CAKE: Convex Adaptive Kernel Density Estimation

Ravi Ganti **Alexander Gray**
College of Computing (CSE), GeorgiaTech

Abstract

In this paper we present a generalization of kernel density estimation called Convex Adaptive Kernel Density Estimation (CAKE) that replaces single bandwidth selection by a convex aggregation of kernels at all scales, where the convex aggregation is allowed to vary from one training point to another, treating the fundamental problem of heterogeneous smoothness in a novel way. Learning the CAKE estimator given a training set reduces to solving a single convex quadratic programming problem. We derive rates of convergence of CAKE like estimator to the true underlying density under smoothness assumptions on the class and show that given a sufficiently large sample the mean squared error of such estimators is optimal in a minimax sense. We also give a risk bound of the CAKE estimator in terms of its empirical risk. We empirically compare CAKE to other density estimators proposed in the statistics literature for handling heterogeneous smoothness on different synthetic and natural distributions.

1 Introduction

The problem of density estimation is as follows: Given n i.i.d. points $x_1, \ldots, x_n$ sampled from the distribution with a density function f, the task is to construct a density estimator $\hat{f} : \mathbb{R}^d \times (\mathbb{R}^d)^n \to \mathbb{R}$ which provably converges to the true underlying density function f in a suitable sense. Accurate density estimation allows one to build accurate classifiers, regressors and also facilitates data visualization. Parametric approaches to density estimation e.g. fitting a mixture of Gaussians to the data with the expectation-maximization algorithm (Bishop et al., 2006) require strong parametric assumptions on the true underlying density function which are seldom known. A variety of non-parametric approaches for density estimation such as kernel density estimation (KDE) and wavelet based methods exist. Non parametric methods require only the weaker assumption that the underlying function that we are trying to recover belongs to a smooth class of functions, and hence are suitable in many domains where no knowledge of appropriate parametric form is available.

Kernel density estimation (KDE) (Parzen, 1962) is the most popular non-parametric method for density estimation in part because other approaches such as wavelets do not extend well beyond one or two dimensions. KDE involves fitting smoothing kernels, which is a symmetric probability density function (PDF), at the different training points. The density at a point x is then simply the sum of kernel contributions due to all training points x_i at x. This yields an estimator of the form

$$\hat{f}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} k\left(\frac{x - x_i}{h}\right). \quad (1)$$

k is a smoothing kernel that is chosen priori. Examples of smoothing kernels include Gaussian and Epanechnikov kernels. The task in KDE is to estimate the bandwidth h. Some common approaches to estimating h include maximizing the leave-one-out likelihood cross validation and minimizing least squares cross validation error (LSCV) (An excellent survey of bandwidth selection methods can be found in (Jones et al., 1996), (Hall et al., 1995)). Though kernel density estimators are consistent (Tsybakov, 2009) they are not good at modeling distributions which have spatially varying smoothness. This affects the problem visually and also leads to slower rates of asymptotic convergence. In this paper we will present a generalization of KDE called *Convex Adaptive Kernel Density Estimation* (CAKE). The basic idea of CAKE is to use a set of base kernels with different bandwidths to fit kernels at different training points by a convex aggregation (CA) of these base kernels. However, the trick is to allow this CA's to vary across the different training points. This turns out to be equivalent to fitting

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15.

a kernel at the different training point with the bandwidth being a function of the coefficients of the CA, the training point and the test point where we need to estimate the density. By doing so we are able to learn a density estimator that adapts well to varying levels of smoothness of the true density.

Previous Work. One of the proposed techniques to solve the problem of learning density functions with spatially varying levels of smoothness is to learn the optimal smoothing kernel by solving a variational problem that minimizes the variance or mean integrated squared error (MISE) of the density estimates via Legendre's polynomials (Gasser et al., 1985). Such kernels are data independent and are difficult to generalize to higher dimensions. In the "variable kernel" density estimation method (VKDE) the bandwidth varies with the p^{th} nearest neighbour of the training points to the remaining $n-1$ training points. VKDE with smoothing parameter h is defined as

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(hd_{p,i})^d} k\left(\frac{x-x_i}{hd_{p,i}}\right). \quad (2)$$

where $d_{p,i}$ is the distance of the p^{th} nearest neighbour of x_i in the dataset, and h is a universal smoothing factor. Estimators obtained by VKDE are probability density functions (PDF), and inherit the smoothness properties of the kernel k, but still are not very good at capturing complex distributions and require optimization over a continuous variable h and a discrete variable p. Nearest neighbour KDE (NNKDE) fits kernels of bandwidth h at all training points where h varies with the p^{th} nearest neighbour of the test point x where we need to estimate the density. NNKDE is known to exhibit rough tails (Silverman, 1986) that are discontinuous and do not necessarily integrate to 1. A generalization of the "variable kernel" method which is known to work well for 1-d problems is the adaptive kernel density estimation (AKDE) method (Breiman et al., 1977), where the width of the kernel varies according to the training points. AKDE works by first choosing a bandwidth h, to get a pilot density estimate at different training points, and then scales the bandwidth of each training point by θ_i, giving larger bandwidths at training points where the density is small and smaller bandwidths where the density is large. The estimator is

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(h\theta_i)^d} k\left(\frac{x-x_i}{h\theta_i}\right). \quad (3)$$

where θ_i is the local bandwidth scaling factor. Girolami *et al* (Girolami & He, 2003) proposed the reduced set density estimator (RSDE) where the kernel contributions due to the different training points were scaled by different weights and the density is estimated as

$$\hat{f}(x) = \sum_{i=1}^{n} \frac{\gamma_i}{h^d} k\left(\frac{x-x_i}{h}\right). \quad (4)$$

The weights γ_i were then learnt by solving a convex QP which minimizes the ISE under convexity constraints. Due to the structure of the optimization problem the vector γ turns out to be sparse, and hence only the reduced set (non-zero γ_i values) matters. However it is clear that for points that are far away from the reduced set the density estimate of RSDE is an underestimate.

