Continuity of the Distribution Function of the arg max of a Gaussian Process*

Matias D. Cattaneo† Gregory Fletcher Cox‡ Michael Jansson§
Kenichi Nagasawa¶

January 22, 2025

Abstract

An increasingly important class of estimators has members whose asymptotic distribution is non-Gaussian, yet characterizable as the arg max of a Gaussian process. This paper presents high-level sufficient conditions under which such asymptotic distributions admit a continuous distribution function. The plausibility of the sufficient conditions is demonstrated by verifying them in three prominent examples, namely maximum score estimation, empirical risk minimization, and threshold regression estimation. In turn, the continuity result buttresses several recently proposed inference procedures whose validity seems to require a result of the kind established herein. A notable feature of the high-level assumptions is that one of them is designed to enable us to employ the celebrated Cameron-Martin theorem. In a leading special case, the assumption in question is demonstrably weak and appears to be close to minimal.

Keywords: Cameron-Martin Theorem, Cube Root Asymptotics, Gaussian Processes, Reproducing Kernel Hilbert Space.

*We are grateful to Boris Hanin, Jason Klusowski, Ulrich Müller, Mathieu Rosenbaum, Misha Shkolnikov, Ronnie Sircar, and Will Underwood for insightful discussions and comments. Cattaneo gratefully acknowledges financial support from the National Science Foundation through grants SES-1947805, DMS-2210561, and SES-2241575, and Jansson gratefully acknowledges financial support from the National Science Foundation through grant SES-1947662.

†Department of Operations Research and Financial Engineering, Princeton University.
‡Department of Economics, National University of Singapore.
§Department of Economics, University of California at Berkeley.
¶Department of Economics, University of Warwick.

1 Introduction

An increasingly important class of estimators has members whose asymptotic distribution is non-Gaussian, yet characterizable as the arg max of a Gaussian process. To fix ideas, letting $\boldsymbol{\theta}_0 \in \mathbb{R}^d$ denote a parameter (vector) of interest, the estimators $\hat{\boldsymbol{\theta}}_n$ in question satisfy

$$r_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \rightsquigarrow \operatorname*{arg\,max}_{\mathbf{s}\in\mathbb{R}^d} \mathcal{G}(\mathbf{s}), \tag{1}$$

where n is the sample size, r_n is a rate of convergence, $\rightsquigarrow$ denotes weak convergence (as $n \to \infty$), and $\mathcal{G}$ is a Gaussian process admitting a unique maximizer (over $\mathbb{R}^d$) whose distribution is non-Gaussian. The seminal work of Kim and Pollard (1990) was concerned with (cube root asymptotic) cases where $r_n = \sqrt[3]{n}$ and the mean function of $\mathcal{G}$ is a quadratic form, but subsequent work (e.g., Hansen, 2000; Lai and Lee, 2005; Lee, Liao, Seo, and Shin, 2021; Lee and Pun, 2006; Lee and Yang, 2020; Westling and Carone, 2020; Yu and Fan, 2021) has documented the relevance of allowing for the extra flexibility afforded by the more general formulation in (1).

Letting μ and $\mathcal{C}$ denote the mean function and covariance kernel of $\mathcal{G}$ and defining

$$F_{\hat{\mathbf{s}}}(\mathbf{t}) = \mathbb{P}[\hat{\mathbf{s}} \leq \mathbf{t}], \qquad \hat{\mathbf{s}} = \operatorname*{arg\,max}_{\mathbf{s}\in\mathbb{R}^d} \mathcal{G}(\mathbf{s}), \tag{2}$$

our goal in this paper is to give conditions on μ and $\mathcal{C}$ that imply continuity of $F_{\hat{\mathbf{s}}}$. Continuity of $F_{\hat{\mathbf{s}}}$ is useful when the goal is to use $\hat{\boldsymbol{\theta}}_n$ to construct confidence regions. For instance, van der Vaart (1998, Lemma 23.3) assumes continuity when establishing validity of bootstrap-based confidence intervals; see also Politis, Romano, and Wolf (1999, Section 1.2). Moreover, and relatedly, it follows from Polya's theorem that if (1) holds and if $F_{\hat{\mathbf{s}}}$ is continuous, then the distribution functions of $r_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ converges to $F_{\hat{\mathbf{s}}}$ not only in the bounded Lipschitz metric (or any other metric metrizing weak convergence), but also in the Kolmogorov metric; that is, we have a result of the form

$$\sup_{\mathbf{t}\in\mathbb{R}^d} \left| \mathbb{P}\left[r_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \leq \mathbf{t} \right] - F_{\hat{\mathbf{s}}}(\mathbf{t}) \right| \to 0. \tag{3}$$

When μ is a quadratic form and $\mathcal{C}$ is a bilinear form, the distribution of $\hat{\mathbf{s}}$ is Gaussian. More generally, under mild conditions on μ the distribution of $\hat{\mathbf{s}}$ is that of a transformation of a Gaussian vector when $\mathcal{C}$ is a bilinear form, implying in particular that the properties of $F_{\hat{\mathbf{s}}}$ can be deduced by means of a change of variables argument. Two other special cases where a complete characterization of $F_{\hat{\mathbf{s}}}$ is available are when $d = 1$, $\mathcal{C}$ is the covariance kernel of a two-sided Brownian motion, and μ is proportional to either the absolute value function or the square function. In both cases, the distribution of $\hat{\mathbf{s}}$ is that of a scalar multiple of a random variable with a well-known continuous distribution. Somewhat more generally, Cattaneo, Jansson, and Nagasawa (2024, Lemma A.2) gave conditions on μ under which $F_{\hat{\mathbf{s}}}$ is continuous when $d = 1$ and $\mathcal{C}$ is the covariance kernel of a two-sided Brownian motion. On the other hand, little (if anything) appears to be known about the properties of $F_{\hat{\mathbf{s}}}$ when $d > 1$ and $\mathcal{C}$ is not bilinear.

In this paper we close this gap by presenting sufficient conditions for continuity of $F_{\hat{\mathbf{s}}}$ that do not require $d = 1$ and are applicable (only) when $\mathcal{C}$ is not bilinear. The most notable of these conditions involves both μ and $\mathcal{C}$ and requires that for every $N \in \mathbb{N}$, restriction of $\mathcal{G}$ to $[-N, N]^d$ has a mean function that belongs to the reproducing kernel Hilbert space (RKHS) of its covariance kernel. By the celebrated Cameron-Martin theorem, if a Gaussian process has a mean belonging to the RKHS of its covariance kernel, then its induced probability measure and the probability measure induced by its centered version are mutually absolutely continuous. The proof of our main result uses this fact and an assumed shift equivariance property of the covariance kernel to deduce continuity of $F_{\hat{\mathbf{s}}}$.

The usefulness of our main result is illustrated by applying it to three important estimation examples, namely maximum score estimation, empirical risk minimization, and threshold regression estimation. Each example involves an estimator satisfying (1) with d (possibly) greater than one and a covariance kernel that is not bilinear. Although distinct in important ways, the examples enjoy the common feature that continuity of $F_{\hat{\mathbf{s}}}$ can be shown by verifying the conditions of our main result. In particular, the condition that the mean function belongs to the RKHS of the covariance kernel can be verified by following a general strategy outlined in Lifshits (1995).

Although the main motivation for our paper is statistical, namely to provide easy-to-verify sufficient conditions for a continuity property needed to justify large-sample inference based on distributional approximations of the form (1), our paper also offers a contribution to the probability literature on the distributional properties of the arg max of a Gaussian process. That problem is substantially different from the well-studied problem of understanding the distributional properties of the maximum itself, where the $d = 1$ case is mostly settled (Lifshits, 1995, and references therein) and the multidimensional case is fairly well understood (Azaïs and Wschebor, 2005, and references therein). Indeed, as noted by Samorodnitsky and Shen (2013b, p. 3494), "very little is known about the random location of the supremum" of a Gaussian process.

