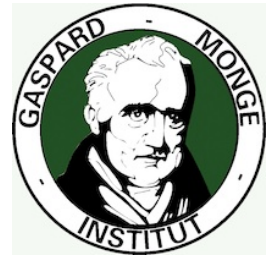
Some applications in unsupervised and semi-supervised learning of Laplacian eigenmaps

Guillaume Obozinski

LIGM/Ecole des Ponts - ParisTech



CGL workshop

Athens Sep 29th- Oct 2nd 2013

A quick return to supervised learning

Decision theoretic framework

- Given some data pairs (X, Y) corresponding to input and output
- Given a loss function $\ell : (a, y) \mapsto \ell(a, y)$

Learn a function f so as to minimize the risk

$$\mathcal{R}(f) = \mathbb{E}[\ell(f(X, Y))]$$

Difficulty: we do not know the distribution of the data. We only have a training set consisting of i.i.d. pairs

$$\mathcal{L} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

Empirical risk minimization principle

$$\text{Solve} \qquad \min_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell(f(x_i), y_i)$$

Empirical Risk Minimization principle

Empirical risk

$$\hat{\mathcal{R}}_n = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)$$

$\hat{\mathcal{R}}_n$ is an unbiased estimate of $\mathcal{R}$:

$$\mathbb{E}[\hat{\mathcal{R}}_n] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\ell(f(X_i), Y_i)] = \mathbb{E}[\ell(f(X_1), Y_1)]$$

Empirical risk minimization

$$\min_{f \in \mathcal{F}} \hat{\mathcal{R}}_n$$

Linear regression

- Learn $f : \mathcal{X} = \mathbb{R}^p \to \mathcal{Y} = \mathbb{R}$
- $\ell : (a, y) \mapsto \frac{1}{2}(a - y)^2$
- $\mathcal{F} = \{f_w : x \mapsto w^\top x \mid w \in \mathbb{R}^p\}$

Risk:

$$\mathcal{R}(f_w) = \frac{1}{2}\mathbb{E}[(Y - X^\top w)^2]$$

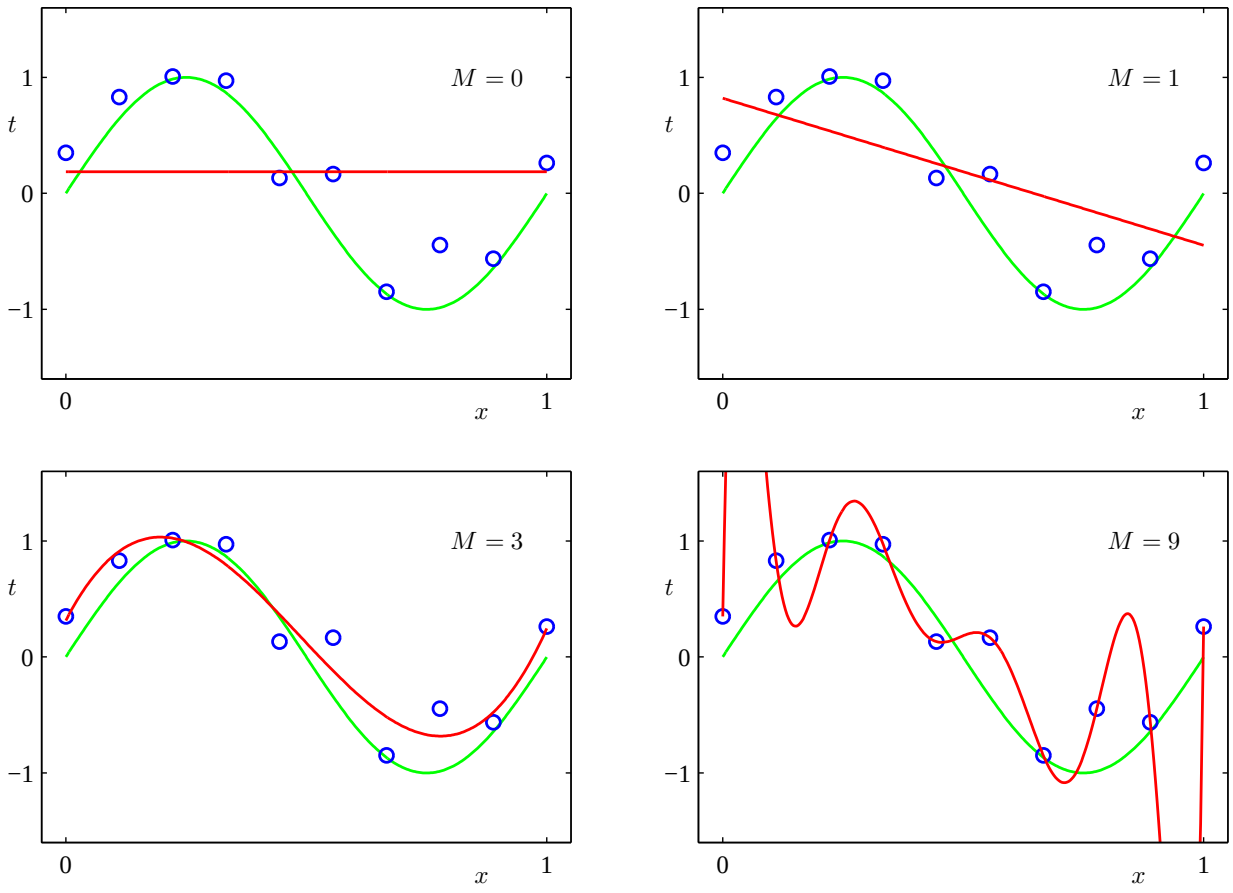
Empirical risk:

$$\hat{\mathcal{R}}_n = \frac{1}{2n}\sum_{i=1}^{n}(y_i - w^\top x_i)^2 = \frac{1}{2n}\left\|\mathbf{y} - \mathbf{X}w\right\|_2^2$$

