

On the Rates of Convergence of Induced Ordered Statistics and their Applications*

Federico A. Bugni

Department of Economics
Northwestern University

federico.bugni@northwestern.edu

Ivan A. Canay

Department of Economics
Northwestern University

iacanay@northwestern.edu

Deborah Kim

Department of Economics
University of Warwick

deborah.kim@warwick.ac.uk

March 7, 2026

Abstract

Induced order statistics (IOS) arise when sample units are reordered according to the value of an auxiliary variable, and the associated responses are analyzed in that induced order. IOS play a central role in applications where the goal is to approximate the conditional distribution of an outcome at a fixed covariate value using observations whose covariates lie closest to that point, including regression discontinuity designs, k -nearest-neighbor methods, and distributionally robust optimization. Existing asymptotic results allow the dimension of the IOS vector to grow with the sample size only under smoothness conditions that are often too restrictive for practical data-generating processes. In particular, these conditions rule out boundary points, which are central to regression discontinuity designs. This paper develops general convergence rates for IOS under primitive and comparatively weak assumptions. We derive sharp marginal rates for the approximation of the target conditional distribution in Hellinger and total variation distances under quadratic mean differentiability and show how these marginal rates translate into joint convergence rates for the IOS vector. Our results are widely applicable: they rely on a standard smoothness condition and accommodate both interior and boundary conditioning points, as required in regression discontinuity and related settings. In the supplementary appendix, we provide complementary results under a Taylor/Hölder remainder condition. Our results reveal a clear trade-off between smoothness and speed of convergence, identify regimes in which Hellinger and total variation distances behave differently, and provide explicit growth conditions on the number of nearest neighbors.

KEYWORDS: total variation distance, Hellinger distance, induced order statistics, rank tests, permutation tests.

JEL classification codes: C12, C14.

*We thank Matias Cattaneo for comments and discussion. All errors are our own.

1 Introduction

Induced order statistics arise when units in a sample are reordered according to the value of an auxiliary variable, and one then studies the associated responses in that induced order. Classical work on induced order statistics (IOS) has developed their basic probabilistic properties, limit distributions, and representations in terms of mixtures and conditional laws; see, for instance, [David and Galambos \(1974\)](#); [Bhattacharya \(1974\)](#); [Reiss \(1989\)](#); [Kaufmann and Reiss \(1992\)](#) and the book-level treatment by [Falk et al. \(2010\)](#). More recently, IOS have proved useful in a number of econometric and statistical applications in which one wishes to approximate the conditional distribution of an outcome given a covariate taking a specific value. A prominent example is the regression discontinuity design, where identifying assumptions are formulated in terms of the distribution of potential outcomes at a cutoff and IOS provide a natural way to approximate this distribution using observations whose running variable lies closest to the cutoff; see, for example, [Canay and Kamat \(2018\)](#); [Goldman and Kaplan \(2018\)](#); [Bugni and Canay \(2021\)](#); [Bugni et al. \(2023, 2025\)](#). Similar constructions appear in k -nearest-neighbor methods, distributionally robust optimization, and related problems where analysis focuses on the subsample formed by observations closest to a point of interest; see, for example, [Esteban-Pérez and Morales \(2022\)](#).

In the vast majority of these applications, the dimension of the IOS vector is kept fixed in the asymptotic approximations. Existing results that allow the dimension to grow with the sample size rely on smoothness assumptions that are often too restrictive for the underlying data-generating process. This raises the question of whether one can obtain general convergence results for IOS under primitive and comparatively weak conditions, in a form that can serve as a reusable toolkit for analyzing tests and estimators based on k nearest neighbors as k grows with the sample size.

The present paper answers this question. We derive convergence rates in the Hellinger and total variation distances for the joint law of the IOS, and we trace how these rates depend on the smoothness of the underlying model. Our main results are widely applicable: they rely on quadratic mean differentiability (QMD) of the conditional densities at the conditioning point, a standard condition in asymptotic statistics, and accommodate both interior and boundary conditioning points—as required in regression discontinuity designs and related settings. Under this condition, we obtain sharp marginal rates for approximating the target conditional distribution and show how these translate into joint convergence rates for the IOS vector, with explicit growth conditions on k . In the supplementary appendix, we provide complementary results under a Taylor/Hölder remainder condition, which yield a wider range of marginal exponents and clarify how rates slow down or fail as smoothness weakens.

Formally, we consider a pair of random vectors (X, Y) taking values in $\mathbf{R}^d \times \mathbf{R}^m$, with joint density f . The conditional law of Y given $X = x_0$ is denoted by P , while for small $r > 0$ the conditional law of Y given $X \in B_r := \{x : \|x - x_0\| < r\}$ is denoted by P_r . Given an i.i.d. sample $\{(X_i, Y_i) : 1 \leq i \leq n\}$, we use Euclidean distance to x_0 to identify the set of the k nearest observations and collect the associated responses into the vector, with components listed in original

sample order,

$$S_n := (S_{n,1}, \dots, S_{n,k}) .$$

The ideal benchmark experiment for many applications would be an i.i.d. sample of size k from P , say $S = (\tilde{Y}_1, \dots, \tilde{Y}_k)$ with $\tilde{Y}_1, \dots, \tilde{Y}_k$ i.i.d. with law P . Our primary object of interest is the discrepancy between the laws $\mathcal{L}(S_n)$ and $\mathcal{L}(S)$ as a function of n and k , measured in Hellinger distance $H(\cdot)$ and total variation distance $\text{TV}(\cdot)$. These metrics control the error of tests and estimators based on S_n relative to their infeasible counterparts based on S via classical variational characterizations, so their rates directly translate into size and risk bounds for procedures built from induced order statistics.

The paper makes three main contributions. First, we provide general high-level results that map marginal approximation rates for the conditional law P_r to joint convergence rates for the IOS vector S_n . Specifically, under a mild Lipschitz condition on the density of X , we show how bounds on the *marginal* rates of the form

$$H(P_r, P) = O(r^{a_h}) \quad \text{and} \quad \text{TV}(P_r, P) = O(r^{a_{tv}}) \tag{1.1}$$

translate directly into *joint* convergence rates for the IOS vector S_n of the form

$$\begin{aligned} H(\mathcal{L}(S_n), \mathcal{L}(S)) &= O(k^{1/2}(k/n)^{a_h/d}) \\ \text{TV}(\mathcal{L}(S_n), \mathcal{L}(S)) &= O\left(\min\left\{k(k/n)^{a_{tv}/d}, k^{1/2}(k/n)^{a_h/d}\right\}\right), \end{aligned} \tag{1.2}$$

as $n \rightarrow \infty$ and $k = k_n$ grows with n . By separating the analysis into (i) a high-level result linking marginal and joint rates and (ii) primitive conditions that determine the marginal exponents (a_h, a_{tv}) , our approach isolates precisely when and how smoothness assumptions affect the behavior of IOS. In addition, the framework treats Hellinger and total variation distances on equal footing, allowing their distinct behaviors to be characterized transparently and under comparable assumptions.

Second, we develop primitive conditions that deliver the required marginal rates under QMD of the conditional densities at x_0 , which is a standard condition that accommodates both interior and boundary points. The supplementary appendix provides a complementary smoothness framework based on differentiability of f in x with a Hölder remainder, which delivers a wider range of marginal exponents (a_h, a_{tv}) and clarifies how rates slow down as smoothness weakens. A key distinction is that the joint Hellinger rate depends solely on the marginal exponent a_h , whereas the joint total variation rate is shaped by both a_h and a_{tv} through the minimum in (1.2). Under Hölder smoothness, $a_{tv} = 2a_h$, so total variation decays faster than Hellinger and this improvement carries over to the joint rates. Yet both metrics impose the same growth condition on k , so the faster total variation rate does not relax the admissible order of k .

Third, we formally compare our assumptions with those of Falk et al. (2010, Theorem 3.5.2) (FHR), which is a benchmark result in this setting. We identify the features of that condition that

generate the unusually fast rate $H(P_r, P) = O(r^2)$, and we show that these features impose strong structure on the joint density, including local exponential-family behavior and stringent restrictions on how the support of Y may vary with X . In contrast, our assumptions allow for substantially more flexible data-generating processes, including boundary points and other non-smooth features that are ruled out by the FHR conditions.

The end result is a unified set of conditions and rates that can serve as a toolkit for analyzing tests and estimators based on induced order statistics across a broad range of applications. We illustrate the relevance of our results through several examples. First, we revisit permutation-based inference in the regression discontinuity design of [Canay and Kamat \(2018\)](#) and show how our theory yields new asymptotic results for that framework. Existing large-sample approximations treat the dimension of S_n as fixed, and commonly used heuristics for selecting k rely on rate calculations that are not consistent with the findings established here. Our analysis provides formally valid growth conditions on k , clarifies how smoothness assumptions shape the behavior of IOS-based procedures, and offers guidance for constructing data-dependent rules for choosing k in practice. We then discuss applications to k -nearest-neighbor estimators and to distributionally robust optimization. In particular, the recent work of [Esteban-Pérez and Morales \(2022\)](#) on distributionally robust optimization relies on the regularity conditions of [Falk et al. \(2010\)](#) to control the discrepancy between empirical and true conditional laws, providing another example in which the tools developed here can be used to relax smoothness requirements and refine asymptotic guarantees. Across these settings, our results provide a unified framework for deriving asymptotic properties of IOS-based methods under transparent and verifiable smoothness conditions.

The remainder of the paper is organized as follows. [Section 2](#) introduces the main notation and revisits the existing result of [Falk et al. \(2010\)](#). [Section 3](#) presents the main formal results. [Section 3.1](#) covers the quadratic-mean differentiability setting, while the Taylor/Hölder-type smoothness framework is developed in the supplementary appendix (see [Appendix S.2](#)). [Section 4](#) illustrates how our results apply to several econometric and statistical problems, with a focus on the properties of permutation-based tests in the regression discontinuity design. [Section 5](#) examines the interplay between these assumptions, explaining why the conditions of [Falk et al. \(2010\)](#) yield faster convergence relative to the ones we present here. The proofs of the main theorems appear in [Appendix A](#). Finally, the supplementary appendix collects auxiliary lemmas, technical proofs, and additional examples. To keep the main text uncluttered, all results in the supplement are numbered with an “S.” prefix (e.g., [Lemma S.1](#), [Theorem S.2](#)).

2 Setup and Notation

Let $X \in \mathcal{X} \subseteq \mathbf{R}^d$ and $Y \in \mathcal{Y} \subseteq \mathbf{R}^m$, and let Q denote the distribution of (X, Y) , assumed to belong to a model class \mathbf{Q} . Assume Q is dominated by the product measure $d\mu := dx \otimes \nu(dy)$, where dx is Lebesgue measure on \mathbf{R}^d and ν is an arbitrary fixed σ -finite measure on \mathbf{R}^m . We denote the joint

density with respect to $d\mu$ by $f(x, y)$ and the marginal density of X by

$$g(x) := \int f(x, y) \nu(dy) .$$

We denote the conditional distribution of Y given $X = x_0$ by P , assumed to belong to a model class \mathbf{P} induced by \mathbf{Q} . For any $x \in \mathbf{R}^d$ with $g(x) > 0$, we denote the density of the conditional distribution of Y given $X = x$ by

$$p_x(y) := \frac{f(x, y)}{g(x)} .$$

This setup allows Y to be continuous, discrete, or mixed, while requiring that X admits a Lebesgue density.

Fix $x_0 \in \mathcal{X}$ and let $r > 0$ be given. We denote the Euclidean distance of X from x_0 by

$$R := \|X - x_0\| \tag{2.1}$$

and the open ball of radius r with center x_0 by

$$B_r := \{x \in \mathbf{R}^d : \|x - x_0\| < r\} .$$

We denote by P_r the distribution that determines the local conditional distributions of Y conditional on the event $X \in B_r$ and let

$$h_r(y) := \frac{\int_{B_r} f(x, y) dx}{\int_{B_r} g(x) dx} .$$

Under the smoothness conditions introduced in the next section, it follows that $h_r(y)$ approximates $p_{x_0}(y)$ as $r \rightarrow 0$. For notational simplicity, the dependence of B_r and $h_r(y)$ on x_0 is kept implicit, since the center x_0 remains fixed throughout the paper.

We observe an i.i.d. sample of size n from $Q \in \mathbf{Q}$ and denote it by

$$\{(X_i, Y_i) : 1 \leq i \leq n\} . \tag{2.2}$$

We also denote the order statistics of $\{R_i : 1 \leq i \leq n\}$ in (2.1) by $R_{(1)} \leq \dots \leq R_{(n)}$. Note that by our assumptions on \mathbf{Q} , ties occur with probability 0. Define the (random) permutation σ_n of $\{1, \dots, n\}$ by

$$\sigma_n(i) = j \iff R_i = R_{(j)} , \quad i, j = 1, \dots, n ,$$

so that $\sigma_n(i)$ is the rank of observation i . The index set of the k observations closest to x_0 is

$$K_n := \{i \in \{1, \dots, n\} : \sigma_n(i) \leq k\} .$$

Let ι_n be the increasing rearrangement of K_n (i.e., $\iota_n(1) < \dots < \iota_n(k)$ and $\{\iota_n(1), \dots, \iota_n(k)\} = K_n$). Finally, for $j = 1, \dots, k$, let $Y_{\iota_n(j)}$ be the response attached to the observation whose distance

from x_0 ranks among the k smallest, with the indices $\iota_n(1) < \dots < \iota_n(k)$ listed in the original sample order. In this paper, these are our *induced order statistics*, listed in original sample order (equivalently, a permutation of distance-ordered concomitants) (see, e.g., [David and Galambos \(1974\)](#); [Bhattacharya \(1974\)](#); [Canay and Kamat \(2018\)](#)).

Collect these *induced order statistics* into

$$S_n := (S_{n,1}, \dots, S_{n,k}) := (Y_{\iota_n(1)}, \dots, Y_{\iota_n(k)}) . \quad (2.3)$$

By definition, S_n is a k -tuple of outcomes of Y corresponding to those X 's that fall closest to x_0 . [Example 1](#) below provides an illustrative example. Under suitable smoothness conditions, one could use this sample to approximate the distribution of Y given $X = x_0$, which we previously denoted by P . This is important because many hypotheses and parameters of interest are formulated in terms of P . For these problems, the “ideal” random sample would be

$$S := (S_1, \dots, S_k) := (\tilde{Y}_1, \dots, \tilde{Y}_k), \quad \text{with } \tilde{Y}_1, \dots, \tilde{Y}_k \text{ i.i.d. with } P . \quad (2.4)$$

In such settings, S_n can serve as an “approximation” of S . [Section 4](#) gives detailed examples.

Example 1. Let $m = d = 1$, $n = 10$, $k = 4$, and suppose $x_0 = 0$. Consider the hypothetical sample

$$\{(-3, 2), (4, -1), (-1, 0), (2, 5), (0.5, 1), (-0.2, 3), (7, 4), (-5, -2), (1.4, 6), (-0.7, 8)\} .$$

In this case $R_i = |X_i|$ and so the rank map σ_n is given by

$$\begin{array}{c|cccccccccc} i & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ \hline \sigma_n(i) & 7 & 8 & 4 & 6 & 2 & 1 & 10 & 9 & 5 & 3 \end{array} .$$

Since $k = 4$, the index set of the k nearest observations is

$$K_n = \{i : \sigma_n(i) \leq 4\} = \{6, 5, 10, 3\} .$$

Applying the re-arrangement determined by ι_n we get $\iota_n = (3, 5, 6, 10)$. Hence the induced order statistics are the Y -values attached to these indices, listed in the original sample order:

$$S_n = (Y_{\iota_n(1)}, Y_{\iota_n(2)}, Y_{\iota_n(3)}, Y_{\iota_n(4)}) = (Y_3, Y_5, Y_6, Y_{10}) = (0, 1, 3, 8) .$$

■

For hypothesis testing, let $\phi : \mathbf{R}^k \rightarrow [0, 1]$ be a measurable test functional. Working with the infeasible test $\phi(S)$ would be ideal, but in practice, we resort to $\phi(S_n)$ instead. Since $\phi \leq 1$, the variational characterization of total variation implies

$$|E_Q[\phi(S_n)] - E_Q[\phi(S)]| \leq \text{TV}(\mathcal{L}(S_n), \mathcal{L}(S)) , \quad (2.5)$$

where $\mathcal{L}(\cdot)$ denotes the law of a random vector and TV denotes the total variation distance. Moreover, the standard comparison between total variation and the Hellinger distance H yields

$$\mathbb{H}^2(\mathcal{L}(S_n), \mathcal{L}(S)) \leq \text{TV}(\mathcal{L}(S_n), \mathcal{L}(S)) \leq \sqrt{2} \mathbb{H}(\mathcal{L}(S_n), \mathcal{L}(S)) . \quad (2.6)$$

Thus, any rate for either distance immediately delivers a rate for the size error in (2.5).

Example 2. Consider the setting in Example 1, where $k = 4$ and $S_n = (0, 1, 3, 8)$. In the setting of Section 4, the Cramér–von Mises (CvM) statistic compares two subsamples: the first corresponds to the first $k/2$ components of S_n , namely $(S_{n,1}, S_{n,2}) = (0, 1)$, and the second to the last $k/2$ components, $(S_{n,3}, S_{n,4}) = (3, 8)$. The associated empirical CDFs are

$$\hat{F}_n^-(t) = \frac{1}{2}(I\{0 \leq t\} + I\{1 \leq t\}) , \quad \text{and} \quad \hat{F}_n^+(t) = \frac{1}{2}(I\{3 \leq t\} + I\{8 \leq t\}) .$$

The CvM test statistic is therefore

$$T(S_n) = \frac{1}{k} \sum_{j=1}^k (\hat{F}_n^-(S_{n,j}) - \hat{F}_n^+(S_{n,j}))^2 = \frac{1}{4} \left(\left(\frac{1}{2}\right)^2 + 1^2 + \left(\frac{1}{2}\right)^2 + 0^2 \right) = \frac{3}{8} . \quad (2.7)$$

■

For estimation, consider an IOS–based estimator $\psi(S_n)$ of a scalar parameter $\theta(P)$, where $\theta(P)$ is defined as a bounded functional of the conditional law P . Examples include the conditional CDF or the conditional mean under $\|Y\| \leq B$ for some constant B . Then

$$E_Q[(\psi(S_n) - \theta(P))^2] \leq 2E[(\psi(\tilde{S}_n) - \psi(\tilde{S}))^2] + 2E_P[(\psi(S) - \theta(P))^2] ,$$

where the expectation in the first term is taken with respect to any joint distribution of (\tilde{S}_n, \tilde{S}) whose marginals satisfy $\tilde{S}_n \sim \mathcal{L}(S_n)$ and $\tilde{S} \sim \mathcal{L}(S)$. The second term is the standard k –sample risk of the infeasible estimator $\psi(S)$ based on k i.i.d. draws from P , and is typically of order $O(1/k)$ for sample means or CDF estimators under mild moment conditions. Under the assumption that $|\psi| \leq B$, a maximal coupling yields

$$EE[(\psi(\tilde{S}_n) - \psi(\tilde{S}))^2] \leq 4B^2 \text{TV}(\mathcal{L}(S_n), \mathcal{L}(S)) . \quad (2.8)$$

Thus, any convergence rate for $\mathbb{H}(\mathcal{L}(S_n), \mathcal{L}(S))$ or $\text{TV}(\mathcal{L}(S_n), \mathcal{L}(S))$ yields an immediate, uniform bound on the mean squared error of $\psi(S_n)$.

Example 3. In the same setting as Example 1 with $k = 4$ and $S_n = (0, 1, 3, 8)$, take the parameter of interest to be the conditional mean $\theta(P) = E_P[Y] = E_Q[Y \mid X = 0]$. The natural estimator based

on k i.i.d. draws from P is $\psi(S) = \frac{1}{k} \sum_{j=1}^k S_j$; the feasible IOS-based estimator is

$$\psi(S_n) = \frac{1}{k} \sum_{j=1}^k S_{n,j} = \frac{1}{4}(0 + 1 + 3 + 8) = 3 .$$

■

Taken together, these two applications demonstrate that the convergence rates developed in this paper provide generic control over the performance of IOS-based procedures, whether for hypothesis testing or estimation. The monograph [Falk et al. \(2010\)](#) provides an influential benchmark for the convergence of induced order statistics in Hellinger distance, obtained under a relatively strong smoothness condition tailored to their framework. Before presenting our results, we briefly summarize the key result in [Falk et al. \(2010\)](#) and its main assumption.

