Markov Random Walk Representations with Continuous Distributions

Chen-Hsiang Yeang
MIT Artificial Intelligence Laboratory
200 Technology Square, Cambridge, MA 02139

Martin Szummer
Microsoft Research Laboratory
Cambridge, UK

Abstract

We propose a framework to extend Markov random walks (Szummer and Jaakkola, 2001) to a continuum of points. In this framework, the transition probability between two points is the integral of the probability density over all paths connecting the two points. Evaluation of this transition probability is equivalent to solving the diffusion equation with a potential term. The solution is a generalization to the heat kernel (Kondor and Lafferty, 2001; Belkin and Niyogi, 2002). The continuation of discrete random walks allows us to incorporate prior knowledge about the manifold shape and the distribution of data. Experiments on a synthetic dataset suggest that continuous random walks capture the distance metric on a manifold more faithfully than discrete random walks.

1 Introduction

Many machine learning problems require evaluating pairwise relations (similarities or distances) between data points. Global distances often fail to capture the true relations if data are distributed on a non-flat manifold. Therefore, an appropriate way of defining the pairwise distance is by patching together local distances between neighboring points. Two distinct examples are geodesic distances (Tenenbaum, 1998) and Markov random walks (Szummer & Jaakkola, 2001).

Markov random walks are a probabilistic framework of relating data points in a metric space. Transition probabilities are specified according to local distances between neighbors. As the random walk progresses, distant points are mixed up and fine structures of the data are destroyed. However, significant structures at the global scale will remain for a long time. Since local relations are propagated to distant points, Markov random walks incorporate the structure of data distribution in the distance metric.

Despite of their power, Markov random walks are limited by the discreteness of data points. In the original setting, random walks take place only at observed data points. This constraint is not desirable because observed data points are only a finite sample of the underlying distribution. In certain contexts, we would like random walks to occur beyond observed data points. For instance, when clustering a dataset where there is an uncertainty in measurements, the data is represented as a scalar field (an image in 2D) rather than a set of points. The field value indicates the density of the data. In this case, random walks should take place in the continuous region where the densities of data do not vanish. Another example is when we have prior knowledge about the data distribution in a continuous region. We would like the transition probabilities of observed data points to take this prior information into account. Moreover, many kernel-based algorithms require the evaluation of the kernel function on unobserved values. Discrete Markov random walks cannot be applied in these algorithms since they allow transitions between observed data points only.

Those examples highlight the importance of extending Markov random walks to a continuum of points. Matrix operations are no longer applicable since there is an uncountable number of entries in the transition matrix. To overcome this problem, we view random walks from the perspective of paths rather than single transitions. We assign probabilities to paths and estimate conditional probabilities by performing path integrals. The path integral formulation is shown to be equivalent to the diffusion equation. Therefore, conditional probabilities are proportional to solutions of the diffusion equation.

Our approach is related to many previous works. Markov random walks on finite data have been applied in semi-supervised learning (Szummer & Jaakkola, 2001) and clustering (Tishby & Slonin, 2001). Re-

cently, diffusion or heat kernels are applied in kernel-based machine learning algorithms (e.g., Kondor & Lafferty, 2002; Belkin & Niyogi, 2002, Lafferty & Lebannon, 2002). The link between path integrals and diffusion equations (Schrödinger's equations) was discovered during the early era of quantum mechanics and quantum field theory (Feynman, 1965). Several works in machine learning (e.g., Bialek, Callan & Strong, 1996; Nemenman & Bialek, 2000; Horn & Gottlieb, 2001) are also inspired by the mathematical techniques used in quantum theory.

The rest of the paper is organized as follows. Section 2 reviews Markov random walks and formulates its continuous version. Section 3 states the relation between the path integral formulation and the diffusion equation. Section 4 reports and discusses experiment results. Section 5 discusses potential machine learning applications which will benefit from continuous Markov random walks. Section 6 draws the conclusion.

