$L_\infty$ Error and Bandwidth Selection for Kernel Density Estimates of Large Data

Yan Zheng and Jeff M. Phillips

University of Utah
Outline

- Introduction and Background
- Problem Statement
- Methods
- Experiment Results
Kernel Density Estimates (KDE)

Point set $P$ of size $n$.

Kernel Density Estimate: $\text{KDE}_P(x) = \frac{\sum_{p \in P} K(p, x)}{|P|}$

Gaussian: $K(x, p) = \exp\left(-\frac{\|p-x\|^2}{2\sigma^2}\right)$. 
Kernel Density Estimates

Point set $P$ of size $n$.

Kernel Density Estimate: $\text{KDE}_P(x) = \frac{\sum_{p \in P} K(p, x)}{|P|}$

Gaussian: $K(x, p) = \exp\left(-\frac{\|p-x\|^2}{2\sigma^2}\right)$. 
Given two point sets \( P, Q \rightarrow \mathbb{R}^d \) and a kernel \( K \), estimate \( L_1(P, Q) \).

\[ \sigma = 5 \]
Given two point sets \( P, Q \subseteq \mathbb{R}^d \) and a kernel \( K \), estimate \( L_1(P, Q) \).

\[
\sigma = 5
\]

\[
\sigma = 2
\]

Bandwidth for KDE
Given two point sets $P, Q \in \mathbb{R}^d$ and a kernel $K$, estimate $L_1(P, Q)$.

Bandwidth for KDE

- $\sigma = 5$
- $\sigma = 2$
- $\sigma = 10$
Approximate Kernel Density Estimates

Approximate $\text{KDE}_P$ with $\text{KDE}_Q$ so that

$$L_{\infty}(P, Q) = \max_{x \in \mathbb{R}^d} |\text{KDE}_P(x) - \text{KDE}_Q(x)| \leq \varepsilon.$$ 

Coloring $\chi : P \to \{-1, +1\}$. Discrepancy:

$$\text{disc}_\chi(P, K) = \max_{x \in \mathbb{R}^d} \left| \sum_{p \in P} \chi(p) K(x, p) \right|$$
Approximate Kernel Density Estimates

Approximate $\text{KDE}_P$ with $\text{KDE}_Q$ so that
$L_\infty(P, Q) = \max_{x \in \mathbb{R}^d} |\text{KDE}_P(x) - \text{KDE}_Q(x)| \leq \varepsilon.$

Coloring $\chi : P \rightarrow \{-1, +1\}$. Discrepancy:
$\text{disc}_\chi(P, \mathcal{K}) = \max_{x \in \mathbb{R}^d} \left| \sum_{p \in P} \chi(p)K(x, p) \right|$
Approximate Kernel Density Estimates

Approximate $KDE_P$ with $KDE_Q$ so that
$L_\infty(P, Q) = \max_{x \in \mathbb{R}^d} |KDE_P(x) - KDE_Q(x)| \leq \varepsilon.$

Coloring $\chi : P \rightarrow \{-1, +1\}$. Discrepancy:
$\text{disc}_\chi(P, \mathcal{K}) = \max_{x \in \mathbb{R}^d} \left| \sum_{p \in P} \chi(p) K(x, p) \right|$
Approximate Kernel Density Estimates

Approximate \( \text{KDE}_P \) with \( \text{KDE}_Q \) so that

\[
L_\infty(P, Q) = \max_{x \in \mathbb{R}^d} |\text{KDE}_P(x) - \text{KDE}_Q(x)| \leq \varepsilon.
\]

Coloring \( \chi : P \to \{-1, +1\} \). Discrepancy:

\[
\text{disc}_\chi(P, \mathcal{K}) = \max_{x \in \mathbb{R}^d} \left| \sum_{p \in P} \chi(p) K(x, p) \right|
\]
Approximate Kernel Density Estimates

Approximate $\text{KDE}_P$ with $\text{KDE}_Q$ so that

$$L_\infty(P, Q) = \max_{x \in \mathbb{R}^d} |\text{KDE}_P(x) - \text{KDE}_Q(x)| \leq \varepsilon.$$ 

Coloring $\chi : P \to \{-1, +1\}$. Discrepancy:

$$\text{disc}_\chi(P, \mathcal{K}) = \max_{x \in \mathbb{R}^d} \left| \sum_{p \in P} \chi(p) K(x, p) \right|$$
Approximate Kernel Density Estimates