2.1 Existing results

[Falk et al. \(2010\)](#) provide an approximation result for S_n by deriving a rate of convergence for $H(\mathcal{L}(S_n), \mathcal{L}(S))$. The intuition of this result goes as follows. Since R has a continuous CDF, it follows from [Kaufmann and Reiss \(1992, Theorem 1\)](#) that, conditional on $R_{(k+1)} = r$, the vector S_n consists of k i.i.d. draws from P_r :

$$\mathcal{L}(S_n \mid R_{(k+1)} = r) = P_r^k .$$

For small $r > 0$, P_r approximates the target conditional law P , so unconditionally we expect

$$\mathcal{L}(S_n) = \int P_r^k \mathcal{L}(R_{(k+1)})(dr) \approx \int P^k \mathcal{L}(R_{(k+1)})(dr) = P^k = \mathcal{L}(S) . \quad (2.9)$$

In words: the Y -values attached to the k nearest X -neighbors of x_0 behave approximately like k independent draws from P . In [Falk et al. \(2010, Theorem 3.5.2\)](#), a formal proof of this approximation is given under the assumption stated below. For completeness, we reproduce their theorem after the assumption.

Assumption 1 (FHR assumption). *The model class \mathbf{Q} consists of all laws Q for which there exists $t_0 > 0$ such that, uniformly for $t \in \mathbf{R}^d$ with $\|t\| \leq t_0$ and all $y \in \mathbf{R}^m$,*

$$f(x_0 + t, y) = f(x_0, y) \left\{ 1 + t^\top \zeta_1(y) + O(\|t\|^2 \zeta_2(y)) \right\} , \quad (2.10)$$

where $\zeta_1 : \mathbf{R}^m \rightarrow \mathbf{R}^d$, $\zeta_2 : \mathbf{R}^m \rightarrow [0, \infty)$, and

$$\int (\|\zeta_1(y)\|^2 + \zeta_2(y)^2) f(x_0, y) \nu(dy) < \infty . \quad (2.11)$$

Theorem 1 (Falk–Hüsler–Reiss, Theorem 3.5.2). *Suppose $g(x_0) > 0$ and Assumption 1 holds. Then, uniformly over $n \in \mathbf{N}$ and $k \in \{1, \dots, n\}$,*

$$H(\mathcal{L}(S_n), \mathcal{L}(S)) = O(k^{1/2}(k/n)^{2/d}) .$$

Theorem 1 implies that any procedure based on S_n behaves asymptotically like the corresponding one based on the i.i.d. vector $S = (S_1, \dots, S_k)$ in (2.4) with common law P . While powerful, this result rests critically on Assumption 1. That assumption imposes substantive restrictions on the joint density $f(x, y)$ and is often too strong for the applications we study. First, it rules out boundary points: the uniform expansion (2.10) must hold for all t in a full neighborhood of the origin, so x_0 is necessarily in the interior of \mathcal{X} . Regression discontinuity designs, however, condition on a known cutoff of the running variable and perform local analysis separately on observations just below and just above that cutoff, so that x_0 is a boundary point of the support of the running variable in each subsample; see Section 4 for more details. Second, Assumption 1 forces the local behavior of $f(x, y)$ in x to take an exponential tilt form in y , so that the model is locally indistinguishable from an exponential family; see Section 5. Third, it requires that the zero-density set $\{y : f(x_0, y) = 0\}$ remain exactly invariant under local perturbations of x , which excludes many models in which the conditional support of Y varies with X . Finally, Theorem 1 yields a rate for $\text{TV}(\mathcal{L}(S_n), \mathcal{L}(S))$ only indirectly via (2.6), and the resulting bound may not be sharp for total variation. These considerations motivate the next section, where we develop convergence rates for both $H(\mathcal{L}(S_n), \mathcal{L}(S))$ and $\text{TV}(\mathcal{L}(S_n), \mathcal{L}(S))$ under weaker primitive conditions that accommodate both interior and boundary points. In the Supplemental Appendix, we also make explicit the trade-off between smoothness and speed of convergence, including regimes in which the rate deteriorates and thresholds beyond which convergence fails.

3 Rates of Convergence of Induced Ordered Statistics

In this section, we study new, weaker conditions under which the laws of the induced order statistics converge to the “ideal” limit experiment. Our rates for $H(\mathcal{L}(S_n), \mathcal{L}(S))$ and $\text{TV}(\mathcal{L}(S_n), \mathcal{L}(S))$ hinge on the *marginal* approximation errors $H(P_r, P)$ and $\text{TV}(P_r, P)$, where P_r is the conditional law of Y given $X \in B_r$ and P is the conditional law of Y given $X = x_0$. Importantly, we allow x_0 to be either an interior point or a boundary point of \mathcal{X} . Throughout the paper, we say that x_0 is an interior point of \mathcal{X} if there exists $\tilde{r} > 0$ such that $B_r \subset \mathcal{X}$ for all $r \in (0, \tilde{r}]$, and we say that x_0 is a boundary point of \mathcal{X} otherwise. This distinction is required for accommodating applications in regression discontinuity designs.

Our analysis proceeds in two steps. We begin with a high-level result that translates bounds on the marginal rate of convergence of P_r to P , together with a mild continuity condition on the density g of X , into bounds on the joint rate of convergence of $\mathcal{L}(S_n)$ to $\mathcal{L}(S)$. We then develop primitive conditions that deliver the required marginal rates under quadratic mean differentiability (QMD)

of the conditional laws at x_0 (see Section 3.1). A complementary smoothness framework based on Taylor/Hölder-type differentiability with explicit remainder control is deferred to the Supplemental Appendix (see Section S.2 therein). This separation clarifies the key ways in which our assumptions differ from, and weaken, the original condition of Falk et al. (2010, Theorem 3.5.2).

We begin with a mild regularity condition on the marginal density g at x_0 .

Assumption 2 (Local regularity of g at x_0). *The model class \mathbf{Q} consists of all laws Q such that:*

(i) *The marginal density g of X satisfies $g(x_0) > 0$ and*

$$|g(x) - g(x_0)| \leq C_g \|x - x_0\|$$

for all $x \in \mathcal{X}$ in a neighborhood of x_0 and some constant $C_g < \infty$.

(ii) *There exist constants $C_{x_0} > 0$ and $r_0 > 0$ such that*

$$\text{Vol}(B_r \cap \mathcal{X}) \geq C_{x_0} r^d \quad \text{for all } 0 < r < r_0 ,$$

where Vol denotes volume.

Continuity of the distribution of X together with positivity $g(x_0) > 0$ are standard primitives in the induced order statistics literature; see, for example, Reiss (1989), Kaufmann and Reiss (1992), and Falk et al. (2010). We strengthen these conditions slightly by imposing local Lipschitz regularity of g at x_0 and a local thickness condition on $B_r \cap \mathcal{X}$, a formulation that remains compatible with x_0 being a boundary point.

Although Assumption 2 may appear unusual at first glance, it is quite mild once one distinguishes between interior and boundary points. If x_0 is an interior point of \mathcal{X} , then for all sufficiently small r we have $B_r \subset \mathcal{X}$, hence $\text{Vol}(B_r \cap \mathcal{X}) = \text{Vol}(B_r)$ is of order r^d , so condition (ii) holds automatically. In this case, Assumption 2 reduces to condition (i), which imposes local Lipschitz regularity of g at x_0 . If x_0 is a boundary point of \mathcal{X} , condition (ii) simply requires that \mathcal{X} occupy a nonvanishing fraction of each small ball centered at x_0 , a property that holds, for example, when \mathcal{X} has a Lipschitz boundary near x_0 . Condition (i) remains a one-sided Lipschitz condition on g at x_0 . As mentioned earlier, allowing for such boundary points is essential for regression discontinuity design applications.

Finally, as discussed in Section 5, Assumption 1 implies Assumption 2 with x_0 necessarily being an interior point in the support \mathcal{X} , so the result below introduces no additional restrictions beyond those already present in Theorem 1.

Theorem 2. *Let Assumption 2 hold. Suppose that $H(P_r, P) = O(r^{a_h})$ for some $a_h \geq 0$ and $\text{TV}(P_r, P) = O(r^{a_{tv}})$ for some $a_{tv} \geq 0$. Then the following bounds hold uniformly for $n \in \mathbf{N}$ and $k \in \{1, 2, \dots, n\}$:*

$$(a) \text{ H}(\mathcal{L}(S_n), \mathcal{L}(S)) = O(k^{1/2}(k/n)^{a_h/d}).$$

$$(b) \text{ TV}(\mathcal{L}(S_n), \mathcal{L}(S)) = O\left(\min\left\{k(k/n)^{a_{tv}/d}, k^{1/2}(k/n)^{a_h/d}\right\}\right).$$

The proof of this result parallels that of Theorem 1, but differs in several important respects. First, Theorem 2 is formulated directly in terms of the *marginal* convergence rates $\text{H}(P_r, P) = O(r^{a_h})$ and $\text{TV}(P_r, P) = O(r^{a_{tv}})$, making explicit how smoothness of the conditional distribution of Y near x_0 feeds into the *joint* rates of convergence of the induced order statistics. Second, the theorem provides an explicit rate for $\text{TV}(\mathcal{L}(S_n), \mathcal{L}(S))$ alongside the Hellinger bound, thereby allowing the two metrics to be compared on equal footing. Third, and more subtly, the joint total variation rate depends on *both* marginal exponents a_h and a_{tv} . This reflects a structural bottleneck: the total variation rate can be controlled either directly through the marginal total variation discrepancy, or indirectly via the inequality $\text{TV} \leq \sqrt{2}\text{H}$ in (2.6). Consequently, even if the marginal total variation distance decays very rapidly (a_{tv} is large), this does not relax the growth restriction on (k) needed for joint total variation convergence beyond what is already imposed by the Hellinger channel. In contrast, the joint Hellinger rate depends only on the marginal exponent a_h and cannot be improved via the inequality $\text{H}^2 \leq \text{TV}$. This distinction is structural and follows from the exact tensorization of the Hellinger affinity over product measures, which implies that joint Hellinger convergence is fully determined by marginal Hellinger rates. Consequently, improvements in marginal total variation smoothness beyond what is implied by a_h do not translate into faster joint Hellinger convergence. The primitive conditions developed in this paper determine when the exponents a_h and a_{tv} equal one, two, or other values, thereby tracing precisely how marginal smoothness translates into joint convergence rates for the induced order statistics.

We note that under Assumption 1 and $g(x_0) > 0$, Lemma S.5 implies $\text{H}(P_r, P) = O(r^2)$ and $\text{TV}(P_r, P) = O(r^2)$. In this case, part (a) of Theorem 2 reproduces exactly the rate in Theorem 1, illustrating that the two results are consistent.

3.1 Differentiability in Quadratic Mean

In this section, we develop marginal approximation rates for the conditional densities p_x at x_0 under QMD. We show that QMD yields linear marginal convergence of P_r to P in both Hellinger and total variation distances. We also study sharpness and the limits of possible improvements over the linear rate, distinguishing what can and cannot be strengthened at boundary points versus interior points. These marginal results feed into joint convergence bounds for $\mathcal{L}(S_n)$ through Theorem 2.

Assumption 3 (QMD at x_0). *The model class \mathbf{Q} consists of all laws Q for which there exists a measurable score $\dot{\ell}_{x_0} : \mathbf{R}^m \rightarrow \mathbf{R}^d$ with*

$$\int \|\dot{\ell}_{x_0}(y)\|^2 p_{x_0}(y) \nu(dy) < \infty ,$$

such that, as $t \rightarrow 0$ in \mathbf{R}^d and, for all $x_0 + t \in \mathcal{X}$,

$$\int \left[\sqrt{p_{x_0+t}(y)} - \sqrt{p_{x_0}(y)} - \frac{1}{2} (t^\top \dot{\ell}_{x_0}(y)) \sqrt{p_{x_0}(y)} \right]^2 \nu(dy) = o(\|t\|^2) .$$

QMD is a central smoothness condition in asymptotic statistics. It underlies the local asymptotic normality (LAN) framework and plays a key role in the asymptotic theory of likelihood-based procedures (see, e.g., [Le Cam \(1986\)](#), [Van der Vaart \(2000\)](#)). Beyond likelihood theory, QMD also supports diffusion/LAN approximations in statistical decision problems, such as optimal design and inference for adaptive experiments and bandits, by justifying Gaussian limit experiments and the associated risk analyses; [Adusumilli \(2025\)](#).

Lemma [S.6](#) in the supplemental appendix shows that Assumption [1](#) implies Assumption [3](#), but not conversely. Section [5](#) provides a detailed comparison of the two sets of primitive conditions, highlighting the distinct restrictions they impose on \mathbf{Q} and the implications for the behavior of P_r . The next theorem gives the linear marginal rates under QMD and characterizes sharpness and the limits of uniform improvements.

Theorem 3. *Suppose Assumptions [2–3](#) hold. Then,*

$$\mathbf{H}(P_r, P) = O(r) \quad \text{and} \quad \text{TV}(P_r, P) = O(r) . \tag{3.1}$$

Moreover:

(i) *Boundary sharpness. If x_0 lies on the boundary of \mathcal{X} , then the rate in [\(3.1\)](#) is sharp in the sense that there exist a data-generating process $Q \in \mathbf{Q}$ and constants $C > 0$ and $\bar{r} > 0$ such that*

$$\mathbf{H}(P_r, P) \geq Cr \quad \text{and} \quad \text{TV}(P_r, P) \geq Cr \quad \text{for all } r \in (0, \bar{r}) .$$

(ii) *Interior points: no uniform polynomial improvement. For any $\tilde{r} > 0$, let $\mathbf{Q}^o(\tilde{r})$ be the set of all distributions Q that satisfy Assumptions [2](#) and [3](#), and for which $B_r \cap \mathcal{X} = B_r$ for all $r \in (0, \tilde{r})$. Then, for any $\tilde{r} > 0$ and $\varepsilon > 0$, there do not exist constants $C < \infty$ and $\bar{r} \in (0, \tilde{r})$ such that*

$$\sup_{Q \in \mathbf{Q}^o(\tilde{r})} \mathbf{H}(P_r, P) \leq Cr^{1+\varepsilon} \quad \text{for all } r \in (0, \bar{r}) , \tag{3.2}$$

and there do not exist constants $C < \infty$ and $\bar{r} \in (0, \tilde{r})$ such that

$$\sup_{Q \in \mathbf{Q}^o(\tilde{r})} \text{TV}(P_r, P) \leq Cr^{1+\varepsilon} \quad \text{for all } r \in (0, \bar{r}) . \tag{3.3}$$

Theorem [3](#) implies that under QMD the marginal discrepancy is of order $O(r)$ in both $\mathbf{H}(P_r, P)$ and $\text{TV}(P_r, P)$, and this order does not hinge on whether x_0 is an interior point or a boundary point of \mathcal{X} . Because sharpness is established via a boundary-point construction, one might suspect that

interior points deliver uniformly faster rates. The theorem clarifies the limits of this intuition: while for interior points allow for $\text{TV}(P_r, P) = o(r)$, no polynomial improvement over the $O(r)$ bound can be obtained uniformly over the QMD model class. This no-polynomial-improvement conclusion also applies to the Hellinger distance, though, in that case, we do not obtain an analogous $o(r)$ refinement.

Combining (3.1) in Theorem 3 with Theorem 2 yields

$$\text{H}(\mathcal{L}(S_n), \mathcal{L}(S)) = O(k^{1/2}(k/n)^{1/d}), \quad \text{TV}(\mathcal{L}(S_n), \mathcal{L}(S)) = O(k^{1/2}(k/n)^{1/d}).$$

Here $a_h = a_{tv} = 1$, so the relevant bound for $\text{TV}(\mathcal{L}(S_n), \mathcal{L}(S))$ is provided by the inequality $\text{TV} \leq \sqrt{2}\text{H}$, and both distances share the same joint rate. This rate directly determines how fast k may grow relative to n while still ensuring convergence of $\mathcal{L}(S_n)$ to $\mathcal{L}(S)$:

$$k^{1/2}(k/n)^{1/d} \rightarrow 0 \quad \iff \quad k = o(n^{2/(2+d)}).$$

In particular, for $d = 1$ this becomes $k = o(n^{2/3})$. This threshold coincides with the growth conditions derived, through different arguments, in Bugni and Canay (2021) and Bugni et al. (2025) for the specific uses of induced order statistics in those papers.

Theorem 3 delivers a simple benchmark rate. While we view this result as clean and widely applicable, it does not describe how stronger or weaker smoothness assumptions translate into faster or slower rates. Appendix S.2 provides complementary results. Under the Taylor/Hölder remainder condition of the joint density in Assumption S.4, Theorem S.1 delivers a range of marginal rates through explicit exponents (a_h, a_{tv}) that depend on the smoothness order $\kappa_s + \kappa_r$. As $\kappa_s + \kappa_r$ decreases, these exponents decrease, so the marginal convergence slows down, and the joint IOS approximation becomes more demanding through Theorem 2. For example, in the Hölder continuity case $\kappa_s = 0$, the theorem gives $\text{H}(P_r, P) = O(r^{\kappa_r/2})$ and $\text{TV}(P_r, P) = O(r^{\kappa_r})$, so the rate can be arbitrarily slow as $\kappa_r \downarrow 0$. In such cases, IOS-based approximations can break down regardless of the speed at which k grows.

Remark 1. *Theorem 3 is our default marginal rate result. It applies under a standard condition and delivers the same $O(r)$ order for $\text{H}(P_r, P)$ and $\text{TV}(P_r, P)$ whether x_0 is an interior point or a boundary point of \mathcal{X} , which is the relevant case in our regression discontinuity applications. Theorem S.1 is included as a complementary result in the supplementary appendix. Under Assumption S.4, the marginal exponents (a_h, a_{tv}) depend on the geometry term η and can be smaller at boundary points than at interior points; thus, when boundary behavior is central, this framework can yield slower marginal rates than QMD. At the same time, Assumptions 3 and S.4 are non-nested, so the Taylor/Hölder result is not presented as imposing weaker restrictions on \mathbf{Q} . Rather, it makes explicit how smoothness controls marginal rates and clarifies when convergence fails altogether. ■*

4 Applications

We start this section by describing how the general framework developed in this paper applies to the permutation test based on induced order statistics proposed in [Canay and Kamat \(2018\)](#). That paper considers a sharp regression discontinuity design (RDD) in which treatment is assigned when a real-valued running variable X crosses a known cutoff (normalized to zero), and focuses on a popular testable implication: that the distribution of baseline covariates is continuous at the cutoff. Formally, their null hypothesis is that the conditional law of a baseline covariate Y is the same when approaching the cutoff from the left and from the right,

$$H_0 : \mathcal{L}(Y | X = 0^-) = \mathcal{L}(Y | X = 0^+) ,$$

which is the distributional analogue of the usual continuity-of-means checks routinely reported in empirical applications of RDD. Here, 0^+ denotes the limit as $x \downarrow 0$ and 0^- the limit as $x \uparrow 0$ of the conditional laws $\mathcal{L}(Y | X = x)$. Under H_0 , the Y observations whose running variable lies closest to the cutoff on either side should be approximately exchangeable, and [Canay and Kamat \(2018\)](#) exploit this by constructing a permutation test based on induced order statistics formed by the q nearest neighbors to the left and the q nearest neighbors to the right of the cutoff.

In the notation of this paper, the $2q$ induced order statistics used by [Canay and Kamat \(2018\)](#) correspond to the special case $x_0 = 0$ and $k = 2q$, where we separately order the observations with $X_i < 0$ and $X_i > 0$ by their distance to the cutoff. Let $\iota_n^-(1), \dots, \iota_n^-(q)$ denote the indices of the q observations with $X_i < 0$ whose values are closest to 0, listed in the original sample order; and similarly let $\iota_n^+(1), \dots, \iota_n^+(q)$ denote the indices of the q observations with $X_i > 0$ closest to 0. The induced order statistics used in the test can then be written as

$$S_n = (S_{n,1}, \dots, S_{n,k}) = (Y_{\iota_n^-(1)}, \dots, Y_{\iota_n^-(q)}, Y_{\iota_n^+(1)}, \dots, Y_{\iota_n^+(q)}) .$$

The test statistic in [Canay and Kamat \(2018\)](#) is the Cramér–von Mises functional $T(S_n)$ in [\(2.7\)](#) that compares the empirical distribution functions constructed from the left and right subsamples. Critical values are obtained via permutation, based on the permuted vectors $S_n^\pi = (S_{n,\pi(1)}, \dots, S_{n,\pi(k)})$. We denote their resulting test by $\phi(S_n)$.

