Leapfrog distance

Re-embedding data to strengthen recovery guarantees of clustering

Tao Jiang*
Cornell University, ORIE

Joint work with Stephen Vavasis** (Waterloo) and Samuel Tan (Cornell)

Clustering

Informally: Given n points $a_1, \dots, a_n \in \mathbb{R}^d$, partition $\{1, 2, \dots, n\}$ into *K* subsets $C_1, C_2, \dots, C_K$

such that for $i \in C_m, i' \in C_{m'}, \text{dist}(a_i, a_j)$ is small iff $m' = m$

Clustering

Informally: Given n points $a_1, \ldots, a_n \in \mathbb{R}^d$ generated by law μ, partition $\{1, 2, \ldots, n\}$ into K subsets $C_1, C_2, \ldots, C_K$ such that for $i \in C_m$ $i' \in C_{m'}$

- given the law is supported on disjoint, compact sets,

  then $m' = m$ if and only if i, i' lie in the same support;

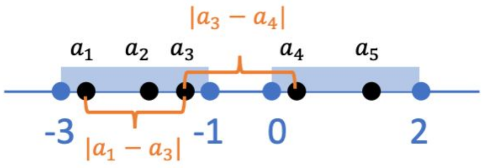


Figure: Example of 1D data generated by a law with disjoint support.

Clustering

Informally: Given n points $a_1, \dots, a_n \in \mathbb{R}^d$ generated by law μ, partition $\{1, 2, \dots, n\}$ into *K* subsets $C_1, C_2, \dots, C_K$ such that for $i \in C_m$ $i' \in C_{m'}$

- given the law is supported on disjoint, compact sets,

  then $m' = m$ if and only if i, i' lie in the same support;
- given the law is a mixture of distributions supported on a single, connected set,

  then $m' = m$ if and only if i, i' are generated by the same distribution.

Clustering

Informally: Given n points $a_1, \ldots, a_n \in \mathbb{R}^d$ generated by law μ, partition $\{1, 2, \ldots, n\}$ into K

subsets $C_1, C_2, \ldots, C_K$ such that for $i \in C_m$ $i' \in C_{m'}$

- given the law is supported on disjoint, compact sets,

  then $m' = m$ if and only if i, i' lie in the same support;

- given the law is a mixture of Gaussians supported on a single, connected set,

  then $m' = m$ if and only if i, i' are within certain standard deviations from the same mean.

Sum-of-norms clustering

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \frac{1}{2}\sum_{i=1}^{n} \|x_i - a_i\|_2^2 + \lambda \sum_{1 \le i<j \le n} \|x_i - x_j\|_2$$

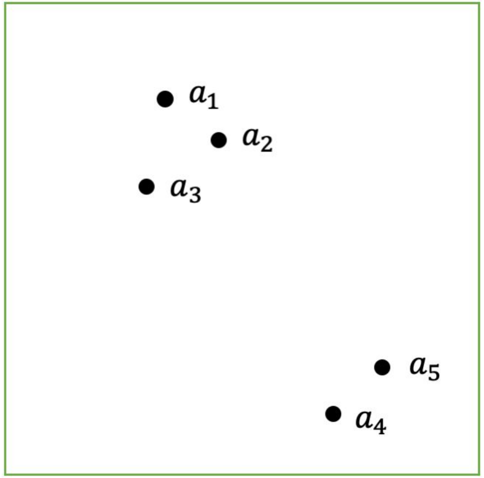


Figure: Given n points.

Sum-of-norms clustering

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \frac{1}{2}\sum_{i=1}^{n} \|x_i - a_i\|_2^2 + \lambda \sum_{1 \le i<j \le n} \|x_i - x_j\|_2$$

Figure: Define unconstrained variable x, which can be interpreted as the cluster “centroid” at optimality.

Sum-of-norms clustering

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \frac{1}{2}\sum_{i=1}^{n} \|x_i - a_i\|_2^2 + \lambda \sum_{1 \le i<j \le n} \|x_i - x_j\|_2$$

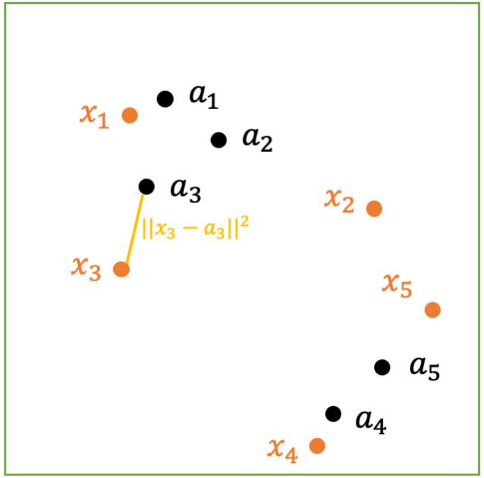


Figure: First term favors xi* close to ai.

Sum-of-norms clustering

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \frac{1}{2}\sum_{i=1}^{n} \|x_i - a_i\|_2^2 + \lambda \sum_{1\le i<j\le n} \|x_i - x_j\|_2$$

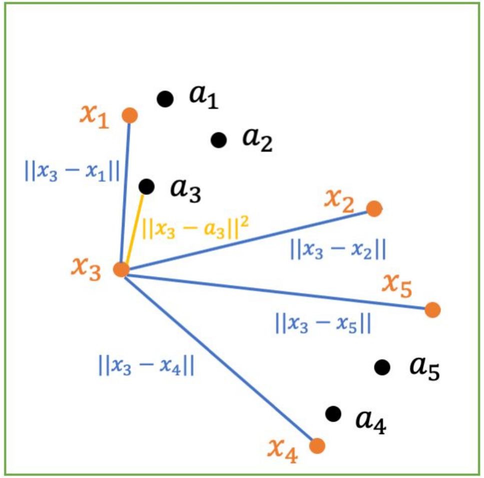


Figure: Second term tends to make many xi* equal to each other

Sum-of-norms clustering

$$\min_{x_1, \ldots, x_n \in \mathbb{R}^d} \frac{1}{2} \sum_{i=1}^{n} \|x_i - a_i\|_2^2 + \lambda \sum_{1 \leq i < j \leq n} \|x_i - x_j\|_2$$

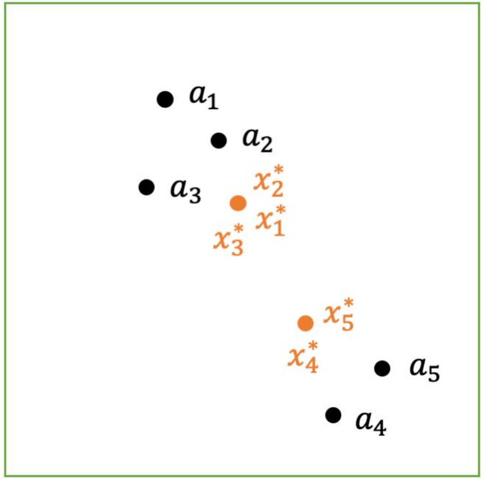


Figure: Optimal solution x* indicates cluster assignments

Good data VS bad data

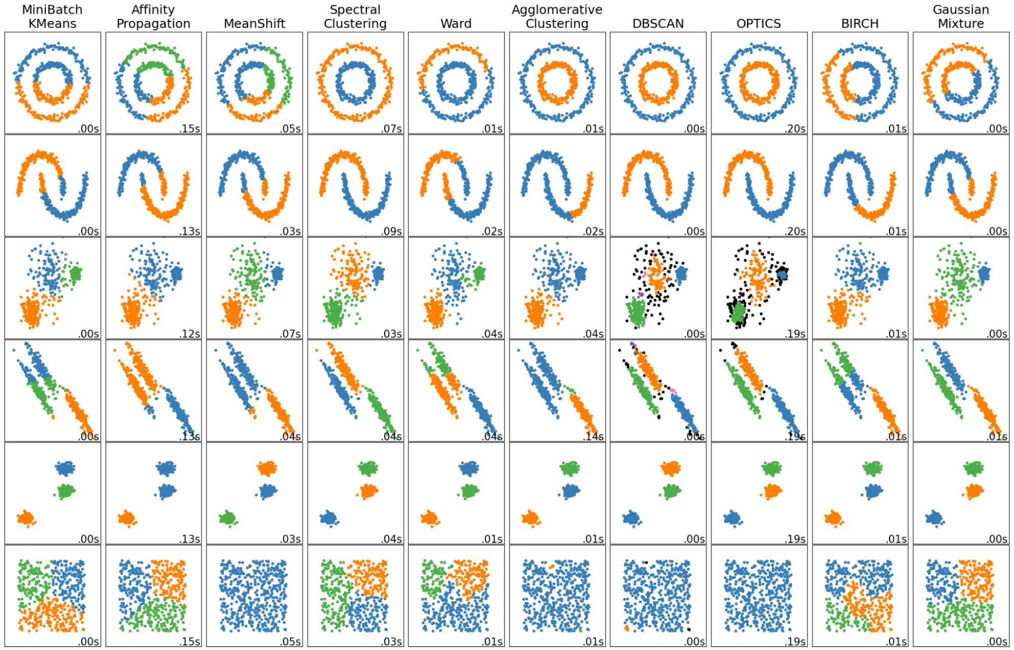
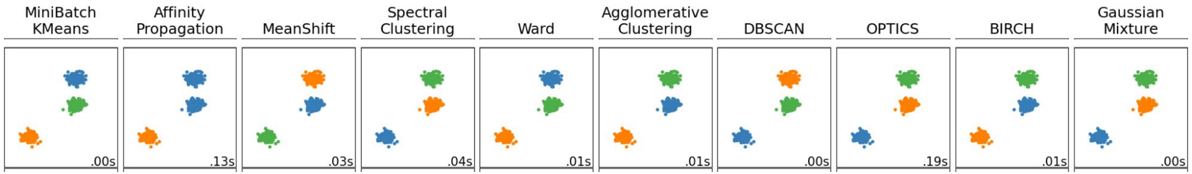


Image taken from https://scikit-learn.org/stable/modules/clustering.html

Good data

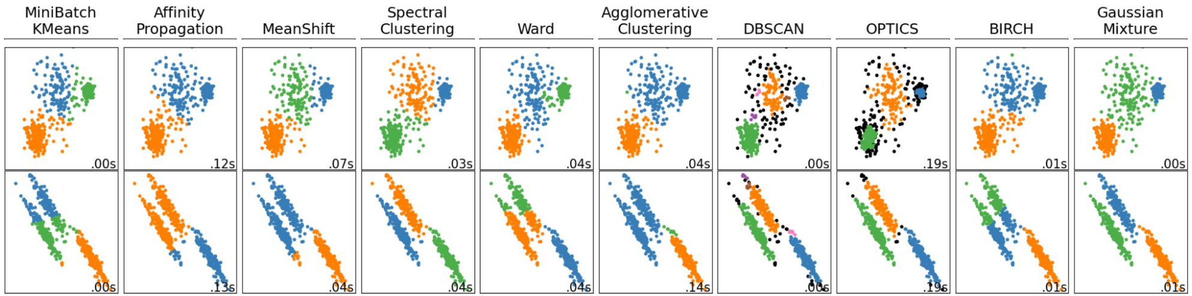


Well-separated blobs: Good!

The intra-cluster distance is small, the inter-cluster distance is huge.

Image taken from https://scikit-learn.org/stable/modules/clustering.html

Bad data

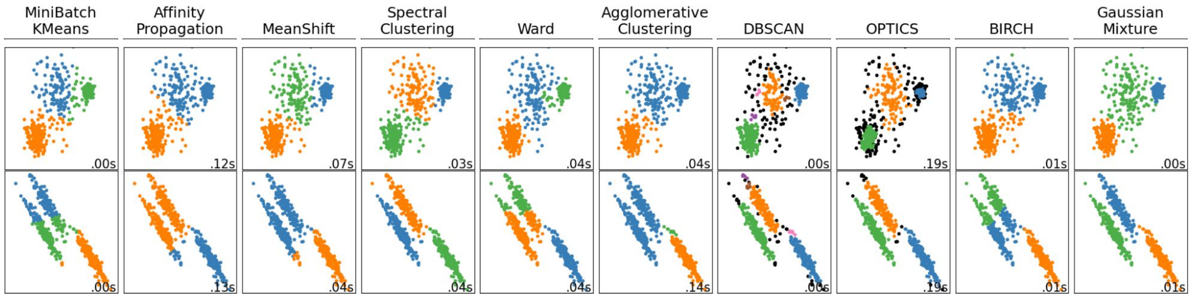


Mixture of anisotropic distributions: Bad!

The inter-cluster distance is too small.

