The Annals of Statistics
1997, Vol. 25, No. 5, 1926–1953

MONTE CARLO SAMPLING IN DUAL SPACE FOR APPROXIMATING THE EMPIRICAL HALFSPACE DISTANCE[1]

BY GUENTHER WALTHER

Stanford University

The Kolmogorov–Smirnov distance is an important tool for constructing confidence sets and tests in univariate problems. In multivariate settings, an analogous role is played by the halfspace distance, which has the merit of being invariant under linear transformations. However, the evaluation of the halfspace distance between two samples is a computationally very intensive combinatorial problem even in moderate dimensions, which severely restricts the use of the halfspace distance, especially in resampling procedures. To approximate this distance in a fast and data-dependent way, the notion of a dual measure is introduced. Based on geometric concepts, it will be shown how the above problem can be put as a density estimation problem using Monte Carlo sampling in a certain dual space. A central limit theorem for the empirical halfspace distance is derived and used as a gauge to compare the new procedure with a traditional random search.

1. Introduction. In a univariate setting, the Kolmogorov–Smirnov distance serves as a standard tool to find confidence sets for a distribution and to construct goodness-of-fit tests.

For d-dimensional probability measures F and G, an analogous role is played by the halfspace distance

$$d(F,G) := \sup_{H \in \{\text{halfspaces in } \mathbf{R}^d\}} |F(H) - G(H)| \tag{1.1}$$

[see, e.g., Beran and Millar (1986, 1989)]. Introduced by Wolfowitz (1954), this distance has the advantage of being invariant under linear transformations, a property not enjoyed by the multivariate Kolmogorov–Smirnov distance based on quarterspaces. A problem with the halfspace distance, however, lies in its evaluation, even if F and G are empirical measures and thus have finite support, a case that arises, for example, when bootstrapping is employed. Looking at all halfspaces in $\mathbf{R}^d$ that have d points lying on their boundaries entails a computational burden that is of the order n^d, where n is the number of points considered. This can easily lead to a prohibitive task even in moderate dimensions, especially if a large number of bootstrap

Received April 1995; revised March 1997.

[1] Research partially supported by NSF Grant DMS-92-24868 at University of California, Berkeley, and by Studienstiftung des deutschen Volkes.

AMS 1991 *subject classifications*. Primary 62G05; secondary 62H40, 65U05.

Key words and phrases. Halfspace distance, dual measure, random search, central limit theorem, Monte Carlo.

replications are involved. If F or G does not have finite support and no tractable analytical expressions are at hand, the measures have to be approximated by a finite sample.

Because of these problems one usually resorts to an approximation to the halfspace distance in lieu of using its exact value. Beran and Millar (1986, 1987, 1989) show how a random search can be used to obtain an approximation that does not detract from vital properties of the underlying confidence sets and tests. This approximation scheme first generates a direction, that is, an element of the unit sphere, at random, then projects the data points on the one-dimensional subspace spanned by this direction and finally computes the one-dimensional Kolmogorov–Smirnov statistic of the projected data. Repeating this procedure for a large number of directions and retaining the maximum separation obtained between the projected data yields an approximation to the halfspace distance. The problem with this approximation scheme is that it wastes time to explore "uninteresting" directions. This problem becomes especially acute in high dimensions and is shared by certain projection pursuit statistics [see, e.g., Li and Cheng (1993)].

Section 3 introduces a new method to approximate the halfspace distance in a fast and data-dependent way. The key concept is that of a dual measure, which is motivated by the notion of a dual set in geometry and defined in Section 2, where also some of its relevant properties are investigated. It is shown how the problem of computing the empirical halfspace distance can be put as a density estimation problem using an auxiliary Monte Carlo sample in a certain dual space, and how this can be exploited by, for example, using Fourier methods. In Section 4 a central limit theorem (CLT) for the empirical halfspace distance is derived and necessary and sufficient conditions are given for any approximation scheme to ensure the validity of the CLT for the approximation to the empirical halfspace distance obtained by that scheme. This result is used in Section 5 to compare the new procedure with a traditional random search on a theoretical basis. A simulation study is presented in Section 6. Most proofs are deferred to Section 7.

2. Dual measures.

2.1. *Geometric preliminaries and notation.* The setting used throughout is Euclidean d-space $\mathbf{R}^d$ equipped with the standard inner product $\langle \cdot , \cdot \rangle$; $|\cdot|$ denotes the d-dimensional Lebesgue measure of a set as well as Euclidean norm in $\mathbf{R}^d$ and absolute value in $\mathbf{R}$, the meaning being clear from the context; $\mathscr{M}^d$ denotes the set of probability measures on $\mathbf{R}^d$; and F_m and G_n are the empirical measures pertaining to samples of size m and n from the probability measures F and G, respectively. For the following definitions and facts see, for example, Stoer and Witzgall (1970).

The *dual* (*polar*) set of a set $A \in \mathbf{R}^d$ is defined as

$$A^* := \{ x \in \mathbf{R}^d : \langle x, a \rangle \leq 1 \text{ for all } a \in A \}.$$

It follows from this definition that $A \subset B$ implies $B^* \subset A^*$ and $A^{**} = \operatorname{conv}(A \cup \{0\})$, where conv denotes the convex hull. Thus, if A is a closed, convex set containing the origin, then $A^{**} = A$ and the mapping $A \mapsto A^*$ is a duality.

As an example, if $B_r(x)$ denotes the closed ball with center x and radius $r > 0$, then $(B_r(0))^* = B_{1/r}(0)$. We will be mainly interested in the following duality between points and halfspaces:

If $a \in \mathbf{R}^d \setminus \{0\}$, then $\{a\}^* = \{x \in \mathbf{R}^d \colon \langle x, a \rangle \leq 1\}$ is the closed halfspace containing 0 in its interior whose bounding hyperplane has normal vector $a/|a|$ and distance $1/|a|$ from 0; $\{a\}^{**}$ is the line segment joining 0 and a, which can be identified with a. Further, $\{0\}^* = \mathbf{R}^d$ and $(\mathbf{R}^d)^* = \{0\}$.

2.2. *Motivation and definition.* Let X be a random variable in $\mathbf{R}^d$. For $a \in \mathbf{R}^d$ consider the inequality

$$\langle a, X \rangle \leq 1, \tag{2.1}$$

which lies at the heart of the duality notion described above. As the inner product $\langle \cdot, \cdot \rangle$ is symmetric, a natural thought is to let a also be random. There is a canonical way in which the distribution of X induces a measure via (2.1) that can be interpreted as a measure on a certain dual space and will hence be called *dual measure*.

DEFINITION 2.1. Let F be a probability measure on $\mathbf{R}^d$. Define the dual measure F^* as the infinite measure given by the density (w.r.t. Lebesgue measure on $\mathbf{R}^d$)

$$f_F^*(x) = F\{x\}^*. \tag{2.2}$$

For convenience later on, in the presence of a random variable X with distribution F, the above definition reads

$$f_F^*(\cdot) = \mathbb{P}(\langle \cdot, X \rangle \leq 1). \tag{2.3}$$

As an illustration, Figures 1–3 show three different dual densities plotted on the unit square in $\mathbf{R}^2$.

2.3. *Basic properties and a representation theorem.* The following properties of a dual density f_F^* will be used in the sequel:

PROPOSITION 2.2. *Let F be a probability measure on* $\mathbf{R}^d$.

(i) *There exists some universal constant* $p > 0$ *such that* $f_F^* \geq p$ *on some nondegenerate convex cone.*

(ii) *The dual density* f_F^* *is upper semicontinuous and the set of discontinuity points of* f_F^* *has Lebesgue measure* 0.

Clearly, $0 \leq f_F^* \leq 1$. So Proposition 2.2 shows that f_F^* is a measurable function and in fact is the density (w.r.t. Lebesgue measure) of an infinite but σ-finite measure F^*.

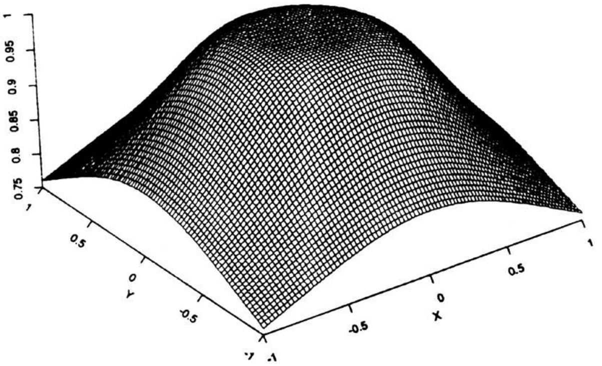

FIG. 1. *Dual density of* $N((0,0), I)$.

PROOF OF PROPOSITION 2.2. Consider the setting (2.3). There exists a quadrant E with $\mathbb{P}(X \in E) \geq p = 2^{-d}$. Then $-E = E^*$ is a convex cone satisfying statement (i): $e \in E$ implies $\{-e\}^* \supset E^{**} = E$ and hence $f_F^*(-e) = \mathbb{P}(X \in \{-e\}^*) \geq \mathbb{P}(X \in E) \geq p$. As for (ii), the linear functional $\mathbf{G}$ defined on $\mathscr{M}^1$ by $\mathbf{G}$: $P \mapsto P((-\infty, 1])$ is upper semicontinuous (w.r.t. the topology of weak convergence) by the Portmanteau theorem. So continuity of the function $\mathbf{H}$ from $\mathbf{R}^d$ to $\mathscr{M}^1$ defined by $\mathbf{H}$: $a \mapsto \mathscr{L}(\langle a, X \rangle)$ implies that the composition $f_F^* = \mathbf{G} \circ \mathbf{H}$ is upper semicontinuous. The second assertion in (ii) is proven in Walther (1994), where further properties of dual measures also are given. □

The basis for sampling from the dual measure as employed in the next section is provided by the following local representation theorem.

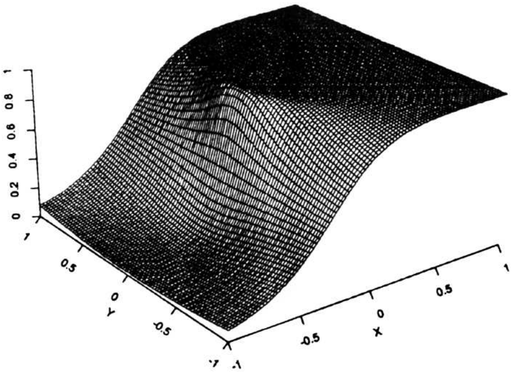

FIG. 2. *Dual density of* $N((-3,0), I)$.

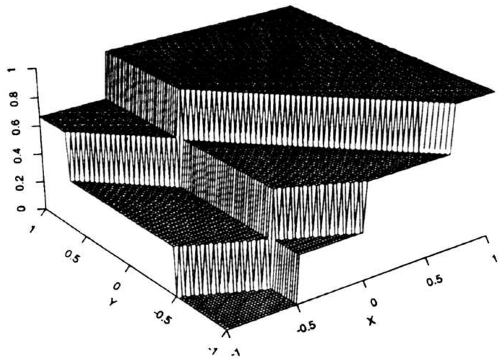

FIG. 3. *Dual density of* $\frac{1}{3}(\delta_{(-2,0)} + \delta_{(-3,-3)} + \delta_{(-0.5,-1)})$.

