Submitted 11/16; Revised 12/17; Published 4/18

# On $b$-bit Min-wise Hashing for Large-scale Regression and Classification with Sparse Data

**Rajen D. Shah** R.SHAH@STATSLAB.CAM.AC.UK
*Statistical Laboratory*
*University of Cambridge*
*Cambridge, CB3 0WB, UK*

**Nicolai Meinshausen** MEINSHAUSEN@STAT.MATH.ETHZ.CH
*Seminar für Statistik*
*ETH Zürich*
*8092 Zürich, Switzerland*

**Editor:** Sanjiv Kumar

## Abstract

Large-scale regression problems where both the number of variables, $p$, and the number of observations, $n$, may be large and in the order of millions or more, are becoming increasingly more common. Typically the data are sparse: only a fraction of a percent of the entries in the design matrix are non-zero. Nevertheless, often the only computationally feasible approach is to perform dimension reduction to obtain a new design matrix with far fewer columns and then work with this compressed data.

$b$-bit min-wise hashing (Li and König, 2011; Li et al., 2011) is a promising dimension reduction scheme for sparse matrices which produces a set of random features such that regression on the resulting design matrix approximates a kernel regression with the resemblance kernel. In this work, we derive bounds on the prediction error of such regressions. For both linear and logistic models, we show that the average prediction error vanishes asymptotically as long as $q\|\boldsymbol{\beta}^*\|_2^2/n \to 0$, where $q$ is the average number of non-zero entries in each row of the design matrix and $\boldsymbol{\beta}^*$ is the coefficient of the linear predictor.

We also show that ordinary least squares or ridge regression applied to the reduced data can in fact allow us fit more flexible models. We obtain non-asymptotic prediction error bounds for interaction models and for models where an unknown row normalisation must be applied in order for the signal to be linear in the predictors.

**Keywords:** large-scale data, min-wise hashing, resemblance kernel, ridge regression, sparse data.

## 1. Introduction

The modern field of high-dimensional statistics has now developed a powerful range of methods to deal with data sets where the number of variables $p$ may greatly exceed the number of variables $n$ (see Bühlmann and van de Geer (2011) for an overview of recent advances). The prototypical example of microarray data, where $p$ may be in the tens of thousands but $n$ is typically not more than a few hundred, has motivated much of this development. Yet not all modern data sets come in this sort of shape and size. The emerging area of 'large-scale data' or the more vaguely defined 'Big Data' is a response to the increasing prevalence of computationally challenging data sets as arise in text analysis or web-scale prediction tasks, to give two examples. Here both $n$ and $p$ can run into the millions or more, particularly if interactions are considered. In these 'large $p$, large $n$' regression scenarios, one can imagine situations where ordinary least squares (OLS) has a competitive performance for prediction, but the sheer size of the data renders it infeasible for computational rather than statistical reasons.

An important feature of many large-scale data sets is that they are sparse: the overwhelming majority of entries in the design matrices are exactly zero. This is not to be confused with signal sparsity, a common assumption in the high-dimensional context. Indeed, when the design matrix is sparse, having only a few variables that contribute to the response would make the expected response values of all observations with no non-zero entries for the important variables exactly the same; one expects that such a property would not be possessed by many data sets. However, similarly to the way in which many high-dimensional techniques exploit sparsity to improve statistical efficiency, one might hope that sparsity in the data could be leveraged to yield both computational and statistical improvements, and indeed we demonstrate in this work that this can be achieved.

Kernel machines are an important class of machine learning methods for which such large-scale data poses particularly serious computational challenges. For example, standard implementations of kernel ridge regression would have computational complexity $O(n^3)$ and a storage cost of $O(n^2)$ when $p$ is considered fixed; a large $p$ will increase these computational costs depending on the kernel to be used. There has therefore been a great deal of work on approximating kernel machines by first randomly mapping the $n \times p$ design matrix $\mathbf{X}$ to a $n \times d$ matrix $\mathbf{S}$ with $d \ll p$ such that dot products between rows of $\mathbf{S}$ approximate the kernel evaluated on the corresponding rows of $\mathbf{X}$. Then a regular ridge regression on $\mathbf{S}$ will resemble a kernel ridge regression on $\mathbf{X}$, for example.

A remarkably effective way of forming $\mathbf{S}$ that is applicable when the design matrix is sparse and binary, is $b$-bit min-wise hashing (Li and König, 2011; Li et al., 2011) which is based on an earlier technique called min-wise hashing (Broder et al., 1998; Cohen et al., 2001; Datar and Muthukrishnan, 2002). Here $\mathbf{S}$ is constructed such that the dot product between any two rows of $\mathbf{S}$, $\mathbf{s}_i^T\mathbf{s}_j$, can approximate the *resemblance* or *Jaccard similarity* or between the corresponding rows of $\mathbf{X}$, defined as $|\mathbf{z}_i \cap \mathbf{z}_j|/|\mathbf{z}_i \cup \mathbf{z}_j|$ where $\mathbf{z}_i = \{k : X_{ik} \neq 0\}$.

The empirical performance of regression and classification procedures following $b$-bit min-wise hashing (Li et al., 2011, 2013) is particularly impressive. Existing theory on $b$-bit min-wise hashing (Li and König, 2011) has focused on the variance and bias in the approximation of the kernel. However, there remain significant gaps in our theoretical understanding of this important procedure when used to approximate a kernel machine:

(a) What sorts of regression models is the resemblance kernel well-suited for and how does sparsity of the design matrix play a role?

(b) What is the loss in prediction accuracy due to the approximation provided by $b$-bit min-wise hashing for different sorts of regression procedures?

(c) What is the overall prediction error incurred by different regression methods following $b$-bit min-wise hashing in different regression models?

An answer to (c) would be the ultimate goal here, and it would appear that in order to tackle this one must first solve (a) and (b). In this paper, we take a very different approach and aim to answer (c) directly: rather than considering what sorts of functions lie in the reproducing kernel Hilbert space (RKHS) associated with the resemblance kernel and have low RKHS norms, we look at the sorts of signals that can be approximated well by linear combinations of columns of the matrix $\mathbf{S}$ constructed by $b$-bit min-wise hashing. In this way, we use the random feature expansions provided by $b$-bit min-wise hashing to understand the predictive properties of the resemblance kernel.

### 1.1 Our contributions and organisation of the paper

In this paper we derive finite-sample bounds on the expected risk of linear and logistic regression following dimension reduction through $b$-bit min-wise hashing under various different models. Our results show that the method, and hence also the resemblance kernel, are particularly suited to sparse data.

We describe the $b$-bit min-wise hashing algorithm in Section 2 and also discuss in greater details the connection to the resemblance kernel. We also introduce a generalisation of $b$-bit min-wise hashing applicable to sparse data with real-valued entries motivated by our theory. Perhaps the simplest sorts of signals that we could hope to be able to fit well are linear signals of the form $\mathbf{X}\boldsymbol{\beta}^*$. In Section 3 we first consider how well a linear combination of columns of $\mathbf{S}$ can approximate such a signal. We then study a much larger class of signals defined by first scaling the rows of $\mathbf{X}$ in different ways depending on their sparsity and then forming a linear signal from a scaled version of $\mathbf{X}$. Some form of row normalisation is often performed on the original data as a pre-processing step, but the optimal normalisation to use is seldom known; our theory shows how $b$-bit min-wise hashing, and hence also the resemblance kernel, is able to automatically discover an appropriate scaling in several settings.

In Section 4.1 we study the performance of ordinary least squares, ridge regression and $\ell_2$-penalised logistic regression using the reduced design matrix it creates. Our results are applicable to both linear signals and nonlinear signals of the sort described above. In the former setting, we show that the expected mean-squared prediction error is bounded by a small constant times $\sqrt{q/n}\,\|\boldsymbol{\beta}^*\|_2$, where $q$ is the average number nonzero entries in the rows of $\mathbf{X}$ and $\boldsymbol{\beta}^*$ is the coefficient vector. We present similar results for logistic regression.

In Section 5 we study another form of nonlinear signal that can be approximated by the $b$-bit minwise hashing and the resemblance kernel: we show that interaction models in the original data can also be captured by main effects regression on the compressed data. Variable importance measures are discussed in Section 6. We conclude with a discussion in Section 7. The appendix contains all proofs, an additional result concerning the implications of our approximation error bound for properties of the RKHS of the resemblance kernel, and an empirical study validating our bounds.

### 1.2 Related work

There has been very little work in understanding properties of the resemblance kernel. One of the few pieces of work in this direction is Bouchard et al. (2013), who show that the kernel matrix with entries given by the Jaccard similarity between different elements of the power set of $\{1, \ldots, p\}$ minus the empty set is positive definite. It follows that the RKHS of the resemblance kernel contains every real-valued function on $p$-dimensional binary vectors (see Section B). However, this result is not informative for understanding which sorts of regression models a kernel ridge regression will perform well for, a question which we provide some answers to through our study of $b$-bit min-wise hashing.

Approximating kernel methods using random feature expansions was pioneered by Rahimi and Recht (2007) who used random Fourier features to approximate translation invariant kernels such as the Gaussian kernel. Sutherland and Schneider (2015) provides bounds on the approximation of the corresponding kernel as well as bounds on the distance between the predictions from regression on the random features and kernel ridge regression in terms of distances between the true kernel and its approximation. Le et al. (2013) introduce a scheme related to random Fourier features that further improves the computational efficiency. Rahimi and Recht (2008) consider more general random feature expansions and study how well they can approximate functions in a family determined by the distribution of feature expansions in terms of a certain form of function norm defined on the family. Rahimi and Recht (2009) provides prediction error bounds for a method that minimises the empirical risk of a weighted sum of random feature expansions where weights are constrained in $\ell_\infty$-norm. Bach (2017) studies how well random feature expansions can approximate elements of their corresponding RKHS in terms of the eigenvalues of the associated kernel integral operator. The Nyström method (Williams and Seeger, 2001) is related and aiming at a computationally efficient low-rank approximation to the full kernel matrix; see (Bach, 2013) and (Rudi et al., 2015) for approximation guarantees.

A distinguishing feature of our work is that bounds are obtained not in terms of the norm of the RKHS of the resemblance kernel, which would be difficult to interpret, but in terms of quantities derived directly from the different models considered (we look at linear models with unknown row scaling and at nonlinear interaction models). We could divide the analysis into two parts: (i) first we could try to understand the predictive accuracy when using exact kernel regression with the resemblance kernel for such true regression functions and then (ii) in a second step understand how much predictive accuracy we lose by using $b$-bit minwise hashing as an approximation to using exact kernel regression with the resemblance kernel. Instead of making these two separate steps, we study here directly how well $b$-bit minwise hashing performs for these model classes.

Properties of $b$-bit min-wise hashing related to similarity search are studied in Li and König (2011). Theory concerning its use for large-scale learning is presented in Li et al. (2011) which quantifies the mean and variance of entries in the Gram matrix $\mathbf{S}\mathbf{S}^T$ and its relationship to the resemblance kernel as well as providing comparisons with random projections and *Vowpal Wabbit*. Random feature expansions for other types of kernels are developed in Shi et al. (2009); Weinberger et al. (2009); Vedaldi and Zisserman (2012); Kar and Karnick (2012); Li (2014); Pennington et al. (2015).

More generally, there is a huge variety of dimension reduction schemes across the statistics and computer science literature. Performing principal component analysis (Jolliffe, 1986) (PCA) and retaining only the first $d$ components is one of the most popular methods. One drawback however in the large-scale data setting is that computing the principal components can be computationally demanding. The method of random projections, motivated by the celebrated Johnson–Lindenstrauss lemma (Johnson and Lindenstrauss, 1984), offers dimension reduction at a low computational cost. In this scheme, $\mathbf{X}$ is mapped to $\mathbf{XA}$, where $\mathbf{A}$ is a $p \times d$ matrix typically with i.i.d. random entries. Efficient implementations are discussed in Achlioptas (2001); Li et al. (2006) and some numerical results on random projections and a wider literature review are in Fradkin and Madigan (2003); Vempala (2005). The software package *Vowpal Wabbit* (Langford et al., 2007) is a popular learning system for large-scale data sets that uses sparse random projections.

A separate line of work has considered pre-multiplying $\mathbf{X}$ with a random matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ to produce a reduced matrix $\mathbf{AX} \in \mathbb{R}^{m \times p}$, known as a *sketch*. Though the dimension $p$ is not reduced, when $n$ is large, performing OLS on the sketched matrix may be possible despite the computational infeasibility of applying least squares directly to $\mathbf{X}$. A number of works have studied properties sketched least squares (see Boutsidis and Drineas (2009); Drineas et al. (2011); Mahoney (2011); Pilanci and Wainwright (2015) and references therein) whilst Pilanci and Wainwright (2014) propose an iterative variant of this scheme. Yang et al. (2017) considers sketching ideas in the context of kernel ridge regression.

## 2. $b$-bit min-wise hashing

Given a sparse design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, the aim of dimension reduction is to map this to a compressed matrix $\mathbf{S} \in \mathbb{R}^{n \times d}$, in a way that is computationally efficient and such that the relevant information in $\mathbf{X}$ is preserved in $\mathbf{S}$. Section 2.2 describes the mapping to $\mathbf{S}$ under $b$-bit min-wise hashing for binary data, as proposed in Li and König (2011) and Li et al. (2011). The construction may seem unintuitive at first sight, but we will try to shed light on why the scheme works for linear and interaction models throughout the manuscript.

### 2.1 Notation

Given a matrix $\mathbf{U}$, we will write $\mathbf{u}_i$ and $\mathbf{U}_j$ for the $i$th row and $j$th column respectively, where both are to be regarded as column vectors. The $ij$th entry will be denoted $U_{ij}$. A vector of 1's will be denoted $\mathbf{1}$.

When the parentheses following probability and expectation signs, $\mathbb{P}$ and $\mathbb{E}$, enclose multiple potential sources of randomness, we will sometimes add subscripts to indicate what is being considered as random. For example, if $U$ and $V$ are random variables, we may write $\mathbb{E}_U(U|V)$ for the conditional expectation of $U$ given $V$, and $\mathbb{E}_{U,V}(U+V)$ for the expected value of $U+V$.

### 2.2 Construction of S with $b$-bit min-wise hashing and binary variables

The compressed matrix $\mathbf{S}$ generated by $b$-bit min-wise hashing consists of blocks of size $2^b$, where we may choose the number of blocks $L$. Each block is created using a random permutation and the blocks of columns form a collection of $L$ i.i.d. random matrices.

There are three steps to the construction.

*Step 1:* Generate a random permutation of the set $\{1, \ldots, p\}$, $\pi_l$, and permute the columns of $\mathbf{X}$ according to this permutation.

*Step 2:* Search along each row of the permuted design matrix (in order of increasing column index) and record in the vector $\mathbf{H}_l \in \mathbb{N}^n$ the indices of the variables (indexed as in the original order) with the first non-zero value or the vector $\mathbf{M}_l \in \mathbb{N}^n$ the indices of the variables (indexed as in the permuted order) with the first non-zero value.

*Step 3:* Form $\mathbf{S}_l \in \{0,1\}^{n \times 2^b}$ with $i$th row given by the last $b$ bits of the binary representation of the $i$th entry of $\mathbf{M}_l$. For example, when $b = 1$, all odd numbers in $\mathbf{M}_l$ map to the vector $(0,1)$, whereas all even numbers map to $(1,0)$.

This construction is illustrated for a toy example in Table 1.

| | | |
|---|---|---|
| $\mathbf{X} = \begin{array}{c} \begin{matrix} 1 & 2 & 3 & 4 \end{matrix} \\ \begin{pmatrix} \cdot & 1 & \cdot & 1 \\ \cdot & \cdot & 1 & 1 \\ 1 & \cdot & 1 & \cdot \\ \cdot & 1 & 1 & \cdot \\ 1 & 1 & \cdot & \cdot \end{pmatrix} \end{array} \overset{\pi_l = 2314}{\mapsto} \begin{array}{c} \begin{matrix} 3 & 1 & 2 & 4 \end{matrix} \\ \begin{pmatrix} \cdot & \cdot & \mathbf{1} & 1 \\ \mathbf{1} & \cdot & \cdot & 1 \\ \mathbf{1} & 1 & \cdot & \cdot \\ \mathbf{1} & \cdot & 1 & \cdot \\ \cdot & \mathbf{1} & 1 & \cdot \end{pmatrix} \end{array}$<br>*Step 1*: non-zero indices whose variable indices will appear in $\mathbf{H}_l$ in Step 2 are in bold. | $\mathbf{H}_l = \begin{pmatrix} 2 \\ 3 \\ 3 \\ 3 \\ 1 \end{pmatrix}, \mathbf{M}_l = \begin{pmatrix} 3 \\ 1 \\ 1 \\ 1 \\ 2 \end{pmatrix}$<br>*Step 2.* | $\mathbf{S}_l = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$<br>*Step 3.* |

Table 1: Steps 1–3 applied to a toy example with $b = 2$. Dots represent zeroes.

We can think of each column of $\mathbf{S}_l$ as representing different categories for the observations. The matrix $\mathbf{S}_l$ itself codes for the assignment of the different rows of $\mathbf{X}$ to the different categories. Different blocks $\mathbf{S}_l$ then represent different random categorisations. Identical rows will always be assigned the same categories and the more different the rows are, the less likely they are to be assigned the same category. The notion of difference here is that of *resemblance*; see Section 2.4

Note that one would not necessarily follow the above steps when implementing $b$-bit min-wise hashing. In practice, one would not store the entire matrix of signs nor all the random permutations. In an implementation, hash functions (Carter and Wegman, 1979) would be used to create the matrix $\mathbf{S}$ deterministically, though it is beyond the scope of this paper to go into the details; see Li et al. (2013) for more information and further computational improvements. With this approach, $\mathbf{S}$ would be created row-by-row, and only a single observation from $\mathbf{X}$ would need to be kept in memory at any one time. Furthermore, many rows could be created in parallel. Other ideas such as one-permutation hashing (Li et al., 2012) can also be used to speed up the pre-processing step.

### 2.3 Continuous data and additional randomisation

For continuous data, we introduce a modification where we replace the map extracting the last $b$ bits by $L$ random maps in the following way. Fix $b$ and let $\mathbf{\Psi} \in \{1, \ldots, 2^b\}^{p \times L}$ be a random matrix with independent entries each having the uniform distribution on the set $\{1, \ldots, 2^b\}$. We then create $\mathbf{S}$ by modifying the previous Step 3 to the following.

*Step 3:* Form $\mathbf{S}_l \in \{0,1\}^{n \times 2^b}$ with $i$th row all zero except component $\Psi_{H_{il} l}$ takes the value 1.

*Step 4:* If $\mathbf{X}$ is not binary, multiply the $i$th row of $\mathbf{S}_l$ by $X_{iH_{il}}$.

This generalisation is motivated by our theoretical results on how well the column space of $\mathbf{S}$ can capture different sorts of signals (see Section 3.1).

