Kernels and Regularization on Graphs

Alexander J. Smola[1] and Risi Kondor[2]

[1] Machine Learning Group, RSISE
Australian National University
Canberra, ACT 0200, Australia
Alex.Smola@anu.edu.au

[2] Department of Computer Science
Columbia University
1214 Amsterdam Avenue, M.C. 0401
New York, NY 10027, USA
risi@cs.columbia.edu

Abstract. We introduce a family of kernels on graphs based on the notion of regularization operators. This generalizes in a natural way the notion of regularization and Greens functions, as commonly used for real valued functions, to graphs. It turns out that diffusion kernels can be found as a special case of our reasoning. We show that the class of positive, monotonically decreasing functions on the unit interval leads to kernels and corresponding regularization operators.

1 Introduction

There has recently been a surge of interest in learning algorithms that operate on input spaces $\mathcal{X}$ other than $\mathbb{R}^n$, specifically, discrete input spaces, such as strings, graphs, trees, automata etc.. Since kernel-based algorithms, such as Support Vector Machines, Gaussian Processes, Kernel PCA, etc. capture the structure of $\mathcal{X}$ via the kernel $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$, as long as we can define an appropriate kernel on our discrete input space, these algorithms can be imported wholesale, together with their error analysis, theoretical guarantees and empirical success.

One of the most general representations of discrete metric spaces are graphs. Even if all we know about our input space are local pairwise similarities between points $x_i, x_j \in \mathcal{X}$, distances (e.g shortest path length) on the graph induced by these similarities can give a useful, more global, sense of similarity between objects. In their work on Diffusion Kernels, Kondor and Lafferty [2002] gave a specific construction for a kernel capturing this structure. Belkin and Niyogi [2002] proposed an essentially equivalent construction in the context of approximating data lying on surfaces in a high dimensional embedding space, and in the context of leveraging information from unlabeled data.

In this paper we put these earlier results into the more principled framework of Regularization Theory. We propose a family of regularization operators (equivalently, kernels) on graphs that include Diffusion Kernels as a special case, and show that this family encompasses all possible regularization operators invariant under permutations of the vertices in a particular sense.

Outline of the Paper: Section 2 introduces the concept of the graph Laplacian and relates it to the Laplace operator on real valued functions. Next we define an extended class of regularization operators and show why they have to be essentially a function of the Laplacian. An analogy to real valued Greens functions is established in Section 3.3, and efficient methods for computing such functions are presented in Section 4. We conclude with a discussion.

2 Laplace Operators

An undirected unweighted graph G consists of a set of vertices V numbered 1 to n, and a set of edges E (i.e., pairs (i,j) where $i,j \in V$ and $(i,j) \in E \Leftrightarrow (j,i) \in E$). We will sometimes write $i \sim j$ to denote that i and j are neighbors, i.e. $(i,j) \in E$. The adjacency matrix of G is an $n \times n$ real matrix W, with $W_{ij} = 1$ if $i \sim j$, and 0 otherwise (by construction, W is symmetric and its diagonal entries are zero). These definitions and most of the following theory can trivially be extended to weighted graphs by allowing $W_{ij} \in [0, \infty)$.

Let D be an $n \times n$ diagonal matrix with $D_{ii} = \sum_j W_{ij}$. The **Laplacian** of G is defined as $L := D - W$ and the **Normalized Laplacian** is $\tilde{L} := D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$. The following two theorems are well known results from spectral graph theory [Chung-Graham, 1997]:

Theorem 1 (Spectrum of $\tilde{L}$). *$\tilde{L}$ is a symmetric, positive semidefinite matrix, and its eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ satisfy $0 \leq \lambda_i \leq 2$. Furthermore, the number of eigenvalues equal to zero equals to the number of disjoint components in G.*

The bound on the spectrum follows directly from Gerschgorin's Theorem.

Theorem 2 (L and $\tilde{L}$ for Regular Graphs). *Now let G be a **regular** graph of degree d, that is, a graph in which every vertex has exactly d neighbors. Then $L = d\,I - W$ and $\tilde{L} = I - \frac{1}{d} W = \frac{1}{d} L$. Finally, $W, L, \tilde{L}$ share the same eigenvectors $\{\mathbf{v}_i\}$, where $\mathbf{v}_i = \lambda_i^{-1} W \mathbf{v}_i = (d - \lambda_i)^{-1} L \mathbf{v}_i = (1 - d^{-1}\lambda_i)^{-1} \tilde{L} \mathbf{v}_i$ for all i.*

L and $\tilde{L}$ can be regarded as linear operators on functions $\boldsymbol{f} : V \mapsto \mathbb{R}$, or, equivalently, on vectors $\boldsymbol{f} = (f_1, f_2, \ldots, f_n)^\top$. We could equally well have defined L by

$$\langle \boldsymbol{f}, L\boldsymbol{f} \rangle = \boldsymbol{f}^\top L \boldsymbol{f} = -\frac{1}{2} \sum_{i \sim j} (f_i - f_j)^2 \text{ for all } \boldsymbol{f} \in \mathbb{R}^n, \tag{1}$$

which readily generalizes to graphs with a countably infinite number of vertices.

The Laplacian derives its name from its analogy with the familiar Laplacian operator $\Delta = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \ldots + \frac{\partial^2}{\partial x_m^2}$ on continuous spaces. Regarding (1) as inducing a semi-norm $\| \boldsymbol{f} \|_L = \langle \boldsymbol{f}, L\boldsymbol{f} \rangle$ on $\mathbb{R}^n$, the analogous expression for Δ defined on a compact space Ω is

$$\| f \|_\Delta = \langle f, \Delta f \rangle = \int_\Omega f\,(\Delta f)\, d\omega = \int_\Omega (\nabla f) \cdot (\nabla f)\ d\omega\,. \tag{2}$$

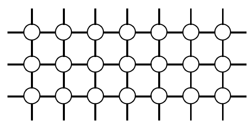
Both (1) and (2) quantify how much f and $\boldsymbol{f}$ vary locally, or how "smooth" they are over their respective domains.

