A Statistical Perspective on Coreset Density Estimation

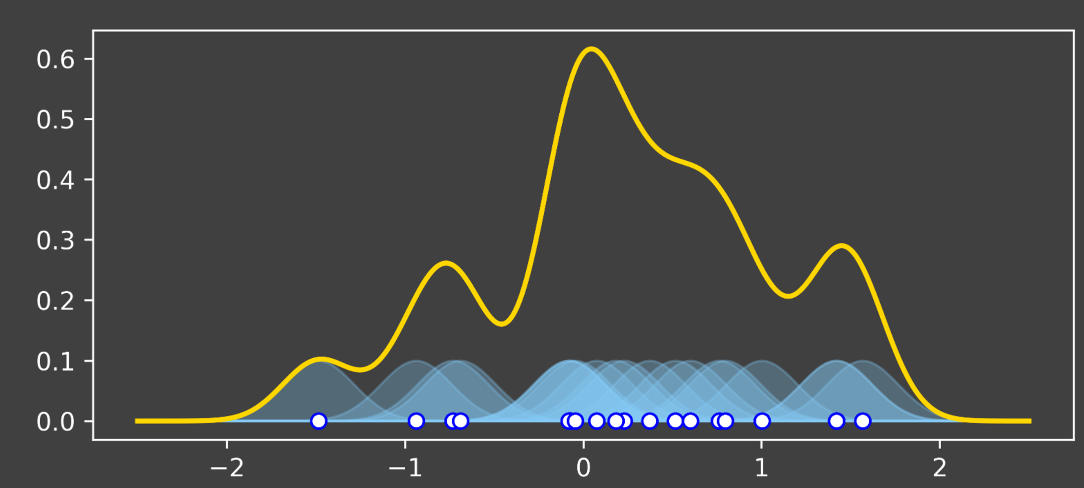
Paxton Turner (MIT), Jingbo Liu (UIUC), Philippe Rigollet (MIT)

Density Estimation

Goal: reconstruct a probability density function from data $X_1, \dots, X_n$

Kernel density estimator (KDE):

$$\hat{f}(y) = \frac{1}{n}\sum_{j=1}^{n} K_h\left(X_j - y\right) = \mathbb{E}_{X\sim \underset{\text{empirical measure}}{\mathbb{P}_n}}[K_h(X-y)]$$



Theorem KDE achieves the minimax rate of estimation $n^{-\frac{\beta}{2\beta+d}}$ over Hölder β densities.

Proof (Fourier analysis of KDE, $d = 1$) Let F denote the Fourier transform. Assume $F[K](\omega) = 1$ near the origin. Controlling the bias:

$$|f(y_0) - \mathbb{E}\,\hat{f}(y_0)|$$

$$= \left|\sum_\omega F[f](\omega)e^{-i\omega y_0} - \mathbb{E}\sum_\omega \underbrace{F[\hat{f}](\omega)}_{F[K_h](\omega)\cdot\frac{1}{n}\sum_{j=1}^{n} e^{i\omega X_j}}e^{-i\omega y_0}\right|$$

$$= \left|\sum_\omega (1 - F[K_h](\omega))F[f](\omega)\,e^{-i\omega y_0}\right|$$

$$\le \sum_{|\omega|>\frac{1}{h}} \left|(1 - F[K](\omega h))\,F[f](\omega)\right| \lesssim h^\beta$$

Bias – variance tradeoff: if $h \asymp n^{-\frac{1}{2\beta+1}}$, then

$$\text{Error}^2 = \text{Bias}^2 + \text{Variance} = h^{2\beta} + \frac{1}{nh} \asymp n^{-\frac{2\beta}{2\beta+1}}$$

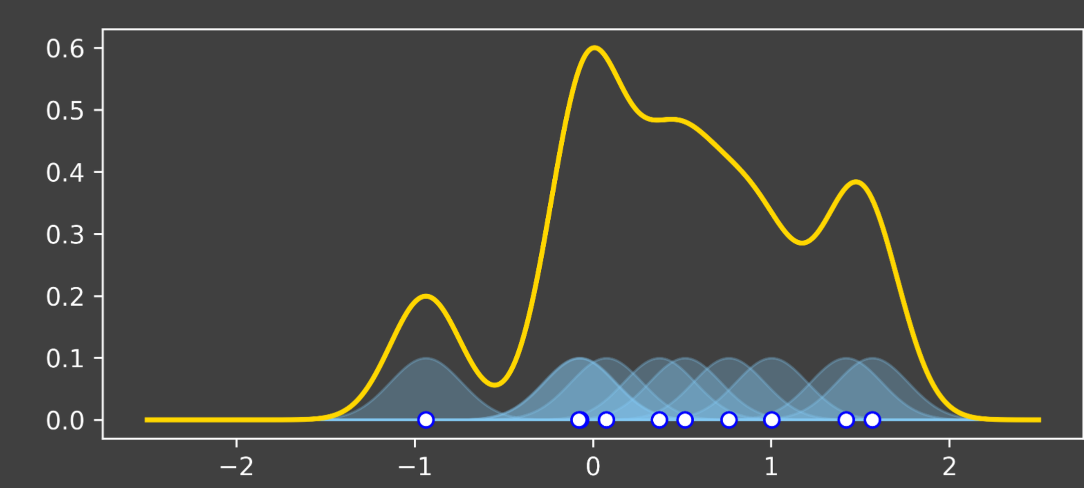
To improve on computational aspects of the KDE, we use coresets.

