Minimax Rates Under Global Overlap Bounds

Jacob Dorn*

August 27, 2026

Abstract

This note presents minimax regression rates for Dorn (2026)'s weak overlap setting. I show that weak overlap degrades the effective outcome smoothness rate by a factor of $1 - 1/\gamma_0$. The bulk of the note is a proof that the minimax sup-norm rate is equal to the minimax pointwise rate, without the usual polylogarithmic penalty. Intuitively, there cannot be too many points with too weak overlap locally without making overlap worse globally. Further, I show the minimax rate is achievable adaptively without knowing or directly estimating the degree of overlap weakness.

1 Introduction

Weak overlap makes outcome regression more difficult. Under weak overlap, the probability of an observation being treated can be near zero or one, so that there are fewer observations to be used in regression.

This paper precisely characterizes the degree to which weak overlap makes outcome regression more difficult. I work under an outcome Hölder smoothness class of order $\beta_\mu > 0$ and a global overlap tail lower bound γ_0, where γ_0 is the number of inverse propensity moments that are assumed to be finite.

Within this class, I show that weak overlap scales the effective outcome smoothness by $1 - 1/\gamma_0$. Equivalently, the effective dimension is increased by $\alpha^{(\mathrm{Mou})} = d/(\gamma_0 - 1)$,

*Artificial intelligence was used to suggest changes and identify potential errors. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-2039656 and by grant T32 HS026116 from the Agency for Healthcare Research and Quality. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation or the Agency for Healthcare Research and Quality.

where $\alpha^{(\mathrm{Mou})}$ is Mou et al. (2023)'s parameter α. This result comes from balancing the effect of the worst possible order of overlap weakness: there are fewer treated observations nearby, so that the optimal estimator uses a larger bandwidth with greater bias. Weaker levels of overlap correspond to slower consistency rates. For any fixed level of overlap weakness, there is some level of Hölder smoothness that achieves a regression consistency rate that is arbitrarily near the parametric rate.

The bulk of this note takes care to show that the optimal sup-norm rate is equal to the optimal pointwise rate, without the usual polylogarithmic factor separating the optimal pointwise and global rates (Stone, 1982). The argument leverages a novel construction partitioning the covariate space into regions with increasingly strong overlap assurances in such a way as the contribution of each region to the worst-case error is half as large as the previous contribution. In this way, the contribution of infinitely many regions to the global error is controlled uniformly.

I further show that the minimax sup-norm rate is achievable adaptively without knowledge or directly estimating the degree of overlap weakness. The approach I take is a grid-within-grid construction: starting with a dense grid that corresponds to strict overlap; estimating the effective overlap weakness at each point based on the optimal bandwidth that balances bias and variance locally; and taking a new coarses grid that ensures that each point is separated by more than the largest bandwidth of any gridpoint, so that each gridpoint can be characterized independently.

The construction requires the local polynomial design to be nondegenerate at each point. I show that nondegeneracy is ensured if the propensity score function is more smooth than $\alpha^{(\mathrm{Mou})}$. If so, local overlap guarantees imply the sufficiently high-order propensity expansion must be nonzero; the propensity score then looks like its Taylor expansion locally, yielding a nondegenerate design. Without this assumption, it is possible for nature to start with the worst level of overlap and add treated mass to concentrate local treated observations on a degenerate curve. Such a construction makes local polynomial estimators more slowly consistent or even inconsistent; however, it is likely that more sophisticated estimators could restore the minimax consistency rate here.

This work relates to the literature on regression with degenerate designs (Stone, 1982; Hall et al., 1997; Gaïffas, 2005; Mou et al., 2023), and most closely to Mou et al. (2023), whose singularity rate for a first-order Sobolev space matches my minimax rate for the implied Hölder class. The result that the sup-norm regression rate comes without a polylogarithmic loss is, to my knowledge, new.

The paper is as follows. Section 2 describes the weak overlap and Hölder smoothness setting. Section 3 shows that propensity smoothness of order $\beta_e > \alpha^{(\text{Mou})}$ ensures a sufficient condition for nondegenerate eigenvalues. Section 4 states the main claim of this note: that the minimax sup-norm rate does not have the usual polylogarithmic loss, and it can be achieved adaptively. Section 4.1, the proof of this claim, is the bulk of the note. Section 5 concludes.

Notation. $E_P[\cdot]$ and $E[\cdot]$ denote expectations under the maintained distribution P over data Z, and $\sup_{P\in A} B$ the supremum of B over distributions $P \in A$ satisfying any maintained restrictions on the distribution and nuisance functions. I write $P(E_n)$ for the probability of an event E_n under the distribution P, leaving the number of draws n implicit where possible. For nonnegative sequences c_n, d_n: $c_n \ll d_n$ and $d_n \gg c_n$ if for all $\epsilon > 0$, $c_n \leq \epsilon d_n$ for n large enough; $c_n \lesssim d_n$ and $d_n \gtrsim c_n$ if $c_n \leq k d_n$ for some $k > 0$ and all n large enough; and $c_n = \Theta(d_n)$ if there are $0 < k_1 \leq k_2 < \infty$ such that $c_n/d_n \in [k_1, k_2]$ for all n large enough. For a sequence of distributions P_n, I write $c_n = O_{P_n}(d_n)$ if for all $\epsilon > 0$ there is a $\delta > 0$ with $\limsup_{n\to\infty} P_n(|c_n|/d_n > \delta) < \epsilon$ and $c_n = o_{P_n}(d_n)$ if the same conclusion holds for all $\epsilon, \delta > 0$. I use $o(d_n)$ and $O(d_n)$ in the usual sense. I use log to refer to the natural logarithm. I write $\|f\|_{L^p(P)} = E_P[|f|^p]^{1/p}$ for finite p and $\|f\|_{L^\infty(P)} = \lim_{p\to\infty} \|f\|_{L^p(P)}$. Following a multivariate version of Tsybakov (2009), $f \in \Sigma(\beta, L)$ if its order-$\ell = \max\{z \in \mathbb{Z} \cap [0, \beta)\}$ derivatives are well-defined and $D^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \cdots}$ satisfies $\|D^\alpha f(x) - D^\alpha f(x')\| \leq L\|x - x'\|^{\beta-\ell}$ for all x, x' and every multiindex α with $\|\alpha\| = \ell$. For such a β, β_μ, or β_e, I write the associated number of derivatives as ℓ, ℓ_μ, or ℓ_e. I use P_n to refer to an arbitrary sequence of distributions for the purposes of computing suprema. I use $\text{Leb}(A)$ to refer to the Lebesgue measure. I write $\lim_{x\to z^+} f(x)$ and $\lim_{x\to z^-} f(x)$ for the right- and left-hand limits of f at z. I use local polynomial regression to refer to the estimator $\hat{\mu}_n(x)(x) = U(0)^T \hat{\beta}_n(x)$, where $\hat{\beta}_n(x)$ solves the local regression problem $\operatorname{argmin}_{\beta\in\mathbb{R}^{\binom{d+\ell_\mu}{d}}} \sum_{i=1}^n D_i\big(Y_i - \beta^T U((X_i - x)/h)\big)^2 K((X_i - x)/h)$, where $U(z)$ is the vector of all monomials in z up to degree ℓ_μ, for simplicity $K(z) = \mathbf{1}\{\|z\| \leq 1\}$, and where when $\hat{\beta}_n(x)$ is not uniquely identified then $\hat{\mu}_n(x)(x) = 0$.

2 Setting

I follow the standard semiparametric model for a binary treatment. I assume the econometrician has access to n samples Z of data (X, D, Y), where $X \in \mathbb{R}^d$ are covariates, $D \in \{0, 1\}$ is a binary treatment, and $Y \in \mathbb{R}$ is an observed outcome.

For simplicity, I focus the theoretical analysis on estimating the average potential outcome $E_P[E[Y \mid X, D = 1]]$. The average treatment effect $E_P[E_P[Y \mid X, D = 1] - E_P[Y \mid X, D = 0]]$ follows as a corollary, and the average treatment effect on the treated $E_P[D(Y(1) - Y(0))]/E_P[D] = (E[Y - E_P[E[Y \mid X, D = 0]]])/E_P[D]$ follows without too much difficulty.

I assume that the propensity score $e(X) = P(D = 1 \mid X)$ is in $(0, 1)$ almost surely, so that the average potential outcome can be identified as $E_P[DY/e(X)]$. I refer to regions of the covariate space in which the propensity can be arbitrarily close to zero as singularities. Strict overlap rules out singularities. However, I do not bound the propensity score away from zero or one. Settings in which singularities are allowed are sometimes called "limited overlap" (Khan and Tamer, 2010; Chaudhuri and Hill, 2024).

I derive uniform convergence rates under bounds on overlap weakness. I follow Ma and Wang (2020), who provide important building blocks in my analysis, and parameterize overlap weakness through a tail parameter γ_0. I extend their results to a model family $\mathscr{P}$ over distributions whose overlap is at least as strong as that tail parameter, in addition to some regularity conditions.

Assumption 1 (Distribution moments and tail bound)**.** Let $\mathscr{P}$ be a nonempty family of distributions, and write $e(X) = P(D = 1 \mid X)$ and $\mu(X) = E_P[Y \mid X, D = 1]$. Then every $P \in \mathscr{P}$ is a distribution over $(X, D, Y) \in \mathbb{R}^d \times \{0, 1\} \times \mathbb{R}$ satisfying the following conditions for some $q > 3, M, \sigma_{\min}, C > 0$, and $\gamma_0 > 1$:

(a) *Conditional moments.* $E[|Y|^q \mid X, D = 1] \leq M^q < \infty$ almost surely.

(b) *Unconditional moments.* $\mathrm{Var}(\mu(X)) \leq M$.

(c) *Residuals.* $\mathrm{Var}(Y \mid X, D) \geq \sigma_{\min}^2$.

(d) *Propensity tail.* $P(e(X) \leq \pi) \leq C\pi^{\gamma_0 - 1}$ for all $\pi \in [0, 1]$.

Assumption 1 is a uniform generalization of Ma and Wang (2020)'s slowly varying tails assumption. Assumptions 1(a) through 1(c) are regularity conditions that rule out cases like perfectly predictable outcomes.

Assumption 1(d) provides the substantial restriction on $\mathscr{P}$: overlap may be weak in the sense that γ_0 is finite, but there is some minimal γ_0 and C that provides an upper bound on the inverse propensity's tail behavior. Intuitively, the tail parameter γ_0 represents the number of moments on $D/e(X)$ that are assumed to be finite. $\gamma_0 \leq 1$ would allow a mass point of $e(X)$ at zero.

I distinguish between three cases of the degree of overlap weakness. I follow Heiler and Kazak (2021) and use *strict overlap* to refer to the case in which the propensity score is bounded away from zero almost surely. Under strict overlap, Assumption 1(d) holds for any finite $\gamma_0 > 1$, and most results here hold after replacing γ_0 with infinity.

I use *somewhat weak overlap* to refer to the case that the inverse propensity distribution may be heavy-tailed, but the inverse propensity weight still has strictly more than two moments, which corresponds to $\gamma_0 > 2$.[1] Heuristically, somewhat weak overlap is the case in $P(e(X) \leq \pi)$ is positive near zero, but goes to zero at an $o(\pi)$ rate. Under somewhat weak overlap, the semiparametric efficiency bound is finite and known to be achievable (Chen et al., 2008; Hirshberg and Wager, 2021; Heiler and Kazak, 2021).

I use *very* weak overlap to refer to the case in which $\gamma_0 < 2$. Under very weak overlap, the semiparametric bound is infinite, so that there is no regular $\sqrt{n}$-consistent estimator (Khan and Tamer, 2010), and IPW with even known propensity scores may fail to be asymptotically normal (Ma and Wang, 2020). A phase transition occurs at $\gamma_0 = 2$, in which case logarithmic terms can emerge.

Figure 1 illustrates behavior for simulated data with various values of γ_0. When the propensity score $e(X)$ has a well-defined density, $\gamma_0 = 2$ corresponds to a roughly uniform distribution of propensity scores (Ma and Wang, 2020). When γ_0 is above two, the density of propensity scores tends to zero at zero; when γ_0 is below two, the density of propensity scores can tend to infinity at zero. Heuristically, there are never too many treated observations with very small propensity scores, but γ_0 governs the degree to which there can be many untreated observations with very small propensity scores.

I characterize optimal nonparametric regression rates under Hölder continuity. For convenience, I fix the covariates to be uniform over a specific hypercube in $\mathbb{R}^d$

[1] Distributions satisfying somewhat weak overlap are sometimes said to satisfy "strict overlap" or "overlap" (Heiler and Kazak, 2021; Bruns-Smith et al., 2024), and distributions that exhibit very weak overlap are sometimes called "heavy tailed" (Chaudhuri and Hill, 2024).

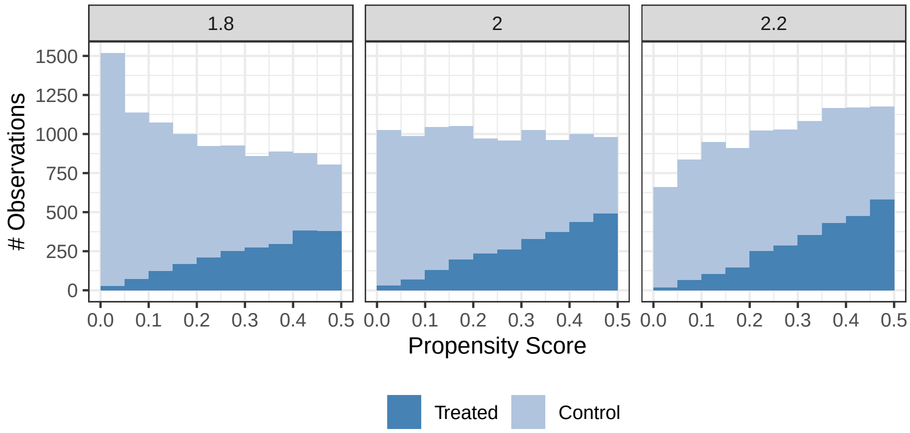

Figure 1: Simulations of 10,000 observations of $e(X)$ with $P(e(X) \leq \pi) = \pi^{\gamma_0 - 1}$ for increasing values of γ_0.

with constant variance.

