Curvature-Adaptive Doubly Robust Estimation for Continuous Treatments*

Hongseok Namkoong
Columbia Business School
namkoong@gsb.columbia.edu

Isaac Scheinfeld
Columbia Business School
ils2124@columbia.edu

Yunbei Xu
National University of Singapore
yunbei@nus.edu.sg

Shunri Zheng
University of Illinois Urbana-Champaign
shunriz2@illinois.edu

Abstract

We study value estimation for deterministic policies with continuous, possibly multivariate treatments. Because a prescribed action has conditional probability zero, estimation must be local. We analyze a doubly robust estimator whose correction evaluates both nuisance functions at the realized treatment. Its exact bias is the sum of a product error and a smoothing error in $(\bar{m} - m_P)f_P/\bar{f}$, rather than in the outcome regression itself. This identity makes the relevant bandwidth depend on the curvature left after first-stage estimation. On explicit reference-assisted classes with curvature radius r, and when the product error is negligible at the target scale, we prove the matching minimax mean-squared error $n^{-1} \vee n^{-4/(d+4)} r^{2d/(d+4)}$, including the parametric transition. We also establish conditional limit theory, feasible studentization, and a cross-fitted oracle approximation. Simulations and a semi-synthetic pricing study illustrate the resulting bias and bandwidth behavior.

Contents

*Authors are listed in alphabetical order.

1 Introduction

Consider the value of a deterministic policy when treatment is continuous. The target depends on the conditional mean outcome at $A = \theta(X)$, yet this event has probability zero. Estimation must therefore use observations whose realized treatments lie near the policy action. The resulting problem is local even when the policy value itself is an average.

Localization changes the design of a debiasing term. In discrete-treatment problems, nuisance functions may be evaluated at the observed or prescribed action without creating a smoothing discrepancy. With a continuous treatment, the two placements lead to different biases. A correction evaluated at the policy action smooths the outcome surface. A correction evaluated at each realized treatment can instead smooth only the error left by the first-stage regression. This distinction is the starting point of the paper.

Let $m_P(a, x) = \mathbb{E}_P(Y \mid A = a, X = x)$ and let $f_P(a \mid x)$ be the conditional treatment density. From auxiliary data, construct references $\bar{m}$ and $\bar{f}$. On an independent evaluation sample, consider the augmented estimating function

$$\psi_h(Z; \bar{m}, \bar{f}) = \bar{m}\{\theta(X), X\} + \frac{K_h\{A - \theta(X)\}}{\bar{f}(A \mid X)}\{Y - \bar{m}(A, X)\}. \tag{1.1}$$

Here $\bar{f}$ denotes the conditional-density function $(a, x) \mapsto \bar{f}(a \mid x)$; $\bar{f}(A_i \mid X_i)$ is simply its value at an observed pair. We use "estimating function" for (1.1) and reserve "generalized propensity" for such conditional-density values.

The placement of the two nuisance functions in (1.1) yields an exact identity. Write

$$\Delta_P = \bar{m} - m_P, \qquad \delta_P = 1 - f_P/\bar{f}, \qquad g_P = \Delta_P f_P/\bar{f}.$$

Conditional on the fitted references, the bias of $\widehat{V}_h = \mathbb{P}_n \psi_h$ is

$$\mathbb{E}_P(\widehat{V}_h \mid \mathcal{D}_0) - V_P(\theta) = \mathbb{E}_P[\Delta_P\{\theta(X), X\}\delta_P\{\theta(X), X\}] + \mathbb{E}_P\left[g_P\{\theta(X), X\} - \int K(u) g_P\{\theta(X) + hu, X\}\, du\right]. \tag{1.2}$$

The first term is the familiar product error. The second smooths g_P, not m_P. Thus (1.1) is unbiased for every bandwidth whenever $\bar{m} = m_P$, regardless of the positive reference density. When $\bar{f} = f_P$, the product term vanishes and the remaining bias depends on the curvature of $\bar{m} - m_P$. With both nuisances estimated, density error enters through the product term and through the ratio $f_P/\bar{f}$ in g_P.

This identity suggests a different measure of local difficulty. Classical second-order kernel theory controls the curvature of the regression itself; on a fixed smoothness ball it yields mean-squared error $n^{-4/(d+4)}$ (Stone, 1980, 1982; Tsybakov, 2009). We instead index the problem by a uniform local bound r on the treatment Hessian of g_P. When the product error is negligible at the target scale, balancing squared smoothing bias against local variance gives

$$\mathfrak{R}_{n,d}(r) = n^{-1} \vee n^{-4/(d+4)} r^{2d/(d+4)}. \tag{1.3}$$

For fixed r, this is the classical second-order rate. When a first-stage learner leaves a shrinking curvature radius, the rate improves and eventually reaches the parametric floor. The comparison is therefore between two statistical classes, not a contradiction of classical lower bounds on a fixed Hölder ball.

1.1 Contributions

The paper develops four aspects of this observation.

(i) *Bias geometry and estimator design.* We give the exact finite-bandwidth identity (1.2) for multivariate treatments and two learned nuisances. An integral Taylor formula shows that the leading smoothing term is governed by the Hessian of g_P. The common neighborhood and integrable derivative conditions are stated explicitly; pointwise Taylor control is not sufficient for a uniform result.

(ii) *Uniform and minimax theory.* We prove a uniform finite-sample risk bound and establish matching upper and lower bounds for (1.3) on two explicit reference-assisted models: one with a known reference and one with a population-projection reference learned on a pilot sample. The bounds include the transition to n^{-1}. They also show how an α-smooth first stage, $\alpha > 2$, can transfer its derivative rate through a second-order outer kernel, subject to the stated product and derivative conditions.

(iii) *Inference with estimated first stages.* We prove conditional central limit theorems for shrinking and fixed bandwidths, give ratio-consistent variance estimators and feasible studentization, and derive a fixed-fold cross-fitted oracle approximation. Following the econometric theory of generated regressors, we separate equicontinuity of the centered empirical process from rate-normalized equicontinuity of the learned treatment derivatives. Sample splitting addresses the former, not the latter.

(iv) *Policy evaluation in finite samples.* We derive a finite-policy Bernstein bound and report a complementary empirical study. Controlled simulations examine bandwidth, dimension, and overlap; a semi-synthetic JD.com pricing study evaluates policies constructed by a PPO-inspired

one-step neural optimization procedure. The experiments illustrate the bias mechanism and its limits, rather than the shrinking-radius minimax exponent itself.

1.2 Related work

The estimator has a precise antecedent. Kallus and Zhou (2018) define the observationwise generalized propensity $Q_i = f_P(A_i \mid X_i)$, note that it may be imputed by conditional-density estimation, and display the scalar version of (1.1). Their input Q_i is the realized value of a density; our nuisance is the full conditional-density function, evaluated over a neighborhood of the policy action. The formal theory controls its contribution through the Hessian of $g_P = (\bar{m} - m_P) f_P / \bar{f}$, rather than through a separate smoothness condition on $\bar{f}$. Their formal bias, variance, and MSE analysis concerns known-density kernel IPW, corresponding here to $\bar{f} = f_P$ and $\bar{m} \equiv 0$.

Kallus and Uehara (2020) subsequently cross-fit learned outcome and conditional-density functions in a general family that contains (1.1); their main bandit theorem analyzes two neighboring kernelizations, one with a smoothed plug-in term and one with a target-evaluated correction. It obtains the fixed-curvature $n^{-4/5}$ rate for a scalar action and leaves a formal minimax result open. Their paper also treats the broader problems of finite-horizon off-policy value and gradient estimation. The distinct contribution here is the curvature-indexed and inferential theory for the unsmoothed target plug-in combined with the realized-action correction. We also make explicit the uniform Taylor expansion and local overlap assumptions required to make the expansion step in the proof of Kallus and Zhou (2018) fully rigorous. In particular, a pointwise $o(h^2)$ remainder cannot be passed through the expectation over X and the kernel integral without uniform integrability; the fixed common neighborhood and the L_1 Hessian modulus in (3.5) supply that control.

Other continuous-treatment estimators make different localization choices. Kennedy et al. (2017) develop nonparametric doubly robust estimators for a dose-response curve. Colangelo and Lee (2026) study a target-evaluated correction and establish inference under uniform derivative conditions. The higher-order construction of Bonvini and Kennedy (2026) gives fast upper rates for dose-response estimation, whereas Mou et al. (2023) develops local minimax theory for linear functionals over RKHS classes. These targets and statistical classes differ from the reference-assisted classes used for our matching lower bounds.

The analysis also belongs to the econometric tradition of two-step estimation and generated regressors (Pagan, 1984; Newey, 1991; Andrews, 1994; Newey, 1994; Newey and McFadden, 1994; Mammen et al., 2012). That literature clarifies why first-stage consistency alone does not control a local random process: derivative behavior and stochastic equicontinuity must be imposed or proved for the learner. Our rate-normalized curvature condition plays this role. From the perspective of nonparametric minimax theory, the main point is equally classical: the rate is determined by the class over which risk is measured. Reference assistance changes that class by restricting the error remaining after the first stage, while leaving the usual fixed-ball theory intact (Brown and Low, 1996).

Organization. Section 2 gives the construction and implementation. Section 3 develops the bias identity, uniform and minimax bounds, and inference. Section 4 treats policy learning and pricing, and Section 5 reports the empirical study. Appendix A contains all proofs; Appendix B contains implementation details and the full set of supplemental figures.

2 Setup and estimator

We observe independent copies of $Z = (X, A, Y)$, where $X \in \mathcal{X}$, $A \in \mathcal{A} \subseteq \mathbb{R}^d$, and Y is the observed outcome. For each $a \in \mathcal{A}$, let $Y(a)$ denote the potential outcome under treatment a. A fixed deterministic policy is a measurable map $\theta : \mathcal{X} \to \mathcal{A}$, with value

$$V_P(\theta) = \mathbb{E}_P\{Y(\theta(X))\}.$$

Under the identification conditions of Section 3,

$$V_P(\theta) = \mathbb{E}_P[m_P\{\theta(X), X\}], \qquad m_P(a, x) = \mathbb{E}_P(Y \mid A = a, X = x),$$

where m_P is the continuous version specified in Assumption 3.1.

Let $K : \mathbb{R}^d \to \mathbb{R}$ be a centered second-order kernel and set $K_h(v) = h^{-d}K(v/h)$. An auxiliary sample is used to construct an outcome reference $\bar{m}(a, x)$ and a conditional-density reference $\bar{f}(a \mid x)$. The evaluation sample is independent of these functions. We study

$$\widehat{V}_h(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left[\bar{m}\{\theta(X_i), X_i\} + \frac{K_h\{A_i - \theta(X_i)\}}{\bar{f}(A_i \mid X_i)}\{Y_i - \bar{m}(A_i, X_i)\}\right]. \tag{2.1}$$

The weighted term is set to zero when its kernel factor is zero. Positivity of $\bar{f}$ is therefore required only on the policy neighborhood reached by the kernel.

2.1 The realized-action correction

The plug-in term in (2.1) is evaluated at the policy action; the correction evaluates both nuisance functions at the observed action. This alignment gives, conditional on X and the references,

$$\mathbb{E}_P\left[\frac{K_h\{A - \theta(X)\}}{\bar{f}(A \mid X)}\{Y - \bar{m}(A, X)\}\,\middle|\, X\right] = -\int K(u)g_P\{\theta(X) + hu, X\}\,du, \tag{2.2}$$

where $g_P = (\bar{m} - m_P)f_P/\bar{f}$. Equation (2.2) is the mechanism behind the exact bias formula of Theorem 3.4. If $\bar{m} = m_P$, the correction has mean zero for every positive $\bar{f}$ and every bandwidth. If $\bar{f} = f_P$, the usual product error disappears and the smoothing bias is determined by $\bar{m} - m_P$.

The placement is substantive. A target-evaluated alternative has the form

$$\bar{m}\{\theta(X), X\} + \frac{K_h\{A - \theta(X)\}}{\bar{f}\{\theta(X) \mid X\}}\big[Y - \bar{m}\{\theta(X), X\}\big]. \tag{2.3}$$

Its regression error and density are fixed at the target action while the kernel ranges over nearby observations. Consequently its leading bias contains curvature of the target regression and density. In (2.1), the corresponding curvature belongs to g_P, the first-stage error after the density change of measure. Section 1 places these constructions in the literature.

2.2 Cross-fitting and bandwidth

For fixed-fold cross-fitting, estimate $\bar{m}$ and $\bar{f}$ off-fold, truncate the fitted density to positive finite bounds on the policy neighborhood, evaluate (2.1) on the held-out fold, and average the resulting pseudo-outcomes. Once the nuisance fits are stored, evaluation requires a single pass through the data.

Prediction error alone does not determine the curvature-sensitive bandwidth. The analysis controls the treatment Hessian of

$$g_P = (\bar{m} - m_P) f_P / \bar{f},$$

including derivatives of $f_P/\bar{f}$ when the density is learned. Given a valid curvature bound $\bar{r}$ and product-error bound $\bar{\rho}$, the rate calculation uses

$$h = h_0 \wedge c_h (n\bar{r}^2)^{-1/(d+4)}.$$

The bound $\bar{r}$ must have the order of the underlying radius to attain the stated rate; a conservative bound remains valid but can lead to excess undersmoothing. Without derivative control, the theory supports only the fixed-radius bandwidth and rate. Thus "curvature-adaptive" describes the radius-dependent analysis and bandwidth choice, not automatic adaptation to an unknown smoothness class.

3 Theory

Throughout, the treatment dimension $d \geq 1$ is fixed. The covariate X may be high dimensional; its dimension enters only through the first-stage method. Unless stated otherwise, $(\bar{m}, \bar{f})$ are fitted on an auxiliary sample $\mathcal{D}_0$ independent of the n observations in (2.1). All conditional statements are conditional on $\mathcal{D}_0$.

3.1 Identification and local regularity

For $h_0 > 0$, define the fixed neighborhood of the policy action

$$\mathcal{N}_\theta(h_0) = \{(a, x) : \|a - \theta(x)\|_\infty \leq h_0\}.$$

Assumption 3.1 (Identification and regular versions)**.** The policy θ is fixed and measurable. Uniformly over the laws under consideration, the following hold for constants $h_0 > 0$ and $0 < \underline{f} < \overline{f} < \infty$.

(a) Consistency holds, $Y = Y(A)$ almost surely, and weak unconfoundedness holds: $Y(a) \perp A \mid X$ for every $a \in \mathcal{A}$.

(b) For P_X-almost every x, $\theta(x) + [-h_0, h_0]^d$ is contained in the interior of the conditional treatment support.

(c) The functions $m_P(a, x) = \mathbb{E}_P(Y \mid A = a, X = x)$, $\mu_P(a, x) = \mathbb{E}_P\{Y(a) \mid X = x\}$, and $f_P(a \mid x)$ denote fixed jointly measurable versions that are continuous in a on $\mathcal{N}_\theta(h_0)$, and

$$\underline{f} \leq f_P(a \mid X) \leq \overline{f} \quad \text{on } \mathcal{N}_\theta(h_0) \quad \text{almost surely.}$$

Because $A = \theta(X)$ is a null event, the target depends on the chosen continuous versions. Consistency and weak unconfoundedness identify $m_P(a, x) = \mu_P(a, x)$ for Lebesgue-almost every local treatment value; positivity and continuity extend this equality throughout the policy neighborhood. Therefore

$$V_P(\theta) = \mathbb{E}_P[m_P\{\theta(X), X\}].$$

Assumption 3.2 (Kernel)**.** The kernel $K : \mathbb{R}^d \to \mathbb{R}$ is bounded and supported on $[-1, 1]^d$, and

$$\int K(u)\,du = 1, \qquad \int uK(u)\,du = 0.$$

For some $\eta > 0$,

$$R_K := \int K(u)^2\,du \in (0, \infty), \qquad \int |K(u)|^{2+\eta}\,du < \infty.$$

Define

$$M_K := \int uu^\top K(u)\,du, \qquad L_{2,K} := \int \|u\|_2^2 |K(u)|\,du < \infty.$$

A product kernel formed from a symmetric univariate kernel satisfies $M_K = \kappa I_d$ and $R_K = R(k)^d$. A general M_K also permits a fixed non-axis-aligned smoothing geometry.

Assumption 3.3 (References and local moments)**.** Conditional on $\mathcal{D}_0$, $\bar{m}$ and $\bar{f}$ are fixed measurable functions. For a constant $\overline{f}_{\mathrm{ref}} < \infty$,

$$\underline{f} \le \bar{f}(a \mid X) \le \overline{f}_{\mathrm{ref}} \quad \text{on } \mathcal{N}_\theta(h_0) \quad \text{almost surely,}$$

and, for a constant $M < \infty$,

$$\mathbb{E}_P |\bar{m}\{\theta(X), X\}|^{2+\eta} \le M.$$

Use jointly measurable regular versions of the conditional moments in the next display. For each $p \in \{2, 2 + \eta\}$,

$$\sup_{\|v\|_\infty \le h_0} \mathbb{E}_{P_X}\left[\frac{f_P\{\theta(X) + v \mid X\}}{\bar{f}\{\theta(X) + v \mid X\}^p} \times \mathbb{E}_P(|Y - \bar{m}\{\theta(X) + v, X\}|^p \mid A = \theta(X) + v, X)\right] \le M. \tag{3.1}$$

The assumptions constrain only a neighborhood of the policy action. Integrable envelopes may replace (3.1); pointwise moment bounds alone do not suffice for the uniform results below.

Write

$$\Delta_P(a, x) = \bar{m}(a, x) - m_P(a, x), \qquad \delta_P(a, x) = 1 - \frac{f_P(a \mid x)}{\bar{f}(a \mid x)},$$

and define the density-ratio-weighted regression error

$$g_P(a, x) = \Delta_P(a, x)\frac{f_P(a \mid x)}{\bar{f}(a \mid x)} = \Delta_P(a, x)\{1 - \delta_P(a, x)\}. \tag{3.2}$$

3.2 Exact bias and a uniform remainder

Let

$$\psi_h(Z; \bar{m}, \bar{f}) = \bar{m}\{\theta(X), X\} + \frac{K_h\{A - \theta(X)\}}{\bar{f}(A \mid X)}\{Y - \bar{m}(A, X)\}, \qquad \widehat{V}_h = \mathbb{P}_n \psi_h.$$

The weighted residual term is defined to be zero whenever its kernel factor is zero.

Theorem 3.4 (Exact conditional bias and Hessian expansion)**.** *Under Assumptions 3.1–3.3, for every* $0 < h \le h_0$*,*

$$\begin{aligned}\beta_{P,h} &:= \mathbb{E}_P(\widehat{V}_h \mid \mathcal{D}_0) - V_P(\theta) \\ &= B_{0,P} + \mathbb{E}_P\left[g_P\{\theta(X), X\} - \int K(u) g_P\{\theta(X) + hu, X\}\, du\right], \end{aligned} \tag{3.3}$$

where

$$B_{0,P} = \mathbb{E}_P[\Delta_P\{\theta(X), X\}\delta_P\{\theta(X), X\}]. \tag{3.4}$$

Suppose additionally that, for P_X*-almost every* x*,* $g_P(\cdot, x)$ *has an absolutely continuous gradient along every line segment in the policy neighborhood. Assume that it has a jointly measurable Hessian version such that, along each such segment* $a + sv$*,*

$$\frac{d}{ds}\nabla_a g_P(a + sv, x) = \nabla_a^2 g_P(a + sv, x)v \quad \text{for almost every } s,$$

and satisfying

$$\sup_{\|v\|_\infty \le h_0} \mathbb{E}_P\|\nabla_a^2 g_P\{\theta(X) + v, X\}\|_{\mathrm{op}} < \infty.$$

Set

$$H_{2,P} = \mathbb{E}_P[\nabla_a^2 g_P\{\theta(X), X\}],$$

and

$$\omega_{2,P}(h) = \sup_{\|v\|_\infty \le h} \mathbb{E}_P\left\|\nabla_a^2 g_P\{\theta(X) + v, X\} - \nabla_a^2 g_P\{\theta(X), X\}\right\|_{\mathrm{op}}. \tag{3.5}$$

Then

$$\beta_{P,h} = B_{0,P} - \frac{h^2}{2}\operatorname{tr}(M_K H_{2,P}) + R_{P,h}, \qquad |R_{P,h}| \le \frac{L_{2,K}h^2}{2}\omega_{2,P}(h). \tag{3.6}$$

If, in addition, the Hessian is absolutely continuous along every local line segment and

$$\mathbb{E}_P\left[\sup_{\|v\|_\infty \le h_0} \|\nabla_a^3 g_P\{\theta(X) + v, X\}\|_{\mathrm{op}}\right] \le M_{3,P},$$

where the norm is the operator norm of the trilinear derivative, then $\omega_{2,P}(h) \le \sqrt{d}\, h M_{3,P}$*.*

For a class $\mathcal{P}$, a uniform expansion requires

$$\lim_{h \downarrow 0} \sup_{P \in \mathcal{P}} \omega_{2,P}(h) = 0.$$

Pointwise twice differentiability alone does not justify integrating an $o(h^2)$ remainder over contexts and kernel arguments.

For risk bounds, define the fixed-neighborhood Hessian seminorm

$$\mathcal{C}_\theta(P; \bar{m}, \bar{f}) = \sup_{\|v\|_\infty \le h_0} \mathbb{E}_P\|\nabla_a^2 g_P\{\theta(X) + v, X\}\|_{\mathrm{op}}. \tag{3.7}$$

Corollary 3.5 (Conditional mean-squared error)**.** *If the assumptions of Theorem 3.4 hold and* $\mathcal{C}_\theta(P; \bar{m}, \bar{f}) \le r$*, then, for every* $0 < h \le h_0$*,*

$$\mathbb{E}_P\{(\widehat{V}_h - V_P)^2 \mid \mathcal{D}_0\} \le C\left\{B_{0,P}^2 + r^2 h^4 + \frac{1}{nh^d} + \frac{1}{n}\right\}, \tag{3.8}$$

where C *depends only on the common constants in the assumptions and on* K*.*

3.3 Uniform risk and the curvature radius

For fixed references $(\bar{m}, \bar{f})$ and $r, \rho \geq 0$, let

$$\mathcal{P}_\theta(r, \rho; \bar{m}, \bar{f}) = \big\{P : P \text{ satisfies Assumptions 3.1–3.3 with common constants,} \\ |B_{0,P}| \leq \rho, \quad \mathcal{C}_\theta(P; \bar{m}, \bar{f}) \leq r\big\}. \tag{3.9}$$

Define

$$\mathfrak{R}_{n,d}(r) = n^{-1} \vee n^{-4/(d+4)} r^{2d/(d+4)}. \tag{3.10}$$

Theorem 3.6 (Uniform upper bound)**.** *For every $0 < h \leq h_0$, uniformly over $P \in \mathcal{P}_\theta(r, \rho; \bar{m}, \bar{f})$,*

$$\mathbb{E}_P\{(\widehat{V}_h - V_P)^2 \mid \mathcal{D}_0\} \leq C\{\rho^2 + r^2 h^4 + (nh^d)^{-1} + n^{-1}\}.$$

With $(nr^2)^{-1/(d+4)} = \infty$ when $r = 0$, set

$$h_{n,r} = h_0 \wedge c_h (nr^2)^{-1/(d+4)}. \tag{3.11}$$

If $\rho^2 \leq C_\rho \mathfrak{R}_{n,d}(r)$, then

$$\sup_{P \in \mathcal{P}_\theta(r,\rho;\bar{m},\bar{f})} \mathbb{E}_P\{(\widehat{V}_{h_{n,r}} - V_P)^2 \mid \mathcal{D}_0\} \leq C' \mathfrak{R}_{n,d}(r).$$

For $d = 1$, this becomes

$$n^{-1} \vee n^{-4/5} r^{2/5}.$$

The risk is on the nonparametric branch when $r \gg n^{-1/2}$ and is of order n^{-1} when $r \lesssim n^{-1/2}$.

