Asymptotic results in nonparametric Bayesian function estimation

Kirichenko, A.

Citation for published version (APA):
Chapter 3

Bayesian function estimation on large graphs

3.1 Introduction

In the previous chapter we have discussed the general mathematical framework that allows us to study the performance of nonparametric function estimation methods on large graphs. We have also obtained minimax rates for two prototypical problems on graphs: the regression and the binary classification. In this chapter we continue studying the same setting and introduce a Bayesian approach to these problems. We develop Bayesian procedures that involve Laplacian regularisation and show that these procedures turn out to be (almost) optimal in a minimax sense.

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], Ando and Zhang [2007], Zhu et al. [2003], Huang et al. [2011]). The graph Laplacian \( L = D - A \), 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, acts on a function \( f \) on the graph as

\[
L f(i) = \sum_{j \sim i} (f(i) - f(j)),
\]

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 normalised Laplacian \( \tilde{L} = D^{-1/2} LD^{-1/2} \). We will continue to work with \( L \), but much of the story goes through if \( L \) is replaced by such a related operator, after minor adaptations.

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 penalisation using this norm, as well as generalisations involving
powers of the Laplacian or other functions, for instance, exponential functions. 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] and Ando and Zhang [2007] consider sub-sampling schemes for estimating a function on a graph.

However, 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.

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 the same assumption on the geometry of the graph and the smoothness of the true function in terms of Sobolev-type balls as in the previous chapter. We show that the proposed methods appear to be (almost) optimal in a minimax sense and adaptive to the regularity of the true function.

Additionally, we consider the case of missing observations, where only a part of observations is available. In order to tackle the problem we assume that the available data is somehow uniformly distributed on the graph. We study two missing mechanisms: coin flipping and sampling with replacement. In the coin flipping mechanism for every vertex we independently flip a coin and see the observation with probability \( p_n \). In the sampling with replacement we sample \( m_n = np_n \) points from the complete data set, allowing repetition of the vertices. We are interested in the case where the available sample size constitutes a small part of the whole set, so we allow \( p_n \to 0 \). Although we only consider part of the data points, we use a distance on functions that takes all \( n \) points into account, i.e. we study a form of generalisation performance.

The remainder of the chapter is organised as follows. In Section 3.2 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. In Section 3.3 these results are used to derive theorems about posterior contraction for the nonparametric regression and the binary classification problems on graphs. In Section 3.4 we treat the missing observation case, introducing the modified families of priors and deriving
posterior contraction rates in the same setting of nonparametric regression and binary classification. Mathematical proofs are given in Section 3.5.

3.2 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. In the next section we 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 (2.2) holds for some \( r \geq 1 \) and the regularity of the function is quantified by Sobolev-type balls defined in (2.4). The first family of priors we consider penalise 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 of the 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 penalised. 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 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. Observe that in the following we denote by \( N(\varepsilon, B, \| \cdot \|) \) the minimal number of balls of \( \| \cdot \| \)-radius \( \varepsilon \) needed to cover \( B \) for some \( \varepsilon > 0 \) and a norm \( \| \cdot \| \) on a set \( B \).

**Theorem 3.2.1** (Power of the Laplacian). Suppose the geometry assumption holds for \( r \geq 1 \). Let \( \alpha > 0 \) be fixed and assume that \( f_0 \in H^\beta(Q) \) for some \( Q > 0 \) and \( 0 < \beta \leq \alpha + r/2 \). Let the random function \( f \) on the graph be defined by

\[
\begin{align*}
c &\sim \text{Exp}(1) \quad (3.2) \\
f | c &\sim \mathcal{N}(0, ((n/c)^{2/r}(L + n^{-2}I)^{\alpha + r/2})^{-1}) \quad (3.3)
\end{align*}
\]

Then there exists a constant \( K_1 > 0 \) and for all \( K_2 > 1 \) there exist Borel measurable
3. Bayesian function estimation on large graphs

subsets $B_n$ of $\mathbb{R}^n$ such that for every sufficiently large $n$

\begin{align}
  \mathbb{P}(\|f - f_0\|_n < \varepsilon_n) &\geq e^{-K_1 n\varepsilon_n^2}, \\
  \mathbb{P}(f \notin B_n) &\leq e^{-K_2 n\varepsilon_n^2}, \\
  \log N(\varepsilon_n, B_n, \|\cdot\|_n) &\leq n\varepsilon_n^2,
\end{align}

where $\varepsilon_n$ 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 $\beta$ in the range $(0, \alpha + r/2]$. In the statistical results presented in Section 3.3 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 $\alpha$. 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 generalisation 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”.

**Theorem 3.2.2** (Exponential of the Laplacian). Suppose the geometry assumption holds for $r \geq 1$. Assume that $f_0 \in H^{\beta}(Q)$ for some $Q > 0$ and $\beta > 0$. Let the random function $f$ on the graph be defined by

\begin{align}
  c &\sim \text{Exp}(1) \\
  f | c &\sim N(0, ne^{-(n/c)^{2/r}L}).
\end{align}

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 $n$

\begin{align}
  \mathbb{P}(\|f - f_0\|_n < \varepsilon_n) &\geq e^{-K_1 n\varepsilon_n^2}, \\
  \mathbb{P}(f \notin B_n) &\leq e^{-K_2 n\varepsilon_n^2}, \\
  \log N(\tilde{\varepsilon}_n, B_n, \|\cdot\|_n) &\leq n\tilde{\varepsilon}_n^2,
\end{align}

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

3.3 Function estimation on graphs

In this section we translate the general Theorems 3.2.1 and 3.2.2 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.

### 3.3.1 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

$$
Y_i = f_0(i) + \sigma \xi_i,
$$

where $f_0$ is the function on $G$ that we want to make inference about, $\xi_i$ are independent standard Gaussian, and $\sigma > 0$. 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 $\sigma$ 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 $\Pi$, 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 $\varepsilon_n$ if for all large enough $M > 0$,

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

as $n \to \infty$. Here the convergence is in probability under the law $P_{f_0}$ corresponding to the data generating model (3.12).

The usual arguments allow us to derive the following statements from Theorems 3.2.1 and 3.2.2. See, for instance, van der Vaart and van Zanten [2008a] or de Jonge and van Zanten [2013] for details.

**Theorem 3.3.1 (Nonparametric regression).** Suppose the geometry assumption holds for $r \geq 1$. Assume that $f_0 \in H^\beta(Q)$ for $\beta, Q > 0$.

(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 $\beta$ 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 $\alpha$ is the hyperparameter 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.
3.3.2 Nonparametric classification

We can derive the analogous results in the classification problem in which 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 \( \rho_0 \), or “soft label function” on the graph, given by

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

We consider priors on \( \rho \) 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 \( \rho = \Psi(f) \), where \( \Psi : \mathbb{R} \to (0, 1) \) is a suitably chosen link function. We again assume that \( \Psi \) is a strictly increasing, differentiable function onto \( (0, 1) \) such that \( \Psi'(\Psi(1 - \Psi)) \) is uniformly bounded. Also in this case we denote the posterior corresponding to a prior \( \Pi \) by \( \Pi(\cdot \mid Y_1, \ldots, Y_n) \) and we say that the posterior contracts around \( \rho_0 \) at the rate \( \varepsilon_n \) if for all large enough \( M > 0 \),

\[
\frac{\Pi(\rho : \|\rho - \rho_0\|_n \geq M\varepsilon_n \mid Y_1, \ldots, Y_n)}{\rho_0} \to 0
\]

as \( n \to \infty \).

To derive the following result from Theorems 3.2.1 and 3.2.2 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 3.3.2 (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 \( \Psi^{-1}(\rho_0) \in H^\beta(Q) \) for \( \beta, Q > 0 \).

