Submitted to Bernoulli

Stratified incomplete local simplex tests for curvature of nonparametric multiple regression

YANGLEI SONG [1], XIAOHUI CHEN[2] and KENGO KATO[3]

[1]*Department of Mathematics and Statistics, Queen's University, Jeffery Hall, Kingston, ON, Canada, K7L 3N6, E-mail:* yanglei.song@queensu.ca
[2]*Department of Statistics, University of Illinois at Urbana-Champaign, 725 S. Wright Street, Champaign, IL 61820, E-mail:* xhchen@illinois.edu
[3]*Department of Statistics and Data Science, Cornell University, 1194 Comstock Hall, Ithaca, NY 14853, E-mail:* kk976@cornell.edu

Principled nonparametric tests for regression curvature in $\mathbb{R}^d$ are often statistically and computationally challenging. This paper introduces the stratified incomplete local simplex (SILS) tests for joint concavity of nonparametric multiple regression. The SILS tests with suitable bootstrap calibration are shown to achieve simultaneous guarantees on dimension-free computational complexity, polynomial decay of the uniform error-in-size, and power consistency for general (global and local) alternatives. To establish these results, we develop a general theory for incomplete U-processes with stratified random sparse weights. Novel technical ingredients include maximal inequalities for the supremum of multiple incomplete U-processes.

Keywords: Nonparametric regression; Curvature testing; Incomplete U-processes; Stratification

1. Introduction

This paper concerns the hypothesis testing problem for curvature (i.e., concavity, convexity, or linearity) of a nonparametric multiple regression function. Testing the validity of such geometric hypothesis is important for performing high-quality subsequent shape-constrained statistical analysis. For instance, there has been considerable effort in nonparametric estimation of a convex (concave) regression function, partly because estimation under convexity constraint requires no tuning parameter as opposed to e.g. standard kernel estimation whose performance depends critically on a user-chosen bandwidth parameter [35, 34, 44, 27, 50, 43, 10, 9, 28, 14, 11, 32, 40]. In empirical studies such as economics and finance, convex (concave) regressions have wide applications in modeling the relationship between wages and education [47], between firm value and product price [5], and between mutual fund return and multiple risk factors [21, 1].

Consider the nonparametric multiple regression model

$$Y = f(V) + \varepsilon, \tag{1}$$

where Y is a scalar response variable, V is a d-dimensional covariate vector, ε is a random error term such that $\mathbb{E}[\varepsilon|V] = 0$ and $\mathrm{Var}(\epsilon) > 0$, and $f : \mathbb{R}^d \to \mathbb{R}$ is the conditional mean (i.e., regression) function. Let P be the joint distribution of $X = (V, Y) \in \mathbb{R}^{d+1}$ and $X_i := (V_i, Y_i), i \in [n] := \{1, \ldots, n\}$ be a sample of independent random vectors with common distribution P. For a given convex, compact subset $\mathcal{V} \subset \mathbb{R}^d$, based on the observations $\{X_i\}_{i=1}^n$, we aim to test the following hypothesis:

$$H_0 : f \text{ is concave on } \mathcal{V}, \tag{2}$$

against some (globally or locally) non-concave alternatives. In this work, we directly leverage the simplex characterization of concave functions, i.e., f is concave on $\mathcal{V}$ if and only if

$$a_1 f(v_1) + \cdots + a_{d+1} f(v_{d+1}) \leqslant f(a_1 v_1 + \cdots + a_{d+1} v_{d+1}), \tag{3}$$

for any $v_1, \ldots, v_{d+1} \in \mathcal{V}$ and nonnegative reals $a_1, \ldots, a_{d+1}$ such that $a_1 + \cdots + a_{d+1} = 1$. Working with this definition allows us to circumvent the need to estimate the regression function f, and thus the resulting tests would be robust to model misspecification. Further, the concavity hypothesis can be quantitatively evaluated on the observed data, which is the idea behind the *simplex statistic* in [1].

Specifically, $d+1$ covariate vectors in $\mathbb{R}^d$ form a simplex. Consider $r := d+2$ data points $x_1 := (v_1, y_1), \ldots, x_r := (v_r, y_r) \in \mathbb{R}^{d+1}$ generated from the model (1). If for all $j \in [r]$, v_j is not in the simplex spanned by $\{v_i : i \neq j\}$, for example the vectors $\{v_1, v_2, v_3, v_5\}$ in Figure 1, then we set $w(x_1, \ldots, x_r) = 0$. Otherwise, there exists a unique j such that v_j can be written as a convex combination of other covariate vectors, i.e., $v_j = \sum_{i \neq j} a_i v_i$ for some $a_i \geqslant 0, \sum_{i \neq j} a_i = 1$; in this case, we compare the response y_j with the same combination of others, $\{y_i : i \neq j\}$, i.e., setting

$$w(x_1, \ldots, x_r) = \sum_{i \neq j} a_i y_i - y_j.$$

For example, in Figure 1, v_4 is in the simplex spanned by $\{v_1, v_2, v_3\}$. We note that the index j and the coefficients $\{a_i : i \neq j\}$ are functions of $\{v_i : i \in [r]\}$, and defer the precise definitions to Section 3.

If f is indeed concave (i.e., in H_0) and ε is symmetric about zero, then $\mathbb{E}\left[\text{sign}(w(X_1, \ldots, X_r)\right] \leqslant 0$ due to (3), where $\text{sign}(t) := \mathbb{1}(t > 0) - \mathbb{1}(t < 0)$ is the sign function. Thus [1] proposes to use the following global U-statistic of all r-tuples from $\{X_i : i \in [n]\}$ and reject the null if the statistic is large:

$$|I_{n,r}|^{-1} \sum_{\iota \in I_{n,r}} \text{sign}\left(w(X_\iota)\right), \text{ with } X_\iota = (X_{i_1}, \ldots, X_{i_r}),$$

where $I_{n,r} := \{\iota = (i_1, \ldots, i_r) : 1 \leqslant i_1 < \ldots < i_r \leqslant n\}$, and $|\cdot|$ denotes the set cardinality.

1.1. Local simplex statistics

Since the above "global" U-statistic is not consistent against general alternatives, e.g., when f is only non-concave in a small region, [1] also proposes the *localized simplex statistics*. Specifically, let $L : \mathbb{R}^d \to \mathbb{R}$ be a function such that $L(z) = 0$ if $\|z\|_\infty := \max_{j \in [d]} |z_j| > 1/2$, and $L_b(\cdot) := b^{-d} L(\cdot / b)$ for $b > 0$. For $x_i := (v_i, y_i) \in \mathbb{R}^{d+1}, i \in [r]$, and a bandwidth parameter $b_n > 0$, define

$$h_v^{\text{sg}}(x_1, \ldots, x_r) := \text{sign}\left(w(x_1, \ldots, x_r)\right) b_n^{d/2} \prod_{k=1}^{r} L_{b_n}(v - v_k), \quad v \in \mathcal{V}. \tag{4}$$

Thus for each $v \in \mathcal{V}$, only nearby data points are utilized in constructing a local statistic. Note that h_v^{sg} depends on b_n, which we omit in most places for simplicity of notations.

Given a finite collection of query (or design) points $\mathscr{V}_n \subset \mathcal{V}$, [1] proposes to reject the null if

$$\sup_{v \in \mathscr{V}_n} U_n(h_v^{\text{sg}}) \text{ is large}, \quad \text{where} \quad U_n(h_v^{\text{sg}}) := |I_{n,r}|^{-1} \sum_{\iota \in I_{n,r}} h_v^{\text{sg}}(X_\iota). \tag{5}$$

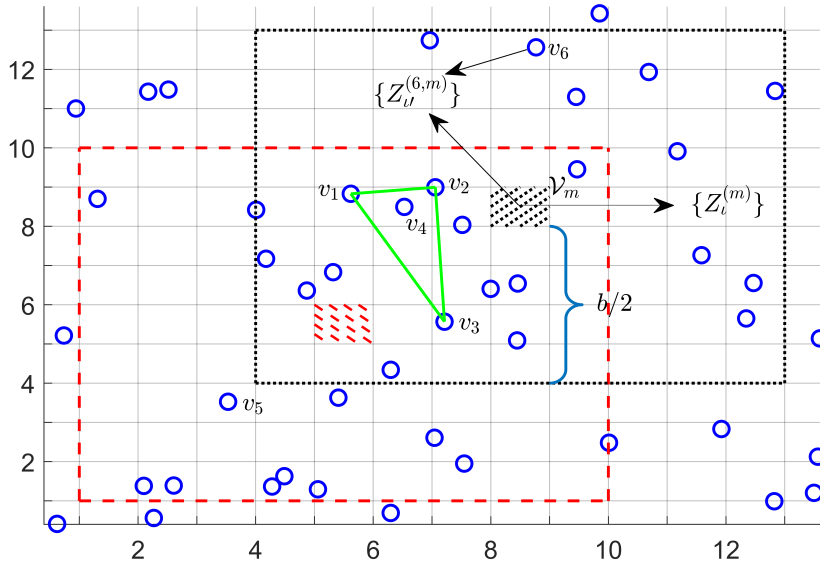

Figure 1: (i). Each circle represents a two-dimensional feature vector (i.e., $d = 2$). For each query point $v \in \mathcal{V}$, a sampling plan is a collection of Bernoulli random variables $\{Z_\iota(v) : \iota \in I_{n,r}\}$, one for each subset of $r = d + 2$ data points. If $Z_\iota(v) = 1$, then $h_v^{\mathrm{sg}}(X_\iota)$ contributes to the average in (5). (ii). The space $\mathcal{V}$ is stratified into disjoint regions (e.g., 1-by-1 squares above). The query points in each region share the same sampling plan (e.g. $\{Z_\iota^{(m)}\}$ for the dotted region $\mathcal{V}_m$), while different regions have independent sampling plans. For example, the indicator for $\iota = (v_1, v_2, v_3, v_4)$ may be one for $\mathcal{V}_m$, but zero for the dashed region. (iii). Due to the localizing kernel (4), for each query point $v \in \mathcal{V}$, it suffices to consider data points that are within $b/2$ distance to v in each coordinate. Thus for $\mathcal{V}_m$ above, e.g., it suffices to consider data points within the dotted square ($b = 8$). The key idea is that query points in a small region share similar nearby data points, and the region-specific sampling plan allows us to allocate the "limited resources" only in "important areas".

In [1], it requires the query points in $\mathscr{V}_n$ to be well separately, i.e., $\|v - v'\|_\infty > b_n$ for each pair of distinct $v, v' \in \mathscr{V}_n$, which is restrictive when $d \geqslant 2$ and b_n cannot be too small. Such a requirement is imposed since [1] uses extreme value theory to obtain the asymptotic distribution of the supremum, for which the convergence of approximation error is known to be logarithmically slow [31].

In [13], a valid jackknife multiplier bootstrap (JMB) is proposed to calibrate the distribution of the supremum of the (local) U-process, $\sup_{v\in\mathcal{V}} U_n(h_v^{\mathrm{sg}})$. Even though JMB tailored to the concavity test problem is statistically consistent, it requires tremendous, if not prohibitive, resources to compute $\sup_{v\in\mathcal{V}} U_n(h_v^{\mathrm{sg}})$, as well as calibrating its distribution via bootstrap, for $d \geqslant 2$. For instance, suppose that V has a Lebesgue density that is bounded away from zero on $\mathcal{V}$. Then the number of data points within the b_n-neighbourhood of $v \in \mathcal{V}$ is on average $O(nb_n^d)$. Thus to compute $U_n(h_v^{\mathrm{sg}})$ for a fixed $v \in \mathcal{V}$, the required number of evaluations of $w(\cdot)$ is on average $O((nb_n^d)^r)$, which is computationally intensive, if $d \geqslant 2$ (thus $r = 4$), and the bandwidth b_n is not too small. In fact, in the numerical study (Section 5), we estimate that for $d = 3, n = 1000, b_n = 0.6$ ($b_n/2$ is the half width), it would take more than 7 days to use bootstrap for calibration even with 40 computer cores.

It is tempting to break the computational bottleneck by using the incomplete version of the U-process $\{U_n(h_v^{\mathrm{sg}}) : v \in \mathcal{V}\}$, which has been studied for high-dimensional U-statistics [12, 53]. Specifically, we may associate each subset of r data points, $\iota \in I_{n,r}$, with an independent Bernoulli random variable Z_ι, and only include $h_v^{\mathrm{sg}}(X_\iota)$ in the average in (5) if $Z_\iota = 1$. Note that this is a "centralized" sampling plan, in the sense that $\{Z_\iota : \iota \in I_{n,r}\}$ is shared by each $v \in \mathcal{V}$. Here, we explain intuitively why such a plan does not solve the computational challenge, and postpone the detailed discussion until Section 4. First, for each $\iota = (i_1, \ldots, i_r) \in I_{n,r}$, if $\|v_{i_j} - v_{i_k}\|_\infty > b_n$ for some $j, k \in [r]$ (e.g., v_5, v_6 in Figure 1), then $h_v^{\mathrm{sg}}(X_\iota) = 0$ for each $v \in \mathcal{V}$. As a result, with a very high probability, a randomly selected r-tuples X_ι is "wasted". Second, if two query points v, v' are not close, in the sense that $\|v - v'\|_\infty > b_n$ (e.g. v_5, v_6 in Figure 1 if they are used as query points), then they share no nearby data points as defined by the localizing kernel in (4), which is a property ignored by the centralized sampling.

1.2. Our contributions

In this paper, we introduce the *stratified incomplete local simplex* (SILS) statistics for testing the concavity assumption in nonparametric multiple regression. We show that SILS tests have simultaneous guarantees on *dimension-free* computational complexity, *polynomial decay* of the uniform error-in-size, and *power consistency* against general alternatives. We elaborate below our contributions, and also refer readers to Figure 1 for a pictorial illustration of key ideas.