In addition to providing results of the form (1), Kim and Pollard (1990) gave sufficient conditions for almost sure uniqueness of the arg max of a multivariate Gaussian process; for more recent contributions along these lines, see Cox (2020) and López and Pimentel (2018) and references therein. In the case of $d = 1$, moments (and sometimes continuity) of the distribution of the arg max of a stochastic process have been studied in special cases by Groeneboom et al. (2015), Pitman and Tang (2018), Samorodnitsky and Shen (2012, 2013a,b), Shen (2018), and references therein, and more recently Cattaneo et al. (2024, Lemma A.2) gave fairly general conditions guaranteeing its continuity. For $d > 1$, there is scarce literature investigating the properties of the distribution of the location of the maximizer of a Gaussian process. We are only aware of Azaïs and Chassan (2020), which shows that the distribution admits a density under the assumption that the sample paths are twice differentiable. Our paper appears to be the first to provide sufficient conditions for continuity of the distribution of the arg max of a multidimensional Gaussian process without assuming differentiability of its sample paths. Furthermore, as we discuss in Section 4, our sufficient conditions are demonstrably weak and appear to be close to minimal in a leading special case.

The remainder of the paper proceeds as follows. Section 2 introduces the three above-mentioned estimation examples. Our main result is presented in Section 3, while Section 4 outlines a general strategy for verifying the conditions of the main result and demonstrates how to apply it in the examples. Finally, Section 5 compares our results with the known results alluded to in the third paragraph of this section.

2 Motivating Examples

The class of estimators satisfying (1) is rich, containing important examples in econometrics, statistics, and many other data science disciplines. To further motivate our work, this section presents three representative examples.

2.1 Maximum Score

Suppose $\{(y_i, w_i, \mathbf{x}_i')'\}_{i=1}^n$ is a random sample from the distribution of a vector $(y, w, \mathbf{x}')'$ generated by the semiparametric binary response model

$$y = \mathbb{1}\{w + \mathbf{x}'\boldsymbol{\theta}_0 \geq u\}, \qquad \text{Median}(u|w, \mathbf{x}) = 0,$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function, $w, u \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^d$ are random variables, and $\boldsymbol{\theta}_0 \in \Theta \subseteq \mathbb{R}^d$ is the parameter of interest. Manski (1975) introduced the celebrated maximum score estimator of $\boldsymbol{\theta}_0$, which is any maximizer $\hat{\boldsymbol{\theta}}_n$ of

$$\sum_{i=1}^n (2y_i - 1)\mathbb{1}\{w_i + \mathbf{x}_i'\boldsymbol{\theta} \geq 0\}$$

with respect to $\boldsymbol{\theta} \in \Theta$. Using the methods of Kim and Pollard (1990), Abrevaya and Huang (2005) gave regularity conditions under which (1) holds with $r_n = \sqrt[3]{n}$ and $\mathcal{G}$ being a Gaussian process whose mean function and covariance kernel take the form

$$\mu(\mathbf{s}) = -\mathbf{s}'\mathbb{E}\left[f_{u|w,\mathbf{x}}(0| - \mathbf{x}'\boldsymbol{\theta}_0, \mathbf{x}) f_{w|\mathbf{x}}(-\mathbf{x}'\boldsymbol{\theta}_0|\mathbf{x})\mathbf{x}\mathbf{x}'\right]\mathbf{s}$$

and

$$\mathcal{C}(\mathbf{s}, \mathbf{t}) = \mathbb{E}\left[f_{w|\mathbf{x}}(-\mathbf{x}'\boldsymbol{\theta}_0|\mathbf{x})\mathcal{C}_{\texttt{BM}}(\mathbf{x}'\mathbf{s}, \mathbf{x}'\mathbf{t})\right],$$

respectively, where $f_{u|w,\mathbf{x}}$ and $f_{w|\mathbf{x}}$ denote conditional (Lebesgue) densities, and where $\mathcal{C}_{\texttt{BM}}$ is the covariance kernel of a two-sided standard Brownian motion; that is,

$$\mathcal{C}_{\texttt{BM}}(s, t) = \min\{|s|, |t|\}\mathbb{1}\{\text{sgn}(s) = \text{sgn}(t)\},$$

with $\text{sgn}(\cdot)$ denoting the sign function.

When $d = 1$, because μ is quadratic and $\mathcal{C}$ is a scalar multiple of $\mathcal{C}_{\texttt{BM}}$, it follows from van der Vaart

and Wellner (1996, Exercise 3.2.5) that the distribution of $\hat{\mathbf{s}}$ is that of a scalar multiple of a random variable with a well-known continuous distribution, namely the Chernoff (1964) distribution. For $d > 1$, on the other hand, it would appear to be an open question whether $F_{\hat{\mathbf{s}}}$ is continuous. We provide an affirmative answer to that question below, hereby buttressing the inference procedures for maximum score estimators proposed by Cattaneo, Jansson, and Nagasawa (2020), Delgado, Rodriguez-Poo, and Wolf (2001), Hong and Li (2020), Jun, Pinkse, and Wan (2015), Lee and Yang (2020), and Patra, Seijo, and Sen (2018) because estimated or bootstrapped quantiles of the asymptotic distribution can be shown to be consistent.

2.2 Empirical Risk Minimization

Mohammadi and van de Geer (2005) considered the classification problem of estimating the minimizer $\boldsymbol{\theta}_0 \in \Theta \subseteq \mathbb{R}^d$ of the classification error $\mathbb{P}[y \neq h_{\boldsymbol{\theta}}(x)]$ with respect to $\boldsymbol{\theta} \in \Theta$, where $y \in \{-1, 1\}$ is a binary outcome, $x \in \mathcal{X} \subseteq \mathbb{R}$ is a scalar feature, and $\{h_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ is a collection of classifiers. Given a random sample $\{(y_i, x_i)\}_{i=1}^n$ from the distribution of (y, x), an empirical risk minimizer is a minimizer $\hat{\boldsymbol{\theta}}_n$ of

$$\sum_{i=1}^{n} \mathbb{1}\{y_i \neq h_{\boldsymbol{\theta}}(x_i)\}.$$

Setting $\mathcal{X} = [0, 1]$ and specializing to the case where the classifiers are of the form

$$h_{\boldsymbol{\theta}}(x) = \sum_{\ell=1}^{d+1} (-1)^{\ell} \mathbb{1}\{\theta_{\ell-1} \leq x < \theta_{\ell}\}$$

for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)' \in \Theta = \{\boldsymbol{\theta} \in [0, 1]^d : 0 = \theta_0 \leq \theta_1 \leq \cdots \leq \theta_d \leq \theta_{d+1} = 1\}$, Mohammadi and van de Geer (2005, Theorem 1) gave conditions under which (1) holds with $r_n = \sqrt[3]{n}$ and $\mathcal{G}$ being a Gaussian process whose mean function and covariance kernel take the form

$$\mu(\mathbf{s}) = \sum_{\ell=1}^{d} \mu_{\ell}(s_{\ell}), \qquad \mu_{\ell}(s_{\ell}) = (-1)^{\ell} p(\theta_{0,\ell}) f(\theta_{0,\ell}) s_{\ell}^2, \tag{4}$$

and

$$\mathcal{C}(\mathbf{s}, \mathbf{t}) = \sum_{\ell=1}^{d} \mathcal{C}_{\ell}(s_{\ell}, t_{\ell}), \qquad \mathcal{C}_{\ell}(s_{\ell}, t_{\ell}) = f(\theta_{0,\ell}) \mathcal{C}_{\mathtt{BM}}(s_{\ell}, t_{\ell}), \tag{5}$$

respectively, where $\boldsymbol{\theta}_0 = (\theta_{0,1}, \ldots, \theta_{0,d})', \mathbf{s} = (s_1, \ldots, s_d)', \mathbf{t} = (t_1, \ldots, t_d)'$, f is a Lebesgue density of x, $p(x) = \mathrm{d}\mathbb{P}[y = 1|x]/\mathrm{d}x$, and where the assumptions imposed on the model ensure that $(-1)^{\ell} p(\theta_{0,\ell}) f(\theta_{0,\ell}) < 0$ for every $\ell = 1, \ldots, d$.

