Diffusion Maps

University of Vienna

Peter Mühlbacher

October 1, 2014

Abstract

Contents

Introduction

The contentual layout follows that of Coifman and Lafon (2006) and Belkin (2003). Some definitions, theorems and examples will be taken from these works as well.

0.1 Dimensionality Reduction

Chapter 1

Diffusion Maps

1.1 Diffusion Kernels

1.1.1 Motivation

In this section it will be elaborated on why global structures need not be preserved and how this leads to diffusion processes.

Preservation of Global Structures

Suppose data points $x_i \in \mathbb{R}^k$ are generated by a low dimensional parameter $\theta_i \in \mathbb{R}^{k'}$, $k' \ll k$ given a map $\Phi : \mathbb{R}^{k'} \to \mathbb{R}^k$. One problem is the (numerical) smoothness of Φ which is not necessarily given. Another problem is that for large k the euclidian distance is no longer a meaningful measure as the volume of the k-dimensional unit ball $\frac{\pi^{\frac{k}{2}}}{\Gamma(\frac{k}{2}+1)}$ converges to 0 as $lim_{k\to\infty}$ which is also known as the *curse of dimensionality*. One may conclude that large distances in the ambient space need not necessarily be preserved as they do not hold much information except that x_i is not "very close" to x_j.

Analogous to Riemannian manifolds (where metric tensors, inducing an inner product on the tangent space and a metric via the exponential map, define the manifold's geometry) we focus solely on local distances in order to recover intrinsic global structures.

A Dual Approach

From inverse problems in spectral geometry (e.g. "Can One Hear the Shape of a Drum?") it is known that much[1] of the geometry of a given set Γ can be derived from the analysis of functions defined on Γ.

In this work eigenvalues and eigenfunctions of averaging operators, i.e., operators whose kernel corresponds to transition probabilities of a Markov process, will be studied in order to define a diffusion map which embeds the data into a Euclidian space where the Euclidian distance is just the diffusion metric.

1.1.2 Construction of a Random Walk on the Data

Definitions

Let $(\Gamma, \mathcal{A}, \mu)$ be a measure space. In practical applications Γ is the given data set consisting of finitely many data points and μ is the counting measure to represent the distribution of the points in the data set. In addition, suppose we are given a symmetric kernel $k : \Gamma \times \Gamma \to \mathbb{R}^+$ which defines the local geometry of Γ.

Examples

Usually Γ is either a subset of the Euclidian space or a weighted graph.

In the first case it seems natural to write k as a function of the Euclidian distance $\nu(||x - y||)$.

In the second case let $b(x, y)$ be the associated adjacancy matrix, that is, $b(x, y) = 1$ if there is an edge going from x to y, and $b(x, y) = 0$ otherwise. The kernel b defines a notion of neighborhood for each point, and also a non-symmetric distance given by $1 - b(x, y)$. Clearly b is not symmetric in general, but we can consider

$$k_1(x, y) := \int_\Gamma b(x, u)b(y, u)d\mu(u)$$

$$k_2(x, y) := \int_\Gamma b(u, x)b(u, y)d\mu(u)$$

where $k_1(x, y)$ counts the number of common neighbors to x and y, while $k_2(x, y)$ counts the number of nodes for which x and y are common neighbors.

[1] the most famous example being Weyl's proof of $\#\{\lambda_k : \lambda_k < \lambda\} \approx \frac{\text{area}(\Gamma)}{2\pi}\lambda$ as $\lambda \to \infty$. (Canzani (2013))

Normalized Graph Laplacian Construction

Generally, such a kernel represents some notion of affinity between points of Γ and thus one can think of the data points as being the nodes of a symmetric graph whose weight function is specified by k. From the graph defined by (Γ, k), one can construct a reversible Markov chain on Γ.

To normalize the kernel, define

Definition 1.1.

$$v^2(x) = \int_\Gamma k(x,y)d\mu(y)$$

and

Definition 1.2.

$$p(x,y) = \frac{k(x,y)}{v^2(x)}.$$

$p(x,y)$ is no longer symmetric, but inherited the positivity and now satisfies a conservation property:

$$\int_\Gamma p(x,y)d\mu(y) = 1$$

As a result the matrix $P := (p(i,j))_{i,j}$ is stochastic and can be interpreted as the transition matrix of a homogeneous Markov process on Γ. In spectral graph theory $\mathbb{I} - P$ is commonly referred to as normalized, weighted graph Laplacian. This naming is justified in A.2.1[2].

better double check that claim in the footnote

To investigate the spectral properties of the corresponding integral operator P defined by $Pf(x) = \int_\Gamma p(x,y)f(y)d\mu(y)$ it is beneficial to examine the symmetric, conjugated Operator A.

Definition 1.3. *Let*

$$a(x,y) = \frac{k(x,y)}{v(x)v(y)} = v(x)p(x,y)\frac{1}{v(y)},$$

then the corresponding diffusion operator A is defined as

$$Af(x) = \int_\Gamma a(x,y)f(y)d\mu(y)$$

Notice that by definition of a one obtains a symmetric form and thus a symmetric operator A.

[2]Let w.l.o.g. $m_0 = 1$ and $m_2 = 2$. Now set $L_\varepsilon = \frac{\varepsilon I - G_\varepsilon}{\varepsilon}$; for uniformly sampled data the potential term ω vanishes and $\lim_{\varepsilon\to 0} L_\varepsilon = \Delta$

1.1.3 Diffusion Kernels

Theorem 1.1.1 (Spectral Properties of the Diffusion Operator)**.** *The diffusion operator A with kernel a is bounded from $L^2(\Gamma, d\mu)$ into itself, symmetric and positive semi-definite.*

Moreover, its norm is

$$||A|| = 1$$

and is taken by the eigenfunction

$$Av = v.$$

Proof. Let $f \in L^2(\Gamma, d\mu)$. We have:

$$\langle Af, f\rangle = \int_{\Gamma^2} k(x,y)\frac{f(x)}{v(x)}\frac{f(y)}{v(y)}d\mu(x)d\mu(y). \tag{1.1}$$

