Electronic Journal of Statistics
Vol. 11 (2017) 891–915
ISSN: 1935-7524
DOI: 10.1214/17-EJS1253

Estimating a smooth function on a large graph by Bayesian Laplacian regularisation

Alisa Kirichenko and Harry van Zanten[*]

Korteweg-de Vries Institute for Mathematics, Science Park 107, 1098 XG Amsterdam, The Netherlands,
e-mail: a.kirichenko@uva.nl; hvzanten@uva.nl

Abstract: We study a Bayesian approach to estimating a smooth function in the context of regression or classification problems on large graphs. We derive theoretical results that show how asymptotically optimal Bayesian regularisation can be achieved under an asymptotic shape assumption on the underlying graph and a smoothness condition on the target function, both formulated in terms of the graph Laplacian. The priors we study are randomly scaled Gaussians with precision operators involving the Laplacian of the graph.

MSC 2010 subject classifications: Primary 62G20, 62C10.
Keywords and phrases: Function estimation on graphs, Laplacian regularisation, nonparametric Bayes.

Received December 2016.

1. Introduction

1.1. Learning a smooth function on a large graph

There are various problems arising in modern statistics that involve making inference about a "smooth" function on a large graph. The underlying graph structure in such problems can have different origins. Sometimes it is given by the context of the problem. This is typically the case, for instance, in the problem of making inference on protein interaction networks (e.g. Sharan et al. (2007)) or in image interpolation problems (Liu et al. (2014)). In other cases the graph is deduced from the data in a preliminary step, as is the case with similarity graphs in label propagation methods (e.g. Zhu and Ghahramani (2002)). Moreover, the different problems that arise in applications can have all kinds of different particular features. For example, the available data can be indexed by the vertices or by the edges of the graph, or both. Also, in some applications only partial data are available, for instance only part of the vertices are labeled (semi-supervised problems). Moreover, both regression problems and classification problems arise naturally in different applications.

[*] Research funded by the Netherlands Organization for Scientific Research (NWO).

Despite all these different aspects, many of these problems and the methods that have been developed to deal with them have a number of important features in common. In many cases the graph is relatively "large" and the function of interest can be viewed as "smoothly varying" over the graph. Consequently, most of the proposed methods view the problem as a high-dimensional or nonparametric estimation problem and employ some regularisation or penalization technique that takes the geometry of the graph into account and that is thought to produce an appropriate bias-variance trade-off.

In this paper we set up the mathematical framework that allows us to study the performance of nonparametric function estimation methods on large graphs. We do not treat all the variants exhaustively, instead we consider two prototypical problems: regression, where the function of interest f is a function on the vertices of the graph that is observed with additive noise, and binary classification, where a label 0 or 1 is observed at each vertex and the object of interest is the "soft label" function f whose value at a vertex v is the probability of seeing a 1 at v. We assume the underlying graph is "large", in the sense that it has n vertices for some "large" n. Our theoretical results deal with the situation that this number n tends to infinity. Although for finite n the graph has a fixed size and we essentially just have to estimate a Euclidean vector in $\mathbb{R}^n$, it is useful to view the problem as high-dimensional or even nonparametric.

Despite the finite structure, it is intuitively clear that the "smoothness" of f, defined in a suitable manner, will have an impact on the difficulty of the problem and on the results that can be attained. Indeed, consider the extreme case of f being a constant function. Then estimating f reduces to estimating a single real number. In the regression setting, for instance, this means that under mild conditions the sample mean gives a $\sqrt{n}$-consistent estimator. In the other extreme case of a completely unrestricted function there is no way of making any useful inference. At best we can say that in view of the James-Stein effect we should employ some degree of shrinking or regularisation. However, if no further assumptions are made, nothing can be said about consistency or rates. We are interested in the question what we should do in the intermediate situation that f has some "smoothness" between these two extremes.

Another aspect that will have a crucial impact on the problem, in addition to the regularity of f, is the geometry of the graph. Indeed, regular grids of different dimensions are special cases of the graphs we shall consider, and we know from existing theory that the best attainable rates for estimating a smooth function on a grid depends on the dimension of the grid. More generally, the geometry of the graph will influence the complexity of the spaces of "smooth" functions on the graph, and hence the performance of statistical or learning methods.

1.2. Laplacian regularisation

Several approaches to learning functions on graphs that have been explored in the literature involve regularisation using the Laplacian matrix associated with the graph (see, for example, Belkin et al. (2004), Smola and Kondor (2003), Hein (2006), Ando and Zhang (2007), Zhu et al. (2003), Huang et al. (2011)).

The graph Laplacian is defined as $L = D - A$, where A is the adjacency matrix of the graph and D is the diagonal matrix with the degrees of the vertices on the diagonal. When viewed as a linear operator, the Laplacian acts on a function f on the graph as

$$Lf(i) = \sum_{j \sim i} \Big(f(i) - f(j) \Big), \tag{1.1}$$

where we write $i \sim j$ if vertices i and j are connected by an edge. Several related operators are routinely employed as well, for instance, the normalized Laplacian $\tilde{L} = D^{-1/2} L D^{-1/2}$. We will continue to work with L in this paper, but much of the story goes through if L is replaced by such a related operator, after minor adaptations.

For a function f on the graph the Laplacian norm is given by $\sum_{j \sim i}(f(i) - f(j))^2$. Clearly, the Laplacian norm of f quantifies how much the function f varies when moving along the edges of the graph. Therefore, several papers have proposed regularisation or penalization using this norm, as well as generalizations involving powers of the Laplacian or other functions, for instance, exponential ones. See, for example, Belkin et al. (2004) or Smola and Kondor (2003) and the references therein. There exist only few papers that study theoretical aspects of the performance of such methods. We mention, for example, Belkin et al. (2004), in which a theoretical analysis of a Tikhonov regularisation method is conducted in terms of algorithmic stability. Johnson and Zhang (2007) consider sub-sampling schemes for estimating a function on a graph.

The existing papers have different viewpoints than ours and do not study how the performance depends on (the combination of) graph geometry and function regularity. Our aim is to develop a framework which makes such a theoretical study of Laplacian regularisation methods possible and to derive some first asymptotic results that exhibit methods that perform well from the point of view of convergence rates and adaptation to regularity.

1.3. Bayesian regularisation

We investigate Bayesian regularisation approaches, where we consider two types of priors on functions on graphs. The first type performs regularisation using a power of the Laplacian. This can be seen as the graph analogue of Sobolev norm regularisation of functions on "ordinary" Euclidean spaces. The second type of priors uses an exponential function of the Laplacian. This can be viewed as the analogue of the popular squared exponential prior on functions on Euclidean space or its extension to manifolds, as studied by Castillo et al. (2014). In both cases we consider hierarchical priors with the aim of achieving automatic adaptation to the regularity of the function of interest.

To assess the performance of our Bayes procedures we take an asymptotic perspective. We let the number of vertices of the graph grow and ask at what rate the posterior distribution concentrates around the unknown function f that generates the data. We make two kinds of assumptions. Firstly, we assume that f has a certain degree of regularity β, defined in suitable manner. The smoothness β is not assumed to be known though, we are aiming at deriving adaptive results.

Secondly, we make an assumption on the asymptotic shape of the graph. In recent years, various theories of graph limits have been developed. Most prominent is the concept of the graphon, e.g. Lovász and Szegedy (2006) or the book of Lovasz (2012). More recently this notion has been extended in various directions, see, for instance, Borgs et al. (2014) and Chung (2014). However, the existing approaches are not immediately suited in the situations we have in mind, which involve graphs that are sparse in nature and are "grid-like" in some sense. Therefore we take an alternative approach and describe the asymptotic shape of the graph through a condition on the asymptotic behaviour of the spectrum of the Laplacian. To be able to derive concrete results we essentially assume that the smallest eigenvalues $\lambda_{n,i}$ of L satisfy

$$\lambda_{n,i}^2 \asymp \left(\frac{i}{n}\right)^{2/r} \tag{1.2}$$

for some $r \geq 1$[1]. Very roughly speaking, this means that asymptotically, or "from a distance", the graph looks like an r-dimensional grid with n vertices. As we shall see, the actual grids are special cases (see Example 2.1), hence our results include the usual statements for regression and classification on these classical design spaces. However, the setting is much more general, since it is really only the *asymptotic* shape that matters. For instance, a 2 by $n/2$ ladder graph asymptotically also looks like a path graph, and indeed we will see that it satisfies our assumption for $r = 1$ as well (Example 2.3). Moreover, the constant r in (1.2) does not need to be a natural number. We will see, for example, at least numerically, that there are graphs whose geometry is asymptotically like that of a grid of non-integer "dimension" r in the sense of condition (1.2).

We stress that we do not assume the existence of a "limiting manifold" for the graph as $n \to \infty$. We formulate our conditions and results purely in terms of intrinsic properties of the graph, without first embedding it in an ambient space. In certain cases in which limiting manifolds do exist (e.g. the regular grid cases) our type of asymptotics can be seen as "infill asymptotics" (Cressie (1993)). For a simple illustration, see Example 3.1. However, in applied settings (see, for instance, Example 2.7) it is typically not clear what a suitable ambient manifold could be, which is why we choose to avoid this issue altogether.