2 Markov random walks in continuum

Suppose there are m data points $\mathcal{B} = \{\mathbf{x}_0, \cdots, \mathbf{x}_{m-1} : \mathbf{x}_i \in R^n, \forall i\}$. We can construct a random walk on $\mathcal{B}$ and compute the conditional probability between each pair of points. Define a random variable x whose states (values) are the points in $\mathcal{B}$. Suppose initially we are certain that x is in state i, i.e., the initial probability $Q_i^0(x) = (0, \cdots, 1, 0, \cdots, 0)^T$ whose i^{th} entry is 1. The transition probability from $\mathbf{x}_i$ to $\mathbf{x}_j$ are inversely related to their Euclidean distance d_{ij}:

$$P_{ji} = \frac{1}{Z} e^{-\beta g(d_{ij})}, \tag{1}$$

where $g(.)$ is an increasing function. A point tends to transition to its neighbors according to equation 1. If $\mathcal{B}$ constitutes vertices of a graph then we set $P_{ji} = P_{ij} = 0$ whenever (i, j) is not an edge of the graph. In our setting we may force the transition probabilities to vanish at non-neighbors in order to avoid distant jumps.

Given the initial state probability $Q_i^0(x)$ and state transition probabilities P_{ji}, it is straightforward to evaluate the state probability at time t:

$$Q_i^t(x) = (P)^t Q_i^0(x). \tag{2}$$

We express $P(.|.)$ as the conditional probability evaluated at time t.

$$P(\mathbf{x}_j, t|\mathbf{x}_i, 0) \equiv P(x(t) = \mathbf{x}_j | x(0) = \mathbf{x}_i) = (Q_i^t(x))_j = (P)_{ji}^t. \tag{3}$$

$P(\mathbf{x}_j, t|\mathbf{x}_i, 0)$ can be viewed as a similarity metric between $\mathbf{x}_i$ and $\mathbf{x}_j$. Different from global metrics such as Euclidean distances or radial basis kernels, it depends on both the locations of end points and the structure of the data manifold. If two points are close but located in a neighborhood where data points are sparse, then the conditional probability is low. The conditional probability also depends on t. If t is long enough and P_{ji} does not vanish, then all the points will mix up at t. The stationary distribution is an eigenvector of P. We want to stop at the stage where significant structures of the data (for example, clusters) are preserved.

We are interested in generalizing the Markov random walk to a continuous limit. In the continuous limit, the state variable can transition to an uncountable number of points, and the time step Δt tends to zero. If the transition probability is isotropic, then the state transition is the Brownian motion, and the conditional probability $P(\mathbf{x}, t|0, 0)$ follows the diffusion (heat) equation:

$$\frac{\partial P}{\partial t} = D\nabla^2 P, \tag{4}$$

where ∇^2 is the Laplacian operator. The solution of equation 4 leads to the heat kernel used in (Kondor & Lafferty, 2002) and (Belkin & Niyogi, 2002). However, in all interesting problems diffusions do not carry out isotropically. Transition probabilities depend not only on neighboring relations but also on our information about data densities. For example, we would like the points close to many observed data points to attract more transitions because they are likely in the region of high density. Incorporating these information in the transition matrix (the kernel function) becomes a very difficult task.

To overcome these problems, we shift perspectives from points to paths. In the case of finite points, the probability that a particular path is realized in the random walk can be computed from the transition probabilities of edges along the path. Conceptually, path probabilities (densities) in continuous case can also be evaluated by multiplying the densities of infinitesimal transitions along the path. However, it is more convenient to directly assign path probabilities because paths can incorporate the global information about regions more easily than single transitions.

How do we assign probabilities to individual paths? We first consider an example of discrete random walks. Let $f(t) = [\mathbf{x_0}, \mathbf{x_{f1}}, \cdots, \mathbf{x_1}]$ be a path connecting $\mathbf{x_0}$ at $t = 0$ and $\mathbf{x_1}$ at $t = T$. If we fix both end points and choose $g(d_{ij}) = d_{ij}^2$ in equation 1, then the path probability is

$$P[f(t)] \propto \prod_t exp(-\beta d^2_{x_{ft}, x_{ft+1}}) = exp(-\beta \sum_t d^2_{x_{ft}, x_{ft+1}}). \tag{5}$$

