ASYMPTOTIC CONSISTENCY ANALYSIS OF (HYPER)GRAPH ALGORITHMS IN SEMI-SUPERVISED LEARNING

Adrien Weihs

SIGMA 2024, Marseille, November 2024

OUTLINE

BACKGROUND IN GRAPH LEARNING

SEMI-SUPERVISED LEARNING ON GRAPHS

- Given n data points $\Omega_n = \{x_i\}_{i=1}^n$ where we assume that $x_i \overset{\text{iid}}{\sim} \mu \in \mathcal{P}(\Omega)$ for $\Omega \subseteq \mathbb{R}^d$ and labels $\{\ell_i\}_{i=1}^N \subset \{0,1\}^N$ with $N \ll n$, we want to find the labels for the remaining points $\{\ell_i\}_{i=N+1}^n$
- Ideally: leverage the **geometric** information of the (unlabelled and labelled) samples
- One possible solution: structure the data in a **weighted** graph and consider the labelling problem on the latter

GRAPH SETTING

- An undirected weighted graph G is a tuple (V, W) where V is the set of vertices and W are the edge weights
- In our case, $V = \Omega_n$ and $W \in \mathbb{R}^{n \times n}$ is symmetric and $w_{ij} \geq 0$
- We say that the vertices x_i and x_j are connected by an edge if $w_{ij} > 0$
- **Intuition**: the "closer" x_i and x_j are, the larger w_{ij} should be

LAPLACE LEARNING I

- Prominent example of labelling on graphs is Laplace learning [46]

- We look for a function $u_n : \Omega_n \to \mathbb{R}$ that satisfies:

$$u_n \in \operatorname*{argmin}_{v:\Omega_n \to \mathbb{R}} \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} \left(v(x_i) - v(x_j) \right)^2 \text{ such that } v(x_i) = \ell_i \text{ for } i \leq N$$

  and we define $\mathcal{E}_n(v) = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} \left(v(x_i) - v(x_j) \right)^2$

- Intuition: vertices x_i and x_j that are close in the graph — i.e. w_{ij} is large — should have similar labels, i.e. **continuity in the graph domain**

LAPLACE LEARNING II

- Since u_n takes values in $\mathbb{R}$, the classification rule for $N < i \leq n$ is:

$$\hat{\ell}_i = \begin{cases} 0 & \text{if } u_n(x_i) < 0.5 \\ 1 & \text{else.} \end{cases}$$

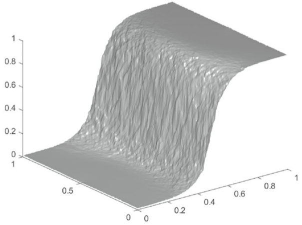


Figure: Possible solution to Laplace learning with points $(0, 0.5)$ and $(1, 0.5)$ labeled 0 and 1 respectively [8]

WHY IS LAPLACE LEARNING CALLED LAPLACE LEARNING?

- For a graph $G = (\Omega_n, W)$, we define the graph Laplacian matrix $\Delta_n = D - W$ where D is the diagonal matrix with entries $d = \sum_{j=1}^{n} w_{ij}$
- We note that

$$\sum_{i,j=1}^{n} w_{ij} \left(u_n(x_i) - u_n(x_j)\right)^2 = u_n^T \Delta_n u_n$$

- **Spectral properties** of Laplacian matrix are crucial in many applications, e.g. spectral clustering [41]

KEY TAKEAWAYS FROM LAPLACE LEARNING

- Laplace learning is a **variational problem** on the graph i.e. functions u_n are minimizers of the functional/energy $\mathcal{E}_n$ (with pointwise constraints)

$\Rightarrow$ Mathematical structure allows for rigorous analysis. Other examples include:

  - Ginzburg-Landau functional on graphs [1], [40]
  - Total variation functional on graphs [18]
  - Mumford-Shah functional on graphs [10]
  - Graph cuts/Spectral clustering [20], [17], [19]

- We want u_n to be somewhat **continuous** for reasonable labelling

ASYMPTOTIC CONSISTENCY ANALYSIS

ASYMPTOTIC CONSISTENCY

- In machine learning, one usually has a finite number n of data points

- However, with our ever-growing data-capturing capabilities we get very large data sets

$\Rightarrow$ Natural question: what happens when $n \to \infty$?

- Desired outcomes:

  – the discrete model *converges* to a continuum model which we can study through classical techniques and we **gain insights** into how to better design the discrete algorithm

  – we want to able to scale algorithms without a **trial and error approach** which is costly

- In this talk: analogue of **scaling laws** in deep learning for graph learning

VARIANTS OF ASYMPTOTIC CONSISTENCY ANALYSIS

- List of asymptotic consistency analysis methods (non-exhaustive):
  - **Pointwise consistency**: for v a continuum function, consider $\mathcal{E}_n(v) \to \mathcal{E}_\infty(v) + \mathrm{error}$ where $\mathcal{E}_n$ and $\mathcal{E}_\infty$ are discrete and continuum energies respectively [6]
  - **Probabilistic/Bayesian consistency**: formulate the discrete and continuum model/learning problems as random processes/as sampling from posteriors and show their convergence [21]/[22]

- First and third parts of this talk, **variational consistency**: convergence of minimizers of discrete problems to minimizers of continuum problems [18], i.e. *convergence after training*

- Second part of this talk, **gradient flow consistency**: you consider convergence of discrete learning trajectory/numerical procedure to continuum one [13], i.e. *convergence of training/numerical procedure*

CASE STUDY OF REGULARITY I

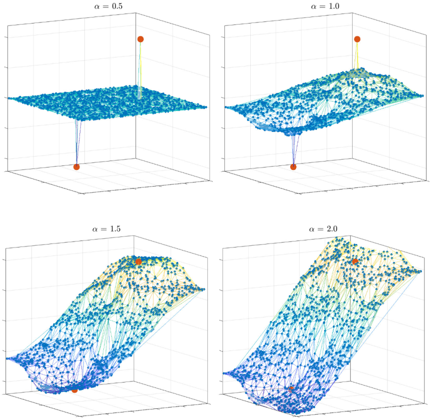


Figure: Function u_n^α minimizing an energy $\mathcal{E}_n^\alpha$ with a parameter α [14].

CASE STUDY OF REGULARITY II

$$\boxed{\alpha \leq 1}$$

$$\mathcal{E}_n^\alpha \to \mathcal{E}_\infty \text{ and } \operatorname{argmin} \mathcal{E}_\infty = \text{constants}$$

$$\Rightarrow \text{For } n \gg 1, u_n = \operatorname{argmin} \mathcal{E}_n \approx \text{constants}$$

$$\boxed{\alpha > 1}$$

$$\mathcal{E}_n^\alpha \to \mathcal{G}_\infty \text{ and } \operatorname{argmin} \mathcal{G}_\infty = \text{regular functions that interpolate the labels}$$

$$\Rightarrow \text{For } n \gg 1, u_n = \operatorname{argmin} \mathcal{E}_n \approx \text{regular functions that interpolate the labels}$$

Figure: Asymptotic hypothesis

TECHNICAL ASPECTS OF ASYMPTOTIC CONSISTENCY

- Two natural questions for the study of asymptotic consistency in graph algorithms:

  1. How does one **adapt the graph setting** to growing data sets?
  2. What is the **limit of our variational problems** $\mathcal{E}_n$ and what can be said about the convergence of the functions u_n?

GRAPH CONSTRUCTION

- As the number of vertices increases, one needs to systematically define weights. We choose:

$$w_{\varepsilon,ij} = \frac{1}{\varepsilon^d}\eta\left(\frac{|x_i - x_j|}{\varepsilon}\right)$$

  for some $\varepsilon > 0$ and non-increasing $\eta : [0,\infty) \mapsto [0,\infty)$

- If $\eta = \mathbb{1}_{[0,1]}$, vertices further apart than ε are not connected by an edge

- $w_{\varepsilon,ij}$ allows to link the extrinsic Euclidean geometry to the intrinsic geometry of the graph: **leverages the geometry of the data**

RANDOM GEOMETRIC GRAPHS

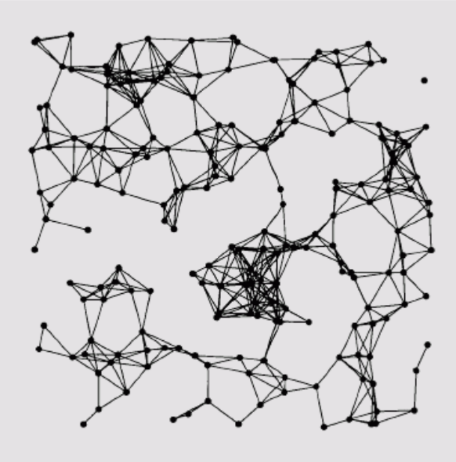


Figure: Visualization of a random geometric graph [35]

SCALED LAPLACIAN MATRIX

- For a graph $G = (\Omega_n, W_{n,\varepsilon})$, the definition of the graph Laplacian matrix is slightly different:

$$\Delta_{n,\varepsilon} = \frac{C}{n\varepsilon^2}\left(D_{n,\varepsilon} - W_{n,\varepsilon}\right)$$

  where $D_{n,\varepsilon}$ is the diagonal matrix with entries $d_{\varepsilon,ii} = \sum_{j=1}^n w_{\varepsilon,ij}$ and C is a constant that depends on η

- We note that

$$\frac{C}{n^2\varepsilon^2}\sum_{i,j=1}^n w_{\varepsilon,ij}\left(u_n(x_i) - u_n(x_j)\right)^2 = \langle u_n, \Delta_{\varepsilon,n} u_n\rangle_{\mathrm{L}^2(\mu_n)}$$

  where $\langle u_n, v_n\rangle_{\mathrm{L}^2(\mu_n)} = \frac{1}{n}\sum_{i=1}^n u_n(x_i)v_n(x_i)$ and μ_n is the empirical measure of Ω_n