[Canay and Kamat \(2018\)](#) establish the asymptotic validity of their approximate permutation test under an asymptotic framework in which k is held fixed as $n \rightarrow \infty$, and they do not analyze regimes in which k grows with n . Since $E_Q[\phi(S)] \leq \alpha$ under their assumptions, the bound in [\(2.5\)](#), together with [Theorems 2 and 3](#) (applied separately to the left and right blocks of size q), and the fact that $d = 1$, yield

$$E_Q[\phi(S_n)] \leq \alpha + \text{TV}(\mathcal{L}(S_n), \mathcal{L}(S)) = \alpha + O(k^{3/2}/n) ,$$

for $k = 2q$. Note that in this application, the conditioning points $x_0 = 0^-$ and $x_0 = 0^+$ lie on the

boundary of the support of X in the corresponding subsamples. This feature, which is intrinsic to RDDs, underscores the importance of Assumption 2 for allowing boundary points. It follows that their permutation test remains asymptotically valid provided

$$q = o(n^{2/3}) , \tag{4.1}$$

recalling that $k = 2q$ in their notation. Hence, the machinery developed in this paper delivers a simple and transparent characterization of how quickly q may grow while maintaining asymptotic size control for the test in Canay and Kamat (2018).

Remark 2. *The rule of thumb proposed in Canay and Kamat (2018) selects q according to the rate $n^{0.9}/\log n$. This choice was motivated by the asymptotic framework in that paper (where q is fixed as $n \rightarrow \infty$) and by a heuristic argument based on the bivariate normal model for (Y, X) . The results developed here clarify how this rule should be modified when one allows q to grow with n . In particular, under Assumptions 2–3, Theorem 3 implies that asymptotic size control requires the rate in (4.1). Consequently, the rule of thumb in Canay and Kamat (2018) does not satisfy the rate restriction needed for validity when q diverges. More importantly, the rate $n^{0.9}$ is not attainable under any of the smoothness regimes considered here, including the alternative smoothness regimes considered in the supplemental appendix. A feasible data-dependent rule may therefore be constructed by replacing the exponent 0.9 with any $\gamma < 2/3$ while retaining the same normalization and constants used in the original paper; that is,*

$$q_{\text{rot}} = C_{\text{ck}} n^\gamma ,$$

where C_{ck} denotes the constant proposed in Canay and Kamat (2018). ■

Remark 3. *Beyond Canay and Kamat (2018), several related papers employ IOS-based procedures to test continuity of a density in RDDs (Bugni and Canay, 2021), to test discontinuities in event studies in (Bugni et al., 2023), or to test for conditional stochastic dominance in RDD and related settings (Bugni et al., 2025; Goldman and Kaplan, 2018). While we do not pursue these applications in detail, our results apply directly to the analysis of such tests and clarify the assumptions required to ensure their validity when the number of induced order statistics k is allowed to grow with the sample size n . ■*

We next illustrate how our main results can be used to justify normal approximations for a broad class of IOS-based estimators (also known as k -nearest-neighbor estimators). Let P denote the conditional law of Y given $X = x_0$, and let $\theta(P) \in \mathbf{R}^\ell$ be the target parameter of interest. Examples include mean-type functionals $\theta(P) = E_P[Y]$ and quantile-type functionals defined by $P(Y \leq \theta(P)) = q$ for a fixed $q \in (0, 1)$. We consider IOS-based estimators of $\theta(P)$ of the form $\psi(S_n)$, where $S_n = (S_{n,1}, \dots, S_{n,k})$ denotes the k induced order statistics and where $\psi : \mathbf{R}^k \rightarrow \mathbf{R}^\ell$ is a Borel-measurable map. Let $S = (S_1, \dots, S_k)$ be the infeasible benchmark consisting of k i.i.d.

draws from P . Under suitable conditions, standard normal approximation arguments imply that

$$\sup_{t \in \mathbf{R}^\ell} \left| P^k(\sqrt{k}(\psi(S) - \theta(P)) \leq t) - \Phi_\Sigma(t) \right| = O(k^{-1/2})$$

where $\Phi_\Sigma(t)$ is the normal distribution function with mean zero and covariance matrix Σ .

Combining the i.i.d. normal approximation for the infeasible benchmark S with a total variation bound yields asymptotic normality of the IOS-based estimator $\psi(S_n)$. To see this, for each $t \in \mathbf{R}^\ell$, let $A_t = (-\infty, t/\sqrt{k} + \theta(P)]$ with the inequality interpreted componentwise. Since ψ is Borel measurable,

$$\left| Q^n(S_n \in \psi^{-1}(A_t)) - P^k(S \in \psi^{-1}(A_t)) \right| \leq \text{TV}(\mathcal{L}(S_n), \mathcal{L}(S))$$

where Q is the law of (X, Y) . Taking the supremum over t and using the $O(k^{-1/2})$ normal approximation for $\psi(S)$ therefore gives

$$\sup_{t \in \mathbf{R}^\ell} \left| Q^n(\sqrt{k}(\psi(S_n) - \theta(P)) \leq t) - \Phi_\Sigma(t) \right| \leq \text{TV}(\mathcal{L}(S_n), \mathcal{L}(S)) + O(k^{-1/2}).$$

Consequently, $\psi(S_n)$ has the same asymptotic normal limit as the infeasible i.i.d. benchmark provided the total variation distance $\text{TV}(\mathcal{L}(S_n), \mathcal{L}(S))$ vanishes as n grows. Under Assumptions 2–3, $\psi(S_n)$ is asymptotically normal provided

$$k = o\left(n^{\frac{2}{d+2}}\right).$$

While our analysis is motivated by IOS-based procedures, the results also speak to settings in which local conditional laws are approximated using shrinking neighborhoods around a conditioning point. One example arises in conditional stochastic optimization with side information, where a decision maker seeks to choose an action $z \in \mathcal{Z}$ (e.g., an order quantity, a portfolio allocation, or a policy parameter) that minimizes the conditional expected cost

$$\inf_{z \in \mathcal{Z}} E_Q[c(z, Y, X) \mid X = x_0] = \inf_{z \in \mathcal{Z}} E_P[c(z, Y, x_0)],$$

with P denoting the conditional law of Y given $X = x_0$. A recent contribution by [Esteban-Pérez and Morales \(2022\)](#) studies this problem using a distributionally robust formulation, in which decisions are chosen to be robust against misspecification of the conditional law within a neighborhood of an empirical approximation constructed from observations with X close to x_0 . Two key tuning parameters in their approach are a trimming fraction α_n and a tolerance radius ρ_n , which controls the size of the distributional neighborhood over which robustness is imposed. The feasibility of their robust optimization problem, and hence the validity of their finite-sample guarantee, requires ρ_n to decay at a rate that is sufficiently slow relative to the discrepancy between the local empirical conditional law and the target law P . A key ingredient in their analysis is a bound on this

discrepancy, measured in Hellinger distance. Their argument implies that if $H(P_r, P) = O(r^a)$ and $r \asymp \alpha_n^{1/m}$, then feasibility requires ρ_n not to decay faster than $\rho_n \gtrsim \alpha_n^{\min\{1, a\}/m}$ (up to lower-order terms). Under the smoothness condition of [Falk et al. \(2010\)](#) one has $a = 2$, while under [Assumption 3](#) one has $a = 1$, so in both cases $\min\{1, a\} = 1$ and the feasibility lower bound scales as $\rho_n \gtrsim \alpha_n^{1/m}$. Thus, QMD preserves the same admissible scaling for ρ_n as the classical framework while requiring substantially weaker local structure (and allowing boundary points). A genuinely different scaling for ρ_n arises only under rougher regimes with $a < 1$, as in the alternative smoothness conditions in the [Supplemental Appendix \(Section S.2\)](#).

Taken together, these examples, RDDs, k -nearest-neighbor methods, and distributionally robust optimization illustrate how our results apply to both IOS-based procedures and to a broader class of problems that rely on local conditional distributions.

5 Understanding the Assumptions

In this section, we take a closer look at the structure of the assumptions introduced earlier, with the goal of isolating the features that determine the achievable rates of convergence. Notably, [Assumption 1](#) in [Falk et al. \(2010\)](#) implies the approximation error $H(P_r, P) = O(r^2)$, a rate that is substantially faster than the feasible rates delivered by [Theorem 3](#), which are sharp within the class of models satisfying its assumptions. Here, we highlight the key factors that determine this discrepancy in rates.

We start with a closer look at [Assumption 1](#). For simplicity, let $d = 1$. Fix y and note that [\(2.10\)](#) implies that

$$\frac{f(x_0 + t, y) - f(x_0, y)}{t} = f(x_0, y) \left\{ \zeta_1(y) + \frac{O(t^2 \zeta_2(y))}{t} \right\} .$$

As $t \rightarrow 0$, the remainder term satisfies $O(t \zeta_2(y)) \rightarrow 0$, so taking the limit yields

$$\frac{\partial f(x_0, y)}{\partial x} = f(x_0, y) \zeta_1(y) \quad \text{and} \quad \frac{\partial}{\partial x} \log f(x_0, y) = \zeta_1(y) .$$

Thus, wherever $f(x_0, y) > 0$, [Assumption 1](#) pins down the local score at x_0 as a function of y alone. Rewriting the expansion more explicitly,

$$\begin{aligned} f(x_0 + t, y) &= f(x_0, y) \exp \left\{ \log \left(1 + t \zeta_1(y) + O(t^2 \zeta_2(y)) \right) \right\} \\ &\stackrel{(1)}{=} f(x_0, y) \exp \left\{ t \zeta_1(y) + O(t^2 \zeta_2(y)) + O(t^2 \zeta_1^2(y)) \right\} , \end{aligned}$$

where (1) holds by the expansion $\log(1 + u) = u + O(u^2)$. Dropping second-order terms gives the leading approximation

$$f(x_0 + t, y) \approx f(x_0, y) \exp\{t \zeta_1(y)\} ,$$

which has the exponential tilt form in y with statistic $\zeta_1(y)$. Assumption 1 thus forces the local joint density to evolve in x *as if* through an exponential tilt in y , making the model locally indistinguishable, to first order, from the exponential family generated by $\zeta_1(y)$.

There are two immediate consequences of this analysis. First, Assumption 1 rules out the possibility that x_0 is a boundary point of the support of X . This is because the uniform expansion in (2.10) must hold for all perturbations t in a full neighborhood of x_0 , which implies that $B_r \subset \mathcal{X}$ for all sufficiently small r . Second, Assumption 1 imposes a local invariance of the zero-density region,

$$\mathcal{N} := \{y \in \mathcal{Y} : f(x_0, y) = 0\} , \tag{5.1}$$

where \mathcal{Y} denotes the unconditional support of Y . If $f(x_0, y) = 0$, then $f(x_0 + t, y) = 0$ for all sufficiently small t . Consequently, any model in which new y -values enter the support of $f(x, \cdot)$ as x moves locally away from x_0 violates Assumption 1, even if it satisfies weaker conditions such as QMD or the alternative smoothness conditions we discuss in the Supplemental Appendix S.2.

In contrast, QMD neither imposes this exact invariance nor rules out boundary points. What QMD does imply is that the conditional law at $x_0 + t$ assigns only $o(\|t\|^2)$ mass to the region where p_{x_0} vanishes. Indeed, on the set $\{y : p_{x_0}(y) = 0\}$, the QMD integrand reduces to $p_{x_0+t}(y)$, so

$$\int_{\{y: p_{x_0}(y)=0\}} p_{x_0+t}(y) \nu(dy) = o(\|t\|^2) .$$

Thus QMD allows $p_{x_0+t}(y)$ to be positive on $\{p_{x_0} = 0\}$ for arbitrarily small t , but controls its total contribution in an integrated sense.

All in all, Assumption 1 is stronger than the conditions required for Theorem 3. In particular, it implies Assumptions 2 (with the additional restriction $x_0 \in \text{int}(\mathcal{X})$) and 3, whereas the converse does not hold; see Lemma S.6 in the supplementary appendix.

6 Concluding Remarks

This paper studies the behavior of induced order statistics when the number of nearest neighbors grows with the sample size. We provide general results that link marginal approximation rates for P_r to joint convergence rates for the IOS vector in Hellinger and total variation distances, and we show how these rates depend on the local behavior of the conditional distribution of interest. In the main text, we develop sharp marginal rates under quadratic mean differentiability (QMD), which delivers a simple benchmark rate while accommodating both interior and boundary points. In the supplementary appendix, we provide complementary results under a Taylor/Hölder remainder condition that yields a wider range of marginal rates. Taken together, these results clarify when IOS-based approximations remain valid, when convergence slows down as smoothness weakens, and when it breaks down altogether. In doing so, we also formalize the sense in which stronger

conditions such as those of [Falk et al. \(2010\)](#) imply comparatively fast convergence by imposing substantial structure on the underlying model.

Beyond providing new results for IOS, the framework developed here is intended as a reusable toolkit for settings in which local conditional distributions are approximated using shrinking neighborhoods around a point of interest. By making explicit the trade-offs between smoothness, boundary behavior, and admissible growth rates for k , the analysis facilitates more transparent asymptotic arguments and helps guide future work on inference and estimation based on IOS. While our results are motivated by IOS-based procedures, most notably those arising in regression discontinuity designs and k -nearest-neighbor methods, they also speak to related problems that rely on local approximations to conditional laws but are not explicitly framed in terms of IOS, such as distributionally robust optimization.

Appendix A Proof of the main results

A.1 Proof of Theorem 2

Part (a). Let r_1 be as defined in Lemma 1. Lemma 1 implies that $F(r) \geq F(r/2) > 0$ for all $r \in (0, r_1]$. Then, [Kaufmann and Reiss \(1992, Theorem 1\)](#) implies that, for all $r \in (0, r_1)$,

$$\mathcal{L}(S_n | R_{(k+1)} = r) = P_r^k, \quad (\text{A-1})$$

where P_r^k denotes the k -fold product measure of P_r .

Consider the following derivation for any $k \in \{1, \dots, n\}$ and $n \in \mathbf{N}$,

$$\begin{aligned} \mathbb{H}^2(\mathcal{L}(S_n), \mathcal{L}(S)) &\stackrel{(1)}{\leq} \int_0^\infty \mathbb{H}^2(\mathcal{L}(S_n | R_{(k+1)} = r), \mathcal{L}(S | R_{(k+1)} = r)) dP_{R_{(k+1)}}(r) \\ &\stackrel{(2)}{=} \int_0^{r_1} \mathbb{H}^2(P_r^k, P^k) dP_{R_{(k+1)}}(r) + \int_{r_1}^\infty \mathbb{H}^2(\mathcal{L}(S_n | R_{(k+1)} = r), \mathcal{L}(S | R_{(k+1)} = r)) dP_{R_{(k+1)}}(r) \\ &\stackrel{(3)}{\leq} k \int_0^{r_1} \mathbb{H}^2(P_r, P) dP_{R_{(k+1)}}(r) + P(R_{(k+1)} \geq r_1), \end{aligned} \quad (\text{A-2})$$

where (1) holds by the convexity lemma (e.g., [Reiss \(1993, Lemma 3.1.3\)](#)), (2) by (A-1), $S \perp R_{(k+1)}$, and $S \sim P^k$, and (3) by $\mathbb{H}^2(P', P'') \leq 1$ and Bernoulli's inequality, which states that $1 - (1-x)^k \leq kx$ for any $x \in [0, 1]$ and $k \in \mathbf{N}$.

Given (A-2), we complete the proof by finding constants $C_1, C_2 \in (0, \infty)$ such that, for all $k \in \{1, \dots, n\}$ and $n \in \mathbf{N}$,

$$\int_0^{r_1} \mathbb{H}^2(P_r, P) dP_{R_{(k+1)}}(r) \leq C_1 (k/n)^{2a_h/d} \quad (\text{A-3})$$

$$P(R_{(k+1)} \geq r_1) \leq C_2 k (k/n)^{2a_h/d}. \quad (\text{A-4})$$

We begin by proving (A-3). By $H(P_r, P) = O(r^{a_h})$, there is $C_3 \in (0, \infty)$ such that, for all $r \geq 0$,

$$H^2(P_r, P) \leq C_3 r^{2a_h} . \quad (\text{A-5})$$

Then, consider the following derivation:

$$\begin{aligned} \int_0^{r_1} H^2(P_r, P) dP_{R_{(k+1)}}(r) &\stackrel{(1)}{\leq} C_3 E[R_{(k+1)}^{2a_h} I\{R_{(k+1)} \leq r_1\}] \\ &\stackrel{(2)}{=} C_3 E[F^{-1}(U_{(k+1:n)})^{2a_h} I\{F^{-1}(U_{(k+1:n)}) \leq r_1\}] \\ &\stackrel{(3)}{\leq} C_3 E[F^{-1}(U_{(k+1:n)})^{2a_h} I\{U_{(k+1:n)} \leq C_L r_1^d\}] \\ &\stackrel{(4)}{\leq} C_3 C_L^{-2a_h/d} E[U_{(k+1:n)}^{2a_h/d} I\{U_{(k+1:n)} \leq C_L r_1^d\}] \\ &\leq C_3 C_L^{-2a_h/d} E[U_{(k+1:n)}^{2a_h/d}] \\ &\stackrel{(5)}{\leq} C_3 C_L^{-2a_h/d} C_4 (k/n)^{2a_h/d} , \end{aligned}$$

where (1) holds by (A-5), (2) by quantile transformation, which gives $R_{(k+1)} = F^{-1}(U_{(k+1:n)})$ for F^{-1} as defined in Lemma 1, (3) by the fact that $F^{-1}(U_{(k+1:n)}) \leq r_1$ implies $U_{(k+1:n)} \leq F(r_1)$ and Lemma 1, which implies that $F(r_1) \leq C_L r_1^d$, (4) by Lemma 1 applied to $U_{(k+1:n)} \leq C_L r_1^d$, and (5) by Lemma S.4 with $m = 2a_h/d$. By setting $C_1 := C_3 C_L^{-2a_h/d} C_4$, (A-3) follows.

We now show (A-4). Let $\tilde{r}_1 \in (0, r_1]$ be an arbitrary continuity point of F . We divide the argument into two cases. First, consider any $k \in \{1, \dots, n\}$ and $n \in \mathbf{N}$ such that $k/n \leq F(\tilde{r}_1)/2$. Let $\mu := k/(n+1) \leq (k/n)(n/(n+1)) \leq F(r_1)/2 < 1$ and $\sigma^2 := \mu(1-\mu)$. Then,

$$\begin{aligned} P(R_{(k+1)} \geq r_1) &\stackrel{(1)}{\leq} P(R_{(k+1)} \geq \tilde{r}_1) \\ &\stackrel{(2)}{\leq} P\left(\frac{\sqrt{n}(U_{(k+1:n)} - \mu)}{\sigma} \geq \frac{\sqrt{n}(F(\tilde{r}_1) - \mu)}{\sigma}\right) \\ &\stackrel{(3)}{\leq} \exp\left(-\frac{n(F(\tilde{r}_1) - \mu)^2}{3(F(\tilde{r}_1) - \mu^2)}\right) \\ &\stackrel{(4)}{\leq} \exp(-nF(\tilde{r}_1)/12) \\ &\stackrel{(5)}{\leq} C_5 (1/n)^{2a_h/d} \\ &\stackrel{(6)}{\leq} C_5 k(k/n)^{2a_h/d} , \end{aligned} \quad (\text{A-6})$$

where (1) holds by $\tilde{r}_1 \leq r_1$, (2) by $\{F^{-1}(U_{(k+1:n)}) \geq \tilde{r}_1\} \subseteq \{U_{(k+1:n)} \geq F(\tilde{r}_1)\}$, which relies on \tilde{r}_1 being a continuity point of F , (3) by $F(\tilde{r}_1) - \mu \geq F(\tilde{r}_1)/2 > 0$ and Reiss (1989, Lemma 3.1.1) with $\varepsilon = \sqrt{n}(F(\tilde{r}_1) - \mu)/\sigma > 0$ and $\sigma^2 = \mu(1-\mu)$, (4) by $\mu \in [0, F(\tilde{r}_1)/2]$, which yields $(F(\tilde{r}_1) - \mu)^2 \geq F(\tilde{r}_1)^2/4$ and $F(\tilde{r}_1) \geq F(\tilde{r}_1) - \mu^2 > 0$, (5) follows from $\exp(-cn) = O(n^{-\delta})$ for any $\delta > 0$, and (6) by $k \geq 1$.

