Asymptotic results in nonparametric Bayesian function estimation

Kirichenko, A.

Chapter 2

The problem of function estimation on large graphs

2.1 Introduction

2.1.1 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 and classification problems arise naturally in different applications.

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 penalisation technique that takes the geometry of the graph into account and that is thought to produce an appropriate bias-variance trade-off.

In this chapter we set up the mathematical framework that allows us to study the performance of such 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.

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 rate 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.

2.1.2 Asymptotic behaviour of estimators

To assess the performance of procedures we take an asymptotic perspective. We let the number of vertices of the graph grow and ask how fast an estimation procedure can "learn" the underlying function of interest. We derive minimax rates for regression and binary classification on the graph by providing a lower bound on them and presenting an estimator that achieves that rate. In order to do that 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 Lovász [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 $\lambda_{i,n}$ of L satisfy

$$\lambda_{i,n}^2 \asymp \left(\frac{i}{n}\right)^{2/r} \tag{2.1}$$

for some $r \geq 1$. Here we write $a_n \asymp b_n$ if $0 < \liminf a_n/b_n \leq \limsup a_n/b_n < \infty$. 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 2.2.1), 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 path graph, and indeed we will see that it satisfies our assumption for $r = 1$ as well (Example 2.2.3). Moreover, the constant r in (2.1) 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 (2.1).

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 2.3.1. However, in applied settings (see, for instance, Example 2.2.7) it is typically not clear what a suitable ambient manifold could be, which is why we choose to avoid this issue altogether.

2.1.3 Organisation

The remainder of the chapter is organised as follows. In the next section we present our geometry assumption and give examples of graphs that satisfy it, either theoretically or numerically. In Section 2.3 we introduce Sobolev-type balls that are used to quantify the regularity of the function. In Section 2.4 we obtain minimax rates for regression and binary classification problems on large graphs. The mathematical proofs are given in Section 2.5.1 and 2.5.2.

2.2 Asymptotic geometry assumption on graphs

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

2.2.1 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

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

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 $\langle \cdot, \cdot \rangle_n$.

2.2.2 Asymptotic geometry assumption

As mentioned in the introduction, we 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 2.2.1 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,

$$C_1\Big(\frac{i}{n}\Big)^{2/r} \leq \lambda_i \leq C_2\Big(\frac{i}{n}\Big)^{2/r}, \qquad \text{for all } i \in \{i_0, \dots, \kappa n\}. \tag{2.2}$$

Note that this condition only restricts a positive fraction κ of the Laplacian eigenvalues, namely the κn smallest ones. Moreover, we do not 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* $i \in \{i_0, \dots, n\}$. 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 $\lambda_2, \dots, \lambda_{i_0}$ as well. Indeed, by Theorem 4.2 of Mohar [1991a] we have

$$\lambda_1 \geq \frac{4}{\operatorname{diam}(G) n} \geq \frac{4}{n^2}. \tag{2.3}$$

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 2.2.1 (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

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

where $i_j = 0, \dots, n^{1/d} - 1$ for $j = 1, \dots, d$. 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 $R = cn^{1/d}\sqrt{\lambda_i}$ 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 2.2.2 (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 condition with parameter $r = d$ as well.

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

Lemma 2.2.1. *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

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

The result follows from this observation. ∎

Example 2.2.3 (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 2.2.1 and Example 2.2.1 it satisfies the geometry condition with parameter $r = 1$.

Example 2.2.4 (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 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 2.2.5 (Two-dimensional grid, numerically)**.** Figure 2.1 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 2.2.1 this line has slope 1.0.

Example 2.2.6 (Watts-Strogatz 'small world' graph)**.** In our second numerical example we consider a graph obtained as a realisation 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 2.2 a particular realisation is shown. Here we have only kept the largest connected component, which has 175 vertices in this case. 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 2.2.7 (Protein interaction graph)**.** In the final example we consider a graph obtained from the protein interaction graph of baker's yeast, as described in

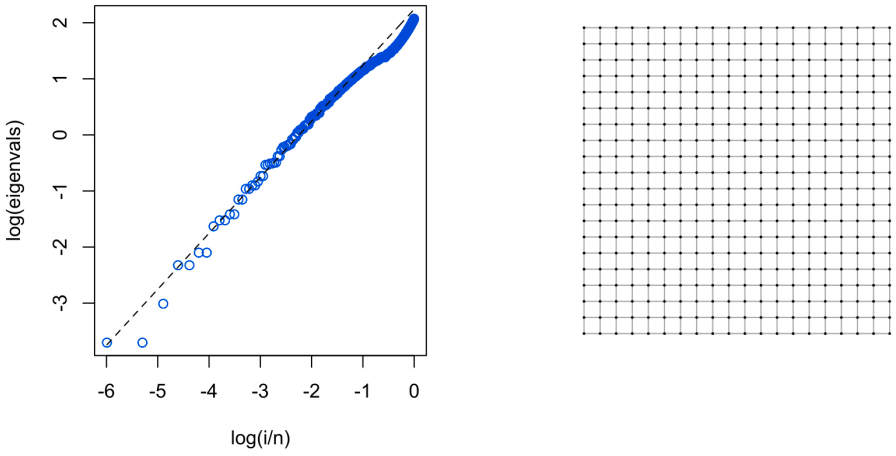

Figure 2.1: Plot of $\log \lambda_i$ against $\log(i/n)$ for the 20×20 grid. Fitted line has slope 1.0, corresponding to $r = 2.0$ in the geometry assumption.