Let $\mathbf{z}_i = \{k : X_{ik} \neq 0\}$ be the set of variable indices whose entries have non-zero values for the $i$th observation. Performing the steps above for all $l = 1, \ldots, L$, we get $n \times L$ matrices $\mathbf{H}$, and $\mathbf{M}$ given by

$$H_{il} = \arg\min_{k \in \mathbf{z}_i} \pi_l(k), \tag{1}$$

$$M_{il} = \min_{k \in \mathbf{z}_i} \pi_l(k) = \pi_l(H_{il}), \tag{2}$$

The matrix $\mathbf{S}$ is a binary $n \times 2^b L$ matrix. With a slight abuse of notation, we will denote by $\mathbf{S}_{ilc}$ the $c$th entry in the $l$th block of $\mathbf{S}$:

$$S_{ilc} := S_{i(c+(l-1)2^b)} = X_{iH_{il}} \mathbb{1}_{\{\Psi_{H_{il}l} = c\}}, \qquad \text{for } c = 1, \ldots, 2^b. \tag{3}$$

If not stated otherwise, we will work with this second randomised variation of $b$-bit min-wise hashing from now on. We emphasise that we do not make the claim this version is to be preferred over the original proposal of Li and König (2011) and Li et al. (2011) when data is binary. We simply introduce the additional randomisation here to simplify the analysis. We note that the two versions are essentially identical for all practical purposes when $b$ is not too large.

### 2.4 The resemblance kernel

We now briefly describe the connection between $b$-bit min-wise hashing and the resemblance kernel alluded to earlier. This is not needed for the rest of the paper, though it provides some intuition for the scheme. A more detailed analysis from this perspective is carried out by Li et al. (2011) and we refer the reader to Hofmann et al. (2008) for a review of kernel methods and the kernel trick.

Suppose $\mathbf{X}$ is binary. Consider the normalised Gram matrix of the compressed design $\mathbf{S}$ from (randomised) $b$-bit min-wise hashing, $\mathbf{S}\mathbf{S}^T/L$. The expected value of the $ij$th component may be calculated as follows.

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\mathbf{s}_i^T \mathbf{s}_j / L) &= \frac{1}{L} \sum_{l=1}^{L} \sum_{c=1}^{2^b} \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\mathbb{1}_{\{\Psi_{H_{il}l}=c\}} \mathbb{1}_{\{\Psi_{H_{jl}l}=c\}}) \\
&= \mathbb{P}(\Psi_{H_{il}l} = \Psi_{H_{jl}l}) \\
&= \mathbb{P}(\Psi_{H_{il}l} = \Psi_{H_{jl}l} | H_{il} = H_{jl}) \mathbb{P}(H_{il} = H_{jl}) \\
&\qquad + \mathbb{P}(\Psi_{H_{il}l} = \Psi_{H_{jl}l} | H_{il} \neq H_{jl}) \{1 - \mathbb{P}(H_{il} = H_{jl})\} \\
&= \frac{|\mathbf{z}_i \cap \mathbf{z}_j|}{|\mathbf{z}_i \cup \mathbf{z}_j|}(1 - 2^{-b}) + 2^{-b}.
\end{aligned}$$

Thus the $ij$th entry is an average of $L$ i.i.d. random variables with expectation a constant plus a constant times the resemblance between the $i$th and $j$th rows of $\mathbf{X}$. If an intercept term is included when regressing on $\mathbf{S}$, the additive constant plays no part, and the scaling would be absorbed into the scaling of the regression coefficients. We also note that when $\mathbf{X}$ is continuous, the resulting kernel is similar to the the CoRE kernels of Li (2014).

Now as the resemblance kernel is positive definite, the theory surrounding the kernel trick tells us that any $\ell_2$-regularised regression on $\mathbf{S}$ is effectively approximating a regularised regression on transformed data $\phi(\mathbf{x}_i)$ where $\phi : \{0,1\}^p \rightarrow \mathcal{H}$ and $\mathcal{H}$ is a high-dimensional inner product space (the feature space). This space may be taken to be a reproducing kernel Hilbert space (RKHS), and then $\phi$ and $\mathcal{H}$ are uniquely defined.

Although this is encouraging, the kernel trick does not guarantee that regression on $\mathbf{S}$ will necessarily have good predictive properties for models of interest. To gain a better understanding, we must study the regularisation properties of the resemblance kernel itself: what characterises those elements of the associated RKHS $\mathcal{H}$ that have low norm and thus will be penalised less?

A direct analysis of the RKHS corresponding to the resemblance kernel in those terms seems challenging. We take a different approach and explicitly construct regression coefficients for $\mathbf{S}$ that approximate signals of interest. By showing that particular signals can be approximated well, we are indirectly discovering elements of $\mathcal{H}$ with low RKHS norm (see also Section B for more details).

## 3. Approximation error

In this section, we present results that bound the expected prediction error when performing regression on the reduced design matrix $\mathbf{S}$ in the contexts of the linear and logistic regression models. Note that throughout the rest of the manuscript, by $b$-bit min-wise hashing we are referring to the randomised variant described in Section 2.3. Let $q_i$ be the number of non-zero entries in the $i$th row of $\mathbf{X}$, and let $\delta_i = q_i/p$ be the row sparsity. We will assume that the signal we wish to approximate for the $i$th observation takes the form

$$\kappa(\delta_i)\mathbf{x}_i^T\boldsymbol{\beta}^*. \tag{4}$$

Here $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is an unknown vector of coefficients and the function $\kappa$ allows the $i$th linear predictor to be scaled in a way which depends on the number of non-zero entries in the $i$th row of $\mathbf{X}$. Some normalisations of special interest include:

(a) $\kappa(\delta)$ constant. This yields standard linear or logistic regression models.

(b) $\kappa(\delta) \propto \delta^{-1/2}$. In text analysis with a bag of words representation of documents, rows of $\mathbf{X}$ are often scaled to have the same $\ell_2$-norm to help balance situations when documents vary greatly in length (Banerjee et al., 2005). When $\mathbf{X}$ is binary, this is exactly achieved by taking $\kappa(\delta) = p^{-1/2}\delta^{-1/2}$, so $\kappa(\delta_i) = q_i^{-1/2}$.

(c) $\kappa(\delta) \propto \delta^{-1}$. This leads to a $\ell_1$-norm scaling as opposed to the $\ell_2$-norm scaling mentioned above.

Throughout we will assume that $\mathbf{X} \in [-1,1]^{n\times p}$, so the entries in $\mathbf{X}$ are bounded. This covers the important case of binary design but also allows for real-valued entries.

The first step in obtaining our prediction error results is to construct a vector $\mathbf{b}^*$ such that $\mathbf{s}_i^T\mathbf{b}^*$ is close to $\kappa(\delta_i)\mathbf{x}_i^T\boldsymbol{\beta}^*$ on average.

### 3.1 Un-scaled signals

We will first consider un-scaled signals where $\kappa(\delta)$ in (4) is a constant. Non-constant row-scaling is treated in more detail in the Section 3.2. To begin with we will assume that $q_i = q \geq 1$ for all $i = 1, \ldots, n$, a restriction which simplifies the results but highlights some interesting properties of $b$-bit min-wise hashing. Unequal row sparsity is treated in detail in the appendix in Section A.4 but a sketch of the results are given just below Theorem 1.

To simplify notation, we first introduce the following norm for $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$\|\boldsymbol{\beta}\|_b^2 := \|\boldsymbol{\beta}\|_2^2 + (2^b - 2) \sum_{k=1}^{p} \frac{\|\mathbf{X}_k\|_2^2}{n} \beta_k^2. \tag{5}$$

For $b = 1$, we have of course that $\|\boldsymbol{\beta}\|_b^2 = 2\|\boldsymbol{\beta}\|_2^2$. For larger values of $b$, the norm is influenced more heavily by the second term which can be seen to be the weighted version of the $\ell_2$-norm, where the weight of each variable is proportional to its squared $\ell_2$-norm. We will first discuss how well the original signal can be approximated with the column space of the matrix $\mathbf{S}$ generated by the $b$-bit min-wise hashing operation.

**Theorem 1** *Let* $\mathbf{S}$ *be the matrix generated by b-bit min-wise hashing. Then there exists a vector* $\mathbf{b}^* \in \mathbb{R}^{2^b L}$ *with the following properties.*

*(i) The approximation is unbiased:* $\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\mathbf{S}\mathbf{b}^*) = \mathbf{X}\boldsymbol{\beta}^*$.

*(ii) The norm is bounded by*

$$\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{b}^*\|_2^2) \leq \frac{(2-\delta)q}{L(1-2^{-b})} \|\boldsymbol{\beta}^*\|_2^2.$$

*(iii) The approximation error is bounded by*

$$\frac{1}{n}\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{S}\mathbf{b}^* - \mathbf{X}\boldsymbol{\beta}^*\|_2^2) \leq \frac{(2-\delta)q}{2^b L(1-2^{-b})} \|\boldsymbol{\beta}^*\|_b^2.$$

*Specifically, for* $b = 1$, $\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{S}\mathbf{b}^* - \mathbf{X}\boldsymbol{\beta}^*\|_2^2)/n \leq (2-\delta)q\|\boldsymbol{\beta}^*\|_2^2/L$.

A form of the approximation error (iii) and the norm bound (ii) continue to be valid in the non-equal sparsity case under a mild restriction on the size of $L$, where we get instead of (iii) the bound

$$\frac{1}{n}\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{S}\mathbf{b}^* - \mathbf{X}\boldsymbol{\beta}^*\|_2^2) \leq \frac{6\bar{q}}{2^b L(1-2^{-b})} \|\boldsymbol{\beta}^*\|_b^2,$$

where $\bar{q}$ is the the average of the $q_i$; see Theorem 12 in the appendix for details.

The results above show that the signal $\mathbf{X}\boldsymbol{\beta}^*$ can be well approximated by a linear combination of the columns in the matrix $\mathbf{S}$ if we generate a sufficiently large number of permutations $L$, especially for sparse data matrices. Another useful property of $\mathbf{b}^*$ here, aside from the approximation accuracy it delivers, is given in (ii): on average, $\|\mathbf{b}^*\|_2^2$ is small when $L$ is large. This proves to be useful when studying the application of ridge regression.

This result has interesting implications for the resemblance kernel and its RKHS $\mathcal{H}$. In particular, it shows that if we constrain the input space to contain those vectors with sparsity $q$, linear functions $f_{\boldsymbol{\beta}}$ defined by coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$ with $\sum_j \beta_j = 0$ have RKHS norm satisfying $\|f_{\boldsymbol{\beta}}\|_{\mathcal{H}}^2 \leq (2-\delta)q\|\boldsymbol{\beta}\|_2^2$. As these properties of the RKHS are not directly used in any subsequent results, we defer formal presentation of these facts to Section B in the appendix.

Whilst the bound on the expectation of $\|\mathbf{b}^*\|_2^2$ is almost constant as $b$ changes, the approximation error bound (iii) does vary with $b$. Consider the case where $\mathbf{X}$ is binary and let $\gamma_k = \|\mathbf{X}_k\|_2^2/n$ be the column sparsity. Typically one would expect $\|\boldsymbol{\beta}^*\|_2^2$ to be significantly larger than $\sum_{k=1}^p \gamma_k \beta_k^{*2}$ and thus increasing $b$ by 1 almost halves the approximation error when $b$ is small.

A proof of Theorem 1 is given in Section A of the appendix; here we briefly sketch some of the main ideas. Note that

$$\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\mathbf{S}\mathbf{b}^*) = \sum_{l=1}^L \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}\biggl(\sum_{c=1}^{2^b} \mathbf{S}_{lc} b_{lc}^*\biggr). \tag{6}$$

We construct $\mathbf{b}^*$ with the following two properties: each of the $L$ blocks of $\mathbf{b}^*$ are i.i.d. with the $l$th block only depending on $\pi_l$ and $\boldsymbol{\Psi}_l$; and each of the $L$ summands in (6) equals $\mathbf{X}\boldsymbol{\beta}^*/L$. With each of the $L$ summands being unbiased in this way, we see that the approximation error is controlled by the variance of the sum; this variance scales as $1/L$ since the summands are i.i.d.

At first sight it may seem surprising that it is possible to exhibit a $\mathbf{b}^*$ with each block having the unbiasedness property discussed above. However, the following construction gives an indication of the possibilities. Using our convention that the $c$th component of the $l$th block of $\mathbf{b}^*$ is indexed as $b_{lc}^* := b_{c+(l-1)2^b}^*$, consider taking

$$b_{lc}^* = \frac{q}{L}\sum_{k=1}^p \beta_k^* \frac{\mathbb{1}_{\{\Psi_{lk}=c\}} - 2^{-b}}{1-2^{-b}}. \tag{7}$$

Then writing $\boldsymbol{\psi} = \boldsymbol{\Psi}_1$, $\pi = \pi_1$, $H_i = H_{i1}$ we have

$$\begin{aligned}\frac{L}{q}\mathbb{E}_{\pi,\boldsymbol{\psi}}\biggl(\sum_{c=1}^{2^b} S_{lc} b_{1c}^*\biggr) &= \mathbb{E}_{\pi,\boldsymbol{\psi}}\biggl(\sum_{c=1}^{2^b}\sum_{j=1}^p X_{ij}\mathbb{1}_{\{H_i=j,\psi_j=c\}}\sum_{k=1}^p \beta_k^* \frac{\mathbb{1}_{\{\psi_k=c\}} - 2^{-b}}{1-2^{-b}}\biggr) \\ &= \mathbb{E}_{\pi,\boldsymbol{\psi}}\biggl(\sum_{j=1}^p X_{ij}\mathbb{1}_{\{H_i=j\}}\sum_{k=1}^p \beta_k^* \frac{\mathbb{1}_{\{\psi_k=\psi_j\}} - 2^{-b}}{1-2^{-b}}\biggr). \end{aligned} \tag{8}$$

Now since $\mathbb{E}_{\boldsymbol{\psi}}\{(\mathbb{1}_{\{\psi_k=\psi_j\}} - 2^{-b})/(1-2^{-b})\} = \mathbb{1}_{\{k=j\}}$ we see the above display equals

$$q\sum_{k=1}^p X_{ik}\beta_k^*\mathbb{P}_\pi(H_i = k) = \mathbf{X}\boldsymbol{\beta}^*.$$

The final line uses the fact that for $k$ with $X_{ik} \neq 0$, $\mathbb{P}_\pi(H_i = k)$ is the reciprocal of the number of non-zero entries in the $i$th row of $\mathbf{X}$; with our simplifying assumption of equal row sparsity, this is precisely $1/q$. Note one could scale the rows of $\mathbf{S}$ according to the number of non-zeroes in each row to achieve unbiasedness in the case of unequal row sparsity. However as shown in Section A.4, it turns out that by incurring some bias one can still keep the approximation error low even in this situation without having to perform any sort of scaling.

The form of $\mathbf{b}^*$ used in the proof of Theorem 1 differs slightly from that in (7) by introducing a random weight multiplying each coefficient that decays as $\pi_l(k)$ increases. This reduces the variance and yields the approximation error in (iii) that has a factor $q$ rather than the factor of $p$ which would be obtained from (7).

### 3.2 Row-scaled signals

We now turn to the more general setting with unequal row sparsity and signal given by (4). We consider the family of scaling functions $\delta \mapsto (\delta_{\min}/\delta)^a$ where $\delta_{\min} = \min_i \delta_i$, for $1/2 \leq a \leq 1$. Including $\delta_{\min}$ in the scaling functions means that were the row sparsity to be equal, the approximation error here would be of the same form as that considered in Theorems 1. We could alternatively replace $\delta_{\min}$ with the average of the $\delta_i$ for the same effect, but using $\delta_{\min}$ helps to simplify the results. Writing $q_{\min} = \min_i q_i$, we have the following results.

**Theorem 2** *Let $L \geq 5$ and assume $\delta_{\min} \leq 1/2$ if $a = 1/2$, and $L > 2/(2a-1)$ if $a > 1/2$. Then there exists $\mathbf{b}^* \in \mathbb{R}^L$ depending on $a$ such that the approximation error satisfies*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}[\{(\delta_{\min}/\delta_i)^a \mathbf{x}_i^T \boldsymbol{\beta}^* - \mathbf{s}_i^T \mathbf{b}^*\}^2] \leq$$
$$\begin{cases} \dfrac{q_{\min}}{2^b L(1-2^{-b})}\|\boldsymbol{\beta}^*\|_b^2 \log\{4\log(L)/\delta_{\min}\} & \text{if } a = 1/2, \\ \\ \dfrac{q_{\min}}{2^b L(1-2^{-b})}\|\boldsymbol{\beta}^*\|_b^2 \dfrac{1}{2a-1}[\log\{2(2a-1)L\}]^{2a-1} & \text{if } 1/2 < a \leq 1, \end{cases}$$

*and the norm of $\mathbf{b}^*$ is bounded in expectation by*

$$\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{b}^*\|_2^2) \leq \begin{cases} \dfrac{q_{\min}\log\{4\log(L)/\delta_{\min}\}}{L(1-2^{-b})}\|\boldsymbol{\beta}^*\|_2^2 & \text{if } a = 1/2, \\ \\ \dfrac{1}{2a-1}\dfrac{q_{\min}[\log\{2(2a-1)L\}]^{2a-1}}{L(1-2^{-b})}\|\boldsymbol{\beta}^*\|_2^2 & \text{if } 1/2 < a \leq 1. \end{cases}$$

The min-wise hashing based dimension reduction scheme appears to be well-suited to approximating signals scaled by a power of the sparsity, with the approximation error only incurring a further multiplicative term involving $\log(L)$ compared to the results of Theorem 1.

We now briefly outline how we construct coefficient vectors $\mathbf{b}^*$ achieving the bounds above. Consider the following refinement of (7):

$$b_{lc}^* = \frac{1}{L}\sum_{k=1}^{p} \beta_k^* \frac{\mathbb{1}_{\{\Psi_{lk}=c\}} - 2^{-b}}{1-2^{-b}} w_{\pi_l(k)},$$

where $\mathbf{w} \in \mathbb{R}^p$ is a vector of non-negative weights. Arguing as in (8) but replacing $q\beta^*_k$ with $\beta^*_k w_{\pi(k)}$ we arrive at

$$L\mathbb{E}_{\pi,\boldsymbol{\psi}}\biggl(\sum_{c=1}^{2^b} S_{lc} b^*_{1c}\biggr) = \sum_{k=1}^{p} X_{ik}\beta^*_k \mathbb{E}_\pi(\mathbb{1}_{\{H_i=k\}} w_{\pi(k)}).$$

Recall that writing $M_i = M_{i1}$, $M_i = \pi(H_i)$, the position of the first non-zero entry in row $i$ under permutation $\pi$. Note that $H_i$ and $M_i$ are independent. Now for large $p$, $M_i$ behaves roughly like a geometric random variable with parameter $\delta_i$. Thus for $k$ with $X_{ik} \neq 0$,

$$\mathbb{E}_\pi(\mathbb{1}_{\{H_i=k\}} w_{\pi(k)}) = \mathbb{E}_\pi(\mathbb{1}_{\{H_i=k\}} w_{M_i}) \approx \frac{1}{p\delta_i}\sum_{\ell=1}^{p} w_\ell \delta_i (1-\delta_i)^{\ell-1} = \frac{1}{p}\sum_{\ell=1}^{p} w_\ell (1-\delta_i)^{\ell-1}.$$

If $w_{\ell+1} = p(-1)^\ell \kappa^{(\ell)}(1)/\ell!$ we see that the RHS resembles a Taylor series of $\kappa(\delta_i)$ about 1. In this way we can approximate a large family of row-scaled signals.

