Research Article
Some Remarks on Diffusion Distances

Maxim J. Goldberg[1] and Seonja Kim[2]

[1] *Theoretical and Applied Science, Ramapo College of NJ, 505 Ramapo Valley Road, Mahwah, NJ 07430, USA*

[2] *Mathematics Department, SUNY Rockland Community College, 145 College Road, Suffern, NY 10901, USA*

Correspondence should be addressed to Maxim J. Goldberg, mgoldber@ramapo.edu

As a diffusion distance, we propose to use a metric (closely related to cosine similarity) which is defined as the L^2 distance between two L^2-normalized vectors. We provide a mathematical explanation as to why the normalization makes diffusion distances more meaningful. Our proposal is in contrast to that made some years ago by R. Coifman which finds the L^2 distance between certain L^1 unit vectors. In the second part of the paper, we give two proofs that an extension of mean first passage time to mean first passage cost satisfies the triangle inequality; we do not assume that the underlying Markov matrix is diagonalizable. We conclude by exhibiting an interesting connection between the (normalized) mean first passage time and the discretized solution of a certain Dirichlet-Poisson problem and verify our result numerically for the simple case of the unit circle.

1. Introduction

Several years ago, motivated by considering heat flow on a manifold, R. Coifman proposed a diffusion distance—both for the case of a manifold and a discrete analog for a set of data points in $\mathbb{R}^n$. In the continuous case, his distance can be written as the L^2 norm of the difference of two specified vectors, each of which has unit L^1 norm. (An analogous situation holds in the discrete case.) Coifman's distance can be successfully used in various applications, including data organization, approximately isometric embedding of data in low-dimensional Euclidean space, and so forth. See, for example, [1–3]. For a unified discussion of diffusion maps and their usefulness in spectral clustering and dimensionality reduction, see [4].

We see a drawback in Coifman's diffusion distance in that it finds the L^2 norm of the distance between two L^1 unit vectors, rather than L^2 unit vectors. As shown by a simple example later in this paper, two vectors (representing two diffusions), which we may want to

consider to be far apart, are actually close to each other in L^2, even though the angle between them is large, because they have small L^2 norm, while still having unit L^1 norm. Additionally, applying Coifman's distance to heat flow in $\mathbb{R}^n$, a factor of a power of time t remains, with the exponent depending on the dimension n. It would be desirable not to have such a factor.

Our main motivation for this paper is to propose an alternate diffusion metric, which finds the L^2 distance between two L^2 unit vectors (with analogous statements for the discrete case). Our distance is thus the length of the chord joining the tips, on the unit hypersphere, of two L^2 normalized diffusion vectors, and is therefore based on cosine similarity (see (4.4) below). Cosine similarity (affinity) is popular in kernel methods in machine learning; see for example, [5, 6] (in particular, Section 3.5.1—Document Clustering Basics) and for a review of kernel methods in machine learning, [7].

In the case of heat flow on $\mathbb{R}^n$, our proposed distance has the property that no dimensionally dependent factor is left. Furthermore, for a general manifold, our diffusion distance gives, approximately, a scaled geodesic distance between two points x and y, when x and y are closer than $\sqrt{t}$, and maximum separation when the geodesic distance between x and y, scaled by $\sqrt{t}$, goes to infinity.

We next give two proofs that the mean first passage cost—defined later in this paper as the cost to visit a particular point for the first time after leaving a specified point—satisfies the triangle inequality. (See Theorem 4.2 in [8] in which the author states that the triangle inequality holds for the mean first passage time.) We give two proofs that do not assume that the underlying Markov matrix is diagonalizable; our proofs do not rely on spectral theory.

We calculate explicitly the normalized limit of the mean first passage time for the unit circle S^1 by identifying the limit as the solution of a specific Dirichlet-Poisson problem on S^1. We also provide numerical verification of our calculation.

The paper is organized as follows. After a section on notation, we discuss R. Coifman's diffusion distance for both the continuous and discrete cases in Section 3. In Section 4, we define and discuss our alternate diffusion distance. In Section 5, we give two proofs of the triangle inequality for mean first passage cost. We conclude the section by exhibiting an interesting connection between the (normalized) mean first passage time and the discretized solution of a certain Dirichlet-Poisson problem and verify our result numerically for the simple case of S^1.

2. Notation and Setup

In this paper, we will present derivations for both the continuous and discrete cases.

In the continuous situation, we assume there is an underlying Riemannian manifold M with measure $d\,x; x, y, u, z, \ldots$ will denote points in M. For $t \geq 0$, $\rho_t(x, y)$ will denote a kernel on $M \times M$, with $\rho_t(x, y) \geq 0$ for all $x, y \in M$, and satisfying the following semigroup property:

$$\int_M \rho_t(x, u)\rho_s(u, y)du = \rho_{t+s}(x, y), \tag{2.1}$$

for all $x, y \in M$, and $s, t \geq 0$. In addition, we assume the following property:

$$\int_M \rho_t(x, y)dx = 1, \tag{2.2}$$

for all $y \in M$ and all $t \geq 0$. The latter convention gives the mass preservation property

$$\int_M T_t f(x) dx = \int_M f(y) dy, \tag{2.3}$$

where

$$T_t f(x) \equiv \int_M \rho_t(x, y) f(y) dy. \tag{2.4}$$

We will often specialize to the case when $\rho_t(x, y) = \rho_t(y, x)$ for all $x, y \in M$ and $t \geq 0$, as in the case of heat flow. Note that when $\rho_t(x, y)$ is the fundamental solution for heat flow, we have $\rho_0(x, u) = \delta_x(u)$, where $\delta_x(u)$ denotes the Dirac delta function centered at x. We will sometimes assume (as in the case of heat flow on a compact manifold) that there exist $0 \leq \lambda_1 \leq \lambda_2 \leq \cdots$, with each λ_j corresponding to a finite dimensional eigenspace, and a complete orthonormal family of L^2 functions $\phi_1, \phi_2, \ldots$, such that

$$\rho_t(x, y) = \sum_{j=1}^{\infty} e^{-\lambda_j t} \phi_j(x) \phi_j(y), \tag{2.5}$$

for $t > 0$. We will also frequently use the following fact: if ρ_t is symmetric in the space variables, then for any $x, y \in M$,

$$\int_M \rho_t(u, x) \rho_t(u, y) du = \int_M \rho_t(x, u) \rho_t(u, y) du = \rho_{2t}(x, y), \tag{2.6}$$

where we have used the symmetry of ρ_t and its semigroup property.

