# A Statistical Perspective on Coreset Density Estimation

**Paxton Turner**
Massachusetts Institute of Technology

**Jingbo Liu**
University of Illinois at Urbana–Champaign

**Philippe Rigollet**
Massachusetts Institute of Technology

## Abstract

Coresets have emerged as a powerful tool to summarize data by selecting a small subset of the original observations while retaining most of its information. This approach has led to significant computational speedups but the performance of statistical procedures run on coresets is largely unexplored. In this work, we develop a statistical framework to study coresets and focus on the canonical task of nonparameteric density estimation. Our contributions are twofold. First, we establish the minimax rate of estimation achievable by coreset-based estimators. Second, we show that the practical coreset kernel density estimators are near-minimax optimal over a large class of Hölder-smooth densities.

## 1 Introduction

The ever-growing size of datasets that are routinely collected has led practitioners across many fields to contemplate effective data summarization techniques that aim at reducing the size of the data while preserving the information that it contains. While there are many ways to achieve this goal, including standard data compression algorithms, they often prevent direct manipulation of data for learning purposes. *Coresets* have emerged as a flexible and efficient set of techniques that permit direct data manipulation. Coresets are well-studied in machine learning (Har-Peled & Kushal, 2007; Feldman *et al.*, 2013; Bachem *et al.*, 2017, 2018; Karnin & Liberty, 2019), statistics (Feldman *et al.*, 2011; Zheng & Phillips, 2017; Munteanu *et al.*, 2018; Huggins *et al.*, 2016; Phillips & Tai, 2018a,b), and computational geometry (Agarwal *et al.*, 2005; Clarkson, 2010; Frahling & Sohler, 2005; Gärtner & Jaggi, 2009; Claici *et al.*, 2020).

Given a dataset $\mathcal{D} = \{X_1, \dots, X_n\} \subset \mathbb{R}^d$ and task (density estimation, logistic regression, etc.) a coreset $\mathcal{C}$ is given by $\mathcal{C} = \{X_i : i \in S\}$ for some subset $S$ of $\{1, \dots, n\}$ of size $|S| \ll n$. A good coreset should suffice to perform the task at hand with the same accuracy as with the whole dataset $\mathcal{D}$.

In this work we study the canonical task of density estimation. Given i.i.d random variables $X_1, \dots, X_n \sim \mathbb{P}_f$ that admit a common density $f$ with respect to the Lebesgue measure over $\mathbb{R}^d$, the goal of density estimation is to estimate $f$. It is well known that the minimax rate of estimation over the $L$-Hölder smooth densities $\mathcal{P}_{\mathcal{H}}(\beta, L)$ of order $\beta$ is given by

$$\inf_{\hat{f}} \sup_{f \in \mathcal{P}_{\mathcal{H}}(\beta,L)} \mathbb{E}_f \|\hat{f} - f\|_2 = \Theta_{\beta,d,L}(n^{-\frac{\beta}{2\beta+d}}), \quad (1)$$

where the infimum is taken over all estimators based on the dataset $\mathcal{D}$. Moreover the minimax rate above is achieved by a kernel density estimator

$$\hat{f}_n(x) := \frac{1}{nh^d} \sum_{j=1}^{n} k\left(\frac{X_i - x}{h}\right) \quad (2)$$

for suitable choices of kernel $k : \mathbb{R}^d \to \mathbb{R}$ and bandwidth $h > 0$ (see e.g. Tsybakov, 2009, Theorem 1.2).

The main goal of this paper is to extend this understanding of rates for density estimation to estimators based on coresets. Specifically we would like to characterize the statistical performance of coresets in terms of their cardinality. To do so, we investigate two families of estimators built on coresets: one that is quite flexible and allows arbitrary estimators to be used on the coreset and another that is more structured and driven by practical considerations; it consists of weighted kernel density estimators built on coresets.

### 1.1 Two statistical frameworks for coreset density estimation

We formally define a coreset as follows. Throughout this work $m = o(n)$ denotes the cardinality of the core-

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. 

set. Given $x \in \mathbb{R}^{d\times n}$, let $S = S(y|x)$ denote a conditional probability measure on the set $\binom{[n]}{m}$ of subsets of $[n] = \{1, 2, \ldots, n\}$ of cardinality $m$. In information theoretic language, $S$ is a channel from $\mathbb{R}^{d\times n}$ to subsets of cardinality $m$. We refer to the channel $S$ as a *coreset scheme* because it designates a data-driven method of choosing a subset of data points. In what follows, we abuse notation and let $S = S(x)$ denote an instantiation of a sample from the measure $S(y|x)$ for $x \in \mathbb{R}^{d\times n}$. A *coreset* $X_S$ is then defined to be the projection of the dataset $X = (X_1, \ldots, X_n)$ onto the subset indicated by $S(X)$: $X_S := \{X_i\}_{i\in S(X)}$.

The first family of estimators that we investigate is quite general and allows the statistician to select a coreset and then employ an estimator that only manipulates data points in the coreset to estimate an unknown density. To study coresets, it is convenient to make the dependence of estimators on observations more explicit than in the traditional literature. More specifically, a density estimator $\hat{f}$ based on $n$ observations $X_1, \ldots, X_n \in \mathbb{R}^d$ is a function $\hat{f} : \mathbb{R}^{d\times n} \to L^2(\mathbb{R}^d)$ denoted by $\hat{f}[X_1, \ldots, X_n](\cdot)$. Similarly, a *coreset-based estimator* $\hat{f}_S$ is constructed from a coreset scheme $S$ of size $m$ and an estimator (measurable function) $\hat{f} : \mathbb{R}^{d\times m} \to L^2(\mathbb{R}^d)$ on $m$ observations. We enforce the additional restriction on $\hat{f}$ that for all $y_1, \ldots, y_m \in \mathbb{R}^d$ and for all bijections $\pi : [m] \to [m]$, it holds that $\hat{f}[y_1, \ldots, y_m](\cdot) = \hat{f}[y_{\pi(1)}, \ldots, y_{\pi(m)}](\cdot)$. Given $S$ and $\hat{f}$ as above, we define the *coreset-based estimator* $\hat{f}_S : \mathbb{R}^{d\times n} \to L^2(\mathbb{R}^d)$ to be the function $\hat{f}_S[X](\cdot) := \hat{f}[X_S](\cdot) : \mathbb{R}^d \to \mathbb{R}$. We evaluate the performance of coreset-based estimators in Section 2 by characterizing their rate of estimation over Hölder classes.[1]

The symmetry restriction on $\hat{f}$ prevents the user from exploiting information about the ordering of data points to their advantage: the only information that can be used by the estimator $\hat{f}$ is contained in the unordered collection of distinct vectors given by the coreset $X_S$.