THREE REASONS WHY $\varepsilon_n \to 0$

- **Geometry**: When $n \to \infty$, it is natural to let $\varepsilon_n \to 0$ as there is increasingly more local information available at each point which allows one to resolve the geometry in the graph at finer scales

- **Numerics**: The numerical cost correlates with the number of neighbours (or the density of the matrix $W_{n,\varepsilon}$) so scaling $\varepsilon_n \to 0$ has the advantage of decreasing computation time

- **Analysis**: Scaling $\varepsilon_n \to 0$ allows us to replace the discrete objective $\mathcal{E}_{n,\varepsilon_n}$ based on finite differences with a continuum objective $\mathcal{E}_\infty$ based on derivatives

CONVERGENCE OF u_n THROUGH Γ-CONVERGENCE

- Functions u_n are all minimizers of functionals $\mathcal{E}_n$
- Framework of choice to deal with **convergence of minimizers of functionals** is Γ-convergence [4]
- Γ-convergence is a property of functionals
- We say that $\mathcal{E}_n$ Γ-converge to $\mathcal{E}_\infty$ in some metric space X if:
  - for all $x_n \to x$ in X, $\liminf_{n\to\infty} \mathcal{E}_n(x_n) \geq \mathcal{E}_\infty(x)$
  - for all $x \in X$, there exists $x_n \to x$ in X such that $\limsup_{n\to\infty} \mathcal{E}_n(x_n) \leq \mathcal{E}_\infty(x)$

CONVERGENCE OF MINIMIZERS

- Fundamental property of Γ-convergence: "compactness of $u_n = \operatorname{argmin} \mathcal{E}_n$ in X" + "Γ-convergence of functionals $\mathcal{E}_n$ to $\mathcal{E}_\infty$" = "convergence in X of u_n to $u_\infty = \operatorname{argmin} \mathcal{E}_\infty$"

- Similar to the direct method in calculus of variations is: "compactness of minimizing sequence" + "lower semi-continuity of functional" = "existence of minimizer"

$\Rightarrow$ We need to find a metric space X in which we can have convergence of u_n to u_∞.

DISCRETE AND CONTINUUM COMPARISONS I

- Intuition: as $n \to \infty$, the discrete sets $\Omega_n \subseteq \Omega$ "converge" to the continuum set Ω

$\Rightarrow$ It is therefore reasonable to assume that $\mathcal{E}_\infty$ is defined for functions $u : \Omega \to \mathbb{R}$

$\Rightarrow$ our metric space X must include functions defined on Ω_n and Ω

$\Rightarrow$ in order to define a metric on X, we need to **compare discrete functions u_n to continuum functions u**

DISCRETE AND CONTINUUM COMPARISONS II

- Since u is not necessarily regular, we need to compare $\int u \, \mathrm{d}\mu$ and $\int u_n \, \mathrm{d}\mu_n$ where μ_n is the empirical measure of Ω_n

$\Rightarrow$ The metric space X will be subset of $\{(\nu, v) \,|\, \nu$ is a measure on $\Omega, \; v \in \mathrm{L}^1(\nu)\}$

- Standard way to compare integrals w.r.t. different measures is through Optimal Transport

- Let $T_n : \Omega \to \Omega_n$ be a function that links μ to μ_n by satisfying the consistency condition $\mu_n(x_i) = \mu(T_n^{-1}\{x_i\})$ for all $x_i \in \Omega_n$

$\Rightarrow$ T_n "projects" Ω to Ω_n by conserving the measure of sets

- if there exists such T_n, we could consider $\int |u - u_n \circ T_n|^p \, \mathrm{d}\mu$ for the metric

SPECIAL CASES OF $\int |u - u_n \circ T_n|^p \, \mathrm{d}\mu \to 0$

- if u is regular enough and we set $u_n = u|_{\Omega_n}$, then

$$\text{“} \int |u - u_n \circ T_n|^p \, \mathrm{d}\mu \to 0 \Leftrightarrow T_n \to \mathrm{Id} \text{”}$$

- if $u = u_n = \mathrm{Id}$, then (by Optimal Transport theory)

$$\text{“} \int |u - u_n \circ T_n|^p \, \mathrm{d}\mu \to 0 \Leftrightarrow \mu_n \text{ converge weakly to } \mu \text{”}$$

$\Rightarrow$ Our convergence definition has to cover these two cases at least

- Does there exist a metric space in which convergence is characterized by all of the above?

METRIC SPACE FOR DISCRETE-TO-CONTINUUM ANALYSIS: THE TL^p-SPACE

- We define the TL^p-space [18] as follows:

$$\mathrm{TL}^p = \{(\nu, v) \,|\, \nu \in \mathcal{P}_p(\Omega), v \in \mathrm{L}^p(\nu)\}$$

- For $(\nu_1, v_1), (\nu_2, v_2) \in \mathrm{TL}^p$, we define the TL^p distance d_{TL^p}:

$$d_{\mathrm{TL}^p}((\nu_1, v_1), (\nu_2, v_2)) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\Omega \times \Omega} |x - y|^p + |u(x) - v(y)|^p \, \mathrm{d}\pi(x, y)$$

  where $\Pi(\nu_1, \nu_2)$ the set of all probability measures on $\Omega \times \Omega$ such that the first marginal is ν_1 and the second marginal is ν_2

METRIC IN TL^p-SPACE

- **d_{TL^p} is equal to the p-Wasserstein distance in a special case**: this allows one to deduce lots of properties of d_{TL^p}

- In particular, we can define convergence between $\{(\nu_n, v_n)\}_{n=1}^\infty$ and (ν, v) in the TL^p-space conveniently [18]:
  - There exists T_n such that $T_{n\#}\nu = \nu_n$ and $\|T_n - \mathrm{Id}\,\|_{\mathrm{L}^\infty} \to 0$
  - ν_n converges weakly to ν
  - $\|v - v_n \circ T_n\|_{\mathrm{L}^p} \to 0$

$\Rightarrow$ We recover **all the requirements** from above

STRATEGY FOR DISCRETE-TO-CONTINUUM ANALYSIS

1. We will consider $\{(\mu_n, u_n)\}_{n=1}^{\infty}$ and (μ, u_∞) as elements of $\mathrm{TL}^p(\Omega)$
   - We note that μ_n converges weakly to μ and, by (for example in Euclidean space) [18, Theorem 2.5], the appropriate transport maps T_n (between μ_n and μ) exist, so showing TL^p-convergence is equivalent to $\|u - u_n \circ T_n\|_{\mathrm{L}^p} \to 0$
2. We (naturally extend $\mathcal{E}_n$ and $\mathcal{E}_\infty$ to TL^p and) show that $\mathcal{E}_n$ Γ-converges to $\mathcal{E}_\infty$ in $\mathrm{TL}^p(\Omega)$
3. We show that $\{(\mu_n, u_n)\}_{n=1}^{\infty}$ is pre-compact and therefore deduce that its limit points are the minimizer(s) of $\mathcal{E}_\infty$

FRACTIONAL LAPLACIAN LEARNING

REGULARITY THROUGH ASYMPTOTIC ANALYSIS

- **Reminder**: functions u_n defined on the discrete set Ω_n are supposed to help in the SSL problem and should satisfy:
  - for all n, $u_n(x_i) = \ell_i$ for all $i \leq N$
  - if the geometry is well-captured in the graph, then $x_j \approx x_k$ implies $u_n(x_j) \approx u_n(x_k)$, i.e. we have some regularity
- Regularity in discrete setting is not so convenient to define, but it is easy in the continuum domain
- **Ideally**: if the problem is well-posed, the functions u_n converge to some u_∞ which is regular and satisfies $u_\infty(x_i) = \ell_i$ for all $i \leq N$

$\Rightarrow$ We would like to have an objective function $\mathcal{E}_\infty$ whose **minimizers are at least continuous in the well-posed case**

DERIVATION OF $\mathcal{E}_\infty$ FOR LAPLACE LEARNING

Let us pick $\eta = \mathbb{1}_{[0,1]}$, ρ the uniform density and $u \in \mathrm{C}^\infty(\Omega)$ with $\frac{\partial u}{\partial n} = 0$ on $\partial\Omega$:

$$\langle u, \Delta_{n,\varepsilon_n} u \rangle_{\mathrm{L}^2(\mu_n)} = \frac{C}{n^2 \varepsilon_n^2} \sum_{i,j=1}^{n} \frac{1}{\varepsilon_n^d} \mathbb{1}_{\{|x_i - x_j| \leq \varepsilon_n\}} \left(u(x_i) - u(x_j)\right)^2$$

$$\xrightarrow{n\to\infty} \frac{C}{\varepsilon_n^{2+d}} \int_\Omega \int_\Omega \mathbb{1}_{\{|y-x|\leq \varepsilon_n\}} (u(y) - u(x))^2 \, \mathrm{d}y \mathrm{d}x$$

$$= \frac{C}{\varepsilon_n^2} \int_\Omega \int_{\{|z|\leq 1\}} (u(x + \varepsilon_n z) - u(x))^2 \, \mathrm{d}z \mathrm{d}x$$

$$= C \int_\Omega |\nabla u(x)|^2 \int_{\{|z|\leq 1\}} z^2 \, \mathrm{d}z \mathrm{d}x + o(\varepsilon_n)$$

$$\xrightarrow{\varepsilon_n \to 0} \int_\Omega |\nabla u(x)|^2 \, \mathrm{d}x = \langle u, \Delta u \rangle_{\mathrm{L}^2(\mu)} \quad \text{where } \Delta \text{ is Laplace operator}$$

REGULARITY CONSIDERATIONS OF LAPLACE LEARNING

- For general η and ρ, $\mathcal{E}_\infty(u) = \langle u, \Delta_\rho u \rangle_{\mathrm{L}^2(\mu)}$ where $\Delta_\rho = -\frac{1}{\rho(x)}\mathrm{div}(\rho^2 \nabla u)$ is the weighted Laplace operator and ρ is the density of μ with respect to Lebesgue measure
- Minimizers u_∞ of $\mathcal{E}_\infty$ are in the Sobolev space $\mathrm{W}^{1,2}(\Omega)$
- In order to get at least continuity, by Sobolev embeddings:

$$u_\infty \begin{cases} \text{is continuous} & \text{if } d = 1 \\ \text{is only in } \mathrm{W}^{1,2} & \text{if } d > 1 \end{cases}$$

$\Rightarrow$ Very **constraining** in practice!