The path probability in equation 5 is completely determined by local relations of points. Very often the path probability also depends on some global information. For instance, we may require that paths traversing a certain region are assigned with high probabilities. As an illustrative example, we assume paths are attracted by two *sources* located at μ_1 and μ_2. These two sources create a mixture of Gaussian density, and the path probability is proportional to the product of point densities along the path:

$$P[f(t)] \propto exp(-\beta \sum_t d^2_{x_{ft}, x_{ft+1}}) \cdot \prod_t G(x_{ft}; \mu_1, \mu_2, \sigma), \tag{6}$$

where $G(x; \mu_1, \mu_2, \sigma)$ is the mixture of Gaussian density function with centers μ_1, μ_2 and variance σ^2.

Extending to a continuous limit, equation 6 becomes

$$P[f(t)] \propto exp(\int_{f(t)} [-\beta(\frac{df(t)}{dt})^2 + \log G(f(t); \mu_1, \mu_2, \sigma)]dt). \tag{7}$$

The path probability depends on two factors. The first term regularizes the smoothness of the path, and the second term incorporates the global information about data distribution. We encode the global information in a *potential function* $V(\mathbf{x})$.

$P[f(t)]$ can be rewritten as

$$P[f(t)] \propto exp(-S[f(t)]), \tag{8}$$

where the loss functional $S[f(t)]$ is

$$S[f(t)] = \int_{f(t)} [\beta(\frac{df(t)}{dt})^2 - V(f(t))]dt. \tag{9}$$

The conditional probability from $(\mathbf{x}_0, t_0)$ to $(\mathbf{x}_1, t_1)$ is the integration of probabilities over all paths connecting these two points:

$$P(\mathbf{x}_1, t_1 | \mathbf{x}_0, t_0) = \int_{\mathbf{x}_0}^{\mathbf{x}_1} \frac{1}{Z} e^{-S[f(t)]} \mathcal{D}[f(t)]. \tag{10}$$

Here Z is a normalization constant, and $\mathcal{D}[f(t)]$ denotes the *path integral* over all functions $f(t)$ which satisfy boundary conditions $f(t_0) = \mathbf{x}_0$ and $f(t_1) = \mathbf{x}_1$. We can view the path integral as summing over an uncountable number of paths.

3 Solving path integral problems

One may argue that equation 8 does not simplify the problem because integrating over all paths is not easier than performing pointwise transition operations. However, evaluating equation 8 turns out to be equivalent to solving a modified diffusion equation.

Theorem 1 Suppose $\mathcal{L} = \beta(\frac{dx}{dt})^2 - V(x,t)$, $P(x_1, t_1; x, t_0) = \int_{x_0}^{x_1} \frac{1}{Z} \exp\{-\int_{t_0}^{t_1} \mathcal{L} dt'\} \mathcal{D}[x(t)]$, and $P(x,t) = \int_{-\infty}^{\infty} P(x,t|x_0,t_0)P(x_0,t_0)dx_0$, then $P(x,t)$ follows the diffusion equation

$$\frac{\partial P(x,t)}{\partial t} = \frac{1}{4\beta} \frac{\partial^2 P(x,t)}{\partial x^2} - V(x,t)P(x,t). \tag{11}$$

An analogous proof of theorem 1 can be found in (Feynman, 1965; Shankar, 1980). A sketch of it is as follows. By discretizing time steps into $t_0, t_0 + \Delta t, \cdots, t_0 + N\Delta t = t_1$, the path integral $\int \mathcal{D}[x(t)]$ can be expressed as the multi-variable integral $\int \prod_i dx^i$, where $x^i = f(t_0 + i\Delta t)$ is the path value at time $t_0 + i\Delta t$. We consider the transition probability $P(x + \Delta x, t + \Delta t; x, t)$ between neighboring points. As $\Delta t \to 0$, we assume the straight line connecting the two points is dominant. The influence of all other paths vanishes due to the quadratic loss functional of the path derivatives. The path integral thus becomes an integral of Gaussian functions. By expressing it in a continuous limit, we obtain equation 11.

