

# Optimal Subsampling for Large Sample Ridge Regression

Yunlu Chen<sup>1</sup> and Nan Zhang<sup>\*2</sup>

<sup>1,2</sup>*Fudan University, Shanghai 200433, China*

## Abstract

Subsampling is a popular approach to alleviating the computational burden for analyzing massive datasets. Recent efforts have been devoted to various statistical models without explicit regularization. In this paper, we develop an efficient subsampling procedure for the large sample linear ridge regression. In contrast to the ordinary least square estimator, the introduction of the ridge penalty leads to a subtle trade-off between bias and variance. We first investigate the asymptotic properties of the subsampling estimator and then propose to minimize the asymptotic-mean-squared-error criterion for optimality. The resulting subsampling probability involves both ridge leverage score and  $\ell_2$  norm of the predictor. To further reduce the computational cost for calculating the ridge leverage scores, we propose the algorithm with efficient approximation. We show by synthetic and real datasets that the algorithm is both statistically accurate and computationally efficient compared with existing subsampling based methods.

**Keywords:** Big data; Ridge regression; Subsampling method; Ridge leverage score.

## 1 Introduction

Linear regression is a popular method to depict the relationship between the response variable  $y \in \mathcal{Y}$  and the covariate  $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^p$ . Observing  $n$  independent and identically distributed data  $\mathcal{F}_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ , we consider the linear model  $y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i$ ,  $i = 1, \dots, n$ , where  $\epsilon_i$  is the

---

\*ZHANG Nan, corresponding author, E-mail: zhangnan@fudan.edu.cn. The authors gratefully acknowledge the support by National Natural Science Foundation of China, Grant Number: 11690014; Science and Technology Commission of Shanghai Municipality, Grant Number: 17JC1420200.

independent and identically distributed error term with mean zero and variance  $\sigma^2$ . The model can be written in the matrix form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where  $\mathbf{y} = (y_1, \dots, y_n)^\top$  is a response vector,  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top$  is an  $n \times p$  design matrix,  $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^\top$  is an error vector. The ordinary least square approach minimizes  $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$  and leads to  $\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$  provided that  $\mathbf{X}^\top \mathbf{X}$  is invertible. Ridge regression, proposed by Hoerl and Kennard (1970) 50 years ago, provides a remedy for ill-conditioned  $\mathbf{X}^\top \mathbf{X}$  in computing the ordinary least square estimator. The ridge regression estimator is defined by adding a ridge on the diagonal of  $\mathbf{X}^\top \mathbf{X}$ , that is,

$$\hat{\boldsymbol{\beta}} = \left( \mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I} \right)^{-1} \mathbf{X}^\top \mathbf{y}, \quad (1.1)$$

where  $\lambda > 0$  is called the ridge parameter.

The optimization function for the ridge regression estimator can be written as

$$\min_{\boldsymbol{\beta}} \{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|^2 \}. \quad (1.2)$$

The ridge penalty introduces bias to the estimator, while the variance is reduced at the same time. It leads to a bias-variance trade-off when we attempt to predict at a new location, see e.g., Hastie et al. (2009); Hastie (2020). Tuning the ridge parameter  $\lambda$  is critical for balancing the bias and variance of the estimator. Typical methods for choosing the ridge parameter include the cross-validation and the generalized cross-validation (Golub et al., 1979).

With massive data, it is often computationally prohibitive to calculate the estimator when either the sample size or the dimension is super large. In recent years, many research efforts have been devoted to addressing the computational issue due to the large data matrix. Kumar et al. (2012) explored the sampling approach for the column subset selection problem by the Nyström method. Dereziński et al. (2020) recently provided an improved theoretical guarantee for low-rank approximations of large datasets. Another popular idea in machine learning is coresets, which constructs estimators based on sub-data. Kacham and Woodruff (2020) utilized the spectral graph sparsification result of Batson et al. (2012) and proposed to merge the coresets obtained from multiple servers. Mahoney (2011); Woodruff (2014) studied matrix sketching to generate smaller datasets with random projections. Wang et al. (2017) addressed the statistical and algorithmic properties

of classical sketch and Hessian sketch. Recently, under the context of ridge regression models, ridge leverage scores, introduced by Alaoui and Mahoney (2015), are defined as the diagonal elements of matrix  $\mathbf{X}(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top$ . Cohen et al. (2017) extended the concept and proposed a low-rank projection-based approach via ridge leverage score sampling. Homrighausen and McDonald (2020) provided the approximated bias and variance for ridge regression but under the special condition that the compression matrix is of a sparse Bernoulli form.