Second, for any $k \in \{1, \dots, n\}$ and $n \in \mathbf{N}$ such that $k/n > F(\tilde{r}_1)/2$, we have

$$P(R_{(k+1)} \geq r_1) \leq 1 \stackrel{(1)}{\leq} C_6(k/n)^{2a_h/d} \stackrel{(2)}{\leq} C_6 k(k/n)^{2a_h/d}, \quad (\text{A-7})$$

where (1) holds for $C_6 := (2/F(\tilde{r}_1))^{2a_h/d}$, and (2) holds since $k \geq 1$. To conclude, note that (A-6) and (A-7) imply (A-4) holds with $C_2 := \max\{C_5, C_6\}$ for all $k \in \{1, \dots, n\}$ and $n \in \mathbf{N}$, as desired.

Part (b). By part (a) and (2.6), we obtain that, for any $k \in \{1, \dots, n\}$ and $n \in \mathbf{N}$, $\text{TV}(\mathcal{L}(S_n), \mathcal{L}(S)) = O(k^{1/2}(k/n)^{a_h/d})$. To complete the proof, it then suffices to show that, for any $k \in \{1, \dots, n\}$ and $n \in \mathbf{N}$, we also have $\text{TV}(\mathcal{L}(S_n), \mathcal{L}(S)) = O(k(k/n)^{a_{tv}/d})$.

Our proof follows the arguments of part (a) very closely. Recall that $a_{tv} \in [a_h, 2a_h]$, which follows again from $H(P_r, P) = O(r^{a_h})$ and (2.6), and note that the arguments used to derive (A-1) continue to apply. Then, for any $k \in 1, \dots, n$ and $n \in \mathbf{N}$, we have

$$\begin{aligned} \text{TV}(\mathcal{L}(S_n), \mathcal{L}(S)) &\stackrel{(1)}{=} \sup_{B \in \mathcal{B}} \left| \int_0^\infty (P(S_n \in B | R_{(k+1)} = r) - P(S \in B | R_{(k+1)} = r)) dP_{R_{(k+1)}}(r) \right| \\ &\leq \int_0^\infty \sup_{B \in \mathcal{B}} |P(S_n \in B | R_{(k+1)} = r) - P(S \in B | R_{(k+1)} = r)| dP_{R_{(k+1)}}(r) \\ &= \int_0^\infty \text{TV}(\mathcal{L}(S_n | R_{(k+1)} = r), \mathcal{L}(S | R_{(k+1)} = r)) dP_{R_{(k+1)}}(r) \\ &\stackrel{(2)}{=} \int_0^{r_1} \text{TV}(P_r^k, P^k) dP_{R_{(k+1)}}(r) + \int_{r_1}^\infty \text{TV}(\mathcal{L}(S_n | R_{(k+1)} = r), \mathcal{L}(S | R_{(k+1)} = r)) dP_{R_{(k+1)}}(r) \\ &\stackrel{(3)}{\leq} k \int_0^{r_1} \text{TV}(P_r, P) dP_{R_{(k+1)}}(r) + P(R_{(k+1)} \geq r_1), \end{aligned} \quad (\text{A-8})$$

where (1) holds by the law of total probability and using \mathcal{B} to denote the common σ -algebra to $\mathcal{L}(S_n)$ and $\mathcal{L}(S)$, and (2) by (A-1), $S \perp R_{(k+1)}$, and $S \sim P^k$, and (3) by [Hoeffding and Wolfowitz \(1958, Eq. \(4.5\)\)](#). Given (A-8), we complete the proof by finding constants $C_1, C_2 \in (0, \infty)$ such that, for all $k \in \{1, \dots, n\}$ and $n \in \mathbf{N}$,

$$\int_0^{r_1} \text{TV}(P_r, P) dP_{R_{(k+1)}}(r) \leq C_1(k/n)^{a_{tv}/d} \quad (\text{A-9})$$

$$P(R_{(k+1)} \geq r_1) \leq C_2 k(k/n)^{a_{tv}/d}. \quad (\text{A-10})$$

Then, (A-9) and (A-10) follow by repeating the arguments used to derive (A-3) and (A-4), with $H(P_r, P) = O(r^{a_h})$ replaced by $\text{TV}(P_r, P) = O(r^{a_{tv}})$.

A.2 Proof of Theorem 3

We start by proving (3.1) and begin with a preliminary result. For any $x \in \mathcal{X}$, we define $D(x)$ as in Lemma 2, and note that $\mathbb{H}^2(P_x, P) = D(x)/2$. Lemma 2 shows that, as $\|x - x_0\| \rightarrow 0$,

$$D(x) = \frac{1}{4}(x - x_0)^\top \left(\int \dot{\ell}_{x_0}(y) \dot{\ell}_{x_0}(y)^\top p_{x_0}(y) \nu(dy) \right) (x - x_0) + o(\|x - x_0\|^2). \quad (\text{A-11})$$

Moreover, Assumption 3 implies

$$\left| \int \dot{\ell}_{x_0}(y) \dot{\ell}_{x_0}(y)^\top p_{x_0}(y) \nu(dy) \right| \leq \int \|\dot{\ell}_{x_0}(y)\|^2 p_{x_0}(y) \nu(dy) < \infty. \quad (\text{A-12})$$

By (A-11) and (A-12), there are constants $C > 0$ and $\bar{r} > 0$ so that, for all $\|x - x_0\| < \bar{r}$,

$$D(x) \leq C\|x - x_0\|^2. \quad (\text{A-13})$$

Let r_1 be as in Lemma 1. By this result, for all $r \in (0, r_1)$,

$$F(r) := \int_{B_r} g(x) dx \geq C_L r^d > 0. \quad (\text{A-14})$$

For any $r \in (0, r_1)$, we can then define the following measure on \mathcal{X} :

$$\mu_r(dx) = \frac{g(x) I\{x \in B_r\}}{F(r)} dx. \quad (\text{A-15})$$

By this definition, note that $h_r(y) = \int p_x(y) \mu_r(dx)$ and $P_r = \int P_x \mu_r(dx)$.

Let $r_2 := \min\{r_1, \bar{r}\}$. For any $r \in (0, r_2)$, consider the following derivation.

$$\begin{aligned} \mathbb{H}^2(P_r, P) &\stackrel{(1)}{\leq} \int \mathbb{H}^2(P_x, P) \mu_r(dx) \\ &\stackrel{(2)}{=} \frac{1}{2F(r)} \int_{B_r} D(x) g(x) dx \\ &\stackrel{(3)}{\leq} \frac{C}{2F(r)} \int_{B_r} \|x - x_0\|^2 g(x) dx \\ &\stackrel{(4)}{\leq} \frac{C}{2C_L r^d} (g(x_0) + C_g r) \int_{B_r} \|x - x_0\|^2 dx \\ &\stackrel{(5)}{=} O(r^2), \end{aligned} \quad (\text{A-16})$$

where (1) holds by the convexity lemma (e.g., Reiss (1993, Lemma 3.1.3)), with measure $\mu_r(dx)$ and $P_r = \int P_x \mu_r(dx)$, (2) by $\mathbb{H}^2(P_x, P) = D(x)/2$ and (A-15), (3) by (A-13), (4) by (A-14) and Assumption 2, and (5) by $\int_{B_r} \|x - x_0\|^2 dx = O(r^{d+2})$. From (A-16), we obtain

$$\mathbb{H}(P_r, P) = O(r).$$

The convergence rate of the total variation distance follows from its relationship with the Hellinger, $\text{TV}(P_r, P) \leq \sqrt{2}\text{H}(P_r, P)$. This completes the proof of (3.1).

Part (i) of the Theorem follows from Lemma S.1. The first statement in Part (ii) follows from Lemma S.2, while the result delivering $\text{TV}(P_r, P) = o(r)$ follows from Lemma S.3. This completes the proof. ■

Appendix B Auxiliary results

Lemma 1. *Under Assumption 2, there exists $r_1 > 0$ and $0 < C_L < C_H < \infty$ such that, for all $r \in [0, r_1]$,*

$$F(r) := \int_{B_r} g(x) dx \quad \text{satisfies} \quad C_L r^d \leq F(r) \leq C_H r^d. \quad (\text{B-1})$$

In addition, for any $u \in (0, C_L r_1^d]$, the (left-continuous) quantile function of F ,

$$F^{-1}(u) := \inf\{r : F(r) \geq u\}, \quad \text{satisfies} \quad F^{-1}(u) \leq (u/C_L)^{1/d}.$$

Proof. Let $r_1 := g(x_0)/(2 \max\{C_g, 1\}) > 0$. For any $x \in \mathcal{X}$ with $\|x - x_0\| \leq r$ and $r \in [0, r_1]$, we have

$$g(x) \stackrel{(1)}{\geq} g(x_0) - C_g r \stackrel{(2)}{\geq} (g(x_0)/2)(2 - C_g/(\max\{C_g, 1\})) \geq g(x_0)/2 \stackrel{(3)}{>} 0, \quad (\text{B-2})$$

where (1) and (3) hold by Assumption 2 and (2) by $r \leq r_1$. Then, for all $r \in [0, r_1]$,

$$F(r) \stackrel{(1)}{\geq} \frac{g(x_0)}{2} \int_{B_r \cap \mathcal{X}} dx \geq \frac{g(x_0)}{2} C_{x_0} r^d \stackrel{(2)}{=} C_L r^d, \quad (\text{B-3})$$

where (1) holds by (B-2) and $\|x - x_0\| \leq r$, and (2) by setting $C_L := \frac{g(x_0)}{2} C_{x_0}$. This proves the first inequality in (B-1).

For any $x \in \mathcal{X}$ with $\|x - x_0\| \leq r$ and $r \in [0, r_1]$, we also have

$$g(x) \stackrel{(1)}{\leq} g(x_0) + C_g r \stackrel{(2)}{\leq} g(x_0) + C_g r_1, \quad (\text{B-4})$$

where (1) holds by Assumption 2 and (2) by $r \leq r_1$. Then, for all $r \in [0, r_1]$,

$$F(r) \stackrel{(1)}{\leq} (g(x_0) + C_g r_1) \text{Vol}(B_r) \stackrel{(2)}{=} C_H r^d,$$

where (1) holds by (B-4) and $\|x - x_0\| \leq r$, and (2) by using that $\text{Vol}(B_r) = \pi^{d/2}/\Gamma(d/2 + 1)r^d$ and setting $C_H := (g(x_0) + C_g r_1)\pi^{d/2}/\Gamma(d/2 + 1)$. This proves the second inequality in (B-1).

For the last claim, take any $r \in (0, r_1]$, and note that

$$F^{-1}(C_L r^d) \stackrel{(1)}{\leq} F^{-1}(F(r)) \stackrel{(2)}{\leq} r, \quad (\text{B-5})$$

where (1) holds by the weak monotonicity of the quantile function and (B-3), and (2) by elementary properties of the quantile function (e.g., Pollard (2002, Eq. 36)). For any $u \in (0, C_L r_1^d]$, applying (B-5) to $r = (u/C_L)^{1/d} \in (0, r_1]$ proves the desired result. ■

Lemma 2. *Let Assumption 3 hold and define, for x in a neighborhood of x_0 ,*

$$D(x) := \int \left(\sqrt{p_x(y)} - \sqrt{p_{x_0}(y)} \right)^2 \nu(dy).$$

Let

$$I(x_0) := \int \dot{\ell}_{x_0}(y) \dot{\ell}_{x_0}(y)^\top p_{x_0}(y) \nu(dy)$$

denote the $d \times d$ matrix induced by the score $\dot{\ell}_{x_0}$. Then, as $x \rightarrow x_0$,

$$D(x) = \frac{1}{4} (x - x_0)^\top I(x_0) (x - x_0) + o(\|x - x_0\|^2).$$

Proof. We begin with some definitions. For any $x \in \mathcal{X} \setminus \{x_0\}$, denote the difference quotient by

$$\Delta(y, x) := \frac{\sqrt{p_x(y)} - \sqrt{p_{x_0}(y)}}{\|x - x_0\|},$$

and the (directional) quadratic mean derivative by

$$\dot{d}(y, x_0; x) := \frac{1}{2} \mathbf{u}_x^\top \dot{\ell}_{x_0}(y) \sqrt{p_{x_0}(y)},$$

where

$$\mathbf{u}_x := \frac{x - x_0}{\|x - x_0\|}.$$

For any $x = x_0$, we define $\Delta(y, x) := 0$ and $\dot{d}(y, x_0; x) := 0$.

Next, consider the following derivation for any $x \in \mathcal{X} \setminus \{x_0\}$,

$$\begin{aligned} D(x) &= \|x - x_0\|^2 \int \Delta(y, x)^2 \nu(dy) \\ &= \|x - x_0\|^2 \int (\dot{d}(y, x_0; x) + (\Delta(y, x) - \dot{d}(y, x_0; x)))^2 \nu(dy) \\ &= \left(\begin{array}{c} \frac{1}{4} (x - x_0)^\top I(x_0) (x - x_0) + \\ \|x - x_0\|^2 \int (\Delta(y, x) - \dot{d}(y, x_0; x))^2 \nu(dy) + \\ 2\|x - x_0\|^2 \int \dot{d}(y, x_0; x) (\Delta(y, x) - \dot{d}(y, x_0; x)) \nu(dy) \end{array} \right). \end{aligned} \quad (\text{B-6})$$

Note that (B-6) extends to $x_0 \in \mathcal{X}$. For all $x \in \mathcal{X}$, we then have

$$\left| D(x) - \frac{1}{4} (x - x_0)^\top I(x_0) (x - x_0) \right| = \|x - x_0\|^2 \left(\begin{array}{c} \int (\Delta(y, x) - \dot{d}(y, x_0; x))^2 \nu(dy) + \\ 2 \left| \int \dot{d}(y, x_0; x) (\Delta(y, x) - \dot{d}(y, x_0; x)) \nu(dy) \right| \end{array} \right).$$

To complete the proof, it suffices to show that for any sequence of $x \in \mathcal{X}$ with $x \rightarrow x_0$,

$$\|x - x_0\|^2 \int (\Delta(y, x) - \dot{d}(y, x_0; x))^2 \nu(dy) = O(\|x - x_0\|^2) \quad (\text{B-7})$$

$$\|x - x_0\|^2 \left| \int \dot{d}(y, x_0; x) (\Delta(y, x) - \dot{d}(y, x_0; x)) \nu(dy) \right| = O(\|x - x_0\|^2). \quad (\text{B-8})$$

To show (B-7), note that for all $x \in \mathcal{X}$,

$$\begin{aligned} & \|x - x_0\|^2 \int (\Delta(y, x) - \dot{d}(y, x_0; x))^2 \nu(dy) \\ & \stackrel{(1)}{=} \int \left(\sqrt{p_x(y)} - \sqrt{p_{x_0}(y)} - \frac{1}{2} ((x - x_0)^\top \dot{\ell}_{x_0}(y)) \sqrt{p_{x_0}(y)} \right)^2 \nu(dy) \\ & \stackrel{(2)}{=} o(\|x - x_0\|^2) \quad \text{as } x \rightarrow x_0, \end{aligned}$$

where (1) holds by definition of $\Delta(y, x)$ and $\dot{d}(y, x_0; x)$ (even if $x = x_0$), and (2) by Assumption 3.

To show (B-8), first note that for all $x \in \mathcal{X}$ with $\|x - x_0\| \neq 0$,

$$\int (\dot{d}(y, x_0; x))^2 \nu(dy) = \frac{1}{4} \int (\mathbf{u}_x^\top \dot{\ell}_{x_0}(y))^2 p_{x_0}(y) \nu(dy) \stackrel{(1)}{\leq} \frac{1}{4} \int \|\dot{\ell}_{x_0}(y)\|^2 p_{x_0}(y) \nu(dy) \stackrel{(2)}{<} \infty, \quad (\text{B-9})$$

where (1) holds by $(\mathbf{u}_x^\top \dot{\ell}_{x_0}(y))^2 \leq \|\dot{\ell}_{x_0}(y)\|^2$, which follows from $\|\mathbf{u}_x\| = 1$ and the Cauchy-Schwarz inequality and (2) by Assumption 3. Also, observe that (B-9) extends to $x_0 \in \mathcal{X}$, as this gives $\int (\dot{d}(y, x_0; x))^2 \nu(dy) = 0$. Then, (B-8) follows from the next derivation for all $x \in \mathcal{X}$,

$$\begin{aligned} & \|x - x_0\|^2 \left| \int \dot{d}(y, x_0; x) (\Delta(y, x) - \dot{d}(y, x_0; x)) \nu(dy) \right| \\ & \stackrel{(1)}{\leq} \|x - x_0\|^2 \left(\int (\Delta(y, x) - \dot{d}(y, x_0; x))^2 \nu(dy) \right)^{1/2} \left(\int \dot{d}(y, x_0; x)^2 \nu(dy) \right)^{1/2} \\ & \stackrel{(2)}{=} o(\|x - x_0\|^2) \quad \text{as } x \rightarrow x_0, \end{aligned}$$

as desired in (B-8), where (1) holds by the Cauchy-Schwarz inequality and (2) by (B-7) and (B-9). This completes the proof. ■

References

ABRAMOWITZ, M. AND I. A. STEGUN (1964): *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, National Bureau of Standards. Applied Mathematics Series.

- ADUSUMILLI, K. (2025): “Risk and Optimal Policies in Bandit Experiments,” *Econometrica*, forthcoming.
- BHATTACHARYA, P. (1974): “Convergence of sample paths of normalized sums of induced order statistics,” *The Annals of Statistics*, 1034–1039.
- BUGNI, F. A. AND I. A. CANAY (2021): “Testing Continuity of a Density via g-order statistics in the Regression Discontinuity Design,” *Journal of Econometrics*, 221, 138–159.
- BUGNI, F. A., I. A. CANAY, AND D. KIM (2025): “Testing Conditional Stochastic Dominance at Target Points,” .
- BUGNI, F. A., J. LI, AND Q. LI (2023): “Permutation-Based Tests for Discontinuities in Event Studies,” *Quantitative Economics*, 14, 37–70.
- CANAY, I. A. AND V. KAMAT (2018): “Approximate Permutation Tests and Induced Order Statistics in the Regression Discontinuity Design,” *The Review of Economic Studies*, 85, 1577–1608.
- DAVID, H. AND J. GALAMBOS (1974): “The asymptotic theory of concomitants of order statistics,” *Journal of Applied Probability*, 762–770.
- DAVID, H. A. AND H. N. NAGARAJA (2003): *Order Statistics*, Wiley Series in Probability and Statistics, third edition ed.
- ESTEBAN-PÉREZ, A. AND J. M. MORALES (2022): “Distributionally robust stochastic programs with side information based on trimmings,” *Mathematical Programming*, 195, 1069–1105.
- FALK, M., J. HÜSLER, AND R.-D. REISS (2010): *Laws of small numbers: extremes and rare events*, Springer Science & Business Media.
- GOLDMAN, M. AND D. M. KAPLAN (2018): “Comparing distributions by multiple testing across quantiles or CDF values,” *Journal of Econometrics*, 206, 143–166.
- HOEFFDING, W. AND J. WOLFOWITZ (1958): “Distinguishability of Sets of Distributions,” *The Annals of Mathematical Statistics*, 29, 700–718.
- KAUFMANN, E. AND R.-D. REISS (1992): “On conditional distributions of nearest neighbors,” *Journal of Multivariate Analysis*, 42, 67–76.
- LE CAM, L. (1986): *Asymptotic Methods in Statistical Decision Theory*, Springer.
- POLLARD, D. (2002): *A User’s Guide to Measure Theoretic Probability*, Cambridge University Press, New York.
- REISS, R.-D. (1989): *Approximate distributions of order statistics: with applications to nonparametric statistics*, Springer-Verlag, New York.