Approximate $\text{KDE}_P$ with $\text{KDE}_Q$ so that

$$L_\infty(P, Q) = \max_{x \in \mathbb{R}^d} |\text{KDE}_P(x) - \text{KDE}_Q(x)| \leq \varepsilon.$$ 

Coloring $\chi : P \rightarrow \{-1, +1\}$. Discrepancy:

$$\text{disc}_\chi(P, K) = \max_{x \in \mathbb{R}^d} \left| \sum_{p \in P} \chi(p) K(x, p) \right|$$
Problem Statement

$L_\infty$ error estimation
Given two point sets $P, Q \subset \mathbb{R}^d$ and a kernel $K$, estimate $L_\infty(P, Q)$.

$L_\infty(P, Q) = \max_{x \in \mathbb{R}^d} |\text{KDE}_P(x) - \text{KDE}_Q(x)| \leq \varepsilon.$
Problem Statement

$L_\infty$ error estimation
Given two point sets $P, Q \subset \mathbb{R}^d$ and a kernel $K$, estimate $L_\infty(P, Q)$.

$L_\infty(P, Q) = \max_{x \in \mathbb{R}^d} |\text{KDE}_P(x) - \text{KDE}_Q(x)| \leq \varepsilon.$

$L_\infty(P, \sigma, Q, \omega) = \max_{x \in \mathbb{R}^d} |\text{KDE}_{P,\sigma}(x) - \text{KDE}_{Q,\omega}(x)| \leq \varepsilon.$
Problem Statement

$L_\infty$ error estimation
Given two point sets $P, Q \subset \mathbb{R}^d$ and a kernel $K$, estimate $L_\infty(P, Q)$.

$L_\infty(P, Q) = \max_{x \in \mathbb{R}^d} |\text{KDE}_P(x) - \text{KDE}_Q(x)| \leq \varepsilon.$

$L_\infty(P, \sigma, Q, \omega) = \max_{x \in \mathbb{R}^d} |\text{KDE}_{P, \sigma}(x) - \text{KDE}_{Q, \omega}(x)| \leq \varepsilon.$

Bandwidth Estimation
Given two point sets $P, Q \subset \mathbb{R}^d$ a kernel $K$ and a bandwidth $\sigma$ estimate $\omega = \arg \min_{\omega} L_\infty(P, \sigma, Q, \omega)$. 
Problem Statement

$L_\infty$ error estimation
Given two point sets $P, Q \subset \mathbb{R}^d$ and a kernel $K$, estimate $L_\infty(P, Q)$.

\[ L_\infty(P, Q) = \max_{x \in \mathbb{R}^d} |\text{KDE}_P(x) - \text{KDE}_Q(x)| \leq \varepsilon. \]

\[ L_\infty(P, \sigma, Q, \omega) = \max_{x \in \mathbb{R}^d} |\text{KDE}_{P, \sigma}(x) - \text{KDE}_{Q, \omega}(x)| \leq \varepsilon. \]

Bandwidth Estimation
Given two point sets $P, Q \subset \mathbb{R}^d$ a kernel $K$ and a bandwidth $\sigma$ estimate $\omega = \arg \min_{\omega} L_\infty(P, \sigma, Q, \omega)$.

Traditional Setting
$\omega = \arg \min_{\omega} \|\mu - \text{KDE}_{Q, \omega}(x)\|_{1,2}$
where $\mu$ is unknown distribution and $Q$ is randomly from $\mu$. 
Why $\sigma$ is given?

Our setting

$Q$ may not randomly from $P$.
The choice of bandwidth may vary largely.

Example: One year temperature data
Why $\sigma$ is given?

Our setting

$Q$ may not randomly from $P$.
The choice of bandwidth may vary largely.

Example: One year temperature data
Why $\sigma$ is given?

Our setting

$Q$ may not randomly from $P$.
The choice of bandwidth may vary largely.

Example: One year temperature data
Why $L_\infty$ error?

**Stronger Bounds**

\[
L_p(P, Q) = \left( \frac{1}{|P|} \sum_{q \in P} |\text{KDE}_P(p) - \text{KDE}_Q(p)|^p \right)^{1/p}.
\]

If $|\text{KDE}_P(x) - \text{KDE}_Q(x)| \leq \varepsilon$ for all $x$, $L_p(P, Q)$ is at most $\varepsilon$.

$L_\infty(P, Q)$ can guarantee the bound.

