Embedding Hamming-metric of Boxes into L_2

Avner Magen

University of Toronto

Hamming-Distance of of sets in R^D

Let $\mathcal{F}$ be a family of sets in $\mathbf{R}^D$. There is a metric-space that is associated with $\mathcal{F}$: For $A, B \in \mathcal{F}$

$d(A, B) = HAMMING(A, B) = \text{vol}(A \backslash B \cup B \backslash A)$

Examples:

1. all measurable sets on the line.

2. intervals on the line.

3. convex bodies in $\mathbf{R}^D$.

4. axis parallel boxes in $\mathbf{R}^D$.

It is L_1. What about L_2?

$L_p(\mathbf{R}^D)$ is the space of continuous functions $f : \mathbf{R}^D \mapsto \mathbf{R}$, and finite norm $\|f\|_p = \left(\int_{\mathbf{R}^D} |f(x)|^p dx\right)^{\frac{1}{p}}$. The Hamming distance is L_1 : take a set A to the characteristic function χ_A, then

$$d(A, B) = \int |\chi_A(x) - \chi_B(x)| dx$$

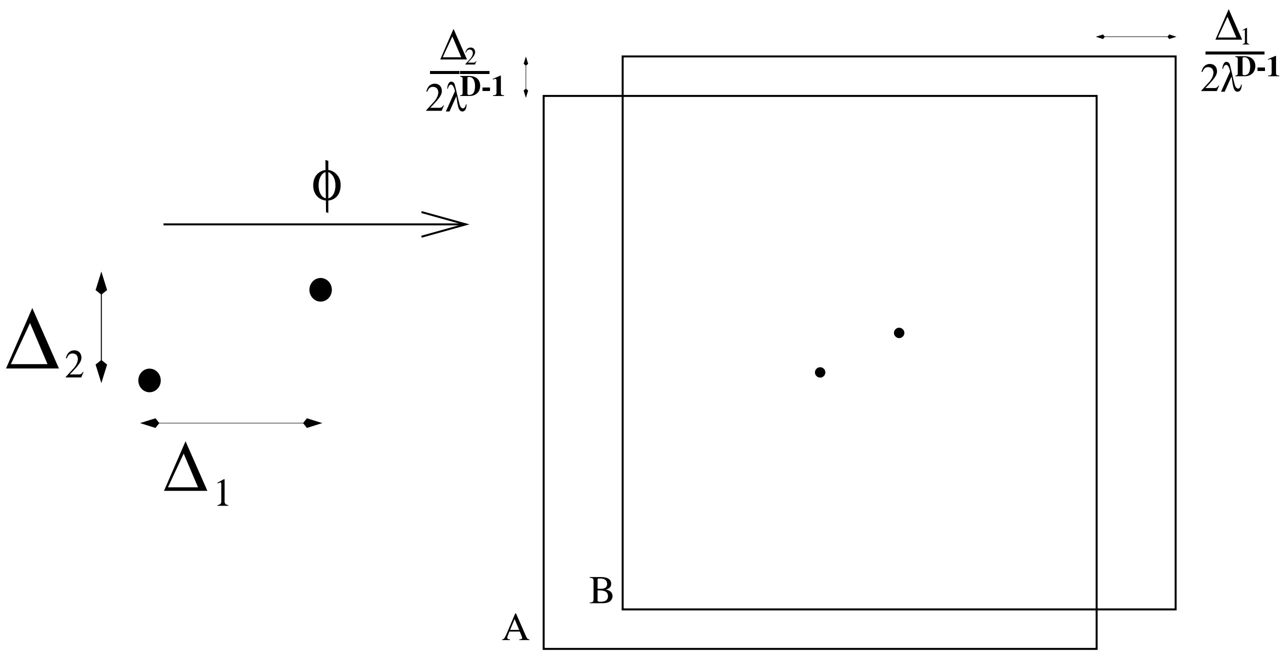
What about the Hilbert space L_2? Here $\|f\| = \sqrt{\int |f(x)|^2 dx}$.

The result

Theorem : There is an embedding from the space $\mathcal{B}_D$ of axis-parallel boxes in $\mathbf{R}^D$ (with Hamming distance) into L_2 with distortion $O(\sqrt{D})$. This is asymptotically optimal.