In the recent paper Hartog and van Zanten (2016) the theoretical results we present in this paper are investigated numerically and serve as a guideline for the tuning of practical Bayesian regularisation methods. Several concrete examples are considered, both for simulated data and for real data problems.

1.4. Organisation

The rest of the paper is organized as follows. In the next section we present our geometry assumption and give examples of graphs that satisfy it, either

[1] We write $a_n \asymp b_n$ if $0 < \liminf a_n/b_n \leq \limsup a_n/b_n < \infty$.

theoretically or numerically. In Section 3 we introduce two families of priors on functions on graphs. We present theorems that quantify the amount of mass that the priors put on neighbourhoods of "smooth" functions and quantify the complexity of the priors in terms of metric entropy. Section 4 contains the proofs of these general results and in Section 5 they are used to derive theorems about posterior contraction in nonparametric regression and binary classification. We end with some concluding remarks in Section 6.

2. Asymptotic geometry assumption on graphs

In this section we formulate our geometry assumption on the underlying graph and give several examples.

2.1. Graphs, Laplacians and functions on graphs

Let G be a connected, simple (i.e. no loops, multiple edges or weights), undirected graph with n vertices labelled $1,\dots,n$. Let A be its adjacency matrix, i.e. A_{ij} is 1 or 0 according to whether or not there is an edge between vertices i and j. Let D be the diagonal matrix with element D_{ii} equal to the degree of vertex i. Let $L = D - A$ be the Laplacian of the graph. We note that strictly speaking, we will be considering sequences of graphs G_n with Laplacians L_n and we will let n tend to infinity. However, in order to avoid cluttered notation, we will omit the subscript n and just write G and L throughout.

A function f on the (vertices of the) graph is simply a function $f : \{1,\dots,n\} \to \mathbb{R}$. Slightly abusing notation we will write f both for the function and for the associated vector of function values $(f(1), f(2),\dots, f(n))$ in $\mathbb{R}^n$. We measure distances and norms of functions using the norm $\|\cdot\|_n$ defined by $\|f\|_n^2 = n^{-1}\sum_{i=1}^n f^2(i)$. The corresponding inner product of two functions f and g is denoted by

$$\langle f, g\rangle_n = \frac{1}{n}\sum_{i=1}^{n} f(i)g(i).$$

Again, in our results n will be varying, so when we speak of a function f on the graph G we are, strictly speaking, considering a sequence of functions f_n. Also, in this case the subscript n will usually be omitted.

The Laplacian L is positive semi-definite and symmetric. It easily follows from the definition that its smallest eigenvalue is 0 (with eigenvector $(1,\dots,1)$). The fact that G is connected implies that the second smallest eigenvalue, the so-called algebraic connectivity, is strictly positive (e.g. Cvetković et al. (2010)). We will denote the Laplacian eigenvalues, ordered my magnitude, by

$$0 = \lambda_{n,0} < \lambda_{n,1} \le \lambda_{n,2} \le \cdots \le \lambda_{n,n-1}.$$

Again we will usually drop the first index n and just write λ_i for $\lambda_{n,i}$. We fix a corresponding sequence of eigenfunctions ψ_i, orthonormal with respect to the inner product $\langle\cdot,\cdot\rangle_n$.

2.2. Asymptotic geometry assumption

As mentioned in the introduction, we will derive results under an asymptotic shape assumption on the graph, formulated in terms of the Laplacian eigenvalues. To motivate the definition we note that the ith eigenvalue of the Laplacian of an n-point grid of dimension d behaves like $(i/n)^{2/d}$ (see Example 2.1 ahead). We will work with the following condition.

Condition. *We say that the* geometry condition *is satisfied with parameter* $r \geq 1$ *if there exist* $i_0 \in \mathbb{N}$, $\kappa \in (0, 1]$ *and* $C_1, C_2 > 0$ *such that for all* n *large enough,*

$$C_1\Big(\frac{i}{n}\Big)^{2/r} \leq \lambda_i \leq C_2\Big(\frac{i}{n}\Big)^{2/r}, \qquad \textit{for all } i \in \{i_0, \ldots, \kappa n\}.$$

Note that this condition only restricts a positive fraction κ of the Laplacian eigenvalues, namely the κn smallest ones. Moreover, we don't need restrictions on the first finitely many eigenvalues. We remark that if the geometry condition is fulfilled, then by adapting the constant C_1 we can ensure that the lower bound holds, in fact, for *all* $i \in \{i_0, \ldots, n\}$. To see this, observe that for n large enough and $\kappa n < i \leq n$ we have

$$\lambda_i \geq \lambda_{\lfloor \kappa n \rfloor} \geq C_1\Big(\frac{\lfloor \kappa n \rfloor}{n}\Big)^{2/r} \geq C_1\Big(\frac{\kappa}{2}\Big)^{2/r}\Big(\frac{i}{n}\Big)^{2/r}.$$

For the indices $i < i_0$ it is useful to note that we have a general lower bound on the first positive eigenvalue λ_1, hence on $\lambda_2, \ldots, \lambda_{i_0}$ as well. Indeed, by Theorem 4.2 of Mohar (1991a) we have

$$\lambda_1 \geq \frac{4}{n \operatorname{diam}(G)} \geq \frac{4}{n^2}. \tag{2.1}$$

Note that this bound also implies that our geometry assumption can not hold with a parameter $r < 1$, since that would lead to contradictory inequalities for λ_{i_0}.

We first confirm that the geometry condition is satisfied for grids and tori of different dimensions.

Example 2.1 (Grids)**.** For $d \in \mathbb{N}$, a regular d-dimensional grid with n vertices can be obtained by taking the Cartesian product of d path graphs with $n^{1/d}$ vertices (provided, of course, that this number is an integer). Using the known expression for the Laplacian eigenvalues of the path graph and the fact that the eigenvalues of products of graphs are the sums of the original eigenvalues, see, for instance, Theorem 3.5 of Mohar (1991b), we get that the Laplacian eigenvalues of the d-dimensional grid are given by

$$4\left(\sin^2\frac{\pi i_1}{2n^{\frac{1}{d}}} + \cdots + \sin^2\frac{\pi i_d}{2n^{\frac{1}{d}}}\right) \asymp \frac{i_1^2 + \cdots + i_d^2}{n^{2/d}},$$

where $i_k = 0, 1, 2, \ldots, n^{1/d} - 1$ for every $k = 1, \ldots, d$. By definition there are $i + 1$ eigenvalues less or equal than the ith smallest eigenvalue λ_i. Hence, for a constant $c > 0$, we have:

$$i + 1 = \sum_{i_1^2 + \cdots + i_d^2 \leq c^2 n^{2/d} \lambda_i} 1.$$

The sum on the right gives the number of lattice points in a sphere of radius $R = c n^{1/d} \sqrt{\lambda_i}$ in $\mathbb{R}^d$. For our purposes it suffices to use crude upper and lower bounds for this number. By considering, for instance, the smallest hypercube containing the sphere and the largest one inscribed in it, it is easily seen that the number of lattice points is bounded from above and below by a constant times R^d. We conclude that for the d-dimensional grid we have $\lambda_i \asymp (i/n)^{2/d}$ for every $i = 0, \ldots, n - 1$. In particular, the geometry condition is fulfilled with parameter $r = d$.

Example 2.2 (Discrete tori)**.** For graph tori we can follow the same line of reasoning as for grids. A d-dimensional torus graph with n vertices can be obtained as a product of d ring graphs with $n^{1/d}$ vertices. Using the known explicit expression of the Laplacian eigenvalues of the ring we find that the d-dimensional torus graph satisfies the geometry conditions with parameter $r = d$ as well.

The following lemma lists a number of operations that can be carried out on a graph without loosing the geometry condition.

Lemma 2.1. *Suppose that $G = G_n$ satisfies the geometry assumption with parameter r. Then the following graphs satisfy the condition with parameter r as well:*

(i) The cartesian product of G with a connected simple graph H with a finite number of vertices (independent of n).
(ii) The graph obtained by augmenting G with finitely many edges (independent of n), provided it is a simple graph.
(iii) The graph obtained from G by deleting finitely many edges (independent of n), provided it is still connected.
(iv) The graph obtained by augmenting G with finitely many vertices and edges (independent of n), provided it is a simple connected graph.

Proof. (i). Say H has m vertices and let its Laplacian eigenvalues be denoted by $0 = \mu_0, \ldots, \mu_m$. Then the product graph has mn vertices and it has Laplacian eigenvalues $\lambda_i + \mu_j$, $i = 0, \ldots, n-1, j = 0, \ldots, m-1$ (see Theorem 3.5 of Mohar (1991b)). In particular, the first n eigenvalues are the same as those of G. Hence, since G satisfies the geometry condition, so does the product of G and H.

(ii) and (iii). These statements follow from the interlacing formula that asserts that if $G + e$ is the graph obtained by adding the edge e to G, then

$$0 \leq \lambda_1(G) \leq \lambda_1(G + e) \leq \lambda_2(G) \leq \cdots \leq \lambda_{n-1}(G) \leq \lambda_{n-1}(G + e).$$

See, for example, Theorem 3.2 of Mohar (1991b) or Theorem 7.1.5 of Cvetković et al. (2010).

