2. \quad (6.3)$$

Scales. For a level $1 \leq a \leq J$, we write $b = J - a$ and define

$$\gamma_a = \frac{\gamma_*}{(b+1)^2}, \quad \lambda_a = \lambda_*(b+1), \quad \eta_a = \frac{\gamma_a 2^{b\alpha_0}}{\lambda_a m}, \quad K_a = \frac{6Q}{\eta_a},$$

where $\gamma_* = \min\{1/4, I_{\text{in}}/(16d)\}$ and $\lambda_* = \max\{1, \sqrt{8Qd\mu_2/3}\}$. Thus $\lambda_a \eta_a = \gamma_a 2^{b\alpha_0}/m$, so that coarse levels carry larger coefficients, and $K_a = K(b)m$ with $K(b) = 6Q\lambda_*(b+1)^3 2^{-b\alpha_0}/\gamma_*$. We set $\bar{K} = \sup_{b \geq 0} K(b)$ and $c_* = 1/(4\bar{K})$, and we require that $m \leq c_* M$, so that

$$4K_a \leq M \quad (1 \leq a \leq J). \quad (6.4)$$

All these constants depend only on d, α, β and ι_{in} , and they are fixed before M is chosen.

The phase and the gate. Let $M \geq 2$ be even with $m \leq c_* M$, and let $h_M = 2\pi/M$. In each pair, let $\xi_{a,q}$, for $1 \leq a \leq J$ and $1 \leq q \leq d$, be independent, with the law of ξ in Lemma 6.1 for $\eta = \eta_a$ and $\lambda = \lambda_a$; by (6.4), the lemma applies. We write G_a and $\widehat{\Xi}_a$ for the corresponding functions G and $\widehat{\Xi}$, and we define

$$S = \prod_{a,q} G_a(\xi_{a,q}), \quad s_{a,q}(x) = \omega_{\gamma_a}(2^a x_q), \quad \phi(x) = \sum_{a=1}^J \sum_{q=1}^d \xi_{a,q} s_{a,q}(x)$$

for $x \in \mathbb{R}^d$. By independence, (6.1), and the inequality $\prod_i (1 - t_i) \geq 1 - \sum_i t_i$ for $t_i \in [0, 1]$, we have

$$ES \geq 1 - \frac{Qd\mu_2}{3\lambda_*^2} \sum_{b \geq 0} (b+1)^{-2} \geq \frac{3}{4}, \quad ES^2 \geq (ES)^2 \geq \frac{9}{16},$$

where we used $\sum_{b \geq 0} (b+1)^{-2} < 2$ and the choice of λ_* . Since $S \leq G_a(\xi_{a,q})$ for every (a, q) and $0 \leq G_a \leq 1$, the last bound in (6.1) implies that, for distinct pairs (a_i, q_i) and integers $r_i \geq 1$ with $\sum_i r_i \leq 2Q - 2$,

$$S \prod_i |\xi_{a_i, q_i}|^{r_i} \leq \prod_i (\lambda_{a_i} \eta_{a_i})^{r_i}. \quad (6.5)$$

Let $\mathcal{U}^\circ$ be the set of $x \in \mathcal{U}$ such that $2^a x_q$ has distance at least γ_a from the integers for all a and q . On $\mathcal{U}^\circ$ we have $s_{a,q}(x) = e_a(x_q) \in \{0, 1\}$, so that $\phi(x)$ is a sum of elements of $h_M \mathbb{Z}$, and therefore

$$\cos(M\phi(x)) = 1 \quad (x \in \mathcal{U}^\circ). \quad (6.6)$$

For each a and q , the set of $x \in \mathcal{U}$ for which $2^a x_q$ has distance less than γ_a from the integers has measure $2\gamma_a$. Since $\sum_a \gamma_a < 2\gamma_*$, a union bound gives

$$|\mathcal{U} \setminus \mathcal{U}^\circ| < 4d\gamma_* \leq I_{\text{in}}/4. \quad (6.7)$$

The priors. In each pair, let Θ be uniform on $[0, 2\pi)$ and independent of $\xi = (\xi_{a,q})_{a,q}$. We call (Θ, ξ) the density parameters of the pair; both blocks of the pair use them, and their law is the same under both priors. On the block of the pair with label ℓ , in block coordinates, we define

$$r(x) = r_0 + (-1)^{\ell-1} h_n \cos(\Theta + \phi(x)), \quad p = p_0 + \iota_{\text{out}}(r - p_0),$$

and $p = p_0$ outside $\mathcal{Q}$. Every realization of ϕ is C^∞ , since it is a finite sum with finite coefficients. Since both blocks of a pair use the same function ϕ in block coordinates, the oscillations of the two blocks cancel in $\int p$, and the mass of p on each pair is $2(v_0/B)\bar{p}$ in every realization, where

$$\bar{p} = p_0 + \bar{\iota}(r_0 - p_0)$$

is the mean value of p over a pair. The total mass is therefore $p_0(1 - v_0) + v_0\bar{p} = p_0 + v_0\bar{\iota}(r_0 - p_0) = 1$. Since ι_{out} vanishes near the block boundaries, p is a C^∞ probability

density. On the support of ι_{out} it is a convex combination of p_0 and $r \in [r_- + \delta_n, r_+ - \delta_n]$, so that $r_- + \delta_n \leq p \leq r_+ - \delta_n$.

Conditional on the density parameters, the signs $\sigma_u, \sigma_v \in \{-1, 1\}$ of a pair have the law

$$P_{\pm}\{\sigma_u = s, \sigma_v = t \mid \Theta, \xi\} = \frac{1}{4}\{1 \pm st \cos(M\Theta)\} \quad (s, t \in \{-1, 1\})$$

under the $\pm$ prior, as in (A5) with $H = \xi$. As in Section 5.2, E averages the density parameters of one pair. Let $\epsilon_u, \epsilon_v \in (0, 1]$. With the amplitudes

$$A_u = \epsilon_u (BR)^{-a_*} m, \quad A_v = \epsilon_v (BR)^{-b_*} m, \quad (6.8)$$

we define

$$u = A_u \sigma_u S \iota_{\text{in}} / p, \quad v = A_v \sigma_v S \iota_{\text{in}} / p \quad (6.9)$$

on both blocks of each pair, and $u = v = 0$ outside $\mathcal{Q}$. The parameters of distinct pairs are independent. Without the gate S and the division by p , these are smooth bumps on the blocks with random signs. Classical lower-bound constructions use such localized bumps with binary coefficients [62, Section 2, p. 1045] or random signs [22, Section 3.5.2], [49, Sections 3–4]; here the signs are correlated with the phase of the density.

Lemma 6.2 (Lattice priors). *Let $J \geq 0$, let $M \geq 2$ be even with $m \leq c_* M$, and let B be the d th power of an even integer. There are constants $C, C_\Gamma \geq 1$ and $C_E \geq 0$, which depend only on $d, \alpha, \beta, r_{\pm}$ and the fixed choices of the construction, and not on $n, B, J, M, \epsilon_u, \epsilon_v$, such that the two priors above have the following properties.*

- (a) *They are admissible with frequency M , with $C_0 = 1/r_-$ and the amplitudes (6.8). Every realization of p is a C^∞ probability density with*

$$r_- + \delta_n \leq p \leq r_+ - \delta_n, \quad (6.10)$$

and the functions u and v vanish outside $\mathcal{Q}$.

- (b) *In every realization, u and v are C^∞ , and*

$$\|u\|_{C^\alpha} \leq C\epsilon_u, \quad \|v\|_{C^\beta} \leq C\epsilon_v, \quad \|u\|_\infty \leq A_u/r_-, \quad \|v\|_\infty \leq A_v/r_-. \quad (6.11)$$

Let K be smooth on a neighborhood N of $(0, 0)$. If $A_u + A_v$ is below a threshold that depends on K , then

$$\begin{aligned} K(0, v) = 0 \text{ on } N &\implies \|K(u, v)\|_{C^\alpha} \leq C_K \epsilon_u, \\ K(u, 0) = 0 \text{ on } N &\implies \|K(u, v)\|_{C^\beta} \leq C_K \epsilon_v, \end{aligned} \quad (6.12)$$

where C_K depends only on K and on the quantities on which C depends.

- (c) *For every pair and every $j \geq 0$, the coefficients Γ_j of Lemma 5.2 satisfy (5.8), that is,*

$$\|\Gamma_j\|_2^2 \leq C_\Gamma^{j+1} A_u^2 A_v^2 e^{C_E M} R^{1-M+(j-M)\sqrt{M}} \mathbf{1}_{j \geq M},$$

where the norm is that of $L^2((\mathcal{U} \times \{1, 2\})^{j+2})$.

(d) With ρ_n as in Proposition 5.3, we have

$$(E_+ - E_-) \int_{[0,1]^d} puv \geq c_2 A_u A_v \rho_n^M, \quad c_2 = \frac{9v_0 I_{\text{in}}}{16r_0}.$$

Block-constant phases. For $J = 0$, we have $\phi = 0$, $S = 1$, $R = m = 1$, and $\mathcal{U}^\circ = \mathcal{U}$, with Θ the only density parameter. On the bump supports, $p = r_0 \pm h_n \cos \Theta$, $pu = A_u \sigma_u \iota_{\text{in}}$, and $p_v = A_v \sigma_v \iota_{\text{in}}$. After removing their signs, these products do not depend on Θ . The M th Fourier coefficient of $1/p$ is of order ρ_n^M by (4.2). The scale conditions of Proposition 5.3 reduce to $1 \leq c_* M$, which holds for all large n . That proposition, applied with Lemma 6.2(a), (c), and (d), yields, for every class $\mathcal{P}$ and target T as in the proposition, a constant $c_1 > 0$ such that, for all large n ,

$$\inf_{\hat{T}} \sup_{P \in \mathcal{P}} P\{|\hat{T} - T(P)| \geq c_1 m^2 a_n\} \geq \frac{3}{8} \quad (6.13)$$

with $m = 1$. The proof of Proposition 7.1 uses the largest J with $m \leq c_* M$, for which $m^2 \asymp \log n$.

6.3 Proof of Lemma 6.2

We use the Hölder norm $\|\cdot\|_{C^t}$ of the model class on $[0,1]^d$, and we write $\|\cdot\|_{C^t(\mathcal{U})}$ for the norm defined in the same way on $\mathcal{U}$, in block coordinates.

Proof of Lemma 6.2(a). We showed above that p is a C^∞ probability density that satisfies (6.10), equals p_0 outside $\mathcal{Q}$, and has the same mass on every pair in every realization, which is (A2). By construction, the parameters of distinct pairs are independent, u and v vanish outside $\mathcal{Q}$, and the signs have the law in (A5). Since $p \geq r_-$ and $0 \leq S, \iota_{\text{in}} \leq 1$, we have $|u| \leq A_u/r_-$ and $|v| \leq A_v/r_-$, which gives (A3) with $C_0 = 1/r_-$. In (A4), we have $H = \xi$, $pu = A_u \sigma_u S \iota_{\text{in}}$, $p_v = A_v \sigma_v S \iota_{\text{in}}$, and, on the block with label ℓ ,

$$p_c = p_0 + \iota_{\text{out}}(r_0 - p_0), \quad p_1 = (-1)^{\ell-1} h_n \iota_{\text{out}} \cos \phi, \quad p_2 = -(-1)^{\ell-1} h_n \iota_{\text{out}} \sin \phi,$$

and $h_u = h_v = S \iota_{\text{in}}$. These functions depend on ξ but neither on Θ nor on the prior. $\square$

Proof of Lemma 6.2(b). We work in block coordinates on a block with label ℓ , and we set $k_{\text{max}} = \lceil \max(\alpha, \beta) \rceil = Q - 2$ and

$$f(y) = \frac{1}{r_0 + (-1)^{\ell-1} h_n \cos(\Theta + y)}, \quad H = S \iota_{\text{in}} f(\phi).$$

Since $p = r$ on the support of ι_{in} , we have $H = S \iota_{\text{in}}/p$ and $|H| \leq 1/r_-$. The derivatives of f through order $k_{\text{max}} + 1$ are bounded by constants that depend only on $r_\pm$ and k_{max} ,

since its denominator lies in $[r_-, r_+]$. Below, constants may depend on $d, \alpha, \beta, r_\pm, \iota_{\text{in}}, Y$ and the fixed constants of the construction, but not on $n, B, J, M, \epsilon_u, \epsilon_v$, or the realization.

Compositions. Let κ be a C^∞ function on $[-1/r_-, 1/r_-]$ with $\kappa(0) = 0$, and let $s_0 \leq t \leq \max(\alpha, \beta)$. We claim that

$$\|\kappa(H)\|_{C^t(\mathcal{U})} \leq C_\kappa(1 + 2^{Jt}/m), \quad (6.14)$$

where C_κ depends on κ only through bounds on its derivatives of order at most $\lceil t \rceil + 1$. We define

$$w_a = \lambda_a \eta_a = \frac{\gamma_a 2^{(J-a)\alpha_0}}{m}, \quad L_a = \frac{2^a}{\gamma_a}, \quad \phi_a = \sum_{c=1}^a \sum_{q=1}^d \xi_{c,q} s_{c,q} \quad (0 \leq a \leq J).$$

Since $\gamma_a \leq 1/4$ and $2^{(J-a)\alpha_0} \leq 2^{Js_0} = m$, we have $w_a \leq 1$ and $L_a \geq 1$. Since $s_{a,q}$ depends only on x_q , we also have $|D^{\mathbf{j}} s_{a,q}| \leq CL_a^{|\mathbf{j}|}$ for every multi-index $\mathbf{j}$, including $\mathbf{j} = 0$. For an integer $1 \leq k \leq \lceil t \rceil$ and $b = J - a$, we have $w_{a-i} L_{a-i}^k / (w_a L_a^k) = \{(b+i+1)/(b+1)\}^{2(k-1)} 2^{-i(k-\alpha_0)}$ for $0 \leq i < a$, and hence

$$\sum_{c=1}^a \sum_{q=1}^d w_c L_c^k \leq d w_a L_a^k \sum_{i \geq 0} (i+1)^{2(k-1)} 2^{-i(k-\alpha_0)} \leq C w_a L_a^k, \quad (6.15)$$

where the series converges because $k \geq 1 > \alpha_0$.

