The Graphical Nadaraya-Watson Estimator in Latent Position Models

Martin GJORGJEVSKI[1] Nicolas KERIVEN[1] Simon BARTHELMÉ[1] Yohann DE CASTRO[2]

[1]GIPSA Lab, CNRS, Grenoble, France

[2]Ecole Centrale Lyon, Lyon, France

Résumé – Étant donné un graphe avec un sous-ensemble de nœuds étiquetés, nous sommes intéressés par la qualité de l'estimateur de la moyenne qui, pour un nœud non étiqueté, prédit la moyenne des observations de ses voisins étiquetés. Nous étudions rigoureusement les propriétés de concentration, les limites de variance et les limites de risque dans ce contexte. Bien que l'estimateur lui-même soit très simple, nous pensons que nos resultats contribueront à la compréhension théorique de méthodes plus sophistiquées telles que les réseaux de neurones de graphes.

Abstract – Given a graph with a subset of labeled nodes, we are interested in the quality of the averaging estimator which for an unlabeled node predicts the average of the observations of its labeled neighbours. We rigorously study concentration properties, variance bounds and risk bounds in this context. While the estimator itself is very simple we believe that our results will contribute towards the theoretical understanding of learning on graphs through more sophisticated methods such as Graph Neural Networks.

1 Introduction

Given a undirected graph on $n+1$ vertices (e.g. nodes represent people) and adjacency matrix $A = [a_{i,j}]$ (e.g. edges represent social relationships) where all but the $(n+1)$-st node have measurements y_i (e.g. salary, living expenses, etc), the graph regression problem adresses prediction of the (continuous valued) measurement y_{n+1} of the remaining node. While there are various sophisticated designs of Graph Neural Networks [6, 9, 16, 17] which can tackle this problem, little has been done in terms of statistical analysis of *any* potential solution in the context where the graph is random. In this paper we will consider the simplest estimator, which for the missing label of node $n+1$ is taking the average over all of its neighbours, i.e.

$$\hat{y}_{n+1} = \frac{\sum_{j=1}^{n} y_j a_{j,n+1}}{\sum_{j=1}^{n} a_{j,n+1}} \tag{1}$$

To our knowledge, the statistical properties of this estimator have *not* been studied in the statistical or machine learning literature. In this article we seek to understand the properties of this estimator in a context when the *graph* is treated as random. To conduct our analysis we will work with a *random graph model* known as Latent Position Model (LPM) [7], where to each node one associates a *latent* position in space, and typically nodes that lie closer together in the latent space are more likely to be linked. Due to the nature of this model, the estimator (1) will resemble the *Nadaraya-Watson* (NW) estimator, a popular regression estimator in the nonparametric estimation literature [15], which just averages measurements in a window centered at the location where signal prediction is to take place. For this reason we decide to title it the **Graphical Nadaraya-Watson** (GNW) estimator. There are two major differences between the NW and the GNW estimator. First, NW is more expressive in the sense that it comes with a tunable parameter (*bandwidth*) h_n, which sets the size of the neighbourhood over which to average. For the GNW there is no tunable parameter, because neighbourhood size is imposed by the graph. Therefore, the performance of GNW depends on a *latent* bandwidth h_n that *is not user-chosen* and determines connectivity patterns in the random graph. Second, in the analysis of GNW we include a parameter α_n which affects the sparsity of the observed graph.

We show that the GNW estimator is strongly related to a NW estimator with bandwidth h_n. Surprisingly, despite the added randomness in the GNW estimator, its error is within a multiplicative constant of the NW error. In addition, our analysis holds in a regime where the graph is asymptotically almost sparse. Whereas most of the methods discussed in the literature require degree of *logarithmic* order i.e. $d_n \geq C\log(n)$ [11, 12], we show that the variance of GNW will converge to zero as soon as $d_n \to \infty$, covering *almost all* regimes of sparsity (excluding the bounded degree regime $d_n \leq D$).

