Consistent Kernel Density Estimation with Non-Vanishing Bandwidth

Efrén Cruz Cortés and Clayton Scott
Electrical Engineering and Computer Science
University of Michigan

Abstract
Consistency of the kernel density estimator requires that the kernel bandwidth tends to zero as the sample size grows. In this paper we investigate the question of whether consistency is possible when the bandwidth is fixed, if we consider a more general class of weighted KDEs. To answer this question in the affirmative, we introduce the fixed-bandwidth KDE (fbKDE), obtained by solving a quadratic program, and prove that it consistently estimates any continuous square-integrable density. We also establish rates of convergence for the fbKDE with radial kernels and the box kernel under appropriate smoothness assumptions. Furthermore, in an experimental study we demonstrate that the fbKDE compares favorably to the standard KDE and the previously proposed variable bandwidth KDE.

1 Introduction
Given an iid sample \( X_1, \ldots, X_n \in \mathbb{R}^d \) drawn according to a probability density \( f \), the kernel density estimator is

\[
f_{\text{KDE}} = \frac{1}{n} \sum_{i=1}^{n} k(\cdot, X_i),
\]

where \( k \) is a kernel function with parameter \( \sigma \). Examples of kernels are functions of the form \( k(x, x') = \sigma^{-d} g(\sigma^{-2} \| x - x' \|^2) \), where \( \int \sigma^{-d} g(\sigma^{-2} \| x - x' \|^2) dx = 1 \) for all \( x' \). A common kernel for density estimation is the Gaussian kernel \( k(x, x') = (2\pi \sigma^2)^{-d/2} \exp\left(-\frac{1}{2} \| x - x' \|^2 / \sigma^2\right) \). Since its inception by Fix and Hodges [1] and development by Rosenblatt and Parzen [2, 3], the KDE has found numerous applications across a broad range of quantitative fields, and has also been the subject of extensive theoretical investigations, spawning several books (see, e.g., [4, 5, 6, 7]) and hundreds of research articles.

A strength of the KDE is that it makes few assumptions about \( f \) and that it is consistent, meaning that it converges to \( f \) as \( n \to \infty \) [8]. Analysis of the KDE stems from the following application of the triangle inequality in some norm of interest, where \( * \) denotes convolution:

\[
\| f_{\text{KDE}} - f \| \leq \| f_{\text{KDE}} - f * k \| + \| f * k - f \|.
\]

Critical to the analysis of the KDE is the dependence of the bandwidth parameter \( \sigma \) on \( n \). The first term tends to zero provided \( n \sigma^d \to \infty \), i.e. the number of data points per unit volume tends to infinity. This is shown using properties of convolutions (since \( f_{\text{KDE}} \) and \( f * k \) may be viewed as convolutions of the kernel with the empirical and true distributions, respectively) and concentration of measure. For the latter term to tend to zero, \( \sigma \to 0 \) is necessary, so that the kernel behaves like a Dirac measure. With additional assumptions on the smoothness of \( f \), the optimal growth of \( \sigma \) as a function of \( n \) can be determined.

The choice of bandwidth determines the performance of the KDE, and automatically selecting the optimal kernel bandwidth remains a difficult problem. Thus, researchers have developed numerous plug-in rules and cross-validation strategies, all of which are successful in some situations. A recent survey counts no fewer than 30 methods in the literature and cites 6 earlier review papers on the topic [9].

As an alternative to the standard KDE, some authors have investigated weighted KDEs, which have the form \( f_n = \sum_{i=1}^{n} \alpha_i k(\cdot, X_i) \). The weight vector \( \alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n \) is learned according to some

This work was supported in part by NSF grant 1422157.
Contact first author at encc@umich.edu for further questions about this work.
criterion, and such weighted KDEs have been shown to yield improved performance over the standard KDE in certain situations \[10, 11, 12, 13\]. Consistency of such estimators has been investigated, but still under the assumption that \( \sigma \to 0 \) with \( n \in \mathbb{N} \) \[14, 15\].

In this work we consider the question of whether it is possible to learn the weights of a weighted KDE such that the resulting density estimator is consistent, for a broad class of \( f \), where the bandwidth \( \sigma \) remains fixed as \( n \to \infty \). This question is of theoretical interest, given that all prior work establishing consistency of a KDE, to our knowledge, requires that the bandwidth shrinks to zero. The question is also of practical interest, since such a density estimator could potentially be less sensitive to the choice of bandwidth than the standard KDE, which, as mentioned above, is the main factor limiting the successful application of KDEs in practice.

In Section 2 below, we introduce a weighted KDE that we refer to as the fixed-bandwidth KDE (fbKDE). Its connection to related work is given in Section 3. The theoretical properties of this estimator, including consistency and rates of convergence with a fixed bandwidth, are established in Section 4. Our analysis relies on reproducing kernel Hilbert spaces, a common structure in machine learning that has seldom been used to understand KDE’s. In Section 5 a simulation study is conducted to compare the fbKDE against the standard KDE and another weighted KDE from the literature. Our results indicate that the fbKDE is a promising alternative to these other methods of density estimation.

## 2 Fixed bandwidth KDE

We start by assuming access to iid data \( X_1, \ldots, X_n \) sampled from an unknown distribution with density \( f \), having support \( S = \text{supp}\{f\} \) contained in the known set \( \mathcal{X} \subseteq \mathbb{R}^d \), and with dominating measure \( \mu \). The set \( \mathcal{X} \) is either compact, in which case \( \mu \) is taken to be the Lebesgue measure, or \( \mathcal{X} = \mathbb{R}^d \), in which case \( \mu \) is a known finite measure. We study a weighted kernel density estimator of the form

\[
\hat{f}_n = \sum_{i=1}^{n} \alpha_i k(\cdot, Z_i)
\]

where \( \alpha := (\alpha_1, \ldots, \alpha_n)^T \in \mathcal{A}_n \subseteq \mathbb{R}^n \), \( Z_i = X_i + \Gamma_i \), and \( \Gamma_i \) is sampled iid from a known distribution with density \( f \). Here, \( \Gamma \) is chosen to ensure \( Z_i \in \mathcal{X} \), but not necessarily in \( S \). Note that \( \alpha \) is defined on \( \mathcal{X} \) and \( k \) on \( \mathcal{X} \times \mathcal{X} \). Throughout, \( \mathcal{A}_n \) is taken to be an \( \ell_1 \) ball in \( \mathbb{R}^n \), that is \( \mathcal{A}_n = \{ \alpha \in \mathbb{R}^n \mid \|\alpha\|_1 \leq R_n \} \), for \( R_n \in \mathbb{R} \). The reason for centering the kernels at \( Z_i = X_i + \Gamma_i \) instead of \( X_i \) is that to accurately approximate \( f \) with a fixed bandwidth, we might need centers outside the support of \( f \).

To measure the error between \( \hat{f}_n \) to \( f \) we may consider the \( L_2(S) \) difference, where \( L_2(S) \) is the space of equivalence classes of square integrable functions, and where (using both \( f \) for the function and its equivalence class) \( \|f\|^2_{L_2(S)} := \int_S f^2 \, d\mu \). However, we do not know the set \( S \) and cannot compute said difference. Hence, we consider the \( L_2(\mathcal{X}) \) difference from \( f \).

To determine the scaling coefficients \( \alpha \) we consider minimizing \( \|f - \hat{f}_n\|^2_{L_2(\mathcal{X})} \) over \( \alpha \in \mathcal{A}_n \). Since \( \|f - \hat{f}_n\|^2_{L_2(\mathcal{X})} = \int_{\mathcal{X}} f^2 \, d\mu - 2 \int_{\mathcal{X}} f \hat{f}_n \, d\mu + \int_{\mathcal{X}} \hat{f}_n^2 \, d\mu \), the term \( \int_{\mathcal{X}} f^2 \) is independent of \( \alpha \), and \( f \) is zero outside \( S \), we can focus on minimizing

\[
J(\alpha) = \int_{\mathcal{X}} f \hat{f}_n \, d\mu - 2 \int_{\mathcal{X}} \hat{f}_n \hat{f}_n \, d\mu.
\]

Define \( H(\alpha) := \int_{\mathcal{S}} f \hat{f}_n \, d\mu = \sum_{i=1}^{n} \alpha_i \int_{\mathcal{S}} k(x, Z_i) \, d\mu(x) =: \sum_{i=1}^{n} \alpha_i h_i \). Since we don’t know \( f \), we don’t know the true form of \( H(\alpha) \) and \( J(\alpha) \). However, the terms \( h_i \) are expectations with respect to \( f \) so we can estimate the term \( H(\alpha) \) using the available data \( \{X_i\}_{i=1}^{n} \). We use the leave-one-out estimator

\[
\hat{H}_n(\alpha) := \sum_{i=1}^{n} \alpha_i \frac{1}{n-1} \sum_{j \neq i} k(X_j, Z_i) =: \sum_{i=1}^{n} \alpha_i \hat{h}_i.
\]

With the aid of \( \hat{H}_n(\alpha) \), we define the function

\[
J_n(\alpha) := \int_{\mathcal{X}} f \hat{f}_n \, d\mu - 2 \hat{H}_n(\alpha)
\]

(2)

to which we have access. Let \( J^* := \inf_{\alpha \in \mathcal{A}_n} J(\alpha) \) and

\[
\alpha(\sigma) := \arg \min_{\alpha \in \mathcal{A}_n} J_n(\alpha).
\]

(3)
The estimator

$$f_{\alpha(n)} = \sum_{i=1}^{n} \alpha_i(k(\cdot, Z_i))$$

(4)

is referred to as the fixed bandwidth kernel density estimator (fbKDE), and is the subject of our study. In the following sections this estimator is shown to be consistent for a fixed kernel bandwidth \(\sigma\) under certain conditions on \(f, k, R_0,\) and \(f^r\).

