Implementing regularization implicitly via approximate eigenvector computation

Michael W. Mahoney * Lorenzo Orecchia †

Abstract

Regularization is a powerful technique for extracting useful information from noisy data. Typically, it is implemented by adding some sort of norm constraint to an objective function and then exactly optimizing the modified objective function. This procedure typically leads to optimization problems that are computationally more expensive than the original problem, a fact that is clearly problematic if one is interested in large-scale applications. On the other hand, a large body of empirical work has demonstrated that heuristics, and in some cases approximation algorithms, developed to speed up computations sometimes have the side-effect of performing regularization implicitly. Thus, we consider the question: What is the regularized optimization objective that an approximation algorithm is exactly optimizing?

We address this question in the context of computing approximations to the smallest non-trivial eigenvector of a graph Laplacian; and we consider three random-walk-based procedures: one based on the heat kernel of the graph, one based on computing the the PageRank vector associated with the graph, and one based on a truncated lazy random walk. In each case, we provide a precise characterization of the manner in which the approximation method can be viewed as implicitly computing the exact solution to a regularized problem. Interestingly, the regularization is not on the usual vector form of the optimization problem, but instead it is on a related semidefinite program.

1 Introduction

Regularization is a powerful technique in statistics, machine learning, and data analysis for learning from or extracting useful information from noisy data [14, 6, 4]. It involves (explicitly or implicitly) making assumptions about the data in order to obtain a "smoother" or "nicer" solution to a problem of interest. The technique originated in integral equation theory, where it was of interest to give meaningful solutions to ill-posed problems for which a solution did not exist [22]. More recently, it has achieved widespread use in statistical data analysis, where it is of interest to achieve solutions that generalize well to unseen data [9]. For instance, much of the work in kernel-based and manifold-based machine learning is based on regularization in Reproducing kernel Hilbert spaces [19].

Typically, regularization is implemented via a two step process: first, add some sort of norm constraint to an objective function of interest; and then, exactly optimize the modified objective function. For instance, one typically considers a loss function $f(x)$ that specifies an empirical penalty depending on both the data and a parameter vector x; and a regularization function $g(x)$ that encodes prior assumptions about the data and that provides capacity control on the vector x. Then, one must solve an optimization problem of the form:

$$\hat{x} = \text{argmin}_x f(x) + \lambda g(x). \tag{1}$$

*Department of Mathematics, Stanford University, Stanford, CA 94305, mmahoney@cs.stanford.edu.

†Computer Science Division, UC Berkeley, Berkeley, CA, 94720 orecchia@eecs.berkeley.edu.

A general feature of regularization implemented in this manner is that, although one obtains solutions that are "better" (in some statistical sense) than the solution to the original problem, one must solve a modified optimization problem that is "worse" (in the sense of being more computationally expensive) than than the original optimization problem.[1] Clearly, this algorithmic-statistical tradeoff is problematic if one is interested in large-scale applications.

On the other hand, it is well-known amongst practitioners that certain heuristics that can be used to speed up computations can sometimes have the side-effect of performing smoothing or regularization implicitly. For example, "early stopping" is often used when a learning model such as a neural network is trained by an iterative gradient descent algorithm; and "binning" is often used to aggregate the data into bins, upon which computations are performed. As we will discuss below, we have also observed a similar phenomenon in the empirical analysis of very large social and information networks [12]. In these applications, the size-scale of the networks renders prohibitive anything but very fast nearly-linear-time algorithms, but the sparsity and noise properties of the networks are sufficiently complex that there is a need to understand the statistical properties *implicit* in these fast algorithms in order to draw meaningful domain-specific conclusions from their output.

Motivated by these observations, we are interested in understanding in greater detail the manner in which algorithms that have superior algorithmic and computational properties either do or do not also have superior statistical properties. In particular, we would like to know:

- To what extent can one formalize the idea that performing an approximate computation can *implicitly* lead to more regular solutions?