Assumption 2 (Hölder continuity and fixed domain)**.** $\mathscr{P}$ satisfies Assumption 1, and for all $P \in \mathscr{P}$, $X \sim Unif([-1,1]^d)$ and $Y \mid X, D \sim \mathcal{N}(D\mu_P(X) + (1-D)\mu'_P(X), \sigma^2_{\min})$, with μ, μ' in the Hölder class $\Sigma(\beta_\mu, L)$ for some fixed $\beta_\mu, L > 0$.

Most of these assumptions are standard assumptions for studying local polynomial regression under strict overlap (Stone, 1982). However, weak overlap allows the density of treated observations to approach zero, in which case outcome regression near the resulting singularity becomes a problem of estimation under a degenerate design (Hall et al., 1997; Gaïffas, 2005). I assume normal outcomes in order to simplify the characterization of the optimal rate. I work within a Hölder class to study the effect of weak overlap, not because the Hölder class should motivate AIPW: many estimators besides AIPW can yield valid t-statistics under Assumption 2 with sufficient smoothness (Hirano et al., 2003; Newey and Robins, 2017; Armstrong and Kolesár, 2021), and higher-order estimators can have better statistical properties (Robins et al., 2008). The distinct advantage of AIPW debiasing is in structural-agnostic models (Balakrishnan et al., 2025). Hölder continuity instead offers a one-parameter family in which the cost of weak overlap can be characterized exactly.

It is known that in this case but under strict overlap, the optimal pointwise rate is $n^{\frac{-\beta_\mu}{2\beta_\mu + d}}$, and the optimal uniform rate $(n/\log(n))^{\frac{-\beta_\mu}{2\beta_\mu + d}}$ has a polylogarithmic penalty, in the sense that the optimal uniform rate is worse by some polynomial factor of $\log(n)$.

3 Sufficient Smoothness

I require that treated observations cannot concentrate in small regions.

Assumption 3 (No adversarial propensity concentration)**.** There are parameters $\rho, \nu > 0$ such that for all $h > 0$ small enough and all $P \in \mathscr{P}$ and $x_0 \in [-1,1]^d$, $P(e(X) \geq \rho \sup_{\|x-x_0\| \leq h} e(x) \mid \|X - x_0\| \leq h) > \nu$.

Dorn (2026) shows that this condition holds if the propensity function is sufficiently smooth. When the propensity function is nonsmooth, it is possible for nature to choose a distribution with better local overlap properties that induces local polynomial degeneracy issues. It is likely that the feasibility proof could bypass Assumption 3 through use of some hypothetical degeneracy-robust estimator, but that would be outside the scope of this work.

Proposition 1 (Sufficient conditions for non-trivial concentration)**.** *Suppose Assumption 2 holds and there is an $L > 0, \beta_e > \frac{d}{\gamma_0 - 1}$ such that $P(D = 1 \mid X) \in \Sigma(\beta_e, L)$ for all $P \in \mathscr{P}$. Then Assumption 3 holds.*

Lemma 1 (Minimal coefficient)**.** *Suppose Assumption 2 holds and $e(X) \in \Sigma(\beta_e, L)$ for some $\beta_e > \frac{d}{\gamma_0 - 1}$. For $0 \leq j < \beta_e$, Let $c_j(v \mid x_0)$ be the coefficient in the j^{th}-order Taylor expansion of $e(x)$ around x_0 applied in the direction of v. Then there is a $c^* > 0$ such that for all x_0, there exists an $x \neq x_0$ and a $0 \leq j \leq \alpha^{(\mathrm{Mou})}$ such that $x_0 + \frac{x - x_0}{\|x - x_0\|} \in [-1,1]^d$ and $c_j\left(\frac{x-x_0}{\|x-x_0\|} \mid x_0\right) \geq c^*$.*

Proof of Lemma 1. If $\beta_e \leq 1$, the claim is immediate by Lemma 3. I therefore proceed assuming $\beta_e > 1$.

Proof by contradiction. Suppose for all $c > 0$, there exists a P and an x_0 such that $c_j\left(\frac{x-x_0}{\|x-x_0\|} \mid x_0\right) < c$ for all $j \leq \alpha^{(\mathrm{Mou})}$.

Define:

$$h^* = \operatorname{argsup}_{h \in (0,1]} \sum_{\alpha^{(\mathrm{Mou})} < j < \beta_e} \bar{c}_j (h)^j + L_e (h)^{\beta_e} \leq \frac{C^{1/d}}{\lceil \alpha^{(\mathrm{Mou})} + 2 \rceil} (h)^{\frac{d}{\gamma_0 - 1}}.$$

This h^* is well-defined because as h tends to zero from above, the left-hand side is of a lower order than the right-hand side. By continuity, $h_n^* \in (0,1]$ and satisfies this weak inequality.

Take $c^* = \min_{0 \le j \le \alpha^{(\mathrm{Mou})}} \frac{C^{\frac{1}{1-\gamma_0}}}{\lceil \alpha^{(\mathrm{Mou})}+2 \rceil} (h^*)^{\frac{d}{\gamma_0 - 1} - j}$. I claim that for this c^*, the content of Lemma 1 holds.

Proof by contradiction. Suppose, not, and there is an $x_0 \in [-1,1]^d$ and an x as above such that $c_j \left(\frac{x - x_0}{\|x - x_0\|} \mid x_0 \right) < c^*$ for all $j \le \alpha^{(\mathrm{Mou})}$. Write $c' = C^{\frac{1}{1-\gamma_0}} \frac{\lceil \alpha^{(\mathrm{Mou})}+1 \rceil}{\lceil \alpha^{(\mathrm{Mou})}+2 \rceil}$. Then for all $x \in [-1,1]^d / x_0$ such that $\|x - x_0\| \le h^*$ and $x_0 + \frac{x - x_0}{\|x - x_0\|} \in [-1,1]^d$:

$$
\begin{aligned}
e(x) &= f(x \mid x_0) + g(x \mid x_0) \\
&= \sum_{0 \le j \le \alpha^{(\mathrm{Mou})}} c_j \left(\frac{x - x_0}{\|x - x_0\|} \mid x_0 \right) \|x - x_0\|^j \\
&\quad + \sum_{\alpha^{(\mathrm{Mou})} < j < \beta_e} c_j \left(\frac{x - x_0}{\|x - x_0\|} \mid x_0 \right) \|x - x_0\|^j + L \|x - x_0\|^{\beta_e} \\
&\le \sum_{0 \le j \le \alpha^{(\mathrm{Mou})}} \frac{C^{\frac{1}{1-\gamma_0}}}{\lceil \alpha^{(\mathrm{Mou})} + 2 \rceil} (h^*)^{\frac{d}{\gamma_0 - 1} - j} \|x - x_0\|^j + \frac{C^{\frac{1}{1-\gamma_0}}}{\lceil \alpha^{(\mathrm{Mou})} + 2 \rceil} (h^*)^{\frac{d}{\gamma_0 - 1}} \le c' \, (h^*)^{\frac{d}{\gamma_0 - 1}} .
\end{aligned}
$$

Then:

$$
\begin{aligned}
P\left(e(X) \le c' \, (h^*)^{\frac{d}{\gamma_0 - 1}} \right) &\ge P\left(x_0 + \frac{X - x_0}{\|X - x_0\|} \in [-1,1]^d, \|X - x_0\| \ge h^* \right) \\
&\ge (h^*)^d = C \left(C^{\frac{1}{1-\gamma_0}} \, (h^*)^{d/(\gamma_0 - 1)} \right)^{\gamma_0 - 1} > C \left(c' \, (h^*)^{\frac{d}{\gamma_0 - 1}} \right)^{\gamma_0 - 1} .
\end{aligned}
$$

But by Assumption 1(d), $P\left(e(X) \le c' \, (h^*)^{\frac{d}{\gamma_0 - 1}} \right) \le C \left(c' \, (h^*)^{\frac{d}{\gamma_0 - 1}} \right)^{\gamma_0 - 1}$. Contradiction.

Therefore, for this $c^* > 0$ and all x_0, there exists an $x \ne x_0$ and a $0 \le j \le \alpha^{(\mathrm{Mou})}$ such that $c_j \left(\frac{x - x_0}{\|x - x_0\|} \mid x_0 \right) \ge c^*$. $\square$

Lemma 2 (Helper claim for Proposition 1)**.** *For any $s \in \{-1, +1\}^d$, define the orthant in direction s intersected with the unit ball as $D_s = \{v : \|v\| \le 1, \; s_i v_i \ge 0 \; \forall i\}$. I claim that there is a constant $\kappa = \kappa(d, \beta_e) > 0$ such that every polynomial q of degree at most ℓ_e that is nonnegative on D_s for some s satisfies $\sup_{v \in D_s} q(v) \ge \kappa \sup_{\|v\| \le 1} |q(v)|$.*

Proof of Lemma 2. Pick some s. Write the number of such coefficients as p. For any coefficients $c \in \mathbb{R}^p$, write the implied polynomial as $q_c(v)$ and define $m(c) = \sup_{\|v\| \le 1} |q_c(v)|$ and $m^+(c) = \sup_{v \in D_s} q_c(v)$. Define $\mathcal{C} = \{c \in \mathbb{R}^p : m(c) = 1 \text{ and } \inf_{v \in D_s} q_c(v) \ge 0\}$. Take $\kappa_s \equiv \inf_{c \in \mathcal{C}} m^+(c)$. It remains to show $\kappa_s > 0$. $\mathcal{C}$ is closed by inspection and abounded by $m(c) = 1$, so compact, and non-empty. The function $m^+(c)$ is continuous by inspection and strictly positive on $\mathcal{C}$ because otherwise $m^+(c) = 0$ and $q_c(v) \geq 0$ on D_s would imply $c = 0 \notin \mathcal{C}$. Therefore $\kappa_s = \inf_{c \in \mathcal{C}} m^+(c) > 0$. Finally, take $\kappa(d, \beta_e) = \min_{s \in \{-1,1\}^d} \kappa_s$. $\kappa > 0$ because there are finitely many options.

Now, take some $v \in D_s$. If $q = q_c$ is a constant zero on D_s, then $c = 0$ by solving for coefficients and $\sup_{v \in D_s} q(v) = \kappa \sup_{\|v\| \leq 1} |q(v)| = 0$. Otherwise, $\sup_{v \in D_s} q(v) \geq \kappa_s \sup_{\|v\| \leq 1} |q(v)| \geq \kappa \sup_{\|v\| \leq 1} |q(v)|$. □

Proof of Proposition 1. Proof by contradiction. Suppose not, and for all $\rho, \nu > 0$, there is an

$$h \in \left(0, (C^{1/d}/(2L))^{1/(\beta_e - d/(\gamma_0 - 1))}\right],$$

$x_0 \in [-1,1]^d$, and $P \in \mathscr{P}$ such that $P(e(X) \geq \rho \sup_{\|x - x_0\| \leq h} e(x) \mid \|X - x_0\| \leq h) \leq \nu$.

Fix $\rho = \kappa/2$ and $\nu = 1/n$, with $\kappa = \kappa(d, \beta_e)$ from Lemma 2. Since the claim was assumed to fail for every $\rho, \nu > 0$, for each n, there is a a bandwidth h_n, a point x_0, and a $P_n \in \mathscr{P}$ with

$$P_n\left(e(X) \geq \rho \sup_{\|x - x_0\| \leq h_n} e(x) \;\middle|\; \|X - x_0\| \leq h_n\right) \leq \gamma_n, \qquad \gamma_n = 1/n \to 0.$$

Because $[-1,1]^d$ is compact and the coefficients are in a compact space, there is a subsequence of n for which x_0 and the local polynomial coefficients of $e(x)$ around x_0 are convergent.

Construct a further subsubsequence as follows. Write $A_n = \{x : \|x - x_0\| \leq h_n\}$ and $\mathcal{V}_n = \{v : \|v\| \leq 1,\ x_0 + h_n v \in [-1,1]^d\}$, so that $V_n \equiv (X - x_0)/h_n$ is uniform on $\mathcal{V}_n$ given $X \in A_n$. Take $s_{n,i} = +1$ if $x_{0,i} \leq 0$ and $s_{n,i} = -1$ otherwise, so that $D_{s_n} \subseteq \mathcal{V}_n \subseteq \{v : \|v\| \leq 1\}$ for all n large enough that $h_n \leq 1$, uniformly in x_0. There are finitely many values of s_n, so pass to a further subsequence along which s_n is constant s. Without loss of generality, suppose this is the sequence.

Let $f_n(\cdot \mid x_0)$ be the order-ℓ_e Taylor polynomial of e at x_0 under P_n. Let $f^*(\cdot \mid x_0)$ for the local polynomial coefficients at the limiting coefficients and write $\tilde{f}_n = f_n - f^*$ and $g_n = e - f_n$, so that $e = f^* + \tilde{f}_n + g_n$.