THEOREM 2.3. *Let $B \in \mathbf{R}^d$ be a compact set, let $X \in \mathbf{R}^d$ be a random variable having distribution F and denote by $F^*|_B$ the restriction of F^* to B. Then*

$$F^*|_B(\cdot) = c_B \int U_B(x, \cdot) G_B(dx),$$

where the scaling constant c_B is given by $c_B = \mathbb{E}|B \cap \{X\}^*|$, the Markov kernel

$$U_B(x, \cdot) = \frac{|B \cap \{x\}^* \cap \cdot|}{|B \cap \{x\}^*|}$$

is the uniform distribution on $B \cap \{x\}^*$ (here $|\cdot|$ denotes Lebesgue measure) and, except in the trivial case $c_B = 0$, the probability measure G_B is absolutely continuous w.r.t. F:

$$\frac{dG_B(x)}{dF(x)} = \frac{|B \cap \{x\}^*|}{\mathbb{E}|B \cap \{X\}^*|}.$$

Theorem 2.3 expounds the following concept: For a one-dimensional distribution function F and real a write

$$F(a) = \int 1_{(-\infty, a]}(x) F(dx) = \int 1_{[x, \infty)}(a) F(dx).$$

Then the last term can formally be read as a mixture of uniform densities (albeit not probability densities).

3. Constructing an estimator in dual space.

This section will show how the dual measures introduced in the previous section can be used to devise a fast and data-dependent approximation to the halfspace distance. It follows from the previous section that, for $a, b \in \mathbf{R}^d$,

$$a \in \{b\}^* \quad \text{iff } b \in \{a\}^*, \tag{3.1}$$

which yields the following duality relation:

$$\mathbb{P}(X \in \{a\}^*) = \mathbb{P}(a \in \{X\}^*). \tag{3.2}$$

As our goal is the halfspace distance, we are interested in the quantity that appears on the left-hand side of (3.2). However, to compute an estimator, we will work in the dual space with the quantity on the right-hand side.

To formulate the problem in the dual space denote by $\mathscr{D}_d^* = \{F\{\cdot\}^* : F \in \mathscr{M}^d\}$ the set of dual densities on $\mathbf{R}^d$ and consider $\mathscr{D}_d^*$ as a subset of the normed linear space $(L_\infty(\mathbf{R}^d), \|\cdot\|_\infty)$. Then the distance on $\mathscr{D}_d^*$ derived from the norm $\|\cdot\|_\infty$ is the halfspace distance:

$$d(F, G) = \|f_F^* - f_G^*\|_\infty \tag{3.3}$$

for $F, G \in \mathscr{M}^d$. This is so because, for $a \in \mathbf{R}^d$, $\{a\}^*$ ranges over all halfspaces with 0 in their interiors, and $\{0\}^* = \mathbf{R}^d$. Using continuity from above for probability measures one concludes

$$d(F, G) = \sup_{a \in \mathbf{R}^d} |F\{a\}^* - G\{a\}^*|$$

and (3.3) follows.

Equation (3.3) shows that the problem of computing the halfspace distance can be interpreted as a density estimation problem in dual space: the halfspace at which $d(F_m, G_n)$ is achieved is the dual of the mode of $|f_{F_m}^* - f_{G_n}^*|$. This motivates the following approach. Draw auxiliary samples of size k from the densities $f_{F_m}^*$ and $f_{G_n}^*$ restricted to a compact set B and rescaled to probability densities there. Then use density estimation to obtain an estimate of $|f_{F_m}^* - f_{G_n}^*|$. There are several possibilities to make use of this estimate: one can start sampling from a density proportional to this estimate of $|f_{F_m}^* - f_{G_n}^*|$, for example, by rejection sampling. Looking at the duals of these points yields halfspaces that will be more concentrated on interesting regions than those obtained by just a uniform random generation, thus giving a data-dependent way to evaluate the halfspace distance.

Instead, we will take the even more promising approach of estimating the mode of $|f_{F_m}^* - f_{G_n}^*|$ and using the dual of the estimated mode as a pilot estimate at which to evaluate the empirical measures.

To see why it is advantageous to take this route, note that generating k auxiliary points in the dual space corresponds to generating k halfspaces in some random way. However, as opposed to the case where the empirical measure of the halfspaces is computed, as in a traditional random search, using density estimation in dual space does not require processing the original sample. Moreover, if one uses the fast Fourier transform for the density estimation, then the auxiliary sample has to be processed only once: note that the computational burden of using the fast Fourier transform to evaluate a density estimate on a grid depends essentially only linearly on the number of grid points as the sample has to be processed only once [see, e.g., Wand (1994) for a detailed analysis]. As each grid point in dual space corresponds to a halfspace in the original space, the computational burden of this estimation scheme in dual space depends essentially only on the number of halfspaces examined. This is in contrast to the traditional random search described in Section 1, where the sample has to be projected anew in each direction examined. A detailed comparison of these two search schemes is given in Section 5.

To summarize the construction of the estimator based on i pilot estimates, here is a description of the algorithm. Recall that we are given a sample $X_1, \dots, X_m$ from F and $Y_1, \dots, Y_n$ from G.

Fix a compact set B containing the origin, for example, a ball or a hypercube.

1. Draw an auxiliary sample of size k from the density proportional to $f^*_{F_m}$ on B; likewise for $f^*_{G_n}$.
2. Choose an evaluation set T (e.g., the auxiliary sample from step 1 or a grid), and compute on T a kernel estimate $\hat{f}_k$ of $f^*_{F_m} - f^*_{G_n}$ based on the auxiliary sample from step 1.
3. Set $\hat{t} = \arg\max_T |\hat{f}_k|$.
4. Repeat steps 1–3 i times and evaluate $|F_m - G_n|$ on the search set of halfspaces $\{\{\hat{t}_1\}^*, \dots, \{\hat{t}_i\}^*\}$.

The maximum value of $|F_m - G_n|$ found gives the estimate of the halfspace distance.

The sampling from $f^*_{F_m}$ in step 1 can be executed in a straightforward way using the local representation Theorem 2.3:

1. Choose an integer i according to the uniform distribution on $\{1, \dots, m\}$. Accept i with probability $|B \cap \{X_i\}^*|/|B|$. Repeat until an integer i has been accepted.
2. Sample from the uniform distribution on $B \cap \{X_i\}^*$, that is, generate a point u from the uniform distribution on B until $\langle u, X_i \rangle \le 1$.

Then u comes from a distribution whose density is proportional to $f^*_{F_m}$ on B. One can easily combine the sampling and density estimation procedures for $f^*_{F_m}$ with those for $f^*_{G_n}$, which saves the time required to compute various proportionality constants. The details, as well as other specifics of the algorithm, will be given in the next sections for the specific situations treated there.

The restriction to the set B means that we are only searching over the range of halfspaces $\{\{a\}^*,\ a \in B\}$. If one does not have a priori knowledge to justify this, one has to shift the data after each repetition in step 4 in a certain direction to cover a different range of halfspaces. It can be arranged that d different shifts (d is the dimension) suffice. To see in detail how this can be done, in the following let $B = B_r(0)$ be a ball centered at 0. Other sets B (e.g., hypercubes) can be dealt with using straightforward modifications.

We will shift the sample from its original position by ce_i, $i = 1, \dots, d$, where $c > 0$ will be determined later and $\{e_1, \dots, e_d\}$ is the standard basis in $\mathbf{R}^d$. Set $e_0 = 0$ to incorporate the case where no shift occurs. Shifting the data by a vector s means shifting the coordinate system by $-s$. Hence, if a point $a \in B_r(0)$ is generated in dual space, then its dual set is described in the coordinate system of the shifted data by the translated set $\{a\}^* - s$. If $a = 0$, this set is all of $\mathbf{R}^d$, otherwise it is a halfspace not necessarily containing 0 in its interior. When an appropriate choice of vectors ce_i is used to shift the data, then the boundaries $\partial(\{a\}^* - ce_i)$ of these halfspaces range over all hyperplanes in $\mathbf{R}^d$:

LEMMA 3.1. *If $cr \ge 2\sqrt{d}$, then*

$$\{\partial(\{a\}^* - ce_i): a \in B_r(0) \setminus \{0\}, i = 0, \dots, d\} = \{\textit{hyperplanes in } \mathbf{R}^d\}.$$

Hence shifting the sample $(\mathbf{X}, \mathbf{Y})$ as described with $c \ge 2\sqrt{d}/r$ guarantees that one searches over all possible separating hyperplanes in $\mathbf{R}^d$ and thus over all possible values of $|F_m(H) - G_n(H)|$, where H ranges over all halfspaces in $\mathbf{R}^d$, as the empirical measures have finite support.

Concerning the implementation of this algorithm, note that the data would of course not really be shifted, but rather the computer program would add an appropriate constant to the data every time they are used.

There are other statistical problems involving the probability content of halfspaces where the duality relation (3.2) may be successfully employed, for example, certain robust location estimators [see, e.g., Nolan (1989, 1992) and Donoho and Gasko (1992)], and projection pursuit statistics [see, e.g., Li and Cheng (1993)]. Those topics will be treated elsewhere.

4. A CLT for the empirical halfspace distance. In the following we will write $Z = F - G$, $Z_{m,n} = F_m - G_n$ and use the usual conventions for signed measures, that is, $f_Z^*(\cdot) = f_F^*(\cdot) - f_G^*(\cdot)$ and so on. The next theorem states a functional CLT for the $f_{Z_{m,n}}^*(\cdot)$-process and then establishes a CLT for the empirical halfspace distance.

THEOREM 4.1. *Let $m = m(n)$ go to infinity together with n such that $\lim_{n\to\infty} n/m(n) = \lambda \ge 0$.*

(a) *$\sqrt{n}(f_{Z_{m,n}}^* - f_Z^*)$ converges weakly, as a random element of $L_\infty(\mathbf{R}^d)$, to a Gaussian process W on $\mathbf{R}^d$ having mean 0 and covariance function*

$$\mathbb{E}W(t)W(t') = (\lambda F + G)(\{x: \langle x, t\rangle \le 1\} \cap \{x: \langle x, t'\rangle \le 1\}) \\ - \lambda f_F^*(t) f_F^*(t') - f_G^*(t) f_G^*(t').$$

(b) *Assume the following*: (i) *there is a unique hyperplane that optimally separates the probabilities F and G in the following sense*: *the function $g(\cdot) := |(F - G)(\{x: \langle x, \cdot\rangle \le 1\})|$ satisfies $\inf_{t \notin N}(g(t_0) - g(t)) > 0$ for every neighborhood N of some point t_0*; (*ii*) *$(F + G)(\{x: \langle x, t_0\rangle = 1\}) = 0$.*

Then the functional $T(\cdot) = \|\cdot\|_\infty$ has a stochastic differential on the set $\{f - g: f, g \in \mathscr{D}_d^*\}$ at f_Z^*, that is, there exists a linear functional $T(f_Z^*; \cdot)$ defined on the space spanned by differences of elements of $\mathscr{D}_d^*$ that satisfies

$$(4.1)\qquad T(f_{Z_{m,n}}^*) - T(f_Z^*) = T(F_Z^*; f_{Z_{m,n}}^* - f_Z^*) \\ + o_p(\|f_{Z_{m,n}}^* - f_Z^*\|_\infty), \qquad n \to \infty$$

and

$$\|f_{Z_{m,n}}^* - f_Z^*\|_\infty \to_P 0.$$

The differential is given by $T(f_Z^*;\cdot) = (\operatorname{sign} f_Z^*(t_0))e_{t_0}(\cdot)$, *where* $e_{t_0}(\cdot)$ *denotes the evaluation operator at* t_0. *In this case the empirical halfspace distance satisfies the CLT*

$$\sqrt{n}\,(d(F_m, G_n) - d(F,G)) \\ \to_d N\big(0, \lambda f_F^*(t_0)(1 - f_F^*(t_0)) + f_G^*(t_0)(1 - f_G^*(t_0))\big).$$

(c) *Without assumption* (ii), (4.1) *continues to hold with the* (*nonlinear*)

$$T(f_Z^*;\cdot) = \lim_{\varepsilon\downarrow 0}\sup_{t\in A_\varepsilon}\big((\operatorname{sign} f_Z^*(t))e_t(\cdot)\big),$$

where $A_\varepsilon = \{t\colon |f_Z^*(t)| \geq \|f_Z^*\|_\infty - \varepsilon\}$, *and the empirical halfspace distance satisfies the CLT*

$$\sqrt{n}\,(d(F_m, G_n) - d(F,G)) \to_d \lim_{\varepsilon\downarrow 0}\sup_{t\in A_\varepsilon}\big((\operatorname{sign}(f_Z^*(t))W(t)\big)$$

Observe that for certain elliptically contoured distributions condition (i) is always satisfied.

