Problem 1

Let $\mathcal{M}$ be a Riemannian manifold with some volume measure $\mu$. For $t \in [0, \infty)$ and $x \in D$, a “well behaved” domain on $\mathcal{M}$ consider the heat equation

$$u_t(t,x) = \Delta u(t,x)$$
$$u(0,x) = f(x)$$

where $\Delta$ is Laplace-Betrami operator (Laplacian) on $\mathcal{M}$. The eigenfunctions $\{\phi_i\}_{i=1}^{\infty}$ of $-\Delta$ form an orthonormal basis on $L^2(\mathcal{M})$ with corresponding eigenvalues $\lambda_i \geq 0$.

The general solution to (1) is given by

$$u(t,x) = \int_D k_t(x,y)f(y)d\mu(y)$$

where $k_t(x,y)$ is referred to as the heat kernel.

(a) Verify that the general solution of (1) with heat kernel

$$k_t(x,y) = \sum_{i=0}^{\infty} \exp(-\lambda_i t) \phi_i(x)\phi_i(y)$$

is indeed a solution.

[Hint: $k_t(x,y)$ satisfies $\lim_{t \to 0} k_t(x,y) = \delta_x(y).$]

(b) Define the diffusion distance function $d_t(x,y)$ at time $t > 0$ by

$$d_t(x,y) = \left(\int_{\mathcal{M}} (k_{t/2}(x,z) - k_{t/2}(z,y))^2 d\mu(z)\right)^{1/2}$$

with $k_t(x,y)$ given by (2). Verify $d_t(x,y)$ is a metric.

[Hint: recall that $L^2$ is a set of equivalence classes of functions that are equal almost everywhere.]

(c) Define an embedding $\Phi : \mathcal{M} \hookrightarrow L^2(\mathcal{M})$ as

$$\Phi(x) = \left(e^{-\lambda_0 t/2}\phi_0(x), e^{-\lambda_1 t/2}\phi_1(x), e^{-\lambda_2 t/2}\phi_2(x), \ldots\right)^T$$

Show that how we can compute $d_t(x,y)$ using and oracle for $\Phi(\cdot)$.

[Hint: First verify that $k_t(x,y)$ satisfies the Chapman-Kolmogorov equations.]
Solution

(a) We substitute (1) into both sides of the heat equation to obtain

\[ u_t(t, x) = \frac{\partial}{\partial t} \int_D k_t(x, y) f(y) dy = \frac{\partial}{\partial t} \int_D \sum_i \exp(-\lambda_i t) \phi_i(x) \phi_i(y) f(y) dy \]

\[ = - \int_D \sum_i \lambda_i \exp(-\lambda_i t) \phi_i(x) \phi_i(y) f(y) dy \]

and

\[ \Delta u(t, x) = \Delta \int_D k_t(x, y) f(y) dy = \int_D \sum_i \exp(-\lambda_i t) \Delta \phi_i(x) \phi_i(y) f(y) dy \]

\[ = - \int_D \sum_i \lambda_i \exp(-\lambda_i t) \phi_i(x) \phi_i(y) f(y) dy \]

since \(-\Delta \phi_i(x) = \lambda_i \phi_i(x)\) and so \(u_t(t, x) = \Delta u(t, x).\) And to verify the initial condition, for a smooth solution we have

\[ u(0, x) = \lim_{t \downarrow 0} \int_D k_t(x, y) f(y) dy = \int_D \lim_{t \downarrow 0} k_t(x, y) f(y) dy = \int_D \delta_x(y) f(y) dy = f(x) \]

using the symmetry of the kernel function.

(b) First note that the kernel is symmetric since

\[ k_t(x, y) = \sum_{i=0}^\infty \exp(-\lambda_i t) \phi_i(x) \phi_i(y) = \sum_{i=0}^\infty \exp(-\lambda_i t) \phi_i(y) \phi_i(x) = k_t(y, x) \]

then to show that \(d_t(x, y)\) is a metric we show the following.