## 4. Prediction error

The approximation error results in the three previous sections allow us to derive bounds on the prediction errors for linear and logistic regression models with potentially row-scaled data. Here we will present results under the assumption of $q$ non-zero entries per row and also where the scaling function $\kappa$ is proportional to the square-root function

$$\kappa_0(\delta) = \sqrt{\delta_{\min}/\delta}. \tag{9}$$

However, all of the approximation error results can be extended to results on prediction error via general theorems on prediction error we present in Section D. In particular, Theorem 12 can be used to show that versions of the equal row sparsity results hold more generally with $q$ replaced by the average number of non-zeroes per row $\bar{q}$ provided $L$ is not excessively large.

### 4.1 Linear regression models

Assume we have the following approximately linear model:

$$Y_i = \alpha^* + \kappa(\delta_i)\mathbf{x}_i^T\boldsymbol{\beta}^* + \varepsilon_i, \qquad 1 = 1, \ldots, n. \tag{10}$$

Here $\alpha^*$ is the intercept and $\mathbf{x}_i \in [-1,1]^p$. We assume that the random noise $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ satisfies $\mathbb{E}(\varepsilon_i) = 0$, $\mathbb{E}(\varepsilon_i^2) = \sigma^2$ and $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$.

Our results here give bounds on a mean-squared prediction error (MSPE) of the form

$$\mathrm{MSPE}((\hat{\alpha}, \hat{\mathbf{b}})) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\pi},\boldsymbol{\Psi}}\{(\alpha^* + \kappa(\delta_i)\mathbf{x}_i^T\boldsymbol{\beta}^* - \hat{\alpha} - \mathbf{S}\hat{\mathbf{b}})^2\} \tag{11}$$

where $\hat{\alpha}$ and $\hat{\mathbf{b}}$ are the estimated intercept and regression coefficients arising from regression on $\mathbf{S}$. Note we consider a denoising-type error: the error on the data used to fit the regression coefficients. Bounds on the prediction error at new observations would require conditions on the distribution of observations and we have avoided making any such assumptions for the results here.

### 4.1.1 Ordinary least squares

Perhaps the simplest way to estimate the linear model is to apply a least squares estimator,

$$(\hat{\alpha}, \hat{\mathbf{b}}) := \operatorname*{arg\,min}_{(\alpha,\mathbf{b})\in\mathbb{R}\times\mathbb{R}^{2^b L}} \|\mathbf{Y} - \alpha\mathbf{1} - \mathbf{S}\mathbf{b}\|_2^2, \tag{12}$$

to the matrix $\mathbf{S}$. We have the following theorem.

**Theorem 3** *Let $(\hat{\alpha}, \hat{\mathbf{b}})$ be the least squares estimator (12). We have the bound*

$$\mathrm{MSPE}((\hat{\alpha}, \hat{\mathbf{b}})) \;\leq\; \frac{C}{2^b L(1-2^{-b})}\;\|\boldsymbol{\beta}^*\|_b^2 + 2^b L\frac{\sigma^2}{n}.$$

*For equal row sparsity $\delta$ we have $C = (2-\delta)q$. For unequal row sparsity, when $\kappa = \kappa_0$ as in* (9)*, the result holds for $C = q_{\min}\log\{4\log(L)/\delta_{\min}\}$.*

An optimal choice $L_b^*$ of $L$ will balance the approximation error and variance contributions (first and second term on the right hand side respectively). In the equal row sparsity we arrive at

$$L_b^* = \frac{\sqrt{(2-\delta)qn}}{2^b\sqrt{1-2^{-b}}}\|\boldsymbol{\beta}^*\|_b$$

which yields an optimal MSPE of the order $\sigma\sqrt{q/n}\|\boldsymbol{\beta}^*\|_b$. If we ignore log terms the rate is analogous in the case of uneven row-sparsity. The slow rate in $n$ seems unavoidable if we do not make stronger conditions on the design. Indeed, a similar error rate is obtained in Theorem 21 of Maillard and Munos (2012) and in Kaban (2014) for OLS following dimension reduction by random projections. More precisely: projecting $K$ times with a random projection, followed by an OLS estimation is shown in Kaban (2014) to lead to a bound on MSPE of

$$\frac{1}{K}\;\|\boldsymbol{\beta}^*\|_\kappa^2 + K\frac{\sigma^2}{n}, \tag{13}$$

where the norm $\|\cdot\|_\kappa$ depends on the eigenvalue structure of the design matrix. In contrast the bound we have above for min-wise hashing depends in contrast on the sparsity $q$ through the constant $C$. The bound (13) is otherwise structurally identical to the bound for $b$-bit min-wise hashing above, and the role of the number $L$ of projections is now taken by the number $K$ of random projections. The optimal values of $K$ and $L$ are both of order $\sqrt{n}$, leading to the same convergence rate of the risk as $n \to \infty$.

To better understand the implications of Theorem 3, it is helpful to fix the size of the signal so that $\|\mathbf{X}\boldsymbol{\beta}^*\|_2^2/n = 1$, and look at whether we can show consistency of the method as both $p, n \to \infty$. If the signal is spread out and all variables have the same sparsity, $\|\boldsymbol{\beta}^*\|_b$ will be of order $\sqrt{p/q}$ and the MSPE will vanish when $p/n \to 0$, which excludes the high-dimensional setting.

However, now assume that the signal is concentrated on a fixed set of variables. The norm $\|\boldsymbol{\beta}^*\|_b$ is then constant as $p$ increases and all that is required for consistency is $q/n \to 0$ (or $q_{\min}/n \to 0$ for the more general case of uneven row-sparsity).

An interesting scenario is one of increasing variable sparseness. In many applications, the more predictor variables are added the sparser they tend to become. In text analysis, the first block of predictor variables might encode the presence of individual words. The next block might code for bigrams and the following, higher order $N$-grams. With this design, predictor variables in each successive block become sparser than the previous. It is then interesting to consider how much the MSPE can increase if we add a block with many sparse variables which contain no additional signal contribution. The result above indicates that the MSPE only increases as $\sqrt{q}$. Adding a block of several million (sparse) bigrams might thus have the same statistical effect as adding several thousand (denser) unigrams (individual words).

We now comment the optimal choice of $L$ and computational complexity. If we assume fixed $\|\boldsymbol{\beta}^*\|_2$ and $n = O(q)$, which is all that would be required to keep the prediction error bounded asymptotically, then the optimal dimension of the min-wise projection scales as $L_b^* = O(q)$, considering $b$ fixed here. This dimension will in general be a substantial reduction over the original dimension of the data, $p$, and would result in a correspondingly large reduction in the computational cost of regression. Indeed, ridge regression or the LAR algorithm (Efron et al., 2004) applied to $\mathbf{X}$ would have complexity $O(q^2p)$, and one would expect that the Lasso (Tibshirani, 1996) would have similar computational cost. In contrast, OLS applied to $\mathbf{S}$ would only require $O(q^3)$ operations, an improvement of $q/p$. The discussion above considered an optimal choice of $L \approx L_b^*$. Even if we cannot afford to work with the optimal dimension $L_b^*$ for computational reasons, the bound will still be useful for smaller values of $L$. The guarantee on prediction accuracy could not be obtained if, for example, simply a random subset of $L$ predictors were chosen and the remaining ones discarded.

The dependence of the bound on $b$ is also interesting: a minimum value occurs for $b = 1$. However, this would imply a larger value of $L_b^*$. Note the memory requirement for storing $\mathbf{S}$ would be $O(nL_b^*b)$ as $b$ bits would be required to store the locations of each of the $nL_b^*$ nonzeroes. We see that with a constraint on $nbL$ or on the number of permutations $L$, larger values of $b$ are more favourable, particularly with high sparsity, as this would tend to make $\|\boldsymbol{\beta}^*\|_b$ not much larger than $\|\boldsymbol{\beta}^*\|_2$. A different perspective on the optimal choice of $b$ based on the variance of inner products of rows of $\mathbf{S}$ is taken in Li and König (2011), with similar conclusions.

### 4.1.2 Ridge regression

Instead of using a least-squares estimator on the transformed data matrix $\mathbf{S}$ we can also apply ridge regression (Hoerl and Kennard, 1970). For a given $\lambda > 0$, the regression coefficients are found by

$$(\hat{\alpha}_\lambda, \hat{\mathbf{b}}_\lambda) := \underset{(\alpha,\mathbf{b})\in\mathbb{R}\times\mathbb{R}^L}{\arg\min} \; \|\mathbf{Y} - \hat{\alpha}\mathbf{1} - \mathbf{S}\mathbf{b}\|_2^2 \;\text{ such that }\; \|\mathbf{b}\|_2^2 \leq \lambda, \tag{14}$$

The theorem below gives a bound on the MSPE of $(\hat{\alpha}, \hat{\mathbf{b}}_\lambda)$.

**Theorem 4** *There exist regularisation parameters $\lambda$ depending on $\boldsymbol{\beta}^*$ and $\mathbf{S}$ such that*

$$\mathrm{MSPE}((\hat{\alpha}_\lambda, \hat{\mathbf{b}}_\lambda)) \;\leq\; \sigma\sqrt{\frac{2C}{(1-2^{-b})n}\|\boldsymbol{\beta}^*\|_2 + \frac{C}{2^b L(1-2^{-b})}\|\boldsymbol{\beta}^*\|_b^2 + \frac{\sigma^2}{n}}.$$

*Here the value of $C$ is defined as in Theorem 3 by $C = (2-\delta)q$ for equal row sparsity $\delta$ and $C = q_{\min}\log\{4\log(L)/\delta_{\min}\}$ for $\kappa = \kappa_0$ and unequal row-sparsity.*

The ridge regression result for large $L$ is similar to that for OLS with an optimal $L_b^*$, though there is a small difference: the leading terms are $\sigma\|\boldsymbol{\beta}^*\|_2\sqrt{q/n}$ and $\sigma\|\boldsymbol{\beta}^*\|_b\sqrt{q/n}$ respectively. Ridge regression takes advantage of the fact that not only do we have a $\mathbf{b}^*$ such that $\mathbf{S}\mathbf{b}^*$ and $\mathbf{X}\boldsymbol{\beta}^*$ are close, we also know that there is a $\mathbf{b}^*$ with this property that has low $\ell_2$-norm. Our bound on the expected squared $\ell_2$-norm of $\mathbf{b}^*$ ((ii) in Theorem 1) does not depend much on $b$. In contrast, OLS only makes use of the approximation error result, (iii) in Theorem 1.

Note that when $L$ is large, regardless of the value of $b$, ridge regression on $\mathbf{S}$ approximates a kernel ridge regression using the resemblance kernel (see Section 2.4). The MSPE of a kernel ridge regression with the resemblance kernel should of course not depend on $b$, and this observation largely agrees with our result.

Another key difference between ridge regression and OLS here is the following: achieving a good prediction error with OLS hinges on a careful choice of $L$. In contrast, with ridge regression, $L$ can (and should) be chosen very large, from a purely statistical point of view. However, the constraint on the $\ell_2$-norm of $\hat{\mathbf{b}}$ needs to be chosen carefully with ridge regression, typically by cross-validation. In practice, the number $L$ of dimensions can be chosen as large as possible according to the available computational budget.

### 4.2 Logistic regression

We give an analogous result to Theorem 4 for classification problems under logistic loss. Let $\mathbf{x}_i \in [-1,1]^p$ and let $\mathbf{Y} \in \{0,1\}^n$ be an associated vector of class labels. We assume the model

$$Y_i \sim \text{Bernoulli}(p_i); \qquad \log\left(\frac{p_i}{1-p_i}\right) = \kappa(\delta_i)\mathbf{x}_i^T\boldsymbol{\beta}^*, \tag{15}$$

with the $Y_i$ independent for $i = 1, \dots, n$. Note that we have omitted the separate intercept term for simplicity.

Here we consider a linear classifier constructed by $\ell_2$-constrained logistic regression. One can obtain a similar result for unconstrained logistic regression based on Lemma 6.6 of Bühlmann and van de Geer (2011), but we do not pursue this further here. Define

$$\hat{\mathbf{b}}_\lambda = \arg\min_{\mathbf{b}} \frac{1}{n}\sum_{i=1}^n \left[-Y_i\mathbf{s}_i^T\mathbf{b} + \log\{1+\exp(\mathbf{s}_i^T\mathbf{b})\}\right] \text{ such that } \|\mathbf{b}\|_2^2 \le \lambda. \tag{16}$$

Let $\mathcal{E}(\hat{\mathbf{b}}_\lambda)$ denote the excess risk of $\hat{\mathbf{b}}_\lambda$ under logistic loss, so

$$\mathcal{E}(\hat{\mathbf{b}}_\lambda) = \frac{1}{n}\sum_{i=1}^n \left[-p_i\mathbf{s}_i^T\hat{\mathbf{b}}_\lambda + \log\{1+\exp(\mathbf{s}_i^T\hat{\mathbf{b}}_\lambda)\}\right] - \frac{1}{n}\sum_{i=1}^n \left[-p_i\kappa(\delta_i)\mathbf{x}_i^T\boldsymbol{\beta}^* + \log\{1+\exp(\kappa(\delta_i)\mathbf{x}_i^T\boldsymbol{\beta}^*)\}\right]. \tag{17}$$

We can now state the analogous result to Theorem 4.

**Theorem 5** *Define $\tilde{p} \in \mathbb{R}$ by*

$$\tilde{p} := \frac{1}{n}\sum_{i=1}^n p_i(1-p_i) \le \frac{1}{2}. \tag{18}$$

*Then we have that there exists a $\lambda$ depending $\boldsymbol{\beta}^*$ and $\mathbf{S}$ such that*

$$\mathbb{E}_{\mathbf{Y},\boldsymbol{\pi},\boldsymbol{\Psi}}\{\mathcal{E}(\hat{\mathbf{b}}_\lambda)\} \leq \sqrt{\frac{2\tilde{p}C}{(1-2^{-b})n}}\|\boldsymbol{\beta}^*\|_2 + \frac{C}{2^{b+2}L(1-2^{-b})}\|\boldsymbol{\beta}^*\|_b^2.$$

*Here the value of $C$ is defined as in Theorem 3 by $C = (2-\delta)q$ for equal row sparsity $\delta$ and $C = q_{\min}\log\{4\log(L)/\delta_{\min}\}$ for $\kappa = \kappa_0$ and unequal row-sparsity.*

The result illustrates that the usefulness of $b$-bit min-wise hashing is not limited to regression problems. In fact, most applications of are classification problems (Li and König, 2011) and our analysis of $b$-bit min-wise hashing here gives a theoretical explanation for its performance in these cases.

## 5. Interaction models

One of the compelling aspects of regression and classification with $b$-bit min-wise hashing is the fact that a particular form of interactions between variables can be fitted. This does not require any change in the procedure other than a possible increase in $L$. To be clear, in order to capture interactions with $b$-bit min-wise hashing, just as in the main effects case, we create a reduced matrix $\mathbf{S}$ and then fit a main effects model to $\mathbf{S}$. The dimension of the compressed data, $2^bL$, can still be substantially smaller than the $O(p^2)$ number of coefficients that would need to be estimated if the interactions were modelled in the conventional way, and so the resulting computational advantage can be very large.

Note that in situations where the number of original predictors, $p$, may be manageable, including interactions explicitly can quickly become computationally infeasible. For example, if we start with, $10^5$ variables, the two-way interactions number more than a billion. For larger values of $p$, even methods such as Random Forest (Breiman, 2001) or Rule Ensembles (Friedman and Popescu, 2008) would suffer similar computational problems.

We now describe a type of interaction model that can be fitted with $b$-bit min-wise hashing. Let $\mathbf{f}^* \in \mathbb{R}^n$ be given by

$$f_i^* = \sum_{k=1}^p X_{ik}\theta_k^{*,(1)} + \sum_{k,k_1=1}^p X_{ik}\mathbb{1}_{\{X_{ik_1}=0\}}\Theta_{k,k_1}^{*,(2)}, \quad i = 1, \ldots, n, \tag{19}$$

where $\boldsymbol{\theta}^{*,(1)} \in \mathbb{R}^p$ is a vector of coefficients for the main effects terms, and $\boldsymbol{\Theta}^{*,(2)} \in \mathbb{R}^{p\times p}$ is a matrix of coefficients for interactions whose diagonal entries are zero. As elsewhere in the paper, throughout this section we will assume that $\mathbf{X} \in [-1,1]^{n\times p}$. Note that if $\mathbf{X}$ were a binary matrix, then (19) parametrises (in fact over-parametrises) all linear combinations of bivariate functions of predictors; that is all possible two-way interactions are included in the model.

In general, the interaction model includes the tensor product of the set of original variables with the columns of an $n \times p$ matrix with $ik$th entry $\mathbb{1}_{\{X_{ik}=0\}}$. The value zero is thus given a special status and the model seems particularly appropriate in the sparse design setting we are considering here.

### 5.1 Approximation error

We will assume that the number of non-zero entries in each row of $\mathbf{X}$ is $q \geq 1$. However, we believe our proof techniques can be extended to the unequal sparsity and unknown row scaling scenario dealt with in Section 3.2. Furthermore, for technical reasons, we assume here that $p \geq 3$.

Let $\mathbf{\Theta}^*$ collect together $\boldsymbol{\theta}^{*,(1)}$ and $\mathbf{\Theta}^{*,(2)}$ and define the following norms analogously to (5):

$$\|\mathbf{\Theta}^*\| := \|\boldsymbol{\theta}^{*,(1)}\|_2 + \left(2(2-\delta)q \sum_{k,k_1,k_2} \left|\Theta^{*,(2)}_{kk_1}\Theta^{*,(2)}_{kk_2}\right|\right)^{1/2}, \tag{20}$$

$$\|\mathbf{\Theta}^*\|_b := \|\boldsymbol{\theta}^{*,(1)}\|_b + \left\{2(2-\delta)q\left(\sum_{k,k_1,k_2} \left|\Theta^{*,(2)}_{kk_1}\Theta^{*,(2)}_{kk_2}\right| + \delta(2^b-2)\sum_{k,k_1,k_2} \frac{\|\mathbf{X}_k\|_2^2}{n}\left|\Theta^{*,(2)}_{kk_1}\Theta^{*,(2)}_{kk_2}\right|\right)\right\}^{1/2}. \tag{21}$$

**Theorem 6** *Suppose we have exactly q non-zero entries in each row of* $\mathbf{X}$. *Then there exists a vector* $\mathbf{b}^* \in \mathbb{R}^{2^b L}$ *with the following properties:*

*(i) The approximation is unbiased,* $\mathbb{E}_{\boldsymbol{\pi},\mathbf{\Psi}}(\mathbf{S}\mathbf{b}^*) = \mathbf{f}^*$.