3 Related work

The use of the \(L^2\) norm as an objective function for kernel density estimation is not new, and in fact, choosing \(\sigma\) to minimize \(J_n\), with \(\alpha_i = 1/n\), is the so-called least-squares leave-one-out cross-validation technique for bandwidth selection. Minimizing \(J_n\) subject to the constraints \(\alpha_i \geq 0\) and \(\sum \alpha_i = 1\) was proposed by [10], and later rediscovered by [11], who also proposed an efficient procedure for solving the resulting quadratic program, and compared the estimator to the KDE experimentally. This same estimator was later analyzed by [14], who established an oracle inequality and consistency under the usual conditions for consistency of the KDE.

Weighted KDEs have also been developed as a means to enhance the standard KDE in various ways. For example, a weighted KDE was proposed in [12] as a form of multiple kernel learning, where, for every data point, multiple kernels of various bandwidths were assigned, and the weights optimized using \(J_n\). A robust kernel density estimator was proposed in [13], by viewing the standard KDE as a mean in a RKHS of \(X\), and for which the weights are constrained to be nonnegative. In contrast, we allow the weights on individual kernels to be negative, and this enables our theoretical analysis below. Finally, the resulting quadratic program, and compared the estimator to the KDE experimentally. This same weights performs as well as the KDE whose bandwidth is chosen by an oracle.

In the above-cited work on weighted KDEs, whenever consistency is shown, it assumes a bandwidth which there is a strictly monotone function \(\sigma_i\) tending to zero. Furthermore, the weights are constrained to be nonnegative. In contrast, we allow the weights on individual kernels to be negative, and this enables our theoretical analysis below. Finally, we remark that the terms “fixed” or “constant” bandwidth have been used previously in the literature to refer to a KDE where each data point is assigned the same bandwidth, as opposed to a “variable bandwidth” KDE where each data point might receive a different bandwidth. We instead use “fixed bandwidth” to mean the bandwidth remains fixed as the sample size grows.

4 Theoretical Properties of the fbKDE

Notation The space \(L^p_n(X)\) is the set of functions on \(X\) for which the \(p\)th power of their absolute value is \(\nu\)-integrable over \(X\). \(L^p_n(X)\) is the set of equivalence classes of \(L^p_n(X)\), where two functions \(g_1\) and \(g_2\) are equivalent if \(\int_X |g_1 - g_2|^p d\nu = 0\). The symbol \(\| \cdot \|_2\) will denote both the Euclidean norm in \(\mathbb{R}^d\) as well as the \(L^2\) norm; which is used will be clear from the context, as the elements of \(\mathbb{R}^d\) will be denoted by letters towards the end of the alphabet (\(x, y,\) and \(z\)). The set \(C(X)\) denotes the space of continuous functions on \(X\). Finally, the support of any function \(g\) is denoted by \(\text{supp} \{g\}\).

We call a function \(k : X \times X \to \mathbb{R}\) a symmetric positive definite kernel (SPD) [22] if \(k\) is symmetric and satisfies the property that for arbitrary \(n\) and \(\{x_i\}_{i=1}^n\) the matrix \([k(x_i, x_j)]_{i,j=1}^n\) is positive semidefinite. Every SPD kernel has an associated Hilbert space \(H_k(X)\) of real-valued functions, called the reproducing kernel Hilbert space (RKHS) of \(k\), which we will denote by \(H\) when \(k\) and \(X\) are clear from the context, and for which \(k(x, x') = \langle k(\cdot, x), k(\cdot, x') \rangle_H\). SPD kernels also exhibit the reproducing property, which states that for any \(h \in H\), \(h(x) = \langle h, k(\cdot, x) \rangle_H\). By a radial SPD kernel we mean an SPD kernel for which there is a strictly monotone function \(g\) such that \(k(x, x') = g(\|x - x'\|_2)\) for any \(x, x' \in X\). Note that the Gaussian kernel is a radial SPD kernel.

If \(k\) is a radial SPD kernel then \(\sup_{x,x' \in X} k(x, x') \leq C_k\) for some \(C_k > 0\). This holds because \(k(x, x') = \langle k(\cdot, x), k(\cdot, x') \rangle \leq \|k(\cdot, x)\| \|k(\cdot, x')\|\) by the reproducing property and Cauchy-Schwarz, and \(\|k(\cdot, x)\| \|k(\cdot, x')\| = \sqrt{k(x, x')} \sqrt{k(x', x')} = g(0)\).

We will make use of the following assumptions:
A0 The set $\mathcal{X}$ is either a compact subset of $\mathbb{R}^d$, in which case $\mu$ is taken to be the Lebesgue measure, or $\mathcal{X} = \mathbb{R}^d$, in which case $\mu$ is a known finite measure.

This assumption is always held throughout the paper. It will not be explicitly stated in the statements of results below, but will be remembered in the proofs when needed.

A1 $X_1, \ldots, X_n$ are sampled independently and identically distributed (iid) according to $f$. $\Gamma_1, \ldots, \Gamma_n$ are sampled iid according to $f_\Gamma$. Furthermore $X_1, \ldots, X_n, \Gamma_1, \ldots, \Gamma_n$ are independent.

Given $\{(X_i, \Gamma_i)\}_{i=1}^n$ as in A1 we define $Z_i := X_i + \Gamma_i$ for $1 \leq i \leq n$. This notation will be kept throughout the paper.

A2 The kernel $k$ is radial and SPD, with $k(x, x) = C_k$ for all $x \in \mathcal{X}$. Furthermore, $k$ is Lipschitz, that is, there exists $L_k > 0$ such that $\|k(x, \cdot) - k(y, \cdot)\|_2 \leq L_k \|x - y\|_2$ holds for all $x, y \in \mathcal{X}$.

Recall from eqn. (3) that $\alpha^{(n)}$ is the minimizer of $J_n$ over $A_n$. To show the consistency of $f_\alpha$ the overall approach of the following sections will be to show that $J(\alpha^{(n)})$ is close to $J^* = \inf_{\alpha \in A_n} \| f - f_\alpha \|_2^2$ with high probability, and then show that $J(\alpha)$ (and therefore $J^*$) can be made arbitrarily small for optimal choice of $\alpha$. We start by stating an oracle inequality relating $J(\alpha^{(n)})$ and $J^*$.

**Lemma 1.** Let $\epsilon > 0$. Let $\{(X_i, \Gamma_i)\}_{i=1}^n$ satisfy assumption A1. Let $k$ satisfy sup$_{x,x' \in \mathcal{X}} k(x, x') \leq C_k$ and $f_\alpha$ be as in Equation (1). Let $\delta = 2n \exp \left(-\frac{(n-1)^2}{8C_k^2} \right)$. With probability $\geq 1 - \delta$,

$$
\| f - f_\alpha^{(n)} \|_2^2 \leq \epsilon + \inf_{\alpha \in A_n} \| f - f_\alpha \|_2^2.
$$

The proof consists of showing the terms $H_n$ of $J_n$ concentrate around the terms $H$ of $J$, and is placed in Section 7. This result allows us to focus on the term $J^* = \inf_{\alpha \in A_n} \| f - f_\alpha \|_2^2$, which we proceed to do in the following sections.

### 4.1 Consistency of $f_\alpha^{(n)}$

We state the consistency of $f_\alpha^{(n)}$ for radial SPD kernels:

**Theorem 1.** Let $\epsilon > 0$. Let $\{(X_i, \Gamma_i)\}_{i=1}^n$ satisfy A1 where $f \in \mathcal{F}(\mathcal{X}) := L^2(\mu) \cap C(\mathcal{X})$ and $\{f_Z\} \supseteq \mathcal{X}$. Let $k$ satisfy A2 and $f_\alpha^{(n)}$ be as in Equation (4). If $A_n$ is such that $R_n \to \infty$ but $R_n^2 \log n / n \to 0$ as $n \to \infty$, then

$$
P_{X, \Gamma} \left\{ \| f - f_\alpha^{(n)} \|_2^2 > \epsilon \right\} \to 0
$$

as $n \to \infty$. \qed

The probability $P_{X, \Gamma} \{ \}$ is the joint probability of $\{(X_i, \Gamma_i)\}_{i=1}^n$. The sketch of the proof is as follows. To analyze the term $J^* = \inf_{\alpha \in A_n} \| f - f_\alpha \|_2^2$ from Lemma 1 we use the fact that $\mathcal{H}$ is dense in $\mathcal{F}$ in the sup-norm sense and that $\mathcal{H}^0 := \left\{ \sum_{j=1}^N c_j k(\cdot, y_j) | y_j \in \mathcal{X}, c_j \in \mathbb{R} \text{and } N \in \mathbb{N} \right\}$ is dense in $\mathcal{H}$ in the $\mathcal{H}$–norm sense. That is, there exists a function $f_{\mathcal{H}} \in \mathcal{H}$ arbitrarily close to $f$ and a function $f_\beta \in \mathcal{H}^0$ arbitrarily close to $f_{\mathcal{H}}$. The function $f_\beta$ has the form

$$
f_\beta := \sum_{j=1}^m \beta_j k(\cdot, y_j),
$$

where $\beta = (\beta_1, \ldots, \beta_m)^T$. Note this is an abuse of notation since the functions $f_\alpha$ and $f_\beta$ do not have the same centers nor necessarily the same number of components. By the triangle inequality we have for any $\alpha \in A_n$:

$$
\| f - f_\alpha \|_2 \leq \| f - f_{\mathcal{H}} \|_2 + \| f_{\mathcal{H}} - f_\beta \|_2 + \| f_\beta - f_\alpha \|_2.
$$

By the above denseness results, the first two terms are small. To make the third term small we need two things: that $R_n$ is large enough so that there is an $\alpha \in A_n$ matching $\beta$, and that there exists centers $\{Z_i\}_{i=1}^n$ of $f_\alpha$ that are close to the centers $\{y_j\}_{j=1}^m$ of $f_\beta$, which will be true with a certain probability. In Section 7 we will prove all these approximations and show that the relevant probability is indeed large and approaches one.
4.2 Convergence rates of \( f_\alpha(n) \) for SPD radial kernels

The rates for radial SPD kernels may be slow, since these kernels are "universal" in that they can approximate arbitrary functions in \( L^2_\lambda(X) \cap C(X) \). To get better rates, we can make stronger assumptions on \( f \). Thus, let \( \mathcal{F}_k = \{ f \mid f \geq 0 \text{ a.e., } \int f(x) \, d\mu(x) = 1, \ f = \int d(x, x) \lambda(x) \, dx, \ \lambda \in L^1(X) \} \), that is, the space of densities expressible as \( k \)-smoothed \( L_1 \) functions. Then we obtain the following convergence rates.