Support vector density estimation (Vapnik & Mukherjee, 1999) fits a cdf to the sampled data by solving an optimization problem which minimizes the ℓ_1 distance between the estimator of cumulative distribution function and its empirical counterpart. Work on similar lines has been done in (Song et al., 2008; Shawe-Taylor & Dolia, 2007). However, these methods to our knowledge lack theoretical results on the bias, variance and consistency of such estimators. Devroye *et al* (Devroye & Lugosi, 2000) investigated learning kernel density estimators in an L_1 framework. They proposed the double kernel method where a pair of kernels k, l are used to learn the bandwidth h which is provably universal. Though promising, to our knowledge no empirical work has been done in this framework. Liu *et al* proposed the RODEO density estimator (Liu et al., 2007) [1] to fit high dimensional distributions. They learn a semi-parametric density estimator where in order to estimate the density at a point x, product kernels are fitted at the different training points and these product kernels are a product of d univariate kernels along different dimensions with different bandwidths. The bandwidths along the different dimensions are learnt by using a test statistic which compares the magnitude of the derivative of the density estimates along different dimensions to the variance of the density estimate. The resulting estimator provably achieves better rates of convergence than KDE, under certain sparsity assumptions.

Our Contributions. The problem of learning an optimal Mercer kernel has been of recent interest in the kernel machines community (Ong et al., 2005; Lanckriet et al., 2004). In this paper we *bridge the two distinct methods of density estimation and kernel machines* and show that by appropriately learning smoothing kernels for the problem of density estimation one can achieve the desired goal of learning density estimators that exhibit varying levels of smoothness in different regions of the space so that even complex distributions can be modeled well. With this goal in mind we propose the *Convex Adaptive Kernel Density Estimation method* (CAKE)

[1] In this paper we shall concern ourselves with local rodeo with a uniform density as the baseline density and KDE as the non-parametric component of the density estimator.

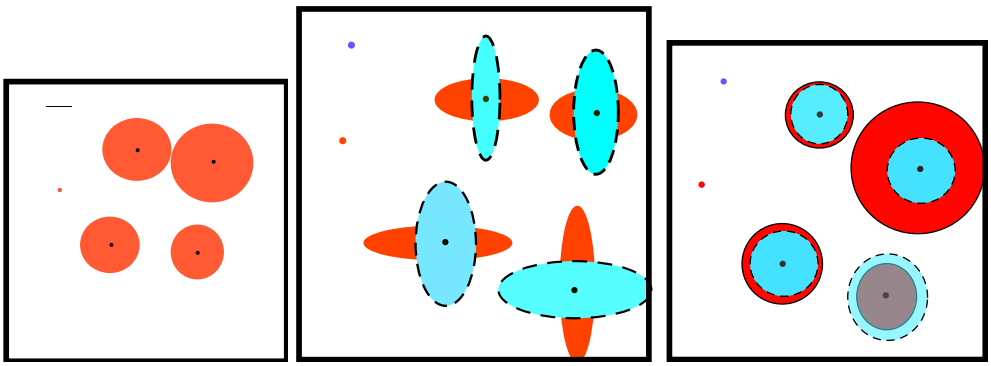

Figure 1: This figure outlines the fundamental difference between the shape of the kernels fitted by different kernel based density estimation methods by AKDE/VKDE (leftmost), RODEO (middle plot), and CAKE (rightmost plot).

method (Section 3). In CAKE a set of base kernels $\mathcal{K}(|\mathcal{K}| = m = O(1))$ is used to learn smoothing kernels at different training points by aggregating them in a convex way. However, the trick is to let these convex aggregations change from one training point to the other. The CAKE density estimator can be written as:

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\alpha_{ij}}{h_j^d}k\left(\frac{x-x_i}{h_j}\right) \tag{5}$$

where $\alpha \in R^{nm} \geq 0$, $\forall i = 1,\ldots,m : \sum_{j=1}^{m}\alpha_{ij} = 1$, where $h_1,\ldots,h_m$ are bandwidths of the m base kernels and are assumed to be known. The constants α_{ij} are learnt by minimizing the regularized LSCV score of $\hat{f}$. The LSCV (Section 2) of $\hat{f}$ is a surrogate of the integrated squared error (ISE) of $\hat{f}$ and can be calculated using the training set. Minimizing the regularized LSCV of the density estimator $\hat{f}$ reduces to solving a quadratic programming problem (QP) over nm variables (Section 3) which can be efficiently solved using a simple variation of the SMO algorithm. The unique power of CAKE as a density estimator (see Figure (1)) stems from the fact that CAKE estimates densities by placing kernels of different bandwidths at different training points (like AKDE and VKDE). In addition, these bandwidths depend on the point x where we need to estimate density (like NNKDE). However, unlike AKDE and RODEO the learning process involves minimization of a risk function, which is the L_2 distance between the estimator $\hat{f}$ and the true density function f. We also show connections to the literature on optimal aggregation of estimators (Nemirovski, 2000; Tsybakov, 2003) and demonstrate how the CAKE is more than a simple convex aggregation of kernel density estimators.

