Min-wise hashing for large-scale regression and classification with sparse data

Rajen D. Shah and Nicolai Meinshausen
University of Cambridge and ETH Zurich

August 7, 2013

Abstract

We study large-scale regression analysis where both the number of variables, p, and the number of observations, n, may be large and in the order of millions or more. This is very different from the now well-studied high-dimensional regression context of "large p, small n". For example, in our "large p, *large* n" setting, an ordinary least squares estimator may be inappropriate for computational, rather than statistical, reasons. In order to make progress, one must seek a compromise between statistical and computational efficiency. Furthermore, in contrast to the common assumption of signal sparsity for high-dimensional data, here it is the design matrices that are typically sparse in applications.

Our approach for dealing with this large, sparse data is based on b-bit min-wise hashing (Li and König, 2011). This can be viewed as a dimensionality reduction technique for a sparse binary design matrix; our variant, which we call min-wise hash random-sign mapping (MRS mapping) also handles the real-valued case, allows for the construction of variable importance measures, and is more amenable to statistical analysis. For both linear and logistic models, we give finite-sample bounds on the prediction error of procedures which perform regression in the new lower-dimensional space after applying MRS mapping. In particular, for the former we show that the average prediction error goes to 0 asymptotically as long as $q\|\boldsymbol{\beta}^*\|_2^2/n \to 0$, where q is the maximal 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 allow us to capture interactions in the original data. A simulation study and some applications help illustrate the circumstances under which we can expect MRS mapping to perform well.

1 Introduction

The modern field of high-dimensional statistics has now developed a powerful range of methods to deal with datasets 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 datasets 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 datasets 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 competitive performance for prediction, but the sheer size of the data renders it computationally infeasible. Thus it makes sense, in this setting, to consider approximations to OLS that can be obtained at a lower computational cost.

Some effort has gone into studying approximations to least squares estimation by using low-rank matrix decompositions including CUR-type decompositions (Drineas et al., 2006, 2011). Here, however, we are interested in developing methodology that can take advantage of sparsity in the design matrix, a property that is often present in large-scale data. 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 datasets. However, similarly to the way in which many high-dimensional techniques exploit sparsity to improve statistical efficiency, here we aim to use sparsity in the data to yield both computational and statistical improvements.

The approach we take uses dimensionality reduction: we first map each of the p-dimensional observations to an L-dimensional space, where we would typically choose $L \ll p$, and then perform regression on the reduced design matrix thus created. Provided L is small enough, this final regression will be computationally feasible. Our dimenionality reduction step is based on a min-wise hashing scheme (Broder et al., 1998; Cohen et al., 2001; Datar and Muthukrishnan, 2002), and is a modification of b-bit min-wise hashing. The latter method can be applied to compress binary design matrices and has been used with SVM-type classifiers (see Li and König (2011); Li et al. (2011); Yu et al. (2012)). Despite its promising empirical results, the theoretical properties of the predictions obtained following a reduction by b-bit min-wise hashing have not been thoroughly explored for linear or logistic regression. Our variant, which we call "**m**in-wise hash **r**andom-**s**ign mapping" (MRS mapping), is more analytically tractable than b-bit min-wise hashing, and has the additional benefit that it can be used on design matrices where some predictors are continuous. In addition, our scheme allows the construction of variable importance measures that can allow one to assess the influence of the individual variables on the predictions.

We describe the MRS mapping algorithm in Section 2. In Section 3 we study the performance of linear and logistic regression using the reduced design matrix created. We show, in particular, that if the original data are well-approximated by a linear model with coefficient vector $\boldsymbol{\beta}^*$, the expected mean-squared prediction error is bounded by a small constant times $\sqrt{q/n}\,\|\boldsymbol{\beta}^*\|_2$, where q is the maximal number of active variables for each observation.

Perhaps more surprisingly, we show in Section 4 that despite the information loss through MRS mapping, a main effects model in the reduced design matrix can also approximate an interaction model in the original data. This does not require a modification of the procedure, though one typically needs a larger dimensionality L of the mapping, to reduce the error in the approximation.

Variable importance measures and other extensions are discussed in Section 5, after which some numerical studies are presented in Section 6. We conclude with a discussion in Section 7, and all proofs are collected in the appendix.

2 MRS mapping and dimension reduction

In this section, we present our MRS mapping methodology for dimension reduction. Given a sparse design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, the aim is to map this to a compressed matrix $\mathbf{S} \in \mathbb{R}^{n \times L}$, in a way that is computationally efficient and such that linear combinations of variables in $\mathbf{X}$ can be well-approximated by linear combinations of columns of $\mathbf{S}$. Section 2.2 describes the mapping to $\mathbf{S}$. The construction may seem rather bizarre at first sight; indeed, our initial motivation for developing MRS mapping was to emulate the equally mysterious b-bit min-wise hashing scheme (Li and König, 2011) in an analytically tractable manner. We explain the connection between the two techniques in Section 2.3. In Section 2.4 we briefly discuss two other dimension reduction schemes that one might consider applying to $\mathbf{X}$, namely principal components and random projections. We begin by establishing some notation.

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}. For a set of indices of columns of $\mathbf{U}$, A, we will write $\mathbf{U}_A$ for the submatrix consisting of the columns of $\mathbf{U}$ indexed by A. A vector of 1's will be denoted $\mathbf{1}$. We also define the following matrix norms

$$|||\mathbf{U}|||_\infty := \max_{i,j} |U_{ij}|$$
$$\|\mathbf{U}\|_F := \left(\sum_{i,j} U_{ij}^2 \right)^{1/2}.$$

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

Here we describe how the l^{th} column of $\mathbf{S}$, $\mathbf{S}_l$, is generated. We will create L columns in total, the entire collection of columns forming a set of i.i.d. random vectors. First let $\mathbf{\Psi} \in \{-1, 1\}^{p \times L}$ be a matrix consisting of i.i.d. random signs, each chosen with probability $1/2$. There are three steps to the construction:

Step 1: Generate a random permutation of the set $\{1, \ldots, p\}$, π_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 design matrix) with the first non-zero value.

Step 3: Form $\mathbf{S}_l \in \mathbb{R}^n$ with i^{th} component given by $\mathbf{\Psi}_{H_{il}l} X_{iH_{il}}$.

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

$\mathbf{X} = \begin{array}{c} \begin{matrix} 1 & 2 & 3 & 4 \end{matrix} \\ \begin{pmatrix} \cdot & 7 & \cdot & 9 \\ \cdot & \cdot & 1 & 4 \\ 1 & \cdot & 2 & \cdot \\ \cdot & 6 & 1 & \cdot \\ 8 & 5 & \cdot & \cdot \end{pmatrix} \end{array} \overset{\pi_l = 2314}{\mapsto} \begin{array}{c} \begin{matrix} 3 & 1 & 2 & 4 \end{matrix} \\ \begin{pmatrix} \cdot & \cdot & \mathbf{7} & 9 \\ \mathbf{1} & \cdot & \cdot & 4 \\ \mathbf{2} & 1 & \cdot & \cdot \\ \mathbf{1} & \cdot & 6 & \cdot \\ \cdot & \mathbf{8} & 5 & \cdot \end{pmatrix} \end{array}$ *Step 1*: non-zero indices whose variable indices will appear in $\mathbf{H}_l$ in Step 2 are in bold.	$\mathbf{H}_l = \begin{pmatrix} 2 \\ 3 \\ 3 \\ 3 \\ 1 \end{pmatrix}$ *Step 2.*	$\mathbf{\Psi}_l = \begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \end{pmatrix} \Rightarrow \mathbf{S}_l = \begin{pmatrix} -7 \\ -1 \\ -2 \\ -1 \\ 8 \end{pmatrix}$ *Step 3.*

Table 1: Steps 1–3 applied to a toy example. Dots represent zeroes.

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

$$H_{il} = \underset{k \in \mathbf{z}_i}{\arg\min}\ \pi_l(k), \tag{2.1}$$

$$S_{il} = \Psi_{H_{il}l} X_{iH_{il}}. \tag{2.2}$$

Suppose $\mathbf{Y}$ is a vector of responses associated with $\mathbf{X}$. Having created matrix $\mathbf{S}$, we now regress $\mathbf{Y}$ on $\mathbf{S}$, rather than the original $\mathbf{X}$. For example, we may perform a linear regression,

$$(\hat{\alpha}, \hat{\mathbf{b}}) \ = \ \underset{(\alpha, \mathbf{b}) \in \mathbb{R} \times D_\lambda}{\arg\min}\ \|\mathbf{Y} - \alpha \mathbf{1} - \mathbf{S}\mathbf{b}\|_2^2,$$

where $D_\lambda = \mathbb{R}^L$ for OLS or, $D_\lambda = \{\mathbf{w} \in \mathbb{R}^L : \|\mathbf{w}\|_2 \leq \lambda\}$ for ridge regression and a given value of the tuning parameter λ.

The estimator $\hat{\mathbf{b}}$ can then be used for predicting the expected responses of the existing observations based on $\mathbf{S}\hat{\mathbf{b}}$. For new observations, a corresponding new $\mathbf{S}$ matrix must be created. To create the $\mathbf{S}$ matrix for new observations we have to use the same permutations and sign-matrix $\mathbf{\Psi}$ as for the original data.

Note that one would not necessarily follow the above steps when implementing the MRS mapping algorithm. 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. (2012b) for more information. 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., 2012a) can also be used to speed up the pre-processing step. For the theoretical analysis in Section 3.1 and following, we will assume that all random inputs in the construction of $\mathbf{S}$ are truly random.

2.3 Connection to b-bit min-wise hashing

Instead of creating the matrix $\mathbf{H}$, b-bit min-wise hashing (Li and König, 2011) calculates a matrix $\mathbf{M} \in \mathbb{N}^{n \times L}$ with entries given by

$$M_{il} = \min_{k \in \mathbf{z}_i} \pi_l(k). \tag{2.3}$$

Using this, a matrix $\tilde{\mathbf{M}} \in \mathbb{N}^{n \times L}$ is created whose entries contain the numeric values of the lowest b bits of the entries in $\mathbf{M}$ when written out in binary form (i.e. the values of the remainders when dividing each entry by 2^b). Finally, each column of $\tilde{\mathbf{M}}$ is expanded into a block of 2^b columns, where each column codes for each of the 2^b possible values. This creates a matrix $\mathbf{T} \in \{0,1\}^{n \times 2^b L}$ with which one can perform regression. Note that no information about the values of of the entries in $\mathbf{X}$ is used other than whether or not they are zero. Thus for design matrices with real-valued entries, some form of quantisation must be performed first.

When $\mathbf{X}$ is binary, MRS mapping is perhaps most closely related to b-bit min-wise hashing with $b = 1$, since the last bit of the value M_{il} can be considered to be a close approximation to a random sign entry. Indeed, in the generation of each column of the final compressed design matrix, both MRS mapping and 1-bit min-wise hashing divide the observations into two groups with membership determined by random signs for MRS mapping and by the parity of the entries in $\mathbf{M}$ for 1-bit min-wise hashing. If for a particular l, two observations, i and i', have $M_{il} = M_{i'l}$, or equivalently $H_{il} = H_{i'l}$, they will be in the same group under both schemes; if not, their assignments to the two groups will be completely random with MRS-mapping, and approximately random for 1-bit min-wise hashing.

Provided one includes an unpenalised intercept term in the regression using $\mathbf{T}$, the fitted values from regression on $\mathbf{T}$ will be the same as when the binary entries in $\mathbf{T}$ are transformed such that all zeroes take the value -1. Thus predictions from MRS mapping should be fairly close to those of 1-bit min-wise hashing.

One could attempt to mimic b-bit min-wise hashing for $b > 1$ by taking b random sign-assignments along with each permutation. The expansion using to code for the different sequences of b signs that can be taken, can be thought of as adding interactions of all order between each block of b columns. However, we do not investigate this further here.

We do not make the claim that MRS mapping performs better when $\mathbf{X}$ is binary. Rather, the random signs used in the former method help to expose the mechanism at work in both schemes that make them successful under certain circumstances. In addition, using random signs allows us to deal with continuous predictors, and makes it easier to compute the variable importance measures to be described in Section 5.2.

2.4 Principal components and random projections

Performing principal component analysis (Jolliffe, 1986) (PCA) and retaining only the fist L components is a popular and effective form of dimension reduction. One drawback in the large-scale data setting is that computing the principal components can be computationally demanding. On the other hand, the generation of $\mathbf{S}$ with MRS mapping has complexity of order L times the number of non-zero entries in $\mathbf{X}$, and thus is almost the best one can hope for.

The method of random projections, motivated by the celebrated Johnson–Lindenstrauss lemma (Johnson and Lindenstrauss, 1984), offers dimension reduction at a similar computational cost to MRS mapping. In this scheme, $\mathbf{X}$ is mapped to $\mathbf{XA}$, where $\mathbf{A}$ is a $p \times L$ matrix with random entries typically chosen to be i.i.d. Gaussians. We compare MRS mapping and random projections from a theoretical perspective in Section 3.2 and empirically in Section 6.2.

One advantage that MRS mapping has over both PCA and random projections, is that interactions between the original predictors in $\mathbf{X}$ can be captured when only performing regression using the main effects in $\mathbf{S}$, as shown in Section 4. In contrast, a regression using main effects in the reduced design matrices obtained through PCA or random projections must necessarily fit a model no more complex than a main effects model in the original data.