Let $F_a = \kappa(S_{\text{lin}} f(\phi_a))$, $V_a = F_a - F_{a-1}$, and $W_a = \phi_a - \phi_{a-1}$ for $1 \leq a \leq J$. By the fundamental theorem of calculus,

$$V_a = S_{\text{lin}} W_a \int_0^1 f'(z_{a,u}) \kappa'(S_{\text{lin}} f(z_{a,u})) du, \quad z_{a,u} = \phi_{a-1} + u W_a,$$

and the arguments of κ lie in $[0, 1/r_-]$. Let $0 \leq |\mathbf{j}| \leq \lceil t \rceil$. By the chain and product rules, $D^{\mathbf{j}} V_a$ is a sum, with a number of terms that depends only on $|\mathbf{j}|$ and d , of products of a power S^ϱ with $\varrho \geq 1$, of derivatives of ι_{in}, f , and κ of order at most $|\mathbf{j}| + 1$, of $D^{\mathbf{j}_0} W_a$, and of $D^{\mathbf{j}_l} z_{a,u}$ for $1 \leq l \leq r$, where $|\mathbf{j}_l| \geq 1$ for $l \geq 1$ and $|\mathbf{j}_0| + \sum_l |\mathbf{j}_l| \leq |\mathbf{j}|$. In particular $r \leq |\mathbf{j}|$. We expand W_a and $z_{a,u}$ over levels and coordinates. The coefficients then enter through products

$$\xi_{a,q_0} D^{\mathbf{j}_0} s_{a,q_0} \prod_{l=1}^r \xi_{c_l,q_l} D^{\mathbf{j}_l} s_{c_l,q_l} \quad (c_l \leq a),$$

with at most $|\mathbf{j}| + 1 \leq k_{\max} + 1 = Q - 1 \leq 2Q - 2$ coefficients, counted with multiplicity. Since $S^\varrho \leq S$, we group repeated coefficients and apply (6.5), which bounds S^ϱ times such a product by $C w_a L_a^{|\mathbf{j}_0|} \prod_l w_{c_l} L_{c_l}^{|\mathbf{j}_l|}$. Summing over q_0 , and over the levels $c_l \leq a$ and the coordinates q_l by (6.15), and using $w_a \leq 1$ and $L_a \geq 1$, we obtain

$$\|D^{\mathbf{j}} V_a\|_\infty \leq C_\kappa w_a L_a^{|\mathbf{j}|} \quad (0 \leq |\mathbf{j}| \leq \lceil t \rceil).$$

With $N = \lceil t \rceil - 1$ and $\vartheta = t - N \in (0, 1]$, we find from the mean value theorem and this bound that, for $|\mathbf{j}| = N$,

$$|D^{\mathbf{j}}V_a(x) - D^{\mathbf{j}}V_a(y)| \leq C_\kappa w_a L_a^N \min\{1, L_a|x - y|\} \leq C_\kappa w_a L_a^t |x - y|^\vartheta,$$

where we used $\min(1, z) \leq z^\vartheta$ for $z \geq 0$. Since $L_a \geq 1$, the derivative suprema in the norm are also at most $C_\kappa w_a L_a^t$, and therefore $\|V_a\|_{C^t(\mathcal{U})} \leq C_\kappa w_a L_a^t$. The function $F_0 = \kappa(S_{t_{\text{in}}}f(0))$ has norm at most C_κ , since $Sf(0) \in [0, 1/r_-]$ is a number. Since $\kappa(H) = F_J = F_0 + \sum_{a=1}^J V_a$, we obtain

$$\|\kappa(H)\|_{C^t(\mathcal{U})} \leq C_\kappa \left\{ 1 + \frac{2^{Jt}}{m} \gamma_*^{1-t} \sum_{b \geq 0} (b+1)^{2(t-1)} 2^{-b(t-\alpha_0)} \right\} \leq C_\kappa (1 + 2^{Jt}/m),$$

where the series is bounded uniformly in $s_0 \leq t \leq \max(\alpha, \beta)$ because $s_0 > \alpha_0$. For $J = 0$ the sum over a is empty. This proves (6.14).

Conclusion. On a block, we have $u = A_u \sigma_u H$ and $v = A_v \sigma_v H$ in block coordinates, which gives the sup-norm bounds. Let g be a C^∞ function on $[0, 1]^d$ that vanishes outside $\mathcal{Q}$ and, on each block, equals g_k in block coordinates, where every g_k vanishes near $\partial\mathcal{U}$. The blocks have side $\ell_B = (v_0/B)^{1/d} \leq 1$. Rescaling multiplies a derivative of order $|\mathbf{j}| \leq t$ by $\ell_B^{-|\mathbf{j}|} \leq \ell_B^{-t}$ and a seminorm of order ϑ of a derivative of order $\lceil t \rceil - 1$ by ℓ_B^{-t} . If x and y do not lie in the same block, the segment from x to y meets the boundary of the block of x at a point x' and the boundary of the block of y at a point y' , where all derivatives of g vanish; the same holds if one of the points lies outside $\mathcal{Q}$. It follows that a difference quotient of order $\vartheta \leq 1$ between x and y is at most twice the largest such quotient within one block. Since the norm adds the derivative suprema and the Hölder seminorm, whose maxima may occur in different blocks, we obtain

$$\|g\|_{C^t([0,1]^d)} \leq 3(B/v_0)^{t/d} \max_k \|g_k\|_{C^t(\mathcal{U})}. \quad (6.16)$$

We apply (6.14) with $\kappa(z) = z$ and $t = \alpha$, and (6.16). Since $2^{J\alpha} = R^{a_*}$, $A_u B^{a_*} = \epsilon_u R^{-a_*} m$, and $m = 2^{Js_0} \leq R^{a_*}$, we obtain

$$\|u\|_{C^\alpha} \leq C A_u B^{a_*} (1 + R^{a_*}/m) = C \epsilon_u (m R^{-a_*} + 1) \leq 2C \epsilon_u.$$

The bound for v is the same with β , b_* , and ϵ_v .

Let K be smooth on a neighborhood of $(0, 0)$ with $K(0, v) = 0$ there, and fix a closed ball N' of radius at most one around $(0, 0)$ inside this neighborhood. On N' we write $K(u, v) = u \tilde{K}(u, v)$ with $\tilde{K}(u, v) = \int_0^1 \partial_u K(su, v) ds$. If $(A_u + A_v)/r_-$ is at most the radius of N' , then on each block $K(u, v) = A_u \sigma_u \kappa(H)$ with $\kappa(z) = z \tilde{K}(A_u \sigma_u z, A_v \sigma_v z)$, and $K(u, v) = 0$ outside the blocks. The derivatives of κ of order at most $k_{\max} + 1$ on $[-1/r_-, 1/r_-]$ are bounded by a constant that depends only on K and N' , because $A_u \leq \epsilon_u B^{-a_*} \leq 1$ and $A_v \leq \epsilon_v B^{-b_*} \leq 1$. The argument above, with this κ and $t = \alpha$, yields $\|K(u, v)\|_{C^\alpha} \leq C_K \epsilon_u$. The case $K(u, 0) = 0$ is symmetric, with β and ϵ_v . $\square$

Proof of Lemma 6.2(c). We fix a pair. Since $pu = A_u \sigma_u S_{l_{\text{in}}}$ and $p_v = A_v \sigma_v S_{l_{\text{in}}}$, (5.2) gives

$$\Gamma_j(x, y, \zeta_1, \dots, \zeta_j) = 2A_u A_v l_{\text{in}}(x) l_{\text{in}}(y) E \left[S^2 \cos(M\Theta) \prod_{i=1}^j \{p(\zeta_i) - \bar{p}\} \right].$$

Since p and $\bar{p}$ lie in $[r_-, r_+]$, we have $|p - \bar{p}| \leq r_+$, so that $\|\Gamma_j\|_\infty \leq 2A_u A_v r_+^j$. Since $\|\Gamma_j\|_2^2 \leq \|\Gamma_j\|_\infty \|\Gamma_j\|_1$, it suffices to prove that

$$\|\Gamma_j\|_1 \leq 8A_u A_v (6e r_+)^j e^{C_E M} R^{1-M+(j-M)\sqrt{M}} \mathbf{1}_{j \geq M}, \quad (6.17)$$

where $e = \exp(1)$; then (c) holds with $C_\Gamma = 16(1 + 6e r_+^2)$. As $(\mathcal{U} \times \{1, 2\})^{j+2}$ has measure 2^{j+2} , we have

$$\|\Gamma_j\|_1 \leq 8A_u A_v (2r_+)^j. \quad (6.18)$$

If $j \geq M + \sqrt{M}$, then $(j - M)\sqrt{M} \geq M$, so that (6.18) implies (6.17).

We expand each factor in Fourier modes in Θ . On the block with label ℓ ,

$$p(\zeta) - \bar{p} = c(\zeta) + w(\zeta) \sum_{o=\pm 1} e^{io(\Theta + \phi(\zeta))},$$

$$c = (l_{\text{out}} - \bar{l})(r_0 - p_0), \quad w = (-1)^{\ell-1} l_{\text{out}} h_n / 2,$$

and $|c| \leq r_+$ and $|w| \leq r_+/2$. We write $\cos(M\Theta) = \frac{1}{2} \sum_{\varepsilon=\pm 1} e^{i\varepsilon M\Theta}$, and we fix ε and a mode $o_i \in \{-1, 0, 1\}$ for each $i \in \{1, \dots, j\}$. We denote the corresponding term of Γ_j by $F_{\varepsilon, o}$; it is $A_u A_v l_{\text{in}}(x) l_{\text{in}}(y) \prod_{o_i=0} c(\zeta_i) \prod_{o_i \neq 0} w(\zeta_i)$ times the expectation of $S^2 e^{i\varepsilon M\Theta} \prod_{o_i \neq 0} e^{io_i(\Theta + \phi(\zeta_i))}$. Since $S^2 e^{i \sum_i o_i \phi(\zeta_i)} = \prod_{a,q} G_a(\zeta_{a,q})^2 e^{i \zeta_{a,q} \Phi_{a,q}}$, integrating over the uniform angle and the independent coefficients, we find that this expectation equals

$$\mathbf{1}_{\left\{ \sum_i o_i = -\varepsilon M \right\}} \prod_{a,q} \widehat{\Xi}_a(\Phi_{a,q}), \quad \Phi_{a,q} = \sum_{i: o_i \neq 0} o_i s_{a,q}(\zeta_i). \quad (6.19)$$

Since $|\sum_i o_i| \leq j$, it follows that $\Gamma_j = 0$ for $j < M$.

Now let $M \leq j < M + \sqrt{M}$. Since there are $2 \cdot 3^j$ terms $F_{\varepsilon, o}$, (6.17) follows from the bound

$$\|F_{\varepsilon, o}\|_1 \leq 4A_u A_v (2e r_+)^j R^{1-M} e^{C_E M}, \quad (6.20)$$

which we prove for a fixed pattern with $\sum_i o_i = -\varepsilon M$. Let $I_- = \{i : o_i = -\varepsilon\}$ and $I_+ = \{i : o_i = \varepsilon\}$, and let $k = |I_+|$. Then $|I_-| = M + k$, and since $|I_-| + |I_+| \leq j$, we have $2k \leq j - M < \sqrt{M} \leq M$. For a digit string $e^* = (e_{a,q}^*) \in \{0, 1\}^{J \times d}$ and $i \in I_- \cup I_+$, we define the target digits $t_{i,a,q} = e_{a,q}^*$ if $i \in I_-$ and $t_{i,a,q} = 1 - e_{a,q}^*$ if $i \in I_+$, and we let $\chi_{i,a,q} = |s_{a,q}(\zeta_i) - t_{i,a,q}|$. We claim that if the product in (6.19) does not vanish, then there is a string e^* such that

$$\sum_{i \in I_- \cup I_+} \chi_{i,a,q} \leq k + K_a \quad \text{for all } a, q. \quad (6.21)$$

Indeed, we fix a and q , and we let $X = -\varepsilon\Phi_{a,q} = \sum_{I_-} s_{a,q}(\zeta_i) - \sum_{I_+} s_{a,q}(\zeta_i)$. Since $0 \leq s_{a,q} \leq 1$, we have $-k \leq X \leq M+k$, so that $-M/2 < X < 3M/2$. Since $\widehat{\Xi}_a(\Phi_{a,q}) \neq 0$, Lemma 6.1 implies that there exists an integer l' with $|\Phi_{a,q} - l'M| \leq K_a$. With $l = -\varepsilon l'$, we have $|X - lM| \leq K_a \leq M/4$ by (6.4), and hence $l \in \{0, 1\}$. We set $e_{a,q}^* = l$. Since $|I_-| = M+k$ and $|I_+| = k$, the sum in (6.21) equals $X+k$ if $l = 0$ and $M+k-X$ if $l = 1$. In both cases it is at most $k + K_a$, which proves the claim. By (6.3), we also have $\chi_{i,a,q} \geq \frac{1}{2}\mathbf{1}\{e_a(\zeta_{i,q}) \neq t_{i,a,q}\}$, where $\zeta_{i,q}$ is the q th block coordinate of ζ_i .

Since the product in (6.19) has modulus at most one, its modulus is at most the sum, over the R strings e^* , of the indicator that the dJ inequalities (6.21) hold. We define $\Omega_a = 2\log d + 4\log(J-a+2)$ for $1 \leq a \leq J$ and $\Omega = d\sum_a \Omega_a$. We bound the indicator of each inequality $X' \leq c'$ by $e^{\Omega_a(c'-X')}$ and multiply over a and q . The indicator for a given e^* is therefore at most

$$e^{\Omega k} \exp\left\{d\sum_a \Omega_a K_a\right\} \prod_{i \in I_- \cup I_+} \exp\left\{-\sum_{a,q} \Omega_a \chi_{i,a,q}\right\}.$$

With $b = J-a$, we have $\Omega_a K_a = \{2\log d + 4\log(b+2)\}K(b)m$, and since $K(b)$ decays geometrically in b , we obtain $d\sum_a \Omega_a K_a \leq C_{\text{gate}}m$ with $C_{\text{gate}} = d\sum_{b \geq 0} \{2\log d + 4\log(b+2)\}K(b) < \infty$. Moreover, $\Omega \leq dJ\{2\log d + 4\log(J+1)\}$ and $m = 2^{J_{s_0}}$, so that $\Omega \leq D\sqrt{m}$ with $D = \sup_{J' \geq 1} dJ'\{2\log d + 4\log(J'+1)\}2^{-J'_{s_0}/2} < \infty$. Since $\bar{K} \geq K(0) = 6Q\lambda_*/\gamma_* > 1$, we have $c_* < 1$ and $m \leq M$. With $2k < \sqrt{M}$, we obtain $\Omega k + d\sum_a \Omega_a K_a \leq C_E M$, where $C_E = C_{\text{gate}} + D$. For $i \in I_- \cup I_+$, by the lower bound on $\chi_{i,a,q}$, the independence of the digits under Lebesgue measure, and $de^{-\Omega_a/2} = (b+2)^{-2}$, we have

$$\int_{\mathcal{U} \times \{1,2\}} |w(\zeta_i)| e^{-\sum_{a,q} \Omega_a \chi_{i,a,q}} d\zeta_i \leq r_+ \prod_{a,q} \frac{1 + e^{-\Omega_a/2}}{2} \leq \frac{r_+}{R} \exp\left\{\sum_{b \geq 0} \frac{1}{(b+2)^2}\right\} \leq \frac{er_+}{R}.$$

Each i with $o_i = 0$ contributes at most $2r_+$, and each of x and y at most 2. Summing over the R strings e^* , we obtain

$$\|F_{\varepsilon,0}\|_1 \leq 4A_u A_v R e^{C_E M} \left(\frac{er_+}{R}\right)^{M+2k} (2r_+)^{j-M-2k} \leq 4A_u A_v (2er_+)^j R^{1-M} e^{C_E M},$$

which proves (6.20). $\square$

Proof of Lemma 6.2(d). Let $I = [r_- + \delta_n, r_+ - \delta_n]$, so that $\rho_I = \rho_n$ in the notation of Lemma 4.1, with $c_I = r_0$, $d_I = h_n$, and $\bar{g}_I = \sqrt{r_0^2 - h_n^2} \leq r_0$. On the block with label ℓ , we have $p_{uv} = A_u A_v \sigma_u \sigma_v S_{\text{in}}^2 l_{\text{in}}^2 / r$ and $r = c_I + d_I \cos(\Theta + \phi + (\ell-1)\pi)$. Since the series (4.2) converges absolutely, multiplying it by $\cos(M\Theta)$ and averaging over Θ with ζ fixed keeps only its term of order M . We write E_Θ for this average. Since M is even, we obtain on both blocks

$$E_\Theta \frac{\cos(M\Theta)}{r} = \frac{\rho_n^M \cos(M\phi)}{\bar{g}_I}.$$

By (5.2), and since Θ is independent of ζ , each of the B blocks has volume v_0/B , and the density parameters of all pairs have the same law, we obtain

$$(E_+ - E_-) \int_{[0,1]^d} p_{uv} = \frac{2v_0 A_u A_v \rho_n^M}{\bar{g}_I} E \left[S^2 \int_{\mathcal{U}} \iota_{\text{in}}^2 \cos(M\phi) \right].$$

By (6.6), $\cos(M\phi) = 1$ on $\mathcal{U}^\circ$. Since $0 \leq \iota_{\text{in}} \leq 1$, (6.7) implies that the integral is at least $I_{\text{in}} - 2|\mathcal{U} \setminus \mathcal{U}^\circ| \geq I_{\text{in}}/2$ in every realization. The expectation is therefore at least $(I_{\text{in}}/2)ES^2 \geq 9I_{\text{in}}/32$, although S and ϕ are dependent. With $\bar{g}_I \leq r_0$, this gives (d) with $c_2 = 9v_0 I_{\text{in}}/(16r_0)$. $\square$

Embedding and proof of the lower bounds

7.1 Embedding the priors in the generic model

We perturb π_0 along scores that shift a, b by $u/w, v/w$ (both by $(u+v)/w$ in the diagonal case). We apply Proposition 5.3 to the resulting finite affine law. The applications use Proposition 7.1 with arbitrary density intervals and scores. As in the lower bounds of [49, Sections 3–4], where the responses are binary, we use a finite response space whose conditional laws are perturbed through the nuisance functions.