Image taken from https://scikit-learn.org/stable/modules/clustering.html

Bad data for sum-of-norms clustering

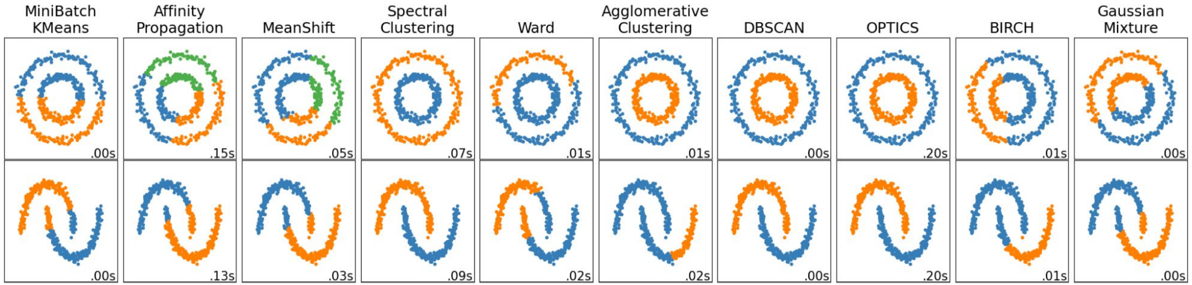


"… if the dataset is made of a large number of independent random variables distributed according to the uniform measure on the union of two disjoint balls of unit radius, and if the balls are sufficiently close to one another, then sum-of-norms clustering will typically fail to recover the decomposition of the dataset into two clusters."

— Dunlap and Mourrat, 2022

Image taken from https://scikit-learn.org/stable/modules/clustering.html

Bad data

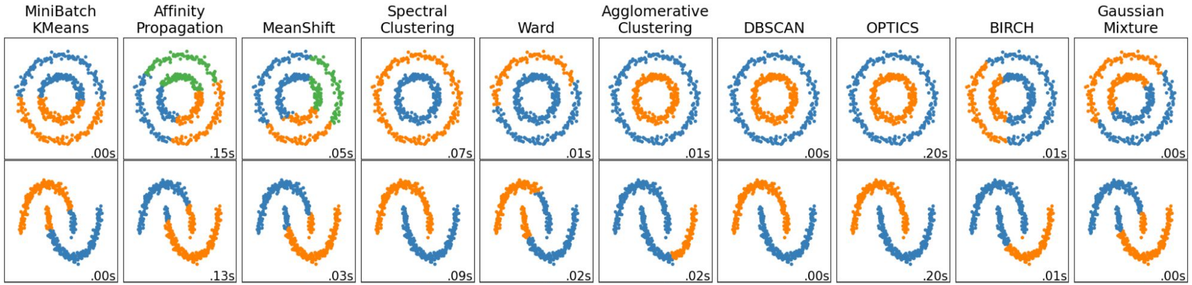


Concentric circles and half-moons: Bad!

The supports are not convex (the convex hull of supports overlap).

Image taken from https://scikit-learn.org/stable/modules/clustering.html

Bad data for sum-of-norms clustering



"Contrary to common expectation, we show that convex clustering can only learn convex clusters, unlike agglomerative clustering."

— Nguyen and Mamitsuka, 2021

Image taken from https://scikit-learn.org/stable/modules/clustering.html

Current solution - exponential weights

- **Good news:** introducing the exponential weights in front of the second term can overcome some of the problems (Toh et al; Dunlap and Mourrat)

Image taken from https://scikit-learn.org/stable/modules/clustering.html

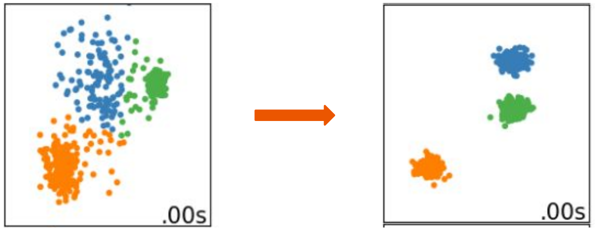
Current solution - exponential weights

- **Good news:** introducing the exponential weights in front of the second term can overcome some of the problems (Toh et al; Dunlap and Mourrat)
- **Bad news:** doing so loses some benefits of sum-of-norms clustering such as agglomerations properties (Chiquet et al), early stopping test (J&V), Sherman-Woodbury-Morrison formula in ADMM updates (Chi and Lange)

Image taken from https://scikit-learn.org/stable/modules/clustering.html

Leading question

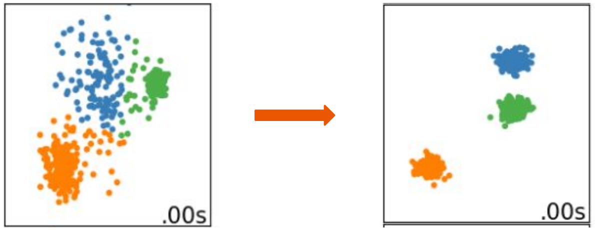
Can we convert the bad datasets into good ones?



Leading question

Can we convert the bad datasets into good ones? Yes!

- ❖ Propose a distance metric that increases the inter-cluster to intra-cluster distance ratio,
- ❖ and reconstruct a new dataset from the distance metric.



Constructing leapfrog distance

Example:

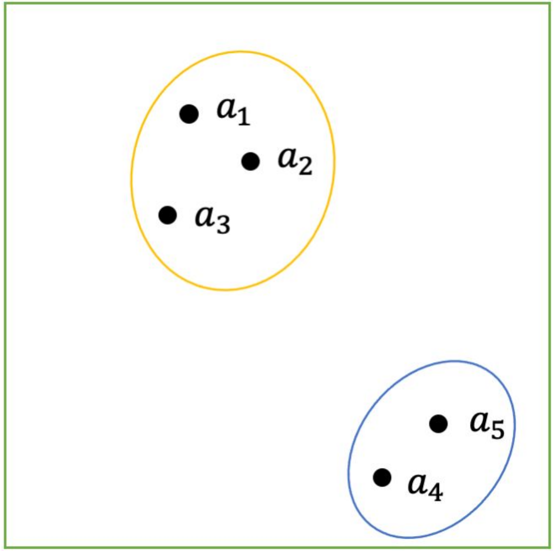


Figure: an example of constructing leapfrog distances for 5 points.

Constructing leapfrog distance

❖ Step 1. Build a complete graph out of the data.

Example:

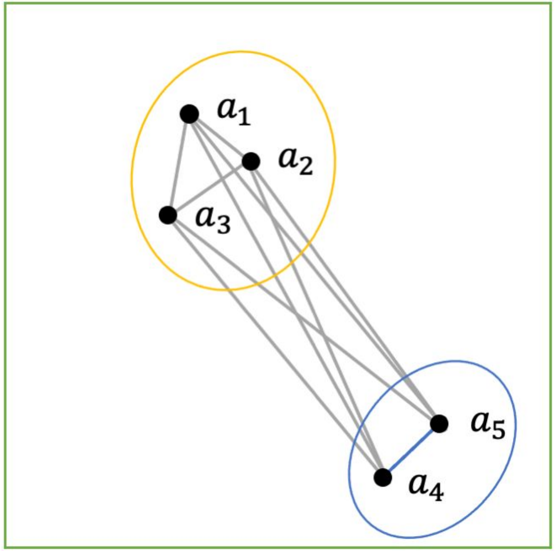


Figure: an example of constructing leapfrog distances for 5 points.

Constructing leapfrog distance

- Step 1. Build a complete graph out of the data.
- Step 2. Assign cost on edge *(i, j)* by $c_{ij} = \|a_i - a_j\|^2$

Example:

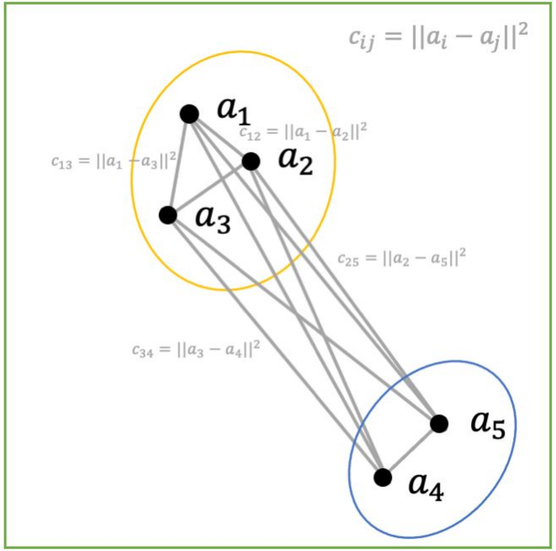


Figure: an example of constructing leapfrog distances for 5 points.

Constructing leapfrog distance

❖ Step 1. Build a complete graph out of the data.

❖ Step 2. Assign cost on edge *(i, j)* by $c_{ij} = \|a_i - a_j\|^2$

❖ Step 3. Define the leapfrog distance between *i, j* by the total cost on the shortest path between *i, j*

Example:

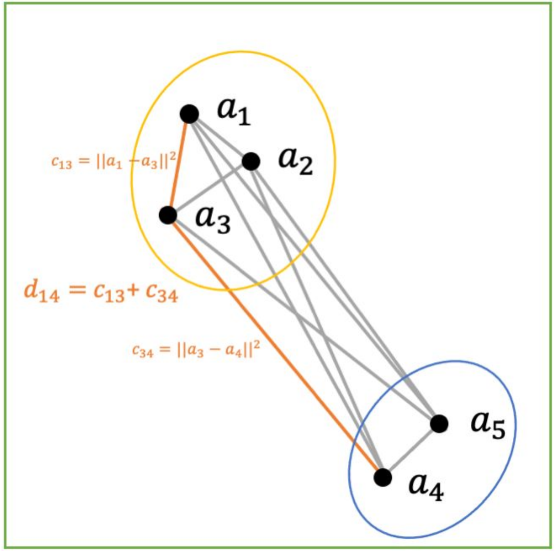


Figure: an example of constructing leapfrog distances for 5 points.

Constructing leapfrog distance

- Step 1. Build a complete graph out of the data.
- Step 2. Assign cost on edge *(i, j)* by $c_{ij} = \|a_i - a_j\|^2$
- Step 3. Define the leapfrog distance between *i, j* by the total cost on the shortest path between *i, j*

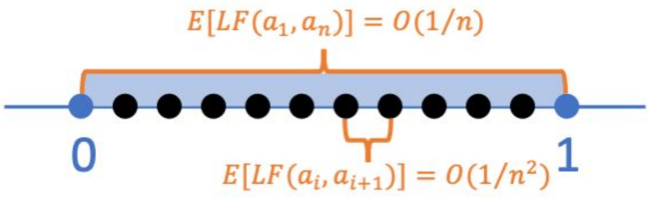


Figure: an example of constructing leapfrog distances for n uniformly distributed points on [0,1].

Example:

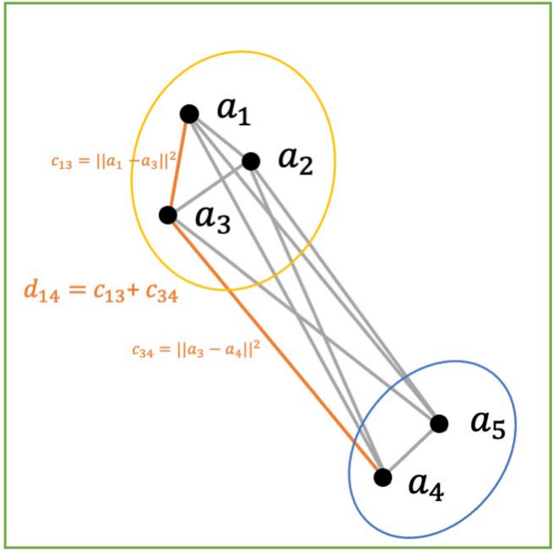


Figure: an example of constructing leapfrog distances for 5 points.

Properties of leapfrog distance

Disjoint, compact supports

Example:

Figure: the expected configuration of data generated by uniform distributions supported on [0,⅓] and [⅓, ⅔]

Assumption:

- the dataset is generated by a law;
- the law is supported on disjoint, compact sets.

Properties of leapfrog distance

Disjoint, compact supports

Example:

Figure: the expected configuration of data generated by uniform distributions supported on [0,⅓] and [⅓, ⅔]

Properties of leapfrog distance

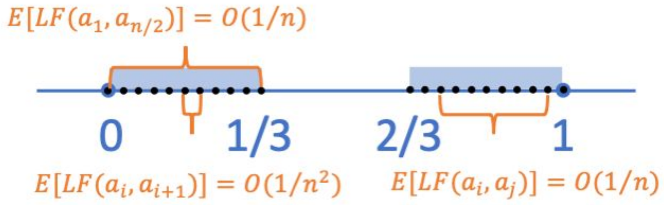
Disjoint, compact supports

Example:

Figure: the expected configuration of data generated by uniform distributions supported on [0,⅓] and [⅓, ⅔]

As the sample size grows…

❖ Property 1. The intra-cluster distance ⟶ 0



Example:

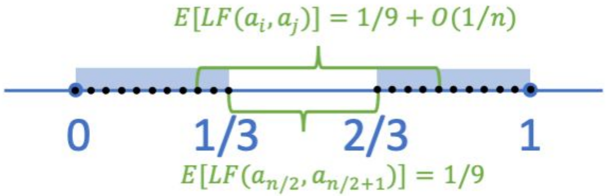
Figure: the expected configuration of data generated by uniform distributions supported on [0,⅓] and [⅓, ⅔]

Properties of leapfrog distance

Disjoint, compact supports

As the sample size grows…

- Property 1. The intra-cluster distance $\longrightarrow$ 0
- Property 2. The inter-cluster distance $\longrightarrow$ c > 0



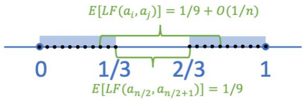
Properties of leapfrog distance

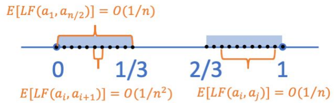
Disjoint, compact supports

Figure: the expected configuration of data generated by uniform distributions supported on [0,⅓] and [⅓, ⅔]

As the sample size grows...