We analyze the MSE of the CAKE like density estimators (Section (4)) and examine its optimal value and show that given enough data (as a function of the true underlying density and the dimensionality) the MSE of CAKE like estimators for the densities in Hölder class $\Sigma(\beta, L)$ is $O(n^{-\frac{2\beta}{2\beta+d}})$. We also provide a bound on the L_1 risk of CAKE in terms of empirical L_1 error using stability arguments. To our knowledge this is the first time stability arguments have been used for density estimation problems (Section (4)).

We empirically compare our density estimator to RSDE, RODEO, AKDE and VKDE on various synthetic and natural datasets (Section (5)). In order to evaluate our density estimator on high dimensional data we use the CAKE density estimator to learn a smoothing kernel based classifier and compare it to smoothing kernel based classifiers learnt using other density estimation methods.

Notation. Vectors are represented in lower case letters and matrices in upper case. We shall use double indexing for vectors of sizes nm. e.g if $v \in R^{nm}$ then v_{pq} refers to the $(n(p-1)+q)^{\text{th}}$ element of the vector v. $\mathbf{1}_n$ refers to a vector of all 1's of size n, and $\mathbf{0}_m$ refers to the vector of all 0's of size m.

2 L_2 Error of a Density Estimator and its Surrogate

We would like to minimize the L_2 error of an estimator also known as the integrated squared error (ISE). One of the prime motivations for choosing ISE as our objective over other objectives such as likelihood is its robustness to outliers (Silverman, 1986). Given any estimator $\hat{f}$ of the underlying density function the ISE of the estimator is

$$ISE(\hat{f}) = \int(\hat{f}(x) - f(x))^2\,\mathrm{d}x \tag{6}$$

Since f^2 is independent of $\hat{f}$, hence minimizing $ISE(\hat{f})$ is equivalent to minimizing

$$LSCV(\hat{f}) = \int \hat{f}^2\,\mathrm{d}x - 2\int f\hat{f}\,\mathrm{d}x. \tag{7}$$

Define $\hat{f}_{-i}(x)$ to be the density estimate at x without taking into account the kernel contribution of the training point x_i, so that $\hat{f}_{-i}(x_i) =$

$\frac{1}{n-1}\frac{1}{h^d}\sum_{\substack{j=1\\ j\neq i}}^{n} k\left(\frac{x_i-x_j}{h}\right)$. We have

$$LSCV(\hat{f}) = \int \hat{f}^2 - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(x_i) \tag{8}$$

Hence $LSCV(\hat{f})$ gives a data-dependent estimator of the ISE. By using a smoothing kernel function say Gaussian kernel with an unknown bandwidth h, one can cross-validate for the optimal bandwidth that gives the smallest value for LSCV. A strong large-sample justification for using the ISE comes from Stone's result (Stone, 1984) which states that asymptotically minimizing $LSCV(\hat{f})$ is equivalent to minimizing $\int_{\mathbb{R}^d}(\hat{f}-f)^2\,\mathrm{d}x$ over all h.

3 Convex Adaptive Kernel Density Estimation Method

The CAKE estimator uses a set of finite number of base kernels and fits a kernel at each training point that is a convex combination of base kernels. However, this convex combination is allowed to vary from one training point to the other. Let $\mathcal{K}$ be a set of finite number of smoothing kernels with known bandwidths $h_1,\ldots,h_m$, where $m = O(1)$ (these bandwidths could have been pre-learnt using the training dataset or could have been provided by an oracle). The CAKE density estimator can be written as

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\alpha_{ij}}{h_j^d}k\left(\frac{x-x_i}{h_j}\right). \tag{9}$$

The problem now reduces to learning the weights α. Our aim is to minimize the L_2 regularized LSCV of the density estimator $\hat{f}$. Using equations (8,9) we get

$$LSCV(\hat{f}) \approx \int_{\mathbb{R}^d}\frac{1}{n^2}\left(\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\alpha_{ij}}{h_j^d}k\left(\frac{x-x_i}{h_j}\right)\right)^2\mathrm{d}x - \frac{2}{n^2}\sum_{i=1}^{n}\sum_{\substack{p=1\\ p\neq i}}^{n}\sum_{j=1}^{m}\frac{\alpha_{pj}}{h_j^d}k\left(\frac{x_i-x_p}{h_j}\right). \tag{10}$$

Define $Z\in\mathbb{R}^{nm\times nm}, \alpha, v\in\mathbb{R}^{nm\times 1}$ as

$$Z[ij,pl] = \int \frac{1}{n^2 h_j^d h_l^d}k\left(\frac{x-x_i}{h_j}\right)k\left(\frac{x-x_p}{h_l}\right)\mathrm{d}x$$

$$v[ij] = \frac{1}{n^2}\sum_{\substack{p=1\\ p\neq i}}^{n}\frac{1}{h_j^d}k\left(\frac{x_i-x_p}{h_j}\right). \tag{11}$$

Hence from Equation (10-11) minimizing L_2 regularized LSCV can be cast as the following optimization problem:

$$P: \min_{\alpha}\ \alpha^T Z\alpha - 2\alpha^T v + \lambda||\alpha||_2^2 \tag{12}$$

$$\text{subject to: } \sum_{j=1}^{m}\alpha_{ij} = 1 \quad \forall i = 1,\ldots n, \alpha \geq 0 \tag{13}$$

where the constraints (13) ensure that Equation (9) indeed is a legal density estimator. It is easy to see that $Z \succeq 0$ and hence the optimization problem (12-13) is a convex QP over nm variables. In order to be able to efficiently solve this QP we can reuse the standard SMO (Keerthi et al., 2001) with the constraint that both the working set variables should come from the same "block" [2] of the α vector over which the convexity constraints have been defined.