*(ii) The* $\ell_2$*-norm is bounded by*

$$\mathbb{E}_{\boldsymbol{\pi},\mathbf{\Psi}}(\|\mathbf{b}^*\|_2^2) \leq \frac{(2-\delta)q}{L(1-2^{-b})}\|\mathbf{\Theta}^*\|^2.$$

*(iii) The approximation error is bounded by*

$$\mathbb{E}_{\boldsymbol{\pi},\mathbf{\Psi}}(\|\mathbf{S}\mathbf{b}^* - \mathbf{f}^*\|_2^2)/n \leq \frac{(2-\delta)q}{2^b L(1-2^{-b})}\|\mathbf{\Theta}^*\|_b^2.$$

The bound on the approximation error in (iii) is most suited to situations where there are a fixed number of interaction terms, so

$$\sum_{k,k_1,k_2} \left|\Theta^{*,(2)}_{kk_1}\Theta^{*,(2)}_{kk_2}\right| = O(1). \tag{22}$$

Then we see that the contribution of the interaction terms to the bound on the approximation error is of order $q^2$. On the other hand, if we are considering a growing number of many small interaction terms, much tighter bounds than that given by (iii) can be obtained. The bounds above show in particular that the form of function given by (19) lies in the RKHS of the resemblance kernel and its RKHS norm is upper bounded by $(2-\delta)q\|\mathbf{\Theta}^*\|^2$; further details are given in the appendix Section B.

The results for interaction models corresponding to Theorems 3, 4 and 5 now follow.

### 5.2 Prediction error

We now present results for linear and logistic regression models where the signal involves interactions.

### 5.2.1 Linear regression models

Assume the model (10) and define the MSPE by (11) but in both cases with $\mathbf{X}\boldsymbol{\beta}^*$ now replaced by $\mathbf{f}^*$ (19). As in the previous section, we will assume that $\mathbf{X}$ has $q$ non-zero entries in each row. When OLS estimation is used, we have the following result.

**Theorem 7** *Let $(\hat{\alpha}, \hat{\mathbf{b}})$ be the least squares estimator (12). Then*

$$\mathrm{MSPE}((\hat{\alpha}, \hat{\mathbf{b}})) \;\leq\; \frac{(2-\delta)q}{2^b L(1-2^{-b})}\|\boldsymbol{\Theta}^*\|_b^2 + 2^b L\frac{\sigma^2}{n}.$$

To interpret the result, consider a situation where there are a fixed number of interaction and main effects of fixed size, so in particular (22) holds. Then treating $b$ as fixed, the optimal $L$, $L^* = O(\sqrt{q^2 n/\sigma})$. If $n, q$ and $p$ increase by collecting new data and adding uninformative variables, then in order for the MSPE to vanish asymptotically, we require $q^2/n \to 0$. Compare this to the corresponding requirement of OLS applied to $\mathbf{X}$, that $p^2/n \to 0$. Particularly in situations of increasing variable sparseness, as discussed in Section 4.1.1, this can amount to a large statistical advantage.

The computational gains can be equally great. If, for example, $n \approx q^2$, then $L^* = O(q^2)$. If ridge regression were applied to $\mathbf{X}$ augmented by $O(p^2)$ interaction terms, the number of operations required would be $O(p^2 q^4)$; OLS using $\mathbf{S}$ has complexity $O(q^6)$. If instead $n \approx p^2$, then regression with explicitly coded interaction terms would have complexity $O(p^6)$, whilst with the compressed data this would be reduced to $O(p^4 q^2)$.

As in the main effects case, the ridge regression result is similar.

**Theorem 8** *Let the ridge regression estimator be given by* (14). *There exists $\lambda$ depending on $\mathbf{f}^*$ and $\mathbf{S}$ such that we have*

$$\mathrm{MSPE}((\hat{\alpha}, \hat{\mathbf{b}})) \;\leq \sigma\sqrt{\frac{(2-\delta)q}{n(1-2^{-b})}}\|\boldsymbol{\Theta}^*\| + \frac{(2-\delta)q}{2^b L(1-2^{-b})}\|\boldsymbol{\Theta}^*\|_b^2 + \frac{\sigma^2}{n}.$$

Similarly to Theorem 4 the result here suggests choosing a large $L$ is always better from a statistical point of view. However, for computational reasons, it may not be possible to take $L$ much larger than $L^*$.

### 5.2.2 Logistic regression

Here we assume the model (15) and define the excess risk by (17), but in both cases with $\mathbf{X}\boldsymbol{\beta}^*$ replaced by $\mathbf{f}^*$.

**Theorem 9** *Define $\tilde{p} \in \mathbb{R}$ as in* (18) *and the $\ell_2$-penalised logistic regression estimator as in* (16). *Then we have that there exists $\lambda$ such that*

$$\mathbb{E}_{\mathbf{Y},\boldsymbol{\pi},\boldsymbol{\Psi}}\{\mathcal{E}(\hat{\mathbf{b}}_\lambda)\} \leq \sigma\sqrt{\frac{\tilde{p}(2-\delta)q}{n(1-2^{-b})}}\|\boldsymbol{\Theta}^*\| + \frac{(2-\delta)q}{2^{b+2}L(1-2^{-b})}\|\boldsymbol{\Theta}^*\|_b^2.$$

One could continue to look at higher-order interaction models by adding three-way interactions in (19) and adapting (20) and (21) in suitable ways. However, being able to show that two-way interaction models can be fitted with $b$-bit min-wise hashing may well be sufficient for most applications.

## 6. Extensions

We now describe some extensions to the methodology.

### 6.1 Variable importance

Typically prediction, rather than model selection, is the primary goal in large-scale applications with sparse data, one reason for this being that we cannot expect a very small subset of variables to approximate the signal well when the design matrix is sparse. Nevertheless, it is often illuminating to study the influence of specific variables or look for the variables that have the largest influence on predictions. Indeed, such study is often undertaken following applications of Random Forest (Breiman, 2001), where several variable importance measures allow practitioners to better interpret the fits produced.

We now describe how importance measures can be obtained for $b$-bit min-wise hashing as described in Section 2.3. Let $\hat{f} : \mathbb{R}^p \to \mathbb{R}$ be the regression function created following regression on $b$-bit min-wise hashed data, and let $\hat{f}_i := \hat{f}(\mathbf{x}_i)$. Furthermore, for $k = 1, \ldots, p$, let $\hat{f}^{(-k)} := \hat{f}(\mathbf{x}_i^{(-k)})$, where $\mathbf{x}_i^{(-k)}$ is equal to $\mathbf{x}_i$ but with $k$th component set to zero.

The vector $\hat{\mathbf{f}} - \hat{\mathbf{f}}^{(-k)}$ is the difference in predictions obtained when fitting to $\mathbf{X}$, and those obtained when fitting to $\mathbf{X}$ with the $k$th column set to zero. When the underlying model in $\mathbf{X}$ contains only main effects (10) and no structural error is present, we might expect that

$$\hat{\mathbf{f}} - \hat{\mathbf{f}}^{(-k)} \approx \boldsymbol{\beta}_k^* \mathbf{X}_k.$$

To obtain a measure of variable importance, one could look at the $\ell_2$-norm of $\hat{\mathbf{f}} - \hat{\mathbf{f}}^{(-k)}$, for example (Breiman, 2001).

The difference in predictions can be computed relatively easily by considering the $n \times 2^b L$ matrix $\tilde{\mathbf{S}}$ with entries given by $\tilde{S}_{ilc} = \tilde{S}_{i(c+(l-1)2^b} = X_{i\tilde{H}_{il}} \mathbb{1}_{\{\Psi_{\tilde{H}_{il}l}=c\}}$, where

$$\tilde{H}_{il} := \underset{k \in \mathbf{z}_i \setminus H_{il}}{\arg\min}\, \pi_l(k).$$

Thus $\tilde{H}_{il}$ is the variable index in $\mathbf{z}_i$ whose value under permutation $\pi_l$ is second smallest among $\{\pi_l(k) : k \in \mathbf{z}_i\}$. If $\mathbf{z}_i \setminus H_{il} = \emptyset$, we simply set $\tilde{S}_{il} = 0$. Then

$$\hat{f}_i - \hat{f}_i^{(-k)} = \sum_{l=1}^{L} \mathbb{1}_{\{H_{il}=k\}} \sum_{c=1}^{2^b} (S_{ilc} - \tilde{S}_{ilc}) \hat{b}_{lc}. \tag{23}$$

Note that we only need to store the $n \times L$ matrix $\mathbf{H}$ and $n \times 2^b L$ matrices $\mathbf{S}$ and $\tilde{\mathbf{S}}$ to compute the variable importance for all variables; moreover the latter matrices only have at most $nL$ non-zero entries each.

Interaction effects are not directly visible, but do manifest themselves in the form of a higher variability among $\{\hat{f}_i - \hat{f}_i^{(-k)} : \mathbf{x}_i \approx \mathbf{x}\}$, for any given value of $\mathbf{x}$, if variable $k$ is involved in an interaction term. In principle, one could attempt to detect this increased variability, but further investigation of this is beyond the scope of the current work.

### 6.2 Other fitting procedures

Here we have only considered OLS, ridge regression and $\ell_2$-penalised logistic regression as prediction methods after reducing the design matrix. However, it is also conceivable that other fitting procedures could be suitable. In particular, it would be interesting to look at matching pursuit, boosting and the Lasso, for which results in (Tropp, 2004; Bühlmann, 2006; Van De Geer, 2008) could be leveraged. Matching pursuit would have the computational advantage that the entire $\mathbf{S}$ matrix would not need to be held in memory. Instead, one could create the columns during the fitting process. Such an approach may be useful for problems where the dimension of the hashing-matrix, $2^b L$, needs to be very large to achieve a desired predictive accuracy.

## 7. Discussion

In this paper we have derived approximation error bounds for $b$-bit min-wise hashing. We were able to show that not only does $b$-bit min-wise hashing take advantage of sparsity in the design matrix computationally, it is also able to exploit this for improved statistical performance. In particular, the MSPE of regression following dimension reduction by $b$-bit min-wise hashing is of the form $\sqrt{q/n}\|\boldsymbol{\beta}^*\|_2$ if the data follow a linear model with coefficient vector $\boldsymbol{\beta}^*$ and $q$ is the average number of non-zero variables for an observation. The linear model can then be well-approximated by the low-dimensional $b$-bit min-wise hashed data if the norm of $\|\boldsymbol{\beta}^*\|_2$ is low, as occurs, for example if the signal is approximately replicated in distinct blocks of variables.

In addition, we have shown that more complicated models such as interaction models can be fitted by a regression on the hashed data matrix that contains only main effects. Though a larger dimension $L$ of the hashed data may be required than when approximating a main effects model, no further changes are needed to the procedure.

These bounds also reveal some of the predictive properties of the resemblance kernel, and provide an insight into the sorts of regression functions that have small norm in its associated RKHS. More generally, we believe that random feature expansions may well be useful as a theoretical tool to understand properties of otherwise intractable kernels. We expect to see more extensions and applications $b$-bit min-wise hashing and other random feature expansions, both as computational and theoretical tools, in the future.

## Acknowledgments

The first author was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1 and an EPSRC programme grant.

## Appendix A. Approximation error results

In this section we prove results on the approximation error presented in the main text (Theorems 1, 2 and 6) as well as an additional result on the approximation error of linear signals when row sparsity is not necessarily equal (Theorem 12).

### A.1 Preliminary results

We will let $q_i$ be the number of non-zeroes in the $i$th row of $\mathbf{X}$ and define $\delta_i = q_i/p$. We will assume that $q_i \geq 1$ for all $i$. For the proofs of results on approximation error in settings with just main effects, we will make use of the following lemma. This lemma formalises the ideas of the discussion at the end of Section 3.2, that the elements of $\mathbf{M}$ behave rather like geometric random variables.

**Lemma 10** *There exist random functions $\{g_l(k)\}_{l=1,\dots,L,\,k=1,\dots,p}$ defined on the same probability space as the permutations $\boldsymbol{\pi}$ with the following properties:*

*(i) The random variables $\{g_1(k)\}_{k=1,\dots,p}, \dots, \{g_L(k)\}_{k=1,\dots,p}$ are i.i.d. and are independent of $\boldsymbol{\Psi}$.*

*(ii) The rank of $g_l(k)$ among $g_l(1), \dots, g_l(p)$ taken in increasing order is $\pi_l(k)$.*

*(iii) Marginally $g_l(k) \sim Geo(p^{-1})$.*

*(iv) $G_{il} := \min_{k\in\mathbf{z}_i} g_l(k) = g_l(H_{il}) \sim Geo(\delta_i)$.*

*(v) $\mathbf{G}$ and $\mathbf{H}$ are independent.*

**Proof** First consider generating permutations $\boldsymbol{\pi}$ in the following way. Let $m \in \mathbb{N}$ and let $\sigma_1^{(m)}, \dots, \sigma_L^{(m)}$ be $L$ i.i.d. random permutations of $\{1, \dots, mp\}$. For $k = 1, \dots, p$, let

$$g_l^{(m)}(k) = \min_{a=0,\dots,m-1} \sigma_l^{(m)}(k + ap).$$

Note that the $g_l^{(m)}(k)$ are all distinct and any ordering of them is equally likely so they define a random permutation of $\{1, \dots, p\}$. Furthermore, for $j = 1, \dots, mp - m + 1$,

$$\mathbb{P}(g_l^{(m)}(k) = j) = \binom{mp-j}{m-1}\Big/\binom{mp}{m} = \frac{1}{p}\left(1 - \frac{1-m^{-1}}{p-m^{-1}}\right)\cdots\left(1 - \frac{1-m^{-1}}{p-(j-1)m^{-1}}\right).$$

Thus

$$\mathbb{P}(g_l^{(m)}(k) = j) \to \frac{1}{p}\left(1 - \frac{1}{p}\right)^{j-1}$$

as $m \to \infty$ for $j = 1, 2, \dots$. Similarly $G_{il}^{(m)} := \min_{k\in\mathbf{z}_i} g_l^{(m)}(k)$ has $\mathbb{P}(G_{il}^{(m)} = j) \to \delta_i(1-\delta_i)^{j-1}$ as $m \to \infty$. Note that $\mathbf{G}^{(m)}$ and $\mathbf{H}$ are independent. Thus

$$\{g_l^{(m)}(k)\}_{l=1,\dots,L,k=1,\dots,p} \xrightarrow{d} \{g_l(k)\}_{l=1,\dots,L,k=1,\dots,p}$$

as $m \to \infty$ with the random variables $g_l(k)$ having the properties given in the statement of the lemma. ■

In the proofs which follow, we will consider the permutations as having been generated as described by Lemma 10. We will let $\pi = \pi_1$, $M_i = M_{i1}$, $g = g_1$, $G_1 = G_{i1}$, $H_i = H_{i1}$ and $\boldsymbol{\psi} = \boldsymbol{\Psi}_1$. Let $C = 2^b$, $\nu = 2^{-b}$.

The next lemma introduces the general form of $\mathbf{b}^*$ that we will use for the main effects results. It also establishes results on the mean and variance of the approximation and gives a bound on $\mathbb{E}(\|\mathbf{b}^*\|_2^2)$; these will form the basis of the theorems to follow.

**Lemma 11** *For a given sequence of weights* $\{w_j\}_{j=1}^{\infty}$*, let* $\tilde{\mathbf{b}}^* \in \mathbb{R}^{LC}$ *be given by*

$$\tilde{b}^*_{lc} = \frac{1}{L}\sum_{k=1}^{p} \beta^*_k \frac{\mathbb{1}_{\{\Psi_{lk}=c\}} - \nu}{1-\nu} w_{g_l(k)}$$

*and let* $\mathbf{b}^* = \mathbb{E}(\tilde{\mathbf{b}}^*|\boldsymbol{\pi})$*. We have the following.*

*(i)*

$$\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\mathbf{s}_i^T\mathbf{b}^*) = \frac{1}{p}\mathbf{x}_i^T\boldsymbol{\beta}^*\sum_{\ell=1}^{\infty}(1-\delta_i)^{\ell-1}w_\ell.$$

*(ii)*

$$\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{b}^*\|_2^2) \le \frac{1}{pL(1-\nu)}\|\boldsymbol{\beta}^*\|_2^2\sum_{\ell=1}^{\infty}w_\ell^2. \tag{24}$$

*(iii)*

$$\mathrm{Var}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\mathbf{s}_i^T\mathbf{b}^*) \le \frac{1}{pL(1-\nu)}\left(\nu\|\boldsymbol{\beta}^*\|_2^2 + (1-2\nu)\sum_{k=1}^{p}X_{ik}^2{\beta_k^*}^2\right)\sum_{\ell=1}^{\infty}w_\ell^2. \tag{25}$$

**Proof** First note that

$$\mathbb{E}\left(\frac{\mathbb{1}_{\{\psi_k=\psi_j\}}-\nu}{1-\nu}\middle|\psi_j\right) = \begin{cases} 1 & \text{if } k=j \\ 0 & \text{otherwise} \end{cases} \tag{26}$$

$$\mathbb{E}\left(\frac{\mathbb{1}_{\{\psi_k=\psi_j\}}-\nu}{1-\nu}\frac{\mathbb{1}_{\{\psi_\ell=\psi_j\}}-\nu}{1-\nu}\middle|\psi_j\right) = \begin{cases} 1 & \text{if } k=\ell=j \\ 0 & \text{if } k\ne\ell \\ \frac{\nu}{1-\nu} & \text{otherwise.} \end{cases} \tag{27}$$

For (i), we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\mathbf{s}_i^T\mathbf{b}^*) &= \mathbb{E}_{g,\boldsymbol{\psi}}\left(\sum_{c=1}^{C}\sum_{j=1}^{p}X_{ij}\mathbb{1}_{\{H_i=j\}}\mathbb{1}_{\{\psi_j=c\}}\sum_{k=1}^{p}\beta^*_k\frac{\mathbb{1}_{\{\psi_k=c\}}-\nu}{1-\nu}w_{g(k)}\right)\\ &= \mathbb{E}_g\left(\sum_{k=1}^{p}X_{ik}\mathbb{1}_{\{H_i=k\}}\beta^*_k w_{g(k)}\right)\\ &= \frac{1}{q_i}\sum_{k=1}^{p}X_{ik}\beta^*_k\mathbb{E}(w_{G_i}),\end{aligned}$$

where to arrive at the second line we used (26).