Computational contributions. The SILS test is proposed to address the computational issue with the test statistic (5), as well as calibrating its distribution. Specifically, we first partition the space $\mathcal{V}$ into disjoint regions $\{\mathcal{V}_m : m \in [M]\}$ for some integer $M \geqslant 1$. Let $N := n^\kappa b_n^{-dr}$ be a computational parameter for some $\kappa > 0$, and for each $m \in [M]$, let $\{Z_\iota^{(m)} : \iota \in I_{n,r}\}$ be a collection of independent Bernoulli random variables with success probability $p_n := N/|I_{n,r}|$, which is called a *sampling plan*. For different regions, the sampling plans are independent. Then we consider the stratified, incomplete version of (5) as our statistic for testing the hypothesis (2):

$$\sup_{m\in[M]} \sup_{v\in\mathcal{V}_m} \left(\sum\nolimits_{\iota\in I_{n,r}} Z_\iota^{(m)} h_v^{\mathrm{sg}}(X_\iota)\right) / \left(\sum\nolimits_{\iota\in I_{n,r}} Z_\iota^{(m)}\right).$$

Similar idea is applied to bootstrap calibration (see Subsection 2.2), which involves another computational parameter $N_2 := n^{\kappa'} b_n^{-dr}$ for some $\kappa' > 0$. Due to the localization by the kernel (4) and the stratification (see Figure 1), we show in Section 4.2 that the overall computational cost is $O(Mn^\kappa \log(n) + Mn^{1+\kappa'} b_n^{-d} \log(n) + BMn)$, where B is the number of bootstrap iterations. Our theory allows κ, κ' to be arbitrarily small, but due to power analysis, we recommend $\kappa = \kappa' = 1$. In addition, M is usually chosen so that $M = O(b_n^{-d})$, and to ensure a non-vanishing number of local data points, we must have $b_n^{-d} = O(n)$; thus the cost is independent of the dimension d.

Further, to alleviate the burden of selecting a single bandwidth, we propose to use the supremum of the statistics associated with *multiple* b_n (Subsection 3.3). Finally, we conduct extensive simulations to demonstrate the computational feasibility of the proposed method, and to corroborate our theory.[1]

Statistical contributions. In addition to the function class $\mathcal{H}^{\mathrm{sg}} := \{h_v^{\mathrm{sg}}\}$, which uses the sign of simplex statistics, we also consider another class of functions $\mathcal{H}^{\mathrm{id}} := \{h_v^{\mathrm{id}}\}$, where h_v^{id} uses $w(\cdot)$ instead of

[1]The implementation can be found on the github (https://github.com/ysong44/Stratified-incomplete-local-simplex-tests).

its sign (see (14)); note that h_v^{id} is unbounded unless ε has bounded support. On one hand, $\mathcal{H}^{\mathrm{sg}}$ requires the observation noise ε to be conditionally symmetric about zero [1], but otherwise is robust to heavy tailed ε. On the other hand, $\mathcal{H}^{\mathrm{id}}$ requires ε to have a light tail, but otherwise imposes no restrictions [13]. For both classes of functions, we establish the size validity, as well as power consistency against general alternatives, for the proposed procedure, under no smoothness assumption on the regression function.

In fact, under fairly general moment assumptions, we derive a unified Gaussian approximation and bootstrap theory for stratified, incomplete U-processes (Section 2 and 6), associated with a general function class $\mathcal{H}$, where the SILS test for regression concavity is an application of the general results.

Technical contributions. The analysis of the stratified, incomplete U-processes requires a strategy different from the coupling approach used for complete U-processes [13]: (i) we establish corresponding results for high dimensional stratified, incomplete U-statistics (Section B of the Supplement Material [52]); (ii) we show that the supremum of the process is well approximated by the supremum over a finite, but diverging, collection of $v \in \mathcal{V}$. The main novelty are local and non-local maximal inequalities to bound the supremum difference between a complete U-process and its stratified, incomplete version (Section A.1 of the Supplement Material [52]), which can also be useful for other applications involving sampling, such as estimating the density of functions of several random variables [25].

We note that the developed maximal inequalities are novel compared to [13] and [12]. First, [13] studies complete U-processes, and neither stratification nor sampling is involved. Second, [12] establishes inequalities for incomplete high dimensional U-vectors, whose proofs are fundamentally different from those for processes, and which does not have the stratification component. See also Remark 3.3 for technical challenges associated with local U-processes.

1.3. Related work

Regression under concave/convex restrictions has a long and rich history dating back to [35]. Traditionally, the literature focused on the univariate ($d = 1$) case [34, 44, 27, 9, 28, 14], but there is a significant recent theoretical progress in the multivariate case [50, 43, 32, 40]; see also [45, 39, 33, 46]. We refer readers to [16, 29] for a review on estimation and inference under shape constraints including concave/convexity constraints.

The literature on testing the hypotheses of regression concavity is relatively scarce, especially for multiple regression, i.e., $d \geqslant 2$. Simplex statistic and its local version are introduced in [1], and the bootstrap calibration (without computational concerns) is investigated in [13]. Several testing procedures based on splines [18, 55, 38] have been proposed, which, however, are only proven to work for the univariate case since they are essentially second-derivative tests at the spline knots. Thus such methods can only test marginal concavity in the presence of multiple covariates, and multi-dimensional spline interpolation is much less understood in the nonparametric regression setting. Further, in the univariate case with a white-noise model, multi-scale testing for qualitative hypotheses is considered in [19], and minimax risks for estimating the L^q distance ($1 \leqslant q < \infty$) between an unknown signal and the cones of positive/monotone/convex functions are established in [37].

A very recent work by [22] proposes a projection framework for testing shape restrictions including concavity, which we call “FS" test. Specifically, the FS test [22] first estimates the regression function f using unconstrained, nonparametric methods (e.g. by sieved splines), and then evaluate and calibrate the L^2 distance between the estimator and the space of concave functions. As discussed in Section E.4 of the Supplement Material [52], the FS test is expected to achieve descent power, but fails to control the size properly when f is not smooth; this is because if f is not smooth enough, there is no

choice of tuning parameter (e.g., the number of terms in sieved B-splines) that can meet its two requirements simultaneously: under-smoothing and strong approximation. In simulation studies (Section 5), we observe that the FS test rejects H_0 with a very large probability when f is concave, piecewise affine. In contrast, for our procedure, the probability of rejecting the null attains the maximum when f is an affine function, as the equality in (3) is achieved if and only if f is affine; thus, the size validity requires no additional assumption on f. Finally, we show in Section 5 that the proposed method achieves a comparable power to the FS test.

We postpone the discussion of related work on the distribution approximation and bootstrap for U-processes until Subsection 6.3.

1.4. Organization of the paper

In Section 2, we introduce stratified, incomplete U-processes, as well as bootstrap calibration, for a general function class $\mathcal{H}$. In Section 3, we apply the general theory to the concavity test application, and establish its size validity and power consistency. In Section 4, we discuss the computational complexity of the proposed procedure as well as its implementation. In Section 5, we present simulation results for $d = 2$, with the cases of $d = 3, 4$ presented in the Supplement Material [52]. In Section 6, we establish the validity of Gaussian approximation and bootstrap for stratified, incomplete U-processes.

1.5. Notation

We denote $X_i, \dots X_{i'}$ by $X_i^{i'}$ for $i \leqslant i'$. For any integer n, we denote by $[n]$ the set $\{1, 2, \dots, n\}$. For $a, b \in \mathbb{R}$, let $\lfloor a \rfloor$ denote the largest integer that does not exceed a, $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. For $a \in \mathbb{R}^d$ and $q \in [1, \infty)$, denote $\|a\|_q = \left(\sum_{i=1}^d |a_i|^q\right)^{1/q}$, and $\|a\|_\infty = \max_{i \in [d]} |a_i|$. For $a, b \in \mathbb{R}^d$, we write $a \leqslant b$ if $a_j \leqslant b_j$ for $1 \leqslant j \leqslant d$, and write $[a, b]$ for the hyperrectangle $\prod_{j=1}^d [a_j, b_j]$ if $a \leqslant b$. For $\beta > 0$, let $\psi_\beta : [0, \infty) \to \mathbb{R}$ be a function defined by $\psi_\beta(x) = e^{x^\beta} - 1$, and for any real-valued random variable ξ, define $\|\xi\|_{\psi_\beta} = \inf\{C > 0 : \mathbb{E}[\psi_\beta(|\xi|/C)] \leqslant 1\}$. Denote by $I_{n,r} := \{\iota = (i_1, \dots, i_r) : 1 \leqslant i_1 < \dots < i_r \leqslant n\}$ the set of all ordered r-tuples of $[n]$ and denote by $|\cdot|$ the set cardinality.

For a nonempty set T, denote $\ell^\infty(T)$ the Banach space of real-valued functions $f : T \to \mathbb{R}$ equipped with the sup norm $\|f\|_T := \sup_{t \in T} |f(t)|$. For a semi-metric space (T, d), denote by $N(T, d, \epsilon)$ its ϵ-covering number, i.e., the minimum number of closed d-balls with radius ϵ that cover T; see [54, Section 2.1]. For a probability space $(T, \mathcal{T}, Q)$ and a measurable function $f : T \to \mathbb{R}$, denote $Qf = \int f dQ$ whenever it is well defined. For $q \in [1, \infty]$, denote by $\|\cdot\|_{Q,q}$ the $L^q(Q)$-seminorm, i.e., $\|f\|_{Q,q} = (Q|f|^q)^{1/q}$ for $q < \infty$ and $\|f\|_{Q,\infty}$ for the essential supremum.

For $k = 0, 1, \dots, r$ and a measurable function $f : (S^r, \mathcal{S}^r) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, let $P^{r-k} f$ denote the function on S^k such that $P^{r-k} f(x_1, \dots, x_k) = \mathbb{E}[f(x_1, \dots, x_k, X_{k+1}, \dots, X_r)]$,, whenever it is well defined. For a generic random variable Y , let $\mathbb{P}_{|Y}(\cdot)$ and $\mathbb{E}_{|Y}[\cdot]$ denote the conditional probability and expectation given Y, respectively. Throughout the paper, we assume that

$$r \geqslant 2, \;\; n \geqslant 4, \;\; N \geqslant 4, \;\; p_n := N/|I_{n,r}| \leqslant 1/2, \;\; N \geqslant n/r \geqslant 1.$$

Also, we assume the probability space is rich enough in the sense that there exists a random variable that has the uniform distribution on $(0, 1)$ and is independent of all other random variables.

2. Stratified incomplete U-processes

In this section, we introduce stratified, incomplete U-processes, as well as bootstrap calibration, for a general function class $\mathcal{H}$. For intuitions, it may help to think $\mathcal{H}$ as the collection of functions h_v^{sg} in (4) indexed by $v \in \mathcal{V}$, and refer to Figure 1.

Thus, let $X_1^n := \{X_1, \ldots, X_n\}$ be independent and identically distributed (i.i.d.) random variables taking value in a measurable space $(S, \mathcal{S})$ with common distribution P. Fix $r \geqslant 2$, and let $\mathcal{H}$ be a collection of symmetric, measurable functions $h : (S^r, \mathcal{S}^r) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Define the U-process and its standardized version as follows: for $h \in \mathcal{H}$,

$$U_n(h) := |I_{n,r}|^{-1} \sum_{\iota \in I_{n,r}} h(X_\iota), \quad \mathbb{U}_n(h) := \sqrt{n}\left(U_n(h) - \mathbb{E}\left[U_n(h)\right]\right),$$

where recall that $X_\iota = (X_{i_1}, \ldots, X_{i_r})$ if $\iota = (i_1, \ldots, i_r) \in I_{n,r}$. The summation in the above complete U-process involves $\sim n^r$ terms, and thus is computationally expensive even for a moderate r (say $\geqslant 3$), which motivates its stratified incomplete version.

2.1. Test statistics

Let $\{\mathcal{H}_m : m \in [M]\}$ be a partition of $\mathcal{H}$, i.e., $\mathcal{H}_{m_1} \cap \mathcal{H}_{m_2} = \emptyset$ for $m_1 \neq m_2$, and $\cup_{m=1}^M \mathcal{H}_m = \mathcal{H}$. The partition, and thus M, may depend on the sample size n. Given a positive integer N, which represents a computational parameter, define

$$\left\{Z_\iota^{(m)} : m \in [M],\ \iota \in I_{n,r}\right\} \overset{i.i.d.}{\sim} \text{Bernoulli}(p_n), \qquad \text{with } p_n := N/|I_{n,r}|,$$

which are independent of the data X_1^n. For $m \in [M]$, denote by $\widehat{N}^{(m)} := \sum_{\iota \in I_{n,r}} Z_\iota^{(m)}$ the total number of sampled r-tuples for the subclass $\mathcal{H}_m$. Further, define a function $\sigma : \mathcal{H} \to \{1, \ldots, M\}$ that maps $h \in \mathcal{H}$ to the index of the partition to which h belongs, i.e., $\sigma(h) = m \Leftrightarrow h \in \mathcal{H}_m$. Finally, we define the *stratified, incomplete U-process* and its standardized version: for $h \in \mathcal{H}$, if $\sigma(h) = m$,

$$U'_{n,N}(h) := \left(\widehat{N}^{(m)}\right)^{-1} \sum_{\iota \in I_{n,r}} Z_\iota^{(m)} h(X_\iota), \quad \mathbb{U}'_{n,N}(h) := \sqrt{n}\left(U'_{n,N}(h) - \mathbb{E}\left[U'_{n,N}(h)\right]\right). \tag{6}$$

An important goal of the paper is to develop bootstrap methods to calibrate the distribution of the supremum of the stratified incomplete U-process, i.e., $\mathbb{M}_n := \sup_{h \in \mathcal{H}} \mathbb{U}'_{n,N}(h)$.

Statistical tests. We will use $\sqrt{n}\sup_{h \in \mathcal{H}} U'_{n,N}(h)$ as the *test statistic*, which can be evaluated given the data X_1^n and sampling plans $\{Z_\iota^{(m)}\}$. If under the null, $P^r h \leqslant 0$ for each $h \in \mathcal{H}$, then

$$\sqrt{n}\sup_{h \in \mathcal{H}} U'_{n,N}(h) \leqslant \sup_{h \in \mathcal{H}} \mathbb{U}'_{n,N}(h) = \mathbb{M}_n.$$

Thus a test based on the α-th upper quantile of $\mathbb{M}_n$ controls the size below α. If, in addition, under certain configuration in the null, $P^r h = 0$ for each $h \in \mathcal{H}$, then the test is *non-conservative*, i.e., controlling the size at α.