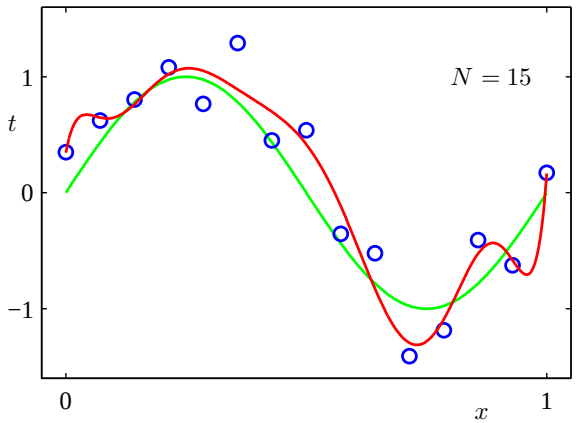
Overfitting: A simple example

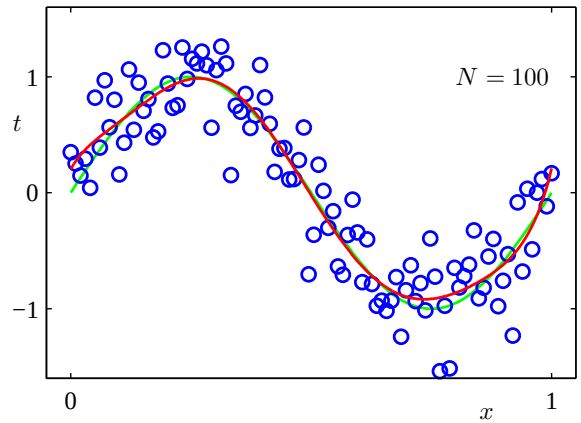
Polynomial regression: $Y = w_0 + w_1 X + w_2 X^2 + \ldots + w_p X^p + \varepsilon$

$$\min_w \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \left(w_0 + w_1 x_i + w_2 x_i^2 + \ldots + w_p x_i^p \right) \right)^2$$



Overfitting: $p \not\ll n$ vs $p \ll n$



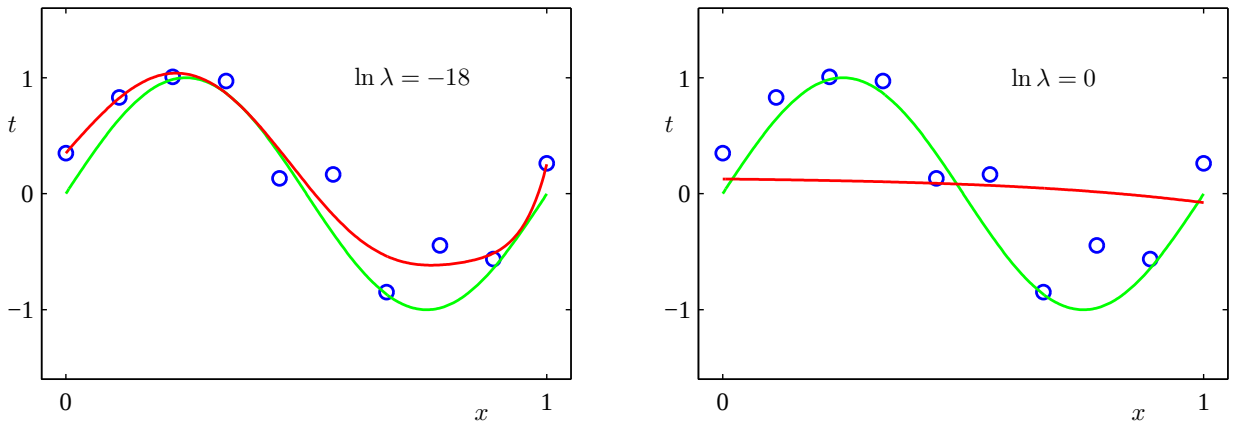


Tikhonov regularization

$$\frac{1}{2n}\sum_{i=1}^{n}\ell(y_i, w^\top\Phi(x_i)) + \frac{\lambda}{2n}\|w\|_2^2$$

- Convex problem of ℓ is convex
- *Shrinkage* method
- called *Ridge regression* in the regression case.

Effect of regularisation $p = 9, n = 10$



How about semi-supervised learning?

- Labeled data

$$\mathcal{L} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

- Unlabeled data

$$\mathcal{U} = \{(x_{n+1}, y_{n+1}), \ldots, (x_{n+m}, y_{n+m})\}$$

How can we use unlabeled data?

Semi-supervised learning is impossible...

