Stochastic graph representations of information

Alfred Hero* **and Abram Magner†**

*Dept of EECS, Dept of Statistics, Dept of BME
Program in Applied and Interdiscimplinary Mathematics
University of Michigan - Ann Arbor

†Dept of Computer Science
SUNY - Albany

Feb 7, 2020

Acknowledgements

Hwang, Sung-Jin, Steven Damelin, Alfred O. Hero, "Shortest path through random points," *Annals of Applied Probability*, Volume 26, Number 5 (2016), 2791-2823.

Acknowledgements

Hwang, Sung-Jin, Steven Damelin, Alfred O. Hero, "Shortest path through random points," *Annals of Applied Probability*, Volume 26, Number 5 (2016), 2791-2823.

Calder, Jeff, Selim Esedoglu, and Alfred O. Hero. "A Hamilton–Jacobi Equation for the Continuum Limit of Nondominated Sorting." *SIAM Journal on Mathematical Analysis*, 46, no. 1 (2014): 603-638.

Acknowledgements

Hwang, Sung-Jin, Steven Damelin, Alfred O. Hero, "Shortest path through random points," *Annals of Applied Probability*, Volume 26, Number 5 (2016), 2791-2823.

Calder, Jeff, Selim Esedoglu, and Alfred O. Hero. "A Hamilton–Jacobi Equation for the Continuum Limit of Nondominated Sorting." *SIAM Journal on Mathematical Analysis*, 46, no. 1 (2014): 603-638.

S. Sekeh, M. Noshad, K. Moon, and AH. "Convergence Rates for Empirical Estimation of Binary Classification Bounds." Entropy, 2019.

Acknowledgements

Hwang, Sung-Jin, Steven Damelin, Alfred O. Hero, "Shortest path through random points," *Annals of Applied Probability*, Volume 26, Number 5 (2016), 2791-2823.

Calder, Jeff, Selim Esedoglu, and Alfred O. Hero. "A Hamilton–Jacobi Equation for the Continuum Limit of Nondominated Sorting." *SIAM Journal on Mathematical Analysis*, 46, no. 1 (2014): 603-638.

S. Sekeh, M. Noshad, K. Moon, and AH. "Convergence Rates for Empirical Estimation of Binary Classification Bounds." Entropy, 2019.

M. Baranwal, A. Magner, P. Elvati, J. Saldinger, A. Violi, A. Hero, "A deep learning architecture for metabolic pathway prediction," Bioinformatics, 2019.

Acknowledgements

Hwang, Sung-Jin, Steven Damelin, Alfred O. Hero, "Shortest path through random points," *Annals of Applied Probability*, Volume 26, Number 5 (2016), 2791-2823.

Calder, Jeff, Selim Esedoglu, and Alfred O. Hero. "A Hamilton–Jacobi Equation for the Continuum Limit of Nondominated Sorting." *SIAM Journal on Mathematical Analysis*, 46, no. 1 (2014): 603-638.

S. Sekeh, M. Noshad, K. Moon, and AH. "Convergence Rates for Empirical Estimation of Binary Classification Bounds." Entropy, 2019.

M. Baranwal, A. Magner, P. Elvati, J. Saldinger, A. Violi, A. Hero, "A deep learning architecture for metabolic pathway prediction," Bioinformatics, 2019.

Magner A, Baranwal M, AH, "The Power of Graph Convolutional Networks to Distinguish Random Graph Models," arXiv preprint arXiv:1910.12954. 2019 Oct 28.

Outline

Entropy and information

Assume that X, Y have joint distribution $f_{X,Y}(x, y)$ on $\mathbb{R}^{d_x \times d_y}$.

Classical definitions of entropy and mutual information Shannon (1948)

- Marginal (Shannon) entropy:

$$H(X) = -\int f_X(x) \log f_X(x) dx$$

- Joint entropy:

$$H(X, Y) = -\int f_{X,Y}(x, y) \log f_{X,Y}(x, y) dx dy$$

- Conditional entropy:

$$H(X|Y) = -\int f_{X,Y}(x, y) \log f_{X|Y}(x|y) dx dy = H(X, Y) - H(Y)$$

- Mutual information:

$$I(X, Y) = \int f_{X,Y}(x, y) \log \frac{f_{X|Y}(x|y)}{f_X(x)} dx dy = H(X) - H(X|Y)$$

Entropy and information

Assume that X, Y have joint distribution $f_{X,Y}(x, y)$ on $\mathbb{R}^{d_x \times d_y}$.

Classical definitions of entropy and mutual information Shannon (1948)

- Marginal (Shannon) entropy:

$$H(X) = -\int f_X(x) \log f_X(x) dx$$

- Joint entropy:

$$H(X, Y) = -\int f_{X,Y}(x, y) \log f_{X,Y}(x, y) dx dy$$

- Conditional entropy:

$$H(X|Y) = -\int f_{X,Y}(x, y) \log f_{X|Y}(x|y) dx dy = H(X, Y) - H(Y)$$

- Mutual information:

$$I(X, Y) = \int f_{X,Y}(x, y) \log \frac{f_{X|Y}(x|y)}{f_X(x)} dx dy = H(X) - H(X|Y)$$

$\Leftrightarrow$ the large-n limit of a certain graph over i.i.d. $\{(X_i, Y_i)\}_{i=1}^n$ from $f_{X,Y}$.

Rényi and Havrda-Charvat-Tsallis (HCT) entropies of order α

- Rényi-α entropy (Rényi (1961)) for $\alpha > 0$:

$$H_\alpha(f) = \frac{1}{1-\alpha}\log \int_{\mathbb{R}^d} f^\alpha(x)dx$$

- Rényi-α information divergence from f to g for $\alpha \in [0,1]$:

$$D_\alpha(f\|g) = \frac{1}{\alpha-1}\log \int_{\mathbb{R}^d} f^\alpha(x)g^{1-\alpha}(x)dx,$$

Property: as $\alpha \to 1$

$$H_\alpha(f) \to -\int f(x)\log f(x)dx \quad (\textit{Shannon entropy})$$

$$D_\alpha(f\|g) \to \int f(x)\log\frac{f(x)}{g(x)}dx \quad (\textit{KL divergence})$$

Rényi and Havrda-Charvat-Tsallis (HCT) entropies of order α

- Rényi-α entropy (Rényi (1961)) for $\alpha > 0$:

$$H_\alpha(f) = \frac{1}{1-\alpha}\log \int_{\mathbb{R}^d} f^\alpha(x)dx$$

- Rényi-α information divergence from f to g for $\alpha \in [0,1]$:

$$D_\alpha(f\|g) = \frac{1}{\alpha-1}\log \int_{\mathbb{R}^d} f^\alpha(x)g^{1-\alpha}(x)dx,$$

Property: as $\alpha \to 1$

$$H_\alpha(f) \to -\int f(x)\log f(x)dx \quad (\textit{Shannon entropy})$$

$$D_\alpha(f\|g) \to \int f(x)\log\frac{f(x)}{g(x)}dx \quad (\textit{KL divergence})$$

- HCT-α entropy (Havrda and Charvat (1967),Tsallis (1988))

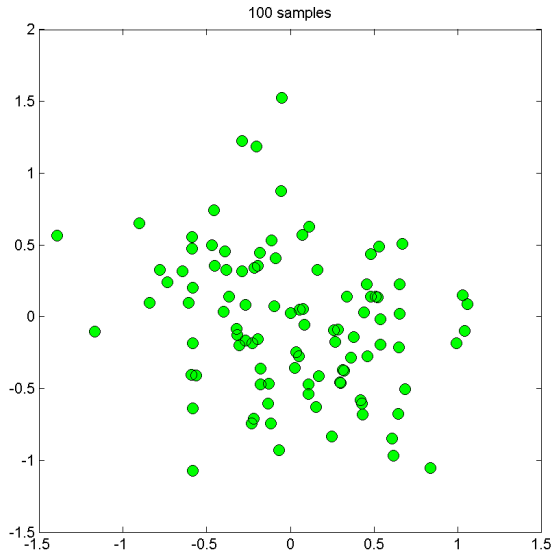
$$\tilde{H}_\alpha(f) = \frac{1}{1-\alpha}\left(\int_{\mathbb{R}^d} f^\alpha(x)dx - 1\right)$$

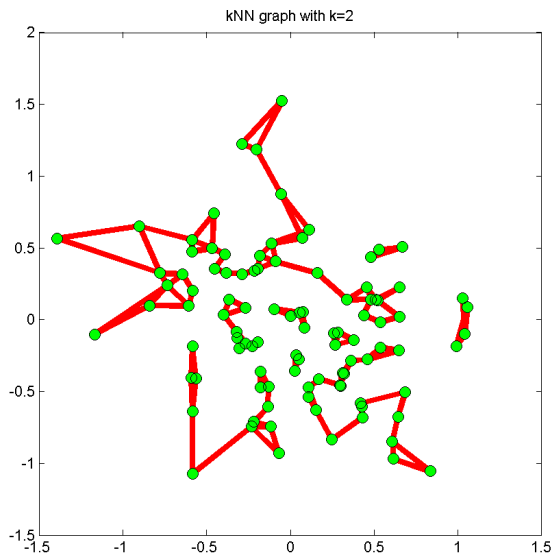
k-nearest neighbor (kNN) graph

- n Euclidean points $\{X_i\}_{i=1}^n$, $X_i \in \mathbb{R}^d$
- $\gamma \in (0, d)$ a parameter
- kNN graph $G = \{V, E\}$

$$
\begin{aligned}
L_\gamma^{kNN}(V) &= \min_{E:\mathbf{A}\underline{1}\geq k\underline{1}} L_\gamma(V, E) \\
&= \min_{E:\mathbf{A}\underline{1}\geq k\underline{1}} \sum_{e_{ij}\in E} |e_{ij}|^\gamma \\
&= \sum_{i=1}^n \sum_{j\in\mathcal{N}_k(X_i)} \|X_i - X_j\|^\gamma
\end{aligned}
$$

- $\mathcal{N}_k(X_i)$ are the k-nearest neighbors of X_i in $\mathcal{X}_n - \{X_i\}$
- Computational complexity is $O(kn\log n)$



