A Statistical Perspective on Coreset Density Estimation

Paxton Turner, Jingbo Liu, and Philippe Rigollet

Jerry Wei

Department of Statistics
University of Washington

Coreset Density Estimation

Given a dataset $\mathcal{D} = \{X_1, \ldots, X_n\} \subset \mathbb{R}^d$ 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 $\{1, \ldots, n\}$ 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 $\mathcal{D}$.

Given i.i.d random variables $X_1, \ldots, X_n \sim \mathbb{P}_f$ 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.

Coreset Density Estimation

The minimax rate of estimation over the L-Hölder smooth densities $\mathcal{P}_{\mathcal{H}}(\beta, L)$ of order β is given by

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

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

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

for suitable choices of kernel $K : \mathbb{R}^d \to \mathbb{R}$ and bandwidth $h > 0$
Main goal: to extend this understanding of rates to two families of estimators based on coresets

Setup and Notation

- A *coreset* X_S is defined to be the projection of the dataset $X = (X_1, \ldots, X_n)$ onto the subset indicated by $S(X)$: $X_S := \{X_i\}_{i \in S(X)}$. (Choice of coreset is data dependent.)
- Define the *cardinality* of S to be $|S| := m$.
- $\hat{f}$ the density estimator based on n observations $X_1, \ldots, X_n \in \mathbb{R}^d$.
- Denote the *coreset-based estimator* as $\hat{f}_S$.
- Let $\mathcal{H}(\beta, L)$ denote the space of Hölder functions. Let $\mathcal{P}_{\mathcal{H}}(\beta, L)$ denote the set of probability density functions contained in $\mathcal{H}(\beta, L)$.
- Define the Sobolev functions $\mathcal{S}(\gamma, L')$ that consist of all $f : \mathbb{R}^d \to \mathbb{R}$ that satisfy

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

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

Framework 1: Coreset-based Estimator

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.

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

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

where the infimum above is over all choices of coreset scheme S of cardinality m and all estimators $\hat{f} : \mathbb{R}^{d \times m} \to L_2(\mathbb{R}^d)$.

Coreset-based Estimator

Theorem

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

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

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.*

Framework 2: Coreset Kernel Density Estimator

weighted coreset kernel density estimators selects a kernel k, bandwidth parameter h, and a coreset X_S of cardinality m 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.

Carathéodory coreset method

For cutoff frequency $T > 0$, define $A = \{\omega \in \frac{\pi}{2}\mathbb{Z}^d : |\omega|_\infty \leq T\}$. Consider the complex vectors $(e^{i\langle X_j, \omega\rangle})_{\omega \in A}$. By Carathéodory's theorem, 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

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

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

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

Algorithmic consideration: 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

Results on Carathéodory coresets

Proposition

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

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

Find Suitable Kernel

There exists a kernel $k_s \in \mathcal{C}^\infty$ that satisfies the conditions above for all β and γ. Let $\psi : [-1, 1] \to [0, 1]$ denote a cutoff function that has the following properties: $\psi \in \mathcal{C}^\infty$, $\psi\big|_{[-1,1]} \equiv 1$, and ψ is compactly 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

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

which implies that k_s is a kernel of order β. 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 ψ is compactly supported, an application of Parseval's identity yields $\kappa_s \in \mathcal{S}(\gamma, c_\gamma)$.

Results on Carathéodory coresets

Applying Proposition 1 to k_s, for the task of density estimation, weighted KDEs built on coresets are nearly as powerful as the coreset-based estimators.

Theorem

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

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

The Carathéodory coreset estimator achieves the minimax rate of estimation with near-optimal coreset size.

Upper bound on Carathéodory coresets

In other words a near-optimal rate of convergence for any coreset size as in Theorem 1.

Corollary

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.

Carathéodory coresets with Gaussian kernel

Gaussian kernel $\phi(x) = (2\pi)^{-d/2} \exp(-\frac{1}{2}|x|_2^2)$. This kernel k_ϕ is a kernel of order $\ell = 1$ and on the full data attains the minimax rate of estimation $c_{d,L} n^{1/(2+d)}$ over the Lipschitz densities $\mathcal{P}_{\mathcal{H}}(1, L)$.

Theorem

Let $\varepsilon > 0$. The Carathéodory coreset estimator $\hat{g}_\phi(y)$ built using the kernel ϕ 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

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

Lower bound on Carathéodory coresets

Nearly matching lower bound to Theorem 2 for coreset KDEs.

Theorem

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

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

Comparison with Uniform Weights

significant gap between the rate of estimation achieved by uniform weighted $\hat{f}_S^{\text{unif}}(y)$ and that of coreset KDEs with general weights.
First case of estimating Lipschitz densities, the class $\mathcal{P}_{\mathcal{H}}(1, L)$.

Theorem

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

$$\inf_{h, S:|S| \leq m} \sup_{f \in \mathcal{P}_{\mathcal{H}}(1,L)} \mathbb{E} \|\hat{f}_S^{\text{unif}} - f\|_2 = \Omega_k\left(\frac{n^{-\frac{1}{3}+\alpha}}{\log n}\right). \tag{6}$$

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

Comparison with Uniform Weights

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 $\varepsilon > 0$.
By this result, if k has lighter than quadratic tails and fast Fourier decay, the error in (6) 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.

Comparison with Uniform Weights

In addition, for very smooth univariate kernels with rapid decay, we have the following lower bound that applies for all $\beta > 0$.

Theorem

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

$$\lim_{s \to +\infty} \frac{1}{s} \log \frac{1}{\int_{|t|>s} k(t)dt} > 0, \tag{7}$$

$$\lim_{\omega \to \infty} \frac{1}{|\omega|} \log \frac{1}{|\mathcal{F}[k](\omega)|} > 0, \tag{8}$$

where we recall that $\mathcal{F}[k]$ denotes the Fourier transform. Let $\hat{f}_S^{\text{unif}}$ 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

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

Comparison with Uniform Weights

Attaining the minimax rate with $\hat{f}_S^{\text{unif}}$ requires $m \geq n^{\frac{\beta+1}{2\beta+1}}$ for such kernels. The 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, expect a lower bound $m = \Omega(n^{\frac{\beta+d}{2\beta+d}})$ to hold for uniformly weighted coreset KDEs attaining the minimax rate.

Comparison to Other Methods

How large does m, the size of the coreset, need to be to guarantee that

$$\sup_{f \in \mathcal{P}_{\mathcal{H}}(\beta, L)} \mathbb{E}_f \|\hat{g}_S - f\|_2 = O_{\beta,d,L}\left(n^{-\frac{\beta}{2\beta+d}}\right)? \tag{9}$$

- Uniform random sampling of a subset of cardinality m yields an i.i.d dataset, so the rate obtained is at least $m^{-\beta/(2\beta+d)}$ and we must take $m = \Omega(n)$ to achieve the minimax rate.
- Frank–Wolfe cab be applied directly in the RKHS corresponding to a positive-semidefinite kernel to approximate the KDE on the full dataset, requires $m = \Omega(n)$. Approximately solve the linear equation (4) using the Frank–Wolfe algorithm with direct implementation again uses $m = \Omega(n)$.
- Utilizes discrepancy theory. Existing result has a algorithm 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

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

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 with nearly optimal for coreset KDEs.

Thanks