(iv). Let v and e be a vertex and an edge that we want to connect to G. Denote G_v a disjoint union of G and v, and by G' the graph obtained by connecting edge e to v and an existing vertex of G. By Theorem 3.1 from Mohar (1991b) we know that the eigenvalues of G_v are $0, 0, \lambda_1(G), \lambda_2(G), \ldots, \lambda_{n-1}(G)$. Using Theorem 3.2 of Mohar (1991b) we see that $0 = \lambda_0(G_v) = \lambda_0(G')$ and

$$0 = \lambda_1(G_v) \le \lambda_1(G') \le \lambda_1(G) \le \lambda_2(G_v) \le \cdots \le \lambda_{n-1}(G) \le \lambda_n(G').$$

The result follows from this observation. □

Example 2.3 (Ladder graph)**.** A ladder graph with n vertices is the product of a path graph with $n/2$ vertices and a path graph with 2 vertices. Hence, by part (i) of Lemma 2.1 and Example 2.1 it satisfies the geometry condition with parameter $r = 1$.

Example 2.4 (Lollipop graph)**.** The so-called lollipop graph $L_{m,n}$ is obtained by attaching a path graph with n vertices with an additional edge to a complete graph with m vertices. If m is constant, i.e. independent of n, then according to parts (ii) and (iv) of the preceding lemma this graph satisfies the geometry condition with $r = 1$.

In the examples considered so far it is possible to verify theoretically that the geometry condition is fulfilled. In a concrete case in which the given graph is not of such a tractable type, numerical investigation of the Laplacian eigenvalues can give an indication as to whether or not the condition is reasonable and provide the appropriate value of the parameter r. A possible approach is to plot $\log \lambda_i$ against $\log(i/n)$. If the geometry condition is satisfied with parameter r, the $\kappa \times 100\%$ left mosts points in this plot should approximately lie on a straight line with slope $2/r$, except possibly a few on the very left.

Our focus in this paper is not on numerics, but it is illustrative to consider a few numerical examples in order to get a better idea of the types of graphs that fit into our framework.

Example 2.5 (Two-dimensional grid, numerically)**.** Figure 1 illustrates the suggested numerical approach for a two-dimensional, 20×20 grid. The dashed line in the left panel is fitted to the left-most 35% of the points in the plot, discarding the first three points on the left. In accordance with Example 2.1 this line has slope 1.0.

Example 2.6 (Watts-Strogatz ‘small world’ graph)**.** In our second numerical example we consider a graph obtained as a realization from the well-known random graph model of Watts and Strogatz (1998). Specifically, we consider in the first step a ring graph with 200 vertices. In the next step every vertex is visited and the edges emanating from the vertex are rewired with probability $p = 1/4$, meaning that with probability 1/4 they are detached from the neighbour of the current vertex and attached to another vertex, chosen uniformly at random. In the right panel of Figure 2 a particular realization is shown. Here we have only kept the largest connected component, which has 175 vertices in this case.

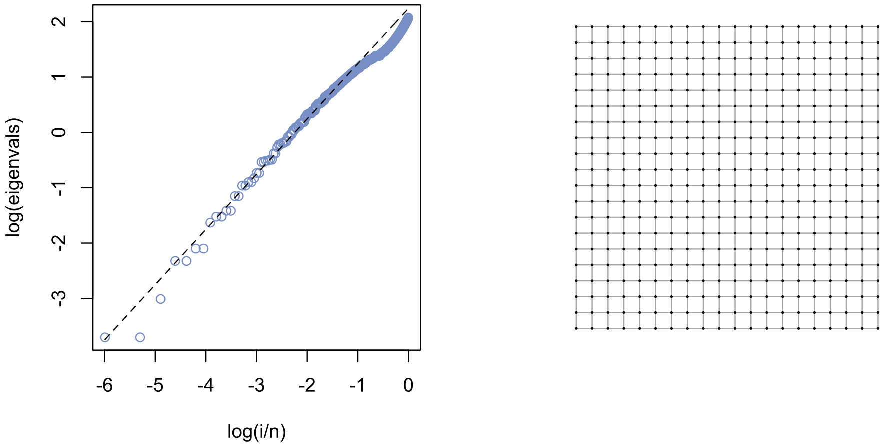

FIG 1. *Plot of* $\log \lambda_i$ *against* $\log(i/n)$ *for the* 20×20 *grid. Fitted line has slope* 1.0*, corresponding to* $r = 2.0$ *in the geometry assumption.*

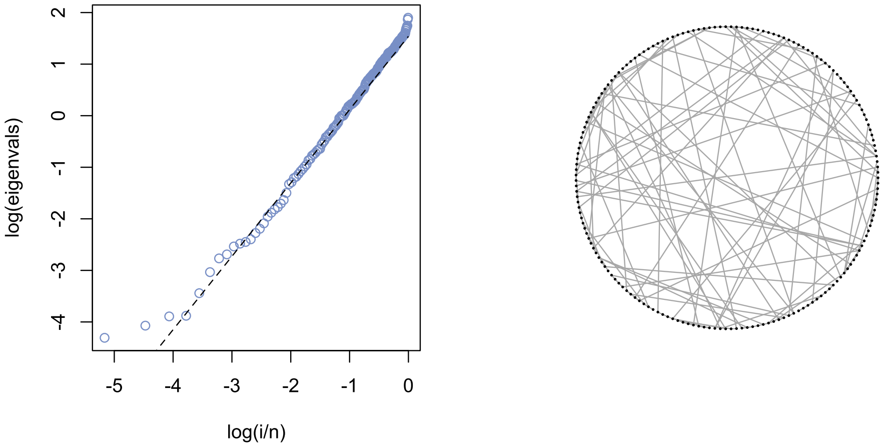

FIG 2. *Plot of* $\log \lambda_i$ *against* $\log(i/n)$ *for the Watts-Strogatz graph in the right panel. Fitted line has slope* 1.42*, corresponding to* $r = 1.4$ *in the geometry assumption.*

On the left we have exactly the same plot as described in the preceding example for the grid case. The plot indicates that it is not unreasonable to assume that the geometry condition holds. The value of the parameter r deduced from the slope of the line equals 1.4 for this graph.

Example 2.7 (Protein interaction graph)**.** In the final example we consider a graph obtained from the protein interaction graph of baker's yeast, as described in detail in Section 8.5 of Kolaczyk (2009). The graph, shown in the right panel of Figure 3, describes the interactions between proteins involved in the commu-

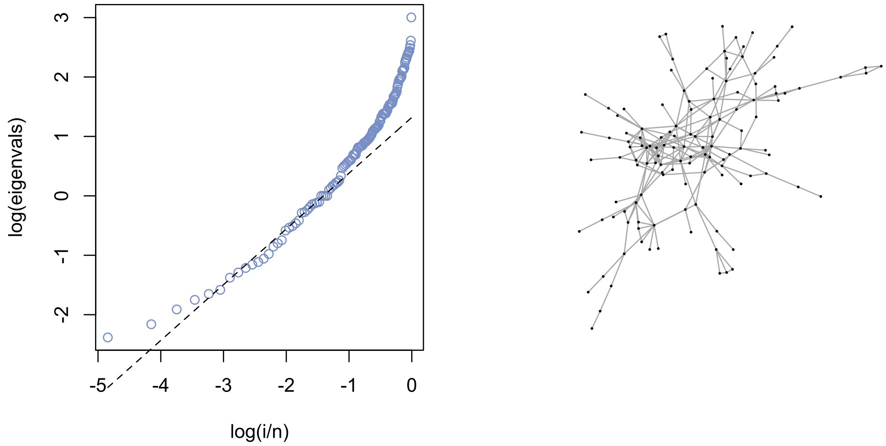

Fig 3. *Plot of* $\log \lambda_i$ *against* $\log(i/n)$ *for the Protein interaction graph in the right panel. Fitted line has slope* 0.94*, corresponding to* $r = 2.1$ *in the geometry assumption.*

nication between a cell and its surroundings. Also for this graph it is true that with a few exceptions, the points corresponding to the 35% smallest eigenvalues lie approximately on a straight line. The same procedure as followed in the other examples gives a value $r = 2.1$ for the parameter in the geometry assumption.

3. General results on prior concentration

We consider two different priors on functions on graphs. The first corresponds to regularisation using a power of the Laplacian, the second one uses an exponential function of the Laplacian. In this section we present two general results which quantify both the mass that these priors give to shrinking $\|\cdot\|_n$-neighbourhoods of a fixed function f_0, and the complexity of the support of the priors, measured in terms of metric entropy[2]. In the next section we will combine these results with known results from Bayesian nonparametrics theory to deduce convergence rates and adaptation for nonparametric regression and classification problems on graphs.

Our results assume that the geometry condition holds for some $r \geq 1$. The mass a prior puts near f_0 will depend on the "regularity" of the function, defined in a suitable manner. Specifically, we will assume it belongs to a Sobolev-type ball of the form

$$H^\beta(C) = \left\{ f : \left\langle f, (I + (n^{\frac{2}{r}} L)^\beta) f \right\rangle_n \leq C^2 \right\} \tag{3.1}$$

for some $\beta, C > 0$ (independent of n). The particular normalisation, which depends on the geometry parameter r, ensures non-trivial asymptotics. This is

[2] For $\varepsilon > 0$ and a norm $\|\cdot\|$ on a set B, we denote by $N(\varepsilon, B, \|\cdot\|)$ the minimal number of balls of $\|\cdot\|$-radius ε needed to cover B.

confirmed in Kirichenko and van Zanten (2017), in which minimax lower bounds are presented which complement the rate results of the present paper.