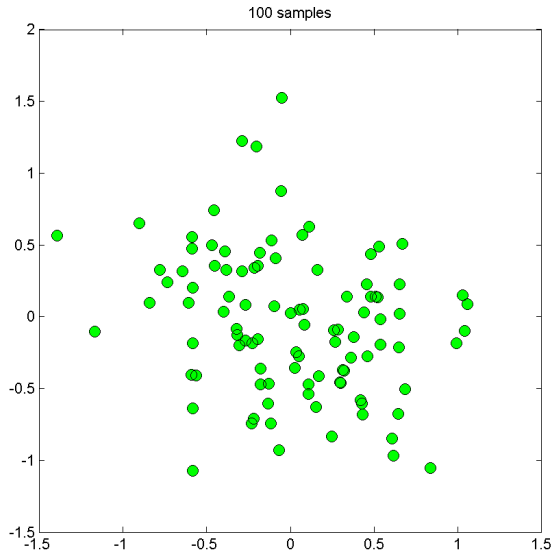


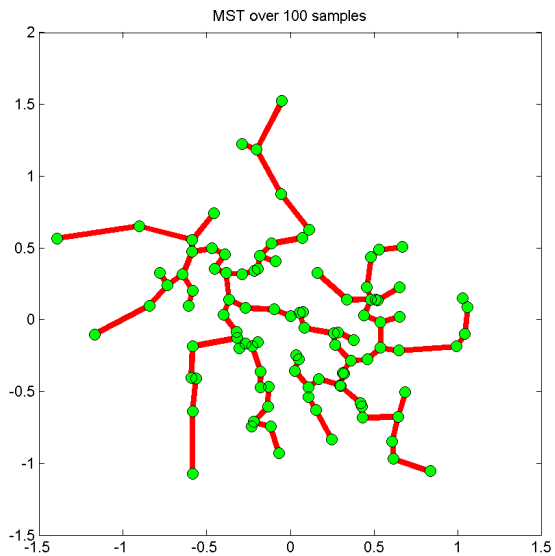
Minimal spanning tree (MST)

- MST is solution of the optimization

$$L_\gamma^{MST}(V) = \min_{E:\mathbf{A}\underline{1}>0} L_\gamma(V,E) = \min_{E:\mathbf{A}\underline{1}>0} \sum_{e_{ij}\in E} |e_{ij}|^\gamma$$

- MST spans all of the vertices V without cycles
- MST has exactly $n-1$ edges
- Computational complexity is $O(n^2 \log n)$





Friedman-Rafsky graph (FR)

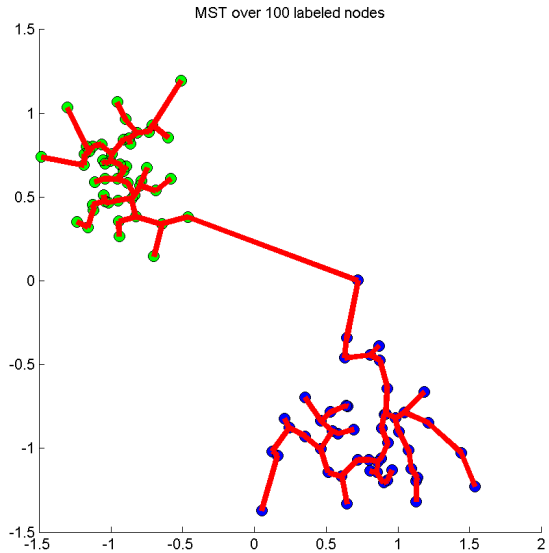
- Two labeled sample sets $\mathcal{X}_n$, $\mathcal{Y}_m$
- Start with MST over $V = \mathcal{X}_n \cup \mathcal{Y}_m$

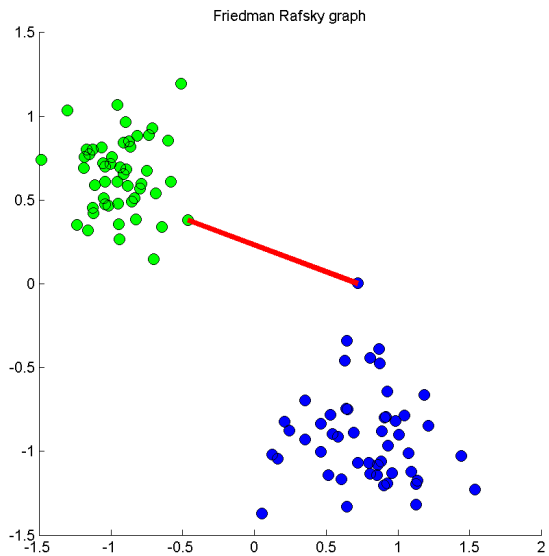
$$L_\gamma^{MST}(V) = \min_{E:\mathbf{A}\underline{1}>0} L_\gamma(V,E)$$
$$= \sum_{e_{ij}\in E^*} |e_{ij}^{XX}|^\gamma + |e_{ij}^{XY}|^\gamma + |e_{ij}^{YY}|^\gamma$$

- FR graph is the set of edges $\{e_{ij}^{XY}\}$
- The length of FR graph is

$$L_\gamma^{FR}(V) = \sum_{e_{ij}^{XY}\in\mathcal{E}^{MST}} |e_{ij}^{XY}|^\gamma$$

- $L_0^{FR}(V)$ was proposed as a multivariate run length statistic to test if $\mathcal{X}_n$ and $\mathcal{Y}_m$ come from the same distribution (Friedman and Rafsky, 1979)





Shortest path (SP) through a graph

- Let G be a graph with $m = |E|$ edges on n vertices V
- $\pi(X_I, X_F)$ a path over G btwn points X_I and X_F

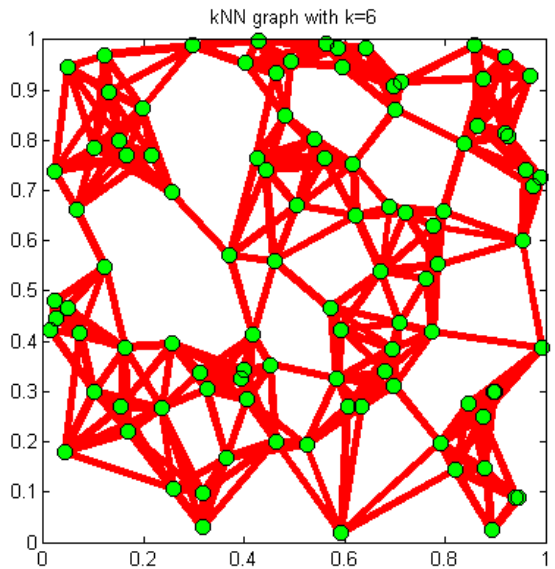
$$\pi(X_I, X_F) = (X_I, X_{i_1}, \ldots, X_{i_l}, X_F)$$

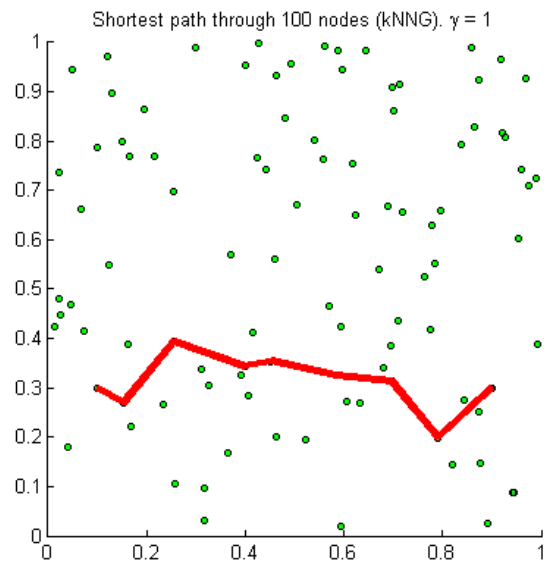
$X_{i_{j+1}}$ is neighbor on G of predecessor X_{i_j} and $X_I = X_{i_0}$, $X_F = X_{i_{l+1}}$

- The shortest path is the solution to

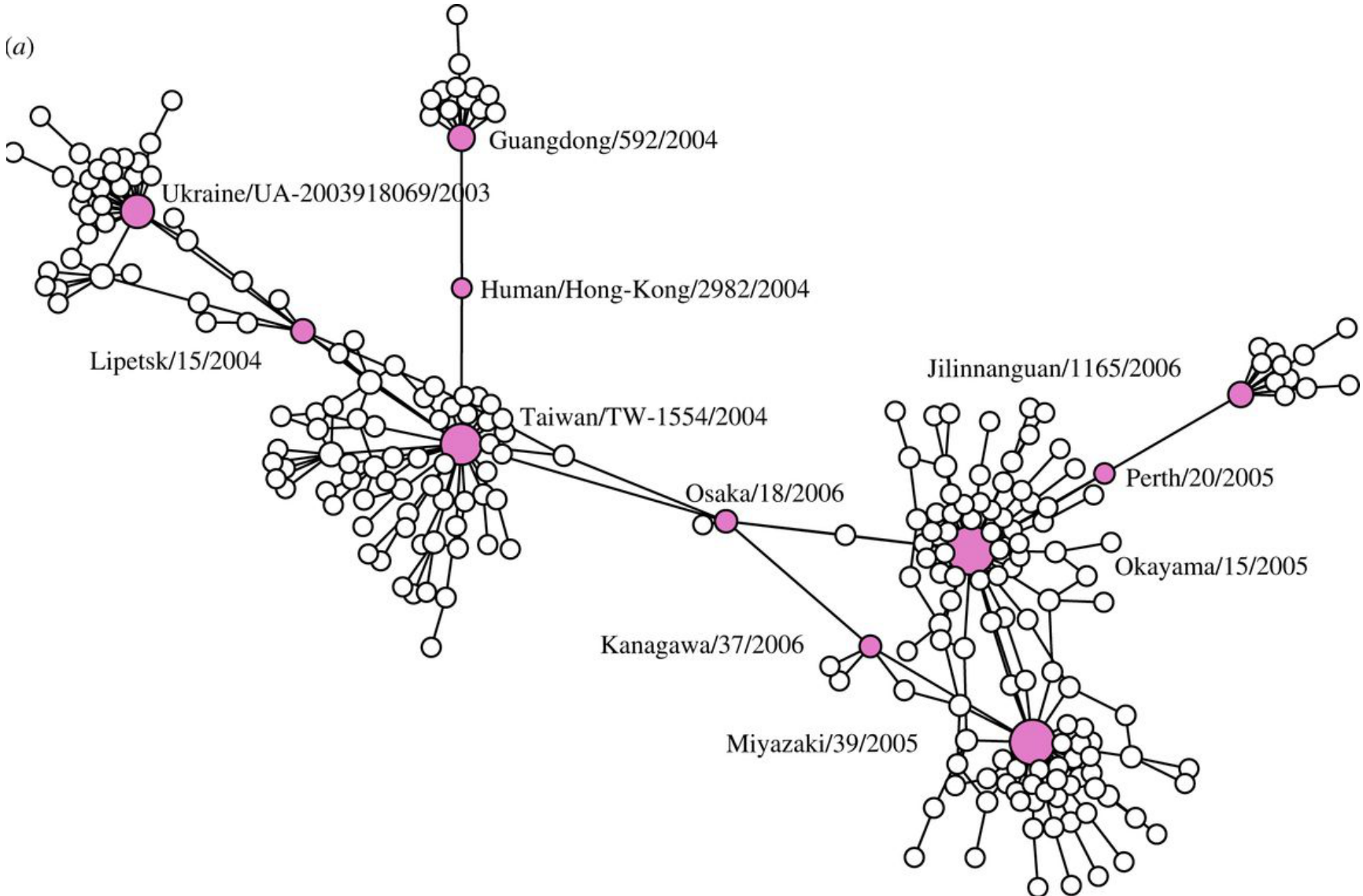
$$L_\gamma^{SP}(V) = \min_{\pi(X_i, X_F)} \sum_{X_i \in \pi(X_I, X_F)} |X_{i_{j+1}} - X_{i_j}|^\gamma$$

