Variational problems on random structures: analysis and applications to data science

Dejan Slepčev
Carnegie Mellon University

Santaló Summer School
Santander
August 13-17, 2018.

Collaborators:

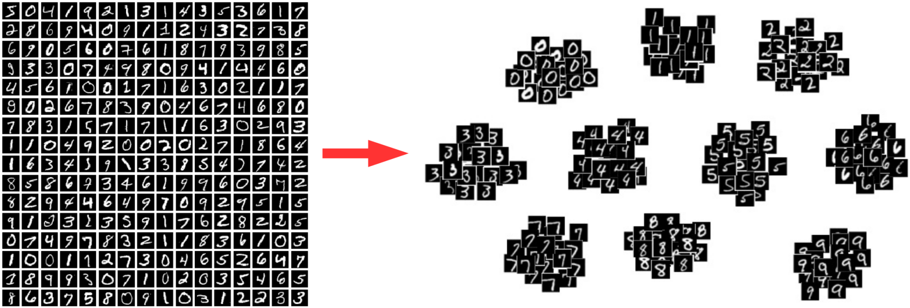
- *Variational approaches to clustering (TV and spectral)*
  Xavier Bresson (NTU Singapore), Nicolás García Trillos (Brown), Thomas Laurent (LMU), James von Brecht (Cal. State, Long Beach)
- *Error rates for graph laplacian*
  Nicolás García Trillos (Brown), Moritz Gerlach (Saarland), Matthias Hein (Saarland)
- *Semi-Supervised Lerning and Regression*
  Marco Caroccia (IST Lisbon), Antonin Chambolle (Ecolé Polytechnique), Matthew Dunlop (Caltech), Matthew Thorpe (Cambridge), Andrew Stuart (Caltech)

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, ...

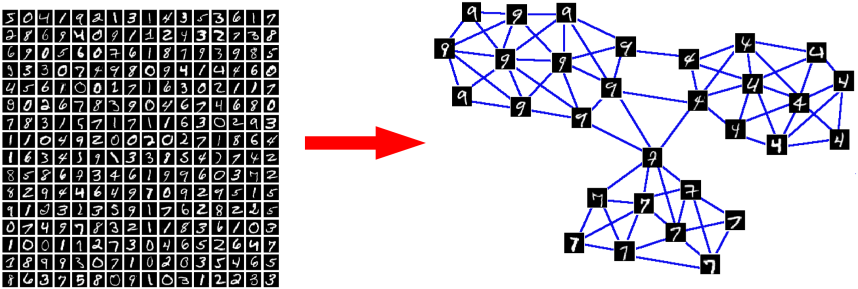
Lectures 1-2

Clustering



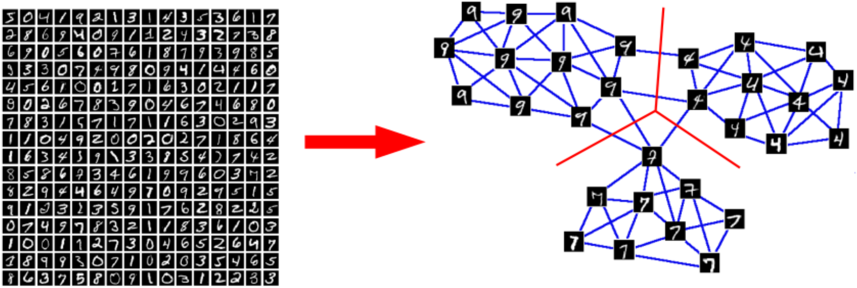
- Partition the data into meaningful groups.

Graph-Based Clustering



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

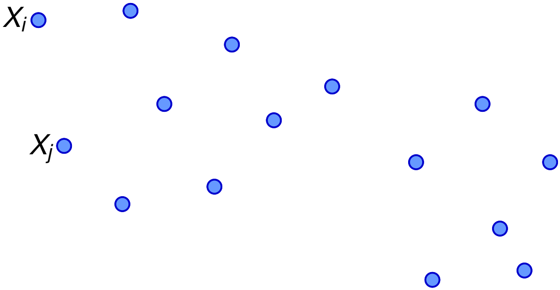
Graph-Based Clustering



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

From point clouds to graphs

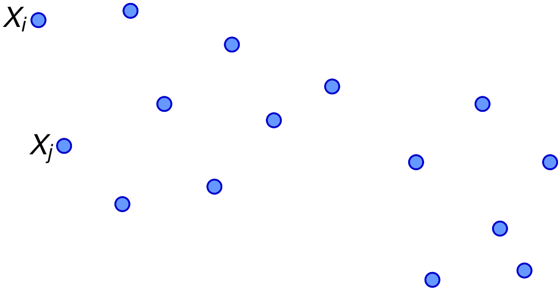
- Let $V = \{X_1, \ldots, X_n\}$ be a point cloud in $\mathbb{R}^d$:



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

From point clouds to graphs

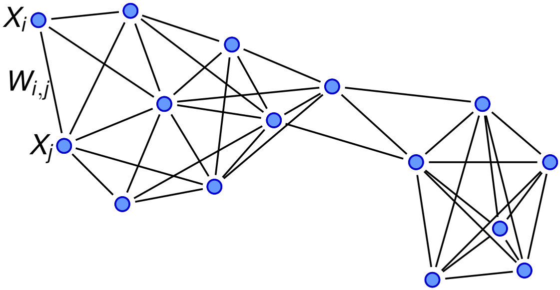
- Let $V = \{X_1, \ldots, X_n\}$ be a point cloud in $\mathbb{R}^d$:



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

From point clouds to graphs

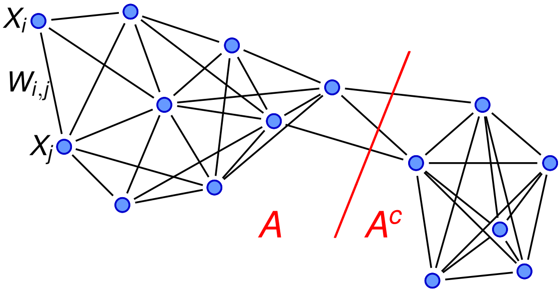
- Let $V = \{X_1, \ldots, X_n\}$ be a point cloud in $\mathbb{R}^d$:



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

Graph cut

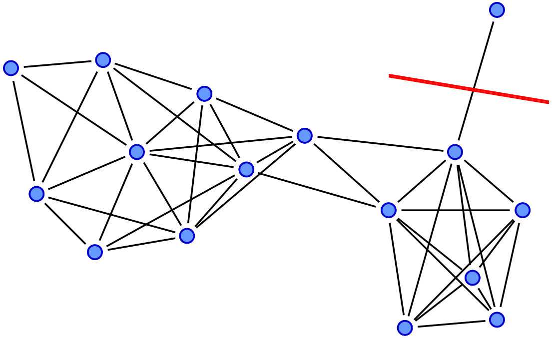
- Let $V = \{X_1, \ldots, X_n\}$ be a point cloud in $\mathbb{R}^d$:



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

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

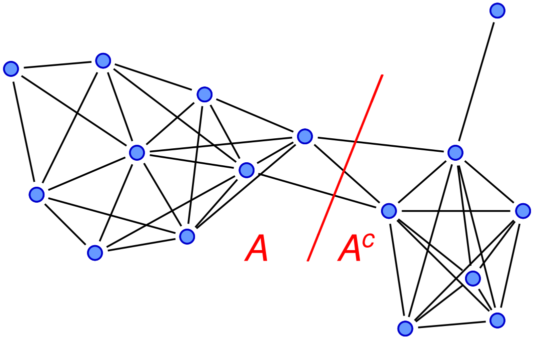
- Let $V = \{X_1, \ldots, X_n\}$ be a point cloud in $\mathbb{R}^d$:



- 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: $A \subset V$.