It is illustrative to consider the assumption in a bit more detail in the simple case of the path graph. The example shows in particular that we have chosen the "correct" normalisation in the definition of the smoothness class.

Example 3.1 (Path graph)**.** Consider a path graph G with n vertices, which we identify with the points i/n in the unit interval, $i = 1, \dots, n$. As seen in Example 2.1, this graph satisfies the geometry condition with parameter $r = 1$. Hence, in this case the collection of functions $H^\beta(C)$ is given by

$$H^\beta(C) = \Big\{ f : \big\langle f, (I + (n^2 L)^\beta) f \big\rangle_n \leq C^2 \Big\}.$$

To understand when a (sequence of) function(s) belongs to this space, say for $\beta = 1$, let f_n be the restriction to the grid $\{i/n, i = 1, \dots n\}$ of a fixed function f defined on the whole interval $[0, 1]$. The assumption that $f_n \in H^1(C)$ then translates to the requirement that

$$\frac{1}{n} \sum_i f^2(i/n) + n \sum_{i \sim j} (f(i/n) - f(j/n))^2 \leq C^2.$$

The first term on the left is a Riemann sum which approximates the integral $\int_0^1 f^2(x)\,dx$. If f is differentiable, then for the second term we have, for large n,

$$n \sum_{i \sim j} (f(i/n) - f(j/n))^2 = n \sum_{i=1}^{n-1} (f((i+1)/n) - f(i/n))^2 \approx \frac{1}{n} \sum_i (f'(i/n))^2,$$

which is a Riemann sum that approximates the integral $\int_0^1 (f'(x))^2\,dx$. Hence in this particular case the space of functions $H^1(C)$ on the graph is the natural discrete version of the usual Sobolev ball

$$\Big\{ f : [0, 1] \to \mathbb{R} : \int_0^1 (f^2(x) + f'^2(x))(x)\,dx \leq C^2 \Big\}.$$

Definition (3.1) is a way of describing "β-regular" functions on a general graph satisfying the geometry condition, without assuming the graph or the function on it are discretised versions of some "continuous limit".

The first family of priors we consider penalize the higher order Laplacian norm of the function of interest. This corresponds to using a Gaussian prior with a power of the Laplacian as precision matrix (inverse covariance). (We note that since the Laplacian always has 0 as an eigenvalue, it is not invertible. We remedy this by adding a small multiple of the identity matrix I to L.) The larger the power of the Laplacian used, the more "rough" functions on the graph are penalized. The power is regulated by a hyperparameter $\alpha > 0$ which can be seen as describing the "baseline regularity" of the prior. To enlarge the range of regularities for which we obtain good contraction rates in the statistical results, we add a multiplicative hyperparameter which we endow with a suitable hyperprior. In (3.2) we assume an exact standard exponential distribution, but inspection of the proof shows that the range of priors for which the result holds is actually larger. To keep the exposition clean we omit these details.

Theorem 3.2 (Power of the Laplacian)**.** *Suppose the geometry assumption holds for $r \geq 1$. Let $\alpha > 0$ be fixed and assume that $f_0 \in H^\beta(C)$ for some $C > 0$ and $0 < \beta \leq \alpha + r/2$. Let the random function f on the graph be defined by*

$$c \sim \mathit{Exp}(1) \tag{3.2}$$

$$f \,|\, c \sim N(0, (((n/c)^{2/r}(L + n^{-2}I))^{\alpha + r/2})^{-1}). \tag{3.3}$$

Then there exists a constant $K_1 > 0$ and for all $K_2 > 1$ there exist Borel measurable subsets B_n of $\mathbb{R}^n$ such that for every sufficiently large n,

$$P(\|f - f_0\|_n < \varepsilon_n) \geq e^{-K_1 n \varepsilon_n^2}, \tag{3.4}$$

$$P(f \notin B_n) \leq e^{-K_2 n \varepsilon_n^2}, \tag{3.5}$$

$$\log N(\varepsilon_n, B_n, \|\cdot\|_n) \leq n\varepsilon_n^2, \tag{3.6}$$

where ε_n is a multiple of $n^{-\beta/(2\beta+r)}$.

Note that in this theorem we obtain the rate $n^{-\beta/(2\beta+r)}$ for all β in the range $(0, \alpha + r/2]$. In the statistical results presented in Section 5 this translates into rate-adaptivity up to regularity level $\alpha + r/2$. So by putting a prior on the multiplicative scale we achieve a degree of adaptation, but only up to an upper bound that is limited by our choice of the hyperparameter α. A possible solution is to consider other functions of the Laplacian instead of using a power of L in the prior specification. Here we consider usage of an exponential function of the Laplacian. We include a "lengthscale" or "bandwidth" hyperparameter that we endow with a prior as well for added flexibility. This prior can be seen as the analogue of the prior used in Castillo et al. (2014) in the context of function estimation on manifolds, which in turn is a generalization of the popular squared exponential Gaussian prior used for estimation functions on Euclidean domains (e.g. Rasmussen and Williams (2006)). However, we stress again that we do not rely on an embedding of our graph in a manifold or the existence of a "limiting manifold".

In the next theorem there is indeed no restriction on the range of the smoothness β. We remark however that we obtain an additional logarithmic factor is the rate. Technically this is a consequence of the larger "complexity" of the support of this prior.

Theorem 3.3 (Exponential of the Laplacian)**.** *Suppose the geometry assumption holds for $r \geq 1$. Assume that $f_0 \in H^\beta(C)$ for some $C > 0$ and $\beta > 0$. Let the random function f on the graph be defined by*

$$c \sim \mathit{Exp}(1) \tag{3.7}$$

$$f \mid c \sim N(0, ne^{-(n/c)^{2/r}L}). \tag{3.8}$$

Then there exists a constant $K_1 > 0$ and for all $K_2 > 1$ there exist Borel subsets B_n of $\mathbb{R}^n$ such that for every sufficiently large n,

$$P(\|f - f_0\|_n < \varepsilon_n) \geq e^{-K_1 n\varepsilon_n^2}, \tag{3.9}$$

$$P(f \notin B_n) \leq e^{-K_2 n\varepsilon_n^2}, \tag{3.10}$$

$$\log N(\tilde{\varepsilon}_n, B_n, \|\cdot\|_n) \leq n\tilde{\varepsilon}_n^2, \tag{3.11}$$

where $\varepsilon_n = (n/\log^{1+r/2} n)^{-\beta/(2\beta+r)}$ and $\tilde{\varepsilon}_n = \varepsilon_n \log^{1/2+r/4} n$.

4. Proofs of Theorems 3.2 and 3.3

Recall that we identify functions on the graph with vectors in $\mathbb{R}^n$. In both cases we have that given c, the random vector f is a centered n-dimensional Gaussian random vector. We view this as a Gaussian random element in the space $(\mathbb{R}^n, \|\cdot\|_n)$. The corresponding RKHS $\mathbb{H}^c$ is the entire space $\mathbb{R}^n$, and the corresponding RKHS-norm is given by

$$\|h\|_{\mathbb{H}^c}^2 = h^T \Sigma_c^{-1} h,$$

where Σ_c is the covariance matrix of $f \mid c$. (See e.g. van der Vaart and van Zanten (2008b) for the definition and properties of the RKHS.) Recall that the ψ_i are the eigenfunctions of L, normalised with respect to the norm $\|\cdot\|_n$. They are then also eigenfunctions of Σ_c^{-1} in both cases. We denote the corresponding eigenvalues by μ_i.

The Gaussian $N(0, \Sigma_c)$ admits the series representation

$$\sum Z_i \psi_i / \sqrt{n\mu_i}, \tag{4.1}$$

where $Z_1, \ldots, Z_n$ are standard normal variables. In particular the functions $\psi_i/\sqrt{n\mu_i}$ form an orthonormal basis of the RKHS $\mathbb{H}^c$. Hence, the ordinary $\|\cdot\|_n$-norm and the RKHS-norm of a function h with expansion $h = \sum h_i \psi_i$ are given by

$$\|h\|_n^2 = \sum_{i=0}^{n-1} h_i^2, \qquad \|h\|_{\mathbb{H}^c}^2 = n \sum_{i=0}^{n-1} \mu_i h_i^2. \tag{4.2}$$

We denote the unit ball of the RKHS by $\mathbb{H}_1^c = \{h \in \mathbb{H}^c : \|h\|_{\mathbb{H}^c} \leq 1\}$.