Subsampling can be viewed as a special case of random projection or sketching. A general subsampling procedure is basically to first select a subsample from the original dataset according to certain subsampling probabilities and then construct an estimator via only the subsample. The efficiency of implementation and nice interpretability make subsampling-based methods attractive. Based on the characteristics of the sampling step, existing methods can be summarized into two categories: deterministic and randomized subsampling. For the deterministic approach, Wang et al. (2019) proposed to select the subsample with extreme values on each dimension of  $\mathbf{X}$  in linear regression such that the information matrix has a well-controlled determinant value. The second approach, the randomized subsampling, assigns subsampling probabilities to each observation and can achieve certain optimality by minimizing various criteria from the theory of experimental design. Drineas et al. (2011), Ma et al. (2015) and Ma et al. (2020) investigated the optimal subsampling for large sample linear regression via leverage scores, i.e., the diagonal elements of  $\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ . Such strategy has inspired further studies on versatile statistical models including logistic regression (Wang et al., 2018), quantile regression (Wang and Ma, 2020) and generalized linear models (Ai et al., 2018).

Our goal in this paper is to alleviate the computational burden for ridge regression with large-scale datasets. In particular, we focus on the case where the full sample size  $n$  is much larger than the dimension  $p$ . Motivated from the idea of subsampling which concerns the asymptotic result (Ma et al., 2020, 2015), we study the bias and variance of the regression coefficient estimator from the subsample. Taking the bias-variance trade-off into consideration, we propose to minimize the asymptotic-mean-squared-error criterion and show that the optimal subsampling probability for each observation depends on not only its ridge leverage score but also the  $\ell_2$  norm of the covariate. Unlike existing subsampling methods for large sample regression models where no penalty term is

involved, it plays an important role to select a proper ridge parameter. Although the derived optimal subsampling probabilities have explicit forms, it is unrealistic to directly apply them because quantities include the ridge parameter and the ridge leverage scores are computationally expensive to calculate. On the one hand, conventional methods for choosing the ridge parameter such as the cross-validation and the generalized cross-validation are time-consuming when applied to the full sample. However, based on the relationship between the best ridge parameter for the full sample and that for the subsample, we can instead apply the cross-validation on the subsample and extrapolate it to the full sample. On the other hand, for efficient approximation, we replace individual ridge leverage scores with their average. As a consequence, the optimal subsampling probabilities are proportional to the  $\ell_2$  norms of predictors. Based on the aforementioned adjustments, our new method exhibits better performance with efficient computation than other sketching and subsampling algorithms over extensive simulation studies, especially when the subsample size is small.

The rest of the paper is organized as follows. Section 2 presents the framework of the subsampling method and explains the details of ridge parameter selection and the optimal subsampling criterion. Section 3 proposes the optimal subsampling algorithm. Section 4 and Section 5 demonstrate the practical effectiveness of our algorithms via simulation and application, respectively.

## 2 Methodology

### 2.1 Subsampling framework

To reduce the computation when dealing with datasets of large sample size  $n$ , the key step of a general subsampling procedure is to select a subsample of size  $r \ll n$  from the original observations according to subsampling probabilities. Extending the weighted estimation algorithm raised in Ma et al. (2015) to the ridge regression, we present the following framework for the ridge regression estimator  $\tilde{\beta}$ . Our proposed algorithms are based on this basic framework with its details shown in Section 3.

Step 1. Construct the subsampling probability for each sample  $\{\pi_i\}_{i=1}^n$ . Draw a subsample  $(\mathbf{X}^*, \mathbf{y}^*)$  of size  $r \ll n$  based on the probability.

Step 2. Determine the ridge parameter  $\tilde{\lambda}$  for the subsample. Calculate the ridge regression estimator using the subsample, i.e.,

$$\tilde{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \|\Phi^* \mathbf{y}^* - \Phi^* \mathbf{X}^* \boldsymbol{\beta}\|^2 + \tilde{\lambda} \|\boldsymbol{\beta}\|^2, \quad (2.1)$$

where  $\Phi^* = \text{diag} \{1/\sqrt{r\pi_k^*}\}_{k=1}^r$ .

Under the above general subsampling framework, two key questions remain to be answered:

1. How to determine the ridge parameter for subsample  $\tilde{\lambda}$ ?
2. What is the optimal subsampling probability for each sample  $\{\pi_i\}_{i=1}^n$ ?

## 2.2 Ridge parameter selection

The regularization in ridge regression plays an essential role in the prediction performance of the estimator. The ridge parameter  $\lambda$  is usually unknown and requires careful tuning. Taking both the bias and variance into consideration, we can view the mean squared error as a function of the ridge parameter, and define the optimal ridge parameter as the one corresponding to the smallest mean squared error. For example, when the design matrix has orthonormal columns, the mean squared error can be derived as  $p\sigma^2(1 + \lambda)^{-2} + \lambda^2(1 + \lambda)^{-2}\boldsymbol{\beta}^\top \boldsymbol{\beta}$ , and thus the optimal ridge parameter is  $p\sigma^2/\boldsymbol{\beta}^\top \boldsymbol{\beta}$ . In practice, the cross-validation and its variants are applied to obtain the optimal ridge parameter. For  $K$ -fold cross-validation, the training data is divided into  $K$  partitions  $\{\mathbf{X}^{(k)}, \mathbf{y}^{(k)}\}_{k=1}^K$  and we denote by  $\hat{\boldsymbol{\beta}}_{\setminus k}(\lambda)$  the estimated coefficient based on all partitions except the  $k$ th one. The optimal ridge parameter is

$$\lambda_{K\text{-fold}} = \arg \min_{\lambda} K^{-1} \sum_{k=1}^K \|\mathbf{y}^{(k)} - \mathbf{X}^{(k)} \hat{\boldsymbol{\beta}}_{\setminus k}(\lambda)\|^2.$$