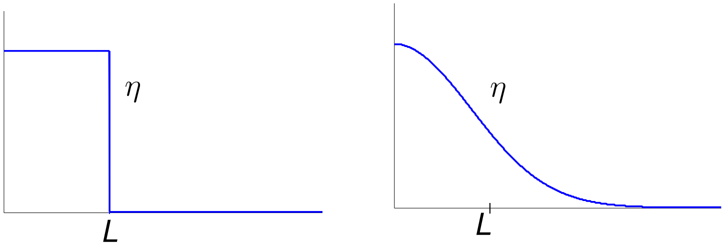
$$Cut(A, A^c) = \sum_{i \in A} \sum_{j \in A^c} W_{i,j}.$$

- Cheeger Cut: Minimize

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

- proximity based graphs

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



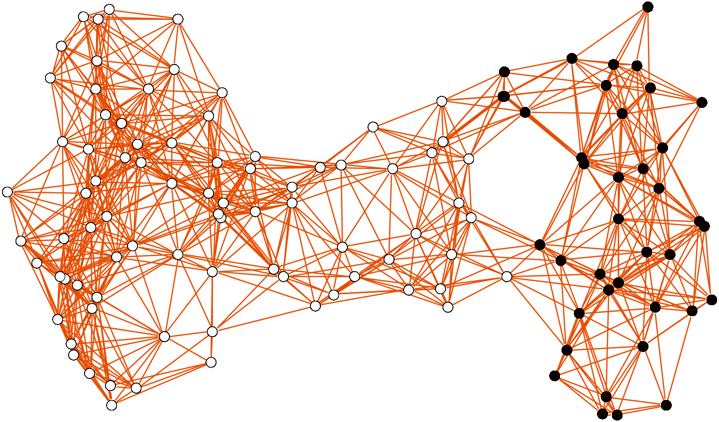
- kNN graphs: Connect each vertex with its k nearest neighbors

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 $u : V \to \mathbb{R}$

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

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

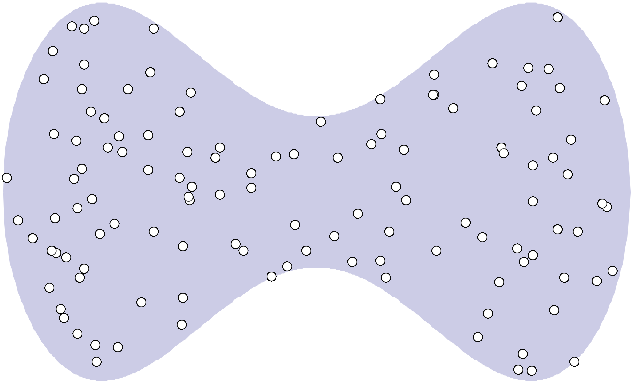
- Assume $u : V \to [0, 1]$. Then $u(x) = \int_0^1 \chi_{\{u \geq \lambda\}}(x) d\lambda$.
- Coarea formula: $GTV_n(u) = \int_0^1 GTV_n(\chi_{\{u \geq \lambda\}}) d\lambda$.
- Convexity $B_n(u) \leq \int_0^1 B_n(\chi_{\{u \geq \lambda\}}) d\lambda$
- If u is a minimizer then for all λ

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

- 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 $d\nu = \rho dx$



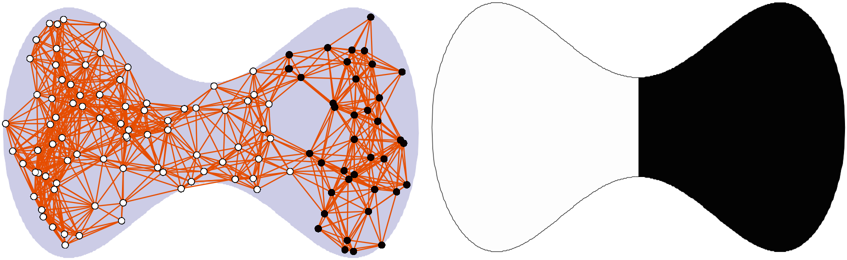
Consistency of Cheeger cut clustering

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 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

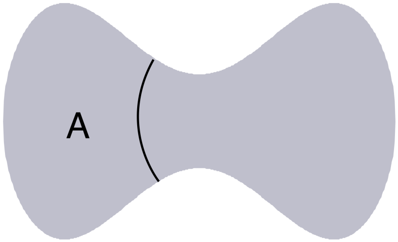
- $d\nu = \rho dx$ probability measure, $\mathrm{supp}(\nu) = D$, $0 < \lambda \leq \rho \leq \frac{1}{\lambda}$ on D.

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

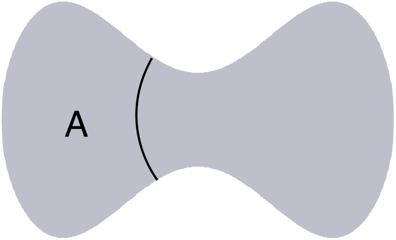
- $d\nu = \rho dx$ probability measure, $\mathrm{supp}(\nu) = D$, $0 < \lambda \leq \rho \leq \frac{1}{\lambda}$ on D.

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 $\mathrm{supp}(\nu) = D$.
- ν has continuous on D density ρ and $0 < \lambda \leq \rho \leq \frac{1}{\lambda}$ on D.

Weighted TV

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

Weighted relative perimeter

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

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

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

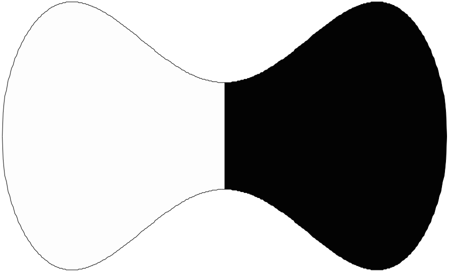
Minimize

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

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

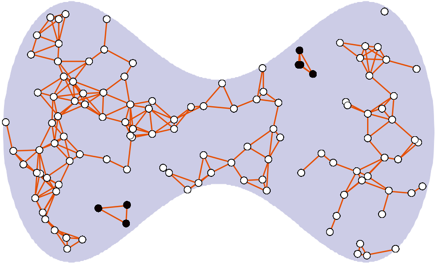
Do the minimizers of

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

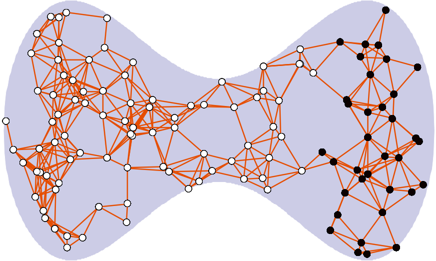
converge as the number of data points $n \to \infty$ to a minimizer of

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

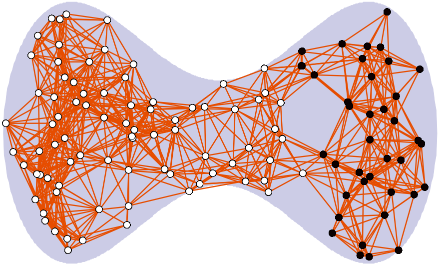
Question 1: For what scaling of $\varepsilon(n)$ can this hold?
Question 2: What is the topology for which $u^n \longrightarrow u$?



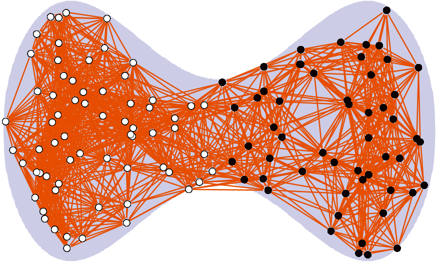
$n = 120, \varepsilon = 0.15$



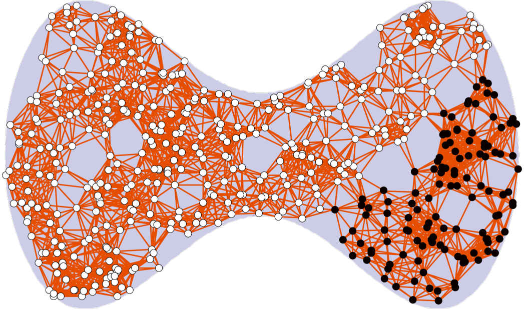
$n = 120, \varepsilon = 0.20$



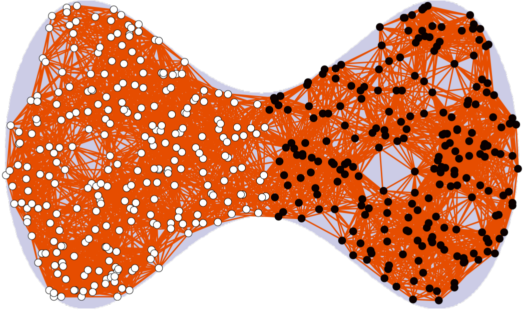
$n = 120, \varepsilon = 0.30$



$n = 120, \varepsilon = 0.40$



$n = 500, \varepsilon = 0.14$



$n = 500, \varepsilon = 0.2$

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

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

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

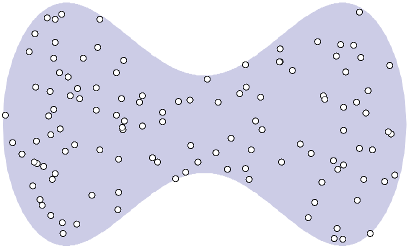
$$GTV_{n,\varepsilon_n}(u^n) = \frac{1}{\varepsilon_n n^2} \sum_{i,j} \eta_{\varepsilon_n}(X_i - X_j) \, |u_i^n - u_j^n|$$

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 $u^n \longrightarrow u$?