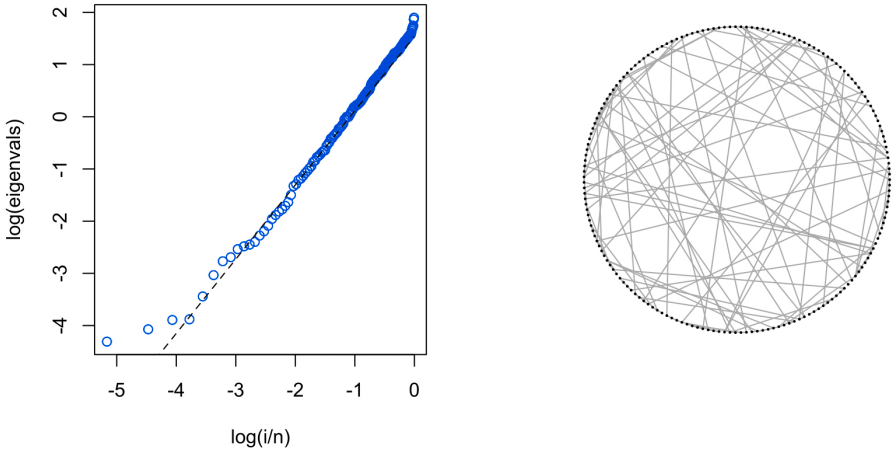

Figure 2.2: Plot of $\log \lambda_i$ against $\log(i/n)$ for the Watts-Strogatz graph in the right panel. Fitted line has slope 1.42, corresponding to $r = 1.4$ in the geometry assumption.

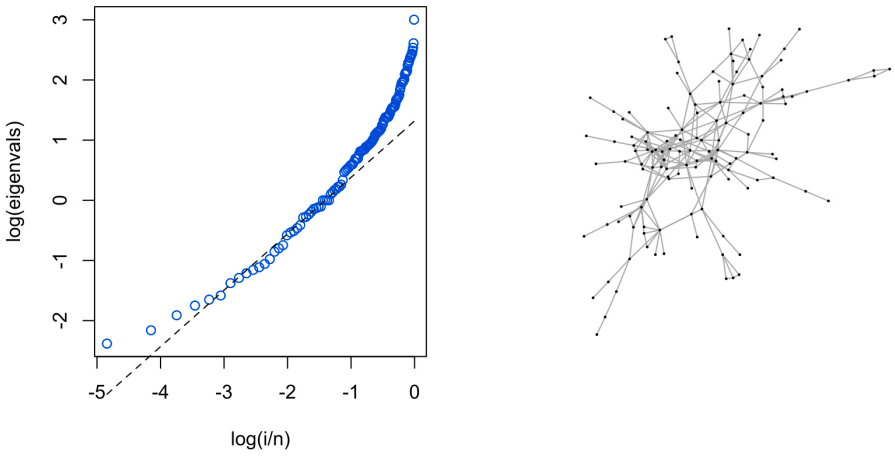

Figure 2.3: Plot of $\log \lambda_i$ against $\log(i/n)$ for the protein interaction graph in the right panel. Fitted line has slope 0.94, corresponding to $r = 2.1$ in the geometry assumption.

detail in Section 8.5 of Kolaczyk [2009]. The graph, shown in the right panel of Figure 2.3, describes the interactions between proteins involved in the communication 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.

2.3 Smoothness assumption

We define the "regularity" of the function using the graph Laplacian. Specifically, we will assume it belongs to a Sobolev-type ball of the form

$$H^\beta(Q) = \left\{ f : \left\langle f, (I + (n^{\frac{2}{r}} L)^\beta) f \right\rangle_n \leq Q^2 \right\} \tag{2.4}$$

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

It is illustrative to look at this assumption in a bit more detail in the simple case of the path graph.

Example 2.3.1 (Path graph)**.** Consider a path graph G with n vertices, which we identify with the points i/n in the unit interval, $i = 1, \ldots, n$. As seen in Example 2.2.1, this graph satisfies the geometry condition with parameter $r = 1$. Hence, in

this case the collection of functions $H^\beta(Q)$ is given by

$$H^\beta(Q) = \left\{ f : \left\langle f, (I + (n^2 L)^\beta) f \right\rangle_n \leq Q^2 \right\}.$$

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 $\{i/n, i = 1, \ldots, n\}$ of a fixed function f defined on the whole interval $[0, 1]$. The assumption that $f_n \in H^1(Q)$ then translates to the requirement that

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

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 $H^1(Q)$ on the graph is the natural discrete version of the usual Sobolev ball

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

Definition (2.4) 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".

2.4 Minimax rates for function estimation problems on graphs

In this section we introduce two prototypical function estimation problems on the graphs: regression and binary classification problems. We assume that the underlying graph satisfies the geometry assumption with some $r \geq 1$. We show that the minimax rates over the balls $H^\beta(Q)$ for both problems depend on the "dimension" of the graph r and the regularity β of the target function. Specifically, we derive that the minimax rate behaves as $n^{-\beta/(2\beta+r)}$, when n is large enough. Naturally the complexity of the problem increases with the growth of the dimension r. There is also the usual dependence on the regularity of the function, meaning that it is easier to estimate a function which varies along the edges of the graph more slowly than a more rough one.