Turning to (ii), note that each component of $\mathbf{b}^*$ has mean zero and so

$$\mathbb{E}({b^*_{lc}}^2) = \mathrm{Var}(b^*_{lc}) = \mathrm{Var}\{\mathbb{E}(\tilde{b}^*_{lc}|\boldsymbol{\pi})\} \le \mathrm{Var}(\tilde{b}^*_{lc}).$$

Now we have

$$\mathbb{E}_{g_1,\ldots,g_L,\boldsymbol{\Psi}}\|\tilde{\mathbf{b}}^*\|_2^2 = \frac{1}{L}\sum_{c=1}^{C}\sum_{k,\ell}\beta_k^*\beta_\ell^*\mathbb{E}\biggl(\frac{\mathbb{1}_{\{\psi_k=c\}}-\nu}{1-\nu}\frac{\mathbb{1}_{\{\psi_\ell=c\}}-\nu}{1-\nu}\biggr)\mathbb{E}(w_{g(k)}w_{g(\ell)})$$

Using (27), we get

$$\mathbb{E}_{g_1,\ldots,g_L,\boldsymbol{\Psi}}\|\tilde{\mathbf{b}}^*\|_2^2 = \frac{1}{L(1-\nu)}\sum_k \beta_k^{*2}\mathbb{E}(w_{g(k)}^2) \leq \frac{1}{pL(1-\nu)}\|\boldsymbol{\beta}^*\|_2^2\sum_{\ell=1}^{\infty}w_\ell^2.$$

For (iii) we argue as follows.

$$\begin{aligned}\mathrm{Var}(\mathbf{s}_i^T\mathbf{b}^*) &\leq \mathrm{Var}(\mathbf{s}_i^T\tilde{\mathbf{b}}^*)\\ &\leq \frac{1}{L}\mathbb{E}_{g,\boldsymbol{\psi}}\biggl(X_{iH_i}^2\sum_{k,\ell}\beta_k^*\beta_\ell^*\frac{\mathbb{1}_{\{\psi_k=\psi_{H_i}\}}-\nu}{1-\nu}\frac{\mathbb{1}_{\{\psi_\ell=\psi_{H_i}\}}-\nu}{1-\nu}w_{g(k)}w_{g(\ell)}\biggr)\end{aligned}$$

Using (27) and the fact that $\mathbf{X}\in[-1,1]^{n\times p}$, we have

$$\begin{aligned}\mathrm{Var}(\mathbf{s}_i^T\mathbf{b}^*) &\leq \frac{1}{L}\mathbb{E}\biggl\{X_{iH_i}^2\biggl(\frac{\nu}{1-\nu}\sum_{k=1}^{p}(\beta_k^*)^2w_{g(k)}^2 + \frac{1-2\nu}{1-\nu}(\beta_{H_i}^*)^2w_{G_i}^2\biggr)\biggr\} \qquad (28)\\ &\leq \frac{1}{L(1-\nu)}\biggl\{\nu\sum_{k=1}^{p}\beta_k^{*2}\mathbb{E}(w_{g(k)}^2) + \frac{1-2\nu}{q_i}\mathbb{E}(w_{G_i}^2)\sum_{k=1}^{p}X_{ik}^2\beta_k^{*2}\biggr\}.\end{aligned}$$

The result then follows as

$$\sum_{\ell=1}^{\infty}w_\ell^2 \geq \mathbb{E}(w_{g(k)}^2) = \frac{1}{p}\sum_{\ell=1}^{\infty}w_\ell^2\biggl(1-\frac{1}{p}\biggr)^{\ell-1} \geq \frac{\delta_i}{q_i}\sum_{\ell=1}^{\infty}w_\ell^2(1-\delta_i)^{\ell-1} = \frac{\mathbb{E}(w_{G_i}^2)}{q_i}.$$

■

### A.2 Proof of Theorem 1

We use a $\mathbf{b}^*$ and $\tilde{\mathbf{b}}^*$ as in Lemma 11 but here we choose the weights $w_\ell$ so as to minimise $\sum_{\ell=1}^{\infty}w_\ell^2$ (a term which features in our upper bounds on the variance and $\mathbb{E}(\|\mathbf{b}^*\|_2^2)$) subject to the unbiasedness constraint (i). The unbiasedness constraint amounts to

$$\sum_{\ell=1}^{\infty}(1-\delta)^{\ell-1}w_\ell = p.$$

Performing the minimisation with this constraint yields

$$w_\ell = p\frac{(1-\delta)^{\ell-1}}{\sum_{\ell=1}^{\infty}(1-\delta)^{2\ell-2}}.$$

With this choice we have

$$\sum_{\ell=1}^{\infty}w_\ell^2 = p^2\biggl(\sum_{\ell=1}^{\infty}(1-\delta)^{2\ell-2}\biggr)^{-1} = p^2\{1-(1-\delta)^2\} = (2-\delta)qp.$$

Substituting into (24) and (25) then yields the result.

### A.3 Proof of Theorem 2

We use a $\mathbf{b}^*$ and $\tilde{\mathbf{b}}^*$ as in Lemma 11 but here we take

$$w_{\ell+1} = p(-1)^\ell \frac{\kappa^{(\ell)}(1)}{\ell!}\{\mathbb{1}_{\{\ell \leq \lfloor m \rfloor\}} + (m - \lfloor m \rfloor)\mathbb{1}_{\{\ell = \lceil m \rceil\}}\}$$

where $m > 0$ is a parameter to be chosen. Thus the weights correspond to coefficients from a truncated Taylor series expansion of $\kappa$ about 1. We have

$$\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}[\{(\delta_{\min}/\delta_i)^a \mathbf{x}_i^T \boldsymbol{\beta}^* - \mathbf{s}_i^T \mathbf{b}^*\}^2] = \{(\delta_{\min}/\delta_i)^a \mathbf{x}_i^T \boldsymbol{\beta}^* - \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\mathbf{s}_i^T \mathbf{b}^*)\}^2 + \mathrm{Var}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\mathbf{s}_i^T \mathbf{b}^*).$$

We first bound the variance term by bounding the squared sum of the sequence of weights. To this end, we note that by Lemma 20

$$\frac{\delta_{\min}^{-2a}}{p^2} \sum_{\ell=1}^{\infty} w_\ell^2 \leq 1 + a^2 + a^2 e^{2a}\Bigg(\sum_{\ell=2}^{\lfloor m \rfloor} \frac{1}{\ell^{2(1-a)}} + \frac{m - \lfloor m \rfloor}{\lceil m \rceil^{2(1-a)}}\Bigg).$$

Now

$$\begin{aligned}\sum_{\ell=2}^{\lfloor m \rfloor} \frac{1}{\ell^{2(1-a)}} + \frac{m - \lfloor m \rfloor}{\lceil m \rceil^{2(1-a)}} &\leq \int_1^m \frac{1}{\ell^{2(a-1)}} d\ell \\ &= \begin{cases} \frac{m^{2a-1}-1}{2a-1} & \text{if } a \neq 1/2 \\ \log(m) & \text{if } a = 1/2. \end{cases}\end{aligned}$$

Let

$$\tau_a(m) = \begin{cases} e\log(me^{5/e})/4 & \text{if } a = 1/2, \\ a^2 e^{2a} m^{2a-1}/(2a-1) & \text{if } 1/2 < a \leq 1. \end{cases}$$

Then

$$\sum_{\ell=1}^{\infty} w_\ell^2 \leq p^2 \delta_{\min}^{2a} \tau_a(m). \tag{29}$$

The variance is then at most

$$\delta_{\min}^{2a} \tau_a(m) \frac{p}{L(1-\nu)} \Bigg(\nu \|\boldsymbol{\beta}^*\|_2^2 + (1-2\nu)\sum_{k=1}^{p} X_{ik}^2 \beta_k^{*2}\Bigg).$$

Turning now to the bias term, note first that by (i) of Lemma 11, this is equal to

$$(\mathbf{x}_i^T \boldsymbol{\beta}^*)^2 \Bigg\{(\delta_{\min}/\delta_i)^a - \frac{1}{p}\sum_{\ell=1}^{\infty}(1-\delta_i)^{\ell-1} w_\ell\Bigg\}^2. \tag{30}$$

We see this is bounded above by

$$\delta_{\min}^{2a}(\mathbf{x}_i^T \boldsymbol{\beta}^*)^2 \Bigg\{ae^a\Bigg(\sum_{\ell=\lceil m \rceil}^{\infty}(1-\delta_i)^\ell \frac{1}{\ell^{1-a}}\Bigg)\Bigg\}^2.$$

Now

$$\sum_{\ell=\lceil m\rceil}^{\infty}(1-\delta_i)^{\ell}\frac{1}{\ell^{1-a}} \le \frac{e^{-\delta_i m}}{m^{1-a}\delta_i}.$$

By the Cauchy–Schwarz inequality (assuming $X_{ij} \in [-1, 1]$)

$$\frac{(\mathbf{x}_i^T\boldsymbol{\beta}^*)^2}{\delta_i} = \frac{1}{\delta_i}\bigg(\sum_{k\in\mathbf{z}_i} X_{ik}\beta_k^*\bigg)^2 \le p\sum_{k=1}^{p} X_{ik}^2\beta_k^{*2} \le p\|\boldsymbol{\beta}^*\|_2^2.$$

Thus the squared bias is at most

$$\frac{p}{1-\nu}\frac{a^2e^{2a}}{m^{1-2a}}\max_{i=1,\dots,n}\left(\frac{e^{-2\delta_i m}}{m\delta_i}\right)\bigg(\nu\|\boldsymbol{\beta}^*\|_2^2 + (1-2\nu)\sum_{k=1}^{p}X_{ik}^2\beta_k^{*2}\bigg).$$

Therefore the MSE (now averaging over the observations) is bounded by the minimum over $m > 0$ of

$$\frac{p}{L(1-2^{-b})}\delta_{\min}^{2a}\bigg\{\tau_a(m) + \frac{a^2e^{2a}}{m^{1-2a}}\max_{i=1,\dots,n}\left(\frac{e^{-2\delta_i m}}{m\delta_i}\right)\bigg\}\|\boldsymbol{\beta}^*\|_b^2.$$

For $a = 1/2$, we set $m = \log(L)/\{2\delta_{\min}\}$. This yields

$$\begin{aligned}\min_{m>0}\bigg\{\tau_{1/2}(m) + \frac{Le}{4}\max_{i=1,\dots,n}\left(\frac{e^{-2\delta_i m}}{m\delta_i}\right)\bigg\} &\le \frac{e}{4}\bigg\{\log\left(\frac{\log(L)e^{5/e}}{2\delta_{\min}}\right) + \frac{2}{\log(L)}\bigg\}\\ &\le \log\{4\log(L)/\delta_{\min}\}\end{aligned}$$

provided $L \ge 10$ and $\delta_{\min} \le 1/2$. Finally the bound for $a > 1/2$ comes from setting

$$m = \frac{1}{2}\log\{2(2a-1)L\}/\delta_{\min}$$

which gives

$$\begin{aligned}\min_{m>0}\bigg\{\tau_a(m) + \frac{La^2e^{2a}}{m^{1-2a}}\max_{i=1,\dots,n}\left(\frac{e^{-2\delta_i m}}{m\delta_i}\right)\bigg\} &\le \frac{\delta_{\min}^{1-2a}a^2e^{2a}}{2^{2a-1}(2a-1)}[\log\{2(2a-1)L\}]^{2a-2}\log\{2(2a-1)eL\}\\ &\le \frac{4\delta_{\min}^{1-2a}}{1-2a}[\log\{2(2a-1)L\}]^{2a-1}\end{aligned}$$

for $L \ge 2/(1-2a)$. Using the bounds on $\tau_a$ with these choices of $m$ and (29), we obtain the bounds on $\mathbb{E}(\|\mathbf{b}^*\|_2^2)$ by substituting into (24).

### A.4 Unequal row sparsity and constant row-scaling

Here we prove results indicated after the presentation of Theorem 1 in Section 3.1. When the scaling function is simply the constant 1, the spread of the $\delta_i$ becomes more critical in determining how well the signal can be approximated. Define

$$\bar{\delta} = \frac{1}{n}\sum_{i=1}^{n}\delta_i,$$

$$\mathcal{V}(\boldsymbol{\delta}) = \frac{1}{\|\mathbf{X}\boldsymbol{\beta}^*\|_2^2}\sum_{i=1}^{n}(\mathbf{x}_i^T\boldsymbol{\beta}^*)^2(\delta_i-\bar{\delta})^2.$$

**Theorem 12** *Suppose*

$$2^b L(1-2^{-b}) \leq \frac{p(2\bar{\delta})^3\|\boldsymbol{\beta}^*\|_b^2}{\|\mathbf{X}\boldsymbol{\beta}^*\|_2^2\mathcal{V}(\boldsymbol{\delta})/n}. \tag{31}$$

*Then there exists* $\mathbf{b}^* \in \mathbb{R}^L$ *such that the approximation error satisfies*

$$\frac{1}{n}\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}\{\|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{S}\mathbf{b}^*\|_2^2\} \leq \frac{6p\bar{\delta}}{2^b L(1-2^{-b})}\|\boldsymbol{\beta}^*\|_b^2, \tag{32}$$

*and*

$$\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{b}^*\|_2^2) \leq \frac{2\bar{q}}{L(1-2^{-b})}\|\boldsymbol{\beta}^*\|_2^2. \tag{33}$$

Provided $2^b L$ is not too large, we recover essentially the same approximation error bound as Theorem 1 up to a constant factor, but with the row sparsity replaced by the average row sparsity $\bar{\delta}$. In the simple situation where the entries of $\mathbf{X}$ are realisations of i.i.d. Bernoulli random variables with probability $\delta$, we would have $\bar{\delta} \approx \delta$, $\|\mathbf{X}\boldsymbol{\beta}^*\|_2^2/n \approx \delta\|\boldsymbol{\beta}^*\|_2^2$ and $\mathcal{V}(\boldsymbol{\delta}) \approx \delta/p$. Substituting these values into the requirement on $2^b L$ shows that the condition reduces to $2^b L \leq 8p^2\delta\{1+(2^b-2)\delta\}$. Note that typically one would choose $2^b L$ of the order $\bar{\delta}p$. More generally, provided $\mathcal{V}(\boldsymbol{\delta})$ and $\|\mathbf{X}\boldsymbol{\beta}^*\|_2^2/\|\boldsymbol{\beta}^*\|_2^2$ are small, we can expect that the bound of Theorem 1 will hold true, up to a constant factor.

Proof of Theorem 12

We use a $\mathbf{b}^*$ and $\tilde{\mathbf{b}}^*$ as in Lemma 11 taking

$$w_\ell = p(1-\bar{\delta})^{\ell-1}\mathbb{1}_{\{\ell\leq m\}}\frac{\bar{\delta}(2-\bar{\delta})}{1-(1-\bar{\delta})^{2m}}.$$

where $m \in \mathbb{N}$ is a parameter to be chosen. This gives

$$\frac{1}{p^2}\sum_{\ell=1}^{\infty} w_\ell^2 = \frac{\bar{\delta}(2-\bar{\delta})}{1-(1-\bar{\delta})^{2m}},$$

which gives us a bound on the variance term.

Lemma 11 (i) gives the expression for the bias term. To bound this, first note that

$$\frac{1}{p}\sum_{\ell=1}^{m}(1-\bar{\delta})^{\ell-1}w_\ell = 1.$$

Next

$$\begin{aligned}\left[\sum_{\ell=1}^{m}(1-\bar{\delta})^{\ell-1}\{(1-\bar{\delta})^{\ell-1}-(1-\delta_i)^{\ell-1}\}\right]^2 &= (\delta_i-\bar{\delta})^2\left[\sum_{\ell=1}^{m}(1-\bar{\delta})^{\ell-1}\sum_{k=0}^{\ell-2}(1-\bar{\delta})^k(1-\delta_i)^{\ell-2-k}\right]^2 \\ &\leq (\delta_i-\bar{\delta})^2\left(\sum_{\ell=1}^{m}(1-\bar{\delta})^{\ell-1}(\ell-1)\right)^2 \\ &= \min\left\{\frac{m(m-1)}{2}, \frac{1}{\bar{\delta}^2}\right\}^2(\delta_i-\bar{\delta})^2.\end{aligned}$$

Also note that as

$$(1-\bar{\delta})^{2m} \leq 1 - 2m\bar{\delta} + m(2m-1)\bar{\delta}^2$$

we have

$$\begin{aligned}\frac{\bar{\delta}(2-\bar{\delta})}{1-(1-\bar{\delta})^{2m}} &\leq \frac{2}{2m - m(2m-1)\bar{\delta}}\mathbb{1}_{\{m \leq 1/(2\bar{\delta})\}} + \frac{2}{1/\bar{\delta} - (1/\bar{\delta}-1)/2}\mathbb{1}_{\{m > 1/(2\bar{\delta})\}} \\ &\leq \max\left(\frac{2}{m+1/2}, \frac{4\bar{\delta}}{1+\bar{\delta}}\right)\end{aligned}$$

and for $m \leq 1/(2\bar{\delta}) + 1/2$,

$$\frac{m(m-1)}{2}\max\left(\frac{2}{m+1/2}, \frac{4\bar{\delta}}{1+\bar{\delta}}\right) \leq (m-1/2)\mathbb{1}_{\{1<m\leq 1/(2\bar{\delta})+1/2\}}.$$

Thus the overall approximation error is bounded above by the minimum over $m = 1, 2, \ldots, \lfloor 1/(2\bar{\delta}) + 1/2 \rfloor$ of

$$\mathbb{1}_{\{m>1\}}(m-1/2)^2\frac{1}{n}\|\mathbf{X}\boldsymbol{\beta}^*\|_2^2\mathcal{V}(\boldsymbol{\delta}) + \max\left(\frac{2}{m+1/2}, 4\bar{\delta}\right)\frac{p}{2^bL(1-2^{-b})}\|\boldsymbol{\beta}^*\|_b^2,$$

which in turn is bounded by the minimum over $m \in [0, 1/(2\bar{\delta})]$ of

$$m^2\frac{1}{n}\|\mathbf{X}\boldsymbol{\beta}^*\|_2^2\mathcal{V}(\boldsymbol{\delta}) + \frac{2}{m}\frac{p}{2^bL(1-2^{-b})}\|\boldsymbol{\beta}^*\|_b^2. \tag{34}$$

Optimising over $m > 0$ in the above then gives

$$m = \min\left\{\left(\frac{p\|\boldsymbol{\beta}^*\|_b^2}{2^bL(1-2^{-b})\|\mathbf{X}\boldsymbol{\beta}^*\|_2^2\mathcal{V}(\boldsymbol{\delta})/n}\right)^{1/3}, \frac{1}{2\bar{\delta}}\right\}.$$

The condition on $L$ (31) ensures that the minimum is achieved at $1/(2\bar{\delta})$. Substituting this value of $m$ into (34) then gives (32). For (33) we note that

$$\sum_{\ell=1}^{\infty} w_\ell^2 \leq 2p^2\bar{\delta};$$

the result follows using Lemma 11 (ii).