We write a finite-support law for Z as $\pi_0 = \sum_{k=1}^{m_0} \pi_{0,k} \delta_{z_k}$, with every $\pi_{0,k} > 0$, we let w_0, a_0, b_0 be as in (2.4), with $w_0 > 0$, and we define $r = (U - a_0 D, V - b_0 D)^\top$, which has mean zero under π_0 . We call bounded functions s_u, s_v on the support of π_0 scores for π_0 if $\mathbb{E}_{\pi_0} s_u = \mathbb{E}_{\pi_0} s_v = 0$ and either

$$\mathbb{E}_{\pi_0}[r(s_u, s_v)] = \text{Id}, \quad \text{or} \quad U = V, \quad s_u = s_v, \quad \mathbb{E}_{\pi_0}[(U - a_0 D)s_u] = 1, \quad (7.1)$$

where Id is the 2×2 identity matrix. We call the second alternative the *diagonal case*; in it we set $r_d = U - a_0 D$, and we call the common score s_d a *diagonal score*. If the residual covariance $\Sigma = \mathbb{E}_{\pi_0}[rr^\top]$ is positive definite, then $(s_u, s_v)^\top = \Sigma^{-1}r$ are scores. If $U = V$ and $\mathbb{E}_{\pi_0} r_d^2 > 0$, then $s_d = r_d / \mathbb{E}_{\pi_0} r_d^2$ is a diagonal score. Given scores, we define

$$\pi_{u,v}(\{z_k\}) = \pi_{0,k} \{1 + u s_u(z_k) + v s_v(z_k)\}. \quad (7.2)$$

These probabilities sum to one because the scores have mean zero, and, since the support is finite, they are at least $\pi_{0,k}/2$ when $|u| + |v|$ is below a fixed threshold. In the notation of (5.1), we have $q_{z_k 0} = \pi_{0,k}$, $q_{z_k u} = \pi_{0,k} s_u(z_k)$, and $q_{z_k v} = \pi_{0,k} s_v(z_k)$.

Let $(\ell_a, \ell_b) = (u, v)$, and $(\ell_a, \ell_b) = (u+v, u+v)$ in the diagonal case, and let $d_u = \mathbb{E}_{\pi_0}[D s_u]$ and $d_v = \mathbb{E}_{\pi_0}[D s_v]$. For every f , we have $\mathbb{E}_{\pi_{u,v}} f = \mathbb{E}_{\pi_0} f + u \mathbb{E}_{\pi_0}[f s_u] + v \mathbb{E}_{\pi_0}[f s_v]$. By $\mathbb{E}_{\pi_0} r = 0$ and (7.1), we obtain $\mathbb{E}_{\pi_{u,v}} r = (\ell_a, \ell_b)^\top$; in the diagonal case, $a_0 = b_0$ and both entries of r equal r_d . As $U = r_1 + a_0 D$ and $V = r_2 + b_0 D$, we find

$$w(u, v) = w_0 + d_u u + d_v v, \quad \mathbb{E}_{\pi_{u,v}} U = a_0 w(u, v) + \ell_a, \quad \mathbb{E}_{\pi_{u,v}} V = b_0 w(u, v) + \ell_b, \quad (7.3)$$

where $w(u, v) = \mathbb{E}_{\pi_{u,v}} D$. The regression ratios are therefore

$$a(u, v) = a_0 + \frac{\ell_a}{w(u, v)}, \quad b(u, v) = b_0 + \frac{\ell_b}{w(u, v)}. \quad (7.4)$$

Suppose that X has density p and the response law at x is $\pi_{u(x),v(x)}$. With $m_U = \mathbb{E}_{\pi_{u,v}} U$, $m_V = \mathbb{E}_{\pi_{u,v}} V$, and $w = w(u, v)$, we have the identity

$$\frac{m_U m_V}{w} = a_0 m_V + b_0 m_U - a_0 b_0 w + \frac{(m_U - a_0 w)(m_V - b_0 w)}{w}.$$

By (2.1) and (7.3), this identity implies that $\psi(P) = \int pF(u, v)$, where

$$F(u, v) = \mathbb{E}_{\pi_{u,v}} \tilde{W} + \lambda \frac{\ell_a \ell_b}{w(u, v)}, \quad \tilde{W} = W + \lambda(a_0 V + b_0 U - a_0 b_0 D). \quad (7.5)$$

The first term is affine in (u, v) , and F is smooth on a fixed neighborhood of the origin on which $w(u, v) \geq w_0/2$. Since $\ell_a \ell_b$ vanishes to second order at the origin, we have $F_{uv}(0, 0) = \lambda \partial_u \partial_v (\ell_a \ell_b) / w_0$, which is λ/w_0 or $2\lambda/w_0$ in the diagonal case, and is nonzero.

Proposition 7.1 (Local lower bounds). *Fix d, α, β with $\theta = (\alpha + \beta)/d$, a finite-support law π_0 for Z with $w_0 > 0$, the functions D, U, V, W , and $\lambda \neq 0$. Suppose that s_u, s_v are scores for π_0 , with $\alpha = \beta$ in the diagonal case, and let $\pi_{u,v}$ and F be as above. Fix $0 < r_- < 1 < r_+$ and $\epsilon > 0$, and let τ be computed from $[r_-, r_+]$ as in Section 5.2.*

- (a) *Suppose that $\theta < 1/2$. For all sufficiently large n there are two priors on laws P of $O = (X, Z)$, under which X has a C^∞ density p and Z given $X = x$ has law $\pi_{u(x),v(x)}$, such that every realization satisfies (6.10), $\|u\|_\infty \leq Cn^{-\alpha/d}$, $\|v\|_\infty \leq Cn^{-\beta/d}$ (so that $\|u\|_\infty + \|v\|_\infty = o(\delta_n)$), and, for every fixed smooth function K on a neighborhood N of $(0, 0)$, for all sufficiently large n depending on K ,*

$$\begin{aligned} K(0, v) = 0 \text{ on } N &\implies \|K(u, v)\|_{C^\alpha} \leq C_K \epsilon, \\ K(u, 0) = 0 \text{ on } N &\implies \|K(u, v)\|_{C^\beta} \leq C_K \epsilon, \end{aligned} \quad (7.6)$$

where C_K does not depend on ϵ or n . We call the laws under these priors the hard laws, and we write $\mathcal{P}_n^h$ for their set. On these laws $\psi(P) = \int pF(u, v)$. If $\mathcal{P}$ is a class that contains $\mathcal{P}_n^h$ and T is a parameter with $T(P) = \psi(P) + t_*$ on $\mathcal{P}_n^h$ for a constant t_* , then there is $c > 0$ such that, for all sufficiently large n ,

$$\inf_{\hat{T}} \sup_{P \in \mathcal{P}} P\{|\hat{T} - T(P)| \geq c a_n \log n\} \geq \frac{3}{8}, \quad r_n(T, \mathcal{P}) \geq \sqrt{3/8} c a_n \log n,$$

with $a_n = a_n(\theta, \tau)$. The infimum includes randomized estimators.

- (b) *For every u_* in a neighborhood of zero, the laws $P_t = \text{Unif}([0, 1]^d) \otimes \pi_{u_*, t}$, for t near zero, have $\psi(P_t) = F(u_*, t)$. For every sufficiently small $u_* \neq 0$ there is $c > 0$, depending also on u_* , such that, on every class $\mathcal{P}$ that contains P_t for all t in an interval about zero, a parameter T with $T(P_t) = \psi(P_t) + t_*$ for a constant t_* satisfies $r_n(T, \mathcal{P}) \geq cn^{-1/2}$ for all sufficiently large n .*

The constants may depend on the fixed quantities above and on ϵ .

Proof. For (a), we apply Proposition 5.3 to the response law (7.2), which has the form (5.1), with the integrand F and the interval $[r_-, r_+]$. Let M be as in that proposition, let c_* be the constant of Section 6.2, and let $s_0 = \min(\alpha, \beta)$. For all large n we have $c_*M \geq 1$, and we set

$$J = \left\lfloor \frac{\log_2(c_*M)}{s_0} \right\rfloor \geq 0, \quad m = 2^{Js_0}, \quad R = 2^{Jd},$$

so that $2^{-s_0}c_*M < m \leq c_*M$. Since $s_0 \leq \alpha, \beta$, we have $m = R^{s_0/d} \leq \min(R^{a_*}, R^{b_*})$, and $\log R = (d/s_0) \log m = O(\log M)$. Let $\epsilon_u = \epsilon_v = \min(\epsilon, 1)$. By Lemma 6.2(a), (c), and (d), for all large n and every $B \geq n$ that is the d th power of an even integer, the priors of Section 6 with these M, J, B, ϵ_u , and ϵ_v satisfy the hypotheses of Proposition 5.3, with constants that do not depend on n or B . By Proposition 5.3, there exists such a B with $A_u \leq \epsilon_u B^{-a_*} \leq n^{-\alpha/d}$ and $A_v \leq \epsilon_v B^{-b_*} \leq n^{-\beta/d}$, and the hard laws are the laws of one observation under the realizations of the two priors for this B . By Lemma 6.2(a) and (b), every realization satisfies (6.10), $\|u\|_\infty \leq A_u/r_-$, and $\|v\|_\infty \leq A_v/r_-$. Since $A_u + A_v \rightarrow 0$, the response probabilities are positive for all large n , and (6.12) implies (7.6) for all large n depending on K , since $\epsilon_u, \epsilon_v \leq \epsilon$. Since $n^{-\min(\alpha, \beta)/d} (\log n)^2 \rightarrow 0$, we have $\|u\|_\infty + \|v\|_\infty = o(\delta_n)$.

By (7.5), $\psi(P) = \int pF(u, v)$ on the hard laws. We apply Proposition 5.3 to $T - t_*$, whose estimation errors equal those of T , and this yields the threshold $c_1 m^2 a_n$ in (5.10). Since $m > 2^{-s_0}c_*M$ and $M \geq \sqrt{\theta(1-2\theta) \log n / \tau} - 1$, we have

$$m^2 \geq 2^{-2s_0-2} c_*^2 \frac{\theta(1-2\theta)}{\tau} \log n$$

for all large n , which gives the testing bound with the threshold $c a_n \log n$ for $c = 2^{-2s_0-2} c_*^2 \theta(1-2\theta) c_1 / \tau$. The mean squared error is at least the squared threshold times the error probability, which gives the RMSE bound. With $J = 0$ in place of this choice, so that $m = R = 1$, these bounds hold with a_n in place of $a_n \log n$, by the same argument.

For (b), the constant perturbations $u = u_*$ and $v = t$, with $p = 1$, give the law P_t , so that $\psi(P_t) = F(u_*, t)$. Let $\varphi(u) = F_v(u, 0)$. Since $\varphi'(0) = F_{uv}(0, 0) \neq 0$, we have $\varphi(u_*) \neq 0$ for every sufficiently small $u_* \neq 0$, by continuity if $\varphi(0) \neq 0$, and since $\varphi(u) = u\{\varphi'(0) + o(1)\}$ otherwise. Relative to $P_* = \text{Unif}([0, 1]^d) \otimes \pi_0$, the law P_t has the density $1 + u_* s_u(Z) + t s_v(Z)$, which is at least $1/2$ when $|u_*| + |t|$ is below a fixed threshold and has the bounded slope $s_v(Z)$ in t . On an open interval about zero on which $P_t \in \mathcal{P}$, the map $t \mapsto T(P_t) = F(u_*, t) + t_*$ has the derivative $\varphi(u_*) \neq 0$ at $t = 0$. The bound now follows from Lemma 5.1. $\square$

7.2 Localization

Every conditional moment of a hard law is affine in (u, v) : $\mathbb{E}_P[f(Z) \mid X] = \mathbb{E}_{\pi_0} f + u \mathbb{E}_{\pi_0}[f s_u] + v \mathbb{E}_{\pi_0}[f s_v]$. Every conditional-response nuisance used below, such as a condi-

tional mean, a ratio of conditional means whose denominator is positive at the baseline, a propensity, or its reciprocal, is therefore of the form $K_P(x) = K(u(x), v(x))$ for a C^∞ function K near the origin.

Lemma 7.2 (Localization and membership). *Adopt the setting of Proposition 7.1. Let K be a C^∞ function on a neighborhood of $(0,0)$, let $k_0 = K(0,0)$, and for a law P with perturbations u, v let $K_P(x) = K(u(x), v(x))$. Let $\gamma_K = \alpha$ if $K(0, v) = k_0$ for all small v , let $\gamma_K = \beta$ if $K(u, 0) = k_0$ for all small u , and let $\gamma_K = \min(\alpha, \beta)$ otherwise. If both identities hold, either value may be used. There is C_K , independent of n and ϵ , such that, for all sufficiently large n , every hard law P satisfies*

$$\|K_P - k_0\|_\infty \leq C_K(\|u\|_\infty + \|v\|_\infty) = o(\delta_n), \quad \|K_P - k_0\|_{C^t} \leq C_K \epsilon \quad (0 < t \leq \gamma_K),$$

and, if $k_0 > 0$, also $r_-k_0 \leq pK_P \leq r_+k_0$. For the laws P_t of Proposition 7.1(b), K_{P_t} is the constant $K(u_*, t)$, which tends to k_0 as $(u_*, t) \rightarrow 0$.

Consider finitely many conditions on (p, u, v) of the following three types, each with its own C^∞ function K and $k_0 = K(0,0)$:

(H) $\|K_P\|_{C^t} \leq H'$, with $0 < t \leq \gamma_K$ and $|k_0| < H'$;

(I) $k_- \leq K_P \leq k_+$, with $k_- < k_0 < k_+$;

(W) $g'_- \leq pK_P \leq g'_+$, with $k_0 > 0$, $g'_- \leq r_-k_0$, and $r_+k_0 \leq g'_+$.

If ϵ is small enough, depending only on these conditions, then for all large n every hard law satisfies all of them. Moreover, for every sufficiently small u_* , the laws P_t , which have $p = 1$, satisfy all of them for t in an interval about zero; this assertion does not use $\theta < 1/2$.