The minimax results we derive consist of two parts. We obtain a lower bound

for the target family of functions and we construct a specific estimator for each problem such that its maximum risk is within a constant factor of the derived lower bound. However, the developed estimators are not rate-adaptive and thus have small applied value. In Chapter 3 we present a Bayesian approach to those problems and introduce the associated priors that have an (almost) minimax rate of convergence of posterior distribution and which are rate-adaptive.

Let G be a connected simple undirected graph with vertices $1, \ldots, n$, satisfying the geometry assumption for $r \geq 1$. In the regression case we assume we have an observation set $Y = (Y_1, \ldots, Y_n)$ on the vertices of the graph such that

$$Y_i \;=\; f_0(i) + \sigma \xi_i, \tag{2.5}$$

where the ξ_i are independent standard Gaussian, $\sigma > 0$ and f_0 is the function of interest.

Under these conditions we derive the following result.

Theorem 2.4.1. *Suppose the geometry assumption holds for $r \geq 1$. Then for every $\beta > 0$*

$$\inf_{\hat{f}} \sup_{f \in H^\beta(Q)} \mathbb{E}_f \left(\|\hat{f} - f\|_n^2 \right) \asymp n^{-2\beta/(2\beta+r)}, \tag{2.6}$$

where the infimum is taken over all estimators $\hat{f}$.

The theorem shows that the minimax rate for the regression problem on the graph is equal to $n^{-\beta/(2\beta+r)}$. We obtain an upper bound on the rate by constructing a projection estimator $\tilde{f}$ for which

$$\sup_{f \in H^\beta(Q)} \mathbb{E}_f \left(\|\tilde{f} - f\|_n^2 \right) \lesssim n^{-2\beta/(2\beta+r)}.$$

However, the choice of our estimator depends on the regularity β of the true function. It means that in real life the estimator is hardly applicable, since we rarely know the smoothness of the target functions in advance. We overcome this problem in the next chapter by developing Bayesian estimators that are rate-adaptive and that have optimal rates of convergence (up to a logarithmic factor).

In the binary classification problem we assume that the data $Y_1, \ldots, Y_n$ 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 ρ_0, or "soft label function" on the graph, given by

$$\rho_0(i) = \mathbb{P}_0(Y_i = 1).$$

We employ a suitably chosen link function $\Psi : \mathbb{R} \to (0, 1)$. We assume that Ψ is a differentiable function onto $(0, 1)$ such that $\Psi'/(\Psi(1 - \Psi))$ is uniformly bounded, and $\Psi'(x) > 0$ for every $x \in \mathbb{R}$. 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 $\Psi^{-1}(\rho_0)$.

Theorem 2.4.2. *Suppose the geometry assumption holds for $r \geq 1$. Consider the orthonormal basis ψ_j of the graph Laplacian (normalised with respect to the $\|\cdot\|_n$–norm). Assume that there exists $C > 0$ such that for every $n \in \mathbb{N}$ and for all $i = 1, \ldots, n$ and $j = 0, \ldots, n-1$ it holds that $|\psi_j(i)| \leq C$. Let $\Psi : \mathbb{R} \to (0,1)$ be a differentiable onto (0,1) such that $\Psi'/(\Psi(1-\Psi))$ is uniformly bounded and $\Psi'(x) > 0$ for every $x \in \mathbb{R}$. Then for $\beta \geq r/2$*

$$\inf_{\hat{\rho}} \sup_{\rho \in \{\rho : \Psi^{-1}(\rho) \in H^\beta(Q)\}} \mathbb{E}_\rho \|\hat{\rho} - \rho\|_n^2 \asymp n^{-2\beta/(2\beta+r)},$$

where the infimum is taken over all estimators $\hat{\rho}$.

According to the theorem the minimax rate for the classification problem is again $n^{-\beta/(2\beta+r)}$. To obtain an upper bound on the minimax risk in the classification setting we use an estimator based on the projection estimator from the proof of the upper bound on the risk in Theorem 2.4.1. Again, the developed estimator is not rate-adaptive, which makes it impractical. In the next chapter we address this issue by devising Bayesian estimators that achieve (almost) minimax rate and which are adaptive.

Observe that the result is obtained for functions with regularity levels $\beta \geq r/2$. Additionally, we put an upper bound on the supremum norm for the normalised eigenvectors of the graph Laplacian. These conditions arise from translation of the norms between the set $H^\beta(Q)$ and the set of soft label functions $\{\rho : \Psi^{-1}(\rho) \in H^\beta(Q)\}$.

We note that for d-dimensional grids the supremum norm of the normalised eigenvectors is bounded. In practical applications one can verify numerically whether the condition is satisfied.

