Variational problems and PDE on random structures

Dejan Slepčev
Carnegie Mellon University

CNA Wroking Group
Nonlocal PDE and variational problems for data
September 22, 2020.

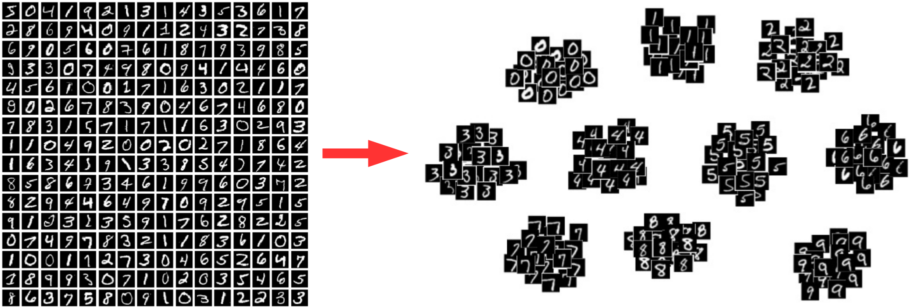
Collaborators:

- *Variational approaches to clustering (TV and spectral)*
  Xavier Bresson (NTU Singapore), Nicolás García Trillos (UW Madison), Thomas Laurent (LMU), James von Brecht (Cal. State, Long Beach)
- *Error rates for graph laplacian*
  Nicolás García Trillos, Moritz Gerlach, Matthias Hein (Tübingen)
- *Semi-Supervised Lerning and Regression*
  Marco Caroccia (Tor Vergata – Rome), Antonin Chambolle (Ecolé Polytechnique), Matthew Dunlop (NYU), Matthew Thorpe (Cambridge), Andrew Stuart (Caltech), Jeff Calder (Minnesota)

Related works

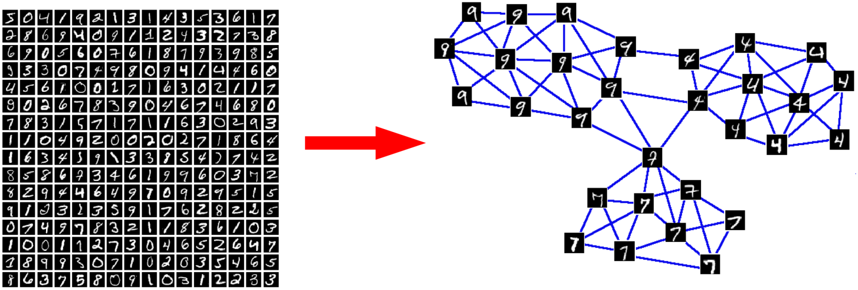
- Belkin and Niyogi, Hein and von Luxburg, Singer and Wu, Li and Shi, Pelletier, Thorpe and Theil, van Gennip, Davis and Sethuramanan, Reeb and Osting, García Trillos and Sanz Alonso, Calder, Müller and Penrose, ...

Clustering



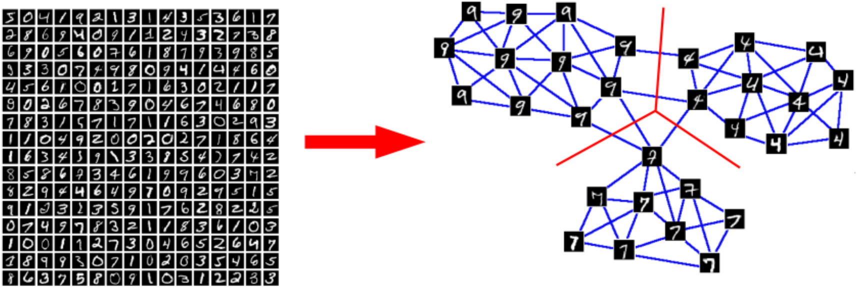
- Partition the data into meaningful groups.

Graph-Based Clustering



- Determine a similarity measure between images
- Construct a graph based on the similarity measure.

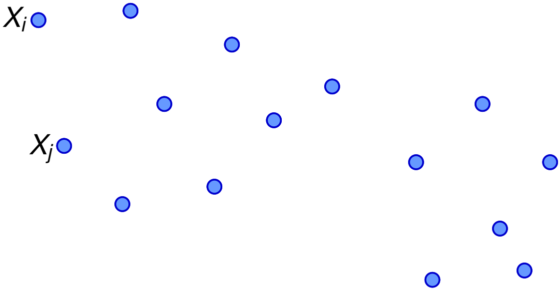
Graph-Based Clustering



- Determine a similarity measure between images
- Construct a graph based on the similarity measure.
- Partition the graph

From point clouds to graphs

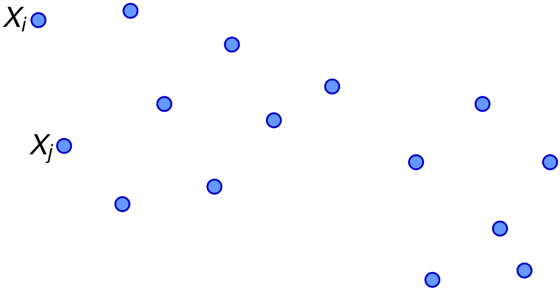
- Let $V = \{X_1, \ldots, X_n\}$ be a point cloud in $\mathbb{R}^d$:



- Connect nearby vertices: Edge weights $W_{i,j}$.

From point clouds to graphs

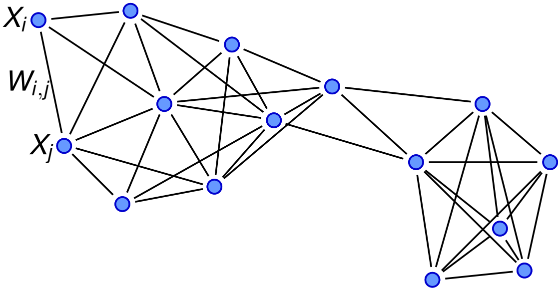
- Let $V = \{X_1, \ldots, X_n\}$ be a point cloud in $\mathbb{R}^d$:



- Connect nearby vertices: Edge weights $W_{i,j}$.

From point clouds to graphs

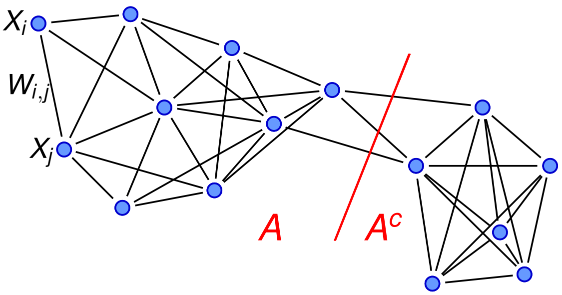
- Let $V = \{X_1, \ldots, X_n\}$ be a point cloud in $\mathbb{R}^d$:



- Connect nearby vertices: Edge weights $W_{i,j}$.

Graph cut

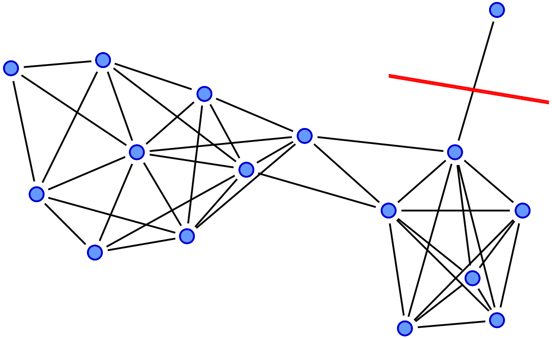
- Let $V = \{X_1, \ldots, X_n\}$ be a point cloud in $\mathbb{R}^d$:



- Connect nearby vertices: Edge weights $W_{i,j}$
- Graph Cut: $A \subset V$.

$$Cut(A, A^c) = \sum_{i \in A} \sum_{j \in A^c} W_{i,j}.$$

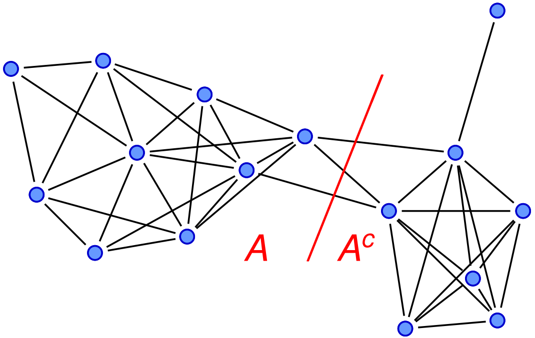
- Let $V = \{X_1, \ldots, X_n\}$ be a point cloud in $\mathbb{R}^d$:



- Connect nearby vertices: Edge weights $W_{i,j}$
- Minimize: $A \subset V$.

$$Cut(A, A^c) = \sum_{i \in A} \sum_{j \in A^c} W_{i,j}.$$

- Let $V = \{X_1, \ldots, X_n\}$ be a point cloud in $\mathbb{R}^d$:



- Graph Cut: $A \subset V$.

$$Cut(A, A^c) = \sum_{i \in A} \sum_{j \in A^c} W_{i,j}.$$

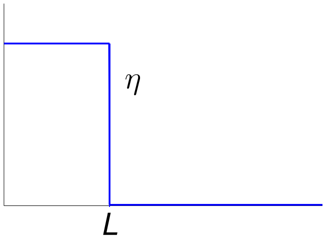
- Cheeger Cut: Minimize

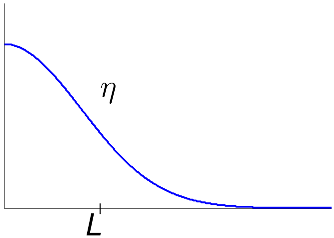
$$GC(A) = \frac{Cut(A, A^c)}{\min\{|A|, |A^c|\}}.$$