4.1. Proof of Theorem 3.2

In this case $\Sigma_c^{-1} = ((n/c)^{2/r}(L + n^{-2}I))^{\alpha + r/2}$ is the precision matrix of f given c and the eigenvalues of Σ_c^{-1} are given by

$$\mu_i = \left(\left(\frac{n}{c}\right)^{2/r}\left(\lambda_i + \frac{1}{n^2}\right)\right)^{\alpha + r/2}.$$

4.1.1. Proof of (3.4)

By Lemma 5.3 of van der Vaart and van Zanten (2008b), it follows from Lemmas 4.1 and 4.2 ahead that under the conditions of the theorem and for $\varepsilon = \varepsilon_n = n^{-\beta/(r+2\beta)}$ and $c = c_n$ satisfying $\sqrt{n}\varepsilon_n^{(\beta-\alpha)/\beta} \leq c_n^{(\alpha+r/2)/r} \leq 2\sqrt{n}\varepsilon_n^{(\beta-\alpha)/\beta}$, we have

$$-\log P(\|f - f_0\|_n \,|\, c) \lesssim \varepsilon_n^{-r/\beta}.$$

By conditioning, it is then seen that

$$P(\|f - f_0\|_n < \varepsilon_n) \geq e^{-K_0\varepsilon_n^{-r/\beta}} \int_{(\sqrt{n}\varepsilon_n^{(\beta-\alpha)/\beta})^{r/(\alpha+r/2)}}^{(2\sqrt{n}\varepsilon_n^{(\beta-\alpha)/\beta})^{r/(\alpha+r/2)}} e^{-x}\,dx \geq e^{-K_1\varepsilon_n^{-r/\beta}},$$

for constants $K_0, K_1 > 0$.

Lemma 4.1. *For n large enough and $\varepsilon > 0$ and $\varepsilon\sqrt{n}/c^{(\alpha+r/2)/r}$ small enough,*

$$-\log P(\|f\|_n \leq \varepsilon \,|\, c) \lesssim \left(\frac{c^{(\alpha+r/2)/r}}{\varepsilon\sqrt{n}}\right)^{\frac{r}{\alpha}}. \tag{4.3}$$

Proof. By the series representation (4.1) we have $P(\|f\|_n \leq \varepsilon \,|\, c) = P(\sum Z_i^2/(n\mu_i) \leq \varepsilon^2)$. Recall from Section 2.2 that we can assume without loss of generality that we have the lower bounds

$$\lambda_i \geq C_1\left(\frac{1}{n}\right)^2, \qquad 1 \leq i \leq i_0, \tag{4.4}$$

$$\lambda_i \geq C_1\left(\frac{i}{n}\right)^{2/r}, \qquad i > i_0. \tag{4.5}$$

These translate into lower bounds for the μ_i from which it follows that for $\varepsilon > 0$,

$$\begin{aligned} P(\|f\|_n^2 \leq 2\varepsilon^2 \,|\, c) &\geq P\Big(\sum_{i\leq i_0} \frac{Z_i^2}{n\mu_i} \leq \varepsilon^2, \sum_{i>i_0} \frac{Z_i^2}{n\mu_i} \leq \varepsilon^2\Big) \\ &\geq P\Big(\sum_{1<i\leq i_0} Z_i^2 \leq (C_1^p c^{-2p/r} n^{(2\alpha+2r-2pr)/r})\varepsilon^2\Big) P\Big(\sum_{i>i_0} \frac{Z_i^2}{i^{2p/r}} \leq C_1^p c^{-2p/r} n\varepsilon^2\Big), \end{aligned}$$

where we write $p = \alpha + r/2$. By Corollary 4.3 from Dunker et al. (1998), the last factor in the last line is bounded form below by

$$\exp\Big(-\operatorname{const} \times (c^{-p/r}\varepsilon\sqrt{n})^{-r/\alpha}\Big),$$

provided $c^{-p/r}\varepsilon\sqrt{n}$ is small enough. By the triangle inequality and independence, the first factor is bounded from below by

$$\Big(P(|Z_1| \leq i_0^{1/2} C_1^{p/2} c^{-p/r} n^{(\alpha+r-pr)/r}\varepsilon)\Big)^{i_0}.$$

Since $r \geq 1$, we have $c^{-p/r}n^{(\alpha+r-pr)/r}\varepsilon = O(c^{-p/r}\varepsilon\sqrt{n})$. Hence, for $c^{-p/r}\varepsilon\sqrt{n}$ small enough the probability is further bounded from below by

$$\operatorname{const} \times \Big(c^{-p/r} n^{(\alpha+r-pr)/r}\varepsilon\Big)^{i_0}.$$

Combining the bounds for the separate factors we find that, for $c^{-p/r}\varepsilon\sqrt{n}$ small enough,

$$-\log P(\|f\|_n^2 \leq 2\varepsilon^2 \,|\, c) \lesssim \log\Big(\frac{c^{p/r}}{n^{(\alpha+r-pr)/r}\varepsilon}\Big) + \Big(\frac{c^{p/r}}{\varepsilon\sqrt{n}}\Big)^{r/\alpha}.$$

Since $r \geq 1$ the first term on the right is smaller than a constant times the second one if $c^{-p/r}\varepsilon\sqrt{n}$ is small enough. □

Lemma 4.2. *Let $f \in H^\beta(C)$ for $\beta \leq \alpha + r/2$. For $\varepsilon > 0$ such that $\varepsilon \to 0$ as $n \to \infty$ and $1/\varepsilon = o(n^{\beta/r})$ and n large enough,*

$$\inf_{h\in\mathbb{H}^c\colon \|h-f\|_n\leq\varepsilon} \|h\|_{\mathbb{H}^c}^2 \lesssim nc^{-(2\alpha+r)/r}\varepsilon^{-\frac{2(\alpha-\beta)+r}{\beta}}. \tag{4.6}$$

Proof. We use an expansion $f = \sum f_i\psi_i$, with ψ_i the orthonormal eigenfunctions of the Laplacian. In terms of the coefficients the smoothness assumption reads $\sum(1+n^{2\beta/r}\lambda_i^\beta)f_i^2 \leq C^2$. Now for I to be determined below, consider $h = \sum_{i\leq I} f_i\psi_i$. In view of (4.4)–(4.5) we have, for I large enough,

$$\|h-f\|_n^2 = \sum_{i>I} f_i^2 \leq \frac{C^2}{1+n^{2\beta/r}\lambda_I^\beta} \leq C^2C_1^{-\beta}I^{-2\beta/r}.$$

Setting $I = \text{const} \times \varepsilon^{-r/\beta}$ we get $\|h-f\|_n \leq \varepsilon$. By (4.2), the RKHS-norm of h satisfies

$$\begin{aligned}\|h\|_{\mathbb{H}^c}^2 &= n\sum_{i\leq I}((n/c)^{2/r}(\lambda_i+n^{-2}))^{\alpha+r/2}f_i^2\\ &\lesssim nc^{-2p/r}C^2 + c^{-2p/r}C^2n^{2+2(\alpha-\beta)/r}\lambda_I^{p-\beta}.\end{aligned}$$

The condition on ε ensures that for the choice of I made above and n large enough, $i_0 \leq I \leq \kappa n$. Hence, by (4.4)–(4.5), $\|h\|_{\mathbb{H}^c}^2$ is bounded by a constant times the right-hand side of (4.6). □

4.1.2. Proof of (3.5) *and* (3.6)

Define $B_n = M_n\mathbb{H}_1^{c_n} + \varepsilon_n\mathbb{B}_1$, where $\mathbb{B}_1$ is the unit ball of $(\mathbb{R}^n, \|\cdot\|_n)$, $\varepsilon_n = n^{-\beta/(r+2\beta)}$ again and c_n, M_n are the sequences to be determined below. By Lemma 4.3 we have

$$\log N(2\varepsilon_n, B_n, \|\cdot\|_n) \leq \log N(\varepsilon_n/M_n, \mathbb{H}_1^{c_n}, \|\cdot\|_n) \lesssim c_n\Big(\frac{M_n}{\varepsilon_n\sqrt{n}}\Big)^{\frac{r}{p}},$$

where $p = \alpha + r/2$ again. For $M_n = M\sqrt{n\varepsilon_n^2}$ and $c_n^{p/r} = N\sqrt{n}\varepsilon_n^{(\beta-\alpha)/\beta}$ this is bounded by a constant times $n\varepsilon_n^2$, which proves (3.6).

It remains to show that for given $K_2 > 1$, the constants M and N can be chosen such that (3.5) holds. We have

$$P(f \notin B_n) \leq \int_0^{c_n} P(f \notin M_n\mathbb{H}_1^{c_n} + \varepsilon_n\mathbb{B}_1 \,|\, c)e^{-c}\,dc + \int_{c_n}^\infty e^{-c}\,dc.$$

For $c \le c_n$ we have the inclusion $\mathbb{H}_1^c \subseteq \mathbb{H}_1^{c_n}$. Hence, by the Borell–Sudakov inequality, we have for $c \le c_n$ that

$$\begin{aligned} P(f \notin B_n \,|\, c) &\le P(f \notin M_n \mathbb{H}_1^c + \varepsilon_n \mathbb{B}_1 \,|\, c) \\ &\le 1 - \Phi(\Phi^{-1}(P(\|f\|_n \le \varepsilon_n \,|\, c) + M_n)) \\ &\le 1 - \Phi(\Phi^{-1}(P(\|f\|_n \le \varepsilon_n \,|\, c_n) + M_n)), \end{aligned}$$

where Φ is the cdf of the standard normal distribution. By Lemma 4.1 the small ball probability in this expression is for $c_n^{p/r} = N\sqrt{n}\varepsilon_n^{(\beta-\alpha)/\beta}$ bounded from below by $\exp(-K\varepsilon_n^{-r/\beta})$ for some $K > 0$. Using the bound $\Phi^{-1}(y) \ge -((5/2)\log(1/y))^{1/2}$ for $y \in (0, 1/2)$, it follows that for $c \le c_n$,

$$P(f \notin B_n \,|\, c) \le 1 - \Phi(M_n - K'\varepsilon_n^{-r/(2\beta)})$$

for some $K' > 0$. For M_n a large enough multiple of that $\varepsilon_n^{-r/(2\beta)}$ this is bounded by $\exp(-K_2\varepsilon_n^{-r/\beta}) = \exp(-K_2 n\varepsilon_n^2)$.