Viewing the $\mathcal{B}_D$ as an extension of ℓ_1^D

Claim : Every bounded region in ℓ_1^D embeds with distortion 1 in $\mathcal{B}_D$.
Proof : Map $x \in \ell_1^D$ to the box $\frac{x}{2\lambda^{D-1}} + [-\frac{1}{2}\lambda, \frac{1}{2}\lambda]^D$.
As λ goes to infinity the distortion goes to 1.



d(A,B) $=\Delta_1 + \Delta_2 + O(1/\lambda)$

But $\mathcal{B}_D$ is much richer than ℓ_1^D. In fact even $\mathcal{B}_1$ cannot embed in ℓ_1^D with any finite distortion, since it contain the metric of K_n for any n, which cannot embed well into ℓ_1^D.

Optimality of the embedding

Corollary : The distortion needed in embedding $\mathcal{B}_D$ into L_2 is at least $\sqrt{D}$). This is since we have shown that the D-dim hamming cube is a subspace of $\mathcal{B}_D$, and since its Euclidean distrotin is $\sqrt{D}$.

Interesting special cases - truncated metrics

- For D constant we get a constant distortion embedding of the box space into L_2.

- For $D = 1$ and considering intervals of the same length $M/2$ we get the metric

$$d(i, j) = \min(|i - j|, M)$$

  (aka **linear-truncated metric**). We can now claim it is almost L_2.

- This metric found use in the context of *metric-labeling-problem* [GT00,CKNZ01].

Truncated metrics - some context

- Schoenberg Showed in 38' that if

$$w(t) = \sqrt{1 - \exp(-ut^2)}$$

then the metric on a Hilbert-Space $d(x,y) = w(\|x - y)\|)$ is isometrically embeddable in L_2. It can be shown that the function of interest $f(t) = \min(t, M)$ is approximated up to a constant by $w(t)$. As a corollary we get that "truncated Euclidean metrics" are also almost Euclidean.

- Recently Mendel and Naor gave a direct proof of the above with an embedding that generalizes to any ℓ_p.

The Embedding of $\mathcal{B}_D$ into L_2

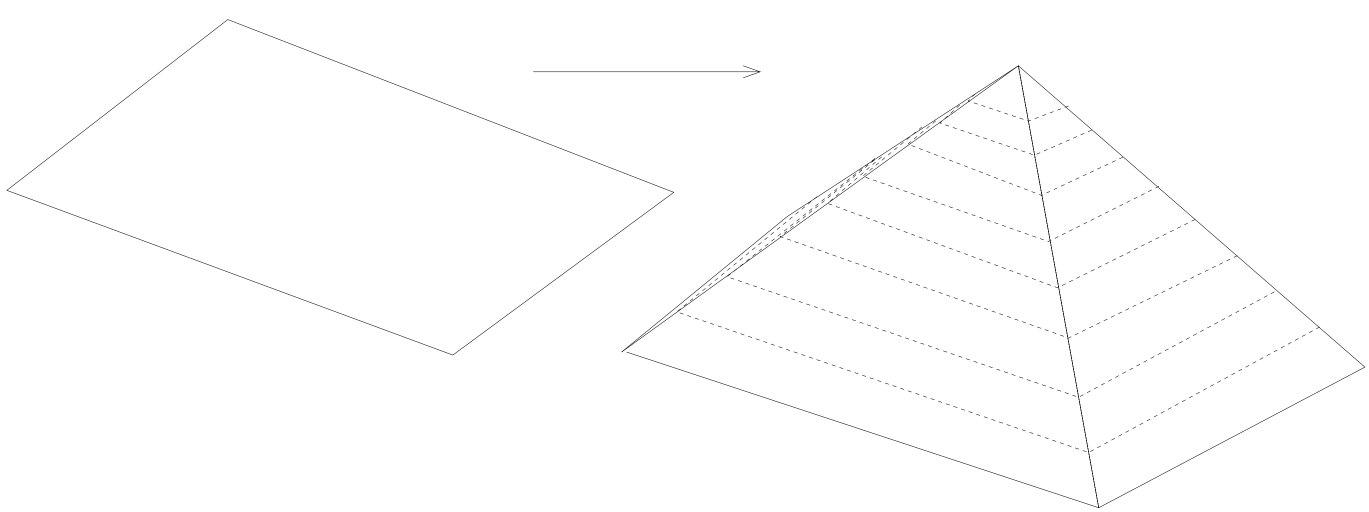
Let A be the box

$$[c_1 - a_1, c_1 + a_1] \times \ldots \times [c_D - a_D, c_D + a_D].$$

Map A to the function f_A.
f_A is supported on A so for $x \notin A, f_A(x) = 0$.
For $x \in A$,

