Convergence and Stability of Graph Convolutional Networks on Large Random Graphs

Nicolas Keriven[*]
CNRS, GIPSA-lab, Grenoble, France
nicolas.keriven@cnrs.fr

Alberto Bietti[*]
NYU Center for Data Science, New York, USA[†]
alberto.bietti@nyu.edu

Samuel Vaiter
CNRS, IMB, Dijon, France
samuel.vaiter@u-bourgogne.fr

Abstract

We study properties of Graph Convolutional Networks (GCNs) by analyzing their behavior on standard models of random graphs, where nodes are represented by random latent variables and edges are drawn according to a similarity kernel. This allows us to overcome the difficulties of dealing with discrete notions such as isomorphisms on very large graphs, by considering instead more natural geometric aspects. We first study the convergence of GCNs to their continuous counterpart as the number of nodes grows. Our results are fully non-asymptotic and are valid for relatively sparse graphs with an average degree that grows logarithmically with the number of nodes. We then analyze the stability of GCNs to small deformations of the random graph model. In contrast to previous studies of stability in discrete settings, our continuous setup allows us to provide more intuitive deformation-based metrics for understanding stability, which have proven useful for explaining the success of convolutional representations on Euclidean domains.

1 Introduction

Graph Convolutional Networks (GCNs [9, 15, 26]) are deep architectures defined on graphs inspired by classical Convolutional Neural Networks (CNNs [28]). In the past few years, they have been successfully applied to, for instance, node clustering [11], semi-supervised learning [26], or graph regression [23, 20], and remain one of the most popular variant of Graph Neural Networks (GNN). We refer the reader to the review papers [7, 46] for more details.

Many recent results have improved the theoretical understanding of GNNs. While some architectures have been shown to be universal [35, 24] but not implementable in practice, several studies have characterized GNNs according to their power to distinguish (or not) graph *isomorphisms* [47, 12, 34] or compute combinatorial graph parameters [13]. However, such notions usually become moot for large graphs, which are almost never isomorphic to each other, but for which GCNs have proved to be successful in identifying large-scale structures nonetheless, *e.g.,* for segmentation or spectral clustering [11]. Under this light, a relevant notion is that of *stability*: since GCNs are trained then tested on different (large) graphs, how much does a change in the graph structure affect its predictions? In the context of signals defined on Euclidean domains, including images or audio, convolutional representations such as scattering transforms or certain CNN architectures have been shown to be stable to *spatial deformations* [33, 5, 39]. However the notion of deformations is not well-defined on

---

[*]Equal contribution.
[†]Work done while AB was at Inria Paris.

discrete graphs, and most stability studies for GCNs use purely discrete metrics that are less intuitive for capturing natural changes in structure [17, 19, 48].

In statistics and machine learning, there is a long history of modelling large graphs with random models, see for instance [6, 21, 27, 36] and references therein for reviews. *Latent space models* represent each node as a vector of latent variables and independently connect the nodes according to a *similarity kernel* applied to their latent representations. This large family of random graph models includes for instance Stochastic Block Models (SBM) [22], graphons [31], random geometric graphs [37], or ε-graphs [10], among many others [36]. A key parameter in such models is the so-called *sparsity factor* α_n that controls the number of edges in $\mathcal{O}(n^2 \alpha_n)$ with respect to the number of nodes n. The *dense* case $\alpha_n \sim 1$ is the easiest to analyze, but often not realistic for real-world graphs. On the contrary, many questions are still open in the *sparse* case $\alpha_n \sim 1/n$ [1]. A middle ground, which will be the setting for our analysis, is the so-called *relatively sparse* case $\alpha_n \sim \log n/n$, for which several non-trivial results are known [29, 25], while being more realistic than the dense case.

Outline and contributions.   In this paper, we analyze the convergence and stability properties of GCNs on large random graphs. We define a "continuous" counterpart to discrete GCNs acting on graph models in Section 2, study notions of invariance and equivariance to isomorphism of random graph models, and give convergence results when the number of nodes grows in Section 3. In particular, our results are fully non-asymptotic, valid for relatively sparse random graphs, and unlike many studies [44, 40] we do not assume that the similarity kernel is smooth or bounded away from zero. In Section 4, we analyze the stability of GCNs to small deformation of the underlying random graph model. Similar to CNNs [33, 5], studying GCNs in the continuous world allows us to define intuitive notions of model deformations and characterize their stability. Interestingly, for GCNs equivariant to permutation, we relate existing discrete notions of distance between graph signals to a Wasserstein-type metric between the corresponding continuous representations, which to our knowledge did not appear in the literature before.

Related work on large-scale random graphs.   There is an long history of studying the convergence of graph-related objects on large random graphs. A large body of works examine the convergence of the eigenstructures of the graph adjacency matrix or Laplacian in the context of spectral clustering [4, 44, 29, 42] or learning with operators [40]. The theory of graphons [31] defines (dense) graph limits for more general metrics, which is also shown to lead to spectral convergence [16]. Closer to our work, notions of Graph Signal Processing (GSP) such as the graph Fourier Transform have been extended to graphons [41] or sampling of general Laplacian operators [30]. Partial results on the capacity of GCNs to distinguish dense graphons are derived in [32], however their analysis based on random walks differs greatly from ours. In general, many of these studies are asymptotic [44, 41], valid only in the dense case [44, 40, 41, 30, 32], or assume kernels that are smooth or bounded away from zero [40], and thus exclude several important cases such as SBMs, ε-graphs, and non-dense graphs altogether. By specifying models of (relatively sparse) random graphs, we derive non-asymptotic, fully explicit bounds with relaxed hypotheses.

Related work on stability.   The study of stability to deformations has been pioneered by Mallat [33] in the context of the scattering transform for signals on Euclidean domains such as images or audio signals [8, 2], and was later extended to more generic CNN architectures [5, 39]. A more recent line of work has studied stability properties of GCNs or scattering representations on discrete graphs, by considering certain well-chosen discrete perturbations and metrics [17–19, 48], which may however have limited interpretability without an underlying model. In contrast, our continuous setup allows us to define more intuitive geometric perturbations based on deformations of random graph models and to obtain deformation stability bounds that are similar to those on Euclidean domains [33]. We note that [30] also considers GCN representations with continuous graph models, but the authors focus on the different notion of "transferability" of graph filters on different discretizations of the *same* underlying continuous graph structure, while we consider *explicit deformations* of this underlying structure and obtain non-asymptotic bounds for the resulting random graphs.

2   Preliminaries

Notations.   The norm $\|\cdot\|$ is the Euclidean norm for vector and spectral norm for matrices, and $\|\cdot\|_F$ is the Frobenius norm. We denote by $\mathcal{B}(\mathcal{X})$ the space of bounded real-valued functions on $\mathcal{X}$

equipped with the norm $\|f\|_\infty = \sup_x |f(x)|$. Given a probability distribution P on $\mathcal{X}$, we denote by $L^2(P)$ the Hilbert space of P-square-integrable functions endowed with its canonical inner product. For multivariate functions $f = [f_1, \ldots, f_d]$ and any norm $\|\cdot\|$, we define $\|f\| = (\sum_{i=1}^d \|f_i\|^2)^{\frac{1}{2}}$. For two probability distributions P, Q on $\mathbb{R}^d$, we define the Wasserstein-2 distance $\mathcal{W}_2^2(P, Q) = \inf\{\mathbb{E}\|X - Y\|^2 \mid X \sim P, Y \sim Q\}$, where the infimum is over all joint distributions of (X, Y). We denote by $f_\sharp P$ the push-forward of P by f, that is, the distribution of $f(X)$ when $X \sim P$.

A graph $G = (A, Z)$ with n nodes is represented by a symmetric adjacency matrix $A \in \{0, 1\}^{n \times n}$ such that $a_{ij} = 1$ if there is an edge between nodes i and j, and a matrix of signals over the nodes $Z \in \mathbb{R}^{n \times d_z}$, where $z_i \in \mathbb{R}^{d_z}$ is the vector signal at node i. We define the normalized Laplacian matrix as $L = L(A) = D(A)^{-\frac{1}{2}} A D(A)^{-\frac{1}{2}}$, where $D(A) = \mathrm{diag}(A1_n)$ is the degree matrix, and $(D(A)^{-\frac{1}{2}})_i = 0$ if $D(A)_i = 0$. The normalized Laplacian is often defined by $\mathrm{Id} - L$ in the literature, however this does not change the considered networks since the filters include a term of order 0.

Graph Convolutional Networks (GCN).   GCNs are defined by alternating filters on graph signals and non-linearities. We use analytic filters (said of order-k if $\beta_\ell = 0$ for $\ell \geqslant k + 1$):

$$h : \mathbb{R} \to \mathbb{R}, \quad h(\lambda) = \sum_{k \geqslant 0} \beta_k \lambda^k. \tag{1}$$

We write $h(L) = \sum_k \beta_k L^k$, *i.e.,* we apply h to the eigenvalues of L when it is diagonalizable.

A GCN with M layers is defined as follows. The signal at the input layer is $Z^{(0)} = Z$ with dimension $d_0 = d_z$ and columns $z_j^{(0)} \in \mathbb{R}^n$. Then, at layer ℓ, the signal $Z^{(\ell)} \in \mathbb{R}^{n \times d_\ell}$ with columns $z_j^{(\ell)} \in \mathbb{R}^n$ is propagated as follows:

$$\forall j = 1, \ldots d_{\ell+1}, \quad z_j^{(\ell+1)} = \rho\left(\sum_{i=1}^{d_\ell} h_{ij}^{(\ell)}(L) z_i^{(\ell)} + b_j^{(\ell)} 1_n\right) \in \mathbb{R}^n, \tag{2}$$

where $h_{ij}^{(\ell)}(\lambda) = \sum_k \beta_{ijk}^{(\ell)} \lambda^k$ are learnable analytic filters, $b_j^{(\ell)} \in \mathbb{R}$ are learnable biases, and the activation function $\rho : \mathbb{R} \to \mathbb{R}$ is applied pointwise. Once the signal at the final layer $Z^{(M)}$ is obtained, the output of the entire GCN is either a signal over the nodes denoted by $\Phi_A(Z) \in \mathbb{R}^{n \times d_{out}}$ or a single vector denoted by $\bar\Phi_A(Z) \in \mathbb{R}^{d_{out}}$ obtained with an additional pooling over the nodes:

$$\Phi_A(Z) \stackrel{\mathrm{def.}}{=} Z^{(M)}\theta + 1_n b^\top, \quad \bar\Phi_A(Z) \stackrel{\mathrm{def.}}{=} \frac{1}{n}\sum_{i=1}^n \Phi_A(Z)_i, \tag{3}$$

where $\theta \in \mathbb{R}^{d_M \times d_{out}}$, $b \in \mathbb{R}^{d_{out}}$ are the final layer weights and bias, and $\Phi_A(Z)_i \in \mathbb{R}^{d_{out}}$ is the output signal at node i. This general model of GCN encompasses several models of the literature, including all spectral-based GCNs [9, 15], or GCNs with order-1 filters [26] which are assimilable to message-passing networks [20], see [46, 7] for reviews. For message-passing networks, note that almost all our results would also be valid by replacing the sum over neighbors by another aggregation function such as $\max$. We assume (true for ReLU, modulus, or sigmoid) that the function ρ satisfies:

$$|\rho(x)| \leqslant |x|, \quad |\rho(x) - \rho(y)| \leqslant |x - y|. \tag{4}$$

Two graphs $G = (A, Z), G' = (A', Z')$ are said to be *isomorphic* if one can be obtained from the other by relabelling the nodes. In other words, there exists a *permutation matrix* $\sigma \in \Sigma_n$, where Σ_n is the set of all permutation matrices, such that $A = \sigma \cdot A' \stackrel{\mathrm{def.}}{=} \sigma A' \sigma^\top$ and $Z = \sigma \cdot Z' \stackrel{\mathrm{def.}}{=} \sigma Z'$, where "$\sigma \cdot$" is a common notation for permuted matrices or signal over nodes. In graph theory, functions that are *invariant* or *equivariant* to permutations are of primary importance (respectively, permuting the input graph does not change the output, or permutes the output). These properties are hard-coded in the structure of GCNs, as shown by the following proposition (proof in Appendix B).

Proposition 1. *We have* $\Phi_{\sigma \cdot A}(\sigma \cdot Z) = \sigma \cdot \Phi_A(Z)$ *and* $\bar\Phi_{\sigma \cdot A}(\sigma \cdot Z) = \bar\Phi_A(Z)$.

Random graphs.   Let $(\mathcal{X}, d)$ be a compact metric space. In this paper, we consider latent space graph models where each node i is represented by an unobserved latent variable $x_i \in \mathcal{X}$, and nodes are connected randomly according to some *similarity kernel*. While the traditional graphon model [31] considers (without lost of generality) $\mathcal{X} = [0, 1]$, it is often more intuitive to allow general spaces to represent meaningful variables [14]. We consider that the observed signal $z_i \in \mathbb{R}^{d_z}$ is a function of the latent variable x_i, without noise for now. In details, a *random graph model* $\Gamma = (P, W, f)$ is

represented by a probability distribution P over $\mathcal{X}$, a symmetric kernel $W : \mathcal{X} \times \mathcal{X} \to [0,1]$ and a bounded function $f : \mathcal{X} \to \mathbb{R}^{d_z}$. A random graph G with n nodes is then generated as follows:

$$\forall j < i \leqslant n : \quad x_i \overset{iid}{\sim} P, \quad z_i = f(x_i), \quad a_{ij} \sim \text{Ber}(\alpha_n W(x_i, x_j)). \tag{5}$$

where Ber is the Bernoulli distribution. We define $d_{W,P} \overset{\text{def.}}{=} \int W(\cdot, x) dP(x)$ the *degree function* of Γ. As outlined in the introduction, the sparsity factor $\alpha_n \in [0,1]$ plays a key role. The so-called *relatively sparse* case $\alpha_n \sim \frac{\log n}{n}$ will be the setting for our analysis.