Graph Constructions

- proximity based graphs

$$W_{i,j} = \eta(X_i - X_j)$$





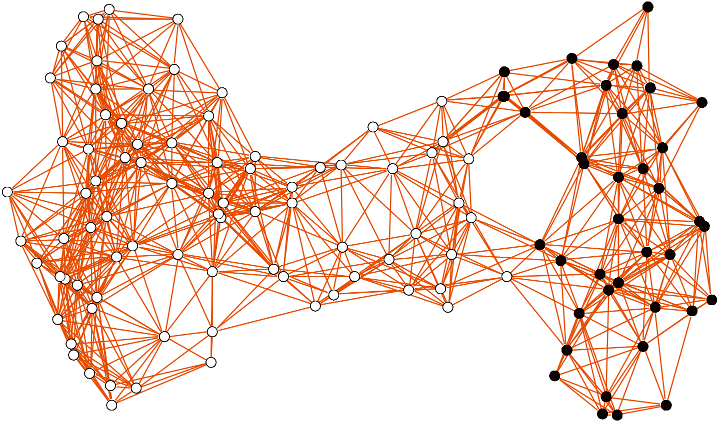
- kNN graphs: Connect each vertex with its k nearest neighbors

Task

Minimize $$GC(A) = \frac{\sum_{i \in A} \sum_{j \in A^c} W_{i,j}}{\min\{|A|, |A^c|\}}$$

Task

Minimize
$$GC(A) = \frac{\sum_{i \in A} \sum_{j \in A^c} W_{i,j}}{\min\{|A|, |A^c|\}}$$



Algorithm of Bresson, Laurent, Uminsky and von Brecht (2013).

Graph Total Variation

Graph total variation

For a function $u : V \to \mathbb{R}$

$$GTV_n(u) = \frac{1}{n^2} \sum_{i,j} W_{i,j} \left| u_i - u_j \right|$$

where $u_i = u(X_i)$.

Note that for a set of vertices $A \subset V$

$$GTV_n(\chi_A) = \frac{1}{n^2} Cut(A, A^c)$$

where χ_A is the characteristic function of A

$$\chi_A(X_i) = \begin{cases} 1 & \text{if } X_i \in A \\ 0 & \text{otherwise.} \end{cases}$$

Relaxed Problem

$$GTV_n(u) = \frac{1}{n^2} \sum_{i,j} W_{i,j} \, |u_i - u_j|.$$

Balance term

$$B_n(u) = \frac{1}{n} \min_{c \in \mathbb{R}} \sum_i |u_i - c|$$

Note that $$B_n(\chi_A) = \frac{1}{n} \min\{|A|, |A^c|\}.$$

Relaxed problem

Minimize $$GC_n(u) = \frac{GTV_n(u)}{B_n(u)}$$

Theorem

Relaxation is exact: There exists a set of vertices A_n such that $u_n = \chi_{A_n}$ minimizes GC_n.

Relaxation is sharp

$GTV_n(u) = \frac{1}{n^2} \sum_{i,j} W_{i,j} |u_i - u_j|, \qquad B_n(u) = \frac{1}{n} \min_{c \in \mathbb{R}} \sum_i |u_i - c|.$

Minimize $\qquad GC_n(u) = \dfrac{GTV_n(u)}{B_n(u)}$

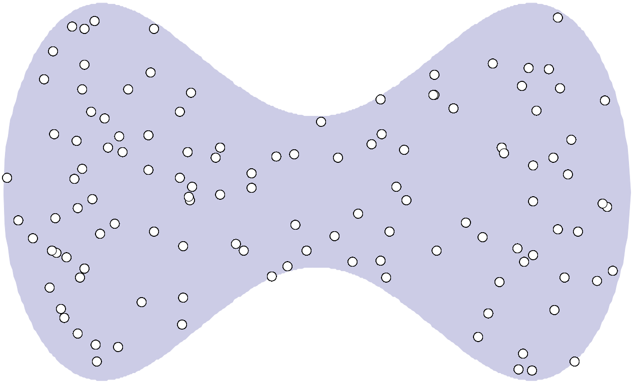
- Assume $u : V \to [0,1]$. Then $u(x) = \int_0^1 \chi_{\{u \geq \lambda\}}(x) d\lambda$.
- Coarea formula: $GTV_n(u) = \int_0^1 GTV_n(\chi_{\{u \geq \lambda\}}) d\lambda$.
- Convexity $B_n(u) \leq \int_0^1 B_n(\chi_{\{u \geq \lambda\}}) d\lambda$
- If u is a minimizer then for all λ

$$\frac{GTV_n(\chi_{\{u \geq \lambda\}})}{B_n(\chi_{\{u \geq \lambda\}})} \geq GTV_n(u) \geq \frac{\int_0^1 GTV_n(\chi_{\{u \geq \lambda\}}) d\lambda}{\int_0^1 B_n(\chi_{\{u \geq \lambda\}}) d\lambda}.$$

- Thus $\{u \geq \lambda\}$ minimizes the Cheeger cut for a.e. λ.

Ground Truth Assumption

Assume points $X_1, X_2, \ldots,$ are drawn i.i.d out of measure $d\nu = \rho dx$



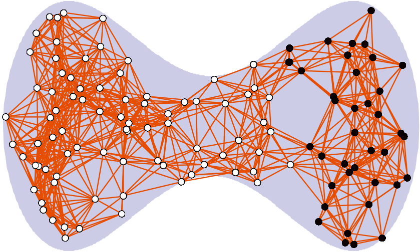
Consistency of Cheeger cut clustering

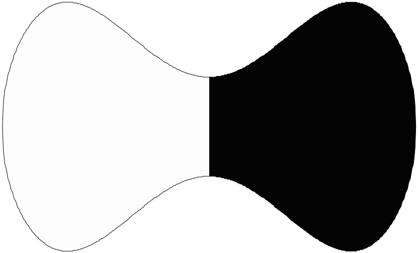
Consistency of clustering

Do the minimizers of

$$GC(A) = \frac{\sum_{i \in A} \sum_{j \in A^c} W_{i,j}}{\min\{|A|, |A^c|\}}$$

converge as the number of data points $n \to \infty$?
Can one characterize the limiting object as a minimizer of a continuum functional?





Localizing the kernel

Localizing the kernel as $n \to \infty$

$$\eta_\varepsilon(z) = \frac{1}{\varepsilon^d}\eta\left(\frac{z}{\varepsilon}\right).$$

Cheeger Cut

$$GC_{n,\varepsilon_n}(u^n) = \frac{\frac{1}{\varepsilon_n n^2}\sum_{i,j}\eta_{\varepsilon_n}(X_i - X_j)\,|u_i^n - u_j^n|}{\frac{1}{n}\min_{c\in\mathbb{R}}\sum_i |u_i^n - c|} =: \frac{GTV_{n,\varepsilon_n}(u^n)}{B_n(u^n)}$$

Question (Consistency) Do minimizers of GC_{n,ε_n} converge as the number of data points $n \to \infty$?

Characterize the limit and the rates $\varepsilon(n)$ for which the asymptotic behavior holds.

Heuristics for the limiting functional

Cheeger Cut

$$GC_{n,\varepsilon_n}(u^n) = \frac{1}{n} \frac{\frac{1}{\varepsilon_n} \sum_{i,j} \eta_{\varepsilon_n}(X_i - X_j)\,|u_i^n - u_j^n|}{\min_{c\in\mathbb{R}} \sum_i |u_i^n - c|} =: \frac{GTV_{n,\varepsilon_n}(u^n)}{B_n(u^n)}$$

Heuristics for **fixed smooth** u. Let $\mu_n = \frac{1}{n}\sum_i \delta_{X_i}$ be the empirical measure

$$\begin{aligned} GTV_{n,\varepsilon}(u) &= \frac{1}{\varepsilon n^2} \sum_{i,j} \eta_{\varepsilon_n}(X_i - X_j)|u(X_i) - u(X_j)| \\ &= \frac{1}{\varepsilon} \iint \eta_\varepsilon(x-y)|u(x)-u(y)|d\mu_n(x)d\mu_n(y) \\ &\overset{n\gg 1}{\approx} \frac{1}{\varepsilon} \iint \eta_\varepsilon(x-y)|u(x)-u(y)|d\mu(x)d\mu(y) =: TV_\varepsilon(u) \\ &\overset{\varepsilon\ll 1}{\approx} \frac{1}{\varepsilon} \iint \eta_\varepsilon(x-y)|\nabla u(x)\cdot(x-y)|d\mu(y)d\mu(x) \\ &\overset{\varepsilon\ll 1}{\approx} \sigma_\eta \int |\nabla u(x)|\rho^2(x)dx. \end{aligned}$$

Total variation in continuum setting

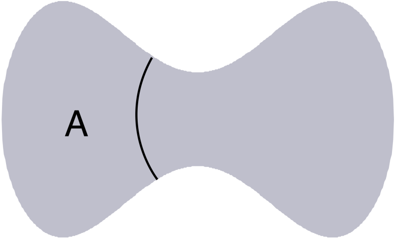
- $d\nu = \rho dx$ probability measure, $\mathrm{supp}(\nu) = D$, $0 < \lambda \leq \rho \leq \frac{1}{\lambda}$ on D.

Weighted relative perimeter

Given $A \subset D$ $$P(A; D, \rho^2) = \int_{D \cap \partial A} \rho^2 dS_{d-1}$$

Weighted TV

$$TV(u, \rho^2) = \int_D |\nabla u| \rho^2 dx$$



Total variation in continuum setting

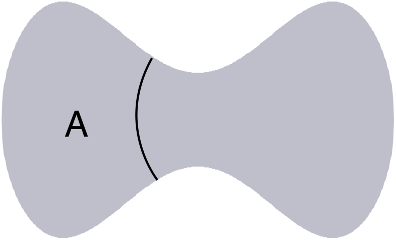
- $d\nu = \rho dx$ probability measure, $\mathrm{supp}(\nu) = D$, $0 < \lambda \leq \rho \leq \frac{1}{\lambda}$ on D.