For random references, the theorem applies on the event that the product, Hessian, and moment bounds hold. High-probability membership gives a conditional O_P risk bound. An unconditional MSE bound further requires uniform integrability or truncation on the complement. The bandwidth in (3.11) uses a valid upper bound on r; choosing it when r is unknown is a separate adaptation problem.

Remark (Why pointwise curvature is insufficient). With a known treatment density, no covariates, $\theta = 0$, and $g(a) = a_1^4$, $\nabla^2 g(0) = 0$, but kernel smoothing has nonzero bias of order h^4. Signed averaging over X gives further counterexamples. Thus a pointwise expected Hessian cannot imply a uniform rh^2 bound. One needs local absolute control such as (3.7), or a directly stated integrated modulus.

3.4 Minimax bounds

We begin with an explicit regression model. Let $b : [-1, 1]^d \to \mathbb{R}$ be known, bounded, and continuous, and define each $u \in W^{2,\infty}([-1, 1]^d)$ by its unique $C^{1,1}$ representative, so that every point evaluation below is unambiguous. Set

$$\mathcal{H}_{2,d}(r, M) = \big\{u \in W^{2,\infty}([-1, 1]^d) : \|u\|_\infty \leq M, \ \|\nabla u\|_\infty \leq M, \\ \|\nabla^2 u\|_{\mathrm{op},\infty} \leq r\big\}.$$

Take $\theta \equiv 0$, fix $h_0 \in (0, 1)$, and let $\mathcal{Q}_d(r; b)$ contain the laws

$$X \equiv 0, \quad A \sim \mathrm{Unif}([-1, 1]^d), \quad Y = b(A) + u(A) + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2), \tag{3.12}$$

where $u \in \mathcal{H}_{2,d}(r, M)$ and ε is independent of A. The target is $V_P = b(0) + u(0)$.

Theorem 3.7 (Reference-assisted minimax rate). *Fix $d \geq 1$, $M > 0$, $\sigma > 0$, $0 \leq r_{\max} < \infty$, and the known function b. There are constants $0 < c < C < \infty$ and n_0 such that, for every $n \geq n_0$ and every possibly n-dependent $r \in [0, r_{\max}]$,*

$$c\mathfrak{R}_{n,d}(r) \leq \inf_{\widetilde{V}} \sup_{P \in \mathcal{Q}_d(r;b)} \mathbb{E}_P(\widetilde{V} - V_P)^2 \leq C\mathfrak{R}_{n,d}(r). \tag{3.13}$$

The infimum is over all estimators based on n observations and allowed to know b, r, M, σ, and d. The upper bound is attained up to constants by (2.1) with $\bar{m} = b$, the known treatment density, and $h \asymp h_0 \wedge (nr^2)^{-1/(d+4)}$.

Theorem 3.7 uses a known reference to isolate the remaining regression error. We next define the reference from the law itself. Let

$$\mathcal{B} = \operatorname{span}\{\phi_1, \ldots, \phi_q\}, \qquad \phi_1 \equiv 1,$$

where q is fixed, each component of ϕ has an absolutely continuous treatment gradient along line segments in the policy neighborhood, and ϕ, $\nabla_a \phi$, and a jointly measurable $\nabla_a^2 \phi$ are uniformly bounded. Set

$$G_P = \mathbb{E}_P\{\phi(A, X)\phi(A, X)^\top\}, \qquad \beta_P = G_P^{-1}\mathbb{E}_P\{\phi(A, X)Y\},$$

and

$$u_P(a, x) = m_P(a, x) - \phi(a, x)^\top \beta_P.$$

Let $\mathcal{P}_{\mathcal{B}}(r)$ be the collection of laws satisfying Assumption 3.1 with common constants and with a known treatment density. For P_X-almost every x, require $u_P(\cdot, x)$ to have an absolutely continuous treatment gradient along policy-neighborhood line segments and a jointly measurable Hessian. Require the eigenvalues of G_P to lie in $[\lambda, \Lambda]$, $\|\beta_P\| \leq M$, and u_P and $\nabla_a u_P$ to be uniformly bounded. In addition, assume

$$\sup_{\|v\|_\infty \leq h_0} \mathbb{E}_P \|\nabla_a^2 u_P\{\theta(X) + v, X\}\|_{\mathrm{F}}^2 \leq r^2. \tag{3.14}$$

Assume also that $\phi(A, X)\{Y - \phi(A, X)^\top \beta_P\}$ is uniformly sub-Gaussian.

On a pilot sample of size N, let

$$\widehat{G} = N^{-1} \sum_{i=1}^{N} \phi_i \phi_i^\top, \qquad \widehat{s} = N^{-1} \sum_{i=1}^{N} \phi_i Y_i,$$

and define stabilized least squares by

$$\widehat{\beta} = \begin{cases} \widehat{G}^{-1}\widehat{s}, & \lambda_{\min}(\widehat{G}) \geq \lambda/2, \\ 0, & \text{otherwise.} \end{cases}$$

Theorem 3.8 (Population-projection residual class). *Use $\bar{m}(a, x) = \phi(a, x)^\top \widehat{\beta}$ and evaluate (2.1) on an independent sample of size n. For every $0 < h \leq h_0$, uniformly over $\mathcal{P}_{\mathcal{B}}(r)$,*

$$\mathbb{E}_P(\widehat{V}_h - V_P)^2 \leq C\{(r^2 + N^{-1})h^4 + (nh^d)^{-1} + n^{-1}\}. \tag{3.15}$$

If $N \asymp n$, the choice

$$h = h_0 \wedge c\{n(r^2 + N^{-1})\}^{-1/(d+4)}$$

gives an upper bound $C\mathfrak{R}_{n,d}(r)$. *Fix* $0 \le r_{\max} < \infty$. *Suppose there is an* x_0 *such that* $\theta(x_0) + [-h_0, h_0]^d \subset (-1,1)^d$ *and*

$$2^{-d} \int_{[-1,1]^d} \phi(a, x_0)\phi(a, x_0)^\top da$$

is positive definite. If $N \asymp n$, *then, uniformly for possibly* n*-dependent* $r \in [0, r_{\max}]$ *and all sufficiently large* n,

$$\inf_{\widetilde{V}} \sup_{P \in \mathcal{P}_{\mathcal{B}}(r)} \mathbb{E}_P(\widetilde{V} - V_P)^2 \asymp \mathfrak{R}_{n,d}(r),$$

where the infimum is over estimators based on the combined samples.

The regression may have order-one curvature in $\mathcal{B}$; only its population-projection residual is restricted by r. The class is therefore defined by P, not by an externally supplied reference.

3.5 Higher smoothness and the minimax perspective

Corollary 3.9 (Transfer of a derivative-learning rate)**.** *Suppose* $N \asymp n$ *and, on events whose probabilities tend to one, the fitted references satisfy*

$$\mathcal{C}_\theta(P; \bar{m}_N, \bar{f}_N) \le Cn^{-(\alpha-2)/(2\alpha+d)}, \qquad \alpha > 2,$$

the product error satisfies

$$B_{0,P}^2 = O_P\{n^{-2\alpha/(2\alpha+d)}\},$$

and the uniform moment conditions hold. Then the conditional MSE of the second-order kernel estimator with (3.11) *is*

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

The corollary assumes rather than derives the derivative-learning rate. With unrestricted continuous covariates, the rate depends on dimension, smoothness, sparsity, or other structure; adaptation to unknown α is separate. The conclusion is that a second-order localization step can inherit a higher-order first-stage rate without requiring a higher-order kernel.

Classical minimax theory measures worst-case risk over a fixed function class (Stone, 1980, 1982; Tsybakov, 2009). For a multi-stage estimator, however, an absolute Hölder ball need not describe the difficulty remaining after the first stage. Here the final bias is governed by the Hessian of $(\bar{m} - m_P)f_P/\bar{f}$, not by that of m_P. The matching bounds identify this curvature radius as the relevant one for the stated classes, in the spirit of local and constrained risk analysis (Brown and Low, 1996).

When r is bounded away from zero, (3.10) recovers the classical second-order exponent. As r shrinks, the rate improves until it reaches the parametric floor. Thus $n^{-4/(d+4)}$ is the fixed-radius case, not a restriction imposed by the second-order kernel. This conclusion refines the criterion while leaving the classical fixed-class result intact.

3.6 Conditional limit theory

For inference, allow the law P_n, pilot $\mathcal{D}_{0,n}$, references $(\bar{m}_n, \bar{f}_n)$, and bandwidth h_n to vary with n. Let $V_n = V_{P_n}(\theta)$ and define

$$\begin{aligned} q_n(v, X) = &\frac{f_{P_n}\{\theta(X) + v \mid X\}}{\bar{f}_n\{\theta(X) + v \mid X\}^2} \\ &\times \mathbb{E}_{P_n}\left([Y - \bar{m}_n\{\theta(X) + v, X\}]^2 \mid A = \theta(X) + v, X\right), \end{aligned} \tag{3.16}$$

and $\Gamma_n = \mathbb{E}_{P_n}\{q_n(0,X)\}$. The conditional second moment in (3.16) is a designated jointly measurable regular version; all variance-continuity statements below refer to that version.

Theorem 3.10 (Shrinking-bandwidth conditional CLT and studentization)**.** *Suppose Assumptions 3.1–3.3 hold with probability tending to one over the pilot, with common constants. Assume*

$$h_n \downarrow 0, \qquad nh_n^d \longrightarrow \infty, \qquad \Gamma_n \longrightarrow_P \Gamma \in (0,\infty),$$

and the local variance-continuity condition

$$\int K(u)^2 \mathbb{E}_{P_n}|q_n(h_n u, X) - q_n(0,X)|\, du \longrightarrow_P 0. \tag{3.17}$$

Then, with β_{n,h_n} denoting the exact conditional bias,

$$\sup_{z\in\mathbb{R}} \left| \mathbb{P}_{P_n}\left(\frac{\sqrt{nh_n^d}(\widehat{V}_{h_n} - V_n - \beta_{n,h_n})}{\sqrt{R_K \Gamma_n}} \le z \,\middle|\, \mathcal{D}_{0,n} \right) - \Phi(z) \right| \longrightarrow_P 0. \tag{3.18}$$

If the moment condition (3.1) also holds with $p = 4+\eta$, define

$$\widehat{\Omega}_n = h_n^d \frac{1}{n}\sum_{i=1}^n \{\psi_{h_n}(Z_i; \bar{m}_n, \bar{f}_n) - \widehat{V}_{h_n}\}^2. \tag{3.19}$$

Then, for every $\varepsilon > 0$,

$$\mathbb{P}_{P_n}\left(\left| \frac{\widehat{\Omega}_n}{R_K\Gamma_n} - 1 \right| > \varepsilon \,\middle|\, \mathcal{D}_{0,n} \right) \longrightarrow_P 0,$$

and (3.18) remains true with $R_K\Gamma_n$ replaced by $\widehat{\Omega}_n$.

If Theorem 3.4's Hessian conditions hold, the exact bias may be replaced by

$$B_{0,n} - \frac{h_n^2}{2}\operatorname{tr}(M_K H_{2,n})$$

provided

$$\sqrt{nh_n^d}\, h_n^2 \omega_{2,n}(h_n) \longrightarrow_P 0. \tag{3.20}$$

It may be omitted if $\mathcal{C}_\theta(P_n; \bar{m}_n, \bar{f}_n) \le r_n$ with probability tending to one and

$$\sqrt{nh_n^d}\{|B_{0,n}| + r_n h_n^2\} \longrightarrow_P 0. \tag{3.21}$$

At an MSE-optimal bandwidth, smoothing bias is generally of the same order as the standard error. A confidence interval centered at V_n therefore requires undersmoothing, consistent bias estimation, or bias correction. On the parametric branch, the optimal bandwidth is bounded away from zero and the normalization changes accordingly.

Theorem 3.11 (Fixed-bandwidth conditional CLT)**.** *Suppose $h_n \to h_\star \in (0, h_0]$ and, conditional on the pilot,*

$$\sigma_n^2 = \operatorname{Var}_{P_n}\{\psi_{h_n}(Z; \bar{m}_n, \bar{f}_n) \mid \mathcal{D}_{0,n}\} \longrightarrow_P \sigma_\star^2 \in (0,\infty).$$

If, for some $\eta > 0$,

$$\mathbb{E}_{P_n}\left[|\psi_{h_n} - \mathbb{E}_{P_n}(\psi_{h_n} \mid \mathcal{D}_{0,n})|^{2+\eta} \mid \mathcal{D}_{0,n} \right] = O_P(1),$$

then the conditional distribution of

$$\frac{\sqrt{n}(\widehat{V}_{h_n} - V_n - \beta_{n,h_n})}{\sigma_n}$$

converges in probability to $N(0,1)$*. The ordinary sample variance of the estimating functions is ratio consistent for* σ_n^2*. Centering at* V_n *requires* $\sqrt{n}\,|\beta_{n,h_n}| \to_P 0$*. Thus a curvature radius and product error both of order* $n^{-1/2}$ *can give parametric MSE, whereas centered root-n inference without bias correction requires total bias* $o(n^{-1/2})$*.*

3.7 Stochastic equicontinuity and estimated first stages

Estimated references place the problem within the econometric theory of generated regressors and semiparametric two-step estimation (Pagan, 1984; Newey, 1991; Andrews, 1994; Newey, 1994; Newey and McFadden, 1994; Mammen et al., 2012). Two forms of stochastic equicontinuity have distinct roles. The first controls treatment derivatives of the realized first-stage error; the second controls the empirical process after the nuisance functions have been fitted.

For the first, let the pilot size be N, allow $(\bar{m}_N, \bar{f}_N)$ to be random, and define

$$H_{N,P}(v)(x) = \nabla_a^2 g_{N,P}\{\theta(x) + v, x\}, \qquad v \in [-h_0, h_0]^d,$$

with $\|H\|_{P,1} = \mathbb{E}_P \|H(X)\|_{\mathrm{op}}$.

Assumption 3.12 (Rate-normalized stochastic equicontinuity)**.** For a deterministic sequence $r_N > 0$:

(a) for every fixed $v \in [-h_0, h_0]^d$, $\|H_{N,P}(v)\|_{P,1} = O_P(r_N)$;

(b) for every $\varepsilon > 0$,

$$\lim_{\delta \downarrow 0} \limsup_{N \to \infty} \mathbb{P}_P \left[\sup_{\substack{v,w \in [-h_0,h_0]^d \\ \|v-w\|_\infty \le \delta}} \frac{\|H_{N,P}(v) - H_{N,P}(w)\|_{P,1}}{r_N} > \varepsilon \right] = 0.$$

Lemma 3.13 (Uniform curvature control)**.** *Under Assumption 3.12,*

$$\sup_{\|v\|_\infty \le h_0} \|H_{N,P}(v)\|_{P,1} = O_P(r_N),$$

and therefore $\mathcal{C}_\theta(P; \bar{m}_N, \bar{f}_N) = O_P(r_N)$*.*

Normalization by r_N is essential: ordinary stochastic equicontinuity gives only an unscaled $o_P(1)$ statement. For a single bandwidth sequence, the weaker local conditions

$$\|H_{N,P}(0)\|_{P,1} = O_P(r_N), \qquad \sup_{\|v\|_\infty \le h_n} \|H_{N,P}(v) - H_{N,P}(0)\|_{P,1} = O_P(r_N)$$

suffice at h_n, but not uniformly over the fixed neighborhood. A random Hölder condition with exponent γ and constant L_N is sufficient locally when $L_N h_n^\gamma = O_P(r_N)$.

The second form concerns the empirical process. An independent pilot removes this issue from Theorem 3.10; cross-fitting or same-sample estimation restores it. For $\eta = (m, f)$, write $\psi_{h,\eta}$ for the estimating function in (2.1) and $\eta_P = (m_P, f_P)$.

Partition the observations into a fixed number J of folds I_j, with $n_j/n \to \pi_j \in (0,1)$, and fit $\widehat{\eta}_{-j}$ without fold I_j. Let

$$\widehat{V}_{\mathrm{cf}} = \sum_{j=1}^{J} \frac{n_j}{n} \mathbb{P}_{n,j} \psi_{h_n, \widehat{\eta}_{-j}}, \qquad \overline{\beta}_n = \sum_{j=1}^{J} \frac{n_j}{n} \{P_n \psi_{h_n, \widehat{\eta}_{-j}} - V_n\}.$$

Theorem 3.14 (Cross-fitted oracle approximation)**.** *Suppose the oracle estimating function $\psi_{h_n, \eta_{P_n}}$ satisfies the local variance, moment, and bandwidth conditions of Theorem 3.10. For $\mathbb{G}_{n,j} = \sqrt{n_j}(\mathbb{P}_{n,j} - P_n)$, assume*

$$\max_{j \le J} \sqrt{h_n^d} \left| \mathbb{G}_{n,j} \{ \psi_{h_n, \widehat{\eta}_{-j}} - \psi_{h_n, \eta_{P_n}} \} \right| = o_P(1). \tag{3.22}$$

Then

$$\frac{\sqrt{n h_n^d}(\widehat{V}_{\mathrm{cf}} - V_n - \overline{\beta}_n)}{\sqrt{\Omega_{0,n}}} \xrightarrow{d} N(0,1),$$

where

$$\Omega_{0,n} = R_K \mathbb{E}_{P_n} \left[\frac{\mathrm{Var}_{P_n}(Y \mid A = \theta(X), X)}{f_{P_n}\{\theta(X) \mid X\}} \right] \longrightarrow \Omega_0 \in (0, \infty).$$

A sufficient foldwise condition for (3.22) is

$$\max_{j \le J} h_n^d P_n \{ \psi_{h_n, \widehat{\eta}_{-j}} - \psi_{h_n, \eta_{P_n}} \}^2 = o_P(1), \tag{3.23}$$

because each held-out fold is independent of its own nuisance fit. Under the oracle version of (3.1) with $p = 4 + \eta$, define

$$\widehat{\Omega}_{\mathrm{cf}} = h_n^d \sum_{j=1}^{J} \frac{n_j}{n} \mathbb{P}_{n,j} \{ \psi_{h_n, \widehat{\eta}_{-j}} - \widehat{V}_{\mathrm{cf}} \}^2.$$

Then (3.23) also gives $\widehat{\Omega}_{\mathrm{cf}}/\Omega_{0,n} \to_P 1$. Centering the cross-fitted statistic directly at V_n requires

$$\sqrt{n h_n^d} \, |\overline{\beta}_n| \longrightarrow_P 0,$$

or a foldwise bias correction that removes $\overline{\beta}_n$ to this order.

Alternatively, (3.22) follows by a Newey-style argument over deterministic nuisance sets containing the fitted functions with probability tending to one. Cross-fitting does not make the fold-specific terms conditionally independent; it produces a common oracle term and a negligible remainder. Nor does sample splitting imply Assumption 3.12, which controls the treatment Hessian of

$$(\bar{m}_N - m_P) f_P / \bar{f}_N$$

and supplies the additional first-stage condition behind the faster rate. Bounded derivatives of the population law alone do not control derivatives of a fitted learner.

4 Policy learning and pricing interpretation

The multivariate formulation has two immediate consequences. Uniform value control yields regret bounds for finite policy classes, while the curvature term has a direct economic interpretation in pricing problems.

4.1 A finite-policy regret bound

Let Θ be a finite collection of deterministic policies and define

$$\theta^* \in \arg\max_{\theta\in\Theta} V_P(\theta), \qquad \widehat{\theta} \in \arg\max_{\theta\in\Theta} \widehat{V}_h(\theta).$$

Fix a deterministic rule for breaking ties in both optimizations. The elementary reduction

$$V_P(\theta^*) - V_P(\widehat{\theta}) \le 2\sup_{\theta\in\Theta} |\widehat{V}_h(\theta) - V_P(\theta)| \tag{4.1}$$

turns uniform value estimation into a regret bound. The linear term in Bernstein's inequality is retained because the local effective sample size may be small.

Proposition 4.1 (Finite-policy Bernstein bound)**.** *Condition on an auxiliary sample, so that $(\bar{m}, \bar{f})$ are fixed, and let $0 < h \le h_0 \wedge 1$. Suppose the assumptions used for Theorem 3.6 hold uniformly over $\theta \in \Theta$, with*

$$\sup_{\theta\in\Theta} |B_{0,P}(\theta)| \le \rho, \qquad \sup_{\theta\in\Theta} \mathcal{C}_\theta(P; \bar{m}, \bar{f}) \le r.$$

Suppose also that

$$|\bar{m}\{\theta(X), X\}| \le M_0$$

almost surely, and, on $\{K_h\{A - \theta(X)\} \ne 0\}$,

$$|Y - \bar{m}(A, X)| \le M_1, \qquad \bar{f}(A \mid X) \ge \underline{f},$$

uniformly over $\theta \in \Theta$. Then there are constants $C_v, C_b < \infty$, depending only on $M_0, M_1, \underline{f}, \overline{f}$, and K, such that, for every $0 < \delta < 1$, with conditional probability at least $1 - \delta$,

$$\begin{aligned} V_P(\theta^*) - V_P(\widehat{\theta}) \le 2\rho + L_{2,K} r h^2 + 2\sqrt{\frac{2C_v \log(2|\Theta|/\delta)}{nh^d}} \\ + \frac{2C_b \log(2|\Theta|/\delta)}{3nh^d}. \end{aligned} \tag{4.2}$$

Consequently, for a universal constant C and $L = \log(2|\Theta|)$,

$$\mathbb{E}_P\Big[V_P(\theta^*) - V_P(\widehat{\theta}) \mid \mathcal{D}_0\Big] \le 2\rho + L_{2,K} r h^2 + C\left\{\sqrt{\frac{C_v L}{nh^d}} + \frac{C_b L}{nh^d}\right\}. \tag{4.3}$$

If $L/(nh^d) \le 1$ and the bandwidth is chosen, up to fixed constants, as

$$h = (h_0 \wedge 1) \wedge \left(\frac{L}{nr^2}\right)^{1/(d+4)},$$

with the second term interpreted as infinity when $r = 0$, then

$$\mathbb{E}_P\Big[V_P(\theta^*) - V_P(\widehat{\theta}) \mid \mathcal{D}_0\Big] \le C'\left[\rho + \left\{\sqrt{\frac{L}{n}} \vee \left(\frac{L}{n}\right)^{2/(d+4)} r^{d/(d+4)}\right\}\right], \tag{4.4}$$

where C' may also depend on h_0 and the fixed problem constants and kernel.

Appendix A.7 proves the result, including the integrated tail bound and the two cases induced by bandwidth truncation.

For infinite policy classes, entropy or another empirical-process argument must replace the union bound. With cross-fitted nuisances one also needs the oracle approximation of Theorem 3.14: sample splitting by itself does not control curvature uniformly over policies.

4.2 Pricing and the kernel moment matrix

For a scalar price p, let $q_P(p,x)$ be demand and $m_P(p,x) = pq_P(p,x)$ be expected revenue. Whenever $m_P(p,x) > 0$, own-price elasticity satisfies

$$\varepsilon(p,x) = -\frac{p\,\partial_p q_P(p,x)}{q_P(p,x)} = 1 - \frac{p\,\partial_p m_P(p,x)}{m_P(p,x)},$$

while

$$\partial_p^2 m_P(p,x) = 2\partial_p q_P(p,x) + p\partial_p^2 q_P(p,x)$$

describes the local change in marginal revenue. Thus, after the density-ratio weighting in (3.2), the radius in Theorem 3.6 measures local error in price sensitivity. It can therefore be read in terms of demand curvature and elasticity (den Boer, 2015; Guelman and Guillén, 2014).

