Ultrametricity indices for the Euclidean and Boolean hypercubes

Patrick Erik Bradley

Institute of Photogrammetry and Remote Sensing (IPF)
Karlsruhe Institute of Technology (KIT)

International conference on *p*-ADIC MATHEMATICAL PHYSICS AND ITS APPLICATIONS
Belgrade, September 07-12, 2015

Introduction

Murtagh observed experimentally:

- Samples which are *sparse* and *random* in $[0,1]^N$ or $\mathbb{F}_2^N$ become more and more ultrametric as $N \to \infty$

His ultrametricity coefficient:

- fraction of triangles which are approximately *isosceles with short base* $(=$ *ultrametric*$)$

1. Ultrametricity indices

Let (X, d) be a finite metric space.

Murtagh:

- $m(X, d) := \frac{\#\ \text{ultrametric}\ \Delta}{\#\ \text{all}\ \Delta}$

topological:

- $t(X, d) := \frac{1}{\text{diam}(X)} \int\limits_0^{\text{diam}(X)} \mu(\Gamma_\epsilon)\, d\epsilon$

2. Topological Ultrametricity Index

Let $\epsilon > 0$. *Vietoris-Rips graph* Γ_ϵ for (X, d):

- Vertices: X
- Edge: (x, y) with $d(x, y) \leq \epsilon$.

Lemma

(X, d) is ultrametric $\Leftrightarrow$ all connected components of all Γ_ϵ are complete.

2. Topological Ultrametricity Index

Let Γ be a finite graph.

- $b_0(\Gamma) :=$ #connected components of Γ
- $c(\Gamma) :=$ #maximal cliques of Γ
- $\mu(\Gamma) := \frac{b_0(\Gamma)}{c(\Gamma)}$
- $t(X, d) = \frac{1}{\mathrm{diam}(X)} \int_0^{\mathrm{diam}(X)} \mu(\Gamma_\epsilon)\, d\epsilon$

  topological ultrametricity index

2. Topological Ultrametricity Index

The subdominant ultrametric

- $\bar{d}(x, y) = \min \{\epsilon \mid x, y \in \text{same connected component of } \Gamma_\epsilon\}$

Lemma

$\bar{d}$ is the subdominant ultrametric associated with d.

2. Topological Ultrametricity index

Let $0 < d_0 \leq \cdots \leq d_n$ be the pairwise positive distances between the points of X.

Lemma

$$t(X,d) = \sum_{i=0}^{n-1} \alpha_i \frac{d_i}{d_n} \quad \textit{with} \sum_{i=0}^{n-1} \alpha_i = 1,$$

$\alpha_0 \in [0,1]$, $\alpha_1, \ldots, \alpha_{n-2} \in (-1,1]$, $\alpha_{n-1} \in (0,1]$.

2. Topological Ultrametricity Index

Proof.

- $\mu(\Gamma_\epsilon) = \mu_{i+1}$ is constant for $d_i < \epsilon \leq d_{i+1}$
- $\mu_0 = \mu(\Gamma_\epsilon) = 1$ for $0 < \epsilon \leq d_0$

$$t(X,d) = \frac{1}{d_n}\left(\mu_0 d_0 + \sum_{i=1}^{n-1} \mu_i (d_i - d_{i-1})\right)$$
$$= \frac{1}{d_n}\left(\sum_{i=0}^{n-2} (\mu_i - \mu_{i+1}) d_i + \mu_{n-1} d_{n-1}\right)$$

with

$$\mu_{n-1} + \sum_{i=0}^{n-2} (\mu_i - \mu_{i+1}) = \mu_0 = 1$$

□

2. Topological Ultrametricity Index

Corollary

$t(X,d)$ is scale invariant: $t(X,d) = t(X, \sigma \cdot d)$ for $\sigma > 0$.

Proof.

$$t(X,d) = \sum_{i=0}^{n-1} \alpha_i \frac{d_i}{d_n} = \sum_{i=0}^{n-1} \alpha_i \frac{\sigma \cdot d_i}{\sigma \cdot d_n} = t(X, \sigma \cdot d)$$

□

3. Sparsity and Randomness

Consider $N \to \infty$.