As evident from the results in Section 2, the information-theoretically optimal coreset estimator does not resemble coreset estimators employed in practice. To remedy this limitation, we also study *weighted coreset kernel density estimators* (KDEs) in Section 3. Here the statistician selects a kernel $k$, bandwidth parameter $h$, and a coreset $X_S$ of cardinality $m$ as defined above and then employs the estimator

$$\hat{f}_S(y) = \sum_{j\in S} \lambda_j h^{-d} k\left(\frac{X_j - y}{h}\right),$$

where the weights $\{\lambda_j\}_{j\in S}$ are nonnegative, sum to one and are allowed to depend on the full dataset.

In the case of uniform weights where $\lambda_j = \frac{1}{m}$ for all $j \in S$, coreset KDEs are well-studied (see e.g. Bach *et al.*, 2012; Harvey & Samadi, 2014; Phillips & Tai, 2018a,b; Karnin & Liberty, 2019). Interestingly, our results show that allowing flexibility in the weights gives a definitive advantage for the task of density estimation. By Theorems 2 and 5, the uniformly weighted coreset KDEs require a much larger coreset than that of weighted coreset KDEs to attain the minimax rate of estimation over univariate Lipschitz densities.

### 1.2 Setup and Notation

We reserve the notation $\|\cdot\|_2$ for the $L^2$ norm and $|\cdot|_p$ for the $\ell^p$-norm. The constants $c, c_{\beta,d}, c_L$, etc. vary from line to line and the subscripts indicate parameter dependences.

Fix an integer $d \geq 1$. For any multi-index $s = (s_1, \ldots, s_d) \in \mathbb{Z}^d_{\geq 0}$ and $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, define $s! = s_1! \cdots s_d!$, $x^s = x_1^{s_1} \cdots x_d^{s_d}$ and let $D^s$ denote the differential operator defined by

$$D^s = \frac{\partial^{|s|_1}}{\partial x_1^{s_1} \cdots \partial x_d^{s_d}}.$$

We reserve the notation $|s|$ for the coordinate-wise application of $|\cdot|$ to the multi-index $s$.

Fix a positive real number $\beta$, and let $\lfloor\beta\rfloor$ denote the maximal integer *strictly* less than $\beta$. Given $L > 0$ we let $\mathcal{H}(\beta, L)$ denote the space of Hölder functions $f : \mathbb{R}^d \to \mathbb{R}$ that are supported on the cube $[-1/2, 1/2]^d$, are $\lfloor\beta\rfloor$ times differentiable, and satisfy

$$|D^s f(x) - D^s f(y)| \leq L\,|x - y|_2^{\beta - \lfloor\beta\rfloor},$$

for all $x, y \in \mathbb{R}^d$ and for all multi-indices $s$ such that $|s|_1 = \lfloor\beta\rfloor$.

Let $\mathcal{P}_{\mathcal{H}}(\beta, L)$ denote the set of probability density functions contained in $\mathcal{H}(\beta, L)$. For $f \in \mathcal{P}_{\mathcal{H}}(\beta, L)$, let $\mathbb{P}_f$ (resp. $\mathbb{E}_f$) denote the probability distribution (resp. expectation) associated to $f$.

For $d \geq 1$ and $\gamma \in \mathbb{Z}_{\geq 0}$, we also define the Sobolev functions $\mathcal{S}(\gamma, L')$ that consist of all $f : \mathbb{R}^d \to \mathbb{R}$ that are $\gamma$ times differentiable and satisfy

$$\|D^\alpha f\|_2 \leq L'$$

for all multi-indices $\alpha$ such that $|\alpha|_1 = \gamma$.

[1] Our notion of coreset-based estimators bares conceptual similarity to various notions of *compression schemes* as studied in the literature, e.g. Littlestone & Warmuth (1986); Ashtiani *et al.* (2020); Hanneke *et al.* (2019).

Given $f \in L^2$, we define the Fourier transform $\mathcal{F}[f] : \mathbb{R}^d \to \mathbb{R}$ by

$$\mathcal{F}[f](\omega) = \int_{\mathbb{R}^d} f(x) e^{-i\langle x,\omega\rangle} \mathrm{d}x.$$

## 2 Coreset-based estimators

In this section we study the performance of coreset-based estimators. Recall that coreset-based estimators are estimators that only depend on the data points in the coreset.

Define the *minimax risk for coreset-based estimators* $\psi_{n,m}(\beta, L)$ over $\mathcal{P}_{\mathcal{H}}(\beta, L)$ to be

$$\psi_{n,m}(\beta, L) = \inf_{\hat{f}, |S|=m} \sup_{f \in \mathcal{P}_{\mathcal{H}}(\beta,L)} \mathbb{E}_f \, \|\hat{f}_S - f\|_2, \quad (3)$$

where the infimum above is over all choices of coreset scheme $S$ of cardinality $m$ and all estimators $\hat{f} : \mathbb{R}^{d\times m} \to L^2(\mathbb{R}^d)$.

Our main result on coreset-based estimators characterizes their minimax risk.

**Theorem 1.** *Fix $\beta, L > 0$ and an integer $d \geq 1$. Assume that $m = o(n)$. Then the minimax risk of coreset-based estimators satisfies*

$$\inf_{\hat{f}, |S|=m} \sup_{f \in \mathcal{P}_{\mathcal{H}}(\beta,L)} \mathbb{E}_f \, \|\hat{f}_S - f\|_2 = \Theta_{\beta,d,L}(n^{-\frac{\beta}{2\beta+d}} + (m \log n)^{-\frac{\beta}{d}}).$$

The above theorem readily yields a characterization of the minimal size $m^*(\beta, d)$ that a coreset can have while still enjoying the minimax optimal rate $n^{-\frac{\beta}{2\beta+d}}$ from (1). More specifically, let $m^* = m^*(n)$ be such that

(i) if $m(n)$ is a sequence such that $m = o(m^*)$, then $\liminf_{n\to\infty} n^{\frac{\beta}{2\beta+d}} \psi_{n,m}(\beta, L) = \infty$, and

(ii) if $m = \Omega(m^*)$ then $\limsup_{n\to\infty} \psi_{n,m}(\beta, L) n^{\frac{\beta}{2\beta+d}} \leq C_{\beta,d,L}$ for some constant $C_{\beta,d,L} > 0$.

Then it follows readily from from Theorem 1 that $m^* = \Theta_{\beta,d,L}(n^{\frac{d}{2\beta+d}} / \log n)$.