**Theorem 2.** Let \( \delta \in (0,1) \). Let \( \mathcal{S} = \mathcal{X}, k \) satisfy \( \square \) \( \{ (X_i, \Gamma_i) \}_{i=1}^n \) satisfy \( \square \) with \( f \in \mathcal{F}_k \), and \( \min_{x \in \mathcal{X}} \{ f_\beta(z) \} > 0 \). Let \( f_\alpha(n) \) be as in Equation (4). If \( d > 4 \) and \( R_n \approx \frac{n^{1/2-d/2}}{} \), then with probability \( \geq 1 - \delta \)

\[
\| f - f_\alpha(n) \|_2 \lesssim \left( \frac{1}{n} \right)^{2/d} \log^{1/2} (n/\delta).
\]

If \( d \leq 4 \) and \( R_n \approx \frac{n^{(1-C)/2}}{} \) for \( C \) an arbitrary constant in \( (0,1) \), then with probability \( \geq 1 - \delta \)

\[
\| f - f_\alpha(n) \|_2 \lesssim \left( \frac{1}{n} \right)^{C/2} \log^{2/d} (n/\delta).
\]

The symbol \( \lesssim \) indicates the first term is bounded by a positive constant (independent of \( n \) and \( d \)) times the second term, and \( \sim \) means they grow at the same rate. Note the condition \( \min_{z \in \mathcal{X}} \{ f_\beta(z) \} > 0 \) is satisfied, for example, if \( \mathcal{X} \) is compact and \( f_\beta \) is Gaussian. The first step in proving Theorem 2 is just a reformulation of the oracle inequality from Lemma 1.

**Lemma 2.** Let \( \delta_1 \in (0,1) \). Let \( \{ (X_i, \Gamma_i) \}_{i=1}^n \) satisfy assumption \( \square \) Let \( k \) satisfy assumption \( \square \) and \( f_\alpha \) be as in Equation (4). Then with probability \( \geq 1 - \delta_1 \)

\[
J(\alpha(n)) \leq \epsilon_1(n) + J^*
\]

where \( \epsilon_1(n) := \sqrt{\mathbb{E} c_k R_n \sqrt{\frac{\log (4n/\delta_1)}{n - 1}}} \).

Now, for the \( J^* \) term in Lemma 2 we introduce the function \( f_\beta \) as in eqn (5) and make the following decomposition, valid for any \( \alpha \in \mathcal{A}_n \):

\[
\| f - f_\alpha \|_2 \leq \| f - f_\beta \|_2 + \| f_\beta - f_\alpha \|_2.
\]

The following lemma concerns the term \( \| f - f_\beta \|_2 \), and is taken from [24].

**Lemma 3.** Let \( f \in \mathcal{F}_k \). For any \( m \in \mathbb{N} \) there are \( m \) points \( \{ y_j \}_{j=1}^m \subset \mathcal{X} \) and \( m \) coefficients \( \{ c_j \}_{j=1}^m \subset \mathbb{R} \) such that

\[
\left\| f - \sum_{j=1}^m c_j \| \lambda \|_1 k(\cdot, y_j) \right\|_\infty \leq \epsilon_2(m)
\]

where \( \epsilon_2(m) := C \| \lambda \|_1 \sqrt{V_k} \sqrt{m} \) for some absolute constant \( C \) and where \( V_k \) is the VC-dimension of the set \( \{ k(\cdot, x) \mid x \in \mathcal{X} \} \).

The VC dimension of a set \( \{ y_i \} \) is the maximum number of points that can be separated arbitrarily by functions of the form \( g_r - r, r \in \mathbb{R} \). For radial kernels, \( V_k = d + 1 \) (see [25, 26]).

Now let \( \beta_j = \frac{\epsilon_2(m)}{m} \) and \( f_\beta = \sum_{j=1}^m \beta_j k(\cdot, y_j) \). For the remaining term \( \| f_\beta - f_\alpha \|_2 \) we will proceed as in the proof of Theorem 1. That is, we need that for all \( y_j \) there is a point \( Z_i \) close to it, and then we can approximate \( f_\beta \) with \( f_{\alpha} = \sum_{i=1}^n \alpha_i k(\cdot, Z_i) \).

**Lemma 4.** Let \( \delta_2 \in (0,1) \), let \( f, m, \) and \( f_\beta \) be as above. Let \( \{ (X_i, \Gamma_i) \}_{i=1}^n \) satisfy assumption \( \square \) With probability \( \geq 1 - \delta_2 \)

\[
\inf_{\alpha \in \mathcal{A}_n} \| f_\beta - f_\alpha \|_2 \leq \epsilon_3(n, m)
\]

where \( \epsilon_3(n, m) := \frac{C}{n^{1/d}} \log^{1/d} (m/\delta_2) \), for \( C \) a constant independent of \( n \) and \( m \).

Putting Lemmas 2, 3, and 4 together and choosing \( m = n^\theta \) for an appropriate \( \theta \) (the exact form is shown in Section 7) we obtain the proof of Theorem 2.
different ways. First we used rules of thumb. For variable bandwidth KDE ([27]), which we refer as the vKDE, has the form

We now show the performance of the fbKDE as compared to the KDE and variable bandwidth KDE. The non-vanishing bandwidth.

box kernel we give a novel argument for approximating Lipschitz functions with box kernels having a

Theorem 3. Thus the fbKDE centered at \( \{ y_i \} \) is \( \tilde{f}_\alpha := \sum_{i=1}^M \alpha_i k(\cdot, y_i) \), where the \( \alpha_i \) weights are learned in the same way as before, the only change being the kernel centers. Let \( F = L_1(\mathcal{X}) \cap L_2(\mathcal{X}) \cap \text{Lip}(\mathcal{X}) \), and let \( L_f \) be the Lipschitz constant of \( f \). Also, let \( k \) be the box kernel \( k(x, y) = \frac{1}{(2\sigma)^d} \mathbf{1}_{\|x-y\|_\infty \leq \sigma} \) defined on \( \mathcal{X} \) and \( \mathcal{X} \), and for simplicity assume \( \sigma = \frac{1}{2q} \) for \( q \) a positive integer. The following theorem is proved in Section 7.

Theorem 3. Let \( f \in F \), supp \( \{ f_Z \} \supset \mathcal{X} \), \( \{(X_i, \Gamma_i)\}_{i=1}^n \) satisfy A1 and \( R_n \sim n^{(d-1)/(2d+2)} \). Let \( \delta \in (0, 1) \). With probability at least \( 1 - \delta \)

Figure 1: Bimodal, triangular, trimodal and kurtotic densities used to evaluate the fbKDE performance.

Figure 2: Bimodal density and kernel estimators with training size 800. The stem subplot indicates the values of \( \alpha \) (centered offset for visualization). Note that some of the \( \alpha \) weights are negative.

4.3 Convergence rates for the box kernel

While the previous theorem considered radial SPD kernels, the oracle inequality applies more generally, and in this section we present rates for a nonradial kernel, the box kernel. We assume \( \mathcal{X} = [0, 1]^d \) and that the kernel centers are predetermined according to a uniform grid. Precise details are given in the proof of Theorem 3. Thus the fbKDE centered at \( \{ y_i \} \) is \( \tilde{f}_\alpha := \sum_{i=1}^M \alpha_i k(\cdot, y_i) \), where the \( \alpha_i \) weights are learned in the same way as before, the only change being the kernel centers. Let \( F = L_1(\mathcal{X}) \cap L_2(\mathcal{X}) \cap \text{Lip}(\mathcal{X}) \), and let \( L_f \) be the Lipschitz constant of \( f \). Also, let \( k \) be the box kernel \( k(x, y) = \frac{1}{(2\sigma)^d} \mathbf{1}_{\|x-y\|_\infty \leq \sigma} \) defined on \( \mathcal{X} \) and \( \mathcal{X} \), and for simplicity assume \( \sigma = \frac{1}{2q} \) for \( q \) a positive integer. The following theorem is proved in Section 7.

Theorem 3. Let \( f \in F \), supp \( \{ f_Z \} \supset \mathcal{X} \), \( \{(X_i, \Gamma_i)\}_{i=1}^n \) satisfy A1 and \( R_n \sim n^{(d-1)/(2d+2)} \). Let \( \delta \in (0, 1) \). With probability at least \( 1 - \delta \)

\[
\left\| f - \tilde{f}_{\alpha(\alpha)} \right\|_2 \leq \frac{\log^{1/2}(n/\delta)}{n^{1/(d+1)}}.
\]

As with the previous results, the stochastic error is controlled via the oracle inequality. Whereas the preceding results leveraged known approximation properties of radial SPD kernels, in the case of the box kernel we give a novel argument for approximating Lipschitz functions with box kernels having a non-vanishing bandwidth.