Relation to aggregation of density estimators. A closely related body of literature is that of optimal aggregation of estimators in least squares regression (Nemirovski, 2000; Tsybakov, 2003) where given m regression estimators the task is to learn an optimal aggregation of these estimators w.r.t a certain model defined by the m estimators. Popular models include convex hull, linear span of the m estimators and the original set of m estimators. The focus has been that of designing estimators whose excess risk w.r.t the optimal aggregation in the model is small. Analogously Rigollet *et al* (Rigollet & Tsybakov, 2007) have investigated optimal aggregation of density estimators. In this work the authors learn the best linear/convex combinations of given base density estimators that minimizes the expected ISE. As an example they consider the case when the m density estimators are all kernel density estimators with Gaussian kernels of different bandwidths. Now if we place additional restrictions on α_{ij}'s so that $\forall i,j : \alpha_{ij} = \gamma_j$ and $\sum_{j=1}^{m}\gamma_j = 1$ then the optimization problem proposed in Equation (12-13) along with these new additional restrictions tries to find an optimal estimator (in the ISE sense) in the convex hull defined of the density estimators $\hat{f}_{n,1},\ldots\hat{f}_{n,m}$, where $\hat{f}_{n,1},\ldots\hat{f}_{n,m}$ are the m kernel density estimators defined by the m base kernels with bandwidths $h_1,\ldots h_m$. However, without the above mentioned restrictions on α vector, our model is richer than the convex aggregation of $\hat{f}_{n,1},\ldots\hat{f}_{n,m}$, and is in some sense a "local convex " aggregation of the these estimators.

[2] Here α vector can be seen as having n blocks of size m each and the convexity constraints in Equation (13) are only over each block of α vector and not across the blocks.

4 MSE for CAKE like density estimators and L_1 risk of CAKE

We are interested in analyzing the MSE and its optimal value for CAKE like density estimator at a point x_0, where $MSE(x_0) \stackrel{\text{def}}{=} \mathbb{E}[\hat{f}(x_0) - f(x_0)]^2$. By CAKE like density estimators we mean estimators of the type

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_{ij}k\Big(\frac{x-x_i}{h_j}\Big) \tag{14}$$

where $\alpha_{ij} \geq 0$, $\sum_{j=1}^{m}\alpha_{ij} = 1$ are fixed constants. We need the following definitions (Nemirovski, 2000) and assumptions.

Definition 1. *Let $\beta, L > 0$. The Hölder class $\Sigma(\beta, L)$ is defined as the set of all functions $f : [0,1]^d \to \mathbb{R}$ which are $l = \lfloor\beta\rfloor$ times differentiable and the derivatives satisfy*

$$|D^l f(x)[\underbrace{h,\ldots h}_{l \text{ times}}] - D^l f(x')[\underbrace{h,\ldots h}_{l \text{ times}}]| \leq L|x-x'|^{\beta-l}|h|^l \ \forall x, x' \in [0,1]^d, h \in \mathbb{R}^d$$

where $\lfloor\beta\rfloor$ is the greatest integer strictly less than β.

Definition 2. *Let $l \geq 1$ be an integer. We say that a kernel $k : \mathbb{R}^d \to \mathbb{R}^d$ has order l if $\forall j_1, \ldots j_d \geq 0$ such that $\sum_{i=1}^{d} j_i \leq l$ we have*

$$\int_{u\in\mathbb{R}^d} k(u)\ \mathrm{d}u = 1, \int_{u\in\mathbb{R}^d} u_1^{j_1}u_2^{j_2}\ldots u_d^{j_d}k(u)\ \mathrm{d}u = 0$$

If $d = 1$, then the above condition becomes $\int_{u\in\mathbb{R}} k(u)\ \mathrm{d}u = 1, \int_{u\in\mathbb{R}} u^j k(u)\ \mathrm{d}u = 0\ \forall j = 1\ldots l$.

Assumption 1 (A1)**.** *The set $\mathcal{K}$ has smoothing kernels whose bandwidths $h_j\ \forall j = 1,\ldots,m$ satisfy the constraint $\frac{h_{j_1}}{h_{j_2}} = c_{j_1 j_2}\ \forall j_1, j_2 = 1\ldots m$ where $0 < c_{j_1 j_2} < \infty$ and $h_j \to 0$ as $n \to \infty\ \ \forall j = 1\ldots m$.*

Assumption 2 (A2)**.** *The true density function f belongs to the Hölder class $\Sigma(\beta, L)$ and all the base kernels are of order $l = \lfloor\beta\rfloor$. Also $C_1 \stackrel{\text{def}}{=} \int_{\mathbb{R}^d} k^2(\theta)\ \mathrm{d}\theta < \infty, C_2 \stackrel{\text{def}}{=} \int_{\mathbb{R}^d} |\theta|^\beta k(\theta)\ \mathrm{d}\theta < \infty$.*

Assumption A1 guarantees that as we see more and more samples the bandwidths all tend to 0 at the same rate. Assumption A2 is satisfied for most commonly used smoothing kernels such as a Gaussian, Epanechnikov kernels. Our main result is that given a large enough sample from the distribution, the optimal MSE of the CAKE like estimator is $O(n^{-\frac{2\beta}{2\beta+d}})$ which is known to be optimal in a minimax sense for the Hölder class of densities $\Sigma(\beta, L)$ (Tsybakov, 2009). The proof of Lemma (1) is similar to the proofs of Propositions (1.1, 1.2) in (Tsybakov, 2009). [3]

[3] Due to lack of space we have postponed the full proofs to the supplementary material.