(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 \( \rho_0 \) at the rate \( n^{-\beta/(r+2\beta)} \).

(ii) (Exponential of the Laplacian.) If the prior on \( \rho \) is given by the law of \( \Psi(f) \), for \( f \) given by (3.7)–(3.8), then the posterior contracts around \( \rho_0 \) at the rate \( n^{-\beta/(r+2\beta)} \log^\kappa n \) for some \( \kappa > 0 \).

3.4 Function estimation on graphs with missing observations

The previous section covered function estimation on large graphs in the full observation setting. However, it is often an intermediate step in function recovery, since a lot of practical applications involve making inference about a function in the situation when only partial data is available. As an example, consider the classification problem on the protein interaction graph in Figure 3.1 studied in
3.4. Function estimation on graphs with missing observations

Figure 3.1: Protein-protein interaction graph. The red vertices in the graph are involved in ICSC, the blue vertices are not involved, and the label of the white vertices is not known.

Hartog and van Zanten [2016]. A vertex of the graph is labelled red if the corresponding protein is involved in so-called intracellular signalling cascades (ICSC), which is a specific form of communication, and blue otherwise. The goal is to recover the missing labels marked white on this graph. Despite the fact that some observations are missing, we would still like to recover function as good as possible with respect to the $\| \cdot \|_n$-norm that takes the values at all the vertices into account. The convergence rates obtained for this case can be then viewed as a quantitative characteristic of generalisation performance. We again assume that the underlying graph satisfies the geometry assumption for some $r \geq 1$ and that the target function has a smoothness level $\beta > 0$ in terms of the Sobolev type balls, defined in Sections 2.2.2 and 2.3.

Certainly, an estimator on the graph will not be able to recover the target function based on incomplete data set as efficiently as in the full observation case. In order to study how having partial data influences the performance of estimators we again take an asymptotic approach and investigate how the rates of convergences are affected by missing data. Suppose we only observe a small proportion of the full data sample, meaning that we see (on average) $np_n$ points out of a complete sample of $n$ points with $p_n$ potentially going to zero as $n \to \infty$. Recall that in the full observation case the minimax rates for regression and binary classification on large graphs were equal to $n^{-\beta/(2\beta+r)}$. It would be natural to expect that in the missing observation case the rates will be transformed into $(np_n)^{-\beta/(2\beta+r)}$, since we only see $np_n$ points. In this chapter we show that the contraction rates for the modifications of the Bayesian estimators developed in Section 3.2 are equal to the aforementioned value in the missing observation case. The minimax rates are not known for the present situation. However, a particular case of our problem with...
3. Bayesian function estimation on large graphs

$r = 1$ can be linked to the regression setting on the interval $[0,1]$ with missing observations, studied in Efroimovich [2011]. As shown in the aforementioned paper, the minimax rate for this problem is $(np_n)^{-\beta/(2\beta+1)}$. Hence, it appears that the rates obtained in this section cannot be fundamentally improved.

The way in which the missing observations are distributed throughout the graph plays a crucial role in estimation. We do not treat all the possible cases but rather consider a prototypical problem that satisfies the following two assumptions. First, we assume that the missingness does not depend on the values of variables in the data (we consider missing completely at random (MCAR) mechanisms, the general classification of types of missing mechanisms is given in Little and Rubin [2002]). We also assume that the available data is somehow uniformly distributed on the graph. We implement these ideas in two missing mechanisms: a coin flipping scheme, where we see the observation at a certain vertex independently of the others with some probability $0 < p_n < 1$, and sampling with replacement of $p_n n$ data points.

In the next section we discuss the missing mechanisms that we study in more detail. Then we present general theorems about the convergence rates of posterior distribution in the regression and binary classification settings with missing observations. The derived theorems are similar and based on the works Ghosal et al. [2000], Ghosal and van der Vaart [2007], and Ghosal and van der Vaart [2001]. However, the missing observation setting requires to control the prior mass and the entropy of balls of larger diameter compared to the full observation case. We introduce a modification of the power prior defined by (3.2)–(3.3) and we consider the exponential prior defined by (3.7)–(3.8) in the missing observation setting. For those priors we derive convergence rates of posterior distributions for regression and binary classification problems on the graph and show that the procedures remain rate-adaptive.

3.4.1 Missing mechanisms

We introduce the missing mechanisms for the problem of recovering a function $g_0$ defined on the vertices of the graph $G$. In the regression setting $g_0$ is equal to the regression function $f_0$, while in the classification setting the function of interest a soft label function $\rho_0$. Consider a full observation set $(Y_1, \ldots, Y_n)$ on the vertices $1, \ldots, n$ of the graph $G$.

For missing at random by coin flipping mechanism we consider $V_i$ to be independent Bernoulli random variables with success probability $p_n > 0$. Next, we define

$$Z_i = \begin{cases} \star, & \text{if } V_i = 0; \\ Y_i, & \text{if } V_i = 1. \end{cases} \quad (3.13)$$

The goal is to estimate $g_0$ from the observations $Z_1, \ldots, Z_n$ as accurately as possible with respect to the $\| \cdot \|_n$ norm

$$\| g_1 - g_2 \|_n^2 = \frac{1}{n} \sum_{i=1}^n (g_1(i) - g_2(i))^2.$$
3.4. Function estimation on graphs with missing observations

For a given prior $\Pi$, the corresponding posterior distribution is denoted by $\Pi(\cdot | Z_1, \ldots, Z_n)$ and we say that the posterior contracts at rate $\varepsilon_n > 0$ if, for every sufficiently large constant $M$, an $n \to \infty$

$$\Pi(g : \|g - g_0\|_n \geq M\varepsilon_n | Z_1, \ldots, Z_n) \xrightarrow{P_{g_0}} 0.$$ 

Here the convergence is in probability under the law $P_{g_0}$ corresponding to the data generating model (3.13).

For the sampling with replacement missing mechanism we only see $m_n = p_n n$ pairs of observations $(X_i, Z_i)$, where

$$X_i \sim \text{Uniform}\{1, \ldots, n\},$$

$$Z_i | X_i \sim Y_{X_i}.$$ (3.14) (3.15)

The goal is again to estimate $g_0$ with as accurately as possible with respect to the $\| \cdot \|_n$-norm, having only $m_n$ pairs of observations available.

For a given prior $\Pi$, the corresponding posterior distribution in this case is denoted by $\Pi(\cdot | (X_1, Z_1), \ldots, (X_{m_n}, Z_{m_n}))$ and we say that the posterior contracts at rate $\varepsilon_n > 0$ if, for every sufficiently large constant $M$, and $n \to \infty$

$$\Pi(g : \|g - g_0\|_n \geq M\varepsilon_n | (X_1, Z_1), \ldots, (X_{m_n}, Z_{m_n})) \xrightarrow{P_{g_0}} 0.$$ 

In the regression problem $g_0 = f_0$ and the observations $Y_i$ satisfy standard regression relation

$$Y_i = f_0(i) + \sigma \xi_i,$$

where $\sigma > 0$ and $\xi_i$ are independent standard Gaussian.

In the binary classification problem we independently observe values $Y_i \in \{0, 1\}$ at the vertices of the graph, and the goal is to estimate the binary regression function $g_0 = \rho_0$, or "soft label function", given by

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

For the classification problem we only consider priors on $\rho$ constructed by first defining a prior on a real-valued function $f$ and then setting $\rho = \Psi(f)$, where $\Psi : \mathbb{R} \to (0, 1)$ is as usual a strictly increasing, differentiable function onto $(0,1)$ such that $\Psi'/\Psi(1 - \Psi)$ is uniformly bounded. Additionally, we denote $f_0$ to be equal to $\Psi^{-1}(\rho_0)$.