Weighted relative perimeter

Given $A \subset D$ $$P(A; D, \rho^2) = \int_{D \cap \partial A} \rho^2 dS_{d-1} = TV(\chi_A, \rho^2)$$

Weighted TV

$$TV(u, \rho^2) = \sup \left\{ \int_D u \, \mathrm{div}(\phi) dx \ : \ |\phi| \leq \rho^2 \ , \ \phi \in C_c^\infty(D, \mathbb{R}^d) \right\}$$



Clustering in continuum setting

- ν probability measure with compact support $\text{supp}(\nu) = D$.
- ν has continuous on D density ρ and $0 < \lambda \leq \rho \leq \frac{1}{\lambda}$ on D.

Weighted TV

$$TV(u, \rho^2) = \sup \left\{ \int_D u \, \text{div}(\phi) dx \ : \ |\phi| \leq \rho^2 \ , \ \phi \in C_c^\infty(D, \mathbb{R}^d) \right\}$$

Weighted relative perimeter

Given $A \subset D$ $\qquad P(A; D, \rho^2) = TV(\chi_A, \rho^2)$

Balance term

$$B(A) = \min\{|A|, 1 - |A|\} \qquad \text{where } |A| = \nu(A).$$

Weighted Cheeger Cut: Minimize

$$C(A) = \frac{P(A; D, \rho^2)}{B(A)}$$

Relaxation in continuum setting

- ν probability measure with compact support $\text{supp}(\nu) = D$.
- ν has continuous on D density ρ and $0 < \lambda \leq \rho \leq \frac{1}{\lambda}$ on D.

Weighted TV

$$TV(u, \rho^2) = \sup \left\{ \int_D u \, \text{div}(\phi) dx \ : \ |\phi| \leq \rho^2 \ , \ \phi \in C_c^\infty(D, \mathbb{R}^d) \right\}$$

Balance term

$$B(u) = \min_{c \in \mathbb{R}} \int_D |u(x) - c| \rho(x) dx$$

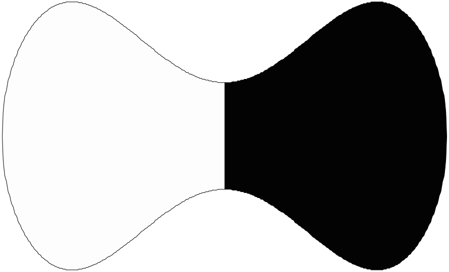
Minimize

$$C(u) = \frac{TV(u, \rho^2)}{B(u)}$$

Clustering in continuum setting

Minimize

$$C(u) = \frac{TV(u, \rho^2)}{B(u)}$$



Localizing the kernel as $n \to \infty$

$$\eta_\varepsilon(z) = \frac{1}{\varepsilon^d} \eta\left(\frac{z}{\varepsilon}\right).$$

Consistency of clustering II

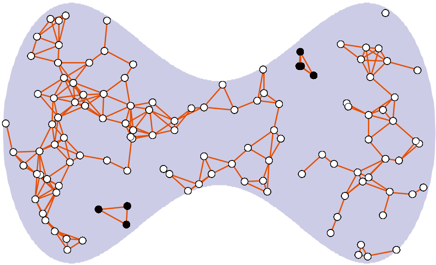
Do the minimizers of

$$GC_{n,\varepsilon_n}(u^n) = \frac{1}{n} \frac{\frac{1}{\varepsilon_n} \sum_{i,j} \eta_{\varepsilon_n}(X_i - X_j) \left|u_i^n - u_j^n\right|}{\min_{c\in\mathbb{R}} \sum_i \left|u_i^n - c\right|}$$

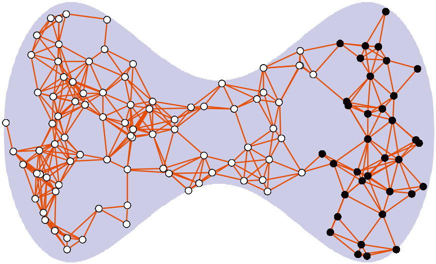
converge as the number of data points $n \to \infty$ to a minimizer of

$$C(u) = \frac{TV(u, \rho^2)}{\min_{c\in\mathbb{R}} \int_D |u(x) - c| \rho(x) dx} \quad ?$$

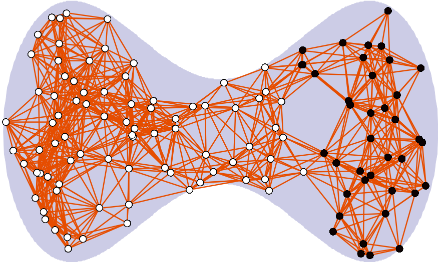
Question 1: For what scaling of $\varepsilon(n)$ can this hold?
Question 2: What is the topology for which $u^n \longrightarrow u$?



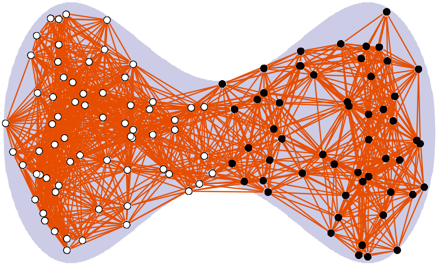
$n = 120, \varepsilon = 0.15$



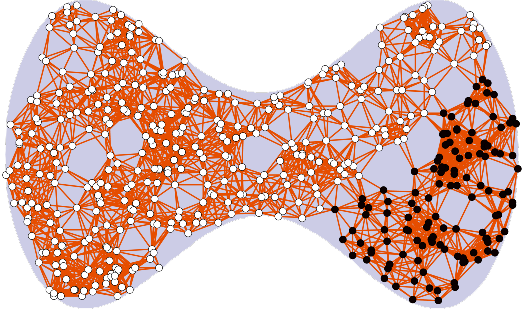
$n = 120, \varepsilon = 0.20$



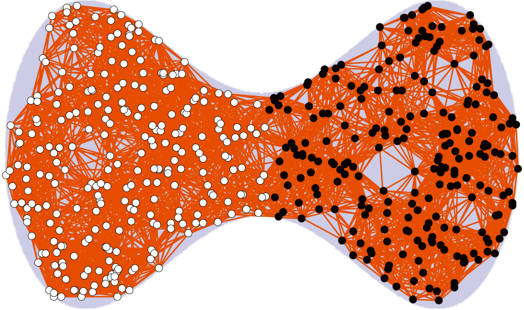
$n = 120, \varepsilon = 0.30$



$n = 120, \varepsilon = 0.40$



$n = 500, \varepsilon = 0.14$



$n = 500, \varepsilon = 0.2$

What was known

Consistency results in statistics/machine learning

- *Arias Castro, Pelletier, and Pudlo 2012* - partial results on the problem
- *Pollard 1981* - k -means
- *Hartigan 1981* - single linkage
- *Belkin and Niyogi 2006* - Laplacian eigenmaps
- *von Luxburg, Belkin, and Bousquet 2004, 2008* - spectral embedding
- *Chaudhuri and Dasgupta 2010* - cluster tree

Consistency results in statistics/machine learning

- *Arias Castro, Pelletier, and Pudlo 2012* - partial results on the problem
- *Pollard 1981* - k -means
- *Hartigan 1981* - single linkage
- *Belkin and Niyogi 2006* - Laplacian eigenmaps
- *von Luxburg, Belkin, and Bousquet 2004, 2008* - spectral embedding
- *Chaudhuri and Dasgupta 2010* - cluster tree

Calculus of Variations

Discrete to continuum for functionals on grids: *Braides 2010, Braides and Yip 2012, Chambolle, Giacomini and Lussardi 2012, Gobbino and Mora 2001, Van Gennip and Bertozzi 2014*

Γ-Convergence

(Y, d_Y) - metric space, $F_n : Y \to [0, \infty]$

Definition

The sequence $\{F_n\}_{n\in\mathbb{N}}$ Γ-**converges** (w.r.t d_Y) to $F : Y \to [0, \infty]$ if:

Liminf inequality: For every $y \in Y$ and whenever $y_n \to y$

$$\liminf_{n\to\infty} F_n(y_n) \geq F(y),$$

Limsup inequality: For every $y \in Y$ there exists $y_n \to y$ such that

$$\limsup_{n\to\infty} F_n(y_n) \leq F(y).$$

Definition (Compactness property)

$\{F_n\}_{n\in\mathbb{N}}$ satisfies the **compactness property** if

$$\left.\begin{array}{l} \{y_n\}_{n\in\mathbb{N}} \text{ bounded and} \\ \{F_n(y_n)\}_{n\in\mathbb{N}} \text{ bounded} \end{array}\right\} \Longrightarrow \{y_n\}_{n\in\mathbb{N}} \text{ has convergent subsequence}$$

Proposition: Convergence of minimizers

Γ-convergence and Compactness imply: If y_n is a minimizer of F_n and $\{y_n\}_{n \in N}$ is bounded in Y then along a subsequence

$$y_n \to y \qquad \text{as } n \to \infty$$

and

$$y \text{ is a minimizer of } F.$$

In particular, if F has a unique minimizer, then a sequence $\{y_n\}_{n \in \mathbb{N}}$ converges to the unique minimizer of F.

Consistency of clustering III

Show that

$$GC_{n,\varepsilon_n}(u^n) = \frac{1}{n} \frac{\frac{1}{\varepsilon_n}\sum_{i,j} \eta_{\varepsilon_n}(X_i - X_j)\,|u_i^n - u_j^n|}{\min_{c\in\mathbb{R}} \sum_i |u_i^n - c|}$$

Γ-converge as the number of data points $n \to \infty$, and $\varepsilon_n \to 0$ at certain rate to

$$F(u) = \frac{\sigma\, TV(u, \rho^2)}{\min_{c\in\mathbb{R}} \int_D |u(x) - c|\rho(x)dx}$$

and show that compactness property holds.