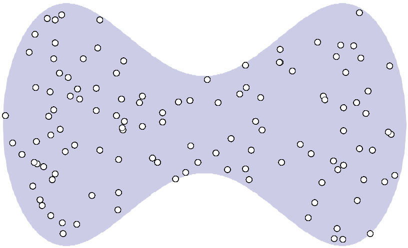
Consider domain D and $V_n = \{X_1, \dots, X_n\}$ random i.i.d points.



- 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.



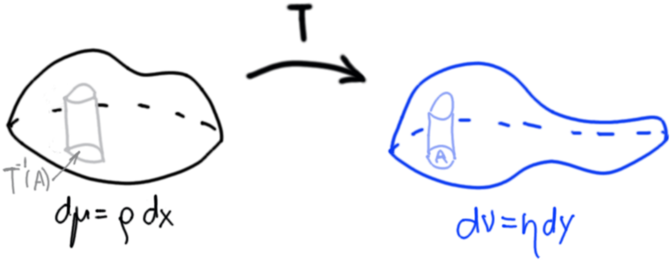
- How to compare $u_n \in L^1(\nu_n)$ and $u \in L^1(D)$ in a way consistent with L^1 topology?

Transport map

- 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. $T : X \to Y$,

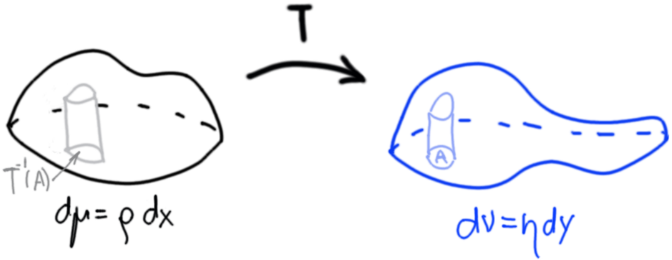
$$T_\sharp \mu = \nu, \qquad \text{that is } \forall A \text{ measurable } \mu(T^{-1}(A)) = \nu(A)$$



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

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

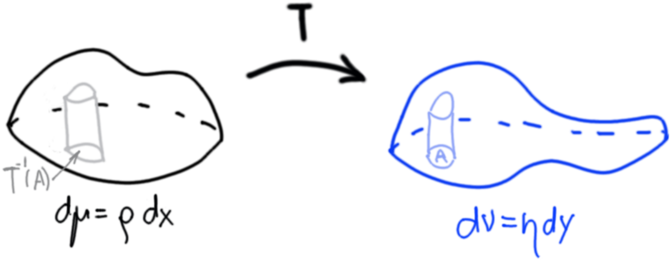
$$T_\sharp \mu = \nu, \qquad \text{that is } \forall A \text{ measurable } \mu(T^{-1}(A)) = \nu(A)$$



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

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

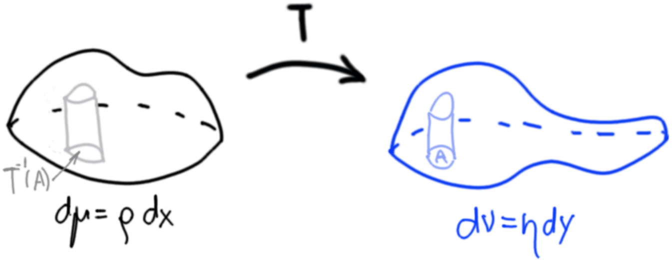
$$T_\sharp \mu = \nu, \qquad \text{that is } \forall A \text{ measurable } \mu(T^{-1}(A)) = \nu(A)$$



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

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

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

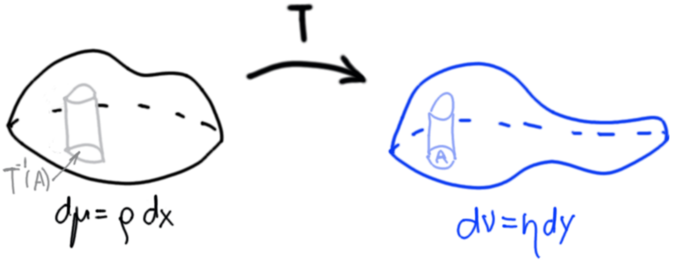


$$\int_{T^{-1}(A)} \rho(x)dx = \int_A \eta(y)dy = \int_{T^{-1}(A)} \eta(T(x))\,|\det(DT(x)|dx$$

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

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

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

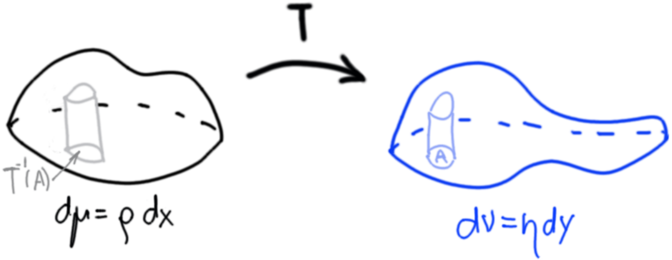


Change of variables: $y = T(x)$, for $f = \chi_A$, using $\chi_{T^{-1}(A)}(x) = \chi_A \circ T(x)$

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

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, $T_\sharp \mu = \nu$

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

Optimal Transport Cost – Monge formulation

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?

Optimal Transport Cost – Monge formulation

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.

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

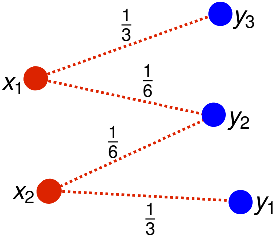
Transport Plan

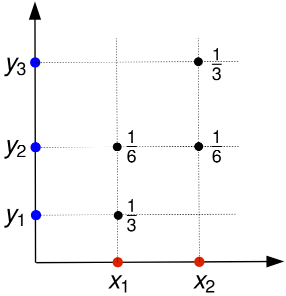
Transport plans, π are probability measures on $X \times Y$ with first marginal μ and second marginal ν:

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

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

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





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

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 $(I \times T)(x) = (x, T(x))$.

- $c(x, y)$ 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.

$\infty-$**transportation distance:**

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

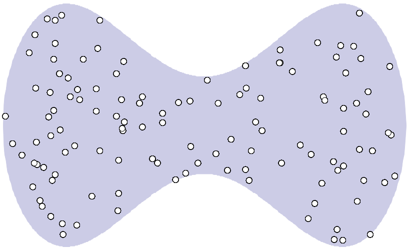
- There exists a minimizer $\pi \in \Pi(\mu, \nu)$.
- If $\mu = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ and $\nu = \frac{1}{n} \sum_{j=1}^n \delta_{y_j}$ then

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

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

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

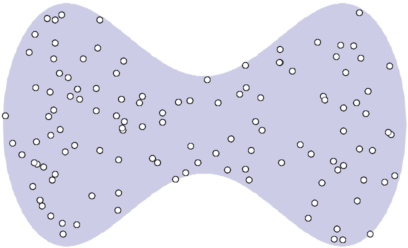
Consider domain D and $V_n = \{X_1, \ldots, X_n\}$ random i.i.d points.