Theorem 1 illustrates two different curses of dimensionality: the first stems from the original estimation problem, and the second stems from the compression problem. As $d \to \infty$, it holds that $m^* \sim n/\log n$, and in this regime there is essentially no compression, as the implicit constant in Theorem 1 grows rapidly with $d$.[2]

[2]In fact, even for the classical estimation problem (1), this constant scales as $d^d$ (see McDonald, 2017, Theorem 3).

Our proof of the lower bound in Theorem 1 first uses a standard reduction from estimation to a multiple hypothesis testing problem over a finite function class. While Fano's inequality is the workhorse of our second step, note that the lower bound must hold only for coreset-based estimators and not *any* estimator as in standard minimax lower bounds. This additional difficulty is overcome by a careful handling of the information structure generated by coreset scheme channels rather than using off-the-shelf results for minimax lower bounds. The full details of the lower bound are in the Supplement.

The estimator achieving the rate in Theorem 1 relies on an encoding procedure. It is constructed by building a dictionary between the subsets in $\binom{[n]}{m}$ and an $\varepsilon$-net on the space of Hölder functions. The key idea is that, for $1 \ll m \leq n/2$, the amount of subsets $\binom{n}{m}$ grows rapidly with $m$, so for $m$ large enough, there is enough information to encode a nearby-neighbor in $L^2(\mathbb{R}^d)$ to the kernel density estimator on the entire dataset.

### 2.1 Proof of the upper bound in Theorem 1

Fix $\varepsilon = c^*(m \log n)^{-\frac{\beta}{d}}$ for $c^*$ to be determined and let $\mathcal{N}_\varepsilon$ denote an $\varepsilon$-net of $\mathcal{P}_{\mathcal{H}}(\beta, L)$ with respect to the $L^2([-\frac{1}{2}, \frac{1}{2}]^d)$ norm. It follows from the classical Kolmogorov-Tikhomorov bound (see, e.g., Theorem XIV of Tikhomirov, 1993) that there exists a constant $C_{\mathsf{KT}}(\beta, d, L) > 0$ such that we can choose $\mathcal{N}_\varepsilon$ with $\log |\mathcal{N}_\varepsilon| \leq C_{\mathsf{KT}}(\beta, d, L)\, \varepsilon^{-d/\beta}$. In particular, there exists $\mathsf{f} \in \mathcal{N}_\varepsilon$ such that $\|\hat{f}_n - \mathsf{f}\|_2 \leq \varepsilon$ where $\hat{f}_n$ is the minimax optimal kernel density estimator defined in (2).

We now develop our encoding procedure for $\mathsf{f}$. To that end, fix an integer $K \geq m$ such that $\binom{K}{m} \geq |\mathcal{N}_\varepsilon|$ and let $\phi : \binom{[K]}{m} \to \mathcal{N}_\varepsilon$ be any surjective map. Our procedure only looks at the first coordinates of the sample $X = \{X_1, \ldots, X_n\}$. Denote these coordinates by $x = \{x_1, \ldots, x_n\}$ and note that these $n$ numbers are almost surely distinct. Let $A$ denote a parameter to be determined, and define the intervals

$$B_{ik} = [(i-1)K^{-1}A + (k-1)A, \\ (i-1)K^{-1}A + (k-1)A + K^{-1}A].$$

For $i = 1, \ldots, K$, define

$$B_i = \bigcup_{k=1}^{1/A} B_{ik}.$$

The next lemma, whose proof is in the Supplement, ensures that with high probability every bin $B_i$ contains the first coordinate $x_i$ of at least one data point.

**Lemma 1.** *Let $K^{-1} = c(\log n)/n$ for $c > 0$ a sufficiently large absolute constant, and let $A = A_{\beta,L,K}$ denote a sufficiently small constant. Then for all $f \in \mathcal{P}_{\mathcal{H}}(\beta, L)$ and $X_1, \dots, X_n \overset{iid}{\sim} \mathbb{P}_f$, the event that for every $j = 1, \dots, K$ there exists some $x_i$ in bin $B_j$ holds with probability at least $1 - O(n^{-2})$.*

In the high-probability event $\mathcal{E}$ that every bin $B_i$ contains the first coordinate of some data point, choose a unique representative $x_j^\circ \in x$ such that $x_j^\circ \in B_j$ and pick any $T_{\mathsf{f}} \in \phi^{-1}(\mathsf{f})$. Then define $S = \{i : x_i = x_j^\circ, j \in T_{\mathsf{f}}\}$. If there exists a bin with no observation, then let $X_S$ consist of two data points lying in the same bin and $m - 2$ random data points. Then set $\hat{f}_S \equiv 0$.

Note that $\hat{f}_S$ is indeed a coreset-based estimator. The function $\hat{f}$ such that $\hat{f}_S = \hat{f}[X_S]$ looks at the $m$ data points in the coreset, and if their first coordinates lie in distinct bins, then $X_S$ is decoded as above to output the corresponding element $\mathsf{f}$ of the net $\mathcal{N}_\varepsilon$. Otherwise, $\hat{f} \equiv 0$.

Next, it suffices to show the upper bound of Theorem 1 in the case when $m \le cn^{d/(2\beta+d)}$ for $c$ a sufficiently small absolute constant. For $c^* = c^*_{\beta,d,L}$ sufficiently large, by Stirling's formula and our choice of $K$ it holds that

$$\log \binom{K}{m} \ge C_{\mathsf{KT}}(\beta, d, L) \left(\frac{1}{\varepsilon}\right)^{\frac{d}{\beta}} \ge \log |\mathcal{N}_\varepsilon|.$$

Hence, the surjection $\phi$ and our encoding estimator $\hat{f}_S$ are well-defined.

Next we have

$$\mathbb{E}_f \|\hat{f}_S - f\|_2 = \mathbb{E}_f\big[\|\mathsf{f} - f\|_2 \mathbb{1}_{\mathcal{E}}\big] + \mathbb{E}_f\big[\|0 - f\|_2 \mathbb{1}_{\mathcal{E}^c}\big].$$

We control the first term as follows using (1) and the fact that $\|\mathsf{f} - \hat{f}_n\|_2 \le \varepsilon$ on $\mathcal{E}$:

$$\begin{aligned}\mathbb{E}_f\big[\|\mathsf{f} - f\|_2 \mathbb{1}_{\mathcal{E}}\big] &\le \mathbb{E}_f \|\hat{f}_n - f\|_2 + \mathbb{E}_f \|\mathsf{f} - \hat{f}_n\|_2 \\ &\le c_{\beta,d,L}\left(n^{\frac{-\beta}{2\beta+d}} + (m \log n)^{-\frac{\beta}{d}}\right).\end{aligned}$$

By the Cauchy-Schwarz inequality,

$$\begin{aligned}\mathbb{E}_f\big[\|0 - f\|_2 \mathbb{1}_{\mathcal{E}^c}\big] &\le \left(\mathbb{E}_f \|f\|_2^2 \, \mathbb{P}(\mathcal{E}^c)\right)^{1/2} \\ &\le c_{\beta,d,L}\, n^{-1}.\end{aligned}$$

Put together, the previous three displays yield the upper bound of Theorem 1.