- Possible choices of G:
  - kNN graph
  - MST
  - Complete graph
- Computational complexity is $O(m + n\log n)$





From local to global structure: virus strain genotyping in epidemiology



A. Wagner, "A genotype network reveals homoplastic cycles of convergent evolution in influenza A (H3N2) haemagglutinin," Proc. Roval Soc. B. Mav 2014.

Local vs global properties of Euclidean graphs

Let $G = \{\mathcal{X}_n, E\}$ be a graph over $\mathcal{X}_n$ with edges E.
Define $L : G \to \mathbb{R}$ be a property of G, e.g., the sum of its edge weights.

Local vs global properties of Euclidean graphs

Let $G = \{\mathcal{X}_n, E\}$ be a graph over $\mathcal{X}_n$ with edges E.
Define $L : G \to \mathbb{R}$ be a property of G, e.g., the sum of its edge weights.

- $L(\mathcal{X}_n)$ is a *global property* of G
- $L(F)$ is a *local property* of G if F is a localized subset of $\mathcal{X}_n$

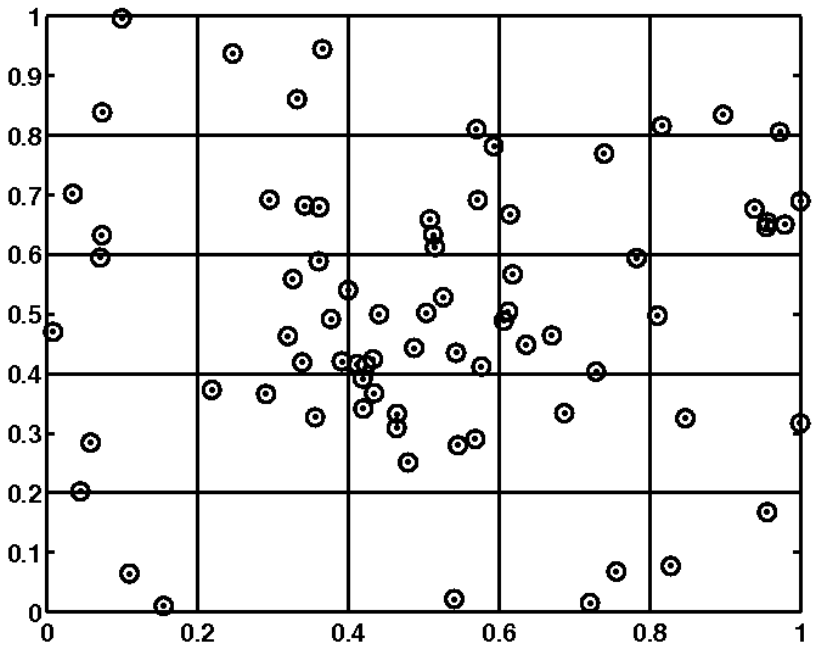
Certain global properties of G are *stable* with respect to local properties

Local vs global properties of Euclidean graphs

Let $G = \{\mathcal{X}_n, E\}$ be a graph over $\mathcal{X}_n$ with edges E.
Define $L : G \to \mathbb{R}$ be a property of G, e.g., the sum of its edge weights.

- $L(\mathcal{X}_n)$ is a *global property* of G
- $L(F)$ is a *local property* of G if F is a localized subset of $\mathcal{X}_n$

Certain global properties of G are *stable* with respect to local properties
$\Rightarrow$ continuous and quasi-additive functionals L



Examples: sum of edges, sum of vertex degrees, degree distribution of kNN and MST
Non-examples: length of k-point MST, lengths of shortest paths in kNN

Continuous and quasi-additive graph functionals (Yukich [1988])

A global property $L_\gamma(F)$ is a continuous quasi-additive graph functional if

- Translation invariance and homogeneity

$$\begin{aligned} \forall\, x \in \mathbb{R}^d, \quad L_\gamma(F+x) &= L_\gamma(F), \quad \text{(translation invariance)} \\ \forall\, c > 0, \quad L_\gamma(cF) &= c^\gamma L_\gamma(F), \quad \text{(homogeneity)} \end{aligned}$$

- Null condition: $L_\gamma(\phi) = 0$, where ϕ is the null set
- Subadditivity: There exists a constant C_1 with the following property: For any uniform resolution $1/m$-partition $\mathcal{Q}^m$

$$L_\gamma(F) \le m^{-\gamma} \sum_{i=1}^{m^d} L_\gamma(m[(F \cap Q_i) - q_i]) + C_1 m^{d-\gamma}$$

- Superadditivity: For same conditions as above, there exists a constant C_2

$$L_\gamma(F) \ge m^{-\gamma} \sum_{i=1}^{m^d} L_\gamma(m[(F \cap Q_i) - q_i]) - C_2 m^{d-\gamma}$$

- Continuity: There exists a constant C_3 such that for all finite subsets F and G of $[0,1]^d$

$$|L_\gamma(F \cup G) - L_\gamma(F)| \le C_3 \left(\mathrm{card}(G)\right)^{(d-\gamma)/d}$$

J. Yukich, "Probability theory of classical Euclidean optimization problems," Springer Lecture Notes in Mathematics, 1998.

kNN and MST length functions converge to the HCT-α entropy

The following theorem holds for any continuous quasi-additive graph, e.g., kNN and MST.

Theorem (Beardwood, Halton&Hammersley 1959, Steele 1997, Yukich 1998)

Let $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ be an i.i.d. realization from a Lebesgue density f supported on compact subset of $\mathbb{R}^d$. If $0 < \gamma < d$

$$\lim_{n\to\infty} L_\gamma^{MST,kNN}(\mathcal{X}_n)/n^{(d-\gamma)/d} \quad = \quad \beta_{\gamma,d} \int f(x)^{(d-\gamma)/d} dx, \qquad (a.s.)$$

Alternatively, letting $\alpha = (d-\gamma)/d$,

$$\frac{1}{1-\alpha}\left(L_\gamma(\mathcal{X}_n)/n^\alpha - 1\right) \quad \to \quad \tilde{H}_\alpha(f) \qquad (a.s.)$$

Steele, *Probability theory and combinatorial optimization*, SIAM 1997.

Beardwood and Halton and Hammersley, "The shortest path through many points," Proc. Cambridge Philosophical Society 1959.

J. Yukich, "Probability theory of classical Euclidean optimization problems," Springer Lecture Notes in Mathematics, 1998.

FR length function converges to an information divergence measure

Let $\mathcal{X} = \{X_1, \ldots, X_n\}$ and $\mathcal{Y} = \{Y_1, \ldots, Y_m\}$ be independent and i.i.d. in $\mathbb{R}^d$ with pdfs f_X and f_Y, respectively. Then

Theorem (Henze (1999), Berisha (2015), Sekeh (2019))

Let n, m converge to infinity in such a way that $n/(n+m) = p$, $p \in [0,1]$. Then

$$1 - L_0^{FR}(\mathcal{X} \cup \mathcal{Y}) \frac{n+m}{2nm} \to D_p(f_x, f_y) \qquad (a.s.)$$

where D_p is Henze-Penrose (HP) divergence

$$D_p(f,g) = (4p(1-p))^{-1} \left(\int \frac{(pf(x) - (1-p)g(x))^2}{pf(x) + (1-p)g(x)} dx - (2p-1)^2 \right)$$

D_p is an information divergence measure that gives a tight bound on Bayes binary classification error.