LAPLACE LEARNING WHEN $d > 1$

- When $d > 1$, the solution of Laplace learning for large n is almost constant and is not useful for SSL

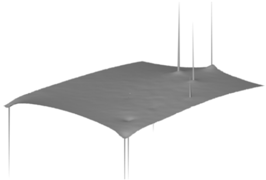


Figure: Spikes in Laplace learning [9]

- It is shown in [39] that in this case, u_n converges (in TL^2) to the minimizers of $\mathcal{E}_\infty = \int_\Omega |\nabla u(x)|^2 \,\mathrm{d}x$ without the pointwise constraints, i.e. a constant

HIGHER ORDER LAPLACE LEARNING

- Recall from general Sobolev inequalities, $\mathrm{W}^{k,p}(\Omega)$ **is embedded in a space of continuous functions if** $k > d/p$

- Other variational problems on graphs have been proposed where the limiting functional $\mathcal{E}_\infty$ **is a higher order Sobolev seminorm** and the conditions to obtain continuous minimizers are less constraining:

  - Pick $k = 1$, but let $p > 1$ (p-Laplacian learning [39] if $p < \infty$ and Lipschitz learning [7] if $p = \infty$)

  - Pick $p = 2$, but let $k > 1$ (fractional Laplacian learning [14]): we will write s instead of k to emphasize that we can pick $s \in \mathbb{R}$ instead of $k \in \mathbb{N}$

p-LAPLACIAN VERSUS FRACTIONAL LAPLACIAN

Attributes	p-Laplacian	fractional Laplacian
Discrete energy	$\frac{1}{n^2 \varepsilon_n^{p+d}} \sum_{i,j=1}^n w_{\varepsilon_n,ij} \lvert u_n(x_i) - u_n(x_j) \rvert^p$	$\langle v, \Delta_{n,\varepsilon_n}^s v \rangle_{\mathrm{L}^2(\mu_n)}$
Solution in SSL	Approximate	Exact
Computation method	Gradient descent	Lagrange multipliers

- Note that Laplace learning is 2-Laplacian learning and fractional Laplacian learning with $s = 1$
- Characterization of well-posed and ill-posed regimes in p-Laplacian learning has been proven in [39]
- Our work deals with the **characterization of the well-posed and ill-posed regimes** in fractional Laplacian learning [43]

FRACTIONAL LAPLACIAN REGULARIZATION IN SSL

- For $s > 0$, in the discrete case, we look for:

$$u_n \in \operatorname*{argmin}_{v:\Omega_n \to \mathbb{R}} \langle v, \Delta^s_{n,\varepsilon_n} v \rangle_{\mathrm{L}^2(\mu_n)} \quad \text{such that } v(x_i) = \ell_i \text{ for } i \leq N$$

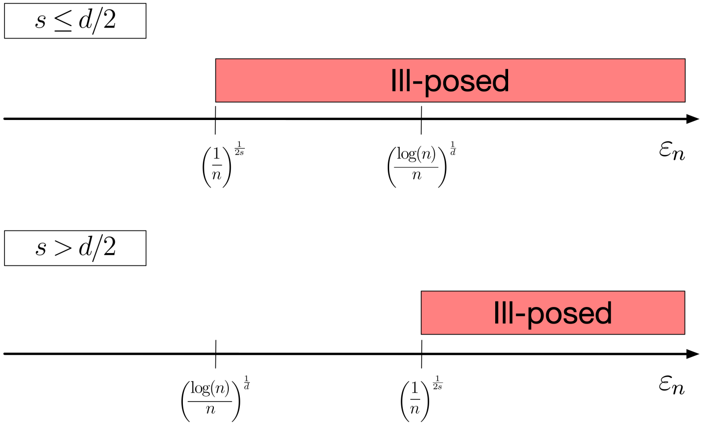
  and we define $\mathcal{E}^{(s)}_{n,\varepsilon_n}(v) = \langle v, \Delta^s_{n,\varepsilon_n} v \rangle_{\mathrm{L}^2(\mu_n)}$

- For $s > 0$, in the continuum, we look for:

$$u_\infty \in \begin{cases} \operatorname{argmin}_{v \in \mathrm{W}^{s,2}(\Omega)} \langle v, \Delta^s_\rho v \rangle_{\mathrm{L}^2(\mu)} & \text{such that } v(x_i) = \ell_i \text{ for } i \leq N, \\ \operatorname{argmin}_{v \in \mathrm{W}^{s,2}(\Omega)} \langle v, \Delta^s_\rho v \rangle_{\mathrm{L}^2(\mu)} \end{cases}$$

- $\mathcal{E}^{(s)}_\infty(v) = \langle v, \Delta^s_\rho v \rangle_{\mathrm{L}^2(\mu)}$ is **equivalent to a $\mathrm{W}^{s,2}$-seminorm** [14] (here we consider fractional Sobolev spaces)

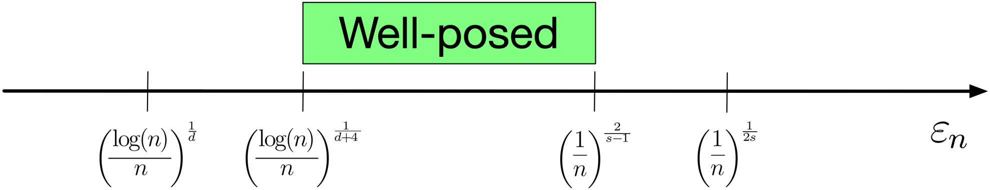
ILL-POSEDNESS CHARACTERIZATION



WELL-POSEDNESS CHARACTERIZATION

- Let α be a constant that determines how much control one has of the L^∞-norm of the discrete eigenvectors $\|\psi_{\varepsilon_n,n,k}\|_{\mathrm{L}^\infty}$ in terms of continuum eigenvalues λ_k for small k, i.e. $\|\psi_{\varepsilon_n,n,k}\|_{\mathrm{L}^\infty} \le C\lambda_k^\alpha$

$$s > \max\{2\alpha + 2 + d/2, 2d + 9\}$$



GEOMETRIC INTERPRETATION OF BOUNDS ON ε_n

- Lower bound: ε_n cannot go to 0 too quickly and it has to be higher than the connectivity threshold of the random geometric graph [35]: $\left(\frac{\log(n)}{n}\right)^{1/d} \ll \left(\frac{\log(n)}{n}\right)^{1/(d+4)} \ll \varepsilon_n$
- Upper bound: we need: $\varepsilon_n \ll \left(\frac{1}{n}\right)^{2/(s-1)}$
- Intuition of the bounds:
  - Lower bound: in order to **capture the geometry of Ω properly**, we need the graph to be connected $\Rightarrow$ intuitive
  - Upper bound: graph **cannot be too densely connected** $\Rightarrow$ more surprising

INTUITION FOR $s > \max\{2\alpha + 2 + d/2, 2d + 9\}$

- We show that we can pick $\alpha = d + 1$ on the flat torus $\mathbb{R}^d \setminus \mathbb{Z}^d$ but, we conjecture that in this setting, actually $\alpha = 0$

- We also believe that the "$+2$" part of $s > 2\alpha + 2 + d/2$ is an artifact of our proof, which if removed (and if $\alpha = 0$) would yield the intuitive condition $s > d/2$ from Sobolev embeddings

- The "$s > 2d + 9$" requirement follows from the fact that we have

$$\left(\frac{\log(n)}{n}\right)^{1/(d+4)} \ll \varepsilon_n \ll \left(\frac{1}{n}\right)^{2/(s-1)}$$

- For the latter to be consistent we need $s/2 - 1/2 > d + 4$ or $s > 2d + 9$

LINK BETWEEN SOBOLEV SPACE INTUITION AND CHARACTERIZATION

- Sobolev space intuition is **relevant**: the ill-posed case is partly characterized by the setting where $\mathrm{W}^{s,2}$ is not embedded in continuous functions, i.e. $s \leq d/2$

- Sobolev space intuition is **not sufficient**: even when $\mathrm{W}^{s,2}$ is embedded in continuous functions, i.e. $s > d/2$, if the graph is too connected, we are still in the ill-posed regime

GAPS IN THE CHARACTERIZATION

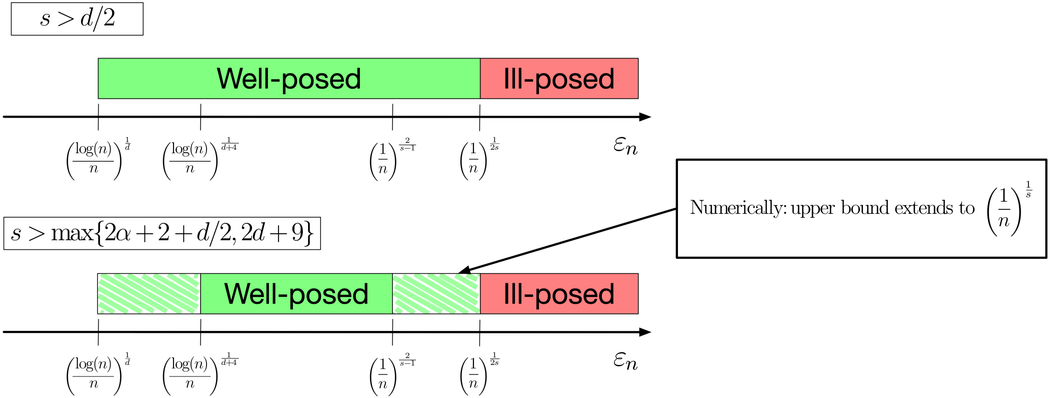


Figure: Conjectured versus proven characterization

OVERVIEW OF THE PROOF: COMPACTNESS I

- Compactness of minimizers in TL^2 (and therefore for the ill-posed case) is proven in [14]
- For the well-posed case, we need to show that there exists a continuous function u_∞ such that $\max_{k \leq n} |u_n(x_k) - u_\infty(x_k)| \to 0$, which ensures that u_∞ satisfies the pointwise constraints
- Using the discrete eigenpairs of Δ_{n,ε_n} to represent u_n, we show Lipschitz regularity of u_n through spectral convergence results between Δ_{n,ε_n} and Δ_ρ