## 3 Coreset kernel density estimators

In this section, we consider the family of weighted kernel density estimators built on coresets and study its rate of estimation over the Hölder densities. In this framework, the statistician first computes a minimax estimator $\hat{f}$ using the entire dataset and then approximates $\hat{f}$ with a weighted kernel density estimator over the coreset. Here we allow the weights to be a measurable function of the entire dataset rather than just the coreset.

As is typical in density estimation, we consider kernels $k : \mathbb{R}^d \to \mathbb{R}$ of the form $k(x) = \prod_{i=1}^d \kappa(x_i)$ where $\kappa$ is an even function and $\int \kappa(x)\,\mathrm{d}x = 1$. Given bandwidth parameter $h$, we define $k_h(x) = h^{-d}\, k(\frac{x}{h})$.

### 3.1 Carathéodory coreset method

Given a KDE with uniform weights and bandwidth $h$ defined by

$$\hat{f}(y) = \frac{1}{n}\sum_{j=1}^n k_h(X_j - y),$$

on a sample $X_1, \dots, X_n$, we define a coreset KDE $\hat{g}_S$ as follows in terms of a cutoff frequency $T > 0$. Define $A = \{\omega \in \frac{\pi}{2}\mathbb{Z}^d : |\omega|_\infty \le T\}$. Consider the complex vectors $(e^{i\langle X_j, \omega\rangle})_{\omega \in A}$. By Carathéodory's theorem (Carathéodory, 1907), there exists a subset $S \subset [n]$ of cardinality at most $2(1 + \frac{4T}{\pi})^d + 1$ and nonnegative weights $\{\lambda_j\}_{j \in S}$ with $\sum_{j \in S} \lambda_j = 1$ such that

$$\frac{1}{n}\sum_{j=1}^n (e^{i\langle X_j, \omega\rangle})_{\omega \in A} = \sum_{j \in S} \lambda_j (e^{i\langle X_j, \omega\rangle})_{\omega \in A}. \quad (4)$$

Then $\hat{g}_S(y)$ is defined to be

$$\hat{g}_S(y) = \sum_{j \in S} \lambda_j k_h(X_j - y).$$

#### 3.1.1 Algorithmic considerations

For a convex polyhedron $P$ with vertices $v_1, \dots, v_n \in \mathbb{R}^D$, the proof of Carathéodory's theorem is constructive and yields a polynomial-time algorithm in $n$ and $D$ to find a convex combination of $D + 1$ vertices that represents a given point in $P$ (Carathéodory, 1907) (see also Hiriart-Urruty & Lemaréchal, 2004, Theorem 1.3.6). For completeness, we describe below this algorithm applied to our problem. Note that, more generally, for a large class of convex bodies, Carathéodory's theorem may be implemented efficiently using standard tools from convex optimization (Grötschel *et al.*, 2012, Chapter 6).

Set $D = 2|A| \le 2(1 + \frac{4T}{\pi})^d$. For $j = 1, \dots, n$, let

$$v_j = (\mathsf{Re}\, e^{i\langle X_j, \omega\rangle}, \mathsf{Im}\, e^{i\langle X_j, \omega\rangle})_{\omega \in A} \in \mathbb{R}^D.$$

Let $M$ denote the matrix with columns $(v_1, 1)^T, \dots, (v_n, 1)^T \in \mathbb{R}^{D+1}$, and let $\Delta_{n-1} \subset \mathbb{R}^n$

denote the standard simplex. Assume without loss of generality that $n \geq D+2$. Next,

1. Find a nonzero vector $w \in \ker(M)$
2. Find $\alpha > 0$ so that $\lambda_1 := \frac{1}{n}\mathbb{1} + \alpha w$ lies on the boundary of $\Delta_{n-1}$

Observe that $M\lambda_1 = (\frac{1}{n}\sum v_i, 1)^T$, and since $\lambda_1 \in \partial\Delta_{n-1}$ the average is now represented using a convex combination of at most $n-1$ of the vertices $v_1, \ldots, v_n$. As long as at least $D+2$ vertices remain, we can continue reducing the number of vertices used to represent $\frac{1}{n}\sum v_j$ by applying steps 1 and 2. Thus after at most $n-D-1$ iterations, we obtain a $(D+1)$-sparse vector $\lambda \in \Delta_{n-1}$ that satisfies $\sum \lambda_j v_j = \frac{1}{n}\sum v_i$, as desired.

### 3.2 Results on Carathéodory coresets

Proposition 1 is key to our results and specifies conditions on the kernel guaranteeing that the Carathéodory method yields an accurate estimator.

**Proposition 1.** *Let $k(x) = \prod_{i=1}^d \kappa(x_i)$ denote a kernel with $\kappa \in \mathcal{S}(\gamma, L')$ such that $|\kappa(x)| \leq c_{\beta,d}\,|x|^{-\nu}$ for some $\nu \geq \beta + d$ and such that the KDE*

$$\hat{f}(y) = \frac{1}{n}\sum_{i=1}^n k_h(X_i - y)$$

*with bandwidth $h = n^{-\frac{1}{2\beta+d}}$ satisfies*

$$\sup_{f\in\mathcal{P}_{\mathcal{H}}(\beta,L)} \mathbb{E}\|f - \hat{f}\|_2 \leq c_{\beta,d,L}\, n^{-\frac{\beta}{2\beta+d}}. \tag{5}$$

*Then the Carathéodory coreset estimator $\hat{g}_S$ constructed from $\hat{f}$ with $T = c_{d,\gamma,L'}\, n^{\frac{d/2+\beta+\gamma}{\gamma(2\beta+d)}}$ satisfies*

$$\sup_{f\in\mathcal{P}_{\mathcal{H}}(\beta,L)} \mathbb{E}\|\hat{g}_S - f\|_2 \leq c_{\beta,d,L}\, n^{-\frac{\beta}{2\beta+d}}.$$