——— (1993): *A Course on Point Processes*, Springer Series in Statistics, New York: Springer.

VAN DER VAART, A. W. (2000): *Asymptotic statistics*, vol. 3, Cambridge university press.

SUPPLEMENTARY APPENDIX FOR
“On the Rate of Convergence of Induced Ordered Statistics
and their Applications”

Federico A. Bugni
Department of Economics
Northwestern University
federico.bugni@northwestern.edu

Ivan A. Canay
Department of Economics
Northwestern University
iacanay@northwestern.edu

Deborah Kim
Department of Economics
University of Warwick
deborah.kim@warwick.ac.uk

Abstract

This document provides supplementary material for the paper “On the Rates of Convergence of Induced Ordered Statistics and their Applications.” It collects auxiliary lemmas and technical proofs omitted from the main text, develops additional marginal rate results under alternative smoothness assumptions, and provides further examples and discussion.

KEYWORDS: hellinger distance, total variation distance, induced order statistics, rank tests, permutation tests.

JEL codes: C12, C14.

Supplementary Contents

Appendix S.1 Supplementary Lemmas	1
Appendix S.2 Results under Alternative Assumptions	14
S.2.1 Auxiliary Results under Alternative Assumptions	17
S.2.2 Proof of Theorem S.1	18

Appendix S.1 Supplementary Lemmas

Lemma S.1. *Suppose Assumptions 2–3 hold. If x_0 lies on the boundary of \mathcal{X} , then there exist a data-generating process $Q \in \mathbf{Q}$ and constants $C > 0$ and $\bar{r} > 0$ such that*

$$\mathbb{H}(P_r, P) \geq Cr \quad \text{and} \quad \text{TV}(P_r, P) \geq Cr \quad \text{for all } r \in (0, \bar{r}) . \quad (\text{S.1.1})$$

Proof. Let $d \geq 1$, $m \geq 1$, and $x_0 = \mathbf{0}_d$. Let $X \sim \text{Unif}([0, 1]^d)$, so that $g(x) = 1[x \in [0, 1]^d]$ and $\mathcal{X} = [0, 1]^d$. Then, Assumption 2 holds with $C_g = 0$, $r_0 = 1$, and $C_{x_0} = (2^{-d}\pi^{d/2}/\Gamma(d/2 + 1))$. To see this, note that for all $r \in (0, 1)$, we have

$$\text{Vol}(B_r \cap \mathcal{X}) = 2^{-d}\text{Vol}(B_r) = \left(2^{-d}\pi^{d/2}/\Gamma(d/2 + 1)\right) r^d . \quad (\text{S.1.2})$$

For any $x \in \mathcal{X}$, we define $\{Y|X = x\} \sim N(\mu(x), \mathbf{I}_m)$ with $\mu(x) := (x_1, 0, \dots, 0)^\top \in \mathbf{R}^m$, so that $p_x(y) = (2\pi)^{-m/2} \exp(-\|y - \mu(x)\|^2/2)$. For any $x \in \mathcal{X}$, it then follows that

$$\nabla_x \ln p_x(y) = (y_1 - x_1, 0, \dots, 0)^\top \in \mathbf{R}^m . \quad (\text{S.1.3})$$

We now show that Assumption 3 holds with

$$\dot{\ell}_{x_0}(y) := \nabla_x \ln p_x(y)|_{x=x_0} = (y_1, 0, \dots, 0)^\top . \quad (\text{S.1.4})$$

First, note that

$$\int \|\dot{\ell}_{x_0}(y)\|^2 p_{x_0}(y) \nu(dy) \stackrel{(1)}{=} \int y_1^2 p_{x_0}(y) \nu(dy) = E[Y_1^2|X = x_0] \stackrel{(2)}{=} 1 < \infty , \quad (\text{S.1.5})$$

where (1) holds by (S.1.4) and (2) by $\{Y|X = x_0\} \sim N(\mathbf{0}_m, \mathbf{I}_m)$. Second, since $p_x(y)$ is positive and differentiable in x , a first-order Taylor expansion of $g(x) := \sqrt{p_x(y)}$ centered at $x = x_0$ gives:

$$\sqrt{p_x(y)} = \sqrt{p_{x_0}(y)} + \frac{1}{2}(p_{\bar{x}}(y))^{-1/2} \nabla_x p_{\bar{x}}(y)^\top (x - x_0) , \quad (\text{S.1.6})$$

where \tilde{x} is between x_0 and x . By evaluating (S.1.6) at $x = x_0 + t \in \mathcal{X}$ and using $(p_{\tilde{x}}(y))^{-1/2} \nabla_x p_{\tilde{x}}(y) = \nabla_x \ln p_{\tilde{x}}(y)$ and (S.1.3), we obtain

$$\sqrt{p_{x_0+t}(y)} = \sqrt{p_{x_0}(y)} + \frac{1}{2} \sqrt{p_{\tilde{x}}(y)} (y_1 - \tilde{x}_1) t_1, \quad (\text{S.1.7})$$

where \tilde{x} is between x_0 and $x_0 + t$. From here, we consider the following derivation:

$$\begin{aligned} & \int \left(\sqrt{p_{x_0+t}(y)} - \sqrt{p_{x_0}(y)} - \frac{1}{2} t^\top \dot{\ell}_{x_0}(y) \sqrt{p_{x_0}(y)} \right)^2 \nu(dy) \\ & \stackrel{(1)}{=} \left(E[(Y_1 - \tilde{x}_1)^2 | X = \tilde{x}] + E[Y_1^2 | X = x_0] - 2 \int \sqrt{p_{\tilde{x}}(y) p_{x_0}(y)} y_1 (y_1 - \tilde{x}_1) \nu(dy) \right) \frac{t_1^2}{4} \\ & \stackrel{(2)}{=} \left(1 - \int \sqrt{p_{\tilde{x}}(y) p_{x_0}(y)} y_1 (y_1 - \tilde{x}_1) \nu(dy) \right) \frac{t_1^2}{2} \\ & = (1 - \exp(-\tilde{x}_1^2/8) (1 - \tilde{x}_1^2/4)) \frac{t_1^2}{2}, \end{aligned} \quad (\text{S.1.8})$$

where (1) holds by (S.1.4) and (S.1.7), and (2) by $\{Y|X = x\} \sim N((x_1, 0, \dots, 0), \mathbf{I}_m)$. As $t \downarrow 0$, $x_0 + t \rightarrow x_0$, and so $\tilde{x} \rightarrow x_0$ and $\exp(-\tilde{x}_1^2/8) (1 - \tilde{x}_1^2/4) \rightarrow 1$. Then, as $t \downarrow 0$, (S.1.8) implies that

$$\int \left(\sqrt{p_{x_0+t}(y)} - \sqrt{p_{x_0}(y)} - \frac{1}{2} t^\top \dot{\ell}_{x_0}(y) \sqrt{p_{x_0}(y)} \right)^2 \nu(dy) = o(t_1^2) = o(\|t\|^2). \quad (\text{S.1.9})$$

By (S.1.5) and (S.1.9), we have verified that Assumption 3 holds.

To complete the proof, it suffices to find $C > 0$ and $\bar{r} > 0$ such that (S.1.1) holds. Let $A := \{y \in \mathbf{R}^m : y_1 > 0\}$, and note that

$$\sqrt{2} \mathbb{H}(P_r, P) \geq \text{TV}(P_r, P) \geq |P_r(A) - P(A)|.$$

Given this result, we complete the proof by finding $C > 0$ and $\bar{r} > 0$ such that

$$|P_r(A) - P(A)| \geq Cr \quad \text{for all } r \in (0, \bar{r}). \quad (\text{S.1.10})$$

We devote the rest of the proof to showing (S.1.10).

Under $P = P_0$, we have $\{Y|X = x_0\} \sim N(\mathbf{0}_m, \mathbf{I}_m)$, and so $P(A) = 1/2$. For any $r \in (0, 1)$,

$$P_r(A) = \int P(Y_1 > 0 | X = x) \mu_r(dx) \stackrel{(1)}{=} E[\Phi(X_1) | X \in B_r(0) \cap [0, 1]^d] \stackrel{(2)}{=} \frac{\int_{B_r(0) \cap [0, 1]^d} \Phi(x_1) dx}{\text{Vol}(B_r(0) \cap [0, 1]^d)}, \quad (\text{S.1.11})$$

where (1) holds by $\{Y|X = x\} \sim N((x_1, 0, \dots, 0), \mathbf{I}_m)$ and (2) by $\{X_1 | X \in B_r(0) \cap [0, 1]^d\} \sim U([0, r])$ for $r \in (0, 1)$. Note that $P_r(A)$ is twice differentiable in $r \in (0, 1)$. A first-order Taylor

expansion of $g(r) := P_r(A)$ centered at $s \in (0, r)$ gives:

$$P_r(A) = P_s(A) + \frac{\partial P_r(A)}{\partial r} \Big|_{r=\tilde{s}}(r-s), \quad (\text{S.1.12})$$

where $\tilde{s} \in [s, r]$. We can obtain

$$\lim_{r \downarrow 0} \frac{\partial P_r(A)}{\partial r} = \frac{\Gamma(d/2+1)}{\sqrt{2\pi}\Gamma((d+3)/2)} > 0. \quad (\text{S.1.13})$$

By (S.1.13) and $\sqrt{2\pi} > 1$, there is $\bar{r} > 0$ such that for all $\tilde{s} \in (0, \bar{r})$,

$$\frac{\partial P_r(A)}{\partial r} \Big|_{r=\tilde{s}} \geq \frac{\Gamma(d/2+1)}{\Gamma((d+3)/2)} > 0. \quad (\text{S.1.14})$$

Then, for s, r with $0 < s < r < \bar{r}$, (S.1.12) and (S.1.14) imply

$$|P_r(A) - P_s(A)| \geq \inf_{\tilde{s} \in [r, s]} \frac{\partial P_r(A)}{\partial r} \Big|_{r=\tilde{s}}(r-s) \geq \frac{\Gamma(d/2+1)}{\Gamma((d+3)/2)}(r-s). \quad (\text{S.1.15})$$

By taking $s \downarrow 0$, $P_s(A) \rightarrow P(A)$, and so (S.1.15) implies that for any $r \in (0, \bar{r})$,

$$|P_r(A) - P(A)| \geq \frac{\Gamma(d/2+1)}{\Gamma((d+3)/2)}r. \quad (\text{S.1.16})$$

Note that (S.1.16) implies (S.1.10) with $C = \Gamma(d/2+1)/\Gamma((d+3)/2)$, concluding the proof. ■

Lemma S.2. *For any $\tilde{r} > 0$, let $\mathbf{Q}^o(\tilde{r})$ be the set of all distributions Q that satisfy Assumptions 2 and 3, and for which $B_r \cap \mathcal{X} = B_r$ for all $r \in (0, \tilde{r})$. Then, for any $\tilde{r} > 0$ and $\varepsilon > 0$, there do not exist constants $C < \infty$ and $\bar{r} \in (0, \tilde{r})$ such that*

$$\sup_{Q \in \mathbf{Q}^o(\tilde{r})} \mathbf{H}(P_r, P) \leq Cr^{1+\varepsilon} \quad \text{for all } r \in (0, \bar{r}), \quad (\text{S.1.17})$$

and there do not exist constants $C < \infty$ and $\bar{r} \in (0, \tilde{r})$ such that

$$\sup_{Q \in \mathbf{Q}^o(\tilde{r})} \text{TV}(P_r, P) \leq Cr^{1+\varepsilon} \quad \text{for all } r \in (0, \bar{r}). \quad (\text{S.1.18})$$

Proof. We prove this result by contradiction. To this end, fix $\tilde{r} > 0$ and $\varepsilon > 0$ arbitrarily, and assume there exists $C < \infty$ and $\bar{r} \in (0, \tilde{r})$ satisfying (S.1.17) or (S.1.18), respectively. For any $Q \in \mathbf{Q}^o(\tilde{r})$, we get

$$\mathbf{H}(P_r, P) \leq Cr^{1+\varepsilon} \text{ and } \text{TV}(P_r, P) \leq Cr^{1+\varepsilon} \text{ for all } r \in (0, \bar{r}). \quad (\text{S.1.19})$$

To complete the proof, it suffices to find one distribution $Q \in \mathbf{Q}^o(\tilde{r})$, such that (S.1.19) fails for all $C < \infty$. We show this in the remainder of the proof.

We define the distribution Q as follows. Let $d \geq 1$, $m \geq 1$, and $x_0 = \mathbf{0}_d$. Let $X \sim \text{Unif}(B_{\tilde{r}})$, so that $g(x) = 1[x \in B_{\tilde{r}}]/\text{Vol}(B_{\tilde{r}})$ and $\mathcal{X} = B_{\tilde{r}}$. Next, let $e_1 := (1, 0, \dots, 0)^\top \in \mathbf{R}^m$ and let

$\delta : \mathbf{R} \rightarrow \mathbf{R}$ be defined as follows: $\delta(u) := u/(1 - \ln u)$ for $u \in (0, \exp(-2))$, and $\delta(u) := 0$ for $u = 0$ or $u > \exp(-2)$. Let $\{Y|X = x\}$ be supported on $\{\mathbf{0}_m, e_1\} \subset \mathbf{R}^m$ with the following distribution:

$$P(Y = e_1|X = x) = \pi(x) := \frac{1}{2} + \delta(\|x\|) \text{ and } P(Y = \mathbf{0}_m|X = x) = 1 - \pi(x) .$$

Since $X \sim \text{Unif}(B_{\tilde{r}})$ and $x_0 = \mathbf{0}_d$, Assumption 2 holds with $C_g = 0$, $r_0 = \tilde{r}$, and $C_{x_0} = \pi^{d/2}/\Gamma(d/2 + 1)$. Also, $B_r \cap \mathcal{X} = B_r$ for all $r \in (0, \tilde{r})$. Next, we show that Assumption 3 holds with $\dot{\ell}_{x_0}(y) := \mathbf{0}_d$ and ν equal to the counting measure on $\{\mathbf{0}_m, e_1\}$. First, note that $\dot{\ell}_{x_0}(y) = \mathbf{0}_d$ implies $\int \|\dot{\ell}_{x_0}(y)\|^2 p_{x_0}(y) \nu(dy) = 0 < \infty$. Second, for any $\|t\| \in (0, \exp(-2))$, we have

$$\begin{aligned} & \int (\sqrt{p_{x_0+t}(y)} - \sqrt{p_{x_0}(y)} - \frac{1}{2} t^\top \dot{\ell}_{x_0}(y) \sqrt{p_{x_0}(y)})^2 \nu(dy) \\ & \stackrel{(1)}{=} (\sqrt{\frac{1}{2} + \delta(\|t\|)} - \sqrt{\frac{1}{2}})^2 + (\sqrt{\frac{1}{2} - \delta(\|t\|)} - \sqrt{\frac{1}{2}})^2 \\ & \stackrel{(2)}{=} 2 - \sqrt{1 + \frac{2\|t\|}{(1-\ln\|t\|)}} - \sqrt{1 - \frac{2\|t\|}{(1-\ln\|t\|)}} , \end{aligned} \tag{S.1.20}$$

where (1) holds by $\dot{\ell}_{x_0}(y) = \mathbf{0}_d$, $p_{x_0+t}(e_1) = 1/2 + \delta(\|x_0 + t\|)$, $p_{x_0+t}(\mathbf{0}_m) = 1/2 - \delta(\|x_0 + t\|)$, and $p_{x_0}(y) = 1/2$ for $y \in \{\mathbf{0}_m, e_1\}$, (2) by $\delta(u) := u/(1 - \ln u)$ for $u \in (0, \exp(-2))$. By a second-order Taylor expansion of $h(u) = 2 - \sqrt{1 + 2u} - \sqrt{1 - 2u}$ centered at $u = 0$, and that $h(0) = h'(0) = 0$,

$$h(u) = ((1 + 2\tilde{u})^{-3/2} + (1 - 2\tilde{u})^{-3/2}) \frac{u^2}{2}$$

with $\tilde{u} \in [0, u]$. Next, note that $\|t\|/(1 - \ln\|t\|) \in (0, \exp(-2)/3)$ for $\|t\| \in (0, \exp(-2))$, and $h''(u) \in (2, 2.1)$ for $u \in (0, \exp(-2)/3)$. By plugging the results on (S.1.20), we conclude that

$$\int (\sqrt{p_{x_0+t}(y)} - \sqrt{p_{x_0}(y)} - \frac{1}{2} t^\top \dot{\ell}_{x_0}(y) \sqrt{p_{x_0}(y)})^2 \nu(dy) = O\left(\frac{\|t\|^2}{(1-\ln\|t\|)^2}\right) .$$

The right-hand side expression of the previous equation is $o(\|t\|^2)$ as $\|t\| \downarrow 0$, which verifies Assumption 3. We then conclude that $Q \in \mathbf{Q}^o(\tilde{r})$.

Next, let $A := \{y \in \mathbf{R}^m : y = e_1\}$. Consider the following argument for any $r \in (0, \min\{\tilde{r}, \exp(-2)\})$.

$$\begin{aligned} \sqrt{2}\mathbb{H}(P_r, P) & \geq \text{TV}(P_r, P) \geq |P_r(A) - P(A)| \\ & \stackrel{(1)}{=} \int_{\|x\| \leq r} \delta(\|x\|) dx / \text{Vol}(B_r) \\ & \stackrel{(2)}{=} \frac{d}{r^d} \int_0^r v^d / (1 - \ln v) dv \\ & \stackrel{(3)}{\geq} \frac{d}{r^d} \int_{r/2}^r v^d / (1 - \ln v) dv \\ & \geq \frac{d}{(1-\ln(r/2))r^d} \int_{r/2}^r v^d dv \\ & = \frac{d(1-2^{-d-1})}{d+1} \frac{r}{1-\ln(r/2)} , \end{aligned} \tag{S.1.21}$$

where (1) holds because $P(Y = e_1 | X = x) = 1/2 + \delta(\|x\|) \geq 1/2$ for all x , $P(Y = e_1 | X = x_0) = 1/2$ since $\delta(0) = 0$, $X \sim \text{Unif}(B_{\tilde{r}})$, and $r < \tilde{r}$, (2) by change of variables to polar coordinates, and (3) by $v^d/(1 - \ln v) \geq 0$ for $v < r < \exp(-2) < 1$. Note that the right-hand side expression in (S.1.21) exceeds $Cr^{(1+\varepsilon)}$ for any $C \in (0, \infty)$ when $r \in (0, \tilde{r})$ becomes sufficiently close to zero. Therefore, (S.1.19) must fail for all $C < \infty$, which concludes the proof. ■

Lemma S.3. *Assume that x_0 is interior of \mathcal{X} , in the sense that for r small enough $B_r \cap \mathcal{X} = B_r$. Then, under Assumptions 2–3,*

$$TV(P_r, P) = o(r) .$$

Proof. First, consider the following argument:

$$\begin{aligned} TV(P_r, P) &= \sup_{|\phi|_\infty \leq 1} \left| \int_y \phi(y)(P_r(dy) - P(dy)) \right| / 2 \\ &\stackrel{(1)}{=} \sup_{|\phi|_\infty \leq 1} \left| \int_y \phi(y) \left[\int_t p_{x_0+t}(y) \mu_r(dt) - p_{x_0}(y) \right] \nu(dy) \right| / 2 \\ &\stackrel{(2)}{=} \sup_{|\phi|_\infty \leq 1} \left| \int_t \left(\int_y \phi(y) p_{x_0+t}(y) \nu(dy) \right) \mu_r(dt) - \int_y \phi(y) p_{x_0}(y) \nu(dy) \right| / 2 \\ &\stackrel{(3)}{=} \sup_{|\phi|_\infty \leq 1} |E[m_\phi(T)] - m_\phi(0)| / 2 , \end{aligned} \tag{S.1.22}$$

where (1) holds by P having density $p_{x_0}(y)$ and P_r having density $\int_t p_{x_0+t}(y) \mu_r(t) dt$ with

$$\mu_r(t) := \frac{g(x_0 + t) I\{\|t\| < r\}}{F(r)} , \tag{S.1.23}$$

(2) by Fubini's Theorem, and (3) by defining the function

$$m_\phi(t) := \int_y \phi(y) p_{x_0+t}(y) \nu(dy)$$

and the random variable T with density μ_r .