The repeated fitting process by using different parts of the original sample leads to a high computational cost, especially when the sample size is large. Golub et al. (1979) proposed the generalized cross-validation to reduce the computation cost of cross-validation. Consider the leave-one-out cross-validation, i.e.,  $K = n$ . It can be shown that

$$\lambda_{\text{LOOCV}} = \arg \min_{\lambda} n^{-1} \sum_{i=1}^n \left\{ \frac{y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}(\lambda)}{1 - h_{ii}(\lambda)} \right\}^2,$$

where  $h_{ii} = \mathbf{x}_i^\top (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{x}_i$  is the diagonal element of the hat matrix  $\mathbf{H} = \mathbf{X} (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top$  where  $i = 1, \dots, n$ . Quantity  $h_{ii}$  measures the influential effect of the  $i$ th data points upon prediction and is called  $\lambda$ -leverage score (Alaoui and Mahoney, 2015). The computational cost of calculating the ridge leverage score is  $\mathcal{O}(np^2)$ . The generalized cross-validation replaces individual leverage scores with their average  $\text{tr}(\mathbf{H})$  to reduce computation, that is,

$$\lambda_{\text{GCV}} = \arg \min_{\lambda} n^{-1} \sum_{i=1}^n \left\{ \frac{y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}(\lambda)}{1 - n^{-1} \text{tr}(\mathbf{H})} \right\}^2. \quad (2.2)$$

The optimal ridge parameter  $\tilde{\lambda}$  for the subsample can be estimated by minimizing the cross-validation criteria but over the subsample. We next provide the rationale of the approximation in the main theorem.

### 2.3 Optimal subsampling

Once the ridge parameter is fixed, we calculate subsampling probabilities for each observation, by which a subset of data points are selected from the full sample with replacement. We anticipate the estimator based on the subsample can achieve some optimality. In the ridge regression, the regularized estimator can perform much better than the ordinary least square estimator if the bias and variances are traded off properly. Therefore, we consider the mean-squared-error type of criterion which involves both bias and variance.

We begin by investigating the difference between the subsampling estimator and its full-sample counterpart. For convenience of analysis, we introduce some notations here to rewrite the subsampling estimator in the form concerning full data. Let  $K_i$  be the number of times the observation  $\mathbf{x}_i$  is sampled and  $(K_1, \dots, K_n)$  thus follows a multinomial distribution. Let  $\mathbf{W} = \mathbf{\Omega} \mathbf{K}$ , where  $\mathbf{K} = \text{diag}\{K_i\}_{i=1}^n$ ,  $\mathbf{\Omega} = \text{diag}\{1/r\pi_i\}_{i=1}^n$ . Simple algebra yields that the ridge regression estimator based on the subsample from (2.1) can be expressed as

$$\tilde{\boldsymbol{\beta}} = \left( \mathbf{X}^{*\top} \Phi^{*2} \mathbf{X}^* + \tilde{\lambda} \mathbf{I} \right)^{-1} \mathbf{X}^{*\top} \Phi^{*2} \mathbf{y}^* = \left( \mathbf{X}^\top \mathbf{W} \mathbf{X} + \tilde{\lambda} \mathbf{I} \right)^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{y}. \quad (2.3)$$

In the following lemma, we demonstrate the difference between the estimator (2.3) and the full-sample estimator (1.1).

**Lemma 1.** If  $0 < \pi_i < 1, i = 1, \dots, n$ ,  $(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top (\mathbf{W} - \mathbf{I}) \mathbf{X} = \mathcal{O}_p(r^{-1/2})$  and  $(\tilde{\lambda} - \lambda)(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} = \mathcal{O}(r^{-1/2})$ , then

$$\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{e} - \tilde{\lambda} (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \hat{\boldsymbol{\beta}} + \mathcal{O}_p(r^{-1}), \quad (2.4)$$

where  $\mathbf{e} = \mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}$ .