Rather than addressing this question in full generality, in this paper we will address it in the context of computing the first nontrivial eigenvector of the graph Laplacian. (Of course, even this special case is of interest since a large body of work in machine learning, data analysis, computer vision, and scientific computation makes use of this vector.) Our main result is a characterization of this implicit regularization in the context of three random-walk-based procedures for computing an approximation to this eigenvector. In particular:

- We consider three random-walk-based procedures—one based on the heat kernel of the graph, one based on computing the the PageRank vector associated with the graph, and one based on a truncated lazy random walk—for computing an approximation to the smallest nontrivial eigenvector of a graph Laplacian, and we show that these approximation procedures may be viewed as implicitly solving a regularized optimization problem exactly.

Interestingly, in order to achieve this identification, we need to relax the standard spectral optimization problem to a semidefinite program. Thus, the variables that enter into the loss function and the regularization term are not unit vectors, as they are more typically in formulations such as Problem (1), but instead they are distributions over unit vectors. This was somewhat unexpected, and the empirical implications of this remain to be explored.

Before proceeding, let us pause to gain an intuition of our results in a relatively simple setting. To do so, consider the so-called Power Iteration Method, which takes as input an $n \times n$ symmetric matrix A and returns as output a number λ (the eigenvalue) and a vector v (the eigenvector)

[1] Perhaps the simplest case is that of ridge regression, which is commonly-used when the matrix A in a linear least-squares objective function is underconstrained or very ill-condtioned. In this case, $f(x) = ||A\hat{x} - b||_2$; $g(x) = ||x||_2^2$ corresponds to making a Gaussian-like assumption on the data; and the solution $\hat{x} = (A^T A + \lambda I)^{-1} A^T b$ can be written in closed form as a slightly larger least-squares optimization problem. More generally, however, even assuming that $f(x)$ and $g(x)$ are convex, one obtains a linear program or convex program that must solved.

such that $Av = \lambda v$.[2] The Power Iteration Method starts with an initial random vector, call it ν_0, and it iteratively computes $\nu_{t+1} = A\nu_t/||A\nu_t||_2$. Under weak assumptions, the method converges to v_1, the dominant eigenvector of A. The reason is clear: if we expand $\nu_0 = \sum_{i=1}^n \gamma_i v_i$ in the basis provided by the eigenfunctions $\{v_i\}_{i=1}^n$ of A, then $\nu_t = \sum_{i=1}^n \gamma_i^t v_i \to v_1$. If we truncate this method after some very small number, say 3, iterations, then the output vector is clearly a suboptimal approximation of the dominant eigen-direction of the particular matrix A; but due to the admixing of information from the other eigenvectors, it may be a better or more robust approximation to the best "ground truth eigen-direction" in the ensemble from which A was drawn. It is this intuition in the context of computing eigenvectors of the graph Laplacian that our main results formalize.

2 Overview of the problem and approximation procedures

For a connected, weighted, undirected graph $G = (V, E)$, let A be its adjacency matrix and D its diagonal degree matrix, *i.e.*, $D_{ii} = \sum_{j:(ij)\in E} w_{ij}$. Let $M = AD^{-1}$ be the natural random walk transition matrix associated with G, in which case $W = (I + M)/2$ is the usual lazy random walk transition matrix. (Thus, we will be *post*-multiplying by *column* vectors.) Finally, let $L = I - D^{-1/2}AD^{-1/2}$ be the normalized Laplacian of G.

We start by considering the standard spectral optimization problem, corresponding to the generalized eigenvector problem $Lx = \lambda Dx$.

$$\begin{aligned} \mathsf{SPECTRAL}: \quad \min \quad & x^T L x \\ \text{s.t.} \quad & x^T x = 1 \\ & x^T D 1 = 0. \end{aligned}$$

In the following, we will assume that this last constraint always holds, effectively limiting ourselves to be in the subspace $\mathbb{R}^n \perp 1$.

Next, we provide a description of three related random-walk-based matrices that arise naturally when considering a graph G.