More explicitly, when $\Omega = \mathbb{R}^m$, up to a constant, $-L$ is exactly the finite difference discretization of Δ on a regular lattice:

$$\Delta f(x) = \sum_{i=1}^{m} \frac{\partial^2}{\partial x_i^2} f \approx \sum_{i=1}^{m} \frac{\frac{\partial}{\partial x_i} f(x + \frac{1}{2}\boldsymbol{e}_i) - \frac{\partial}{\partial x_i} f(x - \frac{1}{2}\boldsymbol{e}_i)}{\delta}$$
$$\approx \sum_{i=1}^{m} \frac{f(x+\boldsymbol{e}_i) + f(x-\boldsymbol{e}_i) - 2f(x)}{\delta^2} =$$
$$\frac{1}{\delta^2} \sum_{i=1}^{m} (f_{x_1,\ldots,x_i+1,\ldots,x_m} + f_{x_1,\ldots,x_i-1,\ldots,x_m} - 2f_{x_1,\ldots,x_m}) = -\frac{1}{\delta^2}[L\boldsymbol{f}]_{x_1,\ldots,x_m},$$

where $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_m$ is an orthogonal basis for $\mathbb{R}^m$ normalized to $\|\,\boldsymbol{e}_i\,\| = \delta$, the vertices of the lattice are at $x = x_1\boldsymbol{e}_1 + \ldots + x_m\boldsymbol{e}_m$ with integer valued coordinates $x_i \in \mathbb{N}$, and $\boldsymbol{f}_{x_1,x_2,\ldots,x_m} = f(x)$.

Moreover, both the continuous and the discrete Laplacians are canonical operators on their respective domains, in the sense that they are invariant under certain natural transformations of the underlying space, and in this they are essentially unique.



Regular grid in two dimensions

The Laplace operator Δ is the unique self-adjoint linear second order differential operator invariant under transformations of the coordinate system under the action of the special orthogonal group SO_m, i.e. invariant under rotations. This well known result can be seen by using Schur's lemma and the fact that SO_m is irreducible on $\mathbb{R}^m$.

We now show a similar result for L. Here the permutation group plays a similar role to SO_m. We need some additional definitions: denote by S_n the group of permutations on $\{1, 2, \ldots, n\}$ with $\pi \in S_n$ being a specific permutation taking $i \in \{1, 2, \ldots n\}$ to $\pi(i)$. The so-called defining representation of S_n consists of $n \times n$ matrices Π_π, such that $[\Pi_\pi]_{i,\pi(i)=1}$ and all other entries of Π_π are zero.

Theorem 3 (Permutation Invariant Linear Functions on Graphs). *Let L be an $n \times n$ symmetric real matrix, linearly related to the $n \times n$ adjacency matrix W, i.e. $L = \mathfrak{T}[W]$ for some linear operator L in a way invariant to permutations of vertices in the sense that*

$$\Pi_\pi^\top \mathfrak{T}[W] \Pi_\pi = \mathfrak{T}\left[\Pi_\pi^\top W \Pi_\pi\right] \tag{3}$$

for any $\pi \in S_n$. Then L is related to W by a linear combination of the following three operations: identity; row/column sums; overall sum; row/column sum restricted to the diagonal of L; overall sum restricted to the diagonal of W.

Proof Let
$$L_{i_1 i_2} = \mathfrak{T}[W]_{i_1 i_2} := \sum_{i_3=1}^{n} \sum_{i_4=1}^{n} T_{i_1 i_2 i_3 i_4} W_{i_3 i_4} \tag{4}$$

with $T \in \mathbb{R}^{n^4}$. Eq. (3) then implies $T_{\pi(i_1)\pi(i_2)\pi(i_3)\pi(i_4)} = T_{i_1 i_2 i_3 i_4}$ for any $\pi \in S_n$.

The indices of T can be partitioned by the equality relation on their values, e.g. $(2,5,2,7)$ is of the partition type $[\,1\,3\,|\,2\,|\,4\,]$, since $i_1 = i_3$, but $i_2 \neq i_1$, $i_4 \neq i_1$ and $i_2 \neq i_4$. The key observation is that under the action of the permutation group, elements of T with a given index partition structure are taken to elements with the same index partition structure, e.g. if $i_1 = i_3$ then $\pi(i_1) = \pi(i_3)$ and if $i_1 \neq i_3$, then $\pi(i_1) \neq \pi(i_3)$. Furthermore, an element with a given index index partition structure can be mapped to any other element of T with the same index partition structure by a suitable choice of π.

Hence, a necessary and sufficient condition for (4) is that all elements of T of a given index partition structure be equal. Therefore, T must be a linear combination of the following tensors (i.e. multilinear forms):

$$
\begin{array}{lll}
A_{i_1 i_2 i_3 i_4} = 1 & & \\
B^{[1,2]}_{i_1 i_2 i_3 i_4} = \delta_{i_1 i_2} & B^{[1,3]}_{i_1 i_2 i_3 i_4} = \delta_{i_1 i_3} & B^{[1,4]}_{i_1 i_2 i_3 i_4} = \delta_{i_1 i_4} \\
B^{[2,3]}_{i_1 i_2 i_3 i_4} = \delta_{i_2 i_3} & B^{[2,4]}_{i_1 i_2 i_3 i_4} = \delta_{i_2 i_4} & B^{[3,4]}_{i_1 i_2 i_3 i_4} = \delta_{i_3 i_4} \\
C^{[1,2,3]}_{i_1 i_2 i_3 i_4} = \delta_{i_1 i_2}\delta_{i_2 i_3} & C^{[2,3,4]}_{i_1 i_2 i_3 i_4} = \delta_{i_2 i_3}\delta_{i_3 i_4} & \\
C^{[3,4,1]}_{i_1 i_2 i_3 i_4} = \delta_{i_3 i_4}\delta_{i_4 i_1} & C^{[4,1,2]}_{i_1 i_2 i_3 i_4} = \delta_{i_4 i_1}\delta_{i_1 i_2} & \\
D^{[1,2][3,4]}_{i_1 i_2 i_3 i_4} = \delta_{i_1 i_2}\delta_{i_3 i_4} & D^{[1,3][2,4]}_{i_1 i_2 i_3 i_4} = \delta_{i_1 i_3}\delta_{i_2 i_4} & D^{[1,4][2,3]}_{i_1 i_2 i_3 i_4} = \delta_{i_1 i_4}\delta_{i_2 i_3} \\
E^{[1,2,3,4]}_{i_1 i_2 i_3 i_4} = \delta_{i_1 i_2}\delta_{i_1 i_3}\delta_{i_1 i_4}\ . & &
\end{array}
$$

The tensor A puts the overall sum in each element of L, while $B^{[1,2]}$ returns the the same restricted to the diagonal of L.