$\Rightarrow$ We deduce equicontinuity of $\tilde{u}_n$ where $\tilde{u}_n = J_{\varepsilon_n} * (u_n \circ T_n)$ and J_{ε_n} is a scaled mollifier

$\Rightarrow$ Through the Ascoli-Arzela theorem, we have that $\tilde{u}_n$ converges uniformly to some u_∞

OVERVIEW OF THE PROOF: COMPACTNESS II

- We write

$$|u_n(x_k) - u_\infty(x_k)| \leq |u_n(x_k) - \tilde{u}_n(x_k)| + |\tilde{u}_n(x_k) - u_\infty(x_k)| =: T_1 + T_2$$

$\Rightarrow$ $T_1 \to 0$ since $\|\mathrm{Id} - T_n\|_{\mathrm{L}^\infty} \to 0$

$\Rightarrow$ $T_2 \to 0$ by uniform convergence

OVERVIEW OF THE PROOF: Γ-CONVERGENCE

- $\liminf$-inequality
  - Well-posed case: depends on discrete Sobolev inequality and above compactness result
  - Ill-posed case: depends on [14]
- $\limsup$-inequality
  - Well-posed case: depends on a bound for the discrete Sobolev seminorm $\mathcal{E}_{n,\varepsilon_n}^{(s)}$ (the discrete $\mathrm{W}^{s,2}$-version of continuum results for $\mathrm{W}^{1,p}$ in [3])
  - Ill-posed case: depends on [14]

RATES OF CONVERGENCE FOR p-LAPLACIAN REGULARIZATION

DISCRETIZING $\mathrm{W}^{s,2}$ SEMINORMS ON GRAPHS

- Results in [43] explain how to appropriately approximate $\mathrm{W}^{s,2}$ seminorms on random geometric graphs by tuning ε_n

- A discretization based on random geometric graphs is convenient: mesh-free and not (so) parametric, i.e. has potential to be easily implemented numerically

$\Rightarrow$ Can we approximate other seminorms on graphs?

DISCRETIZING $\mathrm{W}^{1,p}$ SEMINORMS ON GRAPHS

- Similarity with the $\mathrm{W}^{s,2}$ case: results in [39] explain how to appropriately approximate $\mathrm{W}^{1,p}$ seminorms on random geometric graphs by tuning ε_n

- Difference with the $\mathrm{W}^{s,2}$ case: the proof of $\mathrm{W}^{s,2}$-case relies on eigenpair decomposition and **spectral convergence** results between discrete and continuum operators while the proof of $\mathrm{W}^{1,p}$-case relies on a **nonlocal continuum approximation**

APPROXIMATING $\mathrm{W}^{1,p}$ SEMINORMS IN THE CONTINUUM

- In [3] we find a characterization of $\mathrm{W}^{1,p}$ through the boundedness of the nonlocal formula

$$\int_\Omega \int_\Omega \frac{|f(x) - f(y)|^p}{|x - y|^p} \eta_n(x - y) \, \mathrm{d}x \mathrm{d}y \tag{1}$$

for some kernels η_n

- Modulo slightly changing the kernel, p-Laplacian learning is a discretization of (1):

$$\int_\Omega \int_\Omega |f(x) - f(y)|^p \eta_n(x - y) \, \mathrm{d}x \mathrm{d}y \xrightarrow{\text{discretized}} \frac{1}{n^2} \sum_{i,j=1} \eta_n(x_i - x_j) |f(x_i) - f(x_j)|^p$$

NONLOCAL APPROXIMATIONS ARE CONVENIENT

- Advantage 1, a **straight-forward proof strategy**:
  - discrete $\to$ continuum nonlocal $\to$ continuum local (i.e. the Sobolev seminorm)
  - this is what inspired the proofs in [18] (1-Laplacian learning) and [39] (p-Laplacian learning with $p > 1$)
- Advantage 2, **conceptually simple rates of convergence**:
  - for smooth enough functions, finite-differences can be replaced by derivatives
  - this is similar to how we derived $\mathcal{E}_\infty^{(1)}$, i.e. Laplace learning

p-LAPLACIAN REGULARIZATION PROBLEM

- We want to compute

$$u_\infty \in \underset{v \in \mathrm{W}^{1,p}(\Omega)}{\operatorname{argmin}} \mathcal{F}(v) := \frac{\mu}{p} \|\nabla v\|^p_{\mathrm{L}^p(\Omega)} + \frac{1}{2}\|v - \ell\|^2_{\mathrm{L}^2(\Omega)}$$

- Ideally, we establish **rates of convergence** between a discrete (numerical) solution and u_∞ [42]

- We follow the general strategy: discrete $\to$ continuum nonlocal $\to$ continuum local

STEP 1: LOCAL OPTIMIZATION TO LOCAL GRADIENT FLOW

- Our approach is to consider the gradient flow associated to the optimization problem above:

$$\begin{cases} \frac{\partial}{\partial t} u(t,x) + \mu \Delta_p u(t,x) + u(t,x) = \ell(x), & \text{on } \Omega \times (0,T) \\ |\nabla u(t,x)|^{p-2} \nabla u(t,x) \cdot \overrightarrow{n} = 0, & \text{on } \partial\Omega \times (0,T) \\ u(0,x) = u_0(x) \end{cases}$$

where $\Delta_p u = -\operatorname{div}(|\nabla u|^{p-2} \nabla u)$ is the p-Laplacian operator

STEP 2: LOCAL TO NONLOCAL GRADIENT FLOW

- We approximate Δ_p by the nonlocal p-Laplacian operator

$$\Delta_p^{\varepsilon_n,\eta} u(x) = -\frac{C}{\varepsilon_n^{d+p}} \int_\Omega \eta\left(\frac{|x-y|}{\varepsilon_n}\right) |u(y)-u(x)|^{p-2}(u(y)-u(x))\,\mathrm{d}y$$

  for some η

- This yields the nonlocal gradient flow:

$$\begin{cases} \frac{\partial}{\partial t} u + \mathcal{A}_\ell^{\varepsilon_n}(u) = 0, & \text{on } \Omega \times (0,T) \\ u(0,x) = u_0(x) \end{cases}$$

  where $\mathcal{A}_\ell^{\varepsilon_n}(u) = \mu \Delta_p^{\varepsilon_n,\eta} u + u - \ell$

STEP 3: NONLOCAL TO DISCRETE GRADIENT FLOW

- We approximate $\Delta_p^{\varepsilon_n,\eta}$ by the discrete p-Laplacian operator

$$(\Delta_{p,n}^{\varepsilon_n} u_n)(x_i) = -\frac{C}{n\varepsilon_n^{d+p}} \sum_{j=1}^{n^d} w_{ij} |u_n(x_j) - u_n(x_i)|^{p-2} (u_n(x_j) - u_n(x_i))$$

  for some weights w_{ij}

- With a partition $0 = t^0 < t^1 < \cdots < t^N = T$ and where $\tau^{k-1} = t^k - t^{k-1}$, this yields the discrete gradient flow:

$$\begin{cases} \frac{u_n^k - u_n^{k-1}}{\tau^{k-1}} + \mu \Delta_{p,n}^{\varepsilon_n} u_n^k + u_n^k = (\ell)_n, & \text{for} \quad 1 \leq k \leq N \\ u_n(0) = (u_0)_n \end{cases}$$

  where $(\ell)_n$ is a discretization of ℓ and $(u_0)_n$ is a discretization of u_0

COMMENTS ON DISCRETE-TO-CONTINUUM COMPARISONS

- We partition our space Ω in n^d cells π_i

- We define the projection operator $\mathcal{P}_n : \mathrm{L}^1(\Omega) \mapsto \mathbb{R}^{n^d}$ and the injection operator $\mathcal{I}_n : \mathbb{R}^{n^d} \mapsto \mathrm{L}^1(\Omega)$ as

$$(\mathcal{P}_n u)_i = \frac{1}{|\pi_i|} \int_{\pi_i} u(x)\,\mathrm{d}x \quad \text{and} \quad (\mathcal{I}_n u_n)(x) = \sum_{i=1}^{n^d} u_n \mathbb{1}_{\pi_i}(x)$$

  respectively for $u \in \mathrm{L}^1(\Omega)$ and $i = 1, \ldots, n^d$, $u_n \in \mathbb{R}^{n^d}$ and $x \in \Omega$.

- For example, $(u_0)_n = \mathcal{P}_n u_0$

- Also, $\|\mathcal{I}_n \mathcal{P}_n u_0 - u_0\|_{\mathrm{L}^2(\Omega)}$ depends on the regularity of u_0 and the partition of Ω [12]

RATES OF CONVERGENCE I

- For some $\kappa > 0$, we now set $T = \log(\varepsilon_n^{-\kappa})$, pick $0 = t^0 < t^1 < \cdots < t^{N(n)} = T$ and let τ_n be the maximum step-size of the time-discretization

- We find that for $p \geq 3$,

$$\|\mathcal{I}_n u_n^N - u_\infty\|_{\mathrm{L}^2} \leq C\Bigg(\varepsilon_n^{\kappa/4}(\mathcal{F}(u_0) - \mathcal{F}(u_\infty))^{1/2} + \varepsilon_n \log(\varepsilon_n^{-\kappa})$$

$$+ \varepsilon_n^{-\kappa}\left[\tau_n \frac{\log(\varepsilon_n^{-\kappa})^{2p-3}}{\varepsilon_n^{2(d+p)}} + n^{-\alpha_1} + n^{-\alpha_2} + \frac{\log(\varepsilon_n^{-\kappa})^{(p-1)}}{\varepsilon_n^{d+p+\alpha_3} n^{\alpha_3}}\right]\Bigg)$$

where, $C > 0$ is a constant independent of n, $\kappa > 0$ and $\alpha_i > 0$ are chosen constants depending on the regularity of the initial condition u_0, the data ℓ and the kernel η.