Questions

1. For what scaling of $\varepsilon(n)$ can this hold?
2. What is the topology for $u^n \longrightarrow u$?

Consistency of graph total variation

Show that

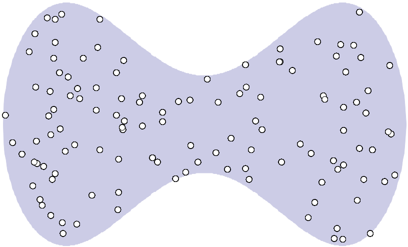
$$GTV_{n,\varepsilon_n}(u^n) = \frac{1}{\varepsilon_n\, n^2} \sum_{i,j} \eta_{\varepsilon_n}(X_i - X_j)\, |u_i^n - u_j^n|$$

Γ-converge to $\sigma TV(u, \rho^2)$, as the number of data points $n \to \infty$, and $\varepsilon_n \to 0$ at certain rate and show that compactness property holds.

Questions

1. For what scaling of $\varepsilon(n)$ can this hold?
2. What is the topology for $u^n \longrightarrow u$?

Consider domain D and $V_n = \{X_1, \ldots, X_n\}$ random i.i.d points.

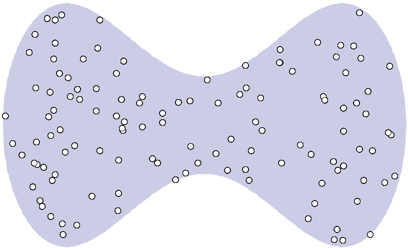


- How to compare $u_n : V_n \to \mathbb{R}$ and $u : D \to \mathbb{R}$ in a way consistent with L^1 topology?

Note that $u \in L^1(\nu)$ and $u_n \in L^1(\nu_n)$, where $\nu_n = \frac{1}{N} \sum_{i=1}^{n} \delta_{X_i}$.

Topology

Consider domain D and $V_n = \{X_1, \ldots, X_n\}$ random i.i.d points.

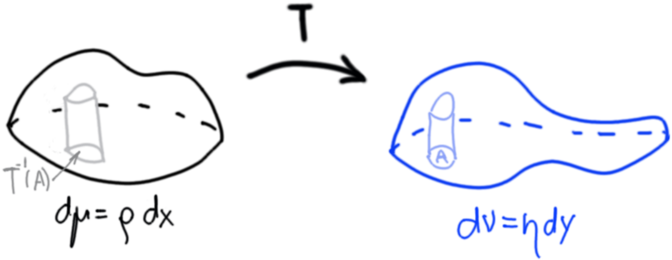


- How to compare $u_n \in L^1(\nu_n)$ and $u \in L^1(D)$ in a way consistent with L^1 topology?

- Let μ and ν be probability measures.
- Assume that all measures are supported in $B(0, R)$ for some large R.
- $X = \text{supp}(\mu)$, $Y = \text{supp}(\nu)$.

Transport map. $T : X \to Y$,

$$T_\sharp \mu = \nu, \qquad \text{that is } \forall A \text{ measurable } \mu(T^{-1}(A)) = \nu(A)$$

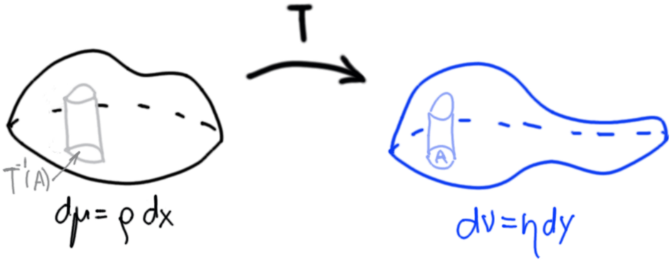


Transport map

- Let μ and ν be probability measures.
- $X = \text{supp}(\mu)$, $Y = \text{supp}(\nu)$.

Transport map. $T : X \to Y$,

$$T_\sharp \mu = \nu, \qquad \text{that is } \forall A \text{ measurable } \mu(T^{-1}(A)) = \nu(A)$$

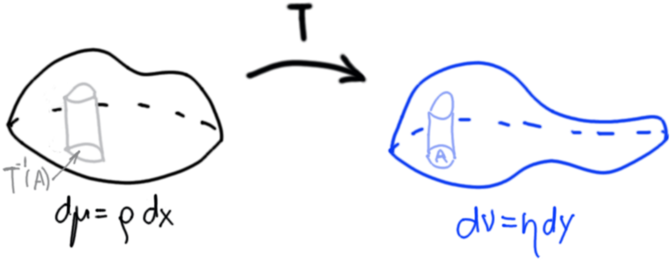


$$\int_{T^{-1}(A)} \rho(x) dx = \int_A \eta(y) dy$$

Transport map

Transport map. $T : X \to Y$,

$$T_\sharp \mu = \nu, \qquad \text{that is } \forall A \text{ measurable } \mu(T^{-1}(A)) = \nu(A)$$

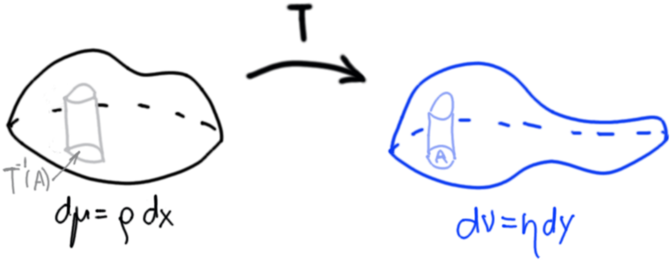


$$\int_{T^{-1}(A)} \rho(x) dx = \int_A \eta(y) dy = \int_{T^{-1}(A)} \eta(T(x)) \, |\det(DT(x)| dx$$

Transport map

Transport map. $T : X \to Y$,

$$T_\sharp \mu = \nu, \qquad \text{that is } \forall A \text{ measurable } \mu(T^{-1}(A)) = \nu(A)$$



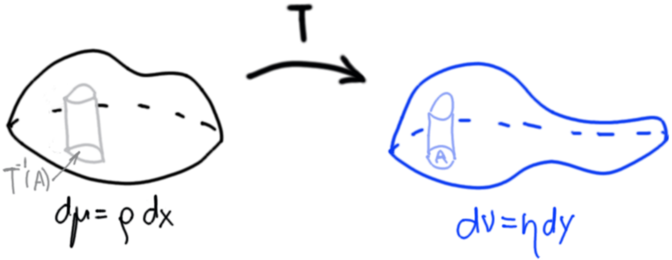
$$\int_{T^{-1}(A)} \rho(x)dx = \int_A \eta(y)dy = \int_{T^{-1}(A)} \eta(T(x))\,|\det(DT(x)|dx$$

$$\rho(x) = \eta(T(x))\,|\det(DT(x)|$$

Transport map

Transport map. $T : X \to Y$,

$$T_\sharp \mu = \nu, \qquad \text{that is } \forall A \text{ measurable } \mu(T^{-1}(A)) = \nu(A)$$

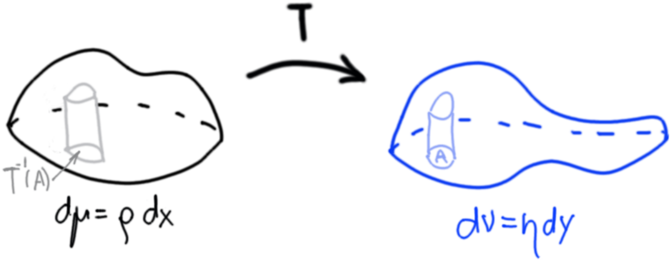


Change of variables: $y = T(x)$, for $f = \chi_A$, using $\chi_{T^{-1}(A)}(x) = \chi_A \circ T(x)$

$$\int_Y f(y) d\nu(y) = \nu(A) = \mu(T^{-1}(A)) = \int_X f(T(x))\, d\mu(x)$$

Transport map. $T : X \to Y$,

$$T_\sharp \mu = \nu, \qquad \text{that is } \forall A \text{ measurable } \mu(T^{-1}(A)) = \nu(A)$$



Change of variables: $y = T(x)$, for all $f \in L^1(d\nu)$

$$\int_Y f(y) d\nu(y) = \int_X f(T(x))\, d\mu(x)$$

- $c(x, y)$ cost of transporting unit mass from x to y
- Assume c is nonnegative and continuous
- Typically $c(x, y) = c(|x - y|)$, in particular $c(x, y) = |x - y|^p$, $p \geq 1$

Transport cost: Let T be a transport map, $T_\sharp \mu = \nu$

$$C(T) = \int_X c(x, T(x))\, d\mu(x)$$

Monge 1781

Optimal Transport Cost: Given μ and ν

$$OT_{c,M}(\mu,\nu) = \inf_{\{T\,:\,T_\sharp\mu=\nu\}} \int_X c(|x - T(x)|)d\mu(x)$$

Q1: Is the set of transport maps, T, nonempty?
Q2: Is infimum a minimum?

Monge 1781

Optimal Transport Cost: Given μ and ν

$$OT_{c,M}(\mu,\nu) = \inf_{\{T\,:\,T_\sharp\mu=\nu\}} \int_X c(|x-T(x)|)d\mu(x)$$

Q1: Is the set of transport maps, T, nonempty? Yes, if $d\mu = \rho dx$.
Q2: Is infimum a minimum?

Monge 1781

Optimal Transport Cost: Given μ and ν

$$OT_{c,M}(\mu,\nu) = \inf_{\{T\,:\,T_\sharp\mu=\nu\}} \int_X c(|x - T(x)|)d\mu(x)$$

Q1: Is the set of transport maps, T, nonempty? Yes, if $d\mu = \rho dx$.
Q2: Is infimum a minimum? Yes, if c is convex.

Transport Plan

Kantorovich 1942

- Let μ and ν be probability measures.
- $X = \text{supp}(\mu)$, $Y = \text{supp}(\nu)$.