Since W has vanishing diagonal, $B^{[3,4]}$, $C^{[2,3,4]}$, $C^{[3,4,1]}$, $D^{[1,2][3,4]}$ and $E^{[1,2,3,4]}$ produce zero. Without loss of generality we can therefore ignore them.

By symmetry of W, the pairs $(B^{[1,3]}, B^{[1,4]})$, $(B^{[2,3]}, B^{[2,4]})$, $(C^{[1,2,3]}, C^{[4,1,2]})$ have the same effect on W, hence we can set the coefficient of the second member of each to zero. Furthermore, to enforce symmetry on L, the coefficient of $B^{[1,3]}$ and $B^{[2,3]}$ must be the same (without loss of generality 1) and this will give the row/column sum matrix $(\sum_k W_{ik}) + (\sum_k W_{kl})$.

Similarly, $C^{[1,2,3]}$ and $C^{[4,1,2]}$ must have the same coefficient and this will give the row/column sum restricted to the diagonal: $\delta_{ij}\,[(\sum_k W_{ik}) + (\sum_k W_{kl})]$.

Finally, by symmetry of W, $D^{[1,3][2,4]}$ and $D^{[1,4][2,3]}$ are both equivalent to the identity map. ■

The various row/column sum and overall sum operations are uninteresting from a graph theory point of view, since they do not heed to the topology of the graph. Imposing the conditions that each row and column in L must sum to zero, we recover the graph Laplacian. Hence, up to a constant factor and trivial additive components, the graph Laplacian (or the normalized graph Laplacian if we wish to rescale by the number of edges per vertex) is the only "invariant" differential operator for given W (or its normalized counterpart $\tilde{W}$). Unless stated otherwise, all results below hold for both L and $\tilde{L}$ (albeit with a different spectrum) and we will, in the following, focus on $\tilde{L}$ due to the fact that its spectrum is contained in $[0,2]$.

3 Regularization

The fact that L induces a semi-norm on $\boldsymbol{f}$ which penalizes the changes between adjacent vertices, as described in (1), indicates that it may serve as a tool to design regularization operators.

3.1 Regularization via the Laplace Operator

We begin with a brief overview of translation invariant regularization operators on continuous spaces and show how they can be interpreted as powers of Δ. This will allow us to repeat the development almost verbatim with $\tilde{L}$ (or L) instead.

Some of the most successful regularization functionals on $\mathbb{R}^n$, leading to kernels such as the Gaussian RBF, can be written as [Smola et al., 1998]

$$\langle f, Pf \rangle := \int |\tilde{f}(\omega)|^2 \, r(\|\omega\|^2) \, d\omega = \langle f, r(\Delta) f \rangle \,. \tag{5}$$

Here $f \in L_2(\mathbb{R}^n)$, $\tilde{f}(\omega)$ denotes the Fourier transform of f, $r(\|\omega\|^2)$ is a function penalizing frequency components $|\tilde{f}(\omega)|$ of f, typically increasing in $\|\omega\|^2$, and finally, $r(\Delta)$ is the extension of r to operators simply by applying r to the spectrum of Δ [Dunford and Schwartz, 1958]

$$\langle f, r(\Delta) f' \rangle = \sum_i \langle f, \psi_i \rangle \, r(\lambda_i) \, \langle \psi_i, f' \rangle$$

where $\{(\psi_i, \lambda_i)\}$ is the eigensystem of Δ. The last equality in (5) holds because applications of Δ become multiplications by $\|\omega\|^2$ in Fourier space. Kernels are obtained by solving the self-consistency condition [Smola et al., 1998]

$$\langle k(x, \cdot), Pk(x', \cdot) \rangle = k(x, x') \,. \tag{6}$$

One can show that $k(x, x') = \kappa(x - x')$, where κ is equal to the inverse Fourier transform of $r^{-1}(\|\omega\|^2)$. Several r functions have been known to yield good results. The two most popular are given below:

	$r(\|\omega\|^2)$	$k(x, x')$	$r(\Delta)$
Gaussian RBF	$\exp\left(\frac{\sigma^2}{2}\|\omega\|^2\right)$	$\exp\left(-\frac{1}{2\sigma^2}\|x - x'\|^2\right)$	$\sum_{i=0}^{\infty} \frac{\sigma^{2i}}{i!} \Delta^i$
Laplacian RBF	$1 + \sigma^2 \|\omega\|^2$	$\exp\left(-\frac{1}{\sigma}\|x - x'\|\right)$	$1 + \sigma^2 \Delta$

In summary, regularization according to (5) is carried out by penalizing $\tilde{f}(\omega)$ by a function of the Laplace operator. For many results in regularization theory one requires $r(\|\omega\|^2) \to \infty$ for $\|\omega\|^2 \to \infty$.

3.2 Regularization via the Graph Laplacian

In complete analogy to (5), we define a class of regularization functionals on graphs as

$$\langle \boldsymbol{f}, P\boldsymbol{f} \rangle := \langle \boldsymbol{f}, r(\tilde{L}) \boldsymbol{f} \rangle \,. \tag{7}$$

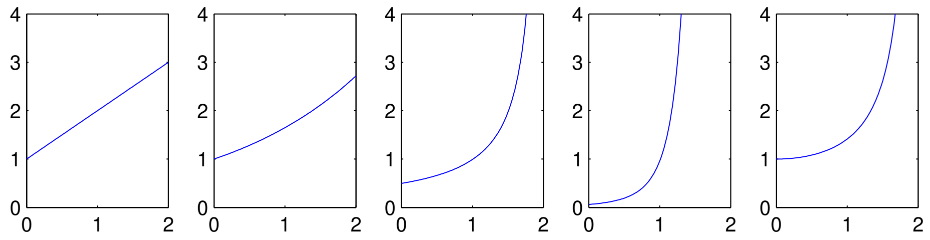


Fig. 1. Regularization function $r(\lambda)$. From left to right: regularized Laplacian ($\sigma^2 = 1$), diffusion process ($\sigma^2 = 1$), one-step random walk ($a = 2$), 4-step random walk ($a = 2$), inverse cosine.