Theorem 1 states the global behavior of averaging over all paths can be depicted by the differential equation of local behavior. Interestingly, this theorem also has a profound impact in physics. In classical mechanics, there are two equivalent formulations of describing the trajectory of a particle. One can start with Newton's laws and express the trajectory in terms of the equation of motion:

$$\frac{\partial}{\partial t}(m\frac{\partial \mathbf{x}}{\partial t}) = \mathbf{F}. \tag{12}$$

Alternatively, one can express the trajectory as the minimizer of the *action* of the system, where the action is a loss functional of paths:

$$S[\mathbf{x}(t)] = \int_{\mathbf{x}(t)} \mathcal{L}[\mathbf{x}(t)]\,dt = \int_{\mathbf{x}(t)} (\frac{m}{2}(\frac{\partial \mathbf{x}}{\partial t})^2 - V(\mathbf{x}, t))\,dt. \tag{13}$$

The solution of equation 12 and the minimizer of equation 13 yield the same trajectory $\mathbf{x}(t)$.

In quantum mechanics, a particle is represented as a wave function $\psi(\mathbf{x}, t)$. The evolution of the wave function is characterized by the Schrödinger's equation (Shankar, 1980):

$$i\hbar \frac{\partial \psi(\mathbf{x}, t)}{\partial t} = -\frac{i\hbar}{2m} \nabla^2 \psi(\mathbf{x}, t) + V(\mathbf{x}, t)\psi(\mathbf{x}, t). \tag{14}$$

Similarly, we can define the action of a particle as classical mechanics. Other than the least action path, a particle is allowed to travel along all possible paths connecting two points. The states at two space-time points are connected by a propagator $U(\mathbf{x}_1, t_1; \mathbf{x}_0, t_0)$:

$$\psi(\mathbf{x}, t) = \int U(\mathbf{x}, t; \mathbf{x}', t')\psi(\mathbf{x}', t')\,d\mathbf{x}', \tag{15}$$

and the propagator is determined by all paths connecting $(\mathbf{x}_0, t_0)$ and $(\mathbf{x}_1, t_1)$.

$$U(\mathbf{x}_1, t_1; \mathbf{x}_0, t_0) = \int_{\mathbf{x}_0}^{\mathbf{x}_1} e^{iS[\mathbf{x}(t)]/\hbar} \mathcal{D}(\mathbf{x}(t)). \quad (16)$$

Analogous to classical mechanics, the wave function whose transition follows equation 16 is a solution of Schrödinger's equation 14 (Feynman 1965; Shankar, 1980).

The propagator in equation 16 closely resembles the conditional probabilities in equation 10. The only difference is the integrand is $e^{iS[\mathbf{x}(t)]}$ instead of $e^{-S[\mathbf{x}(t)]}$. This difference is also reflected in equations 11 and 8. In fact, Schrödinger's equation can be viewed as a diffusion equation in the imaginary time axis.

In many machine learning problems, data are represented by many interrelated parameters. Thus although the data are embedded in a high dimensional Euclidean space, they often "live" on a manifold of much lower dimension. Recently, there are an increasing number of works of uncovering the underlying data manifold (e.g., Tenenbaum, 1998; Roweis & Saul, 2000; Belkin & Niyogim 2002) and constructing heat kernels on particular types of manifolds (e.g., Lafferty and Lebanon, 2002; Kondor and Lafferty, 2002). The diffusion equation in theorem 1 needs to be modified when applying it to a non-Euclidean manifold. First, the smoothness penalty should take the curvature of the manifold (represented as the metric tensor) into account. It becomes the square amplitude of the *tangent velocity* along a path on manifold M:

$$|\mathbf{v}(\mathbf{x}, t)|^2 = g_{ij}(\mathbf{x})(\frac{\mathbf{dx}^i}{dt})(\frac{\mathbf{dx}^j}{dt}), \quad (17)$$

where g_{ij} is the *covariant metric tensor* of M and Einstein's convention of index summation is adopted. Second, if the manifold is bounded, then we may restrict the diffusion on M by creating a barrier potential function:

$$V(\mathbf{x}) = \begin{cases} 0 & \text{when } \mathbf{x} \in M, \\ \infty & \text{when } \mathbf{x} \notin M. \end{cases} \quad (18)$$