3.4.2 General theorems for regression and classification

General results on the convergence rates of posterior distribution, such as Theorem 1.5.2, cannot be directly applied in the case of missing observations. In this section we provide analogous results that are based on the aforementioned theorem with slight modifications allowing us to accommodate the missing data in the case of
regression and classification. We show that in order to derive a rate of convergence the prior mass and the entropy should be controlled for balls of larger diameter than for the same expressions in the full observation case. Observe that these requirements turn out to be more demanding than in the original theorem.

**Theorem 3.4.1.** Consider the sequences $\varepsilon_n, \varepsilon^*_n, \tilde{\varepsilon}_n \to 0$, such that $n\varepsilon^2_n \to \infty$, $\varepsilon^*_n \approx p_n^{-1/2}\varepsilon_n$, $\tilde{\varepsilon}_n \approx \varepsilon_n(\log n)^k$ for $k \geq 0$. Additionally suppose that for some constant $K > 0$ and for all $L > 1$ there exist Borel measurable subsets $B_n$ of $\mathbb{R}^n$ for which the following is true, when $n$ is large enough

$$\Pi(f : \|f - f_0\|_n < \varepsilon^*_n) \geq e^{-Kn\varepsilon^2_n}, \quad (3.16)$$

$$\Pi(f \notin B_n) \leq e^{-Ln\varepsilon^2_n}, \quad (3.17)$$

$$\log N(\varepsilon^*_n, B_n, \|\cdot\|_n) \leq n\tilde{\varepsilon}^2_n. \quad (3.18)$$

Then the posterior contracts at the rate $\varepsilon^*_n(\log n)^k$ for both regression and classification problems and for both missing mechanisms.

The theorem can be proved using general Theorems 1.5.1, 1.5.2. However, the direct application of the theorems is not possible, so we translate the given setting in order for it to match the conditions of the aforementioned theorems. Then we translate the obtained results back to the original setting of the theorem.

### 3.4.3 General results on prior concentration and convergence rates for regression and classification

We again consider two families of priors: the modified power prior and the exponential prior. We present theorems analogous to Theorems 3.2.1, 3.2.2, providing lower bounds on 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. We show that the developed priors satisfy the new, more strict conditions, which we require for obtaining convergence rates in the missing observations setting. In the next section we combine these results with the general theorems from Section 3.4.2 to deduce convergence rates and adaptation for nonparametric regression and classification problems on graphs with missing observations.

Consider a connected undirected simple graph $G$ with vertices labeled $\{1, \ldots, n\}$ that satisfies the geometry assumption with $r \geq 1$.

**Theorem 3.4.2.** Suppose the geometry assumption holds for $r \geq 1$. Let $\alpha > 0$ be fixed and assume that $f_0 \in H^\beta(Q)$ for some $Q > 0$ and $0 < \beta \leq \alpha + r/2$. Let the random function $f$ on the graph be defined by

$$c \sim \text{Exp}(p_n^{r/(2\alpha+r)}) \quad (3.19)$$

$$f | c \sim N(0, (((n/c)^2/(L + n^{-2}I))^{\alpha+r/2})^{-1}). \quad (3.20)$$
3.4. Function estimation on graphs with missing observations

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 $n$,

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

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

$$\log N(\varepsilon_n^*, B_n, \|\cdot\|_n) \leq n \varepsilon_n^2,$$

where $\varepsilon_n \asymp p_n^{r/2(2\beta+r)} n^{-\beta/(2\beta+r)}$ and $\varepsilon_n^* \asymp (np_n)^{-\beta/(2\beta+r)}$.

Observe that the scaling parameter in the power prior is distributed according to $\text{Exp}(p_n^{r/(2\alpha + r)})$, in contrast to the full observation case, where it had a standard exponential distribution. This increases the variance of the prior which is necessary to compensate for the uncertainty arising from the missing data.

Interestingly, as shown in the following theorem, the exponential prior doesn’t require any modifications in order to accommodate missing data.

**Theorem 3.4.3.** Suppose the geometry assumption holds for $r \geq 1$. Assume that $f_0 \in H^\beta(Q)$ for some $Q > 0$ and $\beta > 0$. Let the random function $f$ on the graph be defined by

$$c \sim \text{Exp}(1) \quad (3.24)$$

$$f \mid c \sim N(0, ne^{-\frac{(n/c)^{2/r} L}}) \quad (3.25)$$

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 $n$,

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

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

$$\log N(\varepsilon_n^*, B_n, \|\cdot\|_n) \leq n \varepsilon_n^2,$$

where $\varepsilon_n \asymp p_n^{r/2(2\beta+r)} n^{-\beta/(2\beta+r)} (\log n)^{(1+r/2)/(2\beta+r)}$, $\varepsilon_n^* \asymp p_n^{-1/2} \varepsilon_n$, and $\tilde{\varepsilon}_n \asymp \varepsilon_n (\log n)^k$ for some $k \geq 0$.

### 3.4.4 Bayesian procedure for function estimation on graphs with missing observations

In the regression case we assume the full observation set $Y = (Y_1, \ldots, Y_n)$ on the vertices of the graph to be such that

$$Y_i = f_0(i) + \sigma \xi_i,$$

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

We assume that we effectively see (on average) only $np_n$ data points uniformly distributed on the graph by one of the missing mechanisms described in Section 53.
3. Bayesian function estimation on large graphs

3.4.1. As prior on the regression function $f$ we then employ one of the two priors described by (3.19)–(3.20) or (3.24)–(3.25). If $\sigma$ 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 $f$.

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 $\rho_0$, or “soft label function” on the graph, given by

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

We again assume to see only partial data according to one of the missing mechanisms described in Section 3.4.1. We consider a suitably chosen link function $\Psi : \mathbb{R} \to (0,1)$. We assume that $\Psi$ is a strictly increasing, differentiable function onto $(0,1)$ such that $\Psi'/(\Psi(1 - \Psi))$ is uniformly bounded. We consider priors on $\rho$ constructed by first defining a prior on a real-valued function $f$ by (3.19)–(3.20) or (3.24)–(3.25) and then setting $\rho = \Psi(f)$, where $\Psi : \mathbb{R} \to (0,1)$ is a suitably chosen link function.

Combining the results of the theorems stated in this section we derive the following result.

**Theorem 3.4.4.** Suppose the geometry assumption holds for $r \geq 1$. Let $\alpha > 0$ be fixed and assume that in the regression problem $f_0 \in H^\beta(Q)$ for some $Q > 0$ and $\beta > 0$. For the classification problem we assume $\Psi^{-1}(\rho_0) \in H^\beta(Q)$ for some $Q > 0$ and $\beta > 0$. Then for both problems on the graph

(i) For the power priors based on (3.19)–(3.20) the posterior contracts around the truth at the rate $(np_n)^{-\beta/(2\beta+r)}$ for both missing mechanisms under the additional restriction $\beta \leq \alpha + r/2$.

(ii) For the exponential priors based on (3.24)–(3.25) the posterior contracts around the truth at the rate $(np_n)^{-\beta/(2\beta+r)}(\log n)^{k'}$ for both missing mechanisms for some $k' > 0$ and for all $\beta > 0$.