Here $r(\tilde{L})$ is understood as applying the scalar valued function $r(\lambda)$ to the eigenvalues of $\tilde{L}$, that is,

$$r(\tilde{L}) := \sum_{i=1}^{m} r(\lambda_i)\, \boldsymbol{v}_i \boldsymbol{v}_i^\top, \tag{8}$$

where $\{(\lambda_i, \boldsymbol{v}_i)\}$ constitute the eigensystem of $\tilde{L}$. The normalized graph Laplacian $\tilde{L}$ is preferable to L, since $\tilde{L}$'s spectrum is contained in $[0, 2]$. The obvious goal is to gain insight into what functions are appropriate choices for r.

- From (1) we infer that $\boldsymbol{v}_i$ with large λ_i correspond to rather uneven functions on the graph G. Consequently, they should be penalized more strongly than $\boldsymbol{v}_i$ with small λ_i. Hence $r(\lambda)$ should be monotonically increasing in λ.
- Requiring that $r(\tilde{L}) \succeq 0$ imposes the constraint $r(\lambda) \geq 0$ for all $\lambda \in [0, 2]$.
- Finally, we can limit ourselves to $r(\lambda)$ expressible as power series, since the latter are dense in the space of C_0 functions on bounded domains.

In Section 3.5 we will present additional motivation for the choice of $r(\lambda)$ in the context of spectral graph theory and segmentation. As we shall see, the following functions are of particular interest:

$$r(\lambda) = 1 + \sigma^2 \lambda \quad \text{(Regularized Laplacian)} \tag{9}$$
$$r(\lambda) = \exp\left(\sigma^2/2\lambda\right) \quad \text{(Diffusion Process)} \tag{10}$$
$$r(\lambda) = (aI - \lambda)^{-1} \text{ with } a \geq 2 \quad \text{(One-Step Random Walk)} \tag{11}$$
$$r(\lambda) = (aI - \lambda)^{-p} \text{ with } a \geq 2 \quad \text{(p-Step Random Walk)} \tag{12}$$
$$r(\lambda) = (\cos \lambda\pi/4)^{-1} \quad \text{(Inverse Cosine)} \tag{13}$$

Figure 1 shows the regularization behavior for the functions (9)-(13).

3.3 Kernels

The introduction of a regularization matrix $P = r(\tilde{L})$ allows us to define a Hilbert space $\mathcal{H}$ on $\mathbb{R}^m$ via $\langle f, f \rangle_{\mathcal{H}} := \langle \boldsymbol{f}, P\boldsymbol{f} \rangle$. We now show that $\mathcal{H}$ is a reproducing kernel Hilbert space.

Theorem 4. *Denote by $P \in \mathbb{R}^{m \times m}$ a (positive semidefinite) regularization matrix and denote by $\mathcal{H}$ the image of $\mathbb{R}^m$ under P. Then $\mathcal{H}$ with dot product $\langle f, f \rangle_{\mathcal{H}} := \langle \boldsymbol{f}, P\boldsymbol{f} \rangle$ is a Reproducing Kernel Hilbert Space and its kernel is $k(i,j) = \left[P^{-1}\right]_{ij}$, where P^{-1} denotes the pseudo-inverse if P is not invertible.*

Proof Since P is a positive semidefinite matrix, we clearly have a Hilbert space on $P\mathbb{R}^m$. To show the reproducing property we need to prove that

$$f(i) = \langle f, k(i, \cdot) \rangle_{\mathcal{H}}. \tag{14}$$

Note that $k(i,j)$ can take on at most m^2 different values (since $i,j \in [1:m]$). In matrix notation (14) means that for all $\boldsymbol{f} \in \mathcal{H}$

$$f(i) = \boldsymbol{f}^\top P K_{i,:} \text{ for all } i \iff \boldsymbol{f}^\top = \boldsymbol{f}^\top P K. \tag{15}$$

The latter holds if $K = P^{-1}$ and $\boldsymbol{f} \in P\mathbb{R}^m$, which proves the claim. ■

In other words, K is the Greens function of P, just as in the continuous case. The notion of Greens functions on graphs was only recently introduced by Chung-Graham and Yau [2000] for L. The above theorem extended this idea to arbitrary regularization operators $\hat{r}(\tilde{L})$.

Corollary 1. *Denote by $P = r(\tilde{L})$ a regularization matrix, then the corresponding kernel is given by $K = r^{-1}(\tilde{L})$, where we take the pseudo-inverse wherever necessary. More specifically, if $\{(\boldsymbol{v}_i, \lambda_i)\}$ constitute the eigensystem of $\tilde{L}$, we have*

$$K = \sum_{i=1}^{m} r^{-1}(\lambda_i)\, \boldsymbol{v}_i \boldsymbol{v}_i^\top \text{ where we define } 0^{-1} \equiv 0. \tag{16}$$

3.4 Examples of Kernels

By virtue of Corollary 1 we only need to take (9)-(13) and plug the definition of $r(\lambda)$ into (16) to obtain formulae for computing K. This yields the following kernel matrices:

$$K = (I + \sigma^2 \tilde{L})^{-1} \quad \text{(Regularized Laplacian)} \tag{17}$$

$$K = \exp(-\sigma^2/2\tilde{L}) \quad \text{(Diffusion Process)} \tag{18}$$

$$K = (aI - \tilde{L})^p \text{ with } a \geq 2 \quad \text{(p-Step Random Walk)} \tag{19}$$

$$K = \cos \tilde{L}\pi/4 \quad \text{(Inverse Cosine)} \tag{20}$$

Equation (18) corresponds to the diffusion kernel proposed by Kondor and Lafferty [2002], for which $K(x, x')$ can be visualized as the quantity of some substance that would accumulate at vertex x' after a given amount of time if we injected the substance at vertex x and let it diffuse through the graph along the edges. Note that this involves matrix exponentiation defined via the limit $K = \exp(B) = \lim_{n\to\infty}(I+B/n)^n$ as opposed to component-wise exponentiation $K_{i,j} = \exp(B_{i,j})$.