For the discrete situation, the analog of $\rho_t(x, y)$ is an $N \times N$ matrix A, with $A = \langle a_{ij} \rangle_{i,j=1}^N$, every $a_{ij} \geq 0$. In keeping with the usual convention that A is Markov if each row sum equals 1, that is, $\sum_j a_{ij} = 1$ for all i, the analog of $T_t f(x) = \int_M \rho_t(x, y) f(y) dy$ is $A^T v$, where A^T is the transpose of A, and v is an $N \times 1$ column vector. So the index i corresponds to the second space variable in ρ_t, the index j corresponds to the first space variable in ρ_t, and $t = n$, $n = 1, 2, \ldots$, corresponds to the nth power of A. The obvious analog of ρ_t symmetric in its space variables is a symmetric Markov matrix A, that is, $A = A^T$.

For A as above, not necessarily symmetric, we think of a_{ij} as the probability of transitioning from state s_i to state s_j in $t = 1$ tick of the clock; $S = \{s_1, s_2, \ldots, s_N\}$ is the underlying set of states. For X a subset of the set of states S, the $N \times N$ matrix P_X will denote the following projection: all entries of P_X are 0 except for diagonal entries (k, k), when $s_k \in X$; the latter entries are equal to 1.

Finally, $\mathbf{1}$ will denote the $N \times 1$ column vector where each entry is 1; e_i will denote the $N \times 1$ column vector with the ith component 1, and all others 0, and, for a set of states X, $\overline{X}$ will denote the complement of X with respect to S.

3. A Diffusion Distance Proposed by R. Coifman

Several years ago, R. Coifman proposed a novel diffusion distance based on the ideas of heat flow on a manifold or a discrete analog of heat flow on a set of data points (see, e.g, [1, 2] for a thorough discussion). In this section, we will describe Coifman's distance using our notation, and consider some of its good points, and what we see as some of its drawbacks.

Referring to Section 2, for the continuous case, the unweighted version of Coifman's distance between $x, y \in M$, which we will denote by $d_{C,t}(x, y)$, can be defined as follows:

$$
\begin{aligned}
[d_{C,t}(x, y)]^2 &\equiv \langle T_t(\delta_x - \delta_y), T_t(\delta_x - \delta_y) \rangle \\
&= \langle T_t\delta_x, T_t\delta_x \rangle + \langle T_t\delta_y, T_t\delta_y \rangle - 2\langle T_t\delta_x, T_t\delta_y \rangle.
\end{aligned}
\tag{3.1}
$$

Here,

$$
T_t\delta_z(v) = \int_M \rho_t(v, u)\delta_z(u)\,du = \rho_t(v, z),
\tag{3.2}
$$

for $z \in M$. The $\langle \cdots, \cdots \rangle$ is the usual inner product on $L^2(M)$. (In [1], the authors consider a weighted version of (3.1) which naturally arises when the underlying kernel does not integrate to 1 (in each variable). In terms of data analysis, this corresponds to cases where the data are sampled nonuniformly over the region of interest. For simplicity, we are just using Coifman's unweighted distance.)

Note that we thus have

$$
\begin{aligned}
[d_{C,t}(x, y)]^2 &= \int_M \rho_t(v, x)\rho_t(v, x)\,dv + \int_M \rho_t(v, y)\rho_t(v, y)\,dv - 2\int_M \rho_t(v, x)\rho_t(v, y)\,dv \\
&= \|\rho_t(\cdot, x)\|_2^2 + \|\rho_t(\cdot, y)\|_2^2 - 2\langle \rho_t(\cdot, x), \rho_t(\cdot, y) \rangle.
\end{aligned}
\tag{3.3}
$$

Although Coifman's original definition used a kernel symmetric with respect to the space variable, $d_{C,t}(x, y)$ as given above need not be based on a symmetric ρ_t. Note that, by the defining (3.1), $d_{C,t}(x, y)$ is symmetric in x and y (even if ρ_t is not), and satisfies the triangle inequality. If ρ_t is symmetric in the space variables, from (2.6) we see that:

$$
[d_{C,t}(x, y)]^2 = \rho_{2t}(x, x) + \rho_{2t}(y, y) - 2\rho_{2t}(x, y),
\tag{3.4}
$$

a form matching one of Coifman's formulations for the continuous case.

If, in addition to ρ_t being symmetric in the space variables, we have that (2.5) holds, as in the case of heat flow, we easily see that:

$$
[d_{C,t}(x, y)]^2 = \sum_{j=1}^{\infty} e^{-2\lambda_j t}(\phi_j(x) - \phi_j(y))^2,
\tag{3.5}
$$

the original form proposed by Coifman. Note that the latter expression again explicitly shows that $d_{C,t}(x, y)$ is symmetric in x and y and satisfies the triangle inequality (by considering, for example, the right-hand side as the square of a weighted distance in l^2).

Referring again to Section 2, for the discrete situation, where we start with a set of data points $S = \{s_1, s_2, \ldots, s_N\}$, and A is a Markov matrix specifying the transition probabilities between the "states" of S, the distance between two data points s_i and s_j is given by

$$
\begin{aligned}
[d_{C,1}(s_i, s_j)]^2 &\equiv \left\langle A^T(e_i - e_j), A^T(e_i - e_j) \right\rangle \\
&= \left\langle A^T e_i, A^T e_i \right\rangle + \left\langle A^T e_j, A^T e_j \right\rangle - 2 \left\langle A^T e_i, A^T e_j \right\rangle \\
&= \left\langle AA^T e_i, e_i \right\rangle + \left\langle AA^T e_j, e_j \right\rangle - 2 \left\langle AA^T e_i, e_j \right\rangle \\
&= \left(AA^T \right)_{ii} + \left(AA^T \right)_{jj} - 2 \left(AA^T \right)_{ij},
\end{aligned}
\tag{3.6}
$$

where $\langle \cdots, \cdots \rangle$ is the usual inner product in $\mathbb{R}^N$, and for a matrix B, $(B)_{ij}$ denotes the i, j entry of B. Again, symmetry and the triangle inequality are easily verified. If A is symmetric,

$$
[d_{C,1}(s_i, s_j)]^2 = \left(A^2 \right)_{ii} + \left(A^2 \right)_{jj} - 2 \left(A^2 \right)_{ij}.
\tag{3.7}
$$