This example is similar to the maximum score example insofar as when $d = 1$, the distribution of $\hat{\mathbf{s}}$ is that of a scalar multiple of a random variable with a Chernoff distribution. In fact, also when $d > 1$, the elements of $\hat{\mathbf{s}} = (\hat{s}_1, \ldots, \hat{s}_d)'$ are mutually independent, each having a distribution which is that of a scalar multiple of a random variable with a Chernoff distribution. Indeed, letting $\mathcal{G}_1, \ldots, \mathcal{G}_d$ be mutually independent Gaussian processes with mean functions $\mu_1, \ldots, \mu_d$ and covariance kernels

$\mathcal{C}_1, \ldots, \mathcal{C}_d$, respectively, $\mathcal{G}$ admits the representation $\mathcal{G}(\mathbf{s}) = \sum_{\ell=1}^{d} \mathcal{G}_\ell(s_\ell)$, implying in particular that

$$\hat{s}_\ell = \arg\max_{s_\ell \in \mathbb{R}} \mathcal{G}_\ell(s_\ell) \qquad \text{for each } \ell \in \{1, \ldots, d\}.$$

In other words, this example has special structure that can be used to obtain a definitive characterization of the distribution of $\hat{\mathbf{s}}$ without utilizing new tools. It is nevertheless of interest to explore the ease with which the technology developed in this paper can be deployed to establish continuity of $F_{\hat{\mathbf{s}}}$ in this example. In particular, it is of interest to explore whether the (effective) "dimension reduction" permitted by this example can be leveraged when verifying the conditions of Theorem 1 below.

2.3 Threshold Regression

Consider the threshold regression model

$$y = \mathbf{x}'\boldsymbol{\beta}_0 + \mathbf{x}'\boldsymbol{\delta}_n \mathbb{1}\{q > \mathbf{w}'\boldsymbol{\theta}_0\} + u,$$

where $y \in \mathbb{R}$ is a dependent variable, $\mathbf{x} \in \mathbb{R}^k$ is a (possibly) vector-valued regressor, q is a threshold variable, $\mathbf{w} \in \mathbb{R}^d$ is a (possibly) vector-valued factor governing the threshold cutoff, and where $\boldsymbol{\delta}_n$ is a "threshold effect" whose magnitude vanishes with n. The present model (as well as distinct generalizations thereof) has been studied by Lee, Liao, Seo, and Shin (2021) and Yu and Fan (2021), and differs from the model considered in the seminal work of Hansen (2000) by allowing the factor $\mathbf{w}$ to be non-constant.

Given a random sample $\{(y_i, \mathbf{x}_i', q_i, \mathbf{w}_i')'\}_{i=1}^n$ from the distribution of $(y, \mathbf{x}', q, \mathbf{w}')'$, a least squares estimator $(\hat{\boldsymbol{\beta}}_n', \hat{\boldsymbol{\delta}}_n', \hat{\boldsymbol{\theta}}_n')'$ of $(\boldsymbol{\beta}_0', \boldsymbol{\delta}_n', \boldsymbol{\theta}_0')'$ is a minimizer of

$$\sum_{i=1}^{n} \left(y_i - \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{x}_i'\boldsymbol{\delta}\mathbb{1}\{q_i > \mathbf{w}_i'\boldsymbol{\theta}\}\right)^2$$

over $(\boldsymbol{\beta}', \boldsymbol{\delta}', \boldsymbol{\theta}')' \in \mathbb{R}^{2k+d}$. Assuming $\|\boldsymbol{\delta}_n\| \to 0$, $n\|\boldsymbol{\delta}_n\|^2 \to \infty$, and $\bar{\boldsymbol{\delta}}_n = \boldsymbol{\delta}_n/\|\boldsymbol{\delta}_n\| \to \bar{\boldsymbol{\delta}}$ (for some $\bar{\boldsymbol{\delta}} \in \mathbb{S}^{d-1} = \{\bar{\boldsymbol{\delta}} \in \mathbb{R}^d : \|\bar{\boldsymbol{\delta}}\| = 1\}$), Yu and Fan (2021) gave conditions under which (1) holds with $r_n = n\|\boldsymbol{\delta}_n\|^2$ and $\mathcal{G} = \mathcal{G}(\cdot; \bar{\boldsymbol{\delta}})$ being a Gaussian process whose mean function and covariance kernel take the form

$$\mu(\mathbf{s}) = \mu(\mathbf{s}; \bar{\boldsymbol{\delta}}) = -\frac{1}{2}\bar{\boldsymbol{\delta}}'\mathbb{E}\left[|\mathbf{w}'\mathbf{s}| f_{q|\mathbf{w}}(\mathbf{w}'\boldsymbol{\theta}_0|\mathbf{w})\mathbb{E}_{\cdot|q,\mathbf{w}}[\mathbf{x}\mathbf{x}'|\mathbf{w}'\boldsymbol{\theta}_0, \mathbf{w}]\right]\bar{\boldsymbol{\delta}}$$

and

$$\mathcal{C}(\mathbf{s}, \mathbf{t}) = \mathcal{C}(\mathbf{s}, \mathbf{t}; \bar{\boldsymbol{\delta}}) = \bar{\boldsymbol{\delta}}'\mathbb{E}\left[f_{q|\mathbf{w}}(\mathbf{w}'\boldsymbol{\theta}_0|\mathbf{w})\mathbb{E}_{\cdot|q,\mathbf{w}}[\mathbf{x}\mathbf{x}'u^2|\mathbf{w}'\boldsymbol{\theta}_0, \mathbf{w}]\mathcal{C}_{\mathtt{BM}}(\mathbf{w}'\mathbf{s}, \mathbf{w}'\mathbf{t})\right]\bar{\boldsymbol{\delta}},$$

respectively, where $\mathbb{E}_{\cdot|q,\mathbf{w}}$ denotes conditional expectation.

When $d = 1$, the distribution of $\hat{\mathbf{s}}$ is that of a scalar multiple of a random variable with a (known) continuous distribution (e.g., Section 3.2 of Hansen, 2000). For $d > 1$, on the other hand,

it would appear to be an open question whether $F_{\hat{\mathbf{s}}}$ is continuous. We provide an affirmative answer to that question below.

Remark. As a by-product, the continuity property of the limiting distribution function allows us conclude that if $\|\boldsymbol{\delta}_n\| \to 0$ and if $n\|\boldsymbol{\delta}_n\|^2 \to \infty$, then

$$\sup_{t\in\mathbb{R}^d}\left|\mathbb{P}\left[r_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \le \mathbf{t}\right] - \mathbb{P}\left[\operatorname*{arg\,max}_{\mathbf{s}\in\mathbb{R}^d} \mathcal{G}(\mathbf{s}; \bar{\boldsymbol{\delta}}_n) \le \mathbf{t}\right]\right| \to 0$$

whether or not $\bar{\boldsymbol{\delta}}_n$ is convergent in $\mathbb{S}^{d-1}$.

3 Main Result

As before, let $\mathcal{G}$ be a Gaussian process on $\mathbb{R}^d$ with mean function μ and covariance kernel $\mathcal{C}$. Also, for any $N \in \mathbb{N}$, let $\mathcal{G}_N$ be the restriction of $\mathcal{G}$ to $[-N, N]^d$, let μ_N and $\mathcal{C}_N$ be the mean function and covariance kernel of $\mathcal{G}_N$, and let $\mathscr{H}_N$ be the RKHS of $\mathcal{C}_N$ (as defined in Section 2.6.1 of Giné and Nickl, 2016).

The following high-level assumption holds in each of the examples of Section 2.

Assumption 1.

(i) With probability one, $\mathcal{G}$ has continuous sample paths and admits a maximizer over $\mathbb{R}^d$.

(ii) For any $\mathbf{s}, \mathbf{t} \in \mathbb{R}^d$ and any $\mathbf{h} \in \mathbb{R}^d\backslash\{\mathbf{0}\}$, $\mathcal{C}(\mathbf{h}, \mathbf{h}) > 0$ and

$$\mathcal{C}(\mathbf{h}+\mathbf{s}, \mathbf{h}+\mathbf{t}) - \mathcal{C}(\mathbf{h}+\mathbf{s}, \mathbf{h}) - \mathcal{C}(\mathbf{h}, \mathbf{h}+\mathbf{t}) + \mathcal{C}(\mathbf{h}, \mathbf{h}) = \mathcal{C}(\mathbf{s}, \mathbf{t}).$$

(iii) For any $N \in \mathbb{N}$, $\mu_N \in \mathscr{H}_N$.