***Remark* 2.1.** *A stratification of $\mathcal{H} = \{h_v^{sg} : v \in \mathcal{V}\}$ is equivalent to partitioning $\mathcal{V}$ into sub-regions $\{\mathcal{V}_m : m \in [M]\}$ and letting $\mathcal{H}_m = \{h_v^{sg} : v \in \mathcal{V}_m\}$ (see Figure 1). Query points in $\mathcal{V}_m$ share the*

same sampling plan $\{Z_\iota^{(m)} : \iota \in I_{n,r}\}$*. As we shall see in Section 4, it is computationally important to partition the function class* $\mathcal{H}$ *so that each partition has its individual sampling plan. Our analysis is non-asymptotic, so no stratification* ($M = 1$) *is also allowed.*

2.2. Bootstrap calibration

To operationalize the above test, we use multiplier bootstrap to calibrate the distribution of $\mathbb{M}_n$. To gain intuition, assume for a moment $P^r h = 0$ for $h \in \mathcal{H}$, and observe that

$$(\widehat{N}^{\sigma(h)} N^{-1}) \mathbb{U}'_{n,N}(h) = \mathbb{U}_n(h) + \alpha_n^{1/2} N^{-1/2} \sum_{\iota \in I_{n,r}} \Big(Z_\iota^{(\sigma(h))} - p_n \Big) h(X_\iota), \tag{7}$$

where $\alpha_n := n/N$. The first term on the right is a complete U-statistic, and thus is approximated by its Hájek projection $rn^{-1/2} \sum_{k \in [n]} P^{r-1} h(X_k)$. The second term is due to stratified sampling: conditional on data X_1^r, it is a sum of independent centered Bernoulli random variables, with variance approximately given by $\alpha_n U_n(h^2)$. We will handle these two sources of variation.

The Hájek projection part requires additional notations. Let $\mathcal{D}_n := X_1^n \cup \{Z_\iota^{(m)} : \iota \in I_{n,r}, m \in [M]\}$ be the data involved in the definition of $\mathbb{U}'_{n,N}$ in (6). For each $k \in [n]$, denote by

$$I_{n-1,r-1}^{(k)} := \{\, (i_1, \ldots, i_{r-1}) : 1 \leqslant i_1 < \cdots < i_{r-1} \leqslant n, \quad i_j \neq k \ \text{ for } 1 \leqslant j \leqslant r-1 \,\},$$

the collection of all ordered $r-1$ tuples in the set $\{1, \ldots, n\} \setminus \{k\}$. Let N_2 be another computational budget, and define

$$\Big\{ Z_\iota^{(k,m)} : \ k \in [n], \ m \in [M], \ \iota \in I_{n-1,r-1}^{(k)} \Big\} \overset{i.i.d.}{\sim} \text{Bernoulli}(q_n), \quad q_n := N_2 / |I_{n-1,r-1}|,$$

that are independent of $\mathcal{D}_n$. For example, if $\mathcal{H} = \{h_v^{\text{sg}} : v \in \mathcal{V}\}$, each pair of data point X_k and region $\mathcal{V}_m$ is associated with an independent sampling plan $\{Z_\iota^{(k,m)} : \iota \in I_{n-1,r-1}^{(k)}\}$; see Figure 1.

For $k \in [n]$ and $m \in [M]$, define $\widehat{N}_2^{(k,m)} := \sum_{\iota \in I_{n-1,r-1}^{(k)}} Z_\iota^{(k,m)}$, the number of selected $r-1$ tuples from $[n] \setminus \{k\}$ for X_k and $\mathcal{H}_m$. Further, for $h \in \mathcal{H}$ with $\sigma(h) = m$,

$$\mathbb{G}^{(k)}(h) := \Big(\widehat{N}_2^{(k,m)} \Big)^{-1} \sum_{\iota \in I_{n-1,r-1}^{(k)}} Z_\iota^{(k,m)} h(X_{\iota^{(k)}}), \quad \overline{\mathbb{G}}(h) := n^{-1} \sum_{k=1}^{n} \mathbb{G}^{(k)}(h), \tag{8}$$

where $\iota^{(k)} := \{k\} \cup \iota$. Here, $\mathbb{G}^{(k)}(h)$ is intended as an estimator for the k^{th} term in the Hájek projection, since by definition $\mathbb{E}\left[h(X_{\iota^{(k)}}) \,|X_k \right] = P^{r-1} h(X_k)$.

Multiplier bootstrap. Now let $\Big\{ \xi_k, \ \xi_\iota^{(m)} : \ k \in [n], \ m \in [M], \iota \in I_{n,r} \Big\} \overset{i.i.d.}{\sim} N(0,1)$ be independent standard Gaussian multipliers, independent from the data X_1^n and the sampling plans, i.e.,

$$\mathcal{D}'_n := \mathcal{D}_n \cup \{ Z_\iota^{(k,m)} : k \in [n], \ m \in [M], \ \iota \in I_{n-1,r-1}^{(k)} \}. \tag{9}$$

Define for $h \in \mathcal{H}$ with $\sigma(h) = m$,

$$\begin{aligned} \mathbb{U}^{\#}_{n,A}(h) &:= n^{-1/2} \sum_{k=1}^{n} \xi_k \left(\mathbb{G}^{(k)}(h) - \overline{\mathbb{G}}(h) \right), \\ \mathbb{U}^{\#}_{n,B}(h) &:= \left(\widehat{N}^{(m)} \right)^{-1/2} \sum_{\iota \in I_{n,r}} \xi_\iota^{(m)} \sqrt{Z_\iota^{(m)}} \left(h(X_\iota) - U'_{n,N}(h) \right), \end{aligned} \tag{10}$$

where $0/0$ is interpreted as 0. Note that the multipliers $\{\xi_k\}$ are shared across regions, while $\{\xi_\iota^{(m)}\}$ are region-specific. Further, conditional on $\mathcal{D}'_n$, $\mathbb{U}^{\#}_{n,A}$ and $\mathbb{U}^{\#}_{n,B}$ are centered Gaussian processes with covariance functions $\widehat{\gamma}_A(h,h') := n^{-1} \sum_{k=1}^{n} (\mathbb{G}^{(k)}(h) - \overline{\mathbb{G}}(h))(\mathbb{G}^{(k)}(h') - \overline{\mathbb{G}}(h'))$ and $\widehat{\gamma}_B(h,h') := (\widehat{N}^{(\sigma(h))})^{-1} \sum_{\iota \in I_{n,r}} Z_\iota^{(\sigma(h))} (h(X_\iota) - U'_{n,N}(h))(h'(X_\iota) - U'_{n,N}(h')) \mathbb{1}\{\sigma(h) = \sigma(h')\}$ for any $h, h' \in \mathcal{H}$. In view of (7), we combine these two processes and define

$$\mathbb{U}^{\#}_{n,*}(h) := r\mathbb{U}^{\#}_{n,A}(h) + \alpha_n^{1/2} \mathbb{U}^{\#}_{n,B}(h) \quad \text{for } h \in \mathcal{H}, \qquad \mathbb{M}^{\#}_n := \sup_{h \in \mathcal{H}} \mathbb{U}^{\#}_{n,*}(h). \tag{11}$$

Finally, we estimate the conditional (given $\mathcal{D}'_n$) distribution of $\mathbb{M}^{\#}_n$ by bootstrap, i.e., by repeatedly generating independent realizations of the multipliers $\{\xi_k, \xi_\iota^{(m)}\}$ with the data X_1^n and the sampling plans $\{Z_\iota^{(m)}, Z_\iota^{(k,m)}\}$ fixed, and obtain the critical value for the previous test statistic from the conditional distribution of $\mathbb{M}^{\#}_n$.

2.3. A simplified version of approximation results

To justify the bootstrap procedure, we need to show that conditional on $\mathcal{D}'_n$, the distribution of $\mathbb{M}^{\#}_n$ is close to that of $\mathbb{M}_n$, which is the main result in Section 6. Here we state a simplified version of the approximation results for a uniformly bounded function class $\mathcal{H}$. Note that the bound on $\mathcal{H}$ is allowed to vary with n.

***Definition* 2.2** (VC type function class [13, 15])**.** *A collection, $\mathcal{H}$, of functions on S^r with a measurable envelope function H (i.e. $H \geqslant \sup_{h \in \mathcal{H}} |h|$ pointwise) is said to be VC type with characteristics (A, ν) if $\sup_Q N(\mathcal{H}, \|\cdot\|_{Q,2}, \epsilon \|H\|_{Q,2}) \leqslant (A/\epsilon)^\nu$ for any $\epsilon \in (0,1)$, where $\sup_Q$ is taken over all finitely discrete probability measures on S^r.*

We work with the following assumptions.

(PM). $\mathcal{H}$ is pointwise measurable in the sense that for any $n \in \mathbb{N}$, there exists a countable subset $\mathcal{H}'_n \subset \mathcal{H}$ such that, almost surely, for every $h \in \mathcal{H}$, there exists a sequence $\{h_m\} \subset \mathcal{H}'_n$ with $\lim_m h_m(X_i) = h(X_i)$ for $i \in [n]$.

(VC). $\mathcal{H}$ is VC type with envelope H and characteristics $A \geqslant e \vee (e^{2(r-1)}/16)$ and $\nu \geqslant 1$.

(MB). For some absolute constant $C_0 > 0$, $\log(M) \leqslant C_0 \log(n)$.

(MT-∞). There exist absolute constants $\underline{\sigma} > 0$, $c_0 \in (0,1)$, and a sequence of reals $D_n \geqslant 1$ such that for each $0 \leqslant \ell \leqslant r$ and $1 \leqslant s \leqslant 4$,

$$\mathrm{Var}\left(P^{r-1}h(X_1)\right) \geqslant \underline{\sigma}^2, \qquad \mathrm{Var}\left(h(X_1^r)\right) \geqslant c_0 D_n^{2r-2},$$

$$\|P^{r-\ell}|h|^s\|_{P^\ell,\bar{q}} \leqslant D_n^{2r(s-1)+2\ell-s-2\ell/\bar{q}}, \quad \text{for } \bar{q} \in \{2,3,4\},\ h \in \mathcal{H},$$

$$\|P^{r-\ell}H^s\|_{P^\ell,\bar{q}} \leqslant D_n^{2r(s-1)+2\ell-s-(2\ell-2)/\bar{q}} \text{ for } \bar{q} \in \{2,\infty\}, \quad \|(P^{r-2}H)^{\odot 2}\|_{P^2,\infty} \leqslant D_n^4,$$

where for a function $f : S^2 \to \mathbb{R}$, define $f^{\odot 2}(x_1,x_2) := \int f(x_1,x)f(x_2,x)dP(x)$.

Theorem 2.3. *Assume the conditions (PM), (VC), (MB) and (MT-∞). Then there exists a constant C, depending only on constants $r, \underline{\sigma}, c_0, C_0$, such that with probability at least $1 - C\varrho_n'$,*

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}(\mathbb{M}_n \leqslant t) - \mathbb{P}_{|\mathcal{D}_n'}(\mathbb{M}_n^{\#} \leqslant t) \right| \leqslant C\varrho_n',$$

where $\varrho_n' := \left(\frac{D_n^{2r}K_n^7}{N\wedge N_2}\right)^{1/8} + \left(\frac{D_n^2K_n^7}{n}\right)^{1/8} + \left(\frac{D_n^3K_n^4}{n}\right)^{2/7}$ *and* $K_n := \nu\log(A\vee n)$.

Proof. It follows from Theorem 6.1 and Theorem 6.2. Specifically, (MT-∞) verifies (MT) with $q = \infty$ and $B_n = D_n^r$. Further, we may without loss of generality assume that $\eta_n^{(1)}$, $\eta_n^{(2)}$ and ρ_n, in Theorem 6.1 and 6.2, are bounded by 1, and then it is clear that $\eta_n^{(1)} + \eta_n^{(2)} + \rho_n \leqslant C\varrho_n'$. ■

***Remark* 2.4.** *The condition (MB) requires $\log(M) \leqslant C_0\log(n)$, and the impact of M has been absorbed into K_n, since $K_n \geqslant \log(n)$.*

The condition (MT-∞) are motivated by the application of testing the concavity of a regression function in Section 3. It holds if we use (i). the sign kernel $\{h_v^{sg} : v \in \mathcal{V}\}$ in (4) or (ii). the identity kernel $\{h_v^{sg} : v \in \mathcal{V}\}$ in (14) under the additional assumption that the observation noise ε in (1) is bounded; the more general results in Section 6 are required to remove this assumption.

3. Stratified incomplete local simplex tests: statistical guarantees

In this section, we apply the general theory in Section 2 to the concavity test of a regression function, i.e., H_0 in (2), formally introduce stratified incomplete local simplex tests, and establish the size validity and power consistency. Finally, we propose tests that combine multiple bandwidths.

We first recall the *simplex statistics* proposed in [1]. For $v_1,\ldots,v_{d+1} \in \mathbb{R}^d$, denote by

$$\Delta^\circ(v_1,\ldots,v_{d+1}) := \left\{ \sum_{i=1}^{d+1} a_i v_i : \sum_{i=1}^{d+1} a_i = 1,\ \ a_i > 0 \text{ for } i \in [d+1] \right\}$$

the interior of the simplex spanned by $v_1,\ldots,v_{d+1}$, and define $\mathcal{S} := \bigcup_{j=1}^r \mathcal{S}_j$, where $r := d+2$ and

$$\mathcal{S}_j = \left\{ (v_1,\ldots,v_r) \in \mathbb{R}^{d\times r} : \begin{array}{c} v_1,\ldots,v_{j-1},v_{j+1},\ldots,v_r \text{ are affinely independent} \\ \text{and } v_j \in \Delta^\circ(v_1,\ldots,v_{j-1},v_{j+1},\ldots,v_r) \end{array} \right\}.$$

Clearly, $\mathcal{S}_1,\ldots,\mathcal{S}_r$ are disjoint. To illustrate, in Figure 1, $(v_1,v_2,v_3,v_4) \in \mathcal{S}_4$, but $(v_1,v_2,v_3,v_5) \notin \mathcal{S}$.