Applying the Cauchy-Schwartz inequality we get:

$$\begin{aligned}\left|\int_\Gamma k(x,y)\tfrac{f(y)}{v(y)}d\mu(y)\right| &= \left(\int_\Gamma k(x,y)d\mu(y)\right)^{\frac{1}{2}}\left(\int_\Gamma k(x,y)\tfrac{f(y)^2}{v(y)^2}d\mu(y)\right)^{\frac{1}{2}}\\ &= v(x)(\int_\Gamma k(x,y)\tfrac{f(y)^2}{v(y)^2}d\mu(y))^{\frac{1}{2}}\end{aligned}$$

Hence:

$$\langle Af, f\rangle \leq \int_\Gamma |f(x)|\left(\int_\Gamma k(x,y)\frac{f(y)^2}{v(y)^2}d\mu(y)\right)^{\frac{1}{2}}d\mu(x)$$

and by using the Cauchy-Schwartz inequality once again:

$$\langle Af, f\rangle \leq ||f||\left(\int_{\Gamma^2}\frac{k(x,y)}{v(y)^2}f(y)^2 d\mu(y)d\mu(x)\right)^{\frac{1}{2}} = ||f||^2$$

by symmetry of the kernel which, in combination with (1.1), also implies the positivity of A.

Plugging in v for f it follows immediately that the eigenvalue 1 is actually obtained and v is an eigenfunction. □

Theorem 1.1.2 (Spectral Decomposition of the Diffusion Kernel)**.** *Assuming A is compact*[3] *and $A\phi_l = \lambda_l\phi_l$ we may write the kernel as*

$$a(x,y) = \sum_{l\geq 0}\lambda_l\phi_l(x)\phi_l(y)$$

with $\lambda_0 = 1$ and $\lim_{l\to\infty}\lambda_l = 0$ monotonically.

try to generalize lemma 3.4 of Teschl (2014) (guess it won't work though)

[3] which is no constraint in practice since data is finite

Proof. First, note that A being compact implies that the spectrum is discrete and the sum thus is well defined. By A being symmetric and compact the spectral theorem applies and we get that there exists a sequence of real eigenvalues λ_l converging to 0. The corresponding normalized eigenvectors ϕ_l form an orthonormal set and every $f \in L^2(\Gamma, d\mu)$ can be written as

$$f = \sum_{l \geq 0} \langle \phi_l, f \rangle \phi_l + h$$

where $h \in Ker(A)$.

It follows that

$$Af(x) = \int_\Gamma a(x,y) f(y) d\mu(y) = \sum_{l \geq 0} \lambda_l \int_\Gamma \phi_l(y) f(y) d\mu(y) \ \phi_l(x)$$

which, by linearity of the integral and comparison of components, is just what we were looking for. □

"Komponentenvergleich" auf Englisch finden

From definition of $a(x,y)$ we see that

$$p(x,y) = \sum_{l \geq 0} \lambda_l \underbrace{\frac{\phi_l(x)}{v(x)}}_{=:\psi_l(x)} \phi_l(y) v(y) \tag{1.2}$$

which enables us to efficiently compute tth powers p_t of p.

There are two ways to interpret p_t:

1. p_t has a probabilistic interpretation as the probability for a Markov chain with transition matrix P to reach y from x in t steps.

2. the dual point of view is that of the functions defined on the data. The kernel p_t can be viewed as a bump or more precisely, if $x \in \Gamma$ is fixed, then $p_t(x, \cdot\,)$ is a bump function centered at x and of width increasing with t which intuitively captures the idea of diffusion.

1.1.4 Embedding in the Euclidian Space

Definition 1.4. *Let*

$$D_t(x,y)^2 = ||p_t(x, \cdot\,) - p_t(y, \cdot\,)||^2_{L^2(\Gamma, d\mu/v)} = \int_\Gamma \big(p_t(x,u) - p_t(y,u)\big)^2 \frac{d\mu(u)}{v(u)}$$

be the family of diffusion distances parameterized by t.

For a fixed value of t D_t defines a distance on the set Γ which is small only if there is a large number of small paths connecting x and y (i.e. if there is a large probability of getting from x to y in t steps). It thus emphasizes the notion of a cluster.

Another property following from the summation over all possible paths is that this distance is very robust to noise perturbation (in contrast to the geodesic distance).

Theorem 1.1.3 (A Numerically Feasible Representation)**.**

$$D_t(x,y) = \left(\sum_{l \geq 0} \left(\lambda_l^t (\psi_l(x) - \psi_l(y)) \right)^2 \right)^{\frac{1}{2}}$$

Proof. $\{\phi_l\}_{l \geq 0}$ forming an orthonormal basis for $L^2(\Gamma, d\mu)$ implies that $\{\phi_l v\}_{l \geq 0}$ is an orthonormal basis for $L^2(\Gamma, d\mu/v)$ and thus for x fixed (1.2) may be seen as orthogonal expansion of the function $y \mapsto p_t(x,y)$ into the basis $\{\phi_l v\}_{l \geq 0}$. The coefficients are given by $\{\lambda_l^t \psi_l(x)\}_{l \geq 0}$. The statement follows directly using the Pythagorean theorem. □

An imidiate consequence is that the diffusion distance is well approximable and that it converges towards a function of (numerical) rank 1 as $t \to \infty$ because of the vanishing influence of all eigenvectors with eigenvalues < 1.

One possible interpretation is that $D_t(x,y)$ measures the distance between bumps of "magnitude" t being centered around two points x and y. As t gets larger so does the size of the supports and the number of eigenfunctions needed to calculate $D_t(x,y)$ decreases. This number is related to the minimum number of bumps necessary to cover the set X (like in Weyl's asymptotic law for the decay of the spectrum).

In order to calculate $D_t(x,y)$ to a preset accuracy $\delta > 0$ with a finite number of terms we set

$$s_t(\delta) = \max\{l \in \mathbf{N} : \lambda_l^t \ > \delta \lambda_1^t\}$$

so that, up to relative precision δ

$$D_t(x,y) = \left(\sum_{l=0}^{s_t(\delta)} \left(\lambda_l^t (\psi_l(x) - \psi_l(y)) \right)^2 \right)^{\frac{1}{2}}. \tag{1.3}$$

Definition 1.5. *Let* $\{\Psi_t\}_{t\in\mathbf{N}}$,

$$\Psi_t(x) = \begin{pmatrix} \lambda_1^t\psi_1(x) \\ \lambda_2^t\psi_2(x) \\ \vdots \\ \lambda_{s_t(\delta)}^t\psi_{s_t(\delta)}(x) \end{pmatrix}$$

be the family of diffusion maps. Each component of $\Psi_t(x)$ *is termed diffusion coordinate.*

According to (1.3) diffusion maps embed data in a Euclidian space in such a way that the Euclidian distance equals the diffusion distance up to a relative error δ.