Lemma 1. *Consider the CAKE like density estimator as shown in Equation (14). Let $b(x_0)$ and $\sigma^2(x_0)$ denote the bias and variance of CAKE like density estimators. Then the estimator $\hat{f}$ under assumptions A1, A2 satisfies: $\sigma^2(x_0) \leq \frac{C_1 f_{max}}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\alpha_{ij}^2}{h_j^d}$, $|b(x_0)| \leq \frac{C_2 L}{n!}\sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_{ij}h_j^\beta$, $MSE(x_0) \leq \alpha^T M\alpha$ where f_{max} is the maximum value of the underlying density and $C_3 = \frac{1}{n^2}(\frac{C_2 L}{l!})^2, C_4 = \frac{C_1 f_{max}}{n^2}$ and $|\cdot|$ is the standard Euclidean norm on $\mathbb{R}^d$, and $M \in \mathbb{R}^{nm\times nm}$ is defined as*

$$M[ij,pl] = \begin{cases} C_3 h_j^{2\beta} + \frac{C_4}{h_j^d} & \text{if } i = p \text{ and } j = l \\ C_3 h_j^\beta h_l^\beta & \text{otherwise.} \end{cases} \tag{15}$$

Lemma 2. *Consider the optimization problem*

$$P1 : \min_{\alpha\in R^{nm\times 1}} \alpha^T M\alpha$$
$$\text{subject to} : \sum_{j=1}^{m}\alpha_{ij} = 1\ \forall i = 1,\ldots,n, \quad \alpha \geq 0.$$

Under assumptions A1, A2 and for $n \geq n_0(f_{max}, \beta, d, L)$ the optimal value of the objective is $\mathbf{1}_n^T(AM^{-1}A^T)^{-1}\mathbf{1}_n$, where $A \in R^{n\times nm}$ and the r^{th} row of the matrix A is given by $[\underbrace{\mathbf{0}_m,\ldots,\mathbf{0}_m}_{r-1 \text{ times}}, \mathbf{1}_m, \underbrace{\mathbf{0}_m,\ldots\mathbf{0}_m}_{nm-r \text{ times}}]^T$. Also the optimal value of $MSE(x_0) = O(n^{-\frac{2\beta}{2\beta+d}})$ is attained when $h_j = \Theta(n^{-\frac{1}{2\beta+d}})$.

Proof Sketch. *Let P4 be the optimization problem P1 but without the positivity constraints. Lemmas (4-9) in the supplement establish the equivalence of problems P1 and P4 under assumption A1, and large enough n. The solution of problem P4 is derived in Lemma (10) using Lagrangian. The second part of the proof requires rewriting the upper bound on MSE as $\mathbf{1}_n^T B\mathbf{1}_n$ for an appropriate matrix B (Lemma (11)), followed by a spectral analysis in Lemmas (12-15).*

Theorem 3. *Under assumptions A1, A2 and $h_j = \Theta(n^{-\frac{1}{2\beta+d}})$, $\forall n \geq n_0(f_{max}, \beta, d, L)$ the CAKE estimator satisfies*

$$\sup_{x_0\in\mathbb{R}^d}\sup_{f\in\Sigma(\beta,L)} \mathbb{E}_{\mathcal{D}^n}\Big[(\hat{f}(x_0) - f(x_0))^2\Big] = O(n^{-\frac{2\beta}{2\beta+d}}).$$

Proof. It is enough to prove that $\forall f \in \Sigma(\beta, L), x_0 \in \mathbb{R}^d : f_{\max} < C < \infty$, for some universal C. Then the result follows from Lemma (2). Choose bounded smoothing kernels with $h_j = 1$. Now we have from the Lemma (1) that

$$f(x_0) \leq \frac{C_2 L}{l!} + \int K(x-z)f(z)\ \mathrm{d}z \leq \frac{C_2 L}{l!} + K_{\max} < \infty.$$

Since the R.H.S. is independent of f, x_0, one can choose C to be the R.H.S. of the above equation. □

Our next result is regarding the L_1 risk of CAKE in terms of its empirical L_1 risk.