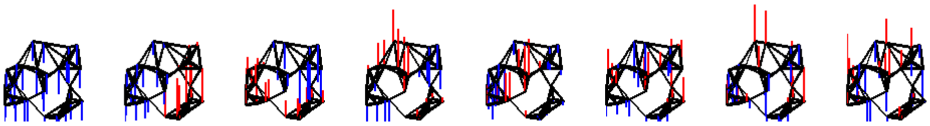
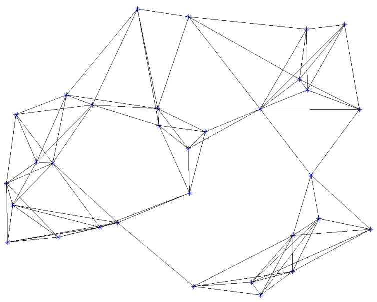


Fig. 2. The first 8 eigenvectors of the normalized graph Laplacian corresponding to the graph drawn above. Each line attached to a vertex is proportional to the value of the corresponding eigenvector at the vertex. Positive values (red) point up and negative values (blue) point down. Note that the assignment of values becomes less and less uniform with increasing eigenvalue (i.e. from left to right).

For (17) it is typically more efficient to deal with the inverse of K, as it avoids the costly inversion of the sparse matrix $\tilde{L}$. Such situations arise, e.g., in Gaussian Process estimation, where K is the covariance matrix of a stochastic process [Williams, 1999].

Regarding (19), recall that $(aI - \tilde{L})^p = ((a-1)I + \tilde{W})^p$ is up to scaling terms equivalent to a p-step random walk on the graph with random restarts (see Section A for details). In this sense it is similar to the diffusion kernel. However, the fact that K involves only a finite number of products of matrices makes it much more attractive for practical purposes. In particular, entries in K_{ij} can be computed cheaply using the fact that $\tilde{L}$ is a sparse matrix.



A nearest neighbor graph.

Finally, the inverse cosine kernel treats lower complexity functions almost equally, with a significant reduction in the upper end of the spectrum. Figure 2 shows the leading eigenvectors of the graph drawn above and Figure 3 provide examples of some of the kernels discussed above.

3.5 Clustering and Spectral Graph Theory

We could also have derived $r(\tilde{L})$ directly from spectral graph theory: the eigenvectors of the graph Laplacian correspond to functions partitioning the graph into clusters, see e.g., [Chung-Graham, 1997, Shi and Malik, 1997] and the references therein. In general, small eigenvalues have associated eigenvectors which vary little between adjacent vertices. Finding the smallest eigenvectors of $\tilde{L}$ can be seen as a real-valued relaxation of the min-cut problem.[3]

For instance, the smallest eigenvalue of $\tilde{L}$ is 0, its corresponding eigenvector is $D^{\frac{1}{2}} 1_n$ with $1_n := (1, \ldots, 1) \in \mathbb{R}^n$. The second smallest eigenvalue/eigenvector pair, also often referred to as the **Fiedler**-vector, can be used to split the graph

---

[3] Only recently, algorithms based on the celebrated semidefinite relaxation of the min-cut problem by Goemans and Williamson [1995] have seen wider use [Torr, 2003] in segmentation and clustering by use of spectral bundle methods.

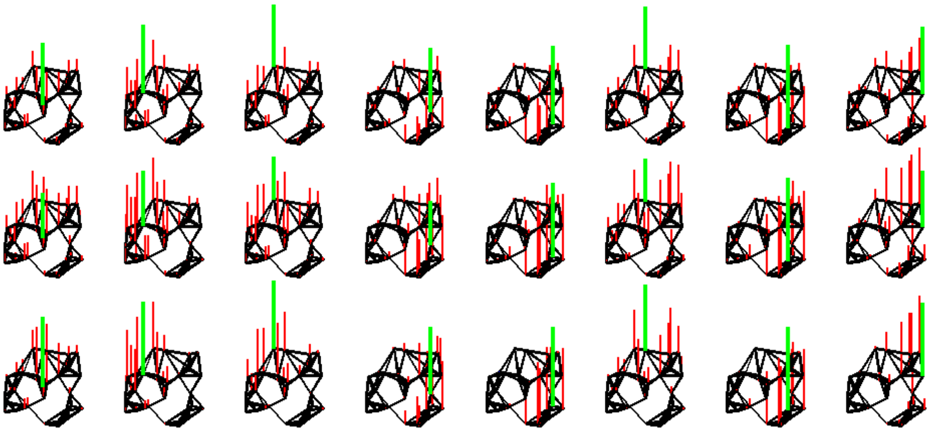


Fig. 3. Top: regularized graph Laplacian; Middle: diffusion kernel with $\sigma = 5$, Bottom: 4-step random walk kernel. Each figure displays K_{ij} for fixed i. The value K_{ij} at vertex i is denoted by a bold line. Note that only adjacent vertices to i bear significant value.

into two distinct parts [Weiss, 1999, Shi and Malik, 1997], and further eigenvectors with larger eigenvalues have been used for more finely-grained partitions of the graph. See Figure 2 for an example.

Such a decomposition into functions of increasing complexity has very desirable properties: if we want to perform estimation on the graph, we will wish to bias the estimate towards functions which vary little over large homogeneous portions [4]. Consequently, we have the following interpretation of $\langle f, f\rangle_{\mathcal{H}}$. Assume that $\boldsymbol{f} = \sum_i \beta_i \boldsymbol{v}_i$, where $\{(\boldsymbol{v}_i, \lambda_i)\}$ is the eigensystem of $\tilde{L}$. Then we can rewrite $\langle f, f\rangle_{\mathcal{H}}$ to yield

$$\langle \boldsymbol{f}, r(\tilde{L})\boldsymbol{f}\rangle = \left\langle \sum_i \beta_i \boldsymbol{v}_i, \sum_j r(\lambda_j)\boldsymbol{v}_j \boldsymbol{v}_j^\top \sum_l \beta_l \boldsymbol{v}_l \right\rangle = \sum_i \beta_i^2 r(\lambda_i). \qquad (21)$$

This means that the components of f which vary a lot over coherent clusters in the graph are penalized more strongly, whereas the portions of f, which are essentially constant over clusters, are preferred. This is exactly what we want.