Part *(i)* guarantees existence (with probability one) of a maximizer of $\mathcal{G}$ over $\mathbb{R}^d$, and over any compact set $S \subset \mathbb{R}^d$. By Kim and Pollard (1990, Lemma 2.6), these maximizers are unique provided $\mathbb{V}[\mathcal{G}(\mathbf{s}) - \mathcal{G}(\mathbf{t})] \neq 0$ for every $\mathbf{s} \neq \mathbf{t}$. Part *(ii)* gives sufficient conditions for this non-degeneracy condition to hold and furthermore ensures that the centered process $\mathcal{G}^\mu = \mathcal{G} - \mu$ is shift equivariant in the sense that the law of the process $\mathcal{G}^\mu(\mathbf{h} + \cdot) - \mathcal{G}^\mu(\mathbf{h})$ if the same for every $\mathbf{h} \in \mathbb{R}^d$. By the Cameron-Martin theorem (e.g., Theorem 2.6.13 of Giné and Nickl, 2016), part *(iii)* ensures that the probability measures associated with $\mathcal{G}_N$ and $\mathcal{G}_N^\mu = \mathcal{G}_N - \mu_N$ are mutually absolutely continuous for any $N \in \mathbb{N}$, a property that plays a key role in the proof of the following result.

Theorem 1. *Under Assumption 1, $F_{\hat{\mathbf{s}}}$ in (2) is continuous.*

3.1 Proof of Theorem 1

For any compact set $S \subset \mathbb{R}^d$, parts *(i)* and *(ii)* of Assumption 1 guarantee existence (with probability one) of a unique maximizer of $\mathcal{G}^\mu$ over S. This observation will be used repeatedly without

further mention. The proof further utilizes the following implication of Assumption 1 *(i)–(ii)*.

Lemma 1. *Suppose Assumption 1 (i)–(ii) holds. For any* $\mathbf{h} \in \mathbb{R}^d$*, any measurable set* $T \subseteq \mathbb{R}^d$*, and any compact set* $S \subset \mathbb{R}^d$*,*

$$\mathbb{P}\left[\arg\max_{\mathbf{s}\in S} \mathcal{G}^\mu(\mathbf{s}) \in T\right] = \mathbb{P}\left[\arg\max_{\mathbf{s}\in S+\mathbf{h}} \mathcal{G}^\mu(\mathbf{s}) \in T + \mathbf{h}\right].$$

Proof of Lemma 1. By change of variables and shifting by a constant,

$$\arg\max_{\mathbf{s}\in S+\mathbf{h}} \mathcal{G}^\mu(\mathbf{s}) = \arg\max_{\mathbf{s}\in S} \mathcal{G}^\mu(\mathbf{h}+\mathbf{s}) + \mathbf{h} = \arg\max_{\mathbf{s}\in S}\{\mathcal{G}^\mu(\mathbf{h}+\mathbf{s}) - \mathcal{G}^\mu(\mathbf{h})\} + \mathbf{h}.$$

The desired conclusion follows from shift equivariance of $\mathcal{G}^\mu$. □

The (joint) distribution function of $\hat{\mathbf{s}} = (\hat{s}_1, \ldots, \hat{s}_d)'$ is continuous if and only if each of its marginal distribution functions is continuous. Fixing $\ell \in \{1, \ldots, d\}$ and $t \in \mathbb{R}$, the proof can therefore be completed by showing that

$$\mathbb{P}\left[\hat{s}_\ell = t\right] = 0. \tag{6}$$

Defining $N_t = \lceil |t| \rceil + 1$, letting $\mathbf{e}_\ell$ denote the ℓth standard basis vector of $\mathbb{R}^d$, and noting that

$$\{\hat{s}_\ell = t\} = \left\{\mathbf{e}_\ell' \arg\max_{\mathbf{s}\in\mathbb{R}^d} \mathcal{G}(\mathbf{s}) = t\right\} \subseteq \bigcup_{N=N_t}^{\infty} \left\{\mathbf{e}_\ell' \arg\max_{\mathbf{s}\in[-N,N]^d} \mathcal{G}(\mathbf{s}) = t\right\},$$

a sufficient condition for (6) to hold is that

$$\mathbb{P}\left[\mathbf{e}_\ell' \arg\max_{\mathbf{s}\in[-N,N]^d} \mathcal{G}(\mathbf{s}) = t\right] = 0 \qquad \text{for every } N \geq N_t.$$

By the Cameron-Martin theorem, under Assumption 1 *(iii)* the displayed condition is equivalent to

$$\mathbb{P}\left[\mathbf{e}_\ell' \arg\max_{\mathbf{s}\in[-N,N]^d} \mathcal{G}^\mu(\mathbf{s}) = t\right] = 0 \qquad \text{for every } N \geq N_t. \tag{7}$$

Fixing $N \geq N_t$ and $J \geq 2$, let

$$S_j = \left\{(s_1, \ldots, s_d) \in [-N, N]^d : -N_t + \frac{1}{2}\frac{j-1}{J-1} \leq s_\ell \leq N_t + \frac{1}{2}\left(\frac{j-1}{J-1} - 1\right)\right\}$$

for $j \in \{1, \ldots, J\}$. Noting that

$$\begin{aligned}[-N, N]^d \supseteq \bar{S} &= \{(s_1, \ldots, s_d) \in [-N, N]^d : -N_t \leq s_\ell \leq N_t\} = \cup_{j=1}^J S_j \\ &\supset \cap_{j=1}^J S_j = \{(s_1, \ldots, s_d) \in [-N, N]^d : -N_t + 1/2 \leq s_\ell \leq N_t - 1/2\} = \underline{S},\end{aligned}$$

we have

$$\begin{aligned}
\mathbb{P}\left[\mathbf{e}_\ell' \operatorname*{arg\,max}_{\mathbf{s}\in[-N,N]^d} \mathcal{G}^\mu(\mathbf{s}) = t\right] &\le \mathbb{P}\left[\mathbf{e}_\ell' \operatorname*{arg\,max}_{\mathbf{s}\in\bar{S}} \mathcal{G}^\mu(\mathbf{s}) = t\right] \le \mathbb{P}\left[\mathbf{e}_\ell' \operatorname*{arg\,max}_{\mathbf{s}\in S_1} \mathcal{G}^\mu(\mathbf{s}) = t\right] \\
&= \frac{1}{J}\sum_{j=1}^{J} \mathbb{P}\left[\mathbf{e}_\ell' \operatorname*{arg\,max}_{\mathbf{s}\in S_j} \mathcal{G}^\mu(\mathbf{s}) = t + \frac{1}{2}\frac{j-1}{J-1}\right] \\
&\le \frac{1}{J}\sum_{j=1}^{J} \mathbb{P}\left[\mathbf{e}_\ell' \operatorname*{arg\,max}_{\mathbf{s}\in \underline{S}} \mathcal{G}^\mu(\mathbf{s}) = t + \frac{1}{2}\frac{j-1}{J-1}\right] \\
&= \frac{1}{J}\mathbb{P}\left[\mathbf{e}_\ell' \operatorname*{arg\,max}_{\mathbf{s}\in \underline{S}} \mathcal{G}^\mu(\mathbf{s}) \in \left\{t + \frac{1}{2}\frac{j-1}{J-1} : 1 \le j \le J\right\}\right] \\
&\le \frac{1}{J},
\end{aligned}$$

where the first equality uses Lemma 1 and the second equality uses uniqueness of the maximizer of $\mathcal{G}^\mu$ over $\underline{S}$. Since $J \ge 2$ was arbitrary, (7) follows. □

Remark. In the proof, N_t can be replaced by any (natural) number exceeding it and the fractions $\{(j-1)/(J-1) : 1 \le j \le J\}$ appearing in the definition of the sets $\{S_j\}$ (and elsewhere) can be replaced by any J distinct members of the unit interval.