2 Regression in the Latent Position Model

2.1 Background on the LPM

The Latent Position Model (LPM) [7] is a generative model which samples a random graph on n nodes in two stages. First, a sample of n i.i.d. *latent* variables $X_i \in \mathbb{R}^d$ with density p is drawn. Second, for each pair of nodes i, j a Bernoulli variable with parameter $k(X_i, X_j)$ determines if there is an edge between nodes i and j. Here, k is a symmetric kernel on $\mathbb{R}^d$, taking values in $[0,1]$. The edge generating Bernoulli variables are conditionally independent given the latent variables. Intuitively we are more likely to observe an edge between two nodes with positions that are similar with respect to k.

When k is a similarity kernel (e.g. radial basis function) as in the NW estimator (4), edges are likely to occur between nodes whose latent positions are nearby in the latent space. As an example, when $k(x,z) = \mathbb{I}(||x-z|| \leq h)$, NW and GNW coincide. This is the random geometric graph [13]. By allowing for discontinous kernels, LPMs can instantiate a

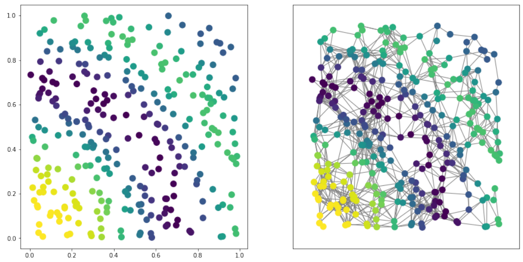

Figure 1 – Sampling a LPM: Left - generating uniformly 300 latent positions on $[0,1]^2$. Right: generating a graph with gaussian kernel $k(x,y) = \exp\left(-\frac{||x-y||^2}{h^2}\right)$

model with intrinsic community structure known as Stochastic Block Model (SBM) [8]. In the framework of LPMs, the graph classification problem been addressed in [14], which employed the method of adjacency spectral embedding. In contrast, we use a local averaging approach.

As large graphs in the real world tend to be sparse [1], a significant effort in the community detection literature is dedicated to understanding statistical properties of graphs with low expected degrees [2, 10, 11, 12]. In such frameworks one considers asymptotic regimes where $k_n(x,z) = \alpha_n K(x,z)$ with K being a fixed kernel and $\alpha_n \to 0$ as $n \to \infty$. The scaling α_n determines the sparsity of the graph where $n\alpha_n$ is the expected degree of the graph, up to a multiplicative constant. The parameter α_n is **not user-chosen** (and not user-known). In this sense, we will instead consider asymptotic regimes with

$$k_n(x,z) = \alpha_n K\left(\frac{x-z}{h_n}\right) \tag{2}$$

where $K\colon \mathbb{R}^d \to [0,1]$ is compactly supported, $0 < \alpha_n \leq 1$ and $h_n > 0$, with α_n, h_n being parameters that are *not user-chosen.*

Once again, it is worth highlighting the key distinction between the setups for NW (4) and GNW (1): the user has the freedom to select the bandwidth h_n in the case of NW, whereas in GNW, this choice is not within the user's control. As $k_n(x,z) \leq \alpha_n$, we see that smaller values of α_n will generate sparser graphs.

The benefit of Latent Position Models is that they allow us to perform graph analysis in the familiar setting of Euclidean geometry. We will use that correspondence to relate signal prediction of graphs to classical nonparametric regression.

2.2 Nonparametric Regression

The regression problem can be stated as estimating a *regression* function $f\colon \mathbb{R}^d \to \mathbb{R}$ given noisy measurements

$$Y_i = f(X_i) + \epsilon_i \tag{3}$$

where $f\colon \mathbb{R}^d \to \mathbb{R}$ with $||f||_\infty \leq B$, ϵ_i additive centered noise with finite variance. One is also given the data points $X_1, ..., X_n$ which can be either deterministic (fixed design) or random i.i.d. samples from a distribution with density p (random design). A classical approach is the weighted average *Nadaraya-Watson* estimator [3, 5, 15]