The goal of learning is to minimize the risk

$$\mathbb{E}[\ell(f(X), Y)]$$

If the function f can be independently optimized for each value of x this requires to minimize

$$\mathbb{E}[\ell(f(x), Y)|X = x] = \int \ell\big(f(x), y\big)\, p(y|x)\, dy, \quad \text{for all } x.$$

- This only requires to know $P_{Y|X}$ and not P_X.

So knowing P_X is useless.

- In fact not quite, but $p_X(x)$ is not very informative if we do not know $p_{Y|X}(y|x)$
- Semi-supervised learning can only be marginally useful...
- These ideas have been used successfully for *covariate shift*.

Estimating P_X is impossible anyway

Curse of dimensionality

Density estimation is hopeless in high-dimension. (Unless the support of the data is low-dimensional and this information is leveraged).

But maybe some characteristics of p_X or some characteristics of its support can still be estimated...

Key assumptions of semi-supervised learning

P_X carries some information about $P_{Y|X}$

Semi-supervised classification

Cluster assumption

Points in the same cluster are likely to be in the same class

Low density separation

The decision boundary should be in a low density region.

Assumptions for the more general semi-supervised case

Smoothness assumption

If two points are close, so should be their corresponding outputs.

Manifold assumption

The data lies (roughly) on a low dimensional manifold of $\mathbb{R}^p$

Define a more informed hypothesis class

Instead of controlling the smoothness of the function learn based on a Sobolev or RKHS norm, require that the functions should be smooth with respect to the geometry of the data.

Transductive learning *vs* Semi-supervised learning

Transductive learning

- Test points are available at train time
- Only prediction at those test points is computed

Semi-supervised learning "beyond transduction"

- Some unlabelled points are available at train time.
- The predictor learned should generalize to new unseen points.

What is a cluster?

- Round?
  - k-means
- Ellipsoidal?
  - Gaussian mixtures
- Convex?
  - Mixture of log concave densities see Cule et al. (2010); Chang and Walther (2007)
- Connected?
  - Spectral Clustering

Goal of clustering

Supervised learning has well-specified formulations, the most common being to minimize the some expected risk.
The density...

- ... is a mixture of Gaussians $\rightarrow$ Recover the Gaussians
- ... is a sum of log-concave densities $\rightarrow$ Recover components
- ... has several modes
  $\rightarrow$ Recover the modes / Recover regions of density above noise level / Singular points of the density
- ... has a support which is made of several connected components + background noise.
  $\rightarrow$ Recover the different components of the support.

Why is this not more addressed in the literature?

- Hard to specify/model beyond a parametric setting

$\rightarrow$ Need more well-specified formulations

- Analysis hard because hard non-convex problems...

Transductive binary classification with min-cut

(Blum and Chawla, 2001)

Construct a weighted graph where each node corresponds to a label or unlabeled data point and solve.

$$\min_{f \in \mathbb{R}^{n+m}} \sum_{(i,j) \in E} w_{ij} |f_i - f_j| \quad \text{s.t.} \quad \forall i \in \{1, \ldots, n\},\ f_i = y_i.$$

The problem can be reformulated as min-cut problem:

- Add a source node and connect all positive examples to it with an infinite capacity
- Add a sink node and connect all negative examples to it with an infinite capacity
- Set to capacity of the edges to be equal to the weights w_{ij}

Solve the min-cut problem.

Spectral Clustering

(see the tutorial of Von Luxburg (2007))

Graph cuts

Let $G = (V, W)$ a weighted graph and $A, B \subset V$ two disjoint subsets.

$$\mathrm{cut}(A, B) = \sum_{i \in A, j \in B} w_{ij}$$

A natural way to partition a graph is to minimize

$$\mathrm{cut}(A_1, A_2, \ldots, A_k) = \sum_{i=1}^{k} \mathrm{cut}(A_i, \bar{A}_i)$$

Balanced cuts

Problem: we want the cutst to be somehow "balanced".

$$\text{RatioCut}(A_1, A_2, \ldots, A_k) = \sum_{i=1}^{k} \frac{\text{cut}(A_i, \bar{A}_i)}{|A_i|}$$

$$\text{NormalizedCut}(A_1, A_2, \ldots, A_k) = \sum_{i=1}^{k} \frac{\text{cut}(A_i, \bar{A}_i)}{\text{vol}(A_i)}$$

with $\text{vol}(A) = \sum_{i \in A} d_i$.

- Problem: the problem is now **NP-hard !** (Wagner and Wagner, 1993)
- Spectral clustering is a heuristic based on relaxations of these problems.

Laplacians

- Given a graph $G = (V, E)$ with $|V| = n$,
- let $W \in \mathbb{R}^{n \times n}$ a matrix of edge weights, i.e. with $w_{ij} = 0$ for all $(i, j) \notin E$.
- Let D the diagonal matrix with $D_{ii} = d_i = \sum_{i'=1}^{n} w_{ii'}$.