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

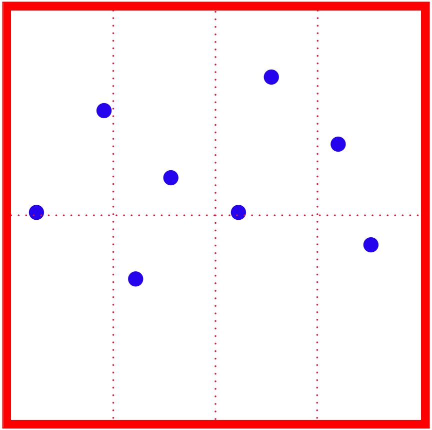
Note that $u \in L^1(\nu)$ and $u_n \in L^1(\nu_n)$, where $\nu_n = \frac{1}{N}\sum_{i=1}^{n} \delta_{X_i}$.

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

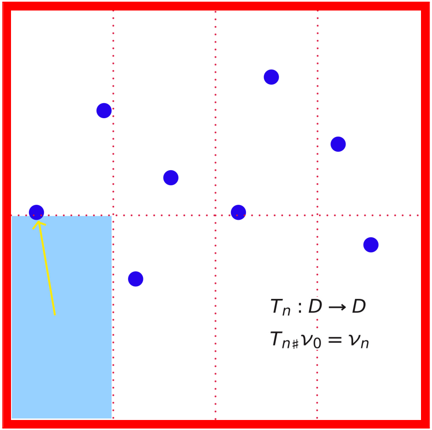


- How to compare $u_n \in L^1(\nu_n)$ and $u \in L^1(D)$ in a way consistent with L^1 topology?

An idea: Divide the domain D into n sets of the same ν measure and to each piece associate a point X_i. That is, consider a map $T_n : D \to D$ such that $T_{\#}\nu = \nu_n$.

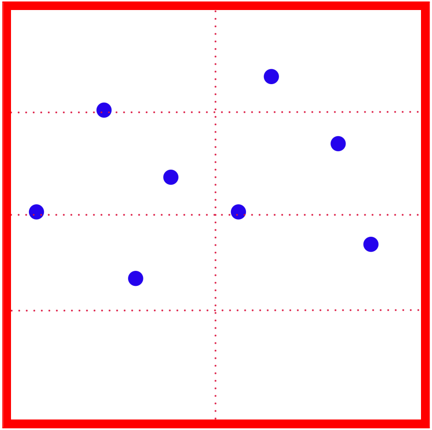


Divide the domain D into n pieces and to each piece associate a point X_i. That is, consider a map $T_n : D \to D$ such that $T_{n\sharp}\nu = \nu_n$.

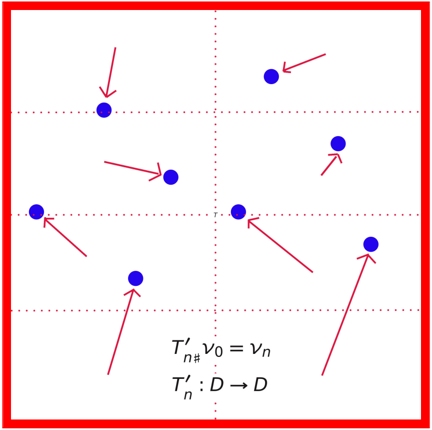


To compare $u \in L^1(\nu)$ and $u_n \in L^1(\nu_n)$ we compare $u_n \circ T_n$ and u in $L^1(\nu)$.

A different partition:

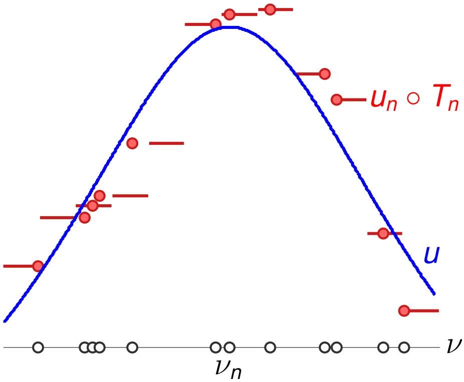


A different partition:



Topology

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



- Let T_n be a transportation map from ν to ν_n

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

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

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^1 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: $\int_{D\times D} |x-y| d\pi_n(x,y) \longrightarrow 0$

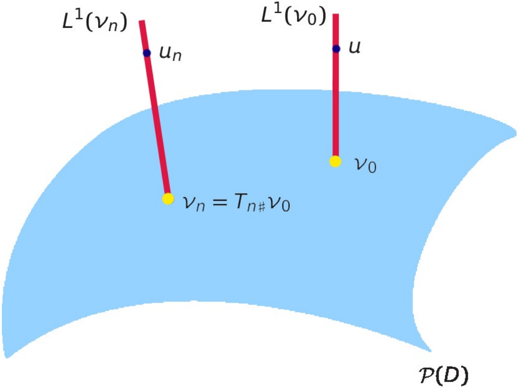
TFAE:

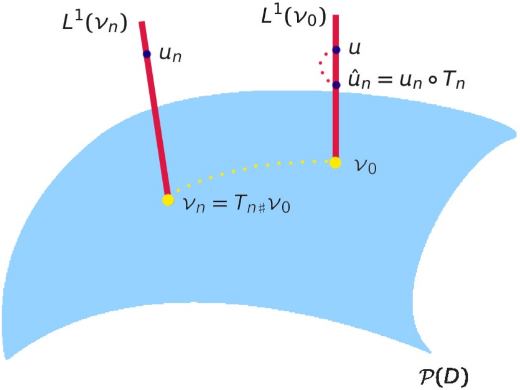
1. $(\nu_n, f_n) \xrightarrow{TL^p} (\nu, f)$ as $n \to \infty$.
2. $\nu_n \rightharpoonup \nu$ and **there exists** a stagnating sequence of transportation plans $\{\pi_n\}_{n\in\mathbb{N}}$ for which

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

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^1 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 (García Trillos and S.)

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

$$\lim_{n\to\infty} \frac{(\log n)^{3/4}}{n^{1/2}} \frac{1}{\varepsilon_n} = 0 \;\; \textit{if } d = 2,$$
$$\lim_{n\to\infty} \frac{(\log n)^{1/d}}{n^{1/d}} \frac{1}{\varepsilon_n} = 0 \;\; \textit{if } d \geq 3.$$

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.

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

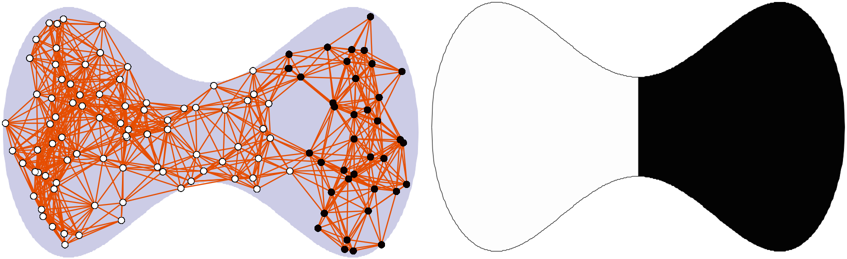
$$\sup_{n\in\mathbb{N}} GTV_{n,\varepsilon_n}(u_n) < \infty,$$

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

Recall:

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

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



Comment of ε_n

- We require

$$\lim_{n\to\infty} \frac{(\log n)^{3/4}}{n^{1/2}} \frac{1}{\varepsilon_n} = 0 \text{ if } d = 2,$$
$$\lim_{n\to\infty} \frac{(\log n)^{1/d}}{n^{1/d}} \frac{1}{\varepsilon_n} = 0 \text{ if } d \geq 3.$$

- Note that for $d \geq 3$ this means that typical degree $\gg \log(n)$.
- Does convergence hold if fewer than $\log(n)$ neighbors are connected to?

Comment of ε_n

- We require

$$\lim_{n\to\infty} \frac{(\log n)^{3/4}}{n^{1/2}} \frac{1}{\varepsilon_n} = 0 \ \textit{if } d = 2,$$
$$\lim_{n\to\infty} \frac{(\log n)^{1/d}}{n^{1/d}} \frac{1}{\varepsilon_n} = 0 \ \textit{if } d \geq 3.$$

- Note that for $d \geq 3$ this means that typical degree $\gg \log(n)$.
- Does convergence hold if fewer than $\log(n)$ neighbors are connected to?

  No. There exists $c > 0$ such that $\varepsilon_n < c\frac{\log(n)^{1/d}}{n^{1/d}}$ then with probability one the random geometric graph is asymptotically disconnected. *Penrose (1999); Gupta and Kumar (1999); Goel,Rai and Krishnamachari (2004).*
  This implies that for large enough n, $\min GC_{n,\varepsilon_n} = 0$. While $\inf C > 0$.