Example 2.4.3 (Grids)**.** By Merris [1998] the eigenvectors of the graph Laplacian of the path graph are as follows

$$\tilde{\psi}_j(i) = \cos\left(\pi ij/n - \pi j/2n\right).$$

The $\|\cdot\|_n$–norm of the jth eigenvector is given by

$$\|\tilde{\psi}_j\|_n^2 = \frac{1}{n}\sum_{i=1}^{n} \cos^2\left(\pi ij/n - \pi j/2n\right) =$$

$$= \frac{1}{2} + \frac{1}{4n\sin(\pi j/n)} \sum_{i=1}^{n} 2\sin(\pi j/n)\cos((2i-1)\pi j/n).$$

Using basic trigonometric computation we have for any $x \in \mathbb{R}$

$$\sum_{i=1}^{n} 2\sin x \cos(2ix - x) = \sum_{i=1}^{n} (\sin 2ix - \sin(2ix - 2x)) = \sin 2nx.$$

Hence, for any $j = 1, \dots, n-1$

$$\|\tilde{\psi}_j\|_n^2 = \frac{1}{2} + \frac{1}{4n} \frac{\sin 2\pi j}{\sin \pi j/n} = \frac{1}{2}.$$

Notice that $\|\tilde{\psi}_0\|_n^2 = 1$. Then for the normalised with respect to the $\|\cdot\|_n$–norm eigenvectors ψ_j of the graph Laplacian of the path graph there exists a constant $C > 0$ such that $|\psi_i(j)| \le C$ for all $n \in \mathbb{N}$, $i = 1, \dots, n$ and $j = 0, \dots, n-1$.

From e.g. Merris [1998] we know that the eigenvectors of a Cartesian product of two graphs are equal to the Kronecker products of pairs of eigenvectors associated with the Laplacians of those graphs. Then the eigenvectors of the graph Laplacian for the 2–dimensional grid with n^2 vertices are given by $\psi_i \psi_j$, $i, j = 1, \dots, n$. Hence, the supremum norm of the normalised eigenvectors is bounded. The grids of higher dimensions retain the property, since they are equal to Cartesian products of path graphs.

2.5 Proofs

2.5.1 Proof of Theorem 2.4.1

2.5.1.1 Preliminaries

Let $Y = (Y_1, \dots, Y_n)$, $\xi = (\xi_1, \dots, \xi_n)$, and let ψ_i be the orthonormal eigenfunctions of the graph Laplacian. Denote $\tilde{\xi}_i = \langle \xi, \psi_i \rangle_n$ and observe that the $\tilde{\xi}_i$ are centred Gaussian with

$$\mathbb{E}\tilde{\xi}_i \tilde{\xi}_j = \frac{1}{n} \delta_{ij}.$$

The inner products $Z_i = \langle Y, \psi_i \rangle_n$ satisfy the following relation for $i = 0, \dots, n-1$

$$Z_i = \langle Y, \psi_i \rangle_n = f_i + \sigma \tilde{\xi}_i,$$

where f_i are coefficients in the decomposition of the target function $f_0 = \sum_{i=0}^{n-1} f_i \psi_i$. Additionally, consider the decomposition of an estimator $\hat{f} = \sum_{i=0}^{i-1} \hat{f}_i \psi_i$. Then

$$\|\hat{f} - f_0\|_n^2 = \left\langle \sum_{i=0}^{n-1} (\hat{f}_i - f_i)\psi_i, \sum_{i=0}^{n-1} (\hat{f}_i - f_i)\psi_i \right\rangle_n = \sum_{i=0}^{n-1} (\hat{f}_i - f_i)^2.$$

Hence, the minimax rates for the original problem are of the same order as the minimax rates for the problem of recovering $f = (f_0, \dots, f_{n-1})$, given the observations

$$Z_i = f_i + \varepsilon \zeta_i, \tag{2.7}$$

where ζ_i are independent standard Gaussian and $\varepsilon = \frac{\sigma}{\sqrt{n}}$. To avoid confusion we define general ellipsoids on the space of coefficients for an arbitrary sequence $a_j > 0$

and some finite constant $Q > 0$

$$B_n(Q) = \{f \in \mathbb{R}^n : \sum_{j=0}^{n-1} a_j^2 f_j^2 \leq Q\}. \tag{2.8}$$

For a function f in the Sobolev-type ball $H^\beta(Q)$ its vector of coefficients belongs to $B_n(Q)$ with

$$a_j^2 = 1 + \lambda_j^{2\beta/r} n^{2\beta/r},\ j = 0, \ldots, n-1.$$

In order to prove the theorem it is sufficient to show that

$$\inf_{\hat{f}} \sup_{f \in B_n(Q)} \mathbb{E}_f \left(\sum_{i=0}^{n-1} (\hat{f}_i - f_i)^2 \right) \asymp n^{-2\beta/(2\beta+r)}. \tag{2.9}$$

We are going to follow the proof of Pinker's theorem (see for example Tsybakov [2009]) which studies a similar case in the setting of the Gaussian white noise model on the interval $[0, 1]$. The proof requires some modifications arising from the nature of our problem. The main differences from the Pinsker's result are that we only have n observations and that the definition of $B_n(Q)$ is slightly different than usual.