- ❖ Property 1. The intra-cluster distance ⟶ 0
- ❖ Property 2. The inter-cluster distance ⟶ c > 0





Properties of leapfrog distance

Disjoint, compact supports

Example:

Figure: the expected configuration of data generated by uniform distributions supported on [0,⅓] and [⅓, ⅔]

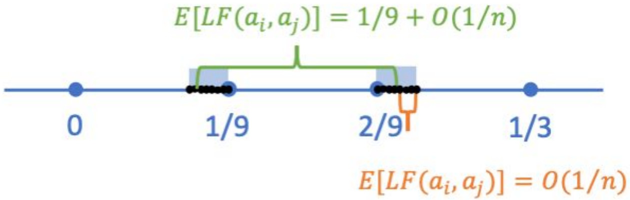
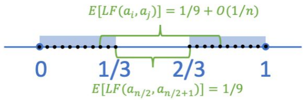


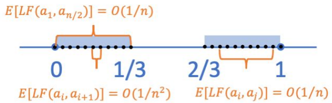
Figure: the equivalent configuration of figure 1 using leapfrog distance

As the sample size grows...

❖ Property 1. The intra-cluster distance ⟶ 0

❖ Property 2. The inter-cluster distance ⟶ c > 0





Properties of leapfrog distance

~~Disjoint, compact supports~~ ?

Example:

Figure: the expected configuration of data generated by uniform distributions supported on [0,⅓] and [⅓, ⅔]

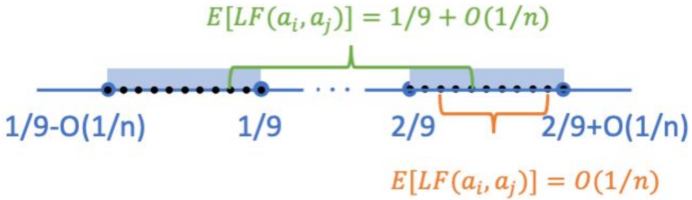
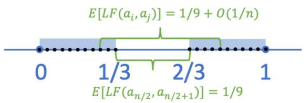


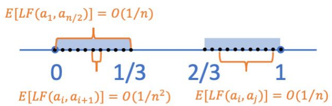
Figure: the equivalent configuration of figure 1 using leapfrog distance

As the sample size grows…

❖ Property 1. The intra-cluster distance ⟶ 0

❖ Property 2. The inter-cluster distance ⟶ c > 0



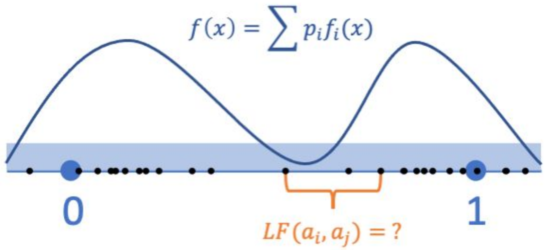


Properties of leapfrog distance

A single connected support in 1D

Assumption:

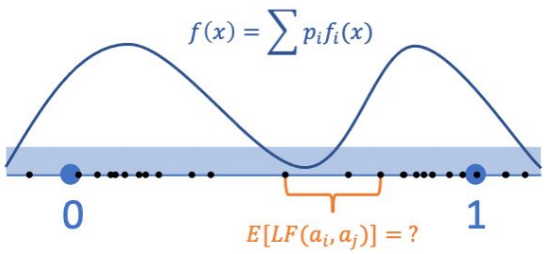
- the dataset is generated by a mixture of distributions;
- the law is supported on a single connected set in 1D.



Properties of leapfrog distance

A single connected support in 1D

Lemma: $$E[\mathrm{LF}(\mathrm{a},\mathrm{b})] = \int_a^b \frac{2}{(n+1)f(x)}dx + o(1/n)$$



Properties of leapfrog distance

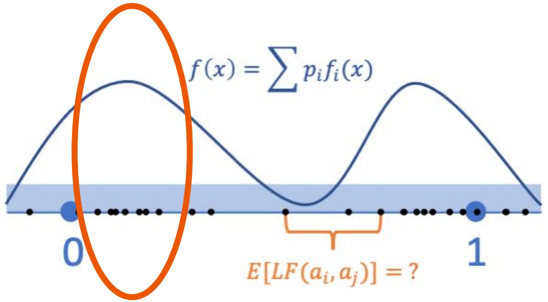
A single connected support in 1D

Lemma: $E[\mathrm{LF}(\mathrm{a}, \mathrm{b})] = \int_a^b \frac{2}{(n+1)f(x)} dx + o(1/n)$

Peak area: higher f $\Rightarrow$ more samples

$\Rightarrow$ shortest path consists of many small steps

$\Rightarrow$ smaller leapfrog distance.



Properties of leapfrog distance

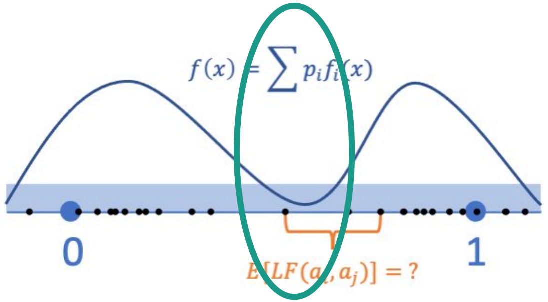
A single connected support in 1D

Lemma: $E[\mathrm{LF}(\mathrm{a},\mathrm{b})] = \int_a^b \frac{2}{(n+1)f(x)} dx + o(1/n)$

Valley area: lower f $\Rightarrow$ fewer samples

$\Rightarrow$ shortest path consists of a few big leaps

$\Rightarrow$ larger leapfrog distance.



Properties of leapfrog distance

A single connected support in 1D

Lemma: $E[\mathrm{LF}(\mathrm{a},\mathrm{b})] = \int_a^b \frac{2}{(n+1)f(x)}dx + o(1/n)$

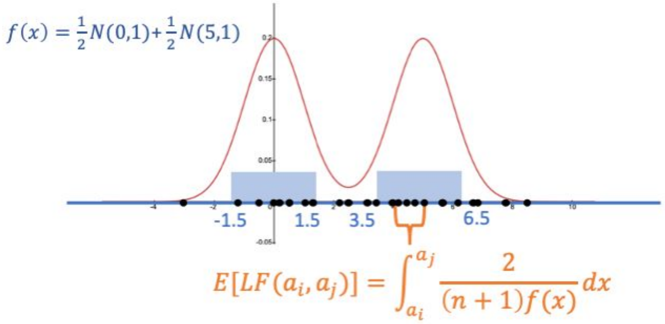
Peak area: higher f $\Rightarrow$ more samples

$\Rightarrow$ smaller leapfrog distance.

Valley area: lower f $\Rightarrow$ fewer samples

$\Rightarrow$ larger leapfrog distance.

Example:



Properties of leapfrog distance

A single connected support in 1D

Lemma: $E[\mathrm{LF(a, b)}] = \int_a^b \frac{2}{(n+1)f(x)} dx + o(1/n)$

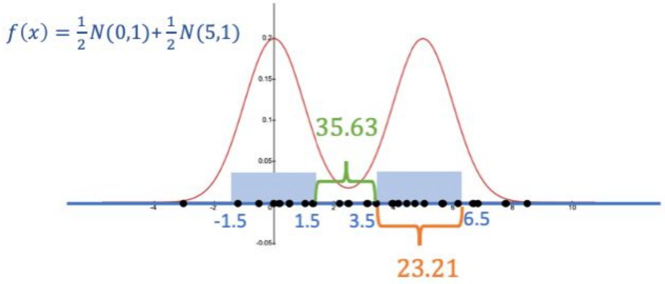
Peak area: higher f $\Rightarrow$ more samples

$\Rightarrow$ smaller leapfrog distance.

Valley area: lower f $\Rightarrow$ fewer samples

$\Rightarrow$ larger leapfrog distance.

Example:



Properties of leapfrog distance

A single connected support in 1D

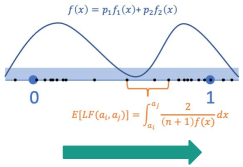
Lemma: $$E[\mathrm{LF}(\mathrm{a},\mathrm{b})] = \int_a^b \frac{2}{(n+1)f(x)}dx + o(1/n)$$

Peak area: higher f $\Rightarrow$ more samples

$\Rightarrow$ smaller leapfrog distance.

Valley area: lower f $\Rightarrow$ fewer samples

$\Rightarrow$ larger leapfrog distance.



Concentration inequalities

Properties of leapfrog distance

A single connected support in 1D

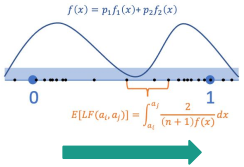
Lemma: $E[\mathrm{LF}(\mathrm{a},\mathrm{b})] = \int_a^b \frac{2}{(n+1)f(x)}dx + o(1/n)$

Peak area: higher *f* $\Rightarrow$ more samples

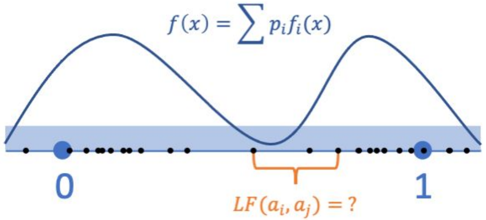
$\Rightarrow$ smaller leapfrog distance.

Valley area: lower *f* $\Rightarrow$ fewer samples

$\Rightarrow$ larger leapfrog distance.



Concentration inequalities

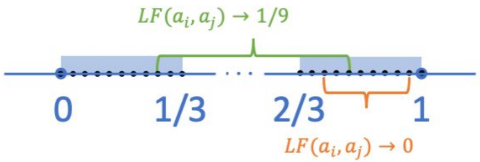


High probability bound for $\mathrm{LF}(a, b)$

Properties of leapfrog distance

Summary

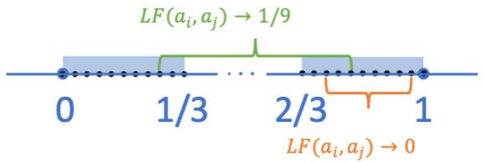
- ❖ If the law is supported on disjoint, compact sets.



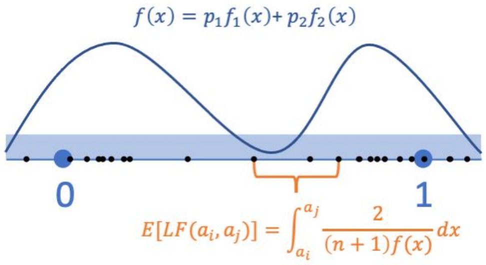
Properties of leapfrog distance

Summary

- ❖ If the law is supported on disjoint, compact sets.



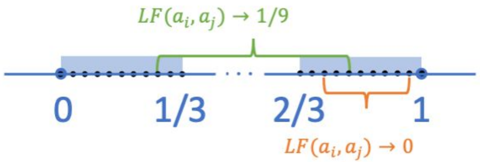
- ❖ If the law is supported on a single connected set in 1D.



Properties of leapfrog distance

Summary

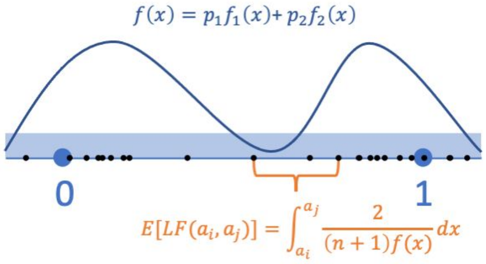
- ❖ If the law is supported on disjoint, compact sets.



Distance

$$\begin{bmatrix} LF(a_1, a_1) & \cdots & LF(a_1, a_n) \\ LF(a_1, a_2) & & LF(a_2, a_n) \\ \vdots & \ddots & \vdots \\ LF(a_1, a_n) & \cdots & LF(a_n, a_n) \end{bmatrix}$$

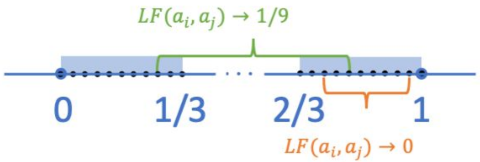
- ❖ If the law is supported on a single connected set in 1D.



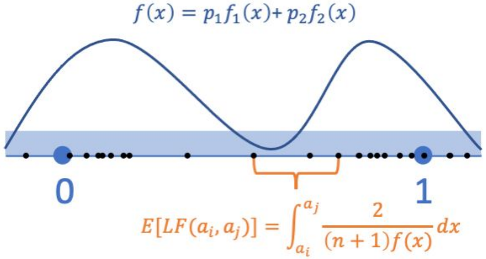
Properties of leapfrog distance

Summary

❖ If the law is supported on disjoint, compact sets.



❖ If the law is supported on a single connected set in 1D.