Chapter 2

Connection with the Fokker-Planck Equation

2.1 A Family of Anisotropic Diffusion Maps

define FPE, infinitesimal generators, SDEs, etc.

In this section we investigate the probability space $(\Omega, \mathcal{A}, \mu)$ and therefore assume an infinite number of data points whose distribution is determined by the probability measure $\mu(x) = e^{-U(x)}$. U may be interpreted as a potential function. A motivation for the choice of μ will be given with (2.8).

2.1.1 Construction of the Family of Diffusions

Start with a Gaussian kernel $k_\varepsilon(x, y) = e^{-\frac{||x-y||^2}{\varepsilon}}$ and let $\alpha > 0$ being a parameter indexing this family of diffusions.

We can estimate the local density q_ε by

Definition 2.1.

$$q_\varepsilon(x) = \int_\Omega k_\varepsilon(x, y) d\mu(y)$$

Now consider the family of (generally) anisotropic kernels

Definition 2.2.

$$k_\varepsilon^{(\alpha)}(x, y) = \frac{k_\varepsilon(x, y)}{q_\varepsilon^\alpha(x) q_\varepsilon^\alpha(y)}$$

We compute the normalization factor[1] $d_\varepsilon^{(\alpha)}$

[1] Note that this factor is the continuous analogue to definition 1.1.

$$d_\varepsilon^{(\alpha)}(x) = \int_\Omega k_\varepsilon^{(\alpha)}(x,y)d\mu(y)$$

and define the forward and symmetric ($p^{(\alpha)}$ and $a^{(\alpha)}$ respectively) transition probability kernels:

Definition 2.3.

$$p_\varepsilon^{(\alpha)}(x|y) = \frac{k_\varepsilon^{(\alpha)}(x,y)}{d_\varepsilon^{(\alpha)}(y)} = Pr\big(x(t+\varepsilon) = x \,|\, x(t) = y\big)$$
$$a_\varepsilon^{(\alpha)}(x|y) = \frac{k_\varepsilon^{(\alpha)}(x,y)}{\sqrt{d_\varepsilon^{(\alpha)}(x)d_\varepsilon^{(\alpha)}(y)}}$$

Now we define the forward, backward and symmetric Chapman-Kolmogorov operators on functions defined on this probability space, as follows:

Definition 2.4.

$$T_{f,\varepsilon}^{(\alpha)}[\phi](x) = \int_\Omega p_\varepsilon^{(\alpha)}(x|y)\phi(y)d\mu(y)$$
$$T_{b,\varepsilon}^{(\alpha)}[\phi](x) = \int_\Omega p_\varepsilon^{(\alpha)}(y|x)\phi(y)d\mu(y)$$
$$T_{s,\varepsilon}^{(\alpha)}[\phi](x) = \int_\Omega a_\varepsilon^{(\alpha)}(x,y)\phi(y)d\mu(y)$$

If $\phi(x)$ is the probability of finding the system at location x at time $t = 0$, then $T_{f,\varepsilon}^{(\alpha)}[\phi]$ is the evolution of this probability to time $t = \varepsilon$. Similarly, if $\psi(z)$ is some function on the space, then $T_{b,\varepsilon}^{(\alpha)}[\psi](x)$ is the mean value of that function at time ε for a random walk that started at x, and so $\left(T_{b,\varepsilon}^{(\alpha)}\right)^m[\psi](x)$ is the average value of the function at time $t = m\varepsilon$.

By definition, the operators $T_{f,\varepsilon}^{(\alpha)}$ and $T_{b,\varepsilon}^{(\alpha)}$ are adjoint under the inner product with weight μ, while the operator $T_{s,\varepsilon}^{(\alpha)}$ is self adjoint under this inner product.

Moreover, just like in the discrete case, since $T_{s,\varepsilon}^{(\alpha)}$ is obtained via conjugation of the kernel $p^{(\alpha)}$ all three operators share the same eigenvalues. The corresponding eigenfunctions can be found via conjugation by $\sqrt{d_\varepsilon^{(\alpha)}}$; i.e.: If $T_{s,\varepsilon}^{(\alpha)}\phi_s = \lambda\phi_s$, then the corresponding eigenfunctions for $T_{f,\varepsilon}^{(\alpha)}$ and $T_{b,\varepsilon}^{(\alpha)}$ are $\phi_f = \sqrt{q_\varepsilon}\phi_s$ and $\phi_b = \frac{\phi_s}{\sqrt{q_\varepsilon}}$, respectively.

double check that section

Since $\sqrt{q_\varepsilon}$ is the first eigenfunction with $\lambda_0 = 1$ of $T_s^{(\alpha)}$, the steady state of the forward operator is simply $q_\varepsilon(x)$, while for the backward operator it is the uniform density, $\phi_b \equiv 1$.

Furthermore, note that the eigenvalues and eigenvectors of the discrete Markov chain described in the previous section are discrete approximations to the eigenvalues and eigenfunctions of these continuous operators.

2.1.2 Transition to Diffusion Processes

Interpreting ε as a timestep[2] and recalling the asymptotic expansions derived in the appendix it is instructive to look at the limit $\varepsilon \to 0$. In this case, the transition probability densities of the Markov chain (that is continuous in space, but discrete in time) converge to those of a diffusion process, whose time evolution is described by a differential equation

$$\frac{\partial \phi}{\partial t} = \mathcal{H}_f^{(\alpha)} \phi$$

where $\mathcal{H}_f^{(\alpha)}$ is the infinitesimal generator or propagator of the forward operator, defined as

Definition 2.5.

$$\mathcal{H}_f^{(\alpha)} = \lim_{\varepsilon \to 0} \frac{T_{f,\varepsilon}^{(\alpha)} - I}{\varepsilon}$$

Similarly, the inifinitesimal operator of the backward operator is given by

Definition 2.6.

$$\mathcal{H}_b^{(\alpha)} = \lim_{\varepsilon \to 0} \frac{T_{b,\varepsilon}^{(\alpha)} - I}{\varepsilon}$$

and as shown in the appendix (theorem A.2.2)

$$\mathcal{H}_b^{(\alpha)} \psi = \Delta \psi - 2(1-\alpha) \nabla \psi \cdot \nabla U$$

as well as

$$\mathcal{H}_f^{(\alpha)} \psi = \Delta \psi - 2\alpha \nabla \psi \cdot \nabla U + (2\alpha - 1)\psi(\nabla U \cdot \nabla U - \Delta U)$$

Now we consider the following three choices of α:

[2] While in the case of a finite amount of data, ε must remain finite so as to have enough neighbors in a ball of radius $\mathcal{O}(\varepsilon^{\frac{1}{2}})$ near each point x, as the number of samples goes to infinity, we can take smaller and smaller values of ε.