Since the developed priors do not use knowledge of the regularity $\beta$ of the true function, we again obtained rate-adaptive results with the same remarks about the log factor as in the full observation case. For the power prior the range of regularities that we can adapt to is bounded by $\alpha + r/2$, where $\alpha$ is the hyperparameter 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. Lower bounds for the rates of estimation are not known for the present situation. However, a particular case of regression on a path graph can be related to the case of the regression problem on the $[0,1]$ interval with missing data, studied in Efromovich [2011]. The minimax rate obtained in that paper is equal to $(np_n)^{-\beta/(2\beta+1)}$, which coincides with our rates for $r = 1$. Thus, it appears that the rates cannot be fundamentally improved. However, a detailed investigation of lower bounds in this setting is necessary to obtain proper insight into this matter.
3.5 Proofs

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 centred $n$-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

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

where $\Sigma_c$ 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 $\psi_i$ are the eigenfunctions of $L$, normalised with respect to the norm $\| \cdot \|_n$. They are then also eigenfunctions of $\Sigma^{-1}_c$ in both cases. We denote the corresponding eigenvalues by $\mu_i$.

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

$$\sum \zeta_i \psi_i / \sqrt{n \mu_i},$$

where $\zeta_1, \ldots, \zeta_n$ 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

$$\| h \|_n^2 = \sum_{i=0}^{n-1} h_i^2, \quad \| h \|_{\mathbb{H}^c}^2 = n \sum_{i=0}^{n-1} \mu_i h_i^2. \quad (3.31)$$

We denote the unit ball of the RKHS by $\mathbb{H}^c_1 = \{ h \in \mathbb{H}^c : \| h \|_{\mathbb{H}^c} \leq 1 \}$.

3.5.1 Proof of Theorem 3.2.1

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

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

3.5.1.1 Proof of (3.4)

By Lemma 5.3 of van der Vaart and van Zanten [2008b], it follows from Lemmas 3.5.1 and 3.5.2 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

$$- \log \mathbb{P}(\| f - f_0 \|_n \mid c) \lesssim \varepsilon_n^{-r/\beta}.$$
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 3.5.1.** For \(n\) large enough and \(\varepsilon > 0\) and \(\varepsilon\sqrt{n}/c^{(\alpha + r/2)/r}\) small enough,

\[
- \log P(\|f\|_n \leq \varepsilon \mid c) \lesssim \left(\frac{c^{(\alpha + r/2)/r}}{\varepsilon} \right) \frac{\varepsilon}{n}.
\]

**Proof.** By the series representation (3.30) we have

\[
P(\|f\|_n \leq \varepsilon \mid c) = P\left(\sum_{i \leq i_0} \frac{\zeta_i^2}{n\mu_i} \leq \varepsilon^2\right).
\]

Recall from Section 2.2.2 that we can assume without loss of generality that we have the lower bounds

\[
\lambda_i \geq C_1 \left(\frac{i}{n}\right)^2, \quad 1 \leq i \leq i_0, \quad (3.33)
\]

\[
\lambda_i \geq C_1 \left(\frac{i}{n}\right)^{2/r}, \quad i > i_0.
\]

These translate into lower bounds for the \(\mu_i\) from which it follows that for \(\varepsilon > 0\),

\[
P(\|f\|_n^2 \leq 2\varepsilon^2 \mid c) \geq P\left(\sum_{i \leq i_0} \frac{\zeta_i^2}{n\mu_i} \leq \varepsilon^2, \sum_{i > i_0} \frac{\zeta_i^2}{n\mu_i} \leq \varepsilon^2\right)
\]

\[
\geq P\left(\sum_{1 < i \leq i_0} \zeta_i^2 \leq (C_1^p e^{-2p/r} n^{(2\alpha + 2r - 2pr)/r})\varepsilon^2\right) P\left(\sum_{i > i_0} \frac{\zeta_i^2}{n\mu_i} \leq C_1^p e^{-2p/r} n^2\varepsilon^2\right),
\]

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

\[
\exp\left(-\text{const} \times (c^{-p/r}\varepsilon\sqrt{n})^{-r/\alpha}\right),
\]

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

\[
\left(P(|\zeta_1| \leq i_0^{1/2} C_1^{p/2} e^{-p/r} n^{(\alpha + r - pr)/r} \varepsilon}\right)^{i_0}.
\]

Since \(r \geq 1\), we have \(e^{-p/r} n^{(\alpha + r - pr)/r} \varepsilon = O(e^{-p/r} \varepsilon\sqrt{n})\). Hence, for \(c^{-p/r}\varepsilon\sqrt{n}\) small
3.5. Proofs

enough the probability is further bounded from below by

\[ \text{const} \times \left( c^{-p/r} n^{(\alpha + r - pr)/r} \right)^{i_0}. \]

Combining the bounds for the separate factors we find that, for \( c - p/r \varepsilon \) small enough,

\[ -\log P(\|f\|_n^2 \leq 2 \varepsilon^2 / c) \lesssim \log \left( \frac{c^{p/r}}{n^{(\alpha + r - pr)/r}} \right) + \left( \frac{c^{p/r}}{\varepsilon \sqrt{n}} \right)^{r/\alpha}. \]

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 3.5.2. Let \( f \in H^\beta(Q) \) 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,

\[ \inf_{h \in H^c} \|h - f\|_n \leq \varepsilon \|h\|_{H^c} \lesssim n c^{-2(\alpha + r)/r} \varepsilon^{-2(\alpha - \beta)/r}. \]  

(3.35)

Proof. We use the expansion \( f = \sum f_i \psi_i \), with \( \psi_i \) the orthonormal eigenfunctions of the Laplacian. In terms of the coefficients the smoothness assumption reads \( \sum (1 + n^{2\beta/r} \chi_i^\beta) f_i^2 \leq Q^2 \). Now for \( I \) to be determined below, consider \( h = \sum_{i \leq I} f_i \psi_i \).

In view of (3.33)–(3.34) we have, for \( I \) large enough,

\[ \|h - f\|_n^2 = \sum_{i > I} f_i^2 \leq \frac{Q^2}{1 + n^{2\beta/r} \chi_i^\beta} \leq Q^2 c^{-2\beta/r} I^{-2\beta/r}. \]

Setting \( I = \text{const} \times \varepsilon^{-r/\beta} \) we get \( \|h - f\|_n \leq \varepsilon \). By (3.31), the RKHS-norm of \( h \) satisfies

\[ \|h\|_{H^c}^2 = n \sum_{i \leq I} ((n/c)^{2/r} (\lambda_i + n^{-2}))^{\alpha + r/2} f_i^2 \leq n c^{-2p/r} Q^2 + c^{-2p/r} Q^2 n^{2 + 2(\alpha - \beta)/r} \chi_i^{\beta - \beta}. \]

The condition on \( \varepsilon \) ensures that for the choice of \( I \) made above and \( n \) large enough, \( i_0 \leq I \leq \kappa n \). Hence, by (3.33)–(3.34), \( \|h\|_{H^c}^2 \) is bounded by a constant times the right-hand side of (3.35).