4 Verification of Assumption 1

4.1 Assumption 1*(i)*

The continuity part of Assumption 1*(i)* is mild and usually trivial to verify. Under continuity, a high-level sufficient condition for existence of a maximizer of $\mathcal{G}$ over $\mathbb{R}^d$ is that

$$\mathbb{P}\left[\limsup_{\|\mathbf{s}\|\to\infty} \mathcal{G}(\mathbf{s}) < 0\right] = 1 \tag{8}$$

where we use $\mathbb{P}[\mathcal{G}(\mathbf{0}) = 0] = 1$. In turn, proceeding as in the proof of Kim and Pollard (1990, Lemma 2.5) it can be shown that if the covariance kernel satisfies the (self-similarity) property that for some $H > 0$,

$$\mathcal{C}(\tau\mathbf{s}, \tau\mathbf{t}) = \tau^{2H}\mathcal{C}(\mathbf{s}, \mathbf{t}) \qquad \text{for every } \mathbf{s}, \mathbf{t} \in \mathbb{R}^d, \tau > 0, \tag{9}$$

then (8) is implied by the following mild condition on the mean function:

$$\limsup_{\|\mathbf{s}\|\to\infty} \frac{\mu(\mathbf{s})}{\|\mathbf{s}\|^{H+\epsilon}} < 0 \qquad \text{for some } \epsilon > 0. \tag{10}$$

The conditions (9) and (10) are both fairly primitive. Also, by inspection (9) can be seen to hold with $H = 1/2$ in each of the examples of Section 2. Moreover, setting $H = 1/2$, (10) can be seen

to hold with $\epsilon = 3/2$ in the maximum score and empirical risk minimization examples, and with $\epsilon = 1/2$ in the threshold regression example.

To explain why it is no coincidence that the self-similarity property of $\mathcal{C}$ holds in the examples of Section 2, it may be helpful to note that in each case $\mathcal{G}$ is the weak limit of a process of the form

$$\mathbf{s} \mapsto \sqrt{\frac{r_n}{n}} \sum_{i=1}^{n} [m_n(\mathbf{z}_i, \boldsymbol{\theta}_0 + r_n^{-1}\mathbf{s}) - m_n(\mathbf{z}_i, \boldsymbol{\theta}_0)],$$

where $\{\mathbf{z}_i\}_{i=1}^n$ is a random sample and m_n is some function (possibly depending on n). For instance, in the threshold regression example we have

$$m_n(\mathbf{z}, \boldsymbol{\theta}) = -\frac{1}{2\|\boldsymbol{\delta}_n\|} \left(y - \mathbf{x}'\boldsymbol{\beta}_0 - \mathbf{x}'\boldsymbol{\delta}_n \mathbb{1}\{q > \mathbf{w}'\boldsymbol{\theta}\}\right)^2, \qquad \mathbf{z} = (y, \mathbf{x}', q, \mathbf{w}')'.$$

It therefore stands to reason that $\mathcal{C}$ can be characterized as follows:

$$\mathcal{C}(\mathbf{s}, \mathbf{t}) = \lim_{n\to\infty} r_n \mathbb{E}[\{m_n(\mathbf{z}, \boldsymbol{\theta}_0 + r_n^{-1}\mathbf{s}) - m_n(\mathbf{z}, \boldsymbol{\theta}_0)\}\{m_n(\mathbf{z}, \boldsymbol{\theta}_0 + r_n^{-1}\mathbf{t}) - m_n(\mathbf{z}, \boldsymbol{\theta}_0)\}].$$

To further conclude that (9) holds with $H = 1/2$, it suffices to assume that the preceding display admits the following strengthening: for any $\eta_n > 0$ with $\eta_n = O(r_n^{-1})$,

$$\mathcal{C}(\mathbf{s}, \mathbf{t}) = \lim_{n\to\infty} \frac{\mathbb{E}[\{m_n(\mathbf{z}, \boldsymbol{\theta}_0 + \eta_n\mathbf{s}) - m_n(\mathbf{z}, \boldsymbol{\theta}_0)\}\{m_n(\mathbf{z}, \boldsymbol{\theta}_0 + \eta_n\mathbf{t}) - m_n(\mathbf{z}, \boldsymbol{\theta}_0)\}]}{\eta_n}. \tag{11}$$

A characterization of the form (11) is valid in each of the examples of Section 2.

4.2 Assumption 1*(ii)*

By inspection, the displayed part of Assumption 1*(ii)* holds in each of the examples of Section 2. To explain why this is no coincidence, suppose the following "local uniform" version of the characterization (11) is valid: for any $\eta_n > 0$ with $\eta_n = O(r_n^{-1})$ and any $\boldsymbol{\theta}_n = \boldsymbol{\theta}_0 + O(\eta_n)$,

$$\mathcal{C}(\mathbf{s}, \mathbf{t}) = \lim_{n\to\infty} \frac{\mathbb{E}[\{m_n(\mathbf{z}, \boldsymbol{\theta}_n + \eta_n\mathbf{s}) - m_n(\mathbf{z}, \boldsymbol{\theta}_n)\}\{m_n(\mathbf{z}, \boldsymbol{\theta}_n + \eta_n\mathbf{t}) - m_n(\mathbf{z}, \boldsymbol{\theta}_n)\}]}{\eta_n}. \tag{12}$$

Then validity of the displayed part of Assumption 1*(ii)* follows from the fact that

$$\begin{aligned}
&\mathbb{E}[\{m_n(\mathbf{z}, \boldsymbol{\theta}_0 + \eta_n[\mathbf{h} + \mathbf{s}]) - m_n(\mathbf{z}, \boldsymbol{\theta}_0)\}\{m_n(\mathbf{z}, \boldsymbol{\theta}_0 + \eta_n[\mathbf{h} + \mathbf{t}]) - m_n(\mathbf{z}, \boldsymbol{\theta}_0)\}] \\
&\quad - \mathbb{E}[\{m_n(\mathbf{z}, \boldsymbol{\theta}_0 + \eta_n[\mathbf{h} + \mathbf{s}]) - m_n(\mathbf{z}, \boldsymbol{\theta}_0)\}\{m_n(\mathbf{z}, \boldsymbol{\theta}_0 + \eta_n\mathbf{h}) - m_n(\mathbf{z}, \boldsymbol{\theta}_0)\}] \\
&\quad - \mathbb{E}[\{m_n(\mathbf{z}, \boldsymbol{\theta}_0 + \eta_n\mathbf{h}) - m_n(\mathbf{z}, \boldsymbol{\theta}_0)\}\{m_n(\mathbf{z}, \boldsymbol{\theta}_0 + \eta_n[\mathbf{h} + \mathbf{t}]) - m_n(\mathbf{z}, \boldsymbol{\theta}_0)\}] \\
&\quad + \mathbb{E}[\{m_n(\mathbf{z}, \boldsymbol{\theta}_0 + \eta_n\mathbf{h}) - m_n(\mathbf{z}, \boldsymbol{\theta}_0)\}\{m_n(\mathbf{z}, \boldsymbol{\theta}_0 + \eta_n\mathbf{h}) - m_n(\mathbf{z}, \boldsymbol{\theta}_0)\}] \\
&= \mathbb{E}[\{m_n(\mathbf{z}, \boldsymbol{\theta}_n + \eta_n\mathbf{s}) - m_n(\mathbf{z}, \boldsymbol{\theta}_n)\}\{m_n(\mathbf{z}, \boldsymbol{\theta}_n + \eta_n\mathbf{t}) - m_n(\mathbf{z}, \boldsymbol{\theta}_n)\}]
\end{aligned}$$

for any $\mathbf{s}, \mathbf{t}, \mathbf{h} \in \mathbb{R}^d$, where $\boldsymbol{\theta}_n = \boldsymbol{\theta}_0 + \eta_n \mathbf{h}$. Again, a characterization of the form (12) is valid in each of the examples of Section 2.