- $(\mathbb{H}^N, d) := ([0,1]^N, d_E)$ or $(\mathbb{F}_2^N, d_H)$
- $d_E =$ Euclidean distance, $d_H =$ Hamming distance
- $X =$ finite random sample of $\mathbb{H}^N$ of fixed cardinality

Observation.
If $\frac{d_0}{d_n} \xrightarrow{\mathcal{P}} 1$, then $m(X,d), t(X,d) \xrightarrow{\mathcal{P}} 1$

3. Sparsity and Randomness

Theorem

Let X be uniformly distributed. Then $\frac{d_0}{d_n} \xrightarrow{\mathcal{P}} 1$.

3. Sparsity and Randomness

Proof.

Case $\mathbb{H} = [0,1]$. Consider for uniform iid x_i, y_i:

$$z_N = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2$$

$$z_N \xrightarrow{\mathcal{P}} \mathbb{E}(z_1) = \iint_{[0,1]^2} (x-y)^2 \, dxdy = \frac{1}{6}$$

$$\frac{d_0}{\sqrt{N}} = \min\left\{\sqrt{z_{N,0}}, \dots, \sqrt{z_{N,n}}\right\} \xrightarrow{\mathcal{P}} \frac{1}{\sqrt{6}}$$

$$\frac{d_n}{\sqrt{N}} = \max\left\{\sqrt{z_{N,0}}, \dots, \sqrt{z_{N,n}}\right\} \xrightarrow{\mathcal{P}} \frac{1}{\sqrt{6}}$$

$$\Rightarrow \qquad \frac{d_0}{d_n} = \frac{d_0/\sqrt{N}}{d_n/\sqrt{N}} \xrightarrow{\mathcal{P}} \frac{1/\sqrt{6}}{1/\sqrt{6}} = 1$$

3. Sparsity and Randomness

Case $\mathbb{H} = \mathbb{F}_2$.

- x, y uniform r.v. in $\mathbb{F}_2^N \Rightarrow \|x\|_H \sim B(N, \frac{1}{2})$
- Also, $d(x,y) = \|x+y\| \sim B(N, \frac{1}{2})$:

$$\mathbb{P}(\|x+y\| = k,\ x \text{ fixed}) = \frac{\binom{N}{k}}{4^N}$$

$$\mathbb{P}(\|x+y\| = k) = \sum_x \mathbb{P}(\|x+y\| = k,\ x \text{ fixed}) = 2^N \cdot \frac{\binom{N}{k}}{4^N} = \frac{\binom{N}{k}}{2^N}$$

- For normalised positive distances $\frac{d_H(x,y)}{N} > 0$:

$$\mathbb{E}\left(\frac{d_H(x,y)}{N}\right) = \frac{1}{2\left(1 - \frac{1}{2^N}\right)}$$

3. Sparsity and Randomness

- By Chebyshev inequality,

$$\mathbb{P}\left(\left|\frac{d_H(x,y)}{N} - \frac{1}{2\left(1-\frac{1}{2^N}\right)}\right| > \epsilon\right) \leq \frac{1}{\epsilon^2}\,\mathrm{Var}\left(\frac{d_H(x,y)}{N}\right)$$

$$= \frac{1}{\epsilon^2}\left(\frac{1}{4N\left(1-\frac{1}{2^N}\right)} + \frac{1}{4\left(1-\frac{1}{2^N}\right)} - \frac{1}{4\left(1-\frac{1}{2^N}\right)^2}\right)$$

$$\to 0 \quad (N \to \infty)$$

- This means $\frac{d_H(x,y)}{N} \xrightarrow{\mathcal{P}} \frac{1}{2}$
- Hence, $\frac{d_0}{N} \xrightarrow{\mathcal{P}} \frac{1}{2}$, $\frac{d_n}{N} \xrightarrow{\mathcal{P}} \frac{1}{2}$
- Hence, $\frac{d_0}{d_n} \xrightarrow{\mathcal{P}} 1$ □

3. Sparsity and Randomness

Theorem

Let $\mathbb{H} = [0,1]$, *and* $x \in X$ *with independent coordinates* $x_i \sim N(\mu_i, \sigma_i)$ *such that* $\sigma_i^2 \leq b$ *for all* i. *Then* $\frac{d_0}{d_n} \xrightarrow{\mathcal{P}} 1$.