Define

- the **graph Laplacian** as

$$L = D - W,$$

- the (symmetric) **normalized graph Laplacian** as

$$\tilde{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I_n - D^{-\frac{1}{2}} W D^{-\frac{1}{2}},$$

- the **"random walk" Laplacian** as

$$\tilde{L}_{\mathrm{rw}} = D^{-1} L.$$

Properties of Laplacians eigenvectors

- If M is the incidence matrix $M_{(i,j),k} = \delta(i,k) - \delta(j,k)$ then

$$L = M^\top M \succeq 0.$$

- $\mathbf{1} \in \mathrm{Ker}(L), \mathbf{1} \in \mathrm{Ker}(\tilde{L}_{\mathrm{rw}}) = \mathrm{Ker}(D^{-1}L)$ since

$$(L\mathbf{1})_i = d_i - \sum_j w_{ij} = 0.$$

  If $\tilde{L}_{\mathrm{rw}} u = \lambda u$, denoting $\tilde{u} = D^{1/2} u$ then

$$\tilde{L}\tilde{u} = D^{-1/2} L D^{-1/2} D^{1/2} u = \lambda D^{-1/2} D u = \lambda \tilde{u}.$$

- $D^{1/2}\mathbf{1} \in \mathrm{Ker}(\tilde{L})$.
- If the graph has several connected components, the Laplacian is block-diagonal for an appropriate ordering of the Laplacian with blocks $B_1, \ldots, B_k$ and $\mathbf{1}_{B_j} \in \mathrm{Ker}(L)$ and $\mathrm{Ker}(\tilde{L}_{\mathrm{rw}})$.

So clustering can be obtained by a permutation + identification of the blocks...

Expressing RatioCut with the Laplacian for $k = 2$

Define

$$f_i = \begin{cases} \sqrt{\frac{|\bar{A}|}{|A|}} & \text{if } i \in A \\ -\sqrt{\frac{|A|}{|\bar{A}|}} & \text{if } i \in \bar{A}. \end{cases}$$

We then have

$$\begin{aligned}
f^\top L f &= \sum_{i,j=1}^{n} w_{ij} (f_i - f_j)^2 \\
&= \sum_{i \in A, j \in \bar{A}} w_{ij} \left(\sqrt{\tfrac{|\bar{A}|}{|A|}} + \sqrt{\tfrac{|A|}{|\bar{A}|}} \right)^2 + \sum_{i \in \bar{A}, j \in A} w_{ij} \left(-\sqrt{\tfrac{|A|}{|\bar{A}|}} - \sqrt{\tfrac{|\bar{A}|}{|A|}} \right)^2 \\
&= 2 \operatorname{cut}(A, \bar{A}) \left(\frac{|\bar{A}|}{|A|} + \frac{|A|}{|\bar{A}|} + 2 \right) \\
&= 2 \operatorname{cut}(A, \bar{A}) \left(\frac{|\bar{A}| + |A|}{|A|} + \frac{|A| + |\bar{A}|}{|\bar{A}|} \right) \\
&= 2|V| \operatorname{RatioCut}(A, \bar{A}).
\end{aligned}$$

Expressing RatioCut with the Laplacian for $k = 2$ (II)

But if

$$f_i = \begin{cases} \sqrt{\frac{|\bar{A}|}{|A|}} & \text{if } i \in A \\ -\sqrt{\frac{|A|}{|\bar{A}|}} & \text{if } i \in \bar{A}. \end{cases}$$

then

$$\sum_{i=1}^{n} f_i = \sum_{i \in A} \sqrt{\frac{|\bar{A}|}{|A|}} - \sum_{i \in \bar{A}} \sqrt{\frac{|A|}{|\bar{A}|}} = |A| \sqrt{\frac{|\bar{A}|}{|A|}} - |\bar{A}| \sqrt{\frac{|A|}{|\bar{A}|}} = 0$$

and

$$\|f\|^2 = \sum_{i=1}^{n} f_i^2 = |A| \frac{|\bar{A}|}{|A|} + |\bar{A}| \frac{|A|}{|\bar{A}|} = n$$

A relaxation for RatioCut ($k = 2$)

We rewrite $\min_{A \subset V} \text{RatioCut}(A, \bar{A})$ as

$$\min_{f \in \mathbb{R}^n,\, A \subset V} f^\top L f \qquad \text{s.t.} \qquad f_i = \begin{cases} \sqrt{\frac{|\bar{A}|}{|A|}} & \text{if } i \in A \\ -\sqrt{\frac{|A|}{|\bar{A}|}} & \text{if } i \in \bar{A}. \end{cases}$$

Given previous remarks a natural relaxation is

Relaxation of RatioCut as a spectral problem

$$\min_{f \in \mathbb{R}^n} f^\top L f \quad \text{s.t.} \quad f \perp \mathbf{1}_n, \quad \|f\|_2^2 = n.$$

The solution is the so-called *Fiedler vector*, i.e. the eigenvector associated to the second smallest eigenvalue.

RatioCut via the Laplacian ($k \geq 2$)

Define $H \in \mathbb{R}^{n \times k}$ by $H_{ij} = \frac{1}{\sqrt{|A_j|}} \mathbf{1}_{\{i \in A_j\}}$.