Proof. Shrinking the domain of K , we may assume that it is an open square N centered at the origin, so that $(u, 0)$ and $(0, v)$ lie in N whenever (u, v) does, and that K is Lipschitz on N with a constant C_K . By Proposition 7.1(a), we have $\|u\|_\infty + \|v\|_\infty \leq Cn^{-\min(\alpha, \beta)/d} = o(\delta_n)$, uniformly over the hard laws, which gives the bound in the supremum norm.

For the Hölder bounds, we write

$$K(u, v) - k_0 = \{K(u, v) - K(0, v)\} + \{K(0, v) - k_0\}$$

on N . The first brace vanishes when $u = 0$, and the second when $v = 0$. By (7.6), for all large n , their compositions with (u, v) have norms at most $C\epsilon$ in C^α and in C^β , respectively, where C does not depend on ϵ . If $K(0, v) = k_0$, only the first brace remains, which gives the bound in C^α . If $K(u, 0) = k_0$, then $K - k_0$ itself vanishes when $v = 0$, and the bound in C^β follows from (7.6). In general, both braces are bounded in $C^{\min(\alpha, \beta)}$, since $\|f\|_{C^t} \leq (1+d)\|f\|_{C^{t'}}$ for $0 < t \leq t'$ and f on $[0, 1]^d$. This follows from $|x - y| \leq \sqrt{d}$ on $[0, 1]^d$ and, if $\lceil t \rceil < \lceil t' \rceil$, from the mean value theorem, which shows that the derivatives of order $\lceil t \rceil - 1$ are Lipschitz with constant $\sqrt{d}\|f\|_{C^{t'}}$ on the convex set $[0, 1]^d$. Using the same comparison and enlarging C_K , we find that the bounds hold for every $0 < t \leq \gamma_K$.

Let $k_0 > 0$ and $\eta = \|K_P - k_0\|_\infty$. Since $\eta = o(\delta_n)$, for all large n we have $K_P > 0$ and $r_+\eta \leq k_0\delta_n$, and (6.10) gives

$$pK_P \geq (r_- + \delta_n)(k_0 - \eta) \geq r_-k_0 + k_0\delta_n - r_+\eta \geq r_-k_0,$$

and in the same way $pK_P \leq (r_+ - \delta_n)(k_0 + \eta) \leq r_+k_0 - k_0\delta_n + r_+\eta \leq r_+k_0$.

Since a constant has Hölder norm equal to its absolute value, we obtain (H) from $\|K_P\|_{C^t} \leq |k_0| + C_K\epsilon \leq H'$, once $C_K\epsilon \leq H' - |k_0|$. The bound in the supremum norm implies (I), and the bounds on pK_P imply (W). For the laws P_t , we have $p = 1$ and $K_{P_t} = K(u_*, t)$, a constant. The three conditions hold strictly at $(u_*, t) = (0, 0)$, for (W) because $g'_- \leq r_-k_0 < k_0 < r_+k_0 \leq g'_+$. By continuity, they hold for (u_*, t) near zero. $\square$

Proof of Theorem 2.3(b). Since the minimax risk and the minimax error probability do not decrease when the class grows, it suffices to work on $\mathcal{P}_r(\pi_0)$. We use the scores $\Sigma^{-1}r$ in the positive definite case of (2.5), and the diagonal score $r_d/\mathbb{E}_{\pi_0}r_d^2$ in the diagonal case; both exist by (2.5). Let $r_- = g_-/w_0$ and $r_+ = g_+/w_0$, so that $0 < r_- < 1 < r_+$ by (2.5). Since scaling both endpoints of an interval by the same factor leaves ρ unchanged, and since $a_* + b_* = \theta$, the τ and a_n of Proposition 7.1 are those of (2.7) and (2.9).

Under a hard law or a law P_t , the functions w, a, b of Section 2 are K_P for the functions $w(u, v), a(u, v), b(u, v)$ of (7.3) and (7.4), with $k_0 = w_0, a_0, b_0$, and $g = pw$. Membership in $\mathcal{P}_r(\pi_0)$ consists of the conditions $\max(\delta, w_0 - r) \leq w \leq w_0 + r$, $|a - a_0| \leq r$, and $|b - b_0| \leq r$, of type (I); $g_- \leq pw \leq g_+$, of type (W) with $r_\pm w_0 = g_\pm$; and $\|a\|_{C^\alpha} \leq H$ and $\|b\|_{C^\beta} \leq H$, of type (H). By (2.5), the baseline values satisfy the conditions of types (I) and (H) strictly. For (H), we may take $\gamma_K = \alpha$ for a and $\gamma_K = \beta$ for b , since $a(0, v) = a_0$ and $b(u, 0) = b_0$ in the positive definite case, and $\alpha = \beta$ in the diagonal case. We fix ϵ as in Lemma 7.2 for these conditions. Then, for all large n , every hard law belongs to $\mathcal{P}_r(\pi_0)$, and, for every sufficiently small $u_* \neq 0$, so does P_t for every t in an interval about zero.

If $0 < \theta < 1/2$, we apply Proposition 7.1(a) with $\mathcal{P} = \mathcal{P}_r(\pi_0)$, $T = \psi$, and $t_* = 0$, and we find $c_0 > 0$ such that, for all large n , the minimax probability of an error of at least $c_0 a_n \log n$ is at least $3/8$, and $r_n(\psi, \mathcal{P}_r(\pi_0)) \geq \sqrt{3/8} c_0 a_n \log n$. With $c = c_0/2$, this gives (2.11) and the lower bound in (2.13), since $3/8 \geq 1/4$ and $\sqrt{3/8} \geq 1/2$. For every $\theta > 0$, Proposition 7.1(b), with $\mathcal{P} = \mathcal{P}_r(\pi_0)$ and $T = \psi$, implies that there exists $c' > 0$ such that $r_n(\psi, \mathcal{P}_r(\pi_0)) \geq c' n^{-1/2}$ for all large n , which is the lower bound in (2.12) when $\theta \geq 1/2$. For $\theta \geq 1/2$, we take $c = c'$ and the threshold of the root- n argument. For $0 < \theta < 1/2$, we take $c = \min\{c_0/2, c'\}$ and the larger of the two thresholds for n . This gives all the assertions. $\square$

Conclusion

First, when $s < d/4$, the bounds of Theorem 2.3 differ by the factor $(\log n)^{\nu/2-3/4}$, which is $(\log n)^{1/4}$ when $0 < \alpha, \beta \leq 1$. In the estimator, the factor $(\log n)^{1/4}$ comes from the number of resolutions near the dominant one (Section 1.3). Each of our lower-bound constructions, in contrast, places its bumps at a single resolution. Closing this gap would determine the minimax rate up to constants. The power of $\log n$ is also left open in [46].

Second, our estimator uses α, β , the bounds g_-, g_+ , and the remaining class parameters. McClean et al. [43] and Park [46, Corollary 1] attain the rough-design exponent for the expected conditional covariance without knowing the nuisance smoothness, and adaptive results for related functionals under smooth density assumptions were obtained in [39]. Park also reports an estimator that does not use α, β and attains his upper bound up to a larger power of $\log n$, uniformly over smoothness indices in a known compact subset of $\{\alpha + \beta < d/2\}$ [46, Section 3.4]. This estimator is analyzed in Theorem S13 of the supplement of [46]. For the general class, it is open whether the upper bound of Theorem 2.3, including the correction, can be attained without knowledge of α, β . In the setting of [46], it is open whether this is possible without the larger power of $\log n$.

Third, minimax rates remain unresolved when g belongs to a Hölder class of positive smoothness that fails the condition stated before (1.1). Park [46, Section 6.2] bounds the risk of his selected estimator of the expected conditional covariance at each fixed Hölder density, and he notes that this is not a minimax result over a class of smooth densities. The question of how the minimax rate depends on the smoothness of the covariate density is also described as open in [76, Remark 5]. When the two nuisance functions have equal smoothness indices, the conjecture following equation (1.3) of Robins et al. [48] addresses this case; our results confirm the exponent it suggests in the rough-design limit.

Reductions

Most targets of Section 3 are related to a target of Theorem 2.3 through an observable mean, a Lipschitz map, or a transformation of the data that does not depend on the unknown law. We record the lower-bound constructions and these three reductions.

A.1 Hard laws

Definition A.1 (Hardness). Let $\mathcal{P}_n$ be classes of laws of one observation, let T be a real parameter defined on each $\mathcal{P}_n$, and let $t_n > 0$. We say that T is *hard at scale t_n on $(\mathcal{P}_n)$* if

$$\liminf_{n \rightarrow \infty} \inf_{\hat{T}} \sup_{P \in \mathcal{P}_n} \mathbb{P}_P\{|\hat{T} - T(P)| \geq t_n\} \geq \frac{3}{8},$$

where, for each n , the infimum is over all estimators, including randomized ones, based on n independent observations.

Hardness at scale t_n implies hardness at scale ct_n for every fixed $c \in (0, 1]$, and on every sequence of classes $\mathcal{P}'_n \supseteq \mathcal{P}_n$.

Lemma A.2 (From hardness to risk). *Let T be hard at scale t_n on $(\mathcal{P}_n)$ and a real parameter on a class $\mathcal{P}$. Suppose that $\mathcal{P}_n \subseteq \mathcal{P}$ for all large n . Then, for all large n ,*

$$\inf_{\hat{T}} \sup_{P \in \mathcal{P}} \mathbb{P}_P\{|\hat{T} - T(P)| \geq t_n\} \geq \frac{1}{4}, \quad r_n(T, \mathcal{P}) \geq \frac{t_n}{2}. \quad (\text{A.1})$$

The infimum includes randomized estimators. In particular, if $0 < \theta < 1/2$ and $t_n = c_1 \underline{a}_n(\theta, \tau_1)$ for a fixed $c_1 > 0$ and an interval I as in Definition 3.1, then $(T, \mathcal{P})$ satisfies the lower bracket (θ, I) with $c = c_1/2$.

Proof. By the lim inf bound, the minimax probability is at least $1/4$ over $\mathcal{P}_n$, and hence over $\mathcal{P} \supseteq \mathcal{P}_n$, for all large n . For every estimator, we have $\sup_{P \in \mathcal{P}} \mathbb{E}_P(\hat{T} - T(P))^2 \geq t_n^2 \sup_{P \in \mathcal{P}} \mathbb{P}_P\{|\hat{T} - T(P)| \geq t_n\}$, which gives the RMSE bound and, with $t_n = c_1 \underline{a}_n = 2c \underline{a}_n$, the lower bracket. $\square$

We fix the inputs of Proposition 7.1: a finite-support law π_0 for the response, the functions D, U, V, W and λ , scores for π_0 (in the diagonal case, one function s_d with

$\mathbb{E}_{\pi_0} s_d = 0$ and $\mathbb{E}_{\pi_0}[(U - a_0 D)s_d] = 1$, used for both perturbations), smoothness indices α, β with $\theta = (\alpha + \beta)/d < 1/2$, an interval $[r_-, r_+]$ with $0 < r_- < 1 < r_+$, and $\epsilon > 0$. By Proposition 7.1(a), there exists a fixed $c_1 > 0$ such that, for every constant t_* , the parameter $\psi + t_*$ is hard at scale $c_1 a_n(\theta, \tau) \log n$ on the hard laws $(\mathcal{P}_n^h)$, where τ is computed from $[r_-, r_+]$ and ψ is the target of Section 2 for (D, U, V, W, λ) . On a hard law, the response law at x is the perturbation of π_0 by $(u(x), v(x))$, and $\|u\|_\infty \leq Cn^{-\alpha/d}$, $\|v\|_\infty \leq Cn^{-\beta/d}$. For a fixed finite family of restrictions of the types (H), (I), and (W) of Lemma 7.2, we take ϵ sufficiently small. Then, by that lemma, every hard law satisfies them for all large n , and the laws P_t of Proposition 7.1(b) satisfy them for every sufficiently small u_* and all t in an interval about zero. The reductions preserve hardness, and the lower brackets on the classes of Section 3 follow from Lemma A.2.

A.2 Three reductions

Let h be a bounded measurable map from the observation space into $\mathbb{R}^k$, with $k \geq 0$, and write $\mu(P) = \mathbb{E}_P h(O)$. Let $\mathcal{R}$ be a compact rectangle that contains $\mu(P)$ for every law P that we consider, and let $b_h^2 = \sup_P \mathbb{E}_P \|h(O) - \mu(P)\|_2^2$. When $k = 0$, the terms with h and μ are absent and $b_h = 0$. Let $I_1, \dots, I_m$ be compact intervals. A map Φ on $I_1 \times \dots \times I_m \times \mathcal{R}$ is L -Lipschitz if it is Lipschitz with constant $L \geq 1$ for the sum of the absolute differences of the real arguments plus the Euclidean distance in $\mathcal{R}$.

Lemma A.3 (Reductions).

(a) (Combination.) Suppose that $T_1, \dots, T_m$ are real parameters on a class $\mathcal{P}$ with $T_i(P) \in I_i$, and that $T_0(P) = \Phi(T_1(P), \dots, T_m(P), \mu(P))$ for all $P \in \mathcal{P}$, where Φ is L -Lipschitz. Then, for every $n \geq 1$,

$$r_n(T_0, \mathcal{P}) \leq L \sum_{i=1}^m r_n(T_i, \mathcal{P}) + L b_h n^{-1/2}.$$

(b) (Reverse relations.) Let $m = 1$, let $\mathcal{P}'$ be a class, and suppose that $T_0(P) \in I_1$ and

$$|T_1(P) - \Phi(T_0(P), \mu(P))| \leq e \quad (P \in \mathcal{P}'), \quad (\text{A.2})$$

where Φ is L -Lipschitz and $e \geq 0$. Then, for every $n \geq 1$ and every $t > 0$ with $t \geq 2e$,

$$r_n(T_1, \mathcal{P}') \leq L r_n(T_0, \mathcal{P}') + L b_h n^{-1/2} + e, \quad (\text{A.3})$$

$$\inf_{\hat{T}_0} \sup_{P \in \mathcal{P}'} \mathbb{P}_P \left\{ |\hat{T}_0 - T_0(P)| \geq \frac{t}{4L} \right\} \geq \inf_{\hat{T}_1} \sup_{P \in \mathcal{P}'} \mathbb{P}_P \{ |\hat{T}_1 - T_1(P)| \geq t \} - \frac{16L^2 b_h^2}{nt^2}. \quad (\text{A.4})$$

Consequently, suppose that for all large n these hypotheses hold with $\mathcal{P}' = \mathcal{P}'_n$ and $e = e_n \leq t_n/2$, where h , $\mathcal{R}$, I_1 , and Φ do not depend on n . If T_1 is hard at scale t_n on $(\mathcal{P}'_n)$, and if $k = 0$ or $nt_n^2 \rightarrow \infty$, then T_0 is hard at scale $t_n/(4L)$ on $(\mathcal{P}'_n)$.

(c) (Kernels.) Let K be a Markov kernel from the observation space of a class $\mathcal{P}_0$ to that of a class $\mathcal{P}_1$ that does not depend on the unknown law, and let KP be the law of the output when the input has law P . Suppose that $KP \in \mathcal{P}_1$ and $T_1(KP) = T_0(P)$ for every $P \in \mathcal{P}_0$. Then, for every $n \geq 1$ and $t > 0$,

$$\inf_{\hat{T}_0} \sup_{P \in \mathcal{P}_0} \mathbb{P}_P\{|\hat{T}_0 - T_0(P)| \geq t\} \leq \inf_{\hat{T}_1} \sup_{Q \in \mathcal{P}_1} \mathbb{P}_Q\{|\hat{T}_1 - T_1(Q)| \geq t\},$$

and $r_n(T_0, \mathcal{P}_0) \leq r_n(T_1, \mathcal{P}_1)$. In particular, upper bounds for T_1 on $\mathcal{P}_1$ hold for T_0 on $\mathcal{P}_0$, lower bounds and probability bounds for T_0 on $\mathcal{P}_0$ hold for T_1 on $\mathcal{P}_1$, and if T_0 is hard at scale t_n on $(\mathcal{P}_{0,n})$, then T_1 is hard at scale t_n on $(K\mathcal{P}_{0,n})$.