$$\hat{f}_{NW}(x) = \begin{cases} \frac{\sum_{i=1}^n Y_i k(x,X_i)}{\sum_{i=1}^n k(x,X_i)} & \text{if } \sum_{i=1}^n k(x,X_i) \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Here, $k(x,z) = K\left(\frac{x-z}{h}\right)$ depends on function $K\colon \mathbb{R}^d \to \mathbb{R}$ and the *bandwidth* h which controls the scale on which the data is being averaged. This parameter needs to be chosen carefully, as too small values of h produce estimates of high variance, while too large values of h give highly biased estimators, an instance of the *Bias-Variance tradeoff*, a well known phenomenon in statistics. There are two main measures of statistical performance for NW (4), the *pointwise* and *integrated risk*. For a given point $x \in \mathbb{R}^d$, the pointwise risk is given by

$$\mathcal{R}\left(\hat{f}_{NW}(x), f(x)\right) = \mathbb{E}\left[\left(\hat{f}_{NW}(x) - f(x)\right)^2\right] \tag{5}$$

where the expectation is taken over the noise and the data points $X_1, ..., X_n$ for the random design setting (only over the noise for the fixed design). It is also known as *mean squared error (MSE)*. This metric is local in the sense that it only captures statistical information for a particular point. A metric that captures global statistical information is the *integrated risk* given by

$$\mathcal{R}\left(\hat{f}_{NW}, f\right) = \int \mathcal{R}\left(\hat{f}_{NW}(x), f(x)\right) p(x)dx \tag{6}$$

The integrated risk is also known as *mean integrated squared error (MISE)* and can be interpreted as the risk for a new random variable X with density p, independent from the data $X_1, ..., X_n$.

2.3 Framework and Outline

We observe a random graph with $n+1$ nodes sampled according to a LPM and assume that for all nodes but the last there is a label of the form (3). Conditionally on node $n+1$ having latent position $x \in \mathbb{R}^d$, we write $a(x, X_i)$ for the indicator of an edge between the node $n+1$ and node i. With this notation, the GNW estimator (1) becomes

$$\hat{f}_{GNW}(x) = \begin{cases} \frac{\sum_{i=1}^n Y_i a(x,X_i)}{\sum_{i=1}^n a(x,X_i)} & \text{if } \sum_{i=1}^n a(x,X_i) \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

For $x \in \mathbb{R}^d$, we introduce the *local expected degree* $d_n(x)$ by

$$d_n(x) = n \int_{\mathbb{R}^d} k_n(x,z)p(z)dz \tag{8}$$

which acts as local sample size. Indeed, $d_n(x)$ is the expected degree of node $n+1$ given that its latent position is x. Finally, for $x \in \mathbb{R}^d$ and $f\colon \mathbb{R}^d \to \mathbb{R}$ bounded, we introduce the *local averaging operator*

$$b_n(f,x) = \begin{cases} \frac{\int f(z) k_n(x,z) p(z) dz}{\int k_n(x,z)p(z)dz} & \text{if } d_n(x) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

which can be considered as a biased version of $f(x)$ (the sequence of operators $b_n(f,x)$ resembles an approximation of identity and under mild conditions on K and f discussed in Section 4 one has $b_n(f,x) \to f(x)$ if $h_n \to 0$). Note that $b_n(f,x)$ does not depend on α_n in (2). Similarly to the pointwise risk (5) of NW (4), we will consider the *pointwise risk* for GNW (7)

$$\mathcal{R}\left(\hat{f}_{GNW}(x), f(x)\right) = \mathbb{E}\left[\left(\hat{f}_{GNW}(x) - f(x)\right)^2\right] \quad (10)$$

where the expectation is taken over all random variables appearing in the model (*edge randomness, latent positions and noise*), as well as the *integrated risk*

$$\mathcal{R}\left(\hat{f}_{GNW}, f\right) = \int \mathcal{R}\left(\hat{f}_{GNW}(x), f(x)\right) p(x)dx \quad (11)$$

Our main result is a bound on the *integrated risk* (11). The approach taken in this paper is to bound (10) over the support of p and then to integrate the result. To this goal, we follow a bias-variance decomposition inspired approach by introducing the *bias proxy*

$$b_n(x) = b_n(f,x) - f(x) \quad (12)$$

and the *variance proxy*

$$v_n(x) = \mathbb{E}\left[\left(\hat{f}_{GNW}(x) - b_n(f,x)\right)^2\right] \quad (13)$$