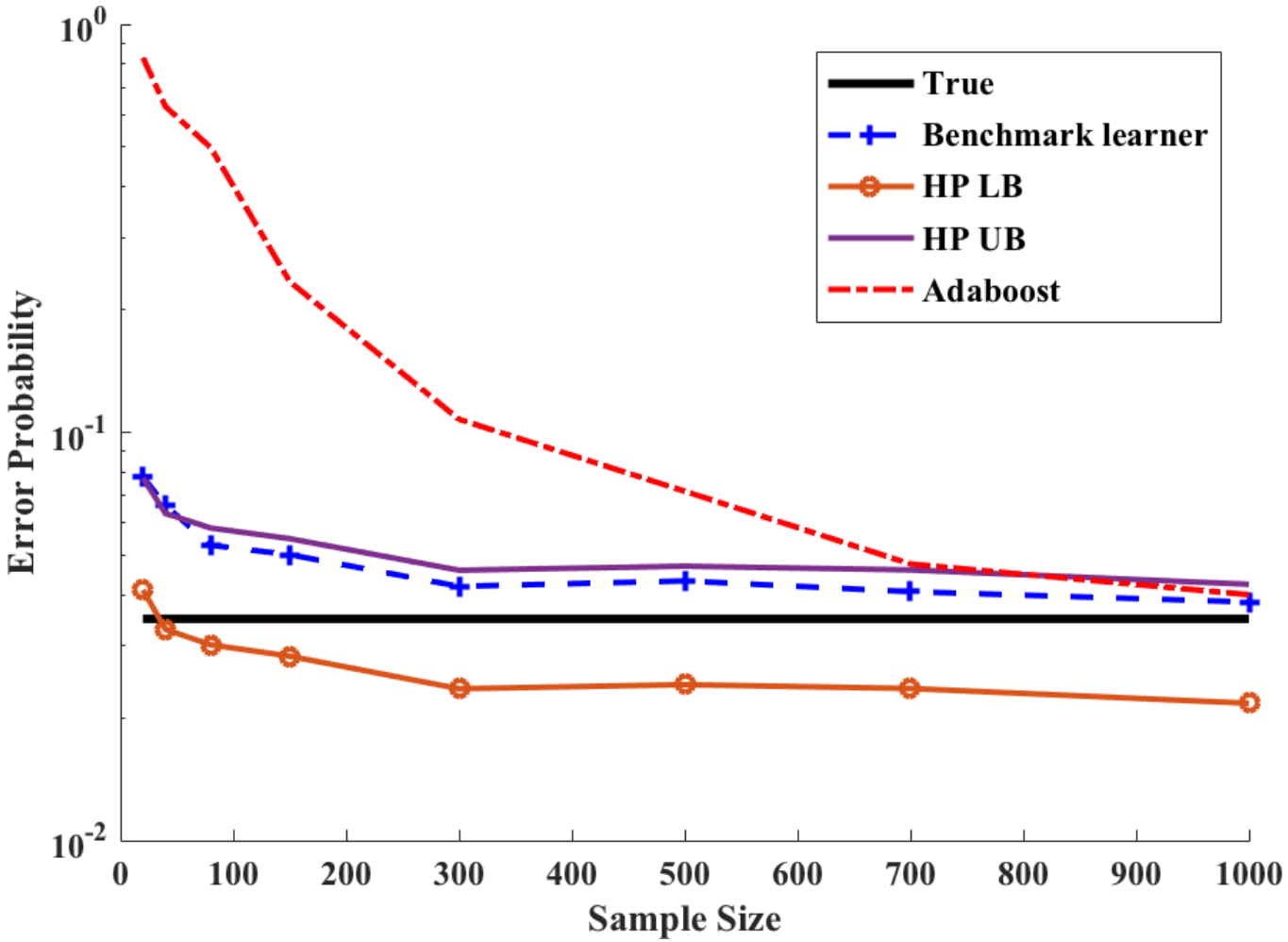
N. Henze and M. Penrose, "On the multivariate runs test," Ann. of Statistics, 1999.

V. Berisha and AH, "Empirical non-parametric estimation of the Fisher Information," IEEE Signal Processing Letters, 2015.

S. Sekeh, M. Noshad, K. Moon, and AH. "Convergence Rates for Empirical Estimation of Binary Classification Bounds." Entropy, 2019.

Application of HP divergence: a minibatch stopping rule (Noshad [2019])

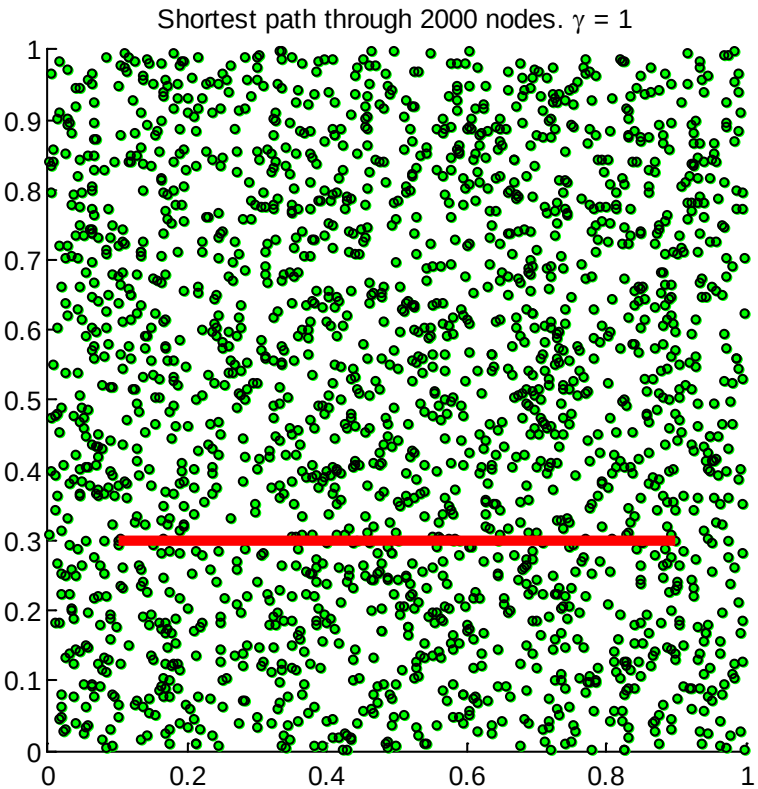
Simulation: classification of 2 mean shifted 10 dim Gaussian densities



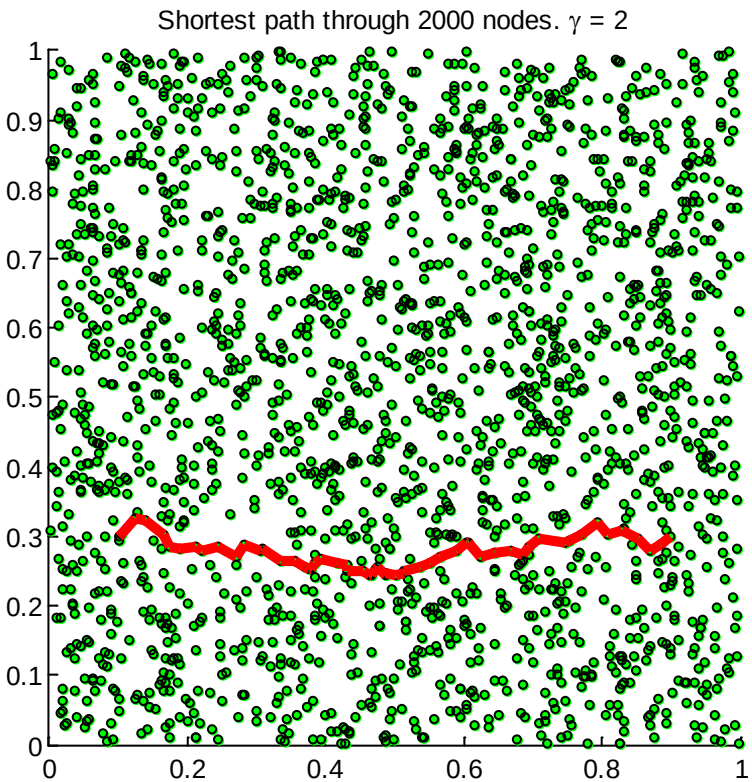
M. Noshad, L. Xu and AH, "Learning to benchmark: determining best achievable misclassification error from trinaing data," arXiv:1909.07192, 2019.

Shortest path between two points: uniform distribution

$$L_\gamma^{SP}(\mathcal{V}) \quad = \quad \min_{\pi(X_i,X_F)} \sum_{X_i \in \pi(X_I,X_F)} |X_{i_{j+1}} - X_{i_j}|^\gamma$$



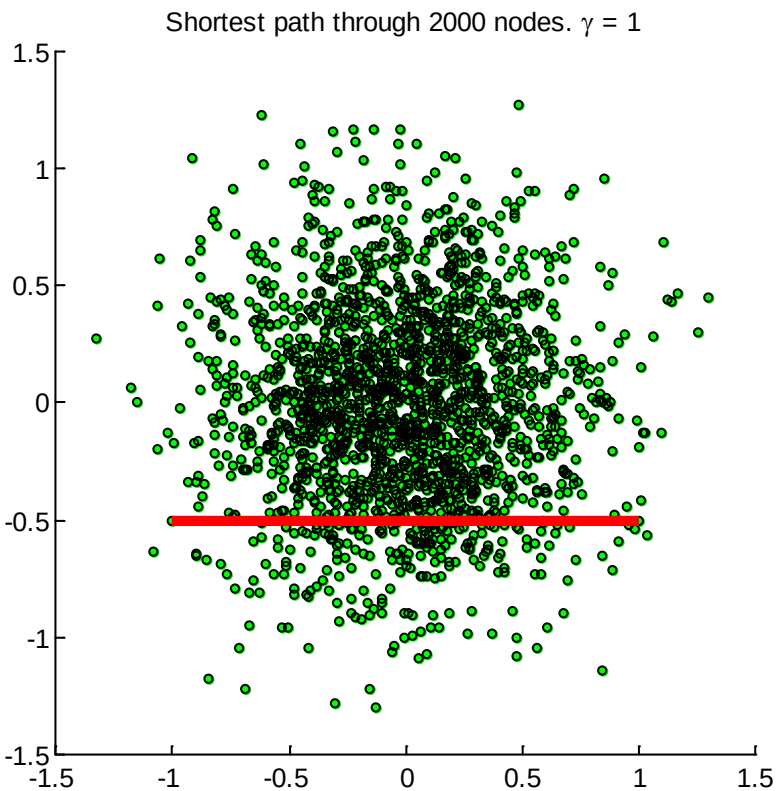
Euclidean distance ($\gamma = 1$)



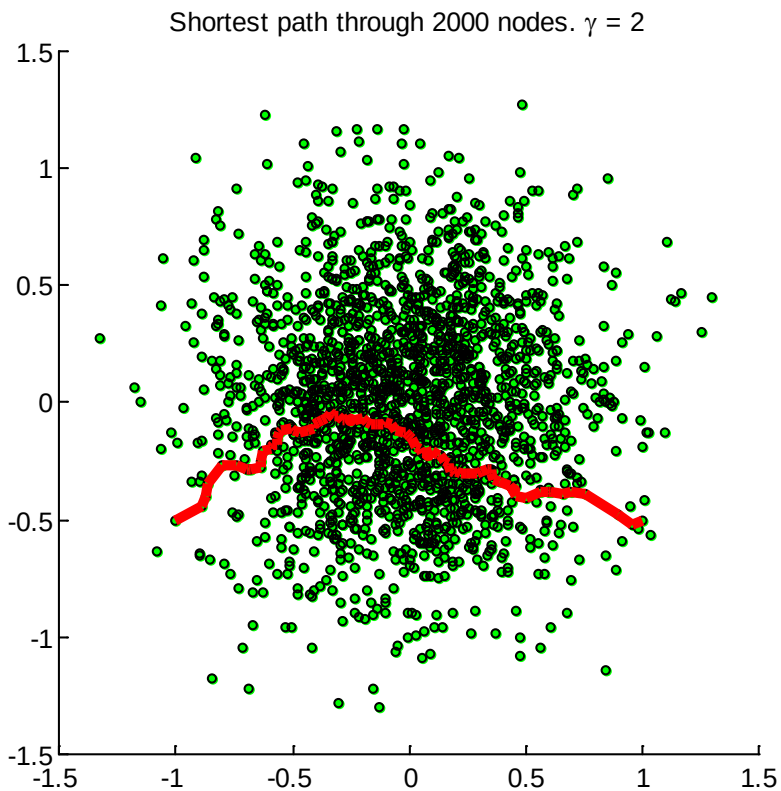
(Euclidean distance)^2 ($\gamma = 2$)

SP between two points: lensing effect of Gaussian distribution

$$L_\gamma^{SP}(\mathcal{V}) = \min_{\pi(X_i,X_F)} \sum_{X_i \in \pi(X_I,X_F)} |X_{i_{j+1}} - X_{i_j}|^\gamma$$



Euclidean ($\gamma = 1$)



(Euclidean)^2 ($\gamma = 2$)

Continuum limit of shortest path through complete graph

Let $\mathcal{X} = \{X_1, \ldots, X_n\}$ be i.i.d. random vectors in $\mathbb{R}^d$ with marginal pdf f hiving support set $\mathcal{S}$. Fix two points x_I and x_F in $\mathbb{R}^d$.