Proof. Let $\hat{\mu}$ be the coordinatewise projection onto $\mathcal{R}$ of the empirical mean of $h(O_1), \dots, h(O_n)$, and let Π_i be the projection onto I_i . Since these projections do not increase distance to points of $\mathcal{R}$ and I_i , we have $\mathbb{E}_P\|\hat{\mu} - \mu(P)\|_2^2 \leq b_h^2/n$. Suppose that real parameters $S, S_1, \dots, S_m$ satisfy $S_i(P) \in I_i$ and $|S(P) - \Phi(S_1(P), \dots, S_m(P), \mu(P))| \leq e$. Given estimators $\hat{S}_i$, the estimator $\hat{S} = \Phi(\Pi_1\hat{S}_1, \dots, \Pi_m\hat{S}_m, \hat{\mu})$ satisfies

$$|\hat{S} - S(P)| \leq L \left\{ \sum_{i=1}^m |\hat{S}_i - S_i(P)| + \|\hat{\mu} - \mu(P)\|_2 \right\} + e.$$

With $S = T_0$, $S_i = T_i$, and $e = 0$, (a) follows from Minkowski's inequality in $L^2(P^n)$ by taking suprema and infima, without independence between estimators. With $m = 1$, $S = T_1$, and $S_1 = T_0$, the same argument on $\mathcal{P}'$ proves (A.3). If moreover $|\hat{S} - T_1(P)| \geq t \geq 2e$, the display implies that $|\hat{T}_0 - T_0(P)| \geq t/(4L)$ or $\|\hat{\mu} - \mu(P)\|_2 \geq t/(4L)$, and by Chebyshev's inequality the latter event has probability at most $16L^2b_h^2/(nt^2)$. Since $\hat{S}$ is an estimator of T_1 , (A.4) follows by taking the supremum over $P \in \mathcal{P}'$ and then the infimum over $\hat{T}_0$. Applying it with $t = t_n$ and taking the $\liminf$ gives the hardness assertion, since the error term vanishes when $k = 0$ and otherwise tends to zero by assumption.

For (c), we apply K independently to the n observations, with randomness independent of the sample, and evaluate $\hat{T}_1$ on the outputs. Since, under P , the outputs have law $(KP)^n$ and $T_1(KP) = T_0(P)$, the resulting error has the law of the error of $\hat{T}_1$ under $KP \in \mathcal{P}_1$. The suprema and infima give both inequalities and the remaining assertions. $\square$

Lemma A.3(c) is the elementary direction of Le Cam's comparison of experiments by randomization [32, Chapter 2]: an experiment is at least as informative as its image under a Markov kernel that does not depend on the unknown law. Composing an estimator for the image with the kernel, we obtain an estimator whose error has the same law. This gives both the inequality for error probabilities and the inequality for risks.

For a ratio we use (a) with $\Phi(x, y) = x/y$ on $[-C_0, C_0] \times [\delta_{\mathcal{D}}, C_0]$, which is L -Lipschitz with $L = \max\{1, \delta_{\mathcal{D}}^{-1}, C_0\delta_{\mathcal{D}}^{-2}\}$, since $|x/y - x'/y'| \leq |x - x'|/\delta_{\mathcal{D}} + C_0|y - y'|/\delta_{\mathcal{D}}^2$. A target

that differs from another by a sign and an additive constant has the same estimation errors, so that all bounds transfer between the two.

Remark A.4 (Combining rates). For fixed τ and ν , and $0 < \theta < 1/2$, we have $\bar{a}_n(\theta, \tau, \nu) = n^{-\theta} e^{O(\sqrt{\log n})}$, since $(\log n)^c = e^{o(\sqrt{\log n})}$ for every fixed c . Hence, if $\theta_1 < \min(\theta_2, 1/2)$, then $\bar{a}_n(\theta_2, \tau, \nu_2) + n^{-1/2} = o(\bar{a}_n(\theta_1, \tau, \nu_1))$, and if $\theta_1 = \theta_2$ and $\nu_1 \leq \nu_2$, then $\bar{a}_n(\theta_1, \tau, \nu_1) \leq \bar{a}_n(\theta_2, \tau, \nu_2)$. For $\theta \geq 1/2$ all these rates are $n^{-1/2}$. It follows from Lemma A.3(a) that if each $(T_i, \mathcal{P})$ satisfies the upper bracket (θ_i, I, ν_i) with a common interval I , then $(T_0, \mathcal{P})$ satisfies the upper bracket (θ, I, ν) with $\theta = \min_i \theta_i$ and $\nu = \max\{\nu_i : \theta_i = \theta\}$.

Missing-data means and treatment effects

The treatment effects inherit lower bounds below the threshold from one MAR family through a kernel with the observed regression in one arm and the constant $1/2$ in the other. Their arm means inherit upper bounds for the MAR mean by keeping one arm.

B.1 The MAR family and two maps

We take $Z = (R, RY)$ with $R, Y \in \{0, 1\}$ and $(D, U, V, W, \lambda) = (R, 1, RY, 0, 1)$, and we let π_0 give $(R, RY) = (0, 0), (1, 0), (1, 1)$ the probabilities $1/2, 1/4, 1/4$. Then $(w_0, a_0, b_0) = (1/2, 2, 1/2)$, and the residual $r = (1 - 2R, RY - R/2)^\top$ takes the values $(1, 0)$, $(-1, -1/2)$, and $(-1, 1/2)$ at the three support points, so that $\Sigma = \text{Cov}_{\pi_0}(r) = \text{diag}(1, 1/8)$. Since Σ is positive definite, the functions $(s_u, s_v)^\top = \Sigma^{-1}r = (1 - 2R, 8RY - 4R)^\top$ are scores for π_0 , as in Section 7.1. The perturbed response law $\pi_{u,v}(\{z\}) = \pi_0(\{z\})\{1 + us_u(z) + vs_v(z)\}$ is

$$\begin{aligned} P(R = 0 \mid X) &= \frac{1+u}{2}, \\ P(R = 1, Y = 0 \mid X) &= \frac{1-u}{4} - v, \quad P(R = 1, Y = 1 \mid X) = \frac{1-u}{4} + v, \end{aligned} \tag{B.1}$$

so that

$$w = \frac{1-u}{2}, \quad \frac{1}{w} = \frac{2}{1-u}, \quad \frac{1}{1-w} = \frac{2}{1+u}, \quad b = \frac{1}{2} + \frac{2v}{1-u}, \tag{B.2}$$

and $\psi(P) = \int abg = \mathbb{E}_P[b(X)]$ is the MAR mean. These observed laws are those of full-data laws in which Y is Bernoulli with mean $b(X)$ and independent of R given X . The lower bound for the missing-data mean in [49, Section 3] uses binary models of this kind.

For indices (α, β') with $(\alpha + \beta')/d < 1/2$ and an interval $[r_-, r_+]$ with $0 < r_- < 1 < r_+$, we call the hard laws of Appendix A.1 for this π_0 and these scores the *MAR family*. The MAR mean is hard on it at scale $c_1 \underline{a}_n((\alpha + \beta')/d, \tau)$, with τ computed from $[r_-, r_+]$. In the notation of Lemma 7.2, the functions $w, 1-w, 1/w$, and $1/(1-w)$ are of the form K_P with $K(0, v)$ constant, so that $\gamma_K = \alpha$, and b is of the form K_P with $K(u, 0) = 1/2$, so that $\gamma_K = \beta'$.

For $j \in \{0, 1\}$, the map that keeps arm j sends (X, A, Y) to (X, R, RY) with $R = \mathbf{1}_{\{A=j\}}$. If the input has law P , then the output has covariate density p , observation probability

w_j , and regression b_j , so that its MAR mean is $\mu_j(P)$. The kernel K_j sends (X, R, RY) to (X, A, Y') with

$$A = R \text{ if } j = 1, \quad A = 1 - R \text{ if } j = 0, \quad Y' = RY + (1 - R)\xi,$$

where $\xi \sim \text{Bernoulli}(1/2)$ is drawn independently. If the input has covariate density p , observation probability w , and regression b , then the output has covariate density p , $P(A = j | X) = w$, $P(A = 1 - j | X) = 1 - w$, $\mathbb{E}[Y' | A = j, X] = b$, and $\mathbb{E}[Y' | A = 1 - j, X] = 1/2$. In particular, $\mu_j(K_j P)$ is the MAR mean of P , $\mu_{1-j}(K_j P) = 1/2$, and

$$\mu_j(K_j P) = \frac{1}{2} + (2j - 1) T_{\text{ATE}}(K_j P). \quad (\text{B.3})$$

If w and b are constant and $p = 1$, then all regressions of $K_j P$ are constant, and

$$T_{\text{ATE}}(K_j P) = T_{\text{ATT}}(K_j P) = T_{\text{ATU}}(K_j P) = (2j - 1) \left(\psi(P) - \frac{1}{2} \right). \quad (\text{B.4})$$

On every law of treatment observations, we write $p_A = \mathbb{E}_P A$. Since $\mathbb{E}_P Y - \mu_0 = \int w(b_1 - b_0)p = p_A T_{\text{ATT}}$, and similarly for the ATU, we have

$$\begin{aligned} T_{\text{ATT}} &= \frac{\mathbb{E}_P Y - \mu_0}{p_A}, \quad \mu_0 = \mathbb{E}_P Y - p_A T_{\text{ATT}}, \\ T_{\text{ATU}} &= \frac{\mu_1 - \mathbb{E}_P Y}{1 - p_A}, \quad \mu_1 = \mathbb{E}_P Y + (1 - p_A) T_{\text{ATU}}. \end{aligned} \quad (\text{B.5})$$

On $\mathcal{T}_\eta$ and $\mathcal{T}^{\text{sep}}$, the overlap condition implies that $p_A \in [\eta, 1 - \eta]$ and $p_A \in [\epsilon, 1 - \epsilon]$, respectively. With $h(O) = (Y, A)$ and $\mathcal{R} = [0, 1] \times [\eta, 1 - \eta]$ or $[0, 1] \times [\epsilon, 1 - \epsilon]$, the maps on the left of (B.5) are Lipschitz on $[0, 1] \times \mathcal{R}$, as functions of μ_j and $(\mathbb{E}_P Y, p_A)$, and those on the right are Lipschitz on $[-1, 1] \times \mathcal{R}$, as functions of the effect and $(\mathbb{E}_P Y, p_A)$.

B.2 A fixed-accuracy pilot

The model $\mathcal{M}^{\text{sep}}$ bounds $g = wp$ only by ϵp_- and $(1 - \epsilon)p_+$. We divide the responses by a pilot estimate $\hat{w}$ of w from an independent half of the sample. Estimating nuisance functions on a separate subsample is standard in this literature; see the training samples of [48] and the cross-fitting of [45]. Let $\mathcal{W} = \{f \in \mathcal{H}^\alpha(H_0) : \epsilon \leq f \leq 1 - \epsilon\}$.

Lemma B.1 (A fixed-accuracy pilot). *Fix d, α, H_0 , $0 < \epsilon < 1/2$, $p_- > 0$, and $\zeta_0 \in (0, 1)$. There are $H_1 < \infty$, $c > 0$, and, for each $n_1 \geq 1$, a map $\hat{w}$ from n_1 observations of (X, R) to C^∞ functions on $[0, 1]^d$, such that $(x, \text{data}) \mapsto \hat{w}(x)$ is measurable, $\epsilon \leq \hat{w} \leq 1 - \epsilon$, $\|\hat{w}\|_{C^\alpha} \leq H_1$, and*

$$\mathbb{P}_P\{\|\hat{w} - w\|_\infty > \zeta_0\} \leq c^{-1} e^{-cn_1}$$

for every law of (X, R) such that $R \in \{0, 1\}$, the covariate density is at least p_- , and $w = \mathbb{E}[R | X] \in \mathcal{W}$.

Proof. Let $\gamma = \min(\alpha, 1)$ and $C_0 = \sqrt{d} H_0$. Every $f \in \mathcal{W}$ satisfies $|f(x) - f(y)| \leq C_0 \|x - y\|_2^\gamma$, by its Hölder bound if $\alpha \leq 1$ and its gradient bound if $\alpha > 1$. We fix an integer k with $C_0(2\sqrt{d}/k)^\gamma \leq \zeta_0/4$, we partition $[0, 1]^d$ into k^d congruent cubes Q , and we let Q^+ be the set of points of $[0, 1]^d$ at ℓ^∞ -distance at most $1/(2k)$ from Q , so that $\text{diam } Q^+ \leq 2\sqrt{d}/k$. We fix a C^∞ partition of unity (φ_Q) of $[0, 1]^d$ with $\varphi_Q \geq 0$ and $\text{supp } \varphi_Q \subseteq Q^+$; for instance, products of one-dimensional ones. Let N_Q be the number of observations with $X_i \in Q$, let $\tilde{w}_Q$ be the average of their R_i , with $\tilde{w}_Q = 1/2$ if $N_Q = 0$, let $\hat{w}_Q$ be $\tilde{w}_Q$ clipped to $[\epsilon, 1 - \epsilon]$, and let $\hat{w} = \sum_Q \hat{w}_Q \varphi_Q$. Then $\hat{w}$ is C^∞ , depends measurably on the data, takes values in $[\epsilon, 1 - \epsilon]$ as a convex combination, and satisfies $\|\hat{w}\|_{C^\alpha} \leq H_1 = \sum_Q \|\varphi_Q\|_{C^\alpha}$, which depends only on k, d , and α .