*Proof.* We first rewrite the subsampling estimator by multiplying  $(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1}$ ,

$$\tilde{\boldsymbol{\beta}} = \left\{ (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} (\mathbf{X}^\top \mathbf{W} \mathbf{X} + \tilde{\lambda} \mathbf{I}) \right\}^{-1} \left\{ (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{y} \right\}. \quad (2.5)$$

For the inverse term, we apply the Taylor series expansion,

$$\begin{aligned} \left\{ (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} (\mathbf{X}^\top \mathbf{W} \mathbf{X} + \tilde{\lambda} \mathbf{I}) \right\}^{-1} &= \left[ \mathbf{I} + (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \{ \mathbf{X}^\top (\mathbf{W} - \mathbf{I}) \mathbf{X} + (\tilde{\lambda} - \lambda) \mathbf{I} \} \right]^{-1} \\ &= \mathbf{I} - (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top (\mathbf{W} - \mathbf{I}) \mathbf{X} - (\tilde{\lambda} - \lambda) (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} + \mathcal{O}_p(r^{-1}). \end{aligned}$$

For the other term in (2.5),

$$\begin{aligned} (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{y} &= (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \left\{ \mathbf{X}^\top \mathbf{y} + \mathbf{X}^\top (\mathbf{W} - \mathbf{I}) \mathbf{y} \right\} \\ &= \hat{\boldsymbol{\beta}} + (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top (\mathbf{W} - \mathbf{I}) \mathbf{y}. \end{aligned}$$

Since  $(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top (\mathbf{W} - \mathbf{I}) \mathbf{y}$  and  $(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top (\mathbf{W} - \mathbf{I}) \mathbf{e}$  are of the same order as  $(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top (\mathbf{W} - \mathbf{I}) \mathbf{X}$ ,

$$\begin{aligned} \tilde{\boldsymbol{\beta}} &= \hat{\boldsymbol{\beta}} + (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top (\mathbf{W} - \mathbf{I}) \mathbf{e} + (\lambda - \tilde{\lambda}) (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \hat{\boldsymbol{\beta}} + \mathcal{O}_p(r^{-1}) \\ &= \hat{\boldsymbol{\beta}} + (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{e} - \tilde{\lambda} (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \hat{\boldsymbol{\beta}} + \mathcal{O}_p(r^{-1}) \end{aligned}$$

The second equation holds due to the normal equation for ridge regression.  $\square$

We investigate the asymptotic mean squared error of  $\tilde{\boldsymbol{\beta}}$ , which is used as our criterion for determining the optimal subsampling probabilities.

**Theorem 1.** If the full sample size  $n$  is fixed,  $\|\mathbf{x}_i\| < \infty, i = 1, \dots, n$ , the sampling probabilities  $\{\pi_i\}_{i=1}^n$  are nonzero, and  $\tilde{\lambda} - \lambda = \mathcal{O}(r^{-1/2})$ , then the asymptotic variance and mean are

1.  $AVar(\tilde{\boldsymbol{\beta}}) = \Sigma_c - \lambda^2 r^{-1} (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \hat{\boldsymbol{\beta}} \hat{\boldsymbol{\beta}}^\top (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1}$ , where  $\Sigma_c = r^{-1} (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} (\sum_{i=1}^n \pi_i^{-1} e_i^2 \mathbf{x}_i \mathbf{x}_i^\top) (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1}$ ,  $e_i = y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$ .

$$2. AE(\tilde{\boldsymbol{\beta}}) = \hat{\boldsymbol{\beta}} + (\lambda - \tilde{\lambda})(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \hat{\boldsymbol{\beta}}.$$

The asymptotic mean squared error of  $\tilde{\boldsymbol{\beta}}$  is therefore

$$AMSE(\tilde{\boldsymbol{\beta}}) = \Sigma_c - \lambda^2 r^{-1} (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \hat{\boldsymbol{\beta}} \hat{\boldsymbol{\beta}}^\top (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} + (\lambda - \tilde{\lambda})^2 (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \hat{\boldsymbol{\beta}} \hat{\boldsymbol{\beta}}^\top (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1}.$$

*Proof.* We begin by the deduction of the result 1. The roadmap of the proof of this part is motivated from the result in Ma et al. (2020) since the variance term in linear regression case is of similar form. We use Cramer-Wold device to establish the asymptotic normality of  $(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{e} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \sum_{j=1}^r \mathbf{X}^\top \boldsymbol{\Omega} \mathbf{K}^{(j)} \mathbf{e}$ . Consider each term in the summation, for any non-zero constant vector  $\mathbf{b} \in \mathbb{R}^p$ , we have

$$\text{var}\{(\mathbf{b}^\top \mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \boldsymbol{\Omega} \mathbf{K}^{(1)} \mathbf{e}\} = r^{-1} \mathbf{a}^\top \left( \sum_{i=1}^n \frac{e_i^2}{\pi_i} \mathbf{x}_i \mathbf{x}_i^\top \right) \mathbf{a} - r^{-1} \mathbf{a}^\top \mathbf{X}^\top \mathbf{e} \mathbf{e}^\top \mathbf{X} \mathbf{a},$$

where  $\mathbf{a} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{b}$ . By applying the normal equation of ridge regression, we have

$$\text{var}\{(\mathbf{b}^\top \mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \boldsymbol{\Omega} \mathbf{K}^{(1)} \mathbf{e}\} = \mathbf{b}^\top \left\{ \Sigma_c + \lambda^2 r^{-1} (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \hat{\boldsymbol{\beta}} \hat{\boldsymbol{\beta}}^\top (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \right\} \mathbf{b}.$$

