Minimum Finite Covering

ML Section Talk
Tong Chen

September 1, 2025

Outline

1. Method
2. Issues
3. Solutions

General definition

ε-covering, ε-coreset (general case)

Let $\mathcal{S} \subseteq (\Omega, d)$, where Ω is a space endowed with distance metric $d : \Omega \times \Omega \to \mathbb{R}$. A set $\mathcal{T} \subseteq \Omega$ is said to be an ε-**covering** of $\mathcal{S}$, if for all $\mathbf{x} \in \mathcal{S}$, there exists $\mathbf{y} \in \mathcal{T}$, such that $d(\mathbf{x}, \mathbf{y}) \leq \varepsilon$, or equivalently,

$$\mathcal{S} \subseteq \bigcup_{\mathbf{y} \in \mathcal{T}} \mathbf{B}(\mathbf{y}, \varepsilon),$$

where $\mathbf{B}(\mathbf{y}, \varepsilon) := \{\mathbf{x} \in \Omega : d(\mathbf{x}, \mathbf{y}) \leq \varepsilon\}$. If $\mathcal{T} \subseteq \mathcal{S}$, we call it an ε-**coreset**.

Minimum ε-coreset of a finite set

Formulation in finte case

Let $\mathcal{S} = \{\mathbf{x}_i\}_{i=1}^N \subseteq (\Omega, d)$. For $\varepsilon > 0$, define the adjacency matrix of $\mathcal{S}$ as

$$\mathbf{A}(\varepsilon) := [a_{ij}(\varepsilon)], \; a_{ij}(\varepsilon) = \begin{cases} 1, & d(\mathbf{x}_i, \mathbf{x}_j) \leq \varepsilon; \\ 0, & \text{otherwise.} \end{cases}$$

Let $\mathcal{T} \subseteq \mathcal{S}$ and define $\mathbf{s} \in \{0,1\}^N$, where $s_i = 1$ if $\mathbf{x}_i \in \mathcal{T}$ otherwise $s_i = 0$. Then

(1) $\mathcal{T}$ is an ε-coreset $\iff \mathbf{A}(\varepsilon) \cdot \mathbf{s} \geq 1$;

(2) $\mathcal{T}$ is an ε-coreset with minimum size: $\min_{\mathbf{s} \in \{0,1\}^N} \{\|\mathbf{s}\|_1 : \mathbf{A}(\varepsilon) \cdot \mathbf{s} \geq 1\}$

Properties of minimum ε-coreset

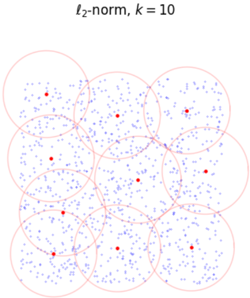
- Converging to original dataset $\mathcal{S}$.
- Relation to Hausdorff distance:

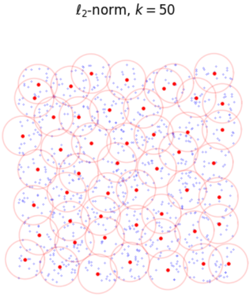
(1) $d_H(\mathcal{S}, \mathcal{T}) = \varepsilon \iff \begin{cases} \mathcal{S} \text{ is an } \varepsilon\text{-covering of } \mathcal{T}; \\ \mathcal{T} \text{ is an } \varepsilon\text{-covering of } \mathcal{S}. \end{cases}$

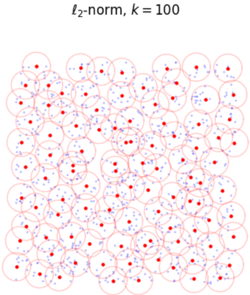
(2) If $\mathcal{T} \subseteq \mathcal{S}$, then $d_H(\mathcal{S}, \mathcal{T}) = \varepsilon \iff \mathcal{T}$ is an ε-covering of $\mathcal{S}$.

- Dimension-free: $\mathbf{A}(\varepsilon)$ is of size N-by-N.
- Distance flexible: $d : \Omega \times \Omega \to \mathbb{R}$.