(i) Since \((k_{t/2}(x, z) + k_{t/2}(z, y))^2 \geq 0\) we have \(d_t(x, y) \geq 0\) which is zero whenever \(x = y.\) Furthermore if \(d_t(x, y) = 0\) then \(k_{t/2}(x, z) - k_{t/2}(z, y) = 0\) almost everywhere. Thus in \(L_2\) for all \(z\) by symmetry of the kernel we have \(k_{t/2}(x, z) = k_{t/2}(z, y)\) and so \(x = y.\)

(ii) Since \((k_{t/2}(x, z) - k_{t/2}(z, y))^2 = (k_{t/2}(y, z) - k_{t/2}(x, z))^2 = (k_{t/2}(y, z) - k_{t/2}(y, x))^2\) where the last step follows by symmetry of the kernel. Thus \(d_t(x, y) = d_t(y, x).\)

(iii) Since \(k_t(x, y)\) has a basis in \(L^2(\mathcal{M})\) we have that \(k_{t/2}(x, z) - k_{t/2}(y, z)\) is in \(L^2(\mathcal{M})\) and the result follows by Minkowski’s inequality since

\[ d_t(x, y) = \|k_{t/2}(x, \cdot) - k_{t/2}(\cdot, y)\|_2 \leq \|k_{t/2}(x, \cdot) - k_{t/2}(\cdot, w) + k_{t/2}(w, \cdot) - k_{t/2}(\cdot, y)\|_2 \]

\[ = \|k_{t/2}(x, \cdot) - k_{t/2}(\cdot, w)\|_2 + \|k_{t/2}(w, \cdot) - k_{t/2}(\cdot, y)\|_2 \]

\[ = d(x, w) + d(w, y) \]

(c) We can show that the kernel satisfies \(k_{t+s}(x, y) = k_t(x, z)k_s(z, y)\) (Chapman-Kolmogorov) as follows.

\[ \int_{\mathcal{M}} k_t(x, z) k_s(z, y) d\mu(z) = \int_{\mathcal{M}} \sum_i \exp(-\lambda_i t) \phi_i(x) \phi_i(z) \sum_j \exp(-\lambda_j s) \phi_j(z) \phi_j(y) d\mu(z) \]

\[ = \int_{\mathcal{M}} \sum_{i,j} \exp(-\lambda_i t - \lambda_j s) \phi_i(x) \phi_i(z) \phi_j(z) \phi_j(y) d\mu(z) \]

\[ = \sum_i \exp(-\lambda_i (t + s)) \phi_i(x) \phi_i(y) \int_{\mathcal{M}} \phi_i^2(z) d\mu(z) \]

\[ = k_{t+s}(x, y) \]
since $\phi_i$’s form an orthonormal basis for $L_2(M)$. Expanding the square in the definition we have

$$
\begin{align*}
\frac{d^2_t(x,y)}{2} &= \int_M k_{t/2}(x,z) k_{t/2}(z,y) d\mu(z) - 2 \int_M k_{t/2}(x,z) k_{t/2}(z,y) d\mu(z) + \int_M k_{t/2}(y,z) k_{t/2}(z,y) d\mu(z) \\
&= k_t(x,x) - 2 k_t(x,y) + k_t(y,y) \\
&= \exp(-\lambda t) (\phi_i(x) \phi_i(x) - 2 \phi_i(x) \phi_i(y) + \phi_i(y) \phi_i(y)) \\
&= \sum_i \exp(-\lambda t) (\phi_i(x) - \phi_i(y))^2 \\
&= \sum_i (\Phi_i(x) - \Phi_i(y))^2 \\
&= \|\Phi_i(x) - \Phi_i(y)\|^2.
\end{align*}
$$

Thus given an oracle for $\Phi(\cdot)$ we can calculate $d_t(x,y)$ using a Euclidean norm.

**Problem 2**

Consider a linear classifier SVM. In class we have shown that by introducing slack variables $\xi = (\xi_1, \ldots, \xi_n)$ we can find a SVM solution by defining an optimization problem:

$$
\min_{w,b} \frac{1}{2} \|w\|^2 + \lambda \|\xi\|_p^p
$$

s.t. $y_i(w^T x_i + b) \geq 1 - \xi_i$, $\xi_i \geq 0$, $y = 1, \ldots, n$

where we’ve used the $L_p$ penalty function to enforce the constraints.

(a) Derive the dual for standard $L_1$ ($p = 1$) penalty function.

(b) Do the same for the $L_2$ penalty. Compare to the dual in part (a).