5 Experimental Results

We now show the performance of the fbKDE as compared to the KDE and variable bandwidth KDE. The variable bandwidth KDE ([27]), which we refer as the vKDE, has the form \( \tilde{f}_{\text{vKDE}} = \frac{1}{n} \sum_{i=1}^n k_\sigma(x, X_i) \) where each data point has an individual bandwidth \( \sigma_i \). [27] proposes to use \( \sigma_i = \sigma \lambda^{1/2} (f_{\text{KDE}}(X_i))^{-1/2} \) where \( \sigma \) is the standard KDE’s bandwidth parameter and \( \lambda \) is the geometric mean of \( \{ f_{\text{KDE}}(X_i) \}_{i=1}^n \). When implementing the fbKDE, there are a few considerations. First, when computing \( \int f \sum_i^2(x) d\mu(x) \), the first term of \( J_\alpha(\alpha) \), we must compute the integral \( \int f k(x, z_i) k(x, z_j) d\mu(x) \). For computational considerations we assume \( \mathcal{X} = \mathbb{R}^d \) and \( \mu \) is the Lebesgue measure. This deviates from our theory, which requires finite \( \mu \) for \( \mathcal{X} = \mathbb{R}^d \). Thus, for the Gaussian kernel, which we use in our experiments, this leads to \( \int f k(x, z_i) k(x, z_j) d\mu(x) = k_\sigma(z_i, z_j) \). To obtain the \( \alpha \) weights we have to solve a quadratic program. We used the ADMM algorithm described in [28], with the aid of the \( \ell_1 \) projection algorithm from [29].

We examine a few benchmark real datasets as well as synthetic data from 1-dimensional densities. The synthetic densities are the triangular density as well as three Gaussian mixtures, a bimodal, a trimodal and a kurtotic, as shown in Figure 1. We computed the parameters \( \sigma, R_n, \) and \( \sigma_\gamma \) in two different ways. First we used rules of thumb. For \( \sigma \), we used Silverman’s rule of thumb ([4]). For \( R_n \)
Table 1: Performance comparison for different datasets and bandwidth selection methods. For the synthetic datasets we drew 1000 samples, \( n = 800 \) of which were used for training.

| Dataset   | Rule of Thumb | Cross-Validation |
|-----------|---------------|------------------|
|           | fbKDE | KDE | vKDE | fbKDE | KDE | vKDE |
| Bimodal   | \(-1.0180\) | \(-0.8110\) | \(-0.8765\) | \(-1.0660\) | \(-0.9785\) | \(-1.0413\) |
| Triangular| \(0.2287\)   | 1.0141         | 0.8468         | 0.1865         | 0.5534         | 0.2782         |
| Trinodal  | \(-1.2889\) | \(-1.2897\) | \(-1.2958\) | \(-1.2332\) | \(-1.2095\) | \(-1.2073\) |
| Kurtotic  | \(-0.3317\) | \(-0.2919\) | \(-0.3025\) | \(-0.3457\) | \(-0.3379\) | \(-0.3456\) |
| Banana    | \(-0.5444\) | \(-0.4735\) | \(-0.5271\) | \(-0.5831\) | \(-0.5647\) | \(-0.5894\) |
| Ringnorm  | \(-0.0838\) | \(-0.0821\) | \(-0.0839\) | \(-0.0821\) | \(-0.0837\) | \(-0.0853\) |
| Thyroid   | \(-0.0317\) | \(-0.2919\) | \(-0.3025\) | \(-0.3457\) | \(-0.3379\) | \(-0.3456\) |
| Diabetes  | \(-0.5444\) | \(-0.4735\) | \(-0.5271\) | \(-0.5831\) | \(-0.5647\) | \(-0.5894\) |
| Waveform  | \(-0.0838\) | \(-0.0821\) | \(-0.0839\) | \(-0.0821\) | \(-0.0837\) | \(-0.0853\) |
| Iris      | \(-0.0317\) | \(-0.2919\) | \(-0.3025\) | \(-0.3457\) | \(-0.3379\) | \(-0.3456\) |

we used, based on the convergence rates, \((n/\log(n))^{1/3}\) for \( d \leq 4 \) and \((n/\log(n))^{(1/2 - 2/d)}\) for \( d > 4 \). For \( \sigma_n \) we used, inspired by the Jaakkola heuristic \([30]\), the median distance to the 5th nearest neighbor. We used a V-fold CV procedure over 100 parameter combinations drawn randomly from \( \Theta_d \), with \( V = 2 \) for \( n > 1000 \) and \( V = 3 \) otherwise (see \([31]\)). \( \Theta_d := [1, 5]_\ell \times [1, 1, 2\sqrt{\log(n)}]_\ell \times [0.01, 1]_\ell \) for \( d \leq 4 \), and \( \Theta_d := [1, 1]_\ell \times [1, 1, 2n^{(1/2 - 2/d)}]_\ell \times [0.01, 1]_\ell \) for \( d > 4 \), where the subscript \( \ell \) indicates logarithmic spacing and where the range is chosen thus since the data is standardized and, for the \( R_n \) range, informed by the convergence rates. Finally, we used \( n = 4/5 \) of the original data size for training and the rest for testing. We compute the value \( J_{n}^T \) as \( \int \frac{1}{n} \sum_{i=1}^{n} r(x_i - z_i) \), where \( \{ x_i \} \) is the test set. Figure 2 shows the bimodal density, the fbKDE, KDE, and vKDE along with the \( \alpha \) values for the fbKDE. Table 2 shows the \( J_n(\alpha^{(n)})^T \) error as well as the \( \| \cdot \|_\infty \) error, where for some function \( g \), \( \| g \|_\infty := \max_{x \in X} \| g(x) \| \).

In Figure 2 the density has two Gaussian elements with different widths. It is difficult for KDE and even for the vKDE to approximate such a density. The fbKDE, however, is able to approximate components of different smoothness by making some weights negative. These weights subtract excess mass in regions of refinement. Note that by doing so the fbKDE may itself overshoot and become negative. A similar behavior is exhibited for other densities, in which smoothness varies across regions.

In Table 4 we report the performance of the three estimators for both CV and rule of thumb. Note that the fbKDE often performs better in terms of the \( \| \cdot \|_\infty \) case and when the bandwidth is chosen according to the rule of thumb. The fbKDE outperforms both the KDE and vKDE in about half of the cases.

Finally we show, for the bimodal density, a comparison of the performance as the sample size grows. We have chosen the parameters according to the rules of thumb discussed above. Table 4 presents the errors. Note that as the sample size grows the KDE and vKDE do not significantly improve, even though the bandwidth is being updated according to Silverman’s rule. The fbKDE leverages new observations and refines its approximation, and this effect is more obvious for the \( \| \cdot \|_\infty \) case. Indeed, the \( \| \cdot \|_\infty \) error for fbKDE is smaller at \( n = 50 \) than at \( n = 2000 \) for KDE and vKDE at \( n = 2050 \). Similar results hold for the other synthetic datasets. This highlights a notable property of the fbKDE, that it can handle densities with differing degrees of smoothness.

6 Conclusion

We have presented a new member of the family of kernel estimators, the fixed bandwidth kernel density estimator, with desirable statistical properties. In particular, we showed the fbKDE is consistent for fixed kernel parameter \( \sigma \), and we provided convergence rates. The fbKDE is a good alternative to the KDE in cases where computing an optimal bandwidth is difficult and for densities that are hard to approximate with inappropriate bandwidths. In these cases and as is shown in the experimental section, the fbKDE
can greatly improve on the KDE. The way in which fbKDE achieves a more refined approximation is by balancing properly placed positive and negative weights, sometimes outside of the original support, which is facilitated by the $\Gamma$ variables, and which is not possible with the standard KDE. A few problems of interest remain open. We have illustrated two possible rate of convergence results, but expect many other such results are possible, depending on the choice of kernel and smoothness assumptions. It also remains an open problem to extend our results to more general domains $\mathcal{X}$ and to dependent data.

7 Proofs

7.1 Oracle Inequality

First recall the oracle inequality lemma:

**Lemma** [1] Let $\epsilon > 0$. Let $\{(X_i, \Gamma_i)\}_{i=1}^n$ satisfy assumption A1. Let $k$ satisfy assumption A2 and $f_\alpha$ be as in Equation (1). Let $\delta = 2n \exp \left(-\frac{(n-1)\epsilon^2}{8C_2R_2^2}\right)$. With probability $\geq 1 - \delta$,

$$\|f - f_{\alpha(\epsilon)}\|^2 \
\leq \epsilon + \inf_{\alpha \in A_n} \|f - f_\alpha\|^2.$$