4.3 Assumption 1*(iii)*

Evaluating Assumption 1*(iii)* is usually straightforward when $\mathscr{H}_N$ is known. More generally, a viable strategy for verifying Assumption 1*(iii)* can be based on Lifshits (1995, Chapters 6 and 9). This subsection first outlines that strategy, and then demonstrates its usefulness by employing it in each of the examples in Section 2.

4.3.1 General Strategy

For $N \in \mathbb{N}$, suppose $\mathscr{E}_N = \{e_N(\cdot; \mathbf{s}) : \mathbf{s} \in [-N, N]^d\}$ is a model of the covariance kernel $\mathcal{C}_N$ (in the terminology of Chapter 6 of Lifshits, 1995); that is, suppose that for some measure space $(\Omega_N, \mathscr{B}_N, \nu_N)$, $\mathscr{E}_N$ is a collection of elements of $L_2(\Omega_N, \mathscr{B}_N, \nu_N)$ satisfying

$$\mathcal{C}_N(\mathbf{s}, \mathbf{t}) = \int e_N(\boldsymbol{\omega}; \mathbf{s}) e_N(\boldsymbol{\omega}; \mathbf{t}) \mathrm{d}\nu_N(\boldsymbol{\omega}) \qquad \text{for all } \mathbf{s}, \mathbf{t} \in [-N, N]^d. \tag{13}$$

Then, as discussed in Lifshits (1995, Chapter 9), the mean function μ_N belongs to $\mathscr{H}_N$ if (and only if) it admits a function $l_N \in L_2(\Omega_N, \mathscr{B}_N, \nu_N)$ satisfying

$$\mu_N(\mathbf{s}) = \int e_N(\boldsymbol{\omega}; \mathbf{s}) l_N(\boldsymbol{\omega}) \mathrm{d}\nu_N(\boldsymbol{\omega}) \qquad \text{for all } \mathbf{s} \in [-N, N]^d. \tag{14}$$

Lifshits (1995) demonstrates how this strategy can be used to characterize $\mathscr{H}_N$ for several examples of Gaussian processes. There is no general blueprint for defining $\mathscr{E}_N$ and l_N satisfying (13)-(14). We modify the arguments used in Lifshits (1995) to cover our examples.

4.3.2 Example: Maximum Score (Continued)

For $N \in \mathbb{N}$, let $\mathscr{B}_N$ be the Borel σ-algebra on $\Omega_N = \mathbb{R}^{1+d}$ and let $\nu_N = \lambda \times \mathbb{P}_{\mathbf{x}}$, where λ is the Lebesgue measure on $\mathbb{R}$ and $\mathbb{P}_{\mathbf{x}}$ is the probability measure induced by $\mathbf{x}$. A direct calculation shows that (13)-(14) hold with $\boldsymbol{\omega} = (\omega_1, \mathbf{x}')'$,

$$e_N(\boldsymbol{\omega}; \mathbf{s}) = \left[\mathbb{1}\{0 \leq \omega_1 \leq \mathbf{x}'\mathbf{s}\} + \mathbb{1}\{\mathbf{x}'\mathbf{s} \leq \omega_1 < 0\}\right] \sqrt{f_{w|\mathbf{x}}(-\mathbf{x}'\boldsymbol{\theta}_0|\mathbf{x})},$$

and

$$l_N(\boldsymbol{\omega}) = -2\mathbb{1}\{|\omega_1| \leq N\sqrt{d}\|\mathbf{x}\|\}|\omega_1| f_{u|w,\mathbf{x}}(0| - \mathbf{x}'\boldsymbol{\theta}_0, \mathbf{x}) \sqrt{f_{w|\mathbf{x}}(-\mathbf{x}'\boldsymbol{\theta}_0|\mathbf{x})}.$$

4.3.3 Example: Empirical Risk Minimization (Continued)

For $N \in \mathbb{N}$, let $\mathscr{B}_N$ be the Borel σ-algebra on $\Omega_N = \mathbb{R}^d$ and let ν_N be the Lebesgue measure on $\mathbb{R}^d$. Then (13)-(14) hold with $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_d)'$,

$$e_N(\boldsymbol{\omega}; \mathbf{s}) = \sqrt{24 N u_N(\boldsymbol{\omega})} \sum_{\ell=1}^{d} \left(\omega_\ell - \frac{\sqrt[3]{s_\ell}}{2} \right) \left[\mathbb{1}\{0 \leq \omega_\ell \leq \sqrt[3]{s_\ell}\} + \mathbb{1}\{\sqrt[3]{s_\ell} \leq \omega_\ell < 0\} \right] \sqrt{f(\theta_{0,\ell})}$$

and

$$l_N(\boldsymbol{\omega}) = \sqrt{\frac{75 N u_N(\boldsymbol{\omega})}{2}} \sum_{\ell=1}^{d} \omega_\ell^3 |\omega_\ell| (-1)^\ell p(\theta_{0,\ell}) \sqrt{f(\theta_{0,\ell})},$$

where u_N is a Lebesgue density of the uniform distribution on $[-N, N]^d$.

The strategy described in Section 4.3.1 and followed in the previous paragraph is general and is not designed to leverage the special structure highlighted in Section 2.2. Because it stands to reason that the additive separability of $\mathcal{C}_N$ induces an analogous simplification of the associated $\mathscr{H}_N$, it seems natural to ask whether such a "tensorization" (is materialized and) can in turn be exploited when verifying Assumption 1*(iii)*.

The special structure (5) of $\mathcal{C}_N$ implies that $h_N \in \mathscr{H}_N$ if and only if it is of the form

$$h_N(\mathbf{s}) = \sum_{\ell=1}^{d} h_{\ell,N}(s_\ell),$$

with $h_{\ell,N}$ belonging to the RKHS of $\mathcal{C}_{\ell,N}$ for each ℓ. Now, proceeding as in Giné and Nickl (2016, Example 2.6.7) it can be shown that the RKHSs of $\mathcal{C}_{1,N}, \ldots, \mathcal{C}_{d,N}$ all (coincide and) consist of those functions on $[-N, N]$ that are zero-preserving and absolutely continuous with a square integrable (weak) derivative. In particular, we see immediately that $\mu_N \in \mathscr{H}_N$ when μ is given by (4).

4.3.4 Example: Threshold Regression (Continued)

For $N \in \mathbb{N}$, let $\mathscr{B}_N$ be the Borel σ-algebra on $\Omega_N = \mathbb{R}^{1+d}$ and let $\nu_N = \lambda \times \mathbb{P}_{\mathbf{w}}$, where λ is the Lebesgue measure on $\mathbb{R}$ and $\mathbb{P}_{\mathbf{w}}$ is the probability measure induced by $\mathbf{w}$. Then (13)-(14) hold with $\boldsymbol{\omega} = (\omega_1, \mathbf{w}')'$,

$$e_N(\boldsymbol{\omega}; \mathbf{s}) = \left[\mathbb{1}\{\mathbf{w}'\mathbf{s} \leq \omega_1 < 0\} + \mathbb{1}\{0 \leq \omega_1 \leq \mathbf{w}'\mathbf{s}\} \right] \sqrt{\mathbb{E}_{\cdot|q,\mathbf{w}}[(\bar{\boldsymbol{\delta}}'\mathbf{x}u)^2 | \mathbf{w}'\boldsymbol{\theta}_0, \mathbf{w}] f_{q|\mathbf{w}}(\mathbf{w}'\boldsymbol{\theta}_0 | \mathbf{w})}$$

and

$$l_N(\boldsymbol{\omega}) = -\frac{1}{2} \mathbb{1}\{|\omega_1| \leq N\sqrt{d}\|\mathbf{w}\|\} \mathbb{E}_{\cdot|q,\mathbf{w}}[(\bar{\boldsymbol{\delta}}'\mathbf{x})^2 | \mathbf{w}'\boldsymbol{\theta}_0, \mathbf{w}] \sqrt{\frac{f_{q|\mathbf{w}}(\mathbf{w}'\boldsymbol{\theta}_0 | \mathbf{w})}{\mathbb{E}_{\cdot|q,\mathbf{w}}[(\bar{\boldsymbol{\delta}}'\mathbf{x}u)^2 | \mathbf{w}'\boldsymbol{\theta}_0, \mathbf{w}]}}.$$

5 Discussion of Assumption 1 *(iii)*

Interpreted as a condition on the mean function μ, the strength of Assumption 1 *(iii)* is inversely related to the richness of the RKHSs $\mathscr{H}_N$ generated by the covariance kernel $\mathcal{C}$. It seems natural to ask, therefore, whether certain covariance kernels are so simple that Theorem 1 is silent about the continuity properties of $F_{\hat{\mathbf{s}}}$ for many (possibly most) interesting mean functions. Revisiting a particularly simple, yet important, covariance kernel, Section 5.1 provides an affirmative answer to that question.