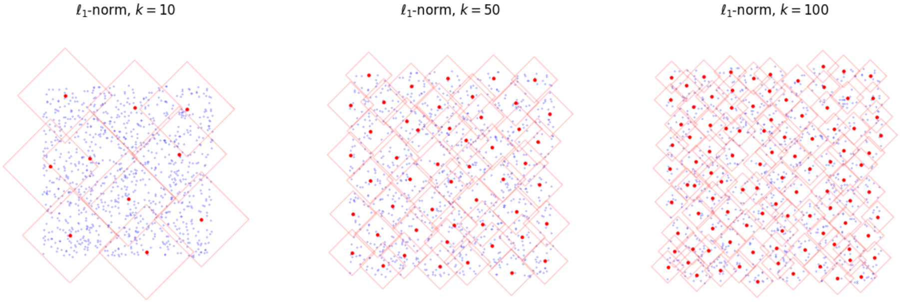
Example: uniform samples



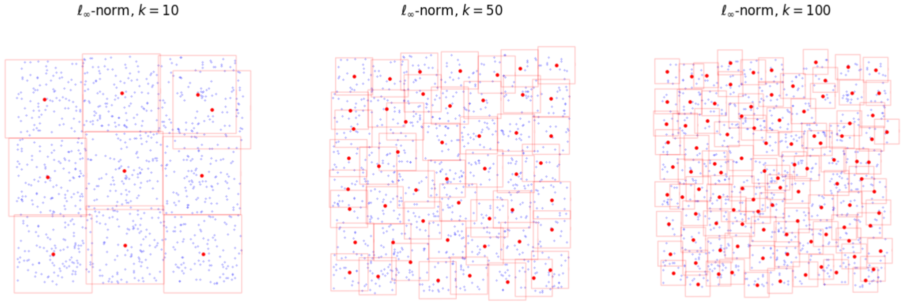




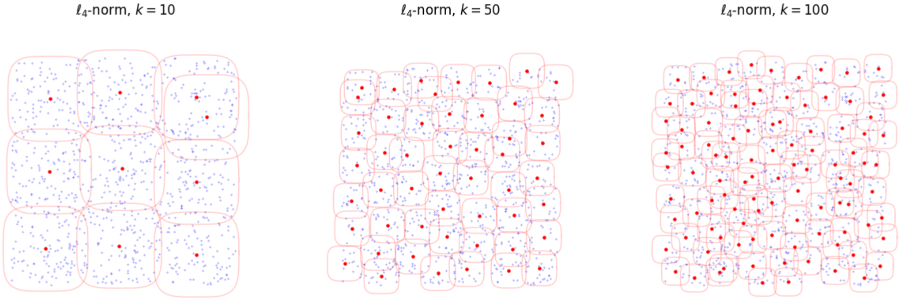
Example: uniform samples



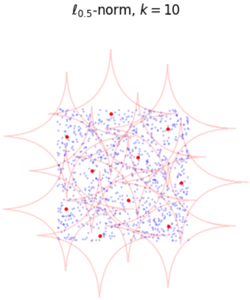
Example: uniform samples

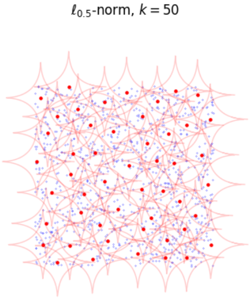


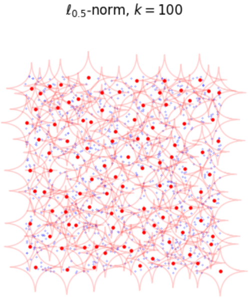
Example: uniform samples



Example: uniform samples







Issues for low-budget regime

- Density of data distribution
- Manifold structure

Issue 1: density

Let p be a probability density function,

- Covering using Euclidean distance: for some $\mathbf{x}_i, \mathbf{x}_j$,

$$\int_{\{\mathbf{x}:\ \|\mathbf{x}-\mathbf{x}_i\|_2 \leq \varepsilon\}} p(\mathbf{x})\mathrm{d}\mathbf{x} \neq \int_{\{\mathbf{x}:\ \|\mathbf{x}-\mathbf{x}_j\|_2 \leq \varepsilon\}} p(\mathbf{x})\mathrm{d}\mathbf{x}$$

- Need some distance metric, such that: for all $\mathbf{x}_i, \mathbf{x}_j$,

$$\int_{\{\mathbf{x}:\ d(\mathbf{x},\mathbf{x}_i) \leq \varepsilon\}} p(\mathbf{x})\mathrm{d}\mathbf{x} = \int_{\{\mathbf{x}:\ d(\mathbf{x},\mathbf{x}_j) \leq \varepsilon\}} p(\mathbf{x})\mathrm{d}\mathbf{x}$$