Distance

$$\begin{bmatrix} LF(a_1, a_1) & \cdots & LF(a_1, a_n) \\ LF(a_1, a_2) & & LF(a_2, a_n) \\ \vdots & \ddots & \vdots \\ LF(a_1, a_n) & \cdots & LF(a_n, a_n) \end{bmatrix}$$

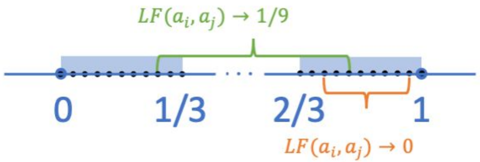
$$\min_{x_1, \ldots, x_n \in \mathbb{R}^d} \frac{1}{2} \sum_{i=1}^{n} \|x_i - a_i\|^2 + \lambda \sum_{1 \le i < j \le n} \|x_i - x_j\|$$

Coordinates a_i

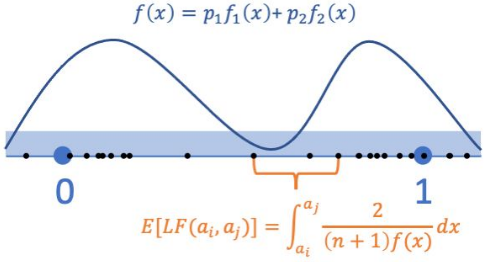
Properties of leapfrog distance

Summary

- If the law is supported on disjoint, compact sets.



- If the law is supported on a single connected set in 1D.



Distance

$$\begin{bmatrix} LF(a_1, a_1) & \cdots & LF(a_1, a_n) \\ LF(a_1, a_2) & & LF(a_2, a_n) \\ \vdots & \ddots & \vdots \\ LF(a_1, a_n) & \cdots & LF(a_n, a_n) \end{bmatrix}$$

?

$$\min_{x_1, \ldots, x_n \in \mathbb{R}^d} \frac{1}{2} \sum_{i=1}^{n} \|x_i - a_i\|^2 + \lambda \sum_{1 \le i < j \le n} \|x_i - x_j\|$$

Coordinates a_i

From distances to coordinates

Trivial embedding in 1D

Distance

$$\begin{bmatrix} LF(a_1, a_1) & \cdots & LF(a_1, a_n) \\ LF(a_1, a_2) & & LF(a_2, a_n) \\ \vdots & \ddots & \vdots \\ LF(a_1, a_n) & \cdots & LF(a_n, a_n) \end{bmatrix}$$

$$\Big\downarrow ?$$

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \frac{1}{2} \sum_{i=1}^{n} \|x_i - a_i\|^2 + \lambda \sum_{1 \le i < j \le n} \|x_i - x_j\|$$

Coordinates a_i

Without loss of generality, we may assume that

$$a_1 \le a_2 \le \cdots \le a_n$$

From distances to coordinates

Trivial embedding in 1D

Distance

$$\begin{bmatrix} LF(a_1,a_1) & \cdots & LF(a_1,a_n) \\ LF(a_1,a_2) & & LF(a_2,a_n) \\ \vdots & \ddots & \vdots \\ LF(a_1,a_n) & \cdots & LF(a_n,a_n) \end{bmatrix}$$

$$\Big\downarrow ?$$

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \frac{1}{2}\sum_{i=1}^{n} \|x_i - a_i\|^2 + \lambda \sum_{1 \le i<j \le n} \|x_i - x_j\|$$

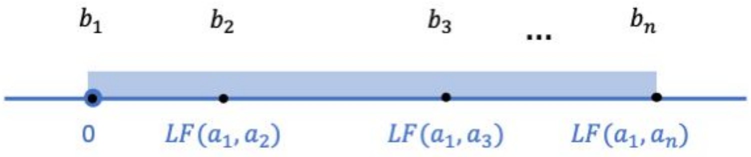
Coordinates a_i

Without loss of generality, we may assume that

$$a_1 \le a_2 \le \cdots \le a_n$$

Then we construct a new dataset $b_1, b_2, \ldots, b_n$ by

❖ Step 1: setting $b_1 = 0$



From distances to coordinates

Trivial embedding in 1D

Distance

$$\begin{bmatrix} LF(a_1,a_1) & \cdots & LF(a_1,a_n) \\ LF(a_1,a_2) & & LF(a_2,a_n) \\ \vdots & \ddots & \vdots \\ LF(a_1,a_n) & \cdots & LF(a_n,a_n) \end{bmatrix}$$

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \frac{1}{2}\sum_{i=1}^{n} \|x_i - a_i\|^2 + \lambda \sum_{1 \le i<j \le n} \|x_i - x_j\|$$

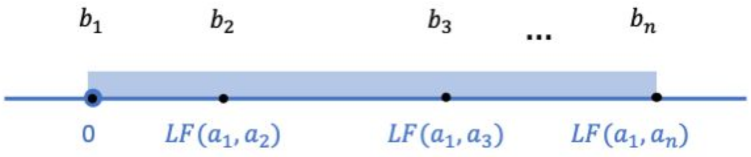
Coordinates a_i

Without loss of generality, we may assume that

$$a_1 \le a_2 \le \cdots \le a_n$$

Then we construct a new dataset $b_1, b_2, \ldots, b_n$ by

- Step 1: setting $b_1 = 0$
- Step 2: setting any point $b_i = \mathrm{LF}(a_1, a_i)$



From distances to coordinates

Trivial embedding in 1D

Distance

$$\begin{bmatrix} LF(a_1,a_1) & \cdots & LF(a_1,a_n) \\ LF(a_1,a_2) & & LF(a_2,a_n) \\ \vdots & \ddots & \vdots \\ LF(a_1,a_n) & \cdots & LF(a_n,a_n) \end{bmatrix}$$

Without loss of generality, we may assume that

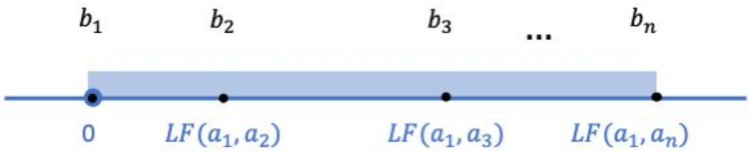
$$a_1 \leq a_2 \leq \cdots \leq a_n$$

Then we construct a new dataset $b_1, b_2, \ldots, b_n$ by

- Step 1: setting $b_1 = 0$
- Step 2: setting any point $b_i = \mathrm{LF}(a_1, a_i)$

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \frac{1}{2}\sum_{i=1}^{n} \|x_i - a_i\|^2 + \lambda \sum_{1 \leq i < j \leq n} \|x_i - x_j\|$$

Coordinates a_i



From distances to coordinates

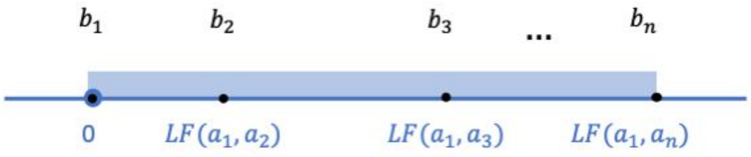
Trivial embedding in 1D

Without loss of generality, we may assume that

$$a_1 \leq a_2 \leq \cdots \leq a_n$$

Then we construct a new dataset $b_1, b_2, \ldots, b_n$ by

- ❖ Step 1: setting $b_1 = 0$
- ❖ Step 2: setting any point $b_i = \mathrm{LF}(a_1, a_i)$



Consequences:

$$\begin{aligned} |b_i - b_j| &= |\mathrm{LF}(a_1, a_i) - \mathrm{LF}(a_1, a_j)| \\ &= \mathrm{LF}(a_i, a_j) \end{aligned}$$

From distances to coordinates

Embedding in arbitrary dimension: disjoint compact supports

Distance

$$D = \begin{bmatrix} LF(a_1,a_1)^2 & \cdots & LF(a_1,a_n)^2 \\ LF(a_1,a_2)^2 & & LF(a_2,a_n)^2 \\ \vdots & \ddots & \vdots \\ LF(a_1,a_n)^2 & \cdots & LF(a_n,a_n)^2 \end{bmatrix}$$

Construct a new dataset $b_1, b_2, \ldots, b_n$ by

multidimensional scaling

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \frac{1}{2} \sum_{i=1}^{n} \|x_i - a_i\|^2 + \lambda \sum_{1 \le i < j \le n} \|x_i - x_j\|$$

Coordinates a_i

From distances to coordinates

Embedding in arbitrary dimension: disjoint compact supports

Distance

$$D = \begin{bmatrix} LF(a_1,a_1)^2 & \cdots & LF(a_1,a_n)^2 \\ LF(a_1,a_2)^2 & & LF(a_2,a_n)^2 \\ \vdots & \ddots & \vdots \\ LF(a_1,a_n)^2 & \cdots & LF(a_n,a_n)^2 \end{bmatrix}$$

Construct a new dataset $b_1, b_2, \ldots, b_n$ by

multidimensional scaling

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \frac{1}{2} \sum_{i=1}^{n} \|x_i - a_i\|^2 + \lambda \sum_{1 \le i<j \le n} \|x_i - x_j\|$$

Coordinates a_i

Consequences:

- $i,\, j \in C \iff \|b_i - b_j\|_2 \to 0$
- $i,\, j \notin C \iff \|b_i - b_j\|_2 \to c > 0$

From distances to coordinates

Summary

❖ If dimension = 1: trivial embedding $b_1, b_2, \ldots, b_n$

$$|b_i - b_j| = \mathrm{LF}(a_i, a_j)$$

From distances to coordinates

Summary

- If dimension = 1: trivial embedding $b_1, b_2, \ldots, b_n$

$$|b_i - b_j| = \mathrm{LF}(a_i, a_j)$$

- If dimension = d with disjoint, compact supports
  - $i,\ j \in C \iff \|b_i - b_j\|_2 \to 0$
  - $i,\ j \notin C \iff \|b_i - b_j\|_2 \to c > 0$

From distances to coordinates

Summary

- If dimension = 1: trivial embedding $b_1, b_2, \ldots, b_n$

$$|b_i - b_j| = \mathrm{LF}(a_i, a_j)$$

- If dimension = d with disjoint, compact supports
  - $i,\ j \in C \iff \|b_i - b_j\|_2 \to 0$
  - $i,\ j \notin C \iff \|b_i - b_j\|_2 \to c > 0$

Distance

$$\begin{bmatrix} LF(a_1, a_1) & \cdots & LF(a_1, a_n) \\ LF(a_1, a_2) & & LF(a_2, a_n) \\ \vdots & \ddots & \vdots \\ LF(a_1, a_n) & \cdots & LF(a_n, a_n) \end{bmatrix}$$

Coordinates $b_1, b_2, \ldots, b_n$

$$\min_{x_1, \ldots, x_n \in \mathbb{R}^d} \frac{1}{2} \sum_{i=1}^{n} \|x_i - a_i\|^2 + \lambda \sum_{1 \le i < j \le n} \|x_i - x_j\|$$

Model

From distances to coordinates

Summary

- If dimension = 1: trivial embedding $b_1, b_2, \ldots, b_n$

$$|b_i - b_j| = \mathrm{LF}(a_i, a_j)$$

- If dimension = d with disjoint, compact supports
  - $i, j \in C \iff \|b_i - b_j\|_2 \to 0$
  - $i, j \notin C \iff \|b_i - b_j\|_2 \to c > 0$

Distance

$$\begin{bmatrix} LF(a_1,a_1) & \cdots & LF(a_1,a_n) \\ LF(a_1,a_2) & & LF(a_2,a_n) \\ \vdots & \ddots & \vdots \\ LF(a_1,a_n) & \cdots & LF(a_n,a_n) \end{bmatrix}$$

$\Downarrow$

Coordinates $b_1, b_2, \ldots, b_n$

$\Downarrow$ Consequences?

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \frac{1}{2}\sum_{i=1}^{n} \|x_i - a_i\|^2 + \lambda \sum_{1 \leq i < j \leq n} \|x_i - x_j\|$$

Model

Recovery of clusters

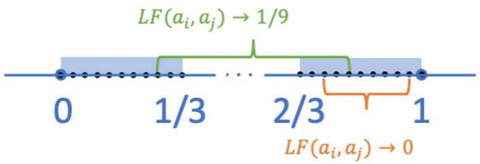
Coordinates $b_1, b_2, \dots, b_n$

Consequences?

$$\min_{x_1, \dots, x_n \in \mathbb{R}^d} \frac{1}{2} \sum_{i=1}^{n} \|x_i - a_i\|^2 + \lambda \sum_{1 \leq i < j \leq n} \|x_i - x_j\|$$

Model

❖ If the underlying law is supported on disjoint, compact sets: perfect recovery because



Recovery of clusters

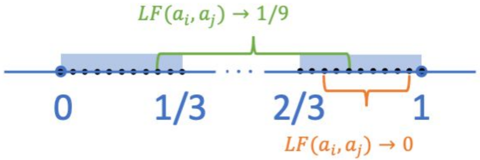
Coordinates $b_1, b_2, \ldots, b_n$

⬇ Consequences?