A related, but arguably more interesting, question is whether Assumption 1 *(iii)* is likely to be "close" to minimal in cases where the covariance kernel generates RKHSs that are sufficiently rich to contain many (possibly most) interesting mean functions. Revisiting another important covariance kernel, Section 5.2 provides evidence suggesting that the answer to that question will be affirmative in certain important special cases.

5.1 Bilinear Covariance Kernel

Suppose $\mathcal{C}(\mathbf{s},\mathbf{t}) = \mathbf{s}'\boldsymbol{\Sigma}\mathbf{t}$ for some (symmetric and) positive definite $\boldsymbol{\Sigma}$; that is, suppose $\mathcal{C}$ is a bilinear form. Then $\mathcal{G}^\mu(\mathbf{s}) = \mathbf{s}'\dot{\mathcal{G}}^\mu$, where $\dot{\mathcal{G}}^\mu = (\mathcal{G}^\mu(\mathbf{e}_1),\ldots,\mathcal{G}^\mu(\mathbf{e}_d))' \sim \mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})$. If also μ is a quadratic form $\mu(\mathbf{s}) = -\mathbf{s}'\boldsymbol{\Gamma}\mathbf{s}/2$ (for some symmetric and positive definite $\boldsymbol{\Gamma}$), then

$$\hat{\mathbf{s}} = \underset{\mathbf{s}\in\mathbb{R}^d}{\arg\max}\left\{-\frac{1}{2}\mathbf{s}'\boldsymbol{\Gamma}\mathbf{s} + \mathbf{s}'\dot{\mathcal{G}}^\mu\right\} = \boldsymbol{\Gamma}^{-1}\dot{\mathcal{G}}^\mu \sim \mathcal{N}(\mathbf{0},\boldsymbol{\Gamma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Gamma}^{-1}).$$

Thus, this case covers asymptotically normal estimators by writing the normal random limit as the argmax of a Gaussian process. More generally, under regularity conditions including invertibility of the gradient $\dot{\mu}$ of μ, we have $\hat{\mathbf{s}} = \dot{\mu}^{-1}(-\dot{\mathcal{G}}^\mu)$, implying in turn that the distributional properties of $\hat{\mathbf{s}}$ can be deduced with the help of standard tools.

In other words, if $\mathcal{C}$ is bilinear, then conditions for continuity of $F_{\hat{\mathbf{s}}}$ can be formulated without invoking the results of this paper. In fact, it turns out that Theorem 1 is completely silent about the case where $\mathcal{C}$ is bilinear because in that case μ satisfies Assumption 1*(iii)* if and only if it is a linear form $\mu(\mathbf{s}) = \mathbf{s}'\dot{\mu}$ (for some $\dot{\mu}\in\mathbb{R}^d$) , in which case Assumption 1*(i)* fails because $\mathcal{G}(\mathbf{s}) = \mathbf{s}'(\dot{\mu}+\dot{\mathcal{G}}^\mu)$ does not admit a maximizer over $\mathbb{R}^d$. To summarize, our results complement existing techniques, Assumption 1*(iii)* being very restrictive precisely when the covariance kernel of $\mathcal{G}$ is so simple that no new methods are needed in order to analyze the distribution of $\hat{\mathbf{s}}$.

5.2 Two-Sided Brownian Motion

The examples in Section 2 all have the common feature that if $d = 1$, then $\mathcal{C}$ is proportional to $\mathcal{C}_{\texttt{BM}}$. More generally, regardless of the value of d the examples all have the feature that $\mathcal{C}$ is a (linear) functional of $\mathcal{C}_{\texttt{BM}}$, a feature which in turn would appear to be shared by most other examples of estimators satisfying (1) with a covariance kernel that is not bilinear. It is therefore of interest

to further investigate the continuity properties of $F_{\hat{\mathbf{s}}}$ in the special case where $\mathcal{G}^\mu$ is a two-sided Brownian motion.

Accordingly, suppose $d = 1$ and suppose $\mathcal{C} = \sigma^2 \mathcal{C}_{\mathtt{BM}}$ for some $\sigma^2 > 0$. Also in this case Assumption 1*(iii)* reduces to a primitive condition on μ. Indeed, proceeding as in Giné and Nickl (2016, Example 2.6.7) it can be shown that Assumption 1*(iii)* holds if and only if μ is zero-preserving and absolutely continuous with a locally square integrable (weak) derivative.

In the leading special case where

$$\mu(s) = -c|s|^\gamma \qquad \text{for some } c, \gamma > 0, \tag{15}$$

Theorem 1 therefore implies that $F_{\hat{\mathbf{s}}}$ is continuous whenever $\gamma > 1/2$. The same condition on γ is necessary and sufficient in order to deduce continuity of $F_{\hat{\mathbf{s}}}$ by applying Cattaneo et al. (2024, Lemma A.2), which replaces Assumption 1*(iii)* with the Brownian motion-specific assumption

$$\lim_{\eta \downarrow 0} \frac{\mu(s+\eta) - \mu(s)}{\sqrt{\eta}} = 0 \qquad \text{for every } s \in \mathbb{R}. \tag{16}$$

Maintaining the assumption that $\mathcal{G}^\mu$ is a two-sided Brownian motion, but looking beyond mean functions of the form (15), the assumption (16) is slightly more general than Assumption 1*(iii)*. To see this, notice on the one hand that if μ is absolutely continuous with (weak) derivative $\dot{\mu}$, then

$$\left| \frac{\mu(s+\eta) - \mu(s)}{\sqrt{\eta}} \right| = \left| \frac{\int_s^{s+\eta} \dot{\mu}(t) \mathrm{d}t}{\sqrt{\eta}} \right| \leq \sqrt{\int_s^{s+\eta} \dot{\mu}(t)^2 \mathrm{d}t}$$

by the Cauchy-Schwarz inequality, so (16) holds if $\dot{\mu}$ is locally square integrable. On the other hand, the function

$$s \mapsto -|s| \sin\left(\frac{\pi/2}{\min\{1, |s|\}} \right)$$

satisfies (16), but is not absolutely continuous (on intervals containing zero).

In other words, in the special case of a two-sided Brownian motion, Assumption 1*(iii)* can be replaced by a slightly weaker assumption, namely (16), which can accommodate departures from absolute continuity. What is less clear, but arguably more interesting, is whether Assumption 1*(iii)* imposes unduly restrictive constraints on γ also when μ is of the form (15) near zero. In the remainder of this section, we attempt to shed light on that question.

When μ is of the form (15), the condition $\gamma > 1/2$ serves dual purposes when verifying the fact that $F_{\hat{\mathbf{s}}}$ is continuous. On the one hand, because (9) holds with $H = 1/2$, the condition $\gamma > 1/2$ ensures that the tail behavior of μ is such that (10) holds. In addition, $\gamma > 1/2$ is necessary to ensure that μ is sufficiently well behaved near zero that the weak derivative of μ is locally square integrable. To shed more light on Assumption 1*(iii)*, we disentangle the dual implications of $\gamma > 1/2$ in (15) by considering the mean function

$$\mu(s) = -c|s| \min\{1, |s|\}^{\gamma - 1} \qquad \text{for some } c, \gamma > 0, \tag{17}$$

which automatically satisfies (10), but satisfies Assumption 1*(iii)* only when $\gamma > 1/2$. The following example shows that every $\gamma < 1/2$ admits a $c = c(\gamma, \sigma^2) > 0$ such that if μ is given by (17), then $F_{\hat{\mathbf{s}}}$ is discontinuous, suggesting in turn that for this canonical covariance kernel at least, Assumption 1*(iii)* is close to minimal in Theorem 1.

