A CONTINUUM LIMIT FOR THE PAGERANK ALGORITHM

Amber Yuan, Jeff Calder, and Braxton Osting

SIAM Workshop on Network Science 2020
July 9–10 · Toronto

Summary

We propose a new framework for rigorously studying continuum limits of unsupervised and semi-supervised machine learning methods on directed graphs. We use the framework to study the PageRank algorithm and show that the corresponding continuum limit problem, which is taken as the number of webpages tends to infinity, is a second-order, possibly degenerate, elliptic equation that contains reaction, diffusion, and advection terms. We prove that the numerical scheme is consistent and stable and compute explicit rates of convergence of the discrete solution to the solution of the continuum limit PDE.

Introduction

Unsupervised and semi-supervised machine learning methods often rely on graphs to model data, stimulating research on theoretical properties of operators on graphs. Due to the ubiquity of graph Laplacians in graph-based learning problems, much work has been devoted to understand and quantify how these matrices can uncover geometric and distributional structure from unlabeled data. One popular approach assumes a random geometric graph with n points and length scale $h > 0$, and considers the limit as $n \to \infty$ and $h \to 0$; see, *e.g.*, [1, 3, 4, 5, 6]. Here, the graph vertices are an *i.i.d.* sample of size n from a density ρ supported on a d-dimensional manifold $\mathcal{M}$ embedded in $\mathbb{R}^D$, and the edge weights are given by

$$\omega_{xy} = \Phi\left(\frac{|x-y|}{h}\right),$$

where $\Phi\colon [0,\infty) \to [0,\infty)$ is a kernel function. Under appropriate assumptions, the pointwise consistency result can be established that a graph Laplacian Ł applied to a test function $\phi \in C^3(\mathcal{M})$ converges to

$$\Delta_\rho \phi = \rho^a \operatorname{div}\left(\rho^b \nabla(\rho^c \phi)\right) \qquad \text{as } n \to \infty \text{ and } h \to 0$$

where Δ_ρ is a weighted Laplace-Beltrami operator with various values of a b, c that depend on the choice of the graph Laplacian. *E.g.*, for the unnormalized graph Laplacian, $a = -1$, $b = 2$, $c = 0$, and for the random walk Laplacian $a = -2$, $b = 2$, $c = 0$. If $h \to 0$ and $n \to \infty$ simultaneously, then the condition $nh^{d+2} \gg \log n$ is required for pointwise consistency; it ensures each vertex has enough neighbors to apply appropriate concentration of measure results. To obtain $O(h)$ pointwise consistency rates, it is required that $nh^{d+4} \gg 1$. We contrast this with the condition $nh^d \gg \log n$ required for graph connectivity.

While most of the existing literature focuses on undirected graphs, directed graphs are very important in practice, giving models for physical, biological, or transportation networks, among many other applications [2]. *In this talk, we discuss a new framework for rigorously studying continuum limits of unsupervised and semi-supervised machine learning methods on directed graphs, developed by the authors in [7]. This framework is applied to establish and study a continuum limit for the PageRank algorithm.*

Setup and results

To study continuum limits for problems on directed graphs, we propose a new model that we call a *random directed geometric graph*. Let $x_1, x_2, \ldots, x_n$ be an *i.i.d.* sample of size n on the torus $\mathbb{T}^d = \mathbb{R}^d/\mathbb{Z}^d$ with density $\rho\colon \mathbb{T}^d \to [0,\infty)$. We define a weight, ω_{xy}, from x to y by

$$\omega_{xy} = \Phi\left(\frac{|B(x)(y - x - \varepsilon b(x))|}{h}\right),$$

where $b\colon \mathbb{T}^d \to \mathbb{R}^d$ and $B\colon \mathbb{T}^d \to \mathbb{R}^{d\times d}$ with $B(x)$ having full rank for every $x \in \mathbb{T}^d$. The parameter $h > 0$ is the bandwidth of the kernel, and $\varepsilon > 0$ is the strength of the directionality. We assume the kernel function Φ is smooth, nonnegative, nonincreasing, and $\int_{B(0,2)} \Phi(|z|)\,dz = 1$. When $B = I$ and $b = 0$ or $\varepsilon = 0$, the weights give a random *undirected* geometric graph; for other choices, the graph weights are directed. The vector field b imparts directionality. The matrix B can be viewed as changing the metric locally and for simplicity we take $B(x) = I$. We define the degree of x by $d_n(x) = \sum_{y\in X_n} \omega_n(x,y)$.

We consider a random walk on the directed graph transitioning from vertices x to y with probability $p_{xy} = d_n(x)^{-1}\omega_n(x,y)$. Denoting the *teleportation probability* by $\alpha \in [0,1]$ and the *teleportation probability distribu-*

tion by $v(x)$, the *PageRank vector*, denoted $r_n\colon X_n \to \mathbb{R}$, satisfies the linear system

$$r_n(x) - (1-\alpha)\sum_{y\in X_n}\frac{\omega_n(y,x)}{d_n(y)}r_n(y) = \alpha v(x),$$

for all $x \in X_n$; see, *e.g.*, [2]. Simplifying the problem, we define the *normalized PageRank vector*, $u_n\colon X_n \to \mathbb{R}$, by

$$u_n(x) = \frac{nh^d}{d_n(x)}r_n(x),$$

which satisfies

$$u_n(x) - \gamma \mathrm{L}_n u_n(x) = \frac{nh^d}{d_n(x)}v(x) \qquad \forall x \in X_n, \tag{1}$$

where $\gamma = (1-\alpha)/\alpha$ and the *PageRank Operator* is defined

$$\mathrm{L}_n u(x) := \frac{1}{d_n(x)}\sum_{y\in X_n}\omega_n(y,x)u(y) - u(x).$$