Lemma 4.3. *For n large enough and $c, \varepsilon > 0$ we have*

$$\log N(\varepsilon, \mathbb{H}_1^c, \|\cdot\|_n) \lesssim c\Big(\frac{1}{\varepsilon\sqrt{n}}\Big)^{\frac{r}{\alpha+r/2}}. \tag{4.7}$$

Proof. By expanding the RKHS elements in the eigenbasis of the Laplacian and taking into account the relations (4.2) we see that the problem is to bound the entropy $\log N(\varepsilon, A, \|\cdot\|)$, where

$$A = \Big\{x \in \mathbb{R}^n : n\sum_{i=0}^{n-1}((n/c)^{2/r}(\lambda_i + n^{-2})^{\alpha+r/2}x_i^2 \le 1\Big\}.$$

Using the bounds (4.4)–(4.5), it follows that with $p = \alpha + r/2$ and $R = c^{p/r}n^{-(\alpha+r)/r}$ we have the inclusions

$$\begin{aligned} A &\subset \Big\{x \in \mathbb{R}^n : \sum_{i=0}^{n-1}\lambda_i^p x_i^2 \le R^2\Big\} \\ &\subset \Big\{x \in \mathbb{R}^n : \sum_{i\le i_0} x_i^2 \le C_1^{-p}n^{2p}R^2, \quad \sum_{i>i_0} i^{2p/r}x_i^2 \le C_1^{-p}n^{2p/r}R^2\Big\}. \end{aligned}$$

By using the well-known entropy bounds for balls in $\mathbb{R}^{i_0}$ and ellipsoids in ℓ^2 we deduce from this that for $\varepsilon > 0$,

$$\log N(2\varepsilon, A, \|\cdot\|) \lesssim \log_+\Big(\frac{n^pR}{\varepsilon}\Big) + \Big(\frac{n^{p/r}R}{\varepsilon}\Big)^{r/p} \lesssim \Big(\frac{n^{p/r}R}{\varepsilon}\Big)^{r/p}.$$

The proof is completed by recalling the expressions for p and R. □

4.2. Proof of Theorem 3.3

In this case the eigenvalues of Σ_c^{-1} are given by

$$\mu_i = n^{-1}e^{(n/c)^{2/r}\lambda_i}.$$

4.2.1. Proof of (3.9)

By Lemma 5.3 of van der Vaart and van Zanten (2008b), it follows from Lemmas 4.4 and 4.5 ahead that under the conditions of the theorem and for $\varepsilon = \varepsilon_n = (n/\log^{1+r/2} n)^{-\beta/(r+2\beta)}$ and $n\varepsilon^2/\log^{1+r/2} n \leq c \leq 2n\varepsilon^2/\log^{1+r/2} n$, we have

$$-\log P(\|f - f_0\|_n \leq \varepsilon \,|\, c) \lesssim c \log^{1+r/2} \frac{c}{\varepsilon^2} + e^{Kc^{-2/r}\varepsilon^{-2/\beta}} \lesssim n\varepsilon^2.$$

By conditioning, similar as in the previous case, we find that $-\log P(\|f-f_0\|_n \leq \varepsilon) \lesssim n\varepsilon^2$ as well.

Lemma 4.4. *If $\varepsilon \to 0$, c is bounded away from 0 and $c/\varepsilon^2 \to \infty$, then*

$$-\log P(\|f\|_n \leq \varepsilon \,|\, c) \lesssim c \log^{1+r/2} \frac{c}{\varepsilon^2}.$$

Proof. Again the series representation of the Gaussian law of $f \,|\, c$ gives $P(\|f\|_n \leq \varepsilon \,|\, c) = P(\sum e^{-(n/c)^{2/r}\lambda_i} Z_i^2 \leq \varepsilon^2)$, where the Z_i are independent standard normal random variables. By the lower bounds (4.4)–(4.5), it follows that

$$\begin{aligned} &P(\|f\|_n \leq 2\varepsilon \,|\, c) \\ &\quad \geq P\Big(\sum_{i \leq i_0} e^{-C_1 n^{(2-2r)/r} c^{-2/r}} Z_i^2 \leq \varepsilon^2\Big) P\Big(\sum_{i \geq 1} e^{-C_1 c^{-2/r} i^{2/r}} Z_i^2 \leq \varepsilon^2\Big). \end{aligned}$$

The first probability in the last line is bounded from below by

$$\Big(P(|Z_1| < i_0^{-1/2} e^{(1/2)C_1 n^{(2-2r)/r} c^{-2/r}} \varepsilon)\Big)^{i_0}.$$

Since the quantity on the right of the inequality in this probability becomes arbitrarily small under de conditions of the lemma, this is further bounded form below by a constant times $\varepsilon^{i_0} \exp(i_0((1/2)C_1 n^{(2-2r)/r} c^{-2/r}))$.

For the second probability we use Theorem 6.1 of Li and Shao (2001). This asserts that if $a_k > 0$ and $\sum a_k < \infty$, then as $\varepsilon \to 0$

$$P(\sum a_i Z_i^2 \leq \varepsilon^2) \sim \frac{1}{\sqrt{4\pi \sum (\frac{a_i\gamma_a}{1+2a_i\gamma_a})^2}} e^{\varepsilon^2\gamma_a - (1/2)\sum \log(1+2a_i\gamma_a)}, \tag{4.8}$$

where $\gamma_a = \gamma_a(\varepsilon)$ is uniquely determined, for $\varepsilon > 0$ small enough, by the equation

$$\varepsilon^2 = \sum \frac{a_i}{1 + 2a_i\gamma_a}. \tag{4.9}$$

We apply (4.8) with $a_i = \exp(-C_1(i/c)^{2/r})$.

We first determine bounds for γ_a. Note that in our case the terms in the sum S on the right of (4.9) are decreasing in i. It follows that we have the bounds

$$\int_1^\infty \frac{1}{e^{C_1(x/c)^{2/r}} + 2\gamma_a}\,dx \leq S \leq \int_0^\infty \frac{1}{e^{C_1(x/c)^{2/r}} + 2\gamma_a}\,dx.$$

A change of variables shows that the integral on the right equals

$$\frac{cr}{2C_1^{r/2}}\int_0^\infty \frac{t^{r/2-1}}{e^t+2\gamma_a}\,dt = c\frac{-r\Gamma(r/2)}{4\gamma_a C_1^{r/2}}\mathrm{Li}_{r/2}(-2\gamma_a),$$

where $\mathrm{Li}_s(z)$ denotes the polylogarithm. By Wood (1992),

$$\frac{\mathrm{Li}_{r/2}(-2\gamma_a)}{\log^{r/2} 2\gamma_a} \to -\frac{1}{\Gamma(r/2+1)}$$

as $\gamma_a \to \infty$. Hence for large γ_a, we have the upper bound $S \leq$ const $\times$ $c{\gamma_a}^{-1}\log^{r/2}\gamma_a$. It is easily seen that we have a lower bound of the same order, so that

$$\varepsilon^2 \asymp \frac{c\log^{r/2}\gamma_a}{\gamma_a}.$$

Under our condition that $\varepsilon^2/c \to 0$ this holds if and only if

$$\gamma_a \asymp \frac{c}{\varepsilon^2}\log^{r/2}\frac{c}{\varepsilon^2}.$$

Next we consider the sums appearing on the right of (4.8). To bound $\sum\log(1+2a_i\gamma_a) \leq \sum\log(1+2\exp(-C_1(i/c)^{2/r})\gamma_a)$ we consider the index $I = c(\log\gamma_a/C_1)^{r/2}$, which is determined such that $a_I\gamma_a = 1$. Note that for $m > 0$, we have $a_{mI}\gamma_a = a_I^{m^{2/r}}\gamma_a = \gamma_a^{1-m^{r/2}}$. We first split up the sum, writing

$$\sum\log(1+2a_i\gamma_a) = \sum_{i<I}\log(1+2a_i\gamma_a) + \sum_{i\geq I}\log(1+2a_i\gamma_a)$$

The first sum on the right is bounded by a multiple of $I\log\gamma_a$. The second one we split into blocks of length I. This gives

$$\sum_{i\geq I}\log(1+2a_i\gamma_a) \leq I\sum_{m\geq 1}\log(1+2\gamma_a^{1-m^{r/2}}) \lesssim I.$$

Hence, we have $\sum\log(1+2a_i\gamma_a) \lesssim c\log^{1+r/2}\gamma_a$. For the other sum appearing in (4.8) we have

$$\sum\Big(\frac{2a_i\gamma_a}{1+2a_i\gamma_a}\Big)^2 \leq \sum\frac{2a_i\gamma_a}{1+2a_i\gamma_a} = 2\gamma_a\varepsilon^2.$$

The proof is completed by combining all the bounds we have found. □

Lemma 4.5. *Suppose that $f \in H^\beta(C)$ for some $\beta, C > 0$. For $\varepsilon > 0$ such that $\varepsilon \to 0$ as $n \to \infty$ and $1/\varepsilon = o(n^{\beta/r})$ and $c > 0$,*

$$\inf_{h\in\mathbb{H}^c:\,\|h-f\|_n\leq\varepsilon}\|h\|^2_{\mathbb{H}^c} \lesssim e^{Kc^{-2/r}\varepsilon^{-2/\beta}} \tag{4.10}$$

for n large enough, where $K > 0$ is a constant.