### A.5 Proof of Theorem 6

We let $\mathbf{b}^* = \mathbf{b}^{*,(1)} + \mathbf{b}^{*,(2)}$ where $\mathbf{b}^{*,(1)}$ is chosen in line with Theorem 1. Explicitly, let $\mathbf{b}^{*,(1)} = \mathbb{E}(\tilde{\mathbf{b}}^*|\boldsymbol{\pi})$ where

$$\tilde{b}_{lc}^* = \frac{p}{L}\sum_{k=1}^{p}\theta_k^{*,(1)}\frac{\mathbb{1}_{\{\Psi_{lk}=c\}}-\nu}{1-\nu}\frac{(1-\delta)^{g_l(k)-1}}{\sum_{\ell=1}^{\infty}(1-\delta)^{2\ell-2}}.$$

We construct $\mathbf{b}^{*,(2)}$ to approximate the interactions as follows. Let

$$b_{lc}^{*,(2)} = \frac{pq}{L}\sum_{k=1}^{p}\frac{\mathbb{1}_{\{\Psi_{lk}=c\}}-\nu}{1-\nu}\sum_{k_1=1}^{p}\Theta_{kk_1}^{*,(2)}\mathbb{1}_{\{\pi_l(k_1)<\pi_l(k)\}}w_{\pi_l(k)},$$

where $\mathbf{w} \in \mathbb{R}^p$ is a vector of weights to be chosen such that

$$\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\mathbf{s}_i^T \mathbf{b}^{*,(2)}) = \sum_{k,k_1} X_{ik} \mathbb{1}_{\{X_{ik_1}=0\}} \Theta^{*,(2)}_{kk_1}. \tag{35}$$

We compute

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\mathbf{s}_i^T \mathbf{b}^{*,(2)}) &= \frac{pq}{L} \sum_{l=1}^{L} \sum_{c=1}^{C} \mathbb{E}_{\pi_l,\Psi_l} \left( S_{ilc} \sum_{k=1}^{p} \frac{\mathbb{1}_{\{\Psi_{kl}=c\}} - \nu}{1-\nu} \sum_{k_1=1}^{p} \Theta^{*,(2)}_{kk_1} \mathbb{1}_{\{\pi_l(k_1)<\pi_l(k)\}} w_{\pi_l(k)} \right) \\
&= pq \mathbb{E}_{\pi,\boldsymbol{\psi}} \left( \sum_{c=1}^{C} \sum_{j=1}^{p} X_{ij} \mathbb{1}_{\{H_i=j\}} \mathbb{1}_{\{\psi_j=c\}} \sum_{k=1}^{p} \frac{\mathbb{1}_{\{\psi_k=c\}} - \nu}{1-\nu} \sum_{k_1=1}^{p} \Theta^{*,(2)}_{kk_1} \mathbb{1}_{\{\pi(k_1)<\pi(k)\}} w_{\pi(k)} \right) \\
&= pq \mathbb{E}_{\pi} \left( \sum_{k=1}^{p} X_{ik} \mathbb{1}_{\{H_i=k\}} \sum_{k_1=1}^{p} \Theta^{*,(2)}_{kk_1} \mathbb{1}_{\{\pi(k_1)<\pi(k)\}} \sum_{\ell=2}^{p} w_\ell \mathbb{1}_{\{\pi(k)=\ell\}} \right).
\end{aligned}$$

where in the final line we have appealed to (26). Now observe that for $k \in \mathbf{z}_i$,

$$\mathbb{1}_{\{H_i=k\}} \mathbb{1}_{\{\pi(k_1)<\pi(k)\}} \mathbb{1}_{\{\pi(k)=\ell\}} = \mathbb{1}_{\{X_{ik_1}=0\}} \mathbb{1}_{\{H_i=k\}} \mathbb{1}_{\{M_i=\ell,\, \pi(k_1)<\ell\}},$$

and $\mathbb{1}_{\{H_i=k\}}$ and $\mathbb{1}_{\{M_i=\ell,\, \pi(k_1)<\ell\}}$ are independent. Thus we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}((\mathbf{S}\mathbf{b}^{*,(2)})_i) &= \sum_{k,k_1} X_{ik} \mathbb{1}_{\{X_{ik_1}=0\}} \Theta^{*,(2)}_{kk_1} \sum_{\ell=1}^{p} p \mathbb{P}_\pi(M_i = \ell, \pi(k_1) < \ell) w_\ell \\
&= \sum_{k,k_1} X_{ik} \mathbb{1}_{\{X_{ik_1}=0\}} \Theta^{*,(2)}_{kk_1} \sum_{\ell=2}^{p} (\ell-1) \mathbb{P}_\pi(M_i = \ell | \, \pi(k_1) < \ell) w_\ell \\
&= \sum_{k,k_1} X_{ik} \mathbb{1}_{\{X_{ik_1}=0\}} \Theta^{*,(2)}_{kk_1} \sum_{\ell=2}^{p} (\ell-1) \frac{\binom{p-\ell}{q-1}}{\binom{p-1}{q}} w_\ell.
\end{aligned}$$

Thus if we choose $\mathbf{w}$ such that

$$\sum_{\ell=2}^{p} (\ell-1) \frac{\binom{p-\ell}{q-1}}{\binom{p-1}{q}} w_\ell = 1, \tag{36}$$

property (35) will be satisfied.

Next we compute

$$\begin{aligned}
\mathbb{E}(\|\mathbf{b}^{*,(2)}\|_2^2) &\leq \frac{p^2q^2}{L(1-\nu)}\sum_{k=1}^{p}\mathbb{E}\bigg\{\bigg(\sum_{k_1=1}^{p}\Theta^{*,(2)}_{kk_1}\mathbb{1}_{\{\pi(k_1)<\pi(k)\}}\bigg)^2 w_{\pi(k)}^2\bigg\} \\
&= \frac{p^2q^2}{L(1-\nu)}\sum_{k=1}^{p}\sum_{\ell=1}^{p} w_\ell^2\bigg(\sum_{k_1}(\Theta^{*,(2)}_{kk_1})^2\mathbb{P}(\pi(k)=\ell,\pi(k_1)<\ell) \\
&\qquad + \sum_{k_1\neq k_2}\Theta^{*,(2)}_{kk_1}\Theta^{*,(2)}_{kk_2}\mathbb{P}(\pi(k)=\ell,\pi(k_1)<\ell,\pi(k_2)<\ell)\bigg) \\
&= \frac{pq^2}{L(1-\nu)}\sum_{k=1}^{p}\sum_{\ell=2}^{p} w_\ell^2\bigg(\frac{\ell-1}{p-1}\sum_{k_1}(\Theta^{*,(2)}_{kk_1})^2 + \frac{(\ell-1)(\ell-2)}{(p-1)(p-2)}\sum_{k_1\neq k_2}\Theta^{*,(2)}_{kk_1}\Theta^{*,(2)}_{kk_2}\bigg) \\
&\leq \frac{pq^2}{(p-1)L(1-\nu)}\sum_{k,k_1,k_2}\Big|\Theta^{*,(2)}_{kk_1}\Theta^{*,(2)}_{kk_2}\Big|\sum_{\ell=2}^{p}(\ell-1)w_\ell^2. \qquad (37)
\end{aligned}$$

Choosing

$$w_\ell = \frac{\binom{p-\ell}{q-1}\Big/\binom{p-1}{q}}{\sum_{\ell'=2}^{p}(\ell'-1)\Big\{\binom{p-\ell'}{q-1}\Big/\binom{p-1}{q}\Big\}^2} \qquad (38)$$

minimises (37) subject to (36) to give

$$\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{b}^{*,(2)}\|_2^2) \leq \frac{pq^2}{(p-1)L(1-\nu)}\sum_{k,k_1,k_2}\Big|\Theta^{*,(2)}_{kk_1}\Theta^{*,(2)}_{kk_2}\Big|\left\{\sum_{\ell=1}^{p-1}\ell\left(\frac{\binom{p-1-\ell}{q-1}}{\binom{p-1}{q}}\right)^2\right\}^{-1}.$$

Finally, Lemma 19 bounds the right-most term from above to yield

$$\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{b}^{*,(2)}\|_2^2) \leq \frac{2\{(2-\delta)q\}^2}{L(1-\nu)}\sum_{k,k_1,k_2}\Big|\Theta^{*,(2)}_{kk_1}\Theta^{*,(2)}_{kk_2}\Big|. \qquad (39)$$

Now we turn to the mean-squared error. Observe that $\mathbf{s}_i^T\mathbf{b}^*$ is a sum of $L$ independent random variables, each having the same distribution as

$$\sum_{c=1}^{C} S_{i1c}b^*_{1c} = \sum_{c=1}^{C} S_{i1c}(b^{*,(1)}_{1c} + b^{*,(2)}_{1c}).$$

Thus

$$\begin{aligned}
\mathrm{Var}(\mathbf{s}_i^T\mathbf{b}^*) &\leq \frac{1}{L}\mathbb{E}\bigg(\sum_{c=1}^{C} S_{i1c}(b^{*,(1)}_{1c} + b^{*,(2)}_{1c})\bigg)^2 \\
&\leq \frac{1}{L}\left[\bigg\{\mathbb{E}\bigg(\sum_{c=1}^{C} S_{i1c}\mathbf{b}^{*,(1)}_{1c}\bigg)^2\bigg\}^{1/2} + \bigg\{\mathbb{E}\bigg(\sum_{c=1}^{C} S_{i1c}\mathbf{b}^{*,(2)}_{1c}\bigg)^2\bigg\}^{1/2}\right]^2,
\end{aligned}$$

where we have used the Cauchy–Schwarz inequality in the final line. Now using the fact that $\|\mathbf{X}\|_\infty \leq 1$, and following the argument that leads to (28), we arrive at

$$\mathbb{E}\biggl(\sum_{c=1}^{C} S_{i1c}\mathbf{b}_{1c}^{*,(2)}\biggr)^2 = p^2q^2\mathbb{E}\biggl\{\frac{\nu}{1-\nu}\sum_{k=1}^{p}\biggl(\sum_{k_1=1}^{p}\Theta^{*,(2)}_{kk_1}\mathbb{1}_{\{\pi(k_1)<\pi(k)\}}w_{\pi(k)}\biggr)^2 + \frac{1-2\nu}{1-\nu}X_{iH_i}^2\biggl(\sum_{k_1=1}^{p}\Theta^{*,(2)}_{H_ik_1}\mathbb{1}_{\{\pi(k_1)<M_i\}}w_{M_i}\biggr)^2\biggr\}. \tag{40}$$

We have

$$\begin{aligned}\mathbb{E}\biggl\{X_{iH_i}^2\biggl(\sum_{k_1=1}^{p}\Theta^{*,(2)}_{H_ik_1}\mathbb{1}_{\{\pi(k_1)<M_i\}}w_{M_i}\biggr)^2\biggr\} &= \frac{1}{q}\sum_{k=1}^{p}\sum_{\ell=1}^{p}X_{ik}^2\mathbb{E}\biggl\{\biggl(\sum_{k_1}\Theta^{*,(2)}_{k,k_1}\mathbb{1}_{\{\pi(k_1)<\ell\}}w_\ell\biggr)^2\mathbb{1}_{\{M_i=\ell\}}\biggr\} \\ &= \sum_{k=1}^{p}X_{ik}^2\sum_{\ell=2}^{p}w_\ell^2\biggl(\sum_{k_1}(\Theta^{*,(2)}_{kk_1})^2\mathbb{P}(M_i=\ell,\pi(k_1)<\ell) + \sum_{k_1\neq k_2}\Theta^{*,(2)}_{kk_1}\Theta^{*,(2)}_{kk_2}\mathbb{P}(M_i=\ell,\pi(k_1)<\ell,\pi(k_2)<\ell)\biggr) \\ &= \sum_{k=1}^{p}X_{ik}^2\sum_{\ell=2}^{p}w_\ell^2\biggl(\frac{\ell-1}{p-1}\frac{\binom{p-\ell}{q-1}}{\binom{p-1}{q}}\sum_{k_1}(\Theta^{*,(2)}_{kk_1})^2 + \frac{(\ell-1)(\ell-2)}{(p-1)(p-2)}\frac{\binom{p-\ell}{q-1}}{\binom{p-2}{q}}\sum_{k_1\neq k_2}\Theta^{*,(2)}_{kk_1}\Theta^{*,(2)}_{kk_2}\biggr).\end{aligned} \tag{41}$$

Now

$$\frac{\binom{p-\ell}{q-1}}{\binom{p-1}{q}} \leq \frac{q}{p-1} \qquad \text{and} \qquad \frac{\ell-2}{p-2}\frac{\binom{p-\ell}{q-1}}{\binom{p-2}{q}} \leq \frac{q}{p-1}.$$

Thus by Lemma 19 the quantity in (41) is at most

$$\frac{2(2-\delta)^2\delta}{p^2}\sum_{k,k_1,k_2}^{p}\Bigl|X_{ik}^2\Theta^{*,(2)}_{kk_1}\Theta^{*,(2)}_{kk_2}\Bigr|.$$

Returning to (40) and using the argument leading to (37) therefore gives us

$$\mathbb{E}\biggl(\sum_{c=1}^{C}S_{i1c}\mathbf{b}_{1c}^{*,(2)}\biggr)^2 \leq 2(2-\delta)^2q^2\biggl(\frac{\nu}{1-\nu}\sum_{k,k_1,k_2}\Bigl|\Theta^{*,(2)}_{kk_1}\Theta^{*,(2)}_{kk_2}\Bigr| + \delta\frac{1-2\nu}{1-\nu}\sum_{k,k_1,k_2}\Bigl|X_{ik}^2\Theta^{*,(2)}_{kk_1}\Theta^{*,(2)}_{kk_2}\Bigr|\biggr),$$

which then gives part (iii) of the result.

## Appendix B. Implications for the RKHS of the resemblance kernel

We first observe the following result that is an immediate consequence of Bouchard et al. (2013).

**Proposition 13** *Consider the resemblance kernel with input space $\mathcal{X} = \{0,1\}^p$ and let $\mathcal{H}$ be the corresponding RKHS. Then $\mathcal{H}$ contains every function $f : \mathcal{X} \to \mathbb{R}$.*

**Proof** Let $\mathbf{X} \in \mathbb{R}^{|\mathcal{X}| \times p}$ be the matrix with each row a different element of $\mathcal{X}$ and let $\mathbf{K} \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{X}|}$ be the matrix with $K_{\mathbf{x}\mathbf{x}'} = k(\mathbf{x}, \mathbf{x}')$ where $k$ is the resemblance kernel. Bouchard et al. (2013) shows that $\mathbf{K}$ is positive definite. Given $f : \mathcal{X} \to \mathbb{R}$, let $\mathbf{f} \in \mathbb{R}^{|\mathcal{X}|}$ be the vector of function evaluations so $f_x = f(x)$. Let $\alpha = \mathbf{K}^{-1}\mathbf{f}$. Then

$$f(\cdot) = \sum_{x \in \mathcal{X}} \alpha_x k(\cdot, x)$$

so $f \in \mathcal{H}$. ■

The following corollary of Theorem 6 derives properties of the RKHS associated with the resemblance kernel from our approximation error bounds.

**Corollary 14** *Let $\mathcal{H}$ be the RKHS of the resemblance kernel $k$ when the input space $\mathcal{X} \subset \{0,1\}^p$ is constrained such that every element has $q$ non-zeroes. Suppose $p \geq 3$. For $\boldsymbol{\theta}^{(1)} \in \mathbb{R}^p$, $\boldsymbol{\Theta}^{(2)} \in \mathbb{R}^{p \times p}$ and $\boldsymbol{\Theta} = (\boldsymbol{\theta}^{(1)}, \boldsymbol{\Theta}^{(2)})$, define $f_{\boldsymbol{\Theta}} : \mathcal{X} \to \mathbb{R}$ by*

$$f_{\boldsymbol{\Theta}}(\mathbf{x}) = \sum_{k=1}^{p} x_k \theta_k^{(1)} + \sum_{k=1}^{p} \sum_{j=1}^{p} x_k (1 - x_j) \Theta_{k,j}^{(2)}.$$

*Suppose $\boldsymbol{\Theta}$ is such that $f_{\boldsymbol{\Theta}}$ is centred so $\sum_{\mathbf{x} \in \mathcal{X}} f_{\boldsymbol{\Theta}}(\mathbf{x}) = 0$ Then $f_{\boldsymbol{\Theta}} \in \mathcal{H}$ and $\|f_{\boldsymbol{\Theta}}\|_{\mathcal{H}}^2 \leq (2 - \delta) q \|\boldsymbol{\Theta}\|^2$. In particular if $\boldsymbol{\Theta}^{(2)} = \mathbf{0}$ then $\|f_{\boldsymbol{\theta}^{(1)}}\|_{\mathcal{H}}^2 \leq (2 - \delta) q \|\boldsymbol{\theta}^{(1)}\|_2^2$.*

**Proof** Let $\mathbf{K} \in \mathbb{R} \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{X}|}$ be the matrix with $K_{xx'} = k(x, x')$. We will make use of the fact that $\mathbf{K}$ is positive definite (Bouchard et al., 2013). Suppose $\mathbf{X} \in \{0,1\}^{|\mathcal{X}| \times p}$ has as each row a different element of $\mathcal{X}$. For $L \in \mathbb{N}$, let $\mathbf{S}_L$ be the matrix formed from 1-bit min-wise hashing applied to $\mathbf{X}$ and let $\mathbf{K}_L = 2\mathbf{S}_L\mathbf{S}_L^T/L - \mathbf{J}$ where $\mathbf{J}$ is a $|\mathcal{X}| \times |\mathcal{X}|$ matrix of 1's. Given $\boldsymbol{\Theta}$, let $\mathbf{b}_L^*$ be as in the proof of Theorem 6 (see Section A.5) constructed using the permutations and $\boldsymbol{\Psi}$ matrix corresponding to $\mathbf{S}_L$.

Let $k_L : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be the random kernel associated with $\mathbf{K}_L$, that is $k(x, x') = K_{L,xx'}$; further let $\mathcal{H}_L$ be the associated RKHS. Let $\tilde{\mathbf{b}}_L$ be a centred version of $\mathbf{b}_L^*$ so $\tilde{\mathbf{b}}_L = \mathbf{b}_L^* - \bar{\mathbf{b}}^*{}_L$. Observe that $\|\tilde{\mathbf{b}}_L\|_2^2 \leq \|\mathbf{b}_L^*\|_2^2$. Let $f_L : \mathcal{X} \to \mathbb{R}$ be given by $f_L(x) = (S_L\tilde{\mathbf{b}}_L)_x$. Then $f_L \in \mathcal{H}_L$ and $\|f_L\|_{\mathcal{H}_L}^2 = L\|\tilde{\mathbf{b}}_L\|_2^2/2$.

Note that the construction of $\mathbf{b}_L^*$ ensures that each component block is i.i.d. Thus as $L \to \infty$, we have that almost surely

$$L\|\tilde{\mathbf{b}}_L\| \leq L\|\mathbf{b}_L^*\|_2^2 \to L^2\mathbb{E}\|(b_{L,1}^*, b_{L,2}^*)^T\|_2^2 \leq 2(2 - \delta)q\|\boldsymbol{\Theta}\|^2$$

(note that the expression on the right hand side of the limit does not in fact depend on $L$). Also $\mathbf{K}_L \to \mathbf{K}$ almost surely by the strong law of large numbers (see Section 2.4).