3 Main effect models

In this section, we present results which 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. Here we assume that the mean response depends on $\mathbf{X}\boldsymbol{\beta}^*$ where $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is a vector of coefficients. Models containing interactions will be treated in Section 4.

The first step in obtaining these results is to construct a vector $\mathbf{b}^* \in \mathbb{R}^L$ such that $\mathbf{Sb}^*$ is close to $\mathbf{X}\boldsymbol{\beta}^*$, on average. We address this in the next section.

3.1 Approximation error

Assume that the number of non-zero entries in each row of $\mathbf{X} \in \mathbb{R}^{n\times p}$ is $q \geq 1$. The simple extension to unequal row sparsity will be dealt with below.

Theorem 1. *Let $\mathbf{b}^* \in \mathbb{R}^L$ be defined by*

$$b_l^* = \frac{q}{L} \sum_{k=1}^{p} \beta_k^* \Psi_{kl} w_{\pi_l(k)},$$

where $\mathbf{w} \in \mathbb{R}^p$ is a vector of weights. Then there exists a choice of weight vector, such that $\mathbf{b}^*$ has the following properties.

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

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

(iii) If $\|\mathbf{X}\|_\infty \leq 1$, then $\frac{1}{n}\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{Sb}^* - \mathbf{X}\boldsymbol{\beta}^*\|_2^2) \leq \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{b}^*\|_2^2)$.

(iv) In general,

$$\frac{1}{n}\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{S}\mathbf{b}^* - \mathbf{X}\boldsymbol{\beta}^*\|_2^2) \leq \frac{2 - q/p}{Ln}\left(\|\mathbf{X}\|_F^2\|\boldsymbol{\beta}^*\|_2^2 + q\sum_{k=1}^{p}\|\mathbf{X}_k\|_2^2\beta_k^{*2}\right).$$

The theorem above shows that there exists a vector $\mathbf{b}^*$ that furnishes two bounds on the approximation error of $\mathbf{X}\boldsymbol{\beta}^*$. The version in (iii) holds when $|||\mathbf{X}|||_\infty \leq 1$. The more general bound in (iv) is often tighter depending on the distribution of the entries in $\mathbf{X}$, though a little more complicated. In the results on prediction to follow, we will use the simpler bound (iii) assuming that $|||\mathbf{X}|||_\infty \leq 1$. However, in all of these results, we could equally well use (iv) to give bounds that are valid when the predictor variables are not necessarily bounded. For example, in Theorem 2 we can achieve this by replacing $\sqrt{q}\|\boldsymbol{\beta}^*\|_2$ by $\left(\|\mathbf{X}\|_F^2\|\boldsymbol{\beta}^*\|_2^2 + q\sum_{k=1}^{p}\|\mathbf{X}_k\|_2^2\beta_k^{*2}\right)^{1/2}$.

One might initially think of simply approximating $\mathbf{X}\boldsymbol{\beta}^*$ by its projection on to $\mathbf{S}$. However, it is not clear how to obtain a bound on the approximation error of the type in (iii) or (iv) directly. On the other hand, the relatively simple form of $\mathbf{b}^*$ gives a bound on the approximation error through relatively elementary calculations.

Another useful property of $\mathbf{b}^*$, aside from the approximation accuracy it delivers, is given in (ii): on average, $\|\mathbf{b}^*\|_2^2$ is small when L is large. In addition, the fact that the components of $\mathbf{b}^*$ are i.i.d. entails that $\|\mathbf{b}^*\|_2^2$ concentrates around its expected value with high probability (see Lemma 9 in the appendix). This property will be useful when studying the application of ridge regression to $\mathbf{S}$.

We finally comment on the issue of unequal row sparsity; a simple solution to the problem is as follows. Let q_i be the number of non-zero entries in $\mathbf{x}_i$. Now form an augmented design matrix $\tilde{\mathbf{X}}$ by adding $\max_i q_i - \min_i q_i$ 'dummy' variables to $\mathbf{X}$, all of whose non-zero values are equal to a nominal value ξ and are arranged in such a way that $|\{k : \tilde{\mathbf{x}}_i \neq 0\}| = \max_i q_i$. Here ξ is to be thought of as an arbitrarily small non-zero value so that the indices of the dummy variables can be chosen as entries in the matrix $\mathbf{H}$. The value of ξ is set to zero after $\mathbf{S}$ has been created, so if H_{il} is the index of a dummy variable, then entry S_{il} will vanish. In this way, we can and will assume that there are exactly q non-zero entries in each row, where $q = \max_i q_i$.

In practice, one does not usually need to consider $\tilde{\mathbf{X}}$ as regressing on the original $\mathbf{S}$ tends to produce a vector of coefficients $\hat{\mathbf{b}}$ that adapts reasonably well to unequal row sparsity, as long as the variation in row sparsity across observations is moderate.

3.2 Linear regression models

Assume we have the following approximately linear model:

$$\mathbf{Y} = \alpha^*\mathbf{1} + \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\gamma} + \boldsymbol{\varepsilon}. \tag{3.1}$$

Here α^* is the intercept and $\mathbf{X}$ is a sparse $n \times p$ design matrix with entries in $[-1, 1]$, of which q are non-zero in each row. See the comments after Theorem 1 for how to obtain results without the restriction $|||\mathbf{X}|||_\infty \leq 1$. We assume that the random noise $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ satisfies $\mathbb{E}_{\boldsymbol{\varepsilon}}(\varepsilon_i) = 0$, $\mathbb{E}_{\boldsymbol{\varepsilon}}(\varepsilon_i^2) = \sigma^2$ and $\mathrm{Cov}_{\boldsymbol{\varepsilon}}(\varepsilon_i, \varepsilon_j) = 0$. The vector $\boldsymbol{\gamma} \in \mathbb{R}^n$ represents deterministic structural error. We do not

require that $\boldsymbol{\gamma}$ be orthogonal to the column space of $\mathbf{X}$. Our bounds on prediction error involve $\|\boldsymbol{\gamma}\|_2^2$ and $\|\boldsymbol{\beta}^*\|_2$, so if a $\boldsymbol{\beta}^*$ is available with low ℓ_2-norm which satisfies (3.1) at the expense of a small increase in $\|\boldsymbol{\gamma}\|_2^2$, it is to be preferred. However, without loss of generality, we will demand that $\mathbf{X}^T\mathbf{1} = \mathbf{0}$ and $\boldsymbol{\gamma}^T\mathbf{1} = 0$.

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

$$\mathrm{MSPE}((\hat{\alpha}, \hat{\mathbf{b}})) := \mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\pi},\boldsymbol{\Psi}}\Big(\|\alpha^*\mathbf{1} + \mathbf{X}\boldsymbol{\beta}^* - \hat{\alpha}\mathbf{1} - \mathbf{S}\hat{\mathbf{b}}\|_2^2\Big)/n. \tag{3.2}$$

Thus we consider a denoising-type or in-sample error: the error on the data used to fit the regression coefficients. In the large-scale setting this is less of a restriction than in small sample studies since most new datapoints will be close to the observed data. 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.

3.2.1 Ordinary least squares

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

$$(\hat{\alpha}, \hat{\mathbf{b}}) := \operatorname*{arg\,min}_{(\alpha,\mathbf{b})\in\mathbb{R}\times\mathbb{R}^L} \|\mathbf{Y} - \alpha\mathbf{1} - \mathbf{S}\mathbf{b}\|_2^2, \tag{3.3}$$

to the matrix $\mathbf{S}$, assuming that $L \in \mathbb{N}$ is smaller than the number of samples n. We have the following theorem.

Theorem 2. *Let $(\hat{\alpha}, \hat{\mathbf{b}})$ be the least squares estimator (3.3) and let $L^* = \sqrt{(2 - q/p)qn}\|\boldsymbol{\beta}^*\|_2/\sigma$. We have*

$$\mathrm{MSPE}((\hat{\alpha}, \hat{\mathbf{b}})) \;\le\; 2\sqrt{2 - \frac{q}{p}}\max\left\{\frac{L}{L^*}, \frac{L^*}{L}\right\}\sigma\sqrt{\frac{q}{n}}\|\boldsymbol{\beta}^*\|_2 \;+\; \frac{\|\boldsymbol{\gamma}\|_2^2}{n} + \frac{\sigma^2}{n}.$$

Considering the case where $\boldsymbol{\gamma} = \mathbf{0}$, the result for an optimal choice of $L \approx L^*$ implies that the MSPE is of order $\sigma\sqrt{q/n}\|\boldsymbol{\beta}^*\|_2$. The slow rate in n, which stems from a balance between the approximation error of order $q\|\boldsymbol{\beta}^*\|_2^2/L$ and an estimation error of order $\sigma^2 L/n$, seems unavoidable if we do not make stronger conditions on the design. Indeed, essentially the same error rate is obtained in Theorem 21 of Maillard and Munos (2012) for OLS following dimension reduction by random projections, though there an extra $\log(n)$ factor is incurred. Furthermore the optimal number of projections in that case is of the same order as that of L^* for MRS mapping here.

To better understand the implications of Theorem 2, it is helpful to assume $\alpha^* = 0$ and fix the size of the signal so that $\|\mathbf{X}\boldsymbol{\beta}^*\|_2^2/n = 1$, and look at a few special cases. Consider the random design setting with independent predictors with roughly equal sparsity. In this case, $\|\boldsymbol{\beta}^*\|_2$ will be of order $\sqrt{p/q}$ and the MSPE will vanish if $p/n \to 0$.

The method becomes more attractive if the signal is associated with directions of $\mathbf{X}$ with larger variance. For example, suppose the variables can be partitioned into B blocks of variables $I_1, \ldots, I_B$ such that within the blocks, predictors are independent but the contributions $\mathbf{s}_b = \mathbf{X}_{I_b}\boldsymbol{\beta}^*_{I_b}$ to the signal of each block $b = 1, \ldots, B$ have positive correlations of at least a constant $\rho > 0$. Suppose further that the ratios of the ℓ_2-norms of the $\mathbf{s}_b$ are bounded and all predictors have roughly equal

sparsity. The signals $\|\mathbf{s}_b\|_2^2/n$ then scale as $1/(B^2\rho)$ and the ℓ_2-norm $\|\boldsymbol{\beta}^*\|_2$ will scale as $\sqrt{p/(\rho B q)}$, making the predictions converge to the true signal as long as $(p/B)/n \to 0$. We see that the overall number of blocks is not relevant to the success of the scheme, and all that matters is the average number of variables within a block, p/B.

We can also consider sparsity in the signal: if the number of important variables is fixed, the ℓ_2-norm of the optimal regression vector will remain a constant if we add additional variables, whilst q and p increase. All that is required for consistency is $q/n \to 0$.

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 $\sqrt{q}$ since the norm of $\|\boldsymbol{\beta}^*\|_2$ would be constant in this case. Adding a block of several million (sparse) bigrams might thus have the same statistical effect as adding several thousand (denser) unigrams (individual words).

If we assume $n = O(q)$, which is all that would be required to keep the prediction error bounded asymptotically, then in this situation, we see that $L^* = O(q)$. This could be a substantial reduction over the original dimension of the data, p, and would result in a corresponding 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 $O(q^2 p)$, 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 $L \approx L^*$. Even if we cannot afford to work with the optimal dimension L^* for computational reasons, the bound will still be useful for smaller values of L. The guarantee on prediction accuracy afforded by MRS mapping could not be obtained if, for example, a random subset of the predictors were chosen and remaining ones discarded, or SIS (Fan and Lv, 2008) was used; the latter would require much stronger conditions on the design and signal to be met for it to work well.

3.2.2 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). Here we will not require $L \leq n-1$ and so a higher-dimensional MRS mapping can be used to create $\mathbf{S}$. This will be especially useful when fitting interaction models where a larger choice of L is needed (see Section 4).

The regression coefficients are found for $\eta > 0$ by

$$\hat{\mathbf{b}}^\eta := \arg\min_{\mathbf{b}} \|\mathbf{Y} - \bar{\mathbf{Y}}\mathbf{1} - \mathbf{S}\mathbf{b}\|_2^2 \text{ such that } \|\mathbf{b}\|_2^2 \leq (1+\eta)\frac{(2-q/p)q\|\boldsymbol{\beta}^*\|_2^2}{L}. \tag{3.4}$$

The theorem below gives a bound on the MSPE of $\hat{\mathbf{b}}^\eta$.

Theorem 3. *Let L^* be defined as in Theorem 2. We have*

$$\text{MSPE}(\hat{\mathbf{b}}^\eta) \;\le\; \sqrt{(2-q/p)q}\,\|\boldsymbol{\beta}^*\|_2 \left(\frac{2\sigma\sqrt{1+\eta}+(L^*/L)}{\sqrt{n}}\right) + \rho\frac{\|\mathbf{X}\boldsymbol{\beta}^*+\boldsymbol{\gamma}\|_2^2}{n} + \frac{\|\boldsymbol{\gamma}\|_2^2}{n} + \frac{\sigma^2}{n},$$

where

$$\rho := \exp\left(-\frac{L\eta^2}{18(2-q/p)q\{18(2-q/p)+\eta\}}\right). \tag{3.5}$$

The ridge regression result above is very similar to that for OLS with the leading term of $\sigma\|\boldsymbol{\beta}^*\|_2\sqrt{q/n}$ being identical in both settings. The main difference 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 ℓ_2-norm of $\hat{\mathbf{b}}$ needs to be chosen carefully with ridge regression. The tuning parameter thus simply appears in a different place.