Let j^* be the lowest-order nonzero coefficient of f^*, which is well-defined and of order at most $\lfloor \alpha^{(\mathrm{Mou})} \rfloor = \lfloor d/(\gamma_0 - 1) \rfloor \leq \ell_e$ by the uniform coefficient lower bound in Lemma 1 and the convergence of coefficients on the subsubsequence by construction.

First, I show that $h_n \to 0$. Suppose not, and $\limsup_{n \to \infty} h_n = h^* > 0$. Without loss of generality, suppose $\lim_{n\to\infty} h_n = h^*$. Write $A = \{x \mid \|x - x_0\| \le h^*\}$. Since f^* has a nonzero coefficient of order $j^* \le \ell_e$, $\sup_{\|x-x_0\|\le h^*} f^*(x \mid x_0) > 0$, and the leading behavior is of order $(h^*)^{j^*}$. Then by continuity of densities and bounds on derivatives, for all n large enough,

$$
\begin{aligned}
\gamma_n &\ge P_n\left(e(X) \ge \frac{\rho}{2}\left(\sup_{\|x-x_0\|\le h^*} e(x)\right) \mid \|X - x_0\| \le h^*\right) \\
&\ge P_n\left(e(X) \ge \frac{3\rho}{8}\left(\sup_{\|x-x_0\|\le h^*} f^*(x \mid x_0)\right) \mid \|X - x_0\| \le h^*\right) \\
&\quad - \mathbf{1}\left\{\sup_{\|x-x_0\|\le h^*} |e(x) - f^*(x \mid x_0)| \ge \frac{\rho}{8} \sup_{\|x-x_0\|\le h^*} f^*(x \mid x_0)\right\} \\
&\ge P_n\left(e(X) \ge \frac{3\rho}{8}\left(\sup_{\|x-x_0\|\le h^*} f^*(x \mid x_0)\right) \mid \|X - x_0\| \le h^*\right) \\
&\quad - \mathbf{1}\left\{\sup_{x\in A} \|f_n(x \mid x_0) - f^*(x \mid x_0)\| + |g_n(x \mid x_0)| \ge \frac{\rho}{16} \sup_{\|x-x_0\|\le h^*} f^*(x \mid x_0)\right\} \\
&= P_n\left(e(X) \ge \frac{3\rho}{8}\left(\sup_{\|x-x_0\|\le h^*} f^*(x \mid x_0)\right) \mid \|X - x_0\| \le h^*\right) - \mathbf{1}\left\{O((h^*)^{\beta_e}) \gtrsim (h^*)^{j^*}\right\} \\
&\ge P_n\left(e(X) \ge \frac{\rho}{4}\left(\sup_{\|x-x_0\|\le h^*} f^*(x \mid x_0)\right) \mid \|X - x_0\| \le h^*\right),
\end{aligned}
$$

which is constant (because h^* is fixed) and positive (because f^* is continuous and not a constant zero). Therefore $\gamma_n \not\to 0$. Contradiction.

Write $\tilde{c}_{j,n}(v) = \sum_{\|\alpha\|=j} D^\alpha e(x_0) v^\alpha$ for the order-j Taylor coefficient function of e at x_0 under P_n, $m_{n,j} \equiv \sup_{\|v\|=1} |\tilde{c}_{j,n}(v)|\, h_n^{j-j^*}$, and $m_n \equiv \max_{0\le j\le j^*} m_{n,j}$.

Now I show over two cases that for every subsequence, there is a further subsubsequence, a nonzero polynomial q of degree at most ℓ_e that is nonnegative on D_s, a sequence of functions $r_n(v)$ with $\sup_{\|v\|\le 1} |r_n(v)| = o(1)$, and a sequence of positive numbers d_n for which $e(x) = d_n\big(q(h_n^{-1}(x - x_0)) + r_n(v)\big)$ uniformly on $\|x - x_0\| \le h_n$. By construction, $e(x) \ge 0$ for all $h_n^{-1}(x - x_0) = v \in D_s$ for all n large enough, so that if this is the case, then $q(v) \ge 0$ for all $v \in D_s$. (Otherwise $e(x_0 + h_n v) < 0$ once $|r_n(v)| < -q(v)$.)

Case 1: there is a further subsequence for which $m_n \to 0$, then on that subsequence, for any $0 < \|x - x_0\| \le h_n$, write $v = h_n^{-1}(x - x_0)$ and $q(v) = c_{j^*}\big(v/\|v\|\big)\|v\|^{j^*}$, where c_{j^*} are the coefficients of f^*. Note that $\tilde{f}_n = o(h_n^{j^*})$ by construction of m_n to capture the lower-than-j^* terms and because $m_n \to 0$. Then

$$e(x) = f^*(x \mid x_0) + \tilde{f}_n(x \mid x_0) + g_n(x \mid x_0) = h_n^{j^*} q(v) + o(h_n^{j^*}),$$

so that the desired result holds with $d_n = h_n^{j^*}$.

Case 2: $\liminf_{n\to\infty} m_n > 0$. Define $d_{j,n}(v) = \tilde{c}_{j,n}(v) h_n^{j-j^*}/m_n$, a rescaled version of the input to $m_{n,j}$. Take the polynomial $q_n(v) = \sum_{0\le j\le j^*} d_{j,n}(v/\|v\|)\|v\|^j$. By construction, $\max_{0\le j\le j^*} \sup_{\|v\|\le 1} |d_{j,n}(v)| = 1$, so that there is a $\bar{j} \le j^*$ with a subsequence for which $\sup_{\|v\|\le 1} |d_{\bar{j},n}(v)| = 1$ for all n. Because $\|d_{j,n}(v)\| \le 1$ on the unit sphere, the coefficients c_n of $d_{j,n}$ are in a compact space, so that there is a further subsequence for which $d_{j,n}(v)$ converges uniformly on the unit sphere and for which $\bar{j}$ is constant. Call that limiting polynomial $d_j^*(v)$. Define the j^*-order polynomial $q(v) = \sum_{0\le j\le j^*} d_j^*(v/\|v\|)\|v\|^j$. Because $\sup_{\|v\|=1} |d_{\bar{j}}^*(v)| = 1$, $q(v)$ is not a constant zero. Now suppose $0 < \|x - x_0\| < h_n$ and write $v = h_n^{-1}(x - x_0)$. Note that $\tilde{f}_n(x_0 + h_n v) - m_n h_n^{j^*} q_n(v)$ is a series of polynomials with bounded coefficients and terms of order at least $j^* + 1$. Then:

$$\begin{aligned} e(x) &= m_n h_n^{j^*} q(v) + m_n h_n^{j^*}\big(q_n(v) - q(v)\big) + \tilde{f}_n(x) - m_n h_n^{j^*} q_n(v) + g_n(x) \\ &= m_n h_n^{j^*} q(v) + o(m_n h_n^{j^*}) + O(h_n^{j^*+1}) + O(h_n^{\beta_e}) = m_n h_n^{j^*} q(v) + o(m_n h_n^{j^*}), \end{aligned}$$

so that the desired result holds with $d_n = m_n h_n^{j^*}$ using $m_n > 0$ on this subsubsequence.

Now, on that subsubsequence, write $m_q = \sup_{\|v\|\le 1} |q(v)| > 0$. Define $G = \{v \in D_s : q(v) > 3/4\kappa m_q\}$, which is constant on the subsubsequence. By construction of the κ in Lemma 2 and by continuity, the Lebesgue measure $\mathrm{Leb}(G) > 0$. Then, for all n large enough and all $v \in G$, and noting that $\rho = \kappa/2$,

$$e(x_0 + h_n v) = d_n\big(q(v) + r_n(v)\big) \ge d_n\big(\tfrac{3}{4}\kappa m_q - \sup_{\|v\|\le 1} |r_n(v)|\big) \ge \rho\, d_n\big(m_q + \sup_{\|v\|\le 1} |r_n(v)|\big) \ge \rho \sup_{x\in A_n} e(x),$$

where the last inequality uses $\sup_{x\in A_n} e(x) = \sup_{v\in\mathcal{V}_n} d_n(q(v) + r_n(v)) \le d_n\big(m_q + \sup_{\|v\|\le 1} |r_n(v)|\big)$. Thus, for all n large enough, $v \in G$ implies $e(x_0 + h_n v) \ge \rho \sup_{x\in A_n} e(x)$. Let $\omega_d = \mathrm{Leb}\left(\{v : \|v\| \le 1\}\right)$. Then

$$P_n\left(e(X) \ge \rho \sup_{x\in A_n} e(x) \mid X \in A_n\right) \ge \frac{\mathrm{Leb}(G)}{\mathrm{Leb}(\mathcal{V}_n)} \ge \frac{\mathrm{Leb}(G)}{\omega_d} > 0$$

which has a lower bound independently of n. Then $\gamma_n \not\to 0$. Contradiction. $\square$

4 Minimax Rates

In the worst case, weak overlap of order $\gamma_0 > 1$ plays a role equivalent to scaling the effective outcome smoothness downward by $(1 - 1/\gamma_0)$.

Theorem 1 (Weak overlap reduces effective outcome smoothness)**.** *Suppose γ_0 is finite, and define $r^*_{\mu,n} = n^{\frac{-\beta^*}{2\beta^*+d}}$, where $\beta^* = \beta_\mu (1 - 1/\gamma_0)$. Then*

(i) $r^*_{\mu,n}$ is a pointwise (and uniform) rate upper bound. *There exists a $c > 0$ and a such that if $\mathscr{P}$ satisfies Assumptions 2 and 3, then*

$$\liminf_{n\to\infty} \inf_{\hat{\mu}} \sup_{P\in\mathscr{P}} P\left(|\hat{\mu}(0) - \mu(0)| > c r^*_{\mu,n}\right) > 0.$$

(ii) $r^*_{\mu,n}$ is an achievable uniform (and pointwise) rate*. Suppose $\mathscr{P}$ satisfies Assumptions 2 and 3. Then there exists an estimator $\hat{\mu}(x)$ such that for all $\epsilon > 0$, there is a finite $c(\epsilon)$ such that*

$$\limsup_{n\to\infty} \sup_{P\in\mathscr{P}} P\left(\|\hat{\mu} - \mu\|_{L^\infty(P)} > c(\epsilon) r^*_{\mu,n}\right) \le \epsilon.$$

Further, the estimator can be computed without knowledge of the overlap bound γ_0.

The intuition for this result is that as γ_0 becomes smaller, the global overlap guarantee becomes weaker and nature can choose a point at which there are increasingly few treated observations nearby to use in regression. The mean-squared-error-optimal response to weak overlap is to choose a wider bandwidth than usual that reduces the variance at the cost of increasing the regression bias. The resulting rate penalty occurs even under somewhat weak overlap.

The minimax-optimal tradeoff is reflected in Theorem 1. The strict overlap optimal pointwise convergence rate $n^{\frac{-\beta_\mu}{2\beta_\mu+d}}$ is recovered as the limit as $\gamma_0 \to \infty$, and smaller levels of γ_0 correspond to slower consistency rates. Theorem 1 shows that weak overlap can be interpreted as increasing the effective dimension by $d/(\gamma_0 - 1)$, or equivalently as degrading the effective smoothness rate from β_μ to $\beta_\mu(1 - 1/\gamma_0)$. The result parallels Mou et al. (2023)'s analysis of the effect of weak overlap in a

Sobolev space: their setting implies Hölder continuity of order $\beta_\mu = 1/2$; my pointwise rate for $\beta_\mu = 1/2$, $d = 1$ matches their pointwise rate at the singularity; and my degraded outcome consistency rate can be interpreted as increasing the effective dimension from d to $d\gamma_0/(\gamma_0 - 1)$, which generalizes the Mou et al. singularity exponent α to higher dimensions. In the limit in which γ_0 tends to one, the rate in Theorem 1 can become arbitrarily poor. For example, the difficulty of estimating a twice continuously differentiable function when $\gamma_0 = 2$ is comparable to the difficulty of estimating a Lipschitz-continuous function under strict overlap.

At a high level, the construction follows a standard framework by constructing a grid over $[-1, 1]^d$, choosing bandwidths for each gridpoint to balance the local polynomial bias (increasing in bandwidth) and variance (decreasing in bandwidth), estimating $\hat{\mu}(x)$ at each gridpoint via local polynomial regression, and then interpolating between gridpoints. However, the construction here avoids knowledge of γ_0 by exploiting that for any fixed x with the worst possible overlap under $\mathscr{P}$, taking h to solve $\sum_i \mathbf{1}\{\|X_i - x\| \leq h, D_i = 1\} = h^{-2\beta_\mu}$ yields a bandwidth on the order of $n^{\frac{-1}{2\beta_\mu + d\gamma_0/(\gamma_0 - 1)}}$. I construct a data-adaptive grid by starting with a hypothetical grid that would be appropriate under strict overlap, but otherwise is too dense. I then use the dense grid's implied weak overlap bound to shrink the gridpoint density until I can ensure that every gridpoint has sufficient treated observations within some hypothetical largest gridpoint. I then adapt each gridpoint's bandwidth downward to solve $\sum_i \mathbf{1}\{\|X_i - x\| \leq h, D_i = 1\} = h^{-2\beta_\mu}$, which allows the estimator to exploit better overlap properties away from singularities.