3.6 Approximate Computation

Often it is not necessary to know all values of the kernel (e.g., if we only observe instances from a subset of all positions on the graph). There it would be wasteful to compute the full matrix $r(L)^{-1}$ explicitly, since such operations typically scale with $O(n^3)$. Furthermore, for large n it is not desirable to compute K via (16), that is, by computing the eigensystem of $\tilde{L}$ and assembling K directly.

[4] If we cannot assume a connection between the structure of the graph and the values of the function to be estimated on it, the entire concept of designing kernels on graphs obviously becomes meaningless.

Instead, we would like to take advantage of the fact that $\tilde{L}$ is sparse, and consequently any operation $\tilde{L}\alpha$ has cost at most linear in the number of nonzero elements of $\tilde{L}$, hence the cost is bounded by $O(|E|+n)$. Moreover, if d is the largest degree of the graph, then computing $L^p \boldsymbol{e}_i$ costs at most $|E| \sum_{i=1}^{p-1} (\min(d+1, n))^i$ operations: at each step the number of non-zeros in the rhs decreases by at most a factor of $d+1$. This means that as long as we can approximate $K = r^{-1}(\tilde{L})$ by a low order polynomial, say $\rho(\tilde{L}) := \sum_{i=0}^{N} \beta_i \tilde{L}^i$, significant savings are possible.

Note that we need not necessarily require a uniformly good approximation and put the main emphasis on the approximation for small λ. However, we need to ensure that $\rho(\tilde{L})$ is positive semidefinite.

Diffusion Kernel: The fact that the series $r^{-1}(x) = \exp(-\beta x) = \sum_{m=0}^{\infty} (-\beta)^m \frac{x^m}{m!}$ has alternating signs shows that the approximation error at $r^{-1}(x)$ is bounded by $\frac{(2\beta)^{N+1}}{(N+1)!}$, if we use N terms in the expansion (from Theorem 1 we know that $\|\tilde{L}\| \leq 2$). For instance, for $\beta = 1$, 10 terms are sufficient to obtain an error of the order of 10^{-4}.

Variational Approximation: In general, if we want to approximate $r^{-1}(\lambda)$ on $[0, 2]$, we need to solve the $L_\infty([0, 2])$ approximation problem

$$\underset{\beta,\epsilon}{\operatorname{minimize}} \ \epsilon \ \text{ subject to } \left| \sum_{i=0}^{N} \beta_i \lambda^i - r^{-1}(\lambda) \right| \leq \epsilon \quad \forall \lambda \in [0, 2] \tag{22}$$

Clearly, (22) is equivalent to minimizing $\sup_{\tilde{L}} \|\rho(\tilde{L}) - r^{-1}(\tilde{L})\|$, since the matrix norm is determined by the largest eigenvalues, and we can find $\tilde{L}$ such that the discrepancy between $\rho(\lambda)$ and $r^{-1}(\lambda)$ is attained. Variational problems of this form have been studied in the literature, and their solution may provide much better approximations to $r^{-1}(\lambda)$ than a truncated power series expansion.

4 Products of Graphs

As we have already pointed out, it is very expensive to compute K for arbitrary $\hat{r}$ and $\tilde{L}$. For special types of graphs and regularization, however, significant computational savings can be made.

4.1 Factor Graphs

The work of this section is a direct extension of results by Ellis [2002] and Chung-Graham and Yau [2000], who study factor graphs to compute inverses of the graph Laplacian.

Definition 1 (Factor Graphs). *Denote by (V, E) and (V', E') the vertices V and edges E of two graphs, then the factor graph $(V_f, E_f) := (V, E) \otimes (V', E')$ is defined as the graph where $(i, i') \in V_f$ if $i \in V$ and $i' \in V'$; and $((i, i'), (j, j')) \in E_f$ if and only if either $(i, j) \in E$ and $i' = j'$ or $(i', j') \in E'$ and $i = j$.*

For instance, the factor graph of two rings is a torus. The nice property of factor graphs is that we can compute the eigenvalues of the Laplacian on products very easily (see e.g., Chung-Graham and Yau [2000]):

Theorem 5 (Eigenvalues of Factor Graphs). *The eigenvalues and eigenvectors of the normalized Laplacian for the factor graph between a regular graph of degree d with eigenvalues $\{\lambda_j\}$ and a regular graph of degree d' with eigenvalues $\{\lambda'_l\}$ are of the form:*

$$\lambda_{j,l}^{\text{fact}} = \frac{d}{d+d'}\lambda_j + \frac{d'}{d+d'}\lambda'_l \tag{23}$$

and the eigenvectors satisfy $e^{j,l}_{(i,i')} = e^j_i e'^l_{i'}$, where e^j is an eigenvector of $\tilde{L}$ and e'^l is an eigenvector of $\tilde{L}'$.

This allows us to apply Corollary 1 to obtain an expansion of K as

$$K = (r(L))^{-1} = \sum_{j,l} r^{-1}(\lambda_{jl})\, e^{j,l} \left(e^{j,l}\right)^\top. \tag{24}$$

While providing an explicit recipe for the computation of K_{ij} without the need to compute the full matrix K, this still requires $O(n^2)$ operations per entry, which may be more costly than what we want (here n is the number of vertices of the factor graph).

Two methods for computing (24) become evident at this point: if r has a special structure, we may exploit this to decompose K into the products and sums of terms depending on one of the two graphs alone and pre-compute these expressions beforehand. Secondly, if one of the two terms in the expansion can be computed for a rather general class of values of $r(x)$, we can pre-compute this expansion and only carry out the remainder corresponding to (24) explicitly.