**Preserve the worst case error**

Twitter data
Why $L_\infty$ error?

Stronger Bounds

\[ L_p(P, Q) = \left( \frac{1}{|P|} \sum_{q \in P} |\text{KDE}_P(p) - \text{KDE}_Q(p)|^p \right)^{1/p}. \]

If \(|\text{KDE}_P(x) - \text{KDE}_Q(x)| \leq \varepsilon \) for all \(x\),

\(L_p(P, Q)\) is at most \(\varepsilon\).

\(L_\infty(P, Q)\) can guarantee the bound.

Preserve the worst case error

Twitter data
Outline

- Introduction and Background
- Problem Statement
- Methods
- Experiment Results
Computing $L_\infty$

Definition

\[ G(x) = |\text{KDE}_P(x) - \text{KDE}_Q(x)|. \]

\[ \text{err}(P, Q) = L_\infty(P, Q) = \max_{x \in \mathbb{R}^d} G(x) \]
Computing $L_\infty$

**Definition**

\[ G(x) = |KDE_P(x) - KDE_Q(x)|. \]

\[ \text{err}(P, Q) = L_\infty(P, Q) = \max_{x \in \mathbb{R}^d} G(x) \]

Generate $X \subset \mathbb{R}^d$, return $\text{err}_X(P, Q) = \max_{x \in X} G(x)$

*converges*: as $|X| \to \infty$ then formally $\text{err}_X(P, Q) \to \text{err}(P, Q)$
Computing $\ell_\infty$

**Definition**

\[ G(x) = |\text{KDE}_P(x) - \text{KDE}_Q(x)|. \]
\[ \text{err}(P, Q) = \ell_\infty(P, Q) = \max_{x \in \mathbb{R}^d} G(x) \]

Generate $X \subset \mathbb{R}^d$, return $\text{err}_X(P, Q) = \max_{x \in X} G(x)$

**Converges:** as $|X| \to \infty$ then formally $\text{err}_X(P, Q) \to \text{err}(P, Q)$

**Two step strategy**

$G(x)$ is Lipschitz-continuous:

\[ \hat{x} \in \mathbb{R}^d \text{ close to the point } x^* = \arg \max_{x \in \mathbb{R}^d} G(x) \]

will also have error close to $\text{err}(P, Q)$.

For any radius $r$,

as $|X| \to \infty$ generate a point $\hat{x} \in X$ so that $\|x^* - \hat{x}\| \leq r$
Theorem 1
For $K_\sigma$ a unit Gaussian kernel, and two point sets $P, Q \in \mathbb{R}^d$, then $x^* = \arg\max_{x \in \mathbb{R}^d} G(x)$ must be in $M$, the Minkowski sum of a ball of radius $\sigma$ and the convex hull of $P \cup Q$. 
Baseline Methods

$\mathcal{B}$: the smallest axis-aligned bounding box that contains $M$.

**Rand**: Choose each point uniformly at random from $\mathcal{B}$

**Orgp**: Choose points uniformly at random from $P$.

**Orgp+N**: Choose points randomly from the original point set $P$, then add Gaussian noise with bandwidth $\sigma$, where $\sigma$ is the bandwidth of $K$.

**Grid**: Place a uniform grid on $\mathcal{B}$ and choose one point in each grid.

**Comb: Rand + Orgp**: The combination of method Rand and Orgp.
Cen\{E[m]\}:
Randomly select one point $p_1$ from the original point set $P$ and randomly choose $m$ neighbor points of $p_1$ within the distance of $3\sigma$. $m$ is chosen through a Exponential process with rate $1/E[m]$. Use the centroid of selected neighbor points as the evaluation point.
\text{Cen}\{E[m]\}:
Randomly select one point $p_1$ from the original point set $P$ and randomly choose $m$ neighbor points of $p_1$ within the distance of $3\sigma$. $m$ is chosen through a Exponential process with rate $1/E[m]$. Use the centroid of selected neighbor points as the evaluation point.
Cen\{E[m]\}:
Randomly select one point $p_1$ from the original point set $P$ and randomly choose $m$ neighbor points of $p_1$ within the distance of $3\sigma$. $m$ is chosen through a Exponential process with rate $1/E[m]$. Use the centroid of selected neighbor points as the evaluation point.
WCen{E[m]}:
Randomly select $p_1$ from $P$ and randomly choose $m$ neighbor points of $p_1$ proportional to the weight $\exp(-\frac{||p_n-p_1||^2}{2\sigma^2})$.
$m$ is chosen through a Exponential process with rate $1/E[m]$.
With probability 0.9, it remains $p_n$, with the remaining probability it is chosen randomly from a ball of radius $\sigma$ centered at $p_n$.
Use the weighted centroid of selected points as the evaluation point.
Bandwidth Selection