- **Heat Kernel.** The Heat Kernel of a connected, undirected graph G can be defined as:

$$H_t = \exp(-tL) = \sum_{k=0}^{\infty} \frac{(-t)^k}{k!} L^k, \tag{2}$$

  where $t \geq 0$ is a time parameter. Alternatively, it can be written as $H_t = \sum_i e^{-\lambda_i t} P_i$, where λ_i is the i-th eigenvalue of L and P_i denotes the projection into the eigenspace associated with λ_i. The Heat Kernel is an operator that satisfies the heat equation $\frac{\partial H_t}{\partial t} = -LH_t$ and thus that describes the diffusive spreading of heat on the graph.

- **PageRank.** The PageRank vector $\pi(\gamma, s)$ associated with a connected, undirected graph G is defined to be the unique solution to

$$\pi(\gamma, s) = \gamma s + (1 - \gamma) M \pi(\gamma, s),$$

  where $\gamma \in (0, 1)$ is the so-called teleportation constant; $s \in \mathbb{R}^n$ is a preference vector, often taken to be (up to normalization) the all-ones vector; and M is the natural random

[2] Our result for the truncated lazy random walk generalizes a special case of the Power Method. Formalizing the regularization implicit in the Power Method more generally, or in other methods such as the Lanczos method or the Conjugate Gradient method, is technically more intricate due to the renormalization at each step, which by construction we will not need.

walk matrix associated with G [10].[3] If we fix γ and s, then it is known that $\pi(\gamma, s) = \gamma \sum_{t=0}^{\infty} (1-\gamma)^t M^t s$, and thus that $\pi(\gamma, s) = R_\gamma s$, where

$$R_\gamma = \gamma \left(I - (1-\gamma) M\right)^{-1}. \tag{3}$$

This provides an expression for the PageRank vector $\pi(\gamma, s)$ as a γ-dependent linear transformation matrix R_γ multiplied by the preference vector s [1].

- **Truncated Lazy Random Walk.** Since $M = AD^{-1}$ is the natural random walk transition matrix associated with a connected, undirected graph G, it follows that

$$W_\alpha = \alpha I + (1-\alpha) M \tag{4}$$

represents one step of the α-lazy random walk transition matrix, in which at each step there is a holding probability $\alpha \in [0, 1]$. Just as M is similar to $M' = D^{-1/2} M D^{1/2}$, which permits the computation of its real eigenvalues and full suite of eigenvectors that can be related to those of M, W_α is similar to $W'_\alpha = D^{-1/2} W_\alpha D^{1/2}$. Thus, iterating the random walk W_α is similar to applying the Power Method to W'_α, except that the renormalization at each step need not be performed since the top eigenvalue is unity.

Each of these three matrices has been used to compute vectors that in applications are then used in place of the smallest nontrivial eigenvector of a graph Laplacian. This is typically achieved by starting with an initial random vector and then applying the Heat Kernel matrix, or the PageRank operator, or truncating a Lazy Random Walk.

Finally, we recall that the solution SPECTRAL can also be characterized as the solution to a semidefinite program (SDP). To see this, consider the following SDP:

$$\begin{aligned} \mathsf{SDP}: \quad \min \quad & L \bullet X \\ \text{s.t.} \quad & \mathrm{Tr}(X) = I \bullet X = 1 \\ & X \succeq 0, \end{aligned}$$

where $\bullet$ stands for the Trace, or matrix inner product, operation, *i.e.*, $A \bullet B = \mathrm{Tr}(AB^T) = \sum_{ij} A_{ij} B_{ij}$ for matrices A and B. SDP is a relaxation of the spectral program SPECTRAL from an optimization over unit vectors to an optimization over distributions over unit vectors, represented by the density matrix X.

To see the relationship between the solution x of SPECTRAL and the solution X of SDP, recall that a density matrix X is a matrix of second moments of a distribution over unit vectors. In this case, $L \bullet X$ is the expected value of $x^T L x$, when x is drawn from a distribution defined by X. If X is rank-1, as is the case for the solution to SDP, then the distribution is completely concentrated on a vector v, and the SDP and vector solutions are the same, in the sense that $X = vv^T$. More generally, as we will encounter below, the solution to an SDP may not be rank-1. In that case, a simple way to construct a vector x from a distribution defined by X is to start with an n-vector ξ with entries drawn i.i.d. from the normal distribution $N(0, 1/n)$, and consider $x = X^{1/2}\xi$. Note that this procedure effectively samples from a Gaussian distribution with second moment X.