**Proof of Lemma** [1] Recalling the following definition from Section 2, we have

$$h_i = h_i(X_i, \Gamma_i) := \int_S k(x, X_i + \Gamma_i) f(x) \, d\mu(x),$$

and

$$\hat{h}_i = \hat{h}_i(X, \Gamma_i) := \frac{1}{n-1} \sum_{j \neq i} k(X_j, X_i + \Gamma_i),$$

did we have used the simplified notation $(X, \Gamma_i)$ to represent $(X_1, \ldots, X_n, \Gamma_i)$. To simplify notation further, we let $X_j$ represent $(X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_n)$ and use $P_{X, \Gamma}(\cdot)$ for $P_{X_1, \ldots, X_n, \Gamma_1, \ldots, \Gamma_n}(\cdot)$, and the same goes for $E_{X, \Gamma}(\cdot)$. We now look at the probability that $H_n(\alpha)$ is close to $H(\alpha)$. We have

$$P_{X, \Gamma}\left\{ \sup_{\alpha \in A_n} |H_n(\alpha) - H(\alpha)| > \epsilon \right\} = P_{X, \Gamma}\left\{ \sup_{\alpha \in A_n} \left| \sum_{i=1}^n \alpha_i h_i(X_i, \Gamma_i) - \sum_{i=1}^n \alpha_i h_i(X_i, \Gamma_i) \right| > \epsilon \right\}$$

$$= P_{X, \Gamma}\left\{ \sup_{\alpha \in A_n} \left| \sum_{i=1}^n \alpha_i (\hat{h}_i(X, \Gamma_i) - h_i(X_i, \Gamma_i)) \right| > \epsilon \right\}$$

$$\leq P_{X, \Gamma}\left\{ \sup_{\alpha \in A_n} \left| \sum_{i=1}^n |\alpha_i| (\hat{h}_i(X, \Gamma_i) - h_i(X_i, \Gamma_i)) \right| > \epsilon \right\}$$

$$\leq P_{X, \Gamma}\left\{ \max_{1 \leq i \leq n} \left| \hat{h}_i(X, \Gamma_i) - h_i(X_i, \Gamma_i) \right| > \epsilon \right\}$$

$$\leq \sum_{i=1}^n P_{X, \Gamma}\left\{ \left| \hat{h}_i(X, \Gamma_i) - h_i(X_i, \Gamma_i) \right| > \frac{\epsilon}{R_n} \right\}.$$
Now let $A_i := \{(x_1, \ldots, x_n, \gamma_i) \in S^n \times \text{supp}(f_{\Gamma}) \mid \hat{h}_i(x, \gamma_i) - h_i(x, \gamma_i) > \frac{\epsilon}{n}\}$ and note that $P_{X, \Gamma}(A_i) = P_{X, \Gamma}(A_i)$. We have

$$P_{X, \Gamma}(\{(X_1, \ldots, X_n, \Gamma_i) \in A_i\}) = \int_{\text{supp}(f_\Gamma)} P_{X|\Gamma}(\{(X_1, \ldots, X_n, \gamma_i) \in A_i \mid \Gamma_i = \gamma_i\}) f_\Gamma(\gamma_i) d\gamma_i$$

$$= \int_{\text{supp}(f_\Gamma)} P_X(\{(X_1, \ldots, X_n, \gamma_i) \in A_i\}) f_\Gamma(\gamma_i) d\gamma_i,$$

by the independence assumption of $\tilde{A}_1$. Furthermore,

$$P_X(\{(X_1, \ldots, X_n, \gamma_i) \in A_i\}) = \int_S P_{X_i|X_i}(\{(X_1, \ldots, x_i, \ldots, X_n, \gamma_i) \in A_i \mid X_i = x_i\}) f(x_i) d\mu(x_i)$$

$$= \int_S P_{X_i}(\{(X_1, \ldots, x_i, \ldots, X_n, \gamma_i) \in A_i\}) f(x_i) d\mu(x_i).$$

We now bound the term inside the integral. Since this quantity only depends on $x_i, \gamma_i$ through $x_i + \gamma_i$, we abbreviate it as $P_{X_i}(\{(X_i, z_i) \in A_i\})$, where $z_i = x_i + \gamma_i$. First, note that $h_i(x, \gamma_i) = E_{X_i}(k(X_j, z_i))$ for any $j \neq i$. Also, by assumption there is a $C_k$ such that $k(x, x') \leq C_k$ for all $x, x'$. Hence,

$$P_{X_i}(\{(X_i, z_i) \in A_i\}) = P_{X_i}(\left\{\frac{1}{n-1} \sum_{j=1, j \neq i}^{n} k(X_j, z_i) - E_{X_i}(\frac{1}{n-1} \sum_{j=1, j \neq i}^{n} k(X_j, z_i)) > \epsilon\right\})$$

$$= P_{X_i}(\left\{\frac{1}{n-1} \sum_{j=1, j \neq i}^{n} k(X_j, z_i) - \frac{1}{n-1} E_{X_i}(\sum_{j=1, j \neq i}^{n} k(X_j, z_i)) > \epsilon\right\})$$

$$= P_{X_i}(\left\{\sum_{j=1, j \neq i}^{n} k(X_j, z_i) - E_{X_i}(\sum_{j=1, j \neq i}^{n} k(X_j, z_i)) > (n-1)\epsilon\right\})$$

$$\leq 2\exp\left(-\frac{2(n-1)\epsilon^2}{C_k^2}\right),$$

where we have used Hoeffding’s inequality. So we obtain

$$P_{X, \Gamma}(\left\{|\hat{h}_i - h_i| > \frac{\epsilon}{R_n}\right\}) = \int_{\text{supp}(f_\Gamma)} \int_S P_{X_i}(\{(X_i, z_i) \in A_i\}) f(x_i) f_\Gamma(\gamma_i) d\mu(x_i) d\gamma_i$$

$$\leq 2\exp\left(-\frac{2(n-1)\epsilon^2}{C_k^2 R_n^2}\right) \left(\int_{\text{supp}(f_\Gamma)} f_\Gamma(\gamma_i) d\gamma_i \int_S f(x_i) d\mu(x_i)\right)$$

$$= 2\exp\left(-\frac{2(n-1)\epsilon^2}{C_k^2 R_n^2}\right).$$

$$P_{X, \Gamma}(\sup_{\alpha \in A_n} |H_n(\alpha) - H(\alpha)| > \epsilon) \leq 2\exp\left(-\frac{2(n-1)\epsilon^2}{C_k^2 R_n^2}\right).$$

Therefore, letting $\delta = 2\exp\left(-\frac{2(n-1)\epsilon^2}{C_k^2 R_n^2}\right)$, for any $\alpha \in A_n$

$$P_{X, \Gamma}(\{|J_n(\alpha) - J(\alpha)| \leq 2\epsilon\}) = P_{X, \Gamma}(\{|H_n(\alpha) - H(\alpha)| \leq \epsilon\})$$

$$= 1 - P_{X, \Gamma}(\{|H_n(\alpha) - H(\alpha)| > \epsilon\})$$

$$\geq 1 - \delta.$$

So, with probability $1 - \delta$, $J_n(\alpha) \leq J(\alpha) + 2\epsilon$, and $J(\alpha) \leq J_n(\alpha) + 2\epsilon$. Recall that $J_n(\alpha^{(n)}) \leq J_n(\alpha)$ for all $\alpha \in A_n$. Then with probability $1 - \delta$, $J(\alpha^{(n)}) \leq \inf_{\alpha \in A_n} J(\alpha) + 4\epsilon$.

If we substitute $\epsilon' = 4\epsilon$ we obtain the desired result.
7.2 Consistency of \( f_{\alpha^{(n)}} \)

In the following we will make use of the fact that, for continuous positive definite radial kernels, the RKHS norm dominates the sup-norm which in turn dominates the \( L^2_\mu(\mathcal{X}) \) norm. We state this as a lemma.

**Lemma 5.** Let \( k, \mathcal{X}, \mu \) satisfy assumptions A0 and A2. Then for any \( h \in \mathcal{H} \) we have

\[
\|h\|_2 \leq \mu^{1/2}(\mathcal{X}) \|h\|_\infty \leq \mu^{1/2}(\mathcal{X}) C_{k}^{1/2} \|h\|_{\mathcal{H}}.
\]

**Proof.** By A0 and A2 \( k \) is bounded and continuous, and by Lemma 4.28 of [23], so is every element of \( \mathcal{H} \). Hence, for \( h \in \mathcal{H} \) and for \( \mathcal{X} \) either compact or all of \( \mathbb{R}^d \) the essential supremum equals the supremum, so we obtain

\[
\|h\|_2 = \left( \int_{\mathcal{X}} |h(x)|^2 \, d\mu(x) \right)^{1/2} \leq \mu^{1/2}(\mathcal{X}) \sup_{x \in \mathcal{X}} \{|h(x)|\} \leq \mu^{1/2}(\mathcal{X}) \|h\|_\infty \leq \mu^{1/2}(\mathcal{X}) C_{k}^{1/2} \|h\|_{\mathcal{H}}
\]

where the penultimate inequality follows from the reproducing property and the last inequality is just Cauchy-Schwarz. \( \square \)

Now to prove Theorem 1 we need a couple intermediate lemmas. First, recall

\[
f_\alpha := \sum_{i=1}^{n} \alpha_i k(\cdot, Z_i),
\]

and

\[
\mathcal{H}^0 := \left\{ \sum_{j}^{N} c_j k(\cdot, y_j) | y_j \in \mathcal{X}, c_i \in \mathbb{R}, \text{and } N \in \mathbb{N} \right\}.
\]

Note that \( \mathcal{H}^0 \) is dense in \( \mathcal{H} \) ([23]). Assume \( \mathcal{H} \) is dense in \( L^2_\mu(\mathcal{X}) \). Then, given \( \epsilon > 0 \) and \( \alpha \in A_n \), there is an \( f_H \in \mathcal{H} \) and \( f_\beta \in \mathcal{H}^0 \) of the form

\[
f_\beta := \sum_{j=1}^{m} \beta_j k(\cdot, y_j)
\]

such that \( \|f - f_H\|_2 < \epsilon \) and \( \|f_H - f_\beta\|_{\mathcal{H}} < \epsilon \). We use these functions to bound \( \|f - f_\alpha\|_2 \):

\[
\|f_\alpha - f\|_2 \leq \|f_\alpha - f_\beta\|_2 + \|f_\beta - f_H\|_2 + \|f_H - f\|_2,
\]

and show the last two terms are small in the following lemma.