1. $\alpha = 0$:
$$\mathcal{H}_f^{(0)}\phi = \Delta\phi - (e^U \Delta e^{-U})\phi$$
Chosing $\alpha = 0$, "ignoring" the potential term (in the process of the normalization), we get the normalized graph Laplacian we already dealt with in the first chapter.

2. $\alpha = \frac{1}{2}$:
$$\mathcal{H}_f^{(\frac{1}{2})}\phi = \mathcal{H}_b^{(\frac{1}{2})}\phi = \Delta\phi - \nabla\phi \cdot \nabla U$$
First, note that the infinitesimal operator of the forward and backward operators coincide and reduce to the backward Fokker-Planck equation with potential U.

3. $\alpha = 1$:
$$\mathcal{H}_b^{(1)}\phi = \Delta\phi$$
The intention of chosing $\alpha = 1$ is to "normalize the potential away" which is successful as the resulting infinitesimal generator is the Laplace-Beltrami operator and thus the corresponding diffusion only captures the geometry of the data, totally neglecting the underlying density e^{-U}. As such, it is optimal for tasks like manifold reconstruction.

2.1.3 Investigation of Stochastic Processes

First we start with the definition of a stochastic process:

Definition 2.7. *Let T be an ordered set. A stochastic process is a collection of random variables $X = \{X_t, t \in T\}$ where, for each $t \in T$, $X_t : E \to \Omega$ is a random variable from a measurable space $(E, \mathcal{F})$ to another one $(\Omega, \mathcal{G})$.*

From now on, let $T = \mathbb{R}^+$ be the time and $\Omega = \mathbb{R}^n$ the *state space*.[3]

Now we want to investigate the time evolution of a system that is described by its state $x(t)$ at time t, $(x(t) \in \Omega)$ and is governed by the following stochastic differential equation

$$\dot{x} = -U(x) + \sqrt{2/\beta}\dot{w} \tag{2.1}$$

where $U(x)$ is the potential energy of a state x, β is some thermal factor which is inversely proportional to a given temperature and $w(t)$ is the Brownian motion in n dimensions.

[3]It is important that $x \in \Omega$ does *not* represent a particular particle, but a certain state of the entire system.

Integration of the SDE (2.1) produces random paths whose distribution defines a time dependent probability distributions on Ω. To study the dynamics of the system, it is convenient to consider the time evolution of these probability distributions as it has been shownthat the transition probability density $p(x,t|x_0,0)$ of finding the system at location x at time t, given an initial location x_0 at time $t=0$ satisfies the forward Fokker-Planck equation

find source

$$\frac{\partial p}{\partial t} = \frac{1}{\beta}\Delta p + \nabla\cdot(p\nabla U) =: \mathcal{L}p \tag{2.2}$$

defined in $(x,t)\in\Omega\times\mathbb{R}^+$ with reflecting boundary conditions on $\partial\Omega$, i.e. $\frac{\partial p}{\partial n}=0$ where $\frac{\partial}{\partial n}$ is the derivative in the outward normal direction. Furthermore the initial condition is given by the Dirac-Delta function, i.e. $\lim_{t\to 0} p(x,t|x_0,0)=\delta(x-x_0)$.

Eigenvalues of the Fokker-Planck Operator

As seen in the first chapter one can compute the eigenvalues $\lambda_{\varepsilon,i}^{(\alpha)}$ of $T_{b,\varepsilon}^{(\alpha)}$ by computing those of $T_{s,\varepsilon}^{(\alpha)}$. From definition of $\mathcal{H}_f^{(\alpha)}$ and $\mathcal{H}_b^{(\alpha)}$ it follows that they share the eigenvalues $\mu_i^{(\alpha)}$ which can be written as

$$\mu_i^{(\alpha)} = \lim_{\varepsilon\to 0}\frac{\lambda_{\varepsilon,i}^{(\alpha)}-1}{\varepsilon}$$

and are discrete again. Remember that for $\alpha=\frac{1}{2}$, $\mathcal{H}_f^{(\alpha)}$ and $\mathcal{H}_b^{(\alpha)}$ coincide and become the backward Fokker-Planck operator $\mathcal{L}^*$

Definition 2.8. $\mathcal{L}^*g = \frac{1}{\beta}\Delta g - \nabla g\cdot\nabla U$

Lemma 2.1.1. *$\mathcal{L}$ and $\mathcal{L}^*$ are adjoint under the standard inner product.*

Proof.

$$\begin{aligned}
\langle\mathcal{L}^*f, g\rangle &= \langle\frac{1}{\beta}\Delta f - \nabla g\cdot\nabla U, g\rangle \\
&= \langle\frac{1}{\beta}\Delta f, g\rangle + \langle -\nabla f\cdot\nabla U, g\rangle \\
&= \langle f, \frac{1}{\beta}\Delta g\rangle + \langle -\nabla f, g\cdot\nabla U\rangle \\
&= \langle f, \frac{1}{\beta}\Delta g\rangle + \langle f, \mathrm{div}(g\cdot\nabla U)\rangle \\
&= \langle f, \frac{1}{\beta}\Delta g\rangle + \langle f, \nabla\cdot(g\cdot\nabla U)\rangle = \langle f, \mathcal{L}g\rangle
\end{aligned}$$

□

$\mathcal{L}$ and $\mathcal{L}^*$ being adjoint implies that $\mu_i^{(\frac{1}{2})}$ are not only the eigenvalues of $\mathcal{L}$, but also of $\mathcal{L}^*$.

Computing the Transition Probability Density

For simplicity's sake the calculations will be done in one dimension for the forward Fokker-Planck operator.