Transport plans, π are probability measures on $X \times Y$ with first marginal μ and second marginal ν:

$$\Pi(\mu, \nu) = \{\pi \in \mathcal{P}(X \times Y) \; : \; \pi(A \times Y) = \mu(A), \pi(X \times A) = \nu(A)\}.$$

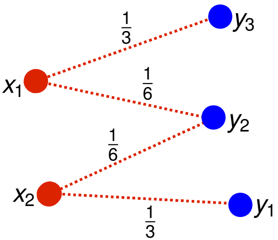
- $\pi(A \times B)$ mass originally in A which is sent to B.
- Unlike with transport maps, the mass can be split
- Note that $\Pi(\mu, \nu)$ is a convex set

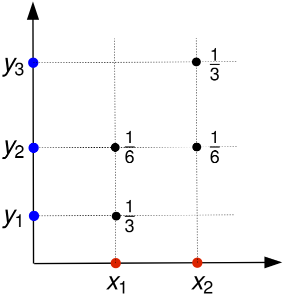
Transport plans, π are probability measures on $X \times Y$ with first marginal μ and second marginal ν:

$$\Pi(\mu,\nu) = \{\pi \in \mathcal{P}(X \times Y) \ : \ \pi(A \times Y) = \mu(A),\, \pi(X \times A) = \nu(A)\}.$$

$\mu = \frac{1}{2}\delta_{x_1} + \frac{1}{2}\delta_{x_2}$,

$\nu = \frac{1}{3}\delta_{y_1} + \frac{1}{3}\delta_{y_2} + \frac{1}{3}\delta_{y_3}$.





$$\pi = \frac{1}{3}\delta_{x_1,y_1} + \frac{1}{6}\delta_{x_1,y_2} + \frac{1}{6}\delta_{x_2,y_2} + \frac{1}{3}\delta_{x_2,y_3}$$

Transport Plan

Transport plans, π are probability measures on $X \times Y$ with first marginal μ and second marginal ν:

$$\Pi(\mu, \nu) = \{\pi \in \mathcal{P}(X \times Y) \ : \ \pi(A \times Y) = \mu(A), \pi(X \times A) = \nu(A)\}.$$

From a map to a plan: Let T be a transport map: $T_\sharp \mu = \nu$. Then $\pi = (I \times T)_\sharp \mu$ is a transport plan. Here $(I \times T)(x) = (x, T(x))$.

Optimal Transport Cost - Kantorovich Formulation

- $c(x, y)$ cost of transporting unit mass from x to y
- Assume c is nonnegative and continuous
- Typically $c(x, y) = c(|x - y|)$, in particular $c(x, y) = |x - y|^p$, $p \geq 1$

Transport cost: Let π be a transport plan, $\pi \in \Pi(\mu, \nu)$

$$C(\pi) = \int_{X \times Y} c(x, y)\, d\pi(x, y)$$

Optimal Transport Cost: Given μ and ν

$$OT_{c,K}(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} c(x, y)\, d\pi(x, y)$$

Q1: Is the set of transport plans, nonempty?

Q2: Is infimum a minimum?

Optimal Transport Cost - Kantorovich Formulation

- $c(x, y)$ cost of transporting unit mass from x to y
- Assume c is nonnegative and continuous
- Typically $c(x, y) = c(|x - y|)$, in particular $c(x, y) = |x - y|^p$, $p \geq 1$

Transport cost: Let π be a transport plan, $\pi \in \Pi(\mu, \nu)$

$$C(\pi) = \int_{X \times Y} c(x, y) \, d\pi(x, y)$$

Optimal Transport Cost: Given μ and ν

$$OT_{c,K}(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} c(x, y) \, d\pi(x, y)$$

Q1: Is the set of transport plans, nonempty? Yes, take $\pi = \mu \times \nu$.
Q2: Is infimum a minimum?

Optimal Transport Cost - Kantorovich Formulation

- $c(x, y)$ cost of transporting unit mass from x to y
- Assume c is nonnegative and continuous
- Typically $c(x, y) = c(|x - y|)$, in particular $c(x, y) = |x - y|^p$, $p \geq 1$

Transport cost: Let π be a transport plan, $\pi \in \Pi(\mu, \nu)$

$$C(\pi) = \int_{X \times Y} c(x, y)\, d\pi(x, y)$$

Optimal Transport Cost: Given μ and ν

$$OT_{c,K}(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} c(x, y)\, d\pi(x, y)$$

Q1: Is the set of transport plans, nonempty? Yes, take $\pi = \mu \times \nu$.
Q2: Is infimum a minimum? Yes. Note $\Pi(\mu, \nu)$ is a convex set, transport cost is a linear function of π.

- Assume $X = \text{supp}(\mu)$, $Y = \text{supp}(\nu)$ are compact

Optimal Transportation Distance: Given μ and ν, and $p \in [1, \infty)$

$$d_p(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} |x - y|^p \, d\pi(x, y) \right)^{\frac{1}{p}}$$

- d_p is a metric on $\mathcal{P}(K)$ for any K compact.
- d_p metrizes weak convergence of measures on $\mathcal{P}(K)$.
- d_2 is known as the Wasserstein distance.

Optimal Transportation for $p = \infty$

$\infty-$transportation distance:

$$d_\infty(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \text{esssup}_\pi \{|x - y| \ : \ x \in X, y \in Y\}$$

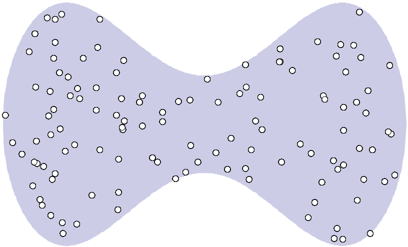
- There exists a minimizer $\pi \in \Pi(\mu, \nu)$.
- If $\mu = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ and $\nu = \frac{1}{n} \sum_{j=1}^n \delta_{y_j}$ then

$$d_\infty(\mu, \nu) = \min_{\sigma-\text{permutation}} \max_i |x_i - y_{\sigma(i)}|.$$

- If μ has density then OT map, T exists (Champion, De Pascale, Juutinen 2008) and then

$$d_\infty(\mu, \nu) = \|T - Id\|_{L^\infty(\mu)}.$$

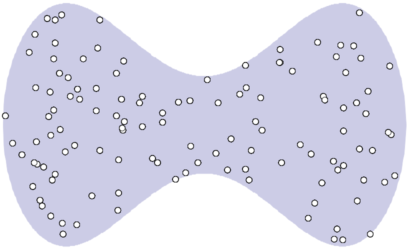
Consider domain D and $V_n = \{X_1, \ldots, X_n\}$ random i.i.d points.



- How to compare $u_n : V_n \to \mathbb{R}$ and $u : D \to \mathbb{R}$ in a way consistent with L^1 topology?

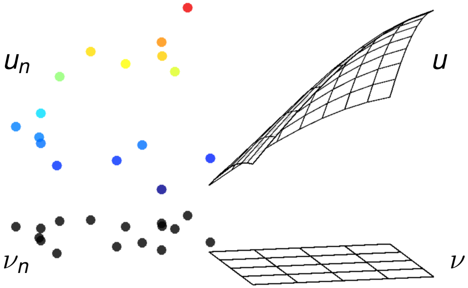
Note that $u \in L^1(\nu)$ and $u_n \in L^1(\nu_n)$, where $\nu_n = \frac{1}{N} \sum_{i=1}^{n} \delta_{X_i}$.

Consider domain D and $V_n = \{X_1, \ldots, X_n\}$ random i.i.d points.



- How to compare $u_n \in L^1(\nu_n)$ and $u \in L^1(D)$ in a way consistent with L^1 topology?

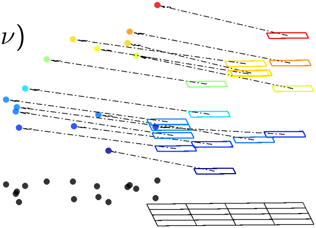
Consider domain D and $V_n = \{X_1, \ldots, X_n\}$ random i.i.d points.



- How to compare $u_n \in L^1(\nu_n)$ and $u \in L^1(D)$ in a way consistent with L^1 topology?

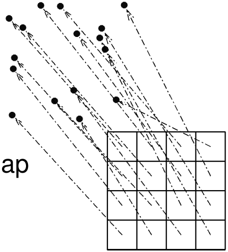
Composition

$u_n \circ T_n \in L^p(\nu)$



Transport map

$T_{n\sharp}\nu = \nu_n$



$$d^p_{TL^p}((\nu, u), (\nu_n, u_n)) = \inf_{T_{n\sharp}\nu = \nu_n} \int_D |u_n(T_n(x)) - u(x)|^p + |T_n(x) - x|^p \rho(x) dx$$

Topology

For $u \in L^1(\nu)$ and $u_n \in L^1(\nu_n)$

$$d((\nu, u), (\nu_n, u_n)) = \inf_{T_{n\sharp}\nu = \nu_n} \int_D |u_n(T_n(x)) - u(x)| + |T_n(x) - x| \rho(x) dx$$

where

$$T_{n\sharp}\nu = \nu_n$$

TL^1 Space

Definition

$$TL^p = \{(\nu, f) \ : \ \nu \in \mathcal{P}(D), \ f \in L^p(\nu)\}$$

$$d^p_{TL^p}((\nu, f), (\sigma, g)) = \inf_{\pi \in \Pi(\nu,\sigma)} \int_{D \times D} |y - x|^p + |g(y) - f(x))|^p d\pi(x, y).$$

where

$$\Pi(\nu, \sigma) = \{\pi \in \mathcal{P}(D \times D) \ : \ \pi(A \times D) = \nu(A), \ \pi(D \times A) = \sigma(A)\}.$$

If $T_\sharp \nu = \sigma$ then $\pi = (I \times T)_\sharp \nu \in \Pi(\nu, \sigma)$ and the integral becomes

$$\int |T(x) - x|^p + |g(T(x)) - f(x)|^p d\nu(x)$$

Lemma

(TL^p, d_{TL^p}) is a metric space.