Solution:

(a) For the $L_1$ penalty the Lagrangian takes the form

$$
L(w, b, \alpha, \beta, \xi) = \frac{1}{2} \sum_{i=1}^n w_i^2 + \lambda \left( \sum_{i=1}^n \xi_i - \sum_{i=1}^n \beta_i \xi_i - \sum_{i=1}^n \alpha_i \left( y_i \left( \sum_{j=1}^n w_j x_i^{(j)} + b \right) - 1 + \xi_i \right) \right)
$$

with $\alpha_i, \beta_i \geq 0$ enforcing the inequality constraints. At the optimal optimal point this satisfies

$$
\begin{align*}
\frac{\partial L}{\partial w_j} &= w_j - \sum_{i=1}^n \alpha_i y_i x_i^{(j)} = 0 \\
\frac{\partial L}{\partial \beta_j} &= \lambda - \beta_j - \alpha_j = 0 \\
\frac{\partial L}{\partial \alpha_j} &= \sum_{i=1}^n \alpha_i y_i = 0
\end{align*}
$$

$\Rightarrow w_j = \sum_{i=1}^n \alpha_i y_i x_i^{(j)} \quad \forall j$

$\Rightarrow \alpha_j = \lambda - \beta_j \quad \forall j$

$\Rightarrow \sum_{i=1}^n \alpha_i y_i = 0$
Then for all $i = 1, \ldots, n$, $\beta_i \geq 0$ implies $0 \leq \alpha_i \leq \lambda$ and

$$L (w, b, \alpha, \beta, \xi) = \frac{1}{2} \sum_{i=1}^{n} \left( \sum_{j=1}^{n} \alpha_i y_i x_i^{(j)} \right)^2 + \sum_{i=1}^{n} \alpha_i \xi_i - \sum_{i=1}^{n} \alpha_i y_i \sum_{j=1}^{n} \left( \sum_{k=1}^{n} \alpha_k y_k x_k^{(j)} \right) x_i^{(j)} + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i \xi_i$$

$$= \frac{1}{2} \sum_{i=1}^{n} \left( \sum_{j=1}^{n} \alpha_i y_i x_i^{(j)} x_i \right) - \sum_{i=1}^{n} \alpha_i y_i \sum_{k=1}^{n} \alpha_k y_k \sum_{j=1}^{n} x_k^{(j)} x_i^{(j)} + \sum_{i=1}^{n} \alpha_i$$

$$= \frac{1}{2} \sum_{i,k} \alpha_i y_i \alpha_k y_k \sum_{j=1}^{n} x_i^{(j)} x_k - \sum_{i,k} \alpha_i y_i \alpha_k y_k \sum_{j=1}^{n} x_k^{(j)} x_i^{(j)} + \sum_{i=1}^{n} \alpha_i$$

$$= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,k} \alpha_i y_i \alpha_k y_k \langle x_i, x_k \rangle$$

Therefore the dual is given by

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,k} \alpha_i y_i \alpha_k y_k \langle x_i, x_k \rangle$$

s.t. $\sum_{i=1}^{n} \alpha_i y_i = 0$

$$0 \leq \alpha_i \leq \lambda \quad \forall i = 1, \ldots, n$$

(b) For the $L_2$ penalty the Lagrangian takes the form

$$L (w, b, \alpha, \beta, \xi) = \frac{1}{2} \sum_{i=1}^{n} w_i^2 + \lambda \sum_{i=1}^{n} \xi_i^2 - \sum_{i=1}^{n} \beta_i \xi_i - \sum_{i=1}^{n} \alpha_i \left( y_i \left( \sum_{j=1}^{n} w_j x_i^{(j)} + b \right) - 1 + \xi_i \right)$$

with $\alpha_i, \beta_i \geq 0$ enforcing the inequality constraints. At the optimal optimal point the only conditions different than in part (a) are

$$\frac{\partial L}{\partial \xi_i} = 2\lambda \xi_i - \beta_j - \alpha_j = 0 \quad \Rightarrow \alpha_j + \beta_j = 2\lambda \xi_i \quad \forall j.$$  

Then

$$L (w, b, \alpha, \beta, \xi) = -\frac{1}{2} \sum_{i,k} \alpha_i y_i \alpha_k y_k \langle x_i, x_k \rangle + \sum_{i=1}^{n} \lambda \xi_i^2 - (\beta_i + \alpha_i) \xi_i + \sum_{i=1}^{n} \alpha_i$$

$$= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,k} \alpha_i y_i \alpha_k y_k \langle x_i, x_k \rangle - \sum_{i=1}^{n} \lambda \xi_i^2$$

$$= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,k} \alpha_i y_i \alpha_k y_k \langle x_i, x_k \rangle - \frac{1}{4\lambda} \sum_{i=1}^{n} (\alpha_i + \beta_i)^2$$

and $\beta \geq 0$ implies $\beta_i = 0$ and the dual is given by

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,k} \alpha_i y_i \alpha_k y_k \langle x_i, x_k \rangle - \frac{1}{4\lambda} \sum_{i=1}^{n} \alpha_i^2$$

s.t. $\sum_{i=1}^{n} \alpha_i y_i = 0$

$$\alpha_i \geq 0 \quad \forall i = 1, \ldots, n$$
Problem 3

In this problem we explore how to use the machinery of Problem 1 in order reduce the dimension of a discrete set of $k$ points $X$. We assume that the points in $X$ are sampled from a Riemannian manifold $\mathcal{M} \subseteq \mathbb{R}^n$ and thus approximate it. A useful fact is that the heat kernel $k_t(x,y)$ is a transitional density function of a Brownian motion on $\mathcal{M}$ stopped at the boundary $\partial D$.