Let us immediately make some assumptions that will hold throughout the paper. We denote by $N(\mathcal{X}, \varepsilon, d)$ the ε-covering numbers (that is, the minimal number of balls of radius ε required to cover $\mathcal{X}$) of $\mathcal{X}$, and assume that they can be written under the form $N(\mathcal{X}, \varepsilon, d) \leqslant \varepsilon^{-d_x}$ for some constant $d_x > 0$ (called the *Minkowski* dimension of $\mathcal{X}$), and diam$(\mathcal{X}) \leqslant 1$. Both conditions can be obtained by a rescaling of the metric d. Let $c_{\min}, c_{\max} > 0$ be constants. A function $f : \mathcal{X} \to \mathbb{R}$ is said to be $(c_{\text{Lip.}}, n_{\mathcal{X}})$-piecewise Lipschitz if there is a partition $\mathcal{X}_1, \ldots, \mathcal{X}_{n_{\mathcal{X}}}$ of $\mathcal{X}$ such that, for all x, x' in the same $\mathcal{X}_i$, we have $|f(x) - f(x')| \leqslant c_{\text{Lip.}} d(x, x')$. All considered random graph models $\Gamma = (P, W, f)$ satisfy that for all $x \in \mathcal{X}$,

$$\|W(\cdot, x)\|_\infty \leqslant c_{\max}, \quad d_{W,P}(x) \geqslant c_{\min}, \quad W(\cdot, x) \text{ is } (c_{\text{Lip.}}, n_{\mathcal{X}})\text{-piecewise Lipschitz}. \tag{6}$$

Unlike other studies [44, 40], we do *not* assume that W itself is bounded away from 0 or smooth, and thus include important cases such as SBMs (piecewise constant W) and ε-graphs (threshold kernels).

Continuous GCNs. Since $d_{W,P} > 0$, we define the normalized Laplacian operator $\mathcal{L}_{W,P}$ by

$$\mathcal{L}_{W,P} f \overset{\text{def.}}{=} \int \frac{W(\cdot, x)}{\sqrt{d_{W,P}(\cdot) d_{W,P}(x)}} f(x) dP(x). \tag{7}$$

Analytic filters on operators are $h(\mathcal{L}) = \sum_k \beta_k \mathcal{L}^k$, with $\mathcal{L}^k = \mathcal{L} \circ \ldots \circ \mathcal{L}$. We do *not* assume that the filters are of finite order (even if they usually are in practice [15]), however we will always assume that $\sum_k k |\beta_k| (2c_{\max}/c_{\min})^k$ converges. Similar to the discrete case, we define continuous GCNs (c-GCN) that act on random graph models, by replacing the input signal Z with f, the Laplacian L by $\mathcal{L}_{W,P}$, and propagating functions instead of node signals *i.e.,* we take $f^{(0)} = f$ the input function with coordinates $f_1^{(0)}, \ldots, f_{d_z}^{(0)}$ and:

$$\forall j = 1, \ldots, d_{\ell+1}, \quad f_j^{(\ell+1)} = \rho \circ \left(\sum_{i=1}^{d_\ell} h_{ij}^{(\ell)}(\mathcal{L}_{W,P}) f_i^{(\ell)} + b_j^{(\ell)} 1(\cdot) \right), \tag{8}$$

where $1(\cdot)$ represent the constant function 1 on $\mathcal{X}$. Once the final layer function $f^{(M)} : \mathcal{X} \to \mathbb{R}^{d_M}$ is obtained, the output of the c-GCN is defined as in the discrete case, either as a multivariate function $\Phi_{W,P}(f) : \mathcal{X} \to \mathbb{R}^{d_{out}}$ or a single vector $\bar{\Phi}_{W,P}(f) \in \mathbb{R}^{d_{out}}$ obtained by pooling:

$$\Phi_{W,P}(f) \overset{\text{def.}}{=} \theta^\top f^{(M)} + b1(\cdot), \quad \bar{\Phi}_{W,P}(f) = \int \Phi_{W,P}(f)(x) dP(x). \tag{9}$$

Hence the *same* parameters $\{(\beta_{ijk}^{(\ell)}, b_j^{(\ell)})_{ijk\ell}, \theta, b\}$ define both a discrete and a continuous GCN, the latter being (generally) not implementable in practice but useful to analyze their discrete counterpart.

For a random graph model $\Gamma = (P, W, f)$, and any invertible map $\phi : \mathcal{X} \to \mathcal{X}$, we define $\phi \cdot W \overset{\text{def.}}{=} W(\phi(\cdot), \phi(\cdot))$ and $\phi \cdot f \overset{\text{def.}}{=} f \circ \phi$. Recalling that $(\phi^{-1})_\sharp P$ is the distribution of $\phi^{-1}(x)$ when $x \sim P$, it is easy to see that $\phi \cdot \Gamma \overset{\text{def.}}{=} ((\phi^{-1})_\sharp P, \phi \cdot W, \phi \cdot f)$ defines the same probability distribution as $\Gamma = (P, W, f)$ over discrete graphs. Therefore, we say that Γ and $\phi \cdot \Gamma$ are *isomorphic*, which is a generalization of isomorphic graphons [31] when P is the uniform measure on $[0,1]$. Note that, technically, ϕ needs only be invertible on the support of P for the above definitions to hold. As with discrete graphs, functions on random graph models can be invariant or equivariant, and c-GCNs satisfy these properties (proof in Appendix B).

Proposition 2. *For all ϕ, $\Phi_{\phi \cdot W, (\phi^{-1})_\sharp P}(\phi \cdot f) = \phi \cdot \Phi_{W,P}(f)$ and $\bar{\Phi}_{\phi \cdot W, (\phi^{-1})_\sharp P}(\phi \cdot f) = \bar{\Phi}_{W,P}(f)$.*

In the rest of the paper, most notation-heavy multiplicative constants are given in the appendix. They depend on $c_{\min}, c_{\max}, c_{\text{Lip.}}$ and the operator norms of the matrices $B_k^{(\ell)} = (\beta_{ijk}^{(\ell)})_{ij}$.

3 Convergence of Graph Convolutional Networks

In this section, we show that a GCN applied to a random graph $G \sim \Gamma$ will be close to the corresponding c-GCN applied to Γ. In the invariant case, $\bar{\Phi}_A(Z)$ and $\bar{\Phi}_{W,P}(f)$ are both vectors in $\mathbb{R}^{d_{out}}$. In the equivariant case, we will show that the output signal $\Phi_A(Z)_i \in \mathbb{R}^{d_{out}}$ at each node is close to the function $\Phi_{W,P}(f)$ evaluated at x_i. To measure this, we consider the (square root of the) Mean Square Error at the node level: for a signal $Z \in \mathbb{R}^{n \times d_{out}}$, a function $f : \mathcal{X} \to \mathbb{R}^{d_{out}}$ and latent variables X, we define $\mathrm{MSE}_X(Z, f) \overset{\text{def.}}{=} (n^{-1} \sum_{i=1}^n \|Z_i - f(x_i)\|^2)^{1/2}$. In the following theorem we use the shorthand $D_{\mathcal{X}}(\rho) \overset{\text{def.}}{=} \frac{c_{\text{Lip.}}}{c_{\min}} \sqrt{d_x} + \frac{c_{\max} + c_{\text{Lip.}}}{c_{\min}} \sqrt{\log \frac{n_{\mathcal{X}}}{\rho}}$.

Theorem 1 (Convergence to continuous GCN). *Let Φ be a GCN and G be a graph with n nodes generated from a model Γ, denote by X its latent variables. There are two universal constants c_1, c_2 such that the following holds. Take any $\rho > 0$, assume n is large enough such that $n \geqslant c_1 D_{\mathcal{X}}(\rho)^2 + \frac{1}{\rho}$, and the sparsity level is such that $\alpha_n \geqslant c_2 c_{\max} c_{\min}^{-2} \cdot n^{-1} \log n$. Then, with probability at least $1 - \rho$,*

$$\mathrm{MSE}_X(\Phi_A(Z), \Phi_{W,P}(f)) \leqslant R_n \overset{\text{def.}}{=} C_1 D_{\mathcal{X}}\left(\frac{\rho}{\sum_{\ell} d_{\ell}}\right) n^{-\frac{1}{2}} + C_2 (n\alpha_n)^{-\frac{1}{2}},$$

$$\left\|\bar{\Phi}_A(Z) - \bar{\Phi}_{W,P}(f)\right\| \leqslant R_n + C_3 \sqrt{\log(1/\rho)} n^{-\frac{1}{2}}.$$

Discussion. The constants C_i are of the form $C_i' \|f\|_\infty + C_i''$ and detailed in the appendix. When the filters are normalized and there is no bias, they are proportional to $M \|f\|_\infty$. In particular, they do not depend on the dimension d_x.

The proof use standard algebraic manipulations, along with two concentration inequalities. The first one exploits Dudley's inequality [43] to show that, for a fixed function f and in the absence of random edges, $\mathcal{L}_{W,P} f$ is well approximated by its discrete counterpart. Note here that we do not seek a *uniform* proof with respect to a functional space, since the c-GCN is fixed. This allows us to obtain non-asymptotic rate while relaxing usual smoothness hypotheses [40]. This first concentration bound leads to the standard rate in $\mathcal{O}(1/\sqrt{n})$.

The second bound uses a fairly involved recent concentration inequality for normalized Laplacians of relatively sparse graphs with random edges derived in [25], which gives the term in $\mathcal{O}(1/\sqrt{\alpha_n n})$. Although this second term has a strictly worse convergence rate except in the dense case $\alpha_n \sim 1$, its multiplicative constant is strictly better, in particular it does not depend on the Minkowski dimension d_x. The condition $n \geqslant 1/\rho$, which suggests a polynomial concentration instead of the more traditional exponential one, comes from this part of the proof.

It is known in the literature that using the normalized Laplacian is often more appropriate than the adjacency matrix. If we where to use the latter, a normalization by $(\alpha_n n)^{-1}$ would be necessary [29]. However, α_n is rarely known, and can change from one case to the other. The normalized Laplacian is adaptative to α_n and does not require any normalization.

Example of applications. Invariant GCNs are typically used for regression or classification at the graph level. Theorem 1 shows that the output of a discrete GCN directly approaches that of the corresponding c-GCN. Equivariant GCNs are typically used for regression at the node level. Consider an ideal function $f^* : \mathcal{X} \to \mathbb{R}^{d_{out}}$ that is well approximated by an equivariant c-GCN $\Phi_{W,P}(f)$ in terms of $L^2(P)$-norm. Then, the error between the output of the discrete GCN $\Phi_A(Z)$ and the sampling of f^* satisfies with high probability $\mathrm{MSE}_X(\Phi_A(Z), f^*) \leqslant \|\Phi_{W,P}(f) - f^*\|_{L^2(P)} + R_n + \mathcal{O}(n^{-\frac{1}{4}})$ using a triangle inequality, Theorem 1 and Hoeffding's inequality.

Noisy or absent signal. Until now, we have considered that the function f was observed without noise. Noise can be handled by considering the Lipschitz properties of the GCN. For instance, in the invariant case, by Lemma 7 in Appendix F, we have $\left\|\bar{\Phi}_A(Z_1) - \bar{\Phi}_A(Z_2)\right\| \lesssim \frac{1}{\sqrt{n}} \|Z_1 - Z_2\|_F$. Hence, if the input signal is the noisy $z_i = f(x_i) + \nu_i$, where ν is centered iid noise, a GCN deviates from the corresponding c-GCN by an additional $n^{-1/2} \|(\nu_i)_i\|_F$, which converges to the standard deviation of the noise. Interestingly, the noise can be filtered out: for instance, if one inputs $\bar{Z} = LZ$ into the GCN, then by a concentration inequality it is not difficult to see that the smoothed noise term converges to 0, and the GCN converges to the c-GCN with smoothed input function $\bar{f} = \mathcal{L}_{W,P} f$.

In some cases such as spectral clustering [11], one does not have an input signal over the nodes, but has only access to the structure of the graph. In this case, several heuristics have been used in the literature, but a definitive answer is yet to emerge. For instance, a classical strategy is to use the (normalized) degrees of the graph $Z = A1_n/(\alpha_n n)$ as input signal [9, 11] (assuming for simplicity that α_n is known or estimated). In this case, using our proofs (Lemma 4 in the appendix) and the spectral concentration in [29], it is not difficult to show that a discrete GCN will converge to its countinuous version with the degree function $f = d_{W,P}$ as input. We will see in Prop. 3 in the next section that this leads to desirable stability properties.

4 Stability of GCNs to model deformations

Stability to deformations is an essential feature for the generalization properties of deep architectures. Mallat [33] studied the stability to small deformations of the wavelet-based scattering transform, which was extended to more generic learned convolutional network, *e.g.,* [5, 39], and tries to establish bounds of the following form for a signal representation $\Phi(\cdot)$:

$$\|\Phi(f_\tau) - \Phi(f)\| \lesssim N(\tau)\|f\|, \tag{10}$$

where $f_\tau(x) = f(x - \tau(x))$ is the deformed signal and $N(\tau)$ quantifies the size of the deformation, typically through norms of its jacobian $\nabla\tau$, such as $\|\nabla\tau\|_\infty = \sup_x \|\nabla\tau(x)\|$. As we have seen in the introduction, it is not clear how to extend the notion of deformation on discrete graphs [17, 19]. We show here that it can be done in the continuous world. We first derive generic stability bounds for discrete random graphs involving a Wasserstein-type metric between the corresponding c-GCNs, then derive bounds of the form (10) for c-GCNs by studying various notions of deformations of random graph models. We note that "spatial" deformations $x \mapsto x - \tau(x)$ are or course not the only possible choice for these models, and leave other types of perturbations to future work.

From discrete to continuous stability. We first exploit the previous convergence result to deport the stability analysis from discrete to continuous GCNs. Let G_1 and G_2 be two random graphs with n nodes drawn from models Γ_1 and Γ_2, and a GCN Φ. In the invariant case, we can directly apply Theorem 1 and the triangle inequality to obtain that $\left\|\bar{\Phi}_{A_1}(Z_1) - \bar{\Phi}_{A_2}(Z_2)\right\| \leqslant \left\|\bar{\Phi}_{W_1,P_1}(f_1) - \bar{\Phi}_{W_2,P_2}(f_2)\right\| + 2R_n$, and study the robustness of $\bar{\Phi}_{W,P}(f)$ to deformations of the model. The equivariant case is more complex. A major difficulty, compared for instance to [30], is that, since we consider two different samplings X_1 and X_2, there are no implicit ordering over the nodes of G_1 and G_2, and one cannot directly compare the output signals of the equivariant GCN *e.g.,* in Frobenius norm. To compare two graph representations, a standard approach in the study of stability (and graph theory in general) has been to define a metric that minimizes over permutations σ of the nodes (*e.g.,* [17, 19]), thus we define $\text{MSE}_\Sigma(Z, Z') \overset{\text{def.}}{=} \min_{\sigma \in \Sigma_n} (n^{-1} \sum_i \|Z_i - Z'_{\sigma(i)}\|^2)^{1/2}$. Theorem 2 relates this to a Wasserstein metric between the c-GCNs (proof in Appendix D).