We emphasize that the bias and variance proxies **do not** correspond to the standard notions of bias and variance of GNW and the standard bias-variance decomposition does not hold. Indeed, in Section 3, Proposition (1) we compute $\mathbb{E}[\hat{f}_{GNW}(x)]$ explicitly. For simplicity we choose to work with the bias (12) and variance (13) *proxies* instead of the standard bias-variance decomposition. It can be shown that the bias and variance proxies are within $O(e^{-2d_n(x)})$ of the *true* bias and variance [4]. Moreover, for bounded functions f with $||f||_\infty \leq B$, we can replace the bias-variance decomposition by the inequality

$$\mathcal{R}\left(\hat{f}_{GNW}(x), f(x)\right) \leq \min\left\{2\left[b_n^2(x) + v_n(x)\right], 4B^2\right\} \quad (14)$$

Hence it suffices to provide bounds on the bias (12) and variance proxies (13). As the variance proxy (13) can be treated in greater generality, we tackle it first in Section 3. The bias proxy (12) is treated in Section 4, along with the pointwise (10) and integrated (11) risks.

3 Statistical properties of GNW

The main result of this section is a bound on the variance proxy (13) of order $1/d_n(x)$. Additionally, we show that under *bounded noise*, $\hat{f}_{GNW}(x)$ (7) concentrates around $b_n(f,x)$ (9) with a rate that is exponentially decaying in $d_n(x)$ (8). Finally, we derive an explicit formula for $\mathbb{E}[\hat{f}_{GNW}(x)]$ in terms of $b_n(f,x)$ and $d_n(x)$. Our key insight is the **decoupling trick**, a novel technique specialised for *Bernoulli* variables which introduces independence into the weights of GNW that otherwise are *ratios* of dependent variables. For proofs we refer to the extended version of this paper [4].

Theorem 3.1 *Suppose that $f\colon \mathbb{R}^d \to \mathbb{R}$ with $||f||_\infty \leq B$ and $\mathbb{E}[\epsilon_1^2] = \sigma^2$. Then*

$$\frac{\sigma^2\left(1 - e^{-d_n(x)}\right)^2}{d_n(x)} \leq v_n(x) \leq \frac{261B^2 + 65\sigma^2}{d_n(x)}$$

For the asymptotic analysis, it states that the variance proxy (13) tends to zero *as soon as* $d_n(x) \to \infty$. Conversely, for bounded degree graphs, in the presence of noise the variance proxy (13) stays bounded away from 0. With stronger assumptions on the distribution on the noise one can prove stronger *concentration* results, as given by the following Theorem.

Theorem 3.2 *Suppose that $f\colon \mathbb{R}^d \to \mathbb{R}$ is bounded with $||f||_\infty \leq B$ and the noise variables satisfy $|\epsilon_i| \leq \sigma$. Then*

$$\mathbb{P}\left(|\hat{f}_{GNW}(x) - b_n(f,x)| \geq \delta\right) \leq 6\exp(-Cd_n(x))$$

Here, $C > 0$ depends on δ,B and σ, but not on the sample size n.

We conclude this section by a formula for $\mathbb{E}[\hat{f}_{GNW}(x)]$. Interestingly, for smooth kernels, an analogous explicit formula for $\mathbb{E}[\hat{f}_{NW}(x)]$ is not available in the kernel regression literature.

Proposition 1 *Suppose that $||f||_\infty \leq B$. Then*

$$\mathbb{E}\left[\hat{f}_{GNW}(x)\right] = b_n(f,x)\left[1 - (1 - \frac{d_n(x)}{n})^n\right]$$

4 Risk of GNW

So far we have bounded the variance proxy (13) in terms of the local expected degree (8) (Theorem 3.1). In this section we bound the pointwise (10) and integrated (11) risk. In view of the bound (14) to bound the risk (10) we need to bound the bias proxy (12), and we also need to understand the relationship between the degree $d_n(x)$ and the parameters α_n and h_n. These problems are addressed in Subsection 4.1. In order to bound the integrated risk (11), we need to control the pointwise risk (10) uniformly over the support of the data distribution p. This problem is addressed in Subsection 4.2.