Next, note that

$$\begin{aligned} m_\phi(t) - m_\phi(0) &= \int_y \phi(y) (p_{x_0+t}(y) - p_{x_0}(y)) \nu(dy) \\ &= \int_y \phi(y) \left((\sqrt{p_{x_0+t}(y)} - \sqrt{p_{x_0}(y)})^2 + 2(\sqrt{p_{x_0+t}(y)} - \sqrt{p_{x_0}(y)}) \sqrt{p_{x_0}(y)} \right) \nu(dy) \\ &\stackrel{(1)}{=} t' a_\phi + R_\phi(t) , \end{aligned} \tag{S.1.24}$$

where (1) holds by defining

$$a_\phi := \int_y \phi(y) \dot{\ell}_{x_0}(y) p_{x_0}(y) \nu(dy)$$

$$R_\phi(t) := \int_y \phi(y) \left((u_t(y))^2 + \frac{1}{4} (t' \dot{\ell}_{x_0}(y))^2 p_{x_0}(y) + u_t(y) (t' \dot{\ell}_{x_0}(y) + 2) \sqrt{p_{x_0}(y)} \right) \nu(dy)$$

with

$$u_t(y) := \sqrt{p_{x_0+t}(y)} - \sqrt{p_{x_0}(y)} - \frac{1}{2} t' \dot{\ell}_{x_0}(y) \sqrt{p_{x_0}(y)} .$$

By (S.1.23) and (S.1.24), we get

$$\begin{aligned} TV(P_r, P) &= \sup_{|\phi|_\infty \leq 1} |E[T]' a_\phi + E[R_\phi(T)]|/2 \\ &\leq \|E[T]\| \sup_{|\phi|_\infty \leq 1} \|a_\phi\|/2 + E[\sup_{|\phi|_\infty \leq 1} |R_\phi(T)|]/2 . \end{aligned} \quad (\text{S.1.25})$$

The desired result is a corollary of (S.1.25) and the next results:

$$\|E[T]\| = O(r^2) \quad (\text{S.1.26})$$

$$\sup_{|\phi|_\infty \leq 1} \|a_\phi\| = O(1) \quad (\text{S.1.27})$$

$$E[\sup_{|\phi|_\infty \leq 1} |R_\phi(T)|] = o(r) . \quad (\text{S.1.28})$$

To show (S.1.26), consider the following argument for any $r \in (0, \tilde{r})$,

$$\begin{aligned} \|E[T]\| &\stackrel{(1)}{=} \left\| \int_t \frac{t g(x_0 + t) I\{\|t\| < r\}}{F(r)} dt \right\| \\ &\stackrel{(2)}{=} \frac{1}{F(r)} \left\| \int_{\{\|t\| < r\}} t (g(x_0 + t) - g(x_0)) dt \right\| \\ &\stackrel{(3)}{\leq} \frac{C_g}{C_L r^d} \int_{\{\|t\| < r\}} \|t\|^2 dt \\ &\stackrel{(4)}{=} O(r^2) , \end{aligned} \quad (\text{S.1.29})$$

as desired, where (1) holds by (S.1.23) and (2) by $r < \tilde{r}$, which implies $B_r \cap \mathcal{X} = B_r$, and so

$$\int_{\{\|t\| < r\} \cap \{x_0 + t \in \mathcal{X}\}} t dt = \int_{\{\|t\| < r\}} t dt = 0 ,$$

(3) by Assumption 2 and Lemma 1, and (4) by $\int_{\{\|t\| < r\}} \|t\|^2 dt \leq O(r^{d+2})$.

To show (S.1.27), consider the following derivation:

$$\sup_{|\phi|_\infty \leq 1} \|a_\phi\| \stackrel{(1)}{\leq} \int_y \|\dot{\ell}_{x_0}(y)\| p_{x_0}(y) \nu(dy) \stackrel{(2)}{\leq} \sqrt{I(x_0)} , \quad (\text{S.1.30})$$

where (1) holds by $|\phi|_\infty \leq 1$, and (2) by the Cauchy–Schwarz inequality and Assumption 3, which implies $I(x_0) =: \int_y \|\dot{\ell}_{x_0}(y)\|^2 p_{x_0}(y) \nu(dy) < \infty$.

Finally, we show (S.1.28). To this end, consider the next argument for any $t \in \mathbb{R}^d$ with $\|t\| \leq r$,

$$\begin{aligned}
\sup_{|\phi|_\infty \leq 1} |R_\phi(t)| &\stackrel{(1)}{\leq} \int_y (u_t(y))^2 \nu(dy) + \frac{\|t\|^2}{4} \int_y \|\dot{\ell}_{x_0}(y)\|^2 p_{x_0}(y) \nu(dy) \\
&\quad + \|t\| \int_y |u_t(y)| \|\dot{\ell}_{x_0}(y)\| \sqrt{p_{x_0}(y)} \nu(dy) + 2 \int_y |u_t(y)| \sqrt{p_{x_0}(y)} \nu(dy) \\
&\stackrel{(2)}{\leq} \int_y (u_t(y))^2 \nu(dy) + \frac{\|t\|^2}{4} \int_y \|\dot{\ell}_{x_0}(y)\|^2 p_{x_0}(y) \nu(dy) \\
&\quad + \left(\int_y (u_t(y))^2 \nu(dy) \right)^{1/2} \left(\|t\| \left(\int_y \|\dot{\ell}_{x_0}(y)\|^2 p_{x_0}(y) \nu(dy) \right)^{1/2} + 2 \right) \\
&\stackrel{(3)}{=} o(\|t\|) \text{ as } \|t\| \rightarrow 0, \tag{S.1.31}
\end{aligned}$$

where (1) holds by $|\phi|_\infty \leq 1$, (2) by the Cauchy–Schwarz inequality, and (3) by Assumption 3, which implies that $(\int_y (u_t(y))^2 \nu(dy))^{1/2} = o(\|t\|)$ as $\|t\| \rightarrow 0$, and $\int_y \|\dot{\ell}_{x_0}(y)\|^2 p_{x_0}(y) \nu(dy) < \infty$. Since T is supported on $\{\|t\| \leq r\}$, (S.1.31) implies (S.1.28), concluding the proof. ■

Lemma S.4. *Let $U_{(k:n)}$ be the k -th order statistic from an i.i.d. sample $U_1, \dots, U_n \sim \text{U}(0, 1)$. For any real $m > 0$,*

$$E[U_{(k:n)}^m] = O((k/n)^m).$$

Proof. For any $m > 0$, $k \in \{1, \dots, n\}$, and $n \in \mathbf{N}$, consider the following derivation:

$$E[U_{(k:n)}^m] \stackrel{(1)}{=} \frac{B(m+k, n-k+1)}{B(k, n-k+1)} \stackrel{(2)}{=} \frac{\Gamma(m+k)/\Gamma(k)}{\Gamma(m+n+1)/\Gamma(n+1)}, \tag{S.1.32}$$

where (1) holds by David and Nagaraja (2003, p. 10-11), Abramowitz and Stegun (1964, 6.2.1, p. 257), where $B(a, b) = \int_0^1 t^{a-1} (1-t)^{b-1} dt : (0, \infty)^2 \rightarrow (0, \infty)$ is the Beta function, and (2) by Abramowitz and Stegun (1964, 6.2.2, p. 257), which gives $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$, where $\Gamma(a) = \int_0^\infty t^{a-1} \exp(-t) dt : (0, \infty) \rightarrow (0, \infty)$ is the Gamma function.

Next, we show that, for all $m > 0$, we can find constants $C_1(m), C_2(m)$ with $0 < C_1(m) < C_2(m) < \infty$ such that for all $t \in \mathbf{N}$,

$$0 < C_1(m) \leq \frac{\Gamma(m+t)}{t^m \Gamma(t)} \leq C_2(m) \leq \infty. \tag{S.1.33}$$

By definition, $\Gamma : (0, \infty) \rightarrow \mathbf{R}$ is continuous and positive. Therefore, for all $t \in \mathbf{N}$, $t^{-m} \Gamma(m+t)/\Gamma(t)$ is positive. Also, Abramowitz and Stegun (1964, 6.1.46, p. 257) implies that $\lim_{t \rightarrow \infty} t^{-m} \Gamma(m+t)/\Gamma(t) = 1$ for any $m \in \mathbf{R}$. Therefore, $\exists t_1 \in \mathbf{N}$ such that for all $t > t_1$, $t^{-m} \Gamma(m+t)/\Gamma(t) \in (1/2, 2)$. Then, for all $t \in \mathbf{N}$,

$$\frac{\Gamma(m+t)}{t^m \Gamma(t)} \leq \max \left\{ \max_{t \leq t_1} \frac{\Gamma(m+t)}{t^m \Gamma(t)}, \max_{t > t_1} \frac{\Gamma(m+t)}{t^m \Gamma(t)} \right\} \leq \max \left\{ \max_{t \leq t_1} \frac{\Gamma(m+t)}{t^m \Gamma(t)}, 2 \right\} := C_2(m) < \infty$$

and

$$\frac{\Gamma(m+t)}{t^m \Gamma(t)} \geq \min \left\{ \min_{t \leq t_1} \frac{\Gamma(m+t)}{t^m \Gamma(t)}, \min_{t > t_1} \frac{\Gamma(m+t)}{t^m \Gamma(t)} \right\} \geq \min \left\{ \min_{t \leq t_1} \frac{\Gamma(m+t)}{t^m \Gamma(t)}, \frac{1}{2} \right\} := C_1(m) > 0 ,$$

as desired in (S.1.33).

To conclude the proof, for any $k \in \{1, \dots, n\}$ and $n \in \mathbf{N}$, by (S.1.32),

$$E[U_{(k:n)}^m] = \frac{\Gamma(m+k)/\Gamma(k)}{\Gamma(m+n+1)/\Gamma(n+1)} .$$

Applying (S.1.33) with $t = k$ in the numerator and with $t = n+1$ in the denominator gives

$$E[U_{(k:n)}^m] \leq \frac{C_2(m)k^m}{C_1(m)(n+1)^m} = \frac{C_2(m)}{C_1(m)} \left(\frac{k}{n+1} \right)^m \leq \frac{C_2(m)}{C_1(m)} \left(\frac{k}{n} \right)^m .$$

Hence $E[U_{(k:n)}^m] = O((k/n)^m)$. ■

Lemma S.5. *Suppose $g(x_0) > 0$ and Assumption 1 holds with additional integrability condition: $\zeta_2 \in L^p(p_{x_0} \cdot \nu)$. Then, for $p \geq 1$,*

$$H(P_r, P) = O(r^{\min\{p, 2\}}) \quad \text{and} \quad \text{TV}(P_r, P) = O(r^2) .$$

Proof. By Assumption 1, there exist $\tilde{\zeta}_{2,v}$ and $\varepsilon_0 > 0, C_1 > 0$ uniformly for all $v \in \mathbf{R}^d$ with $\|v\| \leq \varepsilon_0$ and all $y \in \mathbf{R}^m$,

$$f(x_0 + v, y) = f(x_0, y) [1 + \zeta_1(y)^\top v + \|v\|^2 \tilde{\zeta}_{2,v}(y)], \quad |\tilde{\zeta}_{2,v}(y)| \leq C_1 \zeta_2(y) .$$

Integrating with respect to y gives

$$g(x_0 + v) = g(x_0) [1 + \bar{\zeta}_1^\top v + \bar{\zeta}_{2,v} \|v\|^2] ,$$

where

$$\bar{\zeta}_1 := \int \zeta_1(y) p_{x_0}(y) \nu(dy), \quad \bar{\zeta}_{2,v} := \int \tilde{\zeta}_{2,v}(y) p_{x_0}(y) \nu(dy) .$$

Since $\zeta_2(y) \in L^p(p_{x_0} \cdot \nu)$, $\bar{\zeta}_{2,v}$ is bounded uniformly for $\|v\| \leq \varepsilon_0$:

$$|\bar{\zeta}_{2,v}| \leq \int |\tilde{\zeta}_{2,v}(y)| p_{x_0}(y) \nu(dy) \leq C_1 \int \zeta_2(y) p_{x_0}(y) \nu(dy) =: C_1 \bar{\zeta}_2 < \infty .$$

Choose $r \in (0, \varepsilon_0)$. To obtain $h_r(y)$, we integrate the joint density over B_r . Let $B_r(0)$ be a ball of radius r around the origin. Since $x_0 \in \text{int}(\mathcal{X})$, by symmetry, the numerator is

$$\int_{B_r(0)} f(x_0 + v, y) dv = \text{Vol}(B_r) f(x_0, y) [1 + A_r(y)], \quad \text{where } A_r(y) := \frac{1}{\text{Vol}(B_r)} \int_{B_r(0)} \|v\|^2 \tilde{\zeta}_{2,v}(y) dv .$$

On $B_r(0)$, $\|v\|^2 \leq r^2$, so $A_r(y)$ is bounded from above:

$$|A_r(y)| \leq \frac{1}{\text{Vol}(B_r)} \int_{B_r(0)} \|v\|^2 |\tilde{\zeta}_{2,v}(y)| dv \leq \frac{1}{\text{Vol}(B_r)} \int_{B_r(0)} r^2 C_1 \zeta_2(y) dv = C_1 r^2 \zeta_2(y) . \quad (\text{S.1.34})$$

To get the denominator, integrate $g(x_0 + v)$ over $B_r(0)$. Then, by Tonelli's theorem, we have

$$\int_{B_r(0)} g(x_0 + v) dv = \int_{B_r(0)} \int f(x_0 + v, y) \nu(dy) dv = \text{Vol}(B_r) \int f(x_0, y) [1 + A_r(y)] \nu(dy) .$$

Therefore, we have

$$\int_{B_r(0)} g(x_0 + v) dv = \text{Vol}(B_r) g(x_0) [1 + \bar{A}_r] \quad \text{where} \quad \bar{A}_r := \frac{1}{g(x_0)} \int f(x_0, y) A_r(y) \nu(dy) ,$$

Using the upper bound $|A_r(y)| \leq C_1 r^2 \zeta_2(y)$, we have

$$|\bar{A}_r| \leq \frac{1}{g(x_0)} \int f(x_0, y) |A_r(y)| \nu(dy) \leq C_1 r^2 \int p_{x_0}(y) \zeta_2(y) \nu(dy) \leq C_1 r^2 \bar{\zeta}_2 .$$

Combining the denominator and the numerator, we get

$$h_r(y) = p_{x_0}(y) \frac{1 + A_r(y)}{1 + \bar{A}_r} .$$

Now for simplicity, define

$$\Delta_r(y) := \frac{h_r(y)}{p_{x_0}(y)} - 1 = \frac{A_r(y) - \bar{A}_r}{1 + \bar{A}_r} .$$

From the upper bound of $|\bar{A}_r|$, there exists $\varepsilon_1 \in (0, \varepsilon_0)$ such that, for any $r \in (0, \varepsilon_1)$, $|\bar{A}_r| \leq \frac{1}{2}$, so

$$|\Delta_r(y)| \leq 2(|A_r(y)| + |\bar{A}_r|) \leq 2C_1 r^2 (\zeta_2(y) + \bar{\zeta}_2) =: C_2 r^2 u(y) \quad (\text{S.1.35})$$

where $C_2 := 2C_1 > 0$ and $u(y) := \zeta_2(y) + \bar{\zeta}_2$. Then, the $L^p(p_{x_0} \cdot \nu)$ -integrability condition of ζ_2 implies that $u \in L^p(p_{x_0} \cdot \nu)$, i.e.,

$$\|u\|_{L^p(p_{x_0} \cdot \nu)}^p := \int |\zeta_2(y) + \bar{\zeta}_2|^p p_{x_0}(y) \nu(dy) \leq 2^{p-1} \int (|\zeta_2(y)|^p + |\bar{\zeta}_2|^p) p_{x_0}(y) \nu(dy) < \infty , \quad (\text{S.1.36})$$

as $(\frac{a+b}{2})^p \leq \frac{1}{2}(a^p + b^p)$ holds by convexity for any $p \geq 1$ and $a, b \geq 0$.

Rate of $H(P_r, P)$. Under our assumptions, $h_0(y) := \lim_{r \rightarrow 0} h_r(y) = p_{x_0}(y)$ for all $y \in \mathcal{Y}$. We use the inequality: for $s > -1$, $(\sqrt{1+s} - 1)^2 \leq \min\{s^2, |s|\}$, which gives

$$\left(\sqrt{h_r(y)} - \sqrt{h_0(y)} \right)^2 = \left(\sqrt{1 + \Delta_r(y)} - 1 \right)^2 h_0(y) \leq \min\{\Delta_r^2(y), |\Delta_r(y)|\} h_0(y) . \quad (\text{S.1.37})$$

Case $1 \leq p \leq 2$. For simplicity, let $\tau_r := r^{-2}$. Define $\mathcal{A}_r := \{y \in \mathcal{Y} : u(y) \leq \tau_r\}$ and $\mathcal{A}_r^c := \mathcal{Y} \setminus \mathcal{A}_r$.

By (S.1.37), we have

$$\begin{aligned}
\mathsf{H}(P_r, P)^2 &\stackrel{(1)}{\leq} \frac{1}{2} \int \min\{\Delta_r^2(y), |\Delta_r(y)|\} h_0(y) \nu(dy) \\
&\stackrel{(2)}{\leq} \frac{1}{2} \int_{\mathcal{A}_r} \Delta_r^2(y) h_0(y) \nu(dy) + \frac{1}{2} \int_{\mathcal{A}_r^c} |\Delta_r(y)| h_0(y) \nu(dy) \\
&\stackrel{(3)}{\leq} \frac{C_2^2 r^4}{2} \int_{\mathcal{A}_r} u(y)^2 h_0(y) \nu(dy) + \frac{C_2 r^2}{2} \int_{\mathcal{A}_r^c} u(y) h_0(y) \nu(dy) \\
&\stackrel{(4)}{\leq} \frac{C_2^2 r^4}{2} \int_{\mathcal{A}_r} \tau_r^{2-p} u^p(y) h_0(y) \nu(dy) + \frac{C_2 r^2}{2} \int_{\mathcal{A}_r^c} \tau_r^{1-p} u^p(y) h_0(y) \nu(dy) \\
&\stackrel{(5)}{\leq} \frac{C_2^2 r^{2p}}{2} \int u^p(y) h_0(y) \nu(dy) + \frac{C_2 r^{2p}}{2} \int u^p(y) h_0(y) \nu(dy) = r^{2p} \frac{C_2^2 + C_2}{2} \|u\|_{L^p(p_{x_0} \cdot \nu)}^p,
\end{aligned}$$

where (1) follows from (S.1.37), we have (2) by splitting \mathcal{Y} into \mathcal{A}_r and \mathcal{A}_r^c and by the definition of minimum, (3) follows from (S.1.35), (4) follows from

$$\begin{aligned}
\frac{u(y)}{\tau_r} \leq 1 &\rightarrow \left(\frac{u(y)}{\tau_r}\right)^{2-p} \leq 1 \rightarrow u(y)^2 \leq \tau_r^{2-p} u^p(y) \quad \text{for any } y \in \mathcal{A}_r, \\
\frac{u(y)}{\tau_r} > 1 &\rightarrow \left(\frac{u(y)}{\tau_r}\right)^{1-p} \leq 1 \rightarrow u(y) \leq \tau_r^{1-p} u^p(y) \quad \text{for any } y \in \mathcal{A}_r^c,
\end{aligned}$$

and lastly (5) holds by replacing \mathcal{A}_r and \mathcal{A}_r^c with \mathcal{Y} respectively and by substituting $\tau_r = r^{-2}$. By (S.1.36), $\mathsf{H}(P_r, P) = O(r^p)$.

Case $p \geq 2$. Similar as before, by using (S.1.37) and $(\sqrt{1+s}-1)^2 \leq s^2$, we have

$$\mathsf{H}(P_r, P)^2 \leq \frac{1}{2} \int \Delta_r^2(y) h_0(y) \nu(dy) \leq \frac{C_2^2 r^4}{2} \int u(y)^2 h_0(y) \nu(dy) = \frac{r^4 C_2^2}{2} \|u\|_{L^2(p_{x_0} \cdot \nu)}^2,$$

where the second inequality follows from (S.1.35). Since $p \geq 2$ and $u \in L^p(p_{x_0} \cdot \nu)$ as shown in (S.1.36), the Lyapunov's inequality gives $u \in L^2(p_{x_0} \cdot \nu)$ as well. Thus, $\mathsf{H}(P_r, P) = O(r^2)$. Combining the two cases, we conclude $\mathsf{H}(P_r, P) = O(r^{\min\{p, 2\}})$.