By using Lindeberg-Lévy CLT, we have the variance of the summation,

$$\text{var}\{(\mathbf{b}^\top (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \sum_{j=1}^r \mathbf{X}^\top \boldsymbol{\Omega} \mathbf{K}^{(j)} \mathbf{e})\} = \mathbf{b}^\top \left\{ \Sigma_c + \lambda^2 r^{-1} (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \hat{\boldsymbol{\beta}} \hat{\boldsymbol{\beta}}^\top (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \right\} \mathbf{b}$$

Therefore, we have the result 1 due to the Cramer-Wold device.

Then consider the result 2. Since  $\mathbf{W} = \boldsymbol{\Omega} \mathbf{K}$ , where each element  $K_i$  in  $\mathbf{K} = \text{diag}\{K_i\}_{i=1}^n$  follows the multinomial distribution  $\text{Mult}(r, \{\pi_i\}_{i=1}^n)$ , then  $\mathbb{E}(W_i) = 1$ , where  $W_i$  is the diagonal element of matrix  $\mathbf{W}$ . Thus, we can calculate the expectation

$$\begin{aligned} \mathbb{E}\{(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{e}\} &= (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{e} \\ &= \lambda (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \hat{\boldsymbol{\beta}}, \end{aligned}$$

with the second equation following the normal equation. Consequently, we have  $AE(\tilde{\boldsymbol{\beta}}) = \hat{\boldsymbol{\beta}} + (\lambda - \tilde{\lambda})(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \hat{\boldsymbol{\beta}}$ .  $\square$

The above theorem shows that the controllable part of the criterion is included in the variance term. Considering the expression of variance, only  $\Sigma_c$  depends on the subsampling probability  $\{\pi_i\}_{i=1}^n$ . We resort to minimize the expected trace of  $\Sigma_c$  to obtain the corresponding subsampling probability, as shown in the following theorem.



**Theorem 2.** *When*

$$\pi_i = \frac{\sqrt{(1 - h_{ii})} \|\mathbf{x}_i\|}{\sum_{j=1}^n \sqrt{(1 - h_{jj})} \|\mathbf{x}_j\|},$$

$\mathbb{E}\{\text{tr}(\Sigma_c)\}$  attains its minimum, where  $h_{ii}$  is the ridge leverage score,  $i = 1, \dots, n$ .

*Proof.* Since  $\mathbb{E}\{\text{tr}(\Sigma_c)\} = r^{-1} \sum_{i=1}^n \pi_i^{-1} (1 - h_{ii}) \|\mathbf{x}_i\|^2$ , we can get the following result by applying Hölder's inequality,

$$r^{-1} \sum_{i=1}^n \frac{1 - h_{ii}}{\pi_i} \|\mathbf{x}_i\|^2 = r^{-1} \sum_{i=1}^n \frac{1 - h_{ii}}{\pi_i} \|\mathbf{x}_i\|^2 \sum_{i=1}^n \pi_i \geq \left\{ \sum_{i=1}^n \sqrt{(1 - h_{ii})} \|\mathbf{x}_i\|^2 \right\}^2,$$

with the equality holds if and only if  $\pi_i \propto \sqrt{(1 - h_{ii})} \|\mathbf{x}_i\|$ .  $\square$

In Theorem 2, we obtain the optimal subsampling probability for each observation. It involves both the ridge leverage score and the  $\ell_2$  norm of the predictor. Our subsampling strategy is different from the sketching scheme for ridge regression (Cohen et al., 2017) which only utilized the ridge leverage score. We will compare these methods on simulation and real data.

### 3 Algorithm

Based on the deduction in the above section, we obtain the subsampling probability and the approaching rate between the ridge parameter for subsample  $\tilde{\lambda}$  and that for full sample  $\lambda$ . Integrating these ingredients with the general subsampling procedure,  $\tilde{\lambda} = \lambda$  and  $\pi_i = \frac{\sqrt{(1 - h_{ii})} \|\mathbf{x}_i\|}{\sum_{j=1}^n \sqrt{(1 - h_{jj})} \|\mathbf{x}_j\|}$ ,  $i = 1, \dots, n$ , we can calculate the subsampling estimator. However, there are two quantities whose calculations are still computationally demanding. First, it requires  $\mathcal{O}(np^2)$  to compute the exact ridge leverage scores, which amounts to the same computation as the full-sample estimator. Second, we need to calculate  $\lambda$  first for calculating the ridge leverage score. The ridge parameter  $\tilde{\lambda}$  is then set as  $\lambda$ , which is not desirable since applying the cross-validation to choose  $\lambda$  is time-consuming.