Theorem 2 (Finite-sample stability in the equivariant case). *Adopt the notations of Theorem 1. For $r = 1, 2$, define the distribution $Q_r = \Phi_{W_r,P_r}(f_r)_\sharp P_r$. With probability $1 - \rho$, we have*

$$\text{MSE}_\Sigma\left(\Phi_{A_1}(Z_1), \Phi_{A_2}(Z_2)\right) \leqslant \mathcal{W}_2(Q_1, Q_2) + R_n + C_1\left(n^{-\frac{1}{d_z}} + \left(C_2 + \sqrt[4]{\log\frac{1}{\rho}}\right)n^{-\frac{1}{4}}\right) \tag{11}$$

where C_1 and C_2 are defined in the appendix. When f_1 and f_2 are piecewise Lipschitz, the last terms can be replaced by $C_1'(n^{-1/d_x} + (C_2' + \sqrt[4]{\log(1/\rho)})n^{-1/4})$ for some C_1', C_2'.

In other words, we express stability in terms of a Wasserstein metric between the push-forwards of the measures P_r by their respective c-GCNs. By definition, the l.h.s. of (11) is invariant to permutation of the graphs G_r. Moreover, for $\phi \in \Sigma_P$ by Prop. 2 we have $\Phi_{\phi \cdot W,P}(\phi \cdot f)_\sharp P = \Phi_{W,P}(f)_\sharp(\phi_\sharp P) = \Phi_{W,P}(f)_\sharp P$, and therefore the r.h.s. of (11) is also invariant to continuous permutation ϕ.

We recover the rate R_n from Theorem 1, as well as a term in $1/n^{1/4}$ and a term that depends on the dimension. In the relatively sparse case, the term in $1/\sqrt{\alpha_n n}$ in R_n still has the slowest convergence rate. The proof uses classic manipulations in Optimal Transport [38], as well as concentration results of empirical distributions in Wasserstein norm [45]. In particular, it is known that the latter yields slow convergence rates with the dimension $n^{-1/d}$. While the Q_r's live in $\mathbb{R}^{d_z}$, when the c-GCNs are Lipschitz we can replace d_z by the Minkowski dimension of $\mathcal{X}$, which may be advantageous when $\mathcal{X}$ is a low-dimensional manifold.

In the rest of this section, we analyze the stability of c-GCNs to deformation of random graph models, directly through the Wasserstein bound above (or simple Euclidean norm in the invariant case). Finite-sample bounds are then obtained with Theorem 1 and 2.

Stability of continuous GCNs: assumptions. Assume from now on that $\mathcal{X} \subset \mathbb{R}^d$ equipped with the Euclidean norm. Given a diffeomorphism $\tau : \mathcal{X} \to \mathcal{X}$, we consider spatial deformations of random graph models of the form $(\mathrm{Id} - \tau)$, and aim at obtaining bounds of the form (10) for c-GCNs. Given a reference random graph model $\Gamma = (P, W, f)$, we may consider perturbations to P, W, or f, and thus define $W_\tau \stackrel{\text{def.}}{=} (\mathrm{Id} - \tau) \cdot W$, $P_\tau \stackrel{\text{def.}}{=} (\mathrm{Id} - \tau)_\sharp P$ and $f_\tau \stackrel{\text{def.}}{=} (\mathrm{Id} - \tau) \cdot f$. Of course, after deformation, we still consider that the assumptions on our random graph models (6) are verified. As can be expected, translation-invariant kernels W such as Gaussian kernels or ε-graph kernels are particularly adapted to such deformations, therefore we will often make the following assumption:

$$W(x, x') = w(x - x'), \quad C_{\nabla w} \stackrel{\text{def.}}{=} \sup_{x \in \mathcal{X}} \int \left\| \nabla w \left(\tfrac{x - x'}{2} \right) \right\| \cdot \|x' - x\| \, dP(x') < \infty. \tag{A1}$$

We also define $C_W \stackrel{\text{def.}}{=} \sup_x \int |W(x, x')| dP(x') \leqslant c_{\max}$. While C_W and $C_{\nabla w}$ are easily bounded when $W, \nabla w$ are bounded, they are typically much smaller than such naive bounds when W and ∇w are well localized in space with fast decays, *e.g.,* for the Gaussian kernel or a smooth ε-graph kernel with compact support (for instance, in the latter case, C_W is proportional to $\varepsilon c_{\max}$ instead of $c_{\max}$).

In the case where P is replaced by P_τ, some of our results will be valid beyond translation-invariant kernels. We will instead assume that P_τ has a density with respect to P, close to one: for all x,

$$q_\tau(x) \stackrel{\text{def.}}{=} \tfrac{dP_\tau}{dP}(x), \quad q_\tau(x), q_\tau(x)^{-1} \leqslant C_{P_\tau} < \infty, \quad N_P(\tau) \stackrel{\text{def.}}{=} \|q_\tau - 1\|_\infty. \tag{A2}$$

When $(\mathrm{Id} - \tau) \in \Sigma_P$, then we have $N_P(\tau) = 0$, so that $N_P(\tau)$ measures how much it deviates from such neutral elements and quantifies the size of deformations. In particular, when P is proportional to the Lebesgue measure and $\|\nabla \tau\|_\infty < 1$, we have $q_\tau(x) = \det(I - \nabla \tau(x))^{-1}$; then, for small enough $\|\nabla \tau\|_\infty$, we obtain $N_P(\tau) \lesssim d\|\nabla \tau\|_\infty$, recovering the more standard quantity of Mallat [33]. In this case, we also have the bound $C_{P_\tau} \leqslant 2^d$ if we assume $\|\nabla \tau\|_\infty \leqslant 1/2$.

In the rest of the section, we will assume for simplicity that the considered GCNs Φ have zero bias at each layer. Unless otherwise written, $\|f\|$ refers to $L^2(P)$-norm. All the proofs are in Appendix E.

Deformation of translation-invariant kernels. We first consider applying deformations to the kernel W, which amounts to a perturbation to the edge structure of the graph. For GCNs, this affects the Laplacian operator used for the filters, and could be seen as a perturbation of the "graph shift operator" in the framework of Gama et al. [19]. The following result shows that in this case the stability of GCN representations, both invariant and equivariant, is controlled by $\|\nabla \tau\|_\infty$.

Theorem 3 (Kernel deformation). *Consider a GCN representation Φ with no bias and a random graph $\Gamma = (P, W, f)$. Define $Q = \Phi_{W,P}(f)_\sharp P$ and $Q_\tau = \Phi_{W_\tau, P}(f)_\sharp P$. Assume (A1) and $\|\nabla \tau\|_\infty \leqslant 1/2$. We have*

$$\left. \begin{array}{c} \left\| \bar{\Phi}_{W_\tau, P}(f) - \bar{\Phi}_{W,P}(f) \right\| \\ \mathcal{W}_2(Q, Q_\tau) \end{array} \right\} \leqslant C(C_W + C_{\nabla w}) \|f\| \|\nabla \tau\|_\infty, \tag{12}$$

where C is given in the appendix.

Deformation of the distribution. Let us now consider perturbations of P to P_τ, which corresponds to a change in the node distribution. In practice, this may correspond to several, fairly different, "practical" situations. We describe two different frameworks below.

In shape analysis, P may be supported on a manifold, and P_τ can then represent a deformation of this manifold, *e.g.,* a character that rigidly moves a body part. In this case in particular, we can expect $\|\tau\|_\infty$ to be large, but $\|\nabla \tau\|_\infty$ to be small (*i.e.,* large translation but small deformation). Moreover, if the kernel is translation-invariant, there will be little change in the structure of the generated graph. If additionally the input signal of the c-GCN is approximately deformed along with P, then one can expect the outputs to be stable, which we prove in the following theorem.

Theorem 4 (Distribution deformation, translation-invariant case). *Consider a GCN representation Φ with no bias and a random graph $\Gamma = (P, W, f)$, along with a function f'. Define $Q = \Phi_{W,P}(f)_\sharp P$*

7

and $Q_\tau = \Phi_{W,P_\tau}(f')_\sharp P_\tau$. *Assume* (A1) *and* $\|\nabla\tau\|_\infty \leqslant 1/2$. *We have*

$$\left.\begin{array}{r}\left\|\bar\Phi_{W,P}(f) - \bar\Phi_{W,P_\tau}(f')\right\| \\ \mathcal{W}_2(Q, Q_\tau)\end{array}\right\} \leqslant C(C_W + C_{\nabla w})\|f\|\|\nabla\tau\|_\infty + C'\|f'_\tau - f\|, \qquad (13)$$

where C, C' *are given in the appendix.*

When $f = f'$ are both constant, or when $f' = (\mathrm{Id} - \tau)^{-1} \cdot f$, that is, f' is the mapping of the original signal f on the deformed structure, then we have $\|f'_\tau - f\| = 0$. As mentioned before, in the absence of input signal, a standard choice is to take the degree functions as inputs [9, 11]. The next result shows that this choice also leads to the desired stability.

Proposition 3. *Assume* (A1) *and* $\|\nabla\tau\|_\infty \leqslant 1/2$. *If* $f = d_{W,P}$ *and* $f' = d_{W,P_\tau}$, *then we have* $\|f'_\tau - f\| \leqslant C_{\nabla w}\|\nabla\tau\|_\infty$.

Let us now take a look at the case where W is not translation-invariant. We will then assume that P_τ has a density with respect to P, and in particular that it has the same support: one may for instance imagine a social network with a slightly changing distribution of user preferences, SBMs with changing community sizes, geometric random graphs [37], or graphons [31]. The analysis here being slightly more complex, we focus on invariant c-GCNs.

Theorem 5 (Distribution deformation, non-translation-invariant case). *Consider a GCN representation* Φ *with no bias and a random graph* $\Gamma = (P, W, f)$. *Assume* (A2). *We have*

$$\left\|\bar\Phi_{W,P}(f) - \bar\Phi_{W,P_\tau}(f)\right\| \leqslant \left(CC_{P_\tau}^3 C_W + C'\right)\|f\|N_P(\tau), \qquad (14)$$

where C, C' *are given in the appendix.*

As mentioned above, in the case where P is the Lebesgue measure, *e.g.,* for graphons [31], then we recover the quantity $N_P(\tau) \lesssim d\|\nabla\tau\|_\infty$.

Deformations of the signal. Finally, we consider deformations of the signal on the graph and show a bound similar to the ones in the Euclidean case (10). As can be seen in the proofs, this case is in fact a combination of the previous results (12) and (14); hence we must assume both (A1) and (A2) and obtain a dependence on both $\|\nabla\tau\|_\infty$ and $N_P(\tau)$. Once again we focus on invariant c-GCNs with pooling, similar to classical scattering transform [33].

Proposition 4 (Signal deformation). *Consider a GCN representation* Φ *with no bias and a random graph* $\Gamma = (P, W, f)$. *Assume* (A1), (A2), *and* $\|\nabla\tau\|_\infty \leqslant 1/2$. *We have*

$$\left\|\bar\Phi_{W,P}(f) - \bar\Phi_{W,P}(f_\tau)\right\| \leqslant (CC_{P_\tau}^{1/2}(C_W + C_{\nabla w})\|\nabla\tau\|_\infty + \left(CC_{P_\tau}^3 C_W + C'\right)N_P(\tau))\|f\|, \quad (15)$$

where C, C' *are given in the appendix.*

When P is proportional to the Lebesgue measure, since $N_P(\tau)$ is controlled by $\|\nabla\tau\|_\infty$, the GCN is invariant to translations and stable to deformations, similar to Euclidean domains [33]. We note that studies of stability are often balanced by discussions on how the representation preserves signal (*e.g.,* [33, 5, 18]). In our context, the empirical success of GCNs suggests that these representations maintain good discrimination and approximation properties, though a theoretical analysis of such properties for GCNs is missing and provides an important direction for future work.

5   Numerical experiments

In this section, we provide simple numerical experiments on synthetic data that illustrate the convergence and stability of GCNs. We consider untrained GCNs with random weights in order to assess how these properties result from the choice of architecture rather than learning. The code is accessible at https://github.com/nkeriven/random-graph-gnn.

Convergence. Fig. 1 shows the convergence of an equivariant GCN toward a continuous function on a ε-graph with nodes sampled on a 3-dimensional manifold. We take a constant input signal $f = 1$ here to only assess the effect of the manifold shape. We then examine the effect of the sparsity level on convergence on the corresponding invariant GCN, taking an average of several experiments with high n and $\alpha_n = 1$ as an approximation of the "true" unknown limit value. As expected, the convergence is slower for sparse graphs, however we indeed observe convergence to the *same* output for all values of α_n.
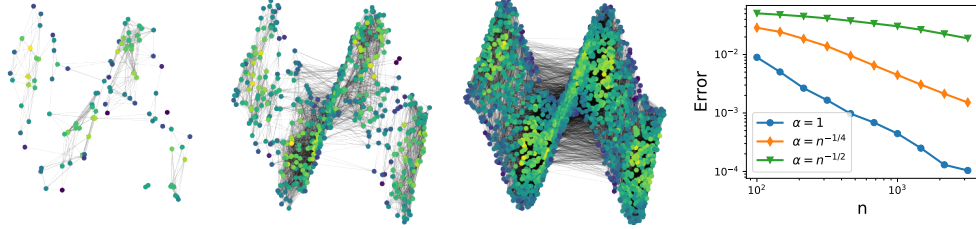
Figure 1: Illustration of convergence of a GCN on random graphs with 3D latent positions and input signal $f = 1$. Left: output signal with growing number of nodes. Right: convergence with different sparsity levels α_n.
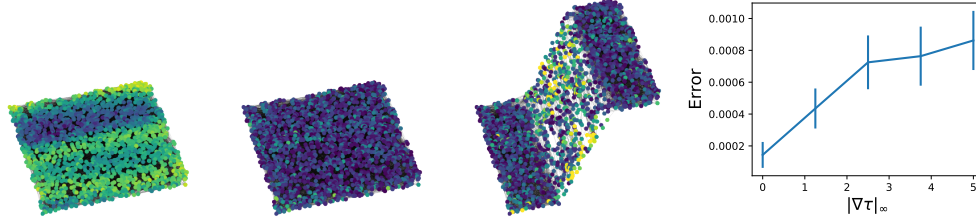

Figure 2: Illustration of stability of a GCN on random graphs with 3D latent positions. From left to right: output signal; difference with the output on the same latent positions but a new drawing of the random edges; difference with the output on deterministically deformed latent positions and corresponding drawing of random edges; difference in output signal of an invariant GCN with respect to the amplitude of the deformation, averaged over 20 experiments.