First of all, notice that in one dimension (2.2) may be written as

$$\begin{aligned}\frac{\partial p}{\partial t} &= \frac{\partial^2 p}{\partial x^2}\frac{1}{\beta} + \frac{\partial}{\partial x}\left(p\frac{\partial U}{\partial x}\right) \\ &= \frac{\partial^2 p}{\partial x^2}\underbrace{\frac{1}{\beta}}_{=:f(x)} + \frac{\partial p}{\partial x}\underbrace{\frac{\partial U}{\partial x}}_{=:g(x)} + p\underbrace{\frac{\partial^2 U}{\partial x^2}}_{=:h(x)}\end{aligned} \tag{2.3}$$

and the reflecting boundary conditions simplify to $p(a,t) = p(b,t) = 0$.

Now apply separation of variables $p(x,t) = X(x)T(t)$ and rewrite (2.3) as

$$\frac{\mathcal{L}X(x)}{X(x)} = \frac{\frac{d}{dt}T(t)}{T(t)}. \tag{2.4}$$

Since both sides of (2.4) only depend on one variable each, they must be equal to a constant, i.e.

$$\mathcal{L}X(x) = \lambda X(x) \tag{2.5}$$

and

$$\frac{d}{dt}T(t) = \lambda T(t).$$

(2.5) is *almost* a regular Sturm-Liouville problem. In order get it into the form

$$\frac{d}{dx}\left[\tilde{f}(x)\frac{d\varphi_n}{dx}\right] + \tilde{h}(x)\varphi_n = \tilde{f}(x)\frac{d^2\varphi_n}{dx^2} + \frac{d\tilde{f}(x)}{dx}\frac{d\varphi_n}{dx} + \tilde{h}(x)\varphi_n = -\lambda_n w(x)\varphi_n$$

we consider the space $L^2([a,b],w)$ with

$$w = \frac{e^{\int \frac{g}{f}}}{f} = \beta e^{\beta U} \tag{2.6}$$

and

$$\langle \phi, \psi \rangle_w = \int_a^b \phi(x)\psi(x)w(x)dx.$$

(2.6) is not only the eigenfunction to the first eigenvalue $\lambda_0 = 0$ (plug it into (2.3)), but, if normalized to absolute value 1, is also known as the Boltzmann equilibrium distribution.

We also get $X(a) = X(b) = 0$. Let φ_n and λ_n be the eigenfunctions[4] and eigenvalues of $\mathcal{L}$ and analogously $\frac{d}{dt}T_n(t) = \lambda_n T(t)$. Note that, as a Sturm-Liouville operator in $L^2([a,b], w(x)dx)$, $\mathcal{L}$ is self-adjoint and its eigenfunctions φ_n are orthogonal and maximal. Without loss of generality assume φ_n to be orthonormal.

find something to cite that this holds for SL problems with appropriate boundary conditions

Now we can write $T_n(t)$ as

$$T_n(t) = a_n e^{-\lambda_n t} \tag{2.7}$$

for some constants a_n that can be written as

$$a_n = \langle \varphi_n, p(\cdot, 0) \rangle_w$$

Note that a_n is a generalized fourier series of the initial distribution $p(\cdot, 0)$, i.e. $p(\cdot, 0) = \sum_k a_k \varphi_n$ on $[a,b]$ because the set of orthonormal eigenfunctions forms a basis. It is chosen such that (2.7) holds true for $t = 0$. As a result

not in this case(?) since $p(x,0) = \delta(x_0 - x)$ and the Dirac-Delta function is not a smooth function - any way to rescue this ansatz?

$$p(x,t) = \sum_{n \geq 0} a_n \varphi_n(x) e^{-\lambda_n t} \tag{2.8}$$

From that representation we get that, as $t \to \infty$, only the first term with $\lambda_0 = 0$ survives which justifies the naming of (2.6) as Boltzmann *equilibrium* distribution.[5].

Analogously, for every smooth $f \in L^2(\Omega)$, one can consider $g(x,t) = \mathbb{E}\{f(x(t)) | x(0)) = x\}$ which satisfies the backward Fokker-Planck equation

$$\frac{\partial g}{\partial t} = \frac{1}{\beta}\Delta g - \nabla g \cdot \nabla U = \mathcal{L}^* g$$

in the domain $(x,t) \in \Omega \times \mathbb{R}^+$, with initial conditions

[4] There is no general analytic (exact) solution to Sturm-Liouville problems, but solving a Sturm-Liouville problem is a easier than solving the whole PDE.

[5] Under the assumption that the system is ergodic, the ergodic theorem tells us that the time average and the space average are the same almost everywhere; as such (and because φ_0 is continuous), regardless of the initial configuration $x_0 \in \Omega$, $\lim_{t\to\infty} p(x,t|x_0,0) = \varphi_0(x)$

$$g(x,0) = f(x)$$

Remember that, as the adjoint[6] of $\mathcal{L}$, $\mathcal{L}^*$ shares the eigenvalues λ_n. It is easy to see that $\psi_0(x) = 1$ is the eigenfunction corresponding to the eigenvalue $\lambda_0 = 0$.

[6] under the standard inner product

Appendix A

Asymptotics for Laplacian Operators

In the following we will deal with a compact manifold $\mathcal{M}$ that is C^∞. Let x be a fixed point, not on the on the boundary, $T_x\mathcal{M}$ be the tangent space to $\mathcal{M}$ at x and $(e_1, \ldots, e_d)$ be a fixed orthonormal basis of $T_x\mathcal{M}$. Furthermore two systems of local coordinates will be introduced:

write introduction + motivation (latter being the justification for naming it graph "Laplacian" earlier)

1. *(Normal coordinates)* The exponential map exp_x generates a set of orthogonal geodesics $(\gamma_1, \ldots, \gamma_d)$ intersecting at x with initial velocity $(e_1, \ldots, e_d)$. Now every point $y \in \mathcal{M}$ in a sufficiently small neighborhood of x has a set of *normal coordinates* $(s_1, \ldots, s_d)$ along these geodesics.

2. *(Tangent coordinates)* Considering the orthogonal projection u of y on $T_x\mathcal{M}$, where $u_i = \langle y - x, e_i \rangle$ in $(e_1, \ldots, e_d)$, we get a system of *tangent coordinates*. The submanifold is now locally parameterized as $y = (u, g(u))$, where $g : \mathbb{R}^d \to \mathbb{R}^{n-d}$. Since $u = (u_1, \ldots, u_d)$ are tangent coordinates, we must have that $\partial g(0) = 0$.