For $j \in [r]$, there exists a unique collection of functions $\{\tau_i^{(j)} : \mathcal{S}_j \to (0,1) : i \in [r] \setminus \{j\}\}$ such that for any $v_1^r := (v_1, \ldots, v_r) \in \mathcal{S}_j$,

$$v_j = \sum_{i \in [r] \setminus \{j\}} \tau_i^{(j)}(v_1^r)\, v_i, \qquad \sum_{i \in [r] \setminus \{j\}} \tau_i^{(j)}(v_1^r) = 1. \tag{12}$$

Now define $w : \mathbb{R}^{(d+1)\times r} \to \mathbb{R}$ as follows: for $x_i := (v_i, y_i) \in \mathbb{R}^{d+1}, i \in [r]$,

$$w(x_1, \ldots, x_r) := \sum_{j=1}^{r} \left(\sum_{i \in [r] \setminus \{j\}} \tau_i^{(j)}(v_1^r)\, y_i - y_j \right) \mathbb{1}\left\{ v_1^r \in \mathcal{S}_j \right\}. \tag{13}$$

It is clear that $\mathcal{S}$ is permutation invariant for $v_1, \ldots, v_r$, and that $w(\cdot)$ is symmetric in its arguments. Key observations are that if the regression function f is concave (i.e. H_0 holds), then $P^r w \leqslant 0$, and that if f is an affine function, $P^r w = 0$, where recall that P is the distribution of $X := (V, Y)$.

Let $L(\cdot)$ be a kernel function and $b_n > 0$ a bandwidth parameter. Recall that $L_b(\cdot) := b^{-d} L(\cdot/b)$ for $b > 0$, and define $\mathcal{H}^{\mathrm{id}} := \{h_v^{\mathrm{id}} : v \in \mathcal{V}\}$, where for each $x_i = (v_i, y_i) \in \mathbb{R}^{d+1}, i \in [r]$,

$$h_v^{\mathrm{id}}(x_1, \ldots, x_r) := w(x_1, \ldots, x_r) b_n^{d/2} \prod_{k=1}^{r} L_{b_n}(v - v_k), \quad v \in \mathcal{V}. \tag{14}$$

Now consider a partition of $\mathcal{V}$, $\{\mathcal{V}_m : m \in [M]\}$, which induces a partition of $\mathcal{H}^{\mathrm{id}}$, i.e., $\mathcal{H}_m^{\mathrm{id}} := \{h_v^{\mathrm{id}} : v \in \mathcal{V}_m\}, m \in [M]$; see Figure 1.

Finally, recall the definitions of $U'_{n,N}(\cdot)$ in (6), $\mathcal{D}'_n$ in (9), and $\mathbb{M}_n^{\#}$ in (11). Given a nominal level $\alpha \in (0,1)$, we propose to reject the null in (2) if and only if

$$\sup_{v \in \mathcal{V}} \sqrt{n} U'_{n,N}(h_v^{\mathrm{id}}) \;\geqslant\; q_\alpha^{\#}, \tag{15}$$

where $q_\alpha^{\#}$ is the $(1-\alpha)$-th quantile of $\mathbb{M}_n^{\#}$ conditional on $\mathcal{D}'_n$.

Sign function. We also consider the function class $\mathcal{H}^{\mathrm{sg}} := \{h_v^{\mathrm{sg}} : v \in \mathcal{V}\}$, where h_v^{sg} is defined in (4). As we shall see, $\{h_v^{\mathrm{sg}} : v \in \mathcal{V}\}$ has the advantage of being bounded, but it requires that the conditional distribution of ε given V is symmetric about zero. On the other hand, $\{h_v^{\mathrm{id}} : v \in \mathcal{V}\}$ imposes no assumption on the shape of the conditional distribution, but requires ε to have a light tail; in this Section, we assume ε to be bounded for $\{h_v^{\mathrm{id}} : v \in \mathcal{V}\}$ so that we can apply Theorem 2.3, and relax this assumption in the Supplement Material [52].

3.1. Assumptions for concavity tests

We assume the distribution of (V, ε) in (1) to be fixed, but allow f to depend on the sample size n, which permits the study of local alternatives. We make the following assumptions: for some absolute constant $C_0 > 1$,

(C1). The kernel $L : \mathbb{R}^d \to \mathbb{R}$ is continuous, of bounded variation, and has support $[-1/2, 1/2]^d$. Or $L(\cdot)$ is the uniform kernel on $[-1/2, 1/2]^d$, i.e., $L(v) = \mathbb{1}\{v \in (-1/2, 1/2)^d\}$ for $v \in \mathbb{R}^d$.

(C2). The number of partitions, M, grows at most polynomially in n, i.e., $\log(M) \leqslant C_0 \log(n)$.

(C3). The bandwidth b_n does not vanish too fast in n, i.e., $1 \leqslant b_n^{-3d/2} \leqslant C_0 n^{1-1/C_0}$.

(C4). V has a Lebesgue density p such that $C_0^{-1} \leqslant p(v) \leqslant C_0$ for $v \in \mathcal{V}^{2b_n}$, where $\mathcal{V}^b := \{v' \in \mathbb{R}^d : \inf_{v'' \in \mathcal{V}} \|v' - v''\|_\infty \leqslant b\}$ is the b-enlargement of $\mathcal{V}$.

(C5). Assume for $* = \text{id}$ or sg and $n \geqslant C_0$, $\inf_{v \in \mathcal{V}} \text{Var}\left(P^{r-1} h_v^*(X_1)\right) \geqslant C_0^{-1}$.

(C6-id') Assume that $\sup_{v \in \mathcal{V}^{b_n}} |f(v)| \leqslant C_0$ and that ε is bounded by C_0 almost surely.

(C6-sg') Assume that $\sup_{v \in \mathcal{V}^{b_n}} |f(v)| \leqslant C_0$, and that ε is independent of V, symmetric about zero, i.e., $\mathbb{P}(\varepsilon > t) = \mathbb{P}(\varepsilon < -t)$ for any $t > 0$, and $\mathbb{P}(\varepsilon > 2C_0) > 0$.

Some comments are in order. (C1) is a standard assumption on the kernel L, which is satisfied by many commonly used kernels. Recall that $\mathcal{V}$ is compact, so (C2) is satisfied if we partition each coordinate into segments of length ηb_n for some small $\eta \in (0, 1)$. (C3) imposes the same condition on the bandwidth b_n as for the procedure using the complete U-process [13, (T5) in Section 4], which holds as long as $n^{-2/(3d)+\eta} \lesssim b_n$ for arbitrarily small $\eta \in (0, 1)$; in comparison, [1] has a (slightly) milder condition on the bandwidth, $n^{-1/d+\eta} \lesssim b_n$ for the discretized U-statistics. (C4) is necessary that for each $v \in \mathcal{V}$, there are enough data points in the b_n-neighbourhood of v. The condition (C6-id') is assumed for the class $\mathcal{H}^{\text{id}}$, while (C6-sg') for $\mathcal{H}^{\text{sg}}$.

Now we focus on (C5) with the function class $\mathcal{H}^{\text{id}}$, as the discussion for $\mathcal{H}^{\text{sg}}$ is similar. For simplicity, assume ε and V in (1) are independent. By a change-of-variable and due to the fact that $\tau_i^{(j)}$ in (12) is invariant under affine transformations (in particular $\tau_i^{(j)}(v - b_n u_1, \ldots, v - b_n u_r) = \tau_i^{(j)}(u_1, \ldots, u_r)$),

$$\mathbb{E}\left[\text{Var}\left(P^{r-1} h_v^{\text{id}}(V_1, Y_1) | V_1\right)\right] = \text{Var}(\varepsilon_1) \int p(v - b_n u_1) L^2(u_1) \mathcal{T}_v^2(u_1) du_1, \quad \text{where}$$

$$\mathcal{T}_v(u_1) = \int \left(\mathbb{1}\{u_1^r \in \mathcal{S}_1\} - \sum_{j=2}^{r} \tau_1^{(j)}(u_1^r) \mathbb{1}\{u_1^r \in \mathcal{S}_j\} \right) \prod_{i=2}^{r} L(u_i) p(v - b_n u_i) du_i.$$

The key observation is that $\text{Var}(P^{r-1} h_v^*(X_1))$ does not vanish as $b_n \to 0$. Then we can find more primitive conditions for (C5). For example, (C5) holds if $L(\cdot) = \mathbb{1}\{\cdot \in (-1/2, 1/2)^d\}$, p is continuous on $\mathcal{V}$, and $\lim_{n \to \infty} b_n = 0$.

***Remark* 3.1.** *In the Supplement Material [52], in Section E.1, we relax the condition (C6-id'), requiring ε to have a light tail, instead of being bounded. Further, in Section E.2, we relax the condition (C6-sg'), allowing ε and V to be dependent.*

3.2. Size validity and power consistency

The following is the master theorem for the statistical guarantees for the stratified incomplete local simplex tests.

Theorem 3.2. *Consider the function class $\mathcal{H}^{id}$ or $\mathcal{H}^{sg}$. Assume that (C1)-(C5) hold and that (C6-id') (resp. (C6-sg')) holds for $\mathcal{H}^{id}$ (resp. $\mathcal{H}^{sg}$). Further, for some $\kappa, \kappa' > 0$,*

$$N := n^{\kappa} b_n^{-dr}, \quad N_2 := n^{\kappa'} b_n^{-dr}. \tag{16}$$

Then there exists a constant C, depending only on C_0, d, κ, κ', such that with probability at least $1 - Cn^{-1/C}$,

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}(\mathbb{M}_n \leqslant t) - \mathbb{P}_{|\mathcal{D}'_n}(\mathbb{M}_n^{\#} \leqslant t) \right| \leqslant Cn^{-1/C}.$$

Proof. In Section E.1 and E.2 of the Supplement Material [52], we show that (PM) and (VC) is implied by (C1), where the latter is due to [26, Proposition 3.6.12]. (MB) is the same as (C2). Further, we verify that (MT-∞) holds with $D_n = Cb_n^{-d/2}$. Then the proof is complete by Theorem 2.3 and due to the requirement on the bandwidth, i.e., (C3). ■

***Remark* 3.3.** *The main challenge in working with $\{h_v^{sg} : v \in \mathcal{V}\}$ (and also with $\mathcal{H}^{id}$) is that the size of the projections of the kernels, $\{P^{r-\ell}|h_v^{sg}| : \ell = 0, 1, \ldots, r\}$ has different orders of magnitude due to localization. The same is true for the absolute moments of $\{|h_v^{sg}|^s\}$ for $s \geqslant 1$. Specifically, in Section E.2 of the Supplement Material [52], we verify (MT-∞) holds for any $\bar{q} \geqslant 1$. Thus for a fixed s, projections onto consecutive levels differ by a factor of $b_n^{-d(1-1/\bar{q})}$. On the other hand, for a fixed ℓ, the second moment ($s = 2$) is greater than the first moment ($s = 1$) by a factor $b_n^{-d(r-1/2)}$.*

The next Corollary establishes the size validity of the proposed procedure. Among all concave functions, affine functions have the (asymptotically) largest rejection probabilities, which attain the nominal levels uniformly over $(0, 1)$ for large n.

Corollary 3.4 (Size validity)**.** *Consider the procedure (15) for testing the hypothesis (2) with $\mathcal{H}^*$ for $* = id$ or sg. Assume the conditions in Theorem 3.2 hold. If the regression function f is concave, i.e., H_0 holds, then for some constant C, depending only on C_0, d, κ, κ',*

$$\mathbb{P}\left(\sup_{v \in \mathcal{V}} \sqrt{n} U'_{n,N}(h_v^*) \ \geqslant \ q_\alpha^{\#} \right) \leqslant \alpha + Cn^{-1/C}, \text{ for any } \alpha \in (0, 1).$$

Further, if f is an affine function, then

$$\sup_{\alpha \in (0,1)} \left| \mathbb{P}\left(\sup_{v \in \mathcal{V}} \sqrt{n} U'_{n,N}(h_v^*) \ \geqslant \ q_\alpha^{\#} \right) - \alpha \right| \leqslant Cn^{-1/C}.$$

Proof. If f is concave, then $P^r h_v^* \leqslant 0$ for $v \in \mathcal{V}$. Further, if f is affine, then $P^r h_v^* = 0$ for $v \in \mathcal{V}$. Then the results follow from Theorem 3.2. ■

The next Corollaries concern the power of the proposed procedure. The proofs can be found in Section E.3 of the Supplement Material [52].

Corollary 3.5 (Power)**.** *Consider the setup as in Corollary 3.4. If in addition*

$$\sqrt{n} P^r h_{v_n}^* \geqslant (C_0)^{-1} n^{\kappa''}, \text{ for some } v_n \in \mathcal{V}, \ \ \kappa'' > \max\{(1 - \kappa)/2, 0\}, \tag{17}$$

then for some constant C, depending only on $C_0, d, \kappa, \kappa', \kappa''$,

$$\mathbb{P}\left(\sup_{v\in\mathcal{V}} \sqrt{n}U'_{n,N}(h^*_v) \;\geqslant\; q^{\#}_{\alpha}\right) \geqslant 1 - Cn^{-1/C}, \text{ for any } \alpha \in (0,1).$$

***Remark* 3.6.** *The condition (17) ensures that the bias $\sqrt{n}P^r h^*_{v_n}$ is significantly larger than the standard deviation of $\mathbb{M}_n$. If $\kappa \geqslant 1$, then κ'' in (17) can be arbitrarily small. Note that due to (C2), the impact of M is absorbed into the constant C.*

Next we provide examples for which (17) holds, and focus on the class $\mathcal{H}^{\mathrm{id}}$. The discussion for $\mathcal{H}^{\mathrm{sg}}$ is similar.