Lipschitz Properties of $h$, where $h(\omega) = \text{err}(P, \sigma, Q, \omega)$

Theorem 2: For any $\omega \geq \sigma \geq 1/A$, $h(\omega)$ is $\beta$-Lipschitz with respect to $\omega$, for
\[ \beta = \frac{1}{|Q|} \sum_{q \in Q} (x^* - q)^2 - 1/\pi |A^3 \]
where $x^* = \arg \max_{x \in \mathbb{R}^2} |\text{KDE}_{P,\sigma}(x) - \text{KDE}_{Q,\omega}(x)|$. 
Random Golden Section Search

Golden Section Search-unimodal function

If $f(x_4) = f_{4a}$ new triple $x_1, x_2, x_4$, $\frac{c}{a} = \frac{a}{b}$

If $f(x_4) = f_{4b}$ new triple $x_2, x_4, x_3$, $\frac{c}{b-c} = \frac{a}{b}$

Eliminating $c$ from these two simultaneous equations yields:

$\left(\frac{b}{a}\right)^2 = \frac{b}{a} + 1$, then $\frac{b}{a} = \phi = \frac{1+\sqrt{5}}{2} = 1.618033988...$
Random Golden Section Search

- Start from range $[l = \sigma, r = 10\sigma]$
- Choose one middle point at $m = \lambda\sigma$ for $\lambda \sim \text{Unif}(1, 10)$.
- If $h(m) > h(r)$, increase $r$ by a factor 10 until $h(m) < h(r)$.
- Repeat with several random values $\lambda$. 

---

![Diagram](image-url)
Outline

- Introduction and Background
- Problem Statement
- Methods
- Experiment Results
Experiment

Data Sets

1 dimension: 1-year hourly temperature data $|P| = 8760$.

$\sigma = 72$ (3 days), $\epsilon = 0.02$, $|Q| = 100$.

2 dimension: OpenStreetMap data from the state of Iowa.

$|P| = 1155102$ with $\sigma = 0.01$ $\epsilon = 0.1$ $|Q| = 1128$
Evaluating Point Generation for $\text{err}_X(P, Q)$

Find point sets $X$ so that $\text{err}_X(P, Q)$ is maximized with $|X|$ small.
Experiment Results-1D

Choosing New Bandwidth Evaluation
10 random trials of random golden section search,
\( \omega = 140 \) vs. \( \sigma = 72 \)
Experiment Results-1D

Choosing New Bandwidth Evaluation

10 random trials of random golden section search, \( \omega = 140 \) vs. \( \sigma = 72 \)
Evaluating Point Generation for $\text{err}_X(P, Q)$

$\text{err}_X(P, Q)$ is maximized with $|X|$ small, $X = 10000$. 

Experiment Results-2D
Evaluating Point Generation for $\text{err}_X(P, Q)$

$\text{err}_X(P, Q)$ is maximized with $|X|$ small, $X = 10000$. 

**Experiment Results-2D**

![Graphs showing the relationship between Num of Evaluation Points and Avg $L_\infty$ for different methods and parameters.](image)
Experiment Results-2D

Choosing New Bandwidth Evaluation

10 random trials of random golden section search, \( \sigma = 0.01 \) vs. \( \omega = 0.01 \)
Experiment Results-2D

Choosing New Bandwidth Evaluation
10 random trials of random golden section search, \( \sigma = 0.01 \) vs. \( \omega = 0.01 \)
Experiment Results-2D

Choosing New Bandwidth Evaluation
10 random trials of random golden section search,
\( \sigma = 0.01 \) vs. \( \omega = 0.024 \)
New Bandwidth for $L_1$ and $L_2$ Error

$\omega = 0.024$ minimize the $L_\infty$ error, $L_1 = 0.130746$, $L_2 = 0.189936$,
$\omega = 0.029$ minimize the $L_1 = 0.127588$ error,
$\omega = 0.025$ minimize the $L_2 = 0.189868$ error,
both are within 1% of the minimum solutions.

![Graph showing error with respect to $\omega$]
Thank you