3.3 Logistic regression

We give an analogous result to Theorem 3 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

$$Y_i \sim \text{Bernoulli}(p_i); \qquad \log\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i^T\boldsymbol{\beta}^* + \gamma_i, \tag{3.6}$$

with the Y_i independent for $1 \le i \le n$. The vector $\boldsymbol{\gamma} \in \mathbb{R}^n$ represents structural error. We have omitted the separate intercept term for simplicity, and we do not require any orthogonality conditions on $\boldsymbol{\gamma}$ or $\mathbf{X}$.

Here we consider a linear classifier constructed by ℓ_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

$$\hat{\mathbf{b}}^\eta = \arg\min_{\mathbf{b}} \frac{1}{n}\sum_{i=1}^n \left[-Y_i\mathbf{s}_i^T\mathbf{b} + \log\{1+\exp(\mathbf{s}_i^T\mathbf{b})\}\right] \;\text{ such that }\; \|\mathbf{b}\|_2^2 \le (1+\eta)\frac{(2-q/p)q\|\boldsymbol{\beta}^*\|_2^2}{L}. \tag{3.7}$$

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

$$\mathcal{E}(\hat{\mathbf{b}}^\eta) = \frac{1}{n}\sum_{i=1}^n \left[-p_i\mathbf{s}_i^T\hat{\mathbf{b}}^\eta + \log\{1+\exp(\mathbf{s}_i^T\hat{\mathbf{b}}^\eta)\}\right] - \frac{1}{n}\sum_{i=1}^n \left[-p_i\mathbf{x}_i^T\boldsymbol{\beta}^* + \log\{1+\exp(\mathbf{x}_i^T\boldsymbol{\beta}^*+\gamma_i)\}\right]. \tag{3.8}$$

We can now state the analogous result to Theorem 3.

Theorem 4. *Define $\tilde{p} \in \mathbb{R}$ by*

$$\tilde{p} = \frac{1}{n}\sum_{i=1}^n p_i(1-p_i) \le \frac{1}{2}. \tag{3.9}$$

Then we have

$$\mathbb{E}_{\mathbf{Y},\boldsymbol{\pi},\boldsymbol{\Psi}}\{\mathcal{E}(\hat{\mathbf{b}}^{\eta})\} \leq \sqrt{(2-q/p)q}\,\|\boldsymbol{\beta}^*\|_2 \left(\frac{\sqrt{(1+\eta)\tilde{p}} + L^*/(4L)}{\sqrt{n}} \right) + \log(2)\rho,$$

where ρ and L^* are defined as in (3.5) *and Theorem 2 respectively.*

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

4 Interaction models

One of the compelling aspects of regression and classification with MRS mapping 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, the dimension of the MRS mapping. To be clear, in order to capture interactions with MRS mapping, 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, L, 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 the type of interaction model that can be fitted with MRS mapping. Let $\mathbf{f}^* \in \mathbb{R}^n$ be given by

$$f_i^* = \sum_{k=1}^{p} X_{ik}\theta_k^{*,(1)} + \sum_{k,k_1=1}^{p} X_{ik}\mathbb{1}_{\{X_{ik_1}=0\}}\Theta_{k,k_1}^{*,(2)}, \quad i = 1,\ldots,n, \tag{4.1}$$

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. Throughout this section we will assume that $|||\mathbf{X}|||_\infty \leq 1$. Note that if $\mathbf{X}$ were a binary matrix, then (4.1) 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.

Now let $\boldsymbol{\Theta}^*$ collect together $\boldsymbol{\theta}^{*,(1)}$ and $\boldsymbol{\Theta}^{*,(2)}$ so that we may define

$$\ell(\boldsymbol{\Theta}^*) := \|\boldsymbol{\theta}^{*,(1)}\|_2 + \left(2(2-q/p)q \sum_{k,k_1,k_2} \left| \Theta^{*,(2)}_{kk_1} \Theta^{*,(2)}_{kk_2} \right| \right)^{1/2}. \tag{4.2}$$

We will show that the results for main effects models with MRS mapping can be transferred to interaction models if we replace $\|\boldsymbol{\beta}^*\|_2$ with $\ell(\boldsymbol{\Theta}^*)$. We proceed exactly as before, first bounding the approximation error and then using this to control the prediction error.

4.1 Approximation error

As in the main effects case, we assume that the number of non-zero entries in each row of $\mathbf{X} \in [-1,1]^{n\times p}$ is $q \geq 1$; see the comments in Section 3.1 for how imbalanced row sparsity can be dealt with. Furthermore, for technical reasons, we assume here that $p \geq 3$.

Theorem 5. *Let* $\mathbf{b}^* \in \mathbb{R}^L$ *be defined by* $\mathbf{b}^* = \mathbf{b}^{*,(1)} + \mathbf{b}^{*,(2)}$, *where*

$$b_l{}^{*,(1)} = \frac{q}{L} \sum_{k=1}^{p} \Psi_{kl} \theta_k^{*,(1)} W_{1\pi_l(k)},$$

$$b_l{}^{*,(2)} = \frac{pq}{L} \sum_{k=1}^{p} \Psi_{kl} \sum_{k_1=1}^{p} \Theta^{*,(2)}_{kk_1} \mathbb{1}_{\{\pi_l(k_1)<\pi_l(k)\}} W_{2\pi_l(k)},$$

and $\mathbf{W}$ *is a* $2 \times p$ *matrix with each row a vector of weights. Then there exists a choice of weight matrix such that* $\mathbf{b}^*$ *has the following properties:*

(i) $\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\mathbf{S}\mathbf{b}^*) = \mathbf{f}^*$;

(ii) $\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{b}^*\|_2^2) \leq (2-q/p)q\ell^2(\boldsymbol{\Theta}^*)/L$;

(iii) $\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{S}\mathbf{b}^* - \mathbf{f}^*\|_2^2)/n \leq \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{b}^*\|_2^2)$.

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

$$\sum_{k,k_1,k_2} \left| \Theta^{*,(2)}_{kk_1} \Theta^{*,(2)}_{kk_2} \right| = O(1). \tag{4.3}$$

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 results for interaction models corresponding to Theorems 2, 3 and 4 now follow.

4.2 Linear regression models

Assume the model (3.1) and define the MSPE by (3.2) but in both cases with $\mathbf{X}\boldsymbol{\beta}^*$ now replaced by $\mathbf{f}^*$ defined in (4.1). Note that where before, the structural error term $\boldsymbol{\gamma}$ necessarily included two-way interaction terms if they were present, here the deterministic error need only include three-way and higher order interactions.

4.2.1 Ordinary least squares

Theorem 6. *Let* $(\hat{\alpha}, \hat{\mathbf{b}}))$ *be the least squares estimator* (3.3) *and let* $L^* = \sqrt{(2-q/p)qn}\,\ell(\mathbf{\Theta}^*)/\sigma$. *We have*

$$\text{MSPE}((\hat{\alpha}, \hat{\mathbf{b}})) \;\le\; 2\sqrt{2-\frac{q}{p}}\max\left\{\frac{L}{L^*}, \frac{L^*}{L}\right\}\sigma\sqrt{\frac{q}{n}}\ell(\mathbf{\Theta}^*) \;+\; \frac{\|\boldsymbol{\gamma}\|_2^2}{n} + \frac{\sigma^2}{n}.$$

To interpret the result, consider a situation where there are a fixed number of interaction and main effects of fixed size, so in particular (4.3) holds. 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 3.2.1, this can amount to a large statistical advantage.

The computational gains can be equally great. In the situation considered here, the optimal dimension $L^* = O(q\sqrt{n})$. 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 $O(p^2q^4)$; 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 $O(p^2q^4)$.

4.2.2 Ridge regression

Here we define the ridge regression estimator $\hat{\mathbf{b}}^\eta$ for $\eta > 0$ by

$$\hat{\mathbf{b}}^\eta = \arg\min_{\mathbf{b}} \|\mathbf{Y} - \bar{\mathbf{Y}} - \mathbf{S}\mathbf{b}\|_2^2 \quad \text{such that} \quad \|\mathbf{b}\|_2^2 \le (1+\eta)\frac{2(2-q/p)q\ell^2(\mathbf{\Theta}^*)}{L}.$$

Note the only difference to (3.4) is that the ℓ_2 norm of $\hat{\mathbf{b}}^\eta$ is allowed to be larger here, for the same value of η.

Theorem 7. *Define*

$$\rho_2 := \exp\left(-\frac{L\eta^2}{18(2-q/p)q\{18(2-q/p)+\eta\}}\right) + \exp\left(-\frac{L\eta^2}{36(2-q/p)^2q^2(18+\eta)}\right), \tag{4.4}$$

and let L^* *be as in Theorem 6. Then we have*

$$\text{MSPE}(\hat{\mathbf{b}}^\eta) \le \sqrt{(2-q/p)q}\,\ell(\mathbf{\Theta}^*)\left(\frac{2\sigma\sqrt{1+\eta}+\sqrt{2}(L^*/L)}{\sqrt{n}}\right) + \rho_2\frac{\|\mathbf{f}^*+\boldsymbol{\gamma}\|_2^2}{n} + \frac{\|\boldsymbol{\gamma}\|_2^2}{n} + \frac{\sigma^2}{n}.$$

As with Theorem 3 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^*.

4.3 Logistic regression

Here we assume the model (3.6) and define the excess risk by (3.8), but in both cases with $\mathbf{X}\boldsymbol{\beta}^*$ replaced by $\mathbf{f}^*$. Let the estimator $\hat{\mathbf{b}}^\eta$ be given for some $\eta > 0$ by

$$\hat{\mathbf{b}}^\eta = \arg\min_{\mathbf{b}} \frac{1}{n}\sum_{i=1}^{n}\left[-Y_i\mathbf{s}_i^T\mathbf{b} + \log\{1+\exp(\mathbf{s}_i^T\mathbf{b})\}\right] \text{ such that } \|\mathbf{b}\|_2^2 \le (1+\eta)\frac{2(2-q/p)q\ell^2(\boldsymbol{\Theta}^*)}{L}.$$

Theorem 8. *Let ρ_2, $\tilde{p}$ and L^* be as in* (4.4), (3.9) *and Theorem 6 respectively. Then we have*

$$\mathbb{E}_{\mathbf{Y},\boldsymbol{\pi},\boldsymbol{\Psi}}\{\mathcal{E}(\hat{\mathbf{b}}^\eta)\} \le \sqrt{(2-q/p)q}\,\ell(\boldsymbol{\Theta}^*)\left(\frac{\sqrt{(1+\eta)\tilde{p}} + (\sqrt{2}L^*)/(4L)}{\sqrt{n}}\right) + \log(2)\rho_2.$$

One could continue to look at higher-order interaction models by adding three-way interactions in (4.1) and adapting the interaction-norm (4.2) in a suitable way. However, being able to show that two-way interaction models can be fitted with MRS mapping may well be sufficient for most applications.

5 Extensions

5.1 Map aggregation

Since the compressed design matrix $\mathbf{S}$ is generated in a random fashion, we can repeat the construction $B > 1$ times to obtain B different $\mathbf{S}$ matrices. In the spirit of bagging (Breiman, 1996; Bühlmann and Yu, 2002) we can then aggregate the predictions obtained from the different random mappings by averaging them. In our experience, there has been a marked improvement when using map aggregation, and L could be chosen much lower than for $B = 1$ to achieve the same predictive accuracy. Since computational cost is typically quadratic in L, this can result in large savings. Furthermore, the computations when $B > 1$ lend themselves to a trivial parallel implementation for both the creation of the different $\mathbf{S}$ matrices and at the model fitting stage.

5.2 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 fits using 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 MRS mapping. Let $\hat{f} : \mathbb{R}^p \to \mathbb{R}$ be the regression function created following MRS mapping, and let $\hat{f}_i := \hat{f}(\mathbf{x}_i)$. Furthermore, for $1 \le k \le p$, 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 (3.1) 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 ℓ_2 norm of $\hat{\mathbf{f}} - \hat{\mathbf{f}}^{(-k)}$, for example (Breiman, 2001).

The difference in predictions can be computed relatively easily by storing an $n \times L$ matrix $\tilde{\mathbf{S}}$ with entries given by $\tilde{S}_{il} = \Psi_{\tilde{H}_{il} l} X_{i\tilde{H}_{il}}$, 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 π_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

$$\hat{f}_i - \hat{f}_i^{(-k)} = \sum_{l=1}^{L} (S_{il} - \tilde{S}_{il}) \mathbb{1}_{\{H_{il}=k\}} \hat{b}_l. \tag{5.1}$$

Note that we only need to store the three $n \times L$ matrices $\mathbf{S}, \tilde{\mathbf{S}}$ and $\mathbf{H}$ to compute the variable importance for all variables.

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.

5.3 Other fitting procedures

Here we have only considered OLS, ridge regression and ℓ_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 mapping, L, needs to be very large to achieve a desired predictive accuracy.