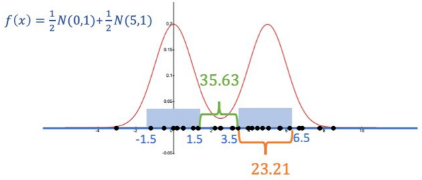
$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \frac{1}{2}\sum_{i=1}^{n} \|x_i - a_i\|^2 + \lambda \sum_{1 \le i < j \le n} \|x_i - x_j\|$$

Model

❖ If the underlying law is supported on disjoint, compact sets: perfect recovery because

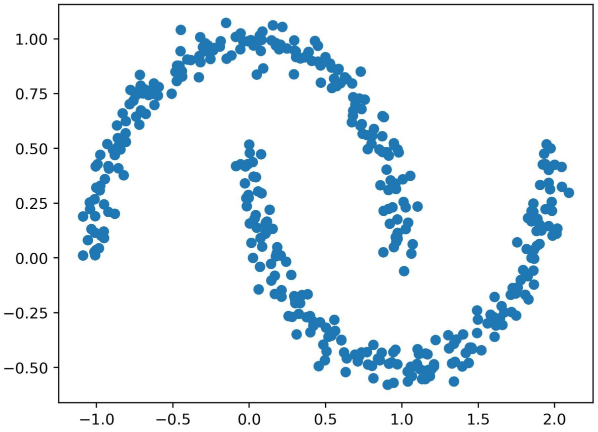


❖ If the underlying law is a mixture of Gaussians supported on a single connected set in 1D: perfect recovery for points close to the means

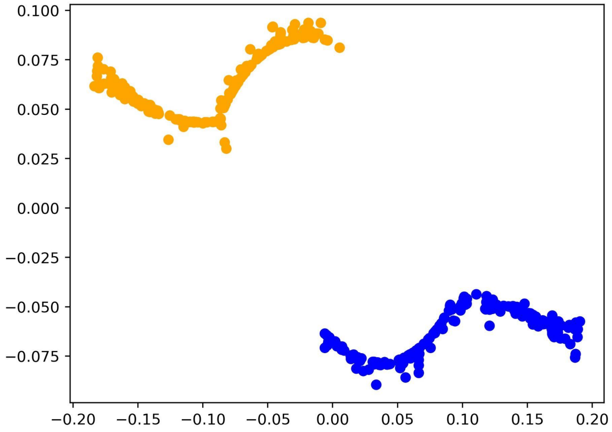


Visualization of MDS embedding + clustering

2 half-moons



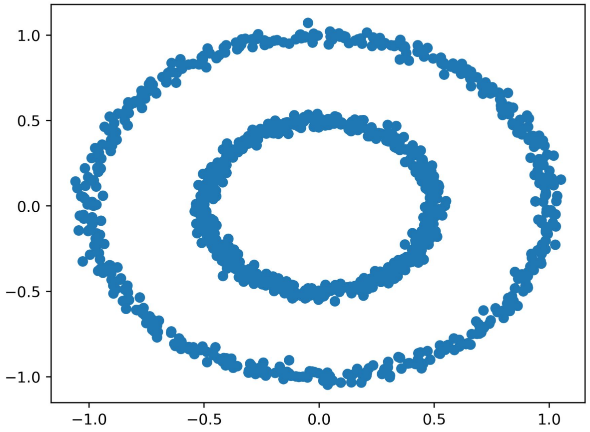
Original coordinates



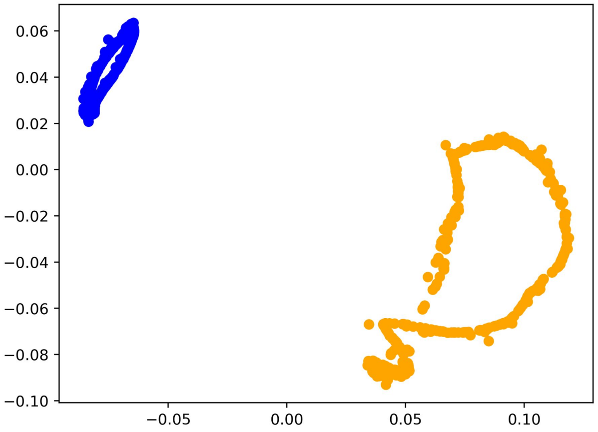
MDS embedding

Visualization of MDS embedding + clustering

Concentric Circles



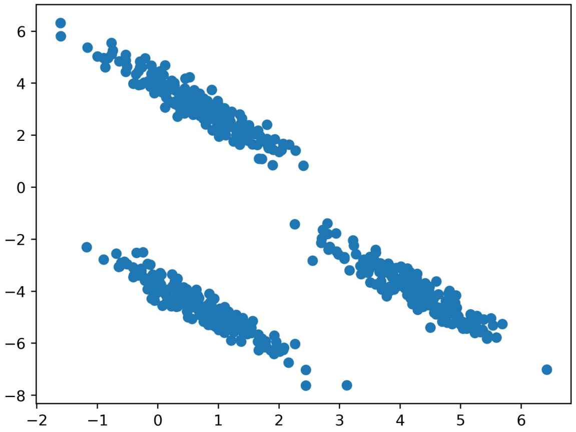
Original coordinates



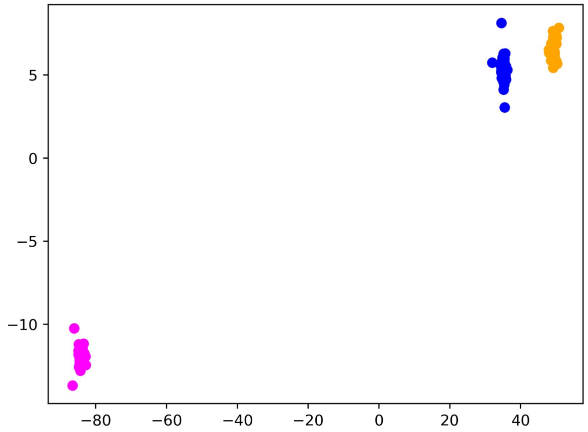
MDS embedding

Visualization of MDS embedding + clustering

Anisotropic mixture



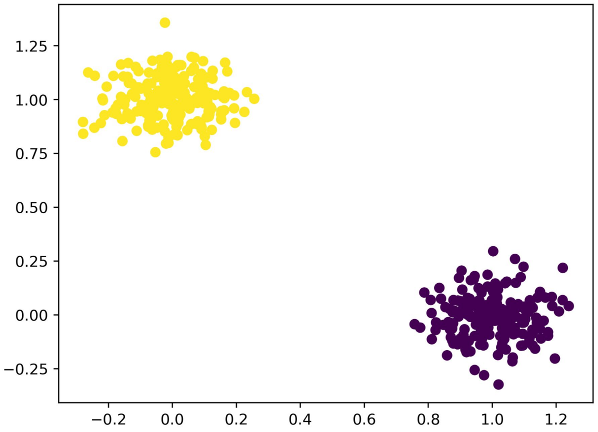
Original coordinates



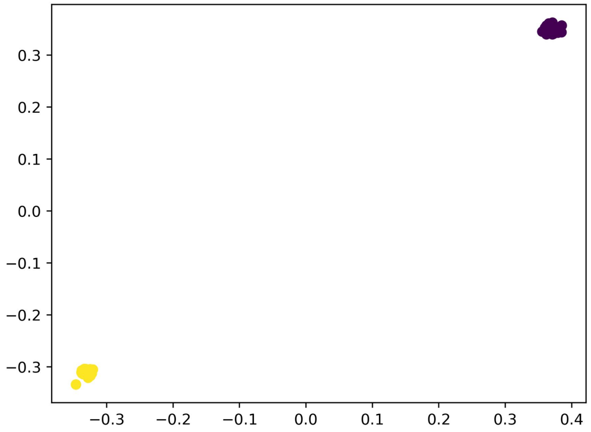
MDS embedding

Visualization of MDS embedding

Mixture of 2 2D Gaussians centered at (0,1) and (1,0): $\sigma = 0.1$



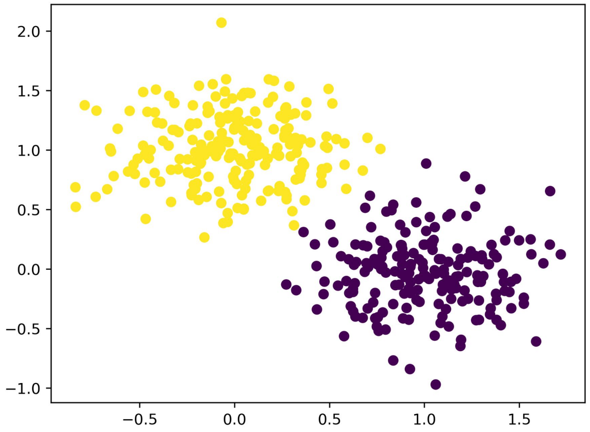
Original coordinates



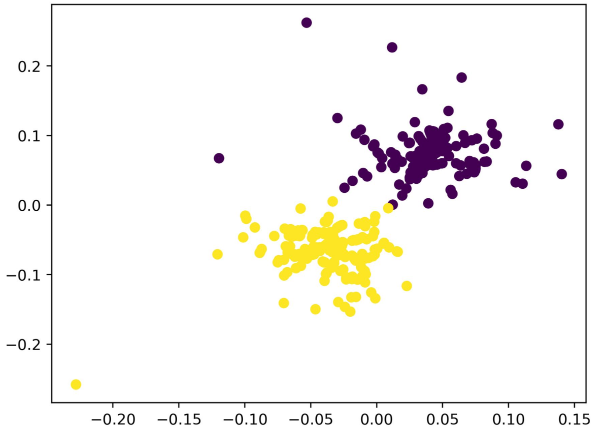
MDS embedding

Visualization of MDS embedding

Mixture of 2 2D Gaussians centered at (0,1) and (1,0): σ = 0.3



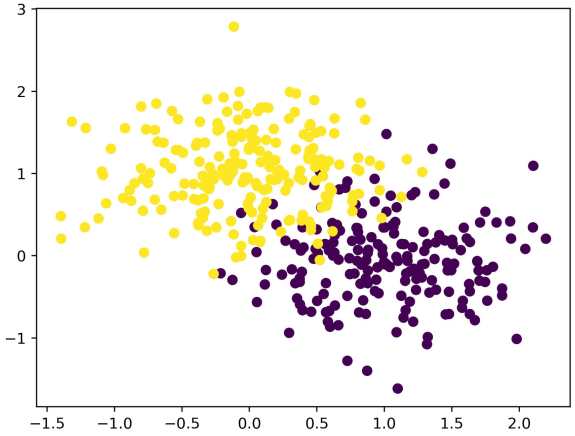
Original coordinates



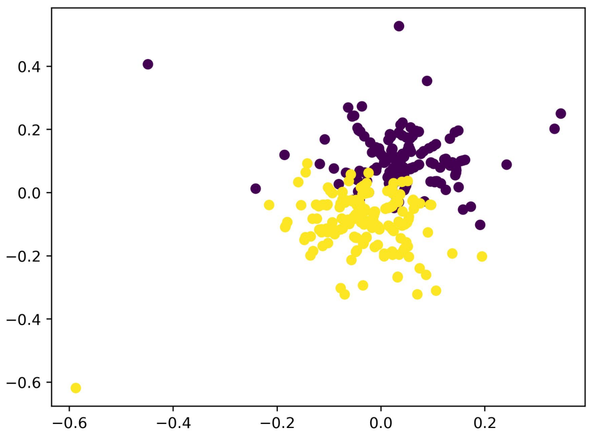
MDS embedding

Visualization of MDS embedding

Mixture of 2 2D Gaussians centered at (0,1) and (1,0): $\sigma = 0.5$



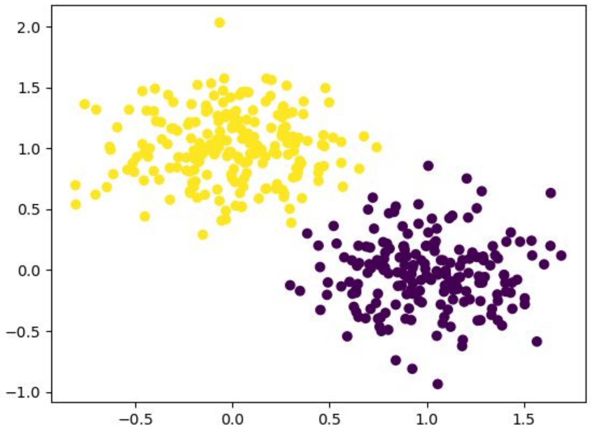
Original coordinates

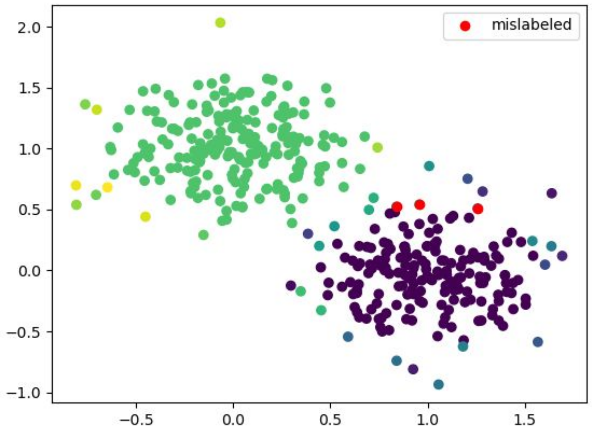


MDS embedding

Visualization of Clustering: No embedding

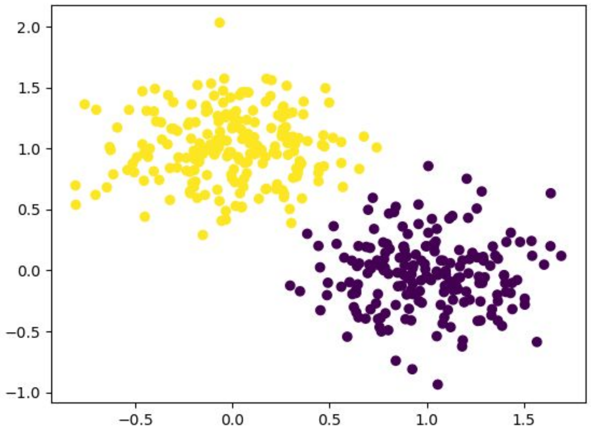
Mixture of 2 2D Gaussians centered at (0,1) and (1,0): $\sigma = 0.29$