Corollary 3.7 (Power - smooth f)**.** *Consider the setup as in Corollary 3.4. Assume that f is fixed and twice continuously differentiable at some $v_0 \in \mathcal{V}$ with a positive definite Hessian matrix at v_0, and that $\lim_{n\to\infty} b_n = 0$. Then $\liminf_{n\to\infty} P^r h^{id}_{v_0} / \left(b_n^{2+d/2}\right) > 0$. Thus if $b_n^{-(d+4)} \leqslant C_0 n^{1-2\kappa''}$ for some $\kappa'' > \max\{(1-\kappa)/2, 0\}$, then for any $\alpha \in (0,1)$, the power converges to one as $n \to \infty$.*

***Remark* 3.8.** *Note that Theorem 7 in [1] establishes the consistency of their test using $\mathcal{H}^{sg}$ (discrete, complete version) under the condition that $nb_n^{d+4}/\log(n) \to \infty$, which is in the same spirit as the requirement on b_n in Corollary 3.7.*

Corollary 3.9 (Power - piecewise affine f)**.** *Consider the setup as in Corollary 3.4. For $j \in \{1,2\}$, let $\theta_{n,j} \in \mathbb{R}^d$ and $\omega_{n,j} \in \mathbb{R}$ such that $\theta_{n,1} \neq \theta_{n,2}$. Let*

$$f(v) = f_n(v) := \max\left\{f_{n,1}(v), f_{n,2}(v)\right\}, \quad \textit{where} \quad f_{n,j}(v) := \theta_{n,j}^T v + \omega_{n,j} \quad \textit{for } j = 1,2.$$

If there exists $v_n \in \mathcal{V}$ such that $f_{n,1}(v_n) = f_{n,2}(v_n)$ for each n, then

$$\liminf_{n\to\infty} P^r h^{id}_{v_n} / \left(b_n^{1+d/2} \|\theta_{n,1} - \theta_{n,2}\|_2\right) > 0.$$

Thus if $b_n^{-(d+2)} \leqslant C_0 n^{1-2\kappa''} \|\theta_{n,1} - \theta_{n,2}\|_2^2$ for some $\kappa'' > \max\{(1-\kappa)/2, 0\}$, then for any $\alpha \in (0,1)$, the power converges to one as $n \to \infty$.

***Remark* 3.10.** *If f does not depend on n, in particular $\theta_{n,j} = \theta_j$ for each n, then the requirement on b_n becomes $b_n^{-(d+2)} \leqslant C_0 n^{1-2\kappa''}$, which is weaker than that for smooth functions f in Corollary 3.7.*

On the other, if we choose $b_n = b > 0$ for each n, then to achieve power consistency, we require $\|\theta_{n,1} - \theta_{n,2}\|_2 \geqslant C_0^{-1} n^{-1/2+\kappa''}$. Observe that f_n is convex if $\theta_{n,1} \neq \theta_{n,2}$, and affine if $\theta_{n,1} = \theta_{n,2}$. Thus this allows "local alternatives" that approach the null at the rate of $n^{-1/2+\kappa''}$.

3.2.1. Discussions

The stratified incomplete local simplex test (SILS) is a least favorable configuration test, with affine functions being (asymptotically) least favorable. This type of test was first proposed for testing the monotonicity of a (univariate) regression function by [24], and then extended to test the (multivariate, coordinate-wise) stochastic monotonicity by [41], and to test the (multivariate) convexity by [1]. See also [13] for the distribution approximation of these test statistics. It is not clear how to extend this idea to test other shape constraints, such as quasi-convexity [38], because it seems difficult to identify the least favourable configuration, or to compute the expectation of test statistics under it.

From Corollary 3.4, the SILS test is asymptotically non-conservative; however, it is non-similar [42], in the sense that for strictly concave functions, the probability of rejection is strictly less than the nominal level α. Being non-similar alone is not evidence again the SILS test (e.g., Z-test for normal means is optimal despite being non-similar), but a least favorable configuration test may be less powerful than alternative tests. The condition (17) requires "local convex curvature" of f for the test to be power consistent; see Corollary 3.7 and 3.9 for examples. The question of how (17) is related to the global L_2 separation rate (see discussions below) is left for future research.

In Section E.4 of the Supplement Material [52], we discuss the L_2 minimax separation rate for concavity test, and an alternative test ("FS" test) [22], which (almost) achieves the minimax rate for smooth functions for $d = 1$ and may do so for $d \geqslant 2$; thus the FS test is expected to have decent power. We note that the validity of our SILS test does not require f being smooth, and that in simulation studies (Section 5) it achieves comparable power to the FS test. In contrast, the FS test fails to control the size properly when f is not smooth (e.g., piecewise affine); this is observed in Section 5, and we also provide a detailed explanation in the Supplement Material [52] (e.g., if $d = 2$, it requires f to be Hölder continuous with smoothness parameter $s > 4$).

3.3. Combining multiple bandwidths

The theory in Subsection 3.2 does not suggest a particular choice for the bandwidth b_n. Since the size validity holds for a wide range of b_n, its selection depends on the targeted alternatives. If the targets are "globally" convex, then b_n should be large in order for the bias, $\{\sqrt{n}P^r h_v^* : v \in \mathcal{V}\}$, to be large. On the other hand, if the targets are only convex in a small region, then b_n should be able to localize those convex regions. See Subsection 5.4 for concrete examples.

One possible remedy is to use multiple bandwidths. Let $\mathcal{B}_n \subset (0, \infty)$ be a *finite* collection of bandwidths. For each $b \in \mathcal{B}_n$, we denote the function h_v^{id} in (14) (resp. h_v^{sg} in (4)) by $h_{v,b}^{\mathrm{id}}$ (resp. $h_{v,b}^{\mathrm{sg}}$) to emphasize the dependence on the bandwidth, and $\mathcal{H}_b^* = \{h_{v,b}^* : v \in \mathcal{V}\}$ for $* = \mathrm{id}$ or sg. Further, for each $b \in \mathcal{B}_n$, let N_b and $N_{2,b}$ be two computational parameters, and consider two independent collections of Bernoulli random variables

$$\mathcal{S}_b := \left\{ Z_\iota^{(m,b)} : m \in [M],\ \iota \in I_{n,r} \right\} \overset{i.i.d.}{\sim} \text{Bernoulli}(p_{n,b}),$$

$$\mathcal{S}_b' := \left\{ Z_\iota^{(k,m,b)} : k \in [n],\ m \in [M],\ \iota \in I_{n-1,r-1}^{(k)} \right\} \overset{i.i.d.}{\sim} \text{Bernoulli}(q_{n,b}),$$

where $p_{n,b} := N_b / |I_{n,r}|$, $q_{n,b} := N_{2,b} / |I_{n-1,r-1}|$, and they are independent of X_1^n. In other words, the sampling plan is independent for each $b \in \mathcal{B}_n$.

Then for each $b \in \mathcal{B}_n$, we denote $U_{n,N}'(h)$ in (6) by $U_{n,N,b}'(h)$ with the sampling plan given by $\mathcal{S}_b$. Similarly, we denote $\mathbb{G}^{(k)}(h)$ and $\bar{\mathbb{G}}(h)$ in (8) by $\mathbb{G}^{(k,b)}(h)$ and $\overline{\mathbb{G}}^{(b)}(h)$ respectively with the sampling plan given by $\mathcal{S}_b'$.

Now let $\mathcal{D}_n' := X_1^n \cup \{\mathcal{S}_b, \mathcal{S}_b' : b \in \mathcal{B}_n\}$, and denote Gaussian multipliers by

$$\{\xi_k : k \in [n]\},\ \left\{ \xi_\iota^{(m,b)} : m \in [M],\ \iota \in I_{n,r}, b \in \mathcal{B}_n \right\} \overset{i.i.d.}{\sim} N(0,1),$$

independent of $\mathcal{D}_n'$. Define for $b \in \mathcal{B}_n$ and $v \in \mathcal{V}_m$,

$$\mathbb{U}_{n,*,b}^{\#}(h_{v,b}^*) := r\mathbb{U}_{n,A,b}^{\#}(h_{v,b}^*) + (n/N_b)^{1/2}\mathbb{U}_{n,B,b}^{\#}(h_{v,b}^*),$$

where $\mathbb{U}^{\#}_{n,A,b}(h^*_{v,b}) := n^{-1/2}\sum_{k=1}^{n}\xi_k(\mathbb{G}^{(k,b)}(h^*_{v,b}) - \overline{\mathbb{G}}^{(b)}(h^*_{v,b}))$, and

$$\mathbb{U}^{\#}_{n,B,b}(h^*_{v,b}) := \left(\sum_{\iota\in I_{n,r}} Z_\iota^{(m,b)}\right)^{-1/2} \sum_{\iota\in I_{n,r}} \xi_\iota^{(m,b)}\sqrt{Z_\iota^{(m,b)}}(h^*_{v,b}(X_\iota) - U'_{n,N,b}(h^*_{v,b})).$$

Finally, for each $\alpha\in(0,1)$, denote by $q^{\#}_\alpha$ the $(1-\alpha)^{th}$ quantile of $\sup_{b\in\mathcal{B}_n, v\in\mathcal{V}}\mathbb{U}^{\#}_{n,*,b}(h^*_{v,b})$, conditional on $\mathcal{D}'_n$. Then we propose to reject the null in (2) if and only if

$$\sup_{b\in\mathcal{B}_n, v\in\mathcal{V}} \sqrt{n}U'_{n,N,b}(h^*_{v,b}) \;\geqslant\; q^{\#}_\alpha. \tag{18}$$

***Remark* 3.11.** *It is possible to allow $\mathcal{B}_n$ to be uncountable, for example, $\mathcal{B}_n := [\ell_n, u_n]$, which corresponds to the uniform in bandwidth results [20, 13]. However, we choose to present the results for a finite $\mathcal{B}_n$ for simplicity, since otherwise we need to also stratify $\mathcal{B}_n$. This approach has a similar spirit to the multi-scale testing of qualitative hypotheses [19].*

To establish the size validity and analyze the power of the test (18), we need a more general theory than those in Section 2 for a function class $\{h_{v,b} : v\in\mathcal{V}, b\in\mathcal{B}_n\}$, where $\mathcal{V}$ is an index set. The key difference is that for each $b\in\mathcal{B}_n$, the computational parameters N_b and $N_{2,b}$ may be of a different order (see, e.g., (16)). The rigorous statements for $\{h_{v,b} : v\in\mathcal{V}, b\in\mathcal{B}_n\}$, which follow from similar arguments as those in Section 6, are not included for simplicity of the presentation. In Subsection 5.4, we conduct a simulation study to investigate the empirical performance of the SILS test with multiple bandwidths.

4. Stratified incomplete local simplex tests: computation

In this section, we discuss the computational complexity and implementation for the stratified incomplete local simplex tests. We focus on $\mathcal{H}^{\mathrm{id}}$ in our discussion and omit the superscript for simplicity. Assume that (C1)-(C6-id') hold, and that the computational parameters N, N_2 are given in (16). Further, as $\mathcal{V}$ is compact, we assume $\mathcal{V}\subset[0,1]^d$ without loss of generality.

For some small $\eta\in(0,1/2)$, let $t := \lfloor 1/(\eta b_n)\rfloor$, and $\tau_i = i\eta b_n$ for $i=0,1,\ldots,t$ and $\tau_{t+1}=1$. Now we partition each coordinate into segments of length ηb_n (except for the rightmost one), i.e., each $\mathcal{V}_m$ is determined by an ordered list $(j_1,\ldots,j_d)$ such that $0\leqslant j_k < t$ for $k\in[d]$ and $\mathcal{V}_m = \{v\in\mathcal{V} : \tau_{j_k}\leqslant v_k < \tau_{j_k+1} \text{ for } k\in[d]\}$. Then the number of partitions $M\leqslant(1+\eta^{-1}b_n^{-1})^d$.

For any $v\in\mathbb{R}^d$ and $A\subset\mathbb{R}^d$, we denote the b_n-neighbourhood by

$$\mathcal{N}(v,b_n) := \{v'\in\mathbb{R}^d : \|v-v'\|_\infty\leqslant b_n/2\},\quad \mathcal{N}(A,b_n) := \bigcup_{v\in A}\mathcal{N}(v,b_n).$$

Denote by $\mathrm{ND}(v,b_n) := \{i\in[n] : V_i\in\mathcal{N}(v,b_n)\}$ and $\mathrm{ND}(A,b_n) := \bigcup_{v\in A}\mathrm{ND}(v,b_n)$ the indices for data points within b_n-neighbourhood of v and A respectively.

As an illustration, in Figure 1 (where $b=8, \eta=1/8$), $\mathcal{V}$ is partitioned into small squares of size 1. For the dotted region $\mathcal{V}_m$, $\mathcal{N}(\mathcal{V}_m,b_n)$ is area encompassed by the big dotted square, so $v_4\in\mathcal{N}(\mathcal{V}_m,b_n)$, but $v_5\notin\mathcal{N}(\mathcal{V}_m,b_n)$. Further, $\mathrm{ND}(\mathcal{V}_m,b_n)$ are indices for data points within the dotted square.

4.1. Stratified sampling

For $m \in [M]$, let $\mathcal{A}(\mathcal{V}_m) := \{\iota = (i_1, \ldots, i_r) \in I_{n,r} : i_j \in \mathrm{ND}(\mathcal{V}_m, b_n) \text{ for } j \in [r]\}$ be the collection of r-tuples whose members are all within b_n-neighbourhood of $\mathcal{V}_m$. For example, in Figure 1, $(v_1, v_2, v_3, v_4) \in \mathcal{A}(\mathcal{V}_m)$, but $(v_1, v_2, v_3, v_5) \notin \mathcal{A}(\mathcal{V}_m)$. Due to the localization by $L(\cdot)$ (cf. (C1)),

$$h_v(x_\iota) = 0, \text{ for any } v \in \mathcal{V}_m \text{ and } \iota \in |I_{n,r}| \setminus \mathcal{A}(\mathcal{V}_m).$$

As a result, the individual values of $\{Z_\iota^{(m)} : \iota \in |I_{n,r}| \setminus \mathcal{A}(\mathcal{V}_m)\}$ are irrelevant, except for their sum, which is a part of $\widehat{N}^{(m)}$. Thus, we generate a Binomial$(|I_{n,r} \setminus \mathcal{A}(\mathcal{V}_m)|, p_n)$ random variable, that accounts for $\sum_{\iota \in |I_{n,r}| \setminus \mathcal{A}(\mathcal{V}_m)} Z_\iota^{(m)}$.

On the other hand, the number of selected r-tuples in $\mathcal{A}(\mathcal{V}_m)$ is on average

$$\mathbb{E}\left[\sum_{\iota \in \mathcal{A}(\mathcal{V}_m)} Z_\iota^{(m)}\right] \lesssim \binom{n(1+\eta)^d b_n^d}{r} \frac{n^\kappa b_n^{-dr}}{|I_{n,r}|} \lesssim (1+\eta)^{dr} n^\kappa,$$

since the $\|\cdot\|_\infty$-diameter of $\mathcal{V}_m$ is ηb_n, and the density of V is bounded (see (C4)). Thus to compute $\sup_{v \in \mathcal{V}_m} \sqrt{n} U'_{n,N}(h_v)$, the number of evaluations of $w(\cdot)$ is on average $\lesssim n^\kappa$, and the computational complexity can be made independent of the dimension d (as η can be chosen to be small).