The "1" appearing in the subscript of $d_{C,1}(s_i, s_j)$ refers to the fact that $A^1 = A$ is used, corresponding to $t = 1$ in the continuous case. As the diffusion along data points flows, after n ticks of the clock, we can successively consider

$$
[d_{C,n}(s_i, s_j)]^2 = \left(A^n \left(A^T \right)^n \right)_{ii} + \left(A^n \left(A^T \right)^n \right)_{jj} - 2 \left(A^n \left(A^T \right)^n \right)_{ij},
\tag{3.8}
$$

which, for a symmetric A, equals

$$
\left(A^{2n} \right)_{ii} + \left(A^{2n} \right)_{jj} - 2 \left(A^{2n} \right)_{ij}.
\tag{3.9}
$$

An important benefit of introducing a diffusion distance as above can be illustrated by considering (3.5). If ρ_t is such that (3.5) holds for a complete orthonormal family $\{\phi_j\}$, we see that as t increases, we are achieving an (approximate) isometric embedding of M into successively lower-dimensional vector spaces (with a weighted norm). More specifically, for $\lambda_j > 0$, if t is large, the terms $e^{-2\lambda_j t}(\phi_j(x) - \phi_j(y))^2$ are nearly 0. So, as t increases, we see that the "heat smeared" manifold M is parametrized by only a few leading ϕ_j's. Thus, "stepping" through higher and higher times, we are obtaining a natural near-parametrization of more and more smeared versions of M, giving rise to a natural ladder of approximations to M.

Analogous considerations hold in the discrete situation for A symmetric, when we easily see that the eigenvalues of A^2 are between 0 and 1 and decrease exponentially for A^{2n}, as n increases (the "heat smeared" data points are now parametrized by a few leading eigenvectors of A, associated to the largest eigenvalues).

See [1–3] for more discussion and examples of the natural embedding discussed above, along with illustrations of its power to organize unordered data, as well as its insensitivity to noise.

We would now like to point out what we see some drawbacks of Coifman's distance, which led us to propose an alternative distance in Section 4.

Let us consider (3.4) for the case where

$$\rho_t(x, y) = (4\pi t)^{-n/2} e^{-|x-y|^2/4t}, \tag{3.10}$$

the fundamental solution to the heat equation in $\mathbb{R}^n$. Then,

$$[d_{C,t}(x, y)]^2 = \frac{2 - 2e^{-|x-y|^2/8t}}{(8\pi t)^{n/2}}. \tag{3.11}$$

If $|x - y|^2/8t$ is small, then to the leading order in $|x - y|^2/4t$,

$$[d_{C,t}(x, y)]^2 = \frac{1}{(8\pi t)^{n/2}} \left(\frac{|x-y|^2}{4t} + \mathcal{O}\left(\left(\frac{|x-y|^2}{4t} \right)^2 \right) \right). \tag{3.12}$$

Thus, if $|x - y| \ll \sqrt{t}$, we do recover the geodesic distance between x and y but, due to the $1/t^{n/2}$ term in front, normalized by a power of t which depends on the dimension n. As pointed out by the reviewer, for $\mathbb{R}^n$ itself, the normalization does depend on n, but is simply a global change of scale, for each t, and thus basically immaterial. Suppose, however, that the data we are considering come in two "clumps", one of dimension n and the other of dimension m, with $n \neq m$. Let us also suppose these clumps are somehow joined together and, far away from the joining region, each clump is basically a flat Euclidean space of the corresponding dimension. Then, far away from the joint, heat diffusion in a particular clump would behave as if it were in $\mathbb{R}^n$, respectively $\mathbb{R}^m$ (until the time that the flowing heat "hits" the joint region). Thus, in the part of each clump that is far from the joint, the diffusion distance would be normalized differently, one normalization depending on n and the other on m. An overall change of scale would not remove this difference, thus we would not recover the usual Euclidean distance in the two clumps simultaneously, as we would like.

The second point of concern is more general in nature. In the continuous case, Coifman's distance involves the L^2 distance between $T_t \delta_z$, when $z = x$, and $T_t \delta_z$ when $z = y$; see (3.1). The L^1 norm of $T_t \delta_z$ is 1, since

$$\int_M T_t \delta_z(v) dv = \int_M \left(\int_M \rho_t(v, u) \delta_z(u) du \right) dv = 1, \tag{3.13}$$

using the mass preservation assumption of (2.2). For the discrete case, $\mathbf{1}^T A^T e_i = 1$, where $\mathbf{1}^T$ is the $1 \times N$ vector of 1's.

So the diffusion distance proposed by Coifman finds the L^2 (resp., l^2) distance between L^1 (resp., l^1) normalized vectors. Let us illustrate by an example for the discrete situation, with $N = 10,000$, in which this may lead to undesired results. Without specifying the matrix

A, suppose that after some time has passed, we have the following two $1 \times 10,000$ vectors giving two different results of diffusion:

$$v_1 = \left(\frac{1}{100}, \ldots, \frac{1}{100}, 0, \ldots, 0 \right), \tag{3.14}$$

where the first one hundred entries are each $1/100$, and the rest $9,900$ entries are 0, and

$$v_2 = \left(\frac{1}{10,000}, \frac{1}{10,000}, \ldots, \frac{1}{10,000} \right), \tag{3.15}$$

where each entry is $1/10,000$.

Note that v_1 and v_2 both have l^1 norm 1. Now, considering two canonical basis vectors e_i^T and e_j^T, $i \neq j$, each of which has l^1 norm 1, we see that $\langle e_i^T - e_j^T, e_i^T - e_j^T \rangle = 2$. So, a distance of $\sqrt{2}$ gives the (in fact, maximum) separation between two completely different (l^1 unit) diffusion vectors. Return to v_1 and v_2, note that v_2 corresponds to total diffusion, while v_1 has only diffused over 1% of the entries. We would thus hope that v_1 and v_2 would be nearly as much separated as e_i^T and e_j^T, that is, have diffusion distance not much smaller than $\sqrt{2}$. But a trivial calculation shows that

$$\sqrt{\langle v_1 - v_2, v_1 - v_2 \rangle} < .1, \tag{3.16}$$

which seems much smaller than what we would like. The problem is that $\sqrt{\langle v_1 - v_2, v_1 - v_2 \rangle}$ is small since the l^2 norm of each of v_1 and v_2 is small, even though the l^1 norm of each is 1.

In the next section, we propose a variant of the diffusion distance discussed in this section. Our version will find the L^2 (resp., l^2) distance between vectors which are normalized to have L^2 (resp., l^2) norm to be 1, rather than L^1 (or l^1) norm 1.

4. An Alternate Diffusion Distance

In this section, we propose a new diffusion distance. Let us first define our alternate diffusion distance for the continuous case. Refer to Section 2 for the definitions of functions and operators used below.