Define $\mathcal{G}$ as the complete graph spanning $\mathcal{X}$

Theorem (Hwang, Damelin and AH 2016)

Assume that $\inf_x f(x) > 0$ over a compact support set $\mathcal{S}$ with pd metric tensor g. For $\gamma > 1$ the shortest path on $\mathcal{G}$ between any two points $x_I, x_F \in \mathcal{S}$ satisfies

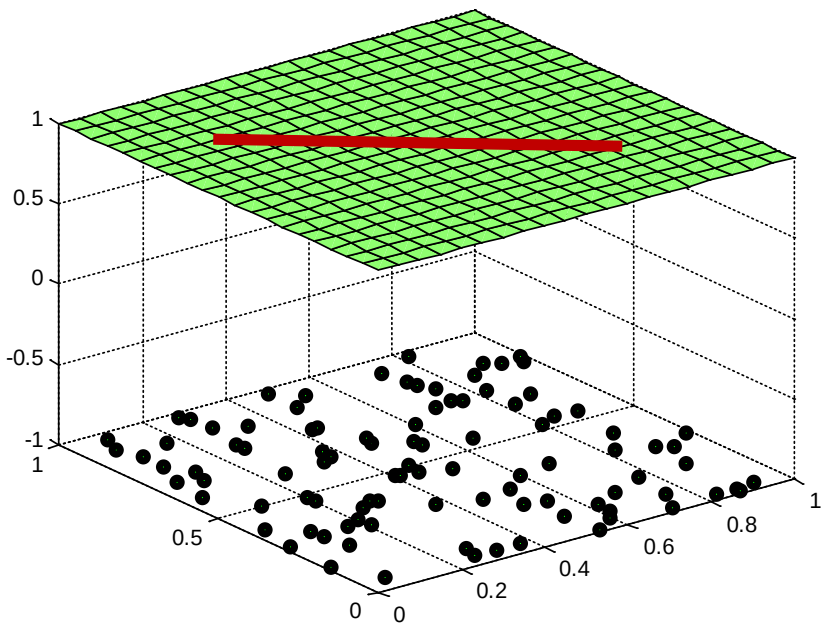
$$L_\gamma^{SP}(\mathcal{X})/n^{(1-\gamma)/d} \to C_{d,\gamma} \underbrace{\inf_\pi \int_0^1 f(\pi_t)^{(1-\gamma)/d} \sqrt{g(\dot{\pi}_t, \dot{\pi}_t)} dt}_{dist_\gamma(x_I, x_F)} \qquad (a.s.)$$

where the infimum is taken over all smooth curves $\pi : [0,1] \to \mathbb{R}^d$ with $\pi_0 = x_I$ and $\pi_1 = x_F$ and $C(d,\gamma)$ is a constant independent of f.

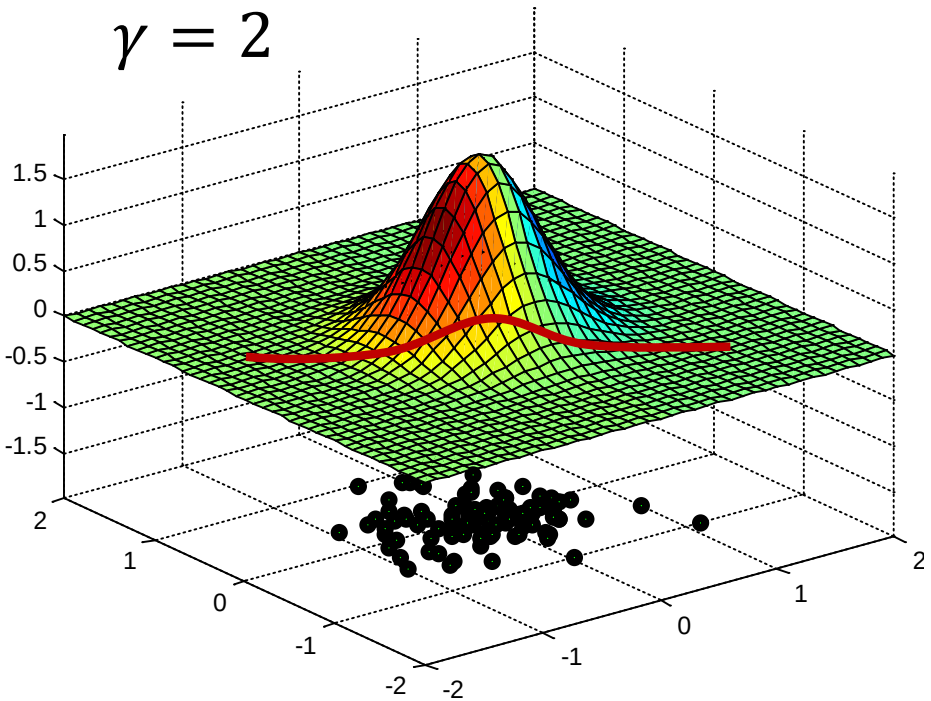
• S.-J. Hwang, S. Damelin, AH, "Shortest path through random points," Annals of Applied Probability, Volume 26, Number 5 (2016), 2791-2823. (arXiv:1202.0045).

The continuum limit of shortest path as $n \to \infty$

$d = 2, \quad \gamma = 2$



Archimedean shortest path



Relativistic shortest path

Continuum limit of shortest path: ODE (Eikonal) variational form

Define

$$F(\pi, \dot{\pi}) = f(\pi)^{(1-\gamma)/d} \sqrt{g(\dot{\pi}, \dot{\pi})}$$

Then Thm. implies normalized shortest path length converges to integral I

$$L_\gamma^{SP}(\mathcal{X})/n^{(1-\gamma)/d} \to I(\pi, \dot{\pi}) = C_{d,\gamma} \inf_\pi \int_0^1 F(\pi_t, \dot{\pi}_t) dt$$

Eikonal form: For initial point $x_I \in \mathbb{R}^d$ consider the distance function $D_{x_I}(x)$ to any other point $x \neq x_I$.

Then, for $\pi = \pi(x_I, x)$ and $g(u, u) = \|u\|^2$, constant contours of integral $I = I(x)$ can be represented as propagating fronts of D_{x_I}.

The distance function D is a viscosity solution of the Eikonal equation

$$\|\nabla D_{x_I}(x)\| = \begin{cases} W(x), & x \in \mathcal{S}/\{x_I\} \\ 0, & o.w. \end{cases}$$

where $W = f^{(1-\gamma)/d}$ (the *speed* of fronts of D).

Eikonal equations can be solved efficiently by Fast Marching (Sethian, 1996) over discretized domain $\mathcal{S}$ of f.

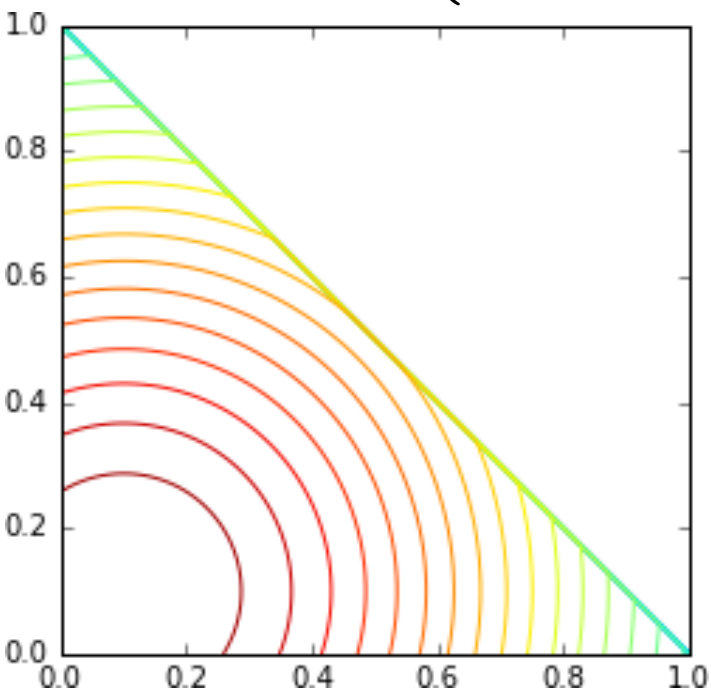
Numerical illustration: shortest path computation