Notice that, locally, any function f on $\mathcal{M}$ may be viewed as $\tilde{f}$ of $(s_1, \ldots, s_d)$ and thus we may write $\Delta f(x) = -\sum_{i=1}^{d} \frac{\partial^2 \tilde{f}}{\partial s_i^2}(0, \ldots, 0)$, where Δ is the Laplace-Beltrami operator on $\mathcal{M}$.

A.1 Comparison of the Geodesic and the Local Projection

In this section we will compute asymptotic expansions for the changes of

The following is basically a citation of Coifman and Lafon (2006).

variable $u \mapsto (s_1, \ldots, s_d)$ and $u \mapsto y$.

In the following, $Q_{x,m}(u)$ denotes a generic homogeneous polynomial of degree m of the variable $u = (u_1, \ldots, u_d)$, whose coefficient depends on x.

Lemma A.1.1. *If $y \in \mathcal{M}$ is in a Euclidean ball of radius $\varepsilon^{\frac{1}{2}}$ around x, then, for ε sufficiently small, there exists:*

$$s_i = u_i + Q_{x,3}(u) + \mathcal{O}(\varepsilon^2) \tag{A.1}$$

Proof. Let γ be the geodesic connection x and y parameterized by arclength. We have $\gamma(0) = x$ and let s be such that $\gamma(s) = y$. If y has normal coordinates $(s_1, \ldots, s_d)$, then we have $s\gamma'(0) = (s_1, \ldots, s_d)$. A Taylor expansion yields

$$\gamma(s) = \gamma(0) + s\gamma'(0) + \frac{s^2}{2}\gamma''(0) + \frac{s^3}{6}\gamma^{(3)}(0) + \mathcal{O}(\varepsilon^2).$$

By definition of a geodesic, the covariant derivative of the velocity is zero, which means that $\gamma''(0)$ is orthogonal to the tangent plane at x. Now since the parameter u_i is defined by $u_i = \langle \gamma(s) - \gamma(0), e_i \rangle$, we obtain that $u_i = s_i + \frac{s^3}{6}\langle \gamma^{(3)}(0), e_i \rangle + \mathcal{O}(\varepsilon^2)$. Iterating this equation yields the result. □

Lemma A.1.2. *Again, let $y \in \mathcal{M}$ be in a Euclidean ball of radius $\varepsilon^{\frac{1}{2}}$ around x; we have*

$$||x - y||^2 = ||u||^2 + Q_{x,4}(u) + Q_{x,4}(u) + \mathcal{O}(\varepsilon^3) \tag{A.2}$$

and

$$det\left(\frac{dy}{du}\right) = 1 + Q_{x,2}(u) + Q_{x,3}(u) + \mathcal{O}(\varepsilon^2). \tag{A.3}$$

Proof. The submanifold is locally parameterized as $u \mapsto (u, g(u))$, where $g : \mathbb{R}^d \to \mathbb{R}^{n-d}$. Writing $g = (g_{i+1}, \ldots, g_n)$ and applying Pythagore's theorem, we obtain

$$||x - y||^2 = ||u||^2 + \sum_{i=d+1}^{n} g_i(u)^2.$$

Using that, by definition, $g_i(0) = 0$ and, as noted before, $\frac{\partial g}{\partial u_i}(0) = 0$. As a consequence $g_i(u) = b_{i,x}(u) + c_{i,x}(u) + \mathcal{O}(\varepsilon^2)$, where $b_{i,x}$ is the Hessian quadratic form of g_i at $u = 0$ and $c_{i,x}$ is the cubic term. This proves (A.2) with

$$Q_{x,4}(u) = \sum_{i=d+1}^{n} b_{i,x}^2(u) \text{ and } Q_{x,5}(u) = 2\sum_{i=d+1}^{n} b_{i,x}(u)c_{i,x}(u).$$

To prove (A.3), observe that $\frac{\partial g}{\partial u_i}(0) = 0$ implies that $\frac{\partial g}{\partial u_i}(0) = \tilde{b}_{i,x}(u) + \tilde{c}_{i,x}(u) + \mathcal{O}(\varepsilon^{\frac{3}{2}})$, where $\tilde{b}_{i,x}(u)$ and $\tilde{c}_{i,x}(u)$ are the linear and quadratic terms in the Taylor expansion of $\frac{\partial g}{\partial u_i}(0)$ at 0. We thus have:

$$\begin{aligned}\frac{\partial y}{\partial u_i}(u) &= \left(v_i, \frac{\partial g}{\partial u_i}(u)\right), \text{ where } v_i = (0, \ldots, 0, 1, 0, \ldots, 0) \in \mathbb{R}^d \\ &= (v_i, \tilde{b}_{i,x}(u) + \tilde{c}_{i,x}(u) + \mathcal{O}(\varepsilon^{\frac{3}{2}}).\end{aligned}$$

The squared volume generated by these d vectors is the determinant of their Gram matrix, i.e.,

$$\left|det\left(\frac{dy}{du}\right)\right|^2 = \sum_{i,j=1}^{d} E_{ij}(u) + \sum_{i,j=1}^{d} F_{ij}(u) + \mathcal{O}(\varepsilon^2),$$

where

$$E_{ij}(u) = \langle \tilde{b}_{i,x}(u), \tilde{b}_{j,x}(u) \rangle \text{ and } F_{ij}(u) = \langle \tilde{b}_{i,x}(u), \tilde{c}_{j,x}(u) \rangle + \langle \tilde{c}_{i,x}(u), \tilde{b}_{j,x}(u) \rangle.$$

Defining

$$Q_{x,2}(u) = \sum_{i,j=1}^{d} E_{ij}(u) \text{ and } Q_{x,3}(u) = \sum_{i,j=1}^{d} F_{ij}(u),$$

we obtain the last result. □

A.2 Infinitesimal Operators for a Family of Graph Laplacians

In this section we present the calculation of the infinitesimal generators for the different diffusion maps characterized by a parameter α.

To start with we first show an asymptotic expansion for diffusion operators G_ε.