***Remark* 4.1.** *Above calculation of complexity does not include the cost of maximizing over $\mathcal{V}_m$. In practice, we select a finite number of query points as in Subsection 4.2. The discussion for the bootstrap part is similar, and we analyze below the complexity of its actual implementation.*

Why stratification? Without stratifying $\mathcal{V}$, each $v \in \mathcal{V}$ share the same sampling plan $\{Z_\iota : \iota \in I_{n,r}\}$. However, we cannot afford to generate all $\{Z_\iota : \iota \in I_{n,r}\}$, as on average there are $N = n^\kappa b_n^{dr}$ non-zero terms. We may attempt to use the above short-cut. For $v_1, v_2 \in \mathcal{V}$, to compute $U'_{n,N}(h_{v_i})$ (for $i = 1, 2$), we only generate $\{Z_\iota : \iota \in \mathcal{A}(\{v_i\})\}$, and the individual values of $\{Z_\iota : \iota \in I_{n,r} \setminus \mathcal{A}(\{v_i\})\}$ are not explicitly generated.

However, the issue is to ensure consistency. (i) In computing $U'_{n,N}(h_{v_1})$, although the individual values of $\{Z_\iota : \iota \in I_{n,r} \setminus \mathcal{A}(\{v_1\})\}$ are irrelevant, we still need to generate a Binomial random variable to account for their sum. However, $\big(I_{n,r} \setminus \mathcal{A}(\{v_1\})\big) \cap \mathcal{A}(\{v_2\})$ in many cases is non-empty, and thus $\sum_{\iota \in |I_{n,r}| \setminus \mathcal{A}(\{v_1\})} Z_\iota^{(m)}$ and $\{Z_\iota : \iota \in \mathcal{A}(\{v_2\})\}$ are not independent. (ii) In many cases, $\mathcal{A}(\{v_1\}) \cap \mathcal{A}(\{v_2\})$ is non-empty, so we cannot independently generate $\{Z_\iota : \iota \in \mathcal{A}(\{v_1\})\}$ and $\{Z_\iota : \iota \in \mathcal{A}(\{v_2\})\}$. Note also that the calculation is needed for multiple $v \in \mathcal{V}$ instead of only v_1, v_2.

***Remark* 4.2.** *In Section E.5 of the Supplement Material [52], we present an algorithm without stratification that addresses the above consistency issue. Its computational complexity is $\lesssim 2^{dr} n^\kappa b_n^{-d}$ evaluations of $w(\cdot)$. If d is fixed, it only loses a b_n^{-d} factor in theory, but 2^{dr} can be very large in practice, and thus it is not computationally feasible (e.g., $2^{dr} = 32768$ if $d = 3$).*

4.2. Implementation of SILS

In practice, instead of taking the supremum over $\mathcal{V}$, we choose a (finite) collection of query points, $\mathscr{V}_n$, one from each partition $\{\mathcal{V}_m : m \in [M]\}$, and approximate the supremum over $\mathcal{V}$ by that over $\mathscr{V}_n$.

As a result, each $v \in \mathscr{V}_n$ has its individual sampling plan ($\{Z_\iota^{(m)} : \iota \in I_{n,r}\}$ if $v \in \mathcal{V}_m$), which can be generated independently for different query points. Further, the test still takes the form of (15) with a finite function class $\mathcal{H} = \{h_v : v \in \mathscr{V}_n\}$.

***Remark* 4.3.** *It is without loss of generality to pick one query point from each region, since we could always decrease η, i.e., making each region smaller. Further, unlike [1], we do not require query points to be well separately, that is, for small η, there are pairs fo queries points $v, v' \in \mathscr{V}_n$, such that $\|v - v'\|_\infty \ll b_n$. Finally, since only one element is picked, if $v \in \mathcal{V}_m$, instead of considering ND$(\mathcal{V}_m, b_n)$, we can focus on ND$(\{v\}, b_n)$.*

***Remark* 4.4.** *In establishing the bootstrap validity for stratified, incomplete U-processes, we first consider the corresponding results for high-dimensional U-statistics, and then approximate the supremum of a U-process by that of its discretized version. Thus the above procedure, which can be viewed as a practical implementation of approximating the supremum of a process, can also be directly justified by Theorems in Section B of the Supplement Material [52].*

Computing the test statistic. In Algorithm 1, we show the pseudo-code to compute, for each $v \in \mathscr{V}_n$, the statistic $U'_{n,N}(h_v)$, and at the same time the conditional (given $\mathcal{D}'_n$) variance $\widehat{\gamma}_B(h_v)$ of $\mathbb{U}^{\#}_{n,B}(h_v)$ in (10); note that we write $\widehat{\gamma}_B(h_v)$ for $\widehat{\gamma}_B(h_v, h_v)$, which is defined following (10). It is well known that sampling T items without replacement from S elements ($S \gg T$) can be done in $O(T \log(T))$ time [30]. Then based on the discussions in the previous subsection, the computational complexity for Algorithm 1 is $O(Mn^\kappa \log(n))$.

Bootstrap. For a fixed $k \in [n]$, notice that $\{\mathbb{G}^{(k)}(h_v) : v \in \mathscr{V}_n\}$ in (8) takes the same form of stratified, incomplete U-processes as the test statistics, and thus we can apply Algorithm 1, with appropriate inputs, to compute it. Since we need to compute $\mathbb{G}^{(k)}$ for each $k \in [n]$, the complexity is

$$n \times M \times \binom{(n-1)b_n^d}{r-1} \frac{n^{\kappa'} b_n^{-dr}}{|I_{n-1,r-1}|} \log(n) \lesssim M n^{1+\kappa'} b_n^{-d} \log(n).$$

Further, as we pick one element from each $\mathcal{V}_m$, given $\mathcal{D}'_n$, $\left\{\mathbb{U}^{\#}_{n,B}(h_v) : v \in \mathscr{V}_n\right\}$ are conditionally independent, with variances $\{\widehat{\gamma}_B(h_v) : v \in \mathscr{V}_n\}$ already computed in Algorithm 1, and we no longer need to generate Gaussian multipliers $\{\xi_\iota^{(m)}\}$ for each summand indexed by ι in (10).

Finally, for independent standard Gaussian multipliers $\{\xi_k, \xi^{(m)} : k \in [n], m \in [M]\}$, we compute for each $v \in \mathscr{V}_n$,

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \xi_k \left(\mathbb{G}^{(k)}(h_v) - \overline{\mathbb{G}}(h_v)\right) + \xi^{(\sigma(h_v))} \sqrt{\widehat{\gamma}_B(h_v)}.$$

Since $\{\mathbb{G}^{(k)} : k \in [n]\}$ and $\{\widehat{\gamma}_B(h_v) : v \in \mathscr{V}_n\}$ have already been computed, the complexity is $O(BMn)$, where B is the number of bootstrap iterations. Hence the overall computational cost is $O(Mn^\kappa \log(n) + Mn^{1+\kappa'} b_n^{-d} \log(n) + BMn)$.

***Remark* 4.5.** *The computational bottleneck is in computing $\{\mathbb{G}^{(k)} : k \in [n]\}$, which, however, is outside the bootstrap iterations. Thus we can afford large B in the bootstrap calibration. The above algorithms can be implemented in a parallel manner using clusters; in particular, $\mathbb{G}^{(k)}$ can be computed separately for each $k \in [n]$. As a result, the efficiency scales linearly in the number of computing cores.*

Input: Observations $\{X_i = (V_i, Y_i) \in \mathbb{R}^{d+1} : i \in [n]\}$, budget N, kernel $L(\cdot)$, bandwidth b_n, query points $\mathscr{V}_n$ (size M).
Output: $U'_{n,N}$, $\widehat{\gamma}_B$: two vectors of length M
Initialization: $p_n = N/\binom{n}{r}$, $U'_{n,N}, \widehat{\gamma}_B$ both set zero ;
for $m \leftarrow 1$ **to** M **do**
  $v = \mathscr{V}_n[m]$;
  Generate $T_1 \sim \text{Binom}(\binom{|\text{ND}(v,b_n)|}{r}, p_n)$, $T_2 \sim \text{Binom}(\binom{n}{r} - \binom{|\text{ND}(v,b_n)|}{r}, p_n)$;
  $\widehat{N} \leftarrow T_1 + T_2$;
  Sample without replacement T_1 terms, $\{\iota_\ell : 1 \leqslant \ell \leqslant T_1\}$, from $\mathcal{A}(\{v\})$;
  for $\ell \leftarrow 1$ **to** T_1 **do**
    $U'_{n,N}[m] \leftarrow U'_{n,N}[m] + h_v(X_{\iota_\ell})$;
    $\widehat{\gamma}_B[m] \leftarrow \widehat{\gamma}_B[m] + (h_v(X_{\iota_\ell}))^2$;
  end
  $U'_{n,N}[m] \leftarrow U'_{n,N}[m]/\widehat{N}$, $\widehat{\gamma}_B[m] \leftarrow \widehat{\gamma}_B[m]/\widehat{N} - (U'_{n,N}[m])^2$
end

Algorithm 1: compute $U'_{n,N}$ and $\widehat{\gamma}_B$ over $\mathscr{V}_n$ for the concavity test.

5. Simulation results

In the simulation studies, we consider setups where the regression function f in (1) is defined on $(0,1)^d$, and the covariates $V = (V_1, \ldots, V_d)$ have a uniform distribution on $(0,1)^d$, for $d = 2, 3, 4$. In this section, the error term ε in (1) has a Gaussian distribution with zero mean and variance σ^2.

***Remark* 5.1.** *The results for $d = 3$ and 4 are qualitatively similar, and presented mostly in Section D of the Supplement Material [52], where we also study asymmetric or heavy tailed distributions for the noise ε (Section D.3).*

We compare our proposed procedure with the method in [22], denoted by "FS".
Proposed procedure. We use the uniform localization kernel $L(\cdot) = \mathbb{1}\{\cdot \in (-1/2, 1/2)^d\}$. The query points are $\mathscr{V}_n := \{0.3, 0.4, 0.5, 0.6, 0.7\}^2$ for $d = 2$, and $\mathscr{V}_n := \{0.3, 0.5, 0.7\}^d$ for $d = 3, 4$. For parameters related to the computational budget, we set $N = 10 \times 25 \times n \times b_n^{-d\times r}$ for $d = 2, 3, 4$, $N_2 = 10^4 \times b_n^{-d\times r}$ for $d = 2, 3$ and $N_2 = 2 \times 10^4 \times b_n^{-d\times r}$ for $d = 4$, and the Bootstrap iterations $B = 1500$. The N is selected so that $\alpha_n := n/N$ is very small, and further increasing it will not improve the power of the test. The estimation of $\{\mathbb{G}^{(k)}(h_v^*) : k \in [n], v \in \mathscr{V}_n\}$ is the computational bottleneck, and empirically we find that further increasing the selected value for N_2 does not improve the accuracy in terms of the size of the proposed procedure. We consider two types of kernels, $\mathcal{H}^{\text{id}}$ and $\mathcal{H}^{\text{sg}}$, and use below "ID" for the former and "SG" for later. For each parameter configuration below, we independently generate (at least) 1,000 datasets, apply our procedure, and estimate the rejection probability.
FS method [22]. We use the implementation provided by the authors[2], where either quadratic or cube splines with j knots in each coordinate are used in constructing an initial estimator for the regression function; we denote the former by FS-Qj and later by FS-Cj. We set the tuning parameter $\gamma_n = 0.01/\log(n)$ and the Bootstrap iteration $B = 200$ as recommended by [22]. Below, the rejection probabilities are estimated based on 1,500 independently generated datasets.

[2]code for [22]: https://www.dropbox.com/s/jmjshxznu31tnn2/ShapeCode.zip?dl=0.

5.1. Running times

The computational savings compared to using the complete U-process, $p_n := N/\binom{n}{r}$ and $q_n = N_2/\binom{n-1}{r-1}$, are listed in Table 2 for several typical configurations. It is clear that for a moderate size dataset (say $n \sim 1000$), using the complete U-process has a very high, if not prohibitive, computational cost (see Table 1 for the running time using the stratified, incomplete U-process). For example, for $d = 3, n = 1000, b_n = 0.6$, it takes on average 5.26 minutes to run our procedure with 40 cores, which implies that with the complete version it would take at least 7.2 days ($= 5.26$ mins$/q_n$).

In contrast, the FS method [22] has a much shorter running time. For example, with $d = 2, n = 1000$, it takes less than 20 seconds with 4 cores (see Table E4 in [22]). For $d \geqslant 3$, it could be challenging to apply the FS method due to the accuracy of estimating the regression function, the projection onto a function space, and the numerical integration needed to compute the distance etc.

$d=2, b_n=0.5$	$d=3, b_n=0.6, \lvert\mathscr{V}_n\rvert=27$			$d=4, b_n=0.7$
$n=1000, \lvert\mathscr{V}_n\rvert=25$	$n=500$	$n=1000$	$n=1500$	$n=2000, \lvert\mathscr{V}_n\rvert=81$
5.06 mins	1.31 mins	5.26 mins	9.12 mins	33.4 mins

Table 1. Running time of the proposed procedure in minutes using 40 computer cores, where N and N_2 are described in the introduction of Section 5.

	p_n			q_n		
	$n=500$	$n=1000$	$n=1500$	$n=500$	$n=1000$	$n=1500$
$d=2, b_n=0.5$	1.2E-2	1.5E-3	4.5E-4	1.2E-1	1.5E-2	4.6E-3
$d=3, b_n=0.6$	1.0e-3	6.4E-5	1.3E-5	8.3E-3	5.1E-4	1.0E-4

Table 2. Computational efficiency for typical configurations, where N and N_2 are described in the introduction of Section 5. For $d = 4, b_n = 0.7, n = 2000$, we have p_n =5.9E-8, q_n = 3.9E-7. Here, sE-t = s$\times 10^{-t}$.