[3] Alternatively, one can define $\pi'(\gamma, s)$ to be the unique solution to $\pi = \gamma s + (1-\gamma) W \pi$, where W is the $1/2$-lazy random walk matrix associated with G. These two vectors are related as $\pi'(\gamma, s) = \pi(\frac{2\gamma}{1+\gamma}, s)$ [1].

3 Approximation procedures and regularized spectral optimization problems

3.1 A simple theorem characterizing the solution to a regularized SDP

Here, we will apply regularization technique to the SDP formulation provided by SDP, and we will show how natural regularization functions yield distributions over vectors which correspond to the diffusion-based or random-walk-based matrices. In order to regularize SDP, we want to modify it such that the distribution is not degenerate on the second eigenvector, but instead spreads the probability on a larger set of unit vectors around v. The regularized version of SDP we will consider will be of the form:

$$\begin{aligned} (\mathsf{F},\eta)-\mathsf{SDP} \quad \min \quad & L\bullet X + {}^1\!/_\eta \cdot F(X) \\ \text{s.t.} \quad & I\bullet X = 1 \\ & X \succeq 0, \end{aligned}$$

where $\eta > 0$ is a trade-off or regularization parameter determining the relative importance of the regularization term $F(X)$, and where F is a real strictly-convex infinitely-differentiable rotationally-invariant function over the positive semidefinite cone. (Think of F as a strictly convex function of the eigenvalues of X.) For example, F could be the negative of the von Neumann entropy of X; this would penalize distributions that are too concentrated on a small measure of vectors. We will consider other possibilities for F below. Note that due to F, the solution X of $(\mathsf{F},\eta)-\mathsf{SDP}$ will in general not be rank-1.

Our main results on implicit regularization via approximate computation will be based on the following structural theorem that provides sufficient conditions for a matrix to be a solution of a regularized SDP of a certain form. Note that the Lagrangian parameter λ and its relationship with the regularization parameter η will play a key role in relating this structural theorem to the three random-walk-based proceudres described previously.

Theorem 1 *Let G be a connected, weighted, undirected graph, with normalized Laplacian L. Then, the following conditions are sufficient for $X^\star$ to be an optimal solution to $(\mathsf{F},\eta)-\mathsf{SDP}$.*

1. $X^\star = (\nabla F)^{-1}\left(\eta\cdot(\lambda^* I - L)\right)$, *for some* $\lambda^*\in\mathbb{R}$,
2. $I\bullet X^\star = 1$,
3. $X^\star \succeq 0$.

Proof: For a general function F, we can write the Lagrangian $\mathcal{L}$ for $(\mathsf{F},\eta)-\mathsf{SDP}$ as follows:

$$\mathcal{L}(X,\lambda,U) = L\bullet X + \frac{1}{\eta}\cdot F(X) - \lambda\cdot(I\bullet X - 1) - U\bullet X$$

where $\lambda\in\mathbb{R}, U\succeq 0$. The dual objective function is

$$h(\lambda,U) = \min_{X\succeq 0}\mathcal{L}(X,\lambda,U).$$

As F is strictly convex, differentiable and rotationally invariant, the gradient of F over the positive semidefinite cone is invertible and the righthand side is minimized when

$$X = (\nabla F)^{-1}(\eta\cdot(-L+\lambda^*\cdot I + U)),$$

where λ^* is chosen such that the second condition in the statement of the theorem is satisfied. Hence,

$$h(\lambda^\star, 0) = L \bullet X^\star + \frac{1}{\eta} \cdot F(X^\star) - \lambda^\star \cdot (I \bullet X^\star - 1) = L \bullet X^\star + \frac{1}{\eta} \cdot F(X^\star).$$

By Weak Duality, this implies that $X^\star$ is an optimal solution to $(\mathsf{F}, \eta) - \mathsf{SDP}$.