This construction avoids the usual polylogarithmic penalty on uniform rates. Intuitively, the lack of polylogarithmic penalty follows because the vast majority of space must exhibit better overlap than the worst-case rate, or else the implied global overlap bound would be even weaker. Formally, I use a peeling device: I partition the gridpoints into strata of increasingly strong overlap guarantees, and therefore larger maximum capacity and smaller bandwidths. By carefully choosing the sequence of stratum capacities (which exhibit a logarithmic rate penalty) and bandwidths (which exhibit a polynomial rate improvement), I am able to derive a geometric stratum sup-norm error guarantee. As a result, the uniform error over the grid is bounded by a convergent series dominated by the first stratum. The key technical requirement for this inductive argument to hold is that the capacity of the initial stratum must be sufficiently large, but finite, to ensure the growth of subsequent strata.

4.1 Proof

I now build up to proving the main claim of this note, Theorem 1.

Lemma 3 (Minimal expected nearby observations). *Suppose the conditions of Theorem 1 hold. Define* $A(x_0 \mid h) = \{x : x \in [-1,1]^d, \|x - x_j\| \le h\}$. *Then there is a constant* $c_0(C, \gamma_0, d) > 0$ *such that*

$$\inf_{x_0 \in [-1,1]^d, P \in \mathscr{P}, h \in (0,1]} E_P\left[D \mid X \in A(x_0 \mid h)\right] \ge c_0(C, \gamma_0, d) h^{d/(\gamma_0 - 1)}.$$

Proof of Lemma 3. Since $X \sim Unif([-1,1]^d)$, there exists a $c_1(d) > 0$ such that $\inf_{x \in [-1,1]^d} P(X \in A(x \mid h)) \ge c_1(d) h^d$ for all $h \in (0,1]$. Now let $\pi(C, d, h, \gamma_0)$ solve $C\pi^{\gamma_0 - 1} = (c_1(d)/2)h^d$, i.e. $\pi(C, d, h, \gamma_0) = (c_1(d)/(2C))^{1/(\gamma_0-1)} h^{d/(\gamma_0-1)}$.

Then, for any $x_0 \in [-1,1]^d$,

$$\begin{aligned} P(X \in A(x_0 \mid h), e(X) \le \pi(C, d, h, \gamma_0)) \le P(e(X) \le \pi(C, d, h, \gamma_0)) \le C\pi(C, d, h, \gamma_0)^{\gamma_0 - 1} \\ = c_1(d)h^d/2 \le P(X \in A(x_0 \mid h))/2. \end{aligned}$$

Therefore

$$\begin{aligned} E_P\left[D \mid X \in A(x \mid h)\right] &\ge E_P\left[D 1\{e(X) \ge \pi(C, d, h, \gamma_0)\} \mid X \in A(x \mid h)\right] \\ &\ge \pi(C, d, h, \gamma_0) P\left(e(X) \ge \pi(C, d, h, \gamma_0) \mid X \in A(x \mid h)\right) \\ &\ge \pi(C, d, h, \gamma_0)/2 = (c_1(d)/(2C))^{1/(\gamma_0-1)}/2 h^{d/(\gamma_0-1)}. \end{aligned}$$

Taking $c_0(C, \gamma_0, d) = (c_1(d)/(2C))^{1/(\gamma_0-1)}/2$ completes the claim. □

Lemma 4 (Minimal eigenvalue technical result). *Suppose the conditions of Theorem 1 hold. There is an* $h' > 0$ *such that*

$$\lim_{\varepsilon \to^+ 0} \sup_{P \in \mathscr{P}, x_0 \in [-1,1]^d, h \in (0, h'], \|v\| = 1} P\left(\|v^T U\| \le \varepsilon \mid D = 1, \|X - x_0\| \le h\right) = 0.$$

Proof of Lemma 4. Take some sequence of $\varepsilon_n \to^+ 0$. Take h', ρ, ν from Assumption 3.

Let P_n be a sequence of distributions in $\mathscr{P}$, let v_n be a sequence of vectors with $\|v\| = 1$, let h_n be a sequence in $(0, h']$, and write $A_n = \{x : \|x - x_0\| \le h_n\}$. Then:

$$P_n\left(\|v_n^T U\| \le \sqrt{\varepsilon_n} \mid D = 1, X \in A_n\right) = \frac{E_{P_n}\left[e(X)\mathbf{1}\left\{\|v_n^T U\| \le \sqrt{\varepsilon_n}\right\} \mid X \in A_n\right]}{E_{P_n}\left[e(X) \mid X \in A_n\right]}$$

$$
\begin{aligned}
&\leq \frac{\left(\sup_{x\in A_n} e(x)\right) P_n\left(\|v_n^T U\| \leq \sqrt{\varepsilon_n} \mid X\in A_n\right)}{P_n\left(e(X)\geq \rho\left(\sup_{x\in A_n} e(x)\right)\right)\rho\left(\sup_{x\in A_n} e(x)\right)} \\
&\leq \frac{\left(\sup_{x\in A_n} e(x)\right) P_n\left(\|v_n^T U\| \leq \sqrt{\varepsilon_n} \mid X\in A_n\right)}{\nu\rho\sup_{x\in A_n} e(x)} \\
&= \frac{P_n\left(\|v_n^T U\| \leq \sqrt{\varepsilon_n} \mid X\in A_n\right)}{\rho\nu} = O\left(\sqrt{\varepsilon_n}\right) = o(1).
\end{aligned}
$$

□

Lemma 5 (Uniform minimal expected eigenvalue)**.** *Suppose the conditions of Theorem 1 hold, and let $U(v)$ be the vector of zero-through-ℓ_μ-order interactions of v. Define:*

$$
\lambda(\bar h) \equiv \inf_{h\in(0,\bar h],\|v\|=1,x_0\in[-1,1]^d,P\in\mathscr{P}} v^T E_P\left[U\left(\frac{X-x_0}{h}\right)U\left(\frac{X-x_0}{h}\right)^T \mid D=1,\|X-x_0\|\leq h\right]v,
$$

$$
\lambda^* \equiv \liminf_{\bar h\to^+0} \lambda(\bar h),
$$

where $U(\cdot)$ is the ℓ_μ-order local polynomial interaction matrix. Then $\lambda^*>0$.

Proof of Lemma 5. Let h' be from Lemma 4. Fix some $\bar h\in(0,h']$. Let h_n be a sequence in $(0,\bar h]$, let v_n be a sequence of vectors with $\|v\|=1$, let $x_{0,n}$ be a sequence of points in $[-1,1]^d$, and let P_n be a sequence of distributions in $\mathscr{P}$. Let $A_n=\{x: \|x-x_{0,n}\|\leq h_n\}$.

By Lemma 4, there is a $\varepsilon>0$ such that:

$$
\inf_{P\in\mathscr{P},x_0\in[-1,1]^d,h\in(0,h_n],\|v\|=1} P\left(\|v^T U\|\geq\varepsilon \mid D=1,\|X-x_0\|\leq h\right)>0.
$$

Call this infimum $\delta>0$. $\|v^T U\|$ is bounded, so δ is finite.

Define:

$$
\begin{aligned}
\lambda_n &\equiv v_n^T E_{P_n}\left[U\left(\frac{X-x_{0,n}}{h_n}\right)U\left(\frac{X-x_{0,n}}{h_n}\right)^T \mid D=1, X\in A_n\right]v_n \\
&= E_{P_n}\left[v_n^T U\left(\frac{X-x_{0,n}}{h_n}\right)\left(v_n^T U\left(\frac{X-x_{0,n}}{h_n}\right)\right)^T \mid D=1, X\in A_n\right] \\
&= E_{P_n}\left[\left(v_n^T U\left(\frac{X-x_{0,n}}{h_n}\right)\right)^2 \mid D=1, X\in A_n\right]
\end{aligned}
$$

$$\geq \varepsilon^2 P_n\left(\left|v_n^T U\left(\frac{X - x_{0,n}}{h_n}\right)\right| \geq \varepsilon \mid D = 1, X \in A_n\right) \geq \varepsilon^2\delta > 0.$$

Therefore, for all $\bar{h} \leq h'$, $\lambda(\bar{h}) \geq \varepsilon^2\delta$. Therefore $\lambda^* \geq \varepsilon^2\delta > 0$. □

Lemma 6 (Uniform functional approximation across multiple gridpoints)**.** *Suppose the conditions of Theorem 1 hold. Let $f_{k_n,n}(v) : [-1,1]^d \to \mathbb{R}$ be a set of k_n functions that are uniformly bounded. Let $\mathcal{S}(m, \underline{h})$ be the set of sets of m tuples (x, h) with $x \in [-1,1]^d$ and $h \in [\underline{h}_n, h']$, where h' comes from Lemma 5's notion of h small enough and where for all $(x_1, h_1), (x_2, h_2) \in S \in \mathcal{S}(m, \underline{h})$ with $x_2 \neq x_1$, there are no points $x \in [-1,1]^d$ with $\|x - x_1\| \leq h_1$ and $\|x - x_2\| \leq h_2$. Write $\tilde{c} = 2^{\frac{2(1+\gamma_0)}{1-\gamma_0}} C^{\frac{2}{1-\gamma_0}} (\sup_v |f(v)|)^{-2}$. Suppose (i) $\underline{h}_n^{\frac{d\gamma_0}{\gamma_0-1}} \gg n^{-1}$; (ii) m_n be a sequence tending to infinity such for all fixed $a > 0$, $m_n \ll \exp\left(-an\underline{h}_n^{\frac{d\gamma_0}{\gamma_0-1}}\right)$; and (iii) and $k_n\left(1 - \left(\frac{e-1}{e}\right)^{\frac{m_n}{\exp\left(\log(1/2)+\tilde{c}^{-1}n^{-1}\underline{h}_n^{-\frac{d\gamma_0}{\gamma_0-1}}\right)}}\right) \to 0$. Then there is a sequence of $\varepsilon_n \to^+ 0$ such that for all $k = 1, \ldots, k_n$:*

$$\sup_{P \in \mathscr{P}, S \in \mathcal{S}(m_n, \underline{h}_n)} P\left(\max_j \left|\frac{\sum_i D_i K\left(\frac{\|X_i - x_j\|}{h_{n,j}}\right) f_{k,n}\left(\frac{X_i - x_j}{h_{n,j}}\right) - E\left[DK\left(\frac{\|X - x_j\|}{h_{n,j}}\right) f_{k,n}\left(\frac{X - x_j}{h_{n,j}}\right)\right]}{nE_{P_n}\left[DK\left(\frac{\|X - x_j\|}{h_{n,j}}\right)\right]}\right| \geq \varepsilon_n\right) = o(1).$$

Proof of Lemma 6. For simplicity, I proceed assuming K is the uniform bandwidth scaled downwards by one-half; the proof would require far more care if the kernel took on nonzero values for an unbounded set. Also, for simplicity, assume that $e(X)$ is bounded above by one-half — the difficulty here comes from small propensity scores, and is already substantial. Let the upper bound of $|f|$ be $b \geq 0$.

Define $R_n \equiv \tilde{c}n\underline{h}_n^{\frac{d\gamma_0}{\gamma_0-1}}$ and

$$V_{n,j,k} \equiv \sum_i D_i K\left(\frac{\|X_i - x_j\|}{h_{n,j}}\right) \frac{f_{k,n}\left(\frac{X_i - x_j}{h_{n,j}}\right) - E\left[f_{k,n}\left(\frac{X - x_j}{h_{n,j}}\right) \mid DK\left(\frac{\|X - x_j\|}{h_{n,j}}\right) = 1\right]}{nE_{P_n}\left[DK\left(\frac{\|X - x_j\|}{h_{n,j}}\right)\right]}.$$

By construction, there is a sequence of $\varepsilon_n^+ \to 0$ such that $k_n\left(1 - (1 - 1/e)^{\frac{m_n}{\exp\left(\log(1/2) + R_n\varepsilon_n^2\right)}}\right) \to 0$.

Note that the events $D_i K\left(\frac{\|X_i - x_j\|}{h_{n,j}}\right)$ are mutually disjoint for a given i across j; therefore write $j(i)$ for the j such that $D_i K\left(\frac{\|X_i - x_{j(i)}\|}{h_{n,j}}\right) = 1$ if feasible, and write

$j(i) = 0$ if no such j exists.

Let P_n be a sequence of distributions in $\mathscr{P}$, let S_n be a sequence of sets in $\mathcal{S}(m_n, \underline{h}_n)$, and let $\{(x_{n,j}, h_{n,j})\}$ be a sequence of associated points and bandwidths. By Lemma 3, there is a constant $c_0(C, \gamma_0, d)$ such that for all j, n,

$$E_{P_n}\left[DK\left(\frac{\|X - x_j\|}{h_{n,j}}\right)\right] \geq c_0(C, \gamma_0, d)\underline{h}_n^{\frac{d\gamma_0}{\gamma_0 - 1}}.$$

Consider the event A of $\{i, j(i)\}$. Note that for any given j, by the Chernoff bound for binomial random variables,

$$P_n\left(\frac{\sum D_i K\left(\frac{\|X_i - x_j\|}{h_{n,j}}\right)}{nE\left[DK\left(\frac{\|X - x_j\|}{h_{n,j}}\right)\right]} \geq 2\right] \leq \exp\left(\frac{-nE\left[DK\left(\frac{\|X - x_j\|}{h_{n,j}}\right)\right]}{3}\right) \leq \exp\left(-\tilde{c}\frac{b^2}{3}n\underline{h}_n^{\frac{d\gamma_0}{\gamma_0 - 1}}\right),$$

$$\text{So that } P_n\left(\max_j \frac{\sum D_i K\left(\frac{\|X_i - x_j\|}{h_{n,j}}\right)}{nE\left[DK\left(\frac{\|X - x_j\|}{h_{n,j}}\right)\right]} \leq 2\right) \leq 1 - \sum_j P_n\left(\frac{\sum D_i K\left(\frac{\|X_i - x_j\|}{h_{n,j}}\right)}{nE\left[DK\left(\frac{\|X - x_j\|}{h_{n,j}}\right)\right]} \geq 2\right)$$

$$\leq 1 - O\left(m_n \exp\left(-\tilde{c}\frac{b^2}{3}n\underline{h}_n^{\frac{d\gamma_0}{\gamma_0 - 1}}\right)\right) = 1 - o(1).$$

I therefore proceed under the high probability event that A is such that $\max_j \frac{\sum D_i K\left(\frac{\|X_i - x_j\|}{h_{n,j}}\right)}{nE\left[DK\left(\frac{\|X - x_j\|}{h_{n,j}}\right)\right]} \leq 2$.