Let $k_\varepsilon(x, y)$ be an isotropic kernel, i.e.:

$$k_\varepsilon(x, y) = h\left(\frac{||x - y||^2}{\varepsilon}\right),$$

where h is assumed to have an exponential decay and let G_ε be the corresponding operator

$$G_\varepsilon f(x) = \frac{1}{\varepsilon^{\frac{d}{2}}} \int_{\mathcal{M}} k_\varepsilon(x,y) f(y) dy.$$

The idea is that, using the previous lemmata, for small ε integrating f against the kernel on the manifold is approximately like integrating on the tangent space.

Theorem A.2.1. *Let $f \in C^3(\mathcal{M})$ and let $0 < \gamma < 1/2$. Then we have, uniformly for all $x \in \mathcal{M}$ at distance larger than ε^γ from $\partial\mathcal{M}$,*

$$G_\varepsilon f(x) = m_0 f(x) + \varepsilon \frac{m_2}{2} \left(\omega(x) f(x) - \Delta f(x)\right) + \mathcal{O}(\varepsilon^2),$$

where

$$m_0 = \int_{\mathbb{R}^d} h(||u||^2) du \text{ and } m_2 = \int_{\mathbb{R}^d} u_1^2 h(||u||^2) du$$

and ω is a potential term depending on the embedding of $\mathcal{M}$.

Proof. Because of the exponential decay of h, the domain of integration can be restricted to the intersection of $\mathcal{M}$ with the ball of radius ε^γ around x. In doing so we generate an error of order

$$\begin{aligned}
\left| \frac{1}{\varepsilon^{\frac{d}{2}}} \int\limits_{\substack{y \in \mathcal{M} \\ ||x-y|| > \varepsilon^\gamma}} h\left(\frac{||x-y||^2}{\varepsilon}\right) f(y) dy \right| &\leq ||f||_\infty \frac{1}{\varepsilon^{\frac{d}{2}}} \int\limits_{\substack{y \in \mathcal{M} \\ ||x-y|| > \varepsilon^\gamma}} \left| h\left(\frac{||x-y||^2}{\varepsilon}\right) \right| dy \\
&\leq ||f||_\infty \int\limits_{\substack{y \in \mathcal{M} \\ ||y|| > \varepsilon^{\gamma - 1/2}}} |h(||y||^2)| dy \\
&\leq C ||f||_\infty Q(\varepsilon^{1/2-\gamma}) e^{-\varepsilon^{\gamma-1/2}},
\end{aligned}$$

where we have used the exponential decay of the kernel and where Q is a polynomial. Since $0 < \gamma < 1/2$, this term is exponentially small and is bounded by $\mathcal{O}(\varepsilon^{\frac{3}{2}})$. Therefore,

$$G_\varepsilon f(x) = \frac{1}{\varepsilon^{\frac{d}{2}}} \int\limits_{\substack{y \in \mathcal{M} \\ ||x-y|| < \varepsilon^\gamma}} h\left(\frac{||x-y||^2}{\varepsilon}\right) f(y) dy + \mathcal{O}(\varepsilon^{\frac{3}{2}}).$$

Now that things are localized around x, we can Taylor-expand the function $(s_1, \dots, s_d) \mapsto f(y(s_1, \dots, s_d))$:

$$f(y) = f(x) + \sum_{i=1}^{d} s_i \frac{\partial \tilde{f}}{\partial s_i}(0) + \frac{1}{2} \sum_{i,j=1}^{d} s_i s_j \frac{\partial^2 \tilde{f}}{\partial s_i \partial s_j}(0) + Q_{x,3}(s_1, \dots, s_d) + \mathcal{O}(\varepsilon^2),$$

where $\tilde{f}(s_1, \dots, s_d) = f(y(s_1, \dots, s_d))$. Invoking (A.1), we obtain

$$f(y) = \tilde{f}(0) + \sum_{i=1}^{d} u_i \frac{\partial \tilde{f}}{\partial s_i}(0) + \frac{1}{2} \sum_{i,j=1}^{d} u_i u_j \frac{\partial^2 \tilde{f}}{\partial s_i \partial s_j}(0) + Q_{x,3}(u) + \mathcal{O}(\varepsilon^2).$$

Likewise, because of (A.2), the Taylor expansion of the kernel is

$$h\left(\frac{||x-y||^2}{\varepsilon}\right) = h\left(\frac{||u||^2}{\varepsilon}\right) + \left(\frac{Q_{x,4}(u) + Q_{x,5}(u)}{\varepsilon}\right) h'\left(\frac{||u||^2}{\varepsilon}\right) + \mathcal{O}(\varepsilon^2).$$

Using (A.3) to change the variable $s \mapsto u$ in the previous integral defining $G_\varepsilon f(x)$ yields:

$$\begin{aligned}
\varepsilon^{\frac{d}{2}} G_\varepsilon f(x) = & \int_{||u|| < \varepsilon^\gamma} \left[h\left(\frac{||u||^2}{\varepsilon}\right) + \left(\frac{Q_{x,4}(u) + Q_{x,5}(u)}{\varepsilon}\right) h'\left(\frac{||u||^2}{\varepsilon}\right) \right] \\
& \times \left[\tilde{f}(0) + \sum_{i=1}^{d} u_i \frac{\partial \tilde{f}}{\partial s_i}(0) + \frac{1}{2} \sum_{i,j=1}^{d} u_i u_j \frac{\partial^2 \tilde{f}}{\partial s_i \partial s_j}(0) + Q_{x,3}(u) \right] \\
& \times \left(1 + Q_{x,2}(u) + Q_{x,3}(u)\right) du + \mathcal{O}\left(\varepsilon^{\frac{d}{2}+2}\right).
\end{aligned}$$