There exists a kernel $k_s \in \mathcal{C}^\infty$ that satisfies the conditions above for all $\beta$ and $\gamma$. We sketch the details here and postpone the full argument to the Proof of Theorem 2 in the Supplement. Let $\psi : \mathbb{R} \to [0,1]$ denote a cutoff function that has the following properties: $\psi \in \mathcal{C}^\infty$, $\psi\big|_{[-1,1]} \equiv 1$, and $\psi$ is supported on $[-2,2]$. Define $\kappa_s(x) = \mathcal{F}[\psi](x)$, and let $k_s(x) = \prod_{i=1}^d \kappa_s(x_i)$ denote the resulting kernel. Observe that for all $\beta > 0$, the kernel $k_s$ satisfies

$$\text{ess sup}_{\omega\neq 0} \frac{|1 - \mathcal{F}[k_s](\omega)|}{|\omega|^\alpha} \leq 1, \quad \forall \alpha \preceq \beta.$$

Using standard results from Tsybakov (2009), this implies that the resulting KDE $\hat{f}_s$ satisfies (5). Since $\psi = \mathcal{F}^{-1}[k_s] \in \mathcal{C}^\infty$, the Riemann–Lebesgue lemma guarantees that $|\kappa_s(x)| \leq c_{\beta,d}\,|x|^{-\nu}$ is satisfied for $\nu = \lceil \beta + d \rceil$. Since $\psi$ is compactly supported, an application of Parseval's identity yields $\kappa_s \in \mathcal{S}(\gamma, c_\gamma)$. Applying Proposition 1 to $k_s$, we conclude that for the task of density estimation, weighted KDEs built on coresets are nearly as powerful as the coreset-based estimators studied in Section 2.

**Theorem 2.** *Let $\varepsilon > 0$. The Carathéodory coreset estimator $\hat{g}_S(y)$ built using the kernel $k_s$ and setting $T = c_{d,\beta,\varepsilon}\, n^{\frac{\varepsilon}{d} + \frac{1}{2\beta+d}}$ satisfies*

$$\sup_{f\in\mathcal{P}_{\mathcal{H}}(\beta,L)} \mathbb{E}_f\|\hat{g}_S - f\|_2 \leq c_{\beta,d,L}\, n^{-\frac{\beta}{2\beta+d}}.$$

*The corresponding coreset has cardinality*

$$m = c_{d,\beta,\varepsilon} n^{\frac{d}{2\beta+d}+\varepsilon}.$$

Theorem 2 shows that the Carathéodory coreset estimator achieves the minimax rate of estimation with near-optimal coreset size. In fact, a small modification yields a near-optimal rate of convergence for any coreset size as in Theorem 1.

**Corollary 1.** *Let $\varepsilon > 0$ and $m \leq c_{\beta,d,\varepsilon}\, n^{\frac{d}{2\beta+d}+\varepsilon}$. The Carathéodory coreset estimator $\hat{g}_S(y)$ built using the kernel $k_s$, setting $h = m^{-\frac{1}{d}+\frac{\varepsilon}{\beta}}$ and $T = c_d\, m^{1/d}$, satisfies*

$$\sup_{f\in\mathcal{P}_{\mathcal{H}}(\beta,L)} \mathbb{E}\|\hat{g}_S - f\|_2 \leq c_{\beta,d,\varepsilon,L}\, \left(m^{-\frac{\beta}{d}+\varepsilon} + n^{-\frac{\beta}{2\beta+d}+\varepsilon}\right),$$

*and the corresponding coreset has cardinality $m$.*

Next we apply Proposition 1 to the popular Gaussian kernel $\phi(x) = (2\pi)^{-d/2}\exp(-\frac{1}{2}|x|_2^2)$. This kernel has rapid decay in the real domain and Fourier space, and is thus amenable to our techniques. Moreover, $\phi$ is a kernel of order $\ell = 1$, (Tsybakov, 2009, Definition 1.3 and Theorem 1.2) and so the standard KDE $\hat{f}_\phi$ on the full dataset attains the minimax rate of estimation $c_{d,L} n^{1/(2+d)}$ over the Lipschitz densities $\mathcal{P}_{\mathcal{H}}(1, L)$.

**Theorem 3.** *Let $\varepsilon > 0$. The Carathéodory coreset estimator $\hat{g}_\phi(y)$ built using the kernel $\phi$ and setting $T = c_{d,\varepsilon}\, n^{\frac{1}{2+d}+\frac{\varepsilon}{d}}$ satisfies*

$$\sup_{f\in\mathcal{P}_{\mathcal{H}}(1,L)} \mathbb{E}\|\hat{g}_\phi - f\|_2 \leq c_{d,L}\, n^{-\frac{1}{2+d}}.$$

*The corresponding coreset has cardinality*

$$m = c_{d,\varepsilon} n^{\frac{d}{2+d}+\varepsilon}.$$

In addition, we have a nearly matching lower bound to Theorem 2 for coreset KDEs. In fact, our lower bound applies to a generalization of coreset KDEs where the vector of weights $\{\lambda_j\}_{j\in S}$ is not constrained to be in the simplex but can range within a hypercube of width that may grow polynomially with $n$.

**Theorem 4.** *Let $A, B \geq 1$. Let $k$ denote a kernel with $\|k\|_2 \leq n$. Let $\hat{g}_S$ denote a weighted coreset KDE with bandwidth $h \geq n^{-A}$ built from $k$ with weights $\{\lambda_j\}_{j\in S}$ satisfying $\max_{j\in S}|\lambda_j| \leq n^B$. Then*

$$\sup_{f\in\mathcal{P}_{\mathcal{H}}(\beta,L)} \mathbb{E}_f\|\hat{g}_S - f\|_2 \geq c_{\beta,d,L}\left[(A+B)^{-\frac{\beta}{d}}(m\log n)^{-\frac{\beta}{d}} + n^{-\frac{\beta}{2\beta+d}}\right].$$

This result is essentially a consequence of the lower bound in Theorem 1 because, in an appropriate sense, coreset KDEs with bounded weights are well-approximated by coreset-based estimators. Hence, in the case of bounded weights, allowing these weights to be measurable functions of the entire dataset rather than just the coreset, as would be required in Section 2, does not make a significant difference for the purpose of estimation. The full details of Theorem 4 are postponed to the Supplement.

### 3.3 Proof sketch of Proposition 1

Here we sketch the proof of Proposition 1, our main tool in constructing effective coreset KDEs. Full details of the argument may be found in the Supplement.