To address the aforementioned issues, we propose an efficient approximation to the optimal subsampling probabilities in Theorem 2. Similar to the idea of generalized cross-validation in (2.2), we approximate the individual ridge leverage score with their average, i.e.,  $n^{-1} \text{tr}(\mathbf{H})$ . It corresponds to the scenario where ridge leverage scores are not highly heterogeneous. Therefore, the subsampling probability reduces to  $\pi_i = \|\mathbf{x}_i\| / \sum_{j=1}^n \|\mathbf{x}_j\|$ ,  $i = 1, \dots, n$ , which involves only the  $\ell_2$  norm of the predictors. Moreover, such an approximation of the subsampling probability

no longer depends on the ridge leverage score, and hence we do not need the full sample  $\lambda$  to calculate  $\pi_i$ . In this way, we can perform the cross-validation or generalized cross-validation to directly calculate  $\tilde{\lambda}$  with the selected subsample at a much lower computational cost. The optimal subsampling ridge regression estimation we propose is summarized in Algorithm 1.

**Algorithm 1.** *Ridge Regression with Optimal Subsampling*

*Step 1. Construct the subsampling probability for each sample  $\pi_i = \|\mathbf{x}_i\| / \sum_{j=1}^n \|\mathbf{x}_j\|, i = 1, \dots, n$ .*

*Draw a subsample  $(\mathbf{X}^*, \mathbf{y}^*)$  of size  $r \ll n$  based on the probability.*

*Step 2. Calculate the ridge regression estimator*

$$\tilde{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \|\Phi^* \mathbf{y}^* - \Phi^* \mathbf{X}^* \boldsymbol{\beta}\|^2 + \tilde{\lambda} \|\boldsymbol{\beta}\|^2, \quad (3.1)$$

*where  $\tilde{\lambda}$  is selected via cross-validation for response vector  $\Phi^* \mathbf{y}^*$  and design matrix  $\Phi^* \mathbf{X}^*$  and  $\Phi^* = \text{diag} \{1/\sqrt{r\pi_k^*}\}_{k=1}^r$ .*

Compared to calculating the exact ridge leverage score, the calculation of subsampling probabilities in Step 1 of Algorithm 1 avoids accessing the full design matrix  $\mathbf{X}$ , and hence the  $\ell_2$  norm of predictors can be obtained in parallel. According to the closeness between  $\tilde{\lambda}$  and  $\lambda$  revealed in Theorem 1, we can extrapolate  $\tilde{\lambda}$  for the subsample to estimate  $\lambda$  for the full sample. It allows us to confirm our theoretical findings with numerical experiments which are presented in the following sections.

## 4 Simulation

In the simulation study, we begin by demonstrating the effectiveness of the approximation of the ridge leverage score. We then compare the proposed methods with other subsampling approaches developed for large sample ridge regression or linear regression, including the ridge leverage score subsampling (Cohen et al., 2017), uniform subsampling for ridge regression, optimal subsampling for linear regression (Ma et al., 2020) and the information-based optimal subdata selection for linear regression (IBOSS) (Wang et al., 2019). Simulated data are generated in six settings. In each simulation, we generate the full data set of size  $n = 10^5$  and dimension  $p = 50$ . Subsample

sizes are set as  $r = 100, 200, 400, 800, 1600, 3200, 6400$ . The design matrix  $\mathbf{X}$  is standardized before being fed into the model. Each experiment is repeated 20 times. We use the mean squared error (MSE) of the estimated coefficient  $\tilde{\beta}$  to evaluate the performance.

The errors  $\epsilon_i, i = 1, \dots, n$  are independently and identically generated from  $N(0, 9)$ . We set the simulations as follows. For a  $q < p$ , let  $\Sigma$  be the  $q \times q$  covariance matrix with element  $\Sigma_{i,j} = 0.5^{1(i \neq j)}, i, j = 1, \dots, n$ . Consider  $q$ -dimensional  $\mathbf{x}_i \sim N(\mathbf{0}, \Sigma), i = 1, \dots, n$  for the true linear model with  $\beta = \mathbf{1}_{q \times 1}$ . An additional  $(p - q)$ -dimensional term  $\mathbf{x}^a$  is generated without being used in the true model, since we want to test if the subsample helps identify the appropriate relationship between the responses and the true covariates.

Case 1.  $q = 10$ ,  $\mathbf{X}_i^a$  follows a multivariate normal distribution, where columns of  $\mathbf{X}_i^a$  are i.i.d. samples from  $N(0, 1)$ .

Case 2.  $q = 10$ ,  $\mathbf{X}_i^a$  follows a multivariate lognormal distribution, where columns of  $\mathbf{X}_i^a$  are i.i.d. samples from  $LN(0, 1)$ .