  So for $d \geq 3$ the condition is optimal in terms of scaling.

Consistency of Cheeger Cuts

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.)

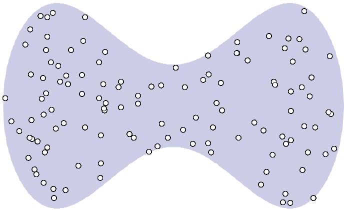
For the same conditions on ε_n as before, with probability one:

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

Moreover, for any sequence of sets $E_n \subseteq \{X_1, \ldots, X_n\}$ 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.

∞-OT between a measure and its random sample

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

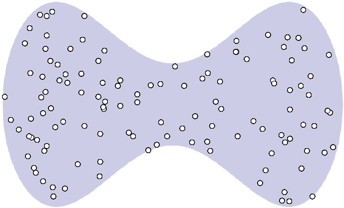


Theorem

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:

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

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:

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

Lectures 3-4

Recall: 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 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. $\int |x - T_n(x)| d\nu(x) \to 0$).

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

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

- $TV_\varepsilon \xrightarrow{\Gamma} TV(\,\cdot\,, \rho^2)$ 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 $T_{n\sharp}\nu = \nu_n$,

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

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

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

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

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

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

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

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

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

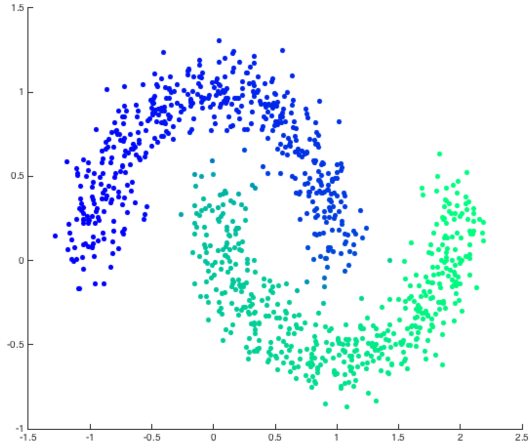
  The weighted degree of a vertex is $d_i = \sum_j W_{i,j}$.
- Dirichlet energy of $u_n : V_n \to \mathbb{R}$ is

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

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

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

Spectral Clustering: Two moons (easy)



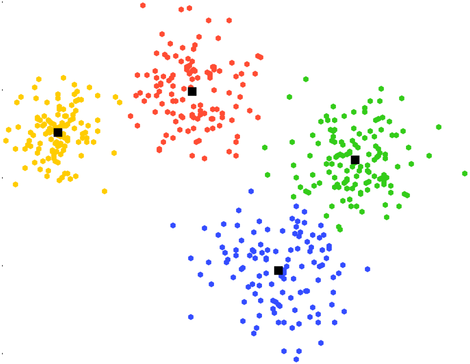
1D embedding: $x_i \mapsto u_2(x_i)$

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

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

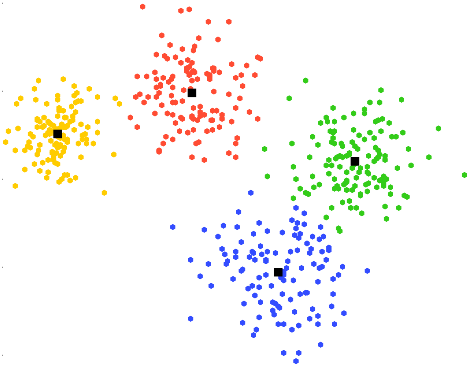
where $\text{dist}(x, A) = \min_{a \in A} |x - a|$.



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

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



Input: Number of clusters k and similarity matrix W.

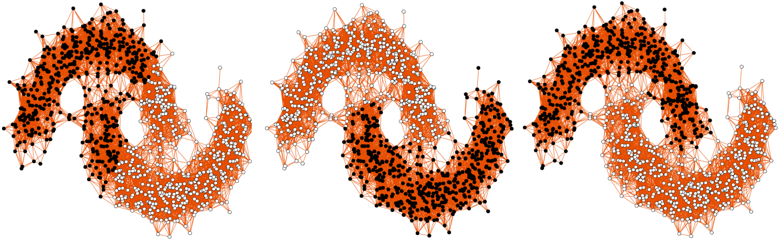
- Construct the graph Laplacian L.
- Compute the eigenvectors $u_1, \ldots, u_k$ of L associated to the k smallest (nonzero) eigenvalues of L.
- For $i = 1, \ldots, n$, let $y_i \in \mathbb{R}^k$ be

$$y_i = [u_1(x_i), \ldots, u_k(x_i)]^T.$$

- Use the k-means algorithm to partition the set of points $\{y_1, \ldots, y_n\}$ into k groups, that we denote by $G_1, \ldots, G_k$.

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

Comparison of Clustering Algorithms



(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*

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

- 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

$$u_2^n \xrightarrow{TL^2} u_2$$

where u_2 is eigenfunction, corresponding to second eigenvalue, of

$$L_c(u_k) := -\frac{\operatorname{div}(\rho^2 \nabla u)}{\rho} = \lambda_2 u \quad \text{in } D$$
$$\frac{\partial u}{\partial n} = 0 \qquad \text{on } \partial D.$$

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

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

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

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.
- For $i = 1, \ldots, n$, let $y_i^n \in \mathbb{R}^k$ be

$$y_i^n = [u_1^n(x_i), \ldots, u_k^n(x_i)]^T.$$

- Use the k-means algorithm to partition the set of points $\{y_1^n, \ldots, y_n^n\}$ 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.

Discrete Spectral Clustering:

- Construct the graph Laplacian L for the geometric graph of the sample
- Compute the eigenvectors $u_1^n, \dots, u_k^n$ of L associated to the k smallest (nonzero) eigenvalues of L.
- For $i = 1, \dots, n$, let $y_i^n \in \mathbb{R}^k$ be

$$y_i^n = [u_1^n(x_i), \dots, u_k^n(x_i)]^T.$$

- Use the k-means algorithm to partition the set of points $\{y_1^n, \dots, y_n^n\}$ into k groups. We denote the resulting partitioning of V_n by $G_1^n, \dots, G_k^n$.

Theorem. Let $G_1^n, \dots G_k^n$ be the clusters above. Let $\nu_i^n = \nu_{n \llcorner G_i^n}$ (the restriction of empirical measure to clusters) for $i = 1, \dots, k$. Then $(\nu_1^n, \dots, \nu_k^n)$ is precompact with respect to weak convergence of measures and converges along a subsequence to $(\nu_1, \dots, \nu_k) = (\nu_{\llcorner G_1}, \dots, \nu_{\llcorner G_k})$ where $G_1, \dots, G_k$ 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 $D = \mathrm{diag}(d_1, \dots, d_n)$.
- 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 $d\mu = \rho dVol_{\mathcal{M}}$.

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

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, $k = 1, \ldots$.

Transportation estimates

Let $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ be the empirical measure of the random i.i.d sample.

Theorem

For any $\beta > 1$ and every $n \in \mathbb{N}$ there exist a transportation map $T_n \colon \mathcal{M} \to X$ and a constant A such that

$$\operatorname{esssup}_{x\in\mathcal{M}} d(x, T_n(x)) \leq \ell := A \begin{cases} \frac{\log(n)^{3/4}}{n^{1/2}}, & \text{if } m = 2, \\ \frac{(\log n)^{1/m}}{n^{1/m}}, & \text{if } m \geq 3, \end{cases}$$

holds with probability at least $1 - C_{K,\mathrm{Vol}(\mathcal{M}),m,i_0} \cdot n^{-\beta}$, where A depends only on K, i_0, R, m, $\mathrm{Vol}(\mathcal{M})$, α and β.