For a vector of prices $A = (p_1, \ldots, p_d)$ and product demands $q_{Pj}(A,x)$,

$$m_P(A,x) = \sum_{j=1}^{d} p_j q_{Pj}(A,x).$$

For $k \neq \ell$,

$$\partial^2_{p_k p_\ell} m_P(A,x) = \partial_{p_\ell} q_{Pk}(A,x) + \partial_{p_k} q_{P\ell}(A,x) + \sum_{j=1}^{d} p_j \partial^2_{p_k p_\ell} q_{Pj}(A,x),$$

so mixed Hessian entries describe error in substitution and cross-price responses. The leading smoothing term is

$$-\frac{h^2}{2}\,\mathrm{tr}(M_K H_{2,P}).$$

For a symmetric product kernel, $M_K = \kappa I_d$, so only diagonal Hessian entries enter this leading scalar bias. Mixed derivatives still affect the operator-norm radius, and enter the leading bias itself when M_K has off-diagonal entries. The kernel moment matrix therefore determines the smoothing geometry for joint pricing (Reibstein and Gatignon, 1984).

5 Numerical experiments

The experiments examine the finite-sample implications of the realized-action correction. Its leading smoothing bias is governed by the interaction of the outcome-regression and conditional-density errors, suggesting both smaller bias and less sensitivity to bandwidth when that interaction is locally flat. Writing

$$m_0(a,x) = \mathbb{E}[Y \mid A = a, X = x], \qquad V(\theta) = \mathbb{E}[m_0\{\theta(X), X\}],$$

we compare the realized-action estimator, denoted TR in the original code and figures, with inverse probability weighting (IPW), the estimator of Colangelo and Lee (2026) (CL), and a vectorized implementation of the local-linear estimator of Kennedy et al. (2017) (KMMS-I).

Two designs serve different purposes. Controlled simulations expose the bias–variance tradeoff with one- and ten-dimensional covariates; a poor-overlap variant serves as a negative control. A semi-synthetic pricing study based on JD.com records places the estimators downstream of learned neural policies. Its candidate policy uses an entropy-regularized adaptive-KL objective inspired by proximal policy optimization (PPO) (Schulman et al., 2017). This is one-step continuous-action policy optimization, not a benchmark in sequential deep reinforcement learning.

5.1 Common implementation

In the controlled simulations, random forests estimate both the outcome regression and the conditional action mean. A Gaussian kernel estimate of the action-error distribution then yields the conditional treatment density. Five-fold cross-validation uses the grids in Appendix B.2.2, and all methods share the same nuisance estimates within a design. The appendix supplies pseudocode, full tuning specifications, and the remaining figures.

5.2 Controlled simulations

5.2.1 Design

Each simulation uses $N_1 = 500$ training observations and $N_2 = 500$ evaluation observations. Each setting has five parallel runs with 20 bootstrap replications per run. We report absolute bias and root mean-squared error (RMSE) over the recorded replications and bandwidth grid.

The one-dimensional design uses

$$m_0(Z, A) = 4\sin(Z) - 4\sin(A) + (A - 4\pi)^2,$$

whereas the ten-dimensional covariate design uses

$$m_0(Z, A) = \frac{2}{5}\sum_{j=1}^{10}\sin(Z_j) - 4\sin(A) + \frac{1}{2}(A - 10\pi)^2.$$

Both response functions are smooth and nonlinear, with nonzero action curvature that makes smoothing bias visible. Appendix B.2 specifies the covariate distributions and the behavior and target rules.

5.2.2 Overlap

Figures 1 and 2 display the one-dimensional favorable-overlap design. The target lies largely in regions of appreciable behavior-policy density, although the local effective sample size varies with the covariate.

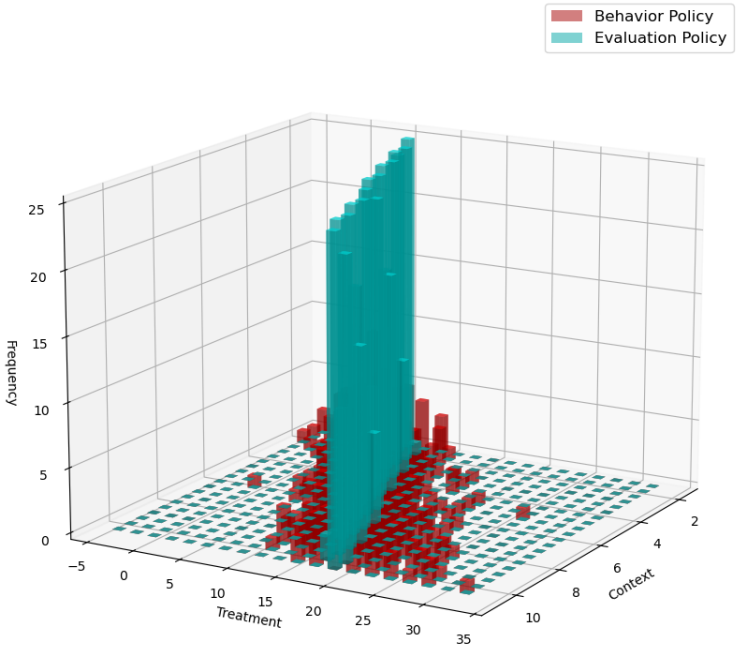

Figure 1: One-dimensional favorable-overlap design: empirical joint distribution of the context and logged action, together with the target pricing rule.

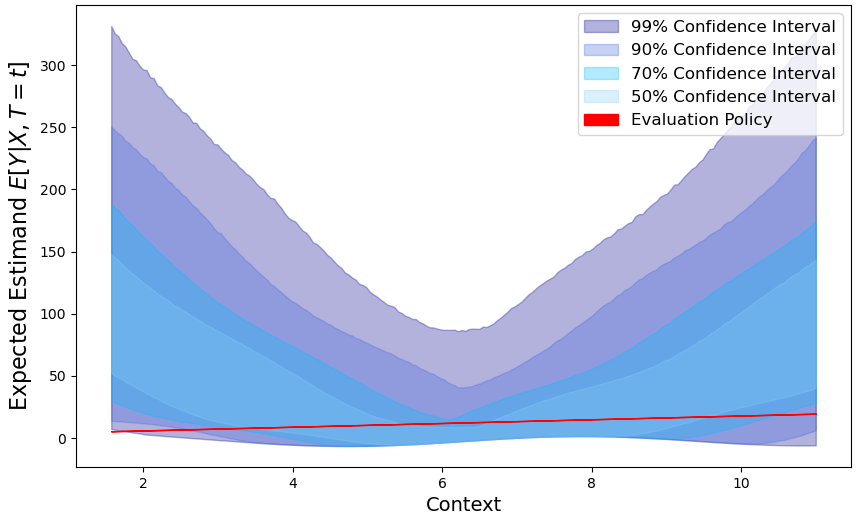

Figure 2: One-dimensional favorable-overlap design: behavior-policy action bands (blue) and target rule (red). The target lies within the central behavior-policy bands over much of the covariate range.

5.2.3 Estimator performance

Figures 3–6 report RMSE and absolute bias as functions of bandwidth. IPW is omitted from these four panels because its errors are much larger on the displayed scale; the complete four-estimator figures appear in Appendix B.2.5.

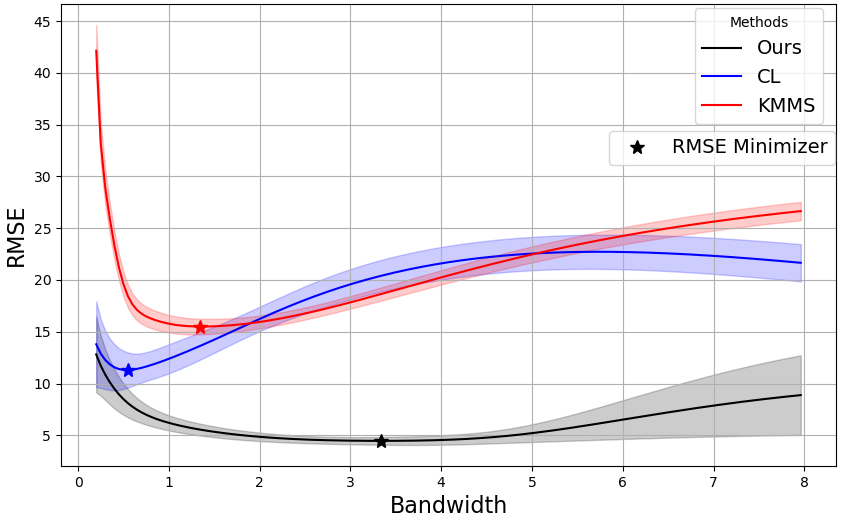

Figure 3: One-dimensional favorable-overlap simulation: RMSE versus bandwidth for TR, CL, and KMMS-I. Stars mark the minimum over the displayed grid. TR attains the lowest minimum and remains near it over a comparatively broad range of bandwidths.

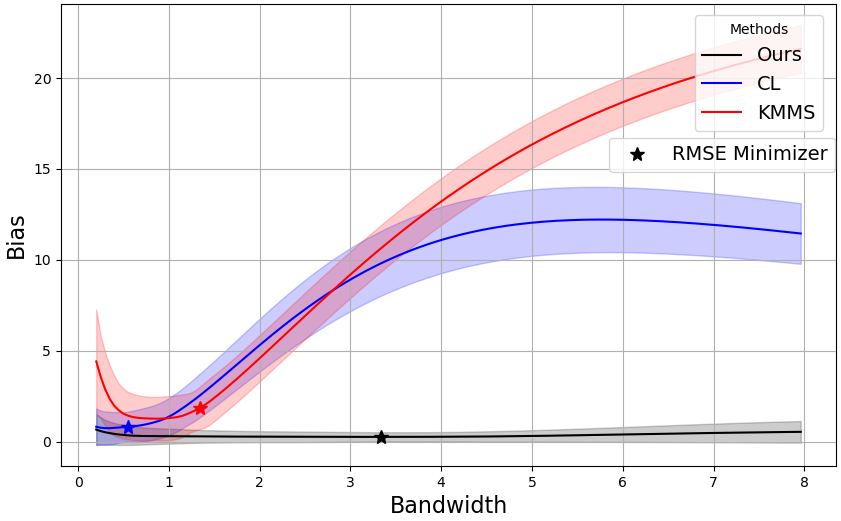

Figure 4: Absolute bias in the one-dimensional favorable-overlap simulation. The TR curve grows more slowly over the bandwidth range in which its RMSE remains near the minimum.

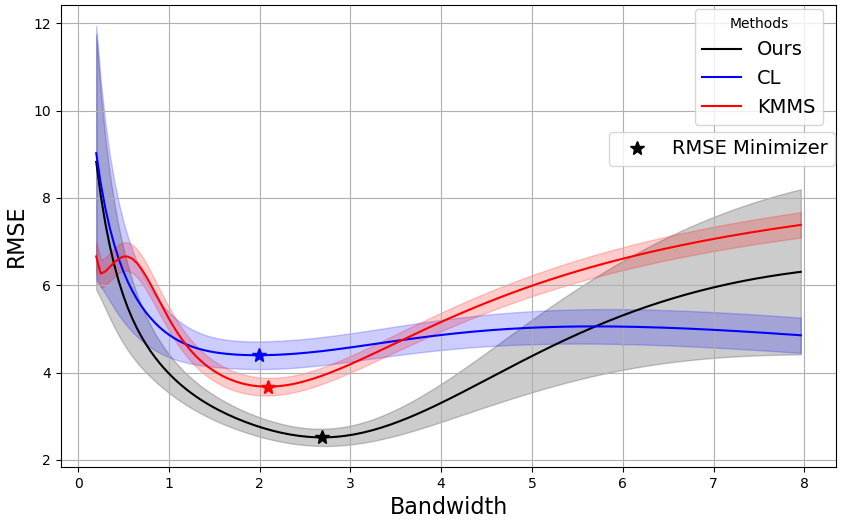

Figure 5: Ten-dimensional favorable-overlap simulation: RMSE versus bandwidth for TR, CL, and KMMS-I. Stars mark the minimum over the displayed grid. TR has the lowest displayed minimum and a comparatively broad low-RMSE region.

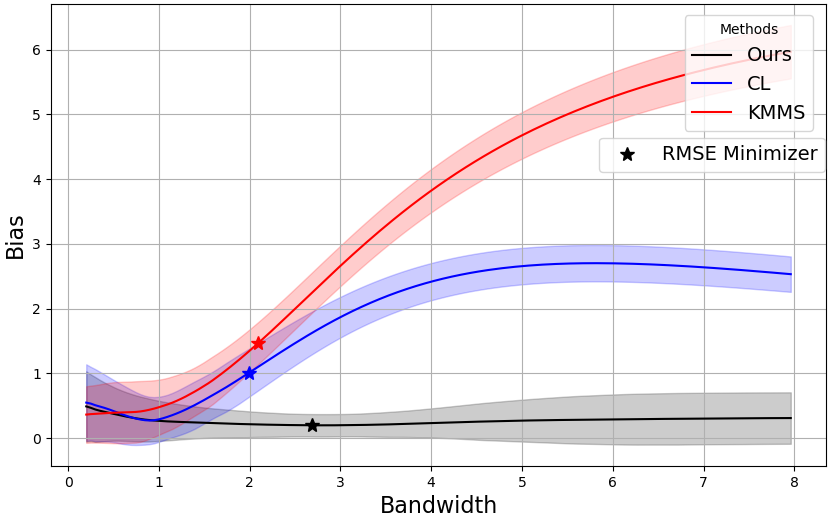

Figure 6: Absolute bias in the ten-dimensional favorable-overlap simulation. Over the reported grid, TR tolerates more smoothing before bias rises sharply.

In both favorable-overlap designs, TR has the lowest reported minimum RMSE and a relatively flat curve near its minimum. This is consistent with the exact bias formula: when first-stage learning leaves a lower-curvature error than the outcome surface, more local observations can be used before smoothing bias dominates. The figures illustrate this mechanism but do not establish an optimal data-driven bandwidth rule.

5.2.4 Poor overlap

The poor-overlap design moves the target away from the support of the behavior policy. Table 1 reports the best value on the same bandwidth grid.

Method	RMSE	Bias	Bandwidth at minimum
TR	**980.758**	**939.354**	**7.945**
CL	980.774	939.366	4.621
IPW	1238.831	1206.441	**7.945**
KMMS-I	1084.013	978.716	4.417

Table 1: Poor-overlap simulation: minimum RMSE and corresponding absolute bias and bandwidth over the reported grid. All four methods have very large errors; TR and CL are essentially tied.

This is a negative control. TR and CL differ by only 0.016 in RMSE and 0.012 in bias, and every method is severely inaccurate. No debiasing formula can compensate for absent support. Appendix B.2.6 gives the behavior-policy action bands and the complete RMSE and bias curves.

5.3 JD.com-based semi-synthetic pricing study

5.3.1 Data and policy-learning pipeline

The pricing study uses the March 2018 JD.com E-Commerce Data Challenge records (Shen et al., 2020). After filtering the focal SKU to the stated price range, the pipeline retains 773 distinct user–context–price combinations. It preserves the observed contexts and prices, fits an OLS outcome model, and repeatedly synthesizes outcomes so that the policy value is known for RMSE and bias calculations. Appendix B.3 records the feature engineering, price filtering, outcome construction, and repeated-sampling procedure.

Gaussian behavior and candidate policies have context-dependent means and variances, represented by fully connected networks for the mean and log standard deviation. Gaussian likelihood fits the behavior policy. The candidate policy maximizes importance-weighted revenue subject to an entropy bonus and an adaptive KL penalty toward the fitted behavior policy. Training stops only after validation revenue has improved by at least 2% and normalized validation effective sample size is at least 0.92. These are optimization and overlap criteria, not claims about deployed revenue. Appendix B.3 gives the objective, KL controller, tuning grids, and optimization parameters.

The neural module supplies a data-adaptive continuous-action policy for the evaluation pipeline. The value estimators use its conditional-mean rule $\theta(x) = \hat{\mu}_\theta(x)$; the fitted scale is used during policy training and overlap control. The theory treats this final rule as fixed and does not analyze the preceding neural optimization.

5.3.2 Estimator comparison

The 800 outcome-generation draws described in the appendix are used to construct the policy-training stage. The value-estimation stage then contains ten repetitions at n = 10,000, 50,000, 100,000, and 200,000. Adaptive KL regularization limits weight concentration, a concern made explicit by the poor-overlap simulation.

At n = 10,000, CL has the lower minimum RMSE (2.293 versus 5.670 for TR). TR has the lowest minimum at each larger sample size: 2.598 versus 2.834 at n = 50,000, 2.436 versus 2.804 at n = 100,000, and 2.091 versus 2.352 at n = 200,000. Relative to CL, these are reductions of 8.3%, 13.1%, and 11.1%, respectively. KMMS-I and IPW have substantially larger minimum RMSE

Sample size	TR	CL	KMMS-I	IPW
10k	5.670	**2.293**	21.949	13.524
50k	**2.598**	2.834	22.188	9.325
100k	**2.436**	2.804	22.259	9.823
200k	**2.091**	2.352	22.129	10.279

Table 2: Minimum RMSE over the displayed bandwidth grid in the JD.com-based semi-synthetic pricing study. Lower is better; bold marks the minimum within each sample size.

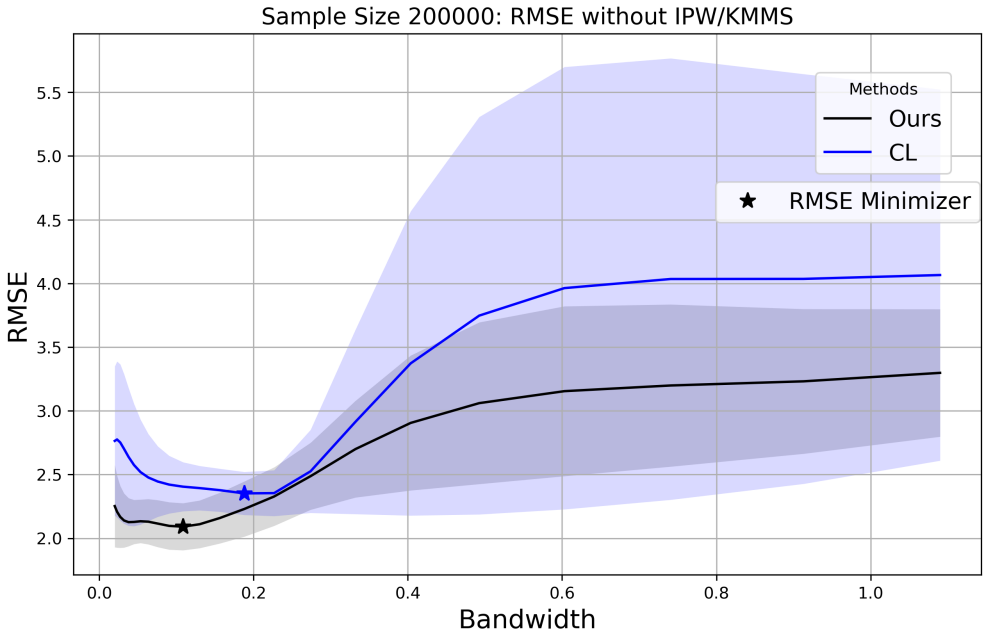

Figure 7: JD.com-based semi-synthetic pricing study, $n = 200{,}000$: RMSE versus bandwidth for TR and CL. Stars mark the minimum over the displayed grid. TR attains 2.091, compared with 2.352 for CL, and remains competitive over a broader bandwidth range. KMMS-I and IPW are reported in the appendix.

throughout. At $n = 200{,}000$, Figures 7 and 8 also show a broader competitive bandwidth range and lower bias for TR at its RMSE-minimizing bandwidth. The recorded evidence therefore suggests a tuning advantage at the three larger sample sizes, not uniform finite-sample dominance.

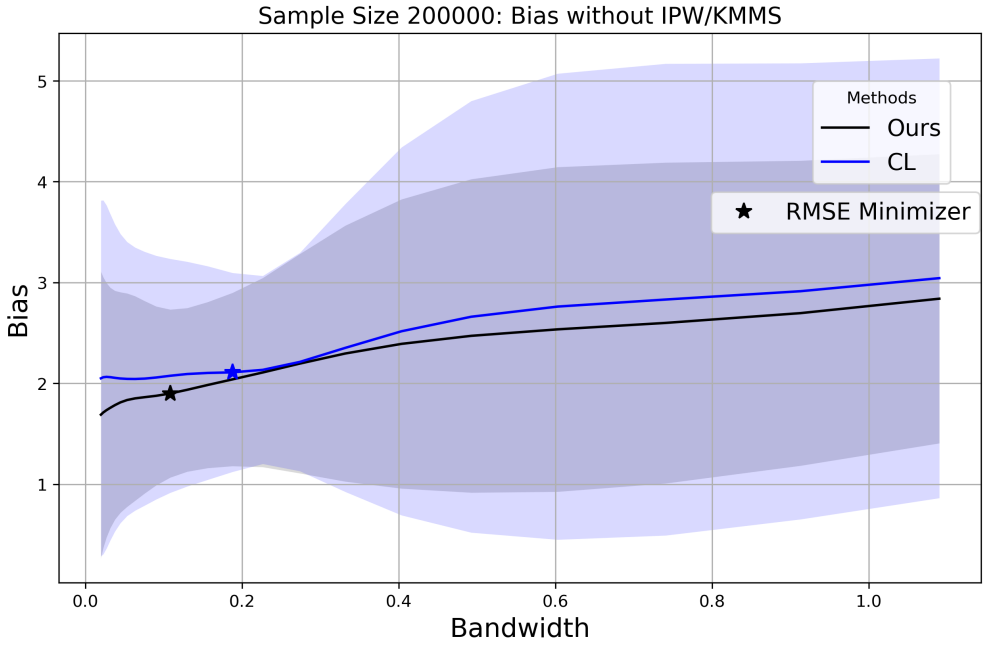

Figure 8: Absolute bias in the same JD.com-based semi-synthetic experiment. Stars identify each estimator's RMSE-minimizing bandwidth. At those choices, TR has lower bias than CL.

6 Conclusion

The central point is that a continuous-treatment augmentation should be constructed at the realized action and paired with a curvature-adaptive bandwidth. The estimator evaluates the learned outcome surface at the policy action, but constructs the outcome error and inverse-density weight at A before localization. Conditional on the first stage, smoothing therefore acts on $(\bar{m} - m_P) f_P / \bar{f}$, rather than on m_P. The curvature of this remaining error, together with the product error, determines the bandwidth and the attainable accuracy. Over the reference-assisted classes considered here, the minimax mean-squared error is

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

and the paper gives matching upper and lower bounds for explicit known- and estimated-reference models.

This result refines, rather than displaces, classical nonparametric theory. Fixed curvature recovers the usual second-order rate. Faster rates arise only when the first stage leaves a shrinking error class. If an α-smooth learner, $\alpha > 2$, controls its Hessian at the corresponding derivative rate, a second-order outer kernel attains the α-smooth rate. Such a transfer requires derivative control, product error, overlap, and enough structure to obtain the asserted first-stage rates; it is not automatic adaptation, nor a minimax claim for an unrestricted dose-response model.