In order to proceed we first consider the problem of obtaining minimax rates in the class of linear estimators. We introduce Pinsker's estimator and present the linear minimax lemma showing that Pinsker's estimator is optimal in the class of linear estimators. The risk of a linear estimator $\hat{f}(l) = (l_1 Z_1, \ldots, l_n Z_n)$ with $l = (l_1, \ldots, l_n) \in \mathbb{R}^n$ is given by

$$R(l, f) = \mathbb{E}_f \sum_{j=0}^{N-1} (\hat{f}_j - f_j)^2 = \sum_{j=0}^{n-1} \left((1 - l_j)^2 f_j^2 + \varepsilon^2 l_j^2 \right).$$

For large n we introduce the following equation with respect to the variable x

$$\frac{\varepsilon^2}{x} \sum_{j=0}^{n-1} a_j (1 - x a_j)_+ = Q. \tag{2.10}$$

Suppose, there exists a unique solution x of (2.10). For such a solution, define a vector of coefficients l' consisting of entries

$$l'_j = (1 - x a_j)_+.$$

The linear estimator $\tilde{f} = \tilde{f}(l')$ is called the Pinsker estimator for the general ellipsoid $B_n(Q)$. The following lemma that appears as Lemma 3.2 in Tsybakov [2009], shows that the Pinsker estimator is a linear minimax estimator.

Lemma 2.5.1 (Linear minimax lemma)**.** *Suppose that $B_n(Q)$ is a general ellipsoid*

defined by (2.8) with $Q > 0$ and a positive set of coefficients $\{a_j\}_{j=1}^n$. Suppose there exists a unique solution x of (2.10) and suppose that

$$S = \varepsilon^2 \sum_{j=0}^{n-1} l'_j < \infty. \tag{2.11}$$

Then the linear minimax risk satisfies

$$\inf_{l\in\mathbb{R}^n} \sup_{f\in B_n(Q)} R(l, f) = \sup_{f\in B_n(Q)} R(l', f) = S. \tag{2.12}$$

In order to follow the steps the proof of Pinsker's Theorem we present a technical lemma that studies the asymptotic properties of the Pinsker's estimator in our setting.

Lemma 2.5.2. *Consider the ellipsoid $B_n(Q)$ defined by (2.8) with $Q > 0$ and*

$$a_j^2 = 1 + \lambda_j^{2\beta/r} n^{2\beta/r},\ j = 0, \dots, n-1.$$

Then, as $n \to \infty$, we have the following

(i) There exists a solution x of (2.10) which is unique and satisfies

$$x \asymp n^{-\beta/(2\beta+r)}.$$

(ii) The weighted sum (2.11) of the coefficients of the Pinsker's estimator satisfies

$$S \asymp n^{-2\beta/(2\beta+r)}.$$

(iii) For $\varepsilon = \frac{\sigma}{\sqrt{n}}$ define $v_j = \frac{\varepsilon^2(1-xa_j)_+}{xa_j}$. Then

$$\max_{j=0,\dots,n-1} v_j^2 a_j^2 = O\left(n^{-r/(2\beta+r)}\right).$$

Proof. (i) According to Lemma 3.1 from Tsybakov [2009] for large enough n and for an increasing positive sequence a_j, $j = 0 < \dots, n-1$ with $a_n \to +\infty$, as $n \to \infty$, there exists a unique solution of (2.10) given by

$$x = \frac{\varepsilon^2 \sum_{j=0}^{N-1} a_j}{Q + \varepsilon^2 \sum_{j=0}^{N-1} a_j^2},$$

where

$$N = \max\left\{ m : \varepsilon^2 \sum_{j=0}^{m-1} a_j(a_m - a_j) < Q \right\} < +\infty.$$

Consider N defined above. Denote

$$A_m = \varepsilon^2 \sum_{j=0}^{m-1} a_j(a_m - a_j) =$$
$$= n^{\beta/r-1}\sigma^2 \sum_{j=0}^{m-1} \sqrt{\left(1+\lambda_j^{2\beta/r} n^{2\beta/r}\right)\left(\lambda_m^{2\beta/r} n^{2\beta/r} - \lambda_j^{2\beta/r} n^{2\beta/r}\right)}.$$

Since the geometry condition on the graph is satisfied for $j = i_0, \dots, \kappa n$ the eigenvalues of the graph Laplacian can be bounded in a following way

$$C_1\Big(\frac{j}{n}\Big)^{2/r} \le \lambda_j \le C_2\Big(\frac{j}{n}\Big)^{2/r}.$$

Then for some $K_1 > 0$

$$A_m \ge n^{-1}\sigma^2 C_1 \sum_{j=1}^{m} j^{\beta/r}\left(C_1 m^{\beta/r} - C_2 j^{\beta/r}\right) \ge K_1 n^{-1} m^{\beta/r} \sum_{j=1}^{m} j^{\beta/r}.$$

Hence, there exists $K_2 > 0$ such that for all

$$m > K_2 n^{r/(2\beta+r)}$$

it holds that $A_m > Q$. In a similar manner we can show that there exists $K_3 > 0$ such that for all

$$m < K_3 n^{r/(2\beta+r)}$$

it holds that $A_m < Q$.

That leads us to the conclusion that $N \asymp n^{r/(2\beta+r)}$. Then equation (2.10) has a unique solution that satisfies

$$x = \frac{\sigma^2}{n\left(Q + \frac{\sigma^2}{n}\sum_{j=0}^{N-1} a_j^2\right)} \sum_{j=0}^{N-1} a_j \asymp \frac{1}{n} N^{1+\beta/r} \asymp n^{-\beta/(2\beta+r)}.$$

.