Stability. In Fig. 2, we illustrate the stability of GCNs to deformations. We first examine the variations in the output of an equivariant GCN when only re-drawing the random edges, with or without modifying the latent positions (in a deterministic manner). We indeed observe that regions that are only translated, such as the "flat" parts of the surface, yield stable output, while deformed regions lead to a deformed output signal. We then verify that a larger deformation leads to a larger distance in output.

6 Conclusion and outlooks

GCNs have proved to be efficient in identifying large-scale structures in graphs and generalizing across graphs of different sizes, which can only partially be explained with discrete graph notions like isomorphisms [47] or stability with permutation-minimizing metrics [18]. In contrast, we have shown that combining them with random models of large graphs allows us to define intuitive notions of deformations and stability in the continuous world like the Euclidean case [33, 5, 39], with direct applications in community-based social networks or shape analysis on point clouds. For this we derived non-asymptotic convergence bounds, valid on relatively sparse random graphs with non-smooth kernels, and new tools like a Wasserstein-type stability bounds on equivariant c-GCNs.

We believe our work to be a first step toward a better understanding of GNNs on large graphs, with many potential outlooks. First, it would be useful to improve the dependence of our bounds on regularity properties of the filters, as done in [19] for the discrete setting, while preserving the mild dependence on the number of filters. In the same vein, finer results may be obtained in particular cases: *e.g.,* the case where $\mathcal{X}$ is a sub-manifold can be studied under the light of Riemannian geometry, stability bounds on SBMs may be expressed with a direct dependence on their parameters, or more explicit stability bounds may be obtained when the (c-)GCN is a structured architecture like the scattering transform on graphs [17]. Convergence results can also be obtained for many other models of random graphs like k-Nearest Neighbor graphs [10]. Finally, while we focus on stability in this paper, as mentioned above the *approximation power* of GCNs (beyond untractable universality [24]) can also be expressed through that of their continuous counterpart, and characterizing which functions are computable by a c-GCN (*e.g.,* with growing width or number of layers) is of foremost importance.

Broader Impact

Graph Neural Networks have been used to many applications, including computer vision, generative models in NLP or protein prediction to cite only a few. Thus, our work is included in a wide literature whose societal impact and ethical considerations are not one-sided. We provide here a theoretical understanding of the behaviour of large random graphs, with the most natural application is community detection in social science [3]. Our contributions, of a theoretical nature, are far from a direct impact in our opinion. We do not see continuous GCNs applied directly in the forseeable future otherwise as a proxy for the study of classic GCNs. Nevertheless, as a stability result, it is a step forward to handle adversarial attacks as highlighted in [19].

Acknowledgements

References

[1] E. Abbe. Community detection and stochastic block models: recent developments. *J. Mach. Learn. Res.*, pages 1–86, 2018.

[2] J. Andén and S. Mallat. Deep scattering spectrum. *IEEE Trans. Signal Process.*, 62(16):4114–4128, 2014.

[3] A.-L. Barabási. *Network Science*. Cambridge University Press, 2016.

[4] M. Belkin and P. Niyogi. Convergence of Laplacian eigenmaps. *NIPS*, pages 129–136, 2007.

[5] A. Bietti and J. Mairal. Group invariance, stability to deformations, and complexity of deep convolutional representations. *J. Mach. Learn. Res.*, 20:1–49, 2019.

[6] B. Bollobas. *Random Graphs*. Cambridge University Press, 2001.

[7] M. M. Bronstein, J. Bruna, Y. Lecun, A. Szlam, and P. Vandergheynst. Geometric Deep Learning: Going beyond Euclidean data. *IEEE Signal Process. Mag.*, 34(4):18–42, 2017.

[8] J. Bruna and S. Mallat. Invariant scattering convolution networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1872–1886, 2013.

[9] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun. Spectral Networks and Locally Connected Networks on Graphs. In *ICLR*, pages 1–14, 2014.

[10] J. Calder and N. G. Trillos. Improved spectral convergence rates for graph Laplacians on epsilon-graphs and k-NN graphs. *arXiv:1910.1347*, pages 1–40, 2019.

[11] Z. Chen, L. Li, and J. Bruna. Supervised community detection with line graph neural networks. In *ICLR*, 2019.

[12] Z. Chen, S. Villar, L. Chen, and J. Bruna. On the equivalence between graph isomorphism testing and function approximation with GNNs. In *NeurIPS*, pages 1–19, 2019.

[13] Z. Chen, L. Chen, S. Villar, and J. Bruna. Can graph neural networks count substructures? *arXiv:2002.04025*, pages 1–36, 2020.

[14] Y. De Castro, C. Lacour, and T. M. P. Ngoc. Adaptive Estimation of Nonparametric Geometric Graphs. *arXiv:1708.02107*, pages 1–39, 2017.

[15] M. Defferrard, X. Bresson, and P. Vandergheynst. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. In *NIPS*, 2016.

[16] P. Diao, D. Guillot, A. Khare, and B. Rajaratnam. Model-free consistency of graph partitioning. *arXiv:1608.03860*, 2016.

[17] F. Gama, J. Bruna, and A. Ribeiro. Diffusion scattering transforms on graphs. In *ICLR*, pages 1–17, 2019.

[18] F. Gama, J. Bruna, and A. Ribeiro. Stability of Graph Scattering Transforms. In *NeurIPS*, pages 1–18, 2019.

[19] F. Gama, J. Bruna, and A. Ribeiro. Stability Properties of Graph Neural Networks. *arXiv:1905.04497*, pages 1–16, 2019.

[20] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl. Neural Message Passing for Quantum Chemistry. In *ICML*, pages 1–14, 2017.

[21] A. Goldenberg, A. X. Zheng, S. E. Fienberg, and E. M. Airoldi. A survey of statistical network models. *Found. Trends Mach. Learn.*, 2(2):129–233, 2009.

[22] P. W. Holland. Stochastic blockmodels: First steps. *Soc. Networks*, 5(2):109–137, 1983.

[23] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, and P. Riley. Molecular graph convolutions: moving beyond fingerprints. *J. Comput. Aided Mol. Des.*, 30(8):595–608, 2016.

[24] N. Keriven and G. Peyré. Universal Invariant and Equivariant Graph Neural Networks. In *NeurIPS*, pages 1–19, 2019.

[25] N. Keriven and S. Vaiter. Sparse and Smooth: improved guarantees for Spectral Clustering in the Dynamic Stochastic Block Model. *ArXiv preprint arXiv:2002.02892*, pages 1–30, 2020.

[26] T. N. Kipf and M. Welling. Semi-Supervised Learning with Graph Convolutional Networks. In *ICLR*, 2017.

[27] E. D. Kolaczyk. *Statistical Analysis of Network Data: Methods and Models*. Springer Series in Statistics, 2010.

[28] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Comput.*, 1(4):541–551, 1989.

[29] J. Lei and A. Rinaldo. Consistency of spectral clustering in stochastic block models. *Ann. Stat.*, 43(1): 215–237, 2015.

[30] R. Levie, W. Huang, L. Bucci, M. M. Bronstein, and G. Kutyniok. Transferability of Spectral Graph Convolutional Neural Networks. *arXiv:1907.12972*, pages 1–41, 2020.

[31] L. Lovász. Large networks and graph limits. *Colloquium Publications*, 60:487, 2012.

[32] A. Magner, M. Baranwal, and A. O. Hero. The Power of Graph Convolutional Networks to Distinguish Random Graph Models. *arXiv:1910.12954*, pages 1–27, 2020.

[33] S. Mallat. Group Invariant Scattering. *Commun. Pure Appl. Math.*, 65(10):1331–1398, 2012.

[34] H. Maron, H. Ben-Hamu, H. Serviansky, and Y. Lipman. Provably Powerful Graph Networks. In *NeurIPS*, pages 1–12, 2019.

[35] H. Maron, E. Fetaya, N. Segol, and Y. Lipman. On the Universality of Invariant Networks. In *ICML*, 2019.

[36] C. Matias and S. Robin. Modeling heterogeneity in random graphs through latent space models: a selective review. *ESAIM: Proc.*, 47:55–74, 2014.

[37] M. Penrose. *Random Geometric Graphs*. 2008.

[38] G. Peyré and M. Cuturi. Computational Optimal Transport. *Found. Trends Mach. Learn.*, 11(5-6):355–206, 2019.

[39] Q. Qiu, X. Cheng, R. Calderbank, and G. Sapiro. DCFNet: Deep Neural Network with Decomposed Convolutional Filters. *ICML*, 9:6687–6696, 2018.

[40] L. Rosasco, M. Belkin, and E. De Vito. On learning with integral operators. *J. Mach. Learn. Res.*, 11: 905–934, 2010.

[41] L. Ruiz, L. F. O. Chamon, and A. Ribeiro. Graphon Signal Processing. *arXiv:2003.05030*, pages 1–13, 2020.

[42] M. Tang and C. E. Priebe. Limit theorems for eigenvectors of the normalized Laplacian for random graphs. *Ann. Stat.*, 46(5):2360–2415, 2018.

[43] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018.

[44] U. Von Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. *Ann. Stat.*, 36(2): 555–586, 2008.

[45] J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4 A):2620–2648, 2019.

[46] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu. A Comprehensive Survey on Graph Neural Networks. *IEEE Trans. Neural Netw. Learn. Syst.*, pages 1–21, 2020.

[47] K. Xu, W. Hu, J. Leskovec, and S. Jegelka. How Powerful are Graph Neural Networks? In *ICLR*, pages 1–15, 2019.

[48] D. Zou and G. Lerman. Graph convolutional neural networks via scattering. *Appl. Comput. Harmon. Anal.*, pages 1–26, 2019.

Supplementary material

In Appendix A, we introduce additional notations and objects that will be used in the proofs. In Appendix B, we study the equivariance of GCNs and prove Props 1 and 2. In Appendix C, we prove Theorem 1 on the convergence of GCNs. In Appendix D, we prove the Wasserstein bound in Theorem 2. In Appendix E, we derive the stability bounds of Section 4. Finally, in Appendix F, we give technical concentration bounds and in Appendix G we provide some third-party results for completeness.

A   Notations

Given a GCN, we define some bounds on its parameters that will be used in the multiplicative constants of the theorem. Recall that the filters are written $h_{ij}^{(\ell)}(\lambda) = \sum_{k=0}^{\infty} \beta_{ijk}^{(\ell)} \lambda^k$. We define $B_k^{(\ell)} = \left(\beta_{ijk}^{(\ell)} \right)_{ji} \in \mathbb{R}^{d_{\ell+1} \times d_\ell}$ the matrix containing the order-k coefficients, and by $B_{k,|\cdot|}^{(\ell)} = \left(\left| \beta_{ijk}^{(\ell)} \right| \right)_{ji}$ the same matrix with absolute value on all coefficients. Then, we define the following bounds:

$$H_2^{(\ell)} = \sum_k \left\| B_k^{(\ell)} \right\| \qquad\qquad H_{\partial,2}^{(\ell)} = \sum_k \left\| B_k^{(\ell)} \right\| k$$

$$H_\infty^{(\ell)} = \sum_k \left\| B_{k,|\cdot|}^{(\ell)} \right\| \left(\frac{2c_{\max}}{c_{\min}} \right)^k \qquad\qquad H_{\partial,\infty}^{(\ell)} = \sum_k \left\| B_k^{(\ell)} \right\| k \left(\frac{2c_{\max}}{c_{\min}} \right)^{k-1}$$

which all converge by our assumptions on the β_k. We may also denote H_2 by $H_{L^2(P)}$ for convenience but this quantity does not depend on P. Note that, only for H_∞, we use the spectral norm of the matrix $B_{k,|\cdot|}$ with non-negative coefficients, which is suboptimal compared to using B_k. This is due to a part of our analysis where we do not operate in a Hilbert space but only in a Banach space $\mathcal{B}(\mathcal{X})$, see Lemma 6. We also define $\left\| b^{(\ell)} \right\| = \sqrt{\sum_j (b_j^{(\ell)})^2}$.

Given X, we define the empirical degree function

$$d_X = d_{W,X} \stackrel{\text{def.}}{=} \frac{1}{n} \sum_i W(\cdot, x_i) \tag{16}$$

Which will be denoted by d_X when the kernel is clear. Although $d_{W,P}$ is bounded away from 0 by the assumption (6), this is not necessarily the case for $d_{W,X}$. This is however true with high probability, as shown by the following Lemma.

Lemma 1. *Let Γ be a model of random graphs. There is a universal constant C such that, if*

$$n \geqslant C D_{\mathcal{X}}(\rho)^2 \tag{17}$$

where $D_{\mathcal{X}}(\rho) = \frac{c_{\text{Lip.}}}{c_{\min}} \sqrt{d_x} + \frac{c_{\max} + c_{\text{Lip.}}}{c_{\min}} \sqrt{\log \frac{n_{\mathcal{X}}}{\rho}}$, then with probability $1 - \rho$, $d_{W,X} \geqslant c_{\min}/2 > 0$.

Proof. Apply Lemma 4 with $f = 1$ to obtain the result. $\qquad\square$

For W and X such that $d_{W,X} > 0$, we define the following empirical Laplacian operator:

$$\mathcal{L}_X f = \mathcal{L}_{W,X} f \stackrel{\text{def.}}{=} \frac{1}{n} \sum_i \frac{W(\cdot, x_i)}{\sqrt{d_X(\cdot) d_X(x_i)}} f(x_i) \tag{18}$$

which we will also denote by $\mathcal{L}_X$ when W is clear. Assuming that $d_X \geqslant c_{\min}/2$, $\mathcal{L}_{W,X}$ is a bounded operator and $\|\mathcal{L}_{W,X}\|_\infty \leqslant \frac{2c_{\max}}{c_{\min}}$.

Given $X = \{x_1, \ldots, x_n\}$ and any dimension d, we denote by S_X the normalized sampling operator acting on functions $f : \mathcal{X} \to \mathbb{R}^d$ defined by $S_X f \stackrel{\text{def.}}{=} \frac{1}{\sqrt{n}} [f(x_1), \ldots, f(x_n)] \in \mathbb{R}^{n \times d}$. The normalizing factor $\frac{1}{\sqrt{n}}$ is natural: we have $\|S_X f\|_F \leqslant \|f\|_\infty$ and by the Law of Large Numbers $\|S_X f\|_F \to \|f\|_{L^2(P)}$ a.s. Finally, given X and W, we define $W(X) \stackrel{\text{def.}}{=} (W(x_i, x_j))_{ij} \in \mathbb{R}^{n \times n}$, and remark that $L(W(X)) \circ S_X = S_X \circ \mathcal{L}_{W,X}$.