Coreset Density Estimation

Coreset: a weighted subset that summarizes the original dataset

Coreset KDE:

$$\hat{f}_C(y) = \sum_{X_j \in C} \lambda_j\, K_h\left(X_j - y\right) = \mathbb{E}_{X\sim \underset{\text{coreset measure}}{\mathbb{P}_C}}[K_h(X-y)]$$



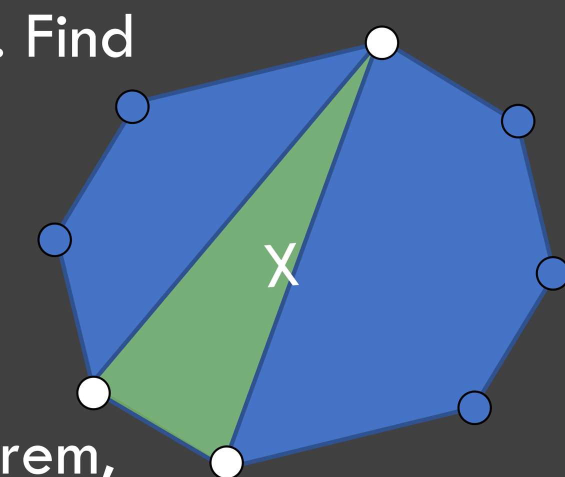
Goal: Establish rate of estimation of coreset KDEs

Carathéodory Coresets

Our construction: ($d = 1$) Let $T \in \mathbb{Z}_{\ge 0}$. Find $\{\lambda_j\}$ and C such that

$$\frac{1}{n}\sum_{j=1}^{n} e^{i\,\omega\,X_j} = \sum_{X_j \in C} \lambda_j e^{i\,\omega\,X_j}$$

for all $|\omega| \le T$. By Carathéodory's theorem, can choose $|C| = O(T)$.



Our analysis: Suppose $F[K](\omega) \lesssim |\omega|^{-\gamma}$. Setting $T \asymp h^{-1-\frac{\beta}{\gamma}}$, we have

$$|\hat{f}(y_0) - \hat{f}_C(y_0)|$$

$$= \left|\sum_\omega F[\hat{f} - \hat{f}_C](\omega)\,e^{-i\omega y_0}\right|$$

$$\lesssim \left|\sum_\omega \left(\frac{1}{n}\sum_{j=1}^{n} e^{i\omega X_j} - \sum_{X_j\in C}\lambda_j e^{i\omega X_j}\right) F[K_h](\omega)\right|$$

$$\lesssim \sum_{|\omega|>T} |F[K](\omega h)| \lesssim |Th|^{-\gamma} \lesssim n^{-\frac{\beta}{2\beta+1}}.$$

Therefore, $|f(y_0) - \hat{f}_C(y_0)| \lesssim n^{-\frac{\beta}{2\beta+1}}$.

Results

Theorem For appropriate kernel, the Carathéodory KDE achieves the minimax rate $n^{-\frac{\beta}{2\beta+d}}$ with a coreset of size

$$|C| = n^{\frac{d}{2\beta+d}+\varepsilon} \qquad \forall\, \varepsilon > 0.$$

Remarks

- We show any coreset procedure requires at least $n^{\frac{d}{2\beta+d}}$ points to achieve minimax rate
- Flexibility in the weights allows Carathéodory KDE to improve upon existing approaches

Open Question

Let $\mathbb{P}_n$ = empirical measure and $\mathbb{P}_C$ = measure on coreset with probabilities $\{\lambda_j\}$. We showed

$$\int K_h(X-y)\mathrm{d}\mathbb{P}_n \approx \int K_h(X-y)\mathrm{d}\mathbb{P}_C\,.$$

Question (coresets for many tasks): Given a class $\mathcal{F}$, what size of C guarantees

$$\sup_{f\in\mathcal{F}} \int f(X)\mathrm{d}\mathbb{P}_n \approx \sup_{f\in\mathcal{F}} \int f(X)\mathrm{d}\mathbb{P}_C\ ?$$

References

Agarwal, Har-Peled, & Varadarajan 2005. "Geometric approximation via coresets". Combinatorial and computational geometry

Karnin & Liberty. 2019. Discrepancy,Coresets, and Sketches in Machine Learning. 32nd Conference on Learning Theory.

Phillips & Tai. 2018. "Near-Optimal Coresets of Kernel Density Estimates". 34th International Symposium on Computational Geometry

Tsybakov. 2009. *Introduction to Non-parametric Estimation*. Springer.