3.5.1.2 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 3.5.3 we have

\[ \log N(2\varepsilon_n, B_n, \|\cdot\|_n) \leq \log N(\varepsilon_n/M_n, \mathbb{H}_1^{c_n}, \|\cdot\|_n) \lesssim c_n \left( \frac{M_n}{\varepsilon_n \sqrt{n}} \right)^{\alpha/2}. \]
where $p = \alpha + r/2$ again. For $M_n = M \sqrt{n \varepsilon^2_n}$ and $c_n^{p/r} = N \sqrt{n \varepsilon_n^{(\beta-\alpha)/\beta}}$ this is bounded by a constant times $n \varepsilon^2_n$, 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

$$\Pr(f \notin B_n) \leq \int_{0}^{c_n} \Pr(f \notin M_n \mathbb{H}^c_1 + \varepsilon_n \mathbb{B}_1 \mid c) e^{-c} \, dc + \int_{c_n}^{\infty} e^{-c} \, dc.$$  

For $c \leq c_n$ we have the inclusion $\mathbb{H}^c_1 \subseteq \mathbb{H}^c_1$. Hence, by the Borell–Sudakov inequality, we have for $c \leq c_n$ that

$$\Pr(f \notin B_n \mid c) \leq \Pr(f \notin M_n \mathbb{H}^c_1 + \varepsilon_n \mathbb{B}_1 \mid c)$$

$$\leq 1 - \Phi(\Phi^{-1}(\Pr(\|f\|_n \leq \varepsilon_n \mid c) + M_n))$$

$$\leq 1 - \Phi(\Phi^{-1}(\Pr(\|f\|_n \leq \varepsilon_n \mid c_n) + M_n)),$$

where $\Phi$ is the distribution function of the standard normal distribution. By Lemma 3.5.1 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 $\Phi^{-1}(y) \geq -((5/2) \log(1/y))^{1/2}$ for $y \in (0, 1/2)$, it follows that for $c \leq c_n$, $\Pr(f \notin B_n \mid c) \leq 1 - \Phi(M_n - K' \varepsilon_n^{-r/(2\beta)})$ for some $K' > 0$. For $M_n$ a large enough multiple of $\varepsilon_n^{-r/(2\beta)}$, this probability is bounded by $\exp(-K_2 \varepsilon_n^{-r/\beta}) = \exp(-K_2 n \varepsilon^2_n)$.

**Lemma 3.5.3.** For $n$ large enough and $c, \varepsilon > 0$ we have

$$\log N(\varepsilon, \mathbb{H}^c_1, \| \cdot \|) \lesssim c \left( \frac{1}{\varepsilon \sqrt{n}} \right)^{\alpha + r/2}. \quad (3.36)$$

**Proof.** By expanding the RKHS elements in the eigenbasis of the Laplacian and taking into account the relations (3.31) we see that the problem is to bound the entropy $\log N(\varepsilon, A, \| \cdot \|)$, where

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

Using the bounds (3.33)–(3.34), it follows that with $p = \alpha + r/2$ and $R = c^{p/r} n^{-(\alpha + r)/r}$ we have the inclusions

$$A \subset \left\{ x \in \mathbb{R}^n : \sum_{i=0}^{n-1} \lambda_i^p x_i^2 \leq R^2 \right\}$$

$$\subset \left\{ x \in \mathbb{R}^n : \sum_{i \leq i_0} x_i^2 \leq C_1^{-p} n^{2p} R^2, \quad \sum_{i > i_0} x_i^{2p/r} \leq C_1^{-p} n^{2p/r} R^2 \right\}.$$  

58
By using the well-known entropy bounds for balls in $\mathbb{R}^{i_0}$ and ellipsoids in $\ell^2$ we deduce from this that for $\varepsilon > 0$,

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

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

### 3.5.2 Proof of Theorem 3.2.2

In this case the eigenvalues of $\Sigma_c^{-1}$ are given by

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

#### 3.5.2.1 Proof of (3.9)

By Lemma 5.3 of van der Vaart and van Zanten [2008b], it follows from Lemmas 3.5.4 and 3.5.5 ahead that under the conditions of the theorem and for $\varepsilon = \varepsilon_n = (n/\log^{1+r/2} n)^{-\beta/(r+2\beta)}$ and $ne^{2}/\log^{1+r/2} n \leq c \leq 2ne^{2}/\log^{1+r/2} n$, we have

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

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

**Lemma 3.5.4.** 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 $P(\|f\|_n \leq \varepsilon | c) = \mathbb{P}(\sum e^{-(n/c)2/r} \lambda_i \zeta_i^2 \leq \varepsilon^2)$, where the $\zeta_i$ are independent standard normal random variables. By the lower bounds (3.33)–(3.34), it follows that

$$P(\|f\|_n \leq 2\varepsilon | c) \
\geq \mathbb{P}\left(\sum_{i \leq i_0} e^{-C_1 n^{(2-2r)/r} \varepsilon^{-2/r}} \zeta_i^2 \leq \varepsilon^2\right) \mathbb{P}\left(\sum_{i \geq 1} e^{-C_1 c^{-2/r} c^{-2/\beta} \varepsilon^{-2/r}} \zeta_i^2 \leq \varepsilon^2\right).$$

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

$$\left(\mathbb{P}(\zeta_1 < i_0^{-1/2} e^{(1/2)C_1 n^{(2-2r)/r} \varepsilon^{-2/r}})\right)^{i_0}.$$  

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

59
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 $\varepsilon \to 0$

$$P\left(\sum a_i \xi_i^2 \leq \varepsilon^2\right) \approx \frac{1}{\sqrt{4\pi}} \frac{e^{-\varepsilon^2\gamma_a - (1/2) \sum \log(1 + 2a_i \gamma_a)}}{\sum (1 + 2a_i \gamma_a)^{\gamma_a}},$$

(3.37)

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

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

(3.38)

We apply (3.37) with $a_i = \exp(-C_1(i/c)^{2/r})$.

We first determine bounds for $\gamma_a$. Note that in our case the terms in the sum $S$ on the right of (3.38) 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

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

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

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

as $\gamma_a \to \infty$. Hence for large $\gamma_a$, we have the upper bound $S \leq \text{const} \times c\gamma_a^{-1} \log^{r/2} \gamma_a$.

It is easily seen that we have a lower bound of the same order, so that

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

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

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

Next we consider the sums appearing on the right of (3.37). To bound $\sum \log(1 + 2a_i \gamma_a)$ 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 $a_{m+1} \gamma_a = a_I^{2m+1} \gamma_a = \gamma_a^{1-m / r^2}$. We first split up the sum, writing

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

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 \( I \).

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

Hence, we have \( \sum \log(1 + 2a_i \gamma_a) \lesssim c \log^{1+r/2} \gamma_a \). For the other sum appearing in (3.37) we have

\[
\sum \left( \frac{2a_i \gamma_a}{1 + 2a_i \gamma_a} \right)^2 \leq \sum \frac{2a_i \gamma_a}{1 + 2a_i \gamma_a} = 2\gamma_a \varepsilon^2.
\]

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

Lemma 3.5.5. Suppose that \( f \in H^\beta(Q) \) for some \( \beta, Q > 0 \). For \( \varepsilon > 0 \) such that \( \varepsilon \to 0 \) as \( n \to \infty \) and \( 1/\varepsilon = o(n^{\beta/r}) \) and \( c > 0 \),

\[
\inf_{h \in H^\varepsilon} \| h - f \|_{\varepsilon} \lesssim e^{Kc^{-2/r} \varepsilon^{-2/\beta}}
\]

for \( n \) large enough, where \( K > 0 \) is a constant.

Proof. We use an expansion \( f = \sum f_i \psi_i \), with \( \psi_i \) the orthonormal eigenfunctions of the Laplacian. We saw in the proof of Lemma 3.5.2 that if we define \( h = \sum_{i \leq I} f_i \psi_i \) for \( I = \text{const} \times \varepsilon^{-r/\beta} \), then \( \| h \|_{\varepsilon} \leq \varepsilon \). By (3.31), the RKHS-norm of \( h \) satisfies in this case

\[
\| h \|_{H^\varepsilon}^2 = \sum_{i \leq I} e^{(n/c)^2/r \lambda_i} f_i^2 \leq Q^2 e^{(n/c)^2/r \lambda_i}.
\]

The condition on \( \varepsilon \) ensures that for the choice of \( I \) made above and \( n \) large enough, \( i_0 \leq I \leq \kappa n \). Hence, by (3.33)–(3.34), \( \| h \|_{H^\varepsilon}^2 \) is bounded by a constant times the right-hand side of (3.39).