13

B Invariance and equivariance

Proof of Prop. 1. The proof is immediate, by observing that $L(\sigma \cdot A) = \sigma \cdot L(A)$, therefore $h(L(\sigma \cdot A))(\sigma \cdot Z) = \sigma \cdot (h(L(A))Z)$, and permutations commute with the pointwise activation function. For the invariant case, we just observe the final pooling on the equivariant case. $\qquad\square$

Proof of Prop. 2. Let us first observe that the degree function is such that

$$d_{W,P}(\phi(x)) = \int W(\phi(x), x')dP(x') = \int (\phi \cdot W)(x, x')d(\phi^{-1})_\sharp P(x') = d_{\phi \cdot W,(\phi^{-1})_\sharp P}(x)$$

Then, we have

$$\phi \cdot (\mathcal{L}_{W,P} f)(x) = \int \frac{W(\phi(x), x')}{\sqrt{d_{W,P}(\phi(x))d_{W,P}(x')}} f(x')dP(x')$$

$$= \int \frac{(\phi \cdot W)(x, x')}{\sqrt{d_{\phi \cdot W,(\phi^{-1})_\sharp P}(x)d_{\phi \cdot W,(\phi^{-1})_\sharp P}(x')}}(\phi \cdot f)(x')d(\phi^{-1})_\sharp P(x')$$

$$= \mathcal{L}_{\phi \cdot W,(\phi^{-1})_\sharp P}(\phi \cdot f)$$

Then, by recursion, we have $\mathcal{L}^k_{\phi \cdot W,(\phi^{-1})_\sharp P}(\phi \cdot f) = \mathcal{L}^{k-1}_{\phi \cdot W,(\phi^{-1})_\sharp P}(\phi \cdot (\mathcal{L}_{W,P} f)) = \ldots = \phi \cdot \mathcal{L}^k_{W,P} f$, and the same is true for filters $h(\mathcal{L})$. We conclude by observing that permutation commutes with pointwise non-linearity: $\rho \circ (\phi \cdot f) = \phi \cdot (\rho \circ f) = \rho \circ f \circ \phi$. The invariant case follows with a final integration against P. $\qquad\square$

C Convergence of GCNs: proof of Theorem 1

We are going to prove Theorem 1 with the following constants:

$$C_1 \propto \frac{c_{\max} + c_{\text{Lip.}}}{c_{\min}} \sum_{\ell=0}^{M-1} C^{(\ell)} H_{\partial,\infty}^{(\ell)} \prod_{s=\ell+1}^{M-1} H_2^{(s)},$$

$$C_2 \propto \frac{c_{\max}}{c_{\min}^2} \sum_{\ell=0}^{M-1} C^{(\ell)} H_{\partial,2}^{(\ell)} \prod_{s=\ell+1}^{M-1} H_2^{(s)},$$

$$C_3 \propto C^{(M)}$$

$$\text{with} \quad C^{(\ell)} \overset{\text{def.}}{=} \|\theta\| \left(\|f\|_\infty \prod_{s=0}^{\ell-1} H_\infty^{(s)} + \sum_{s=0}^{\ell-1} \left\| b^{(s)} \right\| \prod_{p=s+1}^{\ell-1} H_\infty^{(p)} \right) \tag{19}$$

The proof will mainly rely on an application of Dudley's inequality [43, Thm 8.1.6] (Lemma 5 in Appendix F) and a recent spectral concentration inequality for normalized Laplacian in relatively sparse graphs (Theorem 6 in Appendix G).

Proof. We begin the proof by the equivariant case, the invariant case will simply use an additional concentration inequality. Denoting by $Z^{(\ell)}$ (resp. $f^{(\ell)}$) the signal at each layer of the GCN (resp. the function at each layer of the c-GCN), we have

$$\text{MSE}_X \left(\Phi_A(Z), \Phi_{W,P}(f) \right) = \left\| \frac{\Phi_A(Z)}{\sqrt{n}} - S_X \Phi_{W,P}(f) \right\|_F \leqslant \|\theta\| \left\| \frac{Z^{(M)}}{\sqrt{n}} - S_X f^{(M)} \right\|_F$$

where we recall that S_X is the normalized sampling operator (see App. A). We therefore seek to bound that last term.

Assume that the following holds with probability $1 - \rho$: for all $0 \leqslant \ell \leqslant M - 1$

$$\sqrt{\sum_j \left\| \sum_i \left(h_{ij}^{(\ell)}(L)S_X f_i^{(\ell)} - S_X h_{ij}^{(\ell)}(\mathcal{L}_{W,P})f_i^{(\ell)} \right) \right\|^2} \leqslant \Delta^{(\ell)} \tag{20}$$

14

Then, using (4), Lemma 6, and the fact that for unnormalized sampling $\sqrt{n}S_X \circ \rho = \rho \circ (\sqrt{n}S_X)$, we can show by recursion that $\left\| \frac{Z^{(\ell)}}{\sqrt{n}} - S_X f^{(\ell)} \right\|_F \leqslant \varepsilon_\ell$ implies

$$
\left\| \frac{Z^{(\ell+1)}}{\sqrt{n}} - S_X f^{(\ell+1)} \right\|_F
$$

$$
= \left(\sum_j \left\| \frac{1}{\sqrt{n}} \rho \Big(\sum_{i=1}^{d_\ell} h_{ij}^{(\ell)}(L) z_i^{(\ell)} + b_j^{(\ell)} 1_n \Big) - S_X \rho \Big(\sum_{i=1}^{d_\ell} h_{ij}^{(\ell)}(\mathcal{L}_{W,P}) f_i^{(\ell)} + b_j^{(\ell)} 1(\cdot) \Big) \right\|^2 \right)^{\frac{1}{2}}
$$

$$
= \left(\sum_j \frac{1}{n} \left\| \rho \Big(\sum_{i=1}^{d_\ell} h_{ij}^{(\ell)}(L) z_i^{(\ell)} + b_j^{(\ell)} 1_n \Big) - \rho \Big(\sqrt{n} S_X \Big(\sum_{i=1}^{d_\ell} h_{ij}^{(\ell)}(\mathcal{L}_{W,P}) f_i^{(\ell)} + b_j^{(\ell)} 1(\cdot) \Big) \Big) \right\|^2 \right)^{\frac{1}{2}}
$$

$$
\leqslant \left(\sum_j \left\| \sum_{i=1}^{d_\ell} h_{ij}^{(\ell)}(L) \frac{z_i^{(\ell)}}{\sqrt{n}} - S_X h_{ij}^{(\ell)}(\mathcal{L}_{W,P}) f_i^{(\ell)} \right\|^2 \right)^{\frac{1}{2}}
$$

$$
\leqslant \left(\sum_j \left\| \sum_{i=1}^{d_\ell} h_{ij}^{(\ell)}(L) \left(\frac{z_i^{(\ell)}}{\sqrt{n}} - S_X f_i^{(\ell)} \right) \right\|^2 \right)^{\frac{1}{2}}
$$

$$
+ \left(\sum_j \left\| \sum_{i=1}^{d_\ell} h_{ij}^{(\ell)}(L) S_X f_i^{(\ell)} - S_X h_{ij}^{(\ell)}(\mathcal{L}_{W,P}) f_i^{(\ell)} \right\|^2 \right)^{\frac{1}{2}}
$$

$$
\leqslant \varepsilon_{\ell+1} \stackrel{\text{def.}}{=} H_2^{(\ell)} \varepsilon_\ell + \Delta^{(\ell)}
$$

Since $\frac{Z^{(0)}}{\sqrt{n}} = S_X f^{(0)}$ we have $\varepsilon_0 = 0$ and an easy recursion shows that

$$
\left\| \frac{Z^{(M)}}{\sqrt{n}} - S_X f^{(M)} \right\|_F \leqslant \sum_{\ell=0}^{M-1} \Delta^{(\ell)} \prod_{s=\ell+1}^{M-1} H_2^{(s)} \tag{21}
$$

We now need to prove that (20) holds with probability $1 - \rho$ for all ℓ with the appropriate $\Delta^{(\ell)}$. Recall that $L(W(X)) \circ S_X = S_X \circ \mathcal{L}_{W,X}$, and that by (17), with probability $1 - \rho/2$ we have $\|\mathcal{L}_{W,X}\|_\infty \leqslant \frac{2c_{\max}}{c_{\min}}$. Assuming this is satisfied, by Lemma 6 we have

$$
\sqrt{ \sum_j \left\| \sum_i \left(h_{ij}^{(\ell)}(L) S_X f_i^{(\ell)} - S_X h_{ij}^{(\ell)}(\mathcal{L}_{W,P}) f_i^{(\ell)} \right) \right\|^2 }
$$

$$
\leqslant \sqrt{ \sum_j \left\| \sum_i \left(h_{ij}^{(\ell)}(L) - h_{ij}^{(\ell)}(L(W(X))) \right) S_X f_i^{(\ell)} \right\|^2 }
$$

$$
+ \sqrt{ \sum_j \left\| \sum_i S_X \left(h_{ij}^{(\ell)}(\mathcal{L}_{W,X}) - h_{ij}^{(\ell)}(\mathcal{L}_{W,P}) \right) f_i^{(\ell)} \right\|^2 }
$$

$$
\leqslant H_{\partial,2}^{(\ell)} \|L - L(W(X))\| \left\| f^{(\ell)} \right\|_\infty
$$

$$
+ \sum_k \|B_k\| \sqrt{ \sum_i \left(\sum_{\ell=0}^{k-1} \left(\frac{2c_{\max}}{c_{\min}} \right)^\ell \left\| (\mathcal{L}_{W,X} - \mathcal{L}_{W,P}) \mathcal{L}_{W,P}^{k-1-\ell} f_i^{(\ell)} \right\|_\infty \right)^2 } \tag{22}
$$

The first term in (22) is handled with a recent concentration inequality for normalized Laplacian in the relatively sparse graphs with random edges [25], recalled as Theorem 6 in Appendix G. We use the following version.

15

Corollary 1 (of Theorem 6). *Assume* (17) *is satisfied and* $n \geqslant 1/\rho$ *for simplicity. There is a universal constant* C *such that, if*

$$\alpha_n \geqslant \frac{Cc_{\max}}{c_{\min}^2} \cdot \frac{\log n}{n} \tag{23}$$

Then, with probability at least $1 - \rho$, *we have*

$$\|L - L(W(X))\| \lesssim \frac{c_{\max}}{c_{\min}^2} \cdot \frac{1}{\sqrt{\alpha_n n}}$$

Proof. By (17) with the appropriate constant, with probability $1 - \rho/2$ we have $d_X \geqslant c_{\min}/2$, we can therefore apply Theorem 6 to bound $\|L - L(W(X))\|$ conditionally on X using $c \sim 1 + \frac{\log(1/\rho)}{\log(n)} \sim 1$, then use a union bound to conclude. $\qquad\square$

We now bound the second term in (22). Define $\rho_k = \frac{C\rho}{(k+1)^2 \sum_\ell d_\ell}$ with C such that $\sum_{k\ell} d_\ell \rho_k = \rho/4$ (even when the filters are not of finite order). Using an application of Dudley's inequality detailed in Lemma 5 in Appendix F and a union bound, we obtain with probability $1 - \rho/4$ that: for all i, ℓ, k, we have

$$\left\| (\mathcal{L}_{W,X} - \mathcal{L}_{W,P}) \mathcal{L}_{W,P}^k f_i^{(\ell)} \right\|_\infty \lesssim \frac{c_{\max} \left\| \mathcal{L}_{W,P}^k f_i^\ell \right\|_\infty D_\mathcal{X}(\rho_k)}{c_{\min}\sqrt{n}}$$

$$\leqslant \left(\frac{c_{\max}}{c_{\min}} \right)^k \frac{c_{\max} \|f_i^\ell\|_\infty D_\mathcal{X}\left(\frac{C\rho}{(k+1)^2 \sum_\ell d_\ell} \right)}{c_{\min}\sqrt{n}}$$

$$\lesssim \left(\frac{2c_{\max}}{c_{\min}} \right)^k \frac{(c_{\max} + c_{\text{Lip.}}) \|f_i^\ell\|_\infty D_\mathcal{X}\left(\frac{\rho}{\sum_\ell d_\ell} \right)}{c_{\min}\sqrt{n}}$$

Coming back to the second term of (22), with probability $1 - \rho/4$:

$$\sum_k \|B_k\| \sqrt{\sum_i \left(\sum_{\ell=0}^{k-1} \left(\frac{2c_{\max}}{c_{\min}} \right)^\ell \left\| (\mathcal{L}_{W,X} - \mathcal{L}_{W,P}) \mathcal{L}_{W,P}^{k-1-\ell} f_i^{(\ell)} \right\|_\infty \right)^2}$$

$$\lesssim \frac{(c_{\max} + c_{\text{Lip.}}) D_\mathcal{X}\left(\frac{\rho}{\sum_\ell d_\ell} \right)}{c_{\min}\sqrt{n}} \sum_k \|B_k\| k \left(\frac{2c_{\max}}{c_{\min}} \right)^k \sqrt{\sum_i \left\| f_i^{(\ell)} \right\|_\infty^2}$$

$$\leqslant \frac{(c_{\max} + c_{\text{Lip.}}) D_\mathcal{X}\left(\frac{\rho}{\sum_\ell d_\ell} \right)}{c_{\min}\sqrt{n}} H_{\partial,\infty}^{(\ell)} \left\| f^{(\ell)} \right\|_\infty$$

At the end of the day we obtain that with probability $1 - \rho$, (20) is satisfied with

$$\Delta^{(\ell)} \propto \left\| f^{(\ell)} \right\|_\infty \left(\frac{H_{\partial,2}^{(\ell)} c_{\max}}{c_{\min}^2 \sqrt{\alpha_n n}} + \frac{H_{\partial,\infty}^{(\ell)} (c_{\max} + c_{\text{Lip.}}) D_\mathcal{X}\left(\frac{\rho}{\sum_\ell d_\ell} \right)}{c_{\min}\sqrt{n}} \right)$$

We then use Lemma 8 to bound $\left\| f^{(\ell)} \right\|_\infty$ and conclude.