RATES OF CONVERGENCE II

- We note that **each term in the rates corresponds to an approximation step**, namely (from left to right) the gradient flow convergence, the continuum nonlocal-to-local approximation, the discrete-to-continuum nonlocal approximation and discrete-to-continuum approximation of u_0, ℓ and η

- For the error to go to 0,
  - we obtain results similar to CFL-conditions: the time discretization τ_n has to be controlled by the space discretization ε_n
  - ε_n admits a lower bound

RATES OF CONVERGENCE III

- We also show that we can discretize our problem on a random graph models inspired by the study of graphons
- This implies a **random-to-deterministic approximation** in the discrete setting and yields an additional term in the rates of convergence

OVERVIEW OF WELL-POSEDNESS PROOF

- The existence of a solution to the local gradient flow follows from **nonlinear PDE results** [28]
- For the nonlocal gradient flow, we consider the abstract Cauchy problem:
$$\begin{cases} \frac{\partial}{\partial t}u + \mathcal{A}_\ell^{\varepsilon_n}(u) = 0, \quad \text{on } \Omega \times (0,T) \\ u(0,x) = u_0(x) \end{cases}$$

$\Rightarrow$ we show complete accretivity (i.e. a generalization of maximal monotony in Banach spaces) of $\mathcal{A}_\ell^{\varepsilon_n}$ as well as range condition

$\Rightarrow$ we can apply existence results from **nonlinear semigroup theory in Banach spaces** to get solution in terms of semigroups [34]

OVERVIEW OF RATES PROOF I

- For optimization-to-gradient flow rates: standard rates based on convexity
- For continuum nonlocal-to-local gradient flow rates: one needs to consider the error between Δ_p and $\Delta_p^{\varepsilon_n,\eta}$ applied to a regular function and this relies on Taylor expansions

OVERVIEW OF RATES PROOF II

- For discrete-to-continuum gradient flow rates:
  - we show that a time interpolation of $\mathcal{I}_n u_n$ solves a nonlocal gradient flow problem with parameters $\mathcal{I}_n\mathcal{P}_n u_0$, $\mathcal{I}_n\mathcal{P}_n \ell$ and $\mathcal{I}_n\mathcal{P}_n\eta$

  $\Rightarrow$ we use the continuum well-posedness results to obtain a solution in terms of semigroups

  - by considering the error between $\Delta_p^{\varepsilon_n,\eta}$ and $\Delta_p^{\varepsilon_n,\mathcal{I}_n\mathcal{P}_n\eta}$ and contraction properties of semigroups, we obtain: for solutions u_{ε_n} and $\mathcal{I}_n u_n$ to our nonlocal gradient flow with respective parameters u_0, ℓ, η and $\mathcal{I}_n\mathcal{P}_n u_0$, $\mathcal{I}_n\mathcal{P}_n\ell$, $\mathcal{I}_n\mathcal{P}_n\eta$, we have

$$\|u_{\varepsilon_n}(t,\cdot) - \mathcal{I}_n u_n(t,\cdot)\|_{\mathrm{L}^2} \leq Ce^T(\tau_n \frac{T^{2p-3}}{\varepsilon_n^{2(d+p)}} + \|u_0 - \mathcal{I}_n\mathcal{P}_n u_0\|_{\mathrm{L}^2}$$
$$+ \|\ell - \mathcal{I}_n\mathcal{P}_n\ell\|_{\mathrm{L}^2} + \frac{T^{(p-1)}}{\varepsilon_n^{d+p}}\|\eta(\cdot/\varepsilon_n) - \mathcal{I}_n\mathcal{P}_n\eta(\cdot/\varepsilon_n)\|_{\mathrm{L}^2})$$

HYPERGRAPH LEARNING

HYPERGRAPH SETTING

- A hypergraph G is defined as $G = (V, E)$ where V is a set of objects and E a family of subsets e of V with $|e| \geq 2$ (in our case, $V = \Omega_n$)

- **Intuition**: since $|e| \geq 2$, we capture higher order relationships between samples, e.g. similarity of researchers based on paper authorship

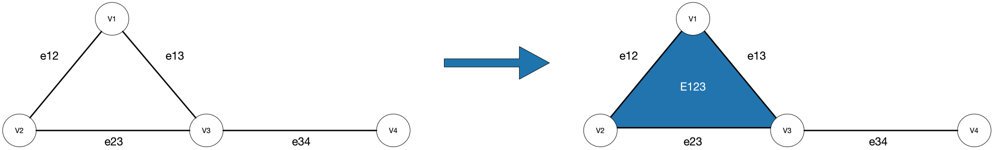


Figure: From graphs to hypergraphs

RELEVANCE OF HYPERGRAPHS

- Learning on hypergraphs is developed in [45, 15, 27]

$\Rightarrow$ How similar are these methodologies with their graph analogues [25, 31, 24, 11]?

- **Ideally**: Hypergraphs should be valuable geometrical models for data compared to graphs due to their additional structure [44, 33]

HYPERGRAPH LEARNING

- The equivalent of Laplace learning on hypergraphs is introduced in [45]

- The idea is to consider the solution to

$$\operatorname*{argmin}_{v:\Omega_n\to\mathbb{R}} \sum_{e\in E} \sum_{\{x_i,x_j\}\subseteq e} \frac{w_0(e,x_i,x_j)}{|e|}(v(x_i)-v(x_j))^2 \text{ such that } v(x_i)=y_i \text{ for } i\leq N \tag{2}$$

  where w_0 is the hyperedge weight function

- **Key observation**: for each hyperedge e, we penalize the smoothness of v between each pair or vertices $\{x_i,x_j\}\subseteq e$

HYPERGRAPH DECOMPOSITION

- Let $q = \max_{e \in E} |e| - 1 \leq n - 1$ and
  $E^{(k)} = \{\{x_i, x_j\} \mid \text{there exists } e \in E \text{ with } |e| = k + 1 \text{ and } \{x_i, x_j\} \in e\}$

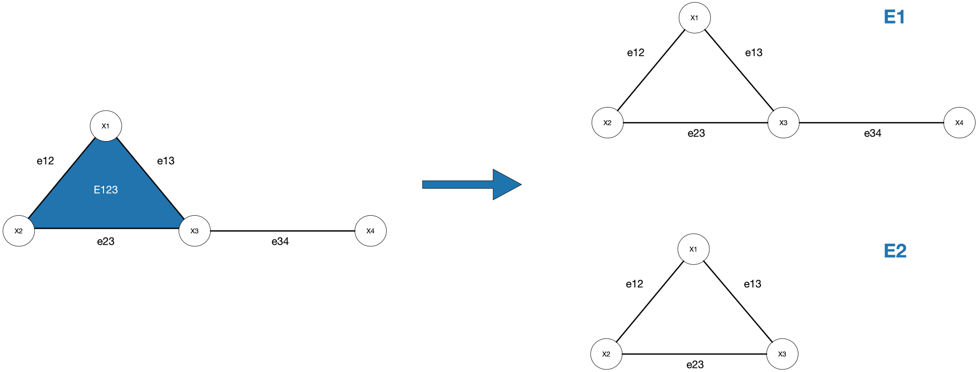


Figure: Skeleton graphs

HYPERGRAPH LEARNING AS A SUM OF LAPLACE LEARNING ON SUBGRAPHS

- **Idea**: order the hyperedges by size in the hypergraph energy (2):

$$\sum_{k=1}^{q} \frac{w_1(k+1)}{k+1} \sum_{\{x_i,x_j\}\in E^{(k)}} w_2(x_i,x_j)(v(x_i)-v(x_j))^2$$

  for some functions w_1, w_2

- **Intuition**: the hypergraph structure can be rewritten as sequence of subgraphs $(V, E^{(k)})$ and hypergraph learning is the sum of Laplace learning on each of theses subgraphs

$\Rightarrow$ We want to perform asymptotic consistency analysis for this model

RANDOM GEOMETRIC HYPERGRAPH WEIGHT MODEL

- For hyperedges of degree $k+1$, we define weights as

$$\prod_{j=1}^{k}\prod_{r=0}^{j-1}\eta\left(\frac{|x_{i_j}-x_{i_r}|}{\varepsilon}\right)$$

- **Intuition**: Whenever η is $\mathbb{1}_{[0,1]}$, the weights are different from 0 if and only if all the $x_{i_0},\dots,x_{i_k}$ are all within the same ball of radius ε

- For $k=1$, this corresponds to the random geometric graph

UPDATED HYPERGRAPH LEARNING OBJECTIVE

- For a fixed $q \geq 1$ and positive coefficients λ_k, we define the hypergraph learning problem as

$$\underset{v_n:\Omega_n \mapsto \mathbb{R}}{\operatorname{argmin}} \sum_{k=1}^{q} \lambda_k \frac{1}{n^{k+1}\varepsilon^{p+kd}} \sum_{i_0,\cdots,i_k=1}^{n} \left[\prod_{j=1}^{k}\prod_{r=0}^{j-1} \eta\left(\frac{|x_{i_j} - x_{i_r}|}{\varepsilon}\right)\right] |v_n(x_{i_1}) - v_n(x_{i_0})|^p \tag{3}$$

  such that $v_n(x_i) = y_i$ for $i \leq N$

- For $q = 1$, this is p-Laplacian learning [39] (and with $q = 1$, $p = 2$, this is Laplace learning)

- **Intuition**: the term $|v(x_{i_1}) - v(x_{i_0})|^p$ is strongly accounted for if all the $x_{i_0}, \ldots, x_{i_k}$ are all **very close**, i.e. in the same hyperedge of size $k+1$