Case 3.  $q = 10$ ,  $\mathbf{X}_i^a$  follows a multivariate  $t$  distribution with degrees of freedom 2, where columns of  $\mathbf{X}_i^a$  are i.i.d. samples from  $t_2(0, 1)$ .

Case 4.  $q = 25$ ,  $\mathbf{X}_i^a$  follows a multivariate normal distribution, where columns of  $\mathbf{X}_i^a$  are i.i.d. samples from  $N(0, 1)$ .

Case 5.  $q = 25$ ,  $\mathbf{X}_i^a$  follows a multivariate lognormal distribution, where columns of  $\mathbf{X}_i^a$  are i.i.d. samples from  $LN(0, 1)$ .

Case 6.  $q = 25$ ,  $\mathbf{X}_i^a$  follows a multivariate  $t$  distribution with degrees of freedom 2, where columns of  $\mathbf{X}_i^a$  are i.i.d. samples from  $t_2(0, 1)$ .

First, we show that subsampling by using the fast approximation of ridge leverage score (ROPT) is similar to that by applying the accurate one (ROPT-acc). Figure 1 displays the comparison result of MSE of the estimators by using the two sampling probabilities. Both methods have similar performance in all 6 cases given different subsample sizes. Therefore, the effectiveness of the  $\ell_2$  approximation of ridge leverage score is demonstrated.

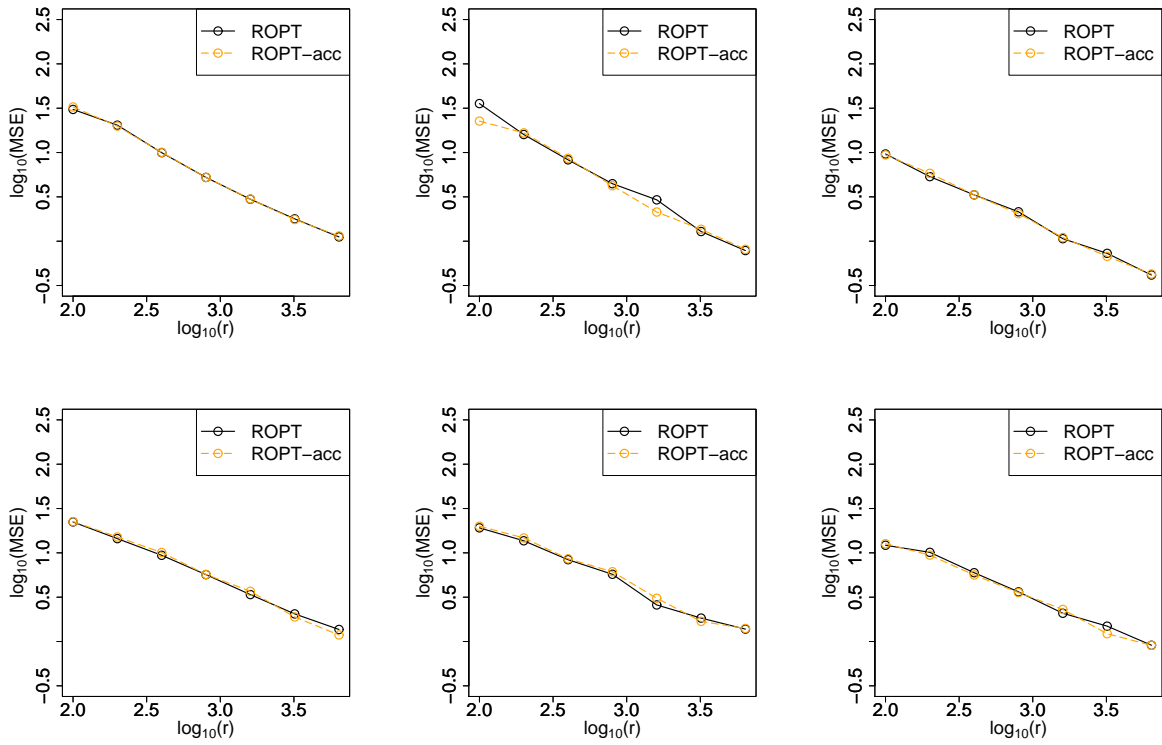


Figure 1: Comparison of different subsampling probabilities:  $x$ -axis is the logarithm of the subsample size,  $y$ -axis is the logarithm of the mean squared error of  $\tilde{\beta}$ .

Then we compare our proposed method with the ridge leverage score subsampling (RLEV), uniform subsampling for ridge regression (RUNIF), optimal subsampling for linear regression (OPT), and the information-based optimal subdata selection for linear regression (IBOSS). The first two competitors are raised for the ridge regression while the rest two are proposed for the linear regression. In Figure 2, our algorithm has the best performance among all the 6 cases when the subsample sizes are small or moderate, while all the methods have similar performance when the subsample sizes are large. First, our method has a great advantage when we use a small subsample. Second, by comparing the two rows of Figure 2, we can find our model preserves its superiority over other models for linear regression even in the cases where the true model favors less for introducing the ridge penalty.