If $H = [h_1, \ldots, h_k]$, then

$$h_j^\top L\, h_j = \frac{1}{|A_j|} \sum_{i,i' \in A_j} L_{ii'} = \frac{1}{|A_j|} \sum_{i \in A_j, i' \notin A_j} L_{ii'} = \frac{\mathrm{cut}(A_j, \bar{A}_j)}{|A_j|}.$$

So that

$$\mathrm{tr}(H^\top L H) = \sum_{j=1}^{k} h_j^\top L h_j = \sum_{j=1}^{k} \frac{\mathrm{cut}(A_j, \bar{A}_j)}{|A_j|} = \mathrm{RatioCut}(A_1, \ldots, A_k).$$

Moreover, $H^\top H = I_k$

Spectral relaxation of Ratio cut

RatioCut can be rewritten as

$$\min_{\substack{H \in \mathbb{R}^{n \times k}, \\ A_1, \ldots, A_k \subset V}} \operatorname{tr}(H^\top L H) \quad \text{s.t.} \quad H^\top H = I_k, \quad \forall (i,j), h_{ij} = \frac{1}{\sqrt{A_j}} \mathbf{1}_{\{i \in A_j\}}.$$

Relaxation of RatioCut as a spectral problem

$$\min_{H \in \mathbb{R}^{n \times k}} \operatorname{tr}(H^\top L H) \quad \text{s.t.} \quad H^\top H = I_k.$$

The solution H^* is composed of the eigenvectors associated with the k smallest eigenvalues of L.

Normalized cut via the Laplacian

Define $H \in \mathbb{R}^{n \times k}$ by $H_{ij} = \frac{1}{\sqrt{\mathrm{vol}(A_j)}} \mathbf{1}_{\{i \in A_j\}}$.

If $H = [h_1, \ldots, h_k]$, then

$$h_j^\top L\, h_j = \frac{1}{\mathrm{vol}(A_j)} \sum_{i,i' \in A_j} L_{ii'} = \frac{1}{\mathrm{vol}(A_j)} \sum_{i \in A_j, i' \notin A_j} L_{ii'} = \frac{\mathrm{cut}(A_j, \bar{A}_j)}{\mathrm{vol}(A_j)}.$$

So that

$$\mathrm{tr}(H^\top L H) = \sum_{j=1}^{k} h_j^\top L h_j = \sum_{j=1}^{k} \frac{\mathrm{cut}(A_j, \bar{A}_j)}{\mathrm{vol}(A_j)} = \mathsf{NormalizedCut}(A_1, \ldots, A_k).$$

Moreover, $h_j^\top D h_j = \frac{1}{\mathrm{vol}(A_j)} \sum_{i \in A_j} d_i = 1$ so that $H^\top D H = I_k$.

Spectral relaxation for RatioCut

$$\min_{H \in \mathbb{R}^{n \times k}} \operatorname{tr}(H^\top L H) \quad \text{s.t.} \quad H^\top D H = I_k.$$

The solution H^* are the eigenvectors associated with the k

smallest eigenvalues for the generalized eigenvalue problem of finding u such that $Lu = \lambda Du$.

Spectral clustering algorithm for Ratio Cut

Algorithm 1 Ratio Cut

1: Let $L = D - W$
2: Let $U = [u_1, \ldots, u_k]$ be the matrix of the k first eigenvectors.
3: Cluster the rows of U

Spectral clustering algorithms for Normalized Cut

Algorithm 2 Diffusion maps (Shi and Malik, 2000)

1: Let $L = D - W$ be the random walk Laplacian.
2: Find the k first generalized eigenvectors solving $Lu = \lambda Du$
3: Let $U = [u_1, \ldots, u_k]$
4: Cluster the rows of U

Algorithm 3 Normalized Cut (Ng et al., 2002)

1: Let $\tilde{L} = I - D^{-1/2} W D^{-1/2}$
2: Let $U = [u_1, \ldots, u_k]$ be the matrix of the k first eigenvectors.
3: Normalize the rows of U so that they have unit Euclidean norm.
4: Cluster the rows of U

NB: These spectral clustering algorithms can be used in the transductive learning setting.

Random walk interpretation of NormalizedCut

- $P = I - \tilde{L}_{\mathrm{rw}} = D^{-1}W$ is a stochastic matrix ($P\mathbf{1} = \mathbf{1}$).
- It defines the probability transition for a random walk on the graph.
- The stationary distribution is $\pi_i = \frac{d_i}{\mathrm{vol}(V)}$.

Indeed, $$\pi^\top P = \frac{1}{\mathrm{vol}(V)}\mathbf{1}^\top DD^{-1}W = \frac{1}{\mathrm{vol}(V)}d^\top = \pi^\top.$$

Proposition

If $X \sim \pi$ and $Y|X \sim P$ then

$$\mathrm{NormalizedCut}(A, \bar{A}) = \mathbb{P}(Y \in \bar{A} \mid X \in A) + \mathbb{P}(Y \in A \mid X \in \bar{A}).$$

Proof.

$$\mathbb{P}(Y \in B, X \in A) = \sum_{i \in A, j \in B} \pi_i p_{ij} = \frac{1}{\mathrm{vol}(V)} \sum_{i \in A, j \in B} w_{ij} = \frac{\mathrm{cut}(A, B)}{\mathrm{vol}(V)},$$

$$\text{and} \quad \mathbb{P}(X \in A) = \sum_{i \in A} \frac{d_i}{\mathrm{vol}(V)} = \frac{\mathrm{vol}(A)}{\mathrm{vol}(V)}.$$