3. Sparsity and Randomness

Proof.

$$z_N = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2$$

$$\mathbb{E}(z_N) = \frac{2}{N}\sum_{i=1}^{N}\sigma_i^2 \leq \frac{2N}{N}b = 2b$$

$$\operatorname{Var} z_N = \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left((x_i - y_i)^2\right) = \frac{1}{N}\mathbb{E}(z_N)$$

Chebyshev inequality:

$$\mathbb{P}(|z_N - \mathbb{E}(z_N)| > \epsilon) \leq \frac{\operatorname{Var} z_N}{\epsilon^2} = \frac{1}{N\epsilon^2}\mathbb{E}(z_N) \leq \frac{2b}{N\epsilon^2} \to 0$$

3. Sparsity and Randomness

- $\mathbb{E}(z_N) > 0$ is increasing and bounded $\Rightarrow \mathbb{E}(z_N) \to \zeta > 0$
- $\Rightarrow z_N \xrightarrow{\mathcal{P}} \zeta > 0$
- $\Rightarrow \frac{d_0}{d_n} \xrightarrow{\mathcal{P}} \frac{\sqrt{\zeta}}{\sqrt{\zeta}} = 1$ □

3. Sparsity and Randomness

Categorial data.

- X in complete disjunctive form
- $x = (\underbrace{0\ldots1\ldots0}_{k_1} \mid \cdots \mid \underbrace{0\ldots1\ldots0}_{k_\ell}) =: (x_1 \mid \cdots \mid x_\ell)$
- *elementary vector* x_i has precisely one 1-entry.
- $d(x_i, y_i) = 2\delta_{x_i,y_i}$
- $d(x,y) = \sum_{i=1}^{\ell} d(x_i, y_i)$
- $\mathbb{P}(d(x_i, y_i) = 2) = 1 - \frac{1}{k_i}$
- $\mathbb{E}\left(\frac{d(x,y)}{\ell}\right) = \frac{2}{\ell}\sum_{i=1}^{\ell}\left(1 - \frac{1}{k_i}\right) = 2\left(1 - \frac{1}{\ell}\sum_{i=1}^{\ell}\frac{1}{k_i}\right)$
- $\mathrm{Var}\left(\frac{d(x,y)}{\ell}\right) = \frac{1}{\ell}\sum_{i=1}^{\ell}\frac{4}{\ell}\left(\frac{1}{k_i} - \frac{1}{k_i^2}\right)$

3. Sparsity and Randomness

Categorial data.

Theorem

Let $k_i \geq 2$. *If* $\frac{1}{\ell}\sum_{i=1}^{\ell}\frac{1}{k_i}$ *converges for* $\ell \to \infty$, *then* $\frac{d_0}{d_n} \xrightarrow{\mathcal{P}} 1$.

3. Sparsity and Randomness

Proof.

Chebyshev inequality:

$$\mathbb{P}\left(\left|\frac{d(x,y)}{\ell} - \mathbb{E}\frac{d(x,y)}{\ell}\right| > \epsilon\right) \leq \frac{\operatorname{Var}\frac{d(x,y)}{\ell}}{\epsilon^2}$$

$$= \frac{1}{\epsilon^2}\cdot\frac{1}{\ell}\sum_{i=1}^{\ell}\frac{4}{\ell}\left(\frac{1}{k_i} - \frac{1}{k_i^2}\right) \to 0 \quad \text{(Cesàro means)}$$

If $\lim \frac{1}{\ell}\sum_{i=1}^{\ell}\frac{1}{k_i} = C$, then $C \leq \frac{1}{2}$ and

$$\frac{d(x,y)}{\ell} \xrightarrow{\mathcal{P}} 2(1-C) > 0$$

$$\Rightarrow \frac{d_0}{d_n} \xrightarrow{\mathcal{P}} \frac{2(1-C)}{2(1-C)} = 1$$

□

4. Ultrametricity Index of $\mathbb{F}_2^N$

- $m_N := m(\mathbb{F}_2^N, d_H)$
- $t_N := t(\mathbb{F}_2^N, d_H)$

Theorem

$$\frac{1}{N} < t_N < m_N < \frac{C}{\sqrt{N}}$$

for $N >> 0$ *with* $C > 0$. *In particular*

$$\lim_{N\to\infty} t_N = \lim_{N\to\infty} m_N = 0$$

4. Ultrametricity Index of $\mathbb{F}_2^N$

Proof that $t_N < \frac{2}{N}$.