As explained in Section 1, for practical use the halfspace distance $d(F_m, G_n)$ is usually replaced by an approximation $\sup_{H\in S_n}|F(H) - G(H)|$ due to computational reasons, where the search set S_n of halfspaces is generated in some deterministic or stochastic way. Proposition 4.2 below gives a necessary and sufficient condition on the search set S_n for the validity of the CLT

$$\sqrt{n}\Big(\sup_{H\in S_n}|F_m(H) - G_n(H)| - d(F,G)\Big) \\ \to_d N\big(0, \lambda f_F^*(t_0)(1 - f_F^*(t_0)) + f_G^*(t_0)(1 - f_G^*(t_0))\big) \tag{4.2}$$

under the assumption of Theorem 4.1(b). Observe that for each halfspace H there exists a halfspace containing 0 in its interior whose bounding hyperplane separates the sample $X_1,\dots,X_m, Y_1,\dots,Y_n$ in the same way as H does and hence results in the same value $|F_m(H) - G_n(H)|$. We will therefore restrict our attention to search sets of halfspaces that contain 0 in their interiors. Otherwise no restrictions whatsoever are placed on the search set. It may be stochastic, obtained in a data-dependent or independent way, be of any size and change arbitrarily with the sample size n.

To simplify notation in the following, assume w.l.o.g. that $f_Z^*(t_0) > 0$ [recall that $f_Z^*(t_0) \neq 0$]. Then

$$\lim_n \sup_t\big(-f_{Z_{m,n}}^*(t)\big) = \sup_t\big(-f_Z^*(t)\big) < f_Z^*(t_0) \quad \text{a.s.}, \tag{4.3}$$

because of (7.2), (i) and (ii), which implies continuity of f_Z^* at t_0.

Hence the CLT (4.2) reads

$$\sqrt{n}\Big(\sup_{t\in\{H^*\colon H\in S_n\}} f_{Z_{m,n}}^*(t) - f_Z^*(t_0)\Big) \\ \to_d N\big(0, \lambda f_F^*(t_0)(1 - f_F^*(t_0)) + f_G^*(t_0)(1 - f_G^*(t_0))\big). \tag{4.4}$$

Also,

$$A_\varepsilon = \{t\colon f_Z^*(t) \geq f_Z^*(t_0) - \varepsilon\} \tag{4.5}$$

for small ε because of (4.3).

For clarity it is helpful in the following to switch occasionally from the underlying probability measure $\mathbb{P}$ and the random variables $X_1, \dots, X_m$ to the m-fold product law F^m on the image space $(\mathbf{R}^d)^m$, and analogously for G^n and $Y_1, \dots, Y_n$. The notation $f_{F_m}^*$, $f_{G_n}^*$ and $F_{Z_{m,n}}^*$ will not be changed as the meanings will always be clear.

PROPOSITION 4.2. *Denote by* μ_{S_n} *the law according to which the search set* S_n *is generated. Then, under the assumption of Theorem* 4.1(b), *for the CLT* (4.2) *to hold under the law* $F^m \otimes G^n \otimes \mu_{S_n}$ *it is necessary and sufficient that there exists a positive sequence* $\{l_n\}$ *with* $l_n = o(n^{-1/2})$, $n \to \infty$, *such that*

$$\mu_{S_n}\big(A_{l_n} \cap \{H^*\colon H \in S_n\} \neq \varnothing\big) \to 1. \tag{4.6}$$

Note that in the case of a deterministic search set, (4.6) requires that the search set in dual space eventually hits A_{l_n}. For certain classes of distributions (see, e.g., the next section), this yields clear-cut recipes for the construction of the search set. The proposition can also be shown to hold in the general case of Theorem 4.1(c), but this will not be needed in the following.

5. Comparison with a traditional random search. The next two subsections investigate conditions under which the CLT (4.2) holds for the traditional uniform random search described in Section 1 and for the search scheme using Monte Carlo sampling in dual space introduced in Section 3. The results thus obtained will allow a comparison of the two schemes in terms of computing time and quality of the approximation in Section 5.3.

5.1. *The CLT for approximations with random search.* In the following, $a_n \ll b_n$ for positive sequences $\{a_n\}, \{b_n\}$ shall mean $\lim_{n\to\infty} b_n/a_n = \infty$, and $a_n \asymp b_n$ shall mean that there exists constants $0 < c_1 < c_2$ with $c_1 \leq b_n/a_n \leq c_2$ for all n.

Recall that for the uniform random search i_n directions $e_1, \dots, e_{i_n} \in S^{d-1}$ are chosen i.i.d. according to the uniform distribution on the unit sphere S^{d-1}. The search set S_n is then given by

$$S_n = \big\{H\colon H = \{x\colon \langle x, e\rangle \leq t\},\, t \in \mathbf{R},\, e \in \{e_1, \dots, e_{i_n}\}\big\}.$$

We are interested in necessary and sufficient conditions on the number of directions i_n for the CLT (4.2) to hold under the assumptions Theorem 4.1(b). Proposition 4.2 shows that no general answer can be given as the behavior of the sets A_{l_n} depends on the underlying distributions. We will therefore restrict ourselves in the following to distributions F and G with f_F^* and f_G^* being twice continuously differentiable in a neighborhood of t_0. This smoothness requirement is not as restrictive as it might look at first sight; it is, for example, not necessary that F have a density in order for f_F^* to meet this condition: for example, let F be the uniform distribution on the unit circle in $\mathbf{R}^2$. Then

$$f_F^*(t) = 1 - \frac{1}{\pi}\arccos\left(\frac{1}{|t|}\right)1(|t| \geq 1),$$

which is twice continuously differentiable except on the unit sphere. The underlying reason is that the projection of F inherent in the computation of f_F^* is a smoothing operation.

If f_F^* and f_G^* are twice continuously differentiable in a neighborhood of t_0, then so is $|f_Z^*| = |f_F^* - f_G^*|$ in some neighborhood of t_0 because $f_Z^*(t_0) \neq 0$, and we get the expansion

$$(5.1) \qquad |f_Z^*(t)| = |f_Z^*(t_0)| + 1/2(t - t_0)'H(t - t_0) + o\left(|t - t_0|^2\right).$$

The linear term does not appear in (5.1) because $|f_Z^*|$ has a maximum at t_0.

THEOREM 5.1. *Assume the conditions of Theorem* 4.1(b) *and further that* f_F^* *and* f_G^* *are twice continuously differentiable in a neighborhood of* t_0 *with the Hessian* H *in* (5.1) *being nonsingular. Then a necessary and sufficient condition for the CLT* (4.2) *to hold under the law* $F^m \otimes G^n \otimes \mu_{S_n}$, *where the search set* S_n *is generated by uniform sampling as described above*, *is that the number of directions* i_n *satisfies*

$$i_n n^{-(d-1)/4} \to \infty, \qquad n \to \infty, d > 1.$$

5.2. *The CLT for approximations with a pilot estimate in dual space.* For the approximation procedure using the pilot estimate in dual space, fix a compact set B containing 0 in its interior and assume that the dual of the unique maximizing halfspace falls into B. Shifting the data as described in Section 3 allows for an analogous treatment in the general case.

Denote by

$$(5.2) \qquad f_F^{*s}(\cdot) = \frac{f_F^*(\cdot)1_B(\cdot)}{\int_B f_F^*}$$

the standardized dual density on B. (Using Proposition 2.2 one readily checks that $\int_B f_F^* > 0$ for the aforementioned B). Likewise, F^{*s} denotes the corresponding probability measure. The dependence on B will be suppressed for notational simplicity.

For the purpose of estimating f_F^* and f_G^* we will use kernels from the class $\mathscr{K}_p$ of all compactly supported pth-order kernels, so that $K \in \mathscr{K}_p$ satisfies $\int K = 1$ and $\int K(t)\prod_{i=1}^j t(i)\,dt = 0$ for all coordinate vectors $t(i) \in \{t_1, \dots, t_d\}$ and $1 \leq j < p$. See, for example, Härdle (1990) or Silverman (1986) for the use of higher-order kernels to reduce the bias.

We will use the abbreviation $\bar{f}^{(\sigma)}(x) = \int \sigma^{-d} K((y-x)/\sigma) f(y)\,dy$ for positive σ and a real-valued function f on $\mathbf{R}^d$. The dependence on K shall be suppressed as the kernel used will be clear from the context. Also write $K_{x,\sigma}(\cdot) = K((\cdot - x)/\sigma)$. Finally, we will later require that $K_{x,\sigma}$ has polynomial discrimination; see, for example, Pollard (1984) for background on empirical process theory.

Recall steps 1 and 2 of the algorithm for the approximation procedures given in Section 3: after drawing an auxiliary sample of size k from the random measure F_m^{*s}, a kernel estimate of $f_{F_m}^*$ can be obtained by computing $I_{F_m}(F_m^{*s})_k \sigma_k^{-d} K_{x,\sigma_k}$, where $(F_m^{*s})_k$ denotes the stochastic empirical in dual space, that is the empirical measure obtained by drawing k points i.i.d. from the dual empirical measure F_m^{*s}, $I_{F_m} := \int_B f_{F_m}^* = (1/m)\sum_{i=1}^m \mathrm{Vol}(B \cap \{X_i\}^*)$, and σ_k is a bandwidth to be chosen. As indicated in Section 3, it is advantageous to combine the auxiliary sampling and density estimation procedures for $f_{F_m}^*$ with that for $f_{G_n}^*$ in order to avoid the computation of the scaling factors I_{F_m} and I_{G_n}. One can obtain an empirical of size k from the signed measure with density proportional to

$$f_{Z_{m,n}}^* = I_{F_m} f_{F_m}^{*s} - I_{G_n} f_{G_n}^s \tag{5.3}$$

on B by way of the following algorithm:

```
Without loss of generality assume $m \le n$, so $m/n \le 1$.
  For $i = 1, \dots, k$ do:
    Sample an integer $p$ uniformly from $\{1, \dots, m+n\}$.
      If $p \le m$, accept $p$ with probability $|B \cap \{X_p\}^*|/|B|$. If $p$ is accepted,
      set $p_i = p$ and sample a point $V$ uniformly from $|B \cap \{X_p\}^*|$.
      If $p > m$, accept $p$ with probability $(m/n)(|B \cap \{Y_{p-m}\}^*|/|B|)$. If $p$ is
      accepted, set $p_i = p$ and sample a point $W$ uniformly from $|B \cap$
      $\{Y_{p-m}\}^*|$.
    Repeat until a point $p$ is accepted.
```

Observe that given $p_i \le m$, V is generated from the density $f_{F_m}^{*s}$. Thus we obtain $V_1, \dots, V_{N^+}$ i.i.d. from the density $f_{F_m}^{*s}$, given $(\mathbf{X}, \mathbf{Y})$, where $N^+ := \sum_{i=1}^k 1(p_i \le m)$ denotes the total number of points V generated. Analogously, $W_1, \dots, W_{N^-}$ are i.i.d. from the density $f_{G_n}^{*s}$, given $(\mathbf{X}, \mathbf{Y})$, where $N^- = k - N^+$.