The inference results preserve this separation. Under the stated score-equicontinuity condition, fitted references may be replaced by their oracle counterparts in the centered empirical process; rate-normalized equicontinuity controls the learned Hessian uniformly. Under the stated moment and variance conditions, the resulting conditional central limit theorems cover both shrinking and fixed bandwidths and permit feasible studentization and cross-fitted oracle approximation. Exact centering, undersmoothing, or bias correction is still required when smoothing bias is visible at the inferential scale.

The experiments accord with these qualifications. TR has the smallest reported minimum RMSE in the favorable-overlap simulations, but every method fails under poor overlap. In the JD.com-based semi-synthetic study, CL is best at the smallest sample size and TR is best from 50,000 observations onward. The semi-synthetic pipeline uses PPO-inspired one-step neural policy optimization. Across the designs, overlap and bandwidth remain consequential even when the leading smoothing bias is reduced.

Several questions remain. Selecting the bandwidth from an unknown curvature radius requires uniform control of the selection step. Inference over rich policy classes requires process-level rather than fixed-policy approximations. A broader account of first-stage learners satisfying derivative equicontinuity, and of minimax behavior beyond the structured reference-assisted models, is also open.

Acknowledgments

Use of artificial intelligence tools. The authors used ChatGPT and Codex for literature searches, proof checking, drafting and language editing, restructuring the manuscript, and organizing existing empirical material. These tools did not generate the empirical results.

A Proofs and supporting calculations

Expectations and probabilities are taken under the law in the corresponding result. When the reference functions are fitted on an auxiliary sample, we first condition on that sample. The references are then fixed and the evaluation observations are independent and identically distributed.

A.1 Exact bias and finite-sample risk

Proof of Theorem 3.4. Put $t = \theta(X)$. The local second-moment assumption and Cauchy–Schwarz justify Fubini's theorem below. Conditional on $A = a$ and X, the outcome error in the estimating function has mean $m_P(a, X) - \bar{m}(a, X) = -\Delta_P(a, X)$. Thus, after the change of variables $a = t + hu$,

$$\begin{aligned} P\psi_h - V_P &= \mathbb{E}_P\{\Delta_P(t, X)\} - \mathbb{E}_P\left[\int K(u)\Delta_P(t + hu, X)\frac{f_P(t + hu \mid X)}{\bar{f}(t + hu \mid X)}\, du\right] \\ &= \mathbb{E}_P\{\Delta_P(t, X)\} - \mathbb{E}_P\left[\int K(u) g_P(t + hu, X)\, du\right]. \end{aligned}$$

Since

$$\Delta_P(t, X) - g_P(t, X) = \Delta_P(t, X)\delta_P(t, X),$$

adding and subtracting $\mathbb{E}_P\{g_P(t, X)\}$ proves (3.3) and (3.4).

For the expansion, absolute continuity of the gradient along line segments gives, for almost every X and every u in the support of K,

$$\begin{aligned} g_P(t + hu, X) = g_P(t, X) &+ h\nabla_a g_P(t, X)^\top u \\ &+ h^2 \int_0^1 (1 - s)u^\top \nabla_a^2 g_P(t + shu, X)u\, ds. \end{aligned} \tag{A.1}$$

The Hessian integrability condition permits integration of this identity. The linear term vanishes because $\int uK(u)\, du = 0$. Add and subtract $\nabla_a^2 g_P(t, X)$ in the last line of (A.1). The non-remainder part is

$$\begin{aligned} -\frac{h^2}{2}\mathbb{E}_P\left[\int K(u)u^\top \nabla_a^2 g_P(t, X)u\, du\right] &= -\frac{h^2}{2}\mathbb{E}_P\left[\operatorname{tr}\{M_K \nabla_a^2 g_P(t, X)\}\right] \\ &= -\frac{h^2}{2}\operatorname{tr}(M_K H_{2,P}). \end{aligned}$$

The remainder is exactly

$$R_{P,h} = -h^2\mathbb{E}_P\left[\int K(u)\int_0^1 (1 - s)u^\top \left\{\nabla_a^2 g_P(t + shu, X) - \nabla_a^2 g_P(t, X)\right\} u\, ds\, du\right]. \tag{A.2}$$

Because $\|shu\|_\infty \le h$,

$$\begin{aligned} |R_{P,h}| &\le h^2 \int |K(u)|\|u\|_2^2\, du \int_0^1 (1 - s)\, ds\, \omega_{2,P}(h) \\ &= \frac{L_{2,K}h^2}{2}\omega_{2,P}(h), \end{aligned}$$

which proves (3.6).

Applying the fundamental theorem of calculus to the Hessian gives

$$\|\nabla_a^2 g_P(t+v,X) - \nabla_a^2 g_P(t,X)\|_{\mathrm{op}} \le \|v\|_2 \sup_{\|w\|_\infty \le h_0} \|\nabla_a^3 g_P(t+w,X)\|_{\mathrm{op}}.$$

Hence, under Euclidean operator norms, $\omega_{2,P}(h) \le \sqrt{d}\, h M_{3,P}$. If the trilinear norm is induced by $\|\cdot\|_\infty$, the factor $\sqrt{d}$ is absent. Since d is fixed, either convention gives the asserted order. □

Proof of Corollary 3.5. The integral Taylor identity (A.1), used without subtracting the Hessian at t, yields

$$|\beta_{P,h}| \le |B_{0,P}| + \frac{L_{2,K}}{2} r h^2. \tag{A.3}$$

Indeed, the expectation of the Hessian norm at every displacement shu is at most r by (3.7).

Write

$$U = \bar{m}\{\theta(X), X\}, \qquad W_h = \frac{K_h\{A-\theta(X)\}}{\bar{f}(A \mid X)}\{Y - \bar{m}(A,X)\}.$$

The moment assumption gives $\mathbb{E}_P(U^2) \le C$. A change of variables and (3.1) with $p=2$ give

$$\begin{aligned}\mathbb{E}_P(W_h^2) &= h^{-d} \int K(u)^2 \mathbb{E}_{P_X}\left[\frac{f_P(t+hu \mid X)}{\bar{f}(t+hu \mid X)^2}\mathbb{E}_P\{(Y-\bar{m})^2 \mid A = t+hu, X\}\right] du \\ &\le C h^{-d}.\end{aligned} \tag{A.4}$$

It follows that $\mathrm{Var}_P(\psi_h \mid \mathcal{D}_0) \le C(1+h^{-d})$. Since the evaluation observations are conditionally independent,

$$\mathbb{E}_P\{(\widehat{V}_h - V_P)^2 \mid \mathcal{D}_0\} = \beta_{P,h}^2 + \frac{1}{n}\mathrm{Var}_P(\psi_h \mid \mathcal{D}_0).$$

Combining this equality with (A.3) and (A.4) proves (3.8). □

Proof of Theorem 3.6. Corollary 3.5, uniformly over the class in (3.9), gives the first display. It remains to evaluate the chosen bandwidth. If the minimum in (3.11) is attained by its second argument, then

$$\begin{aligned} r^2 h_{n,r}^4 &\asymp n^{-4/(d+4)} r^{2d/(d+4)}, \\ (n h_{n,r}^d)^{-1} &\asymp n^{-4/(d+4)} r^{2d/(d+4)}. \end{aligned}$$

If the bandwidth is capped at h_0, then nr^2 is bounded by a constant depending only on h_0 and c_h. Hence $r^2 h_0^4 = O(n^{-1})$ and $(n h_0^d)^{-1} = O(n^{-1})$. The assumed bound on ρ^2 completes the proof. □

A.2 Conditional limit theory

Proof of Theorem 3.10. Let $\mathcal{F}_n = \sigma(\mathcal{D}_{0,n})$. On the event on which the stated uniform conditions hold, define

$$U_n = \bar{m}_n\{\theta(X), X\}, \qquad W_n = \frac{K_{h_n}\{A-\theta(X)\}}{\bar{f}_n(A \mid X)}\{Y - \bar{m}_n(A,X)\},$$

so that $\psi_n = U_n + W_n$. All expectations in the rest of this proof are conditional on $\mathcal{F}_n$ unless otherwise indicated.

We first compute the conditional variance. A change of variables gives

$$h_n^d \mathbb{E}(W_n^2 \mid \mathcal{F}_n) = \int K(u)^2 \mathbb{E}_{P_n}\{q_n(h_n u, X) \mid \mathcal{F}_n\}\, du. \tag{A.5}$$

By (3.17), the right-hand side is $R_K\Gamma_n + o_P(1)$. Moreover, $\mathbb{E}(U_n^2 \mid \mathcal{F}_n) = O_P(1)$ and $\mathbb{E}(W_n^2 \mid \mathcal{F}_n) = O_P(h_n^{-d})$. Cauchy–Schwarz therefore shows that the contribution of U_n, including its covariance with W_n, is $o_P(1)$ after multiplication by h_n^d. Also $\mathbb{E}(|W_n| \mid \mathcal{F}_n) = O_P(1)$, by Cauchy–Schwarz, the localized $p = 2$ bound, and the uniform upper bound on f_{P_n}. Therefore $\mathbb{E}(\psi_n \mid \mathcal{F}_n) = O_P(1)$ and

$$h_n^d \sigma_n^2 := h_n^d \operatorname{Var}(\psi_n \mid \mathcal{F}_n) = R_K\Gamma_n + o_P(1). \tag{A.6}$$

Thus, with probability tending to one, the conditional variance is bounded above and below by positive multiples of h_n^{-d}.

Let $p = 2 + \eta$ and $\xi_{ni} = \psi_n(Z_i) - \mathbb{E}(\psi_n \mid \mathcal{F}_n)$. The localized pth-moment calculation is

$$\begin{aligned}\mathbb{E}(|W_n|^p \mid \mathcal{F}_n) &= h_n^{-d(p-1)} \int |K(u)|^p \mathbb{E}_{P_{n,X}}\left[\frac{f_{P_n}(t + h_n u \mid X)}{\bar{f}_n(t + h_n u \mid X)^p} \mathbb{E}_{P_n}\{|Y - \bar{m}_n|^p \mid A = t + h_n u, X\}\right] du \\ &\le C h_n^{-d(1+\eta)}.\end{aligned}$$

Together with the moment bound on U_n, this yields

$$\mathbb{E}(|\xi_{ni}|^{2+\eta} \mid \mathcal{F}_n) \le C\{1 + h_n^{-d(1+\eta)}\}. \tag{A.7}$$

For the row of n conditionally independent variables, the Lyapunov ratio is therefore

$$\begin{aligned}\frac{\sum_{i=1}^n \mathbb{E}(|\xi_{ni}|^{2+\eta} \mid \mathcal{F}_n)}{\{n\sigma_n^2\}^{1+\eta/2}} &\le C \frac{n h_n^{-d(1+\eta)}}{(n h_n^{-d})^{1+\eta/2}} + o_P(1) \\ &\le C(n h_n^d)^{-\eta/2} + o_P(1) \longrightarrow_P 0.\end{aligned}$$

The conditional Lyapunov theorem and (A.6) give

$$\frac{\sqrt{n h_n^d}(\mathbb{P}_n - P_n)\psi_n}{\sqrt{R_K\Gamma_n}} \implies N(0,1) \quad \text{conditionally in probability.}$$

Because $P_n\psi_n = V_n + \beta_{n,h_n}$, this is (3.18). For completeness, consider any subsequence and then a further subsequence along which the variance and Lyapunov relations hold almost surely. For almost every realization of the pilot, the deterministic triangular-array theorem applies along this further subsequence. Continuity of the normal cdf makes the cdf convergence uniform in z. The subsequence characterization of convergence in probability proves the conditional claim.

We next prove consistency of (3.19). Let $\widetilde{W}_n = W_n - \mathbb{E}(W_n \mid \mathcal{F}_n)$ and $\widetilde{U}_n = U_n - \mathbb{E}(U_n \mid \mathcal{F}_n)$. The strengthened local moment condition supplies the required fourth moment. With $p_4 = 4 + \eta$ and

$$L_n(u, X) = \frac{f_{P_n}(t + h_n u \mid X)}{\bar{f}_n(t + h_n u \mid X)^{p_4}} \mathbb{E}_{P_n}\{|Y - \bar{m}_n|^{p_4} \mid A = t + h_n u, X\},$$

conditional Jensen's and Hölder's inequalities imply

$$\begin{aligned}&\mathbb{E}_{P_{n,X}}\left[\frac{f_{P_n}(t + h_n u \mid X)}{\bar{f}_n(t + h_n u \mid X)^4} \mathbb{E}_{P_n}\{|Y - \bar{m}_n|^4 \mid A = t + h_n u, X\}\right] \\ &\quad \le \mathbb{E}_{P_{n,X}}\left[L_n(u, X)^{4/p_4} f_{P_n}(t + h_n u \mid X)^{1-4/p_4}\right] \le C.\end{aligned}$$

The last inequality uses the localized p_4 bound and the uniform upper bound on f_{P_n}. A change of variables, boundedness of K, and the usual centering inequality give

$$\mathbb{E}(\widetilde{W}_n^4 \mid \mathcal{F}_n) = O_P(h_n^{-3d}).$$

Consequently,

$$\mathrm{Var}\Big\{h_n^d \mathbb{P}_n \widetilde{W}_n^2 \mid \mathcal{F}_n\Big\} \le \frac{h_n^{2d}}{n}\mathbb{E}(\widetilde{W}_n^4 \mid \mathcal{F}_n) = O_P\{(nh_n^d)^{-1}\} = o_P(1).$$

The original $(2+\eta)$ bound on U_n gives $h_n^d \mathbb{P}_n \widetilde{U}_n^2 = o_P(1)$, by Markov's inequality. The cross term is $o_P(1)$ by the empirical Cauchy–Schwarz inequality. It follows that

$$h_n^d \mathbb{P}_n \xi_n^2 = h_n^d \sigma_n^2 + o_P(1) = R_K \Gamma_n + o_P(1).$$

Moreover, $h_n^d (\mathbb{P}_n \xi_n)^2 = O_P(n^{-1})$, so replacing the conditional mean by the sample mean does not change the limit. Conditional Markov and Chebyshev inequalities in the preceding displays show, for every $\varepsilon > 0$,

$$\mathbb{P}_{P_n}\left(\left|\frac{\widehat{\Omega}_n}{R_K \Gamma_n} - 1\right| > \varepsilon \,\middle|\, \mathcal{F}_n\right) \longrightarrow_P 0.$$

The conditional Slutsky theorem yields studentization.

By (3.6), the error made by replacing the exact bias with the displayed second-order expansion is at most $L_{2,K} h_n^2 \omega_{2,n}(h_n)/2$. Condition (3.20) makes its normalized value negligible. Similarly, (A.3) and (3.21) make the entire exact bias negligible when centering at V_n. These observations prove the final assertions. □

Proof of Theorem 3.11. Condition on $\mathcal{F}_n$ and let $\xi_{ni} = \psi_{h_n}(Z_i) - P_n \psi_{h_n}$. The conditional Lyapunov ratio is

$$\frac{n\mathbb{E}(|\xi_{n1}|^{2+\eta} \mid \mathcal{F}_n)}{(n\sigma_n^2)^{1+\eta/2}} = O_P(n^{-\eta/2}) = o_P(1),$$

because σ_n^2 converges in probability to a positive constant. The subsequence argument from the preceding proof gives the conditional normal limit in probability.

To verify the sample variance, take $q = 1+\min(\eta, 2)/2 \in (1, 2]$. Conditional on $\mathcal{F}_n$, $\mathbb{E}|\xi_{n1}^2 - \sigma_n^2|^q = O_P(1)$. The von Bahr–Esseen inequality therefore gives

$$\mathbb{E}\big[\big|\mathbb{P}_n(\xi_n^2 - \sigma_n^2)\big|^q \big| \mathcal{F}_n\big] \le C n^{1-q} O_P(1) = o_P(1).$$

Also $(\mathbb{P}_n \xi_n)^2 = O_P(n^{-1})$. Hence the sample variance is $\sigma_n^2 + o_P(1)$ and is ratio consistent. Since $P_n \psi_{h_n} = V_n + \beta_{n,h_n}$, direct centering at V_n follows when $\sqrt{n}\,\beta_{n,h_n} = o_P(1)$. □

A.3 Two stochastic-equicontinuity arguments

Proof of Lemma 3.13. Let $T = [-h_0, h_0]^d$. Fix $\varepsilon > 0$. Assumption 3.12(b) permits a fixed $\delta > 0$ for which the probability that

$$\sup_{\substack{v,w \in T \\ \|v-w\|_\infty \le \delta}} \|H_{N,P}(v) - H_{N,P}(w)\|_{P,1} > \varepsilon r_N$$

is eventually as small as desired. Compactness gives a finite δ-net $v_1, \ldots, v_L$ of T. On the complement of the preceding event,

$$\sup_{v \in T} \|H_{N,P}(v)\|_{P,1} \le \max_{1 \le \ell \le L} \|H_{N,P}(v_\ell)\|_{P,1} + \varepsilon r_N.$$

For this fixed finite set, part (a) and a union bound show that the maximum is $O_P(r_N)$. Letting the tail probability and then ε decrease proves the first claim. The second follows from the definitions of $H_{N,P}$ and $\mathcal{C}_\theta$. □

Proof of Theorem 3.14. Write $\psi_{0,n} = \psi_{h_n,\eta_{P_n}}$ and $D_{n,j} = \psi_{h_n,\widehat{\eta}_{-j}} - \psi_{0,n}$. By the definition of $\overline{\beta}_n$,

$$\widehat{V}_{\mathrm{cf}} - V_n - \overline{\beta}_n = (\mathbb{P}_n - P_n)\psi_{0,n} + \sum_{j=1}^{J} \frac{n_j}{n}(\mathbb{P}_{n,j} - P_n)D_{n,j}. \tag{A.8}$$

The first term is one empirical average of the common oracle estimating function; no independence assertion across fold-specific fitted functions is needed. Since $\eta_{P_n} = (m_{P_n}, f_{P_n})$, its exact bias is zero and its leading variance is

$$R_K \mathbb{E}_{P_n}\left[\frac{\mathrm{Var}_{P_n}(Y \mid A = \theta(X), X)}{f_{P_n}\{\theta(X) \mid X\}}\right] = \Omega_{0,n}.$$

The shrinking-bandwidth CLT applied to this oracle term yields

$$\frac{\sqrt{nh_n^d}(\mathbb{P}_n - P_n)\psi_{0,n}}{\sqrt{\Omega_{0,n}}} \implies N(0,1).$$

After multiplication by $\sqrt{nh_n^d}$, the second term in (A.8) is

$$\sum_{j=1}^{J} \sqrt{\frac{n_j}{n}}\sqrt{h_n^d}\,\mathbb{G}_{n,j}D_{n,j} = o_P(1)$$

by (3.22), because J is fixed. Slutsky's theorem proves the first conclusion.

For the sufficient condition, condition on all observations outside fold I_j. The function $D_{n,j}$ is then fixed and the observations in I_j are independent of it. Hence

$$\mathbb{E}\Big[h_n^d\{\mathbb{G}_{n,j}D_{n,j}\}^2\Big|\mathcal{D}_{-j}\Big] = h_n^d\,\mathrm{Var}_{P_n}(D_{n,j} \mid \mathcal{D}_{-j}) \le h_n^d P_n D_{n,j}^2 = o_P(1).$$

Conditional Chebyshev's inequality, together with fixed J, proves (3.22).

Define the cross-fitted scaled sample variance by

$$\widehat{\Omega}_{\mathrm{cf}} = h_n^d \sum_{j=1}^{J} \frac{n_j}{n}\mathbb{P}_{n,j}\{\psi_{h_n,\widehat{\eta}_{-j}} - \widehat{V}_{\mathrm{cf}}\}^2.$$

The foldwise independence and (3.23) imply $h_n^d \sum_j (n_j/n)\mathbb{P}_{n,j}D_{n,j}^2 = o_P(1)$. The empirical Cauchy–Schwarz inequality then shows that replacing every fitted term in $\widehat{\Omega}_{\mathrm{cf}}$ by $\psi_{0,n}$ changes it by $o_P(1)$. Under the stated fourth-moment condition, the variance-consistency calculation in the proof of Theorem 3.10 applies to the oracle terms, giving $\widehat{\Omega}_{\mathrm{cf}}/\Omega_{0,n} \to_P 1$. □

A.4 Reference-assisted minimax bounds

Proof of Theorem 3.7: upper bound. In model (3.12), take $\bar{m} = b$ and $\bar{f} = f_P = 2^{-d}$. The estimating function then equals

$$b(0) + 2^d K_h(A)\{u(A) + \varepsilon\}.$$

Use the continuously differentiable representative of u. A $W^{2,\infty}$ function has a Lipschitz gradient, so the linewise integral Taylor identity applies. Kernel centering bounds the absolute bias by $C_K r h^2$; boundedness of u and Gaussianity of the noise bound the variance by $C(1 + h^{-d})/n$. Choose

$$h = h_0 \wedge c(nr^2)^{-1/(d+4)},$$

with the second term interpreted as infinity at $r = 0$. The calculation in the proof of Theorem 3.6 gives the upper bound $C\mathfrak{R}_{n,d}(r)$ uniformly for $0 \le r \le r_{\max}$.

Proof of Theorem 3.7: lower bound. We use two pairs of Gaussian regression models. For the parametric component, let $u_\pm(a) = \pm c_0 n^{-1/2}$. For all sufficiently large n, both functions belong to $\mathcal{H}_{2,d}(r, M)$ for every $r \geq 0$. Their target values differ by $2c_0 n^{-1/2}$, while their n-observation Kullback–Leibler divergence is $2c_0^2/\sigma^2$. Taking c_0 sufficiently small, Le Cam's two-point inequality gives the lower bound c/n.