Towards diffusion maps :a random walk

Stochastic Matrix: $P = D^{-1}W$.

$$P = D^{-1/2}P_sD^{1/2} \quad \text{with} \quad P_s = D^{-1/2}WD^{-1/2} = I - \tilde{L} = V\Lambda V^\top$$

$$P = \Psi\Lambda\Phi^\top \quad \text{with} \quad \Psi = [\psi_1, \ldots, \psi_n],\ \Phi = [\phi_1, \ldots, \phi_n].$$

right eigenvector $\phi_j = D^{1/2}v_j$ left eigenvector: $\psi_j = D^{-1/2}v_j$

Eigenvectors are bi-orthogonal: $\langle \psi_i, \phi_j \rangle = \delta_{i,j}$

The stationary distribution is $\phi_0 = \frac{d}{d^\top \mathbf{1}}$.

Random walk (discrete diffusion)

$$P_{ij} = \mathbb{P}(X_{t+1} = x_j \mid X_t = x_i)$$

$$P^t = D^{-1/2}P_sD^{1/2} = \mathbf{1}\phi_0^\top + \sum_{j=1}^{n} \lambda_j^t \psi_j \phi_j^\top$$

$$p_t(y|x) = \phi_0(y) + \sum_j \lambda_j^t \psi_j(x)\phi_j(y)$$

Diffusion distance and diffusion map

$$D_t^2(x_0, x_1) = \sum_y \sum \left(p_t(y|x_1) - p_t(y|x_0)\right)^2 \frac{1}{\phi_0}$$

$$\begin{aligned} D_t^2 &= P^t D^{-1} (P^t)^\top \mathbf{1}^\top D \mathbf{1} \\ &= D^{-1/2} P_s^t D^{1/2} D^{-1} D^{1/2} P_s^t D^{-1/2} \mathrm{vol}(V) \\ &= D^{-1/2} P_s^{2t} D^{-1/2} \mathrm{vol}(V) \\ &= \Psi \Lambda^{2t} \Psi^\top \mathrm{vol}(V) \\ &= \Psi_t \Psi_t^\top \mathrm{vol}(V) \quad \text{with} \quad \Psi_t = \Psi \Lambda^t \end{aligned}$$

Ψ_t (datapoints $\times$ number of eigenvalues) is the diffusion map

$$D_t^2(x_0, x_1) = \sum_j \lambda_j^{2t} \left(\psi_j(x_0) - \psi_j(x_1)\right)^2$$

A tight "nonspectral" relaxation of balanced cuts when $k = 2$ (Hein and Setzer, 2011)

General formulation for "balanced cut" problems.

$$\min_A \frac{\mathrm{cut}(A, \bar{A})}{\hat{S}(A)}$$

Ratio Cut	$\hat{S}(A) = \lvert A\rvert \, \lvert \bar{A}\rvert \, \lvert V\rvert^{-1}$
Normalized Cut	$\hat{S}(A) = \mathrm{vol}(A)\,\mathrm{vol}(\bar{A})$
Cheeger Cut	$\min(\lvert A\rvert, \lvert \bar{A}\rvert)$

- Ratio of set functions
- Idea: obtain a relaxation by replacing the set function by its Lovász extension.
- take advantage of the fact that $A \mapsto \mathrm{cut}(A, \bar{A})$ is submodular.

Lovász extension

- Let $\hat{S}$ be a set function.
- Let f with $f_{i_1} \geq f_{i_2} \geq \ldots \geq f_{i_n}$.
- Denote $C_k = \{i_1, \ldots, i_k\}$

The Lovász extension of $\hat{S}$ is the function:

$$S(f) = \sum_{k=1}^{n} f_{(k)} \left(\hat{S}(C_k) - \hat{S}(C_{k-1}) \right).$$

For instance, the Lovász extension of the cut function

$$A \mapsto \mathrm{cut}(A, \bar{A}) \qquad \text{is} \qquad f \mapsto \sum_{i=1}^{n} w_{ij} |f_i - f_j|$$

This leads to the formulation:

$$\min_{f : \|f\|_2 = 1} \frac{\sum_{i=1} w_{ij} |f_i - f_j|}{S(f)}$$

Harmonic functions and interpolation on a graph

(Zhu et al., 2003)

- n labelled nodes + m unlabelled nodes
- Compute $L = D - W \in \mathbb{R}^{n+m \times n+m}$

Solve

$$\min f^\top L f \quad \text{s.t.} \quad f_i = y_i, \quad i = 1 \dots n.$$

- For each unlabeled node, the solution satisfies $\hat{f}_i = \frac{1}{d_i} \sum_j w_{ij} f_j$
- **Transductive** method: cannot generalize to new points.
- Corresponds to the conditional means of a joint Gaussian distribution $\propto \exp(-\frac{1}{2} f^\top L f)$.
- Can be computed by Gaussian belief propagation.
- Physical interpretation: equilibrium electric potential for a network of resistors with resistances w_{ij}.
- Useful to regularize the Laplacian $L' = \gamma I + L$.