Combining 17 and 18, the loss functional of a curve $f(t)$ in R^n then becomes

$$S[f(t)] = \int [\beta g_{ij}(f(t))(\frac{df^i}{dt})(\frac{df^j}{dt}) - V(f(t))]dt. \quad (19)$$

Does the equivalence between the path integral and the diffusion equation still hold on a manifold? Theorem 2 extends the results of theorem 1 to a Riemannian manifold. Therefore, in principle we can compute conditional probabilities by solving diffusion equations on a manifold.

Theorem 2: Suppose $\mathcal{L} = \beta g_{ij}(\mathbf{x})(\frac{d\mathbf{x}^i}{dt})(\frac{d\mathbf{x}^j}{dt}) - V(\mathbf{x})$, where $V(\mathbf{x})$ is defined in equation 18, $P(\mathbf{x}, t; \mathbf{x}', 0) = \int_{\mathbf{x}'}^{\mathbf{x}} \frac{1}{Z} \exp\{-\int_0^t \mathcal{L}dt'\}\mathcal{D}[\mathbf{x}(t)]$, and $P(\mathbf{x}, t) = \int P(\mathbf{x}, t; \mathbf{x}', 0)P(\mathbf{x}', 0)d\mathbf{x}'$, then $P(\mathbf{x}, t)$ follows the diffusion equation

$$\frac{\partial P(\mathbf{x}, t)}{\partial t} = Dg^{ij}(\mathbf{x})\frac{\partial^2 P(\mathbf{x}, t)}{\partial \mathbf{x}_i \partial \mathbf{x}_j}, \quad (20)$$

for $\mathbf{x} \in M$, where $g^{ij}(\mathbf{x})$ is the contravariant metric tensor of M, which is the inverse of $g_{ij}(\mathbf{x})$. $P(\mathbf{x}, t) = 0$ for $\mathbf{x} \notin M$. The proof of theorem 2 is in the appendix.

The solution of equation 20 is the diffusion kernel on a Riemannian manifold. On certain types of manifolds, analytic solutions of diffusion kernels have been found. For instance, Lafferty and Lebanon (Lafferty & Lebannon, 2002) transformed data into parametric models and applied the diffusion kernels on the model manifold. Analytical solutions on multinomial and spherical Gaussian model manifolds were described in their paper.

The shape of the manifold is a special case of the data distribution. We can incorporate the information about data distribution in the potential function. For example, use log mixture of Gaussians to represent the influence of a finite number of sources.

4 Experiments

Suppose we have prior information that the data is distributed continuously as in figure 1. There are two high-density narrow regions embedded in the space which has low data density. For machine learning tasks, we would like to construct a similarity measure that respects the data density. For discrete data, transition probabilities given by Markov random walks can be used for this purpose. Here, we employ transition probabilites given by diffusion.

We define a potential function that is constant and low in the high-density regions ($V(x, t)$=0) and high in the low-density region ($V(x, t)$=12). This allows quick diffusion and high transition probabilities within the high-density region, and low transition probabilites outside and between regions. We use a diffusion constant D=0.1, and run the diffusion for a period in the range $t = [0, 10]$.

The solution of the diffusion equation is calculated via a finite element method that discretizes space into triangular patches (as implemented by the Matlab parabolic command, using 600 nodes).

When started from a point in the lower right region, the diffusion begins by expanding radially (figure 2, t=0.09). It then spreads to occupy the right region (t=1.9), and if run for longer spills over to both regions (t=20). When started from the left (figure 3)

Figure 1: Swiss-roll data

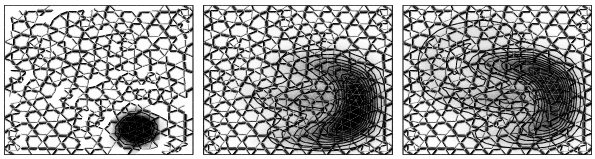

Figure 2: Diffusions on swiss-roll data

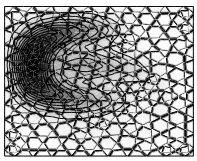

Figure 3: Diffusions on swiss-roll data

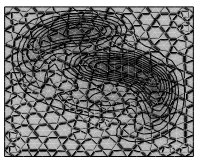

Figure 4: Diffusions off the swiss-roll manifold

region it spreads to cover that part of space (t=14), and when started from the low-density space it reaches both parts (t=14), but is generally more spread out over the whole space. For clustering and classification purposes, the intermediate timescale (t=14) is appropriate for exposing the two contiguous regions[1].