For the local component, fix a nonzero function $\varphi \in C_c^\infty((-1/2, 1/2)^d)$ with $\varphi(0) = 1$, and denote finite bounds for its function, gradient, and Hessian norms by L_0, L_1, L_2. Consider the regime $r \geq c_r n^{-1/2}$ and set

$$h = c_h(nr^2)^{-1/(d+4)}, \qquad \Delta = c_\Delta r h^2, \qquad u_0 = 0, \qquad u_1(a) = \Delta\varphi(a/h).$$

The constants c_h, c_Δ can be chosen, depending only on the fixed model constants, so that $h \leq 1$, the bump is supported in $[-1, 1]^d$, and

$$\|u_1\|_\infty \leq M, \qquad \|\nabla u_1\|_\infty \leq M, \qquad \|\nabla^2 u_1\|_{\mathrm{op},\infty} \leq r.$$

Thus both laws belong to $\mathcal{Q}_d(r; b)$. Their divergence is

$$\begin{aligned} D_{\mathrm{KL}}(P_1^n \| P_0^n) &= \frac{n}{2\sigma^2}\mathbb{E}\{u_1(A)^2\} \\ &= \frac{n\Delta^2 h^d}{2^{d+1}\sigma^2}\int \varphi(v)^2\, dv = C_\varphi c_\Delta^2 c_h^{d+4}. \end{aligned}$$

Choose the constants so that the divergence is bounded by a sufficiently small absolute constant. The target separation is Δ, so Le Cam's inequality yields

$$\begin{aligned} \inf_{\widetilde{V}} \sup_{P \in \{P_0, P_1\}} \mathbb{E}_P(\widetilde{V} - V_P)^2 &\geq c\Delta^2 \\ &= cn^{-4/(d+4)} r^{2d/(d+4)}. \end{aligned}$$

If $r < c_r n^{-1/2}$, the local expression is at most a constant multiple of n^{-1}, and the constant-shift construction supplies the maximum. If $r \geq c_r n^{-1/2}$, the local expression is at least a constant multiple of n^{-1} and supplies the maximum. The lower bound is therefore uniform over possibly n-dependent $r \in [0, r_{\max}]$. □

A.5 The population-projection class

A pilot least-squares bound. Let B_ϕ be a uniform bound on $\|\phi(A, X)\|_2$, and put

$$Z_i = \phi(A_i, X_i)\{Y_i - \phi(A_i, X_i)^\top \beta_P\}.$$

By the definition of β_P, $\mathbb{E}_P Z_i = 0$. Matrix Hoeffding's inequality, fixed q, and the lower eigenvalue bound imply

$$\sup_{P \in \mathcal{P}_{\mathcal{B}}(r)} P\{\lambda_{\min}(\widehat{G}) < \lambda/2\} \leq C_1 e^{-C_2 N}. \tag{A.9}$$

On the complementary event,

$$\widehat{\beta} - \beta_P = \widehat{G}^{-1}\left(N^{-1}\sum_{i=1}^N Z_i\right), \qquad \|\widehat{G}^{-1}\|_{\mathrm{op}} \leq 2/\lambda.$$

The uniform sub-Gaussian condition gives $\mathbb{E}\|N^{-1}\sum_i Z_i\|_2^2 \leq C/N$. On the event in (A.9), stabilization sets $\widehat{\beta} = 0$, while $\|\beta_P\| \leq M$. Therefore

$$\sup_{P \in \mathcal{P}_{\mathcal{B}}(r)} \mathbb{E}_P\|\widehat{\beta} - \beta_P\|_2^2 \leq C/N. \tag{A.10}$$

Proof of Theorem 3.8: upper bound. Because the treatment density is known, $B_{0,P} = 0$, and conditional on the pilot the density-ratio-weighted reference error is

$$g_P(a,x) = \phi(a,x)^\top(\widehat{\beta} - \beta_P) - u_P(a,x).$$

Uniform boundedness of the basis Hessians and (3.14) give, for every $\|v\|_\infty \le h_0$,

$$\begin{aligned}\mathbb{E}_{P_X}\|\nabla_a^2 g_P\{\theta(X)+v, X\}\|_{\mathrm{op}} &\le C\|\widehat{\beta} - \beta_P\|_2 + \left[\mathbb{E}_{P_X}\|\nabla_a^2 u_P\{\theta(X)+v, X\}\|_{\mathrm{F}}^2\right]^{1/2} \\ &\le C\|\widehat{\beta} - \beta_P\|_2 + r.\end{aligned}$$

It follows from the integral Taylor formula that the squared conditional bias, averaged over the pilot, is at most

$$Ch^4\{r^2 + N^{-1}\}. \tag{A.11}$$

It remains to bound the variance without a deterministic bound on the realized least-squares coefficient. Since $\phi_1 \equiv 1$, the first coordinate of Z_i is $Y_i - \phi(A_i, X_i)^\top\beta_P$ and has a uniformly bounded second moment by the sub-Gaussian assumption. Decompose

$$Y - \phi(A,X)^\top\widehat{\beta} = Y - \phi(A,X)^\top\beta_P - \phi(A,X)^\top(\widehat{\beta} - \beta_P).$$

Using boundedness of ϕ, the known-density overlap bound, and (A.10) in the same change-of-variables calculation as (A.4), we obtain

$$\mathbb{E}_{\mathcal{D}_0}\{\mathrm{Var}_P(\psi_h \mid \mathcal{D}_0)\} \le C(1 + h^{-d}).$$

Combining this inequality with (A.11) proves (3.15).

Proof of Theorem 3.8: lower bound. Let $T = N + n \asymp n$ be the total sample size. Take $X \equiv x_0$, $A \sim \mathrm{Unif}([-1,1]^d)$, and $Y = m(A, x_0) + \varepsilon$ with independent Gaussian noise of fixed variance. Write $t_0 = \theta(x_0)$. The treatment law has Gram matrix

$$G_0 = 2^{-d}\int_{[-1,1]^d} \phi(a,x_0)\phi(a,x_0)^\top\, da,$$

which is positive definite by assumption.

For the parametric component, take the two regressions $m_\pm(a,x_0) = \pm c_0 T^{-1/2}$. Since $\phi_1 \equiv 1$, each regression lies in $\mathcal{B}$, its projection residual is zero, and all class conditions hold for small c_0. The preceding Gaussian calculation gives a lower bound $c/T \asymp c/n$.

For the local part, use the smooth function φ from the preceding proof, translated to t_0, and set

$$m_0(a,x_0) = 0, \qquad m_1(a,x_0) = \Delta\varphi\{(a - t_0)/h\},$$

where $h = c_h(Tr^2)^{-1/(d+4)}$ and $\Delta = c_\Delta r h^2$. In the relevant regime, choose the constants so that the support lies in the fixed treatment neighborhood. The population projection coefficient satisfies

$$\|\beta_1\|_2 = \left\| G_0^{-1} 2^{-d} \int \phi(a,x_0) m_1(a,x_0)\, da \right\|_2 \le C\Delta h^d.$$

The projection residual is $u_1 = m_1 - \phi^\top\beta_1$. Hence

$$\sup_a \|\nabla_a^2 u_1(a,x_0)\|_{\mathrm{F}} \le C\Delta h^{-2} + C\Delta h^d \le r$$

after reducing c_Δ. Its function and gradient bounds follow in the same way. The residual u_1 is bounded, so $\phi(A, x_0)\{u_1(A, x_0) + \varepsilon\}$ is uniformly sub-Gaussian. Thus these two laws belong to $\mathcal{P}_{\mathcal{B}}(r)$.

The target separation is $m_1(t_0, x_0) - m_0(t_0, x_0) = \Delta$. Projection defines class membership; it does not alter the regression. The T-observation divergence is $CT\Delta^2 h^d$, which is bounded by a small constant by construction. Le Cam's inequality gives $cT^{-4/(d+4)} r^{2d/(d+4)}$. Combining this with the constant-shift lower bound and using $T \asymp n$ proves the lower assertion. The upper assertion at the radius-dependent bandwidth follows from (3.15), because $N^{-1}h^4 \le Cn^{-1}$ when $N \asymp n$. □

A.6 Derivative-rate transfer

Proof of Corollary 3.9. Let

$$r_n = n^{-(\alpha-2)/(2\alpha+d)}.$$

Then $nr_n^2 = n^{(d+4)/(2\alpha+d)}$, so the bandwidth in (3.11) has order $h_n = n^{-1/(2\alpha+d)}$. Consequently,

$$r_n^2 h_n^4 \asymp (nh_n^d)^{-1} \asymp n^{-2\alpha/(2\alpha+d)}.$$

The term n^{-1} is smaller. On the events stated in the corollary, the product-error contribution is also no larger than this order. Applying Theorem 3.6 conditionally on those events proves the $O_P\{n^{-2\alpha/(2\alpha+d)}\}$ claim. □

A.7 Finite-policy regret

Proof of Proposition 4.1. Condition throughout on $\mathcal{D}_0$ and suppress this conditioning from the notation. The reference functions are fixed, and the evaluation observations are independent with law P. For $\theta \in \Theta$, write

$$\psi_{h,\theta}(Z) = \bar{m}\{\theta(X), X\} + \frac{K_h\{A - \theta(X)\}}{\bar{f}(A \mid X)}\{Y - \bar{m}(A, X)\},$$

so that $\widehat{V}_h(\theta) = \mathbb{P}_n \psi_{h,\theta}$.

The exact bias identity and the uniform curvature bound give

$$\sup_{\theta\in\Theta} |P\psi_{h,\theta} - V_P(\theta)| \le \rho + \frac{L_{2,K}}{2} r h^2. \tag{A.12}$$

We next establish uniform variance and envelope bounds for the centered estimating functions. Put

$$U_\theta = \bar{m}\{\theta(X), X\}, \qquad W_{h,\theta} = \frac{K_h\{A - \theta(X)\}}{\bar{f}(A \mid X)}\{Y - \bar{m}(A, X)\}.$$

Since $h \le 1$, the boundedness assumptions imply

$$|W_{h,\theta}| \le \frac{\|K\|_\infty M_1}{\underline{f}}\, h^{-d}$$

almost surely, uniformly over $\theta \in \Theta$. Consequently,

$$|\psi_{h,\theta}| \le M_0 + \frac{\|K\|_\infty M_1}{\underline{f}}\, h^{-d} \le \left(M_0 + \frac{\|K\|_\infty M_1}{\underline{f}}\right) h^{-d},$$

and hence

$$|\psi_{h,\theta} - P\psi_{h,\theta}| \le C_b h^{-d}, \qquad C_b = 2\left(M_0 + \frac{\|K\|_\infty M_1}{\underline{f}}\right). \tag{A.13}$$

Let $t_\theta = \theta(X)$. A change of variables $a = t_\theta + hu$ gives

$$\begin{aligned} PW_{h,\theta}^2 &= h^{-d}\,\mathbb{E}_{P_X}\Bigg[\int K(u)^2 \frac{f_P(t_\theta + hu \mid X)}{\bar{f}(t_\theta + hu \mid X)^2} \\ &\qquad \times \mathbb{E}_P\Big\{(Y - \bar{m}(t_\theta + hu, X))^2 \mid A = t_\theta + hu, X\Big\}\, du\Bigg] \\ &\le h^{-d}\frac{\overline{f} M_1^2}{\underline{f}^2} R_K. \end{aligned}$$

Therefore, using $h \le 1$ again,

$$\begin{aligned} \mathrm{Var}_P(\psi_{h,\theta}) &\le P\psi_{h,\theta}^2 \\ &\le 2M_0^2 + 2h^{-d}\frac{\overline{f} M_1^2}{\underline{f}^2} R_K \le C_v h^{-d}, \end{aligned} \tag{A.14}$$

where

$$C_v = 2M_0^2 + \frac{2\overline{f} M_1^2 R_K}{\underline{f}^2}.$$

For each fixed $\theta \in \Theta$ and $t > 0$, the two-sided Bernstein inequality gives

$$P\left(|(\mathbb{P}_n - P)\psi_{h,\theta}| > \sqrt{\frac{2C_v t}{nh^d}} + \frac{C_b t}{3nh^d} \,\middle|\, \mathcal{D}_0\right) \le 2e^{-t}.$$

A union bound over Θ gives

$$P\left(\sup_{\theta\in\Theta} |(\mathbb{P}_n - P)\psi_{h,\theta}| > \sqrt{\frac{2C_v t}{nh^d}} + \frac{C_b t}{3nh^d} \,\middle|\, \mathcal{D}_0\right) \le 2|\Theta|e^{-t}. \tag{A.15}$$

Set $t = \log(2|\Theta|/\delta)$. Combining (A.12) with (4.1) and using $\widehat{V}_h(\theta) = \mathbb{P}_n\psi_{h,\theta}$ proves (4.2).

We derive the expected bound from the same tail inequality. Let

$$Z_h = \sup_{\theta\in\Theta} |(\mathbb{P}_n - P)\psi_{h,\theta}|, \qquad L = \log(2|\Theta|),$$

and set

$$a = \sqrt{\frac{2C_v}{nh^d}}, \qquad b = \frac{C_b}{3nh^d}.$$

Substituting $t = L + s$ into (A.15) yields, for every $s \ge 0$,

$$P\Big(Z_h > a\sqrt{L+s} + b(L+s) \,\Big|\, \mathcal{D}_0\Big) \le e^{-s}.$$

If $a + b = 0$, then $Z_h = 0$ almost surely and the conclusion is immediate. Otherwise, the function

$$\varphi(s) = a\sqrt{L+s} + b(L+s)$$

is increasing. The tail-integral formula and the change of variables $z = \varphi(s)$ give

$$\begin{aligned}\mathbb{E}_P(Z_h \mid \mathcal{D}_0) &\le \varphi(0) + \int_0^\infty e^{-s}\varphi'(s)\,ds \\ &= a\sqrt{L} + bL + \int_0^\infty e^{-s}\left\{\frac{a}{2\sqrt{L+s}} + b\right\} ds \\ &\le a\sqrt{L} + bL + \frac{a}{2\sqrt{L}} + b.\end{aligned}$$

Since $L \ge \log 2$, the last expression is at most $3\{a\sqrt{L} + bL\}$. The deterministic regret reduction therefore gives

$$\begin{aligned}\mathbb{E}_P\left[V_P(\theta^*) - V_P(\widehat{\theta}) \mid \mathcal{D}_0\right] &\le 2\rho + L_{2,K} r h^2 + 2\mathbb{E}_P(Z_h \mid \mathcal{D}_0) \\ &\le 2\rho + L_{2,K} r h^2 + C\left\{\sqrt{\frac{C_v L}{nh^d}} + \frac{C_b L}{nh^d}\right\},\end{aligned}$$

where, for example, the universal choice $C = 6\sqrt{2}$ is valid. This proves (4.3).

It remains to evaluate the bandwidth. Write

$$H = h_0 \wedge 1, \qquad q = \frac{L}{n}, \qquad h = H \wedge \left(\frac{q}{r^2}\right)^{1/(d+4)},$$

with the second term interpreted as infinity when $r = 0$. Under the assumed condition $q/h^d \le 1$, the linear Bernstein term is bounded by a constant multiple of the square-root term. Hence

$$\mathbb{E}_P\left[V_P(\theta^*) - V_P(\widehat{\theta}) \mid \mathcal{D}_0\right] \le 2\rho + C_0\left\{rh^2 + \sqrt{\frac{q}{h^d}}\right\}, \tag{A.16}$$

where C_0 depends only on the fixed constants in the proposition and on K.

First suppose the radius-dependent bandwidth is not capped. Then $r > 0$ and

$$h = \left(\frac{q}{r^2}\right)^{1/(d+4)}.$$

Then

$$rh^2 = q^{2/(d+4)} r^{d/(d+4)} \qquad \text{and} \qquad \sqrt{\frac{q}{h^d}} = q^{2/(d+4)} r^{d/(d+4)}.$$

Now suppose that the bandwidth is capped at H, including the case $r = 0$. The cap condition gives

$$r^2 \le q H^{-(d+4)},$$

and consequently

$$rH^2 \le H^{-d/2}\sqrt{q}, \qquad \sqrt{\frac{q}{H^d}} = H^{-d/2}\sqrt{q}.$$

Thus, in both cases,

$$rh^2 + \sqrt{\frac{q}{h^d}} \le C_H\left\{\sqrt{q} \ \vee \ q^{2/(d+4)} r^{d/(d+4)}\right\},$$

where C_H depends only on H, and hence only on h_0. Substitution in (A.16), with the factor 2 absorbed into the constant multiplying ρ, proves (4.4). □

B Implementation and additional empirical results

This appendix gives the implementation details and the full empirical record. It begins with the four estimators, common nuisance fits, and tuning rules; then specifies the controlled designs and their favorable- and poor-overlap results. The final part describes the JD.com-based semi-synthetic sample, Gaussian neural policy models, PPO-inspired adaptive-KL training rule, and RMSE and bias curves at all four sample sizes.

B.1 Algorithmic details

B.1.1 Implementation of the estimators

We report IPW, CL, the vectorized implementation KMMS-I, and TR. The last label is retained from the original code for the realized-action estimator.

Inverse Probability Weighting (IPW) The implementation of the IPW estimator (Kallus and Zhou, 2018) is:

Algorithm 1 Inverse Probability Weighting (IPW) Estimator

Input: Data $\{(X_i, A_i, Y_i)\}_{i=1}^n$, target policy θ, bandwidth h, kernel function $K(\cdot)$
Output: IPW estimate $\hat{V}^{\text{IPW}}$
procedure IPW($\{(X_i, A_i, Y_i)\}_{i=1}^n$, θ, h, $K(\cdot)$)
    Estimate conditional treatment density $\hat{f}(A \mid X)$
    for $i = 1$ **to** n **do**
        Compute kernel weights $K_h(\theta(X_i), A_i) = \frac{1}{h} K(\frac{\theta(X_i) - A_i}{h})$
    end for
    $\hat{V}^{\text{IPW}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{K_h(\theta(X_i), A_i)}{\hat{f}(A_i \mid X_i)} Y_i$
end procedure

Double Debiased Machine Learning (CL) The implementation of the CL estimator (Colangelo and Lee, 2026) is:

Algorithm 2 Double Machine Learning (CL) Estimator

Input: Data $\{(X_i, A_i, Y_i)\}_{i=1}^n$, target policy θ, bandwidth h, kernel function $K(\cdot)$
Output: CL estimate $\hat{V}^{\text{CL}}$
procedure CL($\{(X_i, A_i, Y_i)\}_{i=1}^n$, θ, h, $K(\cdot)$)
    Estimate conditional treatment density $\hat{f}(A \mid X)$
    Estimate outcome regression $\hat{m}(A, X)$
    for $i = 1$ **to** n **do**
        Compute kernel weights $K_h(\theta(X_i), A_i) = \frac{1}{h} K(\frac{\theta(X_i) - A_i}{h})$
    end for
    $\hat{V}^{\text{CL}} = \frac{1}{n} \sum_{i=1}^{n} \left[\hat{m}(\theta(X_i), X_i) + \frac{K_h(\theta(X_i), A_i)}{\hat{f}(\theta(X_i) \mid X_i)} \{Y_i - \hat{m}(\theta(X_i), X_i)\} \right]$
end procedure

KMMS KMMS-I implements the local-linear estimator with vectorized matrix operations and optional GPU execution. These changes concern computation, not the estimating equation.

Algorithm 3 Improved Local Linear Kernel Estimator (KMMS-I)

Input: Data $\{(X_i, A_i, Y_i)\}_{i=1}^n$, target policy θ, bandwidth h, kernel function $K(\cdot)$
Output: Improved KMMS estimate $\hat{V}^{\text{KMMS}}_{\text{Improved}}$
procedure ImprovedKMMS($\{(X_i, A_i, Y_i)\}_{i=1}^n$, θ, h, $K(\cdot)$)
    Estimate conditional treatment density $\hat{f}(A \mid X)$
    Estimate outcome regression $\hat{m}_0(A, X)$
    Estimate marginal treatment distribution $\hat{\varpi}(A)$
    Estimate marginal outcome regression $\hat{\mu}(A)$
    for $i = 1$ **to** n **do**
        Compute the pseudo outcome $\xi_i = \frac{Y_i - \hat{m}_0(A_i, X_i)}{\hat{f}(A_i|X_i)} \hat{\varpi}(A_i) + \hat{\mu}(A_i)$
        Kernel weight $K_h(\theta(X_i), A_i) = \frac{K(\frac{\theta(X_i) - A_i}{h})}{h}$
        Feature $g_i(\theta(X_i), A_i) = (1, \frac{\theta(X_i) - A_i}{h})^T$
    end for
    Solve the weighted least squares problem:
    $\hat{\beta} = \arg\min_\beta \sum_{i=1}^n K_h(\theta(X_i), A_i)(\xi_i - g_i(\theta(X_i), A_i)^T \beta)^2$
    Compute the improved KMMS estimate: $\hat{V}^{\text{KMMS}}_{\text{Improved}} = \frac{1}{n} \sum_{i=1}^n g_i(\theta(X_i), A_i)^T \hat{\beta}$
end procedure

Realized-action doubly robust estimator (TR) The realized-action estimator is implemented as follows:

Algorithm 4 Realized-Action Doubly Robust Estimator (TR)

Input: Data $\{(X_i, A_i, Y_i)\}_{i=1}^n$, target policy θ, bandwidth h, kernel function $K(\cdot)$
Output: TR estimate $\hat{V}^{\text{TR}}$
procedure TR($\{(X_i, A_i, Y_i)\}_{i=1}^n$, θ, h, $K(\cdot)$)
    Estimate conditional treatment density $\hat{f}(A \mid X)$
    Estimate outcome regression $\hat{m}(A, X)$
    for $i = 1$ **to** n **do**
        Compute kernel weights $K_h(\theta(X_i), A_i) = \frac{1}{h} K(\frac{\theta(X_i) - A_i}{h})$
    end for
    $\hat{V}^{\text{TR}} = \frac{1}{n} \sum_{i=1}^n \left[\hat{m}(\theta(X_i), X_i) + \frac{K_h(\theta(X_i), A_i)}{\hat{f}(A_i|X_i)} \{Y_i - \hat{m}(A_i, X_i)\} \right]$
end procedure

Nuisance estimates are computed once within each run and reused over the bandwidth grid and bootstrap replications. The same nuisance-learning procedures are used for the estimators that require the corresponding components. KMMS-I changes the numerical implementation of the local-linear calculation through vectorized matrix operations and GPU acceleration; the reported comparison concerns the resulting value estimates.

B.2 Controlled designs and additional results

B.2.1 Data-generating processes

Single-dimensional case Draw $Z_i \sim \text{Uniform}(\pi/2, 7\pi/2)$ and set $X_i = Z_i$. Throughout this section, Normal(μ, s) uses s as the standard deviation. The behavior rule is $\theta_1(X) = 2X + \text{Normal}(0, 4)$

and the target rule is $\theta_2(X) = 1.5X + \pi$. Outcomes satisfy $Y_i = m_0(Z_i, A_i) + \epsilon_i$, with $\epsilon_i \sim \text{Normal}(0, 1)$ and

$$m_0(Z, A) = 4\sin(Z) - 4\sin(A) + (A - 4\pi)^2.$$

Ten-dimensional case Draw $Z_i \sim \text{Uniform}(7\pi/2, 13\pi/2)^{10}$ and set $X_i = Z_i$. The action remains scalar. With $\beta = [0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.11, 0.13, 0.14, 0.15]$, the behavior rule is $\theta_1(X) = 2\beta^\top X + \text{Normal}(0, 2)$ and the target rule is $\theta_2(X) = 1.5\beta^\top X + 5\pi/2$. Outcomes satisfy $Y_i = m_0(Z_i, A_i) + \epsilon_i$, where $\epsilon_i \sim \text{Normal}(0, 1)$ and

$$m_0(Z, A) = \frac{2}{5}\sum_{j=1}^{10} \sin(Z_j) - 4\sin(A) + \frac{1}{2}(A - 10\pi)^2.$$

B.2.2 Nuisance estimation

Outcome regression We fit $\hat{m}(a, x)$ by random-forest regression (Breiman, 2001), using (A, X) as predictors of Y. The implementation uses scikit-learn and five-fold cross-validation.

Conditional treatment density The implementation uses a conditional location model. A separate random forest estimates the conditional action mean $\hat{\theta}_b(x) \approx \mathbb{E}[A \mid X = x]$. A single Gaussian kernel-density estimate is then fitted to the action errors $A_i - \hat{\theta}_b(X_i)$, producing

$$\hat{f}(a \mid x) = \hat{q}\{a - \hat{\theta}_b(x)\}.$$

The conditional mean may vary flexibly with x, while the action-error distribution is common across contexts. Mean-squared error tunes the forest; held-out log likelihood tunes the density bandwidth.

B.2.3 Hyperparameter tuning

The random-forest components are tuned by five-fold grid search. The grid for $\hat{m}$ is:

- `n_estimators`: {100, 200, 300}.
- `max_depth`: {3, 5, 7}.
- `min_samples_split`: {5, 10, 15, 20}.
- `min_samples_leaf`: {2, 4, 6, 8, 10}.

The grid for $\hat{f}$ follows the two-step location-model construction:

1. A random-forest regression estimates the conditional action mean $\hat{\theta}(X)$ over the grid
   - `n_estimators`: {100, 200, 300}
   - `max_depth`: {3, 5, 7}
   - `min_samples_split`: {5, 10, 15, 20}
   - `min_samples_leaf`: {2, 4, 6, 8, 10}
2. A Gaussian kernel estimates the density of $A_i - \hat{\theta}(X_i)$, with bandwidth selected by cross-validation over {1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0}.

Both steps use five folds.