- K – upper bound on absolute value of sectional curvature
- i_0 – injectivity radius
- R – reach of $\mathcal{M}$ is $\mathbb{R}^d$

Consistency of Spectral Clustering in Manifold Setting

Techniques inspired by *Burago, Ivanov, Kurylev*

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

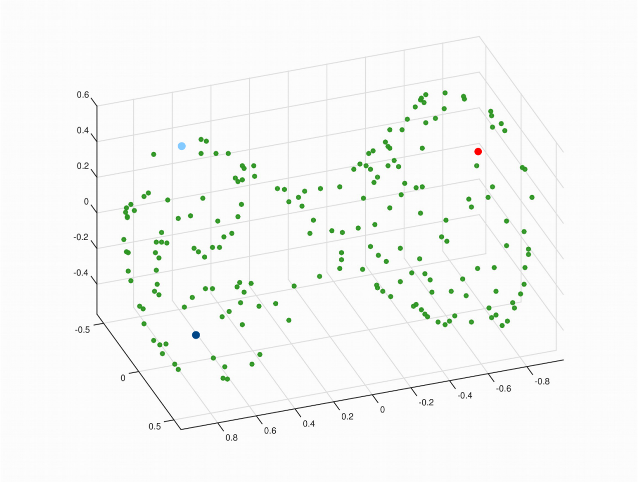
There exists a constant $C_{m,K,\mathrm{Vol}(\mathcal{M}),i_0}$ such that for every $\beta > 1$ and every $n \in \mathbb{N}$ the following holds with probability at least $1 - C_{m,K,\mathrm{Vol}(\mathcal{M}),i_0} \cdot n^{-\beta}$.
For every $k \in \{1, \ldots, n\}$ there exists a constant $C > 0$ depending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

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

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

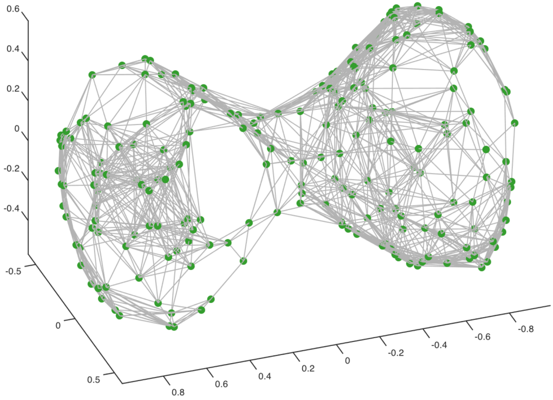
- K – upper bound on absolute value of sectional curvature
- i_0 – injectivity radius
- R – reach of $\mathcal{M}$ is $\mathbb{R}^d$

Semi-supervised learning



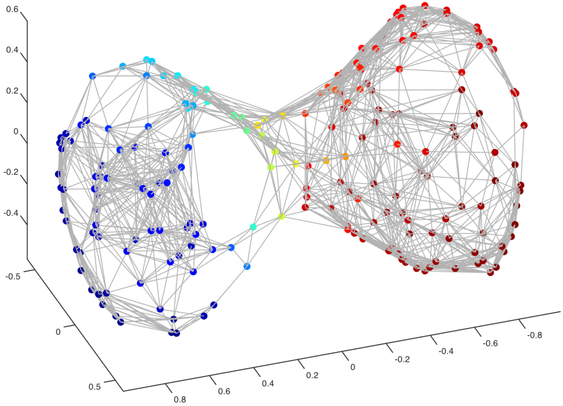
- Colors denote real-valued labels
- Task: Assign real-valued labels to all of the data points

Semi-supervised learning



- Graph is used to represent the geometry of the data set

Semi-supervised learning



- Consider graph-based objective functions which reward the regularity of the estimator and impose agreement with preassigned labels

p-Dirichelt energy

- $V_n = \{x_1, \ldots, x_n\}$, weight matrix W:

$$W_{ij} := \eta\left(|x_i - x_j|\right).$$

- p-Dirichlet energy of $f_n : V_n \to \mathbb{R}$ is

$$E(f_n) = \frac{1}{2}\sum_{i,j} W_{ij}|f_n(x_i) - f_n(x_j)|^p.$$

- For $p = 2$ associated operator is the (unnormalized) graph laplacian

$$L = D - W,$$

where $D = \operatorname{diag}(d_1, \ldots, d_n)$ and $d_i = \sum_j W_{i,j}$.

p-Laplacian semi-supervised learning

Assume we are given k labeled points

$$(x_1, y_1), \ldots (x_k, y_k)$$

and unlabeled points $x_{k+1}, \ldots, x_n$.

Question. How to label the rest of the points?

p-Laplacian SSL

Minimize $$E(f_n) = \frac{1}{2} \sum_{i,j} W_{ij} |f_n(x_i) - f_n(x_j)|^p$$

subject to constraint $$f(x_i) = y_i \quad \text{for } i = 1, \ldots, k.$$

Zhu, Ghahramani, and Lafferty '03 introduced the approach with $p = 2$.
Zhou and Schölkopf '05 consider general p.

p-Laplacian SSL

Minimize $$E(f_n) = \frac{1}{2}\sum_{i,j} W_{ij}|f_n(x_i) - f_n(x_j)|^p$$

subject to constraint $f(x_i) = y_i \quad \text{for } i = 1, \dots, k.$

Questions.

- What happens as $n \to \infty$?
- Do minimizers f_n converge to a solution of a limiting problem?
- In what topology should the question be considered?

Remark.

- We would like to localize η as $n \to \infty$.

p-Laplacian semi-supervised learning: Asymptotics

p-Laplacian SSL

Minimize $$E_n(f_n) = \frac{1}{\varepsilon^p n^2} \sum_{i,j} \eta_\varepsilon(x_i - x_j)|f_n(x_i) - f_n(x_j)|^p$$

subject to constraint $f_n(x_i) = y_i \quad \text{for } i = 1, \ldots, k.$

where

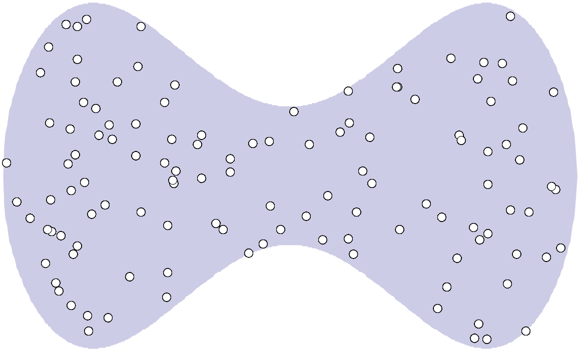
$$\eta_\varepsilon(\,\cdot\,) = \frac{1}{\varepsilon^d}\eta\left(\frac{\cdot}{\varepsilon}\right).$$

Questions.

- Do minimizers f_n converge to a solution of the limiting problem?
- In what topology should the question be considered?
- How shall ε_n scale with n for the convergence to hold?

Ground Truth Assumption

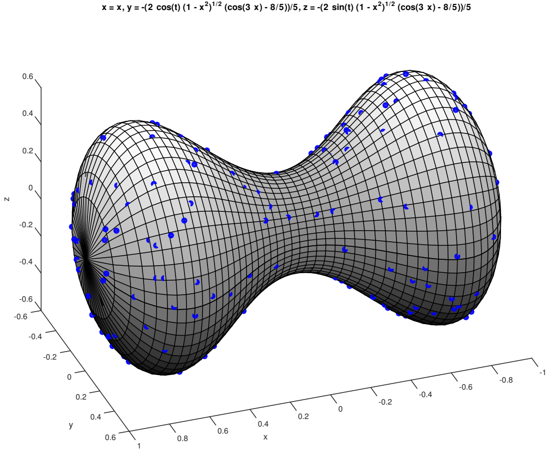
We assume points $x_1, x_2, \ldots,$ are drawn i.i.d out of measure $d\nu = \rho dx$



We also assume ρ is supported on a Lipschitz domain Ω and is bounded above and below by positive constants.

Ground Truth Assumption: Manifold version

Assume points $x_1, x_2, \ldots,$ are drawn i.i.d out of measure $d\nu = \rho d\,\mathrm{Vol}_{\mathcal{M}}$, where $\mathcal{M}$ is a compact manifold without boundary, and $0 < \rho < C$ is continuous.



Harmonic semi-supervised learning

Nadler, Srebro, and Zhou '09 observed that for $p = 2$ the minimizers are spiky as $n \to \infty$. [Also see Wahba '90.]