Denote by $\mu(\mathbf{X}, \mathbf{Y}, k)$ the random measure that generates the auxiliary sample of size k according to the algorithm above, given the original sample $(\mathbf{X}, \mathbf{Y})$. Then

$$\begin{aligned}\mu(\mathbf{X}, \mathbf{Y}, k)(p_i \le m) &= \frac{\sum_{i=1}^m |B \cap \{X_i\}^*|/|B|}{\sum_{i=1}^m |B \cap \{X_i\}^*|/|B| + (m/n)\sum_{i=1}^n |B \cap \{Y_i\}^*|/|B|} \\ &= \frac{I_{F_m}}{I_{F_m} + I_{G_n}}.\end{aligned}$$

Hence the signed empirical measure

$$\frac{1}{k}\left(\sum_{i=1}^{N^+}\delta_{V_i}(\cdot)-\sum_{i=1}^{N^-}\delta_{W_i}(\cdot)\right)$$

comes indeed from a signed measure with density on B proportional to $f^*_{Z_{m,n}}$ in (5.3), given $(\mathbf{X},\mathbf{Y})$.

This leads to the following kernel estimator of $f^*_Z/(I_F+I_G)$ at $x\in B$:

$$\hat{f}_k(x) := \frac{1}{k}\left(\sum_{i=1}^{N^+}\delta_{V_i}-\sum_{i=1}^{N^-}\delta_{W_i}\right)\sigma_k^{-d}K_{x,\sigma_k}. \tag{5.4}$$

In the following we will work with this estimator based on an auxiliary sample of size k originating from the combined sampling scheme just described. The results that will be obtained concerning the quality of the approximation also apply to the simpler sampling and estimation scheme using two auxiliary samples of size k as described earlier in this section and in step 1 of the algorithm in Section 3. As already mentioned before, that procedure is, however, computationally more intensive.

To derive a CLT it is necessary to let k depend on the sample size n, so we will use the notation $k(n)$ in the future. The dependence of $\hat{f}_{k(n)}$ on the original sample $(\mathbf{X},\mathbf{Y})$ is suppressed for notational simplicity, but keep in mind that the auxiliary sample is drawn conditionally on $(\mathbf{X},\mathbf{Y})$.

As described in step 2 of the algorithm in Section 3, the estimator $\hat{f}_{k(n)}$ will be evaluated only on a finite evaluation set $T=T_n$. If the evaluation set T_n is a deterministic grid with too big a mesh size, however, $\arg\max_{T_n}|\hat{f}_{k(n)}|$ may not get close enough to t_0. To overcome this we construct the elements of our search set by

$$\hat{t} = \arg\max_{T_n}\left|\hat{f}_{k(n)}\right| + U_{R_n}, \tag{5.5}$$

where U_{R_n} is an independent random variable uniformly distributed on the ball $B_{R_n}(0)$. Adding a small uniform random variable may not be necessary if T_n is obtained in a random way, but we will pursue a treatment as general as possible here. The search set of halfspaces $\{\{\hat{t}_1\}^*,\dots,\{\hat{t}_{i_n}\}^*\}$ then consists of i_n independent copies of $\{\hat{t}\}^*$ from (5.5), conditional on $(\mathbf{X},\mathbf{Y})$), that is, each copy is computed from a different auxiliary sample in dual space.

THEOREM 5.2. *Let the search set of halfspaces* $S_n=\{\{\hat{t}_1\}^*,\dots,\{\hat{t}_{i_n}\}^*\}$ *be generated as described above, with the evaluation set* T_n *putting at least one point within* $O(r_n)$ *of* t_0 *with probability tending to* 1, *where* r_n *is specified below,* $k(n)\to\infty$ *and* $K\in\mathscr{K}_p$ *such that* $K_{x,\sigma}$ *has polynomial discrimination and is uniformly bounded. Under the conditions of Theorem* 4.1(b) *with* f^*_F *and* f^*_G p *times* $(p\ge 2)$ *continuously differentiable in a neighborhood of* t_0 *and the Hessian* H *in* (5.1) *being nonsingular, a sufficient condition for the CLT* (4.2) *to hold under* $F^m\otimes G^n\otimes(\mu(\mathbf{X},\mathbf{Y},k(n)))^{i_n}$ *is the existence of a* *positive sequence* $r_n \to 0$ *with*

$$(5.6)\qquad \begin{aligned} r_n^{2d/p+4} &\gg \frac{\log k(n)}{k(n)}, \\ n^{-1/2+\varepsilon} \le r_n^2 \textit{ for some } \varepsilon > 0 &\textit{ and } i_n \left(\frac{n^{-1/4}}{r_n} \right)^d \to \infty. \end{aligned}$$

Then R_n *in* (5.5) *can be chosen according to*

$$r_n \ll R_n \le r_n \left(i_n \left(\frac{n^{-1/4}}{r_n} \right)^d \right)^{1/(3d)} .$$

One sees that, unlike in the case of a random search in Theorem 5.1, the smoothness of the dual densities reflects on the size of the search sample (at least in terms of the sufficient conditions of Theorem 5.2; we did not prove necessity of those conditions): a large p allows for r_n shrinking to 0 faster, so i_n can increase more slowly.

5.3. *A comparison of the two search schemes.* Sections 5.1 and 5.2 give a way to compare the performance of the uniform random search with the search scheme using a pilot estimate in dual space. We will look at three different versions of the latter scheme, using different sizes for the auxiliary samples and different evaluation methods for the pilot estimate. All three are set up in such a way that the computational burden of computing the pilot estimate and evaluating the halfspace distance at its dual is of the same order as that of the halfspace evaluation in one direction for the uniform random search. So the complete search procedure of computing the search set $\{\{\hat{t}_1\}^*, \dots, \{\hat{t}_{i_n}\}^*\}$ and evaluating the halfspace distance on this search set entails a computational burden that is of the same order as performing a uniform random search in i_n directions.

Computing the one-dimensional Kolmogorov–Smirnov statistic of the $m + n$ projected data requires sorting, which cannot be done faster than in $O(n \log n)$ steps [see, e.g., Knuth (1973), Chapter 5]. Projecting the data takes $O(n)$ steps, so the overall burden for the uniform random search is $O(n \log n)$ for each of the i_n directions.

First, let $k(n) = n^{1/2}$ and compute $|\hat{f}_{k(n)}|$ at the location of the auxiliary sample only. Evaluating the density estimate at one point takes $O(k(n))$ steps, so the overall burden is $O(k(n)^2) = O(n)$, which also includes generating the auxiliary sample according to the algorithm given in Section 5.2 [burden $O(k(n))$], finding the maximum of $|\hat{f}_{k(n)}|$ at the locations of the auxiliary sample $[O(k(n))]$ and evaluating F_m and G_n at the dual of the computed pilot estimate $[O(n)]$. So the computational burden for each of the i_n elements in the search set is indeed of at most the same order as that of the uniform random search in each of the i_n directions. Shifting the data as described in Section 3 introduces only a constant factor. If we choose $r_n \gg n^{-1/2d}$, one readily checks that, with probability tending to 1, at least one point of the auxiliary sample falls into $B_{r_n}(t_0)$. Hence, by Theorem 5.2 the

CLT (4.2) will hold if i_n satisfies $i_n n^{-(d-2)/4} \to \infty$. So, compared with the uniform sampling of Theorem 5.1, the computational complexity is reduced by one dimension. Condition (5.6) is satisfied for this choice of r_n if $1/p + 2/d < 1/2$ and $-1/2 + \varepsilon < -1/d$, which necessitates $p > 2$ and $d > 4$.

The next scheme is set up to achieve as high a dimension reduction as possible. Let $k(n) = n \log n$ and let T_n be a grid on B (which preferably should be a hypercube here) with mesh size $r_n = n^{-1/d}$ in each direction. Evaluating the densities at the $O(n)$ points of T_n with the fast Fourier transform takes $O(n \log^d(n^{1/d}))$ steps, so the overall burden including generating the auxiliary sample and finding the maximum on T_n is up to log-terms comparable to $O(n \log n)$. Condition (5.6) becomes $i_n n^{-(d-4)/4} \to \infty$ and $2/p + 4/d < 1$, $-1/2 + \varepsilon < -2/d$, which necessitates $p > 2$, $d > 4$. So the complexity is reduced by up to three dimensions.

By choosing r_n larger one can relax the conditions on d and p at the cost of saving less computing time. Letting $k(n) = n$ and T_n be a grid of mesh size $r_n = n^{-1/2d}$ in each direction gives a burden of $O(n)$, which is caused by the generation of the auxiliary sample. Evaluating the densities on T_n by the fast Fourier transform takes only $O(n^{1/2} \log^d(n^{1/2d}))$ steps. By choosing r_n this large we need to satisfy $i_n n^{-(d-2)/4} \to \infty$, $1/p + 2/d < 1$ and $-1/2 + \varepsilon < -1/d$, which now necessitates $p > 1$, $d > 2$.

One also sees that in the case $p = 2$ one may attain computational savings of almost three dimensions for large d by choosing r_n slightly larger than $n^{-1/d}$.

Another informative way to compare the two approximation schemes is to compute the probability of getting close to the empirical halfspace distance. For this purpose we need not impose any smoothness or other assumptions on the underlying distributions F and G. Also, we will use a very simple implementation of the estimate scheme in dual space to see how this case compares with the uniform random search. We take a simple uniform kernel on the unit ball to estimate the dual densities, $K(\cdot) = (1/|B_1(0)|)1_{B_1(0)}(\cdot)$, with a bandwidth $\sigma_{k(n)}$ shrinking to 0 slowly enough: $\sigma^d_{k(n)} \alpha^2_{k(n)} \gg \log k(n)/k(n)$ for a bounded sequence $\{\alpha_n\}$. The evaluation set T_n is taken to be the auxiliary sample. It is not necessary here to add a uniform random variable to $\arg\max_{T_n} |\hat{f}_{k(n)}|$ in (5.5).

We are interested in the event $\sup_{H \in S_n} |F_m(H) - G_n(H)| > d(F_m, G_n) - \varepsilon$, where $\varepsilon > 0$ and S_n is the respective search set.