$\diamond$

3.2 The connection between approximate eigenvector computation and implicit statistical regularization

In this section, we will consider the three diffusion-based or random-walked-based heuristics described in Section 2, and we will show that each may be viewed as solving $(\mathsf{F}, \eta) - \mathsf{SDP}$ for an appropriate value of F and η.

Generalized Entropy and the Heat Kernel. Consider first the Generalized Entropy function:

$$F_H(X) = \mathrm{Tr}(X \log X) - \mathrm{Tr}(X), \tag{5}$$

for which:

$$\begin{aligned} (\nabla F_H)(X) &= \log X \\ (\nabla F_H)^{-1}(Y) &= \exp Y. \end{aligned}$$

Hence, the solution to $(\mathsf{F_H}, \eta) - \mathsf{SDP}$ has the form:

$$X_H^\star = \exp(\eta \cdot (\lambda I - L)), \tag{6}$$

for appropriately-chosen values of λ and η. Thus, we can establish the following lemma.

Lemma 1 *Let $X_H^\star$ be an optimal solution to $(\mathsf{F}, \eta) - \mathsf{SDP}$, when $F(\cdot)$ is the Generalized Entropy function, given by Equation (5). Then*

$$X_H^\star = \frac{H_\eta}{\mathrm{Tr}\left[H_\eta\right]},$$

which corresponds to a "scaled" version of the Heat Kernel matrix with time parameter $t = \eta$.

Proof: From Equation (6), it follows that $X_H^\star = \exp(-\eta \cdot L) \cdot \exp(\eta \cdot \lambda)$, and thus by setting $\lambda = -1/\eta \log(\mathrm{Tr}(\exp(-\eta \cdot L)))$, we obtain the expression for $X_H^\star$ given in the lemma. Thus, $X_H^\star \succeq 0$ and $\mathrm{Tr}(X_H^\star) = 1$, and by Theorem 1 the lemma follows.

$\diamond$

Conversely, given a graph G and time parameter t, the Heat Kernel of Equation (2) can be characterized as the solution to the regularized $(\mathsf{F_H}, \eta) - \mathsf{SDP}$, with the regularization parameter $\eta = t$ (and for the value of the Lagrangian parameter λ as specified in the proof).

Log-determinant and PageRank. Next, consider the Log-determinant function:

$$F_D(X) = -\log\det(X), \tag{7}$$

for which:

$$\begin{aligned} (\nabla F_D)(X) &= -X^{-1} \\ (\nabla F_D)^{-1}(Y) &= -Y^{-1}. \end{aligned}$$

Hence, the solution to $(\mathsf{F_D}, \eta) - \mathsf{SDP}$ has the form:

$$X_D^\star = -(\eta \cdot (\lambda I - L))^{-1}, \tag{8}$$

for appropriately-chosen values of λ and η. Thus, we can establish the following lemma.

Lemma 2 *Let $X_D^\star$ be an optimal solution to $(\mathsf{F}, \eta) - \mathsf{SDP}$, when $F(\cdot)$ is the Log-determinant function, given by Equation (7). Then*

$$X_D^\star = \frac{D^{-1/2} R_\gamma D^{1/2}}{\mathrm{Tr}\left[R_\gamma\right]}$$

which corresponds to a "scaled-and-streached" version of the PageRank matrix R_γ of Equation (3) with teleportation parameter γ depending on η.