Let $k(x) = \prod_{i=1}^d \kappa(x_i)$ denote a kernel, and suppose that $\hat{f}(y) = \frac{1}{n}\sum_{i=1}^n k_h(X_i - y)$ is a good estimator for an unknown density $f$ in that

$$\|f - \hat{f}\|_2 \leq \varepsilon := c_{\beta,d}\, n^{-\frac{\beta}{2\beta+d}}$$

on setting $h = n^{-1/(2\beta+d)}$. Our goal is to find a subset $S \subset [n]$ and weights $\{\lambda_j\}_{j\in S}$ such that

$$\frac{1}{n}\sum_{i=1}^n k_h(X_i - y) \approx \sum_{j\in S}\lambda_j k_h(X_j - y).$$

Suppose for simplicity that $\kappa$ is compactly supported on $[-1/2, 1/2]$. By hypothesis and Parseval's theorem $\kappa \in \mathcal{S}(\gamma, L')$, and we can further show that $k \in \mathcal{S}(\gamma, c_{d,L'})$ and $k_h \in \mathcal{S}(\gamma, c_{d,L'}h^{-d/2-\gamma})$. Let $\bar{\mathcal{F}}[f] = 4^{-2d}\mathcal{F}[f]$ denote the rescaled Fourier transform. Using the Fourier expansion on the interval $[-2,2]^d$ and fast Fourier decay of $k_h$, we have

$$\|k_h(x) - \sum_{|\omega|_\infty < T}\bar{\mathcal{F}}[k_h](\omega)e^{i\langle x,\omega\rangle}\|_2 \leq \varepsilon \qquad (6)$$

when $T = (\frac{c_{d,\gamma,L'}h^{-\frac{d}{2}-\gamma}}{\varepsilon})^{1/\gamma} = c_{d,\gamma,L'}\, n^{\frac{d/2+\beta+\gamma}{\gamma(2\beta+d)}}$. Observe that this matches the setting of $T$ in Proposition 1.

The approximation (6) implies that for $X_i \in [-1/2,1/2]^d$,

$$\hat{f}(y) \approx \sum_{|\omega|_\infty<T}\bar{\mathcal{F}}[k_h](\omega)\left(\frac{1}{n}\sum_{i=1}^n e^{i\langle X_i,\omega\rangle}\right)e^{-i\langle y,\omega\rangle}.$$

Using the Carathéodory coreset and weights $\{\lambda_j\}_{j\in S}$ constructed in Section 3.1, it follows that

$$\sum_{|\omega|_\infty<T}\bar{\mathcal{F}}[k_h](\omega)\left(\frac{1}{n}\sum_{i=1}^n e^{i\langle X_i,\omega\rangle}\right)e^{-i\langle y,\omega\rangle} = \sum_{|\omega|_\infty<T}\bar{\mathcal{F}}[k_h](\omega)\left(\sum_{i=1}^n \lambda_j e^{i\langle X_i,\omega\rangle}\right)e^{-i\langle y,\omega\rangle}.$$

Applying (6) again, we see that the right-hand-side is approximately equal to $\hat{g}_S(y)$, the estimator produced in Section (3.1). By the triangle inequality, we conclude that $\|\hat{g}_S(y) - f\|_2 \leq c_{\beta,d}\,\varepsilon$, as desired.

## 4 Lower bounds for coreset KDEs with uniform weights

In this section we study the performance of univariate uniformly weighted coreset KDEs

$$\hat{f}_S^{\mathsf{unif}}(y) = \frac{1}{m}\sum_{i\in S}k_h(X_i - y),$$

where $X_S$ is the coreset and $|S| = m$. The next results demonstrate that for a large class of kernels, there is significant gap between the rate of estimation achieved by $\hat{f}_S^{\mathsf{unif}}(y)$ and that of coreset KDEs with general weights. First we focus on the particular case of estimating the class $\mathcal{P}_{\mathcal{H}}(1,L)$ of univariate Lipschitz densities. For this class, the minimax rate of estimation (over all estimators) is $n^{-1/3}$, and this can be achieved by a weighted coreset KDE of cardinality $c_\varepsilon n^{1/3+\varepsilon}$ by Theorem 2, for all $\varepsilon > 0$.

**Theorem 5.** *Let $k$ denote a nonnegative kernel satisfying*

$$k(t) = O(|t|^{-(k+1)}), \quad \textit{and} \quad \mathcal{F}[k](\omega) = O(|\omega|^{-\ell})$$

*for some $\ell > 0$, $k > 1$. Suppose that $0 < \alpha < 1/3$. If*

$$m \leq \frac{n^{\frac{2}{3}-2\left(\alpha(1-\frac{2}{\ell})+\frac{2}{3\ell}\right)}}{\log n},$$

*then*

$$\inf_{h,S:|S|\leq m}\sup_{f\in\mathcal{P}_{\mathcal{H}}(1,L)}\mathbb{E}\|\hat{f}_S^{\mathsf{unif}} - f\|_2 = \Omega_k\Big(\frac{n^{-\frac{1}{3}+\alpha}}{\log n}\Big). \quad (7)$$

*The infimum above is over all possible choices of bandwidth $h$ and all coreset schemes $S$ of cardinality at most $m$.*

By this result, if $k$ has lighter than quadratic tails and fast Fourier decay, the error in (7) is a polynomial factor larger than the minimax rate $n^{-1/3}$ when

$m \ll n^{2/3}$. Hence, our result covers a wide variety of kernels typically used for density estimation and shows that the uniformly weighted coreset KDE performs much worse than the encoding estimator or the Carathéodory method. In addition, for very smooth univariate kernels with rapid decay, we have the following lower bound that applies for all $\beta > 0$.

**Theorem 6.** *Fix $\beta > 0$ and a nonnegative kernel $k$ on $\mathbb{R}$ satisfying the following fast decay and smoothness conditions:*

$$\lim_{s\to+\infty} \frac{1}{s} \log \frac{1}{\int_{|t|>s} k(t)dt} > 0, \tag{8}$$

$$\lim_{\omega\to\infty} \frac{1}{|\omega|} \log \frac{1}{|\mathcal{F}[k](\omega)|} > 0, \tag{9}$$

*where we recall that $\mathcal{F}[k]$ denotes the Fourier transform. Let $\hat{f}_S^{\mathsf{unif}}$ be the uniformly weighted coreset KDE. Then there exists $L_\beta > 0$ such that for $L \geq L_\beta$ and any $m$ and $h > 0$, we have*

$$\inf_{h,S:|S|\leq m} \sup_{f\in\mathcal{P}_{\mathcal{H}}(\beta,L)} \mathbb{E}\|\hat{f}_S^{\mathsf{unif}} - f\|_2 = \Omega_{\beta,k}\left(\frac{m^{-\frac{\beta}{1+\beta}}}{\log^{\beta+\frac{1}{2}} m}\right).$$

Therefore attaining the minimax rate with $\hat{f}_S^{\mathsf{unif}}$ requires $m \geq n^{\frac{\beta+1}{2\beta+1}}$ for such kernels. Next, note that the Gaussian kernel satisfies the hypotheses of Theorem 5 and 6. As we show in Theorem 7, results of Phillips & Tai (2018b) imply that our lower bounds are tight up to logarithmic factors: there exists a uniformly weighted Gaussian coreset KDE of size $m = \tilde{O}(n^{2/3})$ that attains the minimax rate $n^{-1/3}$ for estimating univariate Lipschitz densities ($\beta = 1$). In general, we expect a lower bound $m = \Omega(n^{\frac{\beta+d}{2\beta+d}})$ to hold for uniformly weighted coreset KDEs attaining the minimax rate. The proofs of Theorems 5 and 6 can be found in the Supplement.