3.5.2.2 Proof of (3.10)–(3.11)

Define \( B_n = M_n H_1^{c_n} + \varepsilon_n B_1 \), where \( \varepsilon_n \) is as above and \( M_n \) and \( c_n \) are determined below.

For (3.10) we first note again that

\[
\mathbb{P}(f \notin B_n) \leq \int_0^{\varepsilon_n} \mathbb{P}(f \notin M_n H_1^{c_n} + \varepsilon_n B_1 \mid c) e^{-c} \, dc + \int_{\varepsilon_n}^\infty e^{-c} \, dc.
\]

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

\[
\mathbb{P}(f \notin B_n \mid c) \leq 1 - \Phi(\Phi^{-1}(\mathbb{P}(\| f \|_n \leq \varepsilon_n \mid c_n) + M_n)).
\]
3. Bayesian function estimation on large graphs

By Lemma 3.5.4 the small ball probability on the right is lower bounded by
\[
\exp \left( - Kc_n \log^{1+r/2} \frac{c_n}{\varepsilon_n^2} \right).
\]

It follows that for \( c \leq c_n \),
\[
P(f \notin B_n | c) \leq 1 - \Phi \left( M_n - K' c_n \log^{1+r/2} \frac{c_n}{\varepsilon_n^2} \right)
\]
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 \( n \),
\[
P(f \notin B_n) \leq e^{-K'' c_n \log^{1+r/2} \frac{c_n}{\varepsilon_n^2}} + e^{-c_n} \leq 2e^{-c_n}.
\]

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 3.5.6 implies that the entropy satisfies, for any \( \tilde{\varepsilon}_n \geq \varepsilon_n \),
\[
\log N(2\tilde{\varepsilon}_n, B_n, \| \cdot \|_n) \leq \log N(2\varepsilon_n, B_n, \| \cdot \|_n) \lesssim c_n \left( \log \frac{M_n}{\varepsilon_n} \right)^{1+r/2}.
\]

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

**Lemma 3.5.6.** Let \( \varepsilon, c > 0 \) be such that \( c \log^{r/2}(1/\varepsilon) \to \infty \) as \( n \to \infty \). Then for \( n \) large enough,
\[
\log N(\varepsilon, \mathbb{H}_c^1, \| \cdot \|_n) \lesssim c \log^{1+r/2} \left( \frac{1}{\varepsilon} \right).
\]

**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 (3.33)–(3.34) 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 \( \varepsilon \)-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 \( 2\varepsilon \)-net for \( A \). Hence, \( N(2\varepsilon, A, \| \cdot \|) \lesssim \varepsilon^{-I} \). The proof is completed by recalling the expression for \( I \). \( \blacksquare \)
3.5.3 Proof of Theorem 3.4.1

The idea of the proof is to use general results for obtaining convergence rates for posterior distributions formulated in Theorems 1.5.1, 1.5.2. However, straightforward application of the theorems does not provide the desired results, hence we are going to reformulate the setting and the conditions of the theorem, use Theorems 1.5.1, 1.5.2, and then translate the obtained result back into original setting. In the sampling with replacement setting the observed data is identically distributed, which allows us to use Theorem 1.5.1. This is a special case of function estimation with random design and with less data available. For the coin flipping setting, we reformulate the problem in a such a way that it becomes a special case of the function estimation problem with fixed design. That allows us to use Theorem 1.5.2. The main challenge of the proof is to show that the distances employed in Theorems 1.5.1, 1.5.2 can be appropriately related to the $\| \cdot \|_n$--norm that is used in the conditions and conclusions of the theorem.

Observe that Theorems 1.5.1, 1.5.2 can only be applied for the case $k = 0$, when $\varepsilon_n = \tilde{\varepsilon}_n$. When $k \neq 0$, one has to use a generalised version of the Theorem 1.5.2, which can be directly deduced from Theorem 2.1 in Ghosal and van der Vaart [2001]. For simplicity from now on we assume that $k = 0$. Also, we omit the proof for the sampling with replacement missing mechanism in the classification setting, since this case can be handled in the same manner as the regression setting.

3.5.3.1 Proof for sampling with replacement in the regression setting

Since the observations are identically distributed, we can apply Theorem 1.5.1 for the following set up. Consider the space $\{1, \ldots, n\} \times \mathbb{R}$ with the product $\sigma$–algebra. Define the measure $\mu$ such that for $A \subseteq \{1, \ldots, n\}$ and $B \in \mathcal{B}(\mathbb{R})$

$$\mu(A \times B) = \sum_{i=1}^{n} 1_{i \in A} \lambda(B),$$

where $\lambda$ is the standard Lebesgue measure. Then for every $i = 1, \ldots, m_n$ the distribution of $(X_i, Z_i)$ is given by the probability measure $P_{f_0}$, where for a function $f \in \mathbb{R}^n$ on the vertices of the graph the probability measure $P_f$ is given by

$$P_f(A \times B) = \frac{1}{n} \sum_{j \in A} \int_{B} \varphi_{f(j), \sigma^2}(x) dx,$$
3. Bayesian function estimation on large graphs

with \( \varphi_{a,\sigma^2} \) the density of a Gaussian random variable with mean \( a \) and variance \( \sigma^2 \). The Hellinger distance \( d(f_1, f_2) \) between \( P_{f_1} \) and \( P_{f_2} \) satisfies

\[
d^2(P_{f_1}, P_{f_2}) = \int_{\mathbb{R}} \left( \sqrt{dP_{f_1}/d\mu} - \sqrt{dP_{f_2}/d\mu} \right)^2 d\mu = \frac{1}{n} \sum_{i=1}^{n} \int_{\mathbb{R}} \left( \sqrt{\varphi_{f_1(i),\sigma^2}(x)} - \sqrt{\varphi_{f_2(i),\sigma^2}(x)} \right)^2 dx.
\]

Using the properties of \( \varphi_{a,\sigma^2} \) we obtain that

\[
d(P_{f_1}, P_{f_2}) \asymp \|f_1 - f_2\|_n. \tag{3.40}
\]

In order to use Theorem 1.5.1 we consider the Kullback–Leibler divergence given by

\[
-\int \log\left(\frac{dP_f/dP_{f_0}}{dP_{f_0}}\right) dP_{f_0} = -\frac{1}{n} \sum_{i=1}^{n} \int_{\mathbb{R}} \log \frac{\varphi_{f(i),\sigma^2}(x)}{\varphi_{f_0(i),\sigma^2}(x)} \varphi_{f_0(i),\sigma^2}(x) dx \lesssim \|f - f_0\|_n^2.
\]

Additionally, one can see that

\[
\int (\log(dP_f/dP_{f_0}))^2 dP_{f_0} \lesssim \|f - f_0\|_n^2.
\]

For more details on the computations see Ghosal and van der Vaart [2007]. Define the following sets in \( \mathbb{R}^n \)

\[
B(f_0, \varepsilon) = \left\{ f : -\int \log(dP_f/dP_{f_0}) dP_{f_0} \leq \varepsilon^2, \int (\log(dP_f/dP_{f_0}))^2 dP_{f_0} \leq \varepsilon^2 \right\}.
\]

Using (3.16)-(3.18) we obtain that for some constant \( K > 0 \), and for all \( L > 1 \) the following holds

\[
\Pi(B(f_0, \varepsilon)) \geq e^{-Km_n(\varepsilon_n^*)^2}, \tag{3.41}
\]

\[
\Pi(f \notin B_n) \leq e^{-Lm_n(\varepsilon_n^*)^2}, \tag{3.42}
\]

\[
\log N(\varepsilon_n^*, B_n, d) \leq m_n(\varepsilon_n^*)^2. \tag{3.43}
\]