I now apply the Hoeffding inequality to the $\sum DK\left(\frac{\|X - x_j\|}{h_{n,j}}\right) \leq 2E\left[DK\left(\frac{\|X - x_j\|}{h_{n,j}}\right)\right]$ elements of $V_{n,j}$ conditional on A, for all k, n, j

$$P_n\left(|V_{n,j,k}| \geq \varepsilon_n\right) \leq 2\exp\left(\frac{-2\left(nE_{P_n}\left[DK\left(\frac{\|X - x_j\|}{h_{n,j}}\right)\right]\right)^2 \varepsilon_n^2}{\left(\sum D_i K\left(\frac{\|X_i - x_j\|}{h_{n,j}}\right)\right) b^2}\right) \leq 2\exp\left(-R_n \varepsilon_n^2\right).$$

Then:

$$P_n\left(\max_{k=1}^{k_n} \max_{j=1}^{m_n} |V_{n,j,k}| \geq \varepsilon_n \mid A\right) \leq \sum_{k=1}^{k_n}\left(1 - \prod_{j=1}^{m_n}\left(1 - P_n\left(|V_{n,j,k}| \geq \varepsilon_n\right)\right)\right)$$

$$\leq k_n\left(1 - \prod_{j=1}^{m_n}\left(1 - 2\exp\left(-R_n \varepsilon_n^2\right)\right)\right)$$

$$= k_n \left(1 - \left(1 - 2\exp\left(\frac{-1}{16b^2} n \underline{h}_n^{d\frac{\gamma_0}{\gamma_0-1}} \varepsilon_n^2\right)\right)^{m_n}\right)$$

$$= k_n \left(1 - \left(\left(1 - \frac{1}{\exp\left(\log(1/2) + R_n\varepsilon_n^2\right)}\right)^{\exp\left(\log(1/2)+R_n\varepsilon_n^2\right)}\right)^{\frac{m_n}{\exp\left(\log(1/2)+R\right.}}$$

$$= k_n \left(1 - (1 - 1/e - o(1))^{\frac{m_n}{\exp\left(\log(1/2)+R_n\varepsilon_n^2\right)}}\right) = o(1).$$

Therefore $P\left(\max_{k=1}^{k_n} \max_{j=1}^{m_n} |V_{n,j,k}| \geq \varepsilon_n\right) = o(1)$. □

Lemma 7 (Nondegeneracy of local polynomial eigenvalues at estimated bandwidths over gridpoints)**.** *Suppose the conditions of Theorem 1 hold. Fix some $k > 0$ and let $\mathcal{S}_n$ be a set of g_n points $x_{n,j} \in [-1,1]^d$, with $g_n \leq k'n$ for some fixed $k' > 0$. For each $x_{n,j}$, let $h_{n,j} = suph : n\sum_i D_i \mathbf{1}\{\|X_i - x_{n,j}\|\} \leq kh^{-2\beta_\mu}$ and let $h^*_{n,j}$ solve $nE[D\mathbf{1}\{\|X - x_{n,j}\|\}] = kh^{-2\beta_\mu}$. Then there is a $\delta_n \to^+ 0$ such that*

$$\limsup_{n\to\infty} \sup_{P\in\mathscr{P}} P\left(\max_j \left|\frac{\lambda_{\min}\left(\frac{\sum_i D_i K\left(\frac{\|X_i - x_j\|}{h_{n,j}}\right) U\left(\frac{X_i - x_j}{h_{n,j}}\right) U\left(\frac{X_i - x_j}{h_{n,j}}\right)^T}{\sum_i D_i K\left(\frac{\|X_i - x_j\|}{h_{n,j}}\right)}\right)}{\lambda_{\min}\left(\frac{E_P\left[DK\left(\frac{\|X - x_j\|}{h^*_{n,j}}\right) U\left(\frac{X - x_j}{h^*_{n,j}}\right) U\left(\frac{X - x_j}{h^*_{n,j}}\right)^T\right]}{E_P\left[DK\left(\frac{\|X - x_j\|}{h^*_{n,j}}\right)\right]}\right)} - 1\right| \geq \delta_n\right) = o(1).$$

Proof of Lemma 7. For convenience, I proceed assuming K is the uniform bandwidth, scaled downwards by one-half. Let P_n be a sequence of distributions in $\mathscr{P}$.

Apply Lemma 6 to the sequence $\underline{h}_n = n^{\frac{-1}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0-1}}} / \log(n)$ and $m_n = 3g_n \ll \exp\left(an\underline{h}_n^{d\frac{\gamma_0}{\gamma_0-1}}\right)$ for all $a >$, to yield a sequence of $\varepsilon_n^{(a)} \to^+ 0$.

Let $h_{n,j}$ solve $\min_h \left|N_n(h \mid x_{n,j}) - kh^{-2\beta_\mu}\right|$, and let $h^*_{n,j}$ solve $\min_h \left|N^*_n(h \mid x_{n,j}) - kh^{-2\beta_\mu}\right|$, where $N^*_n(h \mid x_{n,j}) = nE\left[e(X)\mathbf{1}\{\|X - x_{n,j}\| \leq h\}\right]$. Further, let $[\underline{h}_{n,j}, \bar{h}_{n,j}]$ be the convex hull of the set of h that solve $N^*_n(h \mid x_{n,j})(1+\varepsilon_n^{(a)})^{-1} = kh^{-2\beta_\mu}$ or $N^*_n(h \mid x_{n,j})(1+\varepsilon_n^{(a)}) = kh^{-2\beta_\mu}$. By Lemma 6, with probability tending to one, $h_{n,j} \in [\underline{h}_{n,j}, \bar{h}_{n,j}]$.

Write:

$$A_{n,j}(h, h') = \frac{E_{P_n}\left[DK\left(\frac{\|X - x_{n,j}\|}{h}\right) U\left(\frac{X - x_{n,j}}{h}\right) U\left(\frac{X - x_{n,j}}{h}\right)^T\right]}{E_{P_n}\left[DK\left(\frac{\|X - x_{n,j}\|}{h'}\right)\right]}$$

$$B_{n,j}(h,h') = \frac{\sum_{i=1}^{n} D_i K\left(\frac{\|X_i - x_{n,j}\|}{h}\right) U\left(\frac{X_i - x_{n,j}}{h}\right) U\left(\frac{X_i - x_{n,j}}{h}\right)^T}{\sum_{i=1}^{n} D_i K\left(\frac{\|X_i - x_{n,j}\|}{h'}\right)}$$

The claim is that there is a $\delta_n \to^+ 0$ such that

$$P_n\left(\left|\frac{\lambda_{\min}\left(B_{n,j}(h_{n,j},h_{n,j})\right)}{\lambda_{\min}\left(A_{n,j}(h^*_{n,j},h^*_{n,j})\right)} - 1\right| \geq \delta_n\right) = o(1).$$

$A^{-1}_{n,j}$ is symmetric, so that $\|A^{-1}_{n,j}\|^2_{(op)}$ is equal to the squared largest eigenvalue of $A^{-1}_{n,j}$, where $\|\cdot\|_{(op)}$ is equal to the operator norm. By Lemma 5, $\|A^{-1}_{n,j}\|^2_{(op)} = \lambda_{\min}(A_{n,j})^{-2}$ is bounded above.

Take $m_n = 3g_n \leq 3k'n$, $k_n = (n+1)$, and $\underline{h}_n = n^{\frac{-1}{2\beta\mu + d\frac{\gamma_0}{\gamma_0 - 1}}} / \log(n)^{\frac{\gamma_0 - 1}{d\gamma_0}}$, so that the first condition of Lemma 6 holds by Lemma 5. The second condition holds because $m_n = O(n)$. The third condition holds by L'Hopital's Rule applied to $n\left(1 - \left(\frac{e-1}{e}\right)^{\frac{an}{\exp(b + cn^d / \log(n))}}\right) \to 0$ for any fixed $a, b, c, d > 0$. Thus, I may apply Lemma 6 to the $k_n = n+1$ bounded functions $f(v) = U_k(v)U_p(v)$ and $\mathbf{1}\left\{\frac{h_{n,j}}{h^*_{n,j}} \leq v \leq \frac{\bar{h}_{n,j}}{h^*_{n,j}}\right\}$ at the bandwidths $\underline{h}_n$. Let the associated ε_n terms be $\varepsilon_n^{(b)}$ Then:

$$\begin{aligned}
&\max_j \left|A_{n,j}(h^*_{n,j},h^*_{n,j})_{k,p} - B_{n,j}(h_{n,j},h_{n,j})_{k,p}\right| \\
\leq &\max_j \left|\begin{array}{cc} & A_{n,j}(h^*_{n,j},h^*_{n,j})_{k,p} \\ - & B_{n,j}(h^*_{n,j},h^*_{n,j})_{k,p}\end{array}\right| + \max_j \left|\begin{array}{cc} & B_{n,j}(h^*_{n,j},h^*_{n,j})_{k,p} \\ - & B_{n,j}(h^*_{n,j},h_{n,j})_{k,p}\end{array}\right| + \max_j \left|\begin{array}{cc} & B_{n,j}(h^*_{n,j},h_{n,j})_{k,p} \\ - & B_{n,j}(h_{n,j},h_{n,j})_{k,p}\end{array}\right| \\
\leq &o_{P_n}(1) + O_{P_n}\left(\max_j \left|A_{n,j}(h^*_{n,j},h^*_{n,j})_{k,p}\right|\right) \max_j \left|\left(1 - \frac{\sum D_i K\left(\frac{\|X_i - x_{n,j}\|}{h^*_{n,j}}\right)}{\sum D_i K\left(\frac{\|X_i - x_{n,j}\|}{h_{n,j}}\right)}\right)\right| \\
&\quad + O\left(\max_j \frac{\sum D_i \mathbf{1}\left\{\|X_i - x_{n,j}\| \in [\underline{h}_{n,j}, \bar{h}_{n,j}]\right\}}{\sum D_i \mathbf{1}\left\{\|X_i - x_{n,j}\| \leq \underline{h}_{n,j}\right\}}\right) \\
= &o_{P_n}(1) + O_{P_n}\left(\varepsilon_n^{(a)}\right) + o\left(\left(1 + \varepsilon_n^{(a)}\right)^2 (1 + \varepsilon_n^{(b)})\right) = o_{P_n}(1).
\end{aligned}$$

Therefore $\max_j \|A_{n,j} - B_{n,j}\|_{(op)} = o(1)$.

Then, by well-known arguments (Horn and Johnson, 2013, p. 381):

$$\begin{aligned}
\max_j \left|\lambda_{\min}(B_{n,j}) - \lambda_{\min}(A_{,nj})\right| &= \max_i \left|\lambda_{\max}(B^{-1}_{n,j}) - \lambda_{\max}(A^{-1}_{,nj})\right| \leq \max_j \|B^{-1}_{n,j} - A^{-1}_{n,j}\|_{(op)} \\
&\leq \max_j \frac{\|A^{-1}_{n,j}\|^2_{(op)} \|A_{n,j} - B_{n,j}\|_{(op)}}{1 - \|A^{-1}_{n,j}(A_{n,j} - B_{n,j})\|_{(op)}} = \max_j \frac{O_{P_n}(1) o_{P_n}(1)}{1 - o_{P_n}(1)} = o_{P_n}(1),
\end{aligned}$$

so that

$$\max_j \left| \frac{\lambda_{\min}(B_{n,j})}{\lambda_{\min}(A_{n,j})} - 1 \right| = \max_j \left| \frac{\lambda_{\min}(B_{n,j}) - \lambda_{\min}(A_{n,j})}{\lambda_{\min}(A_{n,j})} \right| \leq \max_j \frac{|\lambda_{\min}(B_{n,j}) - \lambda_{\min}(A_{,nj})|}{\min_j \lambda_{\min}(A_{n,j})} = o_{P_n}(1),$$

so that the full claim holds. □

Lemma 8 (KL divergence). *For any given $L' > 0$ and $x_0 \in [-1,1]^d$, construct distributions $P_{n,m}$ for $m = 1,2$ as follows. Draw $X \sim U([-1,1]^d)$. Draw $D \mid X \sim Bern\left(C^{-(\gamma_0-1)} P_{n,m}\left(\|X - x_0\| \leq \|x - x_0\|\right)^{1/(\gamma_0-1)}\right)$. Finally, draw $Y \mid X, D \sim \mathcal{N}(D\mu_{n,m}(X), \sigma^2_{\min})$ where $\mu_{n,1}(X) = 0$ and*

$$\mu_{n,2}(X) = \frac{L'}{\exp\left(\frac{-4}{3}\right)} n^{\frac{-\beta_\mu}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0 - 1}}} \exp\left(\frac{-1}{1 - \left(\frac{2\|X - x_0\|}{n^{\frac{-1}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0-1}}}} \right)^2} \right) \mathbf{1}\left\{ \|X - x_0\| \leq \frac{n^{\frac{-1}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0-1}}}}{2} \right\}.$$

Finally, define $\mathscr{P} = \{P_{n,m}\}_{n=1\infty, m=1,2}$. Then (i) if there exists a $\mathscr{P}_0$ satisfying Assumptions 2 and 3, then there exists a fixed L' such that $\mathscr{P}$ satisfies Assumptions 2 and 3. (ii) there is an $\alpha > 0$ finite such that $KL\left(P_{n,1}, P_{n,2}\right) \leq \alpha$.