Rate of $\text{TV}(P_r, P)$. The total variation distance between P_r and P is

$$\begin{aligned}
\text{TV}(P_r, P) &= \frac{1}{2} \int |h_r(y) - h_0(y)| \nu(dy) = \frac{1}{2} \int |\Delta_r(y)| h_0(y) \nu(dy) \\
&\leq \frac{C_2 r^2}{2} \int u(y) h_0(y) \nu(dy) = \frac{C_2 r^2}{2} \|u\|_{L^1(p_{x_0} \cdot \nu)}.
\end{aligned}$$

Since $\|u\|_{L^1(p_{x_0} \cdot \nu)} < \infty$, we have $\text{TV}(P_r, P) = O(r^2)$ as $r \rightarrow 0$. ■

Lemma S.6. *Assumption 1 implies Assumption 2 with $x_0 \in \text{int}(\mathcal{X})$ and Assumption 3. The converse is not true.*

Proof. By Assumption 1, there exist $\tilde{\zeta}_{2,v}$ and $C_0, \varepsilon_0 > 0$ uniformly for all $v \in \mathbf{R}^d$ with $\|v\| \leq \varepsilon_0$

and all $y \in \mathbf{R}^m$,

$$f(x_0 + v, y) = f(x_0, y)[1 + \zeta_1(y)^\top v + \|v\|^2 \tilde{\zeta}_{2,v}(y)], \quad |\tilde{\zeta}_{2,v}(y)| \leq C_0 \zeta_2(y). \quad (\text{S.1.38})$$

Integrating with respect to y gives

$$g(x_0 + v) = g(x_0)[1 + \bar{\zeta}_1^\top v + \bar{\zeta}_{2,v} \|v\|^2] \quad (\text{S.1.39})$$

where

$$\bar{\zeta}_1 := \int \zeta_1(y) p_{x_0}(y) \nu(dy), \quad \bar{\zeta}_{2,v} := \int \tilde{\zeta}_{2,v}(y) p_{x_0}(y) \nu(dy).$$

To see Assumption 2 holds, note that since $x_0 \in \text{int}(\mathcal{X})$, $B_r \subseteq \mathcal{X}$ for all small r , and so $\text{Vol}(B_r \cap \mathcal{X}) = \text{Vol}(B_r)$ is of order r^d . In addition, since $\zeta_2(y) \in L^2(p_{x_0} \cdot \nu)$, $\bar{\zeta}_{2,v}$ is bounded uniformly for $\|v\| \leq \varepsilon_0$:

$$|\bar{\zeta}_{2,v}| \leq C_0 \int \zeta_2(y) p_{x_0}(y) \nu(dy) =: \bar{\zeta}_2 < \infty.$$

Let $\varepsilon_1 := \min\{1, \varepsilon_0\}$. Then, for all v with $\|v\| \leq \varepsilon_1$

$$|g(x_0 + v) - g(x_0)| \stackrel{(a)}{=} |g(x_0) \bar{\zeta}_1^\top v + g(x_0) \bar{\zeta}_{2,v} \|v\|^2| \stackrel{(b)}{\leq} g(x_0) \{ \|\bar{\zeta}_1\| \cdot \|v\| + |\bar{\zeta}_{2,v}| \|v\|^2 \} \stackrel{(c)}{\leq} C_g \|v\|$$

where (a) follows from (S.1.39), (b) holds by Cauchy-Schwarz and the triangular inequality and (c) holds because $\|v\|^2 \leq \|v\|$ and by setting $C_g := g(x_0)(\|\bar{\zeta}_1\| + \bar{\zeta}_2)$.

Next to show that Assumption 3 holds, let

$$\dot{\ell}_{x_0}(y) := \zeta_1(y) - \bar{\zeta}_1, \quad a_v(y) := \zeta_1(y)^\top v + \tilde{\zeta}_{2,v}(y) \|v\|^2, \quad b_v := \sqrt{\frac{1}{1 + \bar{\zeta}_1^\top v + \bar{\zeta}_{2,v} \|v\|^2}}.$$

Combining (S.1.38) and (S.1.39), we obtain $\sqrt{p_{x_0+v}(y)} = \sqrt{p_{x_0}(y)} b_v \sqrt{1 + a_v(y)}$ holds. Define

$$\Delta_v(y) := \frac{\sqrt{p_{x_0+v}(y)} - \sqrt{p_{x_0}(y)}}{\|v\|} - \frac{1}{2} \dot{\ell}_{x_0}(y)^\top \frac{v}{\|v\|} \sqrt{p_{x_0}(y)} = \sqrt{p_{x_0}(y)} \{T_{1,v}(y) + T_{2,v}(y)\},$$

where

$$T_{1,v}(y) := \frac{b_v}{\|v\|} \left(\sqrt{1 + a_v(y)} - 1 - \frac{1}{2} a_v(y) \right) \quad (\text{S.1.40})$$

$$T_{2,v}(y) := \frac{b_v - 1}{\|v\|} + \frac{1}{2} \bar{\zeta}_1^\top \frac{v}{\|v\|} + \frac{b_v}{2\|v\|} a_v(y) - \frac{1}{2} \zeta_1(y)^\top \frac{v}{\|v\|}. \quad (\text{S.1.41})$$

Below, we show that the integrals of the squares of $\sqrt{p_{x_0}} T_{2,v}$ and $\sqrt{p_{x_0}} T_{1,v}$ are $o(1)$.

Term $T_{2,v}$. Applying the Taylor expansion $(1 + t)^{-\frac{1}{2}} = 1 - \frac{1}{2}t + O(t^2)$ around 0 to b_v gives

$$b_v = 1 - \frac{1}{2}(\bar{\zeta}_1^\top v + \bar{\zeta}_{2,v} \|v\|^2) + O((\bar{\zeta}_1^\top v + \bar{\zeta}_{2,v} \|v\|^2)^2) = 1 - \frac{1}{2}(\bar{\zeta}_1^\top v + \bar{\zeta}_{2,v} \|v\|^2) + O(\|v\|^2).$$

The second equality holds because, for some constants $\bar{C}, \bar{\varepsilon} > 0$ uniformly for $\|v\| < \min(\varepsilon_1, \bar{\varepsilon})$,

$$O((\bar{\zeta}_1^\top v + \bar{\zeta}_{2,v}\|v\|^2)^2) \leq 2\bar{C}((\bar{\zeta}_1^\top v)^2 + (\bar{\zeta}_{2,v}\|v\|^2)^2) \leq 2\bar{C}(\|\bar{\zeta}_1\|^2\|v\|^2 + |\bar{\zeta}_{2,v}|^2\|v\|^4) = O(\|v\|^2)$$

where the two inequalities follow from $(a+b)^2 \leq 2(a^2+b^2)$ and the Cauchy-Schwarz inequality, and the last equality holds because $\|\bar{\zeta}_1\|^2 < \infty$ and $|\bar{\zeta}_{2,v}|^2 < \bar{\zeta}_2^2$ uniformly for $\|v\| < \min(\varepsilon_1, \bar{\varepsilon})$. Thus, there exist $C_1, C_2 > 0$, and $\varepsilon_2 \in (0, \varepsilon_1)$ such that, uniformly for all v with $\|v\| < \varepsilon_2$

$$\left| \frac{b_v - 1}{\|v\|} + \frac{1}{2}\bar{\zeta}_1^\top \frac{v}{\|v\|} \right| \leq \frac{1}{2}|\bar{\zeta}_{2,v}|\|v\| + O(\|v\|) \leq C_1\|v\|, \quad (\text{S.1.42})$$

$$|b_v - 1| \leq \frac{1}{2}(\|\bar{\zeta}_1\|\|v\| + |\bar{\zeta}_{2,v}|\|v\|^2) + O(\|v\|^2) \leq C_2\|v\|, \quad (\text{S.1.43})$$

$$b_v^2 \leq \max\{(1 + C_2\varepsilon_2)^2, (1 - C_2\varepsilon_2)^2\} =: \bar{b}. \quad (\text{S.1.44})$$

The term $T_{2,v}(y)^2$ is bounded uniformly for all v with $\|v\| < \varepsilon_2$ as follows:

$$\begin{aligned} T_{2,v}(y)^2 &\leq 2 \left(\frac{b_v - 1}{\|v\|} + \frac{1}{2}\bar{\zeta}_1^\top \frac{v}{\|v\|} \right)^2 + 2 \left(\frac{b_v}{2\|v\|}a_v(y) - \frac{1}{2}\zeta_1(y)^\top \frac{v}{\|v\|} \right)^2 \\ &\leq 2C_1^2\|v\|^2 + 2 \left(\frac{(b_v - 1)}{2\|v\|}\zeta_1(y)^\top v + \frac{b_v\|v\|}{2}\tilde{\zeta}_{2,v}(y) \right)^2 \\ &\leq 2C_1^2\|v\|^2 + (b_v - 1)^2\|\zeta_1(y)\|^2 + b_v^2\|v\|^2\tilde{\zeta}_{2,v}(y)^2 \\ &\leq C_3\|v\|^2 \{1 + \|\zeta_1(y)\|^2 + C_0^2\zeta_2(y)^2\} \end{aligned}$$

where the first inequality holds by $(a+b)^2 \leq 2a^2 + 2b^2$, the second inequality holds by (S.1.42) and by plugging in $a_v(y)$, the third inequality holds by $(a+b)^2 \leq 2a^2 + 2b^2$ and the Cauchy-Schwarz inequality, and the final inequality follows from (S.1.43), (S.1.44), setting $C_3 := \max\{2C_1^2, C_2^2, \bar{b}\}$, and using $\tilde{\zeta}_{2,v}(y)^2 \leq C_0^2\zeta_2(y)^2$. Therefore, uniformly for all v with $\|v\| < \varepsilon_2$

$$\int T_{2,v}(y)^2 p_{x_0}(y) \nu(dy) \leq C_3 \left\{ 1 + \int (\|\zeta_1(y)\|^2 + C_0^2\zeta_2(y)^2) p_{x_0}(y) \nu(dy) \right\} \|v\|^2 =: C_4\|v\|^2.$$

Term $T_{1,v}$. For $u \geq -1$ the following inequality holds:

$$\left| \sqrt{1+u} - 1 - \frac{1}{2}u \right| \leq C_5|u| \min\{|u|, 1\} \quad (\text{S.1.45})$$

for some constant $C_5 > 0$. We apply this with $u = a_v(y)$ since $a_v(y) = f(x_0+v, y)/f(x_0, y) - 1 > -1$ for all y such that $p_{x_0}(y) > 0$. Partition \mathcal{Y} into

$$E_1 := \{y \in \mathcal{Y} : |a_v(y)| \leq \delta_v\}, \quad E_2 := \{y \in \mathcal{Y} : |a_v(y)| > \delta_v\}, \quad \text{with} \quad \delta_v := \sqrt{\|v\|}.$$

On E_1 , by applying (S.1.45) and $\min\{|a_v(y)|, 1\} \leq |a_v(y)|$

$$T_{1,v}(y)^2 = \frac{b_v^2}{\|v\|^2} \left| \sqrt{1+a_v(y)} - 1 - \frac{1}{2}a_v(y) \right|^2 \leq \frac{b_v^2 C_5^2}{\|v\|^2} |a_v(y)|^4 \leq C_5^2 b_v^2 \delta_v^2 \left| \frac{a_v(y)}{\|v\|} \right|^2.$$

By the triangle inequality and Cauchy-Schwarz,

$$\left| \frac{a_v(y)}{\|v\|} \right| = \left| \zeta_1(y)^\top \frac{v}{\|v\|} + \tilde{\zeta}_{2,v}(y) \|v\| \right| \leq \|\zeta_1(y)\| + |\tilde{\zeta}_{2,v}(y)| \|v\| \leq \|\zeta_1(y)\| + C_0 |\zeta_2(y)| \|v\|, \quad (\text{S.1.46})$$

uniformly for all v with $\|v\| < \varepsilon_2$. Then for all v with $\|v\| < \varepsilon_2$

$$\int_{E_1} T_{1,v}(y)^2 p_{x_0}(y) \nu(dy) \leq C_5^2 b_v^2 \|v\| \int 2(\|\zeta_1(y)\|^2 + C_0^2 |\zeta_2(y)|^2 \|v\|^2) p_{x_0}(y) \nu(dy) \leq C_6 \|v\| + C_7 \|v\|^3$$

where we use by $(a+b)^2 \leq 2a^2 + 2b^2$ for the first inequality and the second inequality follows from (S.1.44) and setting $C_6 := 2C_5^2 \bar{b} \int \|\zeta_1(y)\|^2 p_{x_0}(y) \nu(dy)$ and $C_7 := 2C_5^2 \bar{b} C_0^2 \int |\zeta_2(y)|^2 p_{x_0}(y) \nu(dy)$.

Since $\min\{|a_v(y)|, 1\} \leq 1$, we have, for all v with $\|v\| < \varepsilon_2$,

$$T_{1,v}(y)^2 \leq C_5^2 b_v^2 \left| \frac{a_v(y)}{\|v\|} \right|^2 \leq 2C_5^2 \bar{b} \{ \|\zeta_1(y)\|^2 + C_0^2 |\zeta_2(y)|^2 \|v\|^2 \},$$

where the second inequality holds by applying $(a+b)^2 \leq 2a^2 + 2b^2$ to (S.1.46) and by using $b_v^2 \leq \bar{b}$ in (S.1.44). Integrating this on E_2 yields that uniformly for all v with $\|v\| < \varepsilon_2$,

$$\int_{E_2} T_{1,v}(y)^2 p_{x_0}(y) \nu(dy) \leq 2C_5^2 \bar{b} \int_{E_2} \|\zeta_1(y)\|^2 p_{x_0}(y) \nu(dy) + C_8 \|v\|^2,$$

where $C_8 := 2C_5^2 \bar{b} C_0^2 \int |\zeta_2(y)|^2 p_{x_0}(y) \nu(dy)$. We finally claim that

$$\int_{E_2} \|\zeta_1(y)\|^2 p_{x_0}(y) \nu(dy) = o(1). \quad (\text{S.1.47})$$

To see this, fix y with $p_{x_0}(y) > 0$. It holds from (S.1.46) that $|a_v(y)|/\delta_v \leq \|\zeta_1(y)\| \|v\|^{1/2} + C_0 |\zeta_2(y)| \|v\|^{3/2} \rightarrow 0$ as $\|v\| \rightarrow 0$. Therefore, $1_{[y \in E_2]} \rightarrow 0$. By dominated convergence, (S.1.47) holds where we used the integrability of $\|\zeta_1\|^2$. Therefore, for an arbitrary $\eta > 0$, there exists $\varepsilon_3 \in (0, \varepsilon_2)$ such that uniformly for all v with $\|v\| < \varepsilon_3$,

$$\int_{E_2} T_{1,v}(y)^2 p_{x_0}(y) \nu(dy) \leq \eta + C_8 \|v\|^2.$$

Combining the two regions, we have, uniformly for all v with $\|v\| < \varepsilon_3$,

$$\int T_{1,v}(y)^2 p_{x_0}(y) \nu(dy) \leq C_6 \|v\| + C_7 \|v\|^3 + C_8 \|v\|^2 + \eta.$$

To summarize, it holds that, uniformly for all v with $\|v\| < \varepsilon_2$,

$$\int \Delta_v(y)^2 \nu(dy) \leq 2 \int T_{1,v}^2(y) p_{x_0}(y) \nu(dy) + 2 \int T_{2,v}^2(y) p_{x_0}(y) \nu(dy) \leq 2C_6 \|v\| + C_9 \|v\|^2 + 2C_7 \|v\|^3 + 2\eta.$$

where $C_9 := 2(C_4 + C_8)$. Letting $\eta \rightarrow 0$, the upper bound converges to 0 as $\|v\| \rightarrow 0$ and so

quadratic mean differentiability holds. Since

$$\int \|\dot{\ell}_{x_0}(y)\|^2 p_{x_0}(y) \nu(dy) \leq 2 \int \|\zeta_1(y)\|^2 p_{x_0}(y) \nu(dy) + 2\bar{\zeta}_1^2 < \infty ,$$

the integrability condition also holds. Therefore, Assumption 1 implies 3.

Finally, unlike Assumption 3, Assumption 1 imposes a support restriction. If $f(x_0, y_*) = 0$ for some y_* , then $f(x_0 + \varepsilon, y_*) = 0$ for all $\|\varepsilon\| \leq \varepsilon_0$. Thus, the converse implication does not hold in general. To provide a simple example, let $d = 1$, $x_0 = 0$, and $y \in \{0, 1\}$, and let ν be the counting measure on $\{0, 1\}$. For $x \in \mathbf{R}$ with $|x|$ small, set

$$p_x(1) := |x|^3, \quad p_x(0) := 1 - |x|^3 .$$

This is a valid pmf (nonnegative and summing to one) for small $|x|$. To see that Assumption 3 is satisfied, take $\dot{\ell}_{x_0}(y) := 0 \in \mathbf{R}$. Then, writing the $L^2(\nu)$ expression as a sum,

$$\sum_{y \in \{0,1\}} \left[\frac{\sqrt{p_x(y)} - \sqrt{p_{x_0}(y)}}{|x|} - \frac{1}{2} \dot{\ell}_{x_0}(y) \sqrt{p_{x_0}(y)} \right]^2 = \left(\frac{\sqrt{1 - |x|^3} - 1}{|x|} \right)^2 + \left(\frac{|x|^{3/2}}{|x|} \right)^2 .$$

Using $\sqrt{1-t} = 1 - \frac{1}{2}t + O(t^2)$ as $t \rightarrow 0$ with $t = |x|^3$,

$$\left(\frac{\sqrt{1 - |x|^3} - 1}{|x|} \right)^2 = \left(\frac{-\frac{1}{2}|x|^3 + O(|x|^6)}{|x|} \right)^2 = O(|x|^4) \rightarrow 0, \quad \left(\frac{|x|^{3/2}}{|x|} \right)^2 = |x| \rightarrow 0 .$$

Hence the QMD condition holds. Moreover, $\int \|\dot{\ell}_{x_0}(y)\|^2 p_{x_0}(y) \nu(dy) = 0 < \infty$. However, at $y = 1$ we have $f(x_0, 1) = p_{x_0}(1) = 0$, but for every small $x \neq 0$, $f(x, 1) = p_x(1) = |x|^3 > 0$, which is a violation of Assumption 1. ■

Appendix S.2 Results under Alternative Assumptions

The main results of the paper are derived under differentiability in quadratic mean (QMD); see Assumption 3. In this appendix, we study an alternative, perhaps more classical differentiability condition: a Taylor expansion of the joint density $f(\cdot, y)$ around x_0 with a Hölder remainder (Assumption S.4), which allows for a wide range of smoothness levels and remainder controls. This condition may appear similar in spirit to the multiplicative expansion in Assumption 1, since the latter is closely related to the case $(\kappa_s, \kappa_r) = (1, 1)$. We will show, however, that Assumption 1 is much stronger: beyond smoothness and integrability, it imposes a restrictive multiplicative structure (and associated local support restrictions) that is not required under QMD or Assumption S.4. We begin by stating the smoothness with Hölder remainder condition that underpins the results in this section.