## 5 Comparison to other methods

Three methods for constructing coreset kernel density estimators that have previously been explored include random sampling (Joshi *et al.*, 2011; Lopez-Paz *et al.*, 2015), the Frank–Wolfe algorithm (Bach *et al.*, 2012; Harvey & Samadi, 2014; Phillips & Tai, 2018a), and discrepancy-based approaches (Phillips & Tai, 2018b; Karnin & Liberty, 2019). These procedures all result in a uniformly weighted coreset KDE. To compare these results with ours on the problem of density estimation, for each method under consideration we raise the question: How large does $m$, the size of the coreset, need to be to guarantee that

$$\sup_{f\in\mathcal{P}_{\mathcal{H}}(\beta,L)} \mathbb{E}_f\|\hat{g}_S - f\|_2 = O_{\beta,d,L}\left(n^{-\frac{\beta}{2\beta+d}}\right)? \tag{10}$$

Here $\hat{g}_S$ is the resulting coreset KDE and the right-hand-side is the minimax rate over all estimators on the full dataset $X_1, \ldots, X_n$.

Uniform random sampling of a subset of cardinality $m$ yields an i.i.d dataset, so the rate obtained is at least $m^{-\beta/(2\beta+d)}$. Hence, we must take $m = \Omega(n)$ to achieve the minimax rate.

The Frank–Wolfe algorithm is a greedy method that iteratively constructs a sparse approximation to a given element in a convex set (Frank *et al.*, 1956; Bubeck, 2015). Thus Frank–Wolfe may be applied directly in the RKHS corresponding to a positive-semidefinite kernel as shown in Phillips & Tai (2018b) to approximate the KDE on the full dataset. However, due to the shrinking bandwidth in our problem, this approach also requires $m = \Omega(n)$ to guarantee the bound in (10). Another strategy is to approximately solve the linear equation (4) using the Frank–Wolfe algorithm. Unfortunately, a direct implementation again uses $m = \Omega(n)$ data points.

A more effective strategy utilizes discrepancy theory (Phillips, 2013; Phillips & Tai, 2018b; Karnin & Liberty, 2019) (see Matoušek, 1999; Chazelle, 2000, for a comprehensive exposition of discrepancy theory). By the well-known halving algorithm (see e.g. Chazelle & Matoušek, 1996; Phillips & Tai, 2018b) if for all $N \leq n$, the *kernel discrepancy*

$$\mathsf{disc}_k = \sup_{x_1,\ldots,x_N} \min_{\substack{\sigma\in\{-1,+1\}^n \\ \mathbb{1}^T\sigma=0}} \|\sum_{i=1}^N \sigma_i k(x_i - y)\|_\infty$$

is at most $D$, then there exists a coreset $X_S$ of size $\tilde{O}_D(\varepsilon^{-1})$ such that

$$\|\frac{1}{n}\sum_{i=1}^n k(X_i - y) - \frac{1}{m}\sum_{j\in S} k(X_i - y)\|_\infty \leq \varepsilon. \tag{11}$$

The idea of the halving algorithm is to maintain a set of datapoints $\mathcal{C}_\ell$ at each iteration and then set $\mathcal{C}_{\ell+1}$ to be the set of vectors that receive sign $+1$ upon minimizing $\|\sum_{x\in\mathcal{C}_\ell} \sigma_x k(x - y)\|_\infty$. Starting with the original dataset and repeating this procedure $O(\log \frac{n}{m})$ times yields the desired coreset $X_S$ satisfying (11).

Phillips & Tai (2018b, Theorem 4) use a state-of-the-art algorithm from Bansal *et al.* (2018) called the *Gram–Schmidt walk* to give strong bounds on the kernel discrepancy of bounded and Lipschitz kernels $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ that are positive definite and decay rapidly away from the diagonal. With a careful handling of the Lipschitz constant and error in their argument when the bandwidth is set to be $h = n^{-1/(2\beta+d)}$, their techniques yield the following result applied to the kernel $k_s$. For completeness we give details of the argument in the Supplement.

**Theorem 7.** *Let $k_s$ denote the kernel from Section 3.2. The algorithm of Phillips & Tai (2018b) yields in polynomial time a subset $S$ with $|S| = m = \tilde{O}(n^{\frac{\beta+d}{2\beta+d}})$ such that the uniformly weighted coreset KDE $\hat{g}_S$ satisfies*

$$\sup_{f \in \mathcal{P}_{\mathcal{H}}(\beta, L)} \mathbb{E}\|f - \hat{g}_S\|_2 \leq c_{\beta,d,L}\, n^{-\frac{\beta}{2\beta+d}}.$$

This result also applies to more general kernels, for example, the Gaussian kernel when $\beta = 1$. We suspect that this is the best result achievable by discrepancy-based methods. In particular for nonnegative univariate kernels with fast decay in the real and Fourier domains, such as the Gaussian kernel, Theorem 5 implies that this rate is optimal for estimating Lipschitz densities with uniformly weighted coreset KDEs.

In contrast, the Carathéodory coreset KDE as in Theorem 2 only needs cardinality $m = O_\varepsilon(n^{\frac{d}{2\beta+d}+\varepsilon})$ to be a minimax estimator. By Theorem 4, this result is nearly optimal for coreset KDEs with bounded kernels and weights. And as with the other three methods described, our construction is computationally efficient. Hence allowing more general weights results in more powerful coreset KDEs for the problem of density estimation.

## Acknowledgments

We thank Cole Franks for helpful discussions regarding algorithmic aspects of Carathéodory's theorem. We thank the anonymous reviewers for their many helpful comments and suggestions. Philippe Rigollet was supported by NSF awards IIS-1838071, DMS-1712596, DMS-1740751, and DMS-2022448.

## References

Agarwal, Pankaj K, Har-Peled, Sariel, & Varadarajan, Kasturi R. 2005. Geometric approximation via coresets. *Combinatorial and computational geometry*, **52**, 1–30.

Ashtiani, Hassan, Ben-David, Shai, Harvey, Nicholas JA, Liaw, Christopher, Mehrabian, Abbas, & Plan, Yaniv. 2020. Near-optimal sample complexity bounds for robust learning of gaussian mixtures via compression schemes. *Journal of the ACM (JACM)*, **67**(6), 1–42.