Proof of Lemma 8. First, I show (i) that such an L' exists. Because there exists a $P \in \mathscr{P}_0$ in this set and every $P_{n,m}$ has the smallest possible range of $Y - \mu(X) \mid X, D = 1$, it must be that for every $P_{n,m} \in \mathscr{P}$, Assumption 1(a) and Assumption 1(c) hold. Also, if I define $V(x) = P_{n,m}(\|X - x_0\| \leq \|x - x_0\|)$ which is distributed $Unif([0,1])$, then for all $P_{n,m}$:

$$P_{n,m}\left(e(X) \leq \pi\right) = P_{n,m}\left(C^{-(\gamma_0-1)} V(X)^{\frac{1}{\gamma_0-1}} \leq \pi\right) = P_{n,m}\left(V(X) \leq C\pi^{\gamma_0-1}\right) = C\pi^{\gamma_0-1}.$$

Therefore, Assumption 1(d) hold with $C = C'$ and $\underline{\pi}$ small enough. It is also clear that $E_{P_{n,m}}[Y \mid X, D = 0] = E_{P_{n,0}}[Y \mid X, D = 1] \in \Sigma(\beta_\mu, L)$, since these functions are a constant zero.

It remains to show that there is an $L' > 0$ such that for all n, m, $\text{Var}_{P_{n,2}}(\mu_{n,2}(X)) \leq M$ (Assumption 1(b) and completing Assumption 2) and $\mu_{n,2}(X) \in \Sigma(\beta_\mu, L)$ (Assumption 3). For the variance upper bound:

$$\mathrm{Var}_{P_{n,2}}\left(\mu_{n,2}(X)\right) \leq (L')^2 \left(n^{\frac{-2\beta_\mu}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0-1}}}\right) \exp\left(\frac{8}{3} - 2\right) \leq (L')^2 \exp(2/3).$$

so that it suffices to take $L' \leq \sqrt{M \exp(-2/3)}$. For Hölder continuity, write $\mu_{n,2}(X) = \frac{L'}{a} n^{\frac{-\beta_\mu}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0-1}}} g_a\left(\frac{\|X - x_0\|}{n^{\frac{-1}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0-1}}}}\right)$. If $a > 0$ is small enough, then g_a is infinitely differentiable and in $\Sigma(\beta_\mu, 1/2)$. Thus, by standard arguments (Tsybakov, 2009), if L' is small enough, $\mu_{n,2}(X) \in \Sigma(\beta_\mu, L)$. Thus, there is an $L' > 0$ such that $\mathscr{P}$ satisfies Assumptions 2 and 3.

Second, I show the main claim (ii). It is useful to write $h_n = n^{\frac{-1}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0-1}}}$. Then:

$$\begin{aligned}
KL(P_{n,1}, P_{n,2}) &= n P_{n,1}\left(\log \frac{dP_{n,1}}{dP_{n,2}}\right) \\
&\leq n P_{n,1}\left(\|X - x_0\| \leq \frac{h_n}{4}\right) P_{n,1}\left(D = 1 \mid \|X - x_0\| \leq \frac{h_n}{4}\right) \frac{\left(\frac{L'}{\exp\left(\frac{-4}{3}\right)} h_n^{\beta_\mu} \exp\left(\frac{-1}{1-1/4}\right)\right)^2}{2} \\
&= nk2^{-2d-1}(L')^2 \left(\frac{k}{C}\right)^{\frac{1}{\gamma_0-1}} h_n^{d + \frac{d}{\gamma_0-1} + 2\beta_\mu} = k2^{-2d-1}(L')^2\left(\frac{k}{C}\right)^{\frac{1}{\gamma_0-1}} = \text{“}\alpha.\text{”}
\end{aligned}$$

□

Proposition 2 (Inductive grouping)**.** *Fix some $\beta_\mu > 0$, $\gamma_0 > 1$, and $d \geq 1$. Then there is a $\underline{\delta} > 0$ and a $\pi \in (0, 2^{-1/\beta_\mu}]$ such that if $\delta \geq \underline{\delta}$, then the sequence defined by $m_n^{(0)} = \delta$, $m_n^{(1)} = \exp(\delta/2)$, and $m_n^{(k+1)} = \exp\left(\delta 2^{-(k+1)}\left(m_n^{(k)}/m_n^{(0)}\right)^{2\beta_\mu(\gamma_0-1)/(2\beta_\mu + d\gamma_0/(\gamma_0-1))}\right)$ diverges to infinity, and the sequence of $t_n^{(k)} = \left(m_n^{(k-1)}/m_n^{(0)}\right)^{(\gamma_0-1)/(2\beta_\mu+d\gamma_0/(\gamma_0-1))}$ satisfies $t_n^{(k+1)} \geq t_n^{(k)}/\pi$ for $k = 1, 2, \ldots$.*

Proof of Proposition 2. First, I construct such a $\underline{\delta}$. Let δ_I be the smallest value of x such that $(x/\log(x))^{1/(2\beta_\mu)} \geq 4^{1/(2\beta_\mu)}$, $x/\log(x) \geq 4$, and $x \geq 4$. Then take:

$$\underline{\delta} = \sup d \geq \delta_I \text{ s.t. } (d/\log(d))^{\frac{2\beta_\mu + d\gamma_0/(\gamma_0-1)}{2\beta_\mu(\gamma_0-1)}} \geq d^{d/(2\log(d))-1} \text{ or } d = \delta_I$$

This is defined because as d tends to infinity, the inequality does not hold.

Now I show the claim holds. Take $\pi = (\log(\underline{\delta})/\,\underline{\delta})^{1/(2\beta_\mu)}$ and $\underline{r} = (\underline{\delta}/\log(\underline{\delta}))^{\frac{2\beta_\mu + d\gamma_0/(\gamma_0-1)}{2\beta_\mu(\gamma_0-1)}} =$

$\pi^{-\frac{2\beta_\mu+d\gamma_0/(\gamma_0-1)}{\gamma_0-1}}$. Note that by construction, $\pi \in (0, 2^{-1/\beta_\mu}]$.

Let $\delta \geq \underline{\delta}$ and the sequence as above be given.

Take $t_n^{(0)} = (\log(\delta)/\delta)^{1/(2\beta_\mu)}$, so that $\log\left(m_n^{(k)}\right) = \delta 2^{-(k)}\left(t_n^{(k)}\right)^{2\beta_\mu}$ for all $k = 0, 1, \ldots$. Note as a result that $\log\left(m_n^{(k)}\right) = \log(\delta)\left(\prod_{0\leq j<k} \frac{t_n^{(j+1)}/t_n^{(j)}}{2^{1/(2\beta_\mu)}}\right)^{2\beta_\mu}$. In particular, $\log\left(m_n^{(k+1)}/m_n^{(k)}\right) = \log\left(m_n^{(k)}\right)\left(\left(t_n^{(k+1)}/t_n^{(k)}\right)^{2\beta_\mu}/2\right)$ for all $k = 0, 1, \ldots$.

Next, I show by induction that for all $k = 0, 1, \ldots$, $t_n^{(k+1)} \geq t_n^{(k)}/\pi$ and $m_n^{(k+1)} \geq \underline{\rho} m_n^{(k)}$. In the base case, $k = 0$, $t_n^{(k+1)}/t_n^{(k)} = \pi^{-1}$ by construction of π, and $m_n^{(1)}/m_n^{(0)} = \delta^{\delta/(2\log(\delta))-1} \geq \underline{r}$ by construction of $\underline{r}$. In the inductive case,

$$
\begin{aligned}
t_n^{(k+1)}/t_n^{(k)} &= \left(m_n^{(k)}/m_n^{(k-1)}\right)^{(\gamma_0-1)/(2\beta_\mu+d\gamma_0/(\gamma_0-1))} \geq (\underline{r})^{(\gamma_0-1)/(2\beta_\mu+d\gamma_0/(\gamma_0-1))} \\
m_n^{(k+1)}/m_n^{(k)} &= \left(m_n^{(k)}\right)^{\left(t_n^{(k+1)}/t_n^{(k)}\right)^{2\beta_\mu}/2-1} \geq \delta^{\delta/(2\log(\delta))-1} \geq \underline{r}.
\end{aligned}
$$

Thus, for all $k = 0, 1, \ldots$, $t_n^{(k+1)} \geq t_n^{(k)} \geq \pi$ and $m_n^{(k+1)} \geq \underline{r} m_n^{(k)}$ for some $\underline{r} \geq 4$ and $\pi \in (0, 2^{-1/\beta_\mu}]$. The remaining claims hold by inspection. □

Lemma 9 (k_n^* characterization)**.** *Let $\underline{\delta}$, π, $m_n^{(k)}$, and $h_n^{(k)}$ be as in Proposition 2. Let k_n^* be the smallest k for which $\pi^{(k-1)\beta_\mu} \leq 1/\log(n)$. Then $k_n^* = O(\log(\log(n)))$.*

Proof of Lemma 9. $(k_n^* - 1) \leq \frac{\log(1/\pi)}{\beta_\mu}\log(\log(n)) = \Theta(\log(\log(n)))$. □

Lemma 10 (Minimal small-propensity points)**.** *Suppose the conditions of Theorem 1 hold, and there are m_n points $x_j \in [-1,1]^d$ such that $\inf_{j\neq j'} \|x_j - x_{j'}\| \geq h$ and $\max_j nE[D\mathbf{1}\{\|X - x_j\| \leq h/d\}] \leq s(h/d)^{-2\beta_\mu}$. Then there is a universal constant $B \geq 2$ that depends on $C, \gamma_0, \beta_\mu, d, s$ such that $m_n \leq B\left(nh^{2\beta_\mu+d\frac{\gamma_0}{\gamma_0-1}}\right)^{1-\gamma_0}$.*

Proof of Lemma 10. Note that the set of points with $\mathbf{1}\{\|X - x_j\| \leq h/d\}$ are mutually exclusive across x. Note also that for any set A, $E\left[D\mathbf{1}\{X \in A\}\right] \leq \pi$ implies $P(X \in A)/2 \leq P\left(X \in A, e(X) \leq 2\pi\right) \leq P(e(X) \leq 2\pi)$. As a result, for j as in the statement,

$$
\begin{aligned}
E\left[e(X) \mid \|X - x_j\| \leq h/d\right] &\leq \frac{nE[e(X)\mathbf{1}\{\|X - x_j\| \leq h/d\}]}{nP(\|X - x_j\| \leq h/d)} \leq \frac{sd^{d+2\beta_\mu}}{nh^{d+2\beta_\mu}} \\
m_n\left(h/d\right)^d/2 &\leq P\left(e(X) \leq \frac{2sd^{d+2\beta_\mu}}{nh^{d+2\beta_\mu}}\right) \leq C\left(\frac{2sd^{d+2\beta_\mu}}{nh^{d+2\beta_\mu}}\right)^{\gamma_0-1} \\
m_n &\leq C2^{\gamma_0}s^{\gamma_0-1}d^{2\beta_\mu(\gamma_0-1)+d\gamma_0}\left(nh^{2\beta_\mu+d\frac{\gamma_0}{\gamma_0-1}}\right)^{1-\gamma_0}.
\end{aligned}
$$

Finally, if the constant is below 2, without loss of generality set $B = 2$. $\square$

Lemma 11 (Grid width construction)**.** *Suppose the conditions of Theorem 1 hold, and define* $N_n(h \mid x) = \sum_{i=1}^n \mathbf{1}\{D_i = 1, \|X_i - x\| \leq h\}$. *Then there is a* $c > 0$ *and an algorithm based only on* β_μ*,* d*, and the* (X, D) *data that generates a grid width* w_n *such that with probability tending to one,* $n^{-1/(2\beta_\mu+d)} \lesssim w_n \leq dcn^{\frac{-1}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0 - 1}}}$ *and for all* $x \in [-1,1]^d$*,* $N_n\left(cn^{\frac{-1}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0 - 1}}} \mid x_{n,j}\right) \geq 2\left(cn^{\frac{-1}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0 - 1}}}\right)^{-2\beta_\mu}$.

Proof of Lemma 11. First, I construct a $\tilde{h}_n$ that will be used to construct w_n. Let $z = 2$. By Lemma 3, there is a $k'' \geq 0$ such that $E[D\mathbf{1}\{\|X - x\| \leq h] \geq k'' h^{d\gamma_0/(\gamma_0 - 1)}$ for all $h \in (0, 1]$ and all $x \in [-1, 1]^d$. Take $k' = (k''2)^{-1/(2\beta_\mu + d\gamma_0/(\gamma_0-1))}$ and take $\tilde{h}_n = k' n^{-1/(2\beta_\mu + d\gamma_0/(\gamma_0 - 1))}$, so that $k'' n \tilde{h}_n^{d\gamma_0/(\gamma_0-1)} = 2^z \tilde{h}_n^{-2\beta_\mu}$.

To construct a grid without use of γ_0, define a pseudo-grid of width $\bar{w}_n = d/\lfloor (5n)^{1/(2\beta_\mu+d)} \rfloor$. Let n be large enough such that $\bar{w}_n \geq d(n/4)^{-1/(2\beta_\mu+d)}$ for all n. For each pseudo-gridpoint $\bar{x}_{n,j}$, take $\bar{h}_{n,j} = \inf_{h \leq 1} h : N_n(h \mid \bar{x}_{n,j}) \geq 2h^{-2\beta_\mu}$. Take the true grid width as $w_n = \lfloor 1/\left(d \sup \bar{h}_{n,j}\right) \rfloor$.