**Lemma 6.** Let \( \epsilon > 0 \). Let \( k \) satisfy assumption A2 and \( f \in L^2_\mu(\mathcal{X}) \cap C(\mathcal{X}) \). Then

\[
\|f_\alpha - f\|_2 \leq \|f_\alpha - f_\beta\|_2 + 2\mu(\mathcal{X})^{1/2}\epsilon.
\]

**Proof.** If \( \mathcal{X} \) is compact, then \( \mathcal{H} \) is dense in \( C(\mathcal{X}) \) (see [22]). Therefore, for fixed \( \epsilon \), there is an \( f_H \in \mathcal{H} \) such that

\[
\|f_H - f\|_\infty \leq \epsilon,
\]

and by Lemma 6

\[
\|f_H - f\|_2 \leq \mu(\mathcal{X})^{1/2}\epsilon.
\]
If $X = \mathbb{R}^d$, \cite{22} tells us $\mathcal{H}$ is dense in $L^2_0(X)$, so it directly follows that, for any $\epsilon > 0$, there is an $f_\mathcal{H} \in \mathcal{H}$ satisfying

$$\|f_\mathcal{H} - f\|_2 \leq \mu(X)^{1/2}\epsilon.$$  

Similarly, since $\mathcal{H}^0$ is dense in $\mathcal{H}$ \cite{23}, for any fixed $\epsilon$ there is an $f_\beta \in \mathcal{H}^0$ such that

$$\|f_\beta - f_\mathcal{H}\|_\mathcal{H} \leq C^{-1/2}\epsilon,$$

hence, by lemma \cite{5}

$$\|f_\beta - f\|_2 \leq \mu(X)^{1/2}\epsilon.$$  

Therefore:

$$\|f_\alpha - f\|_2 \leq \|f_\alpha - f_\beta\|_2 + \|f_\beta - f_\mathcal{H}\|_2 + \|f_\mathcal{H} - f\|_2 \leq \|f_\alpha - f_\beta\|_2 + 2\mu(X)^{1/2}\epsilon.$$  

\[\Box\]

Note that $f_\beta \in \mathcal{H}^0$ implies $f_\beta = \sum_{j=1}^m \beta_j k(\cdot, y_j)$ for some $m \in \mathbb{N}$ and where $(\beta_j, y_j) \in \mathbb{R} \times X$ for all $1 \leq j \leq m$. To make the first term small, we first quantify the continuity of the kernel $k$. Let $\epsilon' = \epsilon/\|\beta\|_1$ and define

$$\eta_\epsilon := \mu^{1/2}(X) \frac{\epsilon'}{L(k)},$$

where $L(k)$ is the Lipschitz constant of $k$. Then for every $x$ and $y$ in $X$ we have that $\|x - y\|_2 \leq \eta_\epsilon$ implies $\|k(\cdot, x) - k(\cdot, y)\|_2 \leq \mu^{1/2}(X)\epsilon'$.

Recall $f_\alpha = \sum_{i=1}^n \alpha_i k(\cdot, Z_i)$. With the above result in hand we now have to make sure that at least a subset of the centers $\{Z_i\}_{i=1}^n$ of $f_\alpha$ are close to the centers $\{y_j\}_{j=1}^m$ of $f_\beta$ with high probability. First, define $B_j = \{x \in X \mid \|x - y\|_2 \leq \eta_\epsilon\}$ and define $P_z := P_{X \times X}$. Then we obtain the following lemma:

**Lemma 7.** Let $\epsilon > 0$ and $f_\beta$, $f_\alpha$ and $B_j$, $j = 1, \ldots, m$ as above. Then

$$P_z \{\forall B_j \exists Z_{i_j} \in \{Z_i\}_{i=1}^n : Z_{i_j} \in B_j\} \geq 1 - me^{-np_0}$$

for all $n$, where $p_0 = \min_j f_{B_j} f_z(z)dz$.

**Proof of Lemma.** Let the event $A_j^C = \{\forall z \in \{Z_i\}_{i=1}^n, \ z \notin B_j\}$. Then

$$P_z \{A_j^C\} = P_z \{Z_1 \notin B_j, \ldots, Z_n \notin B_j\} = \prod_{i=1}^n P_z \{Z_i \notin B_j\} = \prod_{i=1}^n (1 - P_z \{Z_i \in B_j\}) = \prod_{i=1}^n (1 - p_j) = (1 - p_j)^n,$$

where $p_j = \int_{B_j} f_z(z)dz$. Let $p_0 := \min_j \{p_j\}$ and recall that for $p \leq 1$ we have $1 - p \leq e^{-p}$. Hence $(1 - p_j)^n \leq e^{-np_j} \leq e^{-np_0}$, and

$$P_z \{\cap^m A_j\} = 1 - P_z \{\cup A_j^C\} \geq 1 - \sum_{j=1}^m P_z \{A_j^C\} = 1 - \sum_{j=1}^m (1 - p_j)^n \geq 1 - \sum_{j=1}^m e^{-np_j} = 1 - me^{-np_0}.$$  

11
For this term to approach zero we need \( p_0 := \min \int_{B_1} f_Z(z) \, dz \) to be strictly positive. This follows from the assumption that \( \text{supp} (f_Z) \supseteq \mathcal{X} \). Since \( m \) and \( p_0 \) only depend on \( \epsilon \) and other constants, we get
\[
P_Z \{ \tau^n A_j \} \geq 1 - mc^{-n p_0} \to 1
\]
as \( n \to \infty \).

**Lemma 8.** Let \( \mathcal{X}, \mu \) satisfy \( A \) \( \{(X_i, \Gamma_i)\}_{i=1}^n \) satisfy \( A \) and \( k \) satisfy \( A^2 \). Then, \( \forall \ \epsilon > 0 \ \exists C \) such that
\[
P_{X, \Gamma} \left\{ \inf_{\alpha \in A_n} \| f - f_\alpha \|_2 > \epsilon \right\} \leq Ce^{-n p_0}
\]
for sufficiently large \( n \).

**Proof.** Let \( \delta_2 = m \exp (-np_0) \). With probability \( \geq 1 - \delta_2 \) we have that for every \( j \) there is an \( i_j \) such that \( \| k(\cdot, z_{i_j}) - k(\cdot, y_j) \|_2 \leq \frac{\epsilon}{\| \beta \|_1} \). Then, for \( \alpha^* \) defined as
\[
\alpha^*_i = \begin{cases} \beta_j & : i = i_j \\ 0 & : i \neq i_j \end{cases}
\]
we have
\[
\| f_{\alpha^*} - f_{\beta} \|_2 = \left\| \sum_{i=1}^n \alpha^*_i k(\cdot, z_i) - \sum_{j=1}^m \beta_j k(\cdot, y_j) \right\|_2 \\
\leq \sum_{j=1}^m |\beta_j| \| k(\cdot, z_i) - k(\cdot, y_j) \|_2 \\
\leq \sum_{j=1}^m |\beta_j| \mu^{1/2}(\mathcal{X}) \cdot \frac{\epsilon}{\| \beta \|_1} \\
= \mu^{1/2}(\mathcal{X}) \epsilon
\]
Note that if for two sequences \( \{s_n\}, \{s'_n\} \) we have \( s_n \leq s'_n \) for \( n \geq N_0 \), then \( \lim_{n \to \infty} s_n \leq \lim_{n \to \infty} s'_n \), granted such limits exist. Let \( s_n := P_{X, \Gamma} \{ \inf_{\alpha \in A_n} \| f - f_\alpha \|_2 > 3 \mu^{1/2}(\mathcal{X}) \epsilon \} \) and \( s'_n := \delta_2 \). Note that for \( n \) large enough, say \( n = N_0 \), \( R_n \geq \| \beta \|_1 \) and therefore \( \alpha^* \in A_n \). So we can see that for \( n \geq N_0 \), the inequality
\[
\inf_{\alpha \in A_n} \| f - f_\alpha \|_2 \leq \| f - f_{\alpha^*} \|_2 \\
\leq \| f - f_{\mathcal{H}} \|_2 + \| f_{\mathcal{H}} - f_{\beta} \|_2 + \| f_{\beta} - f_{\alpha^*} \|_2 \\
\leq 3 \mu^{1/2}(\mathcal{X}) \epsilon
\]
holds with probability \( \geq 1 - \delta_2 \). That is, for \( n \geq N_0 \), \( s_n \leq s'_n \), hence since \( s'_n = \delta_2 \to 0 \), we get \( s_n \to 0 \).

**Proof of Theorem**
\[
P_{X, \Gamma} \left\{ J^{(n)} \leq 4 \epsilon + \left( 3 \mu^{1/2}(\mathcal{X}) \epsilon \right)^2 \right\} \leq P_{X, \Gamma} \left\{ J^{(n)} \leq J^* + 4 \epsilon \cap \left\{ J^* \leq \left( 3 \mu^{1/2}(\mathcal{X}) \epsilon \right)^2 \right\} \right\} \leq 1 - P_{X, \Gamma} \left\{ J^{(n)} > J^* + 4 \epsilon \right\} + P_{X, \Gamma} \left\{ J^* > \left( 3 \mu^{1/2}(\mathcal{X}) \epsilon \right)^2 \right\}.
\]
(8)
By Lemma[3] the middle term approaches zero and by Lemma[8] the last term does, so
\[
\lim_{n \to \infty} P_{X, \Gamma} \left\{ J^{(n)} \leq \epsilon' \right\} = 1,
\]
where \( \epsilon' = 4 \epsilon + (3 \mu^{1/2}(\mathcal{X}) \epsilon)^2 \).
7.3 Convergence Rates of $f_\alpha(n)$

The proof of Lemma 3 is found in [24]. Now let $\beta_j = \frac{c_j \lambda_j}{m}$ and $f_\beta = \sum_{j=1}^{m} \beta_j k(\cdot, y_j)$, where $\lambda$, $m$, and $c_j$, $y_j$, for $j = 1, \ldots, m$, are as in Lemma P 3. Recall Lemma P 4.