The outer estimator bandwidths are the grids displayed on the horizontal axes of the preserved figures, and the same grid is used for all estimators within a design. The uploaded source does not contain the underlying numerical arrays; they must therefore be recovered from the analysis code for exact reproduction. The minimizing grid values in the two tables are retained verbatim.

After selecting the hyperparameters, the outcome forest is refitted using (A, X) to predict Y, and the action-mean forest is refitted using X to predict A. The latter is combined with the selected Gaussian density bandwidth as described above; the random forest is not itself used as a density estimator.

B.2.4 Estimated and oracle treatment densities

For nonnegative evaluation weights w_i, we use $\mathrm{ESS}(w) = (\sum_i w_i)^2 / \sum_i w_i^2$ as in Martino et al. (2016). The recorded ESS is approximately 481 of 500 in the one-dimensional design and 401 of 500 in the ten-dimensional design. The uploaded source does not identify the bandwidth used for these two summaries, so they are overlap diagnostics rather than reproducible tuning targets. We also rerun all methods with the known treatment density. Figures 9–12 show the resulting RMSE curves. The displayed minima are similar under the estimated and known densities; no significance test is performed.

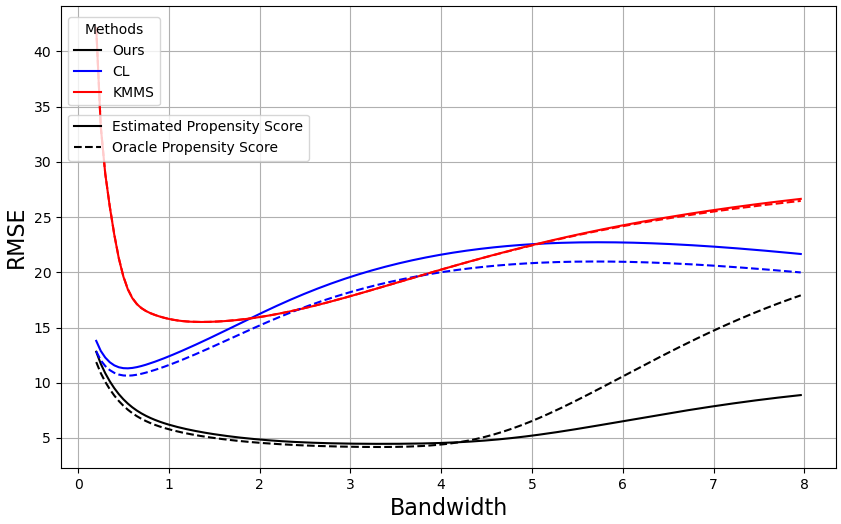

Figure 9: One-dimensional design with the known treatment density, excluding IPW.

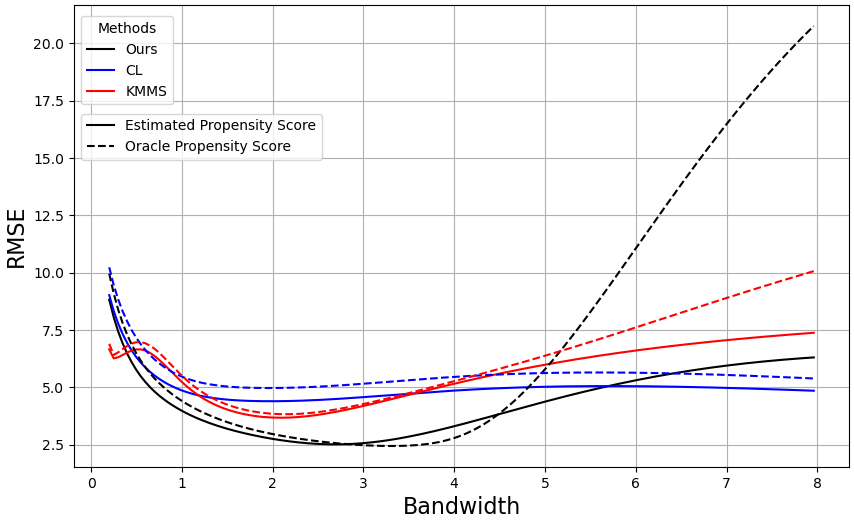

Figure 10: Ten-dimensional covariate design with the known treatment density, excluding IPW.

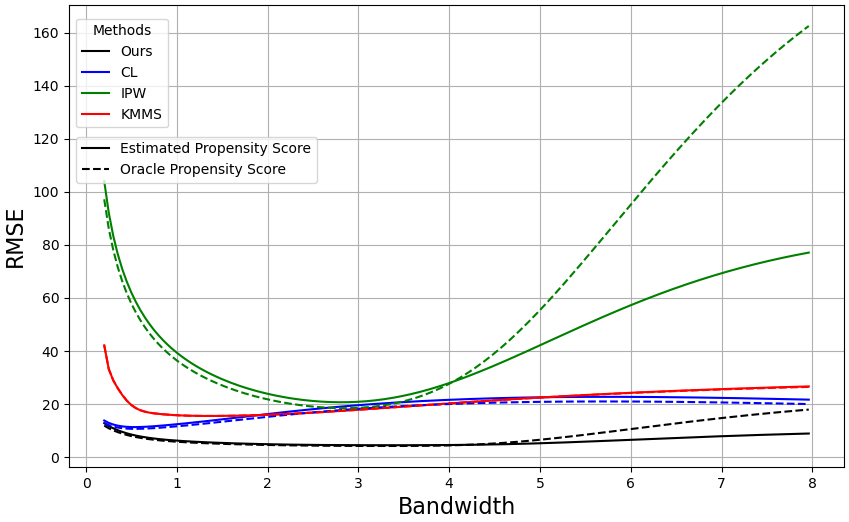

Figure 11: One-dimensional design with the known treatment density.

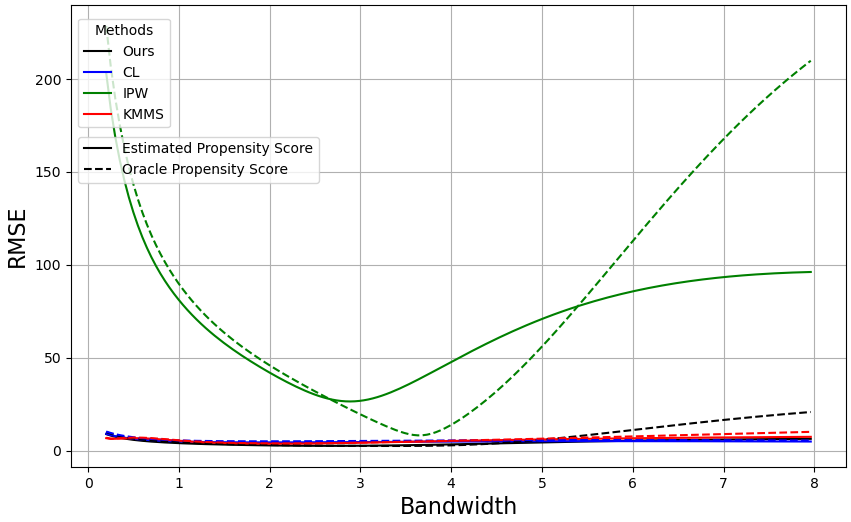

Figure 12: Ten-dimensional covariate design with the known treatment density.

In several panels, an estimated treatment density yields an RMSE minimum close to, and occasionally below, the corresponding known-density minimum. Similar finite-sample behavior for estimated treatment mechanisms has been studied by Robins and Greenland (1992) and Su et al. (2023). The curves here are descriptive: they do not identify a unique mechanism for the difference, and the comparison should not be read as a general preference for an estimated density over the known one.

B.2.5 Favorable overlap: all estimators

The following panels restore IPW to the one- and ten-dimensional comparisons.

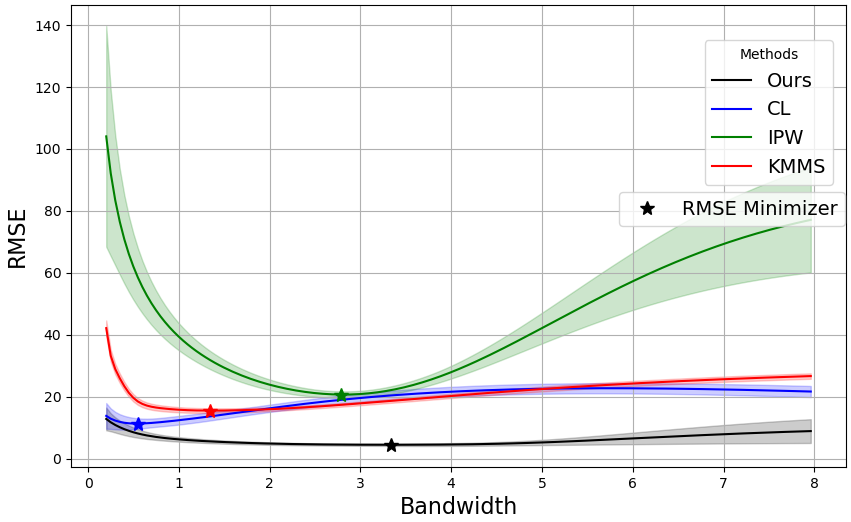

Figure 13: One-dimensional favorable-overlap design: RMSE for all estimators.

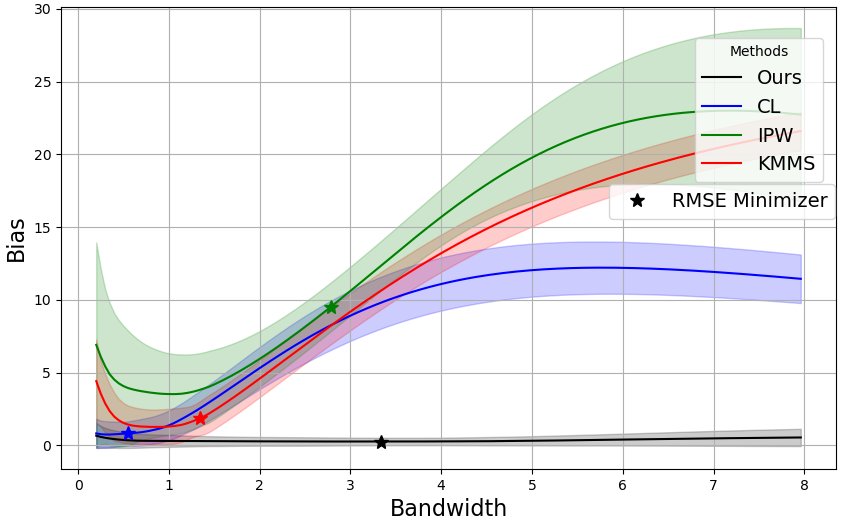

Figure 14: One-dimensional favorable-overlap design: absolute bias for all estimators.

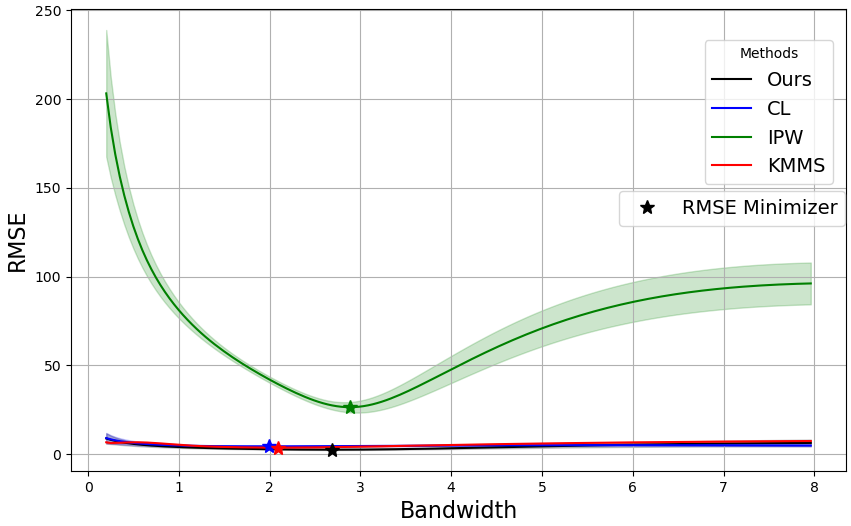

Figure 15: Ten-dimensional favorable-overlap design: RMSE for all estimators.

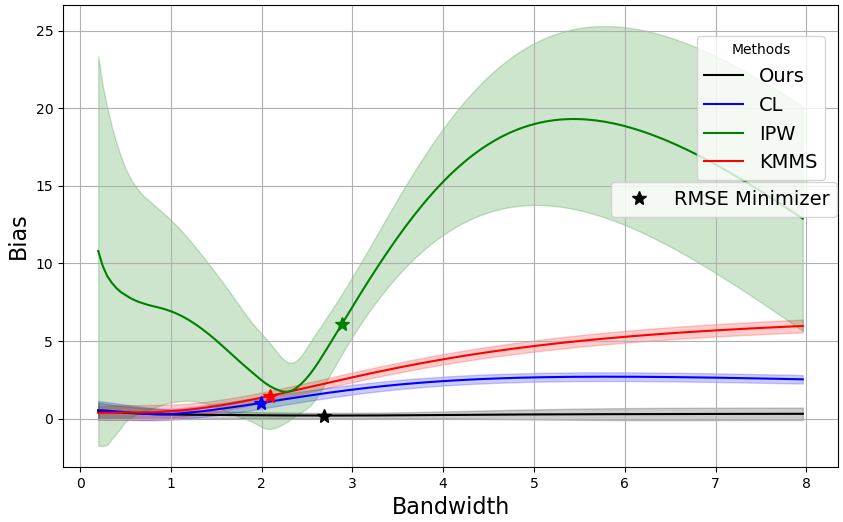

Figure 16: Ten-dimensional favorable-overlap design: absolute bias for all estimators.

Figures 13, 14, 15, and 16 show the relative scale of all four methods. IPW has substantially larger RMSE and bias on the displayed grid. TR has the lowest reported RMSE minimum in both designs and remains near it over a relatively broad range; KMMS-I becomes unstable at the largest bandwidths. These finite-grid comparisons do not establish uniform dominance.

B.2.6 Poor overlap

The negative-control design retains the one-dimensional behavior policy and changes the target to $\theta_2(X) = 1.5X - 10\pi$. This places the target largely outside the region covered by logged actions.

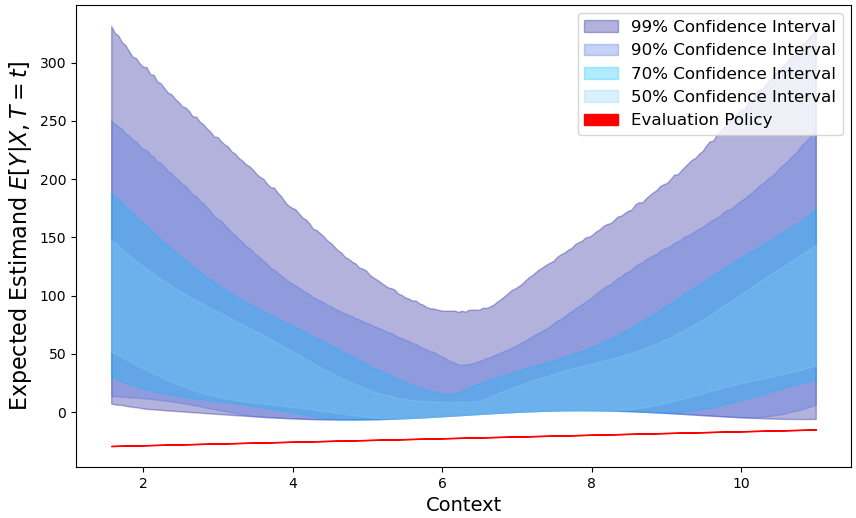

Figure 17: Poor-overlap design: behavior-policy action bands and target rule. The target lies largely outside even the broad behavior-policy bands.

Figure 17 displays central action bands under the behavior policy, not confidence-interval coverage for a value estimator. The target rule lies largely outside even the broad bands, so few logged actions are locally informative for its value.

The remaining panels show the resulting RMSE and bias curves, first for all four methods and then without KMMS-I.

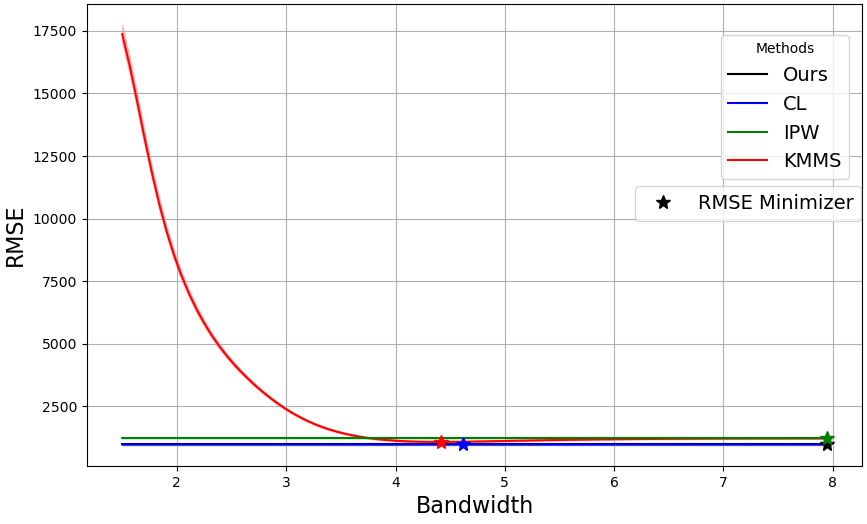

Figure 18: One-dimensional poor-overlap design: RMSE for all estimators.

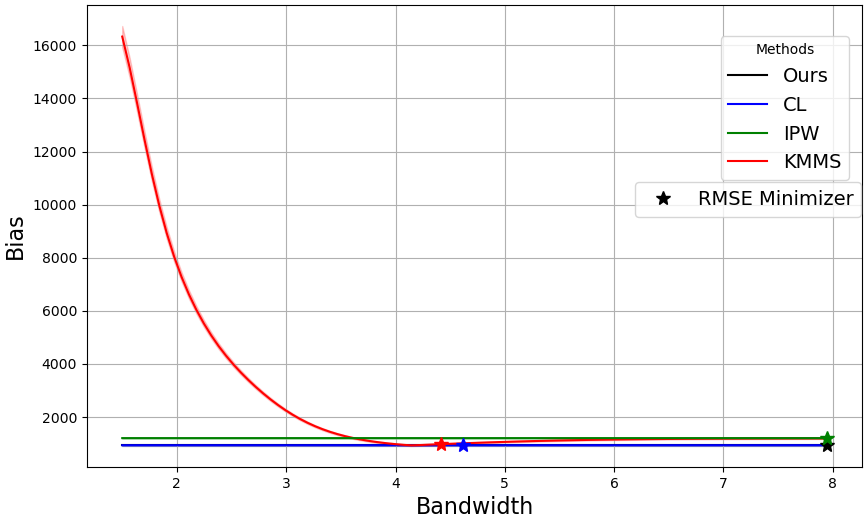

Figure 19: One-dimensional poor-overlap design: absolute bias for all estimators.

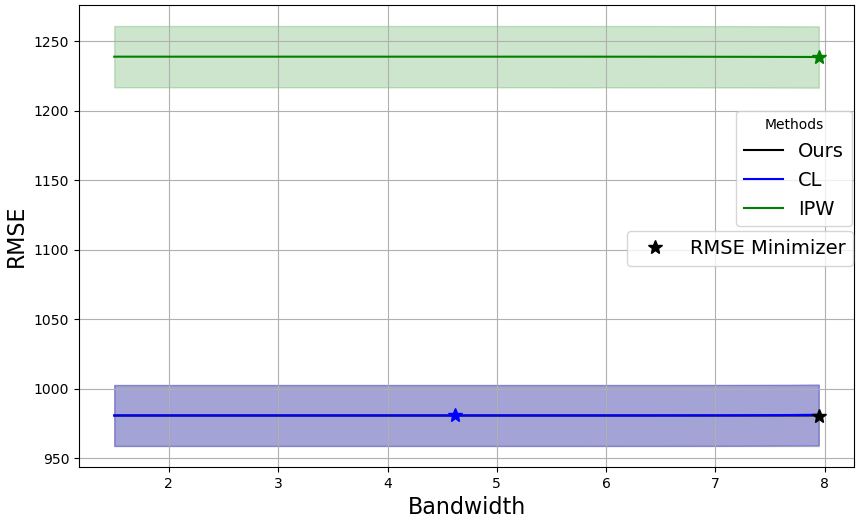

Figure 20: One-dimensional poor-overlap design: RMSE excluding KMMS and KMMS-I.

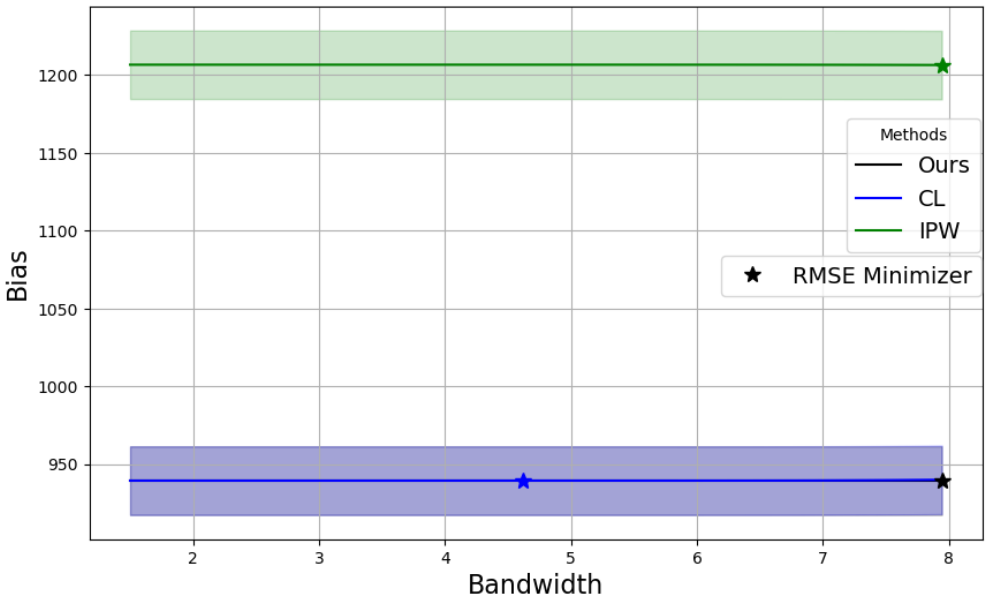

Figure 21: One-dimensional poor-overlap design: absolute bias excluding KMMS and KMMS-I.

Figures 18 and 19 show that every estimator has very large error. At its reported minimizing bandwidth, TR differs from CL by only 0.016 in RMSE and 0.012 in absolute bias. Thus the design is a negative control: the realized-action correction does not restore identification or accurate estimation when the target is unsupported.

B.3 Semi-synthetic design and additional results

B.3.1 Data and features

The semi-synthetic sample is built from the March 2018 JD.com E-Commerce Data Challenge records (Shen et al., 2020). Contexts and prices are observed; the outcomes used to compute repeated-sample RMSE and bias are synthesized as described below.

Feature construction The context vector X contains the following user attributes and behavioral signals:

- **Behavioral features:**
  - `click_count`: number of same-day clicks on the focal SKU;
  - `first_order_month`: month of the user's first order;
  - `months_since_first_order`: elapsed time from the first order to the current interaction.
- **Demographic features** (one-hot encoded):

– `age`: categorical age group, mapped as

$$\text{age\_mapping} = \begin{cases} -1 & \text{U (Unknown age)} \\ 0 & \text{16-25} \\ 1 & \text{26-35} \\ 2 & \text{36-45} \\ 3 & \text{46-55} \\ 4 & \geq 56 \end{cases}$$

– `gender`: user gender;

– `marital_status`: user marital status.