The next proposition compares the two search schemes with regard to this criterion:

PROPOSITION 5.3. *Let $n/m \to \lambda \geq 0$, $n \to \infty$, $k(n) \to \infty$. Then under the uniform random search scheme μ_{S_n} using i_n directions as described in Section 5.1, given* $(\mathbf{X}, \mathbf{Y})$,

$$\mu_{S_n}\Big(\sup_{H \in S_n} |F_m(H) - G_n(H)| > d(F_m, G_n) - \varepsilon\Big) = 1 - (1 - b_{m,n}(\varepsilon))^{i_n},$$

$$\varepsilon > 0,$$

where

$$b_{m,n}(\varepsilon) = \mu^{d-1}\Bigl(\Bigl\{s \in S^{d-1}: \sup_{t\in\mathbf{R}} |F_m(A(s,t)) - G_n(A(s,t))| > d(F_m, G_n) - \varepsilon\Bigr\}\Bigr),$$

$A(s,t) = \{x: \langle x, s\rangle \le t\}$, μ^{d-1} *denotes normalized* $(d-1)$*-dimensional Hausdorff measure on the unit sphere* S^{d-1} *in* $\mathbf{R}^d$, *and*

$$\limsup_{n\to\infty} b_{m,n}(\varepsilon) \le \mu^{d-1}\Bigl(\Bigl\{s \in S^{d-1}: \sup_{t\in\mathbf{R}} |F(A(s,t)) - G(A(s,t))| \ge d(F,G) - \varepsilon\Bigr\}\Bigr) \quad (F^m \otimes G^n)\text{-}a.s.$$

In the case where the search set S_n *is obtained by computing* i_n *copies of the pilot estimate in dual space under the search scheme* $\mu(\mathbf{X},\mathbf{Y},k(n))$ *as described in Section* 5.2 *and above, given* $(\mathbf{X},\mathbf{Y})$,

$$(\mu(\mathbf{X},\mathbf{Y},k(n)))^{i_n}\Bigl(\sup_{H\in S_n} |F_m(H) - G_n(H)| > d(F_m, G_n) - \varepsilon\Bigr) = 1 - (1 - p_{m,n}(\varepsilon))^{i_n}, \qquad \varepsilon > 0,$$

where

$$p_{m,n}(\varepsilon) = \mu(\mathbf{X},\mathbf{Y},k(n))\bigl(|(F_m - G_n)(\{\hat{t}\}^*)| > d(F_m, G_n) - \varepsilon\bigr) \to 1, \qquad n \to \infty, \tag{5.7}$$

$$(F^m \otimes G^n)\text{-}a.s.\ for\ all\ \varepsilon > 0.$$

One sees that the probability of approximating the empirical halfspace distance up to a prescribed error with a given number of trials is much higher with the second scheme if the sample size n is large. This is exactly the situation where one is in need of a computationally efficient method.

6. A simulation study. This section presents a small simulation study to assess the behavior of the two search schemes in a computing environment commonly in use at the time this article is being written.

We will consider situations in two and three dimensions. In $\mathbf{R}^2$, samples are drawn from the distributions

$$F = (1/2)(N((-1,-2), I) + N((-1,2), I))$$

and

$$G = (1/2)(N((1,-2), I) + N((1,2), I)).$$

One then verifies that $d(F,G) = 2\Phi(1) - 1 = 0.682\ldots$ with the optimally separating hyperplane given by $x_1 = 0$. In $\mathbf{R}^3$ we will use

$$F = (1/2)(N((-1,-2,0), I) + N((-1,2,0), I))$$

and

$$G = (1/2)(N((1,-2,0),I) + N((1,2,0),I)).$$

In this situation one also has $d(F,G) = 2\Phi(1) - 1 = 0.682\dots$ and the optimally separating hyperplane satisfies $x_1 = 0$. Using mixture distributions complicates the task of finding an approximation to the halfspace distance, because a halfspace H separating away a component of the mixture gives also a high value of $|F(H) - G(H)|$ and may detract search schemes from finding the correct separating hyperplane.

We will use equal sample sizes $n = m$ and $k(n) = n^{-1/2}$ for the auxiliary sample in dual space, where we restrict ourselves to the compact set $B = B_2(0)$ and employ the shifting mechanism given in Section 3. A Gaussian kernel is used to estimate the dual density, using the points of the auxiliary sample as evaluation set. We let the bandwidth shrink with a rate of $k(n)^{-1/(d+4)}$, as recommended in Silverman (1986). Once a pilot estimate $\hat{t}$ is found in dual space, $|F_m - G_n|$ is evaluated not only at the halfspace $\{\hat{t}\}^*$, but the Kolmogorov–Smirnov statistic is computed for the projected data along the direction $\hat{t}/|\hat{t}|$. This version of the search scheme allows reuse of the code for the uniform random search. The sorting algorithm employed there is the $O(n \log n)$ algorithm M01CAF from the NAG library of FORTRAN subroutines. The random number generators are also taken from there.

Here are the explicit formulae for computing $|B \cap \{X\}^*|$ in the case $d = 2$ and $d = 3$, as required in the algorithm that generates the auxiliary sample in Sections 3 and 5.2. Using Fubini's theorem one computes the volume $|B \cap \{X\}^*|$ in the two-dimensional case with $B = B_r(0)$ as

$$|B \cap \{X\}^*| = \begin{cases} r^2\pi, & \text{if } |X| \le \dfrac{1}{r}, \\ r^2\left(\dfrac{\pi}{2} + \dfrac{1}{|X|}\sqrt{1 - \dfrac{1}{r^2|X|^2}} + \arcsin\dfrac{1}{r|X|}\right), & \text{otherwise,} \end{cases}$$

and in the three-dimensional case as

$$|B \cap \{X\}^*| = \begin{cases} \dfrac{4}{3}r^3\pi, & \text{if } |X| \le \dfrac{1}{r}, \\ \pi\left(\dfrac{2}{3}r^3 + \dfrac{r^2}{|X|} - \dfrac{1}{3|X|^3}\right), & \text{otherwise.} \end{cases}$$

After samples of size n are generated from F and G, a uniform random search is run with 30 random directions. Then 30 pilot estimates are computed via Monte Carlo sampling in dual space and the samples are evaluated in the resulting directions. This whole process, including the generation of samples from F and G, is repeated for 1000 runs. For each of the two search schemes, the 30 evaluations in the directions given by the respective search sets, $ev_k = \sup_{H \in \dots}|F_m(H) - G_n(H)|$, $k = 1,\dots,30$, yield the estimates of the halfspace distance $\hat{d}_{i_n} = \max_{k=1,\dots,i_n} ev_k$, $i_n = 1,\dots,30$, where i_n de-

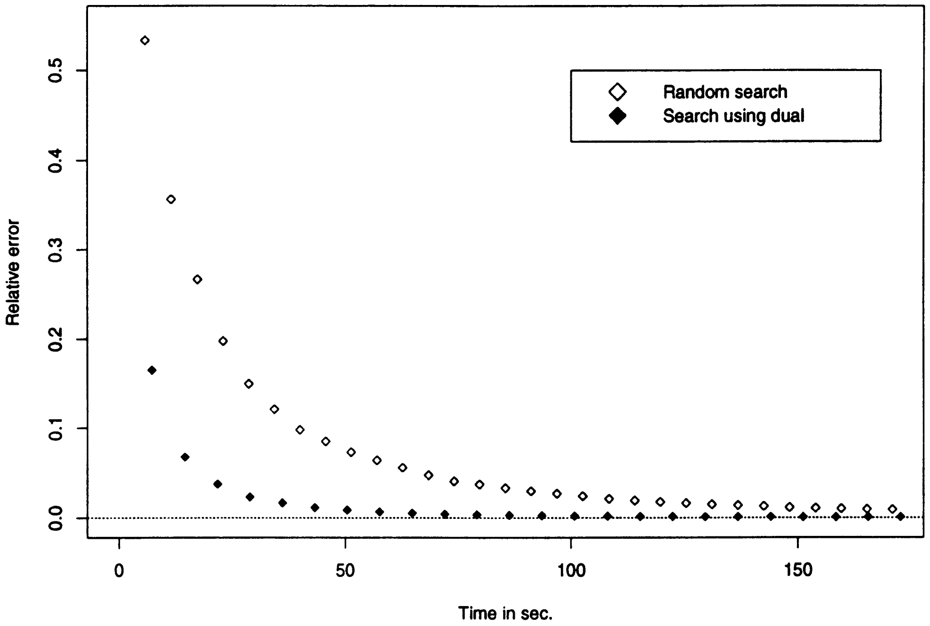

FIG. 4. *Average relative error versus time* ($d = 2$, $n = 100{,}000$).

notes the size of the search set. The relative errors

$$\left(\hat{d}_{i_n} - (2\Phi(1) - 1)\right)/(2\Phi(1) - 1), \qquad i_n = 1,\dots,30,$$

were recorded. The averages of each of these 30 relative errors, obtained over 1000 runs, together with their respective computing times are plotted in Figures 4 and 5. All computing times refer to a SUN SPARCstation 1 + workstation.

The first example in Figure 4 treats the case $d = 2$ and $n = 100{,}000$. Each of the i_n evaluations took about 5.7 seconds for the uniform random search and 7.2 seconds for the search in dual space and the evaluation in the direction found.

For the second case $d = 3$, $n = 100{,}000$ in Figure 5, the respective times are 5.9 seconds and 8.1 seconds.

The examples show that the approximations using a pilot estimate in dual space converge much faster to the halfspace distance than those obtained by a uniform random search.

7. Proofs.

PROOF OF THEOREM 2.3. One checks that all relevant maps are measurable. Using the duality relation (3.1) and Fubini, one obtains the following for

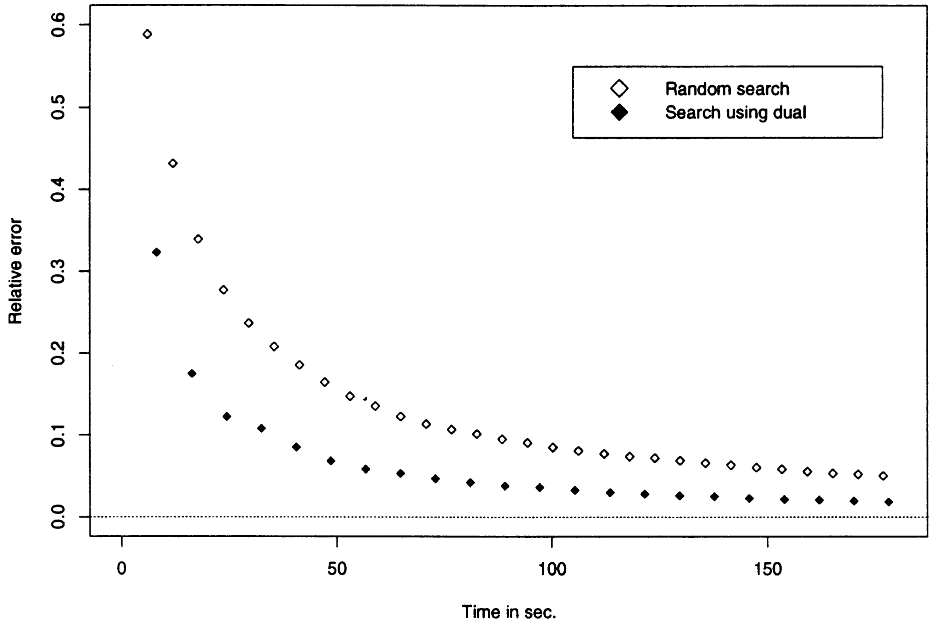

FIG. 5. *Average relative error versus time* ($d = 3$, $n = 100{,}000$).

a measurable set $A \subset \mathbf{R}^d$:

$$\begin{aligned} F^* |_B (A) &= \int_{A \cap B} F\{a\}^* \, da = \int_{A \cap B} \int 1(x \in \{a\}^*) \, dF(x) \, da \\ &= \int \int_{A \cap B} 1(a \in \{x\}^*) \, da \, dF(x) \\ &= c_B \int \frac{|B \cap \{x\}^* \cap A|}{|B \cap \{x\}^*|} \frac{|B \cap \{x\}^*|}{c_B} \, dF(x). \qquad \square \end{aligned}$$

PROOF OF LEMMA 3.1. $\{a\}^*$ is a halfspace if $a \neq 0$, so the set on the LHS consists only of hyperplanes. Conversely, let $H = \{x: \langle x, u \rangle = t\}$ for real t and a unit vector u be a given hyperplane. If $|t| \geq 1/r$, set $a = u/t \in B_r(0) \setminus \{0\}$. Then $\partial(\{a\}^* - ce_0) = \partial\{a\}^* = \partial\{x: \langle x, u/t \rangle \leq 1\} = H$.