TL^p convergence

- $(\nu, f_n) \xrightarrow{TL^p} (\nu, f)$ iff $f_n \xrightarrow{L^1(\nu)} f$
- $(\nu_n, f_n) \xrightarrow{TL^p} (\nu, f)$ iff the measures $(I \times f_n)_\sharp \nu_n$ weakly converge to $(I \times f)_\sharp \nu$. That is if graphs, considered as measures converge weakly.
- The space TL^p is not complete. Its completion are the probability measures on the product space $D \times \mathbb{R}$.

If $(\nu_n, f_n) \xrightarrow{TL^p} (\nu, f)$ then there exists a sequence of transportation plans ν_n such that

$$\int_{D\times D} |x-y|^p d\pi_n(x,y) \longrightarrow 0 \quad \text{as } n \to \infty. \tag{1}$$

We call a sequence of transportation plans $\pi_n \in \Pi(\nu_n, \nu)$ **stagnating** if it satisfies (1).

Stagnating sequence: $\int_{D \times D} |x - y| d\pi_n(x, y) \longrightarrow 0$

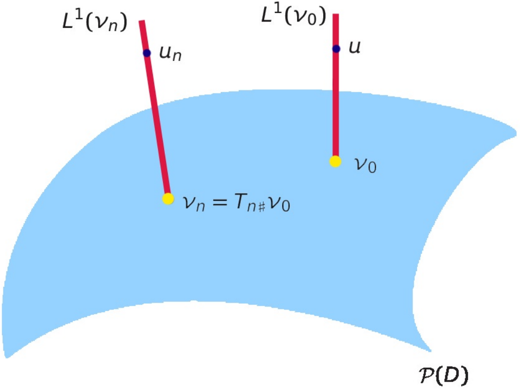
TFAE:

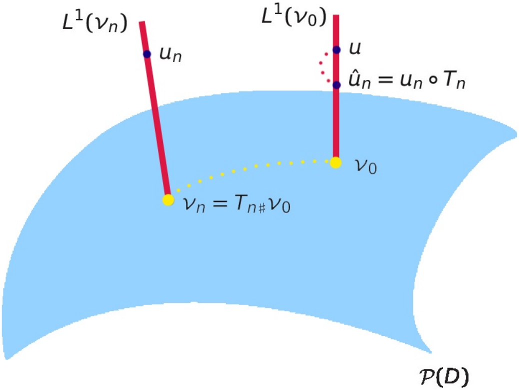
1. $(\nu_n, f_n) \xrightarrow{TL^p} (\nu, f)$ as $n \to \infty$.
2. $\nu_n \rightharpoonup \nu$ and **there exists** a stagnating sequence of transportation plans $\{\pi_n\}_{n \in \mathbb{N}}$ for which

$$\iint_{D \times D} |f(x) - f_n(y)|^p \, d\pi_n(x, y) \to 0, \ as\ n \to \infty. \tag{2}$$

3. $\nu_n \rightharpoonup \nu$ and **for every** stagnating sequence of transportation plans π_n, (2) holds.

Formally $TL^p(D)$ is a fiber bundle over $\mathcal{P}(D)$.




$L^1(\nu_n)$
u_n
$L^1(\nu_0)$
u
$\hat{u}_n = u_n \circ T_n$
ν_0
$\nu_n = T_{n\#}\nu_0$
$\mathcal{P}(D)$

Composition in TL^p space

Lemma

Let $p \geq 1$ and let $\{\nu_n\}_{n\in\mathbb{N}}$ and ν be Borel probability measures on $\mathbb{R}^d$ with finite second moments. Let $F_n \in L^p(\nu_n, \mathbb{R}^d, \mathbb{R}^k)$ and $F \in L^p(\nu, \mathbb{R}^d, \mathbb{R}^k)$. Consider the measures $\tilde{\nu}_n, = F_{n\sharp}\nu_n$ and $\tilde{\nu}, = F_\sharp \nu$. Finally, let $\tilde{f}_n \in L^p(\tilde{\nu}_n, \mathbb{R}^k, \mathbb{R})$ and $\tilde{f} \in L^p(\tilde{\nu}, \mathbb{R}^k, \mathbb{R})$. If

$$(\nu_n, F_n) \xrightarrow{TL^p} (\nu, F) \quad \text{as } n \to \infty,$$

and

$$(\tilde{\nu}_n, \tilde{f}_n) \xrightarrow{TL^p} (\tilde{\nu}, \tilde{f}) \quad \text{as } n \to \infty.$$

Then,

$$(\nu_n, \tilde{f}_n \circ F_n) \xrightarrow{TL^p} (\nu, \tilde{f} \circ F_n) \quad \text{as } n \to \infty.$$

Consistency

$$GTV_{n,\varepsilon_n}(u^n) = \frac{1}{\varepsilon_n n^2} \sum_{i,j} \eta_{\varepsilon_n}(X_i - X_j) \, |u_i^n - u_j^n|$$

Γ-convergence of Total Variation ($d \geq 3$ García Trillos and S. '16, $d = 2$ Penrose and Müller '19, see also Caroccia, Chambolle and S. '20)

Let $\{\varepsilon_n\}_{n\in\mathbb{N}}$ be a sequence of positive numbers converging to 0 satisfying

$$\lim_{n\to\infty} \frac{(\log n)^{1/d}}{n^{1/d}} \frac{1}{\varepsilon_n} = 0 \ \ \textit{if } d \geq 3.$$

Then, GTV_{n,ε_n} Γ-converge to $\sigma \, TV(\,\cdot\,, \rho^2)$ as $n \to \infty$ in the TL^1 sense, where σ depends explicitly on η.

Typical degree $\gg$ log n. If Typical degree $<$ log n then graph becomes disconnected.

Consistency

Γ-convergence of Perimeter

The conclusions hold when all of the functionals are restricted to characteristic functions of sets. That is, the graph perimeters Γ-converge to the continuum perimeter.

Compactness

With the same conditions on ε_n as before, if

$$\sup_{n\in\mathbb{N}} \|u_n\|_{L^1(D,\nu_n)} < \infty,$$

and

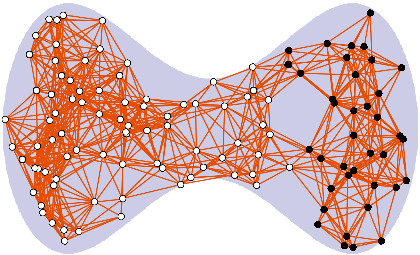
$$\sup_{n\in\mathbb{N}} GTV_{n,\varepsilon_n}(u_n) < \infty,$$

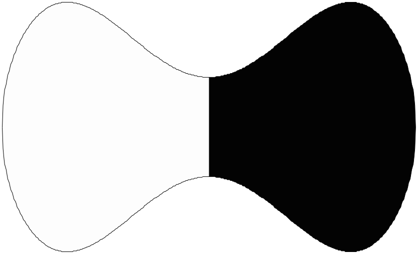
then $\{u_n\}_{n\in N}$ is TL^1-precompact.

Recall:

$$GC_{n,\varepsilon_n}(u^n) = \frac{1}{n} \frac{\frac{1}{\varepsilon_n} \sum_{i,j} \eta_{\varepsilon_n}(X_i - X_j) \left| u_i^n - u_j^n \right|}{\min_{c \in \mathbb{R}} \sum_i \left| u_i^n - c \right|}$$

$$C(u) = \frac{\sigma TV(u, \rho^2)}{\min_{c \in \mathbb{R}} \int_D \left| u(x) - c \right| \rho(x) dx}$$





Recall:

$$GC_{n,\varepsilon_n}(u^n) = \frac{1}{n}\frac{\frac{1}{\varepsilon_n}\sum_{i,j}\eta_{\varepsilon_n}(X_i - X_j)\,|u_i^n - u_j^n|}{\min_{c\in\mathbb{R}}\sum_i |u_i^n - c|}$$

$$C(u) = \frac{\sigma\, TV(u,\rho^2)}{\min_{c\in\mathbb{R}}\int_D |u(x)-c|\rho(x)dx}$$

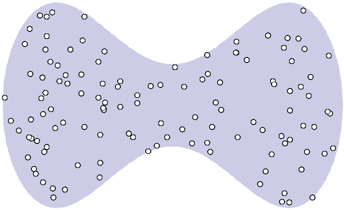
Consistency of Cheeger Cuts (von Brecht, García Trillos, Laurent, S.)

For the same conditions on ε_n as before, with probability one:

$$GC_{n,\varepsilon_n} \xrightarrow{\Gamma} C \qquad \text{w.r.t. } TL^1 \text{ metric.}$$

Moreover, for any sequence of sets $E_n \subseteq \{X_1,\dots,X_n\}$ of almost minimizers of the Cheeger energy, every subsequence has a convergent subsequence (in the TL^1 sense) to a minimizer of the Cheeger energy on the domain D.