Proof: Recall that $L = I - D^{-1/2} A D^{-1/2}$. Since $X_D^\star = {}^1\!/\!_\eta \cdot (L - \lambda I)^{-1}$, by standard manipulations it follows that

$$X_D^\star = \frac{1}{\eta}\left((1-\lambda)I - D^{-1/2} A D^{-1/2}\right)^{-1}.$$

Thus, $X_D^\star \succeq 0$ if $\lambda \leq 0$, and $X_D^\star \succ 0$ if $\lambda < 0$. If we set $\gamma = \frac{\lambda}{\lambda - 1}$ (which varies from 1 to 0, as λ varies from $-\infty$ to 0), then it can be shown that

$$X_D^\star = \frac{-1}{\eta\lambda} D^{-1/2} \gamma \left(I - (1-\gamma) A D^{-1}\right)^{-1} D^{1/2}.$$

By requiring that $1 = \mathrm{Tr}[X_D^\star]$, it follows that

$$\eta = (1-\gamma)\mathrm{Tr}\left[\left(I - (1-\gamma) A D^{-1}\right)^{-1}\right]$$

and thus that $\eta\lambda = -\mathrm{Tr}\left[\gamma (I - (1-\gamma) A D^{-1})^{-1}\right]$. Since $R_\gamma = \gamma (I - (1-\gamma) A D^{-1})^{-1}$, the lemma follows.

$\diamond$

Conversely, given a graph G and teleportation parameter γ, the PageRank of Equation (3) can be characterized as the solution to the regularized $(\mathsf{F_D}, \eta) - \mathsf{SDP}$, with the regularization parameter η as specified in the proof.

Standard p-norm and Truncated Lazy Random Walks. Finally, consider the Standard p-norm function, for $p > 1$:

$$F_p(X) = \frac{1}{p}||X||_p^p = \frac{1}{p}\mathrm{Tr}(X^p), \tag{9}$$

for which:

$$\begin{aligned} (\nabla F_p)(X) &= X^{p-1} \\ (\nabla F_p)^{-1}(Y) &= Y^{1/(p-1)}. \end{aligned}$$

Hence, the solution to $(\mathsf{F_p}, \eta) - \mathsf{SDP}$ has the form

$$X_p^\star = (\eta \cdot (\lambda I - L))^{q-1}, \tag{10}$$

where $q > 1$ is such that ${}^1\!/\!_p + {}^1\!/\!_q = 1$, for appropriately-chosen values of λ and η. Thus, we can establish the following lemma.

Lemma 3 *Let $X_p^\star$ be an optimal solution to $(\mathsf{F}, \eta) - \mathsf{SDP}$, when $F(\cdot)$ is the Standard p-norm function, for $p > 1$, given by Equation (9). Then*

$$X_p^\star = \frac{D^{\frac{-(q-1)}{2}} W_\alpha^{q-1} D^{\frac{q-1}{2}}}{\mathrm{Tr}\left[W_\alpha^{q-1}\right]}$$

which corresponds to a "scaled-and-streached" version of $q-1$ steps of the Truncated Lazy Random Walk matrix W_α of Equation (4) with laziness parameter α depending on η.

Proof: Recall that $L = I - D^{-1/2} A D^{-1/2}$. Since $X_p^\star = \eta^{q-1} \cdot (\lambda I - L)^{q-1}$, by standard manipulations it follows that

$$X_p^\star = \eta^{q-1} \left((\lambda - 1) I + D^{-1/2} A D^{-1/2}\right)^{q-1}$$

Thus, $X_p^\star \succeq 0$ if $\lambda \geq 1$, and $X_p^\star \succ 0$ if $\lambda > 1$. If we set $\alpha = \frac{\lambda - 1}{\lambda}$ (which varies from 0 to 1, as λ varies from 1 to ∞), then it can be shown that

$$X_p^\star = (\eta\lambda)^{q-1} D^{-(q-1)/2} \left(\alpha I - (1-\alpha) A D^{-1}\right)^{q-1} D^{(q-1)/2}.$$

By requiring that $1 = \mathrm{Tr}[X_p^\star]$, it follows that

$$\eta = (1-\alpha) \left\{ \mathrm{Tr}\left[\left(\alpha I + (1-\alpha) A D^{-1}\right)^{q-1} \right] \right\}^{1-p}$$

and thus that $\eta\lambda = \left\{ \mathrm{Tr}\left[\left(\alpha I + (1-\alpha) A D^{-1}\right)^{q-1} \right] \right\}^{1-p}$. Since $W_\alpha = \alpha I + (1-\alpha) A D^{-1}$, the lemma follows.