6 Numerical examples

6.1 Regression: simulation

Here we compare the predictive performance of MRS mapping followed by OLS to ridge regression, the Lasso and Random Forest. We apply the procedures to datasets of moderate size generated

under various simulation settings. We generate data points $i = 1, \ldots, n$ from the model

$$Y_i = \sum_{k=1}^{p} X_{ik}\beta_k^* + \kappa \sum_{(k_1,k_2)\in I} X_{ik_1}X_{ik_2} + \varepsilon_i := f_i^* + \varepsilon_i, \tag{6.1}$$

where $\mathbf{X} \in \{0,1\}^{n\times p}$ is a binary design matrix, I is a collection of indices for interaction terms, and the ε_i are independent $\mathcal{N}(0,\sigma^2)$-distributed. The interaction strength κ controls the amount the interaction terms contribute to the signal. The design matrix is generated in the following way. The components of the first predictor variable are independent Bernoulli(ς) for a sparsity parameter $\varsigma \in (0,1)$. For variables, $k = 2, \ldots, p$, we set the entry $X_{i,k}$ equal to $X_{i,k-1}$ with probability $\rho \in [0,1)$ and equal to a Bernoulli(ς) variable otherwise. The main effects β_k^* are identically 0 except for s non-zero coefficients randomly selected from $\{1, \ldots, p\}$ whose values are chosen independently from a $\mathcal{N}(0,1)$ distribution and rescaled so that $n^{-1}\|\mathbf{X}\boldsymbol{\beta}^*\|_2^2 = 1$. Each variable index within each pair corresponding to an interaction term is drawn at random from $\{1, \ldots, p\}$. We measure predictive performance by the MSPE (3.2) with $\mathbf{X}\boldsymbol{\beta}^*$ replaced by $\mathbf{f}^*$.

Some representative results for different choices of ς, p, n, s, σ, ρ and κ are shown in Figure 1. For MRS mapping, we show the prediction error under progressively larger values of L. A two-fold cross-validated error is computed for each L and once this reaches above 5% of the minimum value, we stop increasing L further (it also stops once a value of $L = 4000$ is reached). To make OLS numerically stable we apply a small ridge penalty by inflating the diagonal values of the covariance matrix of the sign matrix by 1%. Predictions are averaged over $B = 100$ iterations (see 5.1). Five-fold cross-validated ridge, Lasso and Random Forest (Breiman, 2001) prediction error are shown for comparison.

We see that

(i) In the absence of interactions, the MSPE is very similar to that of ridge regression;

(ii) When interactions are present, the MRS mapping procedure is able to fit the interaction terms and predictive performance comes close to that of Random Forest whilst the purely linear procedures fare poorly.

To fit interactions between $p = 1000$ variables in the conventional way we would first have expand the original design matrix to include roughly $5 \cdot 10^5$ interaction terms and then perform a penalised regression in the resulting very high-dimensional model. In contrast, MRS mapping is able to account for a large fraction of the interaction effects when performing regression in a roughly $L = 1000$-dimensional space.

6.2 Regression: text analysis

Kogan et al. (2009) analysed a corpus of financial reports of US companies (10-K filings) to study the extent to which a change in the volatility of the underlying stock can be forecast based on the report. One focus of their work was the change in predictive accuracy over time and tying this change to underlying financial reforms. Here, we take a more simplistic view and use the 16,087 reports provided at `http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets` to try to forecast

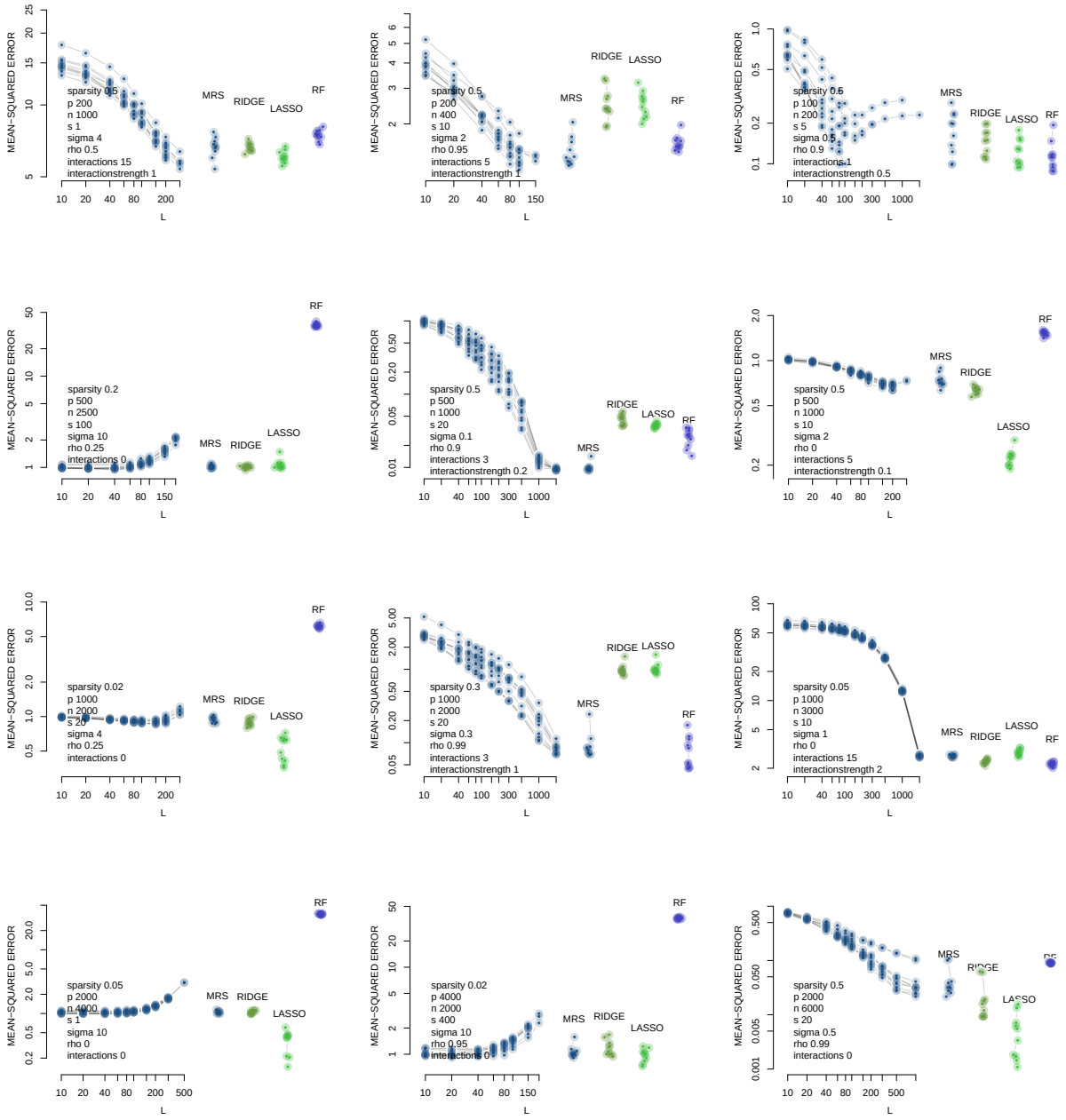


Figure 1: *The MSPE of OLS after MRS mapping (MRS) with different choices of L, under different parameter settings in the model* (6.1) *compared to the MSPEs of ridge regression (RIDGE), the Lasso (LASSO) and Random Forests (RF). For each setting, the MRS mapping procedure looks at increasing values of L until the corresponding cross-validated error curve starts to rise; the MSPE for the selected value of L is shown in the first column to the right of the curve. The different curves and points correspond to one of ten different repetitions of each experiment.*

the change in log-volatility in stock returns (comparing the 12 months after and before the report) based on the predictor variables $\mathbf{X}_k$, $k = 1, \ldots, p = 4,272,227$ of log-scaled term frequencies of unigrams and bigrams. The number of non-zero predictor variables is between a few thousand up to roughly twenty thousand.

As well as using the original response values, we also generate linear and non-linear responses using the same design matrix. For the linear model, we draw each regression coefficient at random from a standard normal distribution, independently for all variables. For the non-linear experiments, we first divide predictor variables randomly into 10 groups. For each group we form a weighted average of the variables with independent standard normal distributed weights. For each observation, we then subtract the median across all weighted averages from each weighted average. Finally, a sign transformation is applied to each resulting value to form a $16,087 \times 10$ transformed design matrix $\mathbf{Z}$ with entries in $\{-1, 0, 1\}$.

In total we consider six different scenarios with the response generated as follows: (a) the log-volatility in the 12 months after the report (not using the pre-report volatility); (b) the change in log-volatility in the underlying stock; (c) the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$ with standard normal distributed coefficients β_k^*, $k = 1, \ldots, p$; (d) a two-way interaction model based on the transformed data $\mathbf{Z}$, $\mathbf{Y} = \sum_{k=1}^{9} \mathbf{Z}_k \mathbf{Z}_{k+1} + \boldsymbol{\varepsilon}$; (e) a three-way interaction model $\mathbf{Y} = \sum_{k=1}^{8} \mathbf{Z}_k \mathbf{Z}_{k+1} \mathbf{Z}_{k+2} + \boldsymbol{\varepsilon}$; and finally (f) a four-way interaction model $\mathbf{Y} = \sum_{k=1}^{7} \mathbf{Z}_k \mathbf{Z}_{k+1} \mathbf{Z}_{k+2} \mathbf{Z}_{k+3} + \boldsymbol{\varepsilon}$. The noise term $\boldsymbol{\varepsilon}$ has independent normally distributed components with mean zero and variance σ^2 times the empirical variance of the signal. We measure predictive accuracy with 5-fold cross-validation and show the resulting correlation between predicted and actual target values in Figure 2.

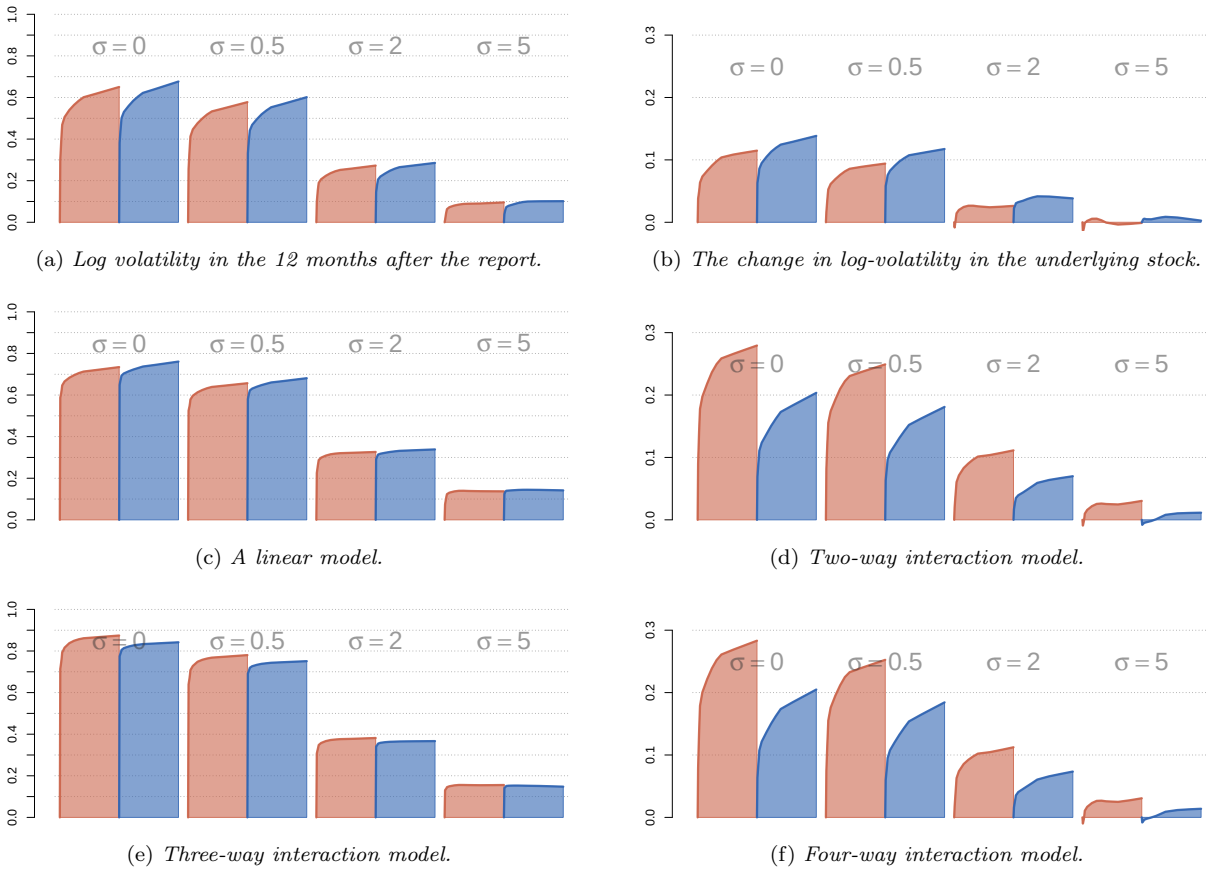
In this example, the dimensionality of the data prohibits computation of penalised regression models using the original design matrix. Thus, for comparison, we use random projections with i.i.d. standard normal entries in the projection matrix (see Section 2.4). Figure 2 shows the results. We see that random projections and MRS mapping perform similarly for the linear model in scenario (c), with the former doing slightly better. The original data (a) and (b) show a very similar pattern across the various noise levels. For the non-linear scenarios (d)–(f), however, MRS mapping outperforms random projections. While latter can only attempt to fit the best linear approximation in these examples, there is more scope to fit interactions with the MRS mapping-based regression.