Notice that (3.41)-(3.43) match one to one the conditions of Theorem 1.5.1. Thus, there exists an \( M > 0 \) such that

\[
\Pi(f : d(P_f, P_{f_0}) \geq M\varepsilon_n^* | (X_1, Z_1), \ldots, (X_{m_n}, Z_{m_n})) \xrightarrow{P_{f_0}} 0.
\]

Using (3.40) we derive the desired result.
3.5. Proofs

3.5.3.2 Proof for the coin flipping mechanism in the regression setting

We are going to apply Theorem 1.5.2 to the following set up. Consider the space $\mathbb{R}^* = \mathbb{R} \cup \{\star\}$ endowed with the $\sigma$-algebra generated by the Borel sets in $\mathbb{R}$ and the set $\{\star\}$. For every measurable subset $B$ of $\mathbb{R}^*$ set

$$\lambda^*(B) = \lambda(B \cap \mathbb{R}) + \delta(B \cap \{\star\}),$$

where $\delta$ is a counting measure and $\lambda$ is the standard Lebesgue measure on $\mathbb{R}$. In this setting $Z_1, Z_2, \ldots, Z_n$ are independent observations distributed according to the following densities (with respect to $\lambda^*$)

$$Z_i \sim g_{0,i}^* = p_n \varphi_{f_0(i), \sigma^2} + (1 - p_n) \mathbb{1}_{\{\star\}},$$

where $\varphi_{a,\sigma^2}$ denotes a Gaussian density of $N(a, \sigma^2)$. For a function $f \in \mathbb{R}^n$ on the vertices of the graph consider a vector of densities $g_f^*$ by setting

$$g_{f,i}^* = p_n \varphi_{f(i), \sigma^2} + (1 - p_n) \mathbb{1}_{\{\star\}}.$$

The average Hellinger distance between such vectors of densities is given by

$$(d_n^*(g_f^*, g_{f'}^*))^2 = \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}^*} (\sqrt{g_{f,i}^*} - \sqrt{g_{f',i}^*})^2 d\lambda^* =
\frac{p_n}{n} \sum_{i=1}^n \int_{\mathbb{R}} (\sqrt{\varphi_f(i, \sigma^2)} - \sqrt{\varphi_{f'}(i, \sigma^2)})^2 d\lambda,$$

where $f, f' \in \mathbb{R}^n$ are two functions on the graph. We obtain that

$$d_n^*(g_f^*, g_{f'}^*) \asymp \sqrt{p_n} \|f - f'\|_n. \quad (3.44)$$

Additionally, we define the following sets

$$B_n^*(f_0, \varepsilon) = \left\{ f \in \mathbb{R}^n : \frac{1}{n} \sum_{i=1}^n K^*(g_{f_0,i}^*, g_{f,i}^*) \leq \varepsilon^2, \frac{1}{n} \sum_{i=1}^n V^*(g_{f_0,i}^*, g_{f,i}^*) \leq C\varepsilon^2 \right\},$$

where for two densities $f^*, g^*$ defined on $\mathbb{R}^*$

$$K^*(f^*, g^*) = \int f^* \log(f^*/g^*) d\lambda^*,
V^*(f^*, g^*) = \int f^* (\log(f^*/g^*) - K(f^*, g^*))^2 d\lambda^*.$$
3. Bayesian function estimation on large graphs

We are going to show that the sets 
\[ G_n^\star = \{ g_f^\star = (g_{f,1}^\star, \ldots, g_{f,n}^\star), \text{for } f \in B_n \} \]

satisfy the following conditions

\[ \log N(\varepsilon_n/36, G_n^\star, d_n^\star) \leq n\varepsilon_n^2, \quad (3.45) \]
\[ \frac{\Pi(f \in \mathbb{R}^n : g_f^\star \notin G_n^\star)}{\Pi(B_n(f_0, \varepsilon_n))} = o\left(e^{-2n\varepsilon_n^2}\right), \quad (3.46) \]
\[ \Pi(B_n(f_0, \varepsilon_n)) \geq e^{-n\varepsilon_n^2}. \quad (3.47) \]

Then by Theorem 1.5.2 we have \[ \Pi(f : d_n^\star(g_f^\star, g_{f_0}^\star) \geq M\varepsilon_n \mid X^{(n)}) \xrightarrow{P_{f_0}} 0 \] for some \( M > 0 \). The statement of the theorem follows from (3.44).

Recall the definitions of the discrepancy measures \( K(\cdot, \cdot) \) and \( V(\cdot, \cdot) \) for two probability densities on \( \mathbb{R} \)

\[ K(f, g) = \int f \log(f/g) d\lambda, \]
\[ V(f, g) = \int (f \log(f/g) - K(f,g))^2 d\lambda. \]

It is known (see for example Ghosal and van der Vaart [2007]) that there exists constants \( A, B \) such that

\[ K^\star(g_{f_0, i}^\star, g_{f, i}^\star) = p_n K(\varphi_{f_0(i), \sigma^2}, \varphi_{f(i), \sigma^2}) \leq A (f_0(i) - f(i))^2, \]
\[ V^\star(g_{f_0, i}^\star, g_{f, i}^\star) = p_n V(\varphi_{f_0(i), \sigma^2}, \varphi_{f(i), \sigma^2}) \leq B (f_0(i) - f(i))^2. \]

Then using (3.16) we obtain

\[ -\log \Pi(B_n^\star(f_0, \varepsilon_n, k)) \gtrsim -\log \Pi(f : \| f - f_0 \|^2_n \leq p_n^{-1}\varepsilon_n^2) \leq Kn\varepsilon_n^2, \quad (3.48) \]

since \( (\varepsilon_n^\star)^2 = p_n^{-1}\varepsilon_n^2 \). Moreover, by (3.44) and (3.18) the entropy can be bounded as follows

\[ \log N(\varepsilon_n, G_n^\star, d_n^\star) \asymp \log N(p_n^{-1/2}\varepsilon_n, B_n, \| \cdot \|_n) \lesssim n\varepsilon_n^2. \quad (3.49) \]

Also, by (3.17)

\[ \Pi(f \in \mathbb{R}^n : g_f^\star \notin G_n^\star) = \Pi(f \notin B_n) \leq e^{-Ln\varepsilon_n^2}. \quad (3.50) \]

Observe that (3.48)-(3.50) match one to one the conditions (3.45)-(3.47). It concludes the proof of the theorem for this case.