In the case $|t| < 1/r$ observe that $|\langle u, e_j \rangle| = |u_j| \geq 1/\sqrt{d}$ for some $j \in \{1, \dots, d\}$, as $\sum_{i=1}^d u_i^2 = 1$. Hence $r|c\langle u, e_j \rangle + t| \geq 1$ and thus

$$a := u/\big(c\langle u, e_j \rangle + t\big) \in B_r(0) \setminus \{0\}.$$

Further,

$$\begin{aligned}\partial\big(\{a\}^* - ce_j\big) &= \partial\left\{x\colon \left\langle x, \frac{u}{c\langle u, e_j\rangle + t}\right\rangle \le 1 - \left\langle ce_j, \frac{u}{c\langle u, e_j\rangle + t}\right\rangle\right\}\\ &= \partial\left\{x\colon \langle x, u\rangle \frac{1}{c\langle u, e_j\rangle + t} \le t\frac{1}{c\langle u, e_j\rangle + t}\right\}\\ &= H. \qquad \square\end{aligned}$$

Proof of Theorem 4.1. Part (a) can be proved similarly to Proposition 1 in Beran and Millar (1986). For later use we rewrite the stochastic equicontinuity property (4.2) in Beran and Millar (1986) for the present setting: for every $\varepsilon > 0$ and $\eta > 0$ there exists $\gamma > 0$ such that

$$\begin{aligned}&\limsup_{n\to\infty} \mathbb{P}\Big(\sup_{G(\gamma)} \Big|\sqrt{n}\,\big(f^*_{Z_{m,n}}(t) - f^*_Z(t)\big)\\ &\qquad\qquad - \sqrt{n}\,\big(f^*_{Z_{m,n}}(t') - f^*_Z(t')\big)\Big| > \eta\Big) < \varepsilon,\end{aligned} \tag{7.1}$$

where $G(\gamma) = \{(t, t')\colon (F + G)(\{x\colon \langle x, t\rangle \le 1\} \vartriangle \{x\colon \langle x, t'\rangle \le 1\}) < \gamma\}$.

We will also use a theorem of Wichura (1970) to the effect that there exist representations $\widetilde{f^*_{Z_{m,n}}}$ and $\tilde{W}$ with $\mathscr{L}(\widetilde{f^*_{Z_{m,n}}}) = \mathscr{L}(f^*_{Z_{m,n}})$, $\mathscr{L}(\tilde{W}) = \mathscr{L}(W)$ and $\|\sqrt{n}\,(\widetilde{f^*_{Z_{m,n}}} - f^*_Z) - \tilde{W}\|_\infty \to 0$ a.s. For part (b) note that assumption (i) implies $|f^*_Z(t_0)| > 0$ and thus $t_0 \neq 0$. Also

$$\varepsilon_n := \|f^*_{Z_{m,n}} - f^*_Z\|_\infty \to 0 \quad \text{a.s., } n \to \infty, \tag{7.2}$$

by the Glivenko–Cantelli theorem for halfspaces.

We will deduce (b) from the general part (c). To see the problem arising in the general case, observe that

$$\|f^*_{Z_{m,n}}\|_\infty = \sup_{A_{2\varepsilon_n}} |f^*_{Z_{m,n}}| \tag{7.3}$$

by (i) and because $\|\,|f^*_{Z_{m,n}}| - |f^*_Z|\,\|_\infty \le \varepsilon_n$, so our analysis can be localized to $A_{2\varepsilon_n}$. In the general case, however, the stochastic equicontinuity property (7.1) is not applicable at t_0 with respect to the neighborhood $A_{2\varepsilon_n}$: $|f^*_Z|$ may be continuous at t_0 while f^*_Z changes sign due to an atom of F or G at $t_0/|t_0|^2$; or mass at that point may cancel in $F - G$ while the discontinuities in the limits of the sample paths of $\sqrt{n}\,(f^*_{F_m} - f^*_F)$ and $\sqrt{n}\,(f^*_{G_n} - f^*_G)$ do not cancel, but add up.

In the general case (c) observe that when $\varepsilon_n < \|f_Z^*\|_\infty/4$ then

$$\begin{aligned} T\big(f_Z^*; f^*_{Z_{m,n}} - f_Z^*\big) &\le \lim_{\varepsilon\downarrow 0}\Big(\sup_{A_\varepsilon}\big(\operatorname{sign} f_Z^*(t) f^*_{Z_{m,n}}(t)\big) - \inf_{A_\varepsilon}\big|f_Z^*(t)\big|\Big)\\ &\le \lim_{\varepsilon\downarrow 0}\Big(\sup_{A_\varepsilon}\big|f^*_{Z_{m,n}}(t)\big| - \|f_Z^*\|_\infty + \varepsilon\Big)\\ &\le T\big(f^*_{Z_{m,n}}\big) - T\big(f_Z^*\big)\\ &= \sup_{A_{2\varepsilon_n}}|f^*_{Z_{m,n}}| - \|f_Z^*\|_\infty \quad [\text{by } (7.3)]\\ &\le \sup_{A_{2\varepsilon_n}}\big(\big|f^*_{Z_{m,n}}(t)\big| - \big|f_Z^*(t)\big|\big)\\ &= \sup_{A_{2\varepsilon_n}}\big(\operatorname{sign} f_Z^*(t)\big(f^*_{Z_{m,n}}(t) - f_Z^*(t)\big)\big). \end{aligned} \tag{7.4}$$

Switch to the representation $(\widetilde{f^*_{Z_{m,n}}}, \tilde{W})$ and set

$$C_n = \sqrt{n}\ \operatorname{sign} f_Z^*(t)\Big(\widetilde{f^*_{Z_{m,n}}}(t) - f_Z^*(t)\Big)$$

and

$$R_n = \sup_{A_{2\varepsilon_n}} C_n - \lim_{\varepsilon\downarrow 0}\sup_{A_\varepsilon} C_n.$$

Then (7.2) and the fact that the convergence of $\sup_{A_\varepsilon} C_n$ is uniform in ε give

$$\limsup_{n\to\infty}\ \sup_{A_{2\varepsilon_n}} C_n \le \lim_{\varepsilon\downarrow 0}\sup_{A_\varepsilon}\big(\operatorname{sign} f_Z^*(t)\tilde{W}\big) = \lim_{n\to\infty}\lim_{\varepsilon\downarrow 0}\sup_{A_\varepsilon} C_n \quad \text{a.s.} \tag{7.5}$$

So $R_n = o_p(1)$ as $R_n \ge 0$, and therefore (7.4) yields

$$T\big(f^*_{Z_{m,n}}\big) - T\big(f_Z^*\big) = T\big(f_Z^*; f^*_{Z_{m,n}} - f_Z^*\big) + o_p\big(n^{-1/2}\big). \tag{7.6}$$

The CLT for the empirical halfspace distance is a consequence of (7.6) and (7.5).

Using the fact that $\mathscr{L}(\|W\|_\infty)$ has a density [see, e.g., Beran and Millar (1986), Proposition 2], one concludes $(\sqrt{n}\,\|f^*_{Z_{m,n}} - f_Z^*\|_\infty)^{-1} = O_p(1)$. Together with (7.6) this proves (4.1) in the general case.

The assertions for the smooth case will follow once it is shown that

$$\begin{aligned} &\lim_{\varepsilon\downarrow 0}\sup_{A_\varepsilon}\Big(\operatorname{sign} f_Z^*(t)\sqrt{n}\,\big(f^*_{Z_{m,n}}(t) - f_Z^*(t)\big)\Big)\\ &\qquad - \operatorname{sign} f_Z^*(t_0)\sqrt{n}\,\big(f^*_{Z_{m,n}}(t_0) - f_Z^*(t_0)\big) = o_p(1). \end{aligned} \tag{7.7}$$

As a consequence of (i), there exists a function $r(\cdot)$ with

$$\lim_{\varepsilon\downarrow 0} r(\varepsilon) = 0 \quad \text{and} \quad A_\varepsilon \subset B_{r(\varepsilon)}(t_0). \tag{7.8}$$

Together with $f_Z^*(t_0) \ne 0$ and the continuity of f_Z^* at t_0, which follows from (ii), this allows us to eliminate sign $f_Z^*(t)$ and sign $f_Z^*(t_0)$ in assertion (7.7). Then (7.7) will follow from (7.1) upon verify that, for given $\gamma > 0$,

$$B_r(t_0) \subset \big\{t\colon (F+G)\big(\{x\colon \langle x, t\rangle \le 1\} \vartriangle \{x\colon \langle x, t_0\rangle \le 1\}\big) < \gamma\big\}$$

if r is small enough. This inclusion is a consequence of (ii). □

PROOF OF PROPOSITION 4.2. The last inclusion in the previous proof and (7.1) show that the continuity assumption (ii) allows us to write a stochastic equicontinuity condition at t_0 for the process $W_{m,n}(t) = \sqrt{n}(f^*_{Z_{m,n}}(t) - f^*_Z(t))$:

For every $\varepsilon > 0$ and $\eta > 0$ there exists a Euclidean neighborhood U of t_0 such that

$$\limsup_{n\to\infty} \mathbb{P}\Big(\sup_{t\in U} |W_{m,n}(t) - W_{m,n}(t_0)| > \eta\Big) < \varepsilon.$$

Therefore

$$W_{m,n}(t_n) - W_{m,n}(t_0) \to 0 \quad \text{in probability (as } n \to \infty) \tag{7.9}$$

for every sequence of random variables $\{t_n\}$ converging to t_0 in probability, as is readily verified or looked up in Pollard [(1984), page 140].