- For $k \leq \epsilon < k+1$ each k-face of $\mathbb{F}_2^N$ is a maximal clique of Γ_ϵ
- Γ_ϵ is connected for $k \geq 1$
- $\#k$-faces of $\mathbb{F}_2^N = 2^{N-k}\binom{N}{k}$

$$
\begin{aligned}
t_N &\leq \frac{1}{N}\left(1 + \sum_{k=1}^{N-1} \frac{1}{2^{N-k}\binom{N}{k}}\right) \leq \frac{1}{N} + \frac{1}{N}\sum_{k=1}^{N-1}\frac{1}{2^{N-k}} \\
&= \frac{1}{N} + \frac{1}{N}\left(1 - \frac{1}{2^{N-1}}\right) = \frac{2}{N}\left(1 - \frac{1}{2^N}\right) \\
&< \frac{2}{N}
\end{aligned}
$$

□

4. Ultrametricity Index of $\mathbb{F}_2^N$

Proof that $t_N > \frac{1}{N}$.

$$t_N = \frac{1}{N}\left(1 + \sum_{k=1}^{N-1} \frac{1}{c(\Gamma_k)}\right) > \frac{1}{N}$$

□

4. Ultrametricity Index of $\mathbb{F}_2^N$

- $\mathcal{U}_N := \{\text{ultrametric } \Delta \text{ in } \mathbb{F}_2^N\}$
- $\mathbb{F}_2^N$ acts without fixed points on $\mathcal{U}_N$ via translations
- $\Rightarrow u_N := \frac{|\mathcal{U}_N|}{2^N} \in \mathbb{N}$

Proposition

$$u_N = \sum_{k=3}^{N} \sum_{\lceil \frac{2k}{3} \rceil \leq i \leq k} \binom{N}{k} \binom{k}{i} \binom{i}{k-i}$$

4. Ultrametricity Index of $\mathbb{F}_2^N$

Proof.

- Via translation: Δ has form $(0, a, b)$
- Side lengths: $\|a\|, \|b\|, \|a+b\|$
- $I := \text{supp}(a)$, $J = \text{supp}(b) \Rightarrow \text{supp}(a+b) = I \triangle J$
- Assume triangle is in a k-face, but not inside one of its faces: $\Rightarrow |I \cup J| = k$
- Δ ultrametric $\Rightarrow |I| = |J|$ and $|I \triangle J| \leq |I|$
- with $\ell = |I \cap J|$ this means:

$$2|I| - \ell = k$$
$$2|I| - 2\ell \leq i$$

- solution: $|I| \geq \left\lceil \frac{2k}{3} \right\rceil$

4. Ultrametricity Index of $\mathbb{F}_2^N$

- number of such triangles:

$$\sum_{\lceil \frac{2k}{3} \rceil \le i \le k} \binom{k}{i} \binom{i}{\ell} = \sum_{\lceil \frac{2k}{3} \rceil \le i \le k} \binom{k}{i} \binom{i}{2i-k}$$

$$= \sum_{\lceil \frac{2k}{3} \rceil \le i \le k} \binom{k}{i} \binom{i}{k-i}$$

- All ultrametric Δ of form $(0, a, b)$:

$$u_N = \sum_{k=3}^{N} \binom{N}{k} \sum_{\lceil \frac{2k}{3} \rceil \le i \le k} \binom{k}{i} \binom{i}{k-i}$$

$\square$

4. Ultrametricity Index of $\mathbb{F}_2^N$

Proof that $m_N \to 0$.