$$f_A(x) = s(x) \cdot \sqrt{D \cdot \text{vol}(A)}$$

where $s(x) = 1 - \max_i \frac{|x_i - c_i|}{a_i}$ is the (normalized) distance of $x \in A$ from ∂A.



Distortion $O(\sqrt{D})$ of the Embedding - Overview of Proof

We show that for any pair of boxes A, B,

$$\Omega(1/\sqrt{D}) \cdot d(A,B) \leq \|f_A - f_B\| \leq O(1) \cdot d(A,B)$$

- Showing first the small distortion of two extreme configurations:
  (i) One of the boxes is $\emptyset$.
  (ii) Almost-identical boxes.

- Showing constant **expansion** for any pair of boxes, using the second case above.

- Showing a **contraction** of $O(\sqrt{D})$
  - when one box is contained in the other.
  - the general case.

Scalability

We notice that the embedding is scalable: if T is a linear transformation then

$$\|f_{TA} - f_{TB}\| = \text{det } T \cdot \|f_A - f_B\|$$

Since

$$\int (f_{TA}(x) - f_{TB}(x))^2 dx =$$

$$\int \left(f_A(y) \sqrt{\frac{\text{vol}(TA)}{\text{vol}(A)}} - f_B(y) \sqrt{\frac{\text{vol}(TB)}{\text{vol}(B)}} \right)^2 \text{det } T \, dy$$

$$= \int (f_A(y) - f_B(y))^2 (\text{det } T)^2 dy.$$

Moreover

$$d(TA, TB) = \text{det } T \cdot d(A, B)$$

so the expansion factor of a pair of boxes $\frac{\|f_A - f_B\|}{d(A,B)}$ does not change under a linear transformation.

small distortion of the distance from $\emptyset$

By scalability, may assume $A = [-1,1]^D$. Let $V = \text{vol}(A)$.

$$\|f_A - f_\emptyset\|^2 = \int f_A(x)^2 dx =$$

$$DV \int_{s=0}^{1} s^2 \cdot D2^D(1-s)^{D-1} ds =$$

$$D^2V^2 \int_{s=0}^{1} s^2(1-s)^{D-1} ds = D^2 2^D B(3,D) =$$

$$D^2V^2 \cdot \frac{2!(D-1)!}{(D+2)!} = \frac{2D}{(D+1)(D+2)} \cdot V^2.$$

Here, $\mathbf{B(m,n)} = \int_{u=0}^{1} u^{m-1}(1-u)^{n-1} du = \frac{(m-1)!(n-1)!}{(m+n-1)!}$ is the Euler Beta function.

So $\|f_A - f_\emptyset\| = \Theta(V/\sqrt{D}) = \Theta(1/\sqrt{D}) \cdot d(A, \emptyset)$.