For any $z \in M$, let

$$\psi_z(u) \equiv \frac{\delta_z(u)}{\|T_t \delta_z\|_2} = \frac{\delta_z(u)}{\sqrt{\int_M \rho_t(v,z)\rho_t(v,z)dv}}. \tag{4.1}$$

Then,

$$T_t \psi_z(u) = \int_M \rho_t(u,w) \frac{\delta_z(w)}{\|T_t \delta_z\|_2} dw = \frac{\rho_t(u,z)}{\|T_t \delta_z\|_2}. \tag{4.2}$$

Note that $T_t \psi_z(\cdot)$ has L^2 norm 1:

$$\int_M [T_t \psi_z(u)]^2 du = \int_M \frac{\rho_t^2(u,z)}{\int_M \rho_t^2(v,z)dv} du = 1. \tag{4.3}$$

For $x, y \in M$, we define our diffusion distance, $d_{2,t}(x,y)$, as follows:

$$\begin{aligned}
[d_{2,t}(x,y)]^2 &\equiv \langle T_t(\psi_x - \psi_y), T_t(\psi_x - \psi_y)\rangle \\
&= 2 - 2\frac{\int_M \rho_t(u,x)\rho_t(u,y)du}{\|\rho_t(\cdot,x)\|_2 \|\rho_t(\cdot,y)\|_2} \\
&= 2 - 2\frac{\langle \rho_t(\cdot,x), \rho_t(\cdot,y)\rangle}{\|\rho_t(\cdot,x)\|_2 \|\rho_t(\cdot,y)\|_2},
\end{aligned} \tag{4.4}$$

where we have used (4.3). Here again, $\langle \cdots, \cdots \rangle$ is the usual inner product on $L^2(M)$. Note the analogy to (3.3).

As is clear from the defining equality in (4.4), $d_{2,t}(x,y)$ is symmetric in x and y and satisfies the triangle inequality:

$$d_{2,t}(x,z) \leq d_{2,t}(x,y) + d_{2,t}(y,z), \tag{4.5}$$

for all $x, y, z \in M$. Geometrically, $d_{2,t}(x,y)$ is the length of the chord joining the tips of the unit vectors $T_t \psi_x$ and $T_t \psi_y$. We have that

$$0 \leq d_{2,t}(x,y) \leq \sqrt{2}, \tag{4.6}$$

for all $x, y \in M$ and $t \geq 0$.

If ρ_t is symmetric in the space variables, by (2.6), we have that

$$[d_{2,t}(x,y)]^2 = 2 - 2\frac{\rho_{2t}(x,y)}{\sqrt{\rho_{2t}(x,x)}\sqrt{\rho_{2t}(y,y)}}. \tag{4.7}$$

As an example, again consider the case where $\rho_t(x,y) = (4\pi t)^{-n/2} e^{-|x-y|^2/4t}$, the fundamental solution to the heat equation in $\mathbb{R}^n$. Then,

$$[d_{2,t}(x,y)]^2 = 2 - 2e^{-|x-y|^2/8t}. \tag{4.8}$$

Note that if $|x-y| \ll \sqrt{t}$, then

$$[d_{2,t}(x,y)]^2 \approx 2 - 2\left(1 - \frac{|x-y|^2}{8t}\right) = \frac{|x-y|^2}{4t}, \tag{4.9}$$

so $d_{2,t}(x, y)$ gives (approximately) the geodesic distance in $\mathbb{R}^n$, in the "near regime" where $|x - y| \ll \sqrt{t}$, and with scale $\sqrt{t}$. Note that unlike (3.12), no $t^{n/2}$ term appears. (Also see the discussion following (3.12).) Also note that if $|x - y|^2/8t$ is large, $d_{2,t}(x, y) \approx \sqrt{2}$ (the greatest possible distance, see (4.6)), so for such t the points x and y are (nearly) maximally separated. Hence, $d_{2,t}(x, y)$, for the case of heat flow in $\mathbb{R}^n$, gives a scaled geodesic distance when x is close to y, with $2\sqrt{t}$ as the unit of length and near maximum separation when x is far from y at the scale $\sqrt{t}$.

For any, say, compact Riemannian manifold M, if $\rho_t(x, y)$ is the fundamental solution to the heat equation on M, we have that

$$\rho_t(x, y) = (4\pi t)^{-n/2} e^{-d^2(x,y)/4t} (1 + \mathcal{O}(t)), \quad \text{as } t \longrightarrow 0^+, \tag{4.10}$$

where $d(x, y)$ is the geodesic distance on M (see [9]). Hence, repeating the expansion in (4.9) for a compact manifold M, with t small, and $d(x, y) \ll \sqrt{t}$, we have that $d_{2,t}(x, y) \approx d(x, y)/2\sqrt{t}$, again recovering (scaled) geodesic distance. (The discussion following (3.12) gives an example for which it would be preferable not to have presented a normalization factor which depends on the dimension.) Exponentially decaying bounds on the fundamental solution of the heat equation for a manifold M (see [9, Chapter XII, Section 12]), suggest that x and y become nearly maximally separated, as given by $d_{2,t}(x, y)$, when $d(x, y)$ (scaled by $\sqrt{t}$) is large, just as in the Euclidean case.

In the discrete situation, where we start with a set of data points $S = \{s_1, s_2, \ldots, s_N\}$, and A is a Markov matrix specifying the transition probabilities between the "states" of S, for $n = 1, 2, \ldots$, we let

$$v_{i,n} = \frac{e_i}{\left\|(A^T)^n e_i\right\|}, \tag{4.11}$$

where e_i is the ith canonical basis vector (see Section 2), and $\|\cdots, \cdots\|$ is the l^2 vector norm. For $s_i, s_j \in S$ and $n = 1, 2, \ldots$, we define $d_{2,n}(s_i, s_j)$ by

$$
\begin{aligned}
\left[d_{2,n}(s_i, s_j)\right]^2 &\equiv \left\langle (A^T)^n (v_{i,n} - v_{j,n}), (A^T)^n (v_{i,n} - v_{j,n}) \right\rangle \\
&= 2 - 2 \frac{\left\langle (A^T)^n e_i, (A^T)^n e_j \right\rangle}{\left\|(A^T)^n e_i\right\| \left\|(A^T)^n e_j\right\|} \\
&= 2 - 2 \frac{\left(A^n (A^T)^n\right)_{ij}}{\sqrt{\left(A^n (A^T)^n\right)_{ii}} \sqrt{\left(A^n (A^T)^n\right)_{jj}}},
\end{aligned}
\tag{4.12}
$$

where $\langle \cdots, \cdots \rangle$ and $\|\cdots, \cdots\|$ are, respectively, the usual inner product and norm in $\mathbb{R}^N$, and, for a matrix B, $(B)_{ij}$ denotes the i, j entry of B.