Laplacian Regularization

(Belkin et al., 2006)

Smooth functions on a manifold

Assume that $\mathcal{M}$ is a smooth compact manifold. Then the Laplacian operator has a countable number of eigenvalues and we then have

$$Lf = \sum_{k=1}^{\infty} \lambda_k \psi_k \langle \psi_k, f \rangle \qquad \text{and} \qquad \langle f, L_{\mathcal{M}} f \rangle = \int_{\mathcal{M}} \|\nabla f(x)\|^2 \, dx.$$

- The eigenbasis of the Laplacian generalizes the Fourier basis to manifolds
- Recover usual Fourier if we consider a torus

Graph Laplacian

The graph Laplacian $L = D - W$ is a discrete counterpart of $L_{\mathcal{M}}$.

Semi-supervised learning setting

- Labeled data

$$\mathcal{L} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

- Unlabeled data

$$\mathcal{U} = \{(x_{n+1}, y_{n+1}), \ldots, (x_{n+m}, y_{n+m})\}$$

How to go beyond transductive methods?

- Learn a function defined on the whole space
- Need to define appropriate space of functions.

Reproducing kernel Hilbert space

Nice space of functions for non-parametric statistics and machine learning? Require that

- the *evaluation functionals* $f \mapsto f(x)$ be $\mathcal{C}^0$ for all $x \in \mathcal{X}$.
- the space should be a Hilbert space $\mathcal{H}$

Then by the Riesz representation theorem, there must exist an element $h_x \in \mathcal{H}$ such that

$$\forall f \in \mathcal{H}, \quad f(x) = \langle h_x, f \rangle_{\mathcal{H}}.$$

But then by definition $h_y(x) = \langle h_x, h_y \rangle_{\mathcal{H}} = h_x(y)$.
Define the *reproducing kernel* as the function

$$K : (x, y) \mapsto \langle h_x, h_y \rangle_{\mathcal{H}}.$$

By definition $h_x(\cdot) = K(x, \cdot)$ so that

$$f(x) = \langle K(x, \cdot), f \rangle_{\mathcal{H}} \quad \text{and} \quad \langle K(x, \cdot), K(y, \cdot) \rangle_{\mathcal{H}} = K(x, y).$$

A reproducing kernel is a positive definite function

The reproducing kernel is necessarily a *symmetric positive definite function* since for all $x_1, \ldots, x_n \in \mathcal{X}$, and all $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$.

$$\sum_{i,j} \alpha_i \alpha_j K(x_i, x_j) = \left\langle \sum_i \alpha_i K(x_i, \cdot), \sum_j \alpha_j K(x_j, \cdot) \right\rangle_{\mathcal{H}} \geq 0,$$

with equality if and only if $\alpha_i = 0$ for all i.

A space with these properties is called a *reproducing kernel Hilbert space* (RKHS).

Converse?

Yes, any symmetric positive definite function is the reproducing kernel of a RKHS (Aronszajn, 1950). We will show it next.

Constructing a RKHS from a symmetric pd function

Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a p.d. function. Consider the vector space of finite linear combinations of $K(x, \cdot)$ for $x \in \mathcal{X}$., i.e. with functions of the form

$$f(\cdot) = \sum_{i=1}^{n} \alpha_i K(x_i, \cdot) \quad \text{and} \quad g(\cdot) = \sum_{j=1}^{m} \beta_j K(x_j, \cdot)$$

Define $\langle K(x, \cdot), K(y, \cdot) \rangle_{\mathcal{H}} = K(x, y)$.

- By linearity, if $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ extends to a bilinear function we should have

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{i,j} \alpha_i \beta_j \langle K(x_i, \cdot), K(x_j, \cdot) \rangle_{\mathcal{H}} = \alpha_i \beta_j K(x_i, x_j).$$

- The value of this expression does not depend on the specific expansions of f and g

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{j} \beta_j f(x_j) = \sum_{i} \alpha_i g(x_i),$$

so this is a valid definition of a bilinear function.

RKHS construction II: Checking that $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is actually a dot product

- Conjugacy: since K is real valued and symmetric we have

$$\langle f, g \rangle_{\mathcal{H}} = \overline{\langle g, f \rangle_{\mathcal{H}}} = \langle g, f \rangle_{\mathcal{H}}$$

- Bilinearity: Satisfied by construction
- Positive-definiteness: Follows immediately from the fact that K is a positive definite function.

So the space of functions we defined $\mathcal{H}$ is a pre-Hilbert space.

RKHS construction III: Checking that $\mathcal{H}$ is complete

Consider a Cauchy sequence of functions $(f_n)_n$ in $\mathcal{H}$, i.e. such that

$$\forall \varepsilon > 0, : \exists N, \ \forall m, n \geq N, \ \|f_n - f_m\|_{\mathcal{H}} \leq \varepsilon.$$

By the Cauchy-Schwarz inequality

$$|f_n(x) - f_m(x)| \leq \|f_n - f_m\|_{\mathcal{H}} K(x, x),$$

so that the sequence of function must converge pointwise to a function f_∞ which satisfies $\|f_\infty\|_{\mathcal{H}} < \infty$ by the triangle inequality. We proved that $\mathcal{H}$ is complete and it is therefore a Hilbert space.