Finally, in the invariant case we have

$$\left\| \bar{\Phi}_A(Z) - \bar{\Phi}_{W,P}(f) \right\| \leqslant \mathrm{MSE}_X \left(\Phi_A(Z), \Phi_{W,P}(f) \right) + \|\theta\| \left\| \frac{1}{n} \sum_i f^{(M)}(x_i) - \mathbb{E}f^{(M)}(X) \right\|$$

Using a vector Hoeffding's inequality (Lemma 11) and a bound on $\left\| f^{(M)} \right\|_\infty$ by Lemma 8 we bound the second term and conclude. $\qquad\square$

D   Wasserstein convergence: proof of Theorem 2

We are going to prove Theorem 2 with the following constants:

$$C_1 \propto \|\theta\| \left(\max_{r=1,2} \|f_r\|_\infty \prod_{s=0}^{\ell-1} H_\infty^{(s)} + \sum_{s=0}^{\ell-1} \left\| b^{(s)} \right\| \prod_{p=s+1}^{\ell-1} H_\infty^{(p)} \right) \qquad C_2 = 27^{d_z/4} \qquad (24)$$

$$C_1' \propto (n_\mathcal{X} n_f)^{\frac{1}{d_x}} \max_{r=1,2} D_r \qquad\qquad\qquad\qquad\qquad\qquad C_2' = 27^{d_x/4}$$

where D_r are defined as (36) with the function f_r as input. The proof will mainly rely on results of the concentration rate of the empirical distribution of iid data to its true value in Wasserstein norm [45] (Theorem 7 in Appendix G).

Proof. For $r = 1,2$, define $y_{r,i} = \Phi_{W_r,P_r}(f_r)(x_{r,i})$ which are drawn iid from $Q_r \overset{\text{def}}{=} \Phi_{W_r,P_r}(f_r)_\sharp P_r$, and we denote by $\hat{Q}_r = n^{-1} \sum_i \delta_{y_{r,i}}$ the empirical distributions. By Theorem 1 and triangle inequality, we have

$$\min_\sigma \sqrt{n^{-1} \sum_i \left\| \Phi_{A_1}(Z_1)_i - \Phi_{A_2}(Z_2)_{\sigma(i)} \right\|^2} \leqslant \min_\sigma \sqrt{\frac{1}{n} \sum_i \left\| y_{1,i} - y_{2,\sigma(i)} \right\|^2} + 2R_n$$

The first term is known, among other appellations, as the so-called "Monge" formulation of optimal transport (OT) [38, Chap. 2]. For uniform weights, as is the case here, it is known that Monge formulation of OT is equivalent to its Kantorovich relaxation, which in turns gives the traditional Wasserstein metric: by [38, Prop. 2.1], we have

$$\sqrt{\frac{1}{n} \sum_i \left\| y_{1,i} - y_{2,\sigma(i)} \right\|^2} = \mathcal{W}_2(\hat{Q}_1, \hat{Q}_2) \leqslant \mathcal{W}_2(Q_1, Q_2) + \sum_{r=1,2} \mathcal{W}_2(\hat{Q}_r, Q_r)$$

We must therefore bound the distance between an empirical distribution $\hat{Q}$ and its true value $Q = g_\sharp P$ for some function $g = \Phi_{W,P}(f)$.

The distribution Q is supported on $g\mathcal{X} \subset \mathbb{R}^{d_z}$, which is bounded by $\|g\|_\infty$ which can be bounded by Lemma 8. Hence its covering numbers are as $N(g\mathcal{X}, \varepsilon, \|\cdot\|) \leqslant (\|g\|_\infty / \varepsilon)^{d_z}$. We can then conclude by Theorem 7.

When f is (c_f, n_f)-piecewise Lipschitz however, by Lemma 9 g is $(C, n_\mathcal{X} n_f)$-Lipschitz where C is defined as (36), and in this case it is easy to see that the covering numbers of $g\mathcal{X}$ also satisfy $N(g\mathcal{X}, \varepsilon, \|\cdot\|) \leqslant n_\mathcal{X} n_f (C/\varepsilon)^{d_x}$. Applying again Theorem 7, we conclude. $\qquad\square$

E   Stability

In this section, norms $\|\cdot\|$ always refer to $L^2(P)$ norms (for functions and operators), and we drop the subscript for simplicity. We denote the pooling operator by $U_P f = \int f dP$. We denote $U = U_P$ and $\mathcal{L} = \mathcal{L}_{W,P}$ for short. For $\tau : \mathcal{X} \to \mathcal{X}$, we denote $W_\tau = (\text{Id} - \tau) \cdot W$, $P_\tau = (\text{Id} - \tau)_\sharp P$, and $f_\tau = (\text{Id} - \tau) \cdot f$. Then we define the shorthands $\mathcal{L}_{W_\tau} = \mathcal{L}_{W_\tau,P}$, $\mathcal{L}_{P_\tau} = \mathcal{L}_{W,P_\tau}$, the composition operator $T_\tau f = (\text{Id} - \tau) \cdot f$ and $U_\tau = U_{P_\tau}$. Finally, we define A and A_τ the integral operators of kernels W and W_τ w.r.t. the measure P, the corresponding diagonal degree operators by D and D_{W_τ}, so that we have $\mathcal{L} = D^{-1/2} A D^{-1/2}$ and similarly for $\mathcal{L}_{W_\tau}$. Similarly, we define D_{P_τ} the diagonal degree operator of W but with respect to P_τ.

We observe that for two functions f, f' and any P, we have

$$\left. \begin{array}{c} \left| \int f dP - \int f' dP \right| \\ \mathcal{W}_2(f_\sharp P, f'_\sharp P) \end{array} \right\} \leqslant \|f - f'\|_{L^2(P)} \qquad (25)$$

The second inequality is immediate by considering the definition $\mathcal{W}_2^2(f_\sharp P, f'_\sharp P) = \inf \left\{ \mathbb{E} \|f(X) - f'(Y)\|^2 \mid X \sim P, Y \sim P \right\}$ and taking $X = Y$ as a coupling. We will therefore manipulate mainly $L^2(P)$ norms.

E.1 Proof of Theorem 3 (deformation change to W)

We are going to prove Theorem 3 with constant

$$C \propto \|\theta\| \, c_{\min}^{-2} \sum_{\ell=0}^{M-1} H_{\partial,2}^{(\ell)} \prod_{s=0,\ s\neq\ell}^{M-1} H_2^{(s)} \tag{26}$$

In this setting, we have two random graph models whose only difference is in the choice of kernel, from W to W_τ, while fixing P and f. This in turn leads to a change in the Laplacian, which is the main quantity that we will need to control.

Since we have assumed the bias to be zero, we have the following Lemma, which we apply with $f = f'$ in this section.

Lemma 2. *We have*

$$\|\Phi_{W,P}(f) - \Phi_{W_\tau,P}(f')\| \leqslant C \, \|f\| \, \|\mathcal{L} - \mathcal{L}_{W_\tau}\| + C' \, \|f - f'\|$$

with

$$C = \|\theta\| \sum_{\ell=0}^{M-1} H_{\partial,2}^{(\ell)} \prod_{s=0,\ s\neq\ell}^{M-1} H_2^{(s)} \tag{27}$$

$$C' = \|\theta\| \prod_{\ell=0}^{M-1} H_2^{(\ell)} \tag{28}$$

Proof. Denoting $f^{(\ell)}$ and $(f^{(\ell)})'$ the functions at each layer, we have using Lemma 6 and (4):

$$\left\| f^{(\ell)} - (f^{(\ell)})' \right\| \leqslant \sqrt{\sum_j \left\| \sum_{i=1}^{d_{\ell-1}} h_{ij}^{(\ell-1)}(\mathcal{L}) f_i^{(\ell-1)} - h_{ij}^{(\ell-1)}(\mathcal{L}_{W_\tau})(f_i^{(\ell-1)})' \right\|^2}$$

$$\leqslant \sqrt{\sum_j \left\| \sum_{i=1}^{d_{\ell-1}} (h_{ij}^{(\ell-1)}(\mathcal{L}) - h_{ij}^{(\ell-1)}(\mathcal{L}_{W_\tau})) f_i^{(\ell-1)} \right\|_*^2}$$

$$+ \sqrt{\sum_j \left\| \sum_{i=1}^{d_{\ell-1}} h_{ij}^{(\ell-1)}(\mathcal{L}_{W_\tau})(f_i^{(\ell-1)} - (f_i^{(\ell-1)})') \right\|^2}$$

$$\leqslant H_{\partial,2}^{(\ell-1)} \|\mathcal{L} - \mathcal{L}_{W_\tau}\| \left\| f^{(\ell-1)} \right\| + H_2^{(\ell-1)} \left\| f^{(\ell-1)} - (f^{(\ell-1)})' \right\|$$

An easy recursion and Lemma 8 give the result. $\square$

The rest of the proof then consists in obtaining a bound on the quantity $\|\mathcal{L}_{W_\tau} - \mathcal{L}\|$ in $L^2(P)$. We have

$$\mathcal{L}_{W_\tau} - \mathcal{L} = D_{W_\tau}^{-\frac{1}{2}} A_\tau D_{W_\tau}^{-\frac{1}{2}} - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$$

$$= D_{W_\tau}^{-\frac{1}{2}} A_\tau (D_{W_\tau}^{-\frac{1}{2}} - D^{-\frac{1}{2}}) + D_{W_\tau}^{-\frac{1}{2}} (A_\tau - A) D^{-\frac{1}{2}} + (D_{W_\tau}^{-\frac{1}{2}} - D^{-\frac{1}{2}}) A D^{-\frac{1}{2}}. \tag{29}$$

We now bound different operators in this decomposition separately.

Bound on $\|A_\tau - A\|$. Define

$$k(x, x') = w(x - \tau(x) - x' + \tau(x')) - w(x - x'), \tag{30}$$

so that $A_\tau - A$ is an integral operator with kernel k. We have, by the fundamental theorem of calculus,

$$k(x, x') = \int_0^1 \langle \nabla w(x - x' + t(\tau(x') - \tau(x))), \tau(x') - \tau(x) \rangle dt.$$

18

Now, note that we have

$$|\tau(x') - \tau(x)| \leqslant \|\nabla\tau\|_\infty \cdot \|x - x'\|$$

$$|x - x' + t(\tau(x') - \tau(x))| \geqslant \frac{1}{2}\|x - x'\|,$$

where the last inequality follows from the reverse triangle inequality and the assumption $\|\nabla\tau\|_\infty \leqslant 1/2$. By Cauchy-Schwarz, and since $\|\nabla w(x)\|$ decreases with $\|x\|$, we have

$$\int |k(x, x')| dP(x') \leqslant \|\nabla\tau\|_\infty \int \|\nabla w((x - x')/2)\| \cdot \|x' - x\|\, dP(x')$$

$$\leqslant C_{\nabla w}\|\nabla\tau\|_\infty.$$

Similarly, we obtain $\int |k(x, x')| dP(x) \leqslant C_{\nabla w}\|\nabla\tau\|_\infty$. Then, Schur's test (Lemma 12) yields

$$\|A_\tau - A\| \leqslant C_{\nabla w}\|\nabla\tau\|_\infty. \tag{31}$$

Bound on $\|D_{W_\tau}^{-1/2} - D^{-1/2}\|$. Define $d = d_{W,P}$ and $d_\tau = d_{W_\tau, P}$. The operator $D_{W_\tau}^{-1/2} - D^{-1/2}$ is diagonal with elements $d_\tau(x)^{-1/2} - d(x)^{-1/2}$, such that $\|D_{W_\tau}^{-1/2} - D^{-1/2}\| \leqslant \left\|d_\tau^{-1/2} - d^{-1/2}\right\|_\infty$. Note that we have

$$|d_\tau(x) - d(x)| = \left|\int k(x, x') dP(x')\right| \leqslant C_{\nabla w}\|\nabla\tau\|_\infty,$$

Then as in the proof of Lemma 5 we have $\left\|d_\tau^{-1/2} - d^{-1/2}\right\|_\infty \leqslant c_{\min}^{-\frac{3}{2}}\|d_\tau - d\|_\infty \leqslant C_{\nabla w}c_{\min}^{-\frac{3}{2}}\|\nabla\tau\|_\infty$.

Final bound. Note that by Schur's test, we have $\|A\| \leqslant C_w$, $\|A_\tau\| \leqslant C_w + C_{\nabla w}\|\nabla\tau\|_\infty$. Further, we have $\|D^{-1/2}\|, \|D_{W_\tau}^{-1/2}\| \leqslant c_{\min}^{-1/2}$. Plugging back into (29), we have the desired bound.

E.2   Proof of Theorem 4 and 5 (deformation change to P)

In this case, we have two random graph models with distributions P and P_τ, while W remains fixed. Depending on the case, the input f will change or not.

Translation-invariant case.   We are going to prove Theorem 4 with C defined as (26) and C' as (28). Using (25) and Prop 2, we obtain that

$$\mathcal{W}_2\left(\Phi_{W,P_\tau}(f')_\sharp P_\tau, \Phi_{W,P}(f)_\sharp P\right) \leqslant \|T_\tau\Phi_{W,P_\tau}(f') - \Phi_{W,P}(f)\| = \|\Phi_{W_\tau,P}(f'_\tau) - \Phi_{W,P}(f)\|$$

In the invariant case, we have similarly that $\left|\bar{\Phi}_{W,P_\tau}(f') - \bar{\Phi}_{W,P}(f)\right| = \left|\bar{\Phi}_{W,P_\tau}(f') - \bar{\Phi}_{W,P}(f)\right| \leqslant \|\Phi_{W_\tau,P}(f'_\tau) - \Phi_{W,P}(f)\|$. Using Lemma 2 and the computation of the previous section, we obtain

$$\|\Phi_{W_\tau,P}(f'_\tau) - \Phi_{W,P}(f)\| \leqslant C(C_W + C_{\nabla w})\|f\|\|\nabla\tau\|_\infty + C'\|f'_\tau - f\|,$$

where C is defined by (26) and C' is defined by (28).

Proof of Prop. 3. When using degree functions as input, using the proof of Prop 2 and the computations of the previous sections we have

$$\|T_\tau d_{W,P_\tau} - d_{W,P}\| = \|d_{W_\tau,P} - d_{W,P}\| \leqslant C_{\nabla w}\|\nabla\tau\|_\infty$$

$\square$

Non Translation-invariant case.   We are going to prove Theorem 5 with constants C defined as (26) and C' defined as (28). By Assumption (A2) we easily have the following:

$$C_{P_\tau}^{-1/2}\|f\|_{L^2(P_\tau)} \leqslant \|f\| \leqslant C_{P_\tau}^{1/2}\|f\|_{L^2(P_\tau)} \tag{32}$$

such that for any k we have $\left\|\mathcal{L}_{P_\tau}^k\right\| \leqslant C_{P_\tau}$ by observing that

$$\|\mathcal{L}_{P_\tau}^k f\| \leqslant C_{P_\tau}^{1/2}\|\mathcal{L}_{P_\tau}^k f\|_{L^2(P_\tau)} \leqslant C_{P_\tau}^{1/2}\|f\|_{L^2(P_\tau)} \leqslant C_{P_\tau}\|f\|$$

We can then prove the following Lemma.