6.3 Classification: URL-identification

Ma et al. (2009) reported on a large-scale classification task of identifying malicious URLs in (near) real time. Each URL is associated with both lexical (derived for example from host name and path tokens) and host-based binary predictor variables (derived for example from WHOIS info and the IP prefix). Data was collected over the course of 4 months. On each day, 20,000 URLs were collected, of which roughly 1/3 were malicious, and the remaining, benign.

For a given URL, a few hundred features will be active, that is, with a nonzero entry. For each day, the total number of active features across all URLs for that day was at least 50,000. Over the course of all days, there are more than $3 \cdot 10^6$ active features.

An important issue is that the distribution of the data is changing over time. Ma et al. (2009)



(a) *Log volatility in the 12 months after the report.*

(b) *The change in log-volatility in the underlying stock.*

(c) *A linear model.*

(d) *Two-way interaction model.*

(e) *Three-way interaction model.*

(f) *Four-way interaction model.*

Figure 2: *Correlation between predicted and actual targets for the text regression example for settings (a)–(f). For each noise level, the correlation is shown as a function of L, which varies between 0 and 250. The red curve corresponds to MRS mapping; the blue, random projections. For linear models and with the original response, the performance of MRS mapping and random projections are similar here. MRS mapping has an advantage when the signal contains stronger non-linearities.*

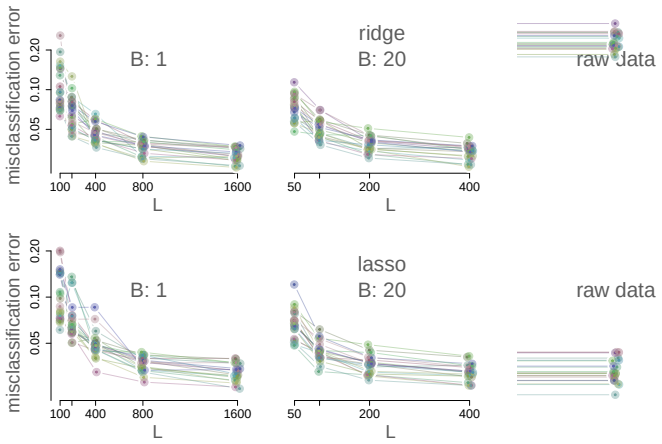


Figure 3: *Prediction accuracy on the test data for the first 30 days of the URL identification dataset. In each plot, the misclassification error is shown as a function of L for $B = 1$ (first panel) and for $B = 20$ (second panel). The results for regression on the original data are shown in the rightmost panel. In the top row, a ridge penalty is used in the logistic model for both the transformed data and the original data, whereas a Lasso cross-validated penalty is used in the bottom row.*

propose a stochastic gradient approach that can learn in real time. Here we do not want to go into all details of the distributional change but simply compare MRS mapping-based logistic regression with Lasso- and ridge-penalised logistic regression using the original data. Due to the size of the dataset, the two latter approaches can only be performed in acceptable computational time in a batch approach. Hence we treat the data from different days as different datasets and for each day, we train on the first $10,000$ URLs, and test on the remaining $10,000$. We use five-fold cross-validation with logistic loss to select the tuning parameters for penalised regressions, with the fits computed using the *glmnet* package Friedman et al. (2010). We then apply these same estimation procedures with the same tuning parameters to the MRS mapped data. The dimension of the mapping L was varied between 100 and 2000. As well as standard MRS mapping-based regression, we also looked at averaging over the predictions given by $B \in \{1, 20\}$ different $\mathbf{S}$ matrices, as described in Section 5.1. To measure performance, we record the misclassification error when the classification threshold is chosen to produce the same error rate in both classes.

Some results are shown in Figure 3. The misclassification error drops as L increases. In all four cases, for L in the thousands, the error approaches that when running Lasso on the original, much higher-dimensional, data. This occurs for lower L when averaging over $B = 20$ predictions, than when just using a single $\mathbf{S}$ matrix. Ridge regression on the original data performs much worse. Note that we can easily extend MRS mapping to use all days as training input, since the dimension L is low enough for subsequent regression to be computationally feasible.

7 Discussion

The large-scale sparse data setting presents many new challenges to statisticians that require novel approaches to overcome them. One could summarise the conventional process by which statistical methodology is developed in two stages: first a procedure that is statistically optimal for the data-generating process of interest is sought out; next, one may attempt to produce fast algorithms for the procedure. In our "large p, large n" context, it often makes more sense to consider computational issues alongside, or even before, statistical ones.

In this work, we have taken b-bit min-wise hashing (Li and König, 2011) as our starting point. From a computational point of view, it is clear that this is very well-suited to large-scale sparse data, and can retain computational feasibility where other dimension reduction techniques, such as those based on PCA, may fail. Its statistical properties, however, are harder to discern immediately.

Rather than studying b-bit min-wise hashing directly, here we considered a variant, MRS mapping. For this latter procedure, we were able to show that not only does it take advantage of sparsity in the design matrix computationally, it also exploits this for improved statistical performance. In particular, the MSPE of regression following dimension reduction by MRS mapping 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 maximal number of non-zero variables for an observation. The linear model can then be well-approximated by the low-dimensional MRS mapped 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 interaction models can be fit by a regression on the MRS mapped data that contains only main effects. Though a larger dimension of the mapped data L may be required than when approximating a main effects model, no further changes are needed to the procedure.

In summary, regression on MRS mapped data with only main effects can be a very powerful prediction engine in settings with millions of predictors and observations. Moreover, the memory footprint and computational cost of the procedure is such that this can be performed with ease on standard computing equipment. We expect to see more extensions and applications of MRS mapping and other methods based on b-bit min-wise hashing in the future.

Acknowledgements

We wish to thank Richard Samworth, for some helpful comments and suggestions.

References

L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

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.

P. Bühlmann and B. Yu. Analyzing bagging. *The Annals of Statistics*, 30:927–961, 2002.

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. Mahoney, and S. Muthukrishnan. Sampling algorithms for ℓ_2 regression and applications. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pages 1127–1136. ACM, 2006.

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.

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B*, 70:849–911, 2008.

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

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:1–22, 2010.

D. J. H. Garling. *Inequalities: A Journey into Linear Analysis*. Cambridge University Press, 2007.

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

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.

S. Kogan, D. Levin, B. Routledge, J. Sagi, and N. Smith. Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 272–280. Association for Computational Linguistics, 2009.

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

P. Li, A. Shrivastava, and C. Konig. Training logistic regression and svm on 200gb data using b-bit minwise hashing and comparisons with vowpal wabbit (vw). *arXiv:1108.3072*, 2011.

P. Li, A. Owen, and C.-H. Zhang. One permutation hashing. In *Advances in Neural Information Processing Systems 25*, pages 3122–3130, 2012a.

P. Li, A. Shrivastava, and C. Konig. b-bit minwise hashing in practice: Large-scale batch and online learning and using gpus for fast preprocessing with simple hash functions. *arXiv preprint arXiv:1205.2958*, 2012b.

J. Ma, L.K. Saul, S. Savage, and G.M. Voelker. Identifying suspicious urls: an application of large-scale online learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 681–688. ACM, 2009.

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

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

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. *The Annals of Statistics*, 36:614–645, 2008.

A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, 1996.

H.F. Yu, C.-J. Hsieh, K.-W. Chang, and C.-J. Lin. Large linear classification when data cannot fit in memory. *ACM Transactions on Knowledge Discovery from Data*, 5:23, 2012.

8 Appendix

In the proofs which follow, we will let $\pi := \pi_1$, $\boldsymbol{\psi} := \boldsymbol{\Psi}^{(1)}$. Note that then π and $\boldsymbol{\psi}$ have the same distribution as π_l and $\boldsymbol{\Psi}_l$ respectively, for any l. Similarly, we will write M_i and H_i for M_{i1} and H_{i1} respectively, where $\mathbf{M}$ is defined as in (2.3) and $\mathbf{H}$ as in (2.1). Furthermore, we will let $\delta := q/p$.

Proof of Theorem 1 There are three steps to the proof: first we determine conditions on $\mathbf{w}$ that are necessary and sufficient for the unbiasedness property (i) to hold; in the second step, we pick $\mathbf{w}$ to minimise $\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{b}^*\|_2^2)$; finally we compute the variance $\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{S}\mathbf{b}^* - \mathbf{X}\boldsymbol{\beta}^*\|_2^2)/n$.

Step 1: We begin by computing

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}((\mathbf{S}\mathbf{b}^*)_i) &= \frac{q}{L}\sum_{l=1}^{L}\mathbb{E}_{\pi_l,\boldsymbol{\Psi}_l}\left(S_{il}\sum_{k=1}^{p}\Psi_{kl}\beta_k^* w_{\pi_l(k)}\right)\\ &= \frac{q}{L}\sum_{l=1}^{L}\mathbb{E}_{\pi,\boldsymbol{\psi}}\left(\sum_{j=1}^{p}X_{ij}\mathbb{1}_{\{H_i=j\}}\psi_j\sum_{k=1}^{p}\beta_k^*\psi_k w_{\pi(k)}\right).\end{aligned}$$

Using the independence of Ψ and π, and the fact that $\mathbb{E}_{\boldsymbol{\psi}}(\psi_j\psi_k) = \mathbb{1}_{\{k=j\}}$, we have that the above display equals

$$q\mathbb{E}_{\pi}\left(\sum_{k=1}^{p}X_{ik}\beta_k^*\mathbb{1}_{\{H_i=k\}}\sum_{\ell=1}^{p}w_\ell\mathbb{1}_{\{\pi(k)=\ell\}}\right).$$

Now observe that

$$\mathbb{1}_{\{H_i=k\}}\mathbb{1}_{\{\pi(k)=\ell\}} = \mathbb{1}_{\{H_i=k\}}\mathbb{1}_{\{M_i=\ell\}},$$

and $\mathbb{1}_{\{H_i=k\}}$ and $\mathbb{1}_{\{M_i=\ell\}}$ are independent. Thus we have

$$\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}((\mathbf{S}\mathbf{b}^*)_i) = \sum_{k=1}^{p}X_{ik}\beta_k^*\sum_{\ell=1}^{p}\mathbb{P}_\pi(M_i=\ell)w_\ell. \tag{8.1}$$

Note that

$$\mathbb{P}_\pi(M_i=\ell) = \binom{p-\ell}{q-1}\Big/\binom{p}{q},$$

so in order for unbiasedness to hold, the inner product in (8.1) must satisfy

$$\sum_{\ell=1}^{p}w_\ell\binom{p-\ell}{q-1}\Big/\binom{p}{q} = 1. \tag{8.2}$$

Step 2: To compute $\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{b}^*\|_2^2)$, we first observe that the components of $\mathbf{b}^*$ are independent and each has expectation 0 as $\mathbb{E}_{\boldsymbol{\Psi}_l}(b_l^*|\pi_l) = 0$. Therefore

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{b}^*\|_2^2) &= \frac{q^2}{L}\mathbb{E}_{\pi,\boldsymbol{\psi}}\left\{\left(\sum_{k=1}^{p}\beta_k^*\psi_k w_{\pi(k)}\right)^2\right\}\\ &= \frac{q^2}{L}\mathbb{E}_{\pi}\left[\mathbb{E}_{\boldsymbol{\psi}}\left\{\left(\sum_{k=1}^{p}\beta_k^*\psi_k\sum_{\ell=1}^{p}w_\ell\mathbb{1}_{\{\pi(k)=\ell\}}\right)^2\middle|\pi\right\}\right]\\ &= \frac{q^2}{L}\sum_{k=1}^{p}\beta_k^{*2}\sum_{\ell=1}^{p}w_\ell^2\mathbb{P}_\pi(\pi(k)=\ell)\\ &= \frac{q^2}{pL}\|\boldsymbol{\beta}^*\|_2^2\|\mathbf{w}\|_2^2.\end{aligned} \tag{8.3}$$

Thus to minimise the $\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{b}^*\|_2^2)$, we must minimise $\|\mathbf{w}\|_2^2$ subject to the constraint (8.2).

Now by the the Cauchy–Schwarz inequality,

$$\|\mathbf{w}\|_2^2 \geq \frac{1}{\sum_{\ell=1}^{p} \binom{p-\ell}{q-1}^2 / \binom{p}{q}^2},$$

with equality if and only if

$$w_\ell = \frac{\binom{p-\ell}{q-1}^2 / \binom{p}{q}^2}{\sum_{\ell'=1}^{p} \binom{p-\ell'}{q-1}^2 / \binom{p}{q}^2}. \tag{8.4}$$

Applying part (i) of Lemma 11 with the sequences

$$b_\ell = \binom{p-\ell}{q-1} \Big/ \binom{p}{q}, \qquad a_\ell = \frac{q}{p}\left(1-\frac{q}{p}\right)^{\ell-1},$$

we conclude that

$$\begin{aligned}
\sum_{\ell=1}^{p-q+1} \binom{p-\ell}{q-1}^2 \Big/ \binom{p}{q}^2 &\geq \frac{q^2}{p^2} \sum_{\ell=1}^{\infty} \left(1-\frac{q}{p}\right)^{2\ell-2} \\
&= \frac{q^2}{p^2} \frac{1}{1-(1-q/p)^2} \\
&= \frac{q}{2p-q}.
\end{aligned} \tag{8.5}$$

This yields

$$\|\mathbf{w}\|_2^2 \leq \frac{2p-q}{q}. \tag{8.6}$$

Substituting into (8.3) then gives part (ii) of the result.