Let $\pi_Q = \mathbb{P}(X \in Q) \geq p_- k^{-d}$ and $m_Q = \mathbb{E}[R \mid X \in Q]$. Since m_Q averages w over Q , we have $m_Q \in [\epsilon, 1 - \epsilon]$ and $|m_Q - w(x)| \leq \zeta_0/4$ for $x \in Q^+$. Since clipping does not increase $|\tilde{w}_Q - m_Q|$, and $\varphi_Q(x) = 0$ unless $x \in Q^+$, we have

$$\|\hat{w} - w\|_\infty \leq \max_Q |\tilde{w}_Q - m_Q| + \zeta_0/4.$$

Conditionally on $N_Q = s \geq 1$, the sum of the R_i with $X_i \in Q$ is Binomial(s, m_Q), and Hoeffding's inequality [20, Theorem 1] implies that $\mathbb{P}\{|\tilde{w}_Q - m_Q| > \zeta_0/2 \mid N_Q = s\} \leq 2e^{-s\zeta_0^2/2}$, a bound that also holds for $s = 0$. Since $N_Q \sim \text{Binomial}(n_1, \pi_Q)$, we obtain

$$\mathbb{P}\{|\tilde{w}_Q - m_Q| > \zeta_0/2\} \leq 2(1 - \pi_Q + \pi_Q e^{-\zeta_0^2/2})^{n_1} \leq 2e^{-n_1 \pi_Q (1 - e^{-\zeta_0^2/2})}.$$

A union bound over the k^d cells gives $\mathbb{P}\{\|\hat{w} - w\|_\infty > \zeta_0\} \leq 2k^d e^{-n_1 p_- k^{-d} (1 - e^{-\zeta_0^2/2})}$, which is at most $c^{-1} e^{-cn_1}$ for a suitable $c > 0$. $\square$

Proof of the upper bracket on $\mathcal{M}^{\text{sep}}$. We fix $\zeta \in (0, 1)$, and we let $\theta = (\alpha + \beta)/d$ and $\nu = q_\alpha + q_\beta$. We apply Lemma B.1 with $\zeta_0 = \zeta\epsilon$ to the first $n_1 = \lfloor n/2 \rfloor$ observations. Conditionally on these, the remaining $N = n - n_1$ observations are independent with law P , and we use them with the observables

$$D' = \frac{R}{\hat{w}(X)}, \quad U' = 1, \quad V' = \frac{RY}{\hat{w}(X)}, \quad W' = 0, \quad \lambda' = 1,$$

which depend on X , as the assertions stated before the proof of Theorem 2.3(a) in Section 4.4 allow. These are bounded by $M'_0 = \epsilon^{-1}$, and

$$w' = \frac{w}{\hat{w}} \geq \frac{\epsilon}{1 - \epsilon}, \quad a' = \frac{\hat{w}}{w}, \quad b' = b, \quad g' = \frac{pw}{\hat{w}}, \quad \int a'b'g' = \int bp.$$

Since $\hat{w} \in \mathcal{H}^\alpha(H_1)$ and $w \in \mathcal{W}$, the function $a' = \hat{w} \cdot (1/w)$ lies in a ball $\mathcal{H}^\alpha(H_2)$, with H_2 depending only on $d, \alpha, H_0, H_1, \epsilon$. Indeed, by the Leibniz and chain rules, each derivative of a' of order at most $\lceil \alpha \rceil - 1$ is a polynomial in the derivatives of $\hat{w}$ and w of these orders

and in $1/w \leq \epsilon^{-1}$, and each factor is bounded and Hölder continuous with exponent $\alpha - \lceil \alpha \rceil + 1$. On the event $\mathcal{A} = \{\|\hat{w} - w\|_\infty \leq \zeta\epsilon\}$, we have $|w/\hat{w} - 1| \leq \zeta$, hence $g' \in I_\zeta$.

Conditionally on a pilot outcome in $\mathcal{A}$, the remaining observations have law in $\mathcal{P}(\alpha, \beta)$ for $(D', U', V', W', \lambda')$ with the constants M'_0 , $\delta' = \epsilon/(1 - \epsilon)$, $H' = \max(H_0, H_2)$, and $[g_-, g_+] = I_\zeta$, and its target is the MAR mean. We let $\hat{\psi}$ be the estimator of Theorem 2.3(a) from Section 4.4, applied with these constants and $\hat{w}$ to the last N observations and clipped to $[0, 1]$. Its tuning depends only on these constants, so that $\hat{\psi}$ is measurable, and on $\mathcal{A}$ its conditional mean squared error is at most $C \bar{a}_N(\theta, \tau_{I_\zeta}, \nu)^2$, with C independent of the pilot. Since its squared error is at most one off $\mathcal{A}$, its mean squared error is at most $C \bar{a}_N(\theta, \tau_{I_\zeta}, \nu)^2 + c^{-1}e^{-cn_1}$. Since $n/2 \leq N \leq n$, the factors $N^{-\theta}$, the powers of $\log N$, and $e^{-\kappa\sqrt{\log N}}$ differ from their values at n by bounded factors, the last because $\sqrt{\log n} - \sqrt{\log N} \leq \log 2 / \sqrt{\log N}$. We conclude that $\bar{a}_N \leq C' \bar{a}_n$. Since $e^{-cn_1} = o(n^{-1})$ and $n^{-1/2} \leq \bar{a}_n$ for large n , we obtain $r_n(\psi, \mathcal{M}^{\text{sep}}) \leq C_\zeta \bar{a}_n(\theta, \tau_{I_\zeta}, \nu)$. $\square$

B.3 Proof of Corollary 3.2 for these targets

Upper bounds. We take the choices of Section 3.1, with the response space $\{(0, 0)\} \cup (\{1\} \times [0, 1])$, $M_0 = 1$, $\delta = \eta$, and $[g_-, g_+] = [\eta, \eta^{-1}]$. Then $\mathcal{M}_\eta(\alpha, \beta, H) \subseteq \mathcal{P}(\alpha, \beta)$, and Theorem 2.3(a) implies the upper bracket for the MAR mean on $\mathcal{M}_\eta$. The upper bracket on $\mathcal{M}^{\text{sep}}$ with I_ζ , for every $H_0 > 1$, was proved in Appendix B.2. The map that keeps arm j sends $\mathcal{T}_\eta(\alpha, \beta_0, \beta_1, H)$ into $\mathcal{M}_\eta(\alpha, \beta_j, H)$, since its output has observation probability w_j with $\eta \leq w_j \leq 1 - \eta$, $1/w_j = a_j \in \mathcal{H}^\alpha(H)$, and $\eta \leq w_j p = g_j \leq \eta^{-1}$, and regression $b_j \in \mathcal{H}^{\beta_j}(H)$. It sends $\mathcal{T}^{\text{sep}}$ into the class $\mathcal{M}^{\text{sep}}$ with indices (α, β_j) and radius $H_0 + 1$, since $\epsilon \leq 1 - w \leq 1 - \epsilon$ and $\|1 - w\|_{C^\alpha} \leq 1 + \|w\|_{C^\alpha}$. Since the MAR mean of the output is μ_j , Lemma A.3(c) shows that μ_j satisfies the upper bracket $((\alpha + \beta_j)/d, I, q_\alpha + q_{\beta_j})$ on $\mathcal{T}_\eta$, with $I = [\eta, \eta^{-1}]$, and on $\mathcal{T}^{\text{sep}}$ with I_ζ . Lemma A.3(a), applied to $T_{\text{ATE}} = \mu_1 - \mu_0$ and to the relations on the left of (B.5), together with Remark A.4, gives the upper brackets of the three effects with the indices (3.7) and $\nu = q_\alpha + q_{\beta_*}$.

Lower bounds. We treat two cases at once. In the first, $\mathcal{M} = \mathcal{M}_\eta$, $\mathcal{T} = \mathcal{T}_\eta$, $I = [\eta, \eta^{-1}]$, and $[r_-, r_+] = [2\eta, 2\eta^{-1}]$. In the second, $\mathcal{M} = \mathcal{M}^{\text{sep}}$, $\mathcal{T} = \mathcal{T}^{\text{sep}}$, and $I = [r_-, r_+] = [p_-, p_+]$. Then $r_- < 1 < r_+$, since $\eta < 1/4$ or by (3.3), and since ρ depends only on r_+/r_- , the value of τ computed from $[r_-, r_+]$ is τ_I . For the MAR mean, let $\beta' = \beta$ and $j = 1$. For the effects, let $\beta' = \beta_j$, where $\beta_j = \min(\beta_0, \beta_1)$ for the ATE, $j = 0$ for the ATT, and $j = 1$ for the ATU, so that $s = (\alpha + \beta')/2$ by (3.7). If P is a law of missing-data observations with $Y \in \{0, 1\}$, then $K_j P$ has propensity w or $1 - w$, regression b in arm j and the constant $1/2$ in the other arm, outcomes in $\{0, 1\}$, and covariate density p . Since the constant $1/2$ lies in every Hölder ball of radius H_0 or H , the law P lies in $\mathcal{M}$, with β' in place of β , and $K_j P$ lies in $\mathcal{T}$, if $1/w, 1/(1 - w) \in \mathcal{H}^\alpha(H)$, $b \in \mathcal{H}^{\beta'}(H)$, $\eta \leq w \leq 1 - \eta$, and $\eta \leq wp, (1 - w)p \leq \eta^{-1}$ in the first case, and if $w, 1 - w \in \mathcal{H}^\alpha(H_0)$, $b \in \mathcal{H}^{\beta'}(H_0)$,

$\epsilon \leq w \leq 1 - \epsilon$, and $p_- \leq p \leq p_+$ in the second. By the paragraph that defines the MAR family, these are conditions of the types (H), (I), and (W) of Lemma 7.2, for functions K with $k_0 = 2$ or $k_0 = 1/2$, or with $K \equiv 1$ for the density. Their hypotheses hold, since $2 < H$, $1/2 < H_0$, $\eta < 1/2 < 1 - \eta$, $\epsilon < 1/2 < 1 - \epsilon$, and $r_{\pm}/2 = \eta^{\mp 1}$ in the first case and $r_{\pm} = p_{\pm}$ in the second.

Suppose first that $s < d/4$. By Lemma 7.2, applied with a small enough fixed value of the input ϵ of Proposition 7.1 (not the overlap bound of $\mathcal{T}^{\text{sep}}$), the MAR family with indices (α, β') and the interval $[r_-, r_+]$ lies in $\mathcal{M}$, and its image $\mathcal{P}'_n = K_j \mathcal{P}_n^h$ lies in $\mathcal{T}$, for all large n . Let $\theta = 2s/d$ and let $t_n = c_1 \underline{a}_n(\theta, \tau_I)$ be the scale at which the MAR mean is hard on the MAR family. For the MAR mean, Lemma A.2 gives the lower bracket on $\mathcal{M}$. For the effects, Lemma A.3(c) shows that μ_j is hard at scale t_n on $(\mathcal{P}'_n)$. By (B.3), T_{ATE} differs from μ_j on $\mathcal{P}'_n$ by a sign and an additive constant, so that it is also hard at scale t_n . For the ATT and the ATU, we apply Lemma A.3(b), with $h(O) = (Y, A)$, to the relations for μ_0 and μ_1 on the right of (B.5). Here $nt_n^2 \rightarrow \infty$, because

$$\log(nt_n^2) = (1 - 2\theta) \log n - 2\kappa \sqrt{\log n} + (\theta + 2) \log \log n + 2 \log c_1,$$

so that the target is hard at scale $t_n/(4L)$ on $(\mathcal{P}'_n)$. In each case, the lower bracket on $\mathcal{T}$ follows from Lemma A.2.

If $s \geq d/4$, we use the laws Q_t under which X is uniform on $[0, 1]^d$ and independent of $Z \sim \pi_{u,v}$ with $(u, v) = (0, t/2)$, for $|t| < 1/8$. They have $p = 1$ and the constant functions $w = 1/2$ and $b = 1/2 + t$, so that they satisfy the conditions above, $Q_t \in \mathcal{M}$, $K_j Q_t \in \mathcal{T}$, and the MAR mean is $\psi(Q_t) = 1/2 + t$. The density of Q_t with respect to Q_0 is $1 + ts_v(Z)/2$, which is affine in t and at least $3/4$, and Lemma 5.1 implies that $r_n(\psi, \mathcal{M}) \geq cn^{-1/2}$. By (B.4), each effect T of $K_j Q_t$ equals $(2j - 1)t$, and Lemma A.3(c), applied to the target $(2j - 1)T + 1/2$ on $\mathcal{T}$, gives $r_n(T, \mathcal{T}) \geq cn^{-1/2}$.

For the MAR mean on $\mathcal{M}_\eta$, the lower bracket also holds on every class between a localized class $\mathcal{P}_r(\pi_0)$ and $\mathcal{M}_\eta$, by Theorem 2.3(b). The baseline π_0 satisfies (2.5): Σ is positive definite, and $0 < \eta < 1/4$ and $H > 2$ give $\max\{\delta, g_-\} = \eta < w_0 = 1/2 < \eta^{-1} = g_+$, $|a_0| = 2 < H$, and $|b_0| = 1/2 < H$. We have $\mathcal{P}_r(\pi_0) \subseteq \mathcal{M}_\eta$ whenever $0 < r < 1/2 - \eta$, since $w \leq 1/2 + r < 1 - \eta$ on $\mathcal{P}_r(\pi_0)$.

The logarithmic asymptotics. Let $\mathcal{P}$ be $\mathcal{M}^{\text{sep}}$ or $\mathcal{T}^{\text{sep}}$, let $s < d/4$ and $\theta = 2s/d$, and let κ_ζ be the κ of (2.9) computed from τ_{I_ζ} . Since $\log a_n(\theta, \tau) = -\theta \log n - \kappa \sqrt{\log n} + (\theta/2) \log \log n$, the lower bracket with $[p_-, p_+]$ and the upper bracket with I_ζ imply that

$$-\kappa_p(\theta) \leq \liminf_{n \rightarrow \infty} \frac{\log r_n + \theta \log n}{\sqrt{\log n}} \leq \limsup_{n \rightarrow \infty} \frac{\log r_n + \theta \log n}{\sqrt{\log n}} \leq -\kappa_\zeta.$$

As $\zeta \rightarrow 0$, we have $\rho_{I_\zeta} \rightarrow \rho_{[p_-, p_+]}$, hence $\tau_{I_\zeta} \rightarrow \tau_p$ and $\kappa_\zeta \rightarrow \kappa_p(\theta)$, which gives (3.4).

Conditional covariances and variances

Bilinear and quadratic targets. We apply Theorem 2.3 with $D = 1$, $M_0 = 1$, $H = H_0$, $[g_-, g_+] = [p_-, p_+]$, and $\delta = 1/2$, with $(W, \lambda) = (0, 1)$ for $\mathbb{E}_P[m_U m_V]$ and Q_2 , and $(W, \lambda) = (UV, -1)$ for $\mathbb{E}_P[\text{Cov}_P(U, V \mid X)]$ and $\bar{\sigma}^2$. Then $w = 1$, $a = m_U$, $b = m_V$, and $g = p$, so that the generic class is $\mathcal{P}_{\text{bil}}(\alpha, \beta)$, or $\mathcal{P}_{\text{quad}}(s)$ when $U = V = Y$ and $\alpha = \beta = s$. If U, V are independent Rademacher variables, then $(w_0, a_0, b_0) = (1, 0, 0)$, the residual covariance in (2.5) is the identity, and the strict inequalities there follow from (3.3) and $H_0 > 1$. A Rademacher variable Y satisfies the diagonal alternative with residual variance one. This gives the brackets of these four targets.

Upper brackets. We write $m_j = \mathbb{E}_P[Y^j]$ for $j = 1, 2$. For a fixed known measurable f_0 with $B_0 = \|f_0\|_\infty < \infty$, we set $\ell_0 = \mathbb{E}_P[f_0(X)^2 - 2Yf_0(X)]$ and $E_{f_0} = \mathbb{E}_P[(f(X) - f_0(X))^2]$. Since $\mathbb{E}_P[Yf_0(X)] = \mathbb{E}_P[f(X)f_0(X)]$, we have

$$E_* = Q_2 - m_1^2, \quad R_*^2 = \frac{Q_2 - m_1^2}{\max\{m_2 - m_1^2, v_{\min}\}}, \quad E_{f_0} = Q_2 + \ell_0,$$

where the middle identity holds on $\mathcal{P}_{\text{quad}}(s; v_{\min})$. In Lemma A.3, we use $h(O) = (Y, Y^2)$, with $\mathcal{R} = [-1, 1] \times [0, 1]$, for E_* and R_*^2 , and $h(O) = f_0(X)^2 - 2Yf_0(X)$, with $\mathcal{R} = [-B_0^2 - 2B_0, B_0^2 + 2B_0]$, for E_{f_0} . Since the denominator is at least $v_{\min}$, the three maps are Lipschitz on $[0, 1] \times \mathcal{R}$. By the bracket of Q_2 and Lemma A.3(a), the three targets satisfy the upper bracket $(2s/d, [p_-, p_+], 2q_s)$ on their classes.