4.1 Bias bound and Pointwise risk of GNW

In the NW literature, it is standard to assume compactly supported kernels [5, 15], an assumption that we adopt as well.

Assumption 1 *There exists $M_1, M_2 > 0$ such that for all $z \in \mathbb{R}^d$ we have $\frac{1}{2}\mathbb{I}\left(||z|| \leq M_1\right) \leq K(z) \leq \mathbb{I}\left(||z|| \leq M_2\right)$*

We denote the *support* of the measure induced by p with Q. In order to control the bias proxy (12), one needs a regularity assumption on f. We work under the assumption of Hölder continuity. For $0 < a \leq 1$ and $L > 0$, we say that f belongs in the **Hölder class** $\Sigma\left(a, L\right)$ on Q if for all $x, z \in Q$, we have $|f(x) - f(z)| \leq L||x - z||^a$.

Assumption 2 *There exist $0 < a \leq 1$, $L > 0$ and $B > 0$ such that $f \in \Sigma\left(a, L\right)$ on Q and $\sup\limits_{x \in Q} |f(x)| \leq B$*

Lemma 4.1 *Suppose Assumptions 1 and 2 hold.*

$$\sup_{x \in Q} |b_n(f,x) - f(x)| \leq LM_2^a h_n^a$$

The classical NW estimator (4) is known to perform poorly near the boundary of the support and in regions where the density function p is low. The following definition describes sets for which the boundary issue can be mitigated. We say that $G \subseteq \mathbb{R}^d$ has (r_0, c_0)-**measure-retaining property** if for all $x \in G$ and all $r \leq r_0$, $m\left(G \cap B_r(x)\right) \geq c_0 m\left(B_r(x)\right)$, where m denotes Lebesgue measure on $\mathbb{R}^d$.

Assumption 3 *There exist $r_0, c_0 > 0$ such that Q has the $(r_0, c_0)-$measure-retaining property.*

Lemma 4.2 *Suppose that Assumption 1, 3 hold. If $M_1 h_n < r_0$ and $x \in Q$ is such that*

$$\inf_{\substack{z \in Q \\ |x-z| \leq M_1 h_n}} p(z) \geq p_0(x) > 0 \tag{15}$$

Then

$$\frac{1}{d_n(x)} \leq \frac{C}{n\alpha_n h_n^d p_0(x)}$$

where C depends on c_0, M_1 and d.

Combining Theorem 3.1 and Lemma 4.2, we see that if Assumptions 1 and 3 hold, then $v_n(x) \leq \frac{C_2}{n\alpha_n h_n^d p_0(x)}$. If, in addition Assumption 2 holds then, combining the last bound with Equation (14) and Lemma 4.1, we get

$$\mathcal{R}\left(\hat{f}_{GNW}(x), f(x)\right) \leq C_1 h_n^{2\alpha} + \frac{C_2}{n\alpha_n h_n^d p_0(x)} \tag{16}$$

where C_1, C_2 depend on the various parameters appearing in the assumptions, and are explicitly given in the extended version of this paper [4].

4.2 Integrated risk of GNW

Finally, to bound the integrated risk (11) using Equation (16), one needs to deal with the quantity $p_0(x)$ (15) appearing in the expression. The simplest way to do this is to assume that for all $x \in Q$, we have $p(x) \geq p_0$. In that case one can take $p_0(x) \equiv p_0$ in Equation (16), giving

$$\mathcal{R}\left(\hat{f}_{GNW}, f\right) \leq C_1 h_n^{2\alpha} + \frac{C_2}{n\alpha_n h_n^d} \tag{17}$$

This rate is classical for the NW estimator (4) under uniform density assumption [5, 15]. One can get a slightly weaker rate for noncompactly supported densities p.