$\Rightarrow$ We emphasize **regularity** more than just on graphs

ASYMPTOTIC CONSISTENCY ANALYSIS

- If $\left(\frac{\log(n)}{n}\right)^{1/d} \ll \varepsilon_n \ll \left(\frac{1}{n}\right)^{1/p}$, then hypergraph learning is well posed and its minimizers converge to the minimizers of

$$\sum_{k=1}^{q} \lambda_k \sigma_\eta^{(k)} \int_\Omega \|\nabla v(x_0)\|_2^p \, \rho(x_0)^{k+1} \, \mathrm{d}x_0 \quad \text{such that } v(x_i) = y_i \text{ for } i \leq N \tag{4}$$

- If $\left(\frac{1}{n}\right)^{1/p} \ll \varepsilon_n$, then hypergraph learning is ill-posed, i.e. its minimizers converge to the minimizers of (4) without pointwise constraints, i.e. constants

- The **closeness of points** captured by hyperedges of size $k+1$ is translated into a power of ρ and high-density regions will be particularly taken into account, i.e. the gradient of v will be small on the latter

- **Observation**: Hypergraph learning is a reweighted variant of p-Laplacian learning and we still only penalize the p-norm of the first derivative of v

OVERVIEW OF THE PROOF

- The proof is based on Γ-convergence and compactness in TL^p-space
- We also use the discrete $\rightarrow$ continuum nonlocal $\rightarrow$ continuum local strategy

HYPERGRAPH LEARNING LAPLACIANS I

- We can show that u minimizing (3) satisfies $\sum_{k=1}^{q} \lambda_k \Delta_{n,\varepsilon_n}^{(k,p)}(u) = 0$ where

$$\Delta_{n,\varepsilon}^{(k,p)}(u)(x_{i_0}) = \frac{1}{n^k \varepsilon^{p+kd}} \sum_{i_1,\dots,i_k=1}^{n} \left[\left[\prod_{j=1}^{k} \prod_{r=0}^{j-1} \eta\left(\frac{|x_{i_j} - x_{i_r}|}{\varepsilon} \right) \right] \right.$$
$$\left. \times |u(x_{i_1}) - u(x_{i_0})|^{p-2} (u(x_{i_1}) - u(x_{i_0})) \right]$$

- For $k = 1$, this is the p-Laplacian operator on graphs

HYPERGRAPH LEARNING LAPLACIANS II

- We get the following **pointwise consistency** result with high probability depending on ε_n and δ:

$$\left| \left(\sum_{k=1}^{q} \lambda_k \Delta_{n,\varepsilon_n}^{(k,p)} \right) (u)(x_{i_0}) - \left(\sum_{k=1}^{q} \lambda_k \Delta_{\infty}^{(k,p)} \right) (u)(x_{i_0}) \right| \leq \mathcal{O}\left(\delta \|u\|_{C^3(\mathbb{R}^d)} \right)$$

where

$$\Delta_{\infty}^{(k,p)}(u)(x_{i_0}) = \Bigg(\|\nabla u(x_{i_0})\|_2^{p-2} \rho(x_{i_0})^k \nabla \rho(x_{i_0}) \cdot \nabla u(x_{i_0}) \times \frac{2(\sigma_\eta^{(k,p)} + (k-1)\sigma_\eta^{(k,p,2)})}{(p-1)\sigma_\eta^{(k,p,1)}}$$
$$+ \rho(x_{i_0})^{k+1} \|\nabla u(x_{i_0})\|_2^{p-2} \Bigg[\mathrm{Tr}(\nabla^2 u(x_{i_0})) + (\frac{\sigma_\eta^{(k)}}{\sigma_\eta^{(k,p,1)}} - 1)$$
$$\times \frac{\nabla u(x_{i_0})^T \nabla^2 u(x_{i_0}) \nabla u(x_{i_0})}{\|\nabla u(x_{i_0})\|_2^2} \Bigg] \Bigg) \frac{\sigma_\eta^{(k,p,1)}(p-1)}{2\rho(x_{i_0})}$$

and constants $\sigma_\eta^{(k,p)}$, $\sigma_\eta^{(k,p,1)}$ and $\sigma_\eta^{(k,p,2)}$

HYPERGRAPH LEARNING LAPLACIANS III

- For $k = 1$, this simplifies to the weighted p-Laplacian operator

$$\Delta_\infty^{(1,p)}(u)(x_{i_0}) = \frac{\sigma_\eta^{(1,p)}}{2\rho(x_{i_0})}\mathrm{div}(\|\nabla u(x_{i_0}))\|_2^{p-2}\nabla u(x_{i_0})\rho(x_{i_0})^2)$$

- We note that the continuum Laplacian of

$$\int_\Omega \|\nabla v(x_0)\|_2^p \, \rho(x_0)^{k+1} \, \mathrm{d}x_0$$

is different from $\Delta_\infty^{(k,p)}(u)$

$\Rightarrow$ This is quite unique for these type of problems and pointwise consistency is not sufficient for discrete-to-continuum analysis in this case!

HIGHER ORDER HYPERGRAPH LEARNING

- We propose the higher order hypergraph learning energy

$$\sum_{k=1}^{q} \lambda_k \langle v, (L_n^{(k)})^k v \rangle_n = \langle v, \sum_{k=1}^{q} \lambda_k (L_n^{(k)})^k v \rangle_n =: \mathcal{F}(u)$$

where $L_n^{(k)}$ is the (regular) Laplacian of the skeleton graphs $G_n^{(k)} = (\Omega_n, E_n^{(k)})$

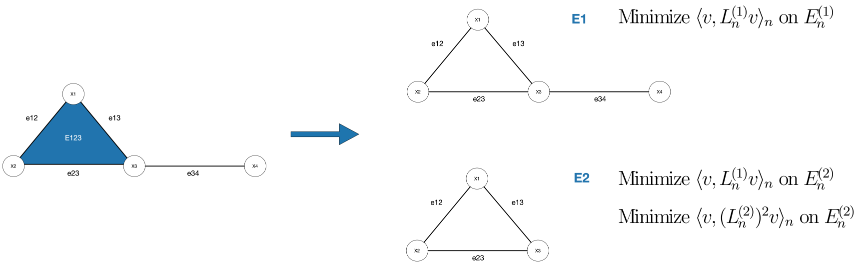


Figure: Higher order hypergraph learning is based on the graph decomposition

INSIGHTS FROM ASYMPTOTIC CONSISTENCY ANALYSIS

- Informally: $L_n \to \Delta$ as $n \to \infty$ where Δ is a weighted Laplace operator

$\Rightarrow$ This implies $\langle v, L_n v\rangle_n \to \int_{\mathbb{R}^d} |\nabla v|^2 \, dx$

$\Rightarrow$ With powers: $\langle v, (L_n)^k v\rangle_n \to \int_{\mathbb{R}^d} |\nabla^k v|^2 \, dx$

- **Intuition** behind (discrete) higher order hypergraph learning: we penalize the k-th derivative of our function on hyperedges of degree $k+1$

LINK BETWEEN HYPEREDGES AND DENSITY

- **Recall**: very close samples $\Rightarrow$ high degree of hyperedge

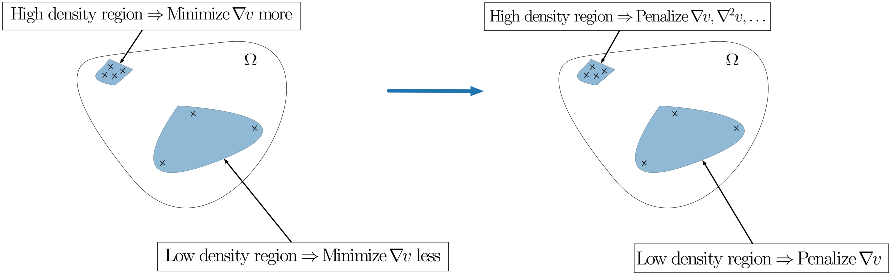


Figure: From hypergraph learning to higher order hypergraph learning

- **Asymptotic Consistency Analysis**: while hypergraph learning converges to a $\mathrm{W}^{1,p}$-seminorm, higher order hypergraph learning should converge to $\mathrm{W}^{q,2}$ norm [43, 14]

MULTISCALE LAPLACE LEARNING

- Higher order Laplace learning corresponds to multiscale Laplace learning [29] on point clouds
- In the latter, subgraphs are constructed directly (without hyperedges)
- However:
  - The hyperedge approach through locality theoretically justifies increasing powers on the Laplacian matrices
  - Higher order Laplace learning can also be formulated on **non point cloud** datasets, i.e. with an inherent hypergraph structure

VARYING THE MAXIMAL HYPEREDGE SIZE $q+1$

Table: Accuracy of various SSL methods on the digits dataset. We pick $\varepsilon_n^{(k)} = 100^{2-k}$ for $1 \le k \le 5$. Proposed methods are in bold.