This identity can be dramatically simplified by identifying odd functions and setting their integral to zero. One is left with

$$\begin{aligned}
\varepsilon^{\frac{d}{2}} G_\varepsilon f(x) = & \tilde{f}(0) \int_{\mathbb{R}^d} h\left(\frac{||u||^2}{\varepsilon}\right) du + \frac{1}{2} \left(\sum_{i=1}^{d} \frac{\partial^2 \tilde{f}}{\partial s_i^2}(0) \right) \int_{\mathbb{R}^d} u_1^2 h\left(\frac{||u||^2}{\varepsilon}\right) du \\
& + \tilde{f}(0) \int_{\mathbb{R}^d} \left[\frac{Q_{x,4}(u)}{\varepsilon} h'\left(\frac{||u||^2}{\varepsilon}\right) + \tilde{Q}_{x,2}(u) h\left(\frac{||u||^2}{\varepsilon}\right) \right] du + \mathcal{O}\left(\varepsilon^{\frac{d}{2}+2}\right),
\end{aligned}$$

where the domain of integration has been extended to $\mathbb{R}^d$ (exponential decay of h). Changing the variable according to $u \mapsto \varepsilon^{\frac{1}{2}} u$,

$$\begin{aligned} G_\varepsilon f(x) =& \tilde{f}(0) \int_{\mathbb{R}^d} h\left(||u||^2\right) du + \frac{\varepsilon}{2} \left(\sum_{i=1}^{d} \frac{\partial^2 \tilde{f}}{\partial s_i^2}(0) \right) \int_{\mathbb{R}^d} u_1^2 h\left(||u||^2\right) du \\ &+ \varepsilon \tilde{f}(0) \int_{\mathbb{R}^d} \left(Q_{x,4}(u) h'\left(||u||^2\right) + Q_{x,2}(u) h\left(||u||^2\right) \right) du + \mathcal{O}(\varepsilon^2), \end{aligned}$$

where we have used the homogeneity of $Q_{x,2}$ and $Q_{x,4}$. Finally, observing that

$$\tilde{f}(0) = f(x) \text{ and } \sum_{i=1}^{d} \frac{\partial^2 \tilde{f}}{\partial s_i^2}(0) = -\Delta f(x),$$

we end up with

$$G_\varepsilon f(x) = m_0 f(x) + \varepsilon \frac{m_2}{2} \left(\omega(x) f(x) - \Delta f(x) \right) + \mathcal{O}(\varepsilon^2),$$

where

$$\omega(x) = \frac{2}{m_2} \int_{\mathbb{R}^d} \left(Q_{x,4}(u) h'\left(||u||^2\right) + Q_{x,2}(u) h\left(||u||^2\right) \right) du.$$

Finally, the uniformity follows from the compactness and smoothness of $\mathcal{M}$. □

One can also show the same result holds for all $x \in \mathcal{M}$ at distance smaller than ε^γ from $\partial\mathcal{M}$. For this proof we refer to Coifman and Lafon (2006) as this would go beyond the scope of the discussion.

Suppose that the data set X consists of a Riemannian manifold with a density $p(x) = e^{-U(x)}$ and let $k_\varepsilon(x, y)$ be a Gaussian kernel (which clearly satisfies the requirements for A.2.1.

Theorem A.2.2. *The infinitesimal generator $\mathcal{H}_b\phi$ of the backward operator $T_{b,\varepsilon}^{(\alpha)}\phi = \int_\Gamma \frac{k_\varepsilon^{(\alpha)}(x,y)}{d_\varepsilon^{(\alpha)}(x)} \phi(y) p(y) dy$ is $\Delta\phi - 2(1-\alpha)\nabla\phi \cdot \nabla U$.*

Proof. From A.2.1 we see that[1]

$$p_\varepsilon(x) = p(x) + \varepsilon(\Delta p(x) + \omega(x) p(x)) + \mathcal{O}(\varepsilon^2)$$

[1] w.l.o.g. we assume $m_0 = 1$ and $m_2 = 2$

and consequently,

$$p_\varepsilon^{-\alpha} = p^{-\alpha}\left(1 - \alpha\varepsilon\left(\frac{\Delta p}{p} + \omega\right)\right)(1 + \mathcal{O}(\varepsilon^2)).$$

Let

$$k_\varepsilon^{(\alpha)}(x,y) = \frac{k_\varepsilon(x,y)}{p_\varepsilon^\alpha(x)p_\varepsilon^\alpha(y)}$$

Then, the normalization factor $d_\varepsilon^{(\alpha)}$ is given by

$$d_\varepsilon^{(\alpha)}(x) = \int_\Gamma k_\varepsilon^{(\alpha)}(x,y)p(y)dy = p_\varepsilon^{-\alpha}(x)p_\varepsilon^{1-\alpha}(x)\left[1 + \varepsilon\left((1-\alpha)\omega - \alpha\frac{\Delta p}{p} + \frac{\Delta p^{1-\alpha}}{p^{1-\alpha}(x)}\right)\right].$$

Therefore, the asymptotic expansion of the backward operator gives

$$T_b^{(\alpha)}\phi = \int_\Gamma \frac{k_\varepsilon^{(\alpha)}(x,y)}{d_\varepsilon^{(\alpha)}(x)}\phi(y)p(y)dy = \phi(x) + \varepsilon\left(\frac{\Delta(\phi p^{1-\alpha})}{p^{1-\alpha}} - \phi\frac{\Delta p^{1-\alpha}}{p^{1-\alpha}}\right)$$

and its infinitesimal generator is

$$\mathcal{H}_b\phi = lim_{\varepsilon\to 0}\frac{T_b - I}{\varepsilon}\phi = \frac{\Delta(\phi p^{1-\alpha})}{p^{1-\alpha}} - \phi\frac{\Delta p^{1-\alpha}}{p^{1-\alpha}}.$$

Plugging in $p = e^{-U}$ into the last equation gives the desired result. □

Similarly, one can show that the infinitesimal generator $\mathcal{H}_f\psi^{(\alpha)}$ of the forward operator $T_{f,\varepsilon}^{(\alpha)}\psi$ is $\Delta\psi - 2\alpha\nabla\psi\cdot\nabla U + (2\alpha-1)\psi(\nabla U\cdot\nabla U - \Delta U)$.

Bibliography

Belkin, M. (2003), *Problems of Learning on Manifolds*. Ph.D. thesis, The University of Chicago.

Canzani, Y. (2013), "Analysis on manifolds via the laplacian.", URL `http://www.math.harvard.edu/~canzani/math253/Laplacian.pdf`.

Coifman, Ronald R. and Stéphane Lafon (2006), "Diffusion maps." *Applied and Computational Harmonic Analysis*, 21, 5 – 30, URL `http://www.sciencedirect.com/science/article/pii/S1063520306000546`. Special Issue: Diffusion Maps and Wavelets.

Teschl, G. (2014), *Topics in Real and Functional Analysis*. unpublished, URL `http://www.mat.univie.ac.at/~gerald/ftp/book-fa/fa.pdf`.