- Histogram data on $(d-1)$-dimensional simplex $\Omega \subset \mathbb{R}^d$

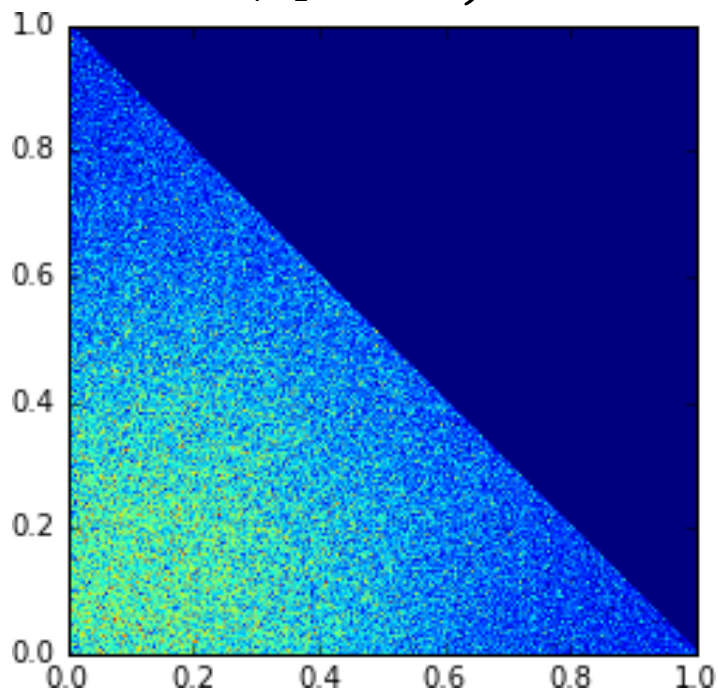
$$\Omega = \left\{ x \in \mathbb{R}^d : x_1, \ldots, x_d \geq 0, \sum_{i=1}^{d} x_i = 1 \right\}.$$

- Equivalent linearly independent representation in hypertriangle $\mathcal{S} \subset \mathbb{R}^{d-1}$:

$$\mathcal{S} = \left\{ x \in \mathbb{R}^{d-1} : x_1, \ldots, x_{d-1} \in [0,1], \sum_{i=1}^{d-1} x_i \leq 1 \right\}$$



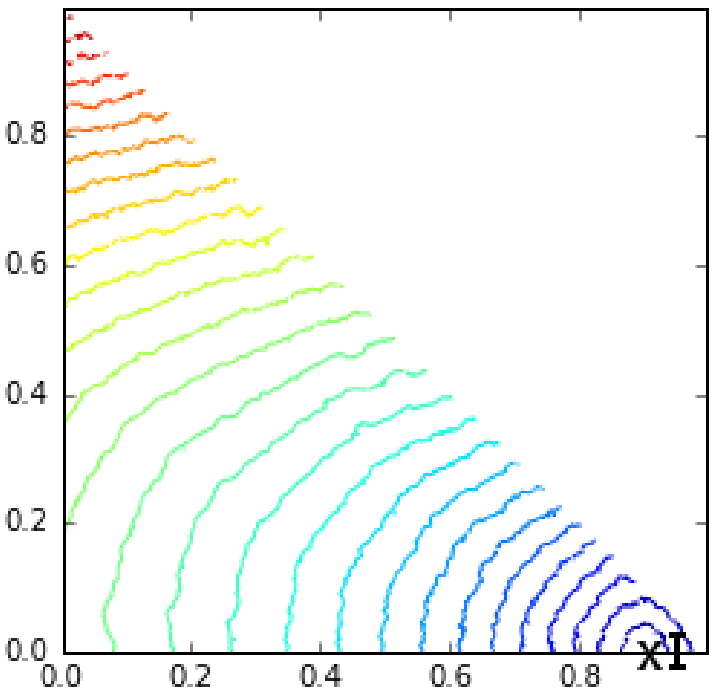
Truncated Gaussian $f(x)$ on $\mathcal{S} \subset \mathbb{R}^2$



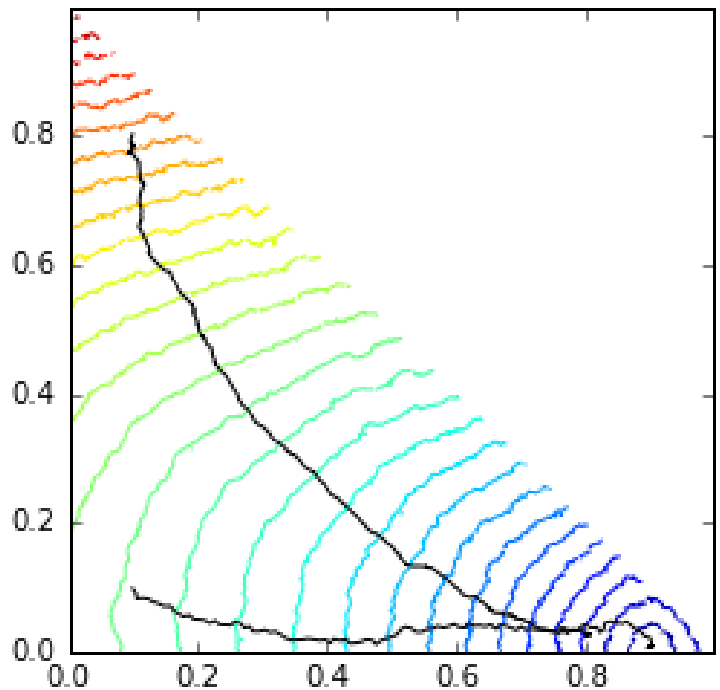
$n = 500,000$ realizations

Numerical illustration: shortest path computation

- Domain $\mathcal{S}$ of distance function D discretized into m^{d-1} cubic cells $\{C_j\}$
- Distance function $D_{x_I} : \mathcal{S} \to \mathbb{R}^+$ computed by FM for an initial point $x_I \in C_j$



Distance function by FM ($\gamma = 2$, $m^2 = 80K$)



Shortest paths by FM

Comparison: Eikonal ODE vs combinatorial Dijkstra

Table: CPU times (secs) for Fast Marching ($n = 500,000$)

cells m^{d-1}	10000	20000	30000	40000	50000	60000	70000
d=2	0.06	0.12	0.17	0.26	0.32	0.37	0.45
d=3	0.16	0.28	0.43	0.65	0.75	0.92	1.12
d=4	0.27	0.7	0.99	1.44	1.92	2.23	3.26
d=5	0.69	1.2	2.03	2.98	3.33	4.66	5.36

Table: CPU times (secs) for Dijkstra

vertices n	1000	2000	3000	4000	5000	6000	7000
d=2	1.08	5.92	11.43	21.42	36.46	108.37	248.19
d=3	1.4	4.84	11.18	20.	32.36	111.48	259.31
d=4	1.14	4.51	10.66	19.14	31.11	113.12	272.03
d=5	1.12	4.54	11.6	21.43	32.87	102.57	247.6

Implementation: Python 3.6.1, Fast Marching from `scikit-fmm 0.0.9`, Dykstra from `NetworkX`

Remarks

Random graph representation of information

- Many information measures have random graph representations.
- Random graphs can induce novel measures of information divergence.

Remarks

Random graph representation of information

- Many information measures have random graph representations.
- Random graphs can induce novel measures of information divergence.
- Graph-based divergence representations can be used to represent MI.
  - HP divergence can be transformed to a MI measure (Sekeh [2019]) :

$$MI_p(X, Y) = D_p(f_{X,Y}, f_X f_Y)$$

  - Thus obtain a *direct* graph estimator of dependency, w/o density estimation.
  - HP dependency shares properties of Shannon MI (Sekeh [2019]).

Remarks

Random graph representation of information

- Many information measures have random graph representations.
- Random graphs can induce novel measures of information divergence.
- Graph-based divergence representations can be used to represent MI.
  - HP divergence can be transformed to a MI measure (Sekeh [2019]) :

$$MI_p(X, Y) = D_p(f_{X,Y}, f_X f_Y)$$

  - Thus obtain a *direct* graph estimator of dependency, w/o density estimation.
  - HP dependency shares properties of Shannon MI (Sekeh [2019]).

From local to global properties

- Random graph representations can elucidate interplay between local and global properties.
- Continuous quasiadditive global properties are stable wrt local perturbations: length of kNN, FR.
  $\Rightarrow$ Global continuum limit is additive integral function over local domains

Remarks

Random graph representation of information

- Many information measures have random graph representations.
- Random graphs can induce novel measures of information divergence.
- Graph-based divergence representations can be used to represent MI.
  - HP divergence can be transformed to a MI measure (Sekeh [2019]) :

$$MI_p(X, Y) = D_p(f_{X,Y}, f_X f_Y)$$

  - Thus obtain a *direct* graph estimator of dependency, w/o density estimation.
  - HP dependency shares properties of Shannon MI (Sekeh [2019]).

From local to global properties

- Random graph representations can elucidate interplay between local and global properties.
- Continuous quasiadditive global properties are stable wrt local perturbations: length of kNN, FR.
  $\Rightarrow$ Global continuum limit is additive integral function over local domains
- Non-Archimedian deviation of shortest path quantifies multiscale interaction
  $\Rightarrow$ Continuum limit of SP is the solution to a Eikonal ode

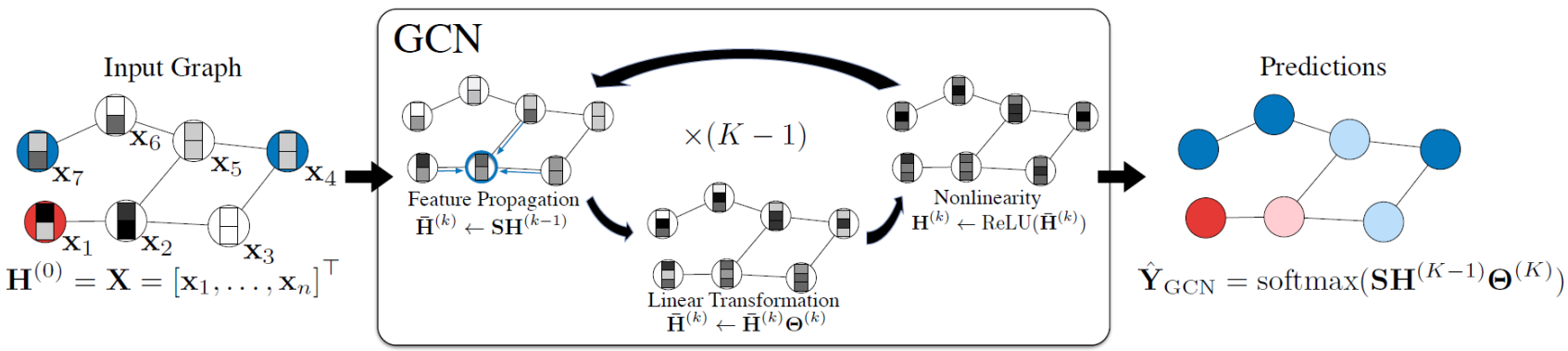
S. Sekeh and AH, "Geometric Estimation of Multivariate Dependency," Entropy, 2019.