q	rate	Laplace	Poisson	**IP-QC**	**CP-QC**	**IP-SC**	**CP-SC**	**IP-CC**	**CP-CC**
2	0.02	11.96 (4.03)	**78.81** (2.98)	24.19 (8.92)	15.81 (5.5)	21.91 (8.44)	14.88 (5.48)	19.55 (7.77)	13.44 (5.08)
	0.05	19.35 (6.62)	**84.87** (1.63)	62.35 (7.28)	34.88 (8.75)	59.01 (7.52)	29.09 (9.18)	53.38 (7.77)	23.86 (7.5)
	0.10	42.87 (7.4)	**87.13** (1.12)	81.84 (3.6)	58.25 (7.34)	80.96 (3.81)	53.24 (6.98)	79.07 (4.3)	49.71 (6.62)
	0.20	68.58 (4.38)	87.61 (0.94)	**89.21** (1.5)	84.77 (2.24)	89.11 (1.48)	81.91 (2.7)	88.86 (1.44)	78.27 (3.3)
	0.30	82.1 (2.02)	87.58 (0.74)	**91.78** (0.86)	90.13 (1.08)	**91.78** (0.88)	88.85 (1.2)	91.74 (0.89)	87.13 (1.3)
	0.50	88.3 (1.11)	87.85 (0.78)	93.87 (0.72)	92.78 (0.87)	93.87 (0.72)	92.01 (0.87)	**93.91** (0.7)	91.08 (0.93)
	0.80	89.73 (1.43)	87.88 (1.42)	**94.96** (0.98)	93.64 (1.16)	94.94 (0.97)	92.86 (1.22)	94.9 (0.96)	91.89 (1.22)
3	0.02	11.96 (4.03)	**78.81** (2.98)	22.57 (9.14)	15.02 (5.8)	20.91 (8.57)	15.46 (5.4)	18.96 (7.82)	13.79 (5.43)
	0.05	19.35 (6.62)	**84.87** (1.63)	61.81 (7.17)	37.24 (7.55)	58.56 (7.5)	31.54 (9.11)	52.93 (7.74)	24.84 (7.85)
	0.10	42.87 (7.4)	**87.13** (1.12)	81.57 (3.51)	60.04 (7.23)	80.78 (3.71)	54.66 (7.07)	78.93 (4.26)	50.4 (6.7)
	0.20	68.58 (4.38)	87.61 (0.94)	**89.12** (1.5)	85.79 (2.17)	89.06 (1.5)	82.83 (2.57)	88.82 (1.47)	79.01 (3.19)
	0.30	82.1 (2.02)	87.58 (0.74)	91.74 (0.87)	90.98 (1.02)	**91.75** (0.87)	89.44 (1.15)	91.73 (0.88)	87.57 (1.28)
	0.50	88.3 (1.11)	87.85 (0.78)	93.87 (0.71)	93.39 (0.81)	**93.89** (0.7)	92.45 (0.86)	**93.89** (0.71)	91.37 (0.92)
	0.80	89.73 (1.43)	87.88 (1.42)	**94.98** (0.99)	94.33 (1.13)	94.96 (0.98)	93.3 (1.2)	94.91 (0.96)	92.18 (1.21)

VARYING THE WEIGHTS λ_k

Table: Accuracy of various SSL methods on the digits dataset. We pick $\varepsilon_n^{(k)} = 100^{2-k}$ for $1 \leq k \leq 5$ and $\lambda_1 = 1$, $\lambda_2 = j^2$, $\lambda_3 = (j+1)^2$. Proposed methods are in bold.

j	rate	Laplace	Poisson	WNLL	Properly	p-Lap	RW	CK	**IP-VQC (2)**	**IP-VQC (3)**
1	0.02	12.2 (4.75)	**79.0** (2.75)	67.07 (6.07)	78.29 (3.14)	77.83 (3.23)	30.17 (11.33)	60.0 (4.17)	20.58 (8.29)	19.66 (8.71)
	0.05	20.42 (7.03)	**84.61** (1.72)	69.2 (4.38)	83.11 (2.08)	82.5 (2.19)	32.0 (5.96)	66.19 (3.73)	53.07 (7.79)	50.55 (8.44)
	0.10	41.62 (6.59)	86.73 (1.36)	80.73 (3.07)	**87.67** (1.45)	87.45 (1.51)	31.95 (5.56)	71.98 (2.73)	78.63 (4.42)	77.94 (4.46)
	0.20	68.47 (4.79)	87.61 (0.99)	86.21 (1.53)	89.04 (0.97)	88.93 (1.0)	40.94 (4.75)	78.25 (1.53)	**89.19** (1.11)	88.97 (1.1)
	0.30	82.17 (2.32)	87.62 (0.8)	88.0 (1.2)	89.81 (0.87)	89.74 (0.89)	44.89 (5.34)	82.11 (0.81)	**91.75** (0.84)	91.67 (0.84)
	0.50	88.18 (1.0)	87.84 (0.96)	89.04 (1.0)	89.98 (1.0)	89.94 (0.99)	37.33 (2.51)	85.67 (0.98)	**93.8** (0.87)	93.77 (0.86)
	0.80	89.65 (1.49)	87.88 (1.4)	89.68 (1.45)	89.97 (1.42)	89.97 (1.41)	33.93 (1.16)	88.34 (1.39)	94.86 (1.0)	**94.89** (1.02)
2	0.02	12.2 (4.75)	**79.0** (2.75)	67.07 (6.07)	78.29 (3.14)	77.83 (3.23)	30.17 (11.33)	60.0 (4.17)	25.16 (9.35)	24.25 (9.65)
	0.05	20.42 (7.03)	**84.61** (1.72)	69.2 (4.38)	83.11 (2.08)	82.5 (2.19)	32.0 (5.96)	66.19 (3.73)	62.69 (6.84)	61.96 (6.85)
	0.10	41.62 (6.59)	86.73 (1.36)	80.73 (3.07)	**87.67** (1.45)	87.45 (1.51)	31.95 (5.56)	71.98 (2.73)	81.51 (3.66)	81.25 (3.61)
	0.20	68.47 (4.79)	87.61 (0.99)	86.21 (1.53)	89.04 (0.97)	88.93 (1.0)	40.94 (4.75)	78.25 (1.53)	**89.49** (1.09)	89.41 (1.1)
	0.30	82.17 (2.32)	87.62 (0.8)	88.0 (1.2)	89.81 (0.87)	89.74 (0.89)	44.89 (5.34)	82.11 (0.81)	**91.83** (0.86)	91.79 (0.83)
	0.50	88.18 (1.0)	87.84 (0.96)	89.04 (1.0)	89.98 (1.0)	89.94 (0.99)	37.33 (2.51)	85.67 (0.98)	**93.79** (0.91)	93.77 (0.9)
	0.80	89.65 (1.49)	87.88 (1.4)	89.68 (1.45)	89.97 (1.42)	89.97 (1.41)	33.93 (1.16)	88.34 (1.39)	94.91 (1.01)	**94.93** (1.0)

HYPERGRAPH LEARNING AS A QUADRATIC FORM

- Since $L_n^{(k)}$ are positive semi-definite, so is $\sum_{k=1}^{q} \lambda_k (L_n^{(k)})^k$ and higher order hypergraph learning is a **quadratic form**
- **Observation**: most extensions of Laplace learning lose this mathematical structure which makes them less convenient to analyze and compute
- **Consequence 1**: we can use spectral truncation to speed up computations
- **Consequence 2**: convenient to perform uncertainty quantification and active learning

BAYESIAN FORMULATION OF HYPERGRAPH LEARNING

- Define a prior for $u \sim \mathcal{N}\left(0, \left(\sum_{k=1}^{q} \lambda_k (L_n^{(k)})^k\right)^{-1}\right)$, a likelihood for $y|u$ proportional to $e^{-\Psi(u,y)}$ for some loss function Ψ

- The posterior $u|y$ is proportional to $e^{-\mathcal{F}(u)-\Psi(u,y)}$ and the maximum à posteriori estimator is minimizer of $\mathcal{F}(u) + \Psi(u,y)$

- Since $\mathcal{F}(u)$ is quadratic, it is easy to sample from prior and consequently from the posterior using the pCN-algorithm [2]: it is possible to perform **uncertainty quantification**

- Since $\mathcal{F}(u)$ is quadratic, the Laplace approximation [37] is precise and we can do **active learning** efficiently [30]

ACTIVE LEARNING

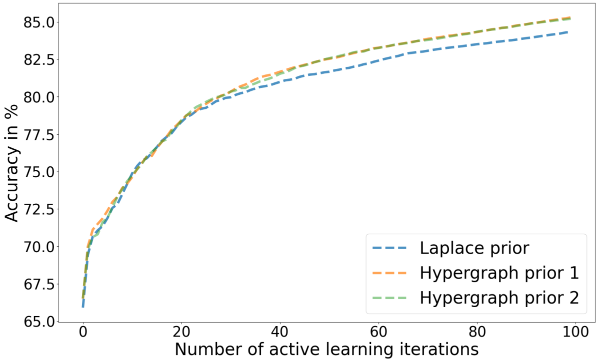


Figure: Accuracy over 100 trials of active learning on the Salinas A dataset using the Laplace prior with $k^{(1)} = 50$, the Hypergraph prior 1 with $k^{(1)} = 50$, $k^{(2)} = 30$, $\lambda_1 = 1$, $\lambda_2 = 2$, $c(1) = 1$, $c(2) = 2$ and the Hypergraph prior 2 with $k^{(1)} = 50$, $k^{(2)} = 30$, $k^{(2)} = 20$, $\lambda_1 = 1$, $\lambda_2 = 2$, $\lambda_3 = 4$, $c(1) = 1$, $c(2) = 2$, $c(3) = 3$. All priors are truncated at $K_n = 100$.

FUTURE RESEARCH DIRECTIONS

GEOMETRIC DIRECTIONS

- Can we consider graph learning problems on other random geometric graph models?
  - **Bidisperse graphs** [32], i.e. local parameter $\varepsilon_n(x_i, x_j)$ of a special kind
  - **Soft random geometric graphs** [36], i.e. a weight exists between x_i and x_j with probability $w_{\varepsilon,ij}$

  $\Rightarrow$ The latter will imply a random-to-deterministic approximation step

- Can we consider **perturbated domains** [5]?

$\Rightarrow$ This will modify our Euler-Lagrange equations and introduce a term linked to the capacity of the perturbated domain

- Can we obtain rates for p-Laplacian regularization on random geometric graphs? Can this be generalized to other inverse problems?

ANALYTIC DIRECTIONS

- Can we find the right way to discretize general $\mathrm{W}^{k,p}$ norms on (hyper)graphs, in particular through **nonlocal formulas** [16]?