Assumption S.4 (Smoothness with Hölder reminder at x_0). Fix $\kappa_s \in \mathbf{N}_0$ and $\kappa_r \in (0, 1]$. The model class **Q** consists of all laws Q such that there exists $\tilde{r} > 0$ for which, for every $y \in \mathbf{R}^m$ and every t with $\|t\| \leq \tilde{r}$ and $x_0 + t \in \mathcal{X}$,

$$f(x_0 + t, y) = \sum_{\|\beta\|_1 \leq \kappa_s} \frac{\partial_x^\beta f(x_0, y)}{\beta!} t^\beta + \zeta_{\kappa_s + \kappa_r}^{\text{rem}}(t, y), \quad |\zeta_{\kappa_s + \kappa_r}^{\text{rem}}(t, y)| \leq M(y) \|t\|^{\kappa_s + \kappa_r}. \quad (\text{S.2.1})$$

Here $\beta = (\beta_1, \dots, \beta_d)$ is a multi-index,

$$\partial_x^\beta f(x_0, y) := \frac{\partial^{\|\beta\|_1} f}{\partial x_1^{\beta_1} \dots \partial x_d^{\beta_d}}(x_0, y), \quad t^\beta := \prod_{j=1}^d t_j^{\beta_j}, \quad \beta! := \prod_{j=1}^d \beta_j!.$$

Moreover,

$$\int M(y) \nu(dy) < \infty, \quad \int |\partial_x^\beta f(x_0, y)| \nu(dy) < \infty \quad \text{for all } \|\beta\|_1 \leq \kappa_s. \quad (\text{S.2.2})$$

Assumption S.4 has several ingredients. First, it requires that the joint density $f(\cdot, y)$ admits a Taylor series expansion of order κ_s around x_0 , with a remainder that is controlled by Hölder continuity of order κ_r . Second, the bound on the remainder term through the envelope $M(y)$ allows the quality of approximation to depend on y , but ensures uniform control in t once $M(y)$ is integrable. Finally, the integrability conditions guarantee that both the derivatives and the envelope are well behaved relative to the dominating measure ν , so that expansions can be integrated and used in distributional approximations. Together, these conditions provide a flexible smoothness requirement that interpolates between differentiability of various orders and Hölder continuity.

There are several features of Assumption S.4 worth highlighting.

Monotonicity in κ_s . For any fixed $\kappa_r \in (0, 1]$, Assumption S.4 becomes stronger as κ_s increases. In particular, if it holds with $(\kappa_s + 1, \kappa_r)$, then it automatically holds with (κ_s, κ_r) .

The case $\kappa_s = 0$. When $\kappa_s = 0$, the condition reduces to Hölder continuity,

$$|f(x, y) - f(x_0, y)| \leq M(y) \|x - x_0\|^{\kappa_r},$$

with $\kappa_r \in (0, 1]$. In particular, $\kappa_r = 1$ yields (local) Lipschitz continuity.

The case $\kappa_s = 1$. When $\kappa_s = 1$, Assumption S.4 takes the form

$$|f(x_0 + t, y) - f(x_0, y) - \nabla_x f(x_0, y)^\top t| \leq M(y) \|t\|^{1 + \kappa_r},$$

i.e., pointwise differentiability of f in x at x_0 with a Hölder remainder of order κ_r . Heuristically, while Assumption 3 imposes differentiability in quadratic mean of $x \mapsto \sqrt{p_x}$ at x_0 , Assumption S.4 requires pointwise differentiability of the *joint* density. Neither assumption implies the other.

The main result of this section is as follows.

Theorem S.1. *Let Assumptions 2 and S.4 hold. Then,*

$$\text{TV}(P_r, P) = O(r^{\min\{1, \kappa_s + \kappa_r\}}) \quad \text{and} \quad \text{H}(P_r, P) = O(r^{\min\{1/2, (\kappa_s + \kappa_r)/2\}}) .$$

Moreover, if x_0 is an interior point of \mathcal{X} ,

$$\text{TV}(P_r, P) = O(r^{\min\{2, \kappa_s + \kappa_r\}}) \quad \text{and} \quad \text{H}(P_r, P) = O(r^{\min\{1, (\kappa_s + \kappa_r)/2\}}) .$$

Theorem S.1 provides the marginal convergence rates implied by Assumption S.4. These rates contrast with those obtained under the QMD condition in Theorem 3: the two sets of primitives generate distinct marginal exponents (a_h, a_{tv}) , which then determine the joint rates through Theorem 2. Under Hölder smoothness, the exponents depend on the combined smoothness order $\kappa_s + \kappa_r$ and on whether x_0 is an interior point of \mathcal{X} . In general,

$$a_h = \min\{1/2, (\kappa_s + \kappa_r)/2\} \quad \text{and} \quad a_{tv} = \min\{1, \kappa_s + \kappa_r\} , \quad (\text{S.2.3})$$

and if x_0 is an interior point of \mathcal{X} ,

$$a_h = \min\{1, (\kappa_s + \kappa_r)/2\} \quad \text{and} \quad a_{tv} = \min\{2, \kappa_s + \kappa_r\} . \quad (\text{S.2.4})$$

In both (S.2.3) and (S.2.4), one has $a_{tv} = 2a_h$, so TV converges (at most) twice as fast as Hellinger.

The effect of whether x_0 is an interior point of \mathcal{X} or not is analogous to the one that occurs in kernel regression estimation under twice continuous differentiability. When x_0 lies in the interior of \mathcal{X} , the first moment is zero and the second term of the Taylor series expansion becomes the leading term. The same phenomenon explains the differences between (S.2.3) and (S.2.4). As $\kappa_s + \kappa_r$ increases, the Hellinger rate saturates at r (interior case) or at $r^{1/2}$ (non-interior case), while the TV rate saturates at r^2 (interior case) or at r (non-interior case). In contrast, under QMD, both metrics necessarily decay at a rate r and do not exhibit this interior/boundary split.

Remark S.4. *The rates in Theorem S.1 can be shown to be sharp by characterizing DGPs that satisfy the conditions of the theorem and lead to the rates described in each of the cases. This sharpness result thus shows that these marginal rates cannot be improved under Assumption S.4. We omit the details here for brevity. ■*

Combining Theorem S.1 with Theorem 2 in the case with $\kappa_s + \kappa_r = 2$ and x_0 belongs in the interior of \mathcal{X} yields

$$\text{H}(\mathcal{L}(S_n), \mathcal{L}(S)) = O(k^{1/2}(k/n)^{1/d}), \quad \text{TV}(\mathcal{L}(S_n), \mathcal{L}(S)) = O(k(k/n)^{2/d}) . \quad (\text{S.2.5})$$

Here $a_{tv} = 2$ and $a_h = 1$, so the direct bound $k(k/n)^{a_{tv}/d}$ from Theorem 2 determines the TV rate, rather than the indirect inequality $\text{TV} \leq \sqrt{2}\text{H}$. As in the marginal case, the TV distance converges

twice as fast as the Hellinger distance. Moreover, both expressions imply the same growth condition on k for joint convergence:

$$k(k/n)^{2/d} \rightarrow 0 \quad \iff \quad k = o(n^{2/(2+d)}) .$$

For $d = 1$ this becomes $k = o(n^{2/3})$, matching the threshold obtained in Section 3.1. If instead $\kappa_s + \kappa_r = 2$ and x_0 does not belong in the interior of \mathcal{X} , the marginal exponents are $a_h = 1/2$ and $a_{tv} = 1$, and Theorem 2 yields the slower joint rates

$$\mathbb{H}(\mathcal{L}(S_n), \mathcal{L}(S)) = O(k^{1/2}(k/n)^{1/(2d)}), \quad \text{TV}(\mathcal{L}(S_n), \mathcal{L}(S)) = O(k(k/n)^{1/d}),$$

which require $k = o(n^{1/(1+d)})$. More generally, as $\kappa_s + \kappa_r$ decreases, the admissible growth of k slows accordingly, and the slowdown is more pronounced in the non-interior case where the linear term does not cancel.

Remark S.5. *The L^1 integrability requirement $\int M(y) \nu(dy) < \infty$ in Assumption S.4 may be strengthened to an L^p condition for some $p > 1$. Such a strengthening does not, by itself, improve the convergence rates for $\mathbb{H}(P_r, P)$ or $\text{TV}(P_r, P)$ obtained in Theorem S.1. This parallels the result in Lemma S.5. ■*

S.2.1 Auxiliary Results under Alternative Assumptions

Assumption 1 in Falk et al. (2010) implies the approximation error $\mathbb{H}(P_r, P) = O(r^2)$, a rate that is substantially faster than the feasible rates delivered by Theorem S.1, which are sharp. Because Assumption 1 is closely related to Assumption S.4 with $\kappa_s = \kappa_r = 1$, it is natural to examine more carefully which aspects of these conditions drive such a discrepancy in rates. In order to keep the discussion as simple as possible, in this section we focus on the univariate case $d = m = 1$ and x_0 in the interior of the support of X .

The main takeaway is that Assumption 1 imposes a multiplicative structure on the local approximation to $f(x, y)$ (or $p_x(y)$) that automatically restricts the behavior of $f(x, y)$ (or $p_x(y)$) in a neighborhood of the zero-density region

$$\mathcal{N} := \{ y \in \mathcal{Y} : f(x_0, y) = 0 \} .$$

Using similar arguments to those in the proof of Theorem S.1, it is possible to show that if Assumption S.4 holds with $\kappa_s = \kappa_r = d = 1$ and $M(y)/f(x_0, y) \in L^2(p_{x_0} \cdot \nu)$, then

$$\mathbb{H}(P_r, P) = \begin{cases} O(r^2) & \text{if } \nu(\mathcal{N}) = 0 , \\ O(r) & \text{otherwise ,} \end{cases} \quad (\text{S.2.6})$$

so that the baseline linear rate can only be improved by strengthening the integrability condition

in Assumption S.4 and imposing $\nu(\mathcal{N}) = 0$. Effectively, these additional assumptions close the gap between Assumption 1 and Assumption S.4, yielding the same fast rate. This highlights the central role of the zero-density region \mathcal{N} in determining whether higher integrability translates into faster convergence. The details behind the result in (S.2.6) are omitted for brevity.

S.2.2 Proof of Theorem S.1

By Assumption 2(ii), for any $r \in (0, r_0)$,

$$\text{Vol}(B_r \cap \mathcal{X}) = \int_{B_r \cap \mathcal{X}} dx \geq C_{x_0} r^d > 0. \quad (\text{S.2.7})$$

Let r_1 denote the constant r in Assumption S.4. For the rest of the proof, let $r_2 = \min\{r_0, r_1\} > 0$.

To get the result for $\text{TV}(P_r, P)$, we derive expansions for any $r \in (0, r_2)$. Let

$$\mathcal{I}_{\kappa_s}^{(1)} := \{\beta \in \mathbf{N}_0^d : 1 \leq \|\beta\|_1 \leq \kappa_s\}, \quad \mathcal{I}_{\kappa_s}^{(2)} := \{\beta \in \mathbf{N}_0^d : 2 \leq \|\beta\|_1 \leq \kappa_s\}.$$

First, we have

$$\begin{aligned} \int_{B_r \cap \mathcal{X}} f(x, y) dx &\stackrel{(1)}{=} \int_{B_r \cap \mathcal{X}} \left(f(x_0, y) + \sum_{\beta \in \mathcal{I}_{\kappa_s}^{(1)}} \frac{\partial_x^\beta f(x_0, y)}{\prod_{s=1}^d \beta_s!} \prod_{j=1}^d (x_j - x_{0,j})^{\beta_j} + \zeta_{\kappa_s + \kappa_r}(x - x_0, y) \right) dx \\ &\stackrel{(2)}{=} \text{Vol}(B_r \cap \mathcal{X}) f(x_0, y) + \sum_{\beta \in \mathcal{I}_{\kappa_s}^{(1)}} \frac{\partial_x^\beta f(x_0, y)}{\prod_{s=1}^d \beta_s!} J_\beta(r) + R(r, y), \end{aligned} \quad (\text{S.2.8})$$

where (1) holds by Assumption S.4, and (2) by setting

$$\begin{aligned} J_\beta(r) &:= \int_{B_r \cap \mathcal{X}} \prod_{j=1}^d (x_j - x_{0,j})^{\beta_j} dx \\ R(r, y) &:= \int_{B_r \cap \mathcal{X}} \zeta_{\kappa_s + \kappa_r}(x - x_0, y) dx. \end{aligned}$$

Second, for any $r \in (0, r_2)$, we have

$$\begin{aligned} \int_{B_r \cap \mathcal{X}} g(x) dx &= \int_{B_r \cap \mathcal{X}} \int_y f(x, y) v(dy) dx \\ &\stackrel{(1)}{=} \int_{B_r \cap \mathcal{X}} \left(g(x_0) + \sum_{\beta \in \mathcal{I}_{\kappa_s}^{(1)}} \frac{\partial_x^\beta \int_y f(x_0, y) v(dy)}{\prod_{s=1}^d \beta_s!} \prod_{j=1}^d (x_j - x_{0,j})^{\beta_j} + \int_y \zeta_{\kappa_s + \kappa_r}(x - x_0, y) v(dy) \right) dx \\ &\stackrel{(2)}{=} \text{Vol}(B_r \cap \mathcal{X}) g(x_0) + \sum_{\beta \in \mathcal{I}_{\kappa_s}^{(1)}} \frac{\partial_x^\beta \int_y f(x_0, y) v(dy)}{\prod_{s=1}^d \beta_s!} J_\beta(r) + R(r), \end{aligned} \quad (\text{S.2.9})$$

where (1) holds by (S.2.8), $g(x_0) = \int_y f(x_0, y)v(dy)$, and by setting

$$R(r) := \int_{B_r \cap \mathcal{X}} \int_y \zeta_{\kappa_s + \kappa_r}(x - x_0, y)v(dy)dx .$$

Finally, for any $r \in (0, r_2)$, we have

$$\begin{aligned} h_r(y) - h_0(y) &= \frac{\int_{B_r \cap \mathcal{X}} f(v, y)dv - h_0(y) \int_{B_r \cap \mathcal{X}} g(x)dx}{\int_{B_r \cap \mathcal{X}} g(x)dx} \\ &\stackrel{(1)}{=} \frac{(\int_{B_r \cap \mathcal{X}} f(v, y)dv - h_0(y) \int_{B_r \cap \mathcal{X}} g(x)dx)/\text{Vol}(B_r \cap \mathcal{X})}{\int_{B_r \cap \mathcal{X}} g(x)dx/\text{Vol}(B_r \cap \mathcal{X})} \\ &\stackrel{(2)}{=} \frac{\sum_{\beta \in \mathcal{I}_{\kappa_s}^{(1)}} \frac{\partial_x^\beta f(x_0, y) - h_0(y) \partial_x^\beta \int_{\tilde{y}} f(x_0, \tilde{y})v(d\tilde{y})}{\prod_{s=1}^d \beta_s!} \frac{J_\beta(r)}{\text{Vol}(B_r \cap \mathcal{X})} + \frac{R(r, y) - h_0(y)R(r)}{\text{Vol}(B_r \cap \mathcal{X})}}{\int_{B_r \cap \mathcal{X}} g(x)dx/\text{Vol}(B_r \cap \mathcal{X})} , \end{aligned} \quad (\text{S.2.10})$$

where (1) holds by (S.2.7), which yields $\text{Vol}(B_r \cap \mathcal{X}) > 0$, and (2) by (S.2.8), (S.2.9), and $f(x_0, y) = h_0(y)g(x_0)$.

For any $x \in \mathcal{X}$ with $\|x - x_0\| \leq r$ and $r \in (0, r_2)$, Assumption 2(i) implies $g(x) \geq g(x_0)/2 > 0$. Then, for any $r \in (0, r_2)$,

$$\int_{B_r \cap \mathcal{X}} g(x)dx \stackrel{(1)}{\geq} (g(x_0)/2)\text{Vol}(B_r \cap \mathcal{X}) > 0 .$$

Then, for all $r \in (0, r_2)$, we get

$$\frac{\int_{B_r \cap \mathcal{X}} g(x)dx}{\text{Vol}(B_r \cap \mathcal{X})} \geq g(x_0)/2 > 0 . \quad (\text{S.2.11})$$

Now there are two cases, depending on whether x_0 is in the interior of \mathcal{X} or not. First, consider the case where x_0 is in the interior of \mathcal{X} . By definition, this implies that there exist \tilde{r} such that $B_r \subset \mathcal{X}$ for all $r \in (0, \tilde{r})$. Then, for any $r \in (0, \min\{r_2, \tilde{r}\})$, we obtain

$$h_r(y) - h_0(y) \stackrel{(1)}{=} \frac{\sum_{\beta \in \mathcal{I}_{\kappa_s}^{(2)}} \frac{\partial_x^\beta f(x_0, y) - h_0(y) \partial_x^\beta \int_{\tilde{y}} f(x_0, \tilde{y})v(d\tilde{y})}{\prod_{s=1}^d \beta_s!} \frac{J_\beta(r)}{\text{Vol}(B_r \cap \mathcal{X})} + \frac{R(r, y) - h_0(y)R(r)}{\text{Vol}(B_r \cap \mathcal{X})}}{\int_{B_r \cap \mathcal{X}} g(x)dx/\text{Vol}(B_r \cap \mathcal{X})} , \quad (\text{S.2.12})$$

where (1) holds by (S.2.10) and $B_r \subset \mathcal{X}$ for all $r \in (0, \tilde{r})$, and so $J_\beta(r) = \int_{B_r} \prod_{j=1}^d (x_j - x_{0,j})^{\beta_j} dx = 0$ for any $\beta \in \{\{0, 1\}^d : \|\beta\|_1 = 1\}$. Then, for any $r \in (0, \min\{r_2, \tilde{r}\})$,

$$\begin{aligned} \text{TV}(P_r, P) &= \frac{1}{2} \int |h_r(y) - h_0(y)|v(dy) \\ &\stackrel{(1)}{\leq} \frac{2 \sum_{\beta \in \mathcal{I}_{\kappa_s}^{(2)}} \frac{\int |\partial_x^\beta f(x_0, y)|v(dy)}{\prod_{s=1}^d \beta_s!} \frac{|J_\beta(r)|}{\text{Vol}(B_r \cap \mathcal{X})} + 2 \frac{\int |R(r, y)|v(dy)}{\text{Vol}(B_r \cap \mathcal{X})}}{g(x_0)} \end{aligned}$$

$$\begin{aligned}
&\stackrel{(2)}{\leq} \sum_{\beta \in \mathcal{I}_{\kappa_s}^{(2)}} O(r^{\|\beta\|_1}) + O(r^{\kappa_s + \kappa_r}) \\
&= O(r^{\min\{2, \kappa_s + \kappa_r\}}),
\end{aligned}$$

where (1) holds by (S.2.12), $\int h_0(y)v(dy) = 1$, $h_0(y) \geq 0$, $|\partial_x^\beta \int_y f(x_0, y)v(dy)| \leq \int |\partial_x^\beta f(x_0, y)|v(dy)$, and $|R(r)| \leq \int |R(r, y)|v(dy)$, and (2) by Assumption S.4, which yields $\int |\partial_x^\beta f(x_0, y)|v(dy) < \infty$,

$$|J_\beta(r)| = \left| \int_{B_r \cap \mathcal{X}} \prod_{j=1}^d (x_j - x_{0,j})^{\beta_j} dx \right| \leq r^{\|\beta\|_1} \text{Vol}(B_r \cap \mathcal{X}),$$

and

$$\int |R(r, y)|v(dy) \leq \int M(y) v(dy) r^{\kappa_s + \kappa_r} \text{Vol}(B_r \cap \mathcal{X}).$$

Next, consider the case where x_0 is not in the interior of \mathcal{X} . The sole difference is that (S.2.12) includes the term $J_\beta(r)$ with $\beta \in \{\{0, 1\}^d : \|\beta\|_1 = 1\}$, which now becomes

$$h_r(y) - h_0(y) \stackrel{(1)}{=} \frac{\sum_{\beta \in \mathcal{I}_{\kappa_s}^{(1)}} \frac{\partial_x^\beta f(x_0, y) - h_0(y) \partial_x^\beta \int_{\tilde{y}} f(x_0, \tilde{y})v(d\tilde{y})}{\prod_{s=1}^d \beta_s!} \frac{J_\beta(r)}{\text{Vol}(B_r \cap \mathcal{X})} + \frac{R(r, y) - h_0(y)R(r)}{\text{Vol}(B_r \cap \mathcal{X})}}{\int_{B_r \cap \mathcal{X}} g(x)dx / \text{Vol}(B_r \cap \mathcal{X})}. \quad (\text{S.2.13})$$

By using this expression and repeating previous arguments, we get

$$\text{TV}(P_r, P) = O(r^{\min\{1, \kappa_s + \kappa_r\}}).$$

The convergence rate of the Hellinger distance follows from its relationship with the total variation distance, $\mathbb{H}^2(P_r, P) \leq \text{TV}(P_r, P)$. ■