The corresponding problem in the continuum is the, possibly degenerate, elliptic PDE on $\mathbb{T}^d$,

$$u + \gamma_\varepsilon \rho^{-2}\mathrm{div}\,(\rho^2 bu) - \frac{1}{2}\sigma_\Phi\gamma_h\rho^{-2}\mathrm{div}\,(\rho^2\nabla u) = \rho^{-1}v, \tag{2}$$

where $\sigma_\Phi = \int \Phi(|z|)z_1^2dz$, $\gamma_\varepsilon = \frac{(1-\alpha)\varepsilon}{\alpha}$, and $\gamma_h = \frac{(1-\alpha)h^2}{\alpha}$. Denote $\eta = \|\rho^{-2}\mathrm{div}\,(\rho^2 b)\|_{L^\infty(\mathbb{T}^d)}$. When $\gamma_h > 0$ and $\eta\gamma_\varepsilon < 1$, a standard result in elliptic PDEs gives that (2) has a unique solution $u \in C^{3,\alpha}(\mathbb{T}^d)$.

Theorem 1 ([7, Theorem 2.3], Convergence of PageRank)**.** *Let $\rho \in C^{2,\alpha}(\mathbb{T}^d)$, $b \in C^{2,\alpha}(\mathbb{T}^d;\mathbb{R}^d)$ and $v \in C^{1,\alpha}(\mathbb{T}^d)$ for any $\alpha \in (0,1)$. Assume that $\gamma_\varepsilon \le 1$, $\gamma_h \in (0,1)$, and $\eta < 1$. Let u_n be the solution to the PageRank problem* (1) *and let $u \in C^3(\mathbb{T}^d)$ be the solution to the PDE* (2)*. Then there exists $C_1, C_2, c_1, c_2 > 0$ with C_1 depending on $\gamma_h > 0$, such that when $\varepsilon + h \le c_1(1-\eta\gamma_\varepsilon)$ we have that*

$$\max_{x\in X_n}|u(x) - u_n(x)| \le C_1(1-\eta\gamma_\varepsilon)^{-1}(\lambda + \varepsilon + h)$$

holds with probability at least $1 - C_2 n\exp(-c_2 nh^{d+2}\lambda^2) - C_2 n\exp\left(-c_2 nh^{d+2}(1-\eta\gamma_\varepsilon)^2\right)$, where $\lambda \in (0,1]$.

Our proof of Theorem 1 first uses a Bernstein type concentration inequality to establish consistency of the PageRank vector and then uses a maximum principle argument to establish convergence.

The continuum PDE (2) has reaction, advection, and diffusion terms. The two reaction terms, u and $\rho^{-1}v$, are due to the teleportation step in PageRank. The advection term, $\mathrm{div}\,(\rho^2 bu)$, describes the advection of the quantity $\rho^2 u$ along the vector field b, and is due to the directional preference in the definition of the weights in a random directed geometric graph. Finally, the weighted diffusion term, $\mathrm{div}\,(\rho^2\nabla u)$, represents diffusion from the random walk step of PageRank.

Theorem 1 is stated as a finite sample size result, where n, ε, h, α, and λ are fixed. If we consider the continuum limit as $n \to \infty$ and $\varepsilon_n, h_n, \alpha_n, \lambda_n \to 0$, then Theorem 1 tells us how to relatively scale the parameters. We assume $\varepsilon_n \le \alpha_n$ and $h_n^2 \le \alpha_n$, so that $\gamma_{\varepsilon_n}, \gamma_{h_n} \le 1$. To ensure the continuum limit holds with probability one, we require

$$\lim_{n\to\infty}\frac{nh_n^{d+2}\lambda_n^2}{\log n} = \infty, \tag{3}$$

which is a standard scaling for pointwise consistency of graph Laplacians. In this case, we have convergence rate of $O(\lambda_n + \varepsilon_n + h_n)$ in Theorem 1 with probability one.

A corollary of Theorem 1 is the asymptotic Lipschitz regularity of the PageRank vector [7, Corollary 2.12], which shows that the PageRank vector does not vary much between vertices.

When $\gamma_h = 0$ or $\gamma_h > 0$ is small, the continuum PDE (2) is approximated by the first order equation

$$u + \gamma_\varepsilon\rho^{-2}\mathrm{div}\,(\rho^2 bu) = \rho^{-1}v \quad \text{ on } \mathbb{T}^d. \tag{4}$$

In this case, an analogous result to Theorem 1 shows that the solution to the PageRank problem (1) converges to the viscosity solution of the PDE (4); see [7, Theorem 2.5].

Finally, using the same techniques, we also obtain the convergence of the probability distribution for the random walk to the solution of a particular reaction-advection-diffusion equation; see [7, Theorem 2.14].

References

[1] M. Belkin and P. Niyogi. Towards a theoretical foundation for Laplacian-based manifold methods. *COLT*, 2005.

[2] D. F. Gleich. PageRank beyond the web. *SIAM Review*, 57(3), 2015.

[3] M. Hein, J.-Y. Audibert, and U. v. Luxburg. Graph Laplacians and their convergence on random neighborhood graphs. *Journal of Machine Learning Research*, 8, 2007.

[4] S. S. Lafon. *Diffusion maps and geometric harmonics*. PhD thesis, Yale University, 2004.

[5] A. Singer. From graph to manifold Laplacian: The convergence rate. *Applied and Computational Harmonic Analysis*, 2006.

[6] D. Ting, L. Huang, and M. Jordan. An analysis of the convergence of graph Laplacians. *ICML*, 2010.

[7] A. Yuan, J. Calder, and B. Osting. A continuum limit for the pagerank algorithm. arXiv:2001.08973, 2020.