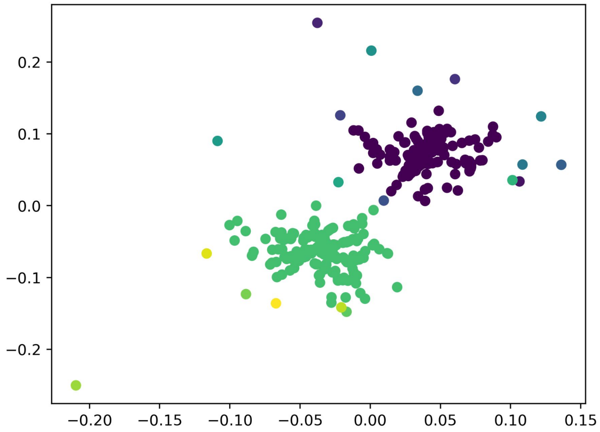




Visualization of Clustering: MDS Embedding

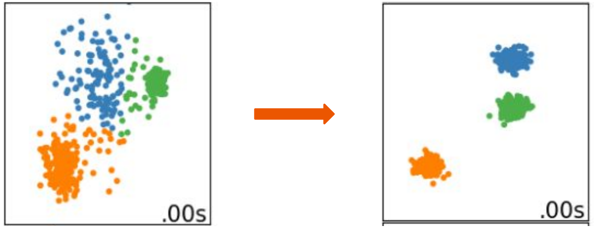
Mixture of 2 2D Gaussians centered at (0,1) and (1,0): σ = 0.29





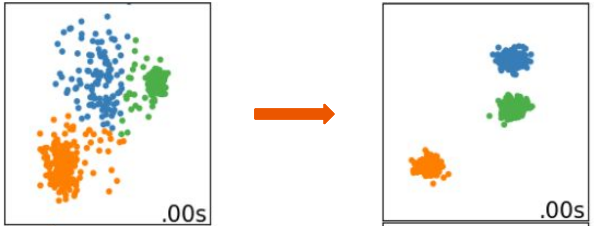
Perfect recovery within 2 standard deviations!

Closing Remarks



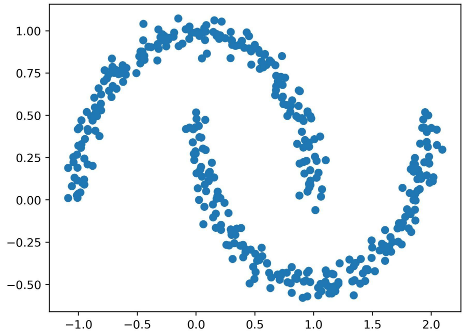
- We converted the bad datasets into good ones by
  - proposing the leapfrog distance which increases the inter-cluster to intra-cluster distance ratio,
  - and reconstructing a new dataset from the distance metric using MDS.
- We proved useful properties of leapfrog distances for data generated by laws supported on disjoint, compact sets and single, connected sets in 1D.
- We are able to improve the performance of sum-of-norms clustering and strengthen the recovery theory.

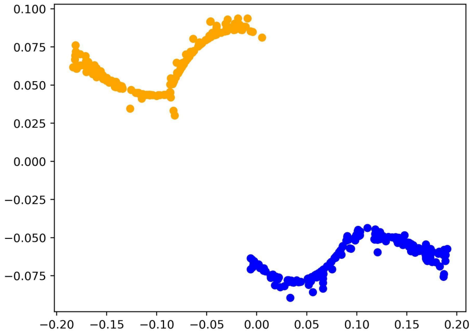
Closing Remarks



- What about other clustering algorithms?
- What about the properties of leapfrog distance for datasets supported on single connected sets in higher dimension?

Thank you!

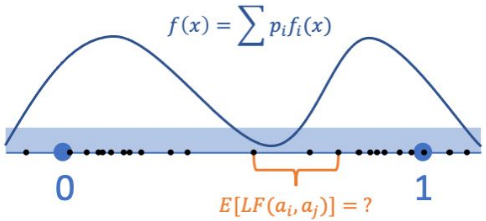

1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0

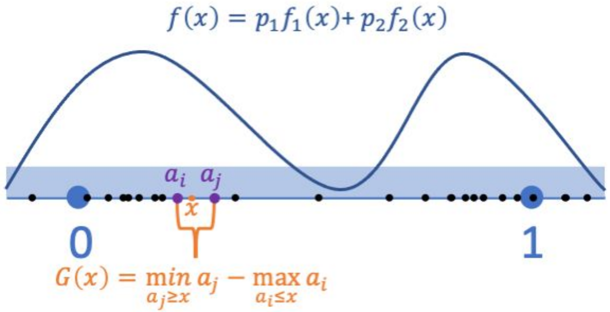

0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
−0.20
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20

Properties of leapfrog distance

A single connected support in 1D

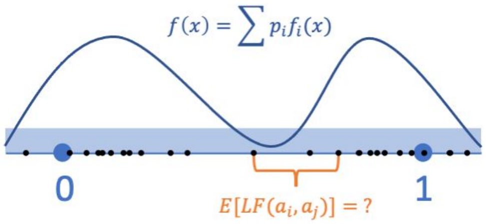
Define a random variable $G(x) = \min_{a_j \geq x} a_j - \max_{a_i \leq x} a_j$.





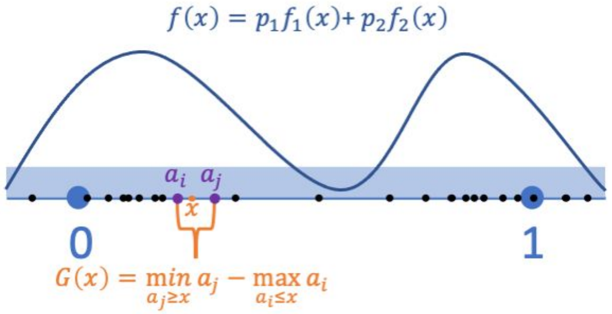
Properties of leapfrog distance

A single connected support in 1D



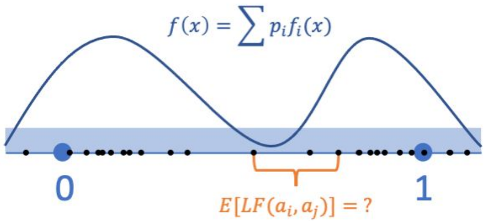
Define a random variable $G(x) = \min_{a_j \geq x} a_j - \max_{a_i \leq x} a_j$.

Claim: $\mathrm{LF}(a,b) = \int_a^b G(x)dx$



Properties of leapfrog distance

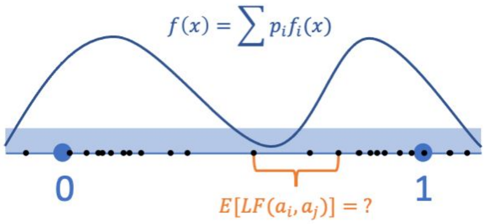
A single connected support in 1D



Claim 1: $\mathrm{LF}(a,b) = \int_a^b G(x)dx$, where $G(x) = \min_{a_j \geq x} a_j - \max_{a_i \leq x} a_j$

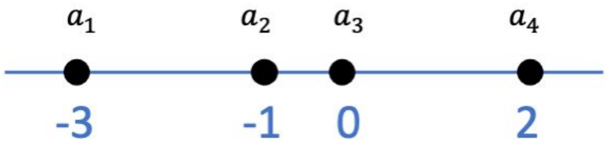
Properties of leapfrog distance

A single connected support in 1D



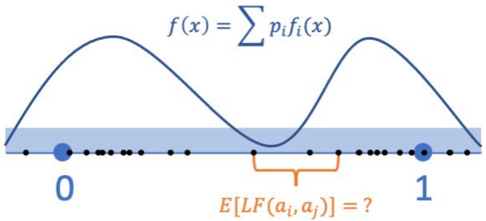
Claim 1: $\mathrm{LF}(a,b) = \int_a^b G(x)dx$, where $G(x) = \min_{a_j \geq x} a_j - \max_{a_i \leq x} a_j$

Example:



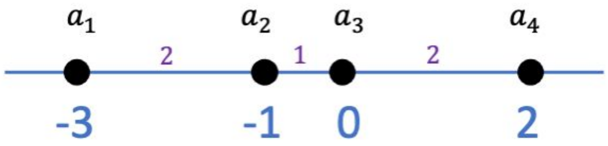
Properties of leapfrog distance

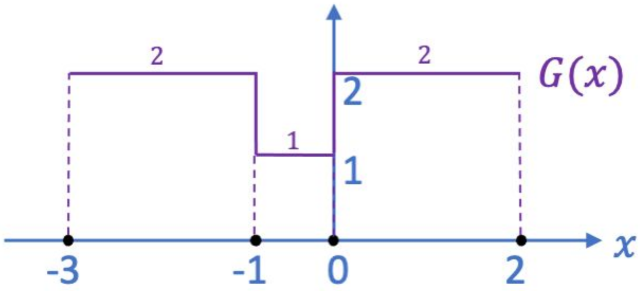
A single connected support in 1D



Claim 1: $\mathrm{LF}(a,b) = \int_a^b G(x)dx$, where $G(x) = \min_{a_j \geq x} a_j - \max_{a_i \leq x} a_j$

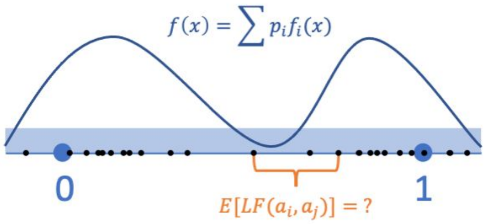
Example:





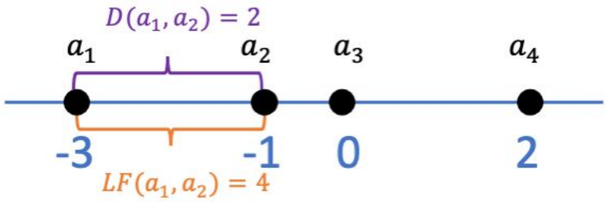
Properties of leapfrog distance

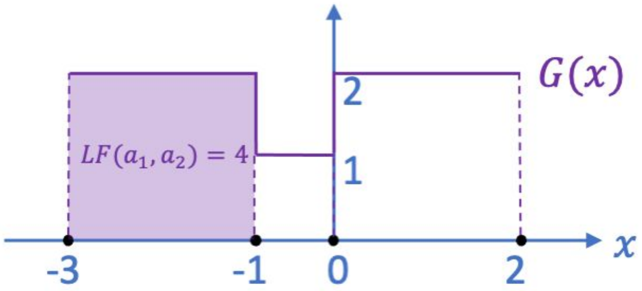
A single connected support in 1D



Claim 1: $\mathrm{LF}(a, b) = \int_a^b G(x)dx$, where $G(x) = \min_{a_j \geq x} a_j - \max_{a_i \leq x} a_j$

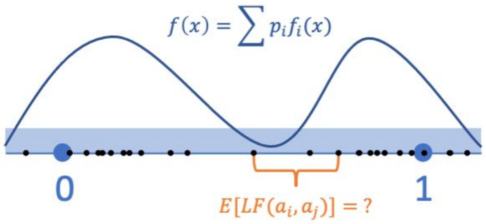
Example:





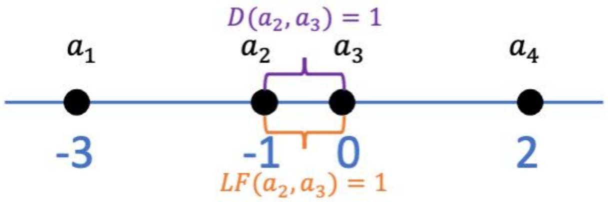
Properties of leapfrog distance

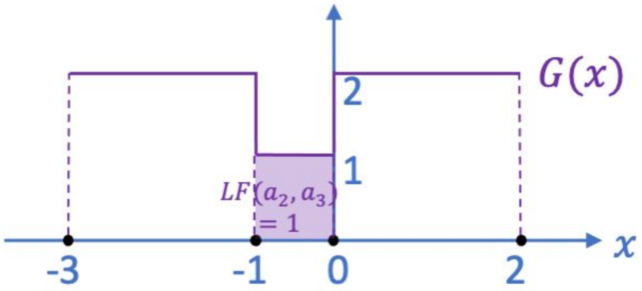
A single connected support in 1D



Claim 1: $\mathrm{LF}(a,b) = \int_a^b G(x)dx$, where $G(x) = \min_{a_j \geq x} a_j - \max_{a_i \leq x} a_j$