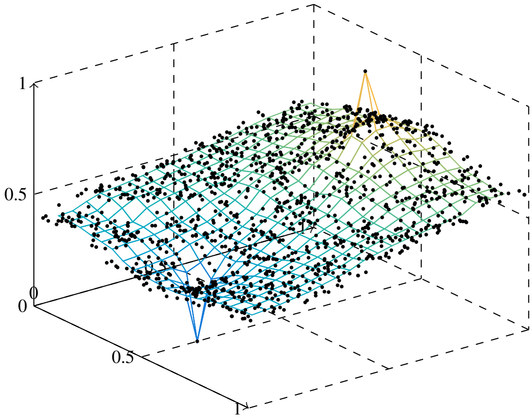


Figure: Graph of the minimizer for $p = 2$, $n = 1280$, i.i.d. data on square; training points $(0.5, 0.2)$ with label 0 and $(0.5, 0.8)$ with label 1.

El Alaoui, Cheng, Ramdas, Wainwright, and Jordan '16, show that spikes can occur for all $p \leq d$ and propose using $p > d$.

Heuristics.

$$E_n^{(p)}(f) = \frac{1}{\varepsilon^p n^2} \sum_{i,j=1}^{n} \eta_\varepsilon(x_i - x_j)|f(x_i) - f(x_j)|^p$$
$$\overset{n\to\infty}{\approx} \iint \eta_\varepsilon(x_i - x_j) \left(\frac{|f(x) - f(y|}{\varepsilon} \right)^p \rho(x)\rho(y)dxdy$$
$$\overset{\varepsilon\to 0}{\approx} \sigma_\eta \int |\nabla f(x)|^p \rho(x)^2 dx$$

Sobolev space $W^{1,p}(\Omega)$ embeds into continuous functions iff $p > d$.

Continuum p-Laplacian semi-supervised learning

μ- measure with density ρ, positive on Ω.

Continuum p-Laplacian SSL

Minimize

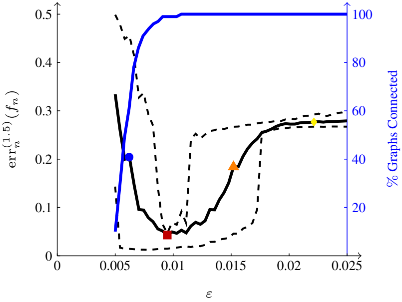
$$E_\infty(f) = \int_\Omega |\nabla f(x)|^p \rho(x)^2 dx$$

subject to constraints that

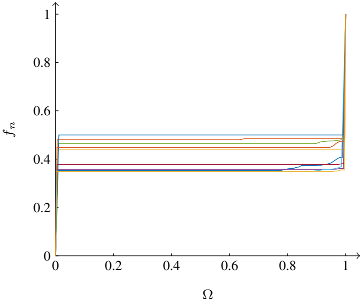
$$f(x_i) = y_i \qquad \text{for all } i = 1, \ldots, k.$$

- The functional is convex
- The problem has a unique minimizer iff $p > d$. The minimizer lies in $W^{1,p}(\Omega)$

Here: *d* = 1 and *p* = 1.5. For $\varepsilon > 0.02$ the minimizers lack the expected regularity.



(a) error for $p = 1.5$ and $d = 1$



(b) minimizers for $\varepsilon = 0.023$, $n = 1280$, ten realizations. Labeled points are $(0, 0)$ and $(1, 1)$.

p-Laplacian semi-supervised learning

Theorem (Thorpe and S. '17)

Let $p > 1$. Let f_n be a sequence of minimizers of $E_n^{(p)}$ satisfying constraints. Let f be a minimizer of $E_\infty^{(p)}$ satisfying constraints.

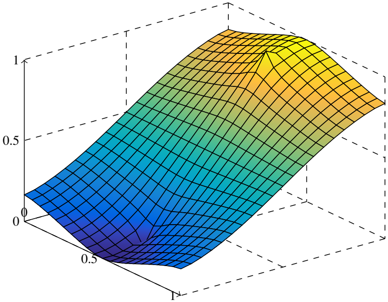
(i) *If $d \geq 3$ and $n^{-\frac{1}{p}} \gg \varepsilon_n \gg \left(\frac{\log n}{n}\right)^{\frac{1}{d}}$ then $p > d$, f is continuous and f_n converges locally uniformly to f, meaning that for any $\Omega' \subset\subset \Omega$*

$$\lim_{n\to\infty} \max_{\{k \leq n \,:\, x_k \in \Omega'\}} |f(x_k) - f_n(x_k)| = 0.$$

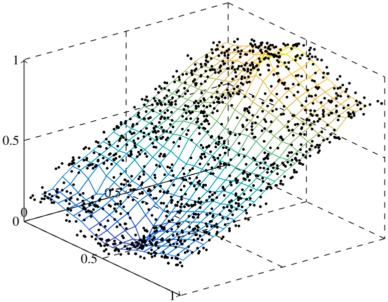
(ii) *If $1 \gg \varepsilon_n \gg n^{-\frac{1}{p}}$; then there exists a sequence of real numbers c_n such that $f_n - c_n$ converges to zero locally uniformly.*

Note that in case (ii) all information about labels is lost in the limit.
The discrete minimizers exhibit spikes.

p-Laplacian semi-supervised learning



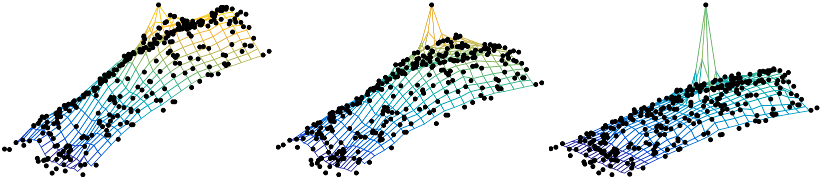
(a) discrete minimizer



(b) continuum minimizer

Minimizer for $p = 4$, $n = 1280$, $\varepsilon = 0.058$ i.i.d. data on square, with training points $(0.2, 0.5)$ and $(0.8, 0.5)$ and labels 0 and 1 respectively.

p-Laplacian semi-supervised learning



(a) $\varepsilon = 0.058$ (b) $\varepsilon = 0.09$ (c) $\varepsilon = 0.2$

$p = 4$ which in 2D is in the well-posed regime

Improved p-Laplacian semi-supervised learning

$p > d$. Labeled points $\{(x_i, y_i) \;:\; i = 1, \ldots, k\}$.

p-Laplacian SSL

Minimize

$$E_n(f_n) = \frac{1}{\varepsilon^2 n^2} \sum_{i,j} \eta_\varepsilon(x_i - x_j)|f_n(x_i) - f_n(x_j)|^p$$

subject to constraint

$$f_n(x_m) = y_i \quad \text{whenever } |x_m - x_i| < 2\varepsilon, \;\; \text{for all } i = 1, \ldots, k.$$

where

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

Theorem (Thorpe and S. '17)

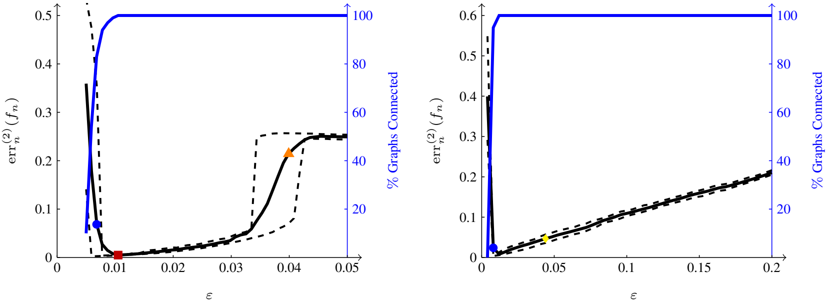
Let $p > d$.

- *f_n be a sequence of minimizers of improved p-Laplacian SSL on n-point sample.*
- *f minimizer of $E_\infty^{(p)}$ satisfying constraints. Since $p > d$ we know f is continuous.*

If $d \geq 3$ and $1 \gg \varepsilon_n \gg \left(\frac{\log n}{n}\right)^{\frac{1}{d}}$ then f_n converges locally uniformly to f, meaning that for any $\Omega' \subset\subset \Omega$

$$\lim_{n\to\infty} \max_{\{k \leq n \,:\, x_k \in \Omega'\}} |f(x_k) - f_n(x_k)| = 0.$$

Here: $d = 1$, $p = 2$, and $n = 1280$. Labeled points are $(0,0)$ and $(1,1)$.