Optimal matchings in dimension $\mathbf{d} \geq \mathbf{3}$: *Ajtai-Komlós-Tusnády (1983), Yukich and Shor (1991), Garcia Trillos and S. (2014)*



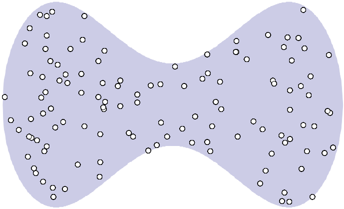
Theorem

There are constants $c > 0$ and $C > 0$ (depending on d) such that with probability one we can find a sequence of transportation maps $\{T_n\}_{n \in \mathbb{N}}$ from ν_0 to ν_n ($T_{n\#}\nu_0 = \nu_n$) and such that:

$$c \leq \liminf_{n \to \infty} \frac{n^{1/d} \| Id - T_n \|_\infty}{(\log n)^{1/d}} \leq \limsup_{n \to \infty} \frac{n^{1/d} \| Id - T_n \|_\infty}{(\log n)^{1/d}} \leq C.$$

∞-OT between a measure and its random sample

Optimal matchings in dimension $\mathbf{d} = \mathbf{2}$: *Leighton and Shor (1986), new proof by Talagrand (2005), Garcia Trillos and S. (2014)*



Theorem

There are constants $c > 0$ and $C > 0$ such that with probability one we can find a sequence of transportation maps $\{T_n\}_{n\in\mathbb{N}}$ from ν_0 to ν_n $(T_{n\#}\nu_0 = \nu_n)$ and such that:

$$c \le \liminf_{n\to\infty} \frac{n^{1/2}\|Id - T_n\|_\infty}{(\log n)^{3/4}} \le \limsup_{n\to\infty} \frac{n^{1/2}\|Id - T_n\|_\infty}{(\log n)^{3/4}} \le C. \tag{3}$$

Consistency: Other point sets

$$GTV_{n,\varepsilon_n}(u^n) = \frac{1}{\varepsilon_n\, n^2} \sum_{i,j} \eta_{\varepsilon_n}(X_i - X_j)\, |u_i^n - u_j^n|$$

Γ-convergence of and Compactness for Graph Total Variation

Assume $d_\infty(\nu_n, \nu) \to 0$ as $n \to \infty$. Let $\{\varepsilon_n\}_{n\in\mathbb{N}}$ be a sequence of positive numbers converging to 0 satisfying

$$\lim_{n\to\infty} \frac{d_\infty(\nu_n, \nu)}{\varepsilon_n} = 0$$

Then, GTV_{n,ε_n} Γ-converge to $\sigma\, TV(\,\cdot\,, \rho^2)$ as $n \to \infty$ in the TL^1 sense, where σ depends explicitly on η.

Furthermore if $\|u_n\|_{L^1(D,\nu_n)}$ and $GTV_{n,\varepsilon_n}(u_n)$ are uniformly bounded the sequence $\{u_n\}_{n\in N}$ is TL^1-precompact.

Hint about the proof

Assume that $u_n \xrightarrow{TL^1} u$ as $n \to \infty$.
There exists $T_{n\sharp}\nu = \nu_n$ stagnating (i.e. $\int |x - T_n(x)| d\nu(x) \to 0$).

$$GTV_{n,\varepsilon_n}(u_n) = \frac{1}{\varepsilon_n} \int_{D\times D} \eta_{\varepsilon_n}(\tilde{x} - \tilde{y}))\, |u_n(\tilde{x}) - u_n(\tilde{y})|\, d\nu_n(\tilde{x}) d\nu_n(\tilde{y})$$
$$= \frac{1}{\varepsilon_n} \int_{D\times D} \eta_{\varepsilon_n}(T_n(x) - T_n(y))\, |u_n \circ T_n(x) - u_n \circ T_n(y)|\, \rho(x)\rho(y) dxdy$$

Define $TV_\varepsilon(u;\rho) := \dfrac{1}{\varepsilon} \displaystyle\int_{D\times D} \eta_\varepsilon(x-y)|u(x) - u(y)|\rho(x)\rho(y) dxdy.$

- $TV_\varepsilon \xrightarrow{\Gamma} TV(\,\cdot\,, \rho^2)$ wrt $L^1(\nu)$ metric.
  (*Alberti-Bellettini, Ponce, Chambolle-Giacomini-Lussardi, Savin-Valdinocci*)
- If $|T_n(x) - x| \ll \varepsilon_n$ then one may be able to compare $GTV_{n,\varepsilon_n}(u_n)$ and $TV_\varepsilon(u_n \circ T_n; \rho)$.

Sketch for liminf part

Assume $\eta = \chi_{B(0,1)}$. Assume $u_n \xrightarrow{TL^1} u$ as $n \to \infty$. Since $T_{n\sharp}\nu = \nu_n$,

$$GTV_{n,\varepsilon_n}(u_n) = \frac{1}{\varepsilon_n}\int_{D^2} \eta_{\varepsilon_n}(T_n(x) - T_n(y))\,|u_n \circ T_n(x) - u_n \circ T_n(y)|\,\rho(x)\rho(y)dxdy.$$

For almost every $(x, y) \in D \times D$ and n large

$$|T_n(x) - T_n(y)| > \varepsilon_n \Rightarrow |x - y| > \tilde{\varepsilon}_n := \varepsilon_n - 2\|Id - T_n\|_\infty > 0.$$

$$\eta\left(\frac{|x-y|}{\tilde{\varepsilon}_n}\right) \le \eta\left(\frac{|T_n(x) - T_n(y)|}{\varepsilon_n}\right).$$

Let $\tilde{u}_n = u_n \circ T_n$. For large enough n

$$\begin{aligned} GTV_{n,\varepsilon_n}(u_n) &\ge \frac{1}{\varepsilon_n^{d+1}}\int_{D\times D} \eta\left(\frac{|x-y|}{\tilde{\varepsilon}_n}\right)|\tilde{u}_n(x) - \tilde{u}_n(y)|\,\rho(x)\rho(y)dxdy \\ &= \left(\frac{\tilde{\varepsilon}_n}{\varepsilon_n}\right)^{d+1} TV_{\tilde{\varepsilon}_n}(\tilde{u}_n;\rho). \end{aligned}$$

Now use $\frac{\tilde{\varepsilon}_n}{\varepsilon_n} \to 1$ and that $u_n \xrightarrow{TL^1} u$ implies $\tilde{u}_n \xrightarrow{L^1(D)} u$ as $n \to \infty$.

Spectral Clustering

- $V_n = \{X_1, \dots, X_n\}$, similarity matrix W, as before:

$$W_{ij} := \frac{1}{\varepsilon^d} \eta \left(\frac{|X_i - X_j|}{\varepsilon} \right).$$

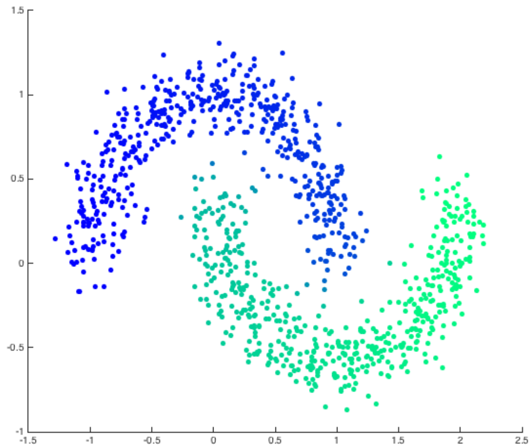
  The weighted degree of a vertex is $d_i = \sum_j W_{i,j}$.
- Dirichlet energy of $u^n : V_n \to \mathbb{R}$ is

$$F(u) = \frac{1}{2} \sum_{i,j} W_{ij} |u^n(X_i) - u^n(X_j)|^2.$$

- Associated operator is the graph laplacian $L = D - W$, where $D = \mathrm{diag}(d_1, \dots, d_n)$.
- To partition the point cloud into two clusters, consider the eigenvector corresponding to second eigenvalue:

$$u_2^n := \arg\min \left\{ \sum_{i,j} W_{ij} |u(X_i) - u(X_j)|^2 \ : \ \sum_i u(X_i) = 0, \ \|u\|_2 = 1 \right\}$$

Spectral Clustering: Two moons (easy)



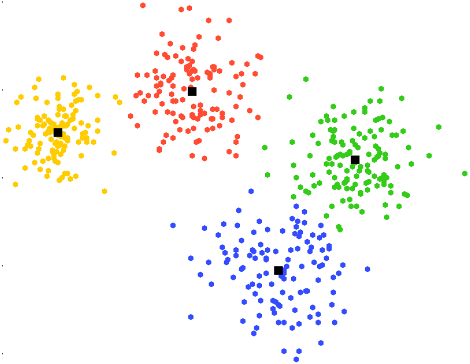
1D embedding: $x_i \mapsto u_2(x_i)$

k-means clustering

Given $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ find a set of k points $A = \{a_1, \ldots, a_k\}$ which minimizes

$$\min_A \frac{1}{n} \sum_{i=1}^{n} \mathrm{dist}(X_i, A)^2$$

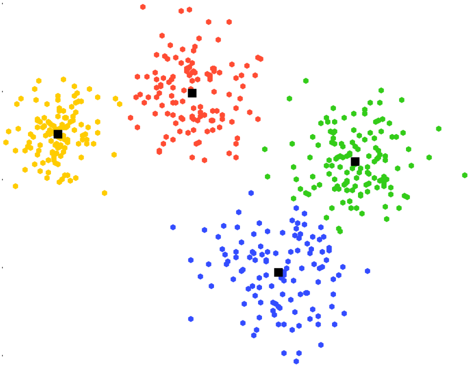
where $\mathrm{dist}(x, A) = \min_{a \in A} |x - a|$.



k-means clustering

Given $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ and $\mu_n = \frac{1}{n}\delta_{x_i}$. Find a set of k points $A = \{a_1, \ldots, a_k\}$ which minimizes

$$\min_{A} \inf_{\mathrm{supp}(\xi)\subseteq A} d_2(\mu_n, \xi).$$



Spectral Clustering

Shi and Malik, 2000, Ng, Jordan, Weiss 2001

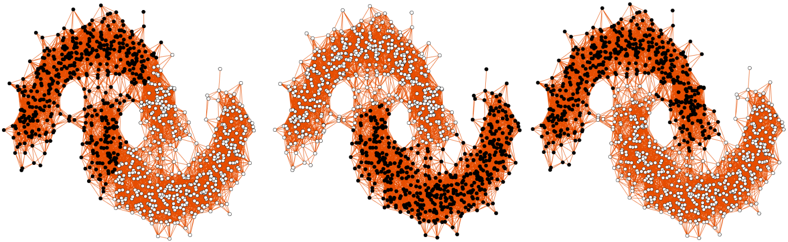
Input: Number of clusters k and similarity matrix W.