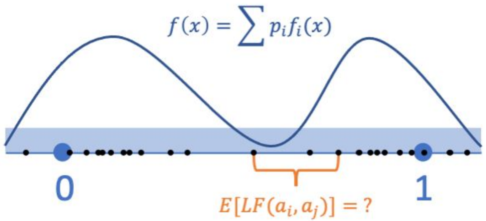
Example:





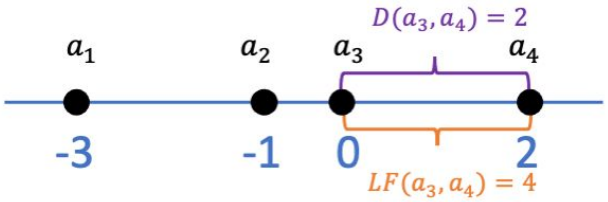
Properties of leapfrog distance

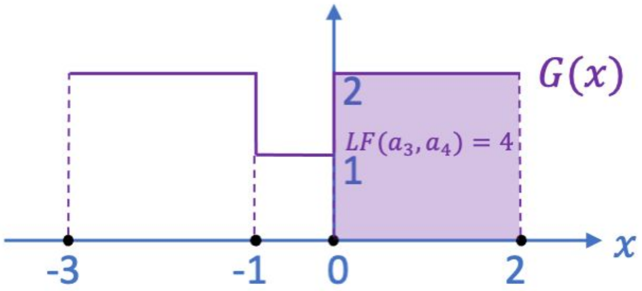
A single connected support in 1D



Claim 1: $\mathrm{LF}(a,b) = \int_a^b G(x)dx$, where $G(x) = \min_{a_j \geq x} a_j - \max_{a_i \leq x} a_j$

Example:



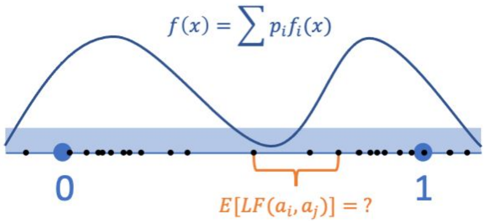


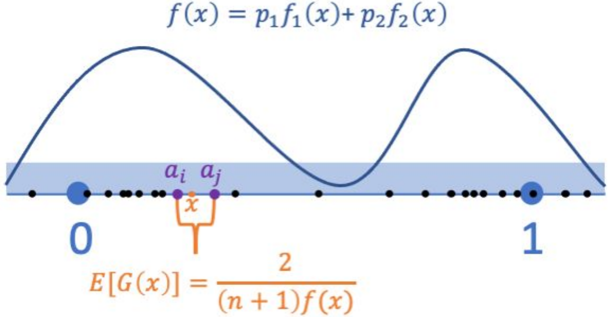
Properties of leapfrog distance

A single connected support in 1D

Claim 1: $\mathrm{LF}(a,b) = \int_a^b G(x)dx$

Claim 2: $E[G(x)] = \dfrac{2}{(n+1)f(x)} + o(1/n)$





Properties of leapfrog distance

A single connected support in 1D

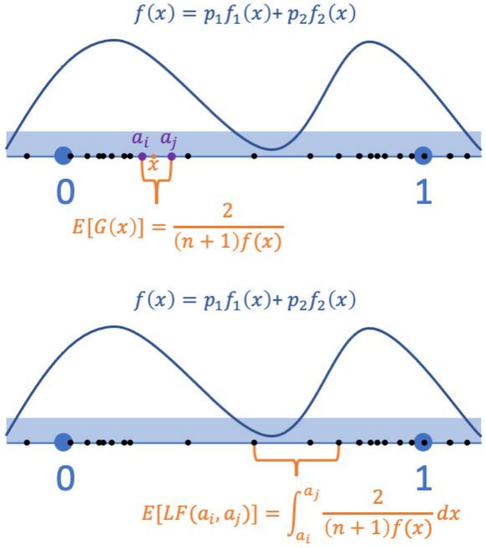
Claim 1: $\mathrm{LF}(a,b) = \int_a^b G(x)dx$

+

Claim 2: $E[G(x)] = \frac{2}{(n+1)f(x)} + o(1/n)$

=

Lemma: $E[\mathrm{LF}(\mathrm{a},\mathrm{b})] = \int_a^b \frac{2}{(n+1)f(x)} dx + o(1/n)$



Properties of leapfrog distance

A single connected support in 1D

Claim 1: $\mathrm{LF}(a,b) = \int_a^b G(x)dx$

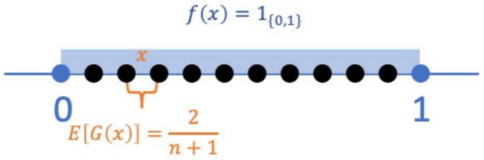
+

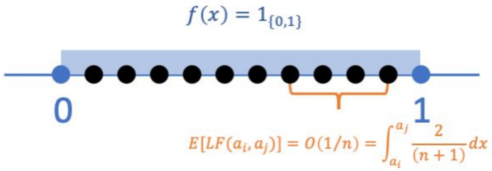
Claim 2: $E[G(x)] = \frac{2}{(n+1)f(x)} + o(1/n)$

=

Lemma: $E[\mathrm{LF}(\mathrm{a},\mathrm{b})] = \int_a^b \frac{2}{(n+1)f(x)}dx + o(1/n)$

Example:





Properties of leapfrog distance

A single connected support in 1D

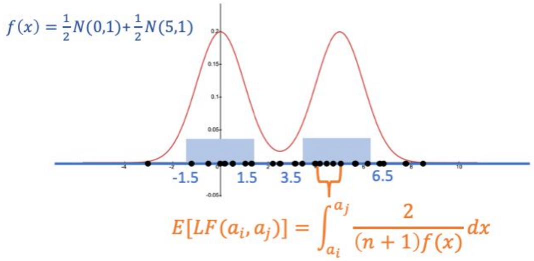
Claim 1: $\mathrm{LF}(a,b) = \int_a^b G(x)dx$

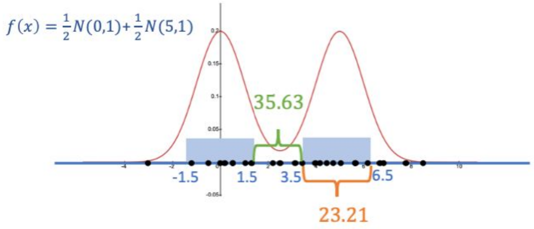
+

Claim 2: $E[G(x)] = \dfrac{2}{(n+1)f(x)} + o(1/n)$

=

Lemma: $E[\mathrm{LF(a,b)}] = \int_a^b \dfrac{2}{(n+1)f(x)} dx + o(1/n)$





Recovery of clusters

Sufficient conditions for recovery

Coordinates $b_1, b_2, \dots, b_n$

Consequences?

$$\min_{x_1,\dots,x_n \in \mathbb{R}^d} \frac{1}{2}\sum_{i=1}^{n} \|x_i - a_i\|^2 + \lambda \sum_{1 \le i < j \le n} \|x_i - x_j\|$$

Model

Theorem: Clusters are correctly recovered if for any cluster C, there holds

$$\max_{i,j \in C} \|b_i - b_j\|_2 \le \min_{i \in C,\, j \notin C} \|b_i - b_j\|_2 \cdot \min_m \frac{|C_m|}{2(n-1)}$$

Recovery of clusters

Sufficient conditions for recovery

Coordinates $b_1, b_2, \ldots, b_n$

Consequences?

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \frac{1}{2}\sum_{i=1}^{n} \|x_i - a_i\|^2 + \lambda \sum_{1 \le i<j \le n} \|x_i - x_j\|$$

Model

Theorem: Clusters are correctly recovered if for any cluster C, there holds

$$\boxed{\max_{i,j\in C} \|b_i - b_j\|_2} \le \min_{i\in C,\, j\notin C} \|b_i - b_j\|_2 \cdot \min_m \frac{|C_m|}{2(n-1)}$$

$\max(\text{intra-cluster distance})$

Recovery of clusters

Sufficient conditions for recovery

Coordinates $b_1, b_2, \ldots, b_n$

Consequences?

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \frac{1}{2}\sum_{i=1}^{n} \|x_i - a_i\|^2 + \lambda \sum_{1 \leq i<j \leq n} \|x_i - x_j\|$$

Model

Theorem: Clusters are correctly recovered if for any cluster C, there holds

$$\boxed{\max_{i,j\in C} \|b_i - b_j\|_2} \leq \boxed{\min_{i \in C,\, j \notin C} \|b_i - b_j\|_2} \cdot \min_m \frac{|C_m|}{2(n-1)}$$

$\max(\text{intra-cluster distance})$ $\min(\text{inter-cluster distance})$

Recovery of clusters

Sufficient conditions for recovery

Coordinates $b_1, b_2, \ldots, b_n$

$\Downarrow$ Consequences?

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \frac{1}{2}\sum_{i=1}^{n} \|x_i - a_i\|^2 + \lambda \sum_{1 \le i < j \le n} \|x_i - x_j\|$$

Model

Theorem: Clusters are correctly recovered if for any cluster C, there holds

$$\boxed{\max_{i,j \in C} \|b_i - b_j\|_2} \le \boxed{\min_{i \in C,\, j \notin C} \|b_i - b_j\|_2} \cdot \boxed{\min_m \frac{|C_m|}{2(n-1)}}$$

$\max(\text{intra-cluster distance})$ $\quad$ $\min(\text{inter-cluster distance})$ $\quad$ $\min\left(\frac{\text{number of points in a cluster}}{2(n-1)}\right)$

Recovery of clusters

Recovery of disjoint, compact supports

Coordinates $b_1, b_2, \ldots, b_n$

⬇ Consequences?

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \frac{1}{2}\sum_{i=1}^{n} \|x_i - a_i\|^2 + \lambda \sum_{1 \le i<j \le n} \|x_i - x_j\|$$

Model

❖ If the underlying law is supported on disjoint, compact sets

➢ $i,\, j \in C \iff \|b_i - b_j\|_2 \to 0$

➢ $i,\, j \notin C \iff \|b_i - b_j\|_2 \to c > 0$

Recovery of clusters

Recovery of disjoint, compact supports

Coordinates $b_1, b_2, \ldots, b_n$

Consequences?

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \frac{1}{2}\sum_{i=1}^{n} \|x_i - a_i\|^2 + \lambda \sum_{1 \le i<j \le n} \|x_i - x_j\|$$

Model

❖ If the underlying law is supported on disjoint, compact sets

$$\boxed{\max_{i,j\in C} \|b_i - b_j\|_2} \le \boxed{\min_{i\in C,\, j\notin C} \|b_i - b_j\|_2} \cdot \boxed{\min_m \frac{|C_m|}{2(n-1)}}$$

$$\max(\text{intra-cluster distance}) \le \min(\text{inter-cluster distance}) \bullet \min\left(\frac{\text{number of points in a cluster}}{2(n-1)}\right)$$

Recovery of clusters

Recovery of disjoint, compact supports

Coordinates $b_1, b_2, \ldots, b_n$

Consequences?

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \frac{1}{2}\sum_{i=1}^{n} \|x_i - a_i\|^2 + \lambda \sum_{1 \le i<j \le n} \|x_i - x_j\|$$

Model

- If the underlying law is supported on disjoint, compact sets

$$\boxed{\max_{i,j\in C} \|b_i - b_j\|_2} \le \boxed{\min_{i\in C,\, j\notin C} \|b_i - b_j\|_2} \cdot \boxed{\min_m \frac{|C_m|}{2(n-1)}}$$

$$\text{max(intra-cluster distance)} \le \text{min(inter-cluster distance)} \bullet \min\left(\frac{\text{number of points in a cluster}}{2(n-1)}\right)$$

$$\|b_i - b_j\|_2 \to 0$$

Recovery of clusters

Recovery of disjoint, compact supports

Coordinates $b_1, b_2, \ldots, b_n$

Consequences?

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \frac{1}{2}\sum_{i=1}^{n} \|x_i - a_i\|^2 + \lambda \sum_{1 \leq i<j \leq n} \|x_i - x_j\|$$

Model

❖ If the underlying law is supported on disjoint, compact sets

$$\boxed{\max_{i,j\in C} \|b_i - b_j\|_2} \leq \boxed{\min_{i \in C,\, j\notin C} \|b_i - b_j\|_2} \cdot \boxed{\min_m \frac{|C_m|}{2(n-1)}}$$

$$\text{max(intra-cluster distance)} \leq \text{min(inter-cluster distance)} \bullet \min\left(\frac{\text{number of points in a cluster}}{2(n-1)}\right)$$

$\|b_i - b_j\|_2 \to 0$

$\|b_i - b_j\|_2 \to c > 0$

Recovery of clusters

Recovery of disjoint, compact supports

Coordinates $b_1, b_2, \ldots, b_n$

Consequences?