$\Rightarrow$ This will be useful to analyze large data limits of energies similar to $\sum_{k=1}^{q} \lambda_k \langle v, (L_n^{(k)})^k v \rangle_n$

$\Rightarrow$ Nonlocal approximations are useful to discretize inverse problems and PDEs on manifolds whose geometry are unknown [23]

MAIN IDEA FOR APPLICATIONS OF HIGHER ORDER HYPERGRAPH LEARNING

- Higher order hypergraph learning behaves like Laplace learning but captures the geometry of the data in a better way

- In particular, **it has the same mathematical structure**

$\Rightarrow$ Replace Laplace learning with higher order hypergraph learning

APPLICATION I: GRAPH NEURAL NETWORKS

- The graph convolution network is defined in [26] through:

$$X' = \text{Normalized Laplacian} \cdot X\Theta$$

$\Rightarrow$ We could try and replace this with our new matrix $\sum_{k=1}^{q} \lambda_k (L_n^{(k)})^k$

$\Rightarrow$ We need to find a way to define normalization appropriately

- We also note that our method can equally be defined on dataset which are **not point clouds**

$\Rightarrow$ We can compare to classical graph/hypergraph deep learning on graph/hypergraph datasets

APPLICATION II: EMBEDDINGS AND SPECTRAL CLUSTERING

- Spectral clustering [41] is a very successful unsupervised clustering method
- It relies on the embedding of data through the eigenvectors of the Laplacian matrix

$\Rightarrow$ What about **spectral clustering** using our new matrix?

- In order to **scale the embedding and to apply it to unseen data**, SpectralNet [38] was developped

$\Rightarrow$ Can we do the same thing with the new Laplacian matrix?

THEORETICAL STUDY OF HIGHER ORDER HYPERGRAPH LEARNING

- Can we prove convergence of posteriors in the large data limit as is done for fractional Laplacian learning [14]?
- Can we prove consistency in semi-supervised learning as is [39, 43]?

College | Physical Sciences
UCLA Mathematics

Adrien Weihs
weihs@math.ucla.edu

REFERENCES

REFERENCES I

[1] A. L. Bertozzi and A. Flenner.
Diffuse interface models on graphs for classification of high dimensional data.
SIAM Review, 58(2):293–328, 2016.

[2] A. L. Bertozzi, X. Luo, A. M. Stuart, and K. C. Zygalakis.
Uncertainty quantification in graph-based classification of high dimensional data.
SIAM/ASA Journal on Uncertainty Quantification, 6(2):568–595, 2018.

[3] J. Bourgain, H. Brezis, and P. Mironescu.
Another look at Sobolev spaces.
In *Optimal Control and Partial Differential Equations*, pages 439–455, 2001.

[4] A. Braides.
Γ*-convergence for Beginners*.
Oxford University Press, 2002.

REFERENCES II

[5] A. Braides.
Chapter 2 a handbook of gamma-convergence.
volume 3 of *Handbook of Differential Equations: Stationary Partial Differential Equations*, pages 101–213. North-Holland, 2006.

[6] J. Calder.
The game theoretic p-Laplacian and semi-supervised learning with few labels.
Nonlinearity, 32(1):301–330, dec 2018.

[7] J. Calder.
Consistency of Lipschitz learning with infinite unlabeled data and finite labeled data.
SIAM Journal on Mathematics of Data Science, 1(4):780–812, 2019.

[8] J. Calder and D. Slepčev.
Properly-weighted graph Laplacian for semi-supervised learning.
Applied Mathematics & Optimization, 82(3):1111–1159, 2020.

REFERENCES III

[9] J. Calder, D. Slepčev, and M. Thorpe.
Rates of convergence for Laplacian semi-supervised learning with low labeling rates.
to appear in Research in the Mathematical Science, preprint arXiv:2006.02765, 2020.

[10] M. Caroccia, A. Chambolle, and D. Slepčev.
Mumford–Shah functionals on graphs and their asymptotics.
Nonlinearity, 33(8):3846–3888, jun 2020.

[11] U. Chitra and B. J. Raphael.
Random walks on hypergraphs with edge-dependent vertex weights.
In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 1172–1181. PMLR, 2019.

[12] R. A. DeVore and G. G. Lorentz.
Constructive Approximation.
Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 1993.

REFERENCES IV

[13] A. Dontchev and F. Lempio.
Difference methods for differential inclusions: A survey.
SIAM Review, 34(2):263–294, 1992.

[14] M. Dunlop, D. Slepcev, A. Stuart, and M. Thorpe.
Large data and zero noise limits of graph-based semi-supervised learning algorithms.
Applied and Computational Harmonic Analysis, 49(2):655–697, 2020.

[15] A. Fazeny, D. Tenbrinck, and M. Burger.
Hypergraph p-laplacians, scale spaces, and information flow in networks.
In L. Calatroni, M. Donatelli, S. Morigi, M. Prato, and M. Santacesaria, editors, *Scale Space and Variational Methods in Computer Vision*, pages 677–690, Cham, 2023. Springer International Publishing.

REFERENCES V

[16] R. Ferreira, C. Kreisbeck, and A. M. Ribeiro.
Characterization of polynomials and higher-order sobolev spaces in terms of functionals involving difference quotients.
Nonlinear Analysis: Theory, Methods and Applications, 112:199–214, 2015.

[17] N. García Trillos, R. Murray, and M. Thorpe.
From graph cuts to isoperimetric inequalities: Convergence rates of cheeger cuts on data clouds.
Archive for Rational Mechanics and Analysis, 244(3):541–598, 2022.

[18] N. García Trillos and D. Slepčev.
Continuum limit of total variation on point clouds.
Archive for Rational Mechanics and Analysis, 220(1):193–241, 2016.

[19] N. García Trillos and D. Slepčev.
A variational approach to the consistency of spectral clustering.
Applied and Computational Harmonic Analysis, 45(2):239–281, 2018.

REFERENCES VI

[20] N. García Trillos, D. Slepčev, J. von Brecht, T. Laurent, and X. Bresson.
Consistency of cheeger and ratio graph cuts.
Journal of Machine Learning Research, 17(181):1–46, 2016.

[21] B. Hanin.
Random neural networks in the infinite width limit as gaussian processes, 2021.

[22] B. Hanin and A. Zlokapa.
Bayesian interpolation with deep linear networks, 2023.

[23] J. Harlim, D. Sanz-Alonso, and R. Yang.
Kernel methods for bayesian elliptic inverse problems on manifolds.
SIAM/ASA Journal on Uncertainty Quantification, 8(4):1414–1445, 2020.

[24] J. Jost and R. Mulas.
Hypergraph laplace operators for chemical reaction networks.
Advances in Mathematics, 351:870–896, 2019.

REFERENCES VII

[25] J. Jost, R. Mulas, and D. Zhang.
p-laplace operators for oriented hypergraphs.
Vietnam Journal of Mathematics, 50(2):323–358, 2022.

[26] T. N. Kipf and M. Welling.
Semi-supervised classification with graph convolutional networks.
In *International Conference on Learning Representations*, 2017.

[27] P. Li and O. Milenkovic.
Submodular hypergraphs: p-laplacians, Cheeger inequalities and spectral clustering.
In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3014–3023. PMLR, 10–15 Jul 2018.

[28] J.-L. Lions.
Quelques méthodes de résolution des problèmes aux limites non linéaires.
Collection études mathématiques. Dunod, 1969.

REFERENCES VIII

[29] E. Merkurjev, D. D. Nguyen, and G.-W. Wei.
Multiscale laplacian learning.
Applied Intelligence, 53(12):15727–15746, nov 2022.

[30] K. S. Miller and A. L. Bertozzi.
Model change active learning in graph-based semi-supervised learning.
Communications on Applied Mathematics and Computation, 6(2):1270–1298, 2024.

[31] R. Mulas, C. Kuehn, T. Böhle, and J. Jost.
Random walks and laplacians on hypergraphs: When do they match?
Discrete Applied Mathematics, 317:26–41, 2022.

[32] S. Nauer, L. Böttcher, and M. A. Porter.
Random-graph models and characterization of granular networks.
Journal of Complex Networks, 8(5):cnz037, 11 2019.

REFERENCES IX

[33] L. Neuhäuser, R. Lambiotte, and M. T. Schaub.
Consensus Dynamics and Opinion Formation on Hypergraphs, pages 347–376.
Springer International Publishing, Cham, 2022.

[34] N. H. Pavel.
Nonlinear Evolution Operators and Semigroups: Applications to Partial Differential Equations.
Lecture Notes in Mathematics. Springer, 1987.

[35] M. D. Penrose.
Random Geometric Graphs.
Oxford University Press, 2003.

[36] M. D. Penrose.
Connectivity of soft random geometric graphs.
The Annals of Applied Probability, 26(2):986–1028, 2016.

REFERENCES X

[37] C. E. Rasmussen and C. K. I. Williams.
Gaussian Processes for Machine Learning.
The MIT Press, 11 2005.

[38] U. Shaham, K. Stanton, H. Li, R. Basri, B. Nadler, and Y. Kluger.
Spectralnet: Spectral clustering using deep neural networks.
In *International Conference on Learning Representations*, 2018.

[39] D. Slepčev and M. Thorpe.
Analysis of p-Laplacian regularization in semisupervised learning.
SIAM Journal on Mathematical Analysis, 51(3):2085–2120, 2019.

[40] Y. van Gennip and A. Bertozzi.
Gamma-convergence of graph Ginzburg–Landau functionals.
Advances in Differential Equations, 17(11–12):1115–1180, 2012.

REFERENCES XI

[41] U. von Luxburg.
A tutorial on spectral clustering.
Statistics and Computing, 2007.

[42] A. Weihs, J. Fadili, and M. Thorpe.
Discrete-to-continuum rates of convergence for p-laplacian regularization, 2023.

[43] A. Weihs and M. Thorpe.
Consistency of fractional graph-laplacian regularization in semisupervised learning with finite labels.
SIAM Journal on Mathematical Analysis, 56(4):4253–4295, 2024.

[44] D. H. Zanette.
Beyond networks: opinion formation in triplet-based populations.
Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 367(1901):3311–3319, 2009.

REFERENCES XII

[45] D. Zhou, J. Huang, and B. Schölkopf.
Learning with hypergraphs: Clustering, classification, and embedding.
In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

[46] X. Zhu, Z. Ghahramani, and J. Lafferty.
Semi-supervised learning using Gaussian fields and harmonic functions.
In *Proceedings of the International Conference on Machine Learning*, 2003.