(ii) Since G satisfies the geometry assumption, we deduce from (i) that for some $K_4 > 0$

$$l'_j \asymp \left(1 - K_4 n^{-\beta/(2\beta+r)} j^{\beta/r}\right)_+, \text{ for } j = i_o, \dots, N.$$

For $j = 0, \dots, i_0 - 1$ we bound l'_j from above by 1. Then

$$S \asymp n^{-1} i_0 + n^{-1} \sum_{j=i_0}^{N-1} \left(1 - K_4 n^{-\beta/(2\beta+r)} j^{\beta/r}\right)_+ \asymp n^{-1} N \asymp n^{-2\beta/(2\beta+r)}.$$

(iii) Note that for $j > N$ we have $v_j^2 = 0$. We also know that $a_N < \frac{1}{x}$. Then

$$v_j^2 a_j^2 = \frac{\sigma^2 a_j (1 - x a_j)_+}{nx} \leq \frac{\sigma^2 a_N}{nx} \leq \frac{\sigma^2}{nx^2}.$$

Hence, as $n \to \infty$,

$$\max_{j=0,\dots,n-1} v_j^2 a_j^2 = O\left(n^{-r/(2\beta+r)}\right).$$

This finishes the proof of the lemma. ∎

2.5.1.2 Proof of the upper bound on the risk

Recall that we only need to provide the upper bound in (2.9). Consider the Pinsker's estimator $\tilde{f} = (\lambda_0' Z_0, \dots, l_{n-1}' Z_{n-1})$ with

$$l_j' = (1 - x a_j)_+,$$

where $a_j^2 = 1 + \lambda_j^{2\beta/r} n^{2\beta/r}$ and x is a unique solution of (2.10). Using Lemma 2.5.1 and Lemma 2.5.2 we conclude that the Pinsker's estimator satisfies

$$\sup_{f \in B_n(Q)} \mathbb{E}_f \left(\sum_{i=0}^{n-1} (\tilde{f}_i - f_i)^2 \right) \lesssim n^{-2\beta/(2\beta+r)}.$$

2.5.1.3 Proof of the lower bound on the risk

We follow the steps of the general reduction scheme for obtaining minimax rates (see e.g. Chapter 3 of Tsybakov [2009] for more details). First, we reduce the considered parameter space. We show that it is sufficient to only take into account the first N coefficients in the decomposition of the target function, where

$$N = \max \left\{ m : \varepsilon^2 \sum_{j=0}^{m-1} a_j (a_m - a_j) < Q \right\}$$

with $a_j^2 = 1 + \lambda_j^{2\beta/r} n^{2\beta/r}$.

Denote R_n to be the minimax risk

$$R_n = \inf_{\hat{f}} \sup_{f \in B_n(Q)} \mathbb{E}_f \left(\sum_{i=0}^{n-1} (\hat{f}_i - f_i)^2 \right).$$

For the coefficients a_j define

$$B_n(Q,N) = \{f^{(N)} = (f_0,\dots,f_{N-1},0,\dots,0) \in \mathbb{R}^n : \sum_{j=0}^{N-1} a_j^2 f_j^2 \le Q\}.$$

Observe that $B_n(Q,N) \subseteq B_n(Q)$. Then

$$R_n \ge \inf_{\hat{f}^{(N)} \in B_n(Q,N)} \sup_{f^{(N)} \in B_n(Q,N)} \mathbb{E}_f \sum_{j=0}^{N-1} (\hat{f}_j - f_j)^2. \tag{2.13}$$

Following the steps of the proof of Pinsker's theorem, we argue that it is sufficient to bound the Bayes risk instead of the minimax risk. Indeed, consider the density $\mu(f^{(N)}) = \prod_{j=0}^{N-1} \mu_{s_j}(f_j)$ with respect to the Lebesgue measure on $\mathbb{R}^N$. Here $s_j = (1-\delta)v_j^2$ for some $\delta \in (0,1)$ and μ_σ denotes the density of the Gaussian distribution with mean 0 and variance σ^2. By (2.13) we can bound the minimax risk from below by the Bayes risk

$$R_n \ge \inf_{\hat{f}^{(N)} \in B_n(Q,N)} \sum_{j=0}^{N-1} \int_{B_n(Q,N)} \mathbb{E}_f (\hat{f}_j - f_j)^2 \mu(f^{(N)}) df^{(N)} \ge I^\star - r^\star, \tag{2.14}$$

where

$$I^\star = \inf_{\hat{f}^{(N)} \in B_n(Q,N)} \sum_{j=0}^{N-1} \int_{\mathbb{R}^N} \mathbb{E}_f (\hat{f}_j - f_j)^2 \mu(f^{(N)}) df^{(N)};$$

$$r^\star = \sup_{\hat{f}^{(N)} \in B_n(Q,N)} \sum_{j=0}^{N-1} \int_{B_n(Q,N)^c} \mathbb{E}_f (\hat{f}_j - f_j)^2 \mu(f^{(N)}) df^{(N)}$$

with $B(Q,N)^c = \mathbb{R}^N \setminus B_n(Q,N)$. From the proof of Pinsker's theorem we get the following bounds

$$I^\star \gtrsim S,$$

$$r^\star \lesssim \exp\left\{-K\left(\max_{j=0,\dots,n-1} v_j^2 a_j^2\right)^{-1}\right\}$$

for some $K > 0$. Using the results of Lemma 2.5.2 we conclude that $R_n \gtrsim n^{-2\beta/(2\beta+r)}$.