- Construct the graph Laplacian L.
- Compute the eigenvectors $u_2^n, \dots, u_k^n$ of L associated to the k smallest (nonzero) eigenvalues of L.
- Nonlinear transformation

$$X_i \mapsto Y_i = [u_1^n(X_i), \dots, u_k^n(X_i)]^T . \in \mathbb{R}^k, \qquad \text{for } i = 1, \dots, n.$$

- Use the k-means algorithm to partition the set of points $\{Y_1, \dots, Y_n\}$ into k groups, that we denote by $G_1, \dots, G_k$.

Output: Clusters $G_1, \dots, G_k$.



(a) k - means (b) spectral (c) Cheeger cut *von Luxburg, Belkin, Bousquet '08, Belkin-Nyogi '07, Ting, Huang, Jordan '10, Singer, Wu '13, Burago, Ivanov, Kurylev '14, Shi, Sun '15*

$$u_k^n = \arg\min \left\{ \sum_{i,j} W_{ij} |u(X_i) - u(X_j)|^2 \ : \ \sum_i u(X_i) u_m^n(X_i) = 0 \ (\forall m < k), \|u\|_2 = 1 \right\}$$

Theorem [García Trillos and S.] Suppose $X_1, \dots, X_n, \dots$ are i.i.d samples of a distribution with density ρ. Then, for $\varepsilon_n \to 0$

$$u_k^n \xrightarrow{TL^2} u^k$$

where u_k is eigenfunction, corresponding to k-th eigenvalue, of

$$L_c(u_k) := -\frac{1}{\rho} \operatorname{div}(\rho^2 \nabla u_k) = \lambda_k u_k \quad \text{in } D$$

$$\frac{\partial u_k}{\partial n} = 0 \qquad \text{on } \partial D.$$

Spectral Convergence of Graph Laplacian II

$$u_k^n = \arg\min \left\{ \sum_{i,j} W_{ij} |u(X_i) - u(X_j)|^2 \; : \; \sum_i u(X_i) u_m^n(X_i) = 0 \; (\forall m < k), \|u\|_2 = 1 \right\}$$

- Suppose $X_1, \ldots, X_n, \ldots$ are i.i.d samples of a distribution with density ρ. Then, for $\varepsilon_n \to 0$ as before

$$u_k^n \xrightarrow{TL^2} u^k$$

where u_k is eigenfunction, corresponding to k-th eigenvalue, of

$$-\frac{1}{\rho} \operatorname{div}(\rho^2 \nabla u_k) = \lambda_k u_k \quad \text{in } D$$

$$\frac{\partial u_k}{\partial n} = 0 \qquad \text{on } \partial D.$$

Consistency of spectral clustering

Discrete Spectral Clustering:

- Construct the graph Laplacian L for the geometric graph of the sample
- Compute the eigenvectors $u_1^n, \ldots, u_k^n$ of L associated to the k smallest (nonzero) eigenvalues of L.
- Nonlinear transformation

$$X_i \mapsto Y_i^n = [u_1^n(X_i), \ldots, u_k^n(X_i)]^T. \in \mathbb{R}^k, \qquad \text{for } i = 1, \ldots, n.$$

- Use the k-means algorithm to partition the set of points $\{Y_1^n, \ldots, Y_k^n\}$ into k groups. We denote the resulting partitioning of V_n by $G_1^n, \ldots, G_k^n$.

Continuum Spectral Clustering:

- Compute the eigenvectors $u_1, \ldots, u_k$ of L_c associated to the k smallest (nonzero) eigenvalues of L_c.
- Consider the measure $\mu = (u_1, \ldots, u_k)_\sharp \nu$.
- Let $\tilde{G}_i \subset \mathbb{R}^k$ be the clusters obtained by k-means clustering of μ.
- $G_i = (u_1, \ldots, u_k)^{-1}(\tilde{G}_i)$ for $i = 1, \ldots, k$ define the *spectral clustering* of ν.

Consistency of spectral clustering

Discrete Spectral Clustering:

- Construct the graph Laplacian L for the geometric graph of the sample
- Compute the eigenvectors $u_1^n, \ldots, u_k^n$ of L associated to the k smallest (nonzero) eigenvalues of L.
- Nonlinear transformation

$$X_i \mapsto Y_i^n = [u_1^n(X_i), \ldots, u_k^n(X_i)]^T. \in \mathbb{R}^k, \qquad \text{for } i = 1, \ldots, n.$$

- Use the k-means algorithm to partition the set of points $\{Y_1^n, \ldots, Y_k^n\}$ into k groups. We denote the resulting partitioning of V_n by $G_1^n, \ldots, G_k^n$.

Theorem (Garciá–Trillos, S. '18)

Let $G_1^n, \ldots G_k^n$ be the clusters above. Let $\nu_i^n = \nu_{n \llcorner G_i^n}$ (the restriction of empirical measure to clusters) for $i = 1, \ldots, k$. Then $(\nu_1^n, \ldots, \nu_k^n)$ is precompact with respect to weak convergence of measures and converges along a subsequence to $(\nu_1, \ldots, \nu_k) = (\nu_{\llcorner G_1}, \ldots, \nu_{\llcorner G_k})$ where $G_1, \ldots, G_k$ is a continuum spectral clustering of ν.

Normalized Graph Laplacian

- As before: $W_{ij} := \frac{1}{\varepsilon^d}\,\eta\left(\frac{|X_i - X_j|}{\varepsilon}\right)$, $d_i = \sum_j W_{i,j} = \sum_j \eta_\varepsilon(|X_i - X_j|)$.
- Dirichlet energy of $u_n : V_n \to \mathbb{R}$ is

$$F(u) = \frac{1}{2}\sum_{i,j} W_{ij}\left(\frac{u_n(X_i)}{\sqrt{d_i}} - \frac{u_n(X_j)}{\sqrt{d_j}}\right)^2 .$$

- Associated operator is the normalized graph laplacian $D^{-1/2}LD^{-1/2} = I - D^{-1/2}WD^{-1/2}$, where $D = \operatorname{diag}(d_1, \ldots, d_n)$.
- To partition the point cloud into two clusters, consider the eigenvector corresponding to second eigenvalue:

$$u_n := \arg\min\left\{\sum_{i,j} W_{ij}\left|\frac{u_n(X_i)}{\sqrt{d_i}} - \frac{u_n(X_j)}{\sqrt{d_j}}\right|^2 \;:\; \sum_i u(X_i) = 0,\ \|u\|_2 = 1\right\}$$

Consistency of Normalized Graph Laplacian

$$u_k^n = \arg\min \left\{ \sum_{i,j} \left| \frac{u_n(X_i)}{\sqrt{d_i}} - \frac{u_n(X_j)}{\sqrt{d_j}} \right|^2 : \sum_i u(X_i) u_m^n(X_i) = 0 \ (\forall m < k), \|u\|_2 = 1 \right\}$$

- Suppose $X_1, \ldots, X_n, \ldots$ are i.i.d samples of a distribution with density ρ. Then, for $\varepsilon_n \to 0$ as before

$$u_k^n \xrightarrow{TL^2} u_k$$

where u_k is eigenfunction, corresponding to k-th eigenvalue, of

$$-\frac{1}{\rho^{3/2}} \nabla \cdot \left(\rho^2 \nabla \left(\frac{u_k}{\sqrt{\rho}} \right) \right) = \lambda_k u_k \quad \text{in } D$$
$$\frac{\partial (u_k / \sqrt{\rho})}{\partial n} = 0 \qquad \text{on } \partial D.$$

Consistency of Spectral Clustering in Manifold Setting

$\mathcal{M}$ compact manifold of dimension m. Data measure μ has density $d\mu = \rho d Vol_{\mathcal{M}}$.

$$\alpha \leq \rho \leq \frac{1}{\alpha} \quad \text{for some } \alpha > 0.$$

The continuum operator is a weighted Laplace-Beltrami operator

$$u \mapsto \frac{1}{\rho} \operatorname{div}_{\mathcal{M}}(\rho^2 \operatorname{grad} u).$$

This operator is symmetric with respect to $L^2(d\mu)$:

$$\|u\|^2_{L^2(d\mu)} = \int_{\mathcal{M}} u^2 d\mu.$$

It has a spectrum

$$0 = \lambda_1 < \lambda_2 \leq \lambda_3 \leq \cdots .$$

with corresponding orthornomal set of eigenfunctions u_k, $k = 1, \ldots$.

Consistency of Spectral Clustering in Manifold Setting

Techniques inspired by *Burago, Ivanov, Kurylev*

Theorem (García Trillos, Gerlach, Hein and S.)

There exists a constant $C_{m,K,\mathrm{Vol}(\mathcal{M}),i_0}$ such that for every $\beta > 1$ and every $n \in \mathbb{N}$ the following holds with probability at least $1 - C_{m,K,\mathrm{Vol}(\mathcal{M}),i_0} \cdot n^{-\beta}$.
For every $k \in \{1, \dots, n\}$ there exists a constant $C > 0$ depending on K, m, ρ, η, R and $\lambda_k(\mathcal{M})$ such that

$$\left| \frac{2}{n\varepsilon^2 \sigma_\eta} \lambda_k(\Gamma_n) - \lambda_k(\mathcal{M}) \right| \leq C \left(\varepsilon + \frac{\ell}{\varepsilon} \right),$$

whenever $\ell < \varepsilon < C^{-1}$.

Recent results by Calder and Garcia Trillos When $\varepsilon \geq n^{-\frac{1}{d+4}}$

$$\left| \frac{2}{n\varepsilon^2 \sigma_\eta} \lambda_k(\Gamma_n) - \lambda_k(\mathcal{M}) \right| \leq C\varepsilon.$$