5.2. Size validity

We start with our proposed procedure, and consider concave functions given by

$$f(v) = v_1^{\kappa_0} + \ldots + v_d^{\kappa_0}, \quad \text{for } v := (v_1, \ldots, v_d) \in (0,1)^d, \tag{19}$$

for $0 < \kappa_0 \leqslant 1$. Here, we consider the rejection probabilities for $\kappa_0 = 1$; that is, f is affine and thus the (asymptotically) least favourable configuration in the null. In Section D.2 of the Supplement Material [52], we present results for strictly concave function with $0 < \kappa_0 < 1$.

For each query point, the average number of data points within its b_n neighbourhood is $n \times b_n^{-d}$. Since a decent size of local points is necessary for the validity of Gaussian approximation, we select b_n so that locally there are at least 150 data points. As we shall see in Subsection 5.3, smaller b_n has a better localization power, while larger b_n is suitable if the targeted alternatives are globally convex.

In Table 3, we list the size for different bandwidth b_n and error variance σ^2 at levels 5% and 10% for f in (19) with $\kappa_0 = 1$. From the Table 3, it is clear that the proposed procedure is consistently on the conservative side. We note that the conservativeness is not due to the stratified sampling. For $d = 1$, we were able to implement the complete version, and observed a similar phenomenon. Further, [13] uses complete U-processes to test regression monotonicity, which are also conservative (see Table 1 therein).

FS method [22]. In Table E2 of [22], we can observe the slight inflation of the empirical size of the FS method when the function is linear. Here, we consider the following *concave*, piecewise affine regression function:

$$f(v_1, v_2) = -|v_1 - 0.8| - |v_2 - 0.8|, \quad \text{for } v_1, v_2 \in (0, 1). \tag{20}$$

The rejection probabilities of the FS method [22] at the nominal level 5% are listed in Table 4. Recall that FS-Qj (resp. FS-Cj) is for using quadratic (resp. cubic) splines with j knots in each coordinate as the initial estimator for the regression function. Except for the global test FS-Q0, which places no interior knots, these probabilities far exceed the nominal level. We provide explanations for the significant size inflation of the FS method [22] in Section E.4 of the Supplement Material [52].

	$n = 500$			$n = 1000$		
$d = 2$, Level = 5%	$b_n = 0.6$	$b_n = 0.55$	$b_n = 0.5$	$b_n = 0.5$	$b_n = 0.45$	$b_n = 0.4$
ID, $\sigma = 0.1$	3.1	2.9	2.8	4.1	2.6	3.4
SG, $\sigma = 0.1$	3.1	4.1	2.8	3.0	3.8	2.7
ID, $\sigma = 0.2$	3.2	3.7	3.0	2.9	3.1	2.4
SG, $\sigma = 0.2$	3.6	3.3	2.9	3.6	3.1	2.5
$d = 2$, Level = 10%	$b_n = 0.6$	$b_n = 0.55$	$b_n = 0.5$	$b_n = 0.5$	$b_n = 0.45$	$b_n = 0.4$
ID, $\sigma = 0.1$	8.3	6.8	6.7	8.1	7.5	7.7
SG, $\sigma = 0.1$	7.4	8.0	6.6	8.1	7.7	6.1
ID, $\sigma = 0.2$	7.8	8.0	6.2	7.6	7.8	6.0
SG, $\sigma = 0.2$	8.7	7.1	7.0	8.6	7.0	6.7

Table 3. Size validity of the proposed procedure for $d = 2$. The probabilities of rejection under affine regression functions, are in the unit of percentage.

Knots j	0	1	2	3	4	5	6	7	8	9
FS-Qj	0	99.5	46.5	9.2	23.6	33.9	16.0	16.0	23.1	20.3
FS-Cj	97.1	50.7	15.3	32.8	23.9	13.6	17.7	21.0	19.9	22.2

Table 4. The rejection probabilities (in percentage) for FS-Qj and FS-Cj [22] at level 5% for the concave (i.e. H_0 holds) function in (20), where $n = 1000$ and $\sigma = 0.1$.

5.3. Power comparison

We study two types of alternatives for the regression function.

Polynomial functions. In the first, we consider f in (19) for $\kappa_0 \in \{1.2, 1.5\}$.

Locally convex functions. For the second, we consider regression functions that are mostly concave over $(0, 1)^d$, but convex in a small region. Specifically, let $\varphi(v) := \exp\left(-\|v\|^2/2\right)$ for $v \in \mathbb{R}^d$, which is concave on the region $\{v \in \mathbb{R}^d : \|v\|_\infty < 1\}$. Then for $c_1, c_2, \omega_1, \omega_2 > 0$ and $\mu_1, \mu_2 \in \mathbb{R}^d$, we consider

$$f(v) = c_1\varphi\left((v - \mu_1)/\omega_1\right) - c_2\varphi\left((v - \mu_2)/\omega_2\right), \text{ for } v \in (0, 1)^d. \tag{21}$$

We let $c_1 = 1, \omega_1 = 1.5$, and $\mu_1 = (0.75, \ldots, 0.75)$ so that without the second term, f would be concave in the entire region $(0,1)^d$. We let $\mu_2 = (0.25, \ldots, 0.25)$, and set c_2 and ω_2 to be small so that f is mostly concave and locally convex in a small neighbourhood of μ_2. In Section D.1 of the Supplement Material [52], we plot the regression function f together with one realization of dataset.

In Table 5 (a) and (b), for the two types of alternatives, we list the power of $\mathcal{H}^{id}$ with different bandwidth parameters b_n, and the FS method [22] using either quadratic (Q) or cubic (C) splines with $j = 0, 1, 2, 5$ knots in each coordinate[3]. For our proposed method, if f is a polynomial function (19), the power increases as b_n increases, as f is globally convex. However, for the locally convex function f (21), the power initially increases as b_n increases, but later drops significantly, as f is only locally convex, but "globally concave". Thus the choice of bandwidth depends on the targeted alternatives. Similar statements can be made about the FS method [22]. Adding knots decreases its power for (19), while a "global" test such as FS-Q0 has little power against (21).

In summary, the proposed procedure has a comparable power to the FS method [22], which however fails to control the size in general. Further, we show next that the issue with selecting b_n can be partly solved by combining multiple bandwidths.

	$\mathcal{H}^{\text{id}}$ with single b_n			FS method [22]							
Level 5%	$b_n = 0.6$	$b_n = 0.8$	$b_n = 1$	Q0	Q1	Q2	Q5	C0	C1	C2	C5
Rej. Prob.	25.6	69.1	96.6	93.8	81.6	57.7	37.7	85.6	60.5	50.7	36.3

(a) Polynomial f (19) with $\kappa_0 = 1.5$

	$\mathcal{H}^{\text{id}}$ with single b_n			FS method [22]							
Level 10%	$b_n = 0.5$	$b_n = 0.6$	$b_n = 0.7$	Q0	Q1	Q2	Q5	C0	C1	C2	C5
Rej. Prob.	20.3	40.3	15.8	7.1	49.7	28.5	30.4	44.3	26.5	27.7	29.4

(b) Locally convex f (21) with $\omega_2 = 0.15, c_2 = 0.3$

	Poly f (19) with $\kappa_0 = 1.5$ at 5%	Locally convex f (21) with $\omega_2 = 0.15, c_2 = 0.3$ at 10%
Rej. Prob.	71.7	39.8

(c) Multiple bandwidths $\{\mathcal{H}_b^{\text{id}} : b \in \{0.6, 0.8, 1\}\}$

Table 5. The rejection probabilities (in percentage) of the proposed method $\mathcal{H}^{\text{id}}$, the FS method [22], and the proposed method with multiple bandwidth $\{\mathcal{H}_b^{\text{id}} : b \in \{0.6, 0.8, 1\}\}$ for $d = 2$, $n = 1000$, $\sigma = 0.5$.

5.4. Combining multiple bandwidths

We consider the procedure (18) in Subsection 3.3 that combines multiple bandwidths, $\{\mathcal{H}_b^{\text{id}} : b \in \mathcal{B}_n\}$ with $\mathcal{B}_n = \{0.6, 0.8, 1\}$. For f in (19) with $\kappa_0 = 1$ (i.e. affine functions), the probability of rejection is 8.4% when the nominal level is 10%; for strictly concave functions with $\kappa_0 < 1$, the probabilities of rejection are listed in Section D.2 of the Supplement Material [52]. In Table 5 (c), we present its power against the two alternatives.

[3] $j = 0, 1$ is used in [22]

With a range of bandwidths, $\{\mathcal{H}_b^{\mathrm{id}} : b \in \mathcal{B}_n\}$ achieves a reasonable power, and is adaptive to the properties of the regression function f, with its computational cost linear in $|\mathcal{B}_n|$. As expected, it is not as powerful as the best performance achievable by $\mathcal{H}_{b_n}^{\mathrm{id}}$ with a single bandwidth b_n, which, however, is unknown in practice.

Thus, we would recommend the procedure (18) with multiple bandwidths $\mathcal{B}_n$. In choosing $\mathcal{B}_n$, one approach is to first decide reasonable lower and upper bounds, $b_{\min}$ and $b_{\max}$, for the bandwidth, and then based on available computational resource, select a few bandwidths in $[b_{\min}, b_{\max}]$ (say equally spaced) to form $\mathcal{B}_n$. As a rule of thumb, one may choose $b_{\min}$ so that there are enough data points in the b_n neighbourhood of each query point (say $\geqslant 120$). On the other, $b_{\max}$ could be decided based on the diameter of the region of interest, $\mathcal{V}$.

6. Gaussian approximation and bootstrap for stratified, incomplete U-processes

In this section, we consider a general function class $\mathcal{H}$, and establish Gaussian approximation and bootstrap results for its associated stratified, incomplete U-processes in Section 2, under the more general moment assumptions (MT) instead of (MT-∞). In particular, the condition (MT) does not require the envelope function H in (VC) to be bounded.

(MT). There exist absolute constants $\underline{\sigma} > 0$, $c_0 \in (0,1)$, $q \in [4,\infty]$, and $B_n \geqslant D_n \geqslant 1$ such that

$$\mathrm{Var}\left(P^{r-1}h(X_1)\right) \geqslant \underline{\sigma}^2, \ \text{ for } h \in \mathcal{H}, \tag{MT-0}$$

$$\sup_{h\in\mathcal{H}} \mathbb{E}\left|P^{r-1}h(X_1) - P^r h\right|^{2+k} \leqslant D_n^k \text{ for } \ k=1,2, \quad \|P^{r-1}H\|_{P,q} \leqslant D_n, \tag{MT-1}$$

$$\begin{aligned}
&\|P^{r-\ell}H^s\|_{P^\ell,2} \leqslant B_n^{2s-2}D_n^{\ell+1-s}, \text{ for } \ell \geqslant 2, s = 1,2,3,4,\\
&\|P^{r-\ell}H^s\|_{P^\ell,q} \leqslant B_n^{2s-2}D_n^{\ell(2-2/q)+2/q-s}, \text{ for } \ell = 1,2,\ s = 2,3,4,\ \ \|P^{r-2}H\|_{P^2,q} \leqslant D_n^{3-2/q},\\
&\|P^{r-\ell}|h|^s\|_{P^\ell,2} \leqslant B_n^{2s-2}D_n^{\ell-s}, \text{ for } \ell = 0,1,2,\ s \in [4] \text{ with } \ell + s > 2,\ h \in \mathcal{H},
\end{aligned} \tag{MT-2}$$

$$\|H\|_{P^r,q} \leqslant B_n^{2-2/q}D_n^{2/q-1}, \quad \|H\|_{P^r,2} \leqslant B_n, \tag{MT-3}$$

$$c_0 B_n^2 D_n^{-2} \leqslant \mathrm{Var}(h(X_1^r)) \leqslant \min\{D_n^{2(r-1)},\ B_n^2 D_n^{-2}\}, \text{ for } h \in \mathcal{H}, \tag{MT-4}$$

$$\sup_{h\in\mathcal{H}} \|P^{r-2}h\|_{P^2,4} \leqslant D_n^2, \quad \|(P^{r-2}H)^{\odot 2}\|_{P^2,q/2} \leqslant D_n^{4-4/q}, \tag{MT-5}$$

where $1/q = 0$ if $q = \infty$, and recall that for a measurable function $f : S^2 \to \mathbb{R}$, define $f^{\odot 2}$ to be a function on S^2 such that $f^{\odot 2}(x_1,x_2) := \int f(x_1,x)f(x_2,x)dP(x)$.

6.1. Gaussian approximation

We first approximate the supremum of the stratified, incomplete U-process (6) by that of an appropriate Gaussian process. Specifically, denote by $P^{r-1}h$ the function on S such that $P^{r-1}f(x_1) := \mathbb{E}[h(x_1, X_2, \ldots, X_r)]$, and $(\ell^\infty(\mathcal{H}), \|\cdot\|_\infty)$ the space of bounded functions indexed by $\mathcal{H}$ equipped

with the supremum norm. Assume there exists a tight Gaussian random variable W_P in $\ell^\infty(\mathcal{H})$ with zero mean and covariance function $\gamma_*(h,h') := \mathrm{Cov}\left(W_P(h), W_P(h')\right) = r^2\gamma_A(h,h') + \alpha_n\gamma_B(h,h')$ for $h,h' \in \mathcal{H}$ where $\alpha_n := n/N$, $\gamma_A(h,h') := \mathrm{Cov}\left(P^{r-1}h(X_1),\ P^{r-1}h'(X_1)\right)$, and $\gamma_B(h,h') := \mathrm{Cov}\left(h(X_1^r),\ h'(X_1^r)\right)\mathbb{1}\{\sigma(h)=\sigma(h')\}$. Note that $\gamma_B(h,h')=0$ if $\sigma(h)\neq\sigma(h')$, which is due to the stratification. The existence of W_P is implied by (VC) and (MT) (see [15][Lemma 2.1]). Further, denote $\widetilde{\mathbb{M}}_n := \sup_{h\in\mathcal{H}} W_P(h)$. We bound the Kolmogorov distance between the two suprema.