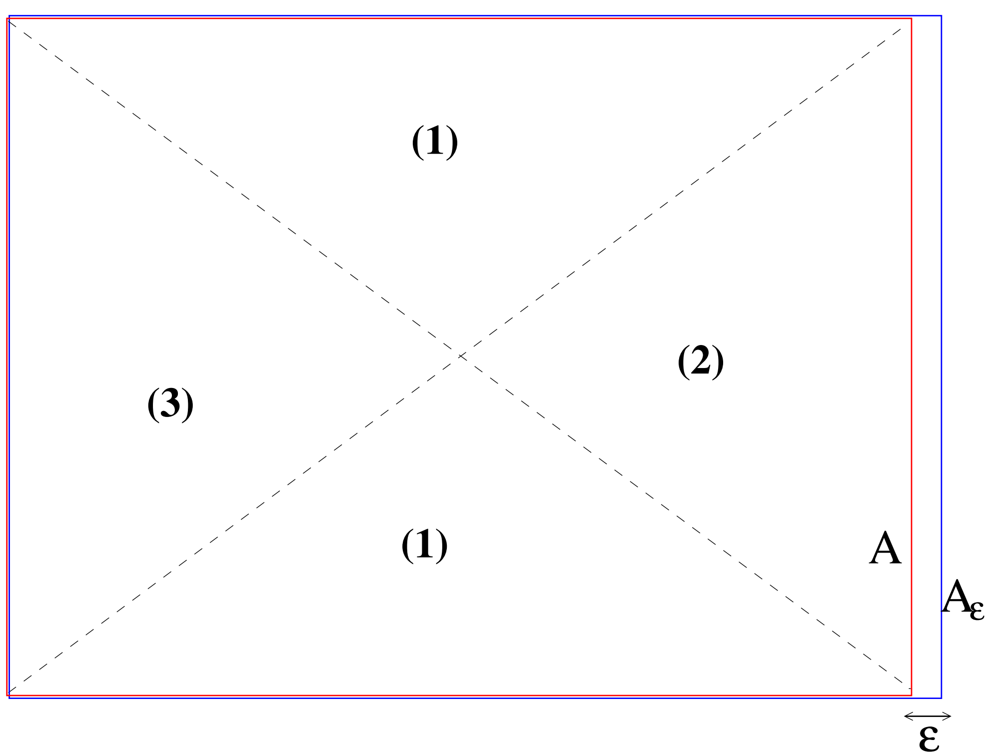
small distortion in almost-indentical boxes

almost-identical boxes are pair of boxes that share a face and are different by at most a factor of ε in the nonidentical axis. Assume $A = [-1, 1]^D$, and $A_\varepsilon = [-1, 1+\varepsilon] \times [-1, 1]^{D-1}$. Three regions to consider:

(1) $|x_1| < |x_j|$ for some j (measure $V(1 - \frac{1}{D})$)

(2) x_1 biggest of all $|x_i|$ (measure $\frac{V}{2D}$)

(3) $-x_1$ biggest of all $|x_i|$ (measure $\frac{V}{2D}$)



small distortion in almost-identical boxes

Can check that $|f_A(x) - f_{A_\varepsilon}(x)|$ is $\Theta(\varepsilon f_A(x))$, $\Theta(\varepsilon\sqrt{DV})$ and $O(\varepsilon\sqrt{DV})$ in the three regions respectively. Therefore