If A is symmetric,

$$[d_{2,n}(s_i, s_j)]^2 = 2 - 2\frac{(A^{2n})_{ij}}{\sqrt{(A^{2n})_{ii}}\,\sqrt{(A^{2n})_{jj}}}. \tag{4.13}$$

As before, n represents the nth tick of the clock.

5. The Mean First Passage Cost Satisfies the Triangle Inequality: An Example of Its Normalized Limit

In this section, we consider a slightly different topic: the mean first passage cost (defined below) between two states as a measure of separation in the discrete situation. We give two explicit proofs showing that the mean first passage cost satisfies the triangle inequality (in [8], the author states this result when all costs are equal to 1 as Theorem 4.2, but the proof is not very explicit in our opinion).

In [10–12], as well as some of the references listed therein, it is shown that the symmetrized mean first passage time and cost are metrics (for mean first passage cost see, in particular, [10]; also, in the above sources the symmetrized mean first passage time is called the commute time). "Symmetrized" refers to the sum of the first cost (time) to reach a specified state from a starting state and to return back to the starting state. This symmetrization is necessary to ensure a quantity symmetric in the starting and destination states. In the sources cited above, the fundamental underlying operator is the graph Laplacian L, which, using the notation of [12], is defined as $L = D - W$. Here, $W = \langle w_{ij} \rangle$ is the adjacency matrix of a graph, and D is the diagonal degree matrix, with the ith entry on the diagonal equaling $\sum_j w_{ij}$. In addition to assuming the nonnegativity of the w_{ij}'s, the authors in the above works assume that W is *symmetric*. The resulting symmetry (and positive semi-definiteness of L) implies the existence of a full set of nonnegative eigenvalues of L, and the diagonalizability of L is used heavily in the proofs that the commute time/cost is a distance. In the random walk interpretation, see, for example, [12], the following normalized Laplacian is relevant: $L_{rw} = I - D^{-1}W$. To make a connection with the notation in the present paper, $D^{-1}W = A$, a Markov matrix giving the transition probabilities of the random walk. Although $D^{-1}W$ is not necessarily symmetric, it is easy to see that $D^{-1}W = D^{-1/2}\{D^{-1/2}WD^{-1/2}\}D^{1/2}$ (see the discussion in [12]). Hence $D^{-1}W$, while not itself symmetric in general, is conjugate to the *symmetric* matrix $D^{-1/2}WD^{-1/2}$, and thus too has a full complement of eigenvalues.

In this section, as in the rest of the paper unless stated otherwise, we are *not* assuming that the Markov matrix A is symmetric or conjugate to a symmetric matrix; hence A may not be diagonalizable (i.e., A may have Jordan blocks of dimension greater than 1). We thus do not have spectral theory available to us. Furthermore, we do not wish to necessarily symmetrize the mean first passage time/cost to obtain a symmetric quantity; we are not actually going to get a distance, but will try to obtain the "most important" property of being a distance, namely, the triangle inequality.

A model example we are thinking about is the following. Suppose we have a map grid and are tracking some localized storm which is currently at some particular location on the grid. We suppose that the storm behaves like a random walk and has a certain (constant in time) probability to move from one grid location to another at each "tick of the clock" (time step). We can thus model the movements of the storm by a Markov matrix A, with

the nth power of A giving the transition probabilities after n ticks of the clock. If there is no overall wind, the matrix A could reasonably be assumed to be symmetric, and we could use spectral theory. But suppose there is an overall wind in some fixed direction, which is making it more probable for the storm to move north, say, rather than south. Then the matrix A is not symmetric; there is a preferred direction of the storm to move in, from one tick of the clock to the next; spectral theory cannot, in general, be used. Furthermore, it may not be reasonable in this situation to consider the commute time—the symmetrized mean first passage time—since we may rather want to know the expected time to reach a certain population center from the current location of the storm, and may not care about the storm's return to the original location. Thus the mean first passage time would be the quantity of interest.

In the first part of this section, we give two proofs that the mean first passage cost/time, associated with a not-necessarily-symmetric Markov matrix A, does indeed satisfy the triangle inequality; our proofs do not rely on spectral theory. We think that satisfying the triangle inequality, while in general failing to be symmetric, is still a very useful property for a bilinear form to have.

We conclude the section by exhibiting a connection between the (normalized) mean first passage time and the discretized solution of a certain Dirichlet-Poisson problem and verify our result numerically for the simple case of the unit circle.

In this section, $S = \{s_1, s_2, \ldots, s_N\}$ is a finite set of states and A is a Markov matrix giving the transition probabilities between states in one tick of the clock (see Section 2). C will denote an $N \times N$ matrix with non-negative entries, $C = \langle c_{ij} \rangle_{i,j=1}^N$. We will think of each $c_{i,j}$ as the "cost" associated with the transition from state s_i to state s_j. By a slight abuse of notation, for $1 \le m, n \le N$, P_n will be the $N \times N$ matrix in which all entries are 0, except the (n, n) entry which is 1 (this corresponds to $X = \{s_n\}$ in Section 2). Also, P_{mn} will be the $N \times N$ matrix in which all entries are 0, except the (m, m) and (n, n) entries each of which is 1 (this corresponds to $X = \{s_m, s_n\}$ in Section 2).

Let Y_{mn} be the random variable which gives the cost accumulated by a particle starting at state s_m until its first visit to state s_n after leaving s_m. In other words, if a particular path of the particle is given by the states $s_m, s_{j_1}, s_{j_2}, \ldots, s_{j_p}, s_n$, the value of Y_{mn} is $c_{mj_1} + c_{j_1 j_2} + \cdots + c_{j_p n}$. We suppose A has the property that for every i, j, there exists an n such that $(A^n)_{ij} > 0$, that is, every state j is eventually reachable from every state i. Then, as is shown in [13] (using slightly different notation), we have the following formula for $\mathbb{E}(Y_{mn})$, which is the expected cost of going from state s_m to state s_n:

$$\mathbb{E}(Y_{mn}) = e_m^T \frac{I}{I - A(I - P_n)} \langle c_{pq} a_{pq} \rangle \mathbf{1}, \tag{5.1}$$

where $\langle c_{pq} a_{pq} \rangle$ is the $N \times N$ matrix with p, q entry equal to $c_{pq} a_{pq}$. (In particular, it is shown in [13] that $I - A(I - P_n)$ is invertible and $\|(A(I - P_n))^k\| \to 0$, as $k \to \infty$.) See [14, 15] for discussion of related expected values, and [8, 10–12, 16–18] for discussion of mean first passage times and related concepts.

We will give two proofs that the expected cost of going from one state to another satisfies the triangle inequality.

Proposition 5.1. $\mathbb{E}(Y_{ij}) \leq \mathbb{E}(Y_{ik}) + \mathbb{E}(Y_{kj})$.

We again note that this proposition, for the case all costs are 1, is stated as Theorem 4.2 in [8], but we feel the proof is not very explicit. (In our proofs below, we assume $j \neq k$; if $j = k$, the inequality in Proposition 5.1 is immediate.)

Proof. (1) Our first proof is probabilistic. Let a random walker start at state s_i and accumulate costs given by the matrix C as he moves from state to state. As soon as the walker reaches state s_j, we obtain a sample value of Y_{ij}. Now, at this point of the walk, there are two possibilities. Either the walker has passed through state s_k before his first visit to s_j after leaving s_i, or he has not. In the first instance, we have obtained sample values of Y_{ik} and Y_{kj} along the way, and $Y_{ij} = Y_{ik} + Y_{kj}$ for this simulation. In the second case, we let the walker continue until he first reaches s_k, to obtain a sample value of Y_{ik}, and walk still more until he reaches s_j for the first time since leaving s_k, thus giving a sample value of Y_{kj} (note that by the memoryless property, this sample value of Y_{kj} is independent of the walker's prior history). In the second case, we thus clearly have that $Y_{ij} < Y_{ik} + Y_{kj}$. Combining the two cases, we have $Y_{ij} \leq Y_{ik} + Y_{kj}$. Repeating the simulation, averaging, and taking the limit as the number of simulations goes to infinity, we obtain that $\mathbb{E}(Y_{ij}) \leq \mathbb{E}(Y_{ik}) + \mathbb{E}(Y_{kj})$.

(2) Our second proof is via explicit matrix computations. Let us define the following two quantities:

$$Q_1 \equiv e_i^T \frac{I}{I - A(I - P_{jk})} \langle c_{pq} a_{pq} \rangle \mathbf{1},$$

$$Q_2 \equiv \left\{ e_i^T \frac{I}{I - A(I - P_{jk})} A P_k \mathbf{1} \right\} e_k^T \frac{I}{I - A(I - P_j)} \langle c_{pq} a_{pq} \rangle \mathbf{1}.$$

(5.2)

(See Section 2 and the paragraphs before the statement of Proposition 5.1.) Now, we have

$$Q_1 \equiv e_i^T \frac{I}{I - A(I - P_{jk})} \langle c_{pq} a_{pq} \rangle \mathbf{1}$$

$$= e_i^T \sum_{m=0}^{\infty} \left[A(I - P_{jk}) \right]^m \langle c_{pq} a_{pq} \rangle \mathbf{1}$$

$$\leq e_i^T \sum_{m=0}^{\infty} \left[A(I - P_k) \right]^m \langle c_{pq} a_{pq} \rangle \mathbf{1}$$

$$= e_i^T \frac{I}{I - A(I - P_k)} \langle c_{pq} a_{pq} \rangle \mathbf{1}$$

$$= \mathbb{E}(Y_{ik});$$

(5.3)

see (5.1). Also,

$$Q_2 \equiv \left\{ e_i^T \frac{I}{I - A(I - P_{jk})} A P_k \mathbf{1} \right\} e_k^T \frac{I}{I - A(I - P_j)} \langle c_{pq} a_{pq} \rangle \mathbf{1}$$

$$= \left\{ e_i^T \frac{I}{I - A(I - P_{jk})} A P_k \mathbf{1} \right\} \mathbb{E}(Y_{kj}). \tag{5.4}$$

But

$$e_i^T \frac{I}{I - A(I - P_{jk})} A P_k \mathbf{1} \leq e_i^T \frac{I}{I - A(I - P_k)} A P_k \mathbf{1}$$

$$= e_i^T \frac{I}{I - A(I - P_k)} (I - A(I - P_k)) \mathbf{1} \tag{5.5}$$

$$= 1,$$

where we have used a series expansion to show the first inequality (all entries are non-negative), and the fact that $A P_k \mathbf{1} = (I - A(I - P_k))\mathbf{1}$, since $A\mathbf{1} = \mathbf{1}$. Thus, $Q_2 \leq \mathbb{E}(Y_{kj})$.

We will finish our second proof by showing that $Q_1 + Q_2 = \mathbb{E}(Y_{ij})$. First note that

$$Q_2 \equiv \left\{ e_i^T \frac{I}{I - A(I - P_{jk})} A P_k \mathbf{1} \right\} e_k^T \frac{I}{I - A(I - P_j)} \langle c_{pq} a_{pq} \rangle \mathbf{1}$$

$$= e_i^T \frac{I}{I - A(I - P_{jk})} A P_k \frac{I}{I - A(I - P_j)} \langle c_{pq} a_{pq} \rangle \mathbf{1}, \tag{5.6}$$

using that $P_k = e_k e_k^T = P_k \mathbf{1} e_k^T$. Thus,

$$Q_1 + Q_2 \quad = e_i^T \frac{I}{I - A(I - P_{jk})} \left\{ I - A(I - P_j) + A P_k \right\} \frac{I}{I - A(I - P_j)} \langle c_{pq} a_{pq} \rangle \mathbf{1}$$

$$= e_i^T \frac{I}{I - A(I - P_{jk})} \left\{ I - A(I - P_j - P_k) \right\} \frac{I}{I - A(I - P_j)} \langle c_{pq} a_{pq} \rangle \mathbf{1}$$

$$= e_i^T \frac{I}{I - A(I - P_{jk})} \left\{ I - A(I - P_{jk}) \right\} \frac{I}{I - A(I - P_j)} \langle c_{pq} a_{pq} \rangle \mathbf{1} \tag{5.7}$$

$$= e_i^T \frac{I}{I - A(I - P_j)} \langle c_{pq} a_{pq} \rangle \mathbf{1}$$

$$= \mathbb{E}(Y_{ij}).$$

Here we have used the fact that $P_j + P_k = P_{jk}$ (as mentioned earlier, we are assuming $j \neq k$; the triangle inequality we are proving holds trivially for the case $j = k$). $\qquad\square$

We would like to point out that the decomposition of $\mathbb{E}(Y_{ij}) = Q_1 + Q_2$ in the second proof above is not a "miraculous" guess. We arrived at this decomposition by writing $\mathbb{E}(Y_{ij})$ as the derivative (evaluated at 0) of the characteristic function (Fourier transform) of Y_{ij} (see [13]), and breaking up the expression to be differentiated into a sum of terms: one term corresponding to the random walk going from s_i to s_j without visiting s_k first, and one term corresponding to visiting s_k before reaching s_j. After differentiation, the resulting six pieces, when suitably combined into two terms, yielded Q_1 and Q_2.

We conclude this section by considering certain (suitably normalized) limiting values of the expected cost of going from state s_i to state s_j, for the first time after leaving s_i, given by (5.1). For this discussion, we will take all the costs to be identically 1, that is, $c_{pq} = 1$ for all p, q. Then, we see from (5.1) that

$$
\begin{aligned}
\mathbb{E}(Y_{ij}) &= e_i^T \frac{I}{I - A(I - P_j)} \mathbf{1} = e_i^T \frac{I}{I - A(I - P_j)} A\mathbf{1} \\
&= e_i^T \frac{I}{I - A(I - P_j)} A(I - P_j)\mathbf{1} + e_i^T \frac{I}{I - A(I - P_j)} AP_j\mathbf{1} \\
&= e_i^T \frac{I}{I - A(I - P_j)} A(I - P_j)\mathbf{1} + e_i^T \frac{I}{I - A(I - P_j)} (I - A(I - P_j))\mathbf{1} \\
&= e_i^T \frac{I}{I - A(I - P_j)} A(I - P_j)\mathbf{1} + 1,
\end{aligned}
\tag{5.8}
$$

where we have used $AP_j\mathbf{1} = (I - A(I - P_j))\mathbf{1}$, since $A\mathbf{1} = \mathbf{1}$.

Now, let us digress a little to describe a stochastic approach to solving certain boundary value problems. The description below follows very closely parts of Chapter 9 in [19]. Some statements are excerpted verbatim from that work, with minor changes in some labels. The background results below are well known and are often referred to as Dynkin's formula (see, e.g, [20]). We are presenting them for the reader's convenience and will use them to exhibit an interesting connection between the mean first passage time and the discretized solution of a certain Dirichlet-Poisson problem; see (5.16).

Let D be a domain in $\mathbb{R}^n$, and let L denote a partial differential operator on $C^2(\mathbb{R}^n)$ of the form:

$$
L = \sum_{i=1}^{n} b_i(x) \frac{\partial}{\partial x_i} + \sum_{i,j=1}^{n} a_{ij}(x) \frac{\partial^2}{\partial x_i \partial x_j},
\tag{5.9}
$$

where $a_{ij}(x) = a_{ji}(x)$. We assume each $a_{ij}(x) \in C^2(D)$ is bounded and has bounded first and second partial derivatives; also, each $b_i(x)$ is Lipschitz. Suppose L is uniformly elliptic in D (i.e., all the eigenvalues of the symmetric matrix $\langle a_{ij}(x) \rangle$ are positive and stay uniformly away from 0 in D). Then, for $g \in C^\alpha(D)$, some $\alpha > 0$, and g bounded, and for $\phi \in C(\partial D)$, the function w defined below solves the following Dirichlet-Poisson problem:

$$
Lw = -g \quad \text{in } D,
$$
$$
\lim_{x \to y, x \in D} w(x) = \phi(y), \quad \text{for all regular points } y \in \partial D.
\tag{5.10}
$$

(Regular points in this context are defined in [19] and turn out to be the same as the regular points in the classical sense, i.e., the points y on ∂D where the limit of the generalized Perron-Wiener-Brelot solution coincides with $\phi(y)$, for all $\phi \in C(\partial D)$.)

Now we define w. We choose a square root $\sigma(x) \in \mathbb{R}^{n \times n}$ of the matrix $2\langle a_{ij}(x) \rangle$, that is,

$$\frac{1}{2}\sigma(x)\sigma^T(x) = \langle a_{ij}(x) \rangle. \tag{5.11}$$

Next, for $b = \langle b_i(x) \rangle$, let X_t be an Itô diffusion solving

$$dX_t = b(X_t)dt + \sigma(X_t)dB_t, \tag{5.12}$$

where B_t is n-dimensional Brownian motion. Then,

$$w(x) = \mathbb{E}_x[\phi(X_\tau)] + \mathbb{E}_x\left[\int_0^\tau g(X_u)du\right], \quad \text{for } x \in D, \tag{5.13}$$

is a solution of (5.10). Here, the expected values are over paths starting from $x \in D$, and τ is the first exit time from D.

Let us transfer the above discussion to, say, a compact manifold M, rather than $\mathbb{R}^n$. We sample M and let the "states" s_k be the sample points. We construct a transition matrix A to give a discretized version of (5.12). Let $\epsilon > 0$ be the approximate separation between the sample points. Fix a sample point s_j, and let $D = M \setminus \overline{B_\epsilon(s_j)}$ be the domain in M consisting of the complement of the closure of the ball of radius ϵ in M, center s_j. Let s_i be a sample point in D. For this situation, in (5.10), let ϕ be the 0 function, and g be the constant 1 function. Then (5.13) becomes:

$$w(x) = \mathbb{E}_{s_i}\left[\int_0^\tau du\right], \tag{5.14}$$

τ is the first exit time from D, that is, first visit time to the ϵ neighborhood of s_j. (Compare with Proposition 8B in [21] which discusses the case of the Dirichlet-Poisson problem (5.10) with $\phi = 0$ and $g = 1$ for a manifold.) As shown in [13] (with slightly different notation), a discrete version of (5.14) is

$$\left(e_i^T \frac{I}{I - A(I - P_j)} \langle \widetilde{c_{pq}} a_{pq} \rangle \mathbf{1}\right) \Delta t, \tag{5.15}$$

where $\widetilde{c_{pj}} = 0$, all p, and $\widetilde{c_{pq}} = 1$, for all q not equal to j, and all p. Thus, $\langle \widetilde{c_{pq}} a_{pq} \rangle = A(I - P_j)$. Combining (5.15) and (5.8), we see that:

$$\mathbb{E}(Y_{ij})\Delta t - \Delta t = \left(e_i^T \frac{I}{I - A(I - P_j)} A(I - P_j)\mathbf{1}\right)\Delta t \approx w(s_j), \tag{5.16}$$

for Δt small.

We thus see a connection between the (normalized) mean first passage time and the solution to the Dirichlet-Poisson problem discussed above.

Let us illustrate the preceding discussion by a simple example: $M = S^1$, the unit circle. We will consider $d^2/d\theta^2$, the Laplacian on S^1, and sample S^1 uniformly. We will let the transition matrix A take the walker from the current state to each of the two immediate neighbor states, with probability $1/2$ for each. The variance is then $(\Delta\theta)^2$. Since $dX_t = \sqrt{2}dB_t$, see (5.12), we must have $(\Delta\theta)^2 = 2\Delta t$, and we should use $(\Delta\theta)^2/2$ as our value of Δt in (5.16). Using symmetry, we can take $s_j = 0$, the 0 angle on S^1, without loss of generality. Let

$$w_\epsilon(\theta) = \frac{1}{2}(\theta - \epsilon)((2\pi - \theta) - \epsilon), \quad \epsilon \leq \theta \leq 2\pi - \epsilon. \tag{5.17}$$

Note that

$$\frac{d^2}{d\theta^2}w_\epsilon(\theta) = -1, \quad \epsilon < \theta < 2\pi - \epsilon,$$
$$w_\epsilon(\epsilon) = w_\epsilon(2\pi - \epsilon) = 0. \tag{5.18}$$

So w_ϵ is the unique solution satisfying (5.10) for our example, on the domain $S^1 \setminus [-\epsilon, \epsilon]$.

To numerically confirm (5.16), we ran numerical experiments in which we discretized S^1 into N equispaced points, with the transition matrix A taking a state to each of its 2 immediate neighbors with probability $1/2$, and used $(\Delta\theta)^2/2$ as the value of Δt in (5.16) to calculate $\mathbb{E}(Y_{ij})\Delta t - \Delta t$. We took s_j to be the angle 0, and s_i to be the closest sample point to the angle with radian measure 1, for example. Letting $\epsilon \to 0$ in (5.17), we compared the value of $\mathbb{E}(Y_{ij})\Delta t - \Delta t$ with $w(\theta) = (1/2\theta)(2\pi - \theta)$. For instance, with $N = 1000$, $s_j = 0$, and s_i the nearest sample point to the angle with radian measure 1, the relative error is less than 0.08%. Note that $w(\theta)$ is, for θ close to 0, essentially a scaled geodesic distance on S^1 (from our base angle 0).

6. Conclusions

The authors have presented a diffusion distance which uses L^2 unit vectors, and which is based on the well-known cosine similarity. They have discussed why the normalization may make diffusion distances more meaningful. We also gave two explicit proofs of the triangle inequality for mean first passage cost, and exhibited a connection between the (normalized) mean first passage time and the discretized solution of a certain Dirichlet-Poisson problem.

Acknowledgments

References

[1] R. R. Coifman and S. Lafon, "Diffusion maps," *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 5–30, 2006.

[2] S. S. Lafon, *Diffusion maps and geometric harmonics*, Ph.D. thesis, Yale University, 2004.

[3] R. R. Coifman and M. Maggioni, "Diffusion wavelets," *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 53–94, 2006.

[4] B. Nadler, S. Lafon, R. R. Coifman, and I. G. Kevrekidis, "Difusion maps, spectral clustering and reaction coordinates of dynamical systems," *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 113–127, 2006.

[5] M. Brand and K. Huang, "A unifying theorem for spectral embedding and clustering," in *Proceedings of the 9th International Workshop on Artificial Intelligence and Statistics*, C. M. Bishop and B. J. Frey, Eds., Key West, Fla, USA, 2003, http://research.microsoft.com/enus/um/cambridge/events/aistats2003/proceedings/papers.htm .

[6] Y. Gong and W. Xu, *Machine Learning for Multimedia Content Analysis*, Springer, New York, NY, USA, 2007.

[7] T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," *The Annals of Statistics*, vol. 36, no. 3, pp. 1171–1220, 2008.

[8] J. J. Hunter, "Stationary distributions and mean first passage times of perturbed Markov chains," *Research Letters in the Information and Mathematical Sciences*, vol. 3, pp. 85–98, 2002.

[9] I. Chavel, *Eigenvalues in Riemannian geometry*, vol. 115 of *Pure and Applied Mathematics*, Academic Press, Orlando, Fla, USA, 1984.

[10] F. Fouss, A. Pirotte, J. -M. Renders, and M. Saerens, "Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 3, pp. 355–369, 2007.

[11] M. Saerens, F. Fouss, L. Yen, and P. Dupont, "The principal components analysis of a graph, and its relationships to spectral clustering," in *Machine Learning: ECML 2004*, vol. 3201 of *Lecture Notes in Artificial Intelligence*, pp. 371–383, Springer, Berlin, Germany, 2004.

[12] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.

[13] M. Goldberg and S. Kim, "Applications of some formulasfor finite state Markov chains," *Applied and Computational Harmonic Analysis*. In press.

[14] J. J. Hunter, "On the moments of Markov renewal processes," *Advances in Applied Probability*, vol. 1, pp. 188–210, 1969.

[15] J. J. Hunter, "Generalized inverses and their application to applied probability problems," *Linear Algebra and its Applications*, vol. 45, pp. 157–198, 1982.

[16] J. J. Hunter, "A survey of generalized inverses and their use in stochastic modelling," *Research Letters in the Information and Mathematical Sciences*, vol. 1, pp. 25–36, 2000.

[17] J. J. Hunter, "Generalized inverses, stationary distributions and mean first passage times with applications to perturbed Markov chains," *Research Letters in the Information and Mathematical Sciences*, vol. 3, pp. 99–116, 2002.

[18] J. G. Kemeny and J. L. Snell, *Finite Markov Chains*, The University Series in Undergraduate Mathematics, Van Nostrand, Princeton, NJ, USA, 1960.

[19] B. Øksendal, *Stochastic Differential Equations: An Introduction with applications*, Springer, Berlin, Germany, 6th edition, 2007.

[20] D. Kulasiri and W. Verwoerd, *"Stochastic Dynamics: Modeling Solute Transport in Porous Media*, vol. 44 of *North-Holland Series in Applied Mathematics and Mechanics*, Elsevier, Amsterdam, The Netherlands, 2002.

[21] K. D. Elworthy, *Stochastic Differential Equations on Manifolds*, vol. 70 of *London Mathematical Society Lecture Note Series*, Cambridge University Press, Cambridge, UK, 1982.