5 Discussion

Continuous Markov random walks are advantageous over discrete Markov random walks in two aspects. First, they are capable of incorporating information from both observed data points and prior beliefs about data distributions. Those information are represented as the potential term in the diffusion equation. We may use the information about continuous regions implicitly in the algorithm. For instance, when performing dimensional reduction on observed data points, we can estimate pairwise distances by performing continuous diffusions. The transition probability between observed data points $\mathbf{x}_i$ and $\mathbf{x}_j$ is the conditional density $P(\mathbf{x}_j, t|\mathbf{x}_i, 0)$ normalized by the sum of conditional densities to all observed data points:

$$Q(\mathbf{x}_j, t|\mathbf{x}_i, 0) = \frac{P(\mathbf{x}_j, t|\mathbf{x}_i, 0)}{\sum_k P(\mathbf{x}_k, t|\mathbf{x}_i, 0)}. \qquad (21)$$

This probability includes the transitions along all paths connecting the two points. In contrast, discrete Markov random walks restrict transitions to occur only among observed points, thus will not be able to incorporate any information beyond these points.

Second, continuous Markov random walks allow us to construct pairwise transition probabilities beyond observed data points. This property makes Markov random walks compatible with all the kernel-based algorithms in machine learning. For example, a support vector machine has the following form:

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}, \mathbf{x_i}). \qquad (22)$$

We can substitute the transition probability density $P(\mathbf{x}, t; \mathbf{x}_i, 0)$ for the kernel function. This kernel function again reflects the information about the structure and the distribution of data on a manifold. Therefore, as suggested in Szummer & Jaakkola, 2001, it will be more suitable than analytic kernels such as radius-based functions or polynomial functions.

To satisfy Mercer's condition the kernel function must to be positive semi-definite. The heat kernel of the diffusion equation without the potential term is positive semi-definite in Euclidean and Riemannian spaces (Itô, 1991). For a diffusion equation with an arbitrary $V(\mathbf{x})$, its kernel is not guaranteed to be positive semi-definite. Nevertheless, if $V(\mathbf{x})$ is non-negative everywhere, then all the eigenvalues of the kernel function are non-negative. The proof of this theorem can be seen in Itô, 1991.

Despite of their advantages, continuous Markov random walks have some unresolved issues. First, it is more costly to perform continuous Markov random walks. To compute the conditional probability density from each point, we need to solve the diffusion equation with an impulse initial condition. Solving differential equations is certainly more involved than matrix multiplications. Second, numerical differential equation solvers, such as finite difference or finite elements methods, discretize the space into small volumes. The number of volumes grow exponentially with the dimension of the embedding space. Since most realistic datasets are embedded in a high dimensional space, the performance of numerical methods may be problematic. One approach to get around this explosion problem is to restrict the support of the density function to a low dimensional subspace. Third, there are

[1]Color figures of diffusion can be accessed on www.ai.mit.edu/people/chyeang/UAI03/

several degrees of freedom in tuning the diffusion equation, including setting the stopping time, adjusting the relative weight between the Laplacian and potential terms, and selecting the potential function. Systematic determination of these free parameters remains to be an open problem.

Continuous Markov random walks are a special case of the Bayesian functional approximation. In the Bayesian functional approximation, we specify the distribution of functions and compute the average quantities – such as the MAP estimation of the function – according to the distribution. Simple distributions such as Gaussian process priors (MacKay, 1997) yield analytical solutions, but other distributions are usually difficult to solve. In this paper, functions (paths) have only one variable (time), and the particular form of the functional distribution transforms the path integral problem into a differential equation. In general, however, such a correspondence does not exist. Advanced techniques in quantum field theory have been applied to solve this general problem, for example, Bialek, 1996.