Theorem 4. *Suppose we are given fixed bandwidths $h_1, \ldots, h_m$, and the underlying density function f is bounded by a constant B. Let $c_d = (\sqrt{2\pi})^d$. Then with probability at least $1-\delta$ over the input training samples, the CAKE estimator $\hat{f}$ given by the problem (12-13) with Gaussian base kernels satisfies the risk bound*

$$\mathbb{E}_{x\sim\mathcal{D}}|\hat{f}(x) - f(x)| \leq \frac{1}{n}\sum_{i=1}^{n}|\hat{f}(x_i) - f(x_i)| + \left[\frac{4}{c_d}\Big[\sum_{j=1}^{m}\frac{1}{h_j^d} + \frac{2\sqrt{2}}{\sqrt{c_d}\sqrt{\lambda}}\sqrt{\sum_{j=1}^{m}\frac{1}{h_j^{2d}}}\sqrt{\sum_{j=1}^{m}\sum_{l=1}^{m}\frac{1}{(\sqrt{h_j^2+h_l^2})^d}}\Big] + B + \sum_{j=1}^{m}\frac{1}{c_d h_j^d}\right]\sqrt{\frac{ln(\frac{1}{\delta})}{2n}}. \quad (16)$$

Proof Sketch. *The proof proceeds by bounding the uniform stability of CAKE w.r.t the loss function $|\hat{f}(x) - f(x)|$ in Lemma (16). Then applying Theorem (15) we get the desired result.*

5 Empirical Results

We implemented all the algorithms in C++ as a part of the open source machine learning toolbox MLPACK (Gray et al., 2009) and compared the estimators on both 1-d and 2-d, synthetic and natural datasets. Marron and Wand (Marron & Wand, 1992) proposed a set of 15 synthetic distributions as a testbed. These mixtures have varying levels of smoothness and modality and serve as an ideal benchmark for us to compare the different density estimators. Due to lack of space we shall investigate the performance of CAKE on four synthetic datasets sampled from skewed unimodal density (SUD), outlier density (OD), bimodal density (BD), trimodal density (TD), and the two famous natural datasets namely Old Faithful geyser dataset and suicide dataset (Silverman, 1986; Sain, 1994). Full experimental results can be found in the supplementary material. Suicide dataset has measurements of duration of hospitalization of attempted suicide patients. Two versions of the Old Faithful dataset are available. The first version has 107 observations (1-d) measuring the eruption length, while the second version has 272 observations (2-d) of both the eruption length as well as the waiting time between eruptions. We used both these datasets for 1-dimensional and multi-dimensional experiments respectively. For our experiments with 1-d distributions we sampled 1600 points for training and tested the final density estimators on another 800 points sampled from the same distribution. To learn the CAKE estimator we used a set of 10 base kernels (all Gaussians) with bandwidths in the range $[\frac{h_p}{3}, 3h_p]$ where h_p is the plugin bandwidth calculated using the equation $h_p = \left(\frac{4}{d+2}\right)^{\frac{1}{d+4}}\hat{\sigma}n^{-\frac{1}{d+4}}$ where $\hat{\sigma}$ is the empirical standard deviation. The bandwidth parameter h used in AKDE, VKDE and RSDE, and the regularization parameter λ used in CAKE were all found by cross-validation. The parameters for local RODEO were chosen as $c_n = log(d), c_0 =$ range of training data, $\beta = 0.9$. These are the settings that were used by Liu *et al* in their paper (Liu et al., 2007). Sain (Sain, 1994) observes that the eruption length of the Old Faithful dataset has 2 modes of approximately equal height separated by a smooth valley, while the Suicide dataset has a unimodal distribution with a long tail. We report the RMSE of different density estimators in Table (2), and show plots of the different density estimators in Figure (2). It is clear from the plots that RSDE tends to over-smooth the distributions and RODEO fails to capture multiple modes in a distribution and gives very rough density estimates in the tails and valleys. CAKE tends to give smoother estimates than AKDE but at the same time captures all the features of the distribution well. VKDE is generally seen to give noisy density estimates. On the Suicide dataset all the density estimators except VKDE show a unimodal structure with a long tail. However, RODEO shows heavy tails and RSDE flattens out the main mode. On the other hand AKDE exaggerates the size of the main mode. The tail behaviour of AKDE and CAKE are better than the other estimators. On the Old Faithful dataset both AKDE and CAKE show a bimodal structure, with CAKE capturing the property of equal mode size better than AKDE. RODEO completely smooths the first mode. RSDE on both these datasets gives ultra smooth density estimates without showing the important features of the distribution. As can be seen from Figure (2) the regularization term helps learn smoother density estimates. The presence of the regularization term is especially important in our problem formulation, because unlike the data-splitting scheme used in aggregation of estimators (Nemirovski, 2000) the same training sample is being used to learn the base kernel bandwidths (by calculating the plug-in bandwidth) and also the final CAKE density estimator. For multidimensional experiment the dataset was whitened for computational purposes. Whitening the dataset is equivalent to working with the original data with the bandwidth of the kernel chosen according to the covariance matrix of the distribution (p.78 of (Silverman, 1986)). On the multi-dimensional version of the Old Faithful dataset