Lemma 3 (Stability in terms of measure change). *We have*

$$\left\| \bar{\Phi}_{W,P_\tau}(f) - \bar{\Phi}_{W,P}(f) \right\| \leqslant C_{P_\tau}^2 C_1 \|f\| \|\mathcal{L}_{P_\tau} - \mathcal{L}\| + C_2 \|f\| \|U_\tau - U\| \tag{33}$$

with C_1 is the same as (27) and $C_2 = \|\theta\| \prod_{\ell=0}^{M-1} H_2^{(\ell)}$.

Proof. From a simple triangle inequality and the estimate $\|U_\tau\| \leqslant C_{P_\tau}$ since $|U_\tau f| = \left| \int f dP_\tau \right| = \left| \int f q_\tau dP \right| \leqslant C_{P_\tau} \|f\|$, we have

$$\begin{aligned}
\left\| \bar{\Phi}_{W,P_\tau}(f) - \bar{\Phi}_{W,P}(f) \right\| &= \left\| U_\tau \Phi_{W,P_\tau}(f) - U \Phi_{W,P}(f) \right\| \\
&\leqslant C_{P_\tau} \left\| \Phi_{W,P_\tau}(f) - \Phi_{W,P}(f) \right\| + \|U_\tau - U\| \left\| \Phi_{W,P}(f) \right\|
\end{aligned}$$

We must now bound the first term. Denoting $f^{(\ell)}$ and $(f^{(\ell)})'$ the functions at each layer, we have using Lemma 6, (4) and the fact that $\left\| \mathcal{L}_{P_\tau}^k \right\| \leqslant C_{P_\tau}$:

$$\begin{aligned}
\left\| f^{(\ell)} - (f^{(\ell)})' \right\| &\leqslant \sqrt{ \sum_j \left\| \sum_{i=1}^{d_{\ell-1}} h_{ij}^{(\ell-1)}(\mathcal{L}) f_i^{(\ell-1)} - h_{ij}^{(\ell-1)}(\mathcal{L}_{P_\tau})(f_i^{(\ell-1)})' \right\|^2 } \\
&\leqslant \sqrt{ \sum_j \left\| \sum_{i=1}^{d_{\ell-1}} (h_{ij}^{(\ell-1)}(\mathcal{L}) - h_{ij}^{(\ell-1)}(\mathcal{L}_{P_\tau}))(f_i^{(\ell-1)})' \right\|^2 } \\
&\quad + \sqrt{ \sum_j \left\| \sum_{i=1}^{d_{\ell-1}} h_{ij}^{(\ell-1)}(\mathcal{L})(f_i^{(\ell-1)} - (f_i^{(\ell-1)})') \right\|^2 } \\
&\leqslant C_{P_\tau} H_{\partial,2}^{(\ell-1)} \|\mathcal{L} - \mathcal{L}_{P_\tau}\| \left\| (f^{(\ell-1)})' \right\| + H_2^{(\ell-1)} \left\| f^{(\ell-1)} - (f^{(\ell-1)})' \right\|
\end{aligned}$$

Then, we use Lemma 8 and the fact that there is no bias to obtain

$$\left\| (f^{(\ell-1)})' \right\| \leqslant C_{P_\tau}^{1/2} \left\| (f^{(\ell-1)})' \right\|_{L^2(P_\tau)} \leqslant C_{P_\tau}^{1/2} \|f\|_{L^2(P_\tau)} \prod_{s=0}^{\ell-1} H_2^{(s)} \leqslant C_{P_\tau} \|f\| \prod_{s=0}^{\ell-1} H_2^{(s)}$$

an easy recursion gives the result. $\qquad\square$

We first bound $\|U_\tau - U\|$ easily, by

$$|U_\tau f - U f| = \left| \int f(x) dP_\tau(x) - \int f(x) dP(x) \right| = \left| \int f(x) \left(q_\tau(x) - 1 \right) dP(x) \right| \leqslant N_P(\tau) \|f\|$$

which is also true for multivariate functions.

We now bound $\|\mathcal{L}_{P_\tau} - \mathcal{L}\|$. If we denote J_τ the diagonal change of variables operator with elements $q_\tau(x)$, we may write

$$\begin{aligned}
\mathcal{L}_{P_\tau} - \mathcal{L} &= D_{P_\tau}^{-1/2} A D_{P_\tau}^{-1/2} J_\tau - D^{-1/2} A D^{-1/2} \\
&= (D_{P_\tau}^{-1/2} - D^{-1/2}) A D_{P_\tau}^{-1/2} J_\tau + D^{-1/2} A (D_{P_\tau}^{-1/2} - D^{-1/2}) J_\tau + D^{-1/2} A D^{-1/2} (J_\tau - \mathrm{Id})
\end{aligned}$$

The following estimates are easily obtained using Schur's test or by a pointwise supremum: $\|A\| \leqslant C_w$, $\|J_\tau\| \leqslant C_{P_\tau}$, $\|J_\tau - \mathrm{Id}\| \leqslant N_P(\tau)$ and $\|D^{-1/2}\|, \|D_{P_\tau}^{-1/2}\| \leqslant c_{\min}^{-1/2}$. It is left to bound and $\|D_{P_\tau}^{-1/2} - D^{-1/2}\|$. As before,

$$\|D_{P_\tau}^{-1/2} - D^{-1/2}\| \leqslant c_{\min}^{-3/2} \|d_\tau - d\|_\infty .$$

and

$$\begin{aligned}
|d_\tau(x) - d(x)| &= \left| \int W(x,x') dP_\tau(x') - \int W(x,x') dP(x') \right| \\
&= \left| \int W(x,x') \left(q_\tau(x') - 1 \right) dP(x') \right| \leqslant C_w N_P(\tau)
\end{aligned}$$

E.3 Proof of Proposition 4 (signal deformations to f)

This is just a triangle inequality, combined with the previous theorems and Prop 2:

$$\left\|\bar{\Phi}_{W,P}(T_\tau f) - \bar{\Phi}_{W,P}(f)\right\| \leqslant \left\|\bar{\Phi}_{W,P}(T_\tau f) - \bar{\Phi}_{W_\tau,P}(T_\tau f)\right\| + \left\|\bar{\Phi}_{W_\tau,P}(T_\tau f) - \bar{\Phi}_{W,P}(f)\right\|$$

$$\leqslant C(C_W + C_{\nabla w})\|\nabla\tau\|_\infty \|T_\tau f\| + \left\|\bar{\Phi}_{W,P_\tau}(f) - \bar{\Phi}_{W,P}(f)\right\|$$

$$\leqslant C(C_W + C_{\nabla w})C_{P_\tau}^{1/2}\|f\|\|\nabla\tau\|_\infty + (CC_{P_\tau}^3 C_W + C')\|f\| N_P(\tau)$$

where C is given by (26) and C' is given by (28).

F Technical Lemma

F.1 Concentration inequalities

Lemma 4 (Chaining on non-normalized kernels). *Consider a kernel W and a probability distribution P satisfying (6), any function $f \in \mathcal{B}(\mathcal{X})$, and $x_1, \ldots, x_n$ drawn iid from P. Then, with probability at least $1 - \rho$,*

$$\left\|\frac{1}{n}\sum_i W(\cdot, x_i)f(x_i) - \int W(\cdot, x)f(x)dP(x)\right\|_\infty \lesssim \frac{\|f\|_\infty \left(c_{\text{Lip.}}\sqrt{d_x} + (c_{\max} + c_{\text{Lip.}})\sqrt{\log\frac{n_\mathcal{X}}{\rho}}\right)}{\sqrt{n}}$$

Proof. Without lost of generality, we do the proof for $\|f\|_\infty \leqslant 1$. For any $x \in \mathcal{X}$, define

$$Y_x = \frac{1}{n}\sum_i W(x, x_i)f(x_i) - \int W(x, x')f(x')dP(x')$$

Since $\|W(\cdot, x)f\|_\infty \leqslant c_{\max}$, for any fixed $x_0 \in \mathcal{X}$, by Hoeffding's inequality we have: with probability at least $1 - \rho$,

$$|Y_{x_0}| \lesssim \frac{c_{\max}\sqrt{\log(1/\rho)}}{\sqrt{n}}$$

Consider any $j \leqslant n_\mathcal{X}$. For any $x_0 \in \mathcal{X}_j$, we have

$$\sup_{x \in \mathcal{X}_j}|Y_x| \leqslant \sup_{x,x' \in \mathcal{X}_j}|Y_x - Y_{x'}| + |Y_{x_0}|$$

The second term is bounded by the inequality above. For the first term, we are going to use Dudley's inequality "tail bound" version [43, Thm 8.1.6]. We first need to check the sub-gaussian increments of the process Y_x. For any $x, x' \in \mathcal{X}_j$, we have

$$\|Y_x - Y_{x'}\|_{\psi_2} \lesssim \frac{1}{n}\left(\sum_{i=1}^n \|(W(x, x_i) - W(x', x_i))f(x_i) - (T_{W,P}f(x) - T_{W,P}f(x'))\|_{\psi_2}^2\right)^{\frac{1}{2}}$$

$$\lesssim \frac{1}{n}\left(\sum_{i=1}^n \|(W(x, x_i) - W(x', x_i))f(x_i)\|_{\psi_2}^2\right)^{\frac{1}{2}}$$

$$\lesssim \frac{1}{n}\left(\sum_{i=1}^n \|(W(x, \cdot) - W(x', \cdot))f(\cdot)\|_\infty^2\right)^{\frac{1}{2}}$$

$$\leqslant \frac{c_{\text{Lip.}}}{\sqrt{n}}d(x, x')$$

where we have used, from [43], Prop. 2.6.1 for the first line, Lemma 2.6.8 for the second, Example 2.5.8 for the third, and the Lipschitz property of W for the last.

Now, we apply Dudley's inequality [43, Thm 8.1.6] to obtain that with probability $1 - \rho$,

$$\sup_{x,x' \in \mathcal{X}_j}|Y_x - Y_{x'}| \lesssim \frac{c_{\text{Lip.}}}{\sqrt{n}}\left(\int_0^1 \sqrt{\log N(\mathcal{X}, d, \varepsilon)}d\varepsilon + \sqrt{\log(1/\rho)}\right)$$

$$\lesssim \frac{c_{\text{Lip.}}}{\sqrt{n}}\left(\sqrt{d_x} + \sqrt{\log(1/\rho)}\right)$$

21

Combining with the decomposition above and applying a union bound over the $\mathcal{X}_j$ yields the desired result. □

Lemma 5 (Chaining on normalized Laplacians). *Consider a kernel W and a probability distribution P satisfying* (6), *any function $f \in \mathcal{B}(\mathcal{X})$, and $x_1, \ldots, x_n$ drawn iid from P. Assume n satisfies* (17). *Then with probability at least $1 - \rho$,*

$$\|(\mathcal{L}_X - \mathcal{L}_P)f\|_\infty \lesssim \frac{c_{\max} \|f\|_\infty D_{\mathcal{X}}(\rho)}{c_{\min}\sqrt{n}} \tag{34}$$

where $D_{\mathcal{X}}(\rho) = \frac{1}{c_{\min}}\left(c_{\text{Lip.}}\sqrt{d_x} + (c_{\max} + c_{\text{Lip.}})\sqrt{\log \frac{n_{\mathcal{X}}}{\rho}}\right)$.

Proof. Again we assume $\|f\|_\infty \leqslant 1$ without lost of generality.

By Lemma 4 with $f = 1$ and (17), with probability $1 - \rho/2$ we have

$$\|d_X - d_P\|_\infty \lesssim \varepsilon_d \overset{\text{def.}}{=} \frac{c_{\text{Lip.}}\sqrt{d_x \log C_{\mathcal{X}}} + (c_{\max} + c_{\text{Lip.}})\sqrt{\log \frac{n_{\mathcal{X}}}{\rho}}}{\sqrt{n}} \leqslant \frac{c_{\min}}{2}$$

and in particular $d_X \geqslant c_{\min}/2$. In this case, for all x, we have

$$\left|\frac{1}{\sqrt{d_X(x)}} - \frac{1}{\sqrt{d_P(x)}}\right| \leqslant \frac{|d_P(x) - d_X(x)|}{\sqrt{d_X(x)d_P(x)}(\sqrt{d_X(x)} + \sqrt{d_P(x)})} \lesssim \frac{\varepsilon_d}{c_{\min}^{3/2}}$$

and for all x, y,

$$\left|\frac{1}{\sqrt{d_X(y)d_X(x)}} - \frac{1}{\sqrt{d_P(y)d_P(x)}}\right| \leqslant \frac{1}{\sqrt{d_X(y)}}\left|\frac{1}{\sqrt{d_X(x)}} - \frac{1}{\sqrt{d_P(x)}}\right|$$

$$+ \frac{1}{\sqrt{d_P(x)}}\left|\frac{1}{\sqrt{d_X(y)}} - \frac{1}{\sqrt{d_P(y)}}\right| \lesssim \frac{\varepsilon_d}{c_{\min}^2}$$

Now, define $\bar{W}(x,y) \overset{\text{def.}}{=} \frac{W(x,y)}{\sqrt{d_P(x)d_P(y)}}$. For all $j \leqslant n_{\mathcal{X}}$ and $x, x' \in \mathcal{X}_j$ and $y \in \mathcal{X}$, we have $\|\bar{W}(\cdot,y)\|_\infty \leqslant \frac{c_{\max}}{c_{\min}}$ and

$$|\bar{W}(x,y) - \bar{W}(x',y)| \leqslant \frac{|W(x,y)|}{\sqrt{d_P(y)}}\left|\frac{1}{\sqrt{d_P(x)}} - \frac{1}{\sqrt{d_P(x')}}\right| + \frac{1}{\sqrt{d_P(x')d_P(y)}}|W(x,y) - W(x',y)|$$

$$\lesssim \frac{c_{\text{Lip.}}c_{\max}}{c_{\min}^2}d(x,x')$$

Hence by applying Lemma 4 we obtain that with probability $1 - \rho/2$,

$$\left\|\frac{1}{n}\sum_i \bar{W}(\cdot,x_i)f(x_i) - \int \bar{W}(\cdot,x)f(x)dP(x)\right\|_\infty$$

$$\lesssim \varepsilon_W \overset{\text{def.}}{=} \frac{\frac{c_{\max}}{c_{\min}}\left(\frac{c_{\text{Lip.}}}{c_{\min}}\sqrt{d_x \log C_{\mathcal{X}}} + \left(1 + \frac{c_{\text{Lip.}}}{c_{\min}}\right)\sqrt{\log \frac{n_{\mathcal{X}}}{\rho}}\right)}{\sqrt{n}}$$