For the necessity part of the proposition, let

$$t_n = \arg\max_{t\in\{H^*:\, H\in S_n\}} f^*_{Z_{m,n}}(t)$$

and suppose first

$$t_n \to t_0 \quad \text{in } \big(F^m \otimes G^n \otimes \mu_{S_n}\big)\text{-probability} \tag{7.10}$$

is not valid. Then, along a subsequence n_s,

$$\Big(F^{m(n_s)} \otimes G^{n_s} \otimes \mu_{S_{n_s}}\Big)\Big(\big\{t_{n_s}\colon |t_{n_s} - t_0| > r\big\}\Big) > \eta \quad \text{for some } r, \eta > 0.$$

Together with $\sup_{t\in B^c_r(t_0)} \sqrt{n}(f^*_{Z_{m,n}}(t) - f^*_Z(t_0)) \to -\infty$ a.s., which is a consequence of (7.2) and (7.8), one sees that along n_s, that LHS in (4.4) converges to $-\infty$ with $(F^{m(n_s)} \otimes G^{n_s} \otimes \mu_{S_{n_s}})$-probability at least η, contradicting the CLT (4.4). So (7.10) holds. Now write

$$\begin{aligned}\sqrt{n}\Big(\sup_{t\in\{H^*:\, H\in S_n\}} f^*_{Z_{m,n}}(t) - f^*_Z(t_0)\Big) &= \sqrt{n}\big(f^*_{Z_{m,n}}(t_n) - f^*_Z(t_n)\big)\\ &\quad + \sqrt{n}\big(f^*_Z(t_n) - f^*_Z(t_0)\big)\end{aligned} \tag{7.11}$$

and suppose

$$\begin{aligned}\limsup_n \big(F^m \otimes G^n \otimes \mu_{S_n}\big)\Big(\big\{t_n\colon \sqrt{n}\big(f^*_Z(t_n) - f^*_Z(t_0)\big) < -\varepsilon\big\}\Big)&\\ > 0 \quad \text{for some } \varepsilon > 0.&\end{aligned} \tag{7.12}$$

The distribution of the first term in (7.11) converges to the law of $W(t_0)$ by (7.9), (7.10) and Theorem 4.1(a). The second term in (7.11) is nonpositive, so (7.12) forbids that the sum of the two terms converges in distribution to $W(t_0)$, contradicting the CLT (4.4) and therefore

$$\lim_{n\to\infty} \big(F^m \otimes G^n \otimes \mu_{S_n}\big)\Big(\big\{t_n\colon \sqrt{n}\big(f^*_Z(t_n) - f^*_Z(t_0)\big) < -\varepsilon\big\}\Big) = 0 \quad \text{for all } \varepsilon > 0.$$

One readily checks that one can replace ε in the above equation with some sequence $\varepsilon_n \downarrow 0$, so

$$\lim_{n\to\infty} \big(F^m \otimes G^n \otimes \mu_{S_n}\big)\big(\{t_n: f_Z^*(t_n) \geq f_Z^*(t_0) - l_n\}\big) = 1 \qquad \text{for some } l_n = o(n^{-1/2})$$

and condition (4.6) follows.

Conversely, if (4.6) holds, let $t_n' = \arg\max_{t\in\{H^*:\, H\in S_n\}} f_Z^*(t)$ and observe

$$\begin{aligned}
&\mu_{S_n}\big(0 \geq \sqrt{n}\,(f_Z^*(t_n') - f_Z^*(t_0)) \geq -\sqrt{n}\,l_n\big)\\
&\quad = \mu_{S_n}\big(t_n' \in A_{l_n}\big)\\
&\quad = \mu_{S_n}\big(A_{l_n} \cap \{H^*: H \in S_n\} \neq \varnothing\big) \to 1.
\end{aligned} \tag{7.13}$$

It follows from (7.8) that $\mu_{S_n}(t_n' \in B_{r(l_n)}(t_0)) \geq \mu_{S_n}(t_n' \in A_{l_n}) \to 1$. Hence

$$t_n' \to t_0 \quad \text{in } \mu_{S_n}\text{-probability.} \tag{7.14}$$

Write

$$\begin{aligned}
&\sqrt{n}\Big(\sup_t f_{Z_{m,n}}^*(t) - f_Z^*(t_0)\Big)\\
&\quad \geq \sqrt{n}\Big(\sup_{t\in\{H^*:\, H\in S_n\}} f_{Z_{m,n}}^*(t) - f_Z^*(t_0)\Big)\\
&\quad \geq \sqrt{n}\,\big(f_{Z_{m,n}}^*(t_n') - f_Z^*(t_n')\big) + \sqrt{n}\,\big(f_Z^*(t_n') - f_Z^*(t_0)\big).
\end{aligned}$$

The term in the first line converges in distribution to $W(t_0)$ by Theorem 4.1(b) and (4.3). The sum in the third line converges also in distribution to $W(t_0)$ by (7.9), (7.14), Theorem 4.1(a) and (7.13), which asserts convergence to 0 in probability of the second term. Hence the term in the second line must also converge in distribution to $W(t_0)$ and the CLT (4.2) as stated in (4.4) follows. □

PROOF OF THEOREM 5.1. For the purpose of evaluating $|F(\cdot) - G(\cdot)|$ the search set S_n is equivalent to

$$\big\{H: H = \{x: \langle x, e\rangle \leq t\},\, t \geq 0,\, e \in \{e_1, \dots, e_{i_n}, -e_1, \dots, -e_{i_n}\}\big\}.$$

The set of duals of these halfspaces is given by $\text{span}\{e_1\} \cup \cdots \cup \text{span}\{e_{i_n}\}$, disregarding the special case of halfspaces with 0 on their boundaries and all of $\mathbf{R}^d$, which do not play a role in the treatment of the CLT.

Reparametrize so that $(1/2)H = -I$, where I denotes the identity matrix. Then by (5.1) there exist positive constants c_1 and c_2 such that $B_{c_1\sqrt{\varepsilon}}(t_0) \subset A_\varepsilon \subset B_{c_2\sqrt{\varepsilon}}$ for small ε. Proposition 4.2 yields as necessary and sufficient condition for the validity of the CLT (4.2) the existence of a positive sequence $l_n = o(n^{-1/4})$ with

$$\mu_{S_n}\Big(B_{l_n}(t_0) \cap \big(\text{span}\{e_1\} \cup \cdots \cup \text{span}\{e_{i_n}\}\big) \neq \varnothing\Big) \to 1.$$

The condition $B_{l_n}(t_0) \cap \operatorname{span}\{e_i\} \neq \varnothing$ is equivalent to e_i belonging to a spherical cap on S^{d-1} whose $(d-1)$-dimensional volume v_n can be shown to satisfy $v_n \asymp l_n^{d-1}$. Hence

$$\begin{aligned}
&\mu_{S_n}\Big(B_{l_n}(t_0) \cap \big(\operatorname{span}\{e_1\} \cup \cdots \cup \operatorname{span}\{e_{i_n}\}\big) \neq \varnothing\Big)\\
&\quad = 1 - (1 - p_n)^{i_n} \quad \big(\text{where } p_n = v_n/|S^{d-1}| \asymp l_n^{d-1}\big)\\
&\quad = 1 - \exp(i_n \log(1 - p_n))\\
&\quad \to 1 \quad \big[\text{iff } i_n \log(1-p_n) \to -\infty\big].
\end{aligned} \tag{7.15}$$

Let $i_n n^{-(d-1)/4} \to \infty$ and set $l_n = (\sqrt{i_n^{-1} n^{(d-1)/4}}\, n^{-(d-1)/4})^{1/(d-1)}$. Then $l_n = o(n^{-1/4})$ and

$$i_n \log(1 - p_n) \leq -i_n p_n \to -\infty \quad \text{as } p_n \asymp \sqrt{i_n^{-1} n^{(d-1)/4}}\, n^{-(d-1)/4}.$$

Conversely, assume (7.15) holds with $p_n = o(n^{-(d-1)/4})$. As $\log(1-x) > -2x$ for small positive x, we get $i_n \log(1-p_n) > -2 i_n p_n$ and hence $i_n p_n \to \infty$, which implies $i_n n^{-(d-1)/4} \to \infty$. □

Proof of Theorem 5.2. The important ideas and statements of the proof will be given. For the details see Walther (1994).

As $K \in \mathscr{K}_p$ and f_F^* is p times continuously differentiable in a neighborhood of t_0, a Taylor expansion shows that there exists a neighborhood N of t_0 where $\sup_N |f_F^{*(\sigma)} - f_F^*| = O(\sigma^p)$, $\sigma \to 0$, and the same holds for f_G^*. Together with Theorem 5.1 in Romano (1988) one can show that $\sigma_{k(n)}^d \alpha_{k(n)}^2 \gg \log k(n)/k(n)$ for $\{\alpha_n\}$ bounded and $k(n) \to \infty$, $n \to \infty$, implies that there exists a neighborhood $N(t_0)$ of t_0 such that

$$\begin{aligned}
&\sup_{x \in N(t_0) \cap B} \Big| \big|\hat{f}_{k(n)}(x)\big| - |f_Z^*(x)|/(I_F + I_G)\Big|\\
&\qquad \ll \max\Big(\alpha_{k(n)}, n^{-1/2+\varepsilon}, k(n)^{-1/2+\varepsilon}, \sigma_{k(n)}^p\Big)
\end{aligned} \tag{7.16}$$

in $\mu(\mathbf{X}, \mathbf{Y}, k(n))$-probability $(F^m \otimes G^n)$-a.s. for all $\varepsilon > 0$, and if in addition $\sigma_{k(n)} \to 0$, then

$$\sup_{x \in (N(t_0))^C \cap B} \big|\hat{f}_{k(n)}(x)\big| < |f_Z^*(t_0)|/(I_F + I_G) - \delta \tag{7.17}$$

with $\mu(\mathbf{X}, \mathbf{Y}, k(n))$-probability tending to 1, $(F^m \otimes G^n)$-a.s. for some $\delta > 0$.

We are now in a position to establish a rate at which $\arg\max_{T_n} |\hat{f}_{k(n)}(x)|$ can converge to t_0 if the evaluation set T_n is chosen appropriately:

Proposition 7.1. *Under the assumptions of Theorem* 5.2, *if* $\sigma_{k(n)}^d \alpha_{k(n)}^2 \gg \log k(n)/k(n)$ *as* $k(n) \to \infty$, *then for all positive sequences* $\{r_n\}, \{R_n\}$ *with* $r_n = o(R_n)$, $r_n \to 0$ *and* $\max(\alpha_{k(n)}, n^{-1/2+\varepsilon}, k(n)^{-1/2+\varepsilon}, \sigma_{k(n)}^p) \leq r_n^2$ *eventu-*

ally, for some $\varepsilon > 0$,

$$\inf_{B_{r_n}(t_0)} \left|\hat{f}_{k(n)}\right| - \sup_{B^C_{R_n}(t_0)} \left|\hat{f}_{k(n)}\right| > 0$$

with $\mu(\mathbf{X}, \mathbf{Y}, k(n))$-probability tending to 1, $(F^m \otimes G^n)$-a.s.