To show that w_n works, take another pseudogrid with pseudo-grid-width $\tilde{w}_n = 1/\lfloor 1/(2d\tilde{h}_n) \rfloor$ and consider hypothetical gridpoints $\tilde{x}_{n,j}$. For simplicity assume $1/(2d\tilde{h}_n)$ is an integer, and in particular that $\tilde{w}_n \leq \tilde{h}_n/d$. Notice also that $n^{-1/d} \ll n^{-1/(2\beta_\mu + d\gamma_0/(\gamma_0-1))} = \Theta(\tilde{h}_n) = \Theta(w_n)$, so that $n^{-1/d} \ll \tilde{h}_n \leq k' d n^{-1/(2\beta_\mu + d\gamma_0/(\gamma_0-1))}$ for all n large enough.

Note also that $\|x - \tilde{x}_{n,j}\| \leq \tilde{h}_n$ and $\|x - \tilde{x}_{n,k}\| \leq \tilde{h}_n$ for $x \in [-1,1]^d$ implies $k = j$, because the gridpoints are separated by at least $d\tilde{h}_n$. There are also $O(\tilde{h}_n^{-d}) = o(n)$ gridpoints. Therefore by Lemma 6, with probability tending to one, $N_{n,j}(\tilde{h}_n \mid \tilde{x}_{n,j}) \geq nE[D\mathbf{1}\{\|X - \tilde{x}_{n,j} \leq \tilde{h}_n\}]/2 \geq \frac{nk''}{2} \tilde{h}_n^{d\gamma_0/(\gamma_0-1)} = 2^{z-1}(\tilde{h}_n)^{-2\beta_\mu}$.

Note that for every $x \in [-1, 1]^d$, there is some $\tilde{x}_{n,j}$ with $\|x - \tilde{x}_{n,j}\| \leq d\tilde{w}_n$. Therefore, with probability tending to one, $N_n((2d+1)\tilde{h}_n \mid x) \geq N_n(d\tilde{w}_n + \tilde{h}_n \mid x) \geq \inf_j N_n(\tilde{h}_n \mid \tilde{x}_{n,j}) \geq 2^{z-1}(\tilde{h}_n)^{-2\beta_\mu} = (2d+1)^{2\beta_\mu} 2^{z-1}((2d+1)\tilde{h}_n)^{-2\beta_\mu} \geq 2^z((2d+1)\tilde{h}_n)^{-2\beta_\mu}$.

Finally, take $c = (2d+1)k'$. With probability tending to one, $w_n \leq d(2d+1)\tilde{h}_n = dcn^{-1/(2\beta_\mu + d\gamma_0/(\gamma_0-1))}$. Note that with probability tending to one, $w_n \leq (2 + 1/d)d\tilde{h}_n$. Note also that by Lemma 6 applied to $h_{n,0} = (n/4)^{-1/(2\beta_\mu+d)}$ at the pseudo-gridpoints $\bar{x}_{n,j}$, with probability tending to one, $N_n(h_{n,0} \mid \bar{x}_{n,j}) \geq n(h_{n,0})^{d\gamma_0}(k''^{\gamma_0} h_{n,0})^{-d/(\gamma_0-1)} \gg 4(h_{n,0})^{-2\beta_\mu}$, so that with probability tending to one, $dcn^{-1/(2\beta_\mu + d\gamma_0/(\gamma_0-1))} \geq w_n \geq h_{n,0} \gtrsim n^{-1/(2\beta_\mu+d)}$.

Finally, note that because each $x_{n,j}$ is in $[-1,1]^d$, by the argument above,

$$N_n\left(cn^{\frac{-1}{2\beta_\mu+d\gamma_0/(\gamma_0-1)}} \mid x_{n,j}\right) = N_n\left((2d+1)\tilde{h}_n \mid x_{n,j}\right) \geq 2\left((2d+1)\tilde{h}_n\right)^{-2\beta_\mu}. \quad \square$$

Lemma 12 (Bounding inclusion probability)**.** *Take c as in Lemma 11; let $\underline{\delta}, \pi$ be as in Proposition 2; take some $\delta \geq \underline{\delta}$; construct $m_n^{(k)}$ and $t_n^{(k)}$ as in Proposition 2; take $h_n^{(k)} = cn^{-1/(2\beta_\mu+d\gamma_0/(\gamma_0-1))}/t_n^{(k)}$; and take k_n^* as in Lemma 9. Let $\{(x_{n,j}, h_{n,j})\}$ be a set of gridpoints $x_{n,j}$ covering $[-1,1]^d$ with edge lengths w_n such that (i) $h_{n,j} = \inf_{h\leq 1} h : N_n(h \mid x_{n,j}) \geq 2h^{-2\beta_\mu}$, (ii) $h_{n,j} \leq h_n^{(1)} \leq w_n/d$, and (iii) $N_n\left(h_n^{(k)} \mid x_{n,j}\right) \geq \frac{n}{2}E\left[D\mathbf{1}\left\{\|X - x_{n,j}\| \leq h_n^{(k)}\right\}\right]$ for all j, k and $h_{n,j} \leq h_n^{(1)} \leq w_n/d$ for all j. Then for every $k = 1, \ldots, k_n^* + 1$, there are at most $m_n^{(k)}$ gridpoints j with $h_{n,j} > h_n^{(k)}$.*

Proof of Lemma 12. Suppose j satisfies $h_{n,j} \geq h_n^{(k)}$. Then by (i) and (ii),

$$\frac{n}{2}E\left[D\mathbf{1}\left\{\|X - x_{n,j}\| \leq h_n^{(k)}\right\}\right] \geq \frac{n\left(h_n^{(k)}\right)^d}{2}E\left[D \mid \|X - x_{n,j}\| \leq h_n^{(k)}\right],$$

so that $nE\left[D\mathbf{1}\left\{\|X - x_{n,j}\| \leq h_n^{(k)}\right\}\right] \leq 4\left(h_n^{(k)}\right)^{-2\beta_\mu}$. Let $\tilde{m}_n^{(k)}$ be the number of gridpoints with $h_{n,j} \geq h_n^{(k)}$. Then by Lemma 10,

$$\tilde{m}_n^{(k)} \leq B\left(n(h_n^{(k)})^{2\beta_\mu+d\frac{\gamma_0}{\gamma_0-1}}\right)^{1-\gamma_0} = B\left(\frac{h_n^{(k)}}{h_n^{(1)}}\right)^{-(\gamma_0-1)\left(2\beta_\mu+d\frac{\gamma_0}{\gamma_0-1}\right)} \leq \exp\left(2^{-k}\delta\left(\frac{h_n^{(k)}}{h_n^{(1)}}\right)^{-2\beta_\mu}\right) = m_n^{(k)}.$$

$\square$

Lemma 13 (Grid characteristics)**.** *Take $c, \pi, \delta, m_n^{(k)}, h_n^{(k)}, m_n^{(k)}, k_n^*$ as in Lemma 12. Then there is a construction of a grid width w_n, gridpoints $x_{n,j}$, and bandwidths $h_{n,j}$ constructed using only β_μ, d, and the (X, D) data, such that with probability tending to one, (i) $N_n\left(h_n^{(1)}\right) \geq 2(h_n^{(1)})^{-2\beta_\mu}$ for all j, (ii) $h_{n,j} \leq h_n^{(1)}$ for all j, (iii) $\|x - x_{n,j}\| \leq h_{n,j}$ and $\|x - x_{n,k}\| \leq h_{n,k}$ implies $k = j$, (iv) $N_n\left(h_n^{(k)} \mid x_{n,j}\right) \geq \frac{n}{2}E\left[D\mathbf{1}\{\|X - x_{n,j}\| \leq h_n^{(k)}\}\right]$ for all j and all $k = 1, \ldots, k_n^* + 1$, (v) the smallest eigenvalue of $\frac{\sum D\mathbf{1}\{\|X-x_{n,j}\|\leq h_{n,j}\}U\left(\frac{X-x_{n,j}}{h_{n,j}}\right)U\left(\frac{X-x_{n,j}}{h_{n,j}}\right)^T}{\sum D\mathbf{1}\{\|X-x_{n,j}\|\leq h_{n,j}\}}$ is at least $\lambda^*/2$ for every j, and (vi) for all $k = 1, \ldots, k_n^* + 1$, there are at most $m_n^{(k)}$ gridpoints j with $h_{n,j} > h_n^{(k)}$.*

Proof of Lemma 13. Let $w_n, x_{n,j}$ be constructed as in Lemma 11. For for every j, take $h_{n,j} = \inf_{h\leq 1} h : N_n(h \mid x_{n,j}) \geq 2h^{-2\beta_\mu}$. Notice that this construction is well-defined and does not use the outcome data and only depends on d, β_μ, and the (X, D) data.

By Lemma 11, $n^{-1/(2\beta_\mu+d)} \lesssim w_n \leq dcn^{-1/(2\beta_\mu+d\gamma_0/(\gamma_0-1))} = dh_n^{(1)}$.

(i) holds by Lemma 11. But then (ii) holds by this construction of $h_{n,j}$.

(iii) holds because $h_{n,j} \leq h_n^{(1)} \leq w_n/d$ by Lemma 11.

(iv) will hold by Lemma 6. Consider the k_n^*+1 functions $\mathbf{1}\{D = 1, \|X-x\| \leq h_n^{(k)}\}$ for $k = 1, \ldots, k_n^* + 1$ evaluated at the $O\big(n^{d/(2\beta_\mu+d)}\big)$ gridpoints $x_{n,j}$ with global bandwidth $h = h_n^{(1)} = \max_k h_n^{(k)}$. I continue on event that the number of gridpoints is $O(1)$ and the grid width is at least $h_n^{(1)}$, which were shown to be arbitrarily high probability earlier in this proof, so that the preconditions of Lemma 6 hold. Recall by Lemma 9 that $k_n^* = O(\log(\log(n)))$. Note that $h_n^{(1)} = \Theta^{n^{-1/(2\beta_\mu+d\gamma_0/(\gamma_0-1))}}$; $n \ll \exp\left(-an(h_n^{(1)})^{d\gamma_0/(\gamma_0-1)}\right)$ by inspection; and $k_n^*\left(1 - \left(\frac{e-1}{e}\right)^{n/\left(\exp\left(\log(1/2)+\tilde{c}^{-1}n^{-1}(h_n^{(1)})^{-d\gamma_0/(\gamma_0-1)}\right)\right)}\right) \to 0$ by L'Hopital's Rule. Therefore by Lemma 6, with probability tending to one, $N_n\left(h_n^{(k)} \mid x_{n,j}\right) \geq \frac{n}{2}E\left[D\mathbf{1}\{\|X - x_{n,j}\| \leq h_n^{(k)}\}\right]$ for all j and all $k = 1, \ldots, k_n^* + 1$.

(v) by strict monotonicty of $h^{-2\beta}$ and weak monotonicty of $N_n(h \mid x)$ in the opposite direction, $h_{n,j} = \inf_{h\leq 1} h : N_n(h \mid x_{n,j}) \geq 2h^{-2\beta_\mu} = \sup_{h\leq 1} h : N_n(h \mid x_{n,j}) \leq 2h^{-2\beta_\mu}$. Thus on the high probability event that the number of gridpoints is $O(n)$, by Lemma 7 and Lemma 5, if n is large enough so that $h_n^{(1)}$ is small enough, then the smallest eigenvalue of the induced design matrices is at least $\lambda^*/2$.

(vi) holds by Lemma 12, with each condition either holding by construction or by claim above. □

Lemma 14 (Maximal local polynomial residual). *Let w_n and $\{x_{n,j}, h_{n,j}\}$ be constructed as in Lemma 13, let Z be the (X, D) data, and define $\bar{\mu}_j = E[\hat{\mu}_j \mid Z]$. Then $E\left[\max_j \left(\hat{\mu}(x_{n,j}) - \bar{\mu}_{n,j}\right)^2 \mid Z\right] = O_P\left(n^{-2\beta_\mu/(2\beta_\mu+d\gamma_0/(\gamma_0-1))}\right)$.*

Proof of Lemma 14. Let $\bar{\mu}_{n,j} = E[\hat{\mu}(x_{n,j}) \mid Z]$. Recall by Lemma 13 that with probability tending to one, the smallest eigenvalue of the realized local polynomial design matrices $\sum DKUU^T/\sum DK$ is at least $\lambda^*/2 > 0$. Note that there are no x with $\|x - x_{n,j}\| \leq h_{n,j}$ and $\|x - x_{n,k}\| \leq h_{n,k}$ by (Lemma 13(iii)), so that $\hat{\mu}(x_{n,j}) \mid Z$ are independent normal draws. Therefore the set of $\hat{\mu}(x_{n,j})-\bar{\mu}(x_{n,j})$ random variables are independent normal, with variance bounded above by $4\lambda^{-2}N_n\left(h_{n,j} \mid x_{n,j}\right)^{-1}\sigma^2_{\max} = cN_n\left(h_{n,j} \mid x_{n,j}\right)^{-1}$ for some universal constant c.