Proof. We use an expansion $f = \sum f_i\psi_i$, with ψ_i the orthonormal eigenfunctions of the Laplacian. We saw in the proof of Lemma 4.2 that if we define $h =$

$\sum_{i\le I} f_i\psi_i$ for $I = \text{const} \times \varepsilon^{-r/\beta}$, then $\|h - f\|_n \le \varepsilon$. By (4.2), the RKHS-norm of h satisfies in this case

$$\|h\|^2_{\mathbb{H}^c} = \sum_{i\le I} e^{(n/c)^{2/r}\lambda_i} f_i^2 \le C^2 e^{(n/c)^{2/r}\lambda_I}.$$

The condition on ε ensures that for the choice of I made above and n large enough, $i_0 \le I \le \kappa n$. Hence, by (4.4)–(4.5), $\|h\|^2_{\mathbb{H}^c}$ is bounded by a constant times the right-hand side of (4.10). □

4.2.2. Proof of (3.10)–(3.11)

Define $B_n := M_n\mathbb{H}_1^{c_n} + \varepsilon_n\mathbb{B}_1$, where ε_n is as above and M_n and c_n are determined below.

For (3.10) we first note again that

$$P(f \notin B_n) \le \int_0^{c_n} P(f \notin M_n\mathbb{H}_1^{c_n} + \varepsilon_n\mathbb{B}_1 \,|\, c)e^{-c}\,dc + \int_{c_n}^\infty e^{-c}\,dc.$$

Exactly as in the proof of (3.5), the Borell–Sudakov inequality implies that for $c \le c_n$,

$$P(f \notin B_n \,|\, c) \le 1 - \Phi(\Phi^{-1}(P(\|f\|_n \le \varepsilon_n \,|\, c_n) + M_n)).$$

By Lemma 4.4 the small ball probability on the right is lower bounded by

$$\exp\Big(-Kc_n \log^{1+r/2}\frac{c_n}{\varepsilon_n^2}\Big).$$

It follows that for $c \le c_n$,

$$P(f \notin B_n \,|\, c) \le 1 - \Phi\Big(M_n - K'\sqrt{c_n \log^{1+r/2}\frac{c_n}{\varepsilon_n^2}}\Big)$$

for some $K' > 0$. For a given $K_2 > 0$, choosing M_n a large multiple of $(c_n \log^{1+r/2}(c_n/\varepsilon_n^2))^{1/2}$ we find that, for large n,

$$P(f \notin B_n) \le e^{-K''c_n \log^{1+r/2}\frac{c_n}{\varepsilon_n^2}} + e^{-c_n} \le 2e^{-c_n}.$$

If $K_2 > 0$ is a given constant, then for c_n a large enough multiple of $n\varepsilon_n^2$, this is bounded by $\exp(-K_2 n\varepsilon_n^2)$.

For these choices of M_n and c_n, Lemma 4.6 implies that the entropy satisfies, for any $\tilde\varepsilon_n \ge \varepsilon_n$,

$$\log N(2\tilde\varepsilon_n, B_n, \|\cdot\|_n) \le \log N(2\varepsilon_n, B_n, \|\cdot\|_n) \lesssim c_n \left(\log\frac{M_n}{\varepsilon_n}\right)^{1+r/2}.$$

This proves that (3.11) holds for $\tilde\varepsilon_n = \varepsilon_n \log^{1/2+r/4} n$.

Lemma 4.6. *Let $\varepsilon, c > 0$ be such that $c\log^{r/2}(1/\varepsilon) \to \infty$ as $n \to \infty$. Then for n large enough,*

$$\log N(\varepsilon, \mathbb{H}_1^c, \|\cdot\|_n) \lesssim c\log^{1+r/2}\Big(\frac{1}{\varepsilon}\Big).$$

Proof. We need to bound the metric entropy of the set

$$A = \{x \in \mathbb{R}^n : \sum_{i=0}^{n-1} e^{(n/c)^{2/r}\lambda_i} x_i^2 \leq 1\},$$

relative to the Euclidean norm $\|\cdot\|$. Set $I = (2/C_1)^{r/2} c \log^{r/2}(1/\varepsilon)$. Under the assumption of the lemma this is larger than i_0, hence by (4.4)–(4.5) we have $\exp(-(n/c)^{2/r}\lambda_I) \leq \varepsilon^2$. It follows that if for $x \in A$ we define the projection x^I by $x^I = (x_1, \ldots, x_I, 0, 0, \ldots)$, then

$$\|x - x^I\|^2 = \sum_{i>I} x_i^2 \leq e^{-(n/c)^{2/r}\lambda_I} \sum_{i>I} e^{(n/c)^{2/r}\lambda_i} x_i^2 \leq \varepsilon^2.$$

Moreover, we have $\|x^I\| \leq 1$. By the triangle inequality, it follows that if the points $x_1, \ldots, x_N$ form an ε-net for the unit ball in $\mathbb{R}^I$, then the points $\bar{x}_1, \ldots, \bar{x}_N$ in $\mathbb{R}^n$ obtained by appending zeros to the x_j form a 2ε-net for A. Hence, $N(2\varepsilon, A, \|\cdot\|) \lesssim \varepsilon^{-I}$. The proof is completed by recalling the expression for I. □

5. Function estimation on graphs

In this section we translate the general Theorems 3.2 and 3.3 into results about posterior contraction in nonparametric regression and binary classification problems on graphs. Since the arguments needed for this translation are very similar to those in earlier papers, we omit full proofs and just give pointers to the literature.

5.1. Nonparametric regression

As before we let G be a connected simple undirected graph with vertices $1, 2, \ldots, n$. In the regression case we assume that we have observations $Y_1, \ldots, Y_n$ at the vertices of the graph, satisfying

$$Y_i = f_0(i) + \varepsilon_i, \tag{5.1}$$

where f_0 is the function on G that we want to make inference about and ε_i are independent $N(0, \sigma^2)$-distributed error variables, for some $\sigma > 0$. We assume that the underlying graph satisfies the geometry assumption with some parameter $r \geq 1$. As prior on the regression function f we then employ one of the two priors described by (3.2)–(3.3) or (3.7)–(3.8). If σ is unknown, we assume it belongs to a compact interval $[a, b] \subset (0, \infty)$ and endow it with a prior with a positive, continuous density on $[a, b]$, independent of the prior on f.

For a given prior Π, the corresponding posterior distribution on f is denoted by $\Pi(\cdot \,|\, Y_1, \ldots, Y_n)$. For a sequence of positive numbers $\varepsilon_n \to 0$ we say that the posterior contracts around f_0 at the rate ε_n if for all large enough $M > 0$,

$$\Pi(f : \|f - f_0\|_n \geq M\varepsilon_n \,|\, Y_1, \ldots, Y_n) \overset{P_{f_0}}{\to} 0$$

as $n \to \infty$. Here the convergence is in probability under the law P_{f_0} corresponding to the data generating model (5.1).

The usual arguments allow us to derive the following statements from Theorems 3.2 and 3.3. See, for instance, van der Vaart and van Zanten (2008a) or de Jonge and van Zanten (2013) for details.

Theorem 5.1 (Nonparametric regression)**.** *Suppose the geometry assumption holds for $r \geq 1$. Assume that $f_0 \in H^\beta(C)$ for $\beta, C > 0$.*

- *(i) (Power of the Laplacian.) If the prior on f is given by (3.2)–(3.3) for $\alpha > 0$ and $\beta \leq \alpha + r/2$, then the posterior contracts around f_0 at the rate $n^{-\beta/(r+2\beta)}$.*
- *(ii) (Exponential of the Laplacian.) If the prior on f is given by (3.7)–(3.8), then the posterior contracts around f_0 at the rate $n^{-\beta/(r+2\beta)} \log^\kappa n$ for some $\kappa > 0$.*

Observe that since the priors do not use knowledge of the regularity β of the regression function, we obtain rate-adaptive results. For the power prior the range of regularities that we can adapt to is bounded by $\alpha + r/2$, where α is the hyper parameter describing the “baseline regularity” of the prior. In the case of the exponential prior the range is unbounded. This comes at the modest cost of having an additional logarithmic factor in the rate.

In Kirichenko and van Zanten (2017) minimax lower bounds are presented which complement the rate results of the present paper. These show that the rates obtained are sharp in the present setting (up to a logarithmic factor in the exponential case). For the regular grid case this is basically also clear from existing lower bound results, since our setup includes the regular grids (Example 2.1) and since our smoothness condition corresponds to ordinary Sobolev regularity in those cases (Example 3.1).