PROOF. Reparametrize so that $(1/2)H = -I$ in (5.1), which then implies that $|f_Z^*(t_0)| - 2|t - t_0|^2 \le |f_Z^*(t)| \le |f_Z^*(t_0)| - \frac{1}{2}|t - t_0|^2$ for t in some neighborhood of t_0. Outside that neighborhood $|f_Z^*(t)| \le |f_Z^*(t_0)| - c$ for some $c > 0$ by assumption (i). Hence $\inf_{B_{r_n}(t_0)}|f_Z^*| - \sup_{B^C_{R_n}(t_0)}|f_Z^*| \ge -2r_n^2 + \min(c, \frac{1}{2}R_n^2) \ge 3r_n^2(I_F + I_G)$ for large n, and therefore

$$\begin{aligned}
&\inf_{B_{r_n}(t_0)} \left|\hat{f}_{k(n)}\right| - \sup_{B^C_{R_n}(t_0)} \left|\hat{f}_{k(n)}\right| \\
&\quad \ge \inf_{B_{r_n}(t_0)} |f_Z^*|/(I_F + I_G) \\
&\qquad - \max\left(\left|f_Z^*(t_0)\right|/(I_F + I_G) - \delta,\ \sup_{B^C_{R_n}(t_0)} |f_Z^*|/(I_F + I_G)\right) \\
&\qquad - 2\max\left(\alpha_{k(n)}, n^{-1/2+\varepsilon}, k(n)^{-1/2+\varepsilon}, \sigma^p_{k(n)}\right) \\
&\qquad\qquad \left[\text{with } \mu(\mathbf{X}, \mathbf{Y}, k(n))\text{-probability tending to 1, } (F^m \otimes G^n)\text{-a.s. by (7.16), (7.17) and because } B_{r_n}(t_0) \subset N(t_0) \text{ eventually}\right] \\
&\quad \ge \min\left(\inf_{B_{r_n}(t_0)} |f_Z^*| - |f_Z^*(t_0)| + \delta(I_F + I_G),\ \inf_{B_{r_n}(t_0)} |f_Z^*| - \sup_{B^C_{R_n}(t_0)} |f_Z^*|\right) \Big/ (I_F + I_G) - 2r_n^2 \\
&\qquad\qquad \text{(eventually)} \\
&\quad \ge 0 \quad \left[\text{eventually, because (ii) implies that } f_Z^* \text{ is continuous at } t_0\right]. \quad \square
\end{aligned}$$

To prove Theorem 5.2 now set $\sigma_{k(n)} = r_n^{2/p}$ and $\alpha_{k(n)} = r_n^2$. Then $r_n = o(R_n)$, $\sigma^d_{k(n)}\alpha^2_{k(n)} \gg \log k(n)/k(n)$ and $\max(\alpha_{k(n)}, n^{-1/2+\varepsilon}, k(n)^{-1/2+\varepsilon}, \sigma^p_{k(n)}) \le r_n^2$ eventually for some $\varepsilon > 0$ [so the condition on r_n in (5.6) is chosen to let r_n decrease to 0 as fast as possible while obeying the conditions of Proposition 7.1]. As the evaluation set T_n has one point within $O(r_n)$ of t_0 with probability tending to 1, Proposition 7.1 gives $\arg\max_{T_n}|\hat{f}_{k(n)}| \in B_{R_n}(t_0)$ with probability tending to 1. Using (5.5) this implies that $\hat{t} \in B_{r_n R_n^{-1} n^{-1/4}}(t_0)$ with probability at least $c(r_n n^{-1/4}/R_n^2)^d$ for some $c > 0$ if $k(n)$ is large enough. So the probability that the set $\{\hat{t}_1, \dots, \hat{t}_{i_n}\}$ hits $B_{r_n R_n^{-1} n^{-1/4}}(t_0)$ is at least $1 - (1 - c(r_n n^{-1/4}/R_n^2)^d)^{i_n}$, which converges to 1, because if R_n is chosen as noted in

Theorem 5.2, then

$$\begin{aligned} i_n \log\left(1 - c\left(\frac{r_n n^{-1/4}}{R_n^2}\right)^d\right) &\le -i_n c\left(\frac{r_n n^{-1/4}}{R_n^2}\right)^d \\ &\le -i_n c\left(\frac{n^{-1/4}}{r_n}\right)^d \left(i_n\left(\frac{n^{-1/4}}{r_n}\right)^d\right)^{-2/3} \\ &= -c\left(i_n\left(\frac{n^{-1/4}}{r_n}\right)^d\right)^{1/3} \to -\infty. \end{aligned}$$

The assertion of the theorem now follows from Proposition 4.2, using (5.1) as in the proof of Theorem 5.1. □

The following lemma will be needed for the proof of Proposition 5.3:

LEMMA 7.2. *Denote by $\mathscr{M}_s^d$ the set of probability measures on $\mathbf{R}^d$ that concentrate on a finite number of singletons. Then*:

(i) *The set $\{|f^*_{F-G}|\colon F, G \in \mathscr{M}_s^d\}$ is dense in $\{|f^*_{F-G}|\colon F, G \in \mathscr{M}^d\}$ in the $\|\cdot\|_\infty$-norm.*

(ii) *The following continuity property holds at each $a \in \mathbf{R}^d$: for every $\varepsilon > 0$, $\eta > 0$ there exists $r > 0$ such that*

$$\inf_{z \in B_\eta(a)} \sup_{x \in B_r(z)} |f_Z^*(a) - f_Z^*(x)| \le \varepsilon. \tag{7.18}$$

(iii) $$\limsup_{r \to 0,\, r > 0} \sup_{x \in A_\varepsilon^C} \frac{\int_{B_r(x)} |f_Z^*(t)|\, dt}{|B_r(x)|} < \|f_Z^*\|_\infty \quad \textit{for all } \varepsilon > 0.$$

PROOF. Part (i) can be deduced from large sample theory: as a consequence of the Glivenko–Cantelli theorem for halfspaces, $\{|f_F^*|\colon F \in \mathscr{M}_s^d\}$ is dense in $\{|f_F^*|\colon F \in \mathscr{M}^d\}$. As for (ii), note that for fixed $\delta > 0$ the map $\mathbf{K}$: $a \mapsto \mathbb{P}(\langle a, X\rangle < 1 + \delta)$ is lower semicontinuous by the Portmanteau theorem. Hence there exists a ball centered at a where $\mathbf{K}(\cdot) \ge \mathbf{K}(a) - \varepsilon$ and therefore on some ball centered at $z = (1+\delta)^{-1}a$: $f_F^*(\cdot) \ge \mathbf{K}((1+\delta)\cdot) \ge \mathbf{K}(a) - \varepsilon \ge f_F^*(a) - \varepsilon$. Using Proposition 2.2, (ii) follows.

As for (iii), use part (i) of the lemma to find $F_s, G_s \in \mathscr{M}_s^d$ that satisfy $\| |f^*_{F-G}| - |f^*_{F_s - G_s}| \|_\infty < \varepsilon/4$. One checks that as F_s and G_s concentrate on a finite number of singletons, those give rise to a partition of $\mathbf{R}^d$ into a finite number of polyhedral sets such that $f^*_{F_s-G_s}$ is constant in the interior of each such set and the values of $f^*_{F_s-G_s}$ on the boundary coincide with those in the interior of one of the adjacent polyhedral sets. This implies the existence of constants $\alpha = \alpha(F_s, G_s) > 0$ and $r_0 = r_0(F_s, G_s)$ such that for all $x \in \mathbf{R}^d$ and $r \le r_0$ one has $|f^*_{F_s-G_s}(\cdot)| \le |f^*_{F_s-G_s}(x)|$ on a subset of $B_r(x)$ with Lebesgue measure at least $\alpha|B_r(x)|$. Thus

$$\sup_{x\in A_\varepsilon^C}\frac{\int_{B_r(x)}|f^*_{F-G}(t)|\,dt}{|B_r(x)|}$$
$$\le \sup_{x\in A_\varepsilon^C}\frac{(|f^*_{F-G}(x)|+\varepsilon/2)\alpha|B_r(x)|+\|f^*_{F-G}\|_\infty(1-\alpha)|B_r(x)|}{|B_r(x)|}$$
$$(r \text{ small enough})$$
$$\le \left(\|f^*_{F-G}\|_\infty-\varepsilon+\frac{\varepsilon}{2}\right)\alpha+\|f^*_{F-G}\|_\infty(1-\alpha)\quad(\text{by the definition of } A_\varepsilon)$$
$$=\|f^*_{F-G}\|_\infty-\frac{\varepsilon\alpha}{2}.\qquad\square$$

Proof of Proposition 5.3. A straightforward calculation gives the assertion for the uniform random search; see also Beran and Millar [(1986), (1.6)]. The inequality concerning $\limsup_{n\to\infty} b_{m,n}$ follows from Fatou's lemma and the Glivenko–Cantelli theorem for halfspaces.

As for the second scheme, note that

$$\sup_{x\in B}\left|\,|\hat{f}_{k(n)}(x)|-\frac{\left|\overline{f_Z^*}^{(\sigma_{k(n)})}(x)\right|}{(I_F+I_G)}\right|\to 0$$

in $\mu(\mathbf{X},\mathbf{Y},k(n))$-probability $(F^m\otimes G^n)$-a.s. This follows from Theorem 5.1 in Romano (1988) and Theorem 4.1(a); see Walther (1994) for the details. So when $|\hat{f}_{k(n)}(x)|$ is evaluated on $A^C_{\varepsilon/2}\cap B$, with $\mu(\mathbf{X},\mathbf{Y},k(n))$-probability tending to 1 the result will be smaller than $\|f_Z^*\|_\infty/(I_F+I_G)-\eta$ for some $\eta>0$, because

$$\limsup_{r\to 0,\, r>0}\ \sup_{x\in A^C_{\varepsilon/2}\cap B}\left|\overline{f_Z^*}^{(r)}(x)\right|\le \limsup_{r\to 0,\, r>0}\ \sup_{x\in A^C_{\varepsilon/2}\cap B}\overline{|f_Z^*|}^{(r)}(x)<\|f_Z^*\|_\infty$$

by Lemma 7.2(iii). On the other hand, by the continuity property (7.18), $|\hat{f}_{k(n)}|$ will eventually be evaluated at a point in $A_{(I_F+I_G)\eta/3}\cap B$ where the result will eventually be bigger than $\|f_Z^*\|_\infty/(I_F+I_G)-\eta/2$. So the evaluation points in $A^C_{\varepsilon/2}\cap B$ will not be chosen as points of maxima and hence $\hat{t}$ falls in $A_{\varepsilon/2}$ with $\mu(\mathbf{X},\mathbf{Y},k(n))$-probability tending to 1. The Glivenko–Cantelli theorem for halfspaces (7.2) completes the proof. □

Acknowledgments. I would like to thank the Editor, an Associate Editor and two referees for their helpful comments and suggestions.

REFERENCES

Beran, R. J. and Millar, P. W. (1986). Confidence sets for a multivariate distribution. *Ann. Statist.* **14** 431–443.

Beran, R. J. and Millar, P. W. (1987). Stochastic estimation and testing. *Ann. Statist.* **15** 1131–1154.

BERAN, R. J. and MILLAR, P. W. (1989). A stochastic minimum distance test for multivariate parametric models. *Ann. Statist.* **17** 125–140.

DONOHO, D. L. and GASKO, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Ann. Statist.* **20** 1803–1827.

HÄRDLE, W. (1990). *Applied Nonparametric Regression.* Cambridge Univ. Press.

KNUTH, D. E. (1973). *The Art of Computer Programming* **3**. Addison-Wesley, Reading, MA.

LI, G.-Y. and CHENG, P. (1993). Some recent developments in projection pursuit in China. *Statist. Sinica* **3** 35–51.

NOLAN, D. (1989). On min-max majority and deepest points. Technical Report 149, Dept. Statistics, Univ. California, Berkeley.

NOLAN, D. (1992). Asymptotics for multivariate trimming. *Stochastic Process. Appl.* **42** 157–169.

POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.

ROMANO, J. P. (1988). On weak convergence and optimality of kernel density estimates of the mode. *Ann. Statist.* **16** 629–647.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

STOER, J. and WITZGALL, C. (1970). *Convexity and Optimization in Finite Dimensions I*. Springer, Berlin.

WALTHER, G. (1994). Statistical applications of geometric duality. Ph.D. dissertation, Univ. California, Berkeley.

WAND, M. P. (1994). Fast computation of multivariate kernel estimators. *J. Comput. Graph. Statist.* **3** 433–445.

WICHURA, M. J. (1970). On the construction of almost uniformly convergent random variables with given weakly convergent image laws. *Ann. Math. Statist.* **41** 284–291.

WOLFOWITZ, J. (1954). Generalization of the theorem of Glivenko–Cantelli. *Ann. Math. Statist.* **25** 131–138.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
SEQUOIA HALL, ROOM 110A
STANFORD, CALIFORNIA 94305-4065
E-MAIL: walther@playfair.stanford.edu