Motivated by this we define a random walk $(X_t : t = i \Delta t)$ for $i \in \mathbb{N}$ on an undirected graph $G(X, E, w)$ which converges to a diffusion process as $n \to \infty$ and $\Delta t \to 0$. Here $w : E \to \mathbb{R}_+$ assigns weights to edges in the graph and we let $W(G)$ be the weighted adjacency matrix i.e. $W_{uv} = w(\{u, v\})$. Furthermore, let $D$ be a diagonal matrix with entries equal to the rows (or equivalently the columns) of $W$ and define the Laplacian of $G$ by $L = D - W$.

(a) Fix some $\Delta t > 0$. Define $P = D^{-1}W$ as the transition matrix for a random walk on the graph $G$ and let $p_t(x,y)$ be the probability of going from $x$ to $y$ in time $t$. Analogously to Problem 1, define a distance between $x, y \in X$ as

$$d_t(x,y) = \left( \sum_{z \in X} \left( p_{t/2}(x,z) - p_{t/2}(z,y) \right)^2 v_t(z) \right)^{1/2}$$

for some $v : X \to \mathbb{R}$. Show that this is not a metric. For what class of graphs is this a semi-metric for all $t > 0$? (i.e. all properties of a metric properties hold except $d_t(x,y) = 0$ does not imply $x = y$).

Modify the definition of $\hat{d}_t(x,y)$ in part (a) in order to obtain a semi-metric $\hat{d}_t(x,y)$ on $X$. Provide and example of a graph for which $d_t(x,y) = 0$ and $x \neq y$ for all $t > 0$.

(b) Quantitatively state why the diffusion distance $d_t(x,y)$ is an appropriate measure of closeness for points $x, y$ on the manifold. Explain why it is not a good choice when we want to cluster a set of points in $\mathbb{R}^n$ based on a measure of closeness given by the standard Euclidean norm $\|x - y\|_2$.

(c) Derive a map $\Phi : X \to \mathbb{R}^m$ such that the diffusion distance over points in $X$ is approximated by the Euclidian norm $\| \cdot \|_2$ over the embedding in $\mathbb{R}^m$. What is the function $v$ and the value $m$?

(d) Relate the map $\Phi$ to an eigenvalue problem in terms of the Laplacian $L$ of the graph $G(X, E, w)$. Suppose you want to reduce $X$ to points in a single dimension. What minimization problem can you solve? Now, given that the minimization problem can be stated as

$$\min_{y \top y = 1} \| ye^\top - e y^\top \|_F$$

where $e$ is a vector of all ones, describe the original point set $X$.

(e) Suppose that the solution $y$ to the optimization problem

$$\min \frac{y^\top Ly}{y^\top Dy}$$

s.t. $y^\top De = 0$

has entries $y_i \in \{-1, 1\}$. What type of graph partitioning problem did we solve? What does the distribution of the $y_i$’s over $\{-1, 1\}$ say about the solution?

Solution

(a) For a general undirected graph $G$ the probability matrix $P$ will not be symmetric (e.g. consider a star graph). So $p_{t/2}(x,z) - p_{t/2}(z,x) \neq 0$ in general which implies $\hat{d}_t(x,y) \neq 0$ for most graphs. However, it is easy to see that for any $d$-regular graph with all edge weights the same, the matrix is symmetric and so $d_t(x,y) = d_t(y,x)$. All other semi-metric properties can be shown to hold as in problem 1.
Since transition probabilities are no symmetric in general we define the distance function to be
\[
{d_t}(x, y) = \left( \sum_{z \in X} \left( p_{t/2}(x, z) - p_{t/2}(y, z) \right)^2 v^2(z) \right)^{1/2}
\]
which is a semi-metric analogously to problem 1. It is not hard to see that a graph with \( V = \{1, 2, 3, 4\} \) and \( E = \{\{1, 2\}, \{2, 3\}, \{3, 4\}, \{4, 1\}\} \) with unit edge weights has \( d_t(1, 3) = 0 \).