Step 3: Turning to the variance,

$$\begin{aligned}
\frac{1}{n}\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{S}\mathbf{b}^* - \mathbf{X}\boldsymbol{\beta}^*\|_2^2) &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}_{\boldsymbol{\pi},\boldsymbol{\Psi}}((\mathbf{S}\mathbf{b}^*)_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{L} \mathrm{Var}_{\pi_l,\boldsymbol{\Psi}_l}(S_{il} b_l^*) \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{L} \mathbb{E}_{\pi_l,\boldsymbol{\Psi}_l}((S_{il} b_l^*)^2) - (\mathbf{x}_i^T \boldsymbol{\beta}^*)^2 / L^2.
\end{aligned}$$

Now if $\|\|\mathbf{X}\|\|_\infty \leq 1$ and hence $\|\mathbf{S}\|_\infty \leq 1$, we see that

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

which is property (iii) after using (ii). For the case where this does not hold, we argue as follows.

$$\begin{aligned}
\frac{L^2}{q^2}\mathbb{E}_{\boldsymbol{\Psi}_l,\pi_l}((S_{il} b_l^*)^2) &= \mathbb{E}_{\pi,\boldsymbol{\psi}} \left\{ \sum_{j=1}^{p} X_{ij}^2 \mathbb{1}_{\{H_i=j\}} \left(\sum_{k=1}^{p} \psi_k \beta_k^* w_{\pi(k)} \right)^2 \right\} \\
&= \mathbb{E}_{\pi} \left\{ \sum_{j\in\mathbf{z}_i} X_{ij}^2 \mathbb{1}_{\{H_i=j\}} \left(\sum_{k\in\mathbf{z}_i} \beta_k^{*2} w_{\pi(k)}^2 + \sum_{k\notin\mathbf{z}_i} \beta_k^{*2} w_{\pi(k)}^2 \right) \right\}.
\end{aligned}$$

We calculate the expectation of the terms involving $k \in \mathbf{z}_i$ and $k \notin \mathbf{z}_i$ separately. For the second set of terms we have

$$\mathbb{E}_\pi \left(\sum_{j \in \mathbf{z}_i} X_{ij}^2 \mathbb{1}_{\{H_i = j\}} \sum_{k \notin \mathbf{z}_i} \beta_k^{*2} \sum_{\ell=1}^{p} w_\ell^2 \mathbb{1}_{\{\pi(k) = \ell\}} \right) = \frac{1}{pq} \|\mathbf{w}\|_2^2 \sum_{j \in \mathbf{z}_i} X_{ij}^2 \sum_{k \notin \mathbf{z}_i} \beta_k^{*2}$$
$$\leq \frac{2 - \delta}{q^2} \sum_{j \in \mathbf{z}_i} X_{ij}^2 \sum_{k \notin \mathbf{z}_i} \beta_k^{*2}. \tag{8.7}$$

For the first, note that when $j \in \mathbf{z}_i$,

$$\mathbb{1}_{\{H_i = j\}} \mathbb{1}_{\{\pi(k) = \ell\}} = \mathbb{1}_{\{j = k\}} \mathbb{1}_{\{H_i = k\}} \mathbb{1}_{\{M_i = \ell\}} + \mathbb{1}_{\{j \neq k\}} \sum_{\ell' < \ell} \mathbb{1}_{\{M_i = \ell'\}} \mathbb{1}_{\{\pi(j) = \ell'\}} \mathbb{1}_{\{\pi(k) = \ell\}}.$$

Taking expectations we get

$$\mathbb{P}_\pi(\{H_i = j\} \cap \{\pi(k) = \ell\}) = \frac{1}{q} \left(\frac{\binom{p-\ell}{q-1}}{\binom{p}{q}} \mathbb{1}_{\{j=k\}} + \sum_{\ell'=1}^{\ell-1} \frac{1}{p - \ell'} \frac{\binom{p-\ell'}{q-1}}{\binom{p}{q}} \mathbb{1}_{\{j \neq k\}} \right).$$

But

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

Thus

$$\begin{aligned}
&\mathbb{E}_\pi \left(\sum_{j \in \mathbf{z}_i} X_{ij}^2 \mathbb{1}_{\{H_i = j\}} \sum_{k \in \mathbf{z}_i} \beta_k^{*2} \sum_{\ell=1}^{p} w_\ell^2 \mathbb{1}_{\{\pi(k) = \ell\}} \right) \\
=&\frac{1}{q} \sum_{j \in \mathbf{z}_i} X_{ij}^2 \sum_{k \in \mathbf{z}_i} \beta_k^{*2} \sum_{\ell=1}^{p} w_\ell^2 \left\{ \mathbb{1}_{\{j=k\}} \frac{\binom{p-\ell}{q-1}}{\binom{p}{q}} + (1 - \mathbb{1}_{\{j=k\}}) \frac{1}{q-1} \left(\frac{q}{p} - \frac{\binom{p-\ell}{q-1}}{\binom{p}{q}} \right) \right\} \\
=&\frac{1}{q-1} \sum_{k \in \mathbf{z}_i} X_{ik}^2 \beta_k^{*2} \sum_{\ell=1}^{p} w_\ell^2 \left(\frac{\binom{p-\ell}{q-1}}{\binom{p}{q}} - \frac{1}{p} \right) + \frac{1}{q-1} \sum_{j \in \mathbf{z}_i} X_{ij}^2 \sum_{k \in \mathbf{z}_i} \beta_k^{*2} \sum_{\ell=1}^{p} w_\ell^2 \left(\frac{1}{p} - \frac{1}{q} \frac{\binom{p-\ell}{q-1}}{\binom{p}{q}} \right) \\
=&\text{(I)} + \text{(II)}.
\end{aligned}$$

As

$$\frac{\binom{p-\ell}{q-1}}{\binom{p}{q}} - \frac{1}{p} \leq \frac{q}{p} - \frac{1}{p} = \frac{q-1}{p},$$

we have that

$$\text{(I)} \leq \frac{2-\delta}{q} \sum_{k=1}^{p} X_{ik}^2 \beta_k^{*2}. \tag{8.8}$$

Turning to (II), note that by Chebyshev's order inequality (Steele, 2004),

$$\sum_{\ell=1}^{p} w_\ell^2 \frac{\binom{p-\ell}{q-1}}{\binom{p}{q}} \geq \frac{1}{p}\|\mathbf{w}\|_2^2,$$

whence

$$\text{(II)} \leq \frac{2-\delta}{q^2} \sum_{j \in \mathbf{z}_i} X_{ij}^2 \sum_{k \in \mathbf{z}_i} \beta_k^{*2}. \tag{8.9}$$

Collecting together equations (8.7), (8.8) and (8.9), we get

$$\frac{L^2}{q^2}\mathbb{E}_{\boldsymbol{\Psi}_l,\pi_l}((S_{il}b_l^*)^2) \leq \frac{2-\delta}{q}\left(\frac{1}{q}\|\mathbf{x}_i\|_2^2\|\boldsymbol{\beta}^*\|_2^2 + \sum_{k=1}^{p} X_{ik}^2\beta_k^{*2}\right),$$

and so

$$\frac{1}{n}\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{S}\mathbf{b}^* - \mathbf{X}\boldsymbol{\beta}^*\|_2^2) \leq \frac{2-\delta}{Ln}\left(\|\mathbf{X}\|_F^2\|\boldsymbol{\beta}^*\|_2^2 + q\sum_{k=1}^{p}\|\mathbf{X}_k\|_2^2\beta_k^{*2}\right).$$

□

Proof of Theorem 2 Let $\mathbf{b}^* \in \mathbb{R}^L$ be as in Theorem 1. Let us write

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

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

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

Now letting $\check{\mathbf{S}} = (\mathbf{1}\ \mathbf{S})$, 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.

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

Hence

$$\begin{aligned}
\text{MSPE}((\hat{\alpha}, \hat{\mathbf{b}})) &= \frac{1}{n}\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\pi},\boldsymbol{\Psi}}(\|(\mathbf{I} - \mathbf{P}_{\check{\mathbf{S}}})(\boldsymbol{\Delta} + \boldsymbol{\gamma}) - \mathbf{P}_{\check{\mathbf{S}}}\boldsymbol{\varepsilon}\|_2^2) \\
&= \frac{1}{n}\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|(\mathbf{I} - \mathbf{P}_{\check{\mathbf{S}}})(\boldsymbol{\Delta} + \boldsymbol{\gamma})\|_2^2) + \frac{1}{n}\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}\{\mathbb{E}_{\boldsymbol{\varepsilon}}(\|\mathbf{P}_{\check{\mathbf{S}}}\boldsymbol{\varepsilon}\|_2^2 \,|\, \boldsymbol{\pi}, \boldsymbol{\Psi})\} \\
&\leq \frac{1}{n}\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\boldsymbol{\Delta}\|_2^2) + \frac{\sigma^2(L+1)}{n} + \frac{1}{n}\|\boldsymbol{\gamma}\|_2^2 \qquad (8.10) \\
&\leq \frac{(2-\delta)q\|\boldsymbol{\beta}^*\|_2^2}{L} + \frac{\sigma^2(L+1)}{n} + \frac{1}{n}\|\boldsymbol{\gamma}\|_2^2,
\end{aligned}$$

where in (8.10) we used the fact that $\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\boldsymbol{\Delta}) = \mathbf{0}$, and the final line follows from property (iii) in Theorem 1 assuming bounded predictor variables with $\|\mathbf{X}\|_\infty \leq 1$. For general predictor variables, we just use property (iv) instead of (iii) in Theorem 1. □

Proof of Theorem 3 Let $\mathbf{b}^* \in \mathbb{R}^L$ be as in Theorem 1. Let the event Λ be defined by

$$\Lambda = \left\{ \|\mathbf{b}^*\|_2^2 < (1+\eta)\frac{(2-\delta)q\|\boldsymbol{\beta}^*\|_2^2}{L} \right\}. \tag{8.11}$$

From the definition of $\hat{\mathbf{b}}$ (dropping the superscript η in the following), we have the following two inequalities:

$$\|\mathbf{Y} - \bar{\mathbf{Y}}\mathbf{1} - \mathbf{S}\hat{\mathbf{b}}\|_2^2 \mathbb{1}_\Lambda \leq \|\mathbf{Y} - \bar{\mathbf{Y}}\mathbf{1} - \mathbf{S}\mathbf{b}^*\|_2^2 \mathbb{1}_\Lambda$$
$$\|\mathbf{Y} - \bar{\mathbf{Y}}\mathbf{1} - \mathbf{S}\hat{\mathbf{b}}\|_2^2 \mathbb{1}_{\Lambda^c} \leq \|\mathbf{Y} - \bar{\mathbf{Y}}\mathbf{1}\|_2^2 \mathbb{1}_{\Lambda^c}.$$

Rearranging gives us

$$\|\alpha\mathbf{1} + \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\gamma} - \bar{\mathbf{Y}}\mathbf{1} - \mathbf{S}\hat{\mathbf{b}}\|_2^2 \mathbb{1}_\Lambda \leq -2\boldsymbol{\varepsilon}^T\mathbf{S}(\hat{\mathbf{b}} - \mathbf{b}^*)\mathbb{1}_\Lambda + \|\alpha\mathbf{1} + \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\gamma} - \bar{\mathbf{Y}}\mathbf{1} - \mathbf{S}\mathbf{b}^*\|_2^2 \mathbb{1}_\Lambda, \tag{8.12}$$

$$\|\alpha\mathbf{1} + \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\gamma} - \bar{\mathbf{Y}}\mathbf{1} - \mathbf{S}\hat{\mathbf{b}}\|_2^2 \mathbb{1}_{\Lambda^c} \leq -2\boldsymbol{\varepsilon}^T\mathbf{S}\hat{\mathbf{b}}\mathbb{1}_{\Lambda^c} + n(\bar{\mathbf{Y}} - \alpha)^2\mathbb{1}_{\Lambda^c} + \|\mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\gamma}\|_2^2\mathbb{1}_{\Lambda^c}, \tag{8.13}$$

where we have appealed to the facts that $\mathbf{X}^T\mathbf{1} = \mathbf{0}$ and $\boldsymbol{\gamma}^T\mathbf{1} = 0$ to arrive at the second equation. Note that both $\mathbf{b}^*$ and $\mathbb{1}_\Lambda$ are independent of $\boldsymbol{\varepsilon}$. Thus adding together (8.12) and (8.13), and then taking expectations yields

$$\begin{aligned}\mathrm{MSPE}(\hat{\mathbf{b}}) = &-\frac{2}{n}\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\pi},\boldsymbol{\Psi}}(\boldsymbol{\varepsilon}^T\mathbf{S}\hat{\mathbf{b}}) + \frac{1}{n}\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\gamma} - \mathbf{S}\mathbf{b}^*\|_2^2\mathbb{1}_\Lambda) + \mathbb{E}_{\boldsymbol{\varepsilon}}\{(\bar{\mathbf{Y}} - \alpha)^2\} \\ &+ \frac{1}{n}\|\mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\gamma}\|_2^2\mathbb{P}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\Lambda^c).\end{aligned} \tag{8.14}$$