6 Conclusion

In this paper, we propose a framework of extending Markov random walks to continuous data. Instead of investigating infinitesimal transitions, we assign probability densities to individual paths. The path probability is determined by the smoothness of the path and the global information regarding data distribution on a manifold. The conditional probability between two points is the integration of probability densities along all connecting paths. The conditional probability turns out to be the solution of the diffusion equation with potential terms. We applied this method on a synthetic dataset and demonstrated the diffusion does faithfully reflect the shape of the data manifold. Our method provides a way to extend Markov random walks in incorporating both observed data and prior information about data distribution.

7 Appendix: the proof of theorem 2

The potential value is 0 on M and ∞ outside M, thus all the points x in this proof are on M and the potential function vanishes.

We consider a transition from $(x', 0)$ to point (x, ϵ) in a small time step ϵ. Because ϵ is very small the time step is indivisible. Thus there is only one path connecting $(x', 0)$ to (x, ϵ), and the path is a straight line in the (x, t) space. The action along this path is the integration of $\mathcal{L}$ from $t = 0$ to $t = \epsilon$, which is $\frac{1}{\epsilon}\beta g_{ij}(x)((x - x')^i(x - x')^j)$, where $(x - x')^i$ is the i^{th} component of the difference vector $(x-x')$, assuming x' is close to x. The conditional probability $P(x, \epsilon; x', 0)$ then becomes

$$P(x, \epsilon; x', 0) = \frac{1}{Z} exp\{-\frac{\beta}{\epsilon} g_{ij}(x - x')^i(x - x')^j\}. \quad (23)$$

By assuming x is close to x' ($x' = x + \eta$), there is a single path connecting x' and x and the equation of transition probability becomes

$$P(x, \epsilon) = \frac{1}{Z} \int exp\{-\frac{\beta}{\epsilon} g_{ij}(x)\eta^i\eta^j\} P(x+\eta, 0)\sqrt{|g_{ij}(x)|}d\eta. \quad (24)$$

Here $\sqrt{g_{ij}(x)}d\eta$ is the volume integral $\sqrt{|g_{ij}(x)|}d\eta_1 \cdots d\eta_m$. To make the Gaussian integral easy to evaluate, we apply the transformation

$$\eta = T\eta', \quad (25)$$

which makes η' diagonal. In other words,

$$g_{ij}(x)\eta^i\eta^j = \eta^T G\eta = \eta'^T T^T GT\eta' = \eta'^T \cdot \eta' = \sum_{i=1}^{m} \eta_i'^2. \quad (26)$$

To make this equation hold $T^T GT = I$. Since both T and G are invertible and symmetric, we can swap the order of matrix multiplication,

$$T^T GT = GT^T T = I. \quad (27)$$

Hence $T = G^{-\frac{1}{2}}$. Expand $P(x + \eta, 0)$ to the second order and write it in matrix form,

$$P(x+\eta, 0) = P(x, 0)+\eta^T \cdot \nabla P(x, 0)+\frac{1}{2}\eta^T H(x)\eta, \quad (28)$$

where $H(x)$ is the Hessian of $P(x, 0)$. After coordinate transformation, $P(x + \eta, 0)$ becomes

$$P(x+\eta', 0) = P(x, 0)+\eta'^T T^T \cdot \nabla P(x, 0)+\frac{1}{2}\eta'^T T^T H(x)T\eta'. \quad (29)$$

The volume element $\sqrt{|G|}d\eta = \sqrt{|G|}\prod_{i=1}^{m} d\eta_i$ is transformed to a new volume element

$$d\eta = \nu(G)d\eta'. \quad (30)$$

where $\nu(G)$ is a scalar function of the metric tensor G. Substituting equations 25, 27 and 28 into 22, we get

$$P(x, \epsilon) = P(x, 0) + \epsilon\frac{\partial P(x,t)}{\partial t} \sim \frac{1}{Z}\int exp\{-\frac{\beta}{\epsilon}\eta'^T \cdot \eta'\} (P(x, 0) + \eta'^T T^T \cdot \nabla P(x, 0) + \frac{1}{2}\eta'^T G^{-1} H(x)\eta')\nu(G)d\eta'. \quad (31)$$