(b) Using the results of problem 1 we see that the diffusion distance approximates the spread of heat on a manifold. Thus the distance between two points is proportional to the time it takes to get from one to the other on the manifold. Two points that are close together with respect to the Euclidian metric may not be close if one is to travel from one to the other on the manifold. Thus the diffusion distance would be a bad choice if we want to cluster points in Euclidian space.

(c) Let \( \phi_j \) and \( \psi_j \) be left and right eigenvectors of \( P \) respectively. Then \( p_s(x, y) \) is given as the \( xy \) entry of \( P^s \) and using the eigen-decomposition of this matrix
\[
p_s(x, y) = \sum_{i=0}^{k-1} \lambda_i^s \psi_i(x) \phi_i(y)
\]
where \( \phi_i(x) \) is the entry corresponding to \( x \) in the \( i \)th left eigenvector (and similarly for \( \psi_i(y) \)). Then
\[
D^{-1}W\psi = \lambda\psi \\
\phi D^{-1}W = \lambda\phi
\]
and its easy to see by substitution that
\[
\phi_i^\top = \frac{1}{\sum_{x \in X} D_{xx}} D\psi_i \quad \text{and} \quad \phi_i \psi_j = \delta_{ij} \quad \text{(*)}
\]
The normalization factor ensures that the left eigenvectors have unit norm. Analogously to problem 1 we write the diffusion distance in terms of the eigen-decomposition of \( p_{t/2}(x, y) \) as
\[
{d_t^2}(x, y) = \sum_{z \in X} \left( p_{t/2}(x, z) - p_{t/2}(y, z) \right)^2 v^2(z)
\]
\[
= \sum_{x \in X} \sum_{i=0}^{k-1} \lambda_i^t (\psi_i(x) - \psi_i(y))^2 \phi_i^2(z) v^2(z)
\]
\[
= \sum_{i=0}^{k-1} \lambda_i^t (\psi_i(x) - \psi_i(y))^2 \sum_{z \in X} \phi_i(z) \phi_i(z) v^2(z)
\]
Letting \( v^2(z) = \sum_{x \in X} D_{xx} \) and using (*) we obtain
\[
{d_t^2}(x, y) = \sum_{i=0}^{k-1} \lambda_i^t (\psi_i(x) - \psi_i(y))^2
\]
\[
= \|\Phi(x) - \Phi(y)\|_2^2
\]
and it follows that the map is given by
\[
\Phi(x) = \left( \lambda_0^{t/2}\psi_0(x), \lambda_1^{t/2}\psi_1(x), \ldots, \lambda_{k-1}^{t/2}\psi_{k-1}(x) \right)^\top
\]
As a convention we can sort the eigenvalues from greatest to least in which case \( \psi_0 = e \), the vector of all ones, with a unit eigenvalue. The first entry is trivially one, and can be ignored so the value of \( m \) is effectively \( k - 1 \).
(d) The laplacian of $G$ is given by $L = D - W = D(I - P)$ in which case

$$L\psi = (D - W)\psi = D(I - P)\psi = D(\psi - \lambda \psi)$$

which yields the eigenvalue problem

$$D^{-1}L\psi = (1 - \lambda)\psi$$

in terms of the right eigenpair $(\lambda, \psi)$ of the transition matrix $P$. The map in terms of the normalized Laplacian $D^{-1}L$ is given by

$$\Phi(x) = \left( (1 - \lambda_1)^{1/2}\psi_1(x), \ldots, (1 - \lambda_{k-1})^{1/2}\psi_{k-1}(x) \right)^\top.$$

In class we’ve shown that minimum eigenvalue is given by the minimization

$$\min_{y^\top y = 1} y^\top Ly = \min_{y^\top y = 1} \frac{1}{2} \sum_{i,j} (y_i - y_j)^2 W_{ij}$$

in which case we take only the first entry of $\Phi$ and project the points down to a line. For a complete graph with $W_{ij} = 1$ this reduces to

$$\min_{y^\top y = 1} \frac{1}{2} \sum_{i,j} \|e y^\top - e^\top y\|_F^2$$

which is equivalent to the given minimization.

(e) The given minimization is the relaxed version of the normalized cut problem as described in Belkin & Niyogi. If the entries of the solution vector are in $\{0, 1\}$ we have solved the problem exactly so that $y_i = 1$ and $y_j = -1$ indicate that vertices $i$ and $j$ are in disjoint partitions. Thus the distribution of the $y_i$’s determines the sizes of the two partitions of the normalized cut.