Now applying the Cauchy–Schwarz inequality and using the fact that

$$\|\hat{\mathbf{b}}\|_2 \leq \sqrt{1+\eta}\frac{\sqrt{(2-\delta)q}\|\boldsymbol{\beta}^*\|_2}{\sqrt{L}},$$

we have

$$\begin{aligned}-\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\pi},\boldsymbol{\Psi}}(\boldsymbol{\varepsilon}^T\mathbf{S}\hat{\mathbf{b}}) &\leq \sqrt{\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{S}^T\boldsymbol{\varepsilon}\|_2^2)\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\hat{\mathbf{b}}\|_2^2)} \\ &\leq \sqrt{\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}\{\mathbb{E}_{\boldsymbol{\varepsilon}}(\|\mathbf{S}^T\boldsymbol{\varepsilon}\|_2^2|\boldsymbol{\pi},\boldsymbol{\Psi})\}}\frac{\sqrt{(1+\eta)(2-\delta)q}\|\boldsymbol{\beta}^*\|_2}{\sqrt{L}}.\end{aligned}$$

But

$$\mathbb{E}_{\boldsymbol{\varepsilon}}(\|\mathbf{S}^T\boldsymbol{\varepsilon}\|_2^2|\boldsymbol{\pi},\boldsymbol{\Psi}) = \mathbb{E}_{\boldsymbol{\varepsilon}}\{\mathrm{Tr}(\boldsymbol{\varepsilon}^T\mathbf{S}\mathbf{S}^T\boldsymbol{\varepsilon})|\boldsymbol{\pi},\boldsymbol{\Psi}\} = \mathbb{E}_{\boldsymbol{\varepsilon}}\{\mathrm{Tr}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\mathbf{S}\mathbf{S}^T)|\boldsymbol{\pi},\boldsymbol{\Psi}\} = \mathrm{Tr}\{\mathbb{E}_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T)\mathbf{S}\mathbf{S}^T\} = \sigma^2 nL,$$

whence

$$-\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\pi},\boldsymbol{\Psi}}(\boldsymbol{\varepsilon}^T\mathbf{S}\hat{\mathbf{b}}) \leq \sigma\sqrt{1+\eta}\sqrt{(2-\delta)qn}\|\boldsymbol{\beta}^*\|_2. \tag{8.15}$$

Meanwhile, by Lemma 9 below, we have that $\mathbb{P}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\Lambda^c) \leq \rho$, with ρ defined as in (3.5). By Theorem 1, property (i), we have

$$\frac{1}{n}\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\gamma} - \mathbf{S}\mathbf{b}^*\|_2^2) = \frac{1}{n}\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{S}\mathbf{b}^*\|_2^2) + \frac{1}{n}\|\boldsymbol{\gamma}\|_2^2, \tag{8.16}$$

whilst property (iii) gives an upper bound on the approximation error $\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{S}\mathbf{b}^*\|_2^2)$. Noting that $\bar{\mathbf{Y}} - \alpha = \bar{\boldsymbol{\varepsilon}}$, we can return to (8.14) with this knowledge to conclude the result. □

Proof of Theorem 4 The proof is very similar to that of Theorem 3. Let the event Λ be defined as in (8.11). By the definition of $\hat{\mathbf{b}}$ (dropping the superscript η), we have

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

Using this, analogously to (8.12) and (8.13) we get,

$$\mathcal{E}(\hat{\mathbf{b}})\mathbb{1}_\Lambda \leq \frac{1}{n}\sum_{i=1}^{n}(Y_i - p_i)\{\mathbf{S}(\hat{\mathbf{b}} - \mathbf{b}^*)\}_i\mathbb{1}_\Lambda + \mathcal{E}(\mathbf{b}^*)\mathbb{1}_\Lambda,$$
$$\mathcal{E}(\hat{\mathbf{b}})\mathbb{1}_{\Lambda^c} \leq \frac{1}{n}\sum_{i=1}^{n}(Y_i - p_i)(\mathbf{S}\hat{\mathbf{b}})_i\mathbb{1}_{\Lambda^c} + \mathcal{E}(\mathbf{0})\mathbb{1}_{\Lambda^c},$$

where $\mathcal{E}(\mathbf{0}) \leq \log(2)$ is the excess risk of the the zero-vector $\mathbf{0} \in \mathbb{R}^L$. Let $\boldsymbol{\varepsilon} := \mathbf{Y} - \mathbf{p}$ be the residual vector. Adding the two equations above and taking expectations yields

$$\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\pi},\boldsymbol{\Psi}}\{\mathcal{E}(\hat{\mathbf{b}})\} \leq \frac{1}{n}\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\pi},\boldsymbol{\Psi}}(\boldsymbol{\varepsilon}^T\mathbf{S}\hat{\mathbf{b}}) + \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}\{\mathcal{E}(\mathbf{b}^*)\mathbb{1}_\Lambda\} + \mathcal{E}(\mathbf{0})\mathbb{P}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\Lambda^c).$$

From Lemma 9, we get $\mathbb{P}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\Lambda^c) \leq \rho$. By the mean value theorem, we have

$$\mathcal{E}(\mathbf{b}^*) \leq \frac{1}{n}\sup_{a\in\mathbb{R}}\left|\frac{e^a}{1+e^a}\left(1 - \frac{e^a}{1+e^a}\right)\right|\|\mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\gamma} - \mathbf{S}\mathbf{b}^*\|_2^2 = \frac{1}{4n}\|\mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\gamma} - \mathbf{S}\mathbf{b}^*\|_2^2.$$

By (8.16), we then have

$$\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\mathcal{E}(\mathbf{b}^*)) \leq \frac{1}{4n}\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{S}\mathbf{b}^*\|_2^2) + \frac{1}{4n}\|\boldsymbol{\gamma}\|_2^2,$$

where (iii) of Theorem 1 give bounds on the quantity in the right-hand-side. Further, the same argument that leads to (8.15) gives

$$\frac{1}{n}\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\pi},\boldsymbol{\Psi}}(\boldsymbol{\varepsilon}^T\mathbf{S}\hat{\mathbf{b}}) \leq \frac{1}{n}\sqrt{\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{S}^T\boldsymbol{\varepsilon}\|_2^2)}\sqrt{1+\eta}\frac{\sqrt{(2-\delta)q}\|\boldsymbol{\beta}^*\|_2}{\sqrt{L}} = \sqrt{\frac{(1+\eta)(2-\delta)\tilde{p}}{n}}\|\boldsymbol{\beta}^*\|_2.$$

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

Proof of Theorem 5 The proof proceeds similarly to that of Theorem 1. As in the latter, we set

$$W_{1\ell} := w_l = \mathbb{P}_\pi(M_i = \ell), \tag{8.17}$$

to ensure $\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\mathbf{Sb}^{*,(1)}) = \mathbf{X}\boldsymbol{\theta}^{*,(1)}$. Now we shall determine conditions on $\mathbf{w}_2$, the second row of $\mathbf{W}$, such that

$$\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}((\mathbf{Sb}^{*,(2)})_i) = \sum_{k,k_1} X_{ik}\mathbb{1}_{\{X_{ik_1}=0\}}\Theta^{*,(2)}_{kk_1}. \tag{8.18}$$

To this end, we compute

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}((\mathbf{Sb}^{*,(2)})_i) &= \frac{pq}{L}\sum_{l=1}^{L}\mathbb{E}_{\pi_l,\boldsymbol{\Psi}_l}\left(S_{il}\sum_{k=1}^{p}\Psi_{kl}\sum_{k_1=1}^{p}\Theta^{*,(2)}_{kk_1}\mathbb{1}_{\{\pi_l(k_1)<\pi_l(k)\}}W_{2\pi_l(k)}\right)\\
&= pq\mathbb{E}_{\pi,\boldsymbol{\psi}}\left(\sum_{j=1}^{p}X_{ij}\mathbb{1}_{\{H_i=j\}}\psi_j\sum_{k=1}^{p}\psi_k\sum_{k_1=1}^{p}\Theta^{*,(2)}_{kk_1}\mathbb{1}_{\{\pi(k_1)<\pi(k)\}}W_{2\pi(k)}\right)\\
&= pq\mathbb{E}_{\pi}\left(\sum_{k=1}^{p}X_{ik}\mathbb{1}_{\{H_i=k\}}\sum_{k_1=1}^{p}\Theta^{*,(2)}_{kk_1}\mathbb{1}_{\{\pi(k_1)<\pi(k)\}}\sum_{\ell=2}^{p}W_{2\ell}\mathbb{1}_{\{\pi(k)=\ell\}}\right).
\end{aligned}$$

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

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

Thus if we choose $\mathbf{w}_2$ such that

$$\sum_{\ell=2}^{p}(\ell-1)\frac{\binom{p-\ell}{q-1}}{\binom{p-1}{q}}W_{2\ell} = 1, \tag{8.19}$$

property (8.18) will be satisfied.

Turning now to the variance, as $\|\mathbf{S}\|_\infty \le \|\mathbf{X}\|_\infty \le 1$, we have

$$\mathrm{Var}_{\boldsymbol{\pi},\boldsymbol{\Psi}}((\mathbf{Sb}^*)_i) \le \frac{1}{L}\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{b}^*\|_2^2) - f_i^{*2}, \tag{8.20}$$

and by the Cauchy–Schwarz inequality,

$$\frac{1}{L}\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{b}^*\|_2^2) = \mathbb{E}_{\pi,\boldsymbol{\psi}}(b_1^{*2}) \leq \left\{ \sqrt{\mathbb{E}_{\pi,\boldsymbol{\psi}}\left(b_1^{*,(1)^2}\right)} + \sqrt{\mathbb{E}_{\pi,\boldsymbol{\psi}}\left(b_1^{*,(2)^2}\right)} \right\}^2$$
$$= \frac{1}{L}\left\{ \sqrt{\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{b}^{*,(1)}\|_2^2)} + \sqrt{\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{b}^{*,(2)}\|_2^2)} \right\}^2. \tag{8.21}$$

We now compute $\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{b}^{*,(2)}\|_2^2)$ as follows.

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{b}^{*,(2)}\|_2^2) &= \frac{p^2q^2}{L}\mathbb{E}_{\pi,\boldsymbol{\psi}}\left\{ \left(\sum_{k=1}^{p} \psi_k \sum_{k_1=1}^{p} \Theta_{kk_1}^{*,(2)} \mathbb{1}_{\{\pi(k_1)<\pi(k)\}} W_{2\pi(k)} \right)^2 \right\} \\
&= \frac{p^2q^2}{L}\mathbb{E}_{\pi}\left[\mathbb{E}_{\boldsymbol{\psi}}\left\{ \left(\sum_{k=1}^{p} \psi_k \sum_{k_1=1}^{p} \Theta_{kk_1}^{*,(2)} \mathbb{1}_{\{\pi(k_1)<\pi(k)\}} \sum_{\ell=2}^{p} W_{2\ell} \mathbb{1}_{\{\pi(k)=\ell\}} \right)^2 \middle| \pi \right\} \right] \\
&= \frac{p^2q^2}{L}\sum_{k=1}^{p}\sum_{\ell=2}^{p} W_{2\ell}^2 \mathbb{E}_{\pi}\left\{ \mathbb{1}_{\{\pi(k)=\ell\}} \left(\sum_{k_1=1}^{p} \Theta_{kk_1}^{*,(2)} \mathbb{1}_{\{\pi(k_1)<\ell\}} \right)^2 \right\} \\
&= \frac{p^2q^2}{L}\sum_{\ell=2}^{p} W_{2\ell}^2 \left(\frac{\ell-1}{p(p-1)} \sum_{k,k_1} (\Theta_{kk_1}^{*,(2)})^2 + \frac{(\ell-1)(\ell-2)}{p(p-1)(p-2)} \sum_{k_1\neq k_2} \sum_{k} \Theta_{kk_1}^{*,(2)} \Theta_{kk_2}^{*,(2)} \right) \\
&\leq \frac{pq^2}{(p-1)L} \sum_{k,k_1,k_2} \left| \Theta_{kk_1}^{*,(2)} \Theta_{kk_2}^{*,(2)} \right| \sum_{\ell=2}^{p} (\ell-1) W_{2\ell}^2.
\end{aligned} \tag{8.22}$$

By the Cauchy–Schwarz inequality, choosing

$$W_{2\ell} = \frac{\binom{p-\ell}{q-1} \Big/ \binom{p-1}{q}}{\sum_{\ell'=2}^{p} (\ell'-1) \left\{ \binom{p-\ell}{q-1} \Big/ \binom{p-1}{q} \right\}^2} \tag{8.23}$$

minimises (8.22) subject to (8.19) to give

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

Finally, Lemma 12 bounds the right-most term from above to yield

$$\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{b}^{*,(2)}\|_2^2) \leq \frac{2\{(2-\delta)q\}^2}{L} \sum_{k,k_1,k_2} \left| \Theta_{kk_1}^{*,(2)} \Theta_{kk_2}^{*,(2)} \right|. \tag{8.24}$$

Using property (ii) of Theorem 1 and substituting (8.24) into (8.21) gives (ii) of Theorem 5. Substituting this into (8.20) then gives property (iii). □