4.2 Product Decomposition of $r(x)$

Central to our reasoning is the observation that for certain $r(x)$, the term $\frac{1}{r(a+b)}$ can be expressed in terms of a product and sum of terms depending on a and b only. We assume that

$$\frac{1}{r(a+b)} = \sum_{m=1}^{M} \rho_n(a)\tilde{\rho}_n(b). \tag{25}$$

In the following we will show that in such situations the kernels on factor graphs can be computed as an analogous combination of products and sums of kernel functions on the terms constituting the ingredients of the factor graph. Before we do so, we briefly check that many $r(x)$ indeed satisfy this property.

$$\exp(-\beta(a+b)) = \exp(-\beta a)\exp(-\beta b) \tag{26}$$

$$(A-(a+b)) = \left(\frac{A}{2}-a\right) + \left(\frac{A}{2}-b\right) \tag{27}$$

$$(A-(a+b))^p = \sum_{n=0}^{p} \binom{p}{n} \left(\frac{A}{2}-a\right)^n \left(\frac{A}{2}-b\right)^{p-n} \tag{28}$$

$$\cos\frac{(a+b)\pi}{4} = \cos\frac{a\pi}{4}\cos\frac{b\pi}{4} - \sin\frac{a\pi}{4}\sin\frac{b\pi}{4} \tag{29}$$

In a nutshell, we will exploit the fact that for products of graphs the eigenvalues of the joint graph Laplacian can be written as the sum of the eigenvalues of the Laplacians of the constituent graphs. This way we can perform computations on ρ_n and $\tilde{\rho}_n$ separately without the need to take the other part of the the product of graphs into account. Define

$$k_m(i,j) := \sum_l \rho_l \left(\frac{d\lambda_l}{d+d'}\right) e_i^l e_j^l \text{ and } \tilde{k}_m(i',j') := \sum_l \tilde{\rho}_l \left(\frac{d\lambda_l}{d+d'}\right) \tilde{e}_{i'}^l \tilde{e}_{j'}^l. \quad (30)$$

Then we have the following composition theorem:

Theorem 6. *Denote by (V,E) and (V',E') connected regular graphs of degrees d with m vertices (and d', m' respectively) and normalized graph Laplacians $\tilde{L}, \tilde{L}'$. Furthermore denote by $r(x)$ a rational function with matrix-valued extension $\hat{r}(X)$. In this case the kernel K corresponding to the regularization operator $\hat{r}(L)$ on the product graph of (V,E) and (V',E') is given by*

$$k((i,i'),(j,j')) = \sum_{m=1}^{M} k_m(i,j)\tilde{k}_m(i',j') \quad (31)$$

Proof Plug the expansion of $\frac{1}{r(a+b)}$ as given by (25) into (24) and collect terms. ■

From (26) we immediately obtain the corollary (see Kondor and Lafferty [2002]) that for diffusion processes on factor graphs the kernel on the factor graph is given by the product of kernels on the constituents, that is $k((i,i'),(j,j')) = k(i,j)k'(i',j')$.

The kernels k_m and $\tilde{k}_m$ can be computed either by using an analytic solution of the underlying factors of the graph or alternatively they can be computed numerically. If the total number of kernels k_n is small in comparison to the number of possible coordinates this is still computationally beneficial.

4.3 Composition Theorems

If no expansion as in (31) can be found, we may still be able to compute kernels by extending a reasoning from [Ellis, 2002]. More specifically, the following composition theorem allows us to accelerate the computation in many cases, whenever we can parameterize $(\hat{r}(L+\alpha I))^{-1}$ in an efficient way. For this purpose we introduce two auxiliary functions

$$K_\alpha(i,j) := \left(\hat{r}\left(\frac{d}{d+d'}L + \frac{\alpha d'}{d+d'}I\right)\right)^{-1} = \sum_l \left(r\left(\frac{d\lambda_l + \alpha d'}{d+d'}\right)\right)^{-1} e_l(i)e_l(j)$$

$$G'_\alpha(i,j) := (L' + \alpha I)^{-1} = \sum_l \frac{1}{\lambda_l + \alpha} e_l(i)e_l(j). \quad (32)$$

In some cases $K_\alpha(i,j)$ may be computed in closed form, thus obviating the need to perform expensive matrix inversion, e.g., in the case where the underlying graph is a chain [Ellis, 2002] and $K_\alpha = G_\alpha$.

Theorem 7. *Under the assumptions of Theorem 6 we have*

$$K((j,j'),(l,l')) = \frac{1}{2\pi i}\int_C K_\alpha(j,l)G'_{-\alpha}(j',l')d\alpha = \sum_v K_{\lambda_v}(j,l)e^v_{j'}e^v_{l'} \tag{33}$$

where $C \subset \mathbb{C}$ is a contour of the $\mathbb{C}$ containing the poles of (V',E') including 0.

For practical purposes, the third term of (33) is more amenable to computation.
Proof From (24) we have

$$\begin{aligned} K((j,j'),(l,l')) &= \sum_{u,v}\left(r\left(\frac{d\lambda_u + d'\lambda_v}{d+d'}\right)\right)^{-1} e^u_j e^u_l e^v_{j'} e^v_{l'} \\ &= \frac{1}{2\pi i}\int_C \sum_u \left(r\left(\frac{d\lambda_u + d'\alpha}{d+d'}\right)\right)^{-1} e^u_j e^u_l \sum_v \frac{1}{\lambda_v - \alpha} e^v_{j'} e^v_{l'} d\alpha \end{aligned} \tag{34}$$

Here the second equality follows from the fact that the contour integral over a pole p yields $\int_C \frac{f(\alpha)}{p-\alpha}d\alpha = 2\pi i f(p)$, and the claim is verified by checking the definitions of K_α and G'_α. The last equality can be seen from (34) by splitting up the summation over u and v. ■

5 Conclusions

We have shown that the canonical family of kernels on graphs are of the form of power series in the graph Laplacian. Equivalently, such kernels can be characterized by a real valued function of the eigenvalues of the Laplacian. Special cases include diffusion kernels, the regularized Laplacian kernel and p-step random walk kernels. We have developed the regularization theory of learning on graphs using such kernels and explored methods for efficiently computing and approximating the kernel matrix.