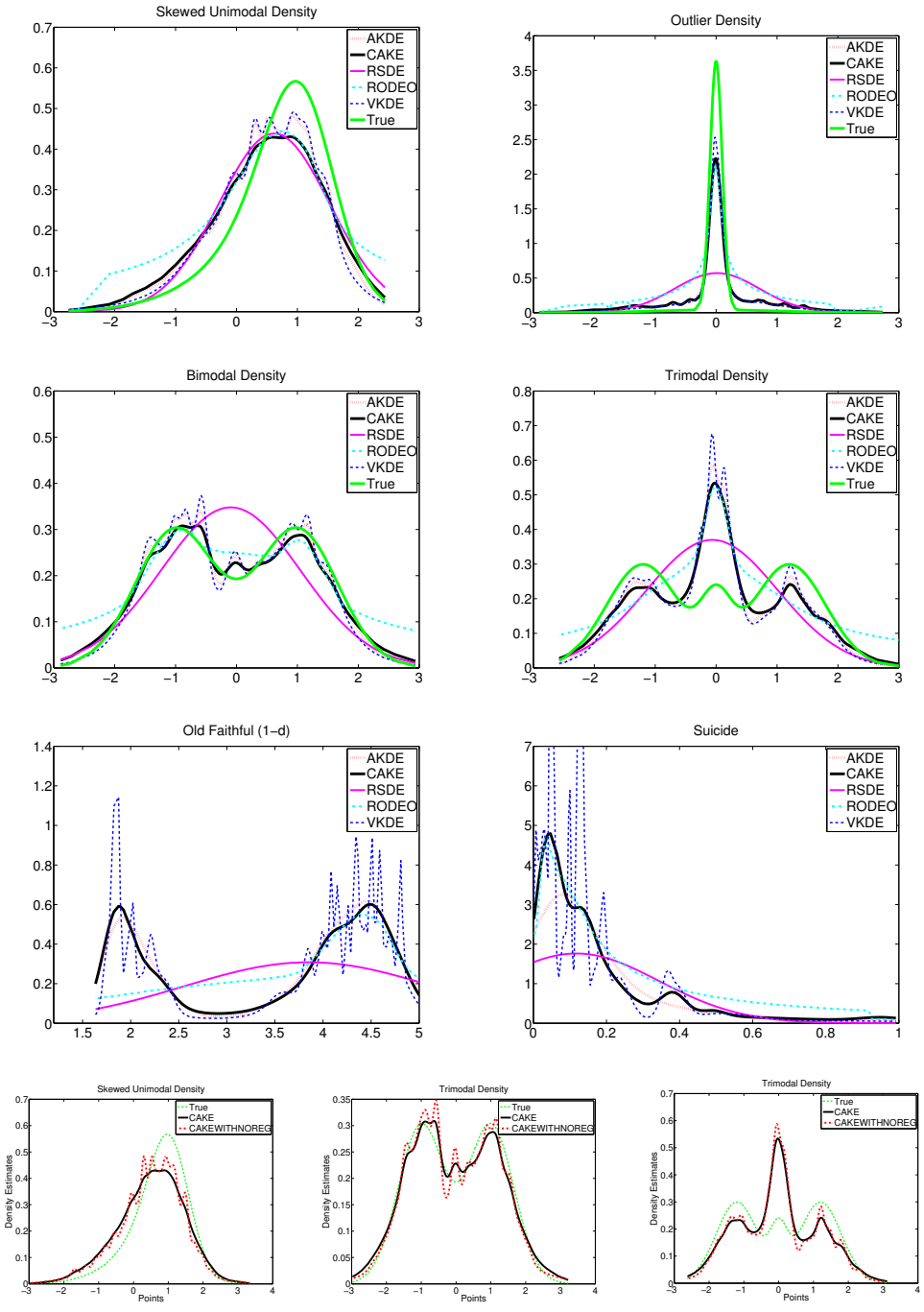

Figure 2: Performance of AKDE, RSDE, CAKE and RSDE on synthetic and natural datasets. The Old Faithful (1-d) models the distribution of eruption lengths of the Old Faithful Geyser. The Suicide dataset models the distribution of(scaled) length of hospitalization of an attempted suicide patient. The old faithful dataset can be obtained from www.stat.cmu.edu/ larry/all-of-statistics/=data/faithful.dat. The Suicide dataset can be obtained from (Silverman, 1986).The plots in the last row show the impact of regularization on the smoothness of the CAKE density estimator.

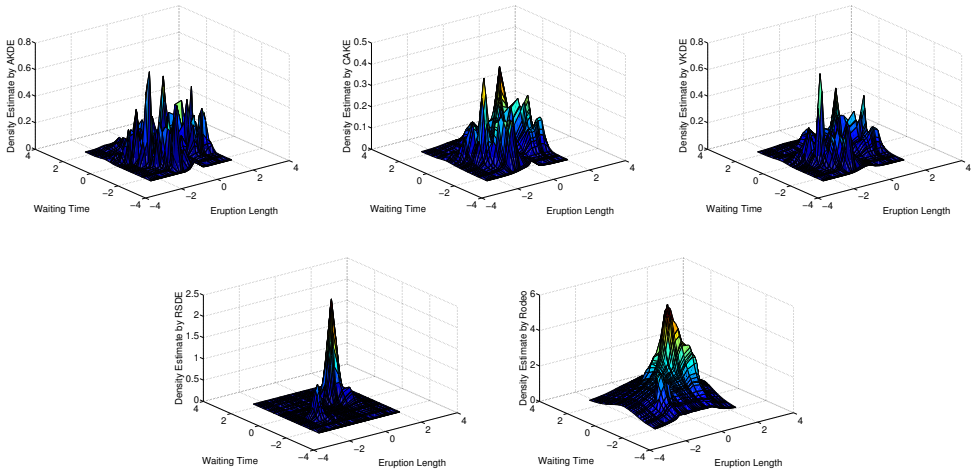

Figure 3: Performance of density estimators on the multi-dimensional Old Faithful dataset.

Dataset	Train/Test size	CAKE	AKDE	Rodeo	SVM
Banana	400/4900	**86.25**	85.4	62.82	88.70
Flare Solar	666/400	**65.0**	55.0	3.25	66.50
Twonorm	400/7000	93.27	**96.51**	50	97.04
Heart	169/99	**79.0**	45.0	32.00	82.0
Titanic	149/2050	**74.12**	74.11	17.21	74.12
Ringnorm	400/7000	**66.10**	50.0	47.63	98.50
German	700/300	**74.00**	18.00	50.00	79.70

Table 1: Comparison of different density estimators when used as classifiers on some UCI datasets. These datasets can be obtained from http://ida.first.fraunhofer.de/projects/bench/benchmarks.htm

CAKE (Figure 3) captures the bimodal nature of the distribution better than the other density estimators. RODEO reflects the bi-modality in the distribution but heavily overestimates the modes.