2.5.2 Proof of Theorem 2.4.2

2.5.2.1 Proof of the upper bound on the risk

We define the estimator that gives us an upper bound on the minimax risk based on the estimator $\tilde{f}$, which has been introduced in subsection 2.5.1.2 of the proof of Theorem 2.4.1. Consider the estimator

$$\tilde{\rho} = \Psi\left(\sum_{i=0}^{N-1} \tilde{f}_i \psi_i\right).$$

By the reasoning given in the aforementioned subsection and using the properties of the link function, we can see that

$$\sup_{\rho\in\{\rho:\Psi^{-1}(\rho)\in H^\beta(Q)\}} \mathbb{E}_\rho \|\tilde{\rho}-\rho\|_n^2 \lesssim \\ \lesssim \sup_{f\in B_n(Q)} \mathbb{E}_f \left(\sum_{i=0}^{n-1} (\tilde{f}_i - f_i)^2\right) \lesssim n^{-2\beta/(2\beta+r)}.$$

2.5.2.2 Proof of the lower bound on the risk

The proof of the lower bound on the risk is based on the corollary of Fano's lemma (for more details see Corollary 2.6 in Tsybakov [2009]). Observe that by Markov's inequality for any soft label functions $\rho_0, \ldots, \rho_M$ there exists $C > 0$ such that

$$\inf_{\hat{\rho}} \sup_{\rho\in\{\rho:\Psi^{-1}(\rho)\in H^\beta(Q)\}} \mathbb{E}_\rho n^{2\beta/(2\beta+r)} \|\hat{\rho}-\rho\|_n^2 \gtrsim \\ \gtrsim \inf_{\hat{\rho}} \max_{\rho\in\{\rho_0,\ldots,\rho_M\}} P_\rho\left(\|\hat{\rho}-\rho\|_n^2 \geq n^{-2\beta/(2\beta+r)}\right).$$

Consider probability measures $P_0, P_1, \ldots, P_M$ corresponding to the soft label functions $\rho_0, \ldots, \rho_M$. For a test $\phi : \mathbb{R}^n \to \{0, 1, \ldots, M\}$ define the average probability of error by

$$\overline{p}_M(\phi) = \frac{1}{M+1} \sum_{j=0}^{M} P_j(\phi \neq j).$$

Additionally, let

$$\overline{p}_M = \inf_\phi \overline{p}_M(\phi).$$

From the general scheme for obtaining lower bounds for minimax risk (for more detail see Chapter 2 of Tsybakov [2009]) we know that if $\rho_0, \ldots, \rho_M$ are such that tor any pair $i, j \in \{0, \ldots, M\}$

$$\|\rho_i - \rho_j\|_n \gtrsim n^{-2\beta/(2\beta+r)}, \text{ when } i \neq j, \tag{2.15}$$

then

$$\inf_{\hat{\rho}} \max_{\rho\in\{\rho_0,\ldots,\rho_M\}} P_\rho\left(\|\hat{\rho}-\rho\|_n^2 \geq Cn^{-2\beta/(2\beta+r)}\right) \gtrsim \overline{p}_M.$$

Also, Corollary 2.6 in Tsybakov [2009] states that if $P_0, P_1, \dots, P_M$ satisfy

$$\frac{1}{M+1}\sum_{j=1}^{M} K(P_j, P_0) \le \alpha \log M, \tag{2.16}$$

for some $0 < \alpha < 1$, then

$$\overline{p}_M \ge \frac{\log(M+1) - \log 2}{\log M} - \alpha.$$

Here $K(\cdot,\cdot)$ is the Kullback–Leibler divergence. Hence, if we construct the probability measures $P_0, P_1, \dots, P_M$ corresponding to some soft label functions $\rho_0, \dots, \rho_M$ for which (2.15), (2.16) holds, we will have

$$\inf_{\hat\rho} \sup_{\rho \in \{\rho : \Psi^{-1}(\rho) \in H^\beta(Q)\}} \mathbb{E}_\rho n^{2\beta/(2\beta+r)} \|\hat\rho - \rho\|_n^2 \gtrsim \frac{\log(M+1) - \log 2}{\log M} - \alpha.$$

If $M \to \infty$, as $n \to \infty$, the required result follows.

Let $N = n^{r/(2\beta+r)}$ and let $\psi_0, \dots, \psi_{n-1}$ to be an orthonormal eigenbasis of the graph Laplacian L with respect to the $\|\cdot\|_n$–norm. For $\delta > 0$ and $\theta = (\theta_1, \dots \theta_N) \in \{\pm 1\}^N$ define

$$f_\theta = \delta N^{-(2\beta+r)/2r} \sum_{j=0}^{N-1} \theta_j \psi_j.$$

We will select M vectors of coefficients $\theta^{(j)}$ such that the probability measures corresponding to $\rho_j = \Psi(f_{\theta^{(j)}})$ will satisfy (2.16), where Ψ is the link function.

Observe that for small enough $\delta > 0$ functions f_θ belong to the class $H^\beta(Q)$. Indeed, using the geometry assumption we obtain

$$\begin{aligned}\left\langle f_\theta, (I + (n^{\frac{2}{r}} L)^\beta) f_\theta \right\rangle_n &= \delta^2 N^{-(2\beta+r)/r} \sum_{j=0}^{N-1} (1 + n^{2\beta/r} \lambda_j^\beta) \le \\ &\le \delta^2 N^{-(2\beta+r)/r} \left(N + C_2 i_0^{(2\beta+r)/r} + C_2 \sum_{j=i_o}^{N} j^{2\beta/r} \right) \le K_1 \delta^2\end{aligned}$$

for some constant $K_1 > 0$.