Acknowledgments This work was supported by a grant of the ARC. The authors thank Eleazar Eskin, Patrick Haffner, Andrew Ng, Bob Williamson and S.V.N. Vishwanathan for helpful comments and suggestions.

A Link Analysis

Rather surprisingly, our approach to regularizing functions on graphs bears resemblance to algorithms for scoring web pages such as PageRank [Page et al., 1998], HITS [Kleinberg, 1999], and randomized HITS [Zheng et al., 2001]. More specifically, the random walks on graphs used in all three algorithms and the stationary distributions arising from them are closely connected with the eigensystem of L and $\tilde{L}$ respectively.

We begin with an analysis of PageRank. Given a set of web pages and links between them we construct a directed graph in such a way that pages correspond to vertices and edges correspond to links, resulting in the (nonsymmetric) matrix W. Next we consider the random walk arising from following each of the links with equal probability in addition to a random restart at an arbitrary vertex with probability ϵ. This means that the probability distribution over states follows the discrete time evolution equation

$$\boldsymbol{p}(t+1) = \left[\epsilon I + (1-\epsilon) W D^{-1}\right] \boldsymbol{p}(t) \tag{35}$$

where D is a diagonal matrix with $D_{ii} = \sum_j W_{ij}$ and $\boldsymbol{p}$ is the vector of probabilities of being on a certain page. The PageRank is then determined from the stationary distribution of $\boldsymbol{p}$. Clearly the largest eigenvalue/eigenvector pair of $\left[\epsilon I + (1-\epsilon) W D^{-1}\right]$ will determine the stationary distribution $\boldsymbol{p}(\infty)$, and the contribution of the other eigenvectors decays geometrically (one may conjecture that in practice only few iterations are needed).

Now consider the same formalism in the context of a 1-step random walk (11): here one computes $aI - \tilde{L} = (a-1)I + D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$. Rescaling by $\frac{1}{a}$ and setting $\epsilon = \frac{1-a}{a}$ yields a matrix with the same spectrum as the linear difference equation (35). Furthermore, for all eigenvectors $\boldsymbol{v}_i$ of $\epsilon I + (1-\epsilon) W D^{-1}$ we can find eigenvectors of $aI - \tilde{L}$ of the form $D^{-\frac{1}{2}} \boldsymbol{v}_i$.

The main difference, however, is that while graphs arising from web pages are directed (following the direction of the link), which leads to asymmetric W, the graphs we studied in this paper are all undirected, leading to symmetric W and $L, \tilde{L}$. We can now view the assignment of a certain PageRank to a page, as achieved via the stationary distribution of the random walk, as a means of finding a "simple" function on the graph of web pages.

In HITS [Kleinberg, 1999] one uses the concept of hubs and authorities to obtain a ranking between web pages. Given the graph G, as represented by W, one seeks to find the largest eigenvalue of the matrix $M := \begin{bmatrix} 0 & W \\ W^\top & 0 \end{bmatrix}$, which can be shown to be equivalent to finding singular value decomposition of W [Zheng et al., 2001] (the latter is also used if we wish to perform latent semantic indexing on the matrix W). More specifically, with $\{\boldsymbol{v}_i, \lambda_i\}$ being the eigensystem of $WW^\top$ (we assume that the eigenvalues are sorted in increasing order), one uses $\boldsymbol{v}_{mj}^2$ as the weight of page j.

This setting was modified by Zheng et al. [2001] to accommodate for a larger subspace (Subspace HITS), which renders the system more robust with respect to small perturbations. More specifically, they use $\sum_{i=1}^m g(\lambda_i) \boldsymbol{v}_{ij}^2$ for some monotonically increasing function $g(\lambda)$ to assess the relevance of page j. The latter, however, is identical to the diagonal entry of $g(W)$. Note the similarity to 7, where we used an essentially rescaled version of W to determine the complexity of the functions under consideration. More specifically, if for regular graphs of order d we set $g(\lambda) = \frac{1}{r(1-\lambda/d)}$ we can see that the HITS rank assigned to pages j is simply the "length" of the corresponding page in "feature space" as given by K_{ii}. In other words, pages with a high HITS rank correspond to unit vectors which are considered simple with respect to the regularizer induced by the underlying graph.

Bibliography

M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. Technical Report TR-2002-01, The University of Chichago, January 2002.

F. Chung-Graham. *Spectral Graph Theory.* Number 92 in CBMS Regional Conference Series in Mathematics. AMS, 1997.

F. Chung-Graham and S. T. Yau. Discrete green's functions. *Journal of Combinatorial Theory*, 91:191–214, 2000.

N. Dunford and J. Schwartz. *Linear operators.* Pure and applied mathematics, v. 7. Interscience Publishers, New York, 1958.

R. Ellis. Discrete green's functions for products of regular graphs. Technical report, University of California at San Diego, 2002. Preliminary Report.

M.X. Goemans and D.P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM*, 42(6):1115–1145, 1995.

J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, November 1999.

R. S. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete structures. In *Proceedings of the ICML*, 2002.

L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, Stanford University, Stanford, CA, USA, November 1998.

J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Conf. Computer Vision and Pattern Recognition*, June 1997.

A. Smola, B. Schölkopf, and K.-R. Müller. The connection between regularization operators and support vector kernels. *Neural Networks*, 11:637–649, 1998.

P.H.S. Torr. Solving Markov random fields using semidefinite programming. In *Artificial Intelligence and Statistics AISTATS*, 2003.

Y. Weiss. Segmentation using eigenvectors: A unifying view. In *International Conference on Computer Vision ICCV*, pages 975–982, 1999.

C. K. I. Williams. Prediction with Gaussian processes: From linear regression to linear prediction and beyond. In Micheal Jordan, editor, *Learning and Inference in Graphical Models*, pages 599–621. MIT Press, 1999.

A. Zheng, A. Ng, and M. Jordan. Stable eigenvector algorithms for link analysis. In W. Croft, D. Harper, D. Kraft, and J. Zobel, editors, *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 258–266, New York, 2001. ACM Press.