Lower brackets below the threshold. Let $s < d/4$. We use the hard laws of Appendix A.1 in the diagonal case, with Rademacher baseline, $(D, U, V, W, \lambda) = (1, Y, Y, 0, 1)$, $s_d = Y$, indices (s, s) , and $[r_-, r_+] = [p_-, p_+]$. Then $P(Y = y \mid X) = \{1 + y(u + v)\}/2$, $f = u + v$, and the target is Q_2 . The conditions of $\mathcal{P}_{\text{quad}}(s)$ are of type (H) for $K = u + v$, with $k_0 = 0$ and $\gamma_K = s$, and of type (W) for $K \equiv 1$. By Lemma 7.2, for a suitable fixed ϵ , the hard laws $\mathcal{P}_n^h$ lie in $\mathcal{P}_{\text{quad}}(s)$ for all large n . Since $Y^2 = 1$ and $|\mathbb{E}_P Y| \leq \|u + v\|_\infty \leq Cn^{-s/d}$, they also lie in $\mathcal{P}_{\text{quad}}(s; v_{\min})$ for all large n . On them Q_2 is hard at scale $t_n = c_1 \underline{a}_n(2s/d, \tau_p)$, and $nt_n^2 \rightarrow \infty$. The exact inverse relations $Q_2 = E_* + m_1^2 = (m_2 - m_1^2)R_*^2 + m_1^2 = E_{f_0} - \ell_0$ are L -Lipschitz on $I_1 \times \mathcal{R}$, with $L = 2, \sqrt{5}, 1$, where $I_1 = [0, 1]$ for E_* and R_*^2 and $I_1 = [0, (1 + B_0)^2]$ for E_{f_0} . By Lemma A.3(b), with $T_1 = Q_2$ and $e_n = 0$, the three targets are hard on $(\mathcal{P}_n^h)$ at scale $t_n/(4L)$, and their lower brackets follow from Lemma A.2.

Root- n lower bounds. We fix $\varphi = \eta \sin(2\pi x_1)$, where $\eta > 0$ depends only on s, H_0 and satisfies $\|\varphi\|_{C^s} \leq H_0$ and $\|\varphi\|_\infty \leq 1/2$. Let P_t have $p = 1$ and $P_t(Y = y \mid X) = \{1 +$

$yt\varphi(X)\}/2$ for $|t| < 1$. These laws lie in $\mathcal{P}_{\text{quad}}(s; v_{\min})$, with $\mathbb{E}_{P_t}Y = 0$ and $\text{Var}_{P_t}(Y) = 1$. We set $A_\varphi = \int \varphi^2 > 0$ and $B_\varphi = \int \varphi f_0$. Then $E_*(P_t) = R_*^2(P_t) = t^2 A_\varphi$ and $E_{f_0}(P_t) = t^2 A_\varphi - 2tB_\varphi + \int f_0^2$. Since the excess-loss derivatives at $t = 1/4$ and $t = 1/2$ differ by $A_\varphi/2$, at one of these points their absolute value is at least $A_\varphi/4$; the derivatives of E_* and R_*^2 there are at least $A_\varphi/2$. The densities relative to the law with uniform design and Rademacher response are $1 + tY\varphi(X) \geq 1/2$, with bounded slope. Lemma 5.1 yields the root- n lower bounds for all three targets. Only the constants for E_{f_0} may also depend on B_0 .

The trial. The map $(X, A, Y_{\text{obs}}) \mapsto (X, (2A - 1)(2Y_{\text{obs}} - 1))$ sends the trial class into $\mathcal{P}_{\text{quad}}(s)$, since its response lies in $[-1, 1]$ and its conditional mean is f . Conversely, we draw an independent $A \sim \text{Bernoulli}(1/2)$ and map (X, Y) to $(X, A, \{1 + (2A - 1)Y\}/2)$. Since the outcome lies in $[0, 1]$ and its arm means are $(1 \pm f)/2$, this kernel sends $\mathcal{P}_{\text{quad}}(s)$ into the trial class. Both kernels preserve p, f , and the two targets. Lemma A.3(c) transfers the upper bounds along the first kernel and the lower and probability bounds along the second.

Wald ratios and the overlap-weighted effect

We recall the causal interpretations of the targets of Section 3.3 and prove Corollary 3.2. The interpretations use identification assumptions; the classes restrict observed laws.

D.1 Wald ratios

Example D.1 (Wald ratio). Let $A(z)$ and $Y(a)$ denote potential treatments and outcomes, and suppose that $A = A(Z)$ and $Y = Y(A)$. If $Z \perp (A(0), A(1), Y(0), Y(1)) \mid X$ and $A(1) \geq A(0)$, then T_W is the complier average treatment effect $\mathbb{E}[Y(1) - Y(0) \mid A(1) > A(0)]$, the local average treatment effect of [2, 21], in the form with covariates given in [11, 16]. Its numerator and denominator are average effects of the instrument, that is, differences of two MAR means.

Example D.2 (Average conditional Wald ratio). Suppose that $P(Z = 1 \mid X) = 1/2$ and $A \leq Z$, so that the instrument is assigned with equal probability and treatment is unavailable when $Z = 0$. Under the assumptions of Example D.1, $b(X)$ is the conditional complier effect. If also $\mathbb{E}[Y(1) - Y(0) \mid X, A(1)] = \mathbb{E}[Y(1) - Y(0) \mid X]$, which says that the covariates account for differences in mean effects between compliers and never-takers, then T_{CW} is the ATE; see [70, Theorem 1 and the following discussion]. The two Wald targets differ in general, since T_W averages $b(X)$ with weights proportional to $q(X)$.

Proof of Corollary 3.2 for the Wald ratios. Wald ratio. We write $\mathcal{N} = \mathbb{E}_P[\Delta_Y(X)] \in [-1, 1]$ and $\mathcal{D} = \mathbb{E}_P[\Delta_A(X)] \in [c_W, 1]$. With the instrument as the treatment, the deterministic maps $(X, Z, A, Y) \mapsto (X, Z, Y)$ and $(X, Z, A, Y) \mapsto (X, Z, A)$ send $\mathcal{I}_{\eta, c_W}(\alpha, \beta, H)$ into $\mathcal{T}_\eta(\alpha, \beta, \beta, H)$, and the ATEs of the images are $\mathcal{N}$ and $\mathcal{D}$. By Corollary 3.2 for the ATE on $\mathcal{T}_\eta(\alpha, \beta, \beta, H)$ and Lemma A.3(c), both satisfy the upper bracket $((\alpha + \beta)/d, [\eta, \eta^{-1}], q_\alpha + q_\beta)$ on $\mathcal{I}_{\eta, c_W}(\alpha, \beta, H)$, and applying Lemma A.3(a) to the ratio, with $C_0 = 1$ and $\delta_{\mathcal{D}} = c_W$, together with Remark A.4, we obtain the upper bracket for T_W . For the lower bracket, the kernel $(X, Z, Y) \mapsto (X, Z, Z, Y)$, which sets $A = Z$, maps $\mathcal{T}_\eta(\alpha, \beta, \beta, H)$ into $\mathcal{I}_{\eta, c_W}(\alpha, \beta, H)$: its output has $m_{A,z} = z$, whose Hölder norm is $z \leq 1 < H$, and $\Delta_A = 1 \geq c_W$, while the remaining conditions are those of $\mathcal{T}_\eta(\alpha, \beta, \beta, H)$ with Z as the treatment. Since T_W equals the ATE of the input on the image, Lemma A.3(c) transfers its lower bracket.

Average conditional Wald ratio. We take the response (Z, A, Y) , the functions $(D, U, V, W, \lambda) = (2A, 1, 2(2Z - 1)Y, 0, 1)$, and $M_0 = 2$. Since $P(Z = 1 \mid X) = 1/2$

and $A \leq Z$, we have $\mathbb{E}_P[2A \mid X] = q$ and $\mathbb{E}_P[2(2Z - 1)Y \mid X] = \Delta_Y$. Then $w = q$, $a = 1/q$, $b = \Delta_Y/q$, $g = qp$, and $\psi = \int (\Delta_Y/q)p = T_{CW}$, so that $\mathcal{J}_\eta(\alpha, \beta, H) \subseteq \mathcal{P}(\alpha, \beta)$ with $\delta = \eta$ and $[g_-, g_+] = [\eta, \eta^{-1}]$, and the upper bracket follows from Theorem 2.3(a). For the lower bracket, the kernel that draws $Z \sim \text{Bernoulli}(1/2)$ independently and maps (X, R, RB) to (X, Z, ZR, ZRB) maps $\mathcal{M}_\eta(\alpha, \beta, H)$ into $\mathcal{J}_\eta(\alpha, \beta, H)$: the output has $A = ZR \leq Z$, $q = P(R = 1 \mid X)$, and $\Delta_Y = qb$ with $b = \mathbb{E}[B \mid R = 1, X]$, so that its conditional Wald ratio is b and its target is the MAR mean of the input. Lemma A.3(c) therefore transfers its lower bracket. $\square$

D.2 The overlap-weighted effect

Example D.3 (Overlap-weighted effect). For fixed γ , minimizing $\mathbb{E}_P[(Y - \gamma A - h(X))^2]$ over h gives $h = m - \gamma w$, and minimizing the resulting quadratic in γ gives (3.8). This definition places no restriction on $\mathbb{E}_P[Y \mid A, X]$, and the numerator and the denominator of (3.8) are expected conditional covariances as in Section 3.2. Under consistency and conditional exchangeability, (3.8) is the overlap-weighted treatment effect [35]; see also [12, Remark 9.3.3] for this interpretation of regression on residuals when effects vary with X .

The weights $w(1 - w)$ appear earlier in Crump et al. [13, Section 5.4, Corollary 5.2], as the weights that minimize the efficiency bound for weighted sample average treatment effects under homoscedasticity.

Robins et al. [48, Example 1c] impose smoothness on $m - \gamma_{OW}w$ instead of on m . Since $|\gamma_{OW}| \leq 1$ and $\mathcal{H}^\alpha(H) \subseteq \mathcal{H}^\beta(CH)$ when $\beta \leq \alpha$, the two conditions agree up to fixed changes of radius when $\beta \leq \alpha$, but they can differ when $\beta > \alpha$. Without smoothness assumptions on the covariate density, and without the assumption that the conditional treatment effect is constant, Robins et al. [48, Remark 4.2] refer to $n^{-2s/d}$, for $s < d/4$, as their conjectured minimax rate for this effect. Here s is the average of the smoothness indices of the propensity and of their regression. For $\beta \leq \alpha$, Corollary 3.2 confirms the exponent $2s/d$, with $s = (\alpha + \beta)/2$, and shows that the minimax rate is smaller than $n^{-2s/d}$ by a factor $e^{-(\kappa_p(2s/d) + o(1))\sqrt{\log n}}$. Concurrently, Park [46, Section 2] reports minimax bounds under a known positive lower bound on the denominator, with effective smoothness $\min(\alpha + \beta, 2\beta)$ in his notation, for the normalized covariance $\mathbb{E}[\text{Cov}(A, Y \mid X)]/\mathbb{E}[\text{Var}(Y \mid X)]$ of two bounded responses, which has the same form as γ_{OW} with the roles of A and Y exchanged; the proof is in the supplement of [46] (Theorem S14), which was not available to us.

Proof of Corollary 3.2 for the overlap-weighted effect. Here $w = P(A = 1 \mid X)$ is the propensity, while the generic D is one. We write $\gamma_{OW} = \mathcal{N}/\mathcal{D}$, where $\mathcal{N} = \mathbb{E}_P[\text{Cov}_P(A, Y \mid X)] \in [-1/4, 1/4]$ and $\mathcal{D} = \mathbb{E}_P[w(1 - w)] \in [\epsilon(1 - \epsilon), 1/4]$, and we let $\theta = \min\{(\alpha + \beta)/d, 2\alpha/d\}$.

Upper bracket. We apply Theorem 2.3(a) with $D = 1$, $M_0 = 1$, $H = H_0$, $\delta = 1/2$, and $[g_-, g_+] = [p_-, p_+]$. The choices $(U, V, W, \lambda) = (A, Y, AY, -1)$ and $(A, A, A, -1)$ give $\mathcal{N}$ and $\mathcal{D}$, with upper brackets $((\alpha + \beta)/d, [p_-, p_+], q_\alpha + q_\beta)$ and $(2\alpha/d, [p_-, p_+], 2q_\alpha)$, respectively. Applying Lemma A.3(a) to the ratio with $C_0 = 1/4$ and $\delta_{\mathcal{D}} = \epsilon(1 - \epsilon)$, together with Remark A.4, we obtain the upper bracket with exponent θ and $\nu = q_\alpha + q_{\min(\alpha, \beta)}$.

A common response family. Let $S_1 = 2A - 1$ and $S_2 = 2Y - 1$. For $c \in \{0, 1/4\}$ we define the baseline and scores by

$$\pi_0^{(c)} = \frac{1 + cS_1S_2}{4}, \quad s_i = \frac{2S_i}{1 + cS_1S_2} \quad (i = 1, 2).$$

Since the baseline factor cancels in score expectations, we have $\mathbb{E}_{\pi_0^{(c)}} s_i = 0$ and $\mathbb{E}_{\pi_0^{(c)}} [(S_j/2)s_i] = \mathbf{1}_{i=j}$. The baseline means of A and Y are $1/2$, and therefore s_1, s_2 are scores for $U = A$, $V = Y$, and s_1 is a diagonal score for $U = V = A$. For small functions ℓ_1, ℓ_2 of x , the law $\pi_0^{(c)}(1 + \ell_1 s_1 + \ell_2 s_2)$ and its moments are

$$P(S_1 = \sigma_1, S_2 = \sigma_2 \mid X) = \frac{1 + c\sigma_1\sigma_2 + 2\ell_1\sigma_1 + 2\ell_2\sigma_2}{4}, \quad w = \frac{1}{2} + \ell_1, \quad m = \frac{1}{2} + \ell_2,$$

$$\mathcal{N} = \frac{c}{4} - \int p\ell_1\ell_2, \quad \mathcal{D} = \frac{1}{4} - \int p\ell_1^2.$$

Lower bracket when $\theta < 1/2$. If $\beta \leq \alpha$, we use the hard laws with indices (α, β) , $c = 0$, $(\ell_1, \ell_2) = (u, v)$, and $(U, V, W, \lambda) = (A, Y, AY, -1)$. If $\beta > \alpha$, we use indices (α, α) , $c = 1/4$, $(\ell_1, \ell_2) = (u + v, 0)$, and $(U, V, W, \lambda) = (A, A, A, -1)$, with the diagonal score s_1 . In both cases $D = 1$, $[r_-, r_+] = [p_-, p_+]$, and the indices give the exponent θ . For membership in $\mathcal{P}_{\text{OW}}(\alpha, \beta)$, we require $p_- \leq p \leq p_+$, $w \in \mathcal{H}^\alpha(H_0)$, $m \in \mathcal{H}^\beta(H_0)$, and $\epsilon \leq w \leq 1 - \epsilon$. The density and overlap conditions are of types (W) and (I). In the family with indices (α, β) , both smoothness conditions are of type (H), with $k_0 = 1/2 < H_0$, $\gamma_K = \alpha$ for w , and $\gamma_K = \beta$ for m . In the family with indices (α, α) , the condition on w is of type (H), with $\gamma_K = \alpha$, while $m = 1/2$ satisfies $\|m\|_{C^\beta} = 1/2 < H_0$ directly. By Lemma 7.2, the hard laws lie in $\mathcal{P}_{\text{OW}}(\alpha, \beta)$ for all large n , and their generic target, $\mathcal{N}$ in the first case and $\mathcal{D}$ in the second, is hard on them at scale $t_n = c_1 a_n(\theta, \tau_p)$.