Now observe that as $\mathbf{f}$ is centred, $\|\mathbf{S}_L\tilde{\mathbf{b}}_L - \mathbf{f}\|_2^2 \leq \|\mathbf{S}_L\mathbf{b}_L^* - \mathbf{f}\|_2^2$. Thus from Theorem 6 (iii) we have that $\mathbf{S}_L\tilde{\mathbf{b}}_L \to \mathbf{f}$ in probability where $\mathbf{f} \in \mathbb{R}^{|\mathcal{X}|}$ has components $f_x = f_{\boldsymbol{\Theta}}(x)$. Therefore there exists a subsequence $L_j$ along which $\mathbf{S}_{L_j}\tilde{\mathbf{b}}_{L_j} \to \mathbf{f}$ almost surely. Thus, there exists a realisation of the random elements above such that simultaneously $\mathbf{K}_{L_j} \to \mathbf{K}$, $f_{L_j}(x) \to f(x)$ as $j \to \infty$ and $\lim_{j\to\infty} \|f_{L_j}\|_{\mathcal{H}_{L_j}}^2 \leq (2 - \delta)q\|\boldsymbol{\Theta}\|^2$. In particular we have that $\|f_{L_j}\|_{\mathcal{H}_{L_j}}$ is bounded for all $j$. Applying Lemma 21 then gives the result. ■

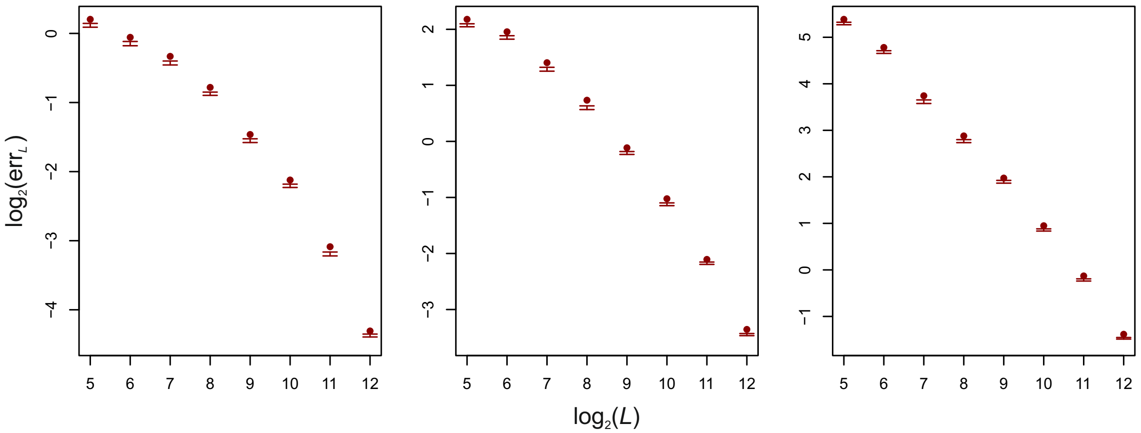

Figure 1: Plot of $\log_2(\mathrm{err}_L)$ against $\log_2(L)$ for $q = 500, 1000, 5000$ going from left to right. The bars give the first and third quartiles of $\log_2(\mathrm{err}_L)$ over the 100 simulations, and the circles give maximum values.

## Appendix C. Empirical verification of Theorem 1

In order to assess whether the scaling in $L$ provided by (iii) of Theorem 1 is in line with what is observed in practice, we looked at several numerical experiments. We generated design matrices $\mathbf{X} \in \{0,1\}^{n\times p}$ with different levels of sparsity $q \in \{500, 1000, 5000\}$ and $(n,p) = (10^4, 10^5)$. Different $\mathbf{S}$ matrices were constructed for each of the three $\mathbf{X}$ matrices with $\log_2 L \in \{5, 6, \dots, 12\}$. We then generated 100 vectors of coefficients $\boldsymbol{\beta}^* \in \mathbb{R}^p$ with $\|\boldsymbol{\beta}^*\|_2 = 1$ for each setting and examined

$$\mathrm{err}_L := \min_{\mathbf{b}\in\mathbb{R}^L} \|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{S}\mathbf{b}\|_2^2/n. \tag{42}$$

Plots of $\log_2(\mathrm{err}_L)$ against $\log_2(L)$ are given in Figure 1. Given the scaling in $L$ suggested by Theorem 1 (iii), we would expect the points to lie on a straight line with slope $-1$.. We see this is indeed approximately the case for lager $L$ and $q$.

Note that Theorem 1 does not make the claim that the $\mathbf{b}^*$ given is optimal in the sense of (42). Indeed it also satisfies unbiasedness and has a low $\ell_2$-norm in expectation: properties not necessarily satisfied by the minimiser of (42). Moreover, the bound must encompass a worst case in terms of $\mathbf{X}$ and the direction of $\boldsymbol{\beta}^*$; tighter bounds may be used at the expense of a more complicated dependence on the precise form of $\mathbf{X}$ and $\boldsymbol{\beta}^*$. However, we see from the empirical study that the scaling in $L$ provided by the result approximately parallels that corresponding to the minimiser of (42).

The details of the simulation study are as follows. The design $\mathbf{X}$ was generated randomly with the first $p/100$ columns containing $q/10$ 1's and the remaining columns containing $9q/10$ 1's. This mimics the setting of increasing variable sparsity described in Section 4.1. The vector of coefficients $\boldsymbol{\beta}^*$ had its first $p/100$ entries generated independently with an Exp(1) distribution and the remaining entries were set to 0; $\boldsymbol{\beta}^*$ was then scaled to have $\ell_2$-norm 1.

## Appendix D. Prediction error results

Here we prove results for the prediction error under linear and logistic regression models. We denote the signal to be estimated by $\mathbf{f}^*$ and assume the existence of a $\mathbf{b}^* \in \mathbb{R}^{2^b L}$ with

$$\begin{aligned}\frac{1}{n}\mathbb{E}(\|\mathbf{f}^* - \mathbf{S}\mathbf{b}^*\|_2^2) &\le c_1/L \\ \mathbb{E}(\|\mathbf{b}^*\|_2^2) &\le c_2/L.\end{aligned}$$

Explicit constructions for such coefficient vectors are provided in the previous section. Using the results here in conjunction with the approximation error results proved in Section A yield Theorems 3–9: for example, substituting (iii) of Theorem 1 immediately gives Theorem 3.

### D.1 Linear regression

We assume the model

$$\mathbf{Y} = \alpha^*\mathbf{1} + \mathbf{f}^* + \boldsymbol{\varepsilon}, \tag{43}$$

where $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$ and our goal is to estimate $\mathbf{f}^*$.

**Theorem 15** *Let $(\hat{\alpha}, \hat{\mathbf{b}})$ be the least squares estimator (12). Then*

$$\mathrm{MSPE}((\hat{\alpha}, \hat{\mathbf{b}})) \;\le\; \frac{c_1}{L} + \frac{\sigma^2\{(2^b-1)L+1\}}{n}.$$

**Proof** Let us write

$$\mathbf{Y} = \alpha^*\mathbf{1} + \mathbf{f}^* + \boldsymbol{\varepsilon} = \alpha^*\mathbf{1} + \mathbf{S}\mathbf{b}^* + \boldsymbol{\Delta} + \boldsymbol{\varepsilon},$$

so $\boldsymbol{\Delta}$ is the approximation error of $\mathbf{S}\mathbf{b}^*$. Then we have

$$\mathrm{MSPE}((\hat{\alpha}, \hat{\mathbf{b}})) = \frac{1}{n}\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\alpha^*\mathbf{1} + \mathbf{f}^* - \hat{\alpha}\mathbf{1} - \mathbf{S}\hat{\mathbf{b}}\|_2^2).$$

Now let $\check{\mathbf{S}} = (\mathbf{1}\ \mathbf{S})$, and $\mathbf{P}_{\check{\mathbf{S}}}$ be the projection on to the column space of $\check{\mathbf{S}}$ (so $\mathbf{P}_{\check{\mathbf{S}}} = \check{\mathbf{S}}\check{\mathbf{S}}^+$, where $\check{\mathbf{S}}^+$ denotes the Moore–Penrose pseudoinverse of $\check{\mathbf{S}}$). We have the following decomposition.

$$\begin{aligned}\alpha^*\mathbf{1} + \mathbf{f}^* - \hat{\alpha}\mathbf{1} - \mathbf{S}\hat{\mathbf{b}} &= \alpha^*\mathbf{1} + \mathbf{f}^* - \mathbf{P}_{\check{\mathbf{S}}}\mathbf{Y} \\ &= \alpha^*\mathbf{1} + \mathbf{S}\mathbf{b}^* + \boldsymbol{\Delta} - \mathbf{P}_{\check{\mathbf{S}}}(\alpha^*\mathbf{1} + \mathbf{S}\mathbf{b}^* + \boldsymbol{\Delta} + \boldsymbol{\varepsilon}) \\ &= (\mathbf{I} - \mathbf{P}_{\check{\mathbf{S}}})\boldsymbol{\Delta} - \mathbf{P}_{\check{\mathbf{S}}}\boldsymbol{\varepsilon}.\end{aligned}$$

Hence

$$\begin{aligned}\mathrm{MSPE}((\hat{\alpha}, \hat{\mathbf{b}})) &= \frac{1}{n}\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\pi},\boldsymbol{\Psi}}(\|(\mathbf{I} - \mathbf{P}_{\check{\mathbf{S}}})\boldsymbol{\Delta} - \mathbf{P}_{\check{\mathbf{S}}}\boldsymbol{\varepsilon}\|_2^2) \\ &= \frac{1}{n}\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|(\mathbf{I} - \mathbf{P}_{\check{\mathbf{S}}})\boldsymbol{\Delta}\|_2^2) + \frac{1}{n}\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}\{\mathbb{E}_{\boldsymbol{\varepsilon}}(\|\mathbf{P}_{\check{\mathbf{S}}}\boldsymbol{\varepsilon}\|_2^2 \,|\, \boldsymbol{\pi}, \boldsymbol{\Psi})\} \\ &\le \frac{1}{n}\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\boldsymbol{\Delta}\|_2^2) + \frac{\sigma^2\{(2^b-1)L+1\}}{n} \qquad (44) \\ &\le \frac{c_1}{L} + \frac{\sigma^2\{(2^b-1)L+1\}}{n},\end{aligned}$$

where in for (44) we have used the fact that rank($\check{\mathbf{S}}$) $\leq (2^b - 1)L + 1$ as each the $L$ blocks sums to a vector of 1's ■

**Theorem 16** *There exists $\lambda$ depending on $\mathbf{f}^*$ and $\mathbf{S}$ such that defining*

$$(\hat{\alpha}, \hat{\mathbf{b}}) := \operatorname*{arg\,min}_{(\alpha, \mathbf{b}) \in \mathbb{R} \times \mathbb{R}^L} \|\mathbf{Y} - \hat{\alpha}\mathbf{1} - \mathbf{S}\mathbf{b}\|_2^2 \;\; \textit{such that} \;\; \|\mathbf{b}\|_2^2 \leq \lambda,$$

*we have*

$$\mathrm{MSPE}((\hat{\alpha}, \hat{\mathbf{b}})) \;\leq\; \sigma\sqrt{\frac{c_2}{n}} + \frac{c_1}{L} + \frac{\sigma^2}{n}.$$

**Proof** We will take $\lambda = \|\mathbf{b}^*\|_2^2$. Let a bar over any vector $\mathbf{v}$ denote the average of the components of $\mathbf{v}$, so $\bar{\mathbf{v}} = \sum_j v_j$. Note that $\hat{\alpha} = \overline{\mathbf{Y} - \mathbf{S}\hat{\mathbf{b}}}$, and define $\hat{a}^* = \overline{\mathbf{Y} - \mathbf{S}\mathbf{b}^*}$. By our choice of $\lambda$, we have that

$$\|\mathbf{Y} - \hat{\alpha}\mathbf{1} - \mathbf{S}\hat{\mathbf{b}}\|_2^2 \leq \|\mathbf{Y} - \hat{a}^*\mathbf{1} - \mathbf{S}\mathbf{b}^*\|_2^2.$$

Noting that for any $\mathbf{v}, \mathbf{u} \in \mathbb{R}^n$, $\mathbf{v}^T(\mathbf{u} - \bar{\mathbf{u}}\mathbf{1}) = (\mathbf{v} - \bar{\mathbf{v}}\mathbf{1})^T\mathbf{u}$, rearranging the inequality above we get

$$\|\alpha^*\mathbf{1} + \mathbf{f}^* - \hat{\alpha}\mathbf{1} - \mathbf{S}\hat{\mathbf{b}}\|_2^2 \leq 2(\boldsymbol{\varepsilon} - \bar{\varepsilon}\mathbf{1})^T\mathbf{S}(\hat{\mathbf{b}} - \mathbf{b}^*) + \|\alpha^*\mathbf{1} + \mathbf{f}^* - \hat{a}^*\mathbf{1} - \mathbf{S}\mathbf{b}^*\|_2^2. \tag{45}$$

Now observe that

$$\begin{aligned}\|\alpha^*\mathbf{1} + \mathbf{f}^* - \hat{a}^*\mathbf{1} - \mathbf{S}\mathbf{b}^*\|_2^2 &= \|\mathbf{f}^* - \overline{\mathbf{f}^*}\mathbf{1} - (\mathbf{S}\mathbf{b}^* - \overline{\mathbf{S}\mathbf{b}^*}\mathbf{1})\|_2^2 + n\bar{\varepsilon}^2 \\ &\leq \|\mathbf{f}^* - \mathbf{S}\mathbf{b}^*\|_2^2 + n\bar{\varepsilon}^2. \end{aligned} \tag{46}$$

As $\mathbf{b}^*$ is independent of $\boldsymbol{\varepsilon}$, taking expectations of (45) yields

$$\mathrm{MSPE}(\hat{\mathbf{b}}) = \frac{2}{n}\mathbb{E}\{(\boldsymbol{\varepsilon} - \bar{\varepsilon}\mathbf{1})^T\mathbf{S}\hat{\mathbf{b}}\} + \frac{1}{n}\mathbb{E}(\|\mathbf{f}^* - \mathbf{S}\mathbf{b}^*\|_2^2) + \frac{\sigma^2}{n}. \tag{47}$$

Now using the fact that $\|\hat{\mathbf{b}}\|_2 \leq \|\mathbf{b}^*\|_2$ and applying the Cauchy–Schwarz inequality we have

$$\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\pi},\boldsymbol{\Psi}}\{(\boldsymbol{\varepsilon} - \bar{\varepsilon}\mathbf{1})^T\mathbf{S}\hat{\mathbf{b}}\} \leq \sqrt{\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\pi},\boldsymbol{\Psi}}\{\|\mathbf{S}^T(\boldsymbol{\varepsilon} - \bar{\varepsilon}\mathbf{1})\|_2^2\}}\sqrt{\mathbb{E}(\|\mathbf{b}^*\|_2^2)}.$$

But

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\varepsilon}}(\|\mathbf{S}^T(\boldsymbol{\varepsilon} - \bar{\varepsilon}\mathbf{1})\|_2^2 | \boldsymbol{\pi}, \boldsymbol{\Psi}) &= \mathbb{E}_{\boldsymbol{\varepsilon}}[\mathrm{Tr}\{(\boldsymbol{\varepsilon} - \bar{\varepsilon}\mathbf{1})^T\mathbf{S}\mathbf{S}^T(\boldsymbol{\varepsilon} - \bar{\varepsilon}\mathbf{1})\} | \boldsymbol{\pi}, \boldsymbol{\Psi}] \\ &= \mathbb{E}_{\boldsymbol{\varepsilon}}[\mathrm{Tr}\{(\boldsymbol{\varepsilon} - \bar{\varepsilon}\mathbf{1})(\boldsymbol{\varepsilon} - \bar{\varepsilon}\mathbf{1})^T\mathbf{S}\mathbf{S}^T\} | \boldsymbol{\pi}, \boldsymbol{\Psi}] \\ &= \mathrm{Tr}[\mathbb{E}_{\boldsymbol{\varepsilon}}\{(\boldsymbol{\varepsilon} - \bar{\varepsilon}\mathbf{1})(\boldsymbol{\varepsilon} - \bar{\varepsilon}\mathbf{1})^T\}\mathbf{S}\mathbf{S}^T] \\ &= \sigma^2\|(\mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^T)\mathbf{S}\|_F^2 \leq \sigma^2\|\mathbf{S}\|_F^2 \leq \sigma^2 nL, \end{aligned}$$

whence

$$\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\pi},\boldsymbol{\Psi}}\{(\boldsymbol{\varepsilon} - \bar{\varepsilon}\mathbf{1})^T\mathbf{S}\hat{\mathbf{b}}\} \leq \sigma\sqrt{c_2 n}. \tag{48}$$

Substituting in to (47) then gives the result. ■

### D.2 Logistic regression

We give an analogous result to Theorem 4 for classification problems under logistic loss. Let $\mathbf{X} \in [-1,1]^{n \times p}$ be the design matrix of predictor variables and let $\mathbf{Y} \in \{0,1\}^n$ be an associated vector of class labels. We assume the model

$$Y_i \sim \text{Bernoulli}(p_i); \qquad \log\left(\frac{p_i}{1-p_i}\right) = f_i,$$

with the $Y_i$ independent for $1 \le i \le n$. Define

$$\hat{\mathbf{b}}_\lambda = \arg\min_{\mathbf{b}} \frac{1}{n}\sum_{i=1}^n \left[-Y_i \mathbf{s}_i^T \mathbf{b} + \log\{1+\exp(\mathbf{s}_i^T\mathbf{b})\}\right] \quad \text{such that} \quad \|\mathbf{b}\|_2^2 \le \lambda.$$

Let $\mathcal{E}(\hat{\mathbf{b}}_\lambda)$ denote the excess risk of $\hat{\mathbf{b}}_\lambda$ under logistic loss, so

$$\mathcal{E}(\hat{\mathbf{b}}_\lambda) = \frac{1}{n}\sum_{i=1}^n \left[-p_i \mathbf{s}_i^T \hat{\mathbf{b}}_\lambda + \log\{1+\exp(\mathbf{s}_i^T\hat{\mathbf{b}}_\lambda)\}\right] - \frac{1}{n}\sum_{i=1}^n \left[-p_i f_i + \log\{1+\exp(f_i)\}\right].$$

**Theorem 17** *Let $\tilde{p} \in \mathbb{R}$ be given by (18). Then we have that there exists $\lambda$ such that*

$$\mathbb{E}_{\mathbf{Y},\boldsymbol{\pi},\boldsymbol{\Psi}}\{\mathcal{E}(\hat{\mathbf{b}}_\lambda)\} \le \frac{c_1}{4L} + \sqrt{\tilde{p}c_2/n}.$$

**Proof** We take $\lambda = \|\mathbf{b}^*\|_2^2$. By the definition of $\hat{\mathbf{b}}$ (dropping the subscript $\lambda$), we have

$$\frac{1}{n}\sum_{i=1}^n \left[-Y_i \mathbf{s}_i^T \hat{\mathbf{b}} + \log\{1+\exp(\mathbf{s}_i^T\hat{\mathbf{b}})\}\right] \le \frac{1}{n}\sum_{i=1}^n \left[-Y_i \mathbf{s}_i^T \mathbf{b}^* + \log\{1+\exp(\mathbf{s}_i^T\mathbf{b}^*)\}\right].$$

Using this, analogously to (45) we get,

$$\mathcal{E}(\hat{\mathbf{b}}) \le \frac{1}{n}\sum_{i=1}^n (Y_i - p_i)\{\mathbf{S}(\hat{\mathbf{b}} - \mathbf{b}^*)\}_i + \mathcal{E}(\mathbf{b}^*).$$

Let $\boldsymbol{\varepsilon} := \mathbf{Y} - \mathbf{p}$ be the residual vector. Since $\mathbf{b}^*$ is independent of $\boldsymbol{\varepsilon}$, after taking expectations we arrive at

$$\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\pi},\boldsymbol{\Psi}}\{\mathcal{E}(\hat{\mathbf{b}})\} \le \frac{1}{n}\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\pi},\boldsymbol{\Psi}}(\boldsymbol{\varepsilon}^T\mathbf{S}\hat{\mathbf{b}}) + \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}\{\mathcal{E}(\mathbf{b}^*)\}.$$

Write $h(a) = \log(1+e^a)$. By the mean value theorem, we have

$$\begin{aligned} |\mathcal{E}(\mathbf{b}^*)| &= \frac{1}{n}\sum_{i=1}^n |h(\mathbf{s}_i^T\mathbf{b}^*) - h(f_i) - (\mathbf{s}_i^T\mathbf{b}^* - f_i)h'(f_i)| \\ &\le \frac{1}{n}\sup_{a\in\mathbb{R}} h''(a)\|\mathbf{f}^* - \mathbf{S}\mathbf{b}^*\|_2^2 \le \frac{c_1}{4L}. \end{aligned}$$

The same argument that leads to (48) gives

$$\frac{1}{n}\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\pi},\boldsymbol{\Psi}}(\boldsymbol{\varepsilon}^T\mathbf{S}\hat{\mathbf{b}}) \le \frac{1}{n}\sqrt{\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{S}^T\boldsymbol{\varepsilon}\|_2^2)}\sqrt{c_2/L} \le \sqrt{\tilde{p}c_2/n}.$$

Collecting together the various inequalities, we get the required result. ■

## Appendix E. Technical lemmas

In this section we collect all technical lemmas used by the results presented earlier.