Proof of Theorem 6 Let $\mathbf{b}^* \in \mathbb{R}^L$ be as in Theorem 5. The proof is almost identical to that of Theorem 2, the only difference being in the approximation error term $\boldsymbol{\Delta}$, here defined by

$$\boldsymbol{\Delta} = \mathbf{f}^* - \mathbf{S}\mathbf{b}^*.$$

Analogously to (8.10), we get

$$\begin{aligned}\mathrm{MSPE}((\hat{\alpha}, \hat{\mathbf{b}})) &\leq \frac{1}{n}\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\boldsymbol{\Delta}\|_2^2) + \frac{\sigma^2(L+1)}{n} + \frac{1}{n}\|\boldsymbol{\gamma}\|_2^2 \\ &\leq \frac{(2-\delta)q\ell^2(\boldsymbol{\Theta}^*)}{L} + \frac{\sigma^2(L+1)}{n} + \frac{1}{n}\|\boldsymbol{\gamma}\|_2^2,\end{aligned}$$

the final line following from property (iii) in Theorem 5. □

Proof of Theorem 7 Let $\mathbf{b}^* \in \mathbb{R}^L$ be as in Theorem 5. The proof is almost identical to that of Theorem 3, the only differences being in the definition of the event Λ and the upper bound on the approximation error term. Here, we take

$$\Lambda = \left\{ \|\mathbf{b}^*\|_2^2 < (1+\eta)\frac{(2-\delta)q\ell^2(\boldsymbol{\Theta}^*)}{L} \right\},$$

and we note that Lemma 10 entails $\mathbb{P}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\Lambda^c) \leq \rho_2$, with ρ_2 defined as in (4.4). The proof then follows through as before, replacing the approximation error term in the main effects setting, $\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{S}\mathbf{b}^*\|_2^2)$, with $\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{f}^* - \mathbf{S}\mathbf{b}^*\|_2^2)$, and using property (iii) in Theorem 5 to bound this from above. □

Proof of Theorem 8 Modifying the proof of Theorem 4 in the same way as proof of Theorem 3 is modified to prove Theorem 7 gives the result. □

Lemma 9. *Let $\mathbf{b}^* \in \mathbb{R}^L$ be as in Theorem 5 with $\mathbf{w}$ defined as in (8.4). Then for $\eta \geq 0$,*

$$\mathbb{P}\left(\|\mathbf{b}^*\|_2^2 \geq (1+\eta)\frac{(2-\delta)q\|\boldsymbol{\beta}^*\|_2^2}{L} \right) \leq \exp\left(-\frac{L\eta^2}{18(2-\delta)q\{18(2-\delta)+\eta\}} \right).$$

Proof. We first bound the moments of b_l^{*2}, in order to later apply Bernstein's inequality (see van der Vaart and Wellner (1996)).

$$\begin{aligned}\mathbb{E}(b_l^{*2m}) &= \frac{q^{2m}}{L^{2m}}\mathbb{E}_\pi\left[\mathbb{E}_{\boldsymbol{\psi}}\left\{ \left(\sum_{k=1}^p \beta_k^* \psi_k w_{\pi_l(k)} \right)^{2m} \middle| \pi \right\} \right] \\ &\leq \frac{(2m)!}{2^m m!}\frac{q^{2m}}{L^{2m}}\mathbb{E}_\pi\left\{ \left(\sum_{k=1}^p \beta_k^{*2} w_{\pi(k)}^2 \right)^m \right\} \qquad (8.25)\end{aligned}$$

by Khintchine's inequality (see (Garling, 2007)). Then

$$\begin{aligned}\mathbb{E}(b_l^{*2m}) &\leq \frac{(2m)!}{2^m m!}\frac{q^{2m}}{L^{2m}}\mathbb{E}_\pi\left(\sum_{k=1}^p \beta_k^{*2} w_{\pi(k)}^2 \right) \max_\pi \left(\sum_{k=1}^p \beta_k^{*2} w_{\pi(k)}^2 \right)^{m-1} \\ &\leq \frac{(2m)!}{2^m m!}\frac{q^{2m}}{L^{2m}}\|\boldsymbol{\beta}^*\|_2^{2m}\|\mathbf{w}\|_\infty^{2m-1}\mathbb{E}_\pi(w_{\pi(k)}^2).\end{aligned}$$

By (8.6),

$$\|\mathbf{w}\|_\infty \le \frac{q/p}{q/(2p-q)} = 2-\delta,$$
$$\mathbb{E}_\pi(w^2_{\pi(k)}) \le \frac{2p-q}{qp} = \frac{2-\delta}{q}.$$

Using the inequalities

$$n! \ge \left(\frac{n}{3}\right)^n,$$
$$n! \le \left(\frac{n}{2}\right)^n \quad \text{for } n \ge 6,$$

we get

$$\frac{(2m)!}{2^m (m!)^2} \le \left(\frac{9}{2}\right)^m, \tag{8.26}$$

so

$$\mathbb{E}(b_l^{*2m}) \le m! \frac{1}{q}\left(\frac{9(2-\delta)^2 q^2 \|\boldsymbol{\beta}^*\|_2^2}{2L^2}\right)^m,$$

whence by Minkowski's inequality,

$$\mathbb{E}[\{b_l^{*2} - \mathbb{E}(b_l^{*2})\}^m] \le m! \frac{1}{q}\left(\frac{9(2-\delta)^2 q^2 \|\boldsymbol{\beta}^*\|_2^2}{L^2}\right)^m.$$

Plugging this into Bernstein's inequality, we finally arrive at

$$\begin{aligned}\mathbb{P}\left(\|\mathbf{b}^*\|_2^2 \ge (1+\eta)\frac{(2-\delta)q\|\boldsymbol{\beta}^*\|_2^2}{L}\right) &\le \mathbb{P}\left(\|\mathbf{b}^*\|_2^2 - \mathbb{E}(\|\mathbf{b}^*\|_2^2) \ge \eta\frac{(2-\delta)q\|\boldsymbol{\beta}^*\|_2^2}{L}\right) \\ &\le \exp\left(-\frac{1}{2}\frac{L\eta^2}{9(2-\delta)q\{18(2-\delta)+\eta\}}\right). \qquad \square\end{aligned}$$

Lemma 10. *Let $\mathbf{b}^* \in \mathbb{R}^L$ be as in Theorem 5 with $\mathbf{W}$ defined by equations* (8.17) *and* (8.23)*. Then*

$$\mathbb{P}\left\{\|\mathbf{b}^*\|_2^2 \ge (1+\eta)\frac{(2-\delta)q}{L}\ell^2(\boldsymbol{\Theta})\right\} \le \exp\left(-\frac{L\eta^2}{18(2-\delta)q\{18(2-\delta)+\eta\}}\right) + \exp\left(-\frac{L\eta^2}{36(2-\delta)^2q^2(18+\eta)}\right).$$

Proof. We proceed as in Lemma 9, which provides a bound on the tail probability of $\|\mathbf{b}^{*,(1)}\|_2^2$. To bound the moments of $(b_l{}^{*,(2)})^2$, we argue as follows.

$$\begin{aligned}\mathbb{E}\{(b_l{}^{*,(2)})^{2m}\} &= \frac{p^{2m}q^{2m}}{L^{2m}}\mathbb{E}_\pi\left[\mathbb{E}_{\boldsymbol{\psi}}\left\{\left(\sum_{k=1}^p \psi_k \sum_{k_1=1}^p \Theta^{*,(2)}_{kk_1}\mathbb{1}_{\{\pi(k_1)<\pi(k)\}}W_{2\pi(k)}\right)^{2m}\Big|\pi\right\}\right] \\ &\le \frac{(2m)!}{2^m m!}\frac{p^{2m}q^{2m}}{L^{2m}}\mathbb{E}_\pi\left[\left\{\sum_{k=1}^p\left(\sum_{k_1=1}^p \Theta^{*,(2)}_{kk_1}\mathbb{1}_{\{\pi(k_1)<\pi(k)\}}\right)^2 W^2_{2\pi(k)}\right\}^m\right]\end{aligned}$$

by Khintchine's inequality. Now

$$\mathbb{E}_\pi\left[\left\{\sum_{k=1}^p\left(\sum_{k_1=1}^p \Theta^{*,(2)}_{kk_1}\mathbb{1}_{\{\pi(k_1)<\pi(k)\}}\right)^2 W^2_{2\pi(k)}\right\}^m\right] \leq \mathbb{E}_\pi\left\{\sum_{k=1}^p\left(\sum_{k_1=1}^p \Theta^{*,(2)}_{kk_1}\mathbb{1}_{\{\pi(k_1)<\pi(k)\}}\right)^2 W^2_{2\pi(k)}\right\}$$
$$\times \max_\pi\left\{\sum_{k=1}^p\left(\sum_{k_1=1}^p \Theta^{*,(2)}_{kk_1}\mathbb{1}_{\{\pi(k_1)<\pi(k)\}}\right)^2 W^2_{2\pi(k)}\right\}^{m-1}$$
$$\leq \frac{2(2-\delta)^2}{p^2}\left(\sum_{k,k_1,k_2}\left|\Theta^{*,(2)}_{kk_1}\Theta^{*,(2)}_{kk_2}\right|\right)^m \|\mathbf{w}_2\|_\infty^{2(m-1)},$$

the final line following from equations (8.22) and (8.24).

Next, appealing to Lemma 12, we see that $\|\mathbf{w}_2\|_\infty \leq 2(2-\delta)^2 q/p$. Thus we have

$$\mathbb{E}\{(b_l{}^{*,(2)})^{2m}\} \leq \frac{(2m)!}{2^m m!}\frac{1}{2(2-\delta)^2q^2}\left(\frac{4(2-\delta)^4q^4}{L^2}\sum_{k,k_1,k_2}\left|\Theta^{*,(2)}_{kk_1}\Theta^{*,(2)}_{kk_2}\right|\right)^m,$$

so

$$\mathbb{E}[\{(b_l{}^{*,(2)})^2 - \mathbb{E}((b_l{}^{*,(2)})^2)\}^m] \leq m!\frac{1}{2(2-\delta)^2q^2}\left(\frac{36(2-\delta)^4q^4}{L^2}\sum_{k,k_1,k_2}\left|\Theta^{*,(2)}_{kk_1}\Theta^{*,(2)}_{kk_2}\right|\right)^m,$$

by (8.26) and Minkowski's inequality. By Bernstein's inequality,

$$\mathbb{P}\left(\|\mathbf{b}^{*,(2)}\|_2^2 - \mathbb{E}(\|\mathbf{b}^{*,(2)}\|_2^2) \geq \eta\frac{2\{(2-\delta)q\}^2}{L}\sum_{k,k_1,k_2}\left|\Theta^{*,(2)}_{kk_1}\Theta^{*,(2)}_{kk_2}\right|\right) \leq \exp\left(-\frac{L\eta^2}{36(2-\delta)^2q^2(18+\eta)}\right).$$

Putting things together, we have

$$\mathbb{P}\left\{\|\mathbf{b}^*\|_2^2 \geq (1+\eta)\frac{(2-\delta)q}{L}\ell^2(\mathbf{\Theta})\right\} \leq \mathbb{P}\left\{\left(\|\mathbf{b}^{*,(1)}\|_2 + \|\mathbf{b}^{*,(2)}\|_2\right)^2 \geq (1+\eta)\frac{(2-\delta)q}{L}\ell^2(\mathbf{\Theta})\right\}$$
$$\leq \mathbb{P}\left(\|\mathbf{b}^{*,(1)}\|_2^2 \geq (1+\eta)\frac{(2-\delta)q\|\boldsymbol{\theta}^{*(1)}\|_2^2}{L}\right)$$
$$+ \mathbb{P}\left(\|\mathbf{b}^{*,(2)}\|_2^2 - \mathbb{E}(\|\mathbf{b}^{*,(2)}\|_2^2) \geq \eta\frac{2\{(2-\delta)q\}^2}{L}\sum_{k,k_1,k_2}\left|\Theta^{*,(2)}_{kk_1}\Theta^{*,(2)}_{kk_2}\right|\right)$$
$$\leq \exp\left(-\frac{L\eta^2}{18(2-\delta)q\{18(2-\delta)+\eta\}}\right) + \exp\left(-\frac{L\eta^2}{36(2-\delta)^2q^2(18+\eta)}\right),$$

using Lemma 9 in the final line. □

Lemma 11. *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*

$$a_i \le b_i \quad \text{for all } i \le i^*,$$
$$a_i \ge b_i \quad \text{for all } i > i^*.$$

Then

(i) If

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

then

$$\sum_{i=1}^{\infty} a_i^2 \le \sum_{i=1}^{\infty} b_i^2.$$

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

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

then

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

Proof. For (i), observe that

$$\sum_{i=1}^{i^*}(b_i^2 - a_i^2) = \sum_{i=1}^{i^*}(b_i + a_i)(b_i - a_i) \ge (b_{i^*} + a_{i^*}) \sum_{i=1}^{i^*}(b_i - a_i)$$
$$\ge (b_{i^*} + a_{i^*}) \sum_{i>i^*}(a_i - b_i) \ge \sum_{i>i^*}(b_i + a_i)(a_i - b_i) = \sum_{i>i^*}({a_i}^2 - {b_i}^2).$$

For (ii) we argue,

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

□

Lemma 12. *Let $q, p \in \mathbb{N}$ with $q \ge 1$, $p \ge \max\{q, 3\}$. We have*

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

Proof. For $1 \le \ell \le p-1$, let $c_\ell = \ell$,

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

Setting all elements with indices at least $\ell = p$ to zero, by (8.5), the sequences $(a_\ell)_{\ell=1}^\infty$, $(b_\ell)_{\ell=1}^\infty$ and $(c_\ell)_{\ell=1}^\infty$ satisfy the hypotheses of Lemma 11. Thus

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

and

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

Noting that

$$1 \ge \frac{1}{2} + \frac{1}{2}\left(\frac{(p-1)^2}{\{2(p-1)-q\}q} - \left\lfloor \frac{(p-1)^2}{\{2(p-1)-q\}q} \right\rfloor \right),$$

after some simplification, we obtain

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

□