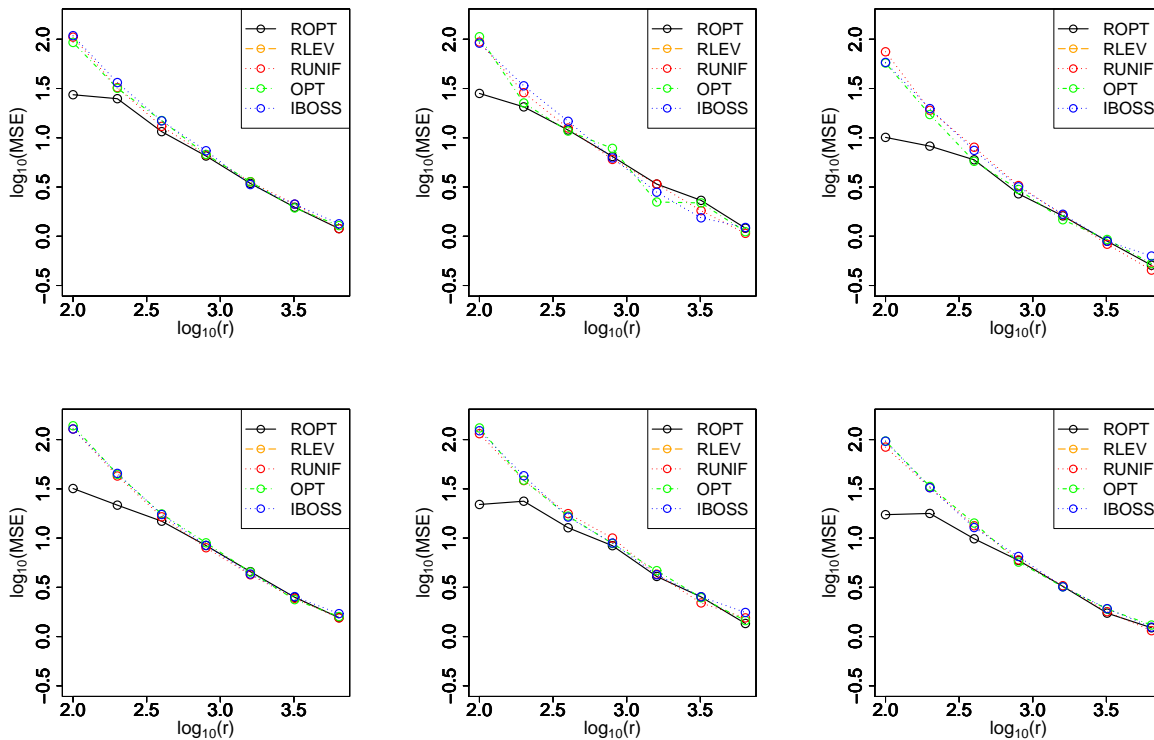


Figure 2: Comparison of different subsampling estimators:  $x$ -axis is the logarithm of the subsample size,  $y$ -axis is the logarithm of the mean squared error of the estimator  $\tilde{\beta}$  compared with true  $\beta$ .

## 5 Real data example

In times of information explosion, people are surrounded by a sea of news from various sources all day and night. For online media, it is critical for them to know what kind of news can attract the public attention, and hence the prediction of the popularity of the news becomes a trendy research topic. To raise the accuracy, numerical features from content, keywords, publish day and earlier popularity of news referenced in the article are extracted and then fed into a regression model to predict the share of the news. We use the open dataset of Online News Popularity Data Set on UC Irvine Machine Learning Repository<sup>1</sup>, which was provided by Fernandes et al. (2015).

The data was collected from Mashable, which is one of the largest news websites, from January 7, 2013, to January 7, 2015. It contains more than 39,000 articles in around 700 days. Except for the two non-predictable features, there is one response, the number of shares, and 58 predictive

<sup>1</sup><http://archive.ics.uci.edu/ml/datasets/Online+News+Popularity>

attributes concerning words, links, media, time, keywords, and natural language processing. Since the number of observations is huge and the number of features is also relatively large, it is prohibitive to allocate the memory for calculating the regression estimator. Therefore, we use the subsampling method to reduce the computation cost. The dataset is randomly divided into 70% for training and 30% for testing. Subsample sizes are set as  $r = 100, 200, 400, 800, 1600, 3200, 6400$ . The design matrix  $\mathbf{X}$  is standardized before being fed into the model. Each experiment is repeated 20 times. Because the true regression coefficient  $\beta$  is unknown, we first compare our estimator  $\tilde{\beta}$  with