5.2. Nonparametric classification

We can derive the analogous results in the classification problem in which we assume that the data $Y_1, \dots, Y_n$ are independent $\{0,1\}$-valued variables, observed at the vertices of the graph. In this case the goal is to estimate the binary regression function p_0, or “soft label function” on the graph, given by

$$p_0(i) = P_0(Y_i = 1).$$

We consider priors on p constructed by first defining a prior on a real-valued function f by (3.2)–(3.3) or (3.7)–(3.8) and then setting $p = \Psi(f)$, where $\Psi : \mathbb{R} \to (0,1)$ is a suitably chosen link function. We will assume that Ψ is a strictly increasing, differentiable function onto $(0,1)$ such that $\Psi'/(\Psi(1-\Psi))$ is uniformly bounded. Note that for instance the sigmoid, or logistic link $\Psi(f) = 1/(1+\exp(-f))$ satisfies this condition. Under our conditions the inverse $\Psi^{-1} : (0,1) \to \mathbb{R}$ is well defined. In this classification setting the regularity condition will be formulated in terms of $\Psi^{-1}(p_0)$. This is natural, since the prior is defined in terms of $\Psi^{-1}(p)$ as well. Also in this case we denote the posterior corresponding to a prior Π by $\Pi(\cdot \,|\, Y_1, \ldots, Y_n)$ and we say that the posterior contracts around p_0 at the rate ε_n if for all large enough $M > 0$,

$$\Pi(p : \|p - p_0\|_n \geq M\varepsilon_n \,|\, Y_1, \ldots, Y_n) \overset{P_0}{\to} 0$$

as $n \to \infty$.

To derive the following result from Theorems 3.2 and 3.3 we can argue along the lines of the proof of Theorem 3.2 of van der Vaart and van Zanten (2008a). Some adaptations are required, since in the present case we have fixed design points. However, the necessary modifications are straightforward and therefore omitted.

Theorem 5.2 (Classification)**.** *Suppose the geometry assumption holds for $r \geq 1$. Let $\Psi : \mathbb{R} \to (0, 1)$ be onto, strictly increasing, differentiable and suppose that $\Psi'/(\Psi(1 - \Psi))$ is uniformly bounded. Assume that $\Psi^{-1}(p_0) \in H^\beta(C)$ for $\beta, C > 0$.*

(i) (Power of the Laplacian.) If the prior on p is given by the law of $\Psi(f)$, for f given by (3.2)–(3.3) for $\alpha > 0$ and $\beta \leq \alpha + r/2$, then the posterior contracts around p_0 at the rate $n^{-\beta/(r+2\beta)}$.

(ii) (Exponential of the Laplacian.) If the prior on p is given by the law of $\Psi(f)$, for f given by (3.7)–(3.8), then the posterior contracts around f_0 at the rate $n^{-\beta/(r+2\beta)} \log^\kappa n$ for some $\kappa > 0$.

6. Concluding remarks

We have introduced a framework for studying the performance of methods for nonparametric function estimation on large graphs. We have proposed assumptions on the geometry of the underlying graph and the regularity of the function formulated in terms of the Laplacian of the graph. Moreover, we have exhibited nonparametric Bayes methods that achieve good convergence rates and that adapt to the unknown regularity of the function of interest.

We have purposely focused on the building up a new framework and deriving a few representative results within that framework and have not yet attempted to explore every possible extension. As a result, extensions and generalizations are possible in a variety of directions.

First of all, it is of interest to study other procedures than just the Bayesian methods with priors (3.2)–(3.3) or (3.7)–(3.8). For instance, empirical Bayes procedures for choosing the hyperparameter c might computationally be more favorable than hierarchical Bayes. Studying the performance of such procedures is possible within the framework of Rousseau and Szabo (2015). In turn, having results for empirical Bayes will allow us to extend the range of priors on c for which we can prove that the hierarchical procedures give good results.

Secondly, results on uncertainty quantification would be valuable. Bayes procedures provide a natural method for quantifying uncertainty through the spread of the posterior distribution. However, it has become clear that in general the question of whether or not Bayesian credible sets can be interpreted as (frequentist) confidence sets is a delicate matter in nonparametric settings (e.g. Szabó et al. (2015)). It would be desirable to have more insight in this issue in the graph setting.

On the level of the geometry assumption, several extensions might be of interest. For instance, instead of a single parameter r governing the "dimension" of the graph it might be interesting to consider frameworks allowing graphs which are less homogenous. When estimating a function on some sub-region of a graph, one would expect that the rates should only depend on the local geometry of the graph in that region. It would be of interest to make such statements mathematically precise and to exhibit procedures with good local properties. More generally, recent numerical work has shown that Bayesian Laplacian regularisation can work quite well in practice on graphs that do not satisfy our geometry assumption, see Hartog and van Zanten (2016). To understand this theoretically our current mathematical results are too limited.

A final possible generalization that we want to mention is to the setting of weighted graphs. This is of interest, since in many applications it is natural to work with weighted graphs to quantify the similarity between vertices. We expect that with additional work our results can be extended to that setting.

References

Ando, R. K. and Zhang, T. (2007). Learning on graph with Laplacian regularization. *Advances in neural information processing systems* **19**, 25.

Belkin, M., Matveeva, I. and Niyogi, P. (2004). Regularization and semi-supervised learning on large graphs. In *COLT*, volume 3120, pp. 624–638. Springer. MR2177939

Borgs, C., Chayes, J. T., Cohn, H. and Zhao, Y. (2014). An l^p theory of sparse graph convergence i: limits, sparse random graph models, and power law distributions. *arXiv:1401.2906*.

Castillo, I., Kerkyacharian, G. and Picard, D. (2014). Thomas Bayes walk on manifolds. *Probability Theory and Related Fields* **158**(3–4), 665–710. MR3176362

Chung, F. (2014). From quasirandom graphs to graph limits and graphlets. *Advances in Applied Mathematics* **56**, 135–174. MR3194210

Cressie, N. (1993). *Statistics for Spatial Data*. Wiley. MR1239641

Cvetković, D., Rowlinson, P. and Simić, S. (2010). An introduction to the theory of graph spectra. *Cambridge*.

de Jonge, R. and van Zanten, J. H. (2013). Semiparametric Bernstein–von Mises for the error standard deviation. *Electron. J. Stat.* **7**, 217–243.

Dunker, T., Lifshits, M. and Linde, W. (1998). Small deviation probabilities of sums of independent random variables. In *High dimensional probability*, pp. 59–74. Springer. MR1652320

Hartog, J. and van Zanten, J. H. (2016). Nonparametric Bayesian label prediction on a graph. *ArXiv e-prints*.

Hein, M. (2006). Uniform convergence of adaptive graph-based regularization. In *International Conference on Computational Learning Theory*, pp. 50–64. Springer. MR2277918

Huang, J., Ma, S., Li, H. and Zhang, C.-H. (2011). The sparse Laplacian shrinkage estimator for high-dimensional regression. *Annals of statistics* **39**(4), 2021.

Johnson, R. and Zhang, T. (2007). On the effectiveness of Laplacian normalization for graph semi-supervised learning. *Journal of Machine Learning Research* **8**(4).

Kirichenko, A. and van Zanten, J. H. (2017). Minimax lower bounds for function estimation on graphs. *In preparation*.

Kolaczyk, E. D. (2009). *Statistical analysis of network data*. Springer Series in Statistics. Springer, New York. Methods and models. MR2724362

Li, W. V. and Shao, Q.-M. (2001). Gaussian processes: inequalities, small ball probabilities and applications. *Stochastic processes: theory and methods* **19**, 533–597.

Liu, X., Zhao, D., Zhou, J., Gao, W. and Sun, H. (2014). Image interpolation via graph-based Bayesian label propagation. *Image Processing, IEEE Transactions on* **23**(3), 1084–1096. MR3172970

Lovasz, L. (2012). *Large networks and graph limits*, volume 60. American Mathematical Soc.

Lovász, L. and Szegedy, B. (2006). Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B* **96**(6), 933–957.

Mohar, B. (1991a). Eigenvalues, diameter, and mean distance in graphs. *Graphs Combin.* **7**(1), 53–64. MR1105467

Mohar, B. (1991b). The Laplacian spectrum of graphs. *Graph theory, combinatorics, and applications* **2**, 871–898.

Rasmussen, C. E. and Williams, C. K. I. (2006). Gaussian processes for machine learning. *MIT Press*.

Rousseau, J. and Szabo, B. (2015). Asymptotic behaviour of the empirical Bayes posteriors associated to maximum marginal likelihood estimator. *arXiv preprint arXiv:1504.04814*.

Sharan, R., Ulitsky, I. and Shamir, R. (2007). Network-based prediction of protein function. *Molecular systems biology* **3**(1), 88.

Smola, A. J. and Kondor, R. (2003). Kernels and regularization on graphs. In *Learning theory and kernel machines*, pp. 144–158. Springer.

Szabó, B., van der Vaart, A. W. and van Zanten, J. H. (2015). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *Ann. Statist.* **43**(4), 1391–1428.

van der Vaart, A. W. and van Zanten, J. H. (2008a). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.* **36**(3), 1435–1463.

van der Vaart, A. W. and van Zanten, J. H. (2008b). Reproducing kernel Hilbert spaces of Gaussian priors. *IMS Collections* **3**, 200–222.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature* **393**(6684), 440–442. MR1716136

Wood, D. (1992). The computation of polylogarithms. Technical Report 15-92, University of Kent, Computing Laboratory, University of Kent, Canterbury, UK.

Zhu, X. and Ghahramani, Z. (2002). Learning from labeled and unlabeled data with label propagation. Technical report.

Zhu, X., Ghahramani, Z., Lafferty, J. et al. (2003). Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, volume 3, pp. 912–919.