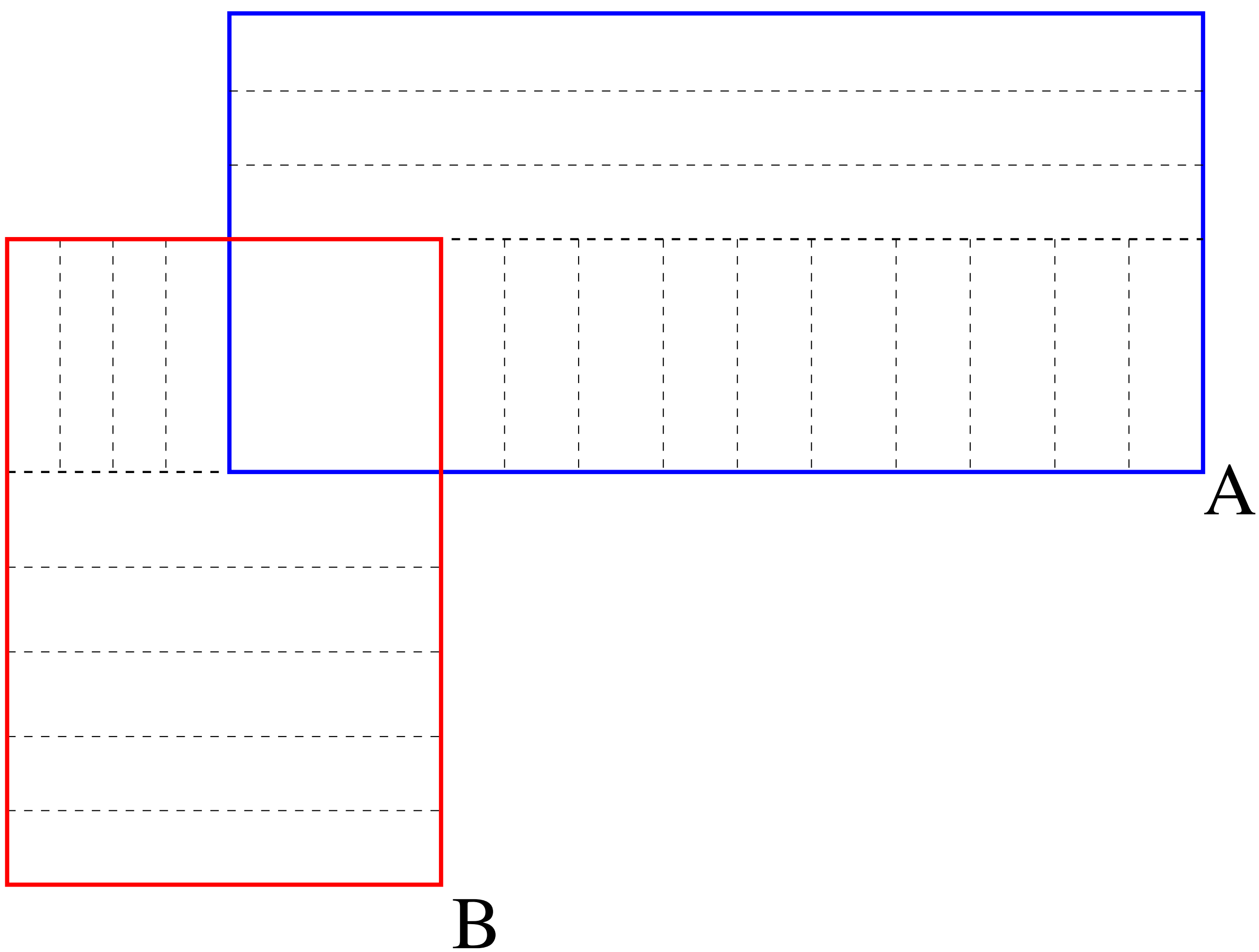
$$\int |f_A(x) - f_{A_\varepsilon}(x)|^2 dx = \Theta\left(\varepsilon^2 \frac{V^2}{D}\right) + \Theta\left(\varepsilon^2 DV \cdot \frac{V}{D}\right)$$

and $\|f_A - f_{A_\varepsilon}\| = \Theta(\varepsilon V) = \Theta(d(A, A_\varepsilon))$.

(Notice that $d(A, A_\varepsilon) = \varepsilon \cdot 2^{D-1} = \varepsilon V/2$)

small expansion in almost-identical boxes
$\rightarrow$ small expansion to any pair of boxes

Between any two boxes there is a geodesic path (via their intersection) where each consecutive pair of boxes is almost-identical, or one of the boxes is empty.
The expansion is no more than the expansion in any of these pairs.