- **Socioeconomic features** (one-hot encoded):

  – `purchase_power`: purchasing-capacity category;

  – `city_level`: city-tier category;

  – `education`: education category.

Price range and sample construction The raw records include discounts below \$10. To keep the learned policies in the operational price range used in the study, we retain final unit prices from \$31.9 to \$89.9 for the focal SKU. This restriction

- removes extreme discounts that can induce negative predictions;
- preserves price variation within the chosen operational range; and
- supports overlap between the fitted behavior and candidate policies.

After filtering and deduplication, 773 distinct user–context–price combinations remain. An ordinary least-squares regression of the outcome on (X, A) defines the outcome model used to synthesize Y. The construction preserves the observed contexts and prices while making the target value available for error calculation. Outcome generation is repeated 800 times to form the semi-synthetic samples used to train the behavior and candidate policies. This policy-construction stage is distinct from the ten value-estimation repetitions conducted at each reported sample size.

The sample sizes in Section 5 refer to observations in each value-estimation run, not to the 773-record base sample.

B.3.2 One-step neural policy construction

Fully connected networks parameterize Gaussian behavior and candidate pricing policies. This module solves a one-step continuous-action optimization problem; there are no states evolving over time, trajectories, or sequential returns.

Gaussian policy network The policy is $\mathcal{N}\{\mu_\theta(X), \sigma_\theta^2(X)\}$. Separate output heads represent its two parameters:

- the mean head returns $\hat{\mu}_\theta(X)$;
- the scale head returns $\log \hat{\sigma}_\theta(X)$, with $\hat{\sigma}_\theta(X) = \exp\{\log \hat{\sigma}_\theta(X)\}$.

The logarithmic parameterization enforces a positive scale. After training, the value estimators take the deterministic target rule to be $x \mapsto \hat{\mu}_\theta(x)$. The scale head enters the training objective and the overlap diagnostics, but not the final deterministic policy value.

Behavior-policy objective The behavior network minimizes three terms.

First, Gaussian negative log likelihood fits observed prices:

$$\mathcal{L}_{\mathrm{NLL}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log \mathcal{N}(A_i; \hat{\mu}_\theta(X_i), \hat{\sigma}^2_\theta(X_i)) \tag{B.1}$$

equivalently,

$$\mathcal{L}_{\mathrm{NLL}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \left[\log \hat{\sigma}_\theta(X_i) + \frac{(A_i - \hat{\mu}_\theta(X_i))^2}{2\hat{\sigma}^2_\theta(X_i)} + \frac{1}{2}\log(2\pi)\right] \tag{B.2}$$

Second, a penalty discourages negative mean prices:

$$\mathcal{L}_{\mathrm{neg}}(\theta) = \lambda_1 \sum_{i=1}^{n} \max(0, -\hat{\mu}_\theta(X_i)) \tag{B.3}$$

where $\lambda_1 = 1.0$.

Third, the average predicted scale is regularized toward the empirical price standard deviation:

$$\mathcal{L}_{\mathrm{std}}(\theta) = \lambda_2 \left| \frac{1}{n}\sum_{i=1}^{n} \hat{\sigma}_\theta(X_i) - \sigma_{\mathrm{data}} \right| \tag{B.4}$$

where σ_{data} is the observed-price standard deviation and $\lambda_2 = 0.1$.

The complete objective is

$$\mathcal{L}_{\mathrm{total}}(\theta) = \mathcal{L}_{\mathrm{NLL}}(\theta) + \mathcal{L}_{\mathrm{neg}}(\theta) + \mathcal{L}_{\mathrm{std}}(\theta) \tag{B.5}$$

Adam minimizes this objective (Kingma and Ba, 2015).

Tuning Validation negative log likelihood selects from the following grid:

- **Network architecture:**
  - Single hidden layer: [64], [128]
  - Two hidden layers: [128, 64], [256, 128]
- **Learning rate:** {1e-3, 5e-4, 1e-4}.
- **Density floor:** predicted conditional densities are clipped below at one of {1e-5, 1e-3, 1e-1}.
- **Regularization:** the final networks use early stopping and no dropout.

The selected specification minimizes validation negative log likelihood.

Candidate-policy objective The behavior policy $\hat{\pi}_{\text{behavior}}(A \mid X)$ is fitted directly to the semi-synthetic sample. The candidate policy instead maximizes the following one-step importance-weighted objective, with a KL penalty and entropy bonus:

$$\mathcal{L}(\theta) = \hat{\mathbb{E}}_{X,A}\left[\frac{\pi_\theta(A|X)}{\hat{\pi}_{\text{behavior}}(A|X)}\hat{R}(X,A)\right] - \beta \cdot \hat{\mathbb{E}}_X\left[D_{\text{KL}}(\hat{\pi}_{\text{behavior}}(\cdot|X)\|\pi_\theta(\cdot|X))\right] + \alpha \cdot \mathcal{H}(\pi_\theta) \quad \text{(B.6)}$$

Here $\pi_\theta(A \mid X) = \mathcal{N}\{A; \mu_\theta(X), \sigma_\theta^2(X)\}$, $\hat{R}(X,A)$ is estimated revenue, β is an adaptive KL coefficient, and α is the entropy weight. The terms have the following roles:

- **Revenue:** importance weighting estimates expected revenue under the candidate policy.
- **KL penalty:** proximity to the behavior policy limits extrapolation beyond logged prices.
- **Entropy bonus:** $\mathcal{H}(\pi_\theta) = \hat{\mathbb{E}}_X[\frac{1}{2}\log\{2\pi e\sigma_\theta^2(X)\}]$ discourages premature concentration.

Adaptive KL controller. At iteration t, let $d_t = \hat{\mathbb{E}}_X[D_{\text{KL}}\{\hat{\pi}_{\text{behavior}}(\cdot \mid X)\|\pi_\theta(\cdot \mid X)\}]$. The controller targets $[d_{\text{lower}}, d_{\text{upper}}]$ through

$$\beta_{t+1} = \begin{cases} \min(\beta_t \cdot \lambda_{\text{inc}}, \beta_{\max}) & \text{if } d_t > d_{\text{upper}} \\ \max(\beta_t/\lambda_{\text{dec}}, \beta_{\min}) & \text{if } d_t < d_{\text{lower}} \\ \beta_t & \text{otherwise} \end{cases} \quad \text{(B.7)}$$

where λ_{inc} and λ_{dec} are adjustment factors.

Optimization parameters. The candidate policy uses:

- 500 training epochs and full batches;
- initial $\beta = 1.4$ and $\alpha = 0.7$;
- Adam learning rate 2e-5;
- KL-controller parameters:
  - KL thresholds: $d_{\text{lower}} = 0.12$, $d_{\text{upper}} = 0.16$
  - Adjustment factors: $\lambda_{\text{inc}} = 1.70$, $\lambda_{\text{dec}} = 1.20$
  - Beta bounds: $\beta_{\min} = 0.80$, $\beta_{\max} = 3.5$
- **Early stopping:** two validation criteria are imposed:
  - validation revenue improves by at least 2%;
  - normalized validation ESS is at least 0.92.

  Training stops when both conditions hold.

The adaptive-KL construction is inspired by PPO (Schulman et al., 2017), but uses neither PPO's clipped surrogate nor a sequential environment. Its purpose here is narrower: to produce a distinct one-step pricing policy while retaining overlap with historical prices.

B.3.3 Results by sample size

Independent samples are drawn from the learned behavior policy at $n \in \{10\text{k}, 50\text{k}, 100\text{k}, 200\text{k}\}$. Each sample is evaluated by IPW, CL, KMMS-I, and TR.

We report

- RMSE $= \{\mathbb{E}[\{\hat{V}(\theta) - V(\theta)\}^2]\}^{1/2}$;
- Bias $= |\mathbb{E}\{\hat{V}(\theta) - V(\theta)\}|$.

For each n, one pair of panels includes all four methods and another shows CL and TR alone on a more informative scale.

Sample size $n = 10\text{k}$ Figures 22 and 23 show the four estimators at $n = 10{,}000$. At this sample size, CL has the lowest reported RMSE minimum.

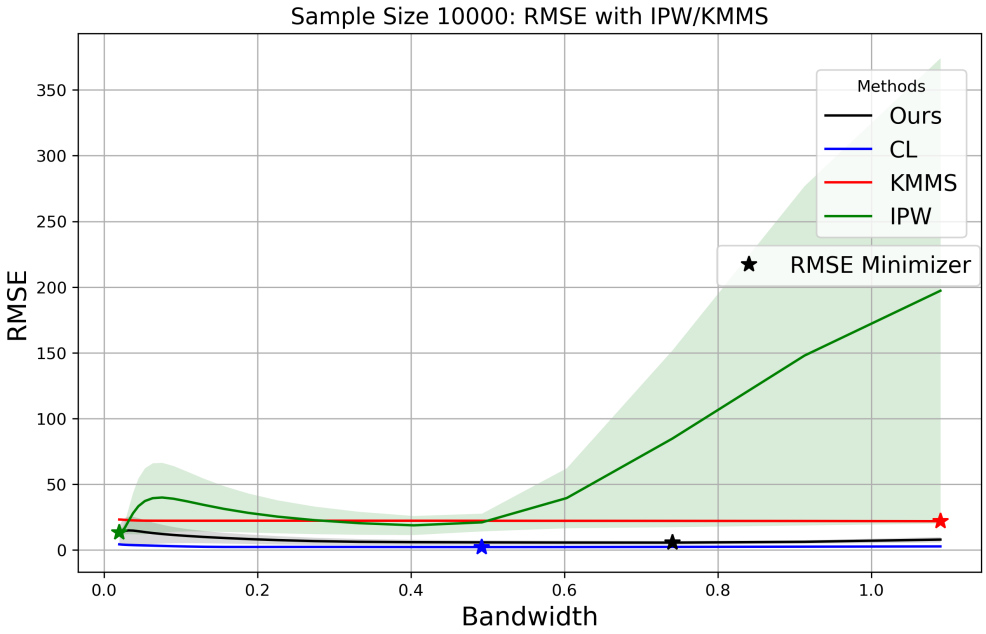

Figure 22: Semi-synthetic JD.com design, $n = 10\text{k}$: RMSE for all estimators.

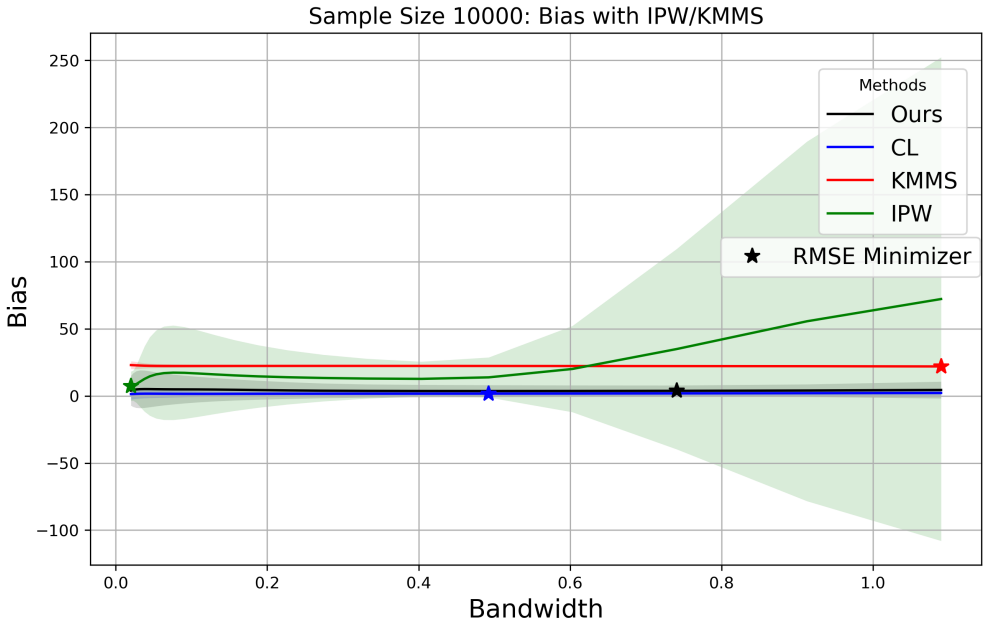

Figure 23: Semi-synthetic JD.com design, $n = 10$k: absolute bias for all estimators.

Figures 24 and 25 show CL and TR alone.

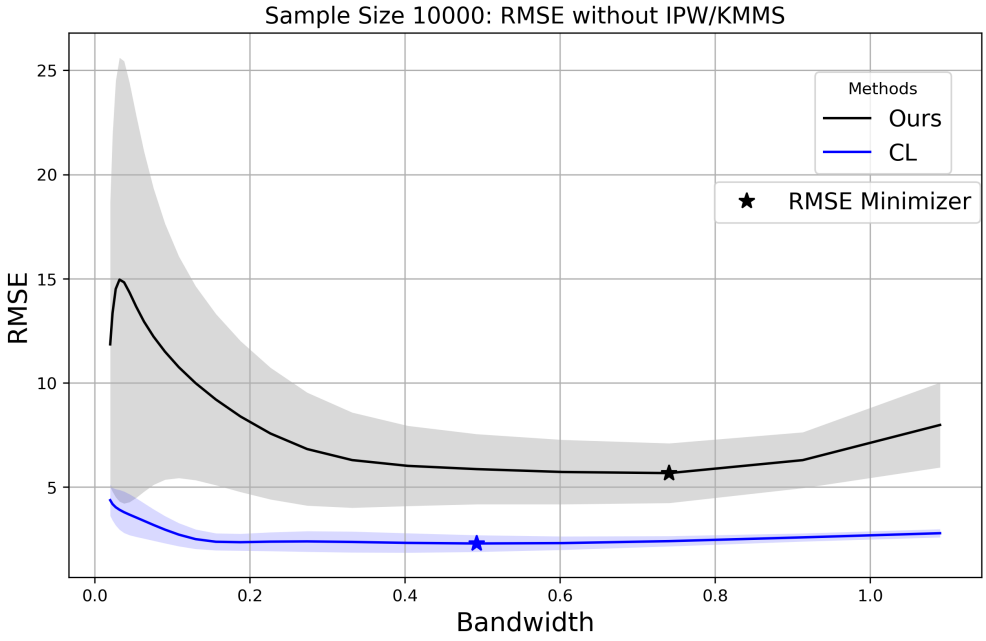

Figure 24: Semi-synthetic JD.com design, $n = 10$k: RMSE for CL and TR.

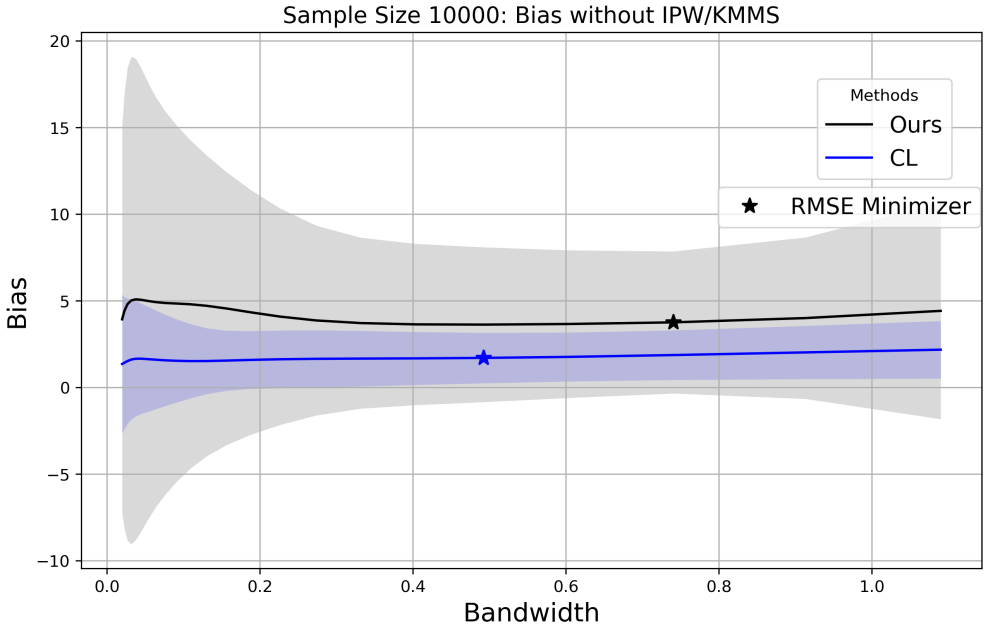

Figure 25: Semi-synthetic JD.com design, $n = 10$k: absolute bias for CL and TR.

Sample size $n = 50$k At $n = 50{,}000$, TR has the lowest reported RMSE minimum. The panels retain the full RMSE and bias curves over bandwidth.

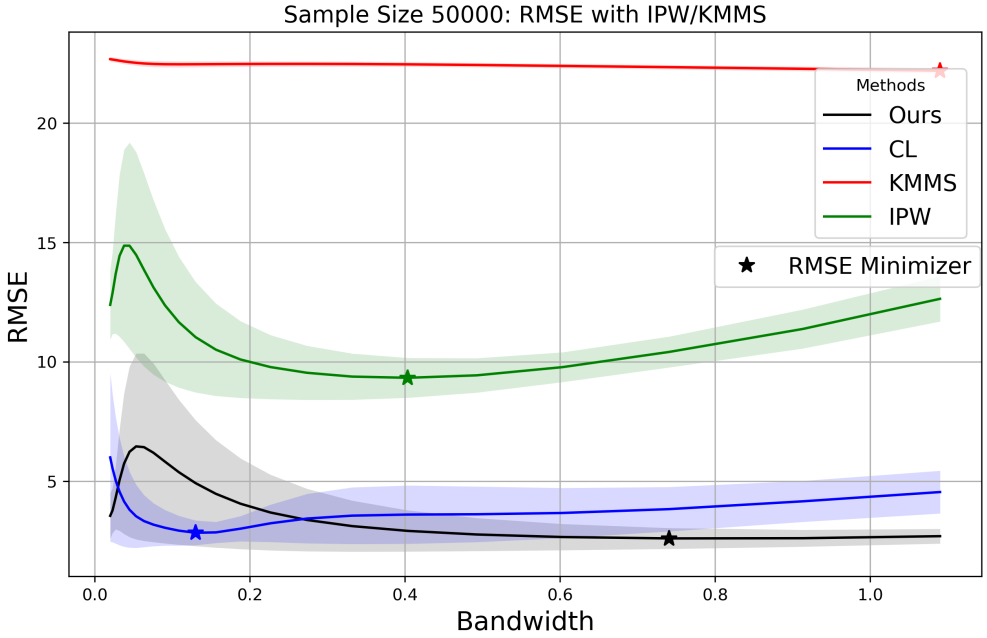

Figure 26: Semi-synthetic JD.com design, $n = 50$k: RMSE for all estimators.

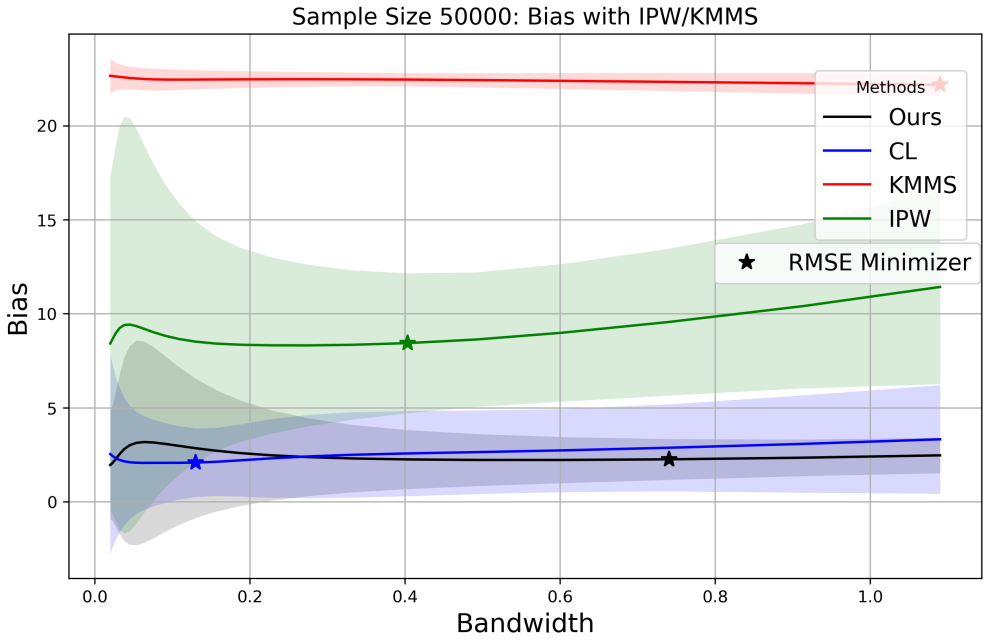

Figure 27: Semi-synthetic JD.com design, $n = 50$k: absolute bias for all estimators.

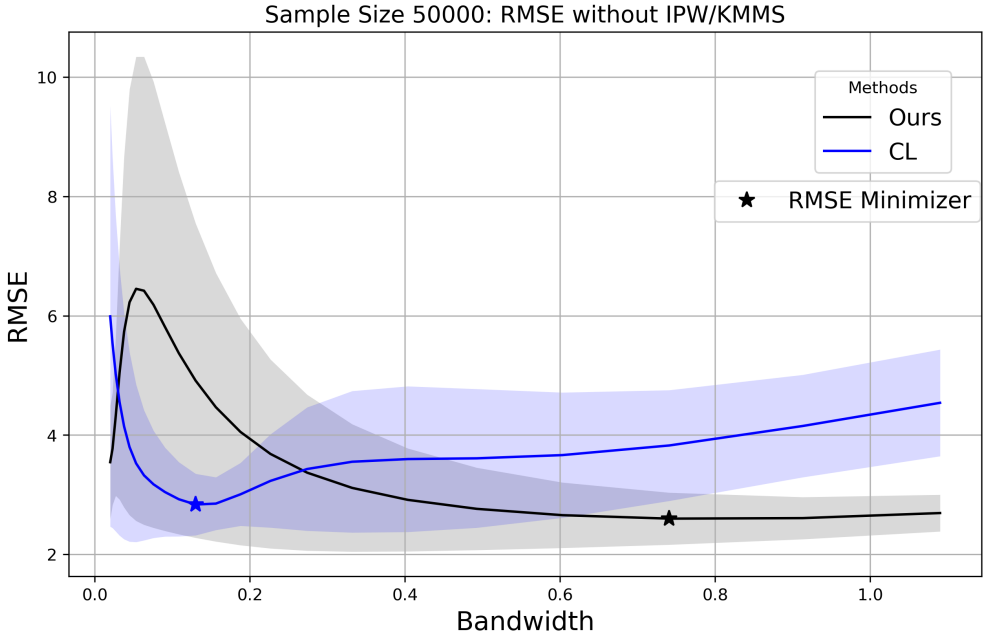

Figure 28: Semi-synthetic JD.com design, $n = 50$k: RMSE for CL and TR.