**Lemma 18** *Let $(a_i)_{i=1}^{\infty}$ and $(b_i)_{i=1}^{\infty}$ be two sequences of non-negative, non-increasing, real numbers such that that there is some $i^* \in \mathbb{N}$ for which*

$$\begin{aligned} a_i &\leq b_i \quad \textit{for all } i \leq i^*, \\ a_i &\geq b_i \quad \textit{for all } i > i^*. \end{aligned}$$

*(i) If*

$$\sum_{i=1}^{\infty} a_i = \sum_{i=1}^{\infty} b_i < \infty,$$

*and $m \geq 1$, then*

$$\sum_{i=1}^{\infty} a_i^m \leq \sum_{i=1}^{\infty} b_i^m.$$

*(ii) If $(c_i)_{i=1}^{\infty}$ is a sequence of non-negative, non-decreasing real numbers and*

$$\sum_{i=1}^{\infty} b_i \leq \sum_{i=1}^{\infty} a_i < \infty, \;\; \sum_{i=1}^{\infty} c_i a_i \,, \sum_{i=1}^{\infty} c_i b_i < \infty,$$

*then*

$$\sum_{i=1}^{\infty} c_i a_i \geq \sum_{i=1}^{\infty} c_i b_i.$$

**Proof** Note that the sequence $(b_i)_{i=1}^{\infty}$ majorises $(a_i)_{i=1}^{\infty}$ (see page 191 of Steele (2004)). Result (i) follows from applying Schur's majorisation inequality (Steele (2004); page 201) with the convex function $x \mapsto x^m$ on $[0, \infty)$.

For (ii) we argue,

$$\sum_{i=1}^{i^*} c_i(b_i - a_i) \leq c_{i^*} \sum_{i=1}^{i^*} (b_i - a_i) \leq c_{i^*} \sum_{i>i^*} (a_i - b_i) \leq \sum_{i>i^*} c_i(a_i - b_i).$$

■

**Lemma 19** *Let $q, p \in \mathbb{N}$ with $q \geq 1$, $p \geq \max\{q, 3\}$. We have*

$$\sum_{\ell=1}^{p-1} \ell \left( \frac{\binom{p-1-\ell}{q-1}}{\binom{p-1}{q}} \right)^2 \geq \frac{1}{2(2 - q/p)^2} \frac{p^2}{(p-1)^2}.$$

**Proof** Let the sequences $(a_\ell)_{\ell=1}^{\infty}$ and $(b_\ell)_{\ell=1}^{\infty}$ be defined by

$$a_\ell = \begin{cases} \left(\frac{\binom{p-1-\ell}{q-1}}{\binom{p-1}{q}}\right)^2 & \text{if } 1 \leq \ell \leq p-1 \\ 0 & \text{otherwise,} \end{cases}$$

$$b_\ell = \begin{cases} \left(\frac{q}{p-1}\right)^2 & \text{if } \ell \leq \left\lfloor \frac{(p-1)^2}{\{2(p-1)-q\}q} \right\rfloor \\ \frac{q}{2(p-1)-q} - \left(\frac{q}{p-1}\right)^2 \left\lfloor \frac{(p-1)^2}{\{2(p-1)-q\}q} \right\rfloor & \text{if } \ell = \left\lfloor \frac{(p-1)^2}{\{2(p-1)-q\}q} \right\rfloor + 1 \\ 0 & \text{otherwise.} \end{cases}$$

Let the sequence $(c_\ell)_{\ell=1}^{\infty}$ be defined by $c_\ell = \ell$. Note the sequences $(a_\ell)_{\ell=1}^{\infty}$, $(b_\ell)_{\ell=1}^{\infty}$ and $(c_\ell)_{\ell=1}^{\infty}$ satisfy the hypotheses of Lemma 18. Thus

$$\sum_{\ell=1}^{p-1} \ell a_\ell \geq \sum_{\ell=1}^{p-1} \ell b_\ell,$$

and

$$\begin{aligned} \sum_{\ell=1}^{p-1} \ell b_\ell = {} & \frac{1}{2}\left(\frac{q}{p-1}\right)^2 \left(\left\lfloor \frac{(p-1)^2}{\{2(p-1)-q\}q} \right\rfloor + 1\right) \left\lfloor \frac{(p-1)^2}{\{2(p-1)-q\}q} \right\rfloor \\ & + \left(\frac{q}{p-1}\right)^2 \left(\frac{(p-1)^2}{\{2(p-1)-q\}q} - \left\lfloor \frac{(p-1)^2}{\{2(p-1)-q\}q} \right\rfloor\right) \left(\left\lfloor \frac{(p-1)^2}{\{2(p-1)-q\}q} \right\rfloor + 1\right). \end{aligned}$$

Letting $x = (p-1)^2/[\{2(p-1)-q\}q]$, we have

$$\begin{aligned} \sum_{\ell=1}^{p-1} \ell b_\ell &= \frac{1}{2}(\lfloor x \rfloor + 1)\lfloor x \rfloor + (x - \lfloor x \rfloor)(\lfloor x \rfloor + 1) \\ &= \frac{1}{2}x(x+1) - \frac{1}{2}\{(x - \lfloor x \rfloor)\lfloor x \rfloor + (x - \lfloor x \rfloor)(x+1)\} + (x - \lfloor x \rfloor)(\lfloor x \rfloor + 1). \end{aligned}$$

Since $1 \geq 1/2 + (x - \lfloor x \rfloor)/2$, we see that

$$(x - \lfloor x \rfloor)(\lfloor x \rfloor + 1) \geq \frac{1}{2}(x - \lfloor x \rfloor)(x + 1 + \lfloor x \rfloor),$$

so

$$\begin{aligned} \sum_{\ell=1}^{p-1} \ell b_\ell &\geq \frac{1}{2}x(x+1) \\ &= \frac{1}{2}\left(\frac{(p-1)^2}{\{2(p-1)-q\}q} + 1\right)\frac{q}{2(p-1)-q} \\ &= \frac{1}{2(p-1)} \frac{p + \{2 - q/(p-1)\}q - 1}{\{2 - q/(p-1)\}^2} \\ &\geq \frac{1}{2(p-1)} \frac{p+q}{\{2 - q/(p-1)\}^2} \\ &\geq \frac{1}{2(2-q/p)^2} \frac{p^2}{(p-1)^2}. \end{aligned}$$

∎

**Lemma 20** *Let $\kappa(\delta) = \delta^{-a}$ where $a \in [0, 1]$. For $\ell \geq 2$,*

$$\left| \frac{\kappa^{(\ell)}(1)}{\ell!} \right| \leq ae^a \frac{1}{\ell^{1-a}}.$$

**Proof**

$$\begin{aligned} \left| \frac{\kappa^{(\ell)}(1)}{\ell!} \right| &= \frac{a(a+1)\cdots(a+\ell-1)}{1\cdot 2\cdots \ell} \\ &= \frac{a}{\ell}\frac{a+1}{1}\frac{a+2}{2}\cdots\frac{a+\ell-1}{\ell-1}. \end{aligned}$$

By Jensen's inequality

$$\begin{aligned} \frac{1}{\ell-1}\bigg\{ \log\left(\frac{a+1}{1}\right) + \log\left(\frac{a+2}{2}\right) + \cdots + \log\left(\frac{a+\ell-1}{\ell-1}\right)\bigg\} \\ \leq \log\left(1 + \frac{a\{1+\log(\ell-1)\}}{\ell-1}\right), \end{aligned}$$

and

$$\left(1 + \frac{a\{1+\log(\ell-1)\}}{\ell-1}\right)^{\ell-1} \leq \exp[a\{1+\log(\ell-1)\}].$$

Thus

$$\left| \frac{\kappa^{(\ell)}(1)}{\ell!} \right| \leq ae^a \frac{(\ell-1)^a}{\ell} \leq ae^a \frac{1}{\ell^{1-a}}.$$

∎

**Lemma 21** *Suppose we have a sequence of positive definite kernels $\{k_L\}_{L=1}^{\infty}$ on a finite input space $\mathcal{X}$. For $L \in \mathbb{N}$, let $\mathbf{K}_L \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$ be the matrix with $K_{L,xx'} = k_L(x, x')$. Suppose that $\mathbf{K}_L \to \mathbf{K}$ where $\mathbf{K}$ is positive definite and corresponds to kernel $k$. Let the RKHS's associated with $k_L$ and $k$ be $\mathcal{H}_L$ and $\mathcal{H}$ respectively. Suppose $f_L \in \mathcal{H}_L$ satisfies $|f_L(x) - f(x)| \to 0$ for some $f : \mathcal{X} \to \mathbb{R}$ and all $x \in \mathcal{X}$, and $\|f_L\|_{\mathcal{H}_L} < C$ for some $C > 0$. Then $f \in \mathcal{H}$ and $\|f_L\|_{\mathcal{H}_L} \to \|f\|_{\mathcal{H}}$ as $L \to \infty$.*

**Proof** Since $\mathcal{X}$ is finite, for each $L$ there exists $\boldsymbol{\alpha}_L \in \mathbb{R}^{|\mathcal{X}|}$ with $(f_L(x))_{x\in\mathcal{X}} = \mathbf{K}_L\boldsymbol{\alpha}_L$. Writing $\mathbf{f} = (f(x))_{x\in\mathcal{X}}$, we have $\mathbf{K}_L\boldsymbol{\alpha}_L \to \mathbf{f}$. Now $\mathbf{f} = \mathbf{K}\boldsymbol{\alpha}$ where $\boldsymbol{\alpha} = \mathbf{K}^{-1}\mathbf{f}$ showing that $f \in \mathcal{H}$.

It remains to show that $\boldsymbol{\alpha}_L^T\mathbf{K}_L\boldsymbol{\alpha}_L \to \boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha}$. Note that $\mathbf{K}_L$ is positive definite for $L$ sufficiently large, so the fact that $\boldsymbol{\alpha}_L^T\mathbf{K}_L\boldsymbol{\alpha}_L < C$ ensures the $\boldsymbol{\alpha}_L$ are bounded. Now suppose, for a contradiction, that there exists $\epsilon > 0$ and a subsequence $L_j$ with

$$|\boldsymbol{\alpha}_{L_j}^T\mathbf{K}_{L_j}\boldsymbol{\alpha}_{L_j} - \boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha}| > \epsilon. \tag{49}$$

Then as the $\boldsymbol{\alpha}_{L_j}$ are bounded, there exists a further subsequence $L_{j_m} = l_m$ such that $\boldsymbol{\alpha}_{l_m} \to \boldsymbol{\alpha}_*$ as $m \to \infty$. But then since the fact that $\mathbf{K}_L \to \mathbf{K}$ implies the maximal eigenvalues of the $\mathbf{K}_L$ are bounded, $\boldsymbol{\alpha}_{l_m}^T \mathbf{K}_{l_m} \boldsymbol{\alpha}_{l_m} \to \boldsymbol{\alpha}_*^T K_{l_m} \boldsymbol{\alpha}_*$ as $m \to \infty$. But then $\boldsymbol{\alpha}_{l_m}^T \mathbf{K}_{l_m} \boldsymbol{\alpha}_{l_m} \to \boldsymbol{\alpha}_*^T \mathbf{K} \boldsymbol{\alpha}_*$, contradicting (49). ■

## References

D. Achlioptas. Database-friendly random projections. In *Proceedings of the twentieth ACM symposium on principles of database systems*, pages 274–281. ACM, 2001.

F. Bach. Sharp analysis of low-rank kernel matrix approximations. In *Conference on Learning Theory*, pages 185–209, 2013.

F. Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18:1–38, 2017.

A. Banerjee, I. Dhillon, J. Ghosh, and S. Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6:1345–1382, sep 2005.

M. Bouchard, A.-L. Jousselme, and P.-E. Dor. A proof for the positive definiteness of the jaccard index matrix. *International Journal of Approximate Reasoning*, 54(5):615 – 626, 2013.

C. Boutsidis and P. Drineas. Random projections for the nonnegative least-squares problem. *Linear Algebra and its Applications*, 431:760–771, 2009.

L. Breiman. Random Forests. *Machine Learning*, 45:5–32, 2001.

A. Broder, M. Charikar, A. Frieze, and M. Mitzenmacher. Min-wise independent permutations. In *Proceedings of the thirtieth annual ACM symposium on theory of computing*, pages 327–336. ACM, 1998.

P. Bühlmann. Boosting for high-dimensional linear models. *Annals of Statistics*, 34:559–583, 2006.

P. Bühlmann and S. van de Geer. *Statistics for high-dimensional data.* Springer, 2011.

J.L. Carter and M.N. Wegman. Universal classes of hash functions. *Journal of Computer and System Sciences*, 18:143–154, 1979.

E. Cohen, M. Datar, S. Fujiwara, A. Gionis, P. Indyk, R. Motwani, J. Ullman, and C. Yang. Finding interesting associations without support pruning. *IEEE Transactions on Knowledge and Data Engineering*, 13:64–78, 2001.

M. Datar and S. Muthukrishnan. Estimating rarity and similarity over data stream windows. *Lecture Notes in Computer Science*, 2461:323, 2002.

P. Drineas, M.W. Michael W Mahoney, S. Muthukrishnan, and T. Sarlós. Faster least squares approximation. *Numerische Mathematik*, 117:219–249, 2011.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–451, 2004.

D. Fradkin and D. Madigan. Experiments with random projections for machine learning. In *Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining*, pages 517–522. ACM, 2003.

J. Friedman and B. Popescu. Predictive learning via rule ensembles. *Annals of Applied Statistics*, 2:916–954, 2008.

A.E. Hoerl and R.W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, pages 55–67, 1970.

T. Hofmann, B. Schölkopf, and A. Smola. Kernel methods in machine learning. *Annals of Statistics*, pages 1171–1220, 2008.

W.B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26(189-206):1, 1984.

I.T. Jolliffe. Principal components in regression analysis. In *Principal component analysis*, pages 129–155. Springer, 1986.

A. Kaban. New bounds on compressive linear least squares regression. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, pages 448–456, 2014.

P. Kar and H. Karnick. Random feature maps for dot product kernels. In Neil D. Lawrence and Mark A. Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22, pages 583–591, 2012.

J. Langford, L. Li, and A. Strehl. Vowpal wabbit online learning project, 2007.

Q. Le, T. Sarlós, and A. Smola. Fastfood-computing hilbert space expansions in loglinear time. In *Proceedings of the 30th International Conference on Machine Learning*, pages 244–252, 2013.

P. Li. Core kernels. *arXiv preprint arXiv:1404.6216*, 2014.

P. Li and A.C. König. Theory and applications of b-bit minwise hashing. *Communications of the ACM*, 54:101–109, 2011.

P. Li, T. Hastie, and K. Church. Very sparse random projections. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 287–296. ACM, 2006.

P. Li, A. Shrivastava, J. Moore, and A. König. Hashing algorithms for large-scale learning. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2011.

P. Li, A. Owen, and C.-H. Zhang. One permutation hashing. In *Advances in Neural Information Processing Systems*, pages 3122–3130, 2012.

P. Li, A. Shrivastava, and A. König. b-bit minwise hashing in practice. In *Proceedings of the 5th Asia-Pacific Symposium on Internetware*, page 13. ACM, 2013.

Michael W Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224, 2011.

O. Maillard and R. Munos. Linear regression with random projections. *Journal of Machine Learning Research*, 13:2735–2772, 2012.

J. Pennington, F. Yu, and S. Kumar. Spherical random features for polynomial kernels. In *Advances in Neural Information Processing Systems*, pages 1846–1854, 2015.

M. Pilanci and M.J. Wainwright. Iterative hessian sketch: Fast and accurate solution approximation for constrained least-squares. *arXiv preprint arXiv:1411.0347*, 2014.

M. Pilanci and M.J. Wainwright. Randomized sketches of convex programs with sharp guarantees. *Information Theory, IEEE Transactions on*, 61(9):5096–5115, 2015.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pages 1177–1184, 2007.

A. Rahimi and B. Recht. Uniform approximation of functions with random bases. In *Communication, Control, and Computing, 2008 46th Annual Allerton Conference on*, pages 555–561. IEEE, 2008.

A. Rahimi and B. Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in neural information processing systems*, pages 1313–1320, 2009.

A. Rudi, R. Camoriano, and L. Rosasco. Less is more: Nyström computational regularization. In *Advances in Neural Information Processing Systems*, pages 1657–1665, 2015.

Q. Shi, J. Petterson, G. Dror, J. Langford, A. Smola, and S. Vishwanathan. Hash kernels for structured data. *Journal of Machine Learning Research*, 10:2615–2637, 2009.

M. Steele. *The Cauchy–Schwarz Master Class*. Cambridge University Press, 2004.

D. J. Sutherland and J. Schneider. On the error of random fourier features. In *UAI*, 2015.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

J.A. Tropp. Greed is good: algorithmic results for sparse approximation. *Information Theory, IEEE Transactions on*, 50:2231–2242, 2004.

S.A. Van De Geer. High-dimensional generalized linear models and the lasso. *Annals of Statistics*, 36:614–645, 2008.

A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(3):480–492, 2012.

S. Vempala. *The random projection method*, volume 65. American Mathematical Society, 2005.

K. Weinberger, A. Dasgupta, J. Langford, A. Smola, and J. Attenberg. Feature hashing for large scale multitask learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1113–1120. ACM, 2009.

C.K.I. Williams and M. Seeger. Using the nyström method to speed up kernel machines. In *Advances in neural information processing systems*, pages 682–688, 2001.

Y. Yang, M. Pilanci, and M.J. Wainwright. Randomized sketches for kernels: Fast and optimal non-parametric regression. *Annals of Statistics*, 25:991–1023, 2017.