CAKE as a classifier. Non parametric kernel classification rule (KCR) (Devroye et al., 1996) learns a binary classifier that labels a point as +1 if the kernel contribution to the density of positively labeled points is larger than those of the negative points which is In this experiment we use different density estimators in KCR and compare the accuracies of the resultant classifier. However, we want to see via suggestive experiments as to how CAKE works in high dimensions when compared to other density estimators. Table (1) suggests good performance of CAKE over other estimators *even though we didn't learn estimators specifically designed for classification task.*

6 Conclusions

We proposed a new kernel density estimator called CAKE which fits kernels at different training points by learning different convex aggregations of base kernels at different training points. We analyzed CAKE theoretically and observed empirically it performs better than most estimators. It would be interesting to see if CAKE with univariate kernels without the convexity constraints can be used in non-parametric regression with group lasso to learn regression functions that are not necessarily globally sparse but are locally sparse. In the present form CAKE requires bandwidths $h_1, \ldots, h_m$. This provides a mechanism for the user to inject domain knowledge. A nice extension of our present framework would involve learning different bandwidths along different dimensions which can be seen as learning the basis set as in sparse coding. Once the bandwidths of the base kernels are learnt we can then use them in the CAKE framework. On the theoretical side we have provided a risk bound that depends on an unknown quantity which is the empirical L_1 distance between our estimator and the true density function. If one can give a data-dependent bound for this quantity then one can use Theorem (4) to provide a completely data-dependent bound on the L_1 distance between the CAKE estimator and the true density. Extension of our analysis to inhomogenous density functions such as Besov spaces is another fruitful direction.

Density	CAKE	Adaptive	Variable	RODEO	RSDE
SUD	0.096	0.083	0.0827	0.098	0.106
OD	0.80	0.85	0.691	0.886	1.70
BD	0.021	0.023	0.197	0.031	0.079
TD	0.141	0.156	0.167	0.145	0.119

Table 2: RMSE values of different density estimators on various synthetic 1-d distribution.

References

Bishop, C., et al. (2006). *Pattern recognition and machine learning*. Springer New York:.

Breiman, L., Meisel, W., & Purcell, E. (1977). Variable kernel estimates of multivariate densities. *Technometrics*.

Devroye, L., Gyrfi, L., & Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer.

Devroye, L., & Lugosi, G. (2000). *Combinatorial methods in density estimation*. Springer-Verlag.

Gasser, T., Mller, H. G., & Mammitzsch, V. (1985). Kernels for nonparametric curve estimation. *JRSS B*.

Girolami, M., & He, C. (2003). Probability density estimation from optimally condensed data samples. *IEEE PAMI*.

Gray, A., et al. (2009). Mlpack. `http://mloss.org/software/view/152/`.

Hall, P., Marron, J., & Titterington, D. (1995). On partial local smoothing rules for curve estimation. *Biometrika*.

Jones, M., Marron, J., & Sheather, S. (1996). A brief survey of bandwidth selection for density estimation. *JASA*.

Keerthi, S., et al. (2001). Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computation*.

Lanckriet, G. R. G., Cristianini, N., Bartlett, P., El Ghaoui, L., & Jordan, M. I. (2004). Learning the Kernel Matrix with Semidefinite Programming. *JMLR*, *5*, 27–72.

Liu, H., Lafferty, J., & Wasserman, L. (2007). Sparse nonparametric density estimation in high dimensions using the rodeo. *AISTATS*.

Marron, J., & Wand, M. (1992). Exact mean integrated squared error. *The Annals of Statistics*, 712–736.

Nemirovski, A. (2000). Topics in non-parametric statistics. *Lectures on probability theory and statistics*.

Ong, C. S., Smola, A., & Williamson, R. C. (2005). Learning the Kernel with Hyperkernels. *JMLR*.

Parzen, E. (1962). On the estimation of a probability density function and mode. *Annals of Mathematical Statistics*.

Rigollet, P., & Tsybakov, A. (2007). Linear and convex aggregation of density estimators. *Mathematical Methods of Statistics*.

Sain, S. (1994). *Adaptive kernel density estimation*. Doctoral dissertation, Rice University.

Shawe-Taylor, J., & Dolia, A. (2007). A Framework for Probability Density Estimation. *AISTATS*.

Silverman, B. (1986). *Density estimation for statistics and data analysis*. Chapman & Hall/CRC.

Song, L., Zhang, X., Gretton, A., Schlkopf, B., Smola, A., & Skolnick, J. (2008). Tailoring density estimation via reproducing kernel moment matching. *ICML*.

Stone, C. J. (1984). An Asymptotically Optimal Window Selection Rule for Kernel Density Estimates. *The Annals of Statistics*, *12*.

Tsybakov, A. (2003). Optimal rates of aggregation. *COLT proceedings* (p. 303).

Tsybakov, A. (2009). *Introduction to nonparametric estimation*. Springer Verlag.

Vapnik, V. N., & Mukherjee, S. (1999). Support vector method for multivariate density estimation. *Advances in NIPS*.