In the first case, eventually $\mathcal{D} \geq 1/8$, and we have

$$\left| \mathcal{N} - \frac{\gamma_{\text{OW}}}{4} \right| = |\gamma_{\text{OW}}| \int p u^2 \leq 8 \|u\|_\infty^3 \|v\|_\infty \leq C n^{-\theta - 2\alpha/d} = o(t_n).$$

The map $\gamma \mapsto \gamma/4$ is 1-Lipschitz on $[-1, 1]$, which contains γ_{OW} . In the second case, $\mathcal{N} = 1/16$ and eventually $\mathcal{D} \in [3/16, 1/4]$, so that $\gamma_{\text{OW}} \in [1/4, 1/3]$ and $\mathcal{D} = 1/(16\gamma_{\text{OW}})$; this map is 1-Lipschitz on $[1/4, 1/3]$. By Lemma A.3(b), with $k = 0$, T_1 the generic target, and $T_0 = \gamma_{\text{OW}}$, the target γ_{OW} is hard at scale $t_n/4$ in both cases, and Lemma A.2 gives the lower bracket.

Root- n lower bound. Let P_t have $p = 1$ and response law $\{1 + tS_1S_2\}/4$ for $|t| < 1/2$. Then $w = m = 1/2$, $\mathcal{N} = t/4$, $\mathcal{D} = 1/4$, and $\gamma_{\text{OW}} = t$. These laws lie in $\mathcal{P}_{\text{OW}}(\alpha, \beta)$, and their densities relative to the law with uniform design and uniform response are $1 + tS_1S_2 \geq 1/2$, with bounded slope. Lemma 5.1 implies the root- n lower bound. $\square$

References

- [1] B. K. Alpert. A class of bases in L^2 for the sparse representation of integral operators. *SIAM Journal on Mathematical Analysis*, 24(1):246–262, 1993. doi: 10.1137/0524016.
- [2] J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996. doi: 10.1080/01621459.1996.10476902.
- [3] P. M. Aronow and P. Lopatto. Nearly-minimax variance estimation under rough random design, 2026. arXiv:2607.13170v3.
- [4] S. Balakrishnan, E. H. Kennedy, and L. Wasserman. The fundamental limits of structure-agnostic functional estimation. *Statistical Science*, 41(3):659–670, 2026. doi: 10.1214/25-STS997.
- [5] P. J. Bickel and Y. Ritov. Estimating integrated squared density derivatives: Sharp best order of convergence estimates. *Sankhyā: The Indian Journal of Statistics, Series A*, 50(3):381–393, 1988.
- [6] L. Birgé and P. Massart. Estimation of integral functionals of a density. *The Annals of Statistics*, 23(1):11–29, 1995. doi: 10.1214/aos/1176324452.
- [7] M. Bonvini, E. H. Kennedy, O. Dukes, and S. Balakrishnan. Doubly-robust inference and optimality in structure-agnostic models with smoothness, 2024. arXiv:2405.08525v2.
- [8] T. T. Cai and M. G. Low. Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *The Annals of Statistics*, 39(2): 1012–1041, 2011. doi: 10.1214/10-AOS849.
- [9] T. T. Cai, M. Levine, and L. Wang. Variance function estimation in multivariate nonparametric regression with fixed design. *Journal of Multivariate Analysis*, 100(1): 126–136, 2009. doi: 10.1016/j.jmva.2008.03.007.
- [10] X. Chen, L. Liu, and R. Mukherjee. Method-of-moments inference for GLMs and doubly robust functionals under proportional asymptotics, 2025. arXiv:2408.06103v3.
- [11] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. doi: 10.1111/ectj.12097.
- [12] V. Chernozhukov, C. Hansen, N. Kallus, M. Spindler, and V. Syrgkanis. *Applied Causal*

- Inference Powered by ML and AI*. Online, 2026. URL <https://causalml-book.org/>. Version 0.1.2, 3 May 2026.
- [13] R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik. Moving the goalposts: Addressing limited overlap in estimation of average treatment effects by changing the estimand. IZA Discussion Paper 2347, Institute for the Study of Labor (IZA), Bonn, 2006. docs.iza.org/dp2347.pdf.
 - [14] E. Dobriban, R. Mukherjee, J. M. Robins, and Z. Wang. Improved variance estimation in homoskedastic nonparametric random-design regression via a two-scale approach, 2026. arXiv:2609.08783v1.
 - [15] D. Evans and A. J. Jones. Non-parametric estimation of residual moments and covariance. *Proceedings of the Royal Society A*, 464(2099):2831–2846, 2008. doi: 10.1098/rspa.2007.0195.
 - [16] M. Frölich. Nonparametric IV estimation of local average treatment effects with covariates. *Journal of Econometrics*, 139(1):35–75, 2007. doi: 10.1016/j.jeconom.2006.06.004.
 - [17] P. R. Halmos. The theory of unbiased estimation. *The Annals of Mathematical Statistics*, 17(1):34–43, 1946. doi: 10.1214/aoms/1177731020.
 - [18] M. A. Hernán and J. M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton, 2020.
 - [19] W. Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948. doi: 10.1214/aoms/1177730196.
 - [20] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.
 - [21] G. W. Imbens and J. D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994. doi: 10.2307/2951620.
 - [22] Yu. I. Ingster and I. A. Suslina. *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*, volume 169 of *Lecture Notes in Statistics*. Springer, New York, 2003. doi: 10.1007/978-0-387-21580-8.
 - [23] J. Jiao, K. Venkat, Y. Han, and T. Weissman. Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 61(5):2835–2885, 2015. doi: 10.1109/TIT.2015.2412945.
 - [24] J. Jin and V. Syrgkanis. Structure-agnostic optimality of doubly robust learning for treatment effect estimation (extended abstract). In *Proceedings of the Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 3159–3160, 2025. Full paper: arXiv:2402.14264v4.
 - [25] J. Jin and V. Syrgkanis. Sharp structure-agnostic lower bounds for general linear functional estimation, 2025. arXiv:2512.17341v2.

- [26] E. H. Kennedy, S. Balakrishnan, and L. Wasserman. Discussion of “On nearly assumption-free tests of nominal confidence interval coverage for causal parameters estimated by machine learning”. *Statistical Science*, 35(3):540–544, 2020. doi: 10.1214/20-STS796.
- [27] G. Kerkycharian and D. Picard. Estimating nonquadratic functionals of a density using Haar wavelets. *The Annals of Statistics*, 24(2):485–507, 1996. doi: 10.1214/aos/1032894450.
- [28] W. Kong and G. Valiant. Estimating learnability in the sublinear data regime. In *Advances in Neural Information Processing Systems*, volume 31, pages 5455–5464, 2018. Extended version: arXiv:1805.01626v3.
- [29] J. K. Kraus, P. S. Vassilevski, and L. T. Zikatanov. Polynomial of best uniform approximation to $1/x$ and smoothing in two-level methods. *Computational Methods in Applied Mathematics*, 12(4):448–468, 2012. doi: 10.2478/cmam-2012-0026.
- [30] G. Last and M. Penrose. *Lectures on the Poisson Process*, volume 7 of *Institute of Mathematical Statistics Textbooks*. Cambridge University Press, Cambridge, 2017. doi: 10.1017/9781316104477.
- [31] G. Last and M. D. Penrose. Poisson process Fock space representation, chaos expansion and covariance inequalities. *Probability Theory and Related Fields*, 150(3–4): 663–690, 2011. doi: 10.1007/s00440-010-0288-5.
- [32] L. Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer Series in Statistics. Springer, New York, 1986. doi: 10.1007/978-1-4612-4946-7.
- [33] O. Lepski, A. Nemirovski, and V. Spokoiny. On estimation of the L_r norm of a regression function. *Probability Theory and Related Fields*, 113(2):221–253, 1999. doi: 10.1007/s004409970006.
- [34] J. Levy, M. van der Laan, A. Hubbard, and R. Pirracchio. A fundamental measure of treatment effect heterogeneity. *Journal of Causal Inference*, 9(1):83–108, 2021. doi: 10.1515/jci-2019-0003.
- [35] F. Li, K. L. Morgan, and A. M. Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018. doi: 10.1080/01621459.2016.1260466.
- [36] L. Li, E. Tchetgen Tchetgen, A. van der Vaart, and J. M. Robins. Higher order inference on a treatment effect under low regularity conditions. *Statistics & Probability Letters*, 81(7):821–828, 2011. doi: 10.1016/j.spl.2011.02.030.
- [37] L. Liu, R. Mukherjee, W. K. Newey, and J. M. Robins. Semiparametric efficient empirical higher order influence function estimators, 2017. arXiv:1705.07577v5.
- [38] L. Liu, R. Mukherjee, and J. M. Robins. Rejoinder: On nearly assumption-free tests of nominal confidence interval coverage for causal parameters estimated by machine learning. *Statistical Science*, 35(3):545–554, 2020. doi: 10.1214/20-STS804.

- [39] L. Liu, R. Mukherjee, J. M. Robins, and E. Tchetgen Tchetgen. Adaptive estimation of nonparametric functionals. *Journal of Machine Learning Research*, 22(99):1–66, 2021. URL <https://jmlr.org/papers/v22/19-892.html>.
- [40] L. Liu, R. Mukherjee, and J. M. Robins. On the asymptotic inadmissibility of double machine learning estimators under structure-agnostic models, 2026. arXiv:2606.22391v2.
- [41] N. Liu, C. Li, Y. Gu, and L. Liu. Stabilized higher-order influence functions: Statistical theory of a class of bilinear forms, 2026. arXiv:2607.04743v3.
- [42] R. J. Mathar. Chebyshev series expansion of inverse polynomials. *Journal of Computational and Applied Mathematics*, 196(2):596–607, 2006. doi: 10.1016/j.cam.2005.10.013.
- [43] A. McClean, S. Balakrishnan, E. H. Kennedy, and L. Wasserman. Double cross-fit doubly robust estimators: Beyond series regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 88(4):1469–1491, 2026. doi: 10.1093/jrsssb/qkag057. Theorem numbering refers to arXiv:2403.15175v3.
- [44] S. McGrath and R. Mukherjee. Nuisance function tuning and sample splitting for optimally estimating a doubly robust functional, 2022. To appear in *The Annals of Statistics*. arXiv:2212.14857v5.
- [45] W. K. Newey and J. M. Robins. Cross-fitting and fast remainder rates for semiparametric estimation, 2018. arXiv:1801.09138v1.
- [46] S. Park. Minimax estimation of the expected conditional covariance under bounds on the covariate density, 2026. arXiv:2610.05006v1.
- [47] T. S. Richardson and A. Rotnitzky. Causal etiology of the research of James M. Robins. *Statistical Science*, 29(4):459–484, 2014. doi: 10.1214/14-STS505.
- [48] J. Robins, L. Li, E. Tchetgen, and A. van der Vaart. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics: Essays in Honor of David A. Freedman*, volume 2 of *IMS Collections*, pages 335–421. Institute of Mathematical Statistics, 2008. doi: 10.1214/193940307000000527.
- [49] J. Robins, E. Tchetgen Tchetgen, L. Li, and A. van der Vaart. Semiparametric minimax rates. *Electronic Journal of Statistics*, 3:1305–1321, 2009. doi: 10.1214/09-EJS479.
- [50] J. M. Robins and Y. Ritov. Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine*, 16(3):285–319, 1997. doi: 10.1002/(SICI)1097-0258(19970215)16:3<285::AID-SIM535>3.0.CO;2-#.
- [51] J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994. doi: 10.1080/01621459.1994.10476818.
- [52] J. M. Robins, L. Li, L. Liu, R. Mukherjee, E. Tchetgen Tchetgen, and A. van der Vaart. Minimax estimation of a functional on a structured high-dimensional model (corrected version), 2015. arXiv:1512.02174v3.

- [53] J. M. Robins, L. Li, R. Mukherjee, E. Tchetgen Tchetgen, and A. van der Vaart. Minimax estimation of a functional on a structured high-dimensional model. *The Annals of Statistics*, 45(5):1951–1987, 2017. doi: 10.1214/16-AOS1515.
- [54] P. M. Robinson. Root- N -consistent semiparametric regression. *Econometrica*, 56(4): 931–954, 1988. doi: 10.2307/1912705.
- [55] A. Rotnitzky, E. Smucler, and J. M. Robins. Characterization of parameters with a mixed bias property. *Biometrika*, 108(1):231–238, 2021. doi: 10.1093/biomet/asaa054.
- [56] D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. doi: 10.1093/biomet/63.3.581.
- [57] A. Sánchez-Becerra. Robust inference for the treatment effect variance in experiments using machine learning, 2023. arXiv:2306.03363v1.
- [58] R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley, New York, 1980. doi: 10.1002/9780470316481.
- [59] Y. Shen, C. Gao, D. Witten, and F. Han. Optimal estimation of variance in nonparametric regression with random design. *The Annals of Statistics*, 48(6):3589–3618, 2020. doi: 10.1214/20-AOS1944.
- [60] X. Song. Batched and complete U-statistics for trace-polynomial estimation from classical shadows, 2026. arXiv:2608.22962v1.
- [61] E. M. Stein and R. Shakarchi. *Fourier Analysis: An Introduction*, volume 1 of *Princeton Lectures in Analysis*. Princeton University Press, Princeton, 2003.
- [62] C. J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982. doi: 10.1214/aos/1176345969.
- [63] L. Tian, A. A. Alizadeh, A. J. Gentles, and R. Tibshirani. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532, 2014. doi: 10.1080/01621459.2014.951443.
- [64] L. N. Trefethen. *Approximation Theory and Approximation Practice*. SIAM, Philadelphia, extended edition, 2019. doi: 10.1137/1.9781611975949.
- [65] L. N. Trefethen and D. Bau, III. *Numerical Linear Algebra*. SIAM, Philadelphia, 1997. doi: 10.1137/1.9780898719574.
- [66] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, 2009. doi: 10.1007/b13794.
- [67] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1998. doi: 10.1017/CBO9780511802256.
- [68] S. Vatedka, N. Kashyap, and A. Thangaraj. Secure compute-and-forward in a bidirectional relay. *IEEE Transactions on Information Theory*, 61(5):2531–2556, 2015. doi: 10.1109/TIT.2015.2412114.

- [69] N. Verzelen and E. Gassiat. Adaptive estimation of high-dimensional signal-to-noise ratios. *Bernoulli*, 24(4B):3683–3710, 2018. doi: 10.3150/17-BEJ975.
- [70] L. Wang and E. Tchetgen Tchetgen. Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(3):531–550, 2018. doi: 10.1111/rssb.12262.
- [71] L. Wang, L. D. Brown, T. T. Cai, and M. Levine. Effect of mean on variance function estimation in nonparametric regression. *The Annals of Statistics*, 36(2):646–664, 2008. doi: 10.1214/009053607000000901.
- [72] B. D. Williamson, P. B. Gilbert, N. R. Simon, and M. Carone. A general framework for inference on algorithm-agnostic variable importance. *Journal of the American Statistical Association*, 118(543):1645–1658, 2023. doi: 10.1080/01621459.2021.2003200.
- [73] Y. Wu and P. Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, 2016. doi: 10.1109/TIT.2016.2548468.
- [74] Y. Wu and P. Yang. Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *The Annals of Statistics*, 47(2):857–883, 2019. doi: 10.1214/17-AOS1665.
- [75] Z. Zeng, S. Balakrishnan, Y. Han, and E. H. Kennedy. Causal inference with high-dimensional discrete covariates, 2024. arXiv:2405.00118v3.
- [76] Y. Zhang, L. Liu, and Z. Zhang. Higher-order debiased estimators for general treatment models. *Econometric Theory*, pages 1–43, 2026. doi: 10.1017/S0266466626100516. First View; remark numbering refers to arXiv:2606.01706v2.