Let $c, \delta, t_n^{(k)}h_n^{(k)}, m_n^{(k)}, k_n^*$ be defined as in Lemma 9, so that $h_n^{(k_n^*)} \leq c' h_n^{(1)}\pi^{(k-1)} \lesssim$

$n^{-1/(2\beta_\mu+d)}\left(1/\log(n)\right)^{1/\beta_\mu}$ by Proposition 2 and the definition of k_n^*. For any fixed $k = 1,\ldots,k_n^*$, by standard arguments,

$$\begin{aligned}
\text{“}M_k\text{”} &= E\left[\max_j \left(\hat{\mu}(x_{n,j}) - \bar{\mu}_{n,j}\right)^2 \mathbf{1}\{h_n^{(k+1)} \le h_{n,j} \le h_n^{(k)}\} \mid Z\right] \\
&\le c'\left(\min_{j:h_n^{(k+1)}\le h_{n,j}\le h_n^{(k)}} N_n\left(h_{n,j} \mid x_{n,j}\right)\right)^{-1} \log\left(\sum_j \mathbf{1}\{h_n^{(k+1)} \le h_{n,j} \le h_n^{(k)}\}\right)
\end{aligned}$$

for some universal constant c'. Thus, with probability tending to one,

$$\begin{aligned}
M_k &\le c'\left(\min_{j:h_n^{(k+1)}\le h_{n,j}\le h_n^{(k)}} h_{n,j}^{-2\beta_\mu}\right)^{-1} \log\left(m_n^{(k)}\right) \le c'\left(h_n^{(k)}\right)^{2\beta_\mu}\log\left(m_n^{(k)}\right) \quad (\text{L12,13(vi)}) \\
&\le c'\delta\left(h_n^{(k)}\right)^{2\beta_\mu} 2^{-(k+1)}\left(t_n^{(k)}\right)^{2\beta_\mu} \qquad (\text{Proposition 2}) \\
&= c'\delta\left(h_n^{(k)}\right)^{2\beta_\mu} 2^{-(k+1)}\left(h_n^{(k)}/h_n^{(1)}\right)^{-2\beta_\mu} = c'\delta\left(h_n^{(1)}\right)^{2\beta_\mu} 2^{-k}.
\end{aligned}$$

Also note that $h_{n,j} < h_n^{(k_n^*)}$ implies $N_n\left(h_{n,j} \mid x_{n,j}\right) = \left(h_{n,j}\right)^{-2\beta_\mu} \ge \left(h_n^{(k_n^*)}\right)^{-2\beta_\mu}$, so that analogously, on the high probability event expressed above that there are $O(n)$ gridpoints and $\max_j h_{n,j} \le h_n^{(1)}$),

$$\begin{aligned}
\text{“}M_{k+1}\text{”} &= E\left[\max_j \left(\hat{\mu}(x_{n,j}) - \bar{\mu}_{n,j}\right)^2 \mathbf{1}\left\{h_{n,j} < h_n^{(k_n^*)}\right\} \mid Z\right] \\
&\le c'\left(h_n^{(k_n^*)}\right)^{2\beta_\mu}\log\left(\frac{1}{n}\right) \lesssim n^{\frac{-2\beta_\mu}{2\beta_\mu + \frac{d\gamma_0}{\gamma_0-1}}}\frac{\log(n)}{\log(n)^2} = o\left(n^{\frac{-2\beta_\mu}{2\beta_\mu+\frac{d\gamma_0}{\gamma_0-1}}}\right).
\end{aligned}$$

Thus, on an event with probability tending to one,

$$\begin{aligned}
E\left[\max_j \left(\hat{\mu}(x_{n,j}) - \bar{\mu}_{n,j}\right)^2 \mid Z\right] &\le \sum_{k=1}^{k_n^*} M_k + M_{k+1} \le \sum_{k=1}^{k_n^*} c'\delta\left(h_n^{(1)}\right)^{2\beta_\mu} 2^{-k} + o\left(n^{\frac{-2\beta_\mu}{2\beta_\mu+d\gamma_0/(\gamma_0-1)}}\right) \\
&\le c'\delta\left(h_n^{(1)}\right)^{2\beta_\mu} + o\left(n^{\frac{-2\beta_\mu}{2\beta_\mu+d\gamma_0/(\gamma_0-1)}}\right) = O\left(n^{\frac{-2\beta_\mu}{2\beta_\mu+d\gamma_0/(\gamma_0-1)}}\right).
\end{aligned}$$

□

Proof of Theorem 1. Recall that $r_{\mu,n}^* = n^{\frac{-\beta_\mu}{2\beta_\mu+d\frac{\gamma_0}{\gamma_0-1}}}$. There are two main directions to show.

Lower bound pointwise rate. Define $x_0 = (0,\ldots,0)$. Let $\mathscr{P}$ be as in Lemma 8, with associated distributions $P_{n,m}$ for $m = 1, 2$. By Lemma 8, $K(P_{n,1}, P_{n,2}) \leq \alpha$. Define the seminorm $d(P, Q) = |E_P[Y \mid X = x_0, D = 1] - E_Q[Y \mid X = x_0, D = 1]|$. By construction,

$$d(P_{n,1}, P_{n,2}) = \overbrace{L' \exp(1/3)}^{\text{“}A/2\text{”}} n^{\frac{-\beta_\mu}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0 - 1}}},$$

where $L' > 0$ is fixed. Therefore $d(P_{n,1}, P_{n,2}) \geq 2s_n$, where $s_n = (A/4)n^{\frac{-\beta_\mu}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0 - 1}}}$. Thus, standard arguments (Tsybakov, 2009) show that for any fixed estimator $\hat{\mu}$ of $E[Y \mid X = x_0, D = 1]$ and all n large enough,

$$\begin{aligned} \sup_{P \in \mathscr{P}} P\left(|\hat{\mu}(x_0) - \mu(x_0)| \geq s_n\right) &\geq \max_{j=1,2} P_j\left(|\hat{\mu}(x_0) - \mu_{n,j}(x_0)| \geq s_n\right) \\ &\geq \max\left(\frac{\exp(-\alpha)}{4}, \frac{1 - \sqrt{\alpha/2}}{2}\right) > 0. \end{aligned}$$

Thus, for all $t > 1$ (see Tsybakov Theorem 2.3),

$$\liminf_{n \to \infty} \inf_{\hat{\mu}} \sup_{P \in \mathscr{P}} P\left(n^{\frac{\beta_\mu}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0 - 1}}} |\hat{\mu}(x_0) - \mu(x_0)| \geq t^{\frac{\beta_\mu}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0 - 1}}}\right) \geq \max\left(\frac{\exp(-ct)}{4}, \frac{1 - \sqrt{ct/2}}{2}\right),$$

where c is a constant that only depends on the parameters of the problem. Thus, $n^{\frac{-\beta_\mu}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0 - 1}}}$ is a lower bound on the *pointwise* rate of convergence.

Achievable uniform rate. This is the more difficult and interesting direction. Thankfully, several lemmas above make the remaining task relatively simple.

By Lemma 13 and Lemma 14, there is a construction of gridpoints $x_{n,j}$ separated by edges of length $w_n = O_P\left(n^{-1/(2\beta_\mu + d\gamma_0/(\gamma_0 - 1))}\right)$ and bandwidths $h_{n,j}$ depending only on β_μ, d, and the (X, D) data Z such that with probability tending to one, $\max_j h_{n,j} = O\left(n^{-1/(2\beta_\mu + d\gamma_0/(\gamma_0 - 1))}\right)$, the smallest eigenvalue of the design matrix used to construct $\hat{\mu}(x_{n,j})$ based on regression of Y on $U\left(\frac{X - x_{n,j}}{h_{n,j}}\right)$ with weights $D\mathbf{1}\{\|X - x_{n,j}\| \leq h_{n,j}\}$ is at least $\lambda^*/2$ for some fixed $\lambda^* > 0$, and $E\left[\max_j \left(\hat{\mu}(x_{n,j}) - \bar{\mu}_{n,j}\right)^2 \mid Z\right] = O\left(n^{-2\beta_\mu/(2\beta_\mu + d\gamma_0/(\gamma_0 - 1))}\right)$ where $\bar{\mu}_{n,j} = E[\hat{\mu}(x_{n,j}) \mid Z]$. But then on this event, $\max_j \hat{\mu}(x_{n,j}) - \mu(x_{n,j})^2 = O\left((\lambda^*)^{-2} \max_j h_{n,j}^{2\beta_\mu}\right) = O\left(n^{-2\beta_\mu/(2\beta_\mu + d\gamma_0/(\gamma_0 - 1))}\right)$. Therefore for this construction, with probability tending to one, $E\left[\max_j \left(\hat{\mu}(x_{n,j}) - \mu(x_{n,j})\right)^2 \mid Z\right] = O\left(n^{-2\beta_\mu/(2\beta_\mu + d\gamma_0/(\gamma_0 - 1))}\right)$. Therefore $\max_j \left(\hat{\mu}(x_{n,j}) - \mu(x_{n,j})\right)^2 = O_P\left(n^{-2\beta_\mu/(2\beta_\mu + d\gamma_0/(\gamma_0 - 1))}\right)$ by Markov's inequality.

Now consider predictions within the grid. For a given point $x \in [-1,1]^d$, if $x = x_{n,j}$ for some j, take $\hat{\mu}(x) = \hat{\mu}(x_{n,j})$. Otherwise, construct $\hat{\mu}(x)$ via local polynomial regression of $\hat{\mu}(x_{n,j})$ on $U\left(\frac{x_{n,j}-x}{(\beta_\mu+1)w_n}\right)$ for gridpoints $x_{n,j}$ with $\|x_{n,j}-x\| \leq (\beta_\mu+1)w_n$. By standard arguments, each such design matrix is nondegenerate so that $(\hat{\mu}(x) - \mu(x))^2 = O\left(w_n^{2\beta_\mu} + \max_j \left(\hat{\mu}(x_{n,j}) - \mu(x_{n,j})\right)^2\right) = O_P\left(n^{-2\beta_\mu/(2\beta_\mu+d\gamma_0/(\gamma_0-1))}\right)$ for all $x \in [-1,1]^d$ simultaneously. Therefore $\hat{\mu}$ achieves the uniform rate $r^*_{\mu,n}$, and with no polylogarithmic penalty.

Completing the proof. An achievable uniform rate of convergence is also an achievable pointwise rate of convergence. Thus, $n^{\frac{-\beta_\mu}{2\beta_\mu+d\frac{\gamma_0}{\gamma_0-1}}}$ is the optimal pointwise rate of convergence. A lower bound on the pointwise rate of convergence is also a lower bound on the uniform rate of convergence. Thus, $n^{\frac{-\beta_\mu}{2\beta_\mu+d\frac{\gamma_0}{\gamma_0-1}}}$ is also the optimal uniform rate of convergence. □

5 Conclusion

This note derives minimax consistency rates under Hölder smoothness with weak overlap. I precisely characterize the degree to which overlap weakness makes outcome regression more difficult. I leave open the performance of Hölder smoothness-based estimators for treatment effects under weak overlap. The analysis requires nondegenerate eigenvalues; a task for future work is whether other more sophisticated estimators can be robust to degenerate eigenvalues introduced by increasing propensity scores relative to the lowest level of overlap allowed under the global bound.

References

Armstrong, T. B. and Kolesár, M. (2021). Finite-sample optimal estimation and inference on average treatment effects under unconfoundedness. *Econometrica*, 89(3):1141–1177.

Balakrishnan, S., Kennedy, E. H., and Wasserman, L. (2025). The fundamental limits of structure-agnostic functional estimation.

Bruns-Smith, D., Dukes, O., Feller, A., and Ogburn, E. L. (2024). Augmented balancing weights as linear regression.

Chaudhuri, S. and Hill, J. B. (2024). Heavy tail robust estimation and inference for average treatment effects. *Econometric Reviews*.

Chen, X., Hong, H., and Tarozzi, A. (2008). Semiparametric efficiency in gmm models of nonclassical measurement errors, missing data and treatment effects. Technical Report 1644, Cowles Foundation.

Dorn, J. (2026). How much weak overlap can doubly robust t-statistics handle?

Gaïffas, S. (2005). Convergence rates for pointwise curve estimation with a degenerate design. *Mathematical Methods of Statistics*, 14(1).

Hall, P., Marron, J. S., Neumann, M. H., and Titterington, D. M. (1997). Curve estimation when the design density is low. *The Annals of Statistics*, 25(2):756 – 770.

Heiler, P. and Kazak, E. (2021). Valid inference for treatment effect parameters under irregular identification and many extreme propensity scores. *Journal of Econometrics*, 222(2):1083–1108.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Hirshberg, D. A. and Wager, S. (2021). Augmented minimax linear estimation. *The Annals of Statistics*, 49(6):3206 – 3227.

Horn, R. A. and Johnson, C. R. (2013). *Matrix Analysis: Second Edition*. Cambridge University Press.

Khan, S. and Tamer, E. (2010). Irregular identification, support conditions, and inverse weight estimation. *Econometrica*, 78(6):2021–2042.

Ma, X. and Wang, J. (2020). Robust inference using inverse probability weighting. *Journal of the American Statistical Association*, 115(532):1851–1860.

Mou, W., Ding, P., Wainwright, M. J., and Bartlett, P. L. (2023). Kernel-based off-policy estimation without overlap: Instance optimality beyond semiparametric efficiency.

Newey, W. K. and Robins, J. M. (2017). Cross-fitting and fast remainder rates for semiparametric estimation. CeMMAP working papers CWP41/17, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Robins, J., Li, L., Tchetgen Tchetgen, E., and van der Vaart, A. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics: Essays in Honor of David A. Freedman*, volume 2 of *IMS Collections*, pages 335–421. Institute of Mathematical Statistics.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric estimators. *The Annals of Statistics*, 10(4):1040–1053.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer.