Baseline zone estimation in two dimensions

Atul Mallik∗, Moulinath Banerjee∗ and Michael Woodroofe

Department of Statistics
University of Michigan
Ann Arbor, Michigan 48109

atuls@umich.edu moulib@umich.edu michaelw@umich.edu

Abstract: We consider the problem of estimating the region on which a non-parametric regression function is at its baseline level in two dimensions. The baseline level typically corresponds to the minimum/maximum of the function and estimating such regions or their complements is pertinent to several problems arising in edge estimation, environmental statistics, fMRI and related fields. We assume the baseline region to be convex and estimate it via fitting a “stump” function to approximate p-values obtained from tests for deviation of the regression function from its baseline level. The estimates, obtained using an algorithm originally developed for constructing convex contours of a density, are studied in two different sampling settings, one where several responses can be obtained at a number of different covariate-levels (dose–response) and the other involving limited number of response values per covariate (standard regression). The shape of the baseline region and the smoothness of the regression function at its boundary play a critical role in determining the rate of convergence of our estimate: for a regression function which is “p-regular” at the boundary of the convex baseline region, our estimate converges at a rate $N^{-2/(4p+3)}$ in the dose–response setting, $N$ being the total budget, and its analogue in the standard regression setting converges at a rate of $N^{-1/(2p+2)}$. Extensions to non-convex baseline regions are explored as well.

1. Introduction

Consider a data generating model of the form $Y = \mu(X) + \epsilon$, where $\mu$ is a function on $[0, 1]^2$ such that

$$\mu(x) = \tau_0 \text{ for } x \in S_0, \text{ and } \mu(x) > \tau_0 \text{ for } x \notin S_0$$

and $\tau_0$ is unknown. The covariate $X$ may arise from a random or a fixed design setting and we assume that $\epsilon$ has mean zero with finite positive variance $\sigma^2$. We are interested in estimating the baseline region $S_0$ beyond which the function deviates from its baseline value. There are several practical motivations behind detecting $S_0$ (or $S_0^c$) which can be thought of as the region of no-signal. For example, in several fMRI studies, one seeks to detect regions of brain activity from cross sectional two-dimensional images. Here, $S_0$ corresponds to the region

∗Supported by NSF Grant DMS-1007751 and a Sokol Faculty Award, University of Michigan
of no-activity in the brain with $S^c_0$ being the region of interest. In LIDAR (light detection and ranging) experiments used for measuring concentration of pollutants in the atmosphere, interest often centers on finding high/low pollution zones (see, for example, Wakimoto and McElroy (1986)); in such contexts, $S_0$ would be the zone of minimal pollution. In dose-response studies, patients may be put on multiple (interacting) drugs (see, for example, Geppetti and Benemei (2009)), and it is of interest to find the dosage levels ($\partial S_0$) at which the effect of the drugs starts kicking in.

The question of detecting $S_0$ is also related to the edge detection problem which involves recovering the boundary of an image. In edge detection, $\mu$ corresponds to the image intensity function with $S^c_0$ being the image and $S_0$ the background. A number of different algorithms in the computer science literature deal with this problem, though primarily in situations where $\mu$ has a jump discontinuity at the boundary of $S_0$; see Qiu (2007) for a review of edge detection techniques. With the exception of work done by Korostelëv and Tsybakov (1993), Mammen and Tsybakov (1995) and a few others, theoretical properties of such algorithms appear to have been rarely addressed. In fact, the study of theoretical properties of such estimates is typically intractable without some regularity assumption on $S_0$; for example, Mammen and Tsybakov (1995) discuss minimax recovery of sets under smoothness assumptions on the boundary.

In this paper, we approach the problem from the point of view of a shape-constraint (typically obtained from background knowledge) on the baseline region. We assume that the region $S_0$ is a closed convex subset of $[0,1]^2$ with a non-empty interior (and therefore, positive Lebesgue measure) and restrict ourselves to the more difficult problem where $\mu$ is continuous at the boundary. Convexity is a natural shape restriction to impose, not only because of analytical tractability, but also as convex boundaries arise naturally in several application areas: see Wang et al. (2007), Ma et al. (2010), Stahl and Wang (2005) and Goldenshluger and Spokoiny (2006) for a few illustrative examples. In the statistics literature, Goldenshluger and Zeevi (2006) provide theoretical analyses of a convex boundary recovery method in a white noise framework. While this has natural connections to our problem, we note that they impose certain conditions (see Definitions 2 and 3 of Goldenshluger and Zeevi (2006) and the associated discussions), which restricts the geometry of the set of interest, $G$, beyond convexity. Hence, their results, particularly on the rate of convergence, are difficult to compare to the ones obtained in our problem. Further, they estimate $G$ through its support function which needs to be estimated along all directions. It is unclear whether an effective algorithm can be devised to adopt this procedure in a regression setting.

Our problem also has connections to the level-sets estimation problem since $S^c_0$ is the “level-set” $\{x : \mu(x) > \tau_0\}$ of the function $\mu$. However, because $\tau_0$ is at the extremity of the range of $\mu$, the typical level-set estimate $\{x : \hat{\mu}(x) > \tau_0\}$, where $\hat{\mu}$ is an estimate of $\mu$, does not perform well unless $\mu$ has a jump at $\partial S_0$ (a situation not considered in this paper). Moreover, this plug-in approach does not account for the pre-specified shape of the level-set. We note that the shape-constrained approach to estimate level-sets has received some attention
in the literature, e.g., Nolan (1991) studied estimating ellipsoidal level-sets in the context of densities, Hartigan (1987) provided an algorithm for estimating convex contours of a density, and Tsybakov (1997) and Cavalier (1997) studied “star-shaped” level-sets of density and regression functions respectively. All the above approaches are based on an “excess mass” criterion (or its local version) that yield estimates with optimal convergence rates (Tsybakov, 1997). It will be seen later that our estimate also recovers the level-set of a transform of $\mu$, but at a level in the interior of the range of the transform. More connections in this regard are explored in Section 5.

In this paper, we extend the approach of Mallik et al. (2011), developed in a simple 1-dimensional setting, to obtain an estimate of $S_0$. We construct $p$-value type statistics which detect the deviation of the function $\mu$ from its baseline value $\tau_0$ at each covariate level and then fit an appropriate “stump” – a piecewise constant function with two levels – to these $p$-values. We study the problem in two distinct sampling settings: the so-called ‘dose–response’ setting where plenty of replicates are available at each covariate value, and the (standard) regression setting where limited (taken to be 1 without loss of generality) responses are available at each covariate level. As mentioned earlier, the regression setting is pertinent to several compelling applications. The dose-response setting is motivated by the minimum effective dose (MED) problems (a one-dimensional version of our problem) where data are available from several patients (multiple replicates) at each dose level (covariate value) and one is interested in finding the lowest dose level where the effect of the concerned drug kicks in. The baseline set in this case is, therefore, an interval $[0, d_0]$ for some unknown $d_0 > 0$. The extension of the dose-response setting to two dimensions not only provides theoretical insight into the behavior of our procedure but is also relevant to pharmacological studies involving drugs that interact.

The smoothness of $\mu$ at its boundary plays a critical role in determining the rate of convergence of our estimate: for a regression function which is “$p$-regular” (formally defined in Section 3) at the boundary of the convex baseline region, our estimate converges at a rate $N^{-2/(4p+3)}$ (Theorem 3 and the following remark) in the dose–response setting, $N$ being the total budget. This coincides with the minimax rate of a related level-set estimation problem; see (Polonik, 1995, Theorem 3.7) and (Tsybakov, 1997, Theorem 2). The analogue of the estimate in the regression setting converges at the slightly slower rate of $N^{-1/(2p+2)}$ (Theorem 6). The difference in the two rates is due to the bias introduced from the use of kernel estimates in the regression setting. A more technical explanation is given in Remark 5. It should be pointed out that our convergence rates are very different from the analogous problem in the density estimation scenario which corresponds to finding the support of a multivariate density. Faster convergence rates (Härdle, Park and Tsybakov, 1995) can be obtained in density estimation due to the simpler nature of the problem: namely, there are no realizations from outside the support of the density.

The main contributions of the paper are the following. We propose a novel and computationally simple approach to estimate baseline sets in two dimensions and deduce consistency and rates of convergence of our estimate in the
two aforementioned settings. Our approach falls at the interface of edge detection and level-set estimation problems as it detects the edge set \( (S_0) \) through a level-set estimate (see Section 5). The proofs require heavy-duty applications of non-standard empirical processes and, along the way, we deduce results which may be of independent interest. For example, we apply a blocking argument which leads to a version of Hoeffding’s inequality for \( m \)-dependent random fields, which is then further extended to an empirical process inequality. This should find usage in spatial statistics and is potentially relevant to approaches based on \( m \)-approximations that answer the central limit question for dependent random fields and their empirical process extensions; see Rosén (1969), Bolthausen (1982) and Wang and Woodroofe (2013) for some work on \( m \)-dependent random fields and \( m \)-approximations. While we primarily address the situation where the baseline set is convex, in the presence of efficient algorithms, our approach is extendible beyond convexity (see Section 5).

The rest of the paper is arranged as follows: we formally define the two settings and describe the estimation procedure in Section 2. Barring \( \mu \) and \( S_0 \), notations are not carried forward from the dose-response setting to the regression setting unless stated otherwise. We list our assumptions in Section 3. We justify consistency and deduce an upper bound on the rate of the convergence of our procedure (assuming a known \( \tau_0 \)) for the dose-response and regression settings in Sections 4.1 and 4.2 respectively. Situations with unknown \( \tau_0 \) are addressed in Section 4.3. We explore extensions to non-convex baseline regions and connections with level-set estimation in Section 5.

2. Estimation Procedure

In this section, we develop a multi-dimensional version of a \( p \)-value procedure originally developed in a one-dimensional setting in Mallik et al. (2011).

2.1. Dose-Response Setting

Consider a data generating model of the form

\[
Y_{ij} = \mu(X_i) + \epsilon_{ij}, \quad j = 1, \ldots, m, \quad i = 1, \ldots, n.
\]

Here \( m = m_n = m_0 n^\beta \) for some \( \beta > 0 \), with \( N = m \times n \) being the total budget. The covariate \( X \) is sampled from a distribution \( F \) with Lebesgue density \( f \) on \([0, 1]^2\) and \( \epsilon \) is independent of \( X \), has mean 0 and variance \( \sigma_0^2 \).

At each level \( X_i = x \), we test the null hypothesis \( H_{0,x} : \mu(x) = \tau_0 \) against the alternative \( H_{1,x} : \mu(x) > \tau_0 \) and use the resulting (approximate) \( p \)-values to construct an estimate of the set \( S_0 \). The non-normalized \( p \)-values are given by

\[
p_{m,n}(x) = 1 - \Phi(\sqrt{m}(\bar{Y}_i - \hat{\tau})),
\]

where \( \bar{Y}_i = \sum_{j=1}^{m} Y_{ij}/m \) and \( \hat{\tau} \) is some suitable estimate of \( \mu \) (to be discussed later). These \( p \)-values asymptotically have mean 1/2 for \( x \in S_0 \) and converge to
zero when \( x \notin S_0 \). This simple observation can be used to construct estimates of \( S_0 \). We fit a stump to the observed \( p \)-values, with levels 1/2 and 0 on either side of the boundary of the set and prescribe the set corresponding to the best fitting stump (in the sense of least squares) as an estimate of \( S_0 \). Formally, we define \( \xi_S(x) = (1/2)1(x \in S) \) and we minimize
\[
\sum_{i=1}^{n} (p_{m,n}(X_i) - \xi_S(X_i))^2 = \sum_{i: X_i \in S} \left( p_{m,n}(X_i) - \frac{1}{2} \right)^2 + \sum_{i: X_i \in S^c} (p_{m,n}(X_i))^2
\]
over choices of \( S \). The above least squares problem can be reduced to minimizing
\[
M_n(S) = \mathbb{P}_n \left\{ \Phi \left( \sqrt{m}(\bar{Y} - \hat{\tau}) \right) - \gamma \right\} 1_S(X),
\]
where \( \mathbb{P}_n \) denotes the empirical measure on \( \{Y_i, X_i\}_{i \leq n} \) and \( \gamma = 3/4 \).

**Remark 1.** Our methodology uses non-normalized \( p \)-values, since the test statistic sitting inside the argument to \( \Phi \) has not been normalized by the estimate of the variance. Alternatively, one could have considered fitting a stump to the normalized \( p \)-values. This alternative version of the procedure would exhibit the same fundamental feature, namely, dichotomous separation over \( S_0 \) and \( S_0^c \) which is why the stump-based procedure works, and produce identical rates of convergence. The non-normalized version is analytically and notationally more tractable as it avoids some routine (but tedious) algebraic justifications required for the normalized version.

The class of sets over which \( M_n \) is minimized should be chosen carefully as very large classes would give uninteresting discrete sets while small classes may not provide a reasonable estimate of \( S_0 \). As we assumed \( S_0 \) to be convex, we minimize \( M_n \) over \( S \), the class of closed convex subsets of \([0,1]^2\). Let \( \hat{S}_n = \text{argmin}_{S \in \mathcal{S}} M_n(S) \). The estimate \( \hat{S}_n \) can be computed by an adaptation of a density level-set estimation algorithm (Hartigan, 1987) which we state below. Note that if a closed convex set \( S^* \) minimizes \( M_n \), the convex hull of \( \{X_i : X_i \in S^*, 1 \leq i \leq n\} \) also minimizes \( M_n \). Hence, it suffices to reduce our search to convex polygons whose vertices belong to the set of \( X_i \)'s. There could be \( 2^n \) such polygons. So, an exhaustive search is computationally expensive.

**Computing the estimate.** We first find the optimal polygon (the convex polygon which minimizes \( M_n \)) for each choice of \( X \) as its leftmost vertex. We use the following notation. Let this particular \( X \) be numbered 1, and let the \( X_i \)'s not to its left be numbered 2, 3, ..., \( r \). The axes are shifted so that 1 is at the origin and the coordinates of point \( i \) are denoted by \( z_i \). The line segment \( az_i + (1 - a)z_j \), \( (0 \leq a \leq 1) \) is written as \([i, j]\). Assume that 1, ..., \( r \) are ordered so that the segments \([1, i]\) move counterclockwise as \( i \) increases and so that \( i \leq j \) if \( i \in [1, j] \). Polygons will be built up from triangles for \( 1 < i < j \leq r \); \( \Delta_{ij} \) is the convex hull of \( [1, i, j] \) excluding \([1, i]\). Note that the segment \([1, i]\) is excluded from \( \Delta_{ij} \) in order to combine triangles without overlap. The quadrilateral with vertices...
Notation for constructing the convex set estimate. An arbitrary vertex is numbered 1, and those not to its left are numbered 2, 3, . . . , 8 in a counterclockwise manner. The triangle $\Delta_{78}$ excludes the line segment $[1, 7]$. The optimal polygon (with measure $M_{67}$) with successive vertices 6, 7 and 1 is depicted as the convex polygon with vertices 1, 4, 6 and 7.

at $1, i, j, k$ for $i < j < k \leq r$ is convex if

$$D_{ijk} = \begin{vmatrix} z'_i & 1 & 1 \\ z'_j & 1 & 1 \\ z'_k & 1 & 1 \end{vmatrix} \geq 0$$

Let $M_{1j}$ be the value of $M_n$ on the line segment $[1, j]$. Further, for $1 < j < k \leq r$, let $M_{jk}$ denote the minimum value of $M_n$ among closed convex polygons with successive counterclockwise vertices $j, k$ and 1. Note that all such convex polygons contain the triangle $\Delta_{jk}$ and hence, $M_n(\Delta_{jk})$, $M_n$ measure of $\Delta_{jk}$, is a common contributing term to the $M_n$ measure of all such polygons. This simple fact forms the basis of the algorithm. It can be shown that

$$M_{jk} = M_{i^*j} + M_n(\Delta_{jk}),$$

where $i^* = I(k, j)$ is chosen to minimize $M_{ij}$ over vertices $i$ with $i < j$, $D_{ijk} \geq 0$, i.e,

$$i^* = I(k, j) = \arg\min_{i : i < j, D_{ijk} > 0} M_{ij}.$$  

Note that $i^*$ could possibly be 1, in which case $M_{jk}$ is simply the $M_n$ measure of the triangle formed by $j, k$ and 1 (including the contribution of line segment $[1, j]$).

One way to construct an optimal polygon with leftmost vertex 1 is to find the minimum among $M_{jk}$, $1 \leq j < k$, where $M_{jk}$’s are computed recursively using 2.1 and 2.2. Hence, one optimal polygon with leftmost vertex 1 has vertices
A. Mallik, M. Banerjee and M. Woodroofe/Baseline zone estimation

Fig 2. An illustration of the procedure in the dose-response setting with $m = 10$ and $n = 100$. The set $S_0$ is a circle centered at $(1, -1)$ with radius 1.

$i_1, i_2, \ldots, i_s = 1$, where either $s = 1$ or $M_{i_2 i_1} = \min_{1 \leq j < k} M_{j k}$, $i_3 = I(i_1, i_2)$, $i_4 = I(i_2, i_3)$, \ldots, $1 = i_s = I(i_{s-2}, i_{s-1})$. Once this is done for each choice of $X$ as the leftmost vertex, the final estimate $\hat{S}_n$ is simply the one with the minimum $M_n$ value among these $n$ constructed polygons.

There are minor modifications to this algorithm which reduce the over-all implementation to $O(n^3)$ computations; see Hartigan (1987, Section 3) for more details.

### 2.2. Regression Setting

Consider a data generating model of the form

$$Y_{kl} = \mu(x_{kl}) + \epsilon_{kl},$$

with $x_{kl} = (u_k, v_l)$, $u_k = k/m$, $v_l = l/m$, $k, l \in \{1, \ldots, m\}$. The total number of observations is thus $n = m^2$. The errors $\epsilon_{kl}$s are independent with mean 0 and variance $\sigma_0^2$. Here, $\mu$ is as defined earlier and we seek to estimate $S_0 = \mu^{-1}(0)$.

As earlier, we test the null hypothesis $H_{0,x} : \mu(x) = \tau_0$ against the alternative $H_{1,x} : \mu(x) > \tau_0$ at each level $x$ and use the resulting $p$-values to construct an estimate of the set $S_0$. For this, let

$$\hat{\mu}(x) = \frac{1}{nh_n^2} \sum_{k,l} Y_{kl} K \left( \frac{x - x_{kl}}{h_n} \right)$$

denote the estimator of $\mu$, with $K$ being a probability density (kernel) on $\mathbb{R}^2$ and $h_n$ the smoothing bandwidth. We take $h_n = h_0 n^{-\beta}$ for $\beta < 1/2$ and $K$ to be the 2-fold product of a symmetric one-dimensional compact kernel, i.e., $K(x_1, x_2) = K_0(x_1)K_0(x_2)$, where $K_0$ is a symmetric probability density on $\mathbb{R}$ with $K_0(x) = 0$ for $|x| \geq L_0$.

The statistic $T(x) = \sqrt{nh_n^2 (\hat{\mu}(x) - \tau_0)}$ converges in distribution to a mean zero normal random variable with variance $\Sigma^2 = \sigma_0^2 \int_{u \in \mathbb{R}^2} K^2(u)d(u)$, when
Notation and Assumptions

We adhere to the setup of Sections 2.1 and 2.2, i.e., we assume the errors to be independent and homoscedastic and consider random and fixed designs respectively for the dose-response and regression settings. A fixed design in the regression setting provides a simpler platform to illustrate the main techniques. In particular, it allows us to treat the kernel estimates as an \( \tilde{m} \)-dependent random field (where \( \tilde{m} \) is specified later) which facilitates obtaining probability bounds on our estimate; see Section 4.2. Also, a random design in the dose–response setting permits the use of empirical process techniques developed for i.i.d. data \((Y_i, X_i)\)’s are i.i.d.). However, we note here that the dose-response model in a fixed (uniform) design setting can be addressed by taking an approach similar (and in fact, simpler due to the absence of smoothing) to that for the regression setting. The results on the rate of convergence of our estimate of \( S_0 \) are identical for the random design and the fixed uniform design dose-response models.

Let \( \lambda \) denote the Lebesgue measure. The precision of the estimates is measured using the metrics

\[
d_F(S_1, S_2) = F(S_1 \Delta S_2) \quad \text{and} \quad d(S_1, S_2) = \lambda(S_1 \Delta S_2)
\]

for the dose–response and the regression settings respectively. The two metrics arise naturally in their respective settings as \( X_i \)'s have distribution \( F \) (in the dose–response setting) and the empirical distribution of the grid points in the regression setting converges to the Uniform distribution on \([0, 1]^2\).
For simplicity, we start assuming $\tau_0$ to be known. It can be shown that our results extend to cases where we impute a $\sqrt{mn}$ (dose-response)/$\sqrt{n}$ (regression) estimate of $\tau$ (more on this in Section 4.3). We summarize the assumptions below:

1. The function $\mu$ is continuous on $[0, 1]^2$. For the standard regression setting, we additionally assume that $\mu$ is Lipschitz continuous of order 1.

2. The function $\mu$ is $p$-regular at $\partial S_0$, i.e., for some $\kappa_0, C_0 > 0$ and for all $x \not\in S_0$ such that $\rho(x, S_0) < \kappa_0$,

$$C_0 \rho(x, S_0)^p \leq |\mu(x) - \tau_0|$$

Here $\rho$ is the $\ell_\infty$ metric in $\mathbb{R}^2$ (for convenience).

3. $S_0 = \mu^{-1}(\tau_0)$ is convex. For some $\epsilon_0 > 0$, $S_0 \subset [\epsilon_0, 1 - \epsilon_0]^2$ and $\lambda(S_0) > 0$.

4. The design density $f$ for the dose-response setting is assumed to be continuous and positive on $[0, 1]^2$.

5. Assumptions on the kernel $K(x) = K_0(x_1)K_0(x_2)$, $x = (x_1, x_2)$, for the standard regression setting:

   (a) $K_0$ is a symmetric probability density.

   (b) $K_0$ is compactly supported, i.e., $K_0(x) = 0$ when $|x| \geq L_0$, for some $L_0 > 0$.

   (c) $K$ is Lipschitz continuous of order 1.

Note that by the uniform continuity of $\mu$ and compactness of $[0, 1]^2$, $\inf\{\mu(x) : \rho(x, S_0) \geq \kappa_0\} > \tau_0$. For a fixed $p, \tau_0, \kappa_0, \delta_0 > 0$, we denote the class of functions $\mu$ satisfying assumptions 1, 2, 3 and

$$\inf\{\mu(x) : \rho(x, S_0) \geq \kappa_0\} - \tau_0 > \delta_0$$

by $F_p = F_p(p, \tau_0, \kappa_0, \delta_0)$.

**Remark 2.** It can be readily seen that if the regularity assumption in (3.1) holds for a particular $p$, it also holds for any $\tilde{p} > p$ as well. We assume that we are working with the smallest $p$ such that (3.1) is satisfied (the set of values $\tilde{p}$ such that (3.1) holds for a fixed $\mu$, $C_0$ and $\kappa_0$ is a closed set and is bounded from below whenever it is non-empty). In level-sets estimation theory, analogous two-sided conditions of the form

$$C_0 \rho(x, S_0)^p \leq |\mu(x) - \tau_0| \leq C_1 \rho(x, S_0)^p$$

are typically assumed (see Tsybakov (1997, Assumptions (4) and (4’)), Cavalier (1997, Assumption (4))). This stronger condition restricts the choice of $p$. However, we note here that the left inequality plays a more significant role as it provides a lower bound on the amount by which $\mu(x)$ differs from $\tau_0$ in the vicinity of $\partial S_0$. Some results in a density level-set estimation problem with a slightly weaker analogue of the left inequality can be found in Polonik (1995). The upper bound (right inequality) is seen to be useful for establishing adaptive properties of certain density level-set estimates (Singh, Scott and Nowak, 2009).
4. Consistency and Rate of Convergence

4.1. Dose-response setting

As \( \tau_0 \) is known, we take \( \tau_0 = 0 \) without loss of generality. Recall that \( \mathbb{M}_n(S) = \mathbb{P}_n \{ \Phi (\sqrt{mY}) - \gamma \} 1_S(X) \). Let \( P_m \) denote the measure induced by \((Y, X)\) and \( M_m(S) = P_m \{ \{ \Phi (\sqrt{mY}) - \gamma \} 1_S(X) \} \).

The process \( M_m \) acts as a population criterion function and can be simplified as follows. Let

\[
Z_{1m} = \frac{1}{\sqrt{m}\sigma_0} \sum_{j=1}^{m} \epsilon_{ij} \tag{4.1}
\]

and \( Z_0 \) be a standard normal random variable independent of \( Z_{1m} \). Then

\[
E [ \Phi (\sqrt{mY}) | X_1 = x ] = E [ \Phi (\sqrt{m\mu(x) + \sigma_0 Z_{1m}}) ]
\]

\[
= E [ E [ 1 (Z_0 < \sqrt{m\mu(x) + \sigma_0 Z_{1m}}) | Z_{1m} ] ]
\]

\[
= P \left[ \frac{Z_0 - \sigma_0 Z_{1m}}{\sqrt{1 + \sigma_0^2}} < \frac{\sqrt{m\mu(x)}}{\sqrt{1 + \sigma_0^2}} \right] = \Phi_m \left( \frac{\sqrt{m\mu(x)}}{\sqrt{1 + \sigma_0^2}} \right),
\]

where \( \Phi_m \) denotes the distribution function of \((Z_0 - \sigma_0 Z_{1m})/\sqrt{1 + \sigma_0^2}\). By Polya’s theorem, \( \Phi_m \) converges uniformly to \( \Phi \) as \( m \to \infty \). Hence, it can be seen that

\[
\lim_{m \to \infty} E [ \Phi (\sqrt{mY}) | X_1 = x ] = \frac{1}{2} 1_{S_0}(x) + 1_{S_0^c}(x).
\]

By the Dominated Convergence Theorem, \( M_m(S) \) converges to \( M(S) \), where

\[
M(S) = M_F(S) = \int_S \left( \frac{1}{2} 1_{S_0}(x) + 1_{S_0^c}(x) - \gamma \right) F(dx)
\]

\[
= (1/2 - \gamma) F(S_0 \cap S) + (1 - \gamma) F(S_0^c \cap S). \tag{4.2}
\]

Note that \( S_0 \) minimizes the limiting criterion function \( M(S) \). An application of the argmin continuous mapping theorem (van der Vaart and Wellner, 1996, Theorem 3.2.2) yields the following result on the consistency of \( \hat{S}_n \).

**Theorem 1.** Assume \( S_0 \) to be a closed convex set and the unique minimizer of \( M(S) \). Then \( \sup_{S \in S} | \mathbb{M}_n(S) - M(S) | \) and \( d_F (\hat{S}_n, S_0) \) converge in outer probability to zero for any \( \gamma \in (0.5, 1) \).

**Remark 3.** We end up proving a stronger result. The consistency is established in terms of the Hausdorff metric which implies consistency with respect to \( d_F \). Moreover, we do not require \( m \) to grow as \( m \alpha n^\beta \), \( \beta > 0 \) for consistency. The condition \( \min(m, n) \to \infty \) suffices. Also, the result extends to higher dimensions as well, i.e., when \( \mu \) is a function from \([0, 1]^d \to \mathbb{R} \) and \( S_0 = \mu^{-1}(0) \) is a closed convex subset of \([0, 1]^d \), then the analogous estimate is consistent. However, an efficient way to compute the estimate is not immediate.
The proof is given in Section A.1 of the Appendix.

We now proceed to deducing the rate of convergence of \( d_F(\hat{S}_n, S_0) \). For this, we study how small the difference \((M_n - M)\) is and how \( M \) behaves in the vicinity of \( S_0 \). We split the difference \((M_n - M)\) into \((M_n - M_m)\) and \((M_m - M)\) and study them separately. The term \( M_n - M_m \) involves an empirical average of centered random variables, efficient bounds on which are derived using empirical process inequalities. We start with establishing a bound on the non-random term \((M_m - M)\) in the vicinity of \( S_0 \). To this end, we first state a fact that gets frequently used in the proofs that follow.

**Fact:** For any \( \delta > 0 \), let \( S^\delta = \{ x : \rho(x, S) < \delta \} \) and \( S_\delta = \{ x : \rho(x, S_0) \geq \delta \} \) denote the \( \delta \)-fattening and \( \delta \)-thinning of the set \( S \). There exists a constant \( c_0 > 0 \) such that for any \( S \in \mathcal{S} \),

\[
\lambda(S^\delta \setminus \delta S) \leq c_0 \delta \text{ and consequently, } F(S^\delta \setminus \delta S) \leq c_0 \delta, \tag{4.3}
\]

with \( c_0 = \|f\|_{\infty} c_0 (\|f\|_{\infty} < \infty, \text{ by Assumption 4}). \) For a proof of the above, see, for example, Dudley (1984, pp. 62–63).

**Lemma 1.** For any \( \delta > 0 \), \( a_n \downarrow 0 \) and \( S \in \mathcal{S} \) such that \( F(S \triangle S_0) < \delta \),

\[
| (M_m - M)(S) - (M_m - M)(S_0) | \leq | \Phi_m(0) - 1/2|\delta + \min(c_0 a_n, \delta) \\
+ \Phi_m \left( \frac{C_0 \sqrt{m a_n^2}}{1 + \sigma_0^2} \right) - 1 |\delta \\
+ \Phi_m \left( \frac{\sqrt{m} \delta_0}{1 + \sigma_0^2} \right) - 1 |\delta.
\]

**Proof.** Note that

\[
M_m(S) - M_m(S_0) = P_m \left[ \left\{ \Phi_m \left( \frac{\sqrt{m} \mu(x)}{\sqrt{1 + \sigma_0^2}} \right) - \gamma \right\} \{ 1_S(x) - 1_{S_0}(x) \} \right]
\]

and

\[
M(S) - M(S_0) = \int \left\{ (1/2)1_{S_0}(x) + 1_{S_\delta}(x) - \gamma \right\} \{ 1_S(x) - 1_{S_0}(x) \} F(dx).
\]

Hence, the expression \(| (M_m - M)(S) - (M_m - M)(S_0) |\) is bounded by

\[
\int_{x \in (S \setminus S_0)} | \Phi_m(0) - 1/2 | F(dx) + \int_{x \in (S_\delta \setminus S_0)} | \Phi_m \left( \frac{\sqrt{m} \mu(x)}{\sqrt{1 + \sigma_0^2}} \right) - 1 | F(dx)
\]

Note that the first term is bounded by \(| \Phi_m(0) - 1/2|\delta \). Further, let \( S_n = \{ x : \rho(x, S_0) \geq a_n \} \). Using (4.3), \( F(S_n \setminus S_0) \leq c_0 a_n \). Also, as \( a_n \downarrow 0 \), \( a_n < \kappa_0 \) for sufficiently large \( n \). Thus, for \( x \in S_n \),

\[
\mu(x) \geq \min(\rho(x, S_0)^p, \delta_0) \geq \min(a_n^p, \delta_0),
\]
using (3.1) and (3.2). Hence, the second sum in (4.4) is bounded by
\[ F(S_n \setminus S_0) + \int_{x \in (S_n \cap S)} \left| \Phi_m \left( \frac{\sqrt{m} \mu(x)}{\sqrt{1 + \sigma_0^2}} \right) - 1 \right| F(dx) \leq \min(\hat{c}_{0a_n}, \delta) \]
\[ + \int_{x \in (S_n \cap S)} \left\{ \left| \Phi_m \left( \frac{C_0 \sqrt{a_n^p} \mu}{\sqrt{1 + \sigma_0^2}} \right) - 1 \right| + \left| \Phi_m \left( \frac{\sqrt{m} \delta_0}{\sqrt{1 + \sigma_0^2}} \right) - 1 \right| \right\} F(dx). \]
As \( F(S_n \cap S) < \delta \), we get the result. □

To control \( M_n - M_m \), we rely on a version of Theorem 5.11 of van de Geer (2000). The result in its original form is slightly general. In their notation, it involves a bound on a special metric \( \rho_K(\cdot) \) (see van de Geer (2000, equation 5.23)) which, in light of Lemma 5.8 of van de Geer (2000), can be controlled by bounding the \( L_2 \)-norm in the case of bounded random variables. This yields the consequence stated below. Here, \( H_B \) denotes the entropy with respect to bracketing numbers.

**Theorem 2.** Let \( \mathcal{G} \) be a class of functions such that \( \sup_{g \in \mathcal{G}} \|g\|_\infty \leq 1 \). For some universal constant \( C > 0 \), let \( C_2, C_3, R \) and \( N > 0 \) satisfy the following conditions:
\[ R \geq \sup_{g \in \mathcal{G}} \|g\|_{L_n(P)}, \]
\[ N \geq C_2 \int_0^R H_B^{1/2}(u, \mathcal{G}, L_2(P))du \vee R \]
\[ C_2^2 \geq C^2(C_3 + 1) \quad \text{and} \]
\[ N \leq C_3 \sqrt{n}R^2. \]
Then
\[ P^* \left[ \sup_{g \in \mathcal{G}} |G_n(g)| > N \right] \leq C \exp \left[ \frac{-N^2}{C^2(C_3 + 1)R^2} \right], \]
where \( P^* \) denotes the outer probability.

We have the following theorem on the rate of convergence of \( \hat{S}_n \).

**Theorem 3.** For any \( \beta > 0 \),
\[ P^* \left( d_F(\hat{S}_n, S_0) > \delta_n \right) \rightarrow 0 \]
for \( \delta_n = K_1 \max\{n^{-2/3}, m^{-1/(2p)}\} \), where \( K_1 > 0 \) is some constant.

**Proof.** Let \( k_n \) be the smallest integer such that \( 2^k \delta_0 \geq 1 \). For \( 0 \leq k \leq k_n \), let \( S_{n,k} = \{ S \in S : 2^k \delta_0 < d_F(S, S_0) \leq 2^{k+1} \delta_0 \} \). As \( \hat{S}_n \) is the minimizer for \( M_n \),
\[ P^* \left( d_F(\hat{S}_n, S_0) > \delta_n \right) \leq \sum_{k=0}^{k_n} P^* \left( \inf_{A \in S_{n,k}} M_n(A) - M_n(S_0) \leq 0 \right). \]
The sum on the right side is bounded by
\[
\sum_{k=0}^{k_n} P^* \left( \sup_{A \in S_{n,k}} |(M_n - M)(S) - (M_n - M)(S_0)| > \inf_{A \in S_{n,k}} (M(S) - M(S_0)) \right).
\]

(4.5)

For \( c(\gamma) = \min(\gamma - 1/2, 1 - \gamma) > 0 \),
\[
M(S) - M(S_0) = (\gamma - 1/2)(F(S_0) - F(S_0 \cap S)) + (1-\gamma)F(S_0 \cap S) \geq c(\gamma)F(S \Delta S_0),
\]
and hence, (4.5) is bounded by
\[
\sum_{k=0}^{k_n} P^* \left( \sup_{A \in S_{n,k}} |(M_n - M)(S) - (M_n - M)(S_0)| > c(\gamma)2^{k-1}\delta_n \right)
\]
\[
+ \sum_{k=0}^{k_n} P^* \left( \sup_{A \in S_{n,k}} |(M_n - M)(S) - (M_n - M)(S_0)| \geq c(\gamma)2^{k-1}\delta_n \right) \quad (4.6)
\]

Note that \( M_n - M \) is a non-random process and hence, each term in the second sum is either 0 or 1. We now show that the second sum in the above display is eventually zero. For this, we apply Lemma 1. Note that
\[
\sup_{A \in S_{n,k}} |(M_n - M)(S) - (M_n - M)(S_0)|
\]
\[
\leq |\Phi_m(0) - 1/2|2^{k+1}\delta_n + \min(\tilde{c}_0a_n, 2^{k+1}\delta_n)
\]
\[
+ |\Phi_m \left( \frac{C_0\sqrt{m\delta_n}}{\sqrt{1 + \sigma_0^2}} \right) - 1|2^{k+1}\delta_n + |\Phi_m \left( \frac{\sqrt{m\delta_n}}{\sqrt{1 + \sigma_0^2}} \right) - 1|2^{k+1}\delta_n
\]
\[
\leq 4 \left| \Phi_m(0) - 1/2 \right| + \left| \Phi_m \left( \frac{\sqrt{m\delta_n}}{\sqrt{1 + \sigma_0^2}} \right) - 1 \right|2^{k-1}\delta_n
\]
\[
+ \left[ \frac{2\tilde{c}_0a_n}{\delta_n} + 4 \left| \Phi_m \left( \frac{C_0\sqrt{m\delta_n}}{\sqrt{1 + \sigma_0^2}} \right) - 1 \right| \right]2^{k-1}\delta_n. \quad (4.7)
\]

Hence, it suffices to show that the coefficient of \( 2^{k-1}\delta_n \) in the above expression is smaller than \( c(\gamma) \). To this end, fix \( 0 < \eta < c(\gamma)/8 \). For large \( m \),
\[
|\Phi_m(0) - 1/2| + \left| \Phi_m \left( \frac{\sqrt{m\delta_n}}{\sqrt{1 + \sigma_0^2}} \right) - 1 \right| \leq \eta.
\]
Choose \( c_0 \) such that \( a_n = c_0m^{-1/(2p)} > \left[ \Phi_m^{-1}(1 - \eta)\sqrt{1 + \sigma_0^2}/(C_0\sqrt{m}) \right]^{1/p} \). For large \( n \), the coefficient of \( 2^{k-1}\delta_n \) in (4.7) is then bounded by
\[
8\eta + \frac{\tilde{c}_0c_0}{K_1} < c(\gamma),
\]
for \( K_1 > (\tilde{c}_0c_0)/(c(\gamma) - 8\eta) \). Hence, each term in the second sum of (4.6) is zero for a suitably large choice of the constant \( K_1 \). Note that the first term in (4.6)
can be written as  
\[ \sum_{k=0}^{k_n} P^* \left( \sup_{A \in S_{n,k}} |G_n g_m(\bar{Y})1_{A \Delta S_0}(X)| > c(\gamma)2^{k-1} \delta_n \sqrt{n} \right), \tag{4.8} \]

where \( g_m(y) = \Phi(\sqrt{m}y) - \gamma \). We are now in a position to apply Theorem 2 to each term of (4.8). In the setup of Theorem 2, \( N = c(\gamma)2^{k-1} \delta_n \sqrt{n} \). The concerned class of functions is \( G_{n,k} = \{ g_m(\bar{Y})1_B(X) : B = A \Delta S_0, B \in S_{n,k} \} \). Note that \( \| g_m1_B \|_{L_2(P)} \leq |E1_B(X)|^{1/2} \leq (2^{k+1} \delta_n)^{1/2} \). So we can pick \( R = R_{n,k} = (2^{k+1} \delta_n)^{1/2} \). As \( S_{n,k} \subset S, N_{\{\}}(u, \{ A \Delta S_0 : A \in S_{n,k} \}, L_2(P)) \leq (N_{\{\}}(u,S,L_2(P))^2 \) for any \( u > 0 \). Also, starting with a bracket \([f_L, f_U] \) for \( \{ A \Delta S_0 : A \in S_{n,k} \} \) containing \( B \) with \( \| f_U - f_L \|_{L_2(P)} \leq u \), we can obtain brackets for the class \( G_{n,k} \) using the inequality

\[ \Phi(\sqrt{m}y) f_L - \gamma f_U \leq g_m(y)1_B(x) \leq \Phi(\sqrt{m}y) f_U - \gamma f_L. \]

As \( \| g_m \|_\infty \leq 1, \)

\[ \| (\Phi(\sqrt{m}y) f_U - \gamma f_L) - (\Phi(\sqrt{m}y) f_L - \gamma f_U) \|_{L_2(P)} \leq u. \]

Hence, \( H_B(u, G_{n,k}, L_2(P)) \leq H_B(u, S, L_2(P)) \). Using the fact that in dimension \( d, H_B(u, S, L_2(P)) = \log(N_{\{\}}(u,S,L_2(P))) \leq A_0 u^{-(d-1)} \) for \( d \geq 2 \) (see Bronštein (1976)), we get

\[ H_B(u, G_{n,k}, L_2(P)) \leq A_0 u^{-1} \]

for some constant \( A_0 > 0 \) (depending only on the design distribution). The conditions of Theorem 2 then translate to

\[ 2^{k-1} c(\gamma) \delta_n \sqrt{n} \geq 2 C_2 \max(A_0, 1)(2^{k+1} \delta_n)^{1/4} \]

\[ C_2^2 \geq C^2(C_3 + 1) \]

\[ c(\gamma)2^{k-1} \delta_n \sqrt{n} \leq C_3 \sqrt{m}2^{k+1} \delta_n. \]

It can be seen that for \( K_1 \geq 2^9 (C_2 \max(A_0, 1)/(c(\gamma))^{4/3}, C_3 = c(\gamma)/4 \) and \( C_2 = \sqrt{5}C/2 \), these conditions are satisfied, and hence, we can bound (4.8) by

\[ \sum_{k=0}^{k_n} C \exp \left\{ -2^{k-3}c^2(\gamma) \delta_n n \right\} \]

As \( \delta_n \gtrsim n^{-2/3} \) (the symbol \( \gtrsim \) is used to denote the corresponding \( \geq \) inequality holding up to some finite positive constant), the term \( \delta_n n \) diverges to \( \infty \) as \( n \to \infty \). Hence, the above display converges to zero. This completes the proof. \( \square \)

**Remark 4.** The result obviously holds for values of \( \delta_n \) larger than the one prescribed above. Hence, it also gives consistency, though it requires \( m \) to grow as \( m_0 n^\beta \). In terms of the total budget, choosing \( \beta = 4p/3 \) corresponds to the optimal
rate in which case \( \delta_n \) is of the order \( n^{-2/3} \) or \( N^{-2/(4p+3)} \). To see this, we just set
\[
n^{-2/3} = m^{-1/2p}
\]
and follow up the implications of this for \( \beta \). The \( N^{-2/(4p+3)} \) rate coincides with the minimax rate obtained for a related density level set problem in Tsybakov (1997, Theorem 2) (see also Polonik (1995, Theorem 3.7)).

Note that the bounds deduced for the two sums in (4.6) depend on \( \mu \) only through \( p \) and \( \delta_0 \), e.g., the exponential bounds from Theorem 2 depend on the class of functions only through their entropy and norm of the envelope which do not change with \( \mu \). Hence, we have the following result which is similar in flavor to the upper bounds deduced for level-set estimates in Tsybakov (1997).

**Corollary 1.** For the choice of \( \delta_n \) given in Theorem 3,
\[
\lim sup_{n \to \infty} \sup_{\mu \in \mathcal{F}_p} \mathbb{E}_{\mu}^* \left[ \delta_n^{-1} d(\hat{S}_n, S_0) \right] < \infty. \tag{4.9}
\]
Here, \( \mathbb{E}_{\mu}^* \) is the expectation with respect to the model with a particular \( \mu \in \mathcal{F}_p \). The other features of the model such as error distribution and the design distribution do not change.

**Proof.** Note that
\[
E_{\mu}^* \left[ \delta_n^{-1} d(\hat{S}_n, S_0) \right] \leq 1 + \sum_{k \geq 0, 2^k \delta_n \leq 1} 2^{k+1} P^* \left( 2^k < \delta_n^{-1} d(\hat{S}_n, S_0) \leq 2^{k+1} \right) \\
\leq 1 + \sum_{k \geq 0, 2^k \delta_n \leq 1} 2^{k} P^* \left( \inf_{A \in \mathcal{S}_{n,k}} M_n(A) - M_n(S_0) \leq 0 \right).
\]
The probabilities \( P^* \left( \inf_{A \in \mathcal{S}_{n,k}} M_n(A) - M_n(S_0) \leq 0 \right) \) can be bounded in an identical manner to that in the proof of the Theorem 3 and hence, we get
\[
\sup_{\mu \in \mathcal{F}_p} E_{\mu}^* \left[ \delta_n^{-1} d(\hat{S}_n, S_0) \right] \leq 1 + \sum_{k=0}^{k_n} C2^{k+1} \exp \left\{ \frac{-2^{k-3} C^2(\gamma)\delta_n n}{C^2(C_3 + 1)} \right\}.
\]
As \( \delta_n n \to \infty \), the right side of the above is bounded and hence, we get the result. \( \square \)

### 4.2. Regression Setting

With \( \tau_0 = 0 \), recall that
\[
M_n(S) = \frac{1}{n} \sum_{k,l:x_{kl} \in \mathcal{I}_n} \left\{ \Phi \left( \sqrt{n} \hat{\mu}(x_{kl}) \right) - \gamma \right\} 1_S(x_{kl}).
\]
For any fixed \( \gamma \in (1/2, 1) \), it can be shown that \( \hat{S}_n \) is consistent for \( S_0 \), i.e.,
\( d(\hat{S}_n, S_0) \) converges in probability to zero.
Theorem 4. Assume $S_0$ to be a closed convex set and the unique minimizer of $M(S)$, where

$$M(S) = (1/2 - \gamma)\lambda(S_0 \cap S) + (1 - \gamma)\lambda(S_0 \cap S).$$

Then, $\sup_{S \in S} |M_n(S) - M(S)|$ converges in probability to zero and $\hat{S}_n$ is consistent for $S_0$ in the sense that $d(\hat{S}_n, S_0)$ converges in probability to zero for any $\gamma \in (0.5, 1)$.

As was the case in the dose-response setting (see Remark 3), a more general result holds and is proved in Section A.2 of the Appendix.

We now deduce a bound on the rate of convergence of $\hat{S}_n$. For notational simplicity, $\kappa_n$ is used to denote a sum over the set $\{k, l : x_{kl} \in I_n\}$ unless stated otherwise. Also, let

$$\tilde{\mu}(x) = \frac{1}{\sqrt{n}h_n} \sum_{k', l'} \mu(x_{k', l'}) K \left( \frac{x - x_{k', l'}}{h_n} \right)$$

Note that $\sqrt{n}h_n^2 \tilde{\mu}(x_{kl}) = \sqrt{n}h_n^2 \tilde{\mu}(x_{kl}) + Z_{kl}$ and $\text{Var}(Z_{kl}) = \Sigma_n^2(x_{kl})$. We have

$$E \left[ \Phi \left( \sqrt{n}h_n^2 \tilde{\mu}(x_{kl}) \right) \right] = E \left[ \Phi \left( \sqrt{n}h_n^2 \tilde{\mu}(x_{kl}) + Z_{kl} \right) \right] = E \left[ 1 \left( Z_0 \leq \sqrt{n}h_n^2 \tilde{\mu}(x_{kl}) + Z_{kl} \right) \right] = \Phi_{kl,n} \left( \frac{\sqrt{n}h_n^2 \tilde{\mu}(x_{kl})}{\sqrt{1 + \Sigma_n^2(x_{kl})}} \right).$$

where $\Phi_{kl,n}$ denotes the distribution function of $(Z_0 - Z_{kl}) / \sqrt{1 + \Sigma_n^2(x_{kl})}$. For $x_{kl} \in I_n$, $\Sigma_n^2(x_{kl})$ and $\Phi_{kl,n}$ do not vary with $k$ and $l$ and are denoted by $\Sigma^2_n$ and $\Phi_n$, respectively, for notational convenience. We get

$$\tilde{M}_n(S) = E\{M_n(S)\} = \frac{1}{n} \sum_{k, l} \left\{ \tilde{\Phi}_n \left( \frac{\sqrt{n}h_n^2 \tilde{\mu}(x_{kl})}{\sqrt{1 + \Sigma_n^2(x_{kl})}} \right) - \gamma \right\} 1_{S(x_{kl})}.\tag{4.11}$$

Also, for $x_{kl} \in I_n$, any $\eta > 0$ and sufficiently large $n$,

$$\frac{1}{nh_n^2\Sigma^2_n} \sum_{k', l'} \sum_{\rho(x_{kl}, x_{k'l'}) \leq L_0nh_n} E \left[ \epsilon_{k'l'} K^2 \left( \frac{x_{kl} - x_{k'l'}}{h_n} \right) 1 \left( \left| \epsilon_{k'l'} \right| (x_{kl} - x_{k'l'}/h_n) \right) > \eta \right]$$
Lemma 2. The distance between $\bar{a}$ and $M$ is driven by the behavior of how small the difference $M - S$ behaves in the vicinity of $\Phi$. Hence, by the Lindeberg–Feller central limit theorem, $\tilde{Z}_{kl}/\Sigma_n$ and consequently, $\tilde{\Phi}_n$ converge weakly to $\Phi$. Further, by Pólya’s theorem, $\Phi_n$ converges uniformly to $\Phi$ as $n \to \infty$, a fact we use in the proof of Lemma 2.

We now consider the distance $d(\bar{S}_n, S_0)$, the rate of convergence of which is driven by the behavior of how small the difference $M_n - M$ is and how $M$ behaves in the vicinity of $S_0$. As before, we split the difference $M_n - M$ into $M_n - \bar{M}_n$ and $\bar{M}_n - M$ and study them separately. We first derive a bound on $\bar{M}_n - M$.

**Lemma 2.** There exist a positive constant $c_1$ such that for any $a_n \downarrow 0$ satisfying $a_n > 2L_0 h_n$, $\delta > 0$ and $\lambda(S \triangle S_0) < \delta$,  

$$
|\langle \bar{M}_n - M \rangle (S) - \langle M_n - M \rangle (S_0)| \leq |\tilde{\Phi}_n(0) - 1/2|\delta + \min(c_0 a_n, \delta) + \phi_n \left( \frac{\sqrt{nh_2^2}C_0(a_n - 2L_0 h_n)^p}{2\sqrt{1 + \Sigma^2}} \right) - 1 |\delta + \phi_n \left( \frac{\sqrt{nh_2^2}d_0}{2\sqrt{1 + \Sigma^2}} \right) - 1 |\delta + c_1 h_n. 
$$

(4.12)

The proof of this lemma is available in Section A.3 of the appendix.

We next consider the term $M_n (S) - \bar{M}_n (S)$. With $\tilde{W}_{kl}$s as defined in (2.3), let $W_{kl} = \tilde{W}_{kl} - E \{ \tilde{W}_{kl} \}$. Then

$$
M_n (S) - \bar{M}_n (S) = \frac{1}{n} \sum_{k,l} W_{kl} 1_S (x_{kl}).
$$

For notational ease, we define $W_{kl} \equiv 0$ whenever $x_{kl} \notin I_n$. As the kernel $K$ is compactly supported, $W_{kl}$ is independent of all $W_{kl'}$s except for those in the set $\{ W_{kl'} : (k', l') \in \{ 1, \ldots, m \}^2, p((k, l), (k', l')) \leq 2L_0 mh_n \}$. The cardinality of this set is at most $m' = 16L_0^2 nh_n^2$. Hence, $\{ W_{kl} \}_{1 \leq k, l \leq m}$ is an ($\sqrt{m'/2}$)-dependent random field. For

$$
k_i = i + k \left\lceil \sqrt{m'} \right\rceil, \quad l_j = j + l \left\lceil \sqrt{m'} \right\rceil \quad \text{and} \quad r_{ij} = \sum_{k, \ell : 1 \leq k, \ell \leq m} 1, \quad (4.13)
$$

let

$$
\tilde{d}_n (S_1, S_2) = \tilde{d}_n (1_{S_1}, 1_{S_2})
$$

$$
= \left\{ \max_{1 \leq i, j \leq \sqrt{m'}} \left[ \frac{1}{r_{ij}} \sum_{k, l : 1 \leq k, l \leq m} (1_{S_1} (x_{k, l}) - 1_{S_2} (x_{k, l}))^2 \right] \right\}^{1/2}
$$

is bounded by

$$
\frac{2 [2L_0 mh_n]^2 \| K \|_2^2}{n h_2^2 (\sigma_0^2 K^2)} E \left[ \varepsilon_{11}^2 \left( \frac{2 \| K \|_\infty}{n h_2^2 (\sigma_0 \sqrt{K^2})} |\varepsilon_{11}| > \eta \right) \right].
$$

which converges to zero. Hence, by the Lindeberg–Feller central limit theorem, $Z_{kl}/\Sigma_n$ and consequently, $\tilde{\Phi}_n$ converge weakly to $\Phi$. Further, by Pólya’s theorem, $\Phi_n$ converges uniformly to $\Phi$ as $n \to \infty$, a fact we use in the proof of Lemma 2.
Lemma 3. For sufficiently large \( n \),

\[
P \left( \frac{1}{n} \sum_{k,l} W_{kl} 1_S(x_{kl}) \geq a \right) \leq 2 \exp \left[ - \frac{na^2}{16m'\|S\|^2_n} \right].
\]

The proof is given in Section A.4 of the Appendix. In fact, such a result holds for general (bounded) \( (\sqrt{m'}/2) \)-dependent random fields \( \{V_{kl} : 1 \leq k, l \leq m\} \) with \( |V_{kl}| \leq 1 \) and weights \( g(x_{kl}) \) (instead of \( 1_S(x_{kl})'s \)) with \( \tilde{d}_n(g_1, g_2) \) defined accordingly, as long as \( n/m' \to \infty \), i.e., it can be shown that

\[
P \left( \frac{1}{n} \sum_{k,l} V_{kl} g(x_{kl}) \geq a \right) \leq 2 \exp \left[ - \frac{na^2}{16m'\|g\|^2_n} \right]. \tag{4.14}
\]

Moreover, we can generalize the above to a probability bound on the supremum of an empirical process.

Theorem 5. Let \( G \) denote a class of weight functions \( g : \{x_{kl} : 1 \leq k, l \leq m\} \to \mathbb{R} \) and \( H \) denote the entropy of this class with respect to covering numbers and the metric \( \tilde{d}_n \). Assume \( \sup_{g \in \mathcal{G}} \|g\|_n \leq R \). Let \( V_{kl}s \) be random variables with \( |V_{kl}| \leq 1 \) such that the inequality (4.14) holds for all \( g \in \mathcal{G} \). Then, there exists a universal constant \( C > 0 \) such that for all \( \delta_1 > \delta_2 \geq 0 \) satisfying

\[
\sqrt{n/m'}(\delta_1 - \delta_2) \geq C \left( \int_{\delta_2/8}^R H^{1/2}(u, \mathcal{G}, \tilde{d}_n)du \vee R \right), \tag{4.15}
\]

we have

\[
P^* \left[ \sup_{g \in \mathcal{G}} \left( \frac{1}{n} \sum_{k,l} V_{kl} g(x_{kl}) \geq \delta_1 - \delta_2 \right) \right] \leq C \exp \left[ - \frac{n(\delta_1 - \delta_2)^2}{Cm'R^2} \right].
\]

The above result states that the supremum of weighted average of (bounded) \( (\sqrt{m'}/2) \)-dependent random fields, where weights belong to a given class, has sub-gaussian tails. As mentioned earlier, we expect this to be useful in \( m \)-approximation approaches that are used for deriving limit theorems for dependent random variables and to obtain their empirical process extensions. Here, we use it to control the centered empirical averages \( \mathbb{M}_n - \bar{M}_n \). The proof of the above result is outlined in Section A.5 of the Appendix.

We are now in a position to deduce a bound on the rate of convergence of \( d(\hat{S}_n, S_0) \).

Theorem 6. Let \( \nu_n = \max \{ h_n, (nh_n^2)^{-1/(2p)} \} \). For some \( K_1 > 0 \), and \( \delta_n = K_1\nu_n \), \( P^* \left( d(\hat{S}_n, S_0) > \delta_n \right) \to 0 \) as \( n \to \infty \).
Proof. Let \( k_n \) be the smallest integer such that \( 2^{k_n+1} \delta_n \geq 1 \). For \( 0 \leq k \leq k_n \), let \( S_{n,k} = \{ S : S \in S, \ 2^k \delta_n < d(S, S_0) \leq 2^{k+1} \delta_n \} \). As, \( \hat{S}_n \) is the minimizer for \( \hat{M}_n \),

\[
P^* \left( d(\hat{S}_n, S_0) > \delta_n \right) \leq \sum_{k=0}^{k_n} P^* \left( \inf_{A \in S_{n,k}} \hat{M}_n(A) - \hat{M}_n(S_0) \leq 0 \right).
\]

The sum on the right side can be written as:

\[
\sum_{k=0}^{k_n} P^* \left( \sup_{A \in S_{n,k}} |(M_n - M)(S) - (M_n - M)(S_0)| > \inf_{A \in S_{n,k}} (M(S) - M(S_0)) \right).
\]

(4.16)

For \( c(\gamma) = \min(\gamma - 1/2, 1-\gamma) \), \( M(S) - M(S_0) \geq c(\gamma) \lambda(S \Delta S_0) \), and hence (4.17) is bounded by

\[
\sum_{k=0}^{k_n} P^* \left( \sup_{A \in S_{n,k}} |(M_n - \hat{M}_n)(S) - (M_n - \hat{M}_n)(S_0)| > c(\gamma)2^{k-1} \delta_n \right)
\]

\[
+ \sum_{k=0}^{k_n} 1 \left[ \sup_{A \in S_{n,k}} |(\hat{M}_n - M)(S) - (M_n - M)(S_0)| \geq c(\gamma)2^{k-1} \delta_n \right].
\]

(4.17)

We first apply Lemma 2 to the second sum in the above display. Note that

\[
\sup_{A \in S_{n,k}} |(\hat{M}_n - M)(S) - (M_n - M)(S_0)|
\]

\[
\leq |\hat{\Phi}_n(0) - 1/2|2^{k+1} \delta_n + \min(h_n, 2^{k+1} \delta_n)
\]

\[
+ |\hat{\Phi}_n \left( \frac{\sqrt{n} h_n \delta_0}{2 \sqrt{1 + \Sigma^2}} \right) - 1|2^{k+1} \delta_n
\]

\[
+ |\hat{\Phi}_n \left( \frac{\sqrt{n} h_n \delta_0}{2 \sqrt{1 + \Sigma^2}} \right) - 1|2^{k+1} \delta_n + c_1 h_n
\]

\[
\leq 4 \left[ |\hat{\Phi}_n(0) - 1/2| + |\hat{\Phi}_n \left( \frac{\sqrt{n} h_n \delta_0}{2 \sqrt{1 + \Sigma^2}} \right) - 1| \right]2^{k-1} \delta_n
\]

\[
+ \frac{2c_0 a_n + c_1 h_n}{\delta_n} + 4 \left[ \hat{\Phi}_n \left( \frac{\sqrt{n} h_n \delta_0}{2 \sqrt{1 + \Sigma^2}} \right) - 1 \right]2^{k-1} \delta_n.
\]

Fix \( 0 < \eta < c(\gamma)/8 \). For large \( n \), \( |\hat{\Phi}_n(0) - 1/2| + c_1/(\sqrt{n} \delta_n) < \eta \). Choose \( c_\eta \) such that \( a_n = c_\eta n > \left[ 2\hat{\Phi}_n^{-1}(1-\eta)\sqrt{1 + \Sigma^2}/(C_0 \sqrt{\delta_n}) \right]^{1/p} + (2L_0)h_n \). Then the coefficient of \( 2^{k-1} \delta_n \) in the above display is bounded by

\[
8\eta + \frac{2c_0 c_\eta + c_1}{K_1} < c(\gamma),
\]

when \( K_1 > (2c_0 c_\eta + c_1)/(c(\gamma) - 8\eta) \). Hence, for a suitably large choice of \( K_1 \), each term in the second sum of (4.17) is zero.
We now apply Theorem 5 to each term in the first sum of (4.17). For this we use the following claim to obtain a bound on the entropy of the class $S_{n,k}$.

**Claim A.** We claim that $\sup_{S_1, S_2 \in S} |d_n^2(S_1, S_2) - \lambda(S_1 \triangle S_2)| = O(h_n)$ and that $H(u, \{B \triangle S_0 : B \in S_{n,k}\}, \tilde{d}_n) \leq A_1(u - c_2 h_n)^{-1}$ for constants $c_2 > 0$ and $A_1 > 0$.

We first use the above claim to prove the result. As a consequence of Claim A, $\sup_{A} \|A\|_n \leq R_{n,k} := (2^{k+1} \delta_n + c_3 h_n)^{1/2}$, for some $c_3 > 0$. Using Theorem 5 with $\delta_1 = c(\gamma) 2^{k-1} \delta_n$, $\delta_2 = 8c_2 h_n$, we arrive at the condition

$$\sqrt{n/m'}(c(\gamma) 2^{k-1} \delta_n - 8c_2 h_n) \gtrsim (R_{n,k} + c h_n)^{1/2} \lor R_{n,k},$$

for some $c_4 > 0$. As $\delta_n \gtrsim \nu_m$, this translates to $(2^{k-1} c(\gamma))^4 \delta_n^3 \gtrsim (2^k + c_5) h_n^4$ for some $c_5 > 0$. This holds for all $k$ when $\delta \gtrsim h_n^{4/3}$ which is true as $\delta_n \gtrsim h_n$.

Hence, we can bound the first sum in (4.17) by

$$\sum_{k=0}^{k_n} 5C \exp \left[ \frac{-n (c(\gamma) 2^{k-1} \delta_n - 8c_2 h_n)^2}{Cn' (2^{k+1} \delta_n + c_3 h_n)^{1/2}} \right].$$

(4.18)

Consequently, the display in (4.18) is bounded by

$$\sum_{k=0}^{\infty} 5C \exp \left[ \frac{-c(\gamma) 2^{k-1} - c_6}{Ch_n^2 (2^{k+1} + c_7)} \right],$$

for some constants $c_6$, $c_7 > 0$. As $\delta_n / h_n^2 \gtrsim h_n^{-1} \to \infty$, we get the result.

**Proof of Claim A.** Note that $d_2^2(S_1, S_2) = \hat{d}_2^2(S_1 \triangle S_2, \phi) = \max_{1 \leq i,j \leq \sqrt{m'}} Q_i^j(S_1 \triangle S_2)$, where $Q_i^j$ is the discrete uniform measure on the points $\{x_{k,l} : k = i + l \sqrt{m'}, l_j = j + l \} \leq m \}$. Note that each $Q_i^j$ approximates Lebesgue measure at rectangle of length $m/m' = O(h_n)$. The rectangles that intersect with the boundary of a set $S$ account for the difference $|Q_i^j(S) - \lambda(S)|$. As argued in the proof of Lemma 2, the error $\sup_{S \in S} \max_{i,j} |Q_i^j(S) - \lambda(S)| \leq \lambda(\{x : \rho(x, \partial S) < O(h_n)\})$, which is $O(h_n)$ using (4.3).

To see that $H(u, \{B \triangle S_0 : B \in S_{n,k}\}, \tilde{d}_n) \leq A_1(u - c_2 h_n)^{-1}$, first note that $H(u, \{B \triangle S_0 : B \in S_{n,k}\}, \tilde{d}_n) \leq H(u, S, \tilde{d}_n)$. For any convex set $S$, it can be shown from arguments analogous to those in the proof for Lemma 2 that for some $c_2 > 0$,

$$\max_{1 \leq i,j \leq \sqrt{m'}} Q_i^j(S^\delta \triangle S) \leq \lambda(S^{(\delta+c_2 h_n)^{1/2}} \triangle (\delta + c_2 h_n) S) \leq c_9(\delta + c_2 h_n).$$

If $S_1, \ldots, S_r$ are the center of the Hausdorff balls with radius $\delta$ that cover $S$ (see (A.1) in the Appendix for a definition of Hausdorff distance $d_H$), then $[S_i, S_i']$, $i \leq r$ form brackets that cover $S$. The sizes of these brackets are $(c_9(\delta + c_2 h_n))^{1/2}$ in terms of the distance $\tilde{d}_n$. Hence,

$$H((c_9(\delta + c_2 h_n))^{1/2}, S, \tilde{d}_n) \leq H_B((c_9(\delta + c_2 h_n))^{1/2}, S, \tilde{d}_n) \leq H(\delta, S, d_H).$$
Letting \( u = c_0(\delta + c_2 h_n)^{1/2} \) and using the fact that \( H(\delta, S, d_H) \leq \delta^{-1/2} \) we get Claim A. \( \square \)

As was the case with Corollary 1, Proposition 6 extends to the following result in an identical manner.

**Corollary 2.** For the choice of \( \delta_n \) given in Proposition 6,

\[
\limsup_{n \to \infty} \sup_{\mu \in F} E_n^* \left[ \delta_n^{-1} d(\hat{S}_n, S_0) \right] < \infty. \tag{4.19}
\]

**Remark 5.** The best rate at which the distance \( d(\hat{S}_n, S_0) \) goes to zero corresponds to \( h_n \sim (nh_n)^{-1/(2p)} \) which yields \( \nu_n \sim h_n = h_0 n^{-1/2(p+1)} \). This is slower than the rate we deduced in the dose-response setting in terms of the total budget \( (N^{-2/(2p+3)}) \). The difference in the rate from the dose-response setting is accounted for by the bias in the smoothed kernel estimates. The regression setting is approximately equivalent to a dose-response model having \( (2Lh_n)^{-2} \) (effectively) independent covariate observations and \( n(2Lh_n)^2 \) (biased) replications. These replications correspond to the number of observations used to compute \( \hat{\mu} \) at a point. If we compare Lemmas 1 and 2, these biased replications add an additional term of order \( h_n \) which is absent in the dose-response setting. This puts a lower bound on the rate at which the set \( S_0 \) can be approximated. In contrast, the rates coincide for the dose-response and the regression settings in the one-dimensional case; see Mallik et al. (2013) and Mallik, Banerjee and Sen (2013). This is due to the fact that in one dimension, the bias to standard deviation ratio \( (b_n/(1/\sqrt{nb_n})) \), where \( b_n = h_n \) is the volume of the bin, is of smaller order compared to that in two dimensions \( (\sqrt{b_n}/(1/\sqrt{nb_n})) \) for estimating \( \hat{\mu} \) (since in the 2d case \( b_n = h_n^2 \), \( h_n \) being the bandwidth). In a nutshell, the curse of dimensionality kicks in at dimension 2 itself in this problem.

### 4.3. Extension to the case of an unknown \( \tau_0 \)

While we deduced our results under the assumption of a known \( \tau_0 \), in real applications \( \tau_0 \) is generally unknown. Quite a few extensions are possible in this situation. For example, in the dose-response setting, if \( S_0 \) can be safely assumed to contain a positive \( F \)-measure set \( U \), then a simple averaging of the \( \bar{Y} \) values realized for \( X \)'s in \( U \) would yield a \( \sqrt{mn} \)-consistent estimator of \( \tau_0 \). If a proper choice of \( U \) is not available, one can obtain an initial estimate of \( \tau_0 \) in the dose–response setting as

\[
\hat{\tau}_{init} = \arg\min_{\tau \in \mathbb{R}} \mathbb{P}_n \left[ \Phi \left( \sqrt{m}(\bar{Y} - \tau) \right) - \frac{1}{2} \right]^2.
\]

This provides a consistent estimate of \( \tau_0 \) under mild assumptions. A \( \sqrt{mn} \)-consistent estimate of \( \tau_0 \) can then be found by using \( \hat{\tau}_{init} \) to compute \( \hat{S}_n \) and then averaging the \( \bar{Y} \) value for the \( X \)'s realized in \( \delta \hat{S}_n \) for a small \( \delta > 0 \). Note that this leads to an iterative procedure where this new estimate of \( \tau \) is used to update the estimate of \( \hat{S}_n \). It can be shown that the rate of convergence remains
unchanged if one imputes a $\sqrt{mn}$-consistent estimate of $\tau_0$. A brief sketch of the following result is given in Section A.6.

**Proposition 1.** Let $\hat{S}_n$ now denote the minimizer of

$$M_n(S, \hat{\tau}) = \mathbb{P}_n \left[ \left\{ \Phi \left( \sqrt{m}(\hat{Y} - \hat{\tau}) \right) \right\} 1_S(X) \right],$$

where $\sqrt{mn}(\hat{\tau} - \tau_0) = O_p(1)$. For $m = m_0 n^\beta$ and $\delta_n$ as defined in Theorem 3, we have $P\left[d(\hat{S}_n, S_0) > \delta_n\right] \to 0$.

In the regression setting as well, an initial consistent estimate of $\tau_0$ can be computed as

$$\hat{\tau}_{\text{init}} = \arg\min_{\tau \in \mathbb{R}} \frac{1}{n} \sum_{k,l} \left[ \Phi \left( \sqrt{n}h_n^2(\hat{\mu}(x_{kl}) - \tau) \right) - \frac{1}{2} \right]^2$$

which can then be used to yield a $\sqrt{n}$-consistent estimate of $\tau_0$ using the iterative approach mentioned above. We have the following result for the rate of convergence of $\hat{S}_n$ in the regression setting.

**Proposition 2.** Let $\hat{S}_n$ now denote the minimizer of

$$M_n(S, \hat{\tau}) = \frac{1}{n} \sum_{k,l} \left[ \Phi \left( \sqrt{n}h_n^2(\hat{\mu}(x_{kl}) - \hat{\tau}) \right) - \gamma \right] 1_S(X)$$

where $\sqrt{n}(\hat{\tau} - \tau_0) = O_p(1)$. For $\delta_n$ as defined in Theorem 6, $P\left[d(\hat{S}_n, S_0) > \delta_n\right] \to 0$.

The proof is outlined in Section A.7 of the Appendix.

5. Discussion

*Extensions to non-convex baseline sets.*** Although we essentially address the situation where the baseline set is convex for dimension $d = 2$, our approach extends past convexity and the two-dimensional setting in the presence of an efficient algorithm and for suitable collections of sets. For example, let $\tilde{S}$ denote such a collection of subsets of $[0, 1]^d$ sets such that

$$\tilde{S}_n = \arg\min_{S \in \tilde{S}} M_n(S)$$

is easy to compute. Here, $\mu$ is a real-valued function from $[0, 1]^d$ and $S_0 = \mu^{-1}(\tau_0)$ is assumed to belong to the class $\tilde{S}$. Then the estimator $\tilde{S}_n$ has the following properties in the dose-response setting.

**Proposition 3.** Assume that $S_0$ is the unique minimizer (up to $F$-null sets) of the population criterion function $M_F$ defined in (4.2). Then $d_F(\tilde{S}_n, S_0)$ converges in probability to zero. Moreover, assume that there exists a constant $\tilde{c} > 0$ such that $F(S^{\setminus \epsilon} \setminus S) \leq \epsilon$ for any $\epsilon > 0$ and $S_0 \in \tilde{S}$, and

$$H_B(u, \tilde{S}, L_2(P)) \leq u^{-r}$$

for some $r < 2$. 
Then, \( P\left(d_F(\tilde{S}_n, S_0) > \tilde{\delta}_n\right) \) converges to zero where \( \tilde{\delta}_n = K_1 \max(n^{-2/(2+r)}, m^{-1/(2p)}) \) for some \( K_1 > 0 \).

The proof follows along lines identical to that for Theorem 3. Note that the condition \( F(S \setminus \epsilon, S) \leq \epsilon \) was needed to derive Lemma 1. This assumption simply rules out sets with highly irregular or non-rectifiable boundaries. Also, the dependence of the rate on the dimension typically comes through \( r \) which usually grows with \( d \). A similar result can be established in the regression setting as well.

Connection with level-set approaches. Note that minimizing \( M_n(S) \) in the dose-response setting is equivalent to minimizing

\[
\tilde{M}_n(S) = M_n(S) - \frac{1}{2} \sum_{i=1}^{N} \left( \frac{1}{4} - p_{m,n}(X_i) \right)
\]

This form is similar to an empirical risk criterion function that is used in Willett and Nowak (2007, equation (7)) in the context of a level-set estimation procedure. It can be deduced that our baseline estimation approach ends up finding the level set \( S_m = \{x : E[p_{m,n}(x)] > 1/4\} \) from i.i.d. data \( \{p_{m,n}(X_i), X_i\}_{i=1}^{n} \) with \( 0 \leq p_{m,n}(X_i) \leq 1 \). As \( m \to \infty \), \( S_m \)'s decrease to \( S_0 \), which is the target set. Hence, any level-set approach could be applied to transformed data \( \{p_{m,n}(X_i), X_i\}_{i=1}^{n} \) to yield an estimate for \( S_m \) which would be consistent for \( S_0 \). Moreover, a similar connection between the two approaches can be made for the regression setting, however the i.i.d. flavor of the observations present in the dose-response setting is lost as \( \{p_n(x_{kl})\}_{1 \leq k,l \leq m} \) are dependent. While the algorithm from Willett and Nowak (2007) can be implemented to construct the baseline set estimate, it is far from clear how the theoretical properties would then translate to our setting given the dependence of the target function \( E[p_{m,n}(x)] \) on \( m \) in the dose-response setting and the dependent nature of the transformed data in the regression setting.

In Scott and Davenport (2007), the approach to the level set estimation problem, using the criterion in Willett and Nowak (2007), is shown to be equivalent to a cost-sensitive classification problem. This problem involves random variables \((X, Y, C) \in \mathbb{R}^d \times \{0,1\} \times \mathbb{R}\), where \( X \) is a feature, \( Y \) a class and \( C \) is the cost for misclassifying \( X \) when the true label is \( Y \). Cost sensitive classification seeks to minimize the expected cost

\[ R(G) = E(C \ 1(G(X) \neq Y)) \]

where \( G \), with a little abuse of notation, refers both to a subset of \( \mathbb{R}^d \) and \( G(x) = 1(x \in G) \). With \( C = |\gamma - Y| \) and \( \tilde{Y} = 1(Y \geq \gamma) \), the objective of the cost-sensitive classification, based on \((X, \tilde{Y}, C)\), can be shown to be equivalent to minimizing the excess risk criterion in Willett and Nowak (2007). So, approaches like support vector machines (SVM) and \( k \)-nearest neighbors (\( k \)-NN),
which can be tailored to solve the cost-sensitive classification problem (see Scott and Davenport (2007)), are relevant to estimating level sets, and thus provide alternative ways to solve the baseline set estimation problem. Since the loss function in (5.1) is not smooth, one might prefer to work with surrogates. Some results in this direction can be found in Scott (2011).

Adaptivity. We have assumed knowledge of the order of the regularity $p$ of $\mu$ at $\partial S_0$, which is required to achieve the optimal rate of convergence, though not for consistency. The knowledge of $p$ dictates the allocation between $m$ and $n$ in the dose-response setting and the choice of the bandwidth $h_n$ in the regression setting for attaining the best possible rates. When $p$ is unknown, the adaptive properties of dyadic trees (see Willett and Nowak (2007) and Singh, Scott and Nowak (2009)) could conceivably be utilized to develop a near-optimal approach. However, this is a hard open problem and will be a topic of future research.

Appendix A: Proofs

A.1. Proof of Theorem 1

Here, we establish consistency with respect to the (stronger) Hausdorff metric,

$$d_H(S_1, S_2) = \max \left[ \sup_{x \in S_1} \rho(x, S_2), \sup_{x \in S_2} \rho(x, S_1) \right].$$

Moreover, we would only require $\min(m, n) \to \infty$ instead of taking $m$ to be of the form $m_0 n^\beta$, $\beta > 0$. To exhibit the dependence on $m$, we will denote $M_n$ by $M_{m,n}$. Recall that $M_m(S) = E[M_{m,n}(S)]$ converges to $M(S)$ for each $S \in S$. Also, $\text{Var}(M_{m,n}(S)) = (1/n)\text{Var}((\Phi(\sqrt{m}Y_1) - \gamma)1_S(X)) \leq 1/n$ which converges to zero. Hence, $M_{m,n}(S)$ converges in probability to $M(S)$ for any $S \in S$, as $\min(m, n) \to \infty$.

The space $(S, d_H)$ is compact (Blaschke Selection theorem) and $M$ is a continuous function on $S$. The desired result will be a consequence of argmin continuous mapping theorem (van der Vaart and Wellner, 1996, Theorem 3.2.2) provided we can justify that $\sup_{S \in S} |M_{m,n}(S) - M(S)|$ converges in probability to zero. To this end, let

$$M^1_{m,n}(S) = M_{m,n}(S) + P_n \gamma 1_X(S) = P_n \Phi(\sqrt{mY_1})1_S(X)$$

and $M^1(S) = M(S) + P \gamma 1_X(S)$. Note that

$$\sup_{S \in S} |M_{m,n}(S) - M(S)| \leq \sup_{S \in S} |(P_n - P)(S)| + \sup_{S \in S} |M^1_{m,n}(S) - M^1(S)|.$$

The first term in the above expression converges in probability to zero (Ranga Rao, 1962). As for the second term, note that $M^1_{m,n}(S)$ converges in probability to $M^1(S)$ for each $S$ and $M^1_{m,n}$ is monotone in $S$, i.e., $M^1_{m,n}(S_1) \leq M^1_{m,n}(S_2)$ whenever $S_1 \subset S_2$. As the space $(S, d_H)$ is compact, there exist
Hence, \( \sup_{S \in \mathcal{S}} |M^1_{m,n}(S) - M^1(S)| \)
\[ = \max_{1 \leq l \leq l(\delta)} \sup_{d_H(S,S(l)) < \delta} |M^1_{m,n}(S) - M^1(S)| \]
\[ \leq 2 \max_{1 \leq l \leq l(\delta)} \sup_{d_H(S,S(l)) < \delta} |M^1_{m,n}(S) - M^1(S(l))| \]
\[ \leq 2 \max_{1 \leq l \leq l(\delta)} (|M^1_{m,n}(S(l)) - M^1(S(l))|, |M^1_{m,n}((S(l))^{\delta}) - M^1(S(l))|). \]

The right side in the above display converges in probability to
\[ 2 \max_{1 \leq l \leq l(\delta)} \max(|M^1(S(l)) - M^1(S(l)|, |M^1((S(l))^{\delta}) - M^1(S(l))|) \]

can be made arbitrarily small by choosing small \( \delta \) (as \( M^1 \) is continuous). Also, as the map \( S \to d_F(S, S_0) \) from \( (S, d_H) \) to \( \mathbb{R} \) is continuous, we have consistency in the \( d_F \) metric as well. This completes the proof. \( \square \)

A.2. Proof of Theorem 4

In light of what has been derived in the proof of Theorem 1, it suffices to show that \( M_n(S) \) converges in probability to \( M(S) \). Note that,

\[ E \left[ \Phi \left( \sqrt{n h_n^2 \mu(x_{kl})} \right) \right] = \Phi_n \left( \frac{\sqrt{n h_n^2 \mu(x_{kl})}}{\sqrt{1 + \Sigma_n^2(x_{kl})}} \right). \]

For \( x \in \{(x_1, x_2): k/m \leq x_1 < (k+1)/m, l/m \leq x_2 < (l+1)/m\} \), let

\[ \hat{f}_n(x) = \hat{W}_{kl} \text{ and } f_n(x) = E \left[ \hat{f}_n(x) \right] = \Phi_n \left( \frac{\sqrt{n h_n^2 \mu(x_{kl})}}{\sqrt{1 + \Sigma_n^2(x_{kl})}} \right) - \gamma. \]

Then \( M_n(S) = \int_S f_n(x) dx \). For any fixed \( x \) in the interior of the set \( S_0 \), \( f_n(x) = \Phi_n(0) \) for sufficiently large \( n \) which converges to \( 1/2 \). As \( \mu \) is continuous, for any fixed \( x \notin S_0, \mu_{x,\delta_x} = \inf \{ \mu(y): \rho(x,y) < \delta_x \} > 0 \) for some \( \delta_x > 0 \). Hence \( f_n(x) \geq \Phi(\sqrt{n h_n^2 \mu_{x,\delta_x}}) \) converges to 1. Also, \( |f_n(x)| \leq 1 \) and hence, \( M_n(S) \) converges to \( M(S) \) by the Dominated convergence theorem.

Moreover,

\[ \text{Var}(M_n(S)) \leq \frac{1}{n^2} \sum_{k,l,k',l'} \text{Cov} \left( \hat{f}_n(x_{k,l}), \hat{f}_n(x_{k',l'}) \right). \]

As \( |\hat{f}_n(x_{k,l})| \leq 1 \), and \( \hat{f}_n(x_{k,l}) \) and \( \hat{f}_n(x_{k',l'}) \) are independent whenever \( \min\{|k-k'|, |l-l'|\} > 2L_0 m h_n \), we have

\[ \sum_{k',l'} \text{Cov} \left( \hat{f}_n(x_{k,l}), \hat{f}_n(x_{k',l'}) \right) \leq (m h_n)^2 = n h_n^2, \]

for any fixed \( k \) and \( l \). Hence, \( \text{Var}(M_n(S)) \) is bounded (up to a constant) by \( n(n h_n^2)/n^2 \) which converges to zero. Hence, \( M_n(S) \) converges in probability to \( M(S) \), which completes the proof. \( \square \)
A.3. Proof of Lemma 2

Let $\text{Bin}_{kl} = \{x = (x_1, x_2) : k/m \leq x_1 < (k+1)/m, l/m \leq x_2 < (l+1)/m\}$. Recall that

$$M(S) - M(S_0)$$

$$= \int \left\{ (1/2)1_{S_0}(x) + 1_{S_0^c}(x) - \gamma \right\} \{1_{S}(x) - 1_{S_0}(x)\} \, dx$$

$$= \sum_{0 \leq k,l \leq (m-1)} \int_{\text{Bin}_{kl}} \left\{ (1/2)1_{S_0}(x) + 1_{S_0^c}(x) - \gamma \right\} \{1_{S}(x) - 1_{S_0}(x)\} \, dx$$

$$= \sum_{0 \leq k,l \leq (m-1)} \left[ \int_{\text{Bin}_{kl} \cap (S_0 \cap S^c)} \left\{ \frac{1}{2} - \gamma \right\} \, dx + \int_{\text{Bin}_{kl} \cap (S_0^c \cap S)} \{1 - \gamma\} \, dx \right]$$

$$= \sum_{k,l : x_{kl} \in \mathcal{I}_n} \left[ \int_{\text{Bin}_{kl} \cap (S_0 \cap S^c)} \left\{ \frac{1}{2} - \gamma \right\} \, dx + \int_{\text{Bin}_{kl} \cap (S_0^c \cap S)} \{1 - \gamma\} \, dx \right] + e_n^{(1)}.$$ 

Here, $e_n^{(1)}$ is the remainder term arising out of replacing the sum of all choices of $k$ and $l$ to sum over $\{(k, l) : x_{kl} \in \mathcal{I}_n\}$. As the integrands in the above sum are bounded by 1, $|e_n^{(1)}| \leq \lambda([0,1]^2 \setminus \mathcal{I}_n) = O(h_n)$. Also,

$$M_n(S) - M_n(S_0)$$

$$= \frac{1}{n} \sum_{k,l} \left\{ \Phi_n \left( \frac{\sqrt{nh_n^2 \bar{\mu}(x_{kl})}}{\sqrt{1 + \Sigma_n^2(x_{kl})}} \right) - \gamma \right\} \{1_{S}(x_{kl}) - 1_{S_0}(x_{kl})\}$$

$$= \sum_{k,l} \int_{x \in \text{Bin}_{kl}} \left\{ \Phi_n \left( \frac{\sqrt{nh_n^2 \bar{\mu}(x_{kl})}}{\sqrt{1 + \Sigma_n^2(x_{kl})}} \right) - \gamma \right\} \{1_{S}(x_{kl}) - 1_{S_0}(x_{kl})\} \, dx.$$ 

Consequently,

$$(\bar{M}_n - M)(S) - (\bar{M}_n - M)(S_0)$$

$$= \sum_{k,l} \int_{x \in \text{Bin}_{kl} \cap (S_0 \cap S^c)} \left( \frac{\sqrt{nh_n^2 \bar{\mu}(x_{kl})}}{\sqrt{1 + \Sigma_n^2(x_{kl})}} - \frac{1}{2} \right) \, dx$$

$$+ \sum_{k,l} \int_{x \in \text{Bin}_{kl} \cap (S_0^c \cap S)} \left( \frac{\sqrt{nh_n^2 \bar{\mu}(x_{kl})}}{\sqrt{1 + \Sigma_n^2(x_{kl})}} - 1 \right) \, dx$$

$$+ e_n^{(2)}, \quad \text{(A.2)}$$

where $e_n^{(2)}$, the contribution of the terms at $\partial(S_0 \cap S)$ along with $e_n^{(1)}$, is bounded by

$$|e_n^{(1)}| + \sum_{k,l : \text{Bin}_{kl} \cap \partial(S_0 \cap S) \neq \emptyset} \int_{\text{Bin}_{kl}} 2 \, dx.$$
Note that $k$ uniformly in $\{Z_n : \rho(x, \partial S_0) < 2/m\}$ which is at most $2c_0(2/m) + 2c_0(2/m) = 8c_0/m$ using (4.3) ($\lambda(x : \rho(x, \partial S) < \alpha)$ for any $\alpha > 0$). Hence, for some $\tilde{c}_1 > 0$,

$$|e_n^{(2)}| \leq O(h_n) + 8c_0/m \leq \tilde{c}_1h_n.$$ 

This contribution is accounted for in the last term of (4.12).

We now study the contribution of the other terms in the right side of (A.2). Note that the integrand in the first sum in the right side of (A.2) is precisely $\Phi_n(0) - 1/2$ whenever $\text{Bin}_{kl} \subset (L_0h_n)$ and $\mu(x_{kl})$ is zero. As the integrand is also bounded by 1, the first sum in the right side of (A.2) is then bounded by

$$|\Phi_n(0) - 1/2| \lambda((S_0 \setminus (L_0h_n)) \cap S) \leq |\Phi_n(0) - 1/2| \delta + \min(c_0L_0h_n, \delta).$$

Choosing $c_1 = \tilde{c}_1 + c_0L_0$, the second term on the right side of the above display is also accounted for in the last term in (4.12). Further, let $S_n = \{x : \rho(x, S_0) > a_n\}$. Note that $\lambda(S_n \setminus S_0) \leq c_0a_n$ using (4.3). Hence, the second sum in (A.2) is bounded by

$$\int_{S_n \setminus S_0} 1dx + \sum_{k,l} \int_{x \in \text{Bin}_{kl} \cap (S_n \cap S)} \left| \Phi_n \left( \frac{\sqrt{n\rho^{2}(x_{kl})}}{\sqrt{1 + \Sigma_n^{2}(x_{kl})}} \right) - 1 \right| dx \leq \min(c_0a_n, \delta) + \sum_{k,l} \int_{x \in \text{Bin}_{kl} \cap (S_n \cap S)} \left| \Phi_n \left( \frac{\sqrt{n\rho^{2}(x_{kl})}}{\sqrt{1 + \Sigma_n^{2}(x_{kl})}} \right) - 1 \right| dx,$$

To bound the second term in right side of the above, note that as $x_{kl} \in \mathcal{I}_n$,

$$\mu(x_{kl}) = E(\mu(x_{kl} + h_nZ_n)) \left\{ \frac{1}{nh_n^2} \sum_{r,s:|r|,|s| \leq L_0mh_n} K \left( \left( \frac{r}{m}, \frac{s}{m} \right) \right) \right\},$$

where $Z_n$ is a discrete random variable supported on $\{r/m, s/m) : |r|, |s| \leq L_0mh_n\}$ with mass function $P[Z_n = (r/m, s/m)] \propto K((r/m, s/m))$. Hence, the argument of $\Phi_n$ can be written as

$$\sqrt{n\rho^2}E(\mu(x_{kl} + h_nZ_n)) \sum_{r,s:|r|,|s| \leq L_0mh_n} K \left( \left( \frac{r}{m}, \frac{s}{m} \right) \right) \frac{1}{nh_n^2\sqrt{1 + \Sigma_n^{2}}}.$$

Note that

$$\sum_{r,s:|r|,|s| \leq L_0mh_n} K \left( \left( \frac{r}{m}, \frac{s}{m} \right) \right) \frac{1}{nh_n^2\sqrt{1 + \Sigma_n^{2}}} = \frac{1}{\sqrt{1 + \Sigma_n^{2}}} + o(1),$$

uniformly in $k$ and $l$ for $x_{kl} \in S \cup S_0$. For $x_{kl} \in S_n \cap S_0$ and $a_n < \kappa_0$, when $\rho(x_{kl} + h_nZ_n, S_0) < \kappa_0$, by triangle inequality,

$$\mu(x_{kl} + h_nZ_n) \geq C_0\rho(x_{kl} + h_nZ_n)^p \geq C_0(\rho(x_{kl}, S_0) - \rho(x_{kl}, x_{kl} + h_nZ_n))^p.$$
As $\rho(x_{kl}, x_{kl} + h_n Z_n) \leq 2L_0 h_n$,

$$\mu(x_{kl} + h_n Z_n) > C_0(a_n - 2L_0 h_n)^p.$$ 

On the other hand, when $\rho(x_{kl} + h_n Z_n, S_0) \geq \kappa_0$, $\mu(x_{kl} + h_n Z_n) > \delta_0$. Consequently, for $x_{kl} \in S_n \cap S_0$, we get

$$\left| \tilde{\Phi}_n \left( \frac{\sqrt{nh_n^2} \mu(x_{kl})}{\sqrt{1 + \sum_n(x_{kl})}} \right) - 1 \right| \leq \left| \tilde{\Phi}_n \left( \frac{\sqrt{nh_n^2} C_0(a_n - 2L_0 h_n)^p}{2\sqrt{1 + \Sigma^2}} \right) - 1 \right| + \left| \tilde{\Phi}_n \left( \frac{\sqrt{nh_n^2} \delta_0}{2\sqrt{1 + \Sigma^2}} \right) - 1 \right| .$$

As $\lambda(S_n \cap S) < \delta$, we get the result.

A.4. Proof of Lemma 3

The sum $\sum_{k,l} W_{kl} 1_S(x_{kl})$ can be written as $\sum_{1 \leq i, j \leq \lceil \sqrt{m'} \rceil} R_{i,j}$ where for $k_i$s and $l_j$s defined as in (4.13), each block

$$R_{i,j} = \sum_{k,l:1 \leq k_i, l_j \leq m} W_{k_i l_j} 1_S(x_{k_i l_j})$$

is a sum of $r_{ij}$ many independent random variables with

$$\left\lfloor \frac{m}{\lceil \sqrt{m'} \rceil} \right\rfloor^2 \leq r_{ij} \leq \left\lceil \frac{m}{\lceil \sqrt{m'} \rceil} \right\rceil^2.$$ 

As $m/\sqrt{m'} \rightarrow \infty$,

$$n/(2m') \leq r_{ij} \leq 2n/m',$$

for large $n$, a fact we use frequently in the proofs. Note that $\sum_{1 \leq i, j \leq \lceil \sqrt{m'} \rceil} r_{ij} = n$ and hence, by convexity of $\exp(\cdot)$,

$$\exp \left( \frac{1}{n} \sum_{1 \leq i, j \leq m} W_{k_i l_j} 1_S(x_{k_i l_j}) \right) \leq \sum_{1 \leq i, j \leq \lceil \sqrt{m'} \rceil} \frac{r_{ij}}{n} \exp \left( \frac{R_{i,j}}{r_{ij}} \right).$$

As $|W_{k_i l_j} 1_S(x_{k_i l_j})| \leq 1_S(x_{k_i l_j})$,

$$P \left( \frac{1}{n} \sum_{k,l} W_{k,l} 1_S(x_{kl}) \geq a \right) \leq \sum_{1 \leq i, j \leq \lceil \sqrt{m'} \rceil} \frac{r_{ij}}{n} E \exp \left( \frac{\lambda R_{i,j}}{r_{ij}} - \lambda a \right) \mbox{ and}$$

$$E \exp \left( \frac{\lambda R_{i,j}}{r_{ij}} \right) \leq \exp \left( \frac{\lambda^2}{8r_{ij}} \sum_{k,l:1 \leq k_i, l_j \leq m} (1_S(x_{k_i l_j}))^2 \right).$$
The second bound in the above display is simply the one used in proving Hoeffding’s inequality for independent sequences (Hoeffding, 1963, equation (4.16)). Consequently,

\[
P \left( \frac{1}{n} \left| \sum_{k,l} W_{kl} 1_S(x_{kl}) \right| \geq a \right) \leq e^{-\lambda a} \exp \left( \frac{\lambda^2}{8(n/2m')^2} \max_{1 \leq i,j \leq \sqrt{m'}} \left[ \sum_{k,l:1 \leq k_i, l_j \leq m} 1_S(x_{k_i l_j}) \right] \right)
\]

Picking

\[
\lambda = \max_{1 \leq i,j \leq \sqrt{m'}} \left[ (1/r_{ij}) \sum_{k,l:1 \leq k_i, l_j \leq m} 1_S(x_{k_i l_j}) \right]
\]

and paralleling the above steps to bound \( P \left( (1/n) \sum_{k,l} W_{kl} 1_S(x_{kl}) \leq a \right) \), we get

\[
P \left( \frac{1}{n} \left| \sum_{k,l} W_{kl} 1_S(x_{kl}) \right| \geq a \right) \leq \frac{2 \exp \left[ \frac{-na^2}{16m' \max_{1 \leq i,j \leq \sqrt{m'}} \left[ (1/r_{ij}) \sum_{k,l:1 \leq k_i, l_j \leq m} 1_S(x_{k_i l_j}) \right] \right]}{16m' \max_{1 \leq i,j \leq \sqrt{m'}} \left[ (1/r_{ij}) \sum_{k,l:1 \leq k_i, l_j \leq m} 1_S(x_{k_i l_j}) \right]}.
\]

Using the definition of \( \tilde{d}_n \), the result follows.

### A.5. Proof of Theorem 5

Let \( S = \min\{s \geq 1 : 2^{-s}R \leq \delta_2/2\} \). By means of condition (4.15), we can choose \( C \) to be a constant large enough so that

\[
\sqrt{n/m'} (\delta_1 - \delta_2) \geq 48 \sum_{s=1}^{S} 2^{-s} R H^{1/2} (2^{-s} R, \mathcal{G}, \tilde{d}_n) \vee (1152 \log 2)^{1/2} (4m') R.
\]

We denote the class of functions \( \mathcal{G} \) by \( \{g_\theta : \theta \in \Theta\} \) for convenience. Let \( \{g^s_j\}_{j=1}^{N_s} \) be a minimal \( 2^{-s} R \)-covering set of \( \tilde{S} \), \( s = 0, 1, \ldots \). So, \( N_s = N(2^{-s} R, \tilde{S}, \tilde{d}_n) \). For any \( \theta \in \Theta \), let \( g^s_\theta \) denote approximation of \( g_\theta \) from the collection \( \{g^s_j\}_{j=1}^{N_s} \). As \( |W_{kl}| \leq 1 \), applying Cauchy-Schwartz to each block \( R_{i,j} \) defined in (A.3) and using (A.4) yields

\[
\left| \frac{1}{n} \sum_{k,l} V_{kl} (g_\theta(x_{kl}) - g^S_\theta(x_{kl})) \right| \leq 2\tilde{d}_n(g_\theta, g^S_\theta) \leq \delta_2.
\]
Hence, it suffices to prove the exponential inequality for
\[ P \left( \max_{j=1, \ldots, N_j} \left| \frac{1}{n} \sum_{k,l} V_{k,l} g_j^n S(x_{kl}) \right| \geq \delta_1 - \delta_2 \right). \]

Next, we use a chaining argument. Note that \( S \sum_{s=1}^S (g^n_\theta - g^{s-1}_\theta) \). By triangle inequality,
\[ \hat{d}_n(g^n_\theta, g^{s-1}_\theta) \leq \hat{d}_n(g^n_\theta, g_{\theta}) + \hat{d}_n(g_{\theta}, g^{s-1}_\theta) \leq 3(2^{-s} R). \]

Let \( \eta_s \) be positive numbers satisfying \( \sum_{s \leq S} \eta_s \leq 1 \). Then,
\[
P^* \left( \sup_{\theta \in \Theta_n} \left| \frac{1}{n} \sum_{k,l} V_{k,l} g^n_\theta(x_{kl}) - g^{s-1}_\theta(x_{kl}) \right| \geq \delta_1 - \delta_2 \right) \leq \sum_{s=1}^S P^* \left( \sup_{\theta \in \Theta_n} \left| \frac{1}{n} \sum_{k,l} V_{k,l} g^n_\theta(x_{kl}) - g^{s-1}_\theta(x_{kl}) \right| \geq (\delta_1 - \delta_2) \eta_s \right) \leq \sum_{s=1}^S 2 \exp \left[ 2H(2^{-s} R, \hat{S}, \tilde{d}_n) - \frac{n(\delta_1 - \delta_2)^2 \eta_s^2}{9(16m')2^{-2s} R^2} \right]. \tag{A.5} \]

We choose \( \eta_s \) to be
\[ \eta_s = \frac{6 \sqrt{16m'}2^{-s} R H^{1/2}(2^{-s} R, \hat{S}, \tilde{d}_n)}{\sqrt{n(\delta_1 - \delta_2)}} \sqrt{\frac{2^{-s} \sqrt{s}}{8}}. \]

The rest of the argument is identical to that Lemma 3.2 of van de Geer (2000). It can be shown that \( \sum_{s \leq S} \eta_s \leq 1 \). Moreover, the above choice of \( \eta_s \) guarantees
\[ H(2^{-s} R, \hat{S}, \tilde{d}_n) \leq \frac{n(\delta_1 - \delta_2)^2 \eta_s^2}{36(16m')2^{-2s} R^2}. \]

Hence the bound in (A.5) is at most
\[ \sum_{s=1}^S 2 \exp \left[ - \frac{n(\delta_1 - \delta_2)^2 \eta_s^2}{18(16m')2^{-2s} R^2} \right]. \]

Next, using \( \eta_s \leq 2^{-s} \sqrt{s}/8 \) and that \( n(\delta_1 - \delta_2^2)/(1152(16m') R^2) \geq \log(2) \), it can be shown that the above display is bounded above by
\[
\sum_{s=1}^\infty 2 \exp \left[ - \frac{n(\delta_1 - \delta_2^2 s^2)}{1152(16m') R^2} \right] \leq 2 \left( 1 - \exp \left[ - \frac{n(\delta_1 - \delta_2^2)}{1152(16m') R^2} \right] \right)^{-1} \exp \left[ - \frac{n(\delta_1 - \delta_2^2 s^2)}{1152(16m') R^2} \right] \leq 4 \exp \left[ - \frac{n(\delta_1 - \delta_2^2 s^2)}{1152(16m') R^2} \right].
\]

This completes the proof.
\textbf{A.6. Proof of Proposition 1}

Note that $\sqrt{m}(\hat{\tau} - \tau_0) = Op(1)$. So, given $\alpha > 0$, there exists $L_\alpha > 0$ such that for \( V_{n,\alpha} = [\tau_0 - L_\alpha/\sqrt{m}, \tau_0 + L_\alpha/\sqrt{m}] \), \( P[\hat{\tau} \in V_{n,\alpha}] > 1 - \alpha \). Let $\hat{S}_n(\tau)$ denote the estimate of $S_0$ based on $M_n(S, \tau)$. Then,

$$P^* \left[ \left| \hat{S}_n(\hat{\tau}), S_0 \right| > \delta_n \right] \leq P^* \left[ d(\hat{S}_n(\hat{\tau}), S_0) > \delta_n, \hat{\tau} \in V_{n,\alpha} \right] + \alpha.$$

Following the arguments for the proof of Proposition 3, the outer probability on the right side can be bounded by

$$\sum_{k=0}^{k_n} P^* \left( \inf_{A \in S_{n,k}} M_n(A, \hat{\tau}) - M_n(S_0, \hat{\tau}) \leq 0, \hat{\tau} \in V_{n,\alpha} \right)$$

The is further bounded by:

$$\sum_{k=0}^{k_n} P^* \left( \sup_{A \in S_{n,k}, \tau \in V_{n,\alpha}} \left| (M_n(S, \tau) - M(S)) - (M_n(S_0, \tau) - M(S_0)) \right| > \inf_{A \in S_{n,k}} (M(S) - M(S_0)) \right). \quad (A.6)$$

As before, $M(S) - M(S_0) = c(\gamma)F(S \Delta S_0)$, and hence, (A.6) is bounded by

$$\sum_{k=0}^{k_n} P^* \left( \sup_{A \in S_{n,k}, \tau \in V_{n,\alpha}} \left| (M_n(S, \tau) - M_m(S_0)) - (M_n(S_0, \tau) - M_m(S_0)) \right| > c(\gamma)2^k \delta_n/3 \right)$$

$$+ \sum_{k=0}^{k_n} \left[ \sup_{A \in S_{n,k}, \tau \in V_{n,\alpha}} \left| (M_m(S, \tau) - M_m(S_0, \tau_0)) - (M_m(S_0, \tau) - M_m(S_0, \tau_0)) \right| \geq c(\gamma)2^k \delta_n/3 \right]$$

$$+ \sum_{k=0}^{k_n} \left[ \sup_{A \in S_{n,k}} \left| (M_m - M)(S, \tau_0) - (M_m - M)(S_0, \tau_0) \right| \geq c(\gamma)2^k \delta_n/3 \right]. \quad (A.7)$$

The third term can be shown to be zero for sufficiently large $n$ in the same manner as in the proof of Proposition 3. Note that the first term can be written as

$$\sum_{k=0}^{k_n} P^* \left( \sup_{A \in S_{n,k}, \tau \in V_{n,\alpha}} \left| \mathbb{G}_n g_{n,\tau}(\bar{Y})1_{A \Delta S_0}(X) \right| > c(\gamma)2^{k-1} \delta_n \sqrt{n} / 3 \right), \quad (A.8)$$

where $g_{n,\tau}(y) = \Phi (\sqrt{n}(y - \tau)) - \gamma$. We are now in a position to apply Theorem 2 to each term of (A.8). In the setup of Theorem 2, $N = 2^{k-1} \delta_n \sqrt{n}$ and the concerned class of functions is $\mathcal{G}_{n,k} = \{ g_{n,\tau}(Y)1_B(X) : B = A \Delta S_0, A \in S_{n,k}, \tau \in V_{n,\alpha} \}$. For $B \in \{A \Delta S_0 : A \in S_{n,k}\}$, \( \| g_{n,\tau}1_B \|_{\mathcal{L}_2(P)} \leq \| E_1B(X) \|_{\mathcal{L}_2} \leq (2^{k+1} \delta_n)^{1/2} \). So, we can choose $R = R_{n,k} = (2^{k+1} \delta_n)^{1/2}$. Also,

$$H_B(u, \{A \Delta S_0 : A \in S_{n,k}\}, L_2(P)) \leq A_0 u^{-1}.$$
for some constant $A_0 > 0$. To bound the entropy of the class of functions $\mathcal{T}_n = \{g_{n,\tau}(\cdot) : \tau \in V_{n,\alpha}\}$, let $t_0 - L_\alpha/\sqrt{mn} = t_0 < t_1 < \ldots < t_{r_n} = \tau_0 + L_\alpha/\sqrt{mn}$ be such that $|t_i - t_{i-1}| \leq u/\sqrt{m}$, for $u > 0$. Note that $r_n \leq 4L_\alpha u/\sqrt{m}$. As $\Phi$ is Lipschitz continuous of order 1 (with Lipschitz constant bounded by 1),

$$ |g_{n,\tau}(\tilde{y}) - g_{n,\tau_i}(\tilde{y})| \leq \sqrt{m}|\tau - \tau_i| \leq u,$$

for $\tau \in [\tau_i, \tau_{i+1}]$. Hence,

$$H_B(u, \mathcal{T}_n, L_2(P)) \leq A_1 \log(u/\sqrt{m}),$$

for some constant $A_1 > 0$ and for small $u > 0$. As the class $\mathcal{G}_{n,k}$ is formed by product of the two classes $\mathcal{S}_{n,k}$ and $\mathcal{T}_n$, the bracketing number for $\mathcal{G}_{n,k}$ is bounded above by,

$$H_B(u, \mathcal{G}_{n,k}, L_2(P)) \leq A_0 u^{-1} + A_1 \log(u/\sqrt{m}) \leq A_2 u^{-1}.$$

In light of the above bound on the entropy, the first term in (A.7) can be shown to go to zero by arguing in the same manner as in the proof of Proposition 3.

For the second term in (A.7), note that $|\Phi(\sqrt{m}(\tilde{y} - \tau)) - \Phi(\sqrt{m}(\bar{y} - \tau_0))| \leq \sqrt{m}|\tau_0 - \tau|$. Hence,

$$|\{(M_m(S, \tau) - M_m(S, \tau_0)) - (M_m(S_0, \tau) - M_m(S_0, \tau_0))\}|

= |P_m[\{\Phi(\sqrt{m}(\tilde{y} - \tau)) - \Phi(\sqrt{m}(\bar{y} - \tau_0))\} \{1_S(X) - 1_{S_0}(X)]\}|

\leq \sqrt{m}|\tau_0 - \tau| |P_m[1_S(X) - 1_{S_0}(X)]|.$$

Thus, the second term in (A.7) is bounded by

$$2(L_\alpha/\sqrt{m}) \sup_{S \in \mathcal{S}_{n,k}} P_m[1_S(X) - 1_{S_0}(X)] \leq L_\alpha 2^{k+2} \delta_n \sqrt{m}/n.$$

This is eventually smaller than $c(\gamma) 2^k \delta_n / 3$ and hence, each term in the second sum of (A.7) is eventually zero. As $\alpha$ is arbitrary, we get the result. \hfill \Box

**A.7. Proof of Proposition 2**

Note that $\sqrt{n}(\hat{\tau} - \tau_0) = O_P(1)$. So, given $\alpha > 0$, there exists $L_\alpha > 0$ such that for $V_{n,\alpha} = [\tau_0 - L_\alpha/\sqrt{n}, \tau_0 + L_\alpha/\sqrt{n}]$, $P[\hat{\tau} \in V_{n,\alpha}] > 1 - \alpha$. Let $\hat{S}_n(\tau)$ denote the estimate of $S_0$ based on $M_n(S, \tau)$. We have,

$$P^*\left[ d(\hat{S}_n(\hat{\tau}), S_0) > \delta_n \right] \leq P^*\left[ \delta_n < d(\hat{S}_n(\hat{\tau}), \hat{\tau} \in V_{n,\alpha}) \right] + \alpha.$$

Following the arguments for the proof of Proposition 6, the first term can be bounded by

$$\sum_{k \geq 0, 2^k \delta_n \leq 1} P^*\left( \inf_{A \in \mathcal{S}_{n,k}} M_n(A, \hat{\tau}) - M_n(S_0, \hat{\tau}) \leq 0, \hat{\tau} \in V_{n,\alpha} \right)$$
This is at most

\[ \sum_{k=0}^{k_n} P^* \left( \sup_{A \in S_{n,k}, \tau \in V_{n,\alpha}} \left| (M_{n}(S, \tau) - M(S)) - (M_{n}(S_0, \tau) - M(S_0)) \right| > \frac{\min \lambda(S \Delta S_0)}{c(\gamma)2^k \delta_n} \right). \]

Note that \( M(S) - M(S_0) \geq c(\gamma) \lambda(S \Delta S_0) \) as earlier, and hence (A.9) is bounded by

\[ \sum_{k=0}^{k_n} P^* \left( \sup_{A \in S_{n,k}, \tau \in V_{n,\alpha}} \left| (M_{n}(S, \tau) - M(S)) - (M_{n}(S_0, \tau) - M(S_0)) \right| > \frac{c(\gamma)2^k \delta_n}{\lambda(S \Delta S_0)} \right). \]

Moreover,

\[
\begin{align*}
\left| (M_{n}(S, \tau) - M(S)) - (M_{n}(S_0, \tau) - M(S_0)) \right| & \leq \left| (M_{n}(S, \tau_0) - M(S)) - (M_{n}(S_0, \tau_0) - M(S_0)) \right| \\
& + \left| (M_{n}(S, \tau) - M(S_0, \tau)) - (M_{n}(S_0, \tau) - M_{n}(S_0, \tau_0)) \right|.
\end{align*}
\]

By the Lipschitz continuity of \( \Phi \), we have

\[
\begin{align*}
\left| (M_{n}(S, \tau) - \hat{M}(S, \tau_0)) - (M_{n}(S_0, \tau) - \hat{M}(S_0, \tau_0)) \right| & \leq \sqrt{n} h_n^2 \left| \tau - \tau_0 \right| \left[ \frac{1}{n} \sum_{k,l} |1_{S}(x_{kl}) - 1_{S_0}(x_{kl})| \right] \\
& \leq \sqrt{n} h_n^2 \left| \tau - \tau_0 \right| \left[ \lambda(S \Delta S_0) + O \left( \frac{1}{\sqrt{n}} \right) \right].
\end{align*}
\]

Here, the last step follows from calculations similar to those in the proof of Lemma 2. Consequently, for sufficiently large \( n \),

\[
\begin{align*}
\sup_{A \in S_{n,k}, \tau \in V_{n,\alpha}} \left| (M_{n}(S, \tau) - \hat{M}(S, \tau_0)) - (M_{n}(S_0, \tau) - \hat{M}(S_0, \tau_0)) \right| & \leq (2L_\alpha) h_n \left[ 2^{k+1} \delta_n + O \left( \frac{1}{\sqrt{n}} \right) \right] < \frac{c(\gamma)}{2} 2^k \delta_n.
\end{align*}
\]

Hence,

\[
\begin{align*}
\sum_{k=0}^{k_n} P^* \left( \sup_{A \in S_{n,k}, \tau \in V_{n,\alpha}} \left| (M_{n}(S, \tau) - M(S)) - (M_{n}(S_0, \tau) - M(S_0)) \right| > c(\gamma)2^k \delta_n \right) & \leq \sum_{k=0}^{k_n} P^* \left( \sup_{A \in S_{n,k}, \tau \in V_{n,\alpha}} \left| (M_{n}(S_0, \tau) - M(S_0)) - (M_{n}(S_0, \tau_0) - M(S_0)) \right| > \frac{c(\gamma)}{2} 2^k \delta_n \right).
\end{align*}
\]

The above is a probability inequality based on the criterion with known \( \tau_0 \). This can be shown to go to zero by calculations identical to those in the proof of Proposition 6. As \( \alpha > 0 \) is arbitrary, we get the result. \( \square \)
References

BOLTHAUSEN, E. (1982). On the central limit theorem for stationary mixing random fields. *Ann. Probab.* **10** 1047–1050.

BRONŠTEIN, E. M. (1976). \(\varepsilon\)-entropy of convex sets and functions. *Sibirsk. Mat. Ž.* **17** 508–514, 715.

CAVALIER, L. (1997). Nonparametric estimation of regression level sets. *Statistics* **29** 131–160.

DUDLEY, R. M. (1984). A course on empirical processes. In *École d’été de probabilités de Saint-Flour, XII—1982. Lecture Notes in Math.* **1097** 1–142. Springer, Berlin.

GEPPETTI, P. and BENEMEI, S. (2009). Pain treatment with opioids: achieving the minimal effective and the minimal interacting dose. *Clin. Drug. Investig.* **29** 3–16.

GOLDENSHLUGER, A. and SPOKOINY, V. (2006). Recovering convex edges of an image from noisy tomographic data. *IEEE Trans. Inform. Theory* **52** 1322–1334.

GOLDENSHLUGER, A. and ZEEVI, A. (2006). Recovering convex boundaries from blurred and noisy observations. *Ann. Statist.* **34** 1375–1394.

HÄRDLE, W., PARK, B. U. and TSYBAKOV, A. B. (1995). Estimation of non-sharp support boundaries. *J. Multivariate Anal.* **55** 205–218.

HARTIGAN, J. A. (1987). Estimation of a convex density contour in two dimensions. *J. Amer. Statist. Assoc.* **82** 267–270.

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30.

KOROSTELÈV, A. P. and TSYBAKOV, A. B. (1993). *Minimax theory of image reconstruction. Lecture Notes in Statistics* **82**. Springer-Verlag, New York.

MA, L., GUO, J., WANG, Y., TIAN, Y. and YANG, Y. (2010). Ship detection by salient convex boundaries. In *Image and Signal Processing (CISP), 2010 3rd International Congress on* **1** 202-205.

MALLIK, A., BANERJEE, M. and SEN, B. (2013). Asymptotics for p-value based threshold estimation in regression settings. submitted, available at [http://www.stat.lsa.umich.edu/~moulib/th-estn-reg-v2.pdf](http://www.stat.lsa.umich.edu/~moulib/th-estn-reg-v2.pdf).

MALLIK, A., SEN, B., BANERJEE, M. and MICHAILIDIS, G. (2011). Threshold estimation based on a p-value framework in dose-response and regression settings. *Biometrika* **98** 887–900.

MALLIK, A., BANERJEE, M., SEN, B. and MICHAILIDIS, G. (2013). Asymptotics for p-value based threshold estimation in dose–response settings. submitted, available at [http://www.stat.columbia.edu/~bodhi/research/th-estn-mnJuly13.pdf](http://www.stat.columbia.edu/~bodhi/research/th-estn-mnJuly13.pdf).

MAMMEN, E. and TSYBAKOV, A. B. (1995). Asymptotical minimax recovery of sets with smooth boundaries. *Ann. Statist.* **23** 502–524.

NOLAN, D. (1991). The excess-mass ellipsoid. *J. Multivariate Anal.* **39** 348–371.

POLONIK, W. (1995). Measuring mass concentrations and estimating density contour clusters—an excess mass approach. *Ann. Statist.* **23** 855–881.
Qiu, P. (2007). Jump surface estimation, edge detection, and image restoration. J. Amer. Statist. Assoc. 102 745–756.

Ranga Rao, R. (1962). Relations between weak and uniform convergence of measures with applications. Ann. Math. Statist. 33 659–680.

Rosén, B. (1969). A note on asymptotic normality of sums of higher-dimensionally indexed random variables. Ark. Mat. 8 33–43 (1969).

Scott, C. (2011). Surrogate losses and regret bounds for cost-sensitive classification with example-dependent costs. In Proceedings of the 28th International Conference on Machine Learning (ICML-11). ICML ’11 153–160. ACM, New York, NY, USA.

Scott, C. and Davenport, M. (2007). Regression level set estimation via cost-sensitive classification. IEEE Trans. Signal Process. 55 2752–2757.

Singh, A., Scott, C. and Nowak, R. (2009). Adaptive Hausdorff estimation of density level sets. Ann. Statist. 37 2760–2782.

Stahl, J. S. and Wang, S. (2005). Convex Grouping Combining Boundary and Region Information. In Proceedings of the Tenth IEEE International Conference on Computer Vision - Volume 2. ICCV ’05 946–953. IEEE Computer Society, Washington, DC, USA.

Tsybakov, A. B. (1997). On nonparametric estimation of density level sets. Ann. Statist. 25 948–969.

van de Geer, S. (2000). Empirical processes in M-estimation. Cambridge University Press.

van der Vaart, A. W. and Wellner, J. A. (1996). Weak convergence and empirical processes: with applications to statistics. Springer Series in Statistics. Springer-Verlag, New York.

Wakimoto, R. M. and McElroy, J. L. (1986). Lidar Observation of Elevated Pollution Layers over Los Angeles. J. Climate Appl. Meteor. 25 1583–1599.

Wang, Y. and Woodroofe, M. (2013). A New Condition for the Invariance Principle for Stationary Random Fields. Statist. Sinica 23. To appear.

Wang, S., Stahl, J., Bailey, A. and Dropps, M. (2007). Global Detection of Salient Convex Boundaries. International Journal of Computer Vision 71 337-359.

Willett, R. M. and Nowak, R. D. (2007). Minimax optimal level set estimation. IEEE Trans. Image Process. 16 2965–2979.