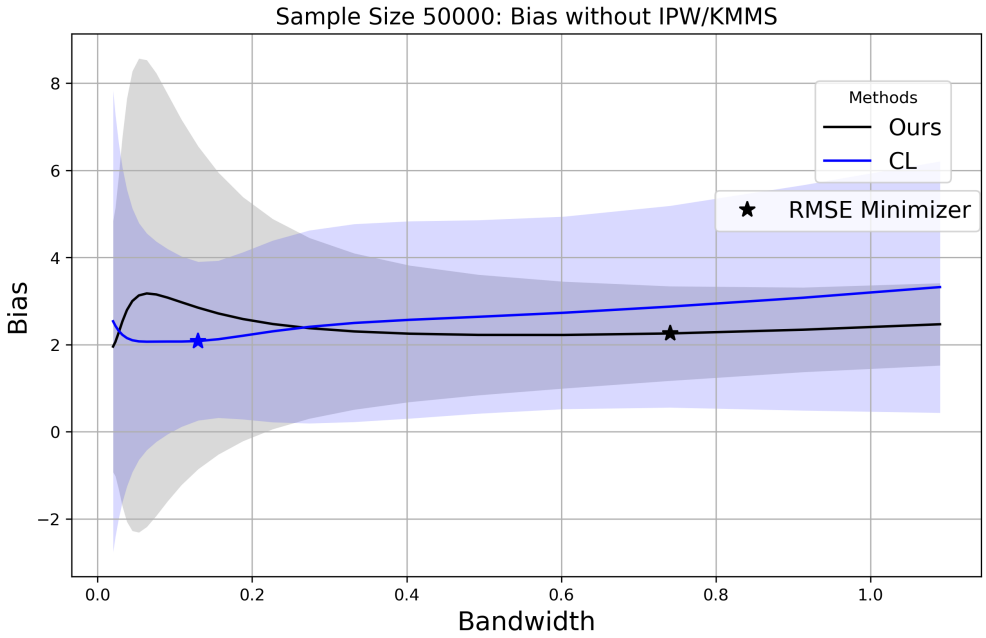

Figure 29: Semi-synthetic JD.com design, $n = 50$k: absolute bias for CL and TR.

Sample size $n = 100$k At $n = 100{,}000$, TR again has the lowest reported RMSE minimum. The panels show the full bandwidth dependence with and without IPW and KMMS-I.

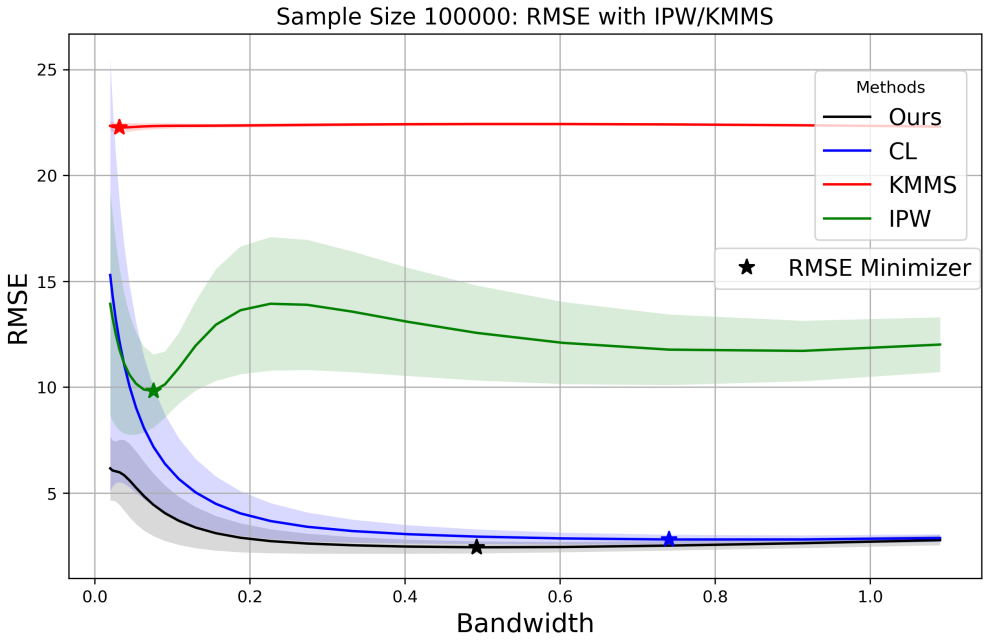

Figure 30: Semi-synthetic JD.com design, $n = 100$k: RMSE for all estimators.

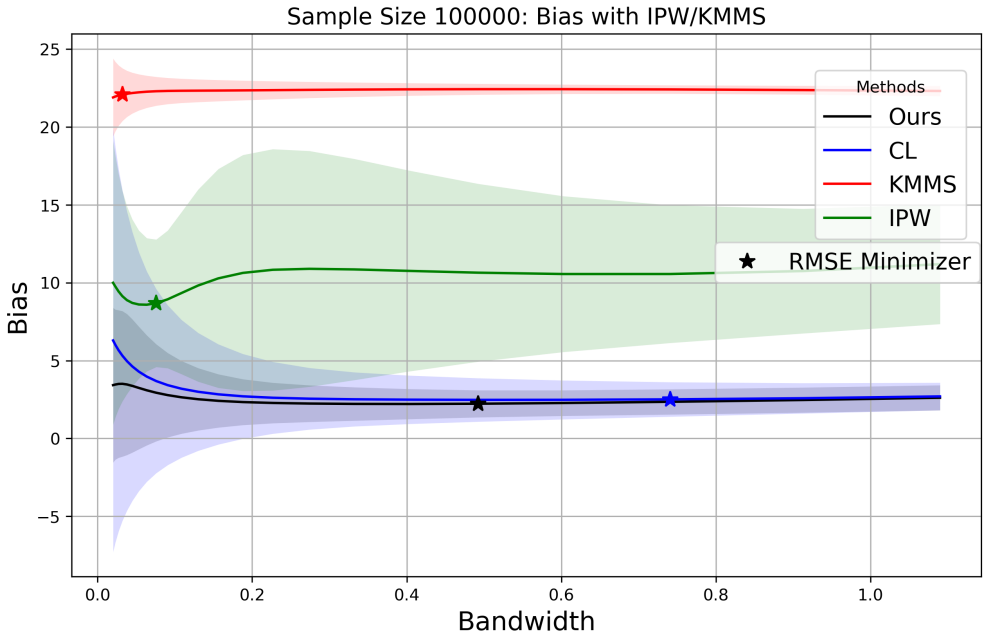

Figure 31: Semi-synthetic JD.com design, $n = 100$k: absolute bias for all estimators.

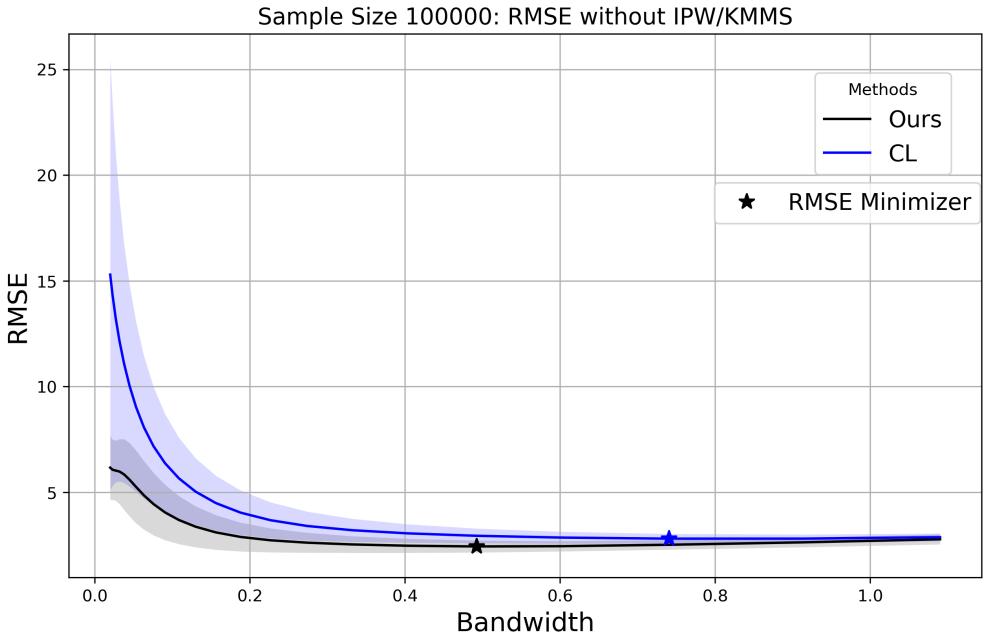

Figure 32: Semi-synthetic JD.com design, $n = 100$k: RMSE for CL and TR.

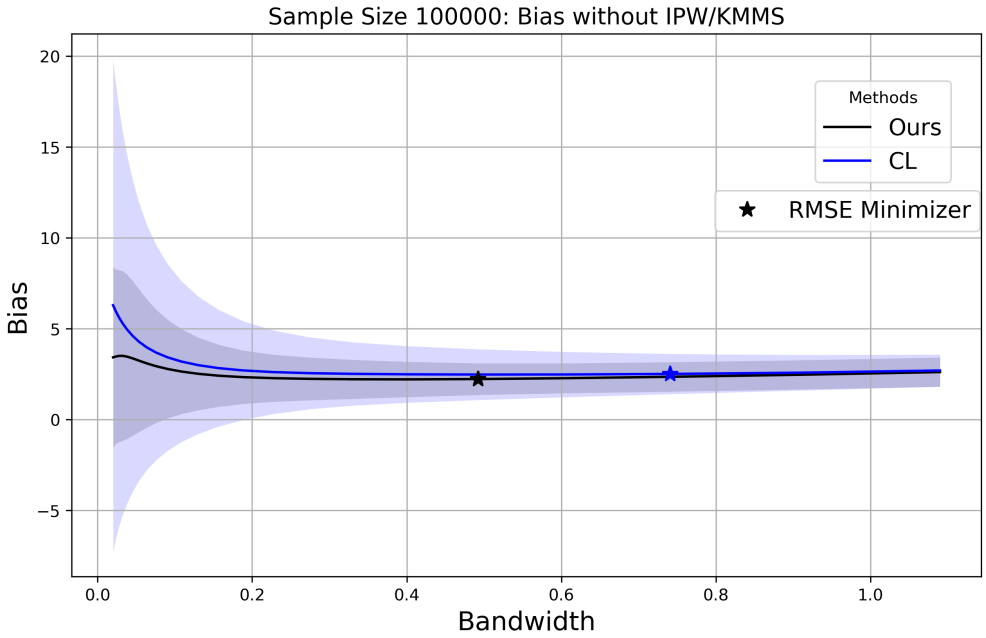

Figure 33: Semi-synthetic JD.com design, $n = 100$k: absolute bias for CL and TR.

Sample size $n = 200$k At $n = 200{,}000$, TR has the lowest reported RMSE minimum and remains competitive with CL over a comparatively broad bandwidth range.

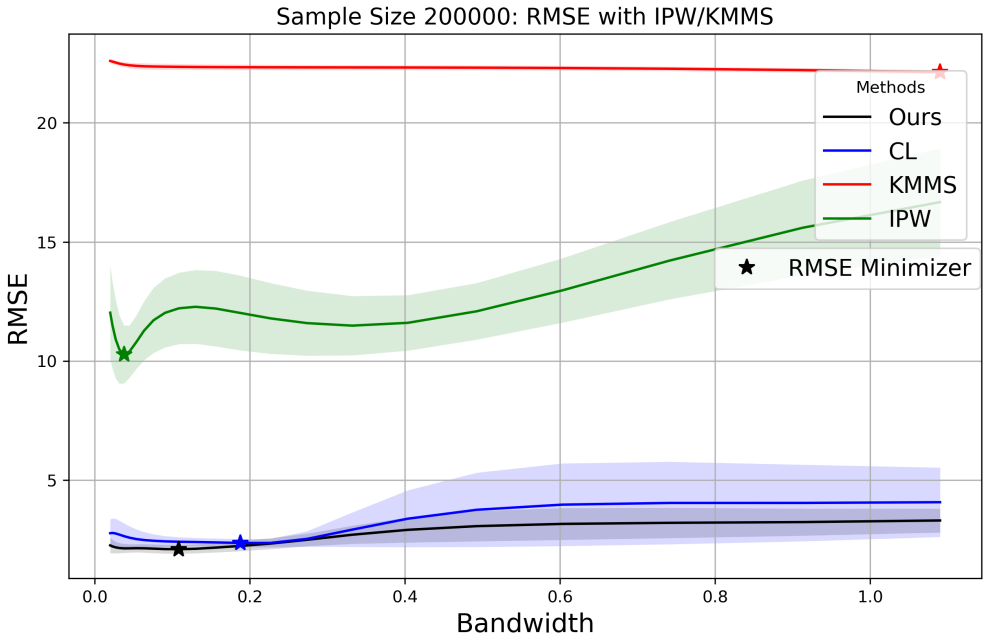

Figure 34: Semi-synthetic JD.com design, $n = 200$k: RMSE for all estimators.

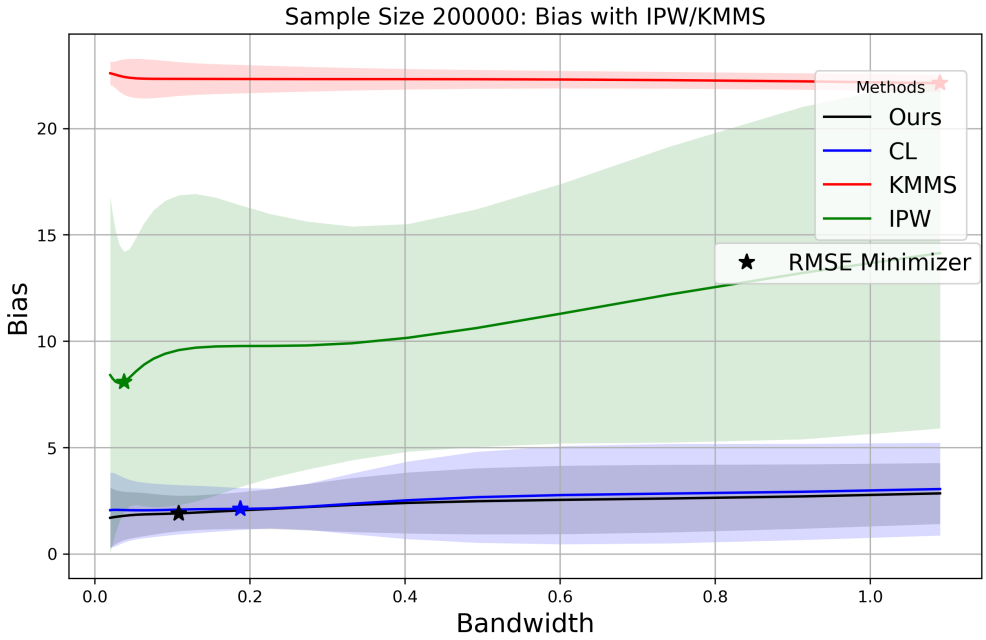

Figure 35: Semi-synthetic JD.com design, $n = 200$k: absolute bias for all estimators.

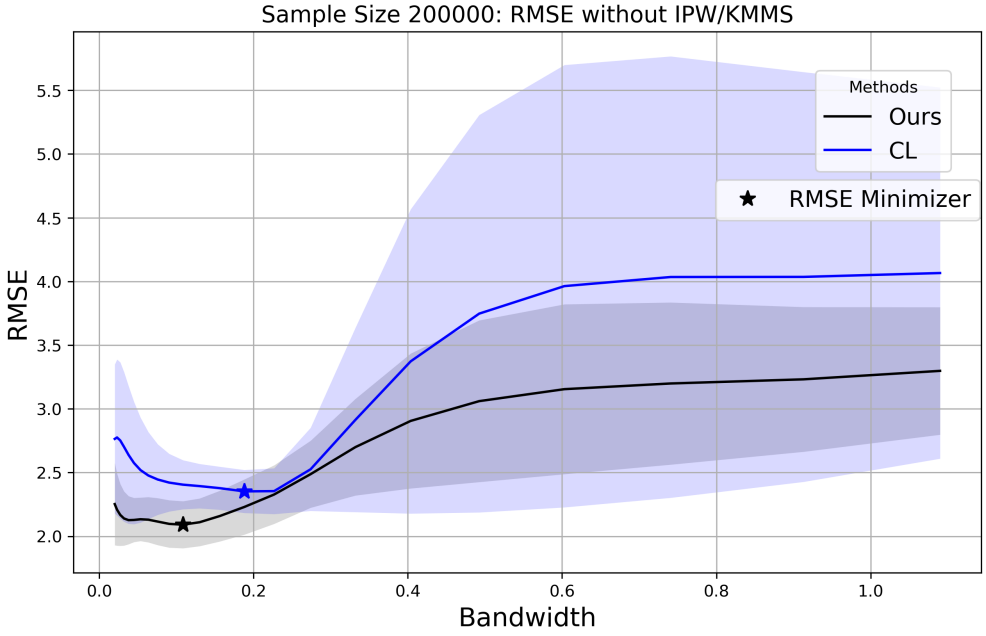

Figure 36: Semi-synthetic JD.com design, $n = 200$k: RMSE for CL and TR.

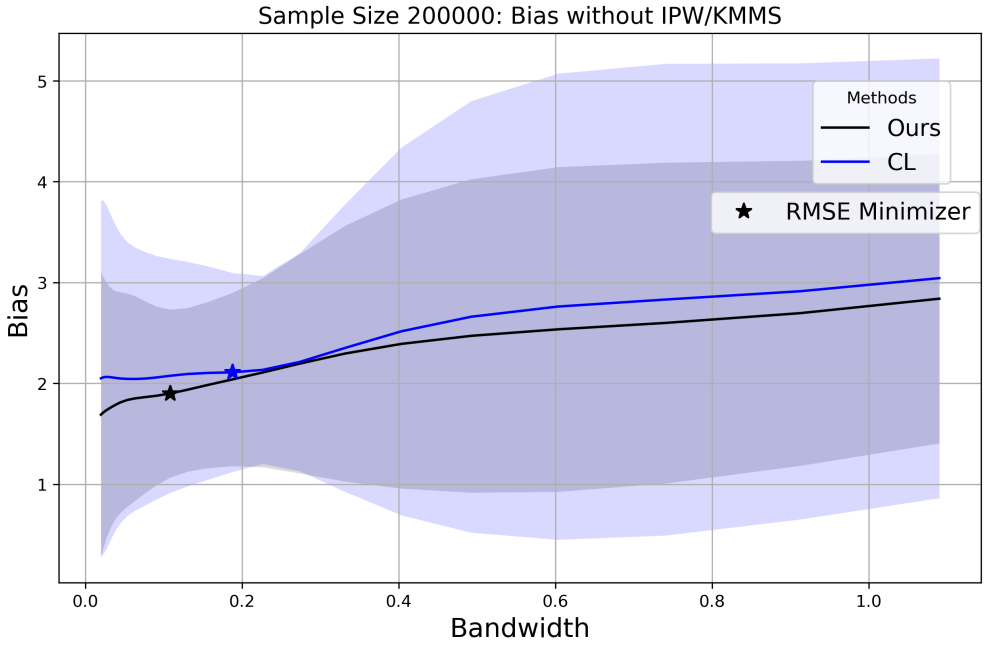

Figure 37: Semi-synthetic JD.com design, $n = 200$k: absolute bias for CL and TR.

Summary across sample sizes There is no uniform winner. CL has the lower minimum RMSE at $n = 10{,}000$; TR has the lower minimum at $n = 50{,}000$, 100,000, and 200,000, reducing CL's minimum by 8.3%, 13.1%, and 11.1%, respectively. The curves echo the controlled designs: TR remains competitive over a relatively broad bandwidth range before bias offsets variance reduction. IPW and KMMS-I have substantially larger errors on the reported grid. Because outcomes are synthesized, the findings concern estimator error in a semi-synthetic one-step pricing design, not deployed revenue.

References

Donald W. K. Andrews. Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica*, 62(1):43–72, 1994. doi: 10.2307/2951475.

Matteo Bonvini and Edward H. Kennedy. Fast convergence rates for dose-response estimation. *arXiv preprint arXiv:2207.11825*, 2026. Revised April 2026.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

Lawrence D Brown and Mark G Low. A constrained risk inequality with applications to nonparametric functional estimation. *The annals of Statistics*, 24(6):2524–2535, 1996.

Kyle Colangelo and Ying-Ying Lee. Double debiased machine learning nonparametric inference with continuous treatments. *Journal of Business & Economic Statistics*, 44(1):67–79, 2026. doi: 10.1080/07350015.2025.2505487.

Arnoud V. den Boer. Dynamic pricing and learning: Historical origins, current research, and new directions. *Surveys in Operations Research and Management Science*, 20(1):1–18, 2015. doi: 10.1016/j.sorms.2015.03.001.

Leo Guelman and Montserrat Guillén. A causal inference approach to measure price elasticity in automobile insurance. *Expert Systems with Applications*, 41(2):387–396, 2014. doi: 10.1016/j.eswa.2013.07.059.

Nathan Kallus and Masatoshi Uehara. Doubly robust off-policy value and gradient estimation for deterministic policies. In *Advances in Neural Information Processing Systems*, volume 33, pages 10420–10430, 2020.

Nathan Kallus and Angela Zhou. Policy evaluation and optimization with continuous treatments. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1243–1251, 2018.

Edward H Kennedy, Zongming Ma, Matthew D McHugh, and Dylan S Small. Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(4):1229–1245, 2017.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the Third International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Enno Mammen, Christoph Rothe, and Melanie Schienle. Nonparametric regression with nonparametrically generated covariates. *The Annals of Statistics*, 40(2):1132–1170, 2012. doi: 10.1214/12-AOS995.

L. Martino, V. Elvira, and F. Louzada. Effective sample size for importance sampling based on discrepancy measures. *arXiv preprint arXiv:1602.03572*, 2016.

Wenlong Mou, Peng Ding, Martin J. Wainwright, and Peter L. Bartlett. Kernel-based off-policy estimation without overlap: Instance optimality beyond semiparametric efficiency. *arXiv preprint arXiv:2301.06240*, 2023.

Whitney K. Newey. Uniform convergence in probability and stochastic equicontinuity. *Econometrica*, 59(4):1161–1167, 1991. doi: 10.2307/2938179.

Whitney K. Newey. Kernel estimation of partial means and a general variance estimator. *Econometric Theory*, 10(2):233–253, 1994. doi: 10.1017/S0266466600008409.

Whitney K. Newey and Daniel McFadden. Large sample estimation and hypothesis testing. In Robert F. Engle and Daniel L. McFadden, editors, *Handbook of Econometrics*, volume 4, chapter 36, pages 2111–2245. Elsevier, 1994. doi: 10.1016/S1573-4412(05)80005-4.

Adrian Pagan. Econometric issues in the analysis of regressions with generated regressors. *International Economic Review*, 25(1):221–247, 1984. doi: 10.2307/2648877.

David J. Reibstein and Hubert Gatignon. Optimal product line pricing: The influence of elasticities and cross-elasticities. *Journal of Marketing Research*, 21(3):259–267, 1984. doi: 10.1177/002224378402100303.

James M Robins and Sander Greenland. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 48(2):479–495, 1992.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Max Shen, Christopher S. Tang, Di Wu, Rong Yuan, and Wei Zhou. Jd.com: Transaction-level data for the 2020 msom data driven research challenge. *Manufacturing & Service Operations Management*, 26(1):2–10, 2020.

Charles J. Stone. Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8(6):1348–1360, 1980. doi: 10.1214/aos/1176345206.

Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982. doi: 10.1214/aos/1176345969.

Fangzhou Su, Wenlong Mou, Peng Ding, and Martin J. Wainwright. High-dimensional analysis and bias correction for continuous treatment effects. *arXiv preprint arXiv:2303.17102*, 2023.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.