Bach, Francis R, Lacoste-Julien, Simon, & Obozinski, Guillaume. 2012. On the Equivalence between Herding and Conditional Gradient Algorithms. *In: ICML*.

Bachem, Olivier, Lucic, Mario, & Krause, Andreas. 2017. Practical coreset constructions for machine learning. *arXiv preprint arXiv:1703.06476*.

Bachem, Olivier, Lucic, Mario, & Krause, Andreas. 2018. Scalable k-means clustering via lightweight coresets. *Pages 1119–1127 of: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Bansal, Nikhil, Dadush, Daniel, Garg, Shashwat, & Lovett, Shachar. 2018. The gram-schmidt walk: a cure for the Banaszczyk blues. *Pages 587–597 of: Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*.

Bubeck, Sébastien. 2015. *Convex optimization: algorithms and complexity*. Now Publishers Inc.

Carathéodory, C. 1907 (Mar.). *Über den Variabilitätsbereich der Koeffizienten von Potenzreihen, die gegebene Werte nicht annehmen*.

Chazelle, & Matoušek. 1996. On linear-time deterministic algorithms for optimization problems in fixed dimension. *Journal of Algorithms*, **21**(3), 579–597.

Chazelle, B. 2000. *The Discrepancy Method: Randomness and Complexity*. Cambridge: Cambridge University Press.

Claici, Sebastian, Genevay, Aude, & Solomon, Justin. 2020. Wasserstein Measure Coresets. *arXiv preprint arXiv:1805.07412*.

Clarkson, Kenneth L. 2010. Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm. *ACM Transactions on Algorithms (TALG)*, **6**(4), 1–30.

Feldman, Dan, Faulkner, Matthew, & Krause, Andreas. 2011. Scalable training of mixture models via coresets. *Pages 2142–2150 of: Advances in neural information processing systems*.

Feldman, Dan, Schmidt, Melanie, & Sohler, Christian. 2013. Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering. *Pages 1434–1453 of: Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*. SIAM.

Frahling, Gereon, & Sohler, Christian. 2005. Coresets in Dynamic Geometric Data Streams. *Pages 209–217 of: Proceedings of the Thirty-Seventh Annual ACM Symposium on Theory of Computing*. STOC '05. New York, NY, USA: Association for Computing Machinery.

Frank, Marguerite, Wolfe, Philip, *et al.* 1956. An algorithm for quadratic programming. *Naval research logistics quarterly*, **3**(1-2), 95–110.

Gärtner, Bernd, & Jaggi, Martin. 2009. Coresets for polytope distance. *Pages 33–42 of: Proceedings of the twenty-fifth annual symposium on Computational geometry*.

Grötschel, Martin, Lovász, László, & Schrijver, Alexander. 2012. *Geometric algorithms and combinatorial optimization*. Vol. 2. Springer Science & Business Media.

Hanneke, Steve, Kontorovich, Aryeh, & Sadigurschi, Menachem. 2019. Sample compression for real-valued learners. *Pages 466–488 of: Algorithmic Learning Theory*.

Har-Peled, Sariel, & Kushal, Akash. 2007. Smaller coresets for k-median and k-means clustering. *Discrete & Computational Geometry*, **37**(1), 3–19.

Harvey, Nick, & Samadi, Samira. 2014. Near-optimal herding. *Pages 1165–1182 of: Conference on Learning Theory*.

Hiriart-Urruty, Jean-Baptiste, & Lemaréchal, Claude. 2004. *Fundamentals of convex analysis*. Springer Science & Business Media.

Huggins, Jonathan, Campbell, Trevor, & Broderick, Tamara. 2016. Coresets for scalable Bayesian logistic regression. *Pages 4080–4088 of: Advances in Neural Information Processing Systems*.

Joshi, Sarang, Kommaraji, Raj Varma, Phillips, Jeff M., & Venkatasubramanian, Suresh. 2011. Comparing Distributions and Shapes Using the Kernel Distance. *Page 47–56 of: Proceedings of the Twenty-Seventh Annual Symposium on Computational Geometry*. SoCG '11. New York, NY, USA: Association for Computing Machinery.

Karnin, Zohar, & Liberty, Edo. 2019. Discrepancy, Coresets, and Sketches in Machine Learning. *Pages 1975–1993 of:* Beygelzimer, Alina, & Hsu, Daniel (eds), *Proceedings of the Thirty-Second Conference on Learning Theory*. Proceedings of Machine Learning Research, vol. 99. Phoenix, USA: PMLR.

Littlestone, Nick, & Warmuth, Manfred. 1986. Relating data compression and learnability. *Unpublished manuscript*.

Lopez-Paz, David, Muandet, Krikamol, Schölkopf, Bernhard, & Tolstikhin, Iliya. 2015. Towards a Learning Theory of Cause-Effect Inference. *Pages 1452–1461 of:* Bach, Francis, & Blei, David (eds), *Proceedings of the 32nd International Conference on Machine Learning*. Proceedings of Machine Learning Research, vol. 37. Lille, France: PMLR.

Matoušek, J. 1999. *Geometric Discrepancy: an Illustrated Guide*. New York: Springer.

McDonald, Daniel. 2017. Minimax Density Estimation for Growing Dimension. *Pages 194–203 of:* Singh, Aarti, & Zhu, Jerry (eds), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. Proceedings of Machine Learning Research, vol. 54. Fort Lauderdale, FL, USA: PMLR.

Munteanu, Alexander, Schwiegelshohn, Chris, Sohler, Christian, & Woodruff, David P. 2018. On Coresets for Logistic Regression. *Pages 6562–6571 of: Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS'18. Red Hook, NY, USA: Curran Associates Inc.

Phillips, Jeff M. 2013. $\varepsilon$-samples for kernels. *Pages 1622–1632 of: Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*. SIAM.

Phillips, Jeff M, & Tai, Wai Ming. 2018a. Improved coresets for kernel density estimates. *Pages 2718–2727 of: Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM.

Phillips, Jeff M., & Tai, Wai Ming. 2018b. Near-Optimal Coresets of Kernel Density Estimates. *Pages 66:1–66:13 of: 34th International Symposium on Computational Geometry, SoCG 2018, June 11-14, 2018, Budapest, Hungary*.

Tikhomirov, V. M. 1993. *$\varepsilon$-Entropy and $\varepsilon$-Capacity of Sets In Functional Spaces*. Dordrecht: Springer Netherlands. Pages 86–170.

Tsybakov, Alexandre B. 2009. *Introduction to Nonparametric Estimation*. Springer series in statistics. Springer.

Zheng, Yan, & Phillips, Jeff M. 2017. Coresets for kernel regression. *Pages 645–654 of: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.