We pick a subset $\{\theta^{(1)}, \dots, \theta^{(M)}\}$ of $\{\pm 1\}^N$ such that for any pair $i, j \in \{1, \dots, M\}$ such that $i \ne j$ the vectors from the subset were sufficiently distant from each other

$$d_h(\theta^{(i)}, \theta^{(j)}) \gtrsim N, \tag{2.17}$$

where $d_h(\theta, \theta') = \sum_{j=0}^{N-1} \mathbb{1}_{\theta_j = \theta'_j}$ is the Hamming distance. By the Varshamov–Gilbert bound (see for example Lemma 2.9 in Tsybakov [2009]) we know that there

exist such a subset $\{\theta^{(1)}, \dots, \theta^{(M)}\}$, and the size M of this subset satisfies

$$M \geq b^N$$

for some $1 < b < 2$. Let $\theta^{(0)} = (0, \dots, 0) \in \mathbb{R}^N$. We define a set of probability measures $\{P_0, \dots, P_M\}$ by setting $P_j = P_{\rho_j}$, where $\rho_j = \Psi(f_{\theta^{(j)}})$.

In order to show that the P_j satisfy (2.16) we present a technical lemma. In the classification setting the Kullback–Leibler divergence $K(\cdot, \cdot)$ satisfies

$$K(P_\rho, P'_\rho) = \sum_{i=1}^{n} \left(\rho(i) \log \frac{\rho(i)}{\rho'(i)} - (1 - \rho(i)) \log \frac{1 - \rho(i)}{1 - \rho'(i)} \right).$$

Lemma 2.5.3. *If $\frac{\Psi'}{\Psi(1-\Psi)}$ is bounded, then there exist $c > 0$ such that for any $v_1, v_2 \in \mathbb{R}^n$ we have*

$$K(P_{\Psi(v_1)}, P_{\Psi(v_2)}) \leq nc||v_1 - v_2||_n^2.$$

Proof. For every $x \in \mathbb{R}$ consider the function $g_x : \mathbb{R} \to \mathbb{R}$ defined as

$$g_x(y) = \Psi(x) \log \frac{\Psi(x)}{\Psi(y)} + (1 - \Psi(x)) \log \frac{1 - \Psi(x)}{1 - \Psi(y)}.$$

We see that $g'_x(y) = \frac{\Psi'(y)}{\Psi(y)(1-\Psi(y))}(\Psi(y) - \Psi(x))$. Then by Taylor's theorem we can see that

$$|g_x(y)| \leq \sup_{v \in [x,y] \cup [y,x]} \left| \frac{\Psi'(v)}{\Psi(v)(1 - \Psi(v))} \right| \sup_{v \in [x,y] \cup [y,x]} |\Psi'(v)|(x - y)^2.$$

The statement of the lemma follows. ∎

By Lemma 2.5.3, we obtain for some $K_2 > 0$

$$K(P_j, P_0) \leq K_2 n \|f_{\theta^{(j)}} - 0\|_n^2 = 4K_2 \delta^2 n N^{-2\beta/r},$$

since

$$\|f_\theta - f_{\theta'}\|_n^2 = 4\delta^2 N^{-(2\beta+r)/r} d_h(\theta, \theta'). \tag{2.18}$$

Observe that this bound does not depend on j. Hence,

$$\frac{1}{M+1} \sum_{j=1}^{M} K(P_j, P_0) \leq K_2 \delta^2 n N^{-2\beta/r} = K_2 \delta^2 \log M.$$

We can choose $\delta > 0$ to be small enough such that the condition (2.16) is satisfied with some $0 < \alpha < 1$.

To finish the proof of the theorem we need to show that $\rho_0, \dots, \rho_M$ satisfy (2.15).

From (2.17) and (2.18) we get

$$\|f_{\theta^{(i)}} - f_{\theta^{(j)}}\|_n \gtrsim n^{-\beta/(2\beta+r)}.$$

Moreover, by the assumption of the theorem we have for any $j = 1, \dots, M$

$$\max_{i=1,\dots,n} |f_{\theta^{(j)}}(i)| \lesssim N^{1-(2\beta+r)/2r} \max_{i=1,\dots,n} |\psi_j(i)| \lesssim N^{(r-2\beta)/2r}.$$

For $\beta \geq r/2$ the norm is then bounded by some constant, which does not depend on n or j. Hence, there exists $K_3 \geq 0$ such that for every $i = 1, \dots, n$ and every $j = 1, \dots, M$

$$|f_{\theta^{(j)}}(i)| \leq K_3.$$

Observe that since $\Psi'(x) \neq 0$ for any $x \in \mathbb{R}$, there exists $K_4 > 0$ such that for any $x, y \in [-K_3, K_3]$

$$|\Psi(x) - \Psi(y)| \geq K_4|x - y|.$$

Thus, for any pair $i, j \in \{1, \dots, M\}$ such that $i \neq j$

$$\|\rho_i - \rho_j\|_n = \|\Psi(f_{\theta^{(i)}}) - \Psi(f_{\theta^{(i)}})\|_n \gtrsim n^{-\beta/(2\beta+r)}.$$

This completes the proof of the theorem.