Common RKHSes for $\mathcal{X} = \mathbb{R}^p$

Linear kernel

- $K(x, y) = x^\top y$
- $\mathcal{H} = \{f_w : x \mapsto w^\top x \mid w \in \mathbb{R}^p\}$
- $\|f_w\|_\mathcal{H} = \|w\|_2$

Polynomial kernel

- $K_h(x, y) = (\gamma + x^\top y)^d$
- $\mathcal{H}$

Radial Basis Function kernel (RBF)

- $K_h(x, y) = \exp\left(-\frac{\|x-y\|_2^2}{2h}\right)$
- $\mathcal{H} =$ Gaussian RKHS

$\|f\|_{\mathcal{H}}$ measures the smoothness of the function f

Indeed:

$$|f(x)-f(x')| = |\left\langle f, K(x,\cdot) - K(x',\cdot)\right\rangle_{\mathcal{H}}| \leq \|f\|_{\mathcal{H}}\|K(x,\cdot)-K(x',\cdot)\|_{\mathcal{H}}$$

- f is Lipschitz with respect to the ℓ_2 distance induced by the RKHS

  $$d(x,x') = \|K(x,\cdot)-K(x',\cdot)\|_{\mathcal{H}} = \sqrt{K(x,x)+K(x',x')-2K(x,x')}$$

- $\|f\|_{\mathcal{H}}$ is the Lispchitz constant

Learning with functions from a RKHS

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i) \; + \; \lambda \, \|f\|_{\mathcal{H}} \tag{P}$$

Theorem (Representer theorem)

The solution of the regularized empirical risk minimization problem lies in the subspace of $\mathcal{H}$ generated by the point x_i, i.e.,

$$f^* = \sum_{i=1}^{n} \alpha_i K(x_i, \cdot) \quad \textit{for some } \alpha_i \in \mathbb{R}. \tag{R}$$

The solution of (P) is therefore of the form (R) with $\alpha \in \mathbb{R}^n$ the solution of

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \ell\Big(\sum_{j=1}^{n} \alpha_j K(x_j, x_i), y_i \Big) \; + \; \lambda \sum_{1 \le i,j \le n} \alpha_i \alpha_j K(x_i, x_j).$$

Regularizing with the Laplacian in the RKHS (Belkin et al., 2006)

$$\mathcal{L} = \{(x_1, y_1), \ldots, (x_n, y_n)\}, \quad \mathcal{U} = \{(x_{n+1}, y_{n+1}), \ldots, (x_{n+m}, y_{n+m})\}$$

Adding Laplacian regularization

$$\min_{f \in \mathcal{H}} \hat{\mathcal{R}}_n(f) + \lambda \|f\|_{\mathcal{H}} + \frac{1}{n+m} \sum_{1 \leq i,j \leq n+m} f(x_i)\, L_{ij}\, f(x_j)$$

New version of the representer theorem:

The optimal solution f^* can be represented as $f^* = \sum_{i=1}^{n+m} \alpha_i K(x_i, \cdot)$

Finite dimensional formulation

The problem then reduces to solving:

$$\min_{\alpha \in \mathbb{R}^{n+m}} \hat{\tilde{\mathcal{R}}}_n(\alpha) + \lambda \alpha^\top \mathbf{K} \alpha + \tfrac{1}{n+m} \alpha^\top \mathbf{K} L \mathbf{K} \alpha,$$

with $\mathbf{K} \in \mathbb{R}^{(n+m)^2}$ the kernel matrix defined by $\mathbf{K}_{ij} = K(x_i, x_j)$.

Challenges ahead

- Better estimation of the Laplacian (Belkin et al., 2009)
- Compress geometric information (Dasgupta and Freund, 2008)
- Go beyond transduction !
- Online algorithms? (Valko et al., 2010)
- Several manifolds
- Beyond manifolds
- ...

References I

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404.

Belkin, M., Niyogi, P., and Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *The Journal of Machine Learning Research*, 7:2399–2434.

Belkin, M., Sun, J., and Wang, Y. (2009). Constructing laplace operator from point clouds in r d. In *Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1031–1040. Society for Industrial and Applied Mathematics.

Blum, A. and Chawla, S. (2001). Learning from labeled and unlabeled data using graph mincuts. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 19–26. Morgan Kaufmann Publishers Inc.

Chang, G. T. and Walther, G. (2007). Clustering with mixtures of log-concave distributions. *Computational Statistics & Data Analysis*, 51(12):6242–6251.

Cule, M., Samworth, R., and Stewart, M. (2010). Maximum likelihood estimation of a multi-dimensional log-concave density. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(5):545–607.

Dasgupta, S. and Freund, Y. (2008). Random projection trees and low dimensional manifolds. In *Proceedings of the 40th annual ACM symposium on Theory of computing*, pages 537–546. ACM.

References II

Hein, M. and Setzer, S. (2011). Beyond spectral clustering-tight relaxations of balanced graph cuts. In *Advances in neural information processing systems*, pages 2366–2374.

Ng, A. Y., Jordan, M. I., Weiss, Y., et al. (2002). On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905.

Valko, M., Kveton, B., Ling, H., Daniel, T., et al. (2010). Online semi-supervised learning on quantized graphs. In *The 26nd Annual Conference on Uncertainty in Artificial Intelligence*.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416.

Wagner, D. and Wagner, F. (1993). *Between min cut and graph bisection*. Springer.

Zhu, X., Ghahramani, Z., Lafferty, J., et al. (2003). Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, volume 3, pages 912–919.