**Lemma 4.** Let $\delta_2 \in (0, 1)$, let $f \in \mathcal{F}_k$ and let $f_\beta$ and $m$ be as above. Let $\{(X_i, \Gamma_i)\}_{i=1}^{n}$ satisfy the assumption $A_1$, then with probability $\geq 1 - \delta_2$

\[
\inf_{\alpha \in \mathbb{A}_n} \|f_\beta - f_\alpha\|_2 \leq c_3(n, m)
\]

where $c_3(n, m) := \frac{C_3}{n^{d_\alpha}} \log^{1/d}(m/\delta_2)$, for $C$ a constant independent of $n$ and $m$. \hfill \square

**Proof of Lemma 4.** Let $\eta = \eta(c_3) := \frac{c_3}{\|f\|_1}$, where $L$ is the Lipschitz constant of the $k$. From Lemmas 7 and 8 we know that with probability $\geq 1 - me^{-np}$, the event that for all $y_j$ there is a data point $Z_{i_j}$ such that $\|y_j - Z_{i_j}\|_2 \leq \eta$ holds. Recall $p_0 = \min_j \int_{B_j} f_Z(z) dz$. Now

\[
p_0 = \min_j \int_{B_j} f_Z(z) dz
\]

where $c'$ is the volume of the $d$-dimensional unit ball.

Now pick the $\alpha$ coefficients as in Lemma 2. With probability $\geq 1 - me^{-nc' \eta^d}$ we have:

\[
\inf_{\alpha \in \mathbb{A}_n} \|f_\beta - f_\alpha\|_2 \leq \left\| \sum_{j=1}^{m} \beta_j \left( k(\cdot, y_i) - k(\cdot, z_{i_j}) \right) \right\|_2
\]

\[
\leq L \mu^{1/2}(X) \|\beta\|_1 \eta
\]

\[
= \mu^{1/2}(X)c_3.
\]

Fixing $\delta_2 \in (0, 1)$ and setting $me^{-nc' \eta^d} = \delta_2$, where $c = c' \min_{x \in X} \{ f^*(x) \}$, we obtain

\[
\epsilon_3 = \frac{L \|\beta\|_1}{(nc)^{1/d}} \log^{1/d}(m/\delta_2).
\]

Finally, noting that $\|\beta\|_1 = \|\lambda\|_1$, where $\lambda$ is as in Lemma 3 yields the desired result. \hfill \square

We now prove Theorem 3.

**Proof of Theorem 3.** We will use the notation and results of Lemmas 2, 3 and 4. First, note that from Lemma 3 we have $\| f - f_\beta \|_2 \leq \mu^{1/2}(X) \| f - f_\beta \|_\infty \leq \mu^{1/2}(X) \epsilon_2(m)$. So putting the three Lemmas together we have that with probability $\geq 1 - (\delta_1 + \delta_2)$

\[
\| f - f_\alpha(n) \|_2 \leq \epsilon_1(n) + \mu(X)(\epsilon_2(m) + \epsilon_3(n, m))^2,
\]

or, taking into account only the dependence on $n$, $m$, $\delta_1$ and $\delta_2$ for the different $\epsilon_i$’s, we have that $\| f - f_\alpha(n) \|_2$ is of the order of

\[
\| f - f_\alpha(n) \|_2 \leq \log^{1/2}(n/\delta_1) \frac{R_n}{n^{1/2}} + \frac{1}{m} + \log^{-2/d}(m/\delta_2) \frac{1}{n^{2/d}}.
\]

Now, we want the number $m$ of centers in $f_\beta$ close to but no larger than the number $n$ of data points, so we set $m = n^\theta$ for some $\theta$ such that $0 < \theta < 1$. Furthermore, we need $R_n$ to grow accordingly, so that $R_n = n^c \theta$, for $c > 0$ a constant possibly dependent on $d$. This yields, ignoring the log terms for now:

\[
\| f - f_\alpha(n) \|_2 \leq \frac{1}{n^{1/2-c\theta}} + \frac{1}{n^{2\theta}} + \frac{1}{n^{2/d}}.
\]

Setting the first two rates equal we obtain $\theta = 1/2(1 + c)$. Note that if $d > 4$ we can match the third term by setting $c = d/4 - 1$ to obtain an overall rate of $2/d$. Otherwise we can set $c$ to any small number to obtain a rate $1/2(1 + c)$ slightly slower than $1/2$. Finally, for the log terms, just let $\delta_1 = \delta_2 = \delta/2$ and note that if $d > 4$ then $\log^{1/2}$ dominates, and if $d \leq 4$ then $\log^{2/d}$ dominates. \hfill \square
7.4 Box rates

First recall the definition of \(\tilde{f}_\alpha\) from Section 4.3. For any \(M\) and \(\epsilon(M)\) let \(f_\beta = \sum_{i=1}^{M} \beta_i k(\cdot, y_i)\) be such that \(\|f - f_\beta\|_2 < \epsilon(M)\), then

\[
\tilde{f}_\alpha := \sum_{i=1}^{M} \alpha_i k(\cdot, y_i).
\]

And, also from section 4.3, \(X = [0, 1]^d, \mathcal{F} = L^1_{\mu}(X) \cap L^2_{\mu}(X) \cap \text{Lip}(X)\), where \(\text{Lip}(X)\) are the Lipschitz functions on \(X\), \(\mu\) is the Lebesgue measure and \(L_f\) is the Lipschitz constant of \(f\). Also, \(k\) is the box kernel \(k(x, y) = \frac{1}{\sqrt{(\pi q)}} \mathbf{1}_{\|x - y\|_{\infty} \leq \sigma}\) defined on \(X \pm \sigma \times X \pm \sigma\), and for simplicity we assume \(\sigma = \frac{1}{2q}\) for \(q\) a positive integer.

**Theorem 3** Let \(f \in \mathcal{F}, \text{supp} \{f_x\} \supset X, \{(X_i, \Gamma_i)\}_{i=1}^{n} \text{ satisfy } A1\) and \(R_n \sim n^{(d-1)/(2d+2)}\). Let \(\delta \in (0, 1)\), then, with probability greater than or equal to \(1 - \delta\)

\[
\|f - f_{\alpha(n)}\|_2 \lesssim \frac{\log(1/\delta)}{n^{1/(d+1)}}.
\]

To prove this theorem we begin by reformulating Lemma 1.

**Lemma 9.** Let \(\delta \in (0, 1)\). Let \(\{(X_i)\}_{i=1}^{n} \text{ satisfy } A1\) and \(k\) satisfy \(\text{sup}_{x, x' \in X} k(x, x') \leq C_k\). If \(\tilde{f}_{\alpha(n)}\) is as above, then with probability \(\geq 1 - \delta\)

\[
J(\alpha(n)) \leq \epsilon_1(n) + \inf_{\alpha \in \mathcal{A}_n} \|f - \tilde{f}_\alpha\|_2
\]

where \(\epsilon_1(n) := \sqrt{8C_kR_M} \frac{\log(AM/\delta)}{n} \). \(\square\)

We now take care of the term \(\inf_{\alpha \in \mathcal{A}_n} \|f - \tilde{f}_\alpha\|_2\).

**Lemma 10.** Let \(f \in \mathcal{F}\). For any \(m \in \mathbb{N}\) there is a function \(f_\beta\) of the form \(f_\beta = \sum_{i=1}^{(mq)^d} \beta_i k(\cdot, y_i)\), where \(\{y_i\}_{i=1}^{(mq)^d} \subset X \pm \sigma\) and \(\|\beta\|_1 \leq (mq)^{d-1} R_{f, \sigma}\) satisfying

\[
\|f - f_\beta\|_2 \leq \epsilon_2(m)
\]

where \(\epsilon_2(m) := \frac{L \sqrt{2q}}{mq}\). \(\square\)

**Proof of Lemma 10.** Let \(i := (i_1, \ldots, i_d)\) be a multi-index with positive elements and associated index \(i\) related by the function \(h:\)

\[
i = h(i) = 1 + \sum_{\ell=1}^{d} (i_\ell - 1)(mq)^{\ell-1}
\]

and its inverse

\[
h^{-1}(i) = \left( \left\lfloor \frac{i}{(mq)^{d-1}} \right\rfloor, \left\lfloor \frac{i}{(mq)^{d-2}} \right\rfloor, \ldots, \left\lfloor \frac{i}{(mq)} \right\rfloor, \left\lfloor \frac{i}{mq} \right\rfloor, \left\lfloor \frac{i}{mq^2} \right\rfloor, \ldots, \left\lfloor \frac{i}{mq^d} \right\rfloor \right).
\]

Divide \(X = [0, 1]^d\) into \((mq)^d\) hypercube regions of equal volume to form the partition \(\Pi_{i=1}^{d} \left[ \left[ \frac{i_\ell - 1}{mq^\ell}, \frac{i_\ell}{mq^\ell} \right[ \right]_{\ell=1}^{d} = \{T_i\}_{i=1}^{(mq)^d}\), where \(T_i = \Pi_{\ell=1}^{d} \left[ \left[ \frac{i_\ell - 1}{mq^\ell}, \frac{i_\ell}{mq^\ell} \right[ \right] \). Now, let

\[
f_m = \sum_{i=1}^{(mq)^d} f(T_i) \mathbf{1}_{\{T_i\}}
\]
where \( \pi_i \in T_i \). Any choice of \( \pi_i \) works but for clarity we choose \( \pi_i = (\ell_1/mq, \ldots, \ell_d/mq) \). Note that \( f_m \) is close to \( f \):

\[
\| f - f_m \|_2^2 \leq \int_X (f(x) - f_m(x))^2 \, dx
\]

\[
= \sum_{i=1}^{(mq)^d} \int_{T_i} (f(x) - f_m(x))^2 \, dx
\]

\[
= \sum_{i=1}^{(mq)^d} \int_{T_i} (f(x) - f(\pi_i))^2 \, dx
\]

\[
\leq \sum_{i=1}^{(mq)^d} \int_{T_i} (L_f \| x - \pi_i \|_2)^2 \, dx
\]

\[
\leq \sum_{i=1}^{(mq)^d} \int_{T_i} \left( \frac{\sqrt{d}}{mq} \right)^2 \, dx
\]

\[
\leq \left( \frac{L_f \sqrt{d}}{mq} \right)^2.
\]