(a) original model (b) improved model

Note that the axes on the error plots for the models are not the same

Techniques

general approach developed with Garcia–Trillos (ARMA '16)

- Γ-convergence. Notion and set of techniques of calculus of variations to consider asymptotics of functionals (here random discrete to continuum)
- TL^p space. Notion of topology based on optimal transportation which allows to compare functions defined on different spaces (here $f_n \in L^p(\mu_n)$ and $f \in L^p(\mu)$)

We also need

- Nonlocal operators and their asymptotics
- In SSL, for constraint to be satisfied we need uniform convergence. This also requires discrete regularity and finer compactness results.

Γ convergence for p-Laplacian

Energy

$$E_n(f_n) = \frac{1}{\varepsilon^2 n^2} \sum_{i,j} \eta_\varepsilon(x_i - x_j) |f_n(x_i) - f_n(x_j)|^p$$

Γ-converges in TL^p space to

$$\sigma E_\infty(f) = \sigma \int_\Omega |\nabla f(x)|^p \rho(x)^2 dx$$

as $n \to \infty$ provided that

$$1 \gg \varepsilon_n \gg \begin{cases} \frac{(\log n)^{\frac{3}{4}}}{\sqrt{n}} & \text{if } d = 2 \\ \left(\frac{\log n}{n}\right)^{\frac{1}{d}} & \text{if } d \geq 3; \end{cases}$$

Role of nonlocal operators

Heuristics.

$$E_n^{(p)}(f) = \frac{1}{\varepsilon^p n^2} \sum_{i,j=1}^{n} \eta_\varepsilon(x_i - x_j)|f(x_i) - f(x_j)|^p$$

$$\overset{n\to\infty}{\approx} \iint \eta_\varepsilon(x_i - x_j) \left(\frac{|f(x) - f(y|}{\varepsilon} \right)^p \rho(x)\rho(y) dxdy$$

$$\overset{\varepsilon\to 0}{\approx} \sigma_\eta \int |\nabla f(x)|^p \rho(x)^2 dx$$

- Discrete problem on graph is closer to a nonlocal functional (with scale ε) than to limiting differential one
- Nonlocal energy does not have the smoothing properties of the differential one.

$$E_n^{(p)}(f) = \frac{1}{\varepsilon^p n^2} \sum_{i,j=1}^{n} \eta_\varepsilon(x_i - x_j)|f(x_i) - f(x_j)|^p.$$

Consider

$$f(x_j) = \begin{cases} 1 & \text{if } j = 1 \\ 0 & \text{else.} \end{cases}$$

Then

$$E_n^{(p)}(f) = \frac{2}{\varepsilon_n^p n^2} \sum_{j=2}^{n} \frac{1}{\varepsilon_n^d} \eta\left(\frac{|x_1 - x_j|}{\varepsilon_n}\right) \sim \frac{1}{\varepsilon_n^p n^2} n\varepsilon_n^d = \frac{1}{\varepsilon_n^p n} \to 0$$

as $n \to \infty$, when $\varepsilon_n^p n \to \infty$.

Manfredi, Oberman, Sviridov, 2012, Calder 2017

The infinity laplacian is defined by

$$L_n^{\infty} f(x_i) = \max_j w_{ij}(f(x_j) - f(x_i)) + \min_j w_{ij}(f(x_j) - f(x_i))$$

and the p-laplacian is defined by

$$L_n^p f = \frac{1}{d} L_n^2 f + \lambda(p-2) L^{\infty} f.$$

PDE based p-Laplacian semi-supervised learning

$$L_n^p f = \frac{1}{d} L_n^2 f + \lambda(p-2) L^\infty f.$$

SSL problem

$$\begin{aligned} L_n^p f &= 0 && \text{on } \Omega \setminus \Omega_L \\ f(x_i) &= y_i && \text{for all } i = 1, \ldots, k. \end{aligned}$$

Theorem (Calder '17)

Assume $p > d$. If $d \geq 3$ and $\varepsilon_n \gg \left(\frac{\log n}{n}\right)^{\frac{1}{3d/2}}$. Then f_n converges uniformly to f, the solution of the limiting problem.

Note that there is no upper bound on ε_n needed.

Higher order regularizations in SSL

with *Dunlop, Stuart, and Thorpe*, model by *Zhou, Belkin '11*.

Random sample $x_1, \dots x_n$. Labels are known if $x_i \in \Omega_L$, open

Using graph laplacian L_n we define

$$A_n = (L_n + \tau^2 I)^\alpha.$$

Power of a symmetric matrix is defined by $M^\alpha = PD^\alpha P^{-1}$ for $M = PDP^{-1}$.

Higher order SSL

Minimize $\quad E(f) = \frac{1}{2}\langle f_n, A_n f_n \rangle_{\mu_n}$

subject to constraint $\quad f_n(x_i) = y_i \quad$ whenever $x_i \in \Omega_L$.

Higher order regularizations in SSL

$$A_n = (L_n + \tau^2 I)^\alpha.$$

Higher order SSL

Minimize $\qquad E(f) = \frac{1}{2}\langle f_n, A_n f_n\rangle_{\mu_n}$

subject to constraint $\qquad f_n(x_i) = y_i \quad \text{whenever } x_i \in \Omega_L.$

Theorem (Dunlop, Stuart, S. Thorpe)

For $\alpha > \frac{d}{2}$, under usual assumptions, minimizers f_n converge in TL^2 to the

minimizer of $\qquad E(f) = \sigma \int_\Omega u(x)(Au)(x)\rho(x)dx$

subject to constraint $\qquad u(x_i) = y_i \quad \text{whenever } x_i \in \Omega_L.$

where $A = (\sigma L_c + \tau I)^\alpha$ and $L_c u = -\frac{1}{\rho}\text{div}(\rho^2 \nabla u)$.

Higher order regularizations in SSL

with *Dunlop, Stuart, and Thorpe*, model by *Zhou, Belkin '11*.

k labeled points, $(x_1, y_1), \dots (x_k, y_k)$, and a random sample $x_{k+1}, \dots x_n$.

Using graph laplacian L_n we define

$$A_n = (L_n + \tau^2 I)^\alpha.$$

Higher order SSL

Minimize $\quad E(f) = \frac{1}{2}\langle f_n, A_n f_n \rangle_{\mu_n}$

subject to constraint $\quad f_n(x_i) = y_i \quad \text{for } i = 1, \dots, k.$

Higher order regularizations

$$A_n = (L_n + \tau^2 I)^\alpha.$$

Higher order SSL

Minimize $\qquad E(f) = \frac{1}{2}\langle f_n, A_n f_n\rangle_{\mu_n}$

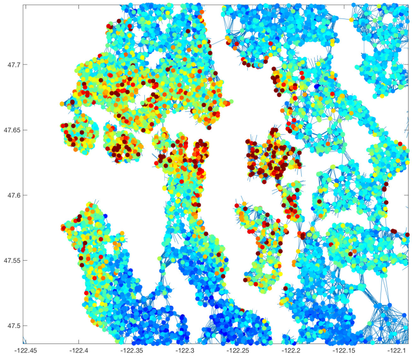
subject to constraint $\qquad f_n(x_i) = y_i \quad \text{for } i = 1, \ldots, k.$

Lemma (Dunlop, Stuart, S. Thorpe)

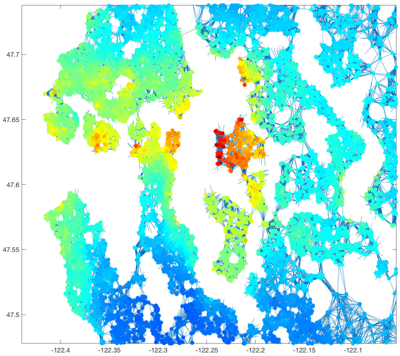
If $1 \gg \varepsilon_n \gg n^{-\frac{1}{2\alpha}}$ then minimizers f_n converge in TL^2 along a subsequence to a constant. That is spikes occur.

Denoising of labels

Housing prices per square foot in Seattle 2015.



() noisy



(a) denoised

Open problems

- Finding better ways to approximate the functional (with Tenbrinck)
- Pointwise assigned labels for higher-order operators
- Regularity of minimizers/PDE on graphs
- Error estimates for consistency of convex functionals (like the Dirichlet functional)
- Error estimates II. In particular why why is the error the smallest for rather coarse graphs? Homogenization?
- Convergence of dynamical models / evolutionary PDE on graphs.
- Convergence of posterior distributions in Bayesian learning.
- Mumford–Shah functional on graphs (with Caroccia and Chambolle)