Outline

Graph convolutional network (GCN) classifiers

GCN: a recently introduced DNN for classifying graph properties (Kipf [2016]). Can perform

- Local classification/interpolation: node label prediction
- Global classification: graph label prediction

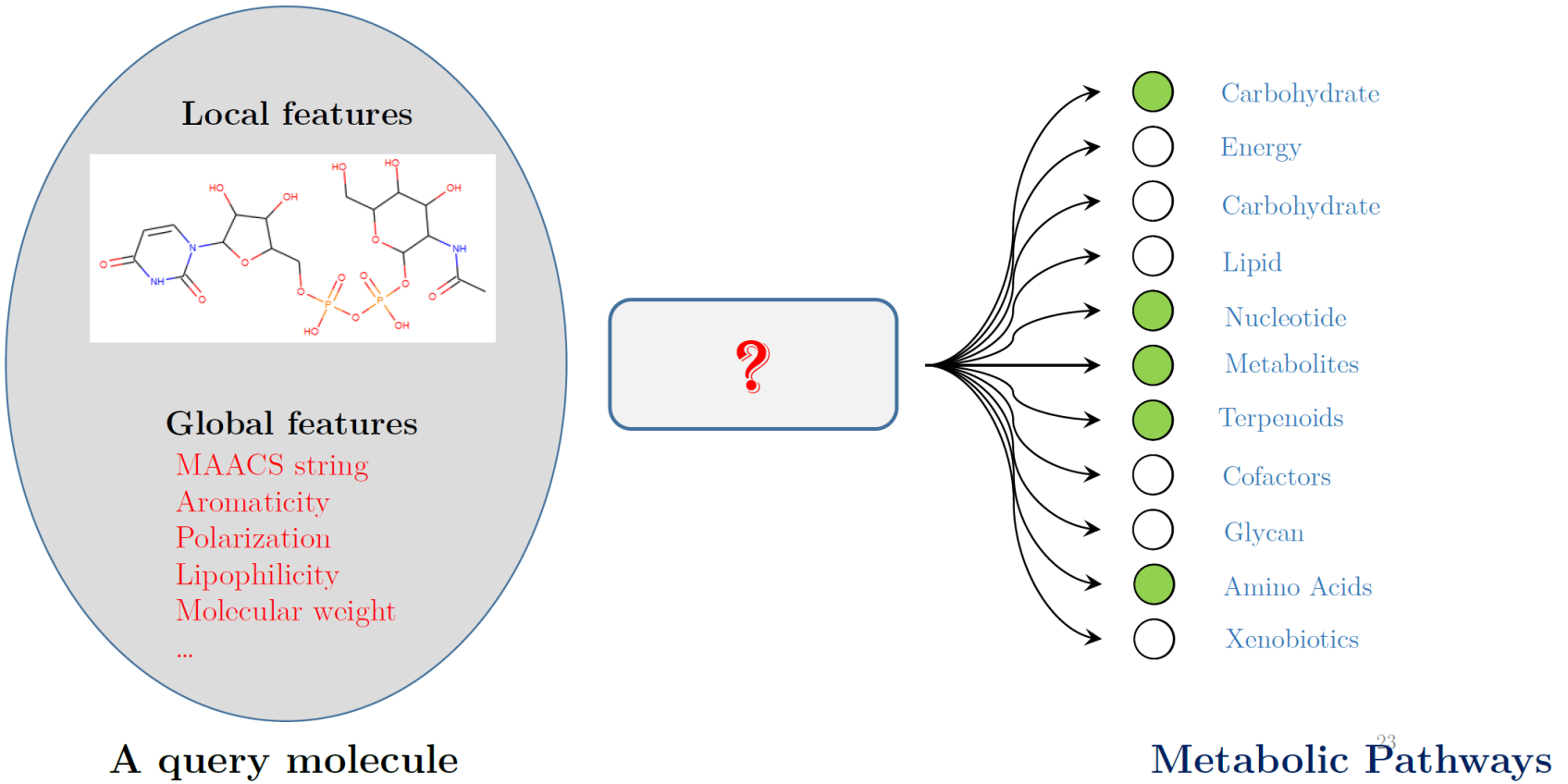


Source: F Wu et al, "Simplifying graph convolutional networks," ICLR 2019.

- **S** is graph kernel matrix that propagates node features
- Θ is matrix of weights that encodes node features

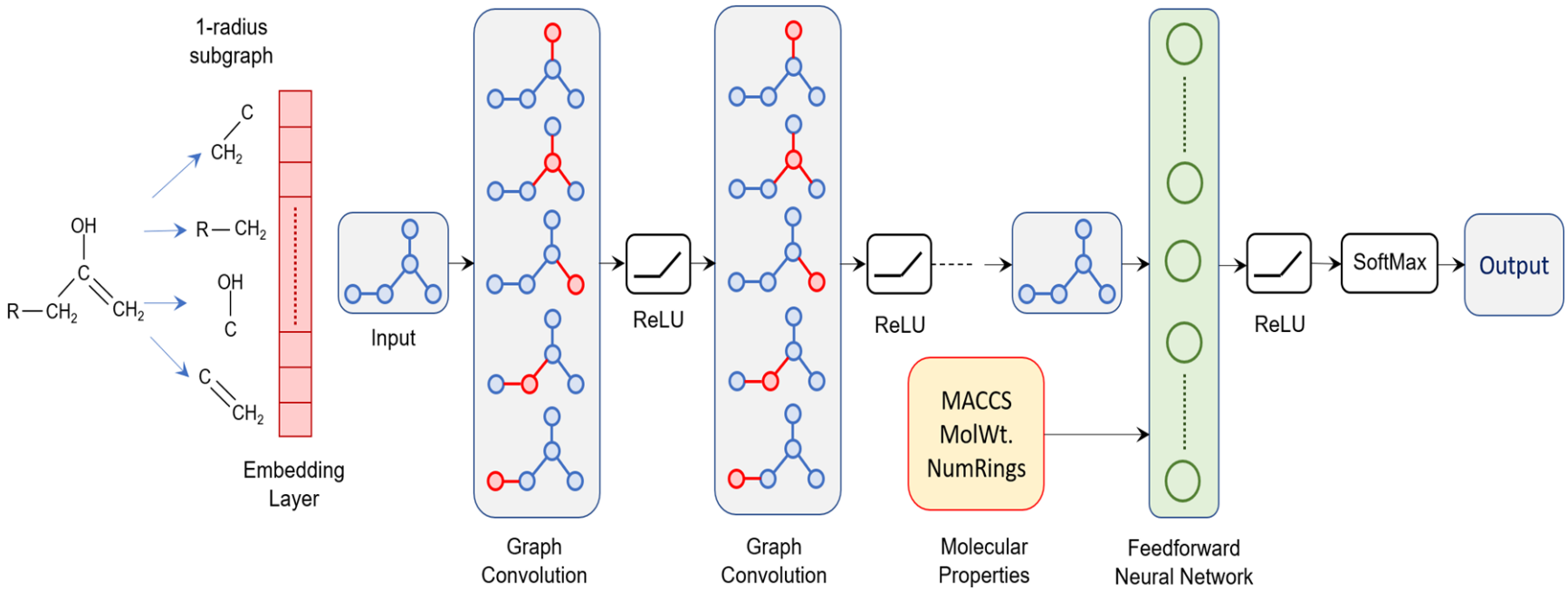
TN Kipf, M Welling, "Semi-supervised classification with graph convolutional networks," ICLR 2017. arXiv preprint arXiv:1609.02907, 2016.

Application: classification of metabolic pathways from molecular features



M. Baranwal, A. Magner, P. Elvati, J. Saldinger, A. Violi, A. Hero, "A deep learning architecture for metabolic pathway prediction," Bioinformatics, 2019.

Proposed GCN architecture



M. Baranwal, A. Magner, P. Elvati, J. Saldinger, A. Violi, A. Hero, "A deep learning architecture for metabolic pathway prediction," Bioinformatics, 2019.

Performance comparisons on Kegg PPI database

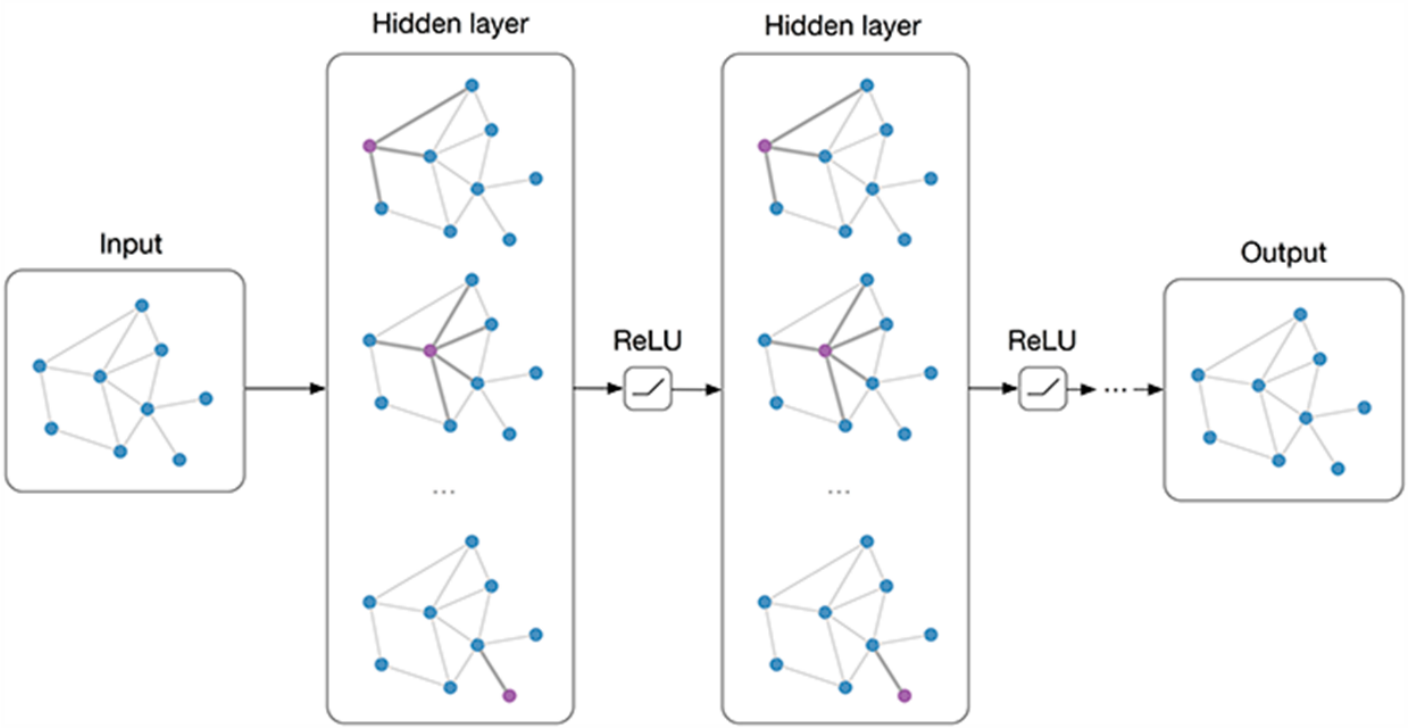
Method	Scores (%)		
	Accuracy	Precision	Recall
Hu *et al.* (2011)	94.64	77.97	67.83
kNN classifier	90.52±.81	56.25±3.2	57.99±2.8
Ensemble logistic regression	85.48±.61	23.68±1.6	18.30±1.5
Independent RFs	97.58±.12	83.69±.78	83.63±.68
GCN + additional features	**97.61**±.12	**91.61**±.52	**92.50**±.44

M. Baranwal, A. Magner, P. Elvati, J. Saldinger, A. Violi, A. Hero, "A deep learning architecture for metabolic pathway prediction," Bioinformatics, 2019.