Example. Fix $\gamma \in (0, 1/2)$ and note that with probability one the sample paths of $\mathcal{G}^\mu$ are γ-Hölder continuous on $[0, 1]$ (e.g., Theorem 14.5 (iii) of Kallenberg, 2021). As a consequence, there exists a constant $c = c(\gamma, \sigma^2)$ such that

$$\mathbb{P}\left[\sup_{s\in(0,1]} s^{-\gamma}\mathcal{G}^\mu(s) \geq c\right] < 1/4. \tag{18}$$

Fixing any such c, let $\mathcal{G} = \mathcal{G}^\mu + \mu$, where μ is defined as in (17).

By (18), we have

$$\mathbb{P}[\hat{\mathbf{s}} \in (0, 1]] \leq \mathbb{P}\left[\sup_{s\in(0,1]} \mathcal{G}(s) \geq 0\right] = \mathbb{P}\left[\sup_{s\in(0,1]} s^{-\gamma}\mathcal{G}^\mu(s) \geq c\right] < 1/4,$$

where the first inequality uses $\mathcal{G}(0) = 0$. Similarly,

$$\begin{aligned}\mathbb{P}[\hat{\mathbf{s}} \in [1, \infty)] &\leq \mathbb{P}\left[\sup_{s\in[1,\infty)} \mathcal{G}(s) \geq 0\right] = \mathbb{P}\left[\sup_{s\in[1,\infty)} s^{-1}\mathcal{G}^\mu(s) \geq c\right] \\ &\leq \mathbb{P}\left[\sup_{s\in[1,\infty)} s^{\gamma-1}\mathcal{G}^\mu(s) \geq c\right] = \mathbb{P}\left[\sup_{s\in(0,1]} s^{1-\gamma}\mathcal{G}^\mu(1/s) \geq c\right] < 1/4,\end{aligned}$$

where the first inequality uses $\mathcal{G}(0) = 0$, the second inequality uses $\gamma \geq 0$, and the third inequality uses (18) and the time inversion property of Brownian motion (e.g., Lemma 14.6 (i) of Kallenberg, 2021). Therefore, $\mathbb{P}[\hat{\mathbf{s}} > 0] \leq \mathbb{P}[\hat{\mathbf{s}} \in (0, 1]] + \mathbb{P}[\hat{\mathbf{s}} \in [1, \infty)] < 1/2$. Likewise, $\mathbb{P}[\hat{\mathbf{s}} < 0] < 1/2$, so $\mathbb{P}[\hat{\mathbf{s}} = 0] = 1 - \mathbb{P}[\hat{\mathbf{s}} < 0] - \mathbb{P}[\hat{\mathbf{s}} > 0] > 0$, implying in particular that $F_{\hat{\mathbf{s}}}$ is discontinuous at zero. ⌟

References

Abrevaya, J. and J. Huang (2005): "On the Bootstrap of the Maximum Score Estimator," *Econometrica*, 73, 1175–1204.

Azaïs, J.-M. and M. Chassan (2020): "Discretization Error for the Maximum of a Gaussian Field," *Stochastic Processes and their Applications*, 130, 545–559.

Azaïs, J.-M. and M. Wschebor (2005): "On the Distribution of the Maximum of a Gaussian Field with d Parameters," *Annals of Applied Probability*, 15, 254–278.

Cattaneo, M. D., M. Jansson, and K. Nagasawa (2020): "Bootstrap-Based Inference for Cube Root Asymptotics," *Econometrica*, 88, 2203–2219.

——— (2024): "Bootstrap-Assisted Inference for Generalized Grenander-type Estimators," *Annals of Statistics*, 52, 1509–1533.

Chernoff, H. (1964): "Estimation of the Mode," *Annals of the Institute of Statistical Mathematics*, 16, 31–41.

Cox, G. (2020): "Almost Sure Uniqueness of a Global Minimum without Convexity," *Annals of Statistics*, 48, 584–606.

Delgado, M. A., J. M. Rodriguez-Poo, and M. Wolf (2001): "Subsampling Inference in Cube Root Asymptotics with an Application to Manski's Maximum Score Estimator," *Economics Letters*, 73, 241–250.

Giné, E. and R. Nickl (2016): *Mathematical foundations of infinite-dimensional statistical models*, Cambridge University Press.

Groeneboom, P., S. Lalley, and N. Temme (2015): "Chernoff's distribution and differential equations of parabolic and Airy type," *Journal of Mathematical Analysis and Applications*, 423, 1804–1824.

Hansen, B. E. (2000): "Sample splitting and threshold estimation," *Econometrica*, 68, 575–603.

Hong, H. and J. Li (2020): "The Numerical Bootstrap," *Annals of Statistics*, 48, 397–412.

Jun, S. J., J. Pinkse, and Y. Wan (2015): "Classical Laplace estimation for $\sqrt[3]{n}$-consistent estimators: Improved convergence rates and rate-adaptive inference," *Journal of Econometrics*, 187, 201–216.

Kallenberg, O. (2021): *Foundations of Modern Probability (Third Edition)*, Springer.

Kim, J. and D. Pollard (1990): "Cube Root Asymptotics," *Annals of Statistics*, 18, 191–219.

Lai, P. Y. and S. M. S. Lee (2005): "An Overview of Asymptotic Properties of Lp Regression under General Classes of Error Distributions," *Journal of the American Statistical Association*, 100, 446–458.

Lee, S., Y. Liao, M. H. Seo, and Y. Shin (2021): "Factor-driven two-regime regression," *Annals of Statistics*, 49, 1656–1678.

Lee, S. M. S. and M. C. Pun (2006): "On m out of n Bootstrapping for Nonstandard M-Estimation With Nuisance Parameters," *Journal of the American Statistical Association*, 101, 1185–1197.

Lee, S. M. S. and P. Yang (2020): "Bootstrap Confidence Regions Based on M-Estimators under Nonstandard Conditions," *Annals of Statistics*, 48, 274–299.

Lifshits, M. A. (1995): *Gaussian Random Functions*, Springer.

López, S. I. and L. P. Pimentel (2018): "On the location of the maximum of a process: Lévy, Gaussian and Random field cases," *Stochastics*, 90, 1221–1237.

Manski, C. F. (1975): "Maximum score estimation of the stochastic utility model of choice," *Journal of Econometrics*, 3, 205–228.

Mohammadi, L. and S. van de Geer (2005): "Asymptotics in Empirical Risk Minimization," *Journal of Machine Learning Research*, 6, 2027–2047.

Patra, R. K., E. Seijo, and B. Sen (2018): "A Consistent Bootstrap Procedure for the Maximum Score Estimator," *Journal of Econometrics*, 205, 488–507.

Pitman, J. and W. Tang (2018): "The argmin process of random walks, Brownian motion and Lévy processes," *Electronic Journal of Probability*, 23, 1–35.

Politis, D. N., J. P. Romano, and M. Wolf (1999): *Subsampling*, Springer.

Samorodnitsky, G. and Y. Shen (2012): "Distribution of the Supremum Location of Stationary Processes," *Electronic Journal of Probability*, 17, 1–17.

——— (2013a): "Intrinsic Location Functionals of Stationary Processes," *Stochastic Processes and their Applications*, 123, 4040–4064.

——— (2013b): "Is the Location of the Supremum of a Stationary Process Nearly Uniformly Distributed?" *Annals of Probability*, 41, 3494–3517.

Shen, Y. (2018): "Location of the Path Supremum for Self-similar Processes with Stationary Increments," *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 54, 2349–2360.

van der Vaart, A. W. (1998): *Asymptotic Statistics*, Cambridge University Press.

van der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes: With Applications to Statistics*, Springer.

Westling, T. and M. Carone (2020): "A Unified Study of Nonparametric Inference for Monotone Functions," *Annals of Statistics*, 48, 1001–1024.

Yu, P. and X. Fan (2021): "Threshold regression with a threshold boundary," *Journal of Business and Economic Statistics*, 39, 953–971.