3.5.3.3 Proof for coin flipping mechanism in the classification setting

We prove the result using Theorem 1.5.2 in the following setting. Consider the space \( \{0, 1, \star\} \) with the counting measure \( \lambda_c^\star \). Notice that \( Z_1, \ldots, Z_n \) are independent
3.5. Proofs

observations distributed according to the following densities

\[ U_i \sim \rho^*_0(i) = p_n \left[ \rho_0(i) \mathbb{1}_{\{1\}} + (1 - \rho_0(i)) \mathbb{1}_{\{0\}} \right] + (1 - p_n) \mathbb{1}_{\{*\}}. \]

For a function \( f \in \mathbb{R}^n \) on the vertices of the graph define densities \( \rho_f, \rho^*_f \) such that

\[ \rho^*_f(i) = p_n \left[ \rho_f(i) \mathbb{1}_{\{1\}} + (1 - \rho_f(i)) \mathbb{1}_{\{0\}} \right] + (1 - p_n) \mathbb{1}_{\{*\}}. \]

Then the average Hellinger distance is given by

\[
\left( d^*_n(\rho^*_f, \rho^*_f') \right)^2 = \frac{1}{n} \sum_{i=1}^{n} \left[ \left( \sqrt{p_n \rho_f(i)} - \sqrt{p_n \rho_f'(i)} \right)^2 + \left( \sqrt{p_n(1 - \rho_f(i))} - \sqrt{p_n(1 - \rho_f'(i))} \right)^2 \right] \approx p_n \| \sqrt{\rho_f} - \sqrt{\rho_f'} \|_n^2.
\]

Thus, using the properties of the link function \( \Psi \) we obtain

\[
\left( d^*_n(\rho^*_f, \rho^*_f') \right)^2 \approx p_n \| \rho_f - \rho_f' \|_n^2.
\] (3.51)

Moreover, for every \( v_1, v_2 \in \mathbb{R}^n \)

\[ \| \sqrt{\Psi(v_1)} - \sqrt{\Psi(v_2)} \|_n \lesssim ||v_1 - v_2||_n. \]

Hence, we have the following relation between \( d^*_n(\rho^*_f, \rho^*_f') \) and \( ||f - f'||_n \)

\[ \left( d^*_n(\rho^*_f, \rho^*_f') \right)^2 \lesssim p_n ||f - f'||_n^2. \]

Applying a similar argument and Lemma 2.5.3 we get that

\[ \frac{1}{n} \sum_{i=1}^{n} K^*(\rho^*_0(i), \rho^*_f(i)) \lesssim p_n \frac{1}{n} \sum_{i=1}^{n} K(\Psi(f_0(i)), \Psi(f(i))) \lesssim p_n ||f_0 - f||_n^2, \]

where \( f_0 = \Psi^{-1}(\rho_0) \). Additionally,

\[ \frac{1}{n} \sum_{i=1}^{n} V^*(\rho^*_0(i), \rho^*_f(i)) \lesssim p_n ||f_0 - f||_n^2. \]

Define the sets \( G^*_n = \{ \rho^*_f, \text{ for } f \in B_n \} \). Also, consider the following sets

\[ B^*_n(\rho_0, \varepsilon) = \Psi \left( \left\{ f : \frac{1}{n} \sum_{i=1}^{n} K^*(\rho^*_0(i), \rho^*_f(i)) \leq \varepsilon^2, \frac{1}{n} \sum_{i=1}^{n} V^*(\rho^*_0(i), \rho^*_f(i)) \leq C\varepsilon^2 \right\} \right). \]
By (3.16)–(3.18) we have
\[
\log N(\varepsilon_n, G^*_n, d^*_n) \lesssim n\varepsilon_n^2,
\]
\[
\Pi(f \in \mathbb{R}^n : \rho^*_f \notin G^*_n) \lesssim e^{-(L-K)n\varepsilon_n^2},
\]
\[
\Pi(B^*_n(\rho_0, \varepsilon_n)) \gtrsim e^{-Kn\varepsilon_n^2}.
\]
Then using Theorem 1.5.2 we obtain
\[
\Pi(f \in \mathbb{R}^n : d^*_n(\rho^*_f, \rho^*_f) \geq M\varepsilon_n | X^{(n)}) \xrightarrow{P_{eq}} 0.
\]
The desired result follows from (3.51).

### 3.5.4 Proofs of Theorem 3.4.2 and Theorem 3.4.3

The proofs of Theorem 3.4.2 and 3.4.3 go along the lines of the proof of Theorem 3.2.1 and 3.2.2 with slight modifications due to the differences in the prior and in the required result. We only provide the outline of the proof of Theorem 3.4.2, since Theorem 3.4.3 can be derived in a similar manner.

#### 3.5.4.1 Proof of (3.21)

We know from Lemma 3.5.1 that for large \( n \) and when \( (\varepsilon \sqrt{n})/c^{(\alpha+r)/r} \) is small enough, there is a following bound on the centred small balls

\[
-\log \mathbb{P}(\|f\|_n \leq \varepsilon) \lesssim \left( \frac{c^{(\alpha+r)/r}}{\varepsilon \sqrt{n}} \right)^{r/\alpha}.
\]

Note that the condition is satisfied for \( n \) large enough, since

\[
\frac{\varepsilon_n^{*} \sqrt{n}}{c^{(\alpha+r)/r}} \asymp (\varepsilon_n^{*})^{\alpha/\beta} \rightarrow 0.
\]

To handle the bias, we recall from Lemma 3.5.2 that for \( \varepsilon \rightarrow 0 \) such that \( \varepsilon^{-1} = o(n^{\beta/r}) \) we obtain

\[
\inf_{h \in \mathbb{H}^c} \|h - f\|_n \leq \varepsilon \|h\|_{\mathbb{H}_c}^2 \lesssim nc^{-(2\alpha+r)/r} \varepsilon^{-(2(\alpha-\beta)+r)/\beta}.
\]

Then for \( c \) such that \( c^{(\alpha+r)/r} \asymp \sqrt{n}(\varepsilon_n^{*})^{1-\alpha/\beta} \) we have that

\[
-\log \mathbb{P}(\|f - f_0\|_n \leq \varepsilon_n^{*} | c) \lesssim n\varepsilon_n^2.
\]

By integrating \( c \) out we deduce that for some constant \( K_1 > 0 \)

\[
\mathbb{P}(\|f - f_0\|_n \leq \varepsilon_n^{*}) \gtrsim e^{-K_1 n\varepsilon_n^2}.
\]
3.5. Proofs

3.5.4.2 Proof of (3.22)-(3.23)

To prove (3.23) we recall that by Lemma 3.5.1 for small $\varepsilon$

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

We use slightly altered sieves

$$B_n = M_n \mathbb{H}_1^{c_n} + \varepsilon_n^* \mathbb{B}_1,$$

where for $M$ and $C_0$ to be determined below

$$c_n = C_0 n^{r/(2\alpha+r)} (\varepsilon_n^*)^{r(1-\alpha/\beta)/(\alpha+r/2)},$$

$$\varepsilon_n^* = (np_n)^{-\beta/(2\beta+r)},$$

$$M_n = M \sqrt{n\varepsilon_n},$$

$$\varepsilon_n = p_n^{r/2(2\beta+r)} n^{-\beta/(2\beta+r)}.$$

Then by Lemma 3.5.3 we get

$$\log (2\varepsilon_n^*, B_n, \|\cdot\|_n) \lesssim c_n \left( \frac{M_n}{\sqrt{n\varepsilon_n^*}} \right)^{r/(\alpha+r/2)} \approx n\varepsilon_n^2.$$ 

In order to prove (3.22) we follow the steps of the proof of Theorem 3.4.2. First, we notice that for $c \leq c_n$ we have $\mathbb{H}_1^c \subset \mathbb{H}_1^{c_n}$. Then for $c \leq c_n$ it holds that

$$\mathbb{P}(f \notin B_n \ | \ c) \leq 1 - \Phi(\Phi^{-1}(\mathbb{P}(\|f\|_n \leq \varepsilon_n^* | c_n)) + M_n) \leq 1 - \Phi(M_n - K'(\varepsilon_n^*)^{-r/2\beta}).$$

We can set $M$ big enough, such that the righthand side is bounded by $e^{-K_2 n\varepsilon_n^2}$. Then we can choose $C_0$ such that

$$\mathbb{P}(f \notin B_n) \leq \int_0^{C_n} \mathbb{P}(f \notin B_n \ | \ c) e^{-p_n^{r/(2\alpha+r)} c} \, dc + \int_{C_n}^{\infty} e^{-p_n^{r/(2\alpha+r)} c} \, dc \leq e^{-K_2 n\varepsilon_n^2}.$$