Assumption 4 *There exist $0 < \beta \leq 1$ and $L > 0$ such that $p \in \Sigma(\beta, L)$ and $\int p^{1/2}(x)dx < \infty$*

Theorem 4.3 *Suppose that Assumptions 1, 2, 3, 4 hold. If $M_1 h_n < r_0$,*

$$\mathcal{R}\left(\hat{f}_{GNW}, f\right) \leq \frac{C_1}{n\alpha_n h_n^{d+\beta}} + C_2 h_n^{\min\{2\alpha, \beta/2\}}$$

where C_1, C_2 depend on $B, \sigma^2, c_0, d, \alpha, \beta, M_1, M_2$ and L.

Typically one optimizes the rates in Equation (17) or Theorem 4.3. However, as the user has no control of h_n, we provide a range of values of h_n depending on $n\alpha_n$ such that for Theorem 4.3 the integrated risk (11) is **uniformly** controlled over the class of regression functions *and* densities appearing in the assumptions of Theorem 4.3. Set $\gamma = \min(2a, \beta)$. There exist values $C_3, C_4 > 0$ depending on the various parameters appearing in the assumptions (explicitly given in [4]) such that if $r \leq \frac{\gamma}{d+\beta+\gamma}$, $h_n < \min(\frac{r_0}{M_1}, 1)$ and

$$(n\alpha_n)^{-\frac{1-r}{d+\beta}} \leq C_3 h_n \leq (n\alpha_n)^{-\frac{r}{\gamma}} \tag{18}$$

then

$$\mathcal{R}\left(\hat{f}_{GNW}, f\right) \leq \frac{C_4}{(n\alpha_n)^r} \tag{19}$$

Note that for $r = \frac{\gamma}{d+\beta+\gamma}$ the bound in (19) is the strongest and the interval in (18) shrinks to a point.

5 Conclusion and perspectives

We showed that both the *pointwise* and *integrated* risk bounds of the risk of $\hat{f}_{GNW}$ are similar to ones of the classical NW estimator. If α_n and h_n fall into the suitable range of values (i.e. $h_n \to 0$ and $n\alpha_n h_n^d \to \infty$) then GNW will perform well. As GNW uses only one-hop neighbourhood information, it does not take advantage of the global graph structure, it would be interesting to compare it with *graph* spectral based regression estimators (such as graphical Kernel Ridge Regression). Furthermore, it would be intriguing to investigate whether utilizing node embeddings could offer statistical benefits for estimation.

Acknowledgement

This work was partly supported by ANR GRandMa (ANR-21- CE23-0006); MIAI@Grenoble Alpes (ANR-19-P3IA-0003) and the Ampère Scholarship of Excellence of ENS de Lyon.

References

[1] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, 01 2002.

[2] Peter J. Bickel, Aiyou Chen, and Elizaveta Levina. The method of moments and degree distributions for network models. *The Annals of Statistics*, 39(5), 10 2011.

[3] Luc P. Devroye. The uniform convergence of the nadaraya-watson regression function estimate. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 6(2):179–191, 1978.

[4] M. Gjorgjevski. The graphical nadaraya watson estimator on latent position models, 2023.

[5] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression.* Springer series in statistics. Springer, 2002.

[6] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[7] Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002.

[8] Paul Holland, Kathryn B. Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5:109–137, 1983.

[9] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[10] Can M. Le, Elizaveta Levina, and Roman Vershynin. Sparse random graphs: regularization and concentration of the laplacian. *CoRR*, abs/1502.03049, 2015.

[11] Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1), feb 2015.

[12] Roberto Imbuzeiro Oliveira. Concentration of the adjacency matrix and of the laplacian in random graphs with independent edges. *arXiv: Combinatorics*, 2009.

[13] M. Penrose. *Random Geometric Graphs*. Oxford studies in probability. Oxford University Press, 2003.

[14] Minh Tang, Daniel L. Sussman, and Carey E. Priebe. Universally consistent vertex classification for latent positions graphs. *The Annals of Statistics*, 41(3), 06 2013.

[15] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008.

[16] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

[17] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.