$exp\{-\frac{\beta}{\epsilon}\eta'^T \cdot \eta'\}$ is a diagonal Gaussian density function (up to a constant) of dimension m. By setting the normalization constant Z appropriately the first term on the right hand side is $P(x, 0)$. The second term on the right hand side is zero since the first moments

of a zero-mean Gaussian distribution are zeros. Only the diagonal entries of $G^{-1}H(x)$ are of interest since the Gaussian integrals of off-diagonal entries are 0 due to the same reason (and separation of components of η'). For a particular diagonal entry $(G^{-1}H)_{ii}$, the Gaussian integral in the third term is

$$\frac{C}{2Z}\int e^{-\frac{\beta}{\epsilon}(\eta_i'^2)}\eta_i'^2(G^{-1}H)_{ii}\nu(G)d\eta_i' = \frac{C'\epsilon}{2Z}(G^{-1}H(x))_{ii}. \tag{32}$$

The third term is the sum over all diagonal entries,

$$D\epsilon tr(G^{-1}H(x)) = D\epsilon(\sum_{i,j} g^{ij}(x)\frac{\partial^2 P}{\partial x_i \partial x_j}), \tag{33}$$

where $g^{ij}(x)$ is entry of G^{-1}. Equation 21 then reduces to

$$\frac{\partial P(x,t)}{\partial t} = Dg^{ij}(x)\frac{\partial^2 P(x,t)}{\partial x_i \partial x_j}. \tag{34}$$

Q.E.D.

References

M. Belkin and P. Niyogi (2002). *Laplacian eigenmaps for dimensionality reduction and data representation. Technical Report TR-2002-01*, Computer Science Department, University of Chicago.

M. Berstein, V. de Silva, J.C. Langford, and J.B. Tenenbaum (2000). *Graph approximations to geodesics on embedded manifolds. Technical Report*, Carnegie Mellon University.

W. Bialek, C.G. Callan, S.P. Strong (1996). *Field theories for learning probability distributions. Physical Review Letters, 77(23)*, 4693-4697.

R.P. Feynman and A.R. Hibbs (1965). *Quantum Mechanics and Path Integrals.* McGraw-Hill, New York.

G.M. Fung, O.L. Mangasarian, J.W Shavlik (2002). *Knowledge-based support vector machine classifiers. Advances in Neural Information Processing Systems, 14.*

D. Horn and A. Gottlieb (2001). *The method of quantum clustering. Advances in Neural Information Processing Systems, 13.*

S. Itô (1991). *Diffusion Equations.* American Mathematical Society.

R.I. Kondor and J. Lafferty (2002). *Diffusion kernels on graphs and other discrete input spaces. Machine Learning: Proceedings of the Ninteenth International Conference.*

J. Lafferty and G. Lebanon (2002). *Information diffusion kernels. Advances in Neural Information Processing Systems, 14.*

D. MacKay (1997). *Introduction to Gaussian processes.* ICANN97.

I. Nemenman and W. Bialek (2000). *Learning continuous distributions: simulations with field theoretic priors. Advances in Neural Information Processing Systems, 12.*

S. Roweis and L. Saul (2000). *Non-linear dimensionality reduction by locally linear embedding. Science, 290* , 2323-2326.

R. Shankar (1980). *Principles of Quantum Mechanics.* Plenum Press, New York.

M. Szummer and T. Jaakkola (2001). *Partially labeled classification with Markov random walks. Advances in Neural Information Processing Systems, 13.*

J. Tenenbaum (1998). *Mapping a manifold of perceptual observations. Advances in Neural Information Processing Systems, 10.*

N. Tishby and N. Slonin (2001). *Data clustering by Markovian relaxation and the information bottleneck method. Advances in Neural Information Processing Systems, 13.*