small contraction of contained boxes

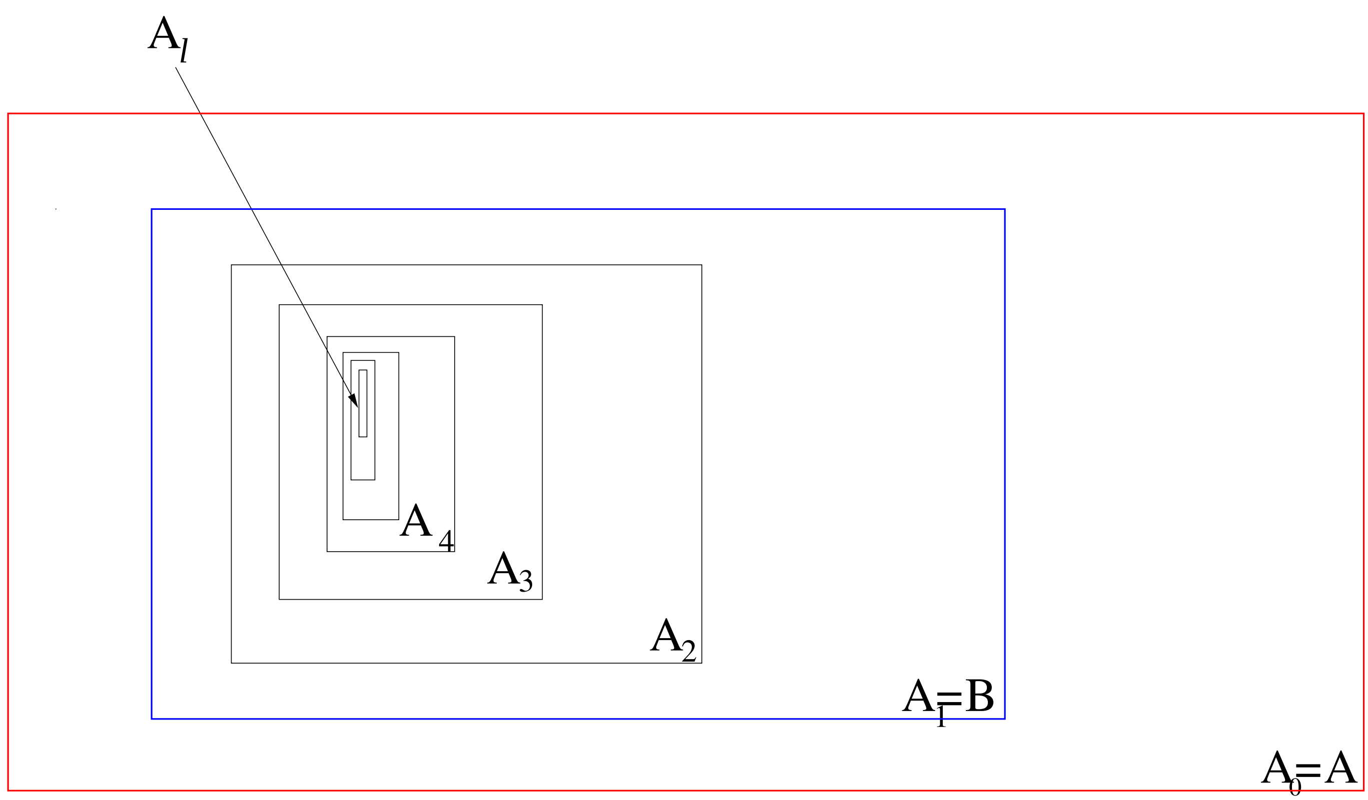
Suppose $B \subset A$. Need to show that

$$\frac{\|f_A - f_B\|}{d(A,B)} \geq 1/\sqrt{D}$$

Let $TA = B$. Look at the decreasing chain

$$A = A_0 \supset B = A_1 \supset A_2 \supset \ldots \supset A_l,$$

where $A_i = T^i A$. .



Now,

$d(A_0, A_l) = d(A_0, A_1) + d(A_1, A_2) + \ldots + d(A_{l-1}, A_l)$.

small contraction of contained boxes

By the geodesic property

$$\frac{\|f_A - f_{A_l}\|}{d(A, A_l)} \leq \max_i \frac{\|f_{A_i} - f_{A_{i-1}}\|}{d(A_i, A_{i-1})}$$

And by scalability

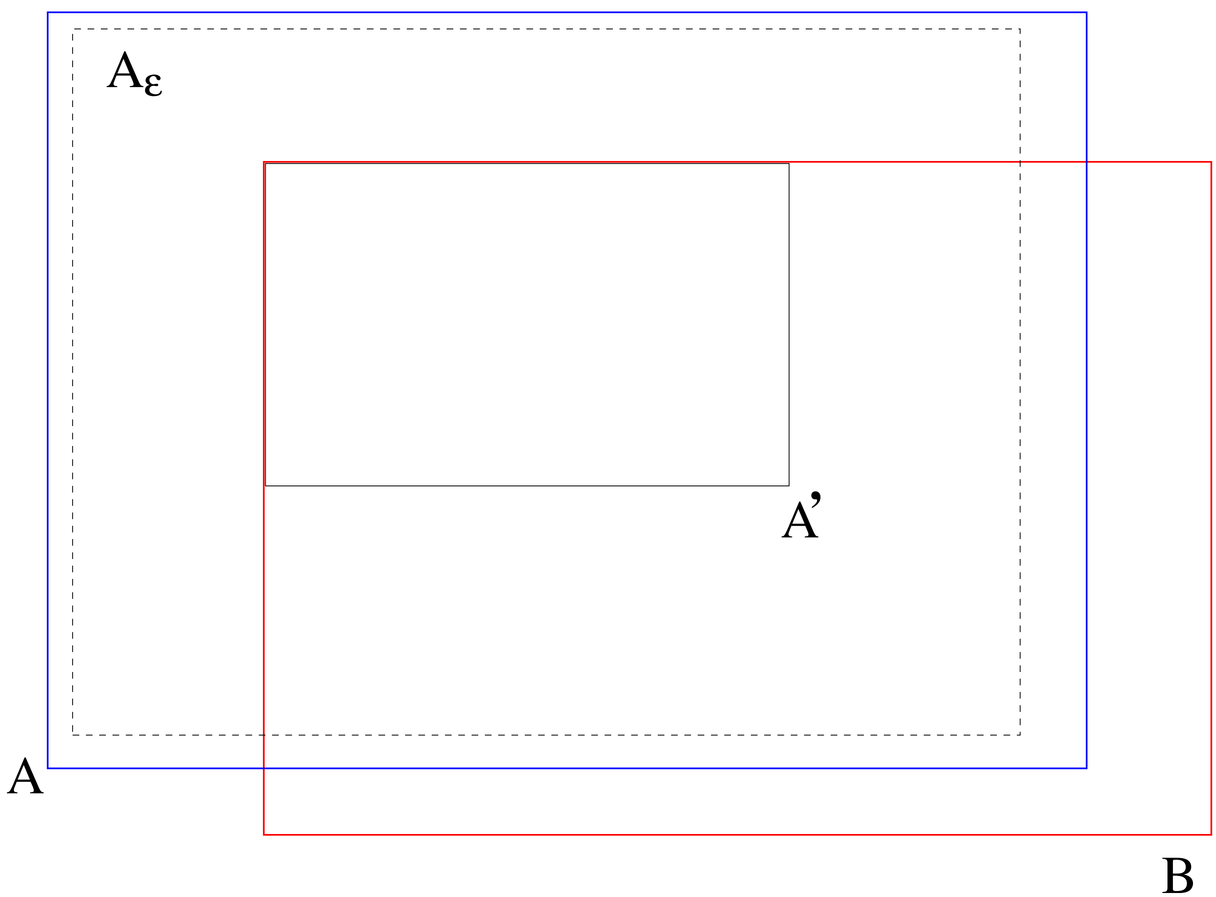
$$\max_i \frac{\|f_{A_i} - f_{A_{i-1}}\|}{d(A_i, A_{i-1})} = \frac{\|f_A - f_B\|}{d(A, B)}$$

So

$$\frac{\|f_A - f_B\|}{d(A, B)} \geq \frac{\|f_A - f_{A_l}\|}{d(A, A_l)} \longrightarrow \frac{\|f_A - f_\emptyset\|}{d(A, \emptyset)} = \Omega(1/\sqrt{D})$$

Small contraction for the general case

If vol(A) $> 2 \cdot$ vol($A \cap B$) can show small contraction by direct integration (like the $\emptyset$ case). Otherwise can transform $A \not\subset B$ to $A' \subset B$ such that $d(A', B) = d(A, B)$. It is enough to show that $\|f_{A'} - f_B\| \leq \|f_A - f_B\|$.
We show this locally by moving from A to A_ε.
claim $\|f_{A_\varepsilon} - f_B\| \leq \|f_A - f_B\|$.



Further work

- Other natural families in $\mathbf{R}^D$.

- Consider the family of sets on the line which consists of at most k intervals. The L_2 distortion of this space is $\Omega(\sqrt{k})$ since the hamming cube $\{0,1\}^k$ is a subspace: To map a vector in $\{0,1\}^k$ take the intervals between $i-1$ and i when the i-th coordinate is 1.
  Conjecture : This is tight. Namely there is an embedding of this space into L_2 with distortion $O(\sqrt{k})$.
  Motivation : All finite ℓ_1 metrics can be regarded as subsets of the cube. k can therefore be regarded as a parameter for ℓ_1-metrics, and if the conjecture is true this parameter indicates quality of Eucldiean embeddability.