We can now conclude, observing that

$$\|(\mathcal{L}_X - \mathcal{L}_P)f\|_\infty = \left\|\frac{1}{n}\sum_i \frac{W(\cdot,x_i)}{\sqrt{d_X(\cdot)d_X(x_i)}}f(x_i) - \int \frac{W(\cdot,x)}{\sqrt{d_P(\cdot)d_P(x)}}f(x)dP(x)\right\|_\infty$$

$$\leqslant \sup_x \frac{1}{n}\sum_i |W(x,x_i)f(x_i)|\left|\frac{1}{\sqrt{d_X(x)d_X(x_i)}} - \frac{1}{\sqrt{d_P(x)d_P(x_i)}}\right|$$

$$+ \left\|\frac{1}{n}\sum_i \bar{W}(\cdot,x_i)f(x_i) - \int \bar{W}(\cdot,x)f(x)dP(x)\right\|_\infty$$

$$\lesssim \frac{c_{\max}}{c_{\min}^2}\varepsilon_d + \varepsilon_W \lesssim \frac{c_{\max}}{c_{\min}^2}\varepsilon_d$$

□

F.2 Misc. bounds

Lemma 6 (Operator norms of filters). *Let $(E, \|\cdot\|_E)$ be a Banach space and $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ be a separable Hilbert space. Let L, L' be two bounded operators on E, and $S : E \to \mathcal{H}$ be a linear operator such that $\|S\|_{\mathcal{H} \to E} \leqslant 1$. For $1 \leqslant i \leqslant d$ and $1 \leqslant j \leqslant d'$, let $h_{ij} = \sum_k \beta_{ijk} \lambda^k$ be a collection of analytic filters, with $B_k = (\beta_{ijk})_{ji} \in \mathbb{R}^{d' \times d}$ the matrix of order-k coefficients, with operator norm $\|B_k\|$. Let $x_1, \ldots, x_d \in E$ be a collection of points. Then:*

$$\sqrt{\sum_j \left\| S \sum_i h_{ij}(L) x_i \right\|_{\mathcal{H}}^2} \leqslant \left(\sum_k \|B_k\| \|L^k\| \right) \sqrt{\sum_i \|x_i\|_E^2}$$

and

$$\sqrt{\sum_j \left\| S \sum_i (h_{ij}(L) - h_{ij}(L')) x_i \right\|_{\mathcal{H}}^2} \leqslant \sum_k \|B_k\| \sqrt{\sum_i \left(\sum_{\ell=0}^{k-1} \|L^\ell\| \|(L - L')(L')^{k-1-\ell} x_i\|_E \right)^2}$$

When $\mathcal{H}$ is only a Banach space, the same results hold with $B_{k,|\cdot|} = (|\beta_{ijk}|)_{ji}$ instead of B_k.

Proof. Let $\{e_\ell\}_{l \geqslant 1}$ be an orthonormal basis for $\mathcal{H}$. For all i, k, decompose $SL^k x_i = \sum_\ell b_{ik\ell} e_\ell$. We have

$$\sqrt{\sum_j \left\| S \sum_i h_{ij}(L) x_i \right\|_{\mathcal{H}}^2} \leqslant \sqrt{\sum_j \left\| \sum_{ik} \beta_{ijk} SL^k x_i \right\|_{\mathcal{H}}^2} \leqslant \sum_k \sqrt{\sum_j \left\| \sum_{i\ell} \beta_{ijk} b_{ik\ell} e_\ell \right\|_{\mathcal{H}}^2}$$

$$\leqslant \sum_k \sqrt{\sum_\ell \sum_j \left(\sum_i \beta_{ijk} b_{ik\ell} \right)^2}$$

$$\leqslant \sum_k \sqrt{\|B_k\|^2 \sum_{i\ell} b_{ik\ell}^2} = \sum_k \|B_k\| \sqrt{\sum_i \|SL^k x_i\|_{\mathcal{H}}^2}$$

$$\leqslant \left(\sum_k \|B_k\| C^k \right) \sqrt{\sum_i \|x_i\|_E^2}$$

The proof of the second claim is obtained in the same way by decomposing $S(L^k - (L')^k) x_i$ in $\mathcal{H}$ and using at the last step:

$$\left\| (L^k - (L')^k) x \right\|_E = \left\| \sum_{\ell=0}^{k-1} L^\ell (L - L')(L')^{k-1-\ell} x \right\|_E \leqslant \sum_{\ell=0}^{k-1} C_\ell \left\| (L - L')(L')^{k-1-\ell} x \right\|_E$$

Finally, when $\mathcal{H}$ is not a Hilbert space, we directly use

$$\sqrt{\sum_j \left\| S \sum_i h_{ij}(L) x_i \right\|_{\mathcal{H}}^2} \leqslant \sqrt{\sum_j \left(\sum_{ik} |\beta_{ijk}| C^k \|x_i\|_E \right)^2} \leqslant \sum_k C^k \sqrt{\sum_j \left(\sum_i |\beta_{ijk}| \|x_i\|_E \right)^2}$$

$$\leqslant \left(\sum_k \|B_{k,|\cdot|}\| C^k \right) \sqrt{\sum_i \|x_i\|_E^2}$$

$\square$

Lemma 7 (Lipschitz property of discrete GCNs). *Let $G_1 = (A, Z_1)$ and $G_2 = (A, Z_2)$ be two graphs with the same structure, and a GCN Φ. Denote by $Z_r^{(M)}$ the signal at the last layer when applying Φ to G_r. We have*

$$\left\| Z_1^{(M)} - Z_2^{(M)} \right\|_F \leqslant \left(\prod_{\ell=0}^{M-1} H_2^{(\ell)} \right) \|Z_1 - Z_2\|_F$$

Proof. For $j \leqslant d_\ell$, using Lemma 6 and (4) we write

$$\left\|Z_1^{(M)} - Z_2^{(M)}\right\|_F \leqslant \sqrt{\sum_j \left\|\sum_{i=1}^{d_{M-1}} h_{ij}^{(M-1)}(L)(z_{1,i}^{(M-1)} - z_{2,i}^{(M-1)})\right\|^2}$$

$$\leqslant H_2^{(M-1)} \left\|Z_1^{(M-1)} - Z_2^{(M-1)}\right\|_F$$

An easy recursion gives the result. $\qquad\square$

Lemma 8 (Bound on c-GCNs). *Apply a c-GCN to a random graph model $\Gamma = (P, W, f)$. Denote by $f^{(\ell)}$ the function at each layer. Then we have*

$$\left\|f^{(\ell)}\right\|_* \leqslant \|f\|_* \prod_{s=0}^{\ell-1} H_*^{(s)} + \sum_{s=0}^{\ell-1} \left\|b^{(s)}\right\| \prod_{p=s+1}^{\ell-1} H_*^{(p)} \tag{35}$$

where $*$ indicates $L^2(P)$ or ∞.

Proof. For $j \leqslant d_\ell$, using Lemma 6 and (4) we write

$$\left\|f^{(\ell)}\right\|_* \leqslant \sqrt{\sum_j \left\|\sum_{i=1}^{d_{\ell-1}} h_{ij}^{(\ell-1)}(\mathcal{L}_{W,P})f_i^{(\ell-1)} + b_j^{(\ell-1)}\right\|_*^2}$$

$$\leqslant \sqrt{\sum_j \left\|\sum_{i=1}^{d_{\ell-1}} h_{ij}^{(\ell-1)}(\mathcal{L}_{W,P})f_i^{(\ell-1)}\right\|_*^2} + \left\|b^{(\ell-1)}\right\|$$

$$\leqslant H_*^{(\ell-1)} \left\|f^{(\ell-1)}\right\|_* + \left\|b^{(\ell-1)}\right\|$$

An easy recursion gives the result. $\qquad\square$

Lemma 9 (Piecewise Lipschitz property of c-GCNs). *Let Γ be a random graph model. Assume that f is piecewise (c_f, n_f)-Lipschitz. Then, $\Phi_{W,P}(f)$ is piecewise $(C, n_f n_\mathcal{X})$-Lipschitz with*

$$C = \|\theta\| \left(c_f \prod_{\ell=0}^{M-1} \left\|B_0^{(\ell)}\right\| + \frac{c_{\text{Lip}}.c_{\max}}{c_{\min}^2} \sum_{\ell=0}^{M-1} H_2^{(\ell)} \left\|f^{(\ell)}\right\|_{L^2(P)} \prod_{s=0}^{\ell-1} \left\|B_0^{(s)}\right\| \right) \tag{36}$$

where $\left\|f^{(\ell)}\right\|_{L^2(P)}$ can be bounded by Lemma 8.

Proof. Define the partition $\mathcal{X}_1', \ldots, \mathcal{X}_{n_f}'$ on which f is Lipschitz, and take $x, x' \in \mathcal{X}_i \cap \mathcal{X}_j'$ for some i, j.

Using the same strategy as in the proof of Lemma 6, we have

$$\left\|f^{(M)}(x) - f^{(M)}(x')\right\| \leqslant \sqrt{\sum_j \left|\sum_i h_{ij}^{(M-1)}(\mathcal{L}_P)f_i^{(M-1)}(x) - h_{ij}^{(M-1)}(\mathcal{L}_P)f_i^{(M-1)}(x')\right|^2}$$

$$\leqslant \sum_k \left\|B_k^{(M-1)}\right\| \sqrt{\sum_i \left|\mathcal{L}_P^k f_i^{(M-1)}(x) - \mathcal{L}_P^k f_i^{(M-1)}(x')\right|^2}$$

Define $\bar{W}(x, y) = \frac{W(x,y)}{\sqrt{d_P(x)d_P(y)}}$. As we have seen in the proof of Lemma 5, $\bar{W}$ is piecewise $\frac{c_{\text{Lip}}.c_{\max}}{c_{\min}^2}$-Lipschitz on the $\mathcal{X}_i$. So, for $k \geqslant 1$, for $x, x' \in \mathcal{X}_i$ by Schwartz inequality we have

$$\left|\mathcal{L}_P^k f_i^{(M-1)}(x) - \mathcal{L}_P^k f_i^{(M-1)}(x')\right| \leqslant \left\|\mathcal{L}_P^{k-1} f_i^{(M-1)}\right\|_{L^2(P)} \frac{c_{\text{Lip}}.c_{\max}}{c_{\min}^2} d(x, x')$$

$$\leqslant \left\|f_i^{(M-1)}\right\|_{L^2(P)} \frac{c_{\text{Lip}}.c_{\max}}{c_{\min}^2} d(x, x')$$

24

And thus

$$\left\| f^{(M)}(x) - f^{(M)}(x') \right\| \leqslant \left\| B_0^{(M-1)} \right\| \left\| f^{(M)}(x) - f^{(M)}(x') \right\|$$
$$+ H_2^{(M-1)} \left\| f^{(M-1)} \right\|_{L^2(P)} \frac{c_{\text{Lip.}} c_{\max}}{c_{\min}^2} d(x, x')$$

A recursion gives the result, with Lemma 8. $\qquad\square$

G Third-party results

Lemma 10 (Hoeffding's inequality). *Let $X_1, \ldots, X_n \in \mathbb{R}$ be independent random variables such that $a \leqslant X_i \leqslant b$ almost surely. Then we have*

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_i (X_i - \mathbb{E}X_i) \right| \geqslant \varepsilon \right) \leqslant 2 \exp\left(-\frac{2\varepsilon^2 n}{(b-a)^2} \right) \tag{37}$$

Lemma 11 (Generalized Hoeffding's inequality [40]). *Let $\mathcal{H}$ be a separable Hilbert space and $\xi_1, \ldots, \xi_n \in \mathcal{H}$ be independent zero-mean random variables such that $\|\xi_i\| \leqslant C$ almost surely. Then with probability at least $1 - \rho$ we have*

$$\left\| \frac{1}{n} \sum_i \xi_i \right\| \leqslant \frac{C\sqrt{2\log(2/\rho)}}{\sqrt{n}} \tag{38}$$

Theorem 6 (Spectral concentration of normalized Laplacian [25, Theorem 4]). *Let A be an adjacency matrix of a graph drawn with independent edges $a_{ij} \sim \text{Ber}(\alpha_n p_{ij})$, where $p_{ij} \leqslant c_{\max}$ and for all i, $\frac{1}{n} \sum_j p_{ij} \geqslant c_{\min} > 0$. Denote by P the $n \times n$ matrix containing the p_{ij}. There is a universal constant C such that:*

$$\mathbb{P}\left(\|L(A) - L(P)\| \geqslant \frac{C(1+c)c_{\max}}{c_{\min}^2 \sqrt{n\alpha_n}} \right) \leqslant e^{-\left(\frac{3c^2}{12+4c} - \log(14) \right)n} + e^{-\frac{3c^2}{12+4c} n\alpha_n + \log(n)}$$
$$+ e^{-\frac{3c_{\min}^2 n\alpha_n}{25 c_{\max}} + \log(n)} + n^{-\frac{c}{4}+6} \tag{39}$$

Theorem 7 (Wasserstein convergence [45]). *Let $(\mathcal{X}, d)$ be a compact metric space with $\text{diam}(\mathcal{X}) \leqslant B$ and $N_\varepsilon(\mathcal{X}) \leqslant (B/\varepsilon)^{d_x}$. Let P be a probability distribution on $\mathcal{X}$, and $x_1, \ldots, x_n$ drawn iid from P. With probability $1 - \rho$,*

$$\mathcal{W}_2(\hat{P}, P) \lesssim B \left(n^{-\frac{1}{d_x}} + \left(27^{\frac{d_x}{4}} + \log(1/\rho)^{\frac{1}{4}} \right) n^{-\frac{1}{4}} \right) \tag{40}$$

where $\hat{P} = n^{-1} \sum_i \delta_{x_i}$.

Proof. The result is obtained by combining Prop. 5 and Prop. 20 in [45] with $\varepsilon' = 1$, with the assumed simplified expression for the covering numbers of $\mathcal{X}$ and a rescaling of the metric such that B disappears from the covering numbers expression. $\qquad\square$

Lemma 12 (Schur's test). *Let T be the integral operator defined by*

$$Tf(x) = \int k(x, x') f(x') d\mu(x').$$

If the kernel k satisfies

$$\sup_x \int |k(x, x')| d\mu(x') \leqslant C \quad \text{and} \quad \sup_{x'} \int |k(x, x')| d\mu(x) \leqslant C,$$

then T is bounded in $L^2(\mu)$, with $\|T\|_{L^2(\mu)} \leqslant C$.