Hence

\[
\| f - f_m \|_2 \leq \frac{L_f \sqrt{d}}{mq}.
\]

Now we note that \( f_m \) can also be expressed as a sum of fixed bandwidth kernels:

\[
f_m = \sum_{i=1}^{(mq)^d} \beta_i(2\sigma)^d k(\cdot, y_i),
\]

where

\[
y_i = \left[ \frac{\ell_1 - 1}{mq} + \sigma, \ldots, \frac{\ell_d - 1}{mq} + \sigma \right]^T
\]

and \( \beta \) is as follows. Let \( \beta_1 = f(\pi_1) \) and

\[
\beta_i = f(\pi_i) - \sum_{s \in S_i} \beta_{h(s)} - f(0) 1_{\{\ell_1 = 1 \forall \ell\}}
\]

for \( 2 \leq i \leq (mq)^d \), where \( s_i = (\max\{1, \ell_1 - (m - 1)\}, \ldots, \max\{1, \ell_d - (m - 1)\}) \) and \( S_i = (\max\{1, \ell_1 - 1\}, \ldots, \max\{1, \ell_d - 1\}) \) are multi-indices. Note that the \( \beta_i \)'s sequentially capture the residual of the function \( f_m \) as we travel along the \( T_i \) regions.

To find \( \| \beta \|_1 \) note that since \( \| \pi_1 - 0 \|_2 \leq L_f \sqrt{d}/(mq) \), we have \( |\beta_i| \leq f(0) + L_f \sqrt{d}/(mq) \). Also note \( \beta_i = f(\pi_i) - f(\pi_{i-1}) \) for \( 2 \leq i \leq m \), hence for \( 2 \leq i \leq m \)

\[
|\beta_i| = |f(\pi_i) - f(\pi_{i-1})|
\]

\[
\leq L_f \| \pi_i - \pi_{i-1} \|_2
\]

\[
\leq L_f \frac{\sqrt{d}}{mq}
\]

For larger \( \ell_1 \) we have \( \beta_i = f(\pi_i) - f(\pi_{i-1}) + \beta_{i-m} - f(0) 1_{\{\ell_1 = 1 \forall \ell\}} \). Note that when \( \ell_1 = m + 1 \) we have lost influence of \( \beta_1 \), so

\[
|\beta_{m+1}| \leq |f(\pi_i) - f(\pi_{i-1})| + |\beta_1| + f(0) \leq 2L_f \frac{\sqrt{d}}{mq} + 2f(0)
\]

and, similarly, \( |\beta_i| \leq 2L_f \frac{\sqrt{d}}{mq} \) for \( m + 2 \leq i \leq 2m \). The process continues such that, in general

\[
|\beta_i| \leq \left( \frac{\ell_1}{m} \right) \left( L_f \frac{\sqrt{d}}{mq} + f(0) 1_{\{\ell_1 = 1 \forall \ell\}} \right)
\]
for $i \leq mq$. Adding these we obtain $4q^2(L_f \sqrt{d} + qf_0)$. Denote this quantity by $e'$. Then this process is repeated over every dimension, having $q$ chunks of multiples of $e'$, the first multiple is $1m$, the second $2m$, and so on. The final sum is then

$$\|\beta\|_1 \leq (mq)^{d-1} \left( \frac{q + 1}{2} \right)^2 (qf(0) + L_f \sqrt{d}) .$$

Therefore

$$(2\sigma)^d \|\beta\|_1 = \frac{(mq)^{d-1}}{q} \left( \frac{q + 1}{2} \right)^2 (qf(0) + L_f \sqrt{d}) .$$

Proof of Theorem 3. This proof is similar to the proof of Theorem 2. Combining Lemmas 9, 10, setting $R_n' \sim m^{(d-1)}$, ignoring the log terms for now and considering only the dependence on $n$ and $m$ we obtain

$$\|f - f_{\alpha(n)}\|^2 \lesssim \frac{m^{d-1}}{n^{1/2}} + \frac{1}{m^2} .$$

Setting these terms equal we obtain $m = n^{1/(2d+2)}$, with an overall rate of $n^{-1/(d+1)}$. Adding the log term we obtain

$$\|f - f_{\alpha(n)}\|^2 \lesssim \frac{\log^{1/2}(n^{d/(2d+2)}/\delta)}{n^{1/(d+1)}} \lesssim \frac{\log^{1/2}(n/\delta)}{n^{1/(d+1)}} .$$

References

[1] E. Fix and J. L. Hodges Jr, “Discriminatory analysis-nonparametric discrimination: consistency properties,” USAF School of Aviation Medicine, Report No. 4, Tech. Rep. 21-49-004, 1951.

[2] M. Rosenblatt, “Remarks on some nonparametric estimates of a density function,” The Annals of Mathematical Statistics, vol. 27, no. 3, pp. 832–837, 1956.

[3] E. Parzen, “On estimation of a probability density function and mode,” The Annals of Mathematical Statistics, vol. 33, no. 3, pp. 1065–1076, 1962.

[4] B. W. Silverman, Density Estimation for Statistics and Data Analysis. London: Chapman and Hall, 1986.

[5] D. W. Scott, Multivariate Density Estimation. New York: Wiley, 1992.

[6] M. P. Wand and M. C. Jones, Kernel Smoothing. Chapman & Hall/CRC, 1994.

[7] L. Devroye and G. Lugosi, Combinatorial Methods in Density Estimation. New York: Springer, 2001.

[8] D. Wied and R. Weiβbach, “Consistency of the kernel density estimator: a survey,” Statistical Papers, vol. 53, no. 1, pp. 1–21, 2012.

[9] N.-B. Heidenreich, A. Schindler, and S. Sperlich, “Bandwidth selection for kernel density estimation: a review of fully automatic selectors,” AStA Advances in Statistical Analysis, vol. 97, no. 4, pp. 403–433, 2013.

[10] D. Kim, “Least squares mixture decomposition estimation,” PhD Thesis, Dept. of Statistics, Virginia Polytechnic Inst. and State Univ., 1995.

[11] M. Girolami and C. He, “Probability density estimation from optimally condensed data samples,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 10, pp. 1253–1264, OCT 2003.

[12] R. Ganti and A. G. Gray, “Cake: Convex adaptive kernel density estimation,” in International Conference on Artificial Intelligence and Statistics, 2011, pp. 498–506.
[13] J. Kim and C. D. Scott, “Robust kernel density estimation,” The Journal of Machine Learning Research, vol. 13, no. 1, pp. 2529–2565, 2012.

[14] ——., “L2 kernel classification,” Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 32, no. 10, pp. 1822–1831, 2010.

[15] P. Rigollet and A. B. Tsybakov, “Linear and convex aggregation of density estimators,” Mathematical Methods of Statistics, vol. 16, no. 3, pp. 260–280, 2007.

[16] E. Cruz Cortés and C. Scott, “Sparse approximation of a kernel mean,” IEEE Transactions on Signal Processing, vol. 65, pp. 1310–1323.

[17] M. Girolami and C. He, “Probability density estimation from optimally condensed data samples,” Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 25, no. 10, pp. 1253–1264, 2003.

[18] J. M. Phillips, “ε-samples for kernels,” in Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms. SIAM, 2013, pp. 1622–1632.

[19] P. Drineas and M. W. Mahoney, “On the Nyström method for approximating a gram matrix for improved kernel-based learning,” The Journal of Machine Learning Research, vol. 6, pp. 2153–2175, 2005.

[20] Y. Chen, M. Welling, and A. Smola, “Super-samples from kernel herding,” in Conference on Uncertainty in Artificial Intelligence (UAI), 2010.

[21] R. Vert and J.-P. Vert, “Consistency and convergence rates of one-class SVM and related algorithms,” J. Machine Learning Research, pp. 817–854, 2006.

[22] C. Scovel, D. Hush, I. Steinwart, and J. Theiler, “Radial kernels and their reproducing kernel hilbert spaces,” Journal of Complexity, vol. 26, no. 6, pp. 641–660, 2010.

[23] I. Steinwart and A. Christmann, Support Vector Machines. Springer, 2008.

[24] G. Gnecco and M. Sanguineti, “Approximation error bounds via Rademacher’s complexity,” Applied Mathematical Sciences, vol. 2, no. 4, pp. 153–176, 2008.

[25] F. Girosi, “Approximation error bounds that use vc-bounds,” in Proc. International Conference on Artificial Neural Networks, F. Fogelman-Soulie and P. Gallinari, editors, vol. 1, 1995, pp. 295–302.

[26] M. A. Kon and L. A. Raphael, “Approximating functions in reproducing kernel Hilbert spaces via statistical learning theory,” Wavelets and Splines: Athens 2005, pp. 271–286, 2006.

[27] D. Comaniciu, V. Ramesh, and P. Meer, “The variable bandwidth mean shift and data-driven scale selection,” in Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on, vol. 1. IEEE, 2001, pp. 438–445.

[28] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” Foundations and Trends® in Machine Learning, vol. 3, no. 1, pp. 1–122, 2011.

[29] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, “Efficient projections onto the ℓ1-ball for learning in high dimensions,” in Proceedings of the 25th international conference on Machine learning. ACM, 2008, pp. 272–279.

[30] T. Jaakkola, M. Diekhans, and D. Haussler, “Using the Fisher kernel method to detect remote protein homologies.” in ISMB, vol. 99, 1999, pp. 149–158.

[31] S. Arlot and M. Lerasle, “Choice of v for v-fold cross-validation in least-squares density estimation,” Journal of Machine Learning Research, vol. 17, no. 208, pp. 1–50, 2016.