- Gosper's approximation:

$$n! = \sqrt{2\pi}\left(\frac{n}{e}\right)^n \sqrt{n+\frac{1}{6}} \cdot Q(n), \qquad \lim_{n\to\infty} Q(n) = 1$$

- $\Delta(\mathbb{F}_2^N, d_H) = \frac{2^N(2^N-1)(2^N-2)}{6}$

$$m_N \sim \frac{6u_N}{4^N} = \frac{6}{4^N}\sum_{k=3}^{N}\sum_{\lceil\frac{2k}{3}\rceil\le i\le k} \frac{N!}{(N-k)!(k-i)!^2(2i-k)!}$$

4. Ultrametricity Index of $\mathbb{F}_2^N$

$$
< \frac{6Q_{\max}}{(2\pi)^{\frac{3}{2}} N^{\frac{3}{2}}} \sum\sum \frac{\left(1+\frac{1}{6N}\right)^{\frac{1}{2}}}{\left(1-\frac{k}{N}+\frac{1}{6N}\right)^{\frac{1}{2}} \left(\frac{k}{N}-\frac{i}{N}+\frac{1}{6N}\right) \left(\frac{2i}{N}-\frac{k}{N}+\frac{1}{6N}\right)^{\frac{1}{2}}}
$$

$$
\cdot \left(\frac{1}{4\left(1-\frac{k}{N}\right)^{1-\frac{k}{N}} \left(\frac{k}{N}-\frac{i}{N}\right)^{2\left(\frac{k}{N}-\frac{i}{N}\right)} \left(\frac{2i}{N}-\frac{k}{N}\right)^{\frac{2i}{N}-\frac{k}{N}}} \right)^N
$$

$$
\sim \frac{6Q_{\max} N^{\frac{1}{2}}}{(2\pi)^{\frac{3}{2}}} \int_{\frac{3}{N}}^{1} \int_{\frac{2}{3}x}^{x} h_N(x,y) e^{Nf(x,y)} \, dydx
$$

4. Ultrametricity Index of $\mathbb{F}_2^N$

$$Q_{\max} = \max\left\{\frac{Q(N)}{Q(N-k)Q(k-i)^2Q(2i-k)}\right\}$$

$$h_N(x,y) = \frac{1}{\left(1-x+\frac{1}{6N}\right)^{\frac{1}{2}}\left(x-y+\frac{1}{6N}\right)\left(2y-x+\frac{1}{6N}\right)^{\frac{1}{2}}}$$

$$f(x,y) = -\log 4 - (1-x\log(1-x) - 2(x-y)\log(x-y) - (2y-x)\log(2y-x)$$

- $(\frac{3}{4}, \frac{1}{2})$ is unique global maximum of $f(x,y)$
- $f(\frac{3}{4}, \frac{1}{2}) = 0$
- Hessian matrix $H(f(\frac{3}{4}, \frac{1}{2}))$ is negative definite

4. Ultrametricity Index of $\mathbb{F}_2^N$

$\Rightarrow$ Laplace method yields:

$$m_N \lesssim \frac{6Q_{\max}N^{\frac{1}{2}}}{(2\pi)^{\frac{3}{2}}}\left(\frac{2\pi}{N}\right)\cdot \det\left(H(f\left(\frac{3}{4},\frac{1}{2}\right))\right)^{-\frac{1}{2}} h_N\left(\frac{3}{4},\frac{1}{2}\right)e^{Nf\left(\frac{3}{4},\frac{1}{2}\right)}$$
$$\approx \frac{2.5651}{N^{\frac{1}{2}}}$$

□

4. Ultrametricity Index of $\mathbb{F}_2^N$

Lower bound for m_N.

- Replace $<$ by $>$ and $Q_{\max}$ by

$$Q_{\min} = \min \left\{ \frac{Q(N)}{Q(N-k)Q(k-i)^2 Q(2i-k)} \right\}$$

$$\Rightarrow m_N \gtrsim \frac{6 Q_{\min}}{(2\pi)^{\frac{1}{2}} N^{\frac{1}{2}}} \approx \frac{2.4}{N^{\frac{1}{2}}}$$

□

Conclusion

- For random samples of fixed size, the ultrametricity indices tend to one if dimension tends to infinity.
- For the discrete hypercube, the ultrametricity indices tend to zero as dimension tends to infinity.
- In particular, the fraction of ultrametric triangles becomes negligible in the discrete hypercube, as dimension tends to infinity.
- Randomness and sparsity pick precisely these, as dimension tends to infinity.