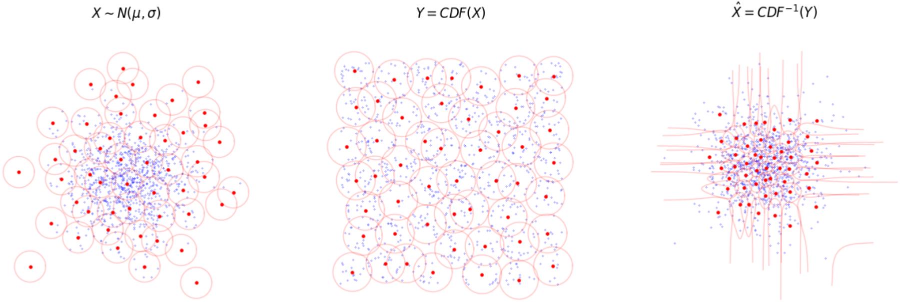
Solution to Issue 1

Let f be the cumulative distribution function (CDF) of p:

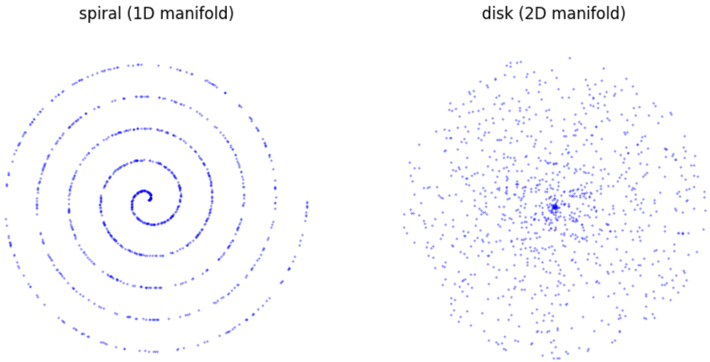
$$f(x) = \int_{-\infty}^{x} p(t)\mathrm{d}t,$$

- If $x \sim p(x)$, then $f(x)$ is **uniform**.
- Define the **pull-back** distance: $d_f(x, y) = |f(x) - f(y)|$.
- $\int_{\{x:\ d_f(x,x_i)\leq\varepsilon\}} p(x)\mathrm{d}x = \int_{\{x:\ d_f(x,x_j)\leq\varepsilon\}} p(x)\mathrm{d}x$

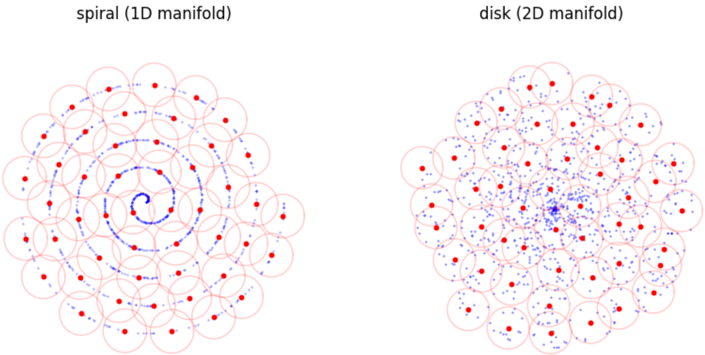
Covering of Gaussian samples



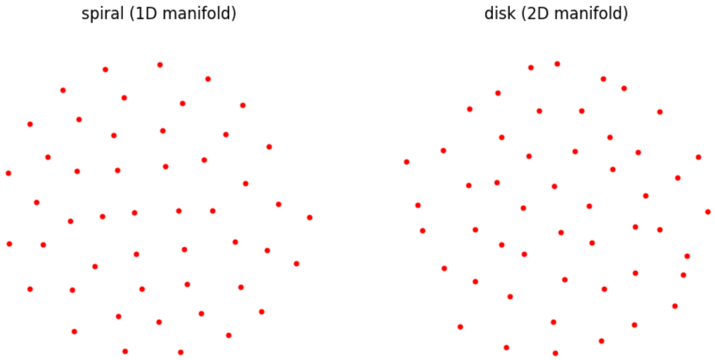
Issue 2: manifold



Covering of manifold



Covering of manifold



Solution to Issue 2

Assume all data is supported on a manifold $\mathcal{M}$, for $\mathbf{x}, \mathbf{y} \in \mathcal{M}$,

- Properly define a curve $\gamma : [0,1] \to \mathcal{M}$, with $\gamma(0) = \mathbf{x}, \gamma(1) = \mathbf{y}$.
- Compute curve length: $L(\gamma) = \int_0^1 \|\gamma'(t)\| \mathrm{d}t$.
- Define the distance by curve length:

$$d_g(\mathbf{x}, \mathbf{y}) = \inf_{\gamma} \{L(\gamma) : \gamma : [0,1] \to \mathcal{M}\}.$$

Summary

- Covering: dimension-free, distance flexible.
- Issues: density and manifold structure.
- Solutions:

  (1) Map non-uniform to uniform (flow-matching), Riemannian to Euclidean (VAE).

  (2) Pull distance back.
- Future work: performance?

Questions?