Theorem 6.1. *Assume the conditions (PM), (VC), (MB), and* (MT-0)-(MT-4)*. Then there exists a constant C, depending only on r, q, ϱ, c_0, C_0, such that*

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}(\mathbb{M}_n \leqslant t) - \mathbb{P}(\widetilde{\mathbb{M}}_n \leqslant t)\right| \leqslant C\eta_n^{(1)} + C\eta_n^{(2)}, \quad \textit{with}$$

$$\eta_n^{(1)} := \left(\frac{D_n^2K_n^7}{n}\right)^{1/8} + \left(\frac{D_n^2K_n^4}{n^{1-2/q}}\right)^{1/4} + \left(\frac{D_n^{3-2/q}K_n^{5/2}}{n^{1-1/q}}\right)^{1/2},$$

$$\eta_n^{(2)} := \left(\frac{B_n^2K_n^7}{N}\right)^{1/8} + \left(\frac{n^{4r/q}K_n^5B_n^{2-8/q}D_n^{8/q}}{N}\right)^{1/4} + \left(\frac{M^{2/q}B_n^{2-4/q}D_n^{4/q}K_n^5}{N^{1-2/q}}\right)^{1/4},$$

where $1/q = 0$ if $q=\infty$.

Proof. The proof can be found in Section C.2 of the Supplement Material [52]. The strategy is to first establish Gaussian approximation results for a finite, yet "dense", subset $\mathcal{H}'$ of $\mathcal{H}$ (Section B), and then approximate the supremum over $\mathcal{H}$ by that over $\mathcal{H}'$, which requires the local maximal inequalities developed in Section A.1. ■

6.2. Bootstrap validity

The next Theorem shows that conditional on $\mathcal{D}'_n$, the maximum of the bootstrap process, $\mathbb{M}_n^\#$ in (11), is well approximated by the maximum of W_P, $\widetilde{\mathbb{M}}_n$, in distribution.

Theorem 6.2. *Assume the conditions (PM), (VC), (MB), and* (MT-0)-(MT-5)*. Let*

$$\varrho_n := \left(\frac{M^{2/q}B_n^{2-4/q}D_n^{4/q}K_n^5}{(N\wedge N_2)^{1-2/q}}\right)^{1/4} + \left(\frac{B_n^2K_n^7}{N\wedge N_2}\right)^{1/8} +$$
$$\left(\frac{D_n^2K_n^7}{n}\right)^{1/8} + \left(\frac{D_n^2K_n^4}{n^{1-2/q}}\right)^{1/4} + \left(\frac{D_n^{8-8/q}K_n^{15}}{n^{3-4/q}}\right)^{1/14} + \left(\frac{D_n^{3-2/q}K_n^4}{n^{1-1/q}}\right)^{2/7}.$$

There exists a constant C, depending only on r, q, ϱ, c_0, C_0, such that with probability at least $1 - C\varrho_n$,

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}_{|\mathcal{D}'_n}(\mathbb{M}_n^\# \leqslant t) - \mathbb{P}(\widetilde{\mathbb{M}}_n \leqslant t)\right| \leqslant C\varrho_n.$$

Proof. The proof can be found in Section C.4 of the Supplement Material [52]. The key steps are to show that given $\mathcal{D}'_n$, the conditional covariance functions $\widehat{\gamma}_A(\cdot,\cdot)$ and $\widehat{\gamma}_B(\cdot,\cdot)$, for $\mathbb{U}_{n,A}^\#$ and $\mathbb{U}_{n,B}^\#$ in (10), are good estimators for $\gamma_A(\cdot,\cdot)$ and $\gamma_B(\cdot,\cdot)$. ■

6.3. Related work

U-processes offer a general framework for many statistical applications such as testing for qualitative features (e.g., monotonicity, curvature) of regression functions [24, 7, 1], testing for stochastic monotonicity [41], nonparametric density estimation [48, 23, 25], and establishing limiting distributions of M-estimators [2, 51]. When indexing function classes are fixed, it is known that the Uniform Central Limit Theorems (UCLTs) [49, 2, 17, 6], as well as limit theorems for bootstrap [3, 56], hold for U-processes under metric (or bracketing) entropy conditions. These references [49, 2, 17, 6, 3, 56] cover both non-degenerate and degenerate U-processes where limiting processes of the latter are Gaussian chaoses rather than Gaussian processes. When the UCLTs do not hold for a possibly changing (in n) indexing function class (i.e., the function class cannot be embedded in any fixed Donsker class), [13] develops a general non-asymptotic theory for approximating the suprema of U-processes, extending the earlier work by [15] on empirical processes. Incomplete U-statistics for a fixed dimension are first considered in [4], and the asymptotic distributions are studied in [8, 36]. In high dimensions, non-asymptotic Gaussian approximation and bootstrap results for randomized incomplete U-statistics are established in [12] for a fixed order and in [53] for diverging orders. The current work considers randomized incomplete (local) U-processes with stratification.

Acknowledgements

The authors would like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments that improved the quality of this paper. Y. Song is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC). This research is enabled in part by support provided by Compute Canada (www.computecanada.ca). X. Chen is supported in part by NSF CAREER Award DMS-1752614, UIUC Research Board Award RB18099, and a Simons Fellowship. X. Chen acknowledges that part of this work is carried out at the MIT Institute for Data, System, and Society (IDSS). K. Kato is partially supported by NSF grants DMS-1952306 and DMS-2014636.

Supplementary Material

Supplement to "Stratified incomplete local simplex tests for curvature of nonparametric multiple regression"
This Supplement Material [52] contains additional simulation results, proofs and discussions. The implementation code for our proposed procedures is included.

References

[1] ABREVAYA, J. and JIANG, W. (2005). A nonparametric approach to measuring and testing curvature. *Journal of Business & Economic Statistics* **23** 1–19.
[2] ARCONES, M. A. and GINE, E. (1993). Limit theorems for U-processes. *The Annals of Probability* 1494–1542.
[3] ARCONES, M. A. and GINÉ, E. (1994). U-processes indexed by Vapnik-Červonenkis classes of functions with applications to asymptotics and bootstrap of U-statistics with estimated parameters. *Stochastic Processes and their Applications* **52** 17–38.
[4] BLOM, G. (1976). Some properties of incomplete U-statistics. *Biometrika* **63** 573-580.

[5] BORENSTEIN, S. and FARRELL, J. (2007). Do investors forecast fat firms? Evidence from the gold-mining industry. *RAND Journal of Economics* **38** 626-647.

[6] BOROVSKIKH, Y. V. (1996). *U-Statistics in Banach Spaces*. V.S.P. Intl Science.

[7] BOWMAN, A., JONES, M. and GIJBELS, I. (1998). Testing monotonicity of regression. *Journal of computational and Graphical Statistics* **7** 489–500.

[8] BROWN, B. M. and KILDEA, D. G. (1978). Reduced U-statistics and the Hodges-Lehmann estimator. *Annals of Statistics* **6** 828-835.

[9] CAI, T. T. and LOW, M. G. (2015). A FRAMEWORK FOR ESTIMATION OF CONVEX FUNCTIONS. *Statistica Sinica* **25** 423-456.

[10] CHATTERJEE, S. (2014). A new perspective on least squares under convex constraint. *Ann. Statist.* **42** 2340–2381.

[11] CHATTERJEE, S. (2016). An improved global risk bound in concave regression. *Electron. J. Statist.* **10** 1608–1629.

[12] CHEN, X. and KATO, K. (2019). Randomized incomplete U-statistics in high dimensions. *The Annals of Statistics* **47** 3127-3156.

[13] CHEN, X. and KATO, K. (2020). Jackknife multiplier bootstrap: finite sample approximations to the U-process supremum with applications. *Probability Theory and Related Fields* **176** 1097-1163.

[14] CHEN, Y. and WELLNER, J. A. (2016). On convex least squares estimation when the truth is linear. *Electron. J. Statist.* **10** 171–209.

[15] CHERNOZHUKOV, V., CHETVERIKOV, D., KATO, K. et al. (2014). Gaussian approximation of suprema of empirical processes. *The Annals of Statistics* **42** 1564–1597.

[16] CHETVERIKOV, D., SANTOS, A. and SHAIKH, A. M. (2018). The econometrics of shape restrictions. *Annual Review of Economics* **10** 31–63.

[17] DE LA PENA, V. and GINÉ, E. (2012). *Decoupling: from dependence to independence*. Springer Science & Business Media.

[18] DIACK, C. A. and THOMAS-AGNAN, C. (1998). A nonparametric test of the non-convexity of regression. *Journal of Nonparametric Statistics* **9** 335–362.

[19] DÜMBGEN, L. and SPOKOINY, V. G. (2001). Multiscale testing of qualitative hypotheses. *Annals of Statistics* 124–152.

[20] EINMAHL, U., MASON, D. M. et al. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *The Annals of Statistics* **33** 1380–1403.

[21] FAMA, E. F. and FRENCH, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* **33** 3 - 56.

[22] FANG, Z. and SEO, J. (2021). A Projection Framework for Testing Shape Restrictions That Form Convex Cones. *Econometrica, forthcoming (arXiv:1910.07689)*.

[23] FREES, E. W. (1994). Estimating densities of functions of observations. *Journal of American Statistical Association* **89** 517-525.

[24] GHOSAL, S., SEN, A. and VAN DER VAART, A. W. (2000). Testing monotonicity of regression. *Ann. Statist.* **28** 1054–1082.

[25] GINÉ, E., MASON, D. M. et al. (2007). On local U-statistic processes and the estimation of densities of functions of several sample variables. *The Annals of Statistics* **35** 1105–1145.

[26] GINÉ, E. and NICKL, R. (2016). *Mathematical foundations of infinite-dimensional statistical models* **40**. Cambridge University Press.

[27] GROENEBOOM, P., JONGBLOED, G. and WELLNER, J. A. (2001). Estimation of a Convex Function: Characterizations and Asymptotic Theory. *Ann. Statist.* **29** 1653–1698.

[28] GUNTUBOYINA, A. and SEN, B. (2015). Global risk bounds and adaptation in univariate convex regression. *Probab. Theory Related Fields* **163** 379-411.

[29] GUNTUBOYINA, A. and SEN, B. (2018). Nonparametric shape-restricted regression. *Statistical Science* **33** 568–594.
[30] GUPTA, P. and BHATTACHARJEE, G. (1984). An efficient algorithm for random sampling without replacement. In *International Conference on Foundations of Software Technology and Theoretical Computer Science* 435–442. Springer.
[31] HALL, P. (1991). On convergence rates of suprema. *Probability Theory and Related Fields* **89** 447–455.
[32] HAN, Q. and WELLNER, J. A. (2016). Multivariate convex regression: global risk bounds and adaptation. *arXiv preprint arXiv:1601.06844*.
[33] HANNAH, L. A. and DUNSON, D. B. (2013). Multivariate convex regression with adaptive partitioning. *The Journal of Machine Learning Research* **14** 3261–3294.
[34] HANSON, D. L. and PLEDGER, G. (1976). Consistency in Concave Regression. *Ann. Statist.* **4** 1038–1050.
[35] HILDRETH, C. (1954). Point estimates of ordinates of concave functions. *Journal of the American Statistical Association* **49** 598–619.
[36] JANSON, S. (1984). The asymptotic distributions of incomplete U-statistics. *Z, Wahrscheinlichkeitstheorie verw. Gebiete* **66** 495-505.
[37] JUDITSKY, A. and NEMIROVSKI, A. (2002). On nonparametric tests of positivity/monotonicity/convexity. *Ann. Statist.* **30** 498–527.
[38] KOMAROVA, T. and HIDALGO, J. (2019). Testing nonparametric shape restrictions. *arXiv preprint arXiv:1909.01675*.
[39] KUOSMANEN, T. (2008). Representation theorem for convex nonparametric least squares. *The Econometrics Journal* **11** 308–325.
[40] KUR, G., DAGAN, Y. and RAKHLIN, A. (2019). Optimality of Maximum Likelihood for Log-Concave Density Estimation and Bounded Convex Regression. *arXiv preprint arXiv:1903.05315*.
[41] LEE, S., LINTON, O. and WHANG, Y.-J. (2009). Testing for stochastic monotonicity. *Econometrica* **77** 585–602.
[42] LEHMANN, E. L. and ROMANO, J. P. (2006). *Testing statistical hypotheses*. Springer Science & Business Media.
[43] LIM, E. and GLYNN, P. W. (2012). Consistency of multidimensional convex regression. *Operations Research* **60** 196–208.
[44] MAMMEN, E. (1991). Nonparametric Regression Under Qualitative Smoothness Assumptions. *Ann. Statist.* **19** 741–759.
[45] MATZKIN, R. (1991). Semiparametric estimation of monotone and concave utility functions for polychotomous choice models. *Econometrica* **59** 1315-1327.
[46] MAZUMDER, R., CHOUDHURY, A., IYENGAR, G. and SEN, B. (2019). A Computational Framework for Multivariate Convex Regression and Its Variants. *Journal of the American Statistical Association* **114** 318-331.
[47] MURPHY, K. M. and WELCH, F. (1990). Empirical Age-Earnings Profiles. *Journal of Labor Economics* **8** 202-229.
[48] NOLAN, D. and POLLARD, D. (1987). U-processes: rates of convergence. *The Annals of Statistics* 780–799.
[49] NOLAN, D. and POLLARD, D. (1988). Functional limit theorems for U-processes. *The Annals of Probability* **16** 1291–1298.
[50] SEIJO, E., SEN, B. et al. (2011). Nonparametric least squares estimation of a multivariate convex regression function. *The Annals of Statistics* **39** 1633–1657.
[51] SHERMAN, R. P. (1994). Maximal inequalities for degenerate U-processes with applications to optimization estimators. *The Annals of Statistics* 439–459.

[52] SONG, Y., CHEN, X. and KATO, K. Supplement to "Stratified incomplete local simplex tests for curvature of nonparametric multiple regression". *DOI: 10.1214/[provided by typesetter].*

[53] SONG, Y., CHEN, X. and KATO, K. (2019). Approximating high-dimensional infinite-order U-statistics: statistical and computational guarantees. *Electronic Journal of Statistics* **13** 4794-4848.

[54] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes With Applications to Statistics*. Springer-Verlag New York.

[55] WANG, J. C. and MEYER, M. C. (2011). Testing the monotonicity or convexity of a function using regression splines. *Canadian Journal of Statistics* **39** 89–107.

[56] ZHANG, D. (2001). Bayesian bootstraps for U-processes, hypothesis tests and convergence of Dirichlet U-processes. *Statistica Sinica* 463–478.