$\diamond$

Conversely, given a graph G, a laziness parameter α, and a number of steps $q' = q - 1$, the Truncated Lazy Random Walk of Equation (4) can be characterized as the solution to the regularized $(\mathsf{F_p}, \eta) - \mathsf{SDP}$, with the regularization parameter η as specified in the proof.

4 Discussion and Conclusion

There is a large body of empirical and theoretical work with a broadly similar flavor to ours. Here, we provide just a few citations that most informed our approach.

- In machine learning, Belkin, Niyogi, and Sindhwan describe a geometrically-motivated framework within which semi-supervised learning algorithms can be constructed [3]; Saul and Roweis (and many others, but less explicitly) observe that adding a regularization term to improve numerical properties also "acts to penalize large weights that exploit correlations beyond some level of precision in the data sampling process" [18]; Rosasco, De Vito, and Verri describe how a large class of regularization methods designed for solving ill-posed inverse problems gives rise to novel learning algorithms [17]; Zhang and Yu show that in boosting, early stopping (as opposed to waiting for full convergence) leads to regularization and hence better prediction [23]; Shi and Yu describe statistical aspects of binning in Gaussian kernel regularization [20]; and Bishop observes that training with noise can be equivalent to Tikhonov regularization [5].

- In numerical linear algebra, O'Leary, Stewart, and Vandergraft have described issues that arise in estimating the largest eigenvalue of a positive definite matrix with the power method [15]; and Parlett, Simon, and Stringer have described convergence issues that arise when estimating the largest eigenvalue with an iterative method [16].

- In the theory of algorithms, Spielman and Teng describe how to perform local graph partitioning using truncated random walks [21]; Andersen, Chung, and Lang describe an improved local graph partitioning algorithm PageRank vectors [1]; and Chung describes how to perform similar operations by using the heat kernel and viewing it as the so-called pagerank of a graph [8].

- In internet data analysis, Andersen and Lang use these methods to try to find communities in large networks [2]; Leskovec, Lang, and Mahoney use these and other methods to show that there do not exist good large communities in large social and information networks [11, 12]; and Lu *et al.* empirically evaluate implicit regularization constraints for improved online review quality prediction [13].

None of this work, however, takes the approach we have adopted of asking: What is the regularized optimization objective that a heuristic or approximation algorithm is exactly optimizing?

We should note that one can interpret our main results from one of two alternate perspectives. From the perspective of worst-case analysis, we provide a simple characterization of several related methods for approximating the smallest nontrivial eigenvector of a graph Laplacian as solving a related optimization problem. By adopting this view, it should perhaps be less surprising that these methods have Cheeger-like inequalities, with related algorithmic consequences, associated with them [21, 1, 8, 7]. From a statistical perspective, one could imagine one method or another being more or less appropriate as a method to compute robust approximations to the smallest nontrivial eigenvector of a graph Laplacian, depending on assumptions being made about the data. By adopting this view, it should perhaps be less surprising that these methods have performed well at identifying structure in sparse and noisy networks [2, 11, 12, 13].

The particular results that motivated us to ask this question had to do with recent empirical work on characterizing the clustering and community structure in very large social and information networks [11, 12]. As a part of that line of work, Leskovec, Lang, and Mahoney (LLM) [12] were interested in understanding the artifactual properties induced in output clusters as a function of different approximation algorithms for a given objective function (that formalized the community concept). LLM observed a severe tradeoff between the objective function value and the "niceness" of the clusters returned by different approximation algorithms. This phenomenon is analogous to the bias-variance tradeoff that is commonly-observed in statistics and machine learning, except that LLM did not perform any explicit regularization—instead, they observed this phenomenon as a function of different approximation algorithms to compute approximate solutions to the intractable graph partitioning problem.