Tuning the GCN

There is little understanding of the factors affecting GCN performance
Selecting the number of layers in the GCN is especially difficult

- Too few layers → ignore global graph topology → poor global sensitivity
- Too many layers → over-diffusion of local features → poor local sensitivity



M. Baranwal, A. Magner, P. Elvati, J. Saldinger, A. Violi, A. Hero, "A deep learning architecture for metabolic pathway prediction," Bioinformatics, 2019.

Outline

Fundamental limits on GCN classification capacity

Theory for representational capacity of the GCN is just starting to appear

- Xu K, Hu W, Leskovec J, Jegelka S, "How powerful are graph neural networks?," ICLR 2019.
- Magner A, Baranwal M, AH, "The Power of Graph Convolutional Networks to Distinguish Random Graph Models," arXiv preprint arXiv:1910.12954. 2019 Oct 28. .

This theory seeks to reveal factors that enable or disable accurate GCN performance

Ultimate aim is to provide principles to guide reliable GCN design

The power of GCNs to distinguish random graph models

Ingredients for our main result

- Graph G on n vertices are realizations drawn i.i.d. from a graphon $W \in \mathcal{W}$
- Geometrize the space of graphons $\mathcal{W}$ with a metric: cut-distance
- Constrain distance δ between degree distributions of W_0 and W_1
- Formulate K layer GCN as a test between $H_0 : G \sim W_0$, vs $H_1 : G \sim W_1$
- Mixing time characterization of random walks on graphon samples
- Apply concentration inequalities to bound misclassification error

The power of GCNs to distinguish random graph models

Ingredients for our main result

- Graph G on n vertices are realizations drawn i.i.d. from a graphon $W \in \mathcal{W}$
- Geometrize the space of graphons $\mathcal{W}$ with a metric: cut-distance
- Constrain distance δ between degree distributions of W_0 and W_1
- Formulate K layer GCN as a test between $H_0 : G \sim W_0$, vs $H_1 : G \sim W_1$
- Mixing time characterization of random walks on graphon samples
- Apply concentration inequalities to bound misclassification error

The following holds if the GCN has "nice" activation functions and bounded weight matrices.

Theorem (Magner (2019) Theorem 1)

Assume that $K > D\log n$, for some constant D possibly depending on W_0 and W_1. Assume the mean degree distributions of inputs $G_0 \sim W_0$ and $G_1 \sim W_1$ are separated by a small distance $\leq \delta$. Then, with high probability the corresponding K-th GCN layer outputs $\hat{H}^{(0,K)}$ and $\hat{H}^{(1,K)}$ are indistinguishable, i.e.,

$$\|\hat{H}^{(0,K)} - \hat{H}^{(1,K)}\|_\infty \leq \frac{\delta}{n}\left(1 + O(n^{-1/2})\right)$$

A. Magner, M. Baranwal, AH, "The power of graph convolutional networks to distingusih between random graph models," arXiv:1910.12954. 2019 Oct 28

The power of GCNs to distinguish random graph models

Consider the case of noise-regularized GCN for which the output of each neuron has additive uniform noise over $[-\epsilon_{res}, \epsilon_{res}]$.

Using Theorem 1, and concentration arguments + Le Cam's method:

Theorem (Magner (2019) Theorem 2)

Assume that $D\log n < K \ll n^{1/2-\epsilon_0}$ *and that* $\epsilon_{res} > \frac{\delta}{2n}$. *Assume the mean degree distributions of inputs* $G_0 \sim W_0$ *and* $G_1 \sim W_1$ *are separated by a small distance* $\leq \delta$. *Then the probability of error of any test for distinguishing between* W_0 *and* W_1 *based on the* K*-th GCN layer output is at least*

$$\left(1 - \frac{\delta}{2\epsilon_{res} n}\right)^n$$

A. Magner, M. Baranwal, AH, "The power of graph convolutional networks to distingusih between random graph models," arXiv:1910.12954. 2019 Oct 28

The power of GCNs to distinguish random graph models

We obtain a converse to Theorems 1 and 2:

Theorem (Magner (2019) Theorem 3)

Let W_0 and W_1 be δ-separated graphons. Then there exists a test the distinguishes with probability $1 - o(1)$ between samples $G_0 \sim W_0$ and $G_1 \sim W_1$ based on the output of the K-th GCN layer, with identity weight matrices and activation functions, provided that $K > D\log n$ for sufficiently large D and $\epsilon_{res} \leq \frac{\delta}{2n}$.

I.e., a simple, linear GCN is sufficient for distinguishing δ-separated graphons.

Recovers empirical results of (Wu [2018]).

F Wu, T ZHang, A de Souza, C Fifty, T Yu, KQ Weinberger, "Simplifying graph convolutional networks," ICML 2019.

A. Magner, M. Baranwal, AH, "The power of graph convolutional networks to distinguish between random graph models," arXiv:1910.12954. 2019 Oct 28

Example: Indistinguishable stochastic block models

We exhibit a concrete family of graphons that are

- well-separated from each other in cut distance and
- 0-separated in terms of degree distribution:

Consider 2-block SBM with density parameters $P_* = (p_1^*, p_2^*, q^*)$. Define the parameter set

$$\mathcal{P} = \{P \ : \ (0,0,0) \prec P = P_* + \tau \cdot (1,1,-1) \preceq (1,1,1)\} \tag{1}$$

and consider SBMs with parameters coming from $\mathcal{P}$.

Theorem (Magner (2019) Theorem 4)

For any pair W_0, W_1 parameterized by $\mathcal{P}$, assume $K > D\log n$ as before. Then with high probability the corresponding K-th GCN layer outputs $\hat{H}^{(0,K)}$ and $\hat{H}^{(1,K)}$ are indistinguishable, i.e.,

$$\|\hat{H}^{(0,K)} - \hat{H}^{(1,K)}\|_\infty = O(n^{-3/2+const}). \tag{2}$$

An analogous error probability lower bound holds.

A. Magner, M. Baranwal, AH, "The power of graph convolutional networks to distinguish between random graph models,"

arXiv:1910.12954. 2019 Oct 28

Elements of the proofs

Start with a linear GCN with identity weight matrices: the K-th GCN layer outputs the embedding matrix

$$\hat{M}^{(K)} = \hat{A}^K, \tag{3}$$

where $\hat{A}$ is the normalized adjacency matrix of the input graph.

If K is close to the ϵ-total variation mixing time of the random walk on G, then the rows of $\hat{A}$ are close to the stationary distribution – a function of the vertex degrees.

Mixing time is $\Theta(\log(n/\epsilon))$ for graphons satisfying mild assumptions.

Several analytic details allow us to extend our analysis to a class of nice activation functions and non-identity weight matrices.

Remarks on GCN theory

- The problem of distinguishing random graph models from representations of samples can be used as a canonical downstream task for evaluating/comparing representation learning methods
- This is first result we know of that quantifies performance limitations of GCN's over graph classes
- Theorems are tight for distinct SBM's having the same degree distribution (Magner [2019], Thm 4).
- Characterization of dependence of D on W_1 and W_0 could provide guidelines for selecting K
- Our theory is limited to dense graphs.
- Extensions to sparse graphs, e.g. graphex's, would be worthwhile next step

Outline

Summary remarks

- Local and global information can be studied using random graphs.
- Random graphs can induce novel measures of information divergence.
- Graphon random graph models can enable sharp results on representation and classification of graphs by neural networks.

Mayank Baranwal, Abram Magner, Paolo Elvati, Jacob Saldinger, Angela Violi, and Alfred O Hero. A deep learning architecture for metabolic pathway prediction. *Bioinformatics*, 2015.

Visar Berisha and A Hero. Empirical non-parametric estimation of the fisher information. *IEEE Signal Processing Letters*, 22(7), 2014.

J Havrda and F Charvat. Concept of structural α-entropy. *Kybernetika*, 3:30–35, 1967.

N. Henze and M. Penrose. On the multivariate runs test. *Annals of Statistics*, 27: 290–298, 1999.

Sung Jin Hwang, Steven B Damelin, and Alfred O Hero III. Shortest path through random points. *arXiv preprint arXiv:1202.0045*, 2012.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Abram Magner, Mayank Baranwal, and Alfred O Hero III. The power of graph convolutional networks to distinguish random graph models. *arXiv preprint arXiv:1910.12954*, 2019.

Morteza Noshad, Li Xu, and Alfred Hero. Learning to benchmark: Determining best achievable misclassification error from training data. *arXiv preprint arXiv:1909.07192*, 2019.

A. Rényi. On measures of entropy and information. In *Proc. 4th Berkeley Symp. Math. Stat. and Prob.*, volume 1, pages 547–561, 1961.

Salimeh Yasaei Sekeh, Morteza Noshad, Kevin R Moon, and Alfred O Hero. Convergence rates for empirical estimation of binary classification bounds. *Entropy*, 21(12):1144, 2019.

C.E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. Journ.*, 27: 379–423, 1948.

Constantino Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of statistical physics*, 52(1-2):479–487, 1988.

Salimeh Yasaei Sekeh and Alfred O Hero. Geometric estimation of multivariate dependency. *Entropy*, 21(8):787, 2019.

J. E. Yukich. *Probability theory of classical Euclidean optimization*, volume 1675 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1998.