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \frac{1}{2}\sum_{i=1}^{n} \|x_i - a_i\|^2 + \lambda \sum_{1 \le i < j \le n} \|x_i - x_j\|$$

Model

❖ If the underlying law is supported on disjoint, compact sets

$$\boxed{\max_{i,j \in C} \|b_i - b_j\|_2} \le \boxed{\min_{i \in C,\, j \notin C} \|b_i - b_j\|_2} \cdot \boxed{\min_m \frac{|C_m|}{2(n-1)}}$$

$$\max(\text{intra-cluster distance}) \le \min(\text{inter-cluster distance}) \bullet \min\left(\frac{\text{number of points in a cluster}}{2(n-1)}\right)$$

$\|b_i - b_j\|_2 \to 0$

$\|b_i - b_j\|_2 \to c > 0$

$\min(\text{mixture probability } p_i)$

Positive constant, independent of n

Coordinates $b_1, b_2, \ldots, b_n$

Consequences?

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \frac{1}{2}\sum_{i=1}^{n} \|x_i - a_i\|^2 + \lambda \sum_{1 \le i<j \le n} \|x_i - x_j\|$$

Model

Recovery of clusters

Recovery of disjoint, compact supports

❖ If the underlying law is supported on disjoint, compact sets

$$0 \leftarrow \boxed{\max_{i,j \in C} \|b_i - b_j\|_2} \le \boxed{\min_{i \in C,\, j \notin C} \|b_i - b_j\|_2} \cdot \boxed{\min_m \frac{|C_m|}{2(n-1)}} \rightarrow c > 0$$

$$\text{max(intra-cluster distance)} \le \text{min(inter-cluster distance)} \bullet \min\left(\frac{\text{number of points in a cluster}}{2(n-1)}\right)$$

$\|b_i - b_j\|_2 \rightarrow 0$

$\|b_i - b_j\|_2 \rightarrow c > 0$

$\min(\text{mixture probability } p_i)$

Positive constant, independent of n

Recovery of clusters

Recovery of a mixture of Gaussians in 1D

Coordinates $b_1, b_2, \ldots, b_n$

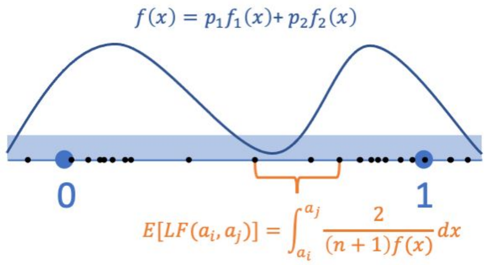
⬇ Consequences?

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \frac{1}{2}\sum_{i=1}^{n} \|x_i - a_i\|^2 + \lambda \sum_{1 \leq i < j \leq n} \|x_i - x_j\|$$

Model

❖ If the underlying law is a mixture of Gaussians supported on a single connected set in 1D

$$|b_i - b_j| = \mathrm{LF}(a_i, a_j)$$



Recovery of clusters

Recovery of a mixture of Gaussians in 1D

Coordinates $b_1, b_2, \ldots, b_n$

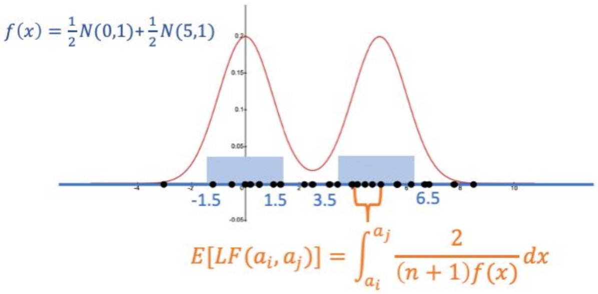
Consequences?

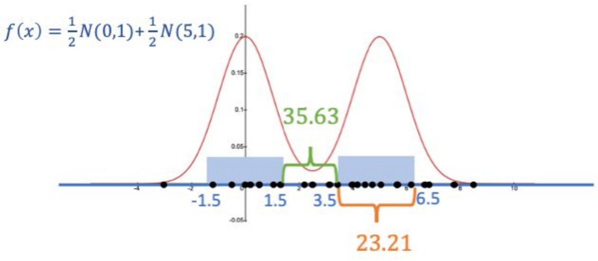
$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \frac{1}{2} \sum_{i=1}^{n} \|x_i - a_i\|^2 + \lambda \sum_{1 \le i < j \le n} \|x_i - x_j\|$$

Model

- If the underlying law is a mixture of Gaussians supported on a single connected set in 1D

$$|b_i - b_j| = \mathrm{LF}(a_i, a_j)$$





Recovery of clusters

Recovery of disjoint, compact supports

Coordinates $b_1, b_2, \ldots, b_n$

$\Downarrow$ Consequences?

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \frac{1}{2}\sum_{i=1}^{n} \|x_i - a_i\|^2 + \lambda \sum_{1 \le i < j \le n} \|x_i - x_j\|$$

Model

❖ If the underlying law is a mixture of Gaussians supported on a single connected set in 1D

$$\boxed{\max_{i,j \in C} \|b_i - b_j\|_2} \le \boxed{\min_{i \in C,\, j \notin C} \|b_i - b_j\|_2} \cdot \boxed{\min_m \frac{|C_m|}{2(n-1)}}$$

$$\text{max(intra-cluster distance)} \le \text{min(inter-cluster distance)} \bullet \min\left(\frac{\text{number of points in a cluster}}{2(n-1)}\right)$$

Recovery of clusters

Recovery of disjoint, compact supports

Coordinates $b_1, b_2, \ldots, b_n$

Consequences?

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \frac{1}{2}\sum_{i=1}^{n} \|x_i - a_i\|^2 + \lambda \sum_{1 \le i<j \le n} \|x_i - x_j\|$$

Model

❖ If the underlying law is a mixture of Gaussians supported on a single connected set in 1D

$$\boxed{\max_{i,j\in C} \|b_i - b_j\|_2} \le \boxed{\min_{i\in C,\, j\notin C} \|b_i - b_j\|_2} \cdot \boxed{\min_m \frac{|C_m|}{2(n-1)}}$$

$$\max(\text{intra-cluster distance}) \le \min(\text{inter-cluster distance}) \bullet \min\left(\frac{\text{number of points in a cluster}}{2(n-1)}\right)$$

$$\|b_i - b_j\|_2 \approx \frac{c}{f(x)}$$

for some x close to the mean

Recovery of clusters

Recovery of disjoint, compact supports

Coordinates $b_1, b_2, \ldots, b_n$

Consequences?

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \frac{1}{2}\sum_{i=1}^{n} \|x_i - a_i\|^2 + \lambda \sum_{1 \le i<j \le n} \|x_i - x_j\|$$

Model

❖ If the underlying law is a mixture of Gaussians supported on a single connected set in 1D

$$\boxed{\max_{i,j\in C} \|b_i - b_j\|_2} \le \boxed{\min_{i \in C,\, j\notin C} \|b_i - b_j\|_2} \cdot \boxed{\min_m \frac{|C_m|}{2(n-1)}}$$

$$\max(\text{intra-cluster distance}) \le \min(\text{inter-cluster distance}) \bullet \min\left(\frac{\text{number of points in a cluster}}{2(n-1)}\right)$$

$$\|b_i - b_j\|_2 \approx \frac{c}{f(x)}$$

for some x close to the mean

$$\|b_i - b_j\|_2 \approx \frac{c'}{f(x')}$$

for some x' far from the mean

Recovery of clusters

Recovery of disjoint, compact supports

Coordinates $b_1, b_2, \ldots, b_n$

Consequences?

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \frac{1}{2}\sum_{i=1}^{n} \|x_i - a_i\|^2 + \lambda \sum_{1 \leq i<j \leq n} \|x_i - x_j\|$$

Model

- ❖ If the underlying law is a mixture of Gaussians supported on a single connected set in 1D

$$\boxed{\max_{i,j\in C} \|b_i - b_j\|_2} \leq \boxed{\min_{i\in C,\, j\notin C} \|b_i - b_j\|_2} \cdot \boxed{\min_m \frac{|C_m|}{2(n-1)}}$$

$$\max(\text{intra-cluster distance}) \leq \min(\text{inter-cluster distance}) \bullet \min\left(\frac{\text{number of points in a cluster}}{2(n-1)}\right)$$

$$\|b_i - b_j\|_2 \approx \frac{c}{f(x)}$$

for some x close to the mean

$$\|b_i - b_j\|_2 \approx \frac{c'}{f(x')}$$

for some x' far from the mean

$\min(\text{mixture probability } p_i)$

Positive constant, independent of n

Recovery of clusters

Recovery of disjoint, compact supports

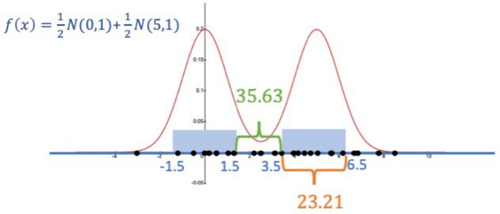
Coordinates $b_1, b_2, \ldots, b_n$

Consequences?

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \frac{1}{2}\sum_{i=1}^{n} \|x_i - a_i\|^2 + \lambda \sum_{1 \le i<j \le n} \|x_i - x_j\|$$

Model

❖ If the underlying law is a mixture of Gaussians supported on a single connected set in 1D



$$\boxed{\max_{i,j\in C} \|b_i - b_j\|_2} \le \boxed{\min_{i \in C,\, j \notin C} \|b_i - b_j\|_2} \cdot \boxed{\min_m \frac{|C_m|}{2(n-1)}}$$

$$\max(\text{intra-cluster distance}) \le \min(\text{inter-cluster distance}) \bullet \min\left(\frac{\text{number of points in a cluster}}{2(n-1)}\right)$$

$$\|b_i - b_j\|_2 \approx \frac{c}{f(x)}$$

for some x close to the mean

$$\|b_i - b_j\|_2 \approx \frac{c'}{f(x')}$$

for some x’ far from the mean

$\min(\text{mixture probability } p_i)$

Positive constant, independent of n

Recovery of clusters

Recovery of disjoint, compact supports

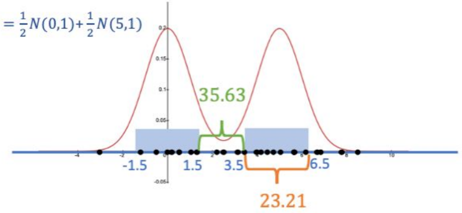
Coordinates $b_1, b_2, \ldots, b_n$

Consequences?

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \frac{1}{2}\sum_{i=1}^{n} \|x_i - a_i\|^2 + \lambda \sum_{1 \le i<j \le n} \|x_i - x_j\|$$

Model

- If the underlying law is a mixture of Gaussians supported on a single connected set in 1D



$$\boxed{\max_{i,j \in C} \|b_i - b_j\|_2} \le \boxed{\min_{i \in C,\, j \notin C} \|b_i - b_j\|_2} \cdot \boxed{\min_m \frac{|C_m|}{2(n-1)}} \to c \max_{i,j \in C} \|b_i - b_j\|,\ c > 1$$

$$\text{max(intra-cluster distance)} \le \text{min(inter-cluster distance)} \bullet \min\left(\frac{\text{number of points in a cluster}}{2(n-1)}\right)$$

$$\|b_i - b_j\|_2 \approx \frac{c}{f(x)}$$

for some x close to the mean

$$\|b_i - b_j\|_2 \approx \frac{c'}{f(x')}$$

for some x' far from the mean

$\min(\text{mixture probability}\, p_i)$

Positive constant, independent of n

Recovery of clusters

Summary

Coordinates $b_1, b_{2,} \ldots, b_n$

$\Downarrow$ Consequences?

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \frac{1}{2}\sum_{i=1}^{n} \|x_i - a_i\|^2 + \lambda \sum_{1 \le i < j \le n} \|x_i - x_j\|$$

Model

Recovery of clusters

Summary

Coordinates $b_1, b_2, \ldots, b_n$

Consequences?

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \frac{1}{2}\sum_{i=1}^{n} \|x_i - a_i\|^2 + \lambda \sum_{1 \leq i<j \leq n} \|x_i - x_j\|$$

Model

❖ If the underlying law is supported on disjoint, compact sets: perfect recovery because

$$0 \leftarrow \max_{i,j \in C} \|b_i - b_j\|_2 \leq \min_{i \in C,\, j \notin C} \|b_i - b_j\|_2 \cdot \min_m \frac{|C_m|}{2(n-1)} \rightarrow c > 0$$

Recovery of clusters

Summary

Coordinates $b_1, b_2, \dots, b_n$

$\Downarrow$ Consequences?

$$\min_{x_1,\dots,x_n \in \mathbb{R}^d} \frac{1}{2}\sum_{i=1}^{n} \|x_i - a_i\|^2 + \lambda \sum_{1 \le i<j \le n} \|x_i - x_j\|$$

Model

- If the underlying law is supported on disjoint, compact sets: perfect recovery because

$$0 \leftarrow \max_{i,j \in C} \|b_i - b_j\|_2 \le \min_{i \in C,\, j \notin C} \|b_i - b_j\|_2 \cdot \min_m \frac{|C_m|}{2(n-1)} \quad \rightarrow c > 0$$

- If the underlying law is a mixture of Gaussians supported on a single connected set in 1D: perfect recovery for points close to the means

$$\max_{i,j \in C} \|b_i - b_j\|_2 \le \min_{i \in C,\, j \notin C} \|b_i - b_j\|_2 \cdot \min_m \frac{|C_m|}{2(n-1)} \quad \rightarrow c \max_{i,j \in C} \|b_i - b_j\|,\ c > 1$$