Although we have focused in this paper simply on the problem of computing an eigenvector, one is typically interested in computing eigenvectors in order to perform some downstream data analysis or machine learning task. For instance, one might be interested in characterizing the clustering properties of the data. Alternatively, the goal might be to perform classification or regression or ranking. It would, of course, be of interest to understand how the concept of *implicit regularization via approximate computation* extends to the output of algorithms for these problems. More generally, though, it would be of interest to understand how this concept of implicit regularization via approximate computation extends to intractable graph optimization problems (that are not obviously formulatable as vector space problems) that are more popular in computer science. That is: What is the (perhaps implicitly regularized) optimization problem that an approximation algorithm for an intractable optimization problem is implicitly optimizing? Such graph problems arise in many applications, but the the formulation and solution of these graph problems tends to be quite different than that of matrix problems that are more popular in machine learning and statistics. Recent empirical and theoretical evidence, however, clearly suggests that regularization will be fruitful in this more general setting.

References

[1] R. Andersen, F.R.K. Chung, and K. Lang. Local graph partitioning using PageRank vectors. In *FOCS '06: Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pages 475–486, 2006.

[2] R. Andersen and K. Lang. Communities from seed sets. In *WWW '06: Proceedings of the 15th International Conference on World Wide Web*, pages 223–232, 2006.

[3] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

[4] P. Bickel and B. Li. Regularization in statistics. *TEST*, 15(2):271–344, 2006.

[5] C. M. Bishop. Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, 7(1):108–116, 1995.

[6] Z. Chen and S. Haykin. On different facets of regularization theory. *Neural Computation*, 14(12):2791–2846, 2002.

[7] F.R.K Chung. Four proofs of Cheeger inequality and graph partition algorithms. In *Proceedings of ICCM*, 2007.

[8] F.R.K. Chung. The heat kernel as the pagerank of a graph. *Proceedings of the National Academy of Sciences of the United States of America*, 104(50):19735–19740, 2007.

[9] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer-Verlag, New York, 2003.

[10] A. N. Langville and C. D. Meyer. Deeper inside PageRank. *Internet Mathematics*, 1(3):335–380, 2004.

[11] J. Leskovec, K.J. Lang, A. Dasgupta, and M.W. Mahoney. Statistical properties of community structure in large social and information networks. In *WWW '08: Proceedings of the 17th International Conference on World Wide Web*, pages 695–704, 2008.

[12] J. Leskovec, K.J. Lang, and M.W. Mahoney. Empirical comparison of algorithms for network community detection. In *WWW '10: Proceedings of the 19th International Conference on World Wide Web*, pages 631–640, 2010.

[13] Y. Lu, P. Tsaparas, A. Ntoulas, and L. Polanyi. Exploiting social context for review quality prediction. In *WWW '10: Proceedings of the 19th International Conference on World Wide Web*, pages 691–700, 2010.

[14] A. Neumaier. Solving ill-conditioned and singular linear systems: A tutorial on regularization. *SIAM Review*, 40:636–666, 1998.

[15] D. P. O'Leary, G. W. Stewart, and J. S. Vandergraft. Estimating the largest eigenvalue of a positive definite matrix. *Mathematics of Computation*, 33(148):1289–1292, 1979.

[16] B. N. Parlett, H. Simon, and L. M. Stringer. On estimating the largest eigenvalue with the Lanczos algorithm. *Mathematics of Computation*, 38(157):153–165, 1982.

[17] L. Rosasco, E. De Vito, and A. Verri. Spectral methods for regularization in learning theory. Technical Report DISI-TR-05-18, 2005.

[18] L.K. Saul and S.T. Roweis. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, 4:119–155, 2003.

[19] B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.

[20] T. Shi and B. Yu. Binning in Gaussian kernel regularization. *Statistica Sinica*, 16:541–567, 2005.

[21] D.A. Spielman and S.-H. Teng. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *STOC '04: Proceedings of the 36th annual ACM Symposium on Theory of Computing*, pages 81–90, 2004.

[22] A.N. Tikhonov and V.Y. Arsenin. *Solutions of Ill-Posed Problems*. W.H. Winston, Washington, D.C., 1977.

[23] T. Zhang and B. Yu. Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 33:1538–1579, 2005.