Bi-stochastic kernels via asymmetric affinity functions[☆]

Ronald R. Coifman, Matthew J. Hirn[*]

Yale University
Department of Mathematics
P.O. Box 208283
New Haven, Connecticut 06520-8283
USA

Abstract

In this short letter we present the construction of a bi-stochastic kernel p for an arbitrary data set X that is derived from an asymmetric affinity function α. The affinity function α measures the similarity between points in X and some reference set Y. Unlike other methods that construct bi-stochastic kernels via some convergent iteration process or through solving an optimization problem, the construction presented here is quite simple. Furthermore, it can be viewed through the lens of out of sample extensions, making it useful for massive data sets.

Keywords: bi-stochastic kernel; Nyström extension

1. Introduction

Given a positive, symmetric kernel (matrix) k, the question of how to construct a bi-stochastic kernel derived from k has been of interest in certain applications such as data clustering. Various algorithms for this task exist. One of the best known is the Sinkhorn-Knopp algorithm [1], in which one alternately normalizes the rows and columns of k to sum to one. A symmetrization of this algorithm is given in [2], and is subsequently used to cluster data. In both cases, an infinite number of iterations are needed for the process to converge to a bi-stochastic matrix. In another application of data clustering, the authors in [3] solve a quadratic programming problem to obtain what they call the Bregmanian bi-stochastication of k. Common to these algorithms and others is the complexity in solving for (or approximating) the bi-stochastic matrix.

Also related to the goal of organizing data, over the last decade we have seen the development of a class of research that utilizes nonlinear mappings into low dimensional spaces in order to organize potentially high dimensional data. Examples include locally linear embedding (LLE) [4], ISOMAP [5], Hessian LLE [6], Laplacian eigenmaps [7], and diffusion maps [8]. In many applications, these data sets are not only high dimensional, but massive. Thus there has been the need to develop complementary methods by which these nonlinear mappings can be computed

[☆] *Applied and Computational Harmonic Analysis*, volume 35, number 1, pages 177-180, 2013. arXiv:1209.0237.
[*] Corresponding author
Email addresses: coifman@math.yale.edu (Ronald R. Coifman), matthew.hirn@yale.edu (Matthew J. Hirn)
URL: www.math.yale.edu/~mh644 (Matthew J. Hirn)

efficiently. The Nyström extension is one early such example; in [9] several out of sample extensions are given for various nonlinear mappings, while [10] utilizes geometric harmonics to extend empirical functions.

In this letter we present an extremely simple bi-stochastic kernel construction that can also be implemented to handle massive data sets. Let X be the data set. The entire construction is derived not from a kernel on X, but rather an asymmetric affinity function $\alpha : X \times Y \to \mathbb{R}$ between the given data and some reference set Y. The key to realizing the bi-stochastic nature of the derived kernel is to apply the correct weighted measure on X. The eigenfunctions (or eigenvectors) of this bi-stochastic kernel on X are also easily computable via a Nyström type extension of the eigenvectors of a related kernel on Y. Since the reference set can usually be taken to much smaller than the original data set, these eigenvectors are simple to compute.

2. A simple bi-stochastic kernel construction

We take our data set to be a measure space (X, μ), in which μ represents the distribution of the points. We also assume that we are given, or able to compute, a finite reference set $Y \triangleq \{y_1, \ldots, y_n\}$. Note that one can take X to be discrete or finite as well; in particular, one special case is when $X = Y$ and $\mu = \frac{1}{n}\sum_{i=1}^{n} \delta_{y_i}$.

2.1. Affinity functions and densities

Let $\alpha : X \times Y \to \mathbb{R}$ be a positive affinity function that measures the similarity between the data set X and the reference set Y. Larger values of $\alpha(x, y_i)$ indicate that the two data points are very similar, while those values closer to zero imply that x and y_i are quite different. The function α serves as a generalization of the traditional kernel function $k : X \times X \to \mathbb{R}$, in which $k(x, x')$ measures the similarity between two points $x, x' \in X$ (just as with α, the larger $k(x, x')$, the more similar the two points). Kernel functions have been successfully used in applied mathematics and machine learning for various data driven tasks. Certain kernel functions can be viewed as an inner product $k(x, x') = \Phi(x) \cdot \Phi(x')$, after the data set X has been mapped nonlinearly into a higher dimensional space via Φ. The idea is that if the kernel k is constructed carefully, than the mapping Φ will arrange the data set X so that certain tasks (such as data clustering) can be done more easily (say using linear methods). We show that the more general function α can be used similarly, with the added benefits of deriving a bi-stochastic kernel that is amenable to an out of sample Nyström type extension.

We derive two densities from the affinity function α, which we shall then use to normalize it. The first of these is the density $\Omega : X \to \mathbb{R}$ on the data set X; we take it as

$$\Omega(x) \triangleq \sum_{i=1}^{n} \alpha(x, y_i), \quad \text{for all } x \in X.$$

We also have a density $\omega : Y \to \mathbb{R}$ on the reference set, which we define as:

$$\omega(y_i) \triangleq \left(\int_X \alpha(x, y_i)\, \Omega(x)\, d\mu(x) \right)^{\frac{1}{2}}, \quad \text{for all } y_i \in Y.$$

Assumption 1. We make the following simple assumptions concerning α:

1. For each $y_i \in Y$, the function $\alpha(\cdot, y_i) : X \to \mathbb{R}$ is square integrable, i.e.,

$$\alpha(\cdot, y_i) \in L^2(X, \mu). \tag{1}$$

2. The densities Ω and ω are finite and strictly positive:

$$0 < \Omega(x) < \infty, \quad \text{for all } x \in X, \tag{2}$$

$$0 < \omega(y_i) < \infty, \quad \text{for all } y_i \in Y. \tag{3}$$

The L^2 integrability condition is necessary to make sure that functions and operators related to X make sense. The upper bounds on the densities simply put a finite limit on how close any point x or y_i is to either Y or X, respectively. Meanwhile, the lower bounds state that each point in X has some relation to the reference set Y, and likewise that each reference point y_i has some similarity to at least part of X.

Using α and the two densities Ω and ω, we define a normalized affinity $\beta : X \times Y \to \mathbb{R}$ as

$$\beta(x, y_i) \triangleq \frac{\alpha(x, y_i)}{\Omega(x)\,\omega(y_i)}, \quad \text{for all } (x, y_i) \in X \times Y.$$

From this point forward we will use the weighted measure $\Omega^2\mu$ on X. This measure is the "correct" measure in the sense that it is the measure for which we can define a bi-stochastic kernel on X in a natural, simple way. Using Assumption 1, one can easily show that for each $y_i \in Y$, the function $\beta(\cdot, y_i) : X \to \mathbb{R}$ is well behaved under this measure:

$$\beta(\cdot, y_i) \in L^2(X, \Omega^2\mu), \quad \text{for all } y_i \in Y.$$

We note that affinity functions similar to β were first considered in [11] in the context of out of sample extensions for independent components analysis (ICA). It has also been utilized in [12] in the context of filtering. The connection to bi-stochastic kernels, though, has until now gone unnoticed.

2.2. The bi-stochastic kernel

To construct the bi-stochastic kernel on X we utilize the β affinity function. Let $p : X \times X \to \mathbb{R}$ denote the kernel, and define it as:

$$\begin{aligned} p(x, x') &\triangleq \langle \beta(x, \cdot), \beta(x', \cdot) \rangle_{\mathbb{R}^n} \\ &= \sum_{i=1}^{n} \beta(x, y_i)\,\beta(x', y_i), \quad \text{for all } (x, x') \in X \times X. \end{aligned}$$

The following proposition summarizes the main properties of p.

Proposition 2. *If Assumption 1 holds, then:*

1. *The kernel p is square integrable under the weighted measure $\Omega^2\mu$, i.e.,*

$$p \in L^2(X \times X, \Omega^2\mu \otimes \Omega^2\mu).$$

2. *The kernel p is bi-stochastic under the weighted measure $\Omega^2\mu$, i.e.,*

$$\int_X p(x, x')\,\Omega(x')^2\mu(x') = \int_X p(x', x)\,\Omega(x')^2\mu(x') = 1, \quad \textit{for all } x \in X.$$

Proof. We begin with the L^2 integrability condition. Using the definition of p, expanding its L^2 norm, and applying Hölder's theorem gives:

$$\begin{aligned}\|p\|^2_{L^2(X\times X,\Omega^2\mu\otimes\Omega^2\mu)} &= \int_X \int_X \sum_{i,j=1}^{n} \beta(x,y_i)\,\beta(x',y_i)\,\beta(x,y_j)\,\beta(x',y_j)\,\Omega(x')^2\,\Omega(x)^2\,d\mu(x)d\mu(x') \\ &\le n^2 \max_{y_i\in Y} \|\beta(\cdot,y_i)\|^4_{L^2(X,\Omega^2\mu)} \\ &< \infty.\end{aligned}$$

Now we show that p is bi-stochastic:

$$\begin{aligned}\int_X p(x,x')\,\Omega(x')^2\,d\mu(x') &= \int_X \sum_{i=1}^{n} \frac{\alpha(x,y_i)\,\alpha(x',y_i)}{\Omega(x)\,\Omega(x')\,\omega(y_i)^2}\,\Omega(x')^2\,d\rho(y)d\mu(x') \\ &= \sum_{i=1}^{n} \frac{\alpha(x,y_i)}{\Omega(x)\,\omega(y_i)^2} \int_X \alpha(x',y_i)\,\Omega(x')\,d\mu(x') \\ &= \sum_{i=1}^{n} \frac{\alpha(x,y_i)}{\Omega(x)} \\ &= 1.\end{aligned}$$

Since p is clearly symmetric, this completes the proof. □

Since $p \in L^2(X\times X, \Omega^2\mu\otimes\Omega^2\mu)$, one can define the integral operator $P : L^2(X,\Omega^2\mu) \to L^2(X,\Omega^2\mu)$ as:

$$(Pf)(x) \triangleq \int_X p(x,x')\,f(x')\,\Omega(x')^2\,d\mu(x'), \quad \text{for all } f \in L^2(X,\Omega^2\mu).$$

Given the results of Proposition 2, we see that P is a Hilbert-Schmidt, self-adjoint, diffusion operator. In terms of data organization and clustering, it is usually the eigenfunctions and eigenvalues of P that are of interest; see, for example, [8]. The fact that P is bi-stochastic though, as opposed to merely row stochastic, could make it particularly interesting for these types of applications in which $I - P$ is used as an approximation of the Laplacian.

2.3. A Nyström type extension

The affinity β can also be used to construct an $n \times n$ matrix A as follows:

$$\begin{aligned}A[i,j] &\triangleq \langle \beta(\cdot,y_i), \beta(\cdot,y_j)\rangle_{L^2(X,\Omega^2\mu)} \\ &= \int_X \beta(x,y_i)\,\beta(x,y_j)\,\Omega(x)^2\,d\mu(x), \quad \text{for all } i,j = 1,\ldots,n.\end{aligned}$$

The matrix A is useful for computing the eigenfunctions and eigenvalues of P. The following proposition is simply an interpretation of the singular value decomposition (SVD) in this context.

Proposition 3. *If Assumption 1 holds, then:*

1. *Let $\lambda \in \mathbb{R} \setminus \{0\}$. Then λ is an eigenvalue of P if and only if it is an eigenvalue of A.*
2. *Let $\lambda \in \mathbb{R} \setminus \{0\}$. If $\psi \in L^2(X, \Omega^2\mu)$ is an eigenfunction of P with eigenvalue λ and $v \in \mathbb{R}^n$ is the corresponding eigenvector of A, then:*

$$\psi(x) = \frac{1}{\sqrt{\lambda}} \sum_{i=1}^{n} \beta(x, y_i)\, v[i],$$
$$v[i] = \frac{1}{\sqrt{\lambda}} \int_X \beta(x, y_i)\, \psi(x)\, \Omega(x)^2\, d\mu(x).$$

3. Acknowledgements

This research was supported by Air Force Office of Scientific Research STTR FA9550-10-C-0134 and by Army Research Office MURI W911NF-09-1-0383. We would also like to thank the anonymous reviewer for his or her helpful comments and suggestions.

References

[1] R. Sinkhorn, P. Knopp, Concerning nonnegative matrices and doubly stochastic matrices, Pacific Journal of Mathematics 21 (1967) 343–348.
[2] R. Zass, A. Shashua, A unifying approach to hard and probabilistic clustering, in: Proceedings of the 10th International Conference on Computer Vision (2005), volume 1, Beijing, China, pp. 294–301.
[3] F. Wang, P. Li, A. C. König, M. Wan, Improving clustering by learning a bi-stochastic data similarity matrix, Knowledge and Information Systems 32 (2012) 351–382.
[4] S. T. Roweis, L. K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.
[5] J. B. Tenenbaum, V. de Silva, J. C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (2000) 2319–2323.
[6] D. L. Donoho, C. Grimes, Hessian eigenmaps: new locally linear embedding techniques for high-dimensional data, Proceedings of the National Academy of Sciences of the United States of America 100 (2003) 5591–5596.
[7] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Computation 15 (2003) 1373–1396.
[8] R. R. Coifman, S. Lafon, Diffusion maps, Applied and Computational Harmonic Analysis 21 (2006) 5–30.
[9] Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N. L. Roux, M. Ouimet, Out-of-sample extensions for LLE, isomap, MDS, eigenmaps, and spectral clustering, in: Advances in Neural Information Processing Systems, MIT Press, 2004, pp. 177–184.
[10] R. R. Coifman, S. Lafon, Geometric harmonics: A novel tool for multiscale out-of-sample extension of empirical functions, Applied and Computational Harmonic Analysis 21 (2006) 31–52.
[11] D. Kushnir, A. Haddad, R. R. Coifman, Anisotropic diffusion on sub-manifolds with application to earth structure classification, Applied and Computational Harmonic Analysis 32 (2012) 280–294.
[12] A. Haddad, D. Kushnir, R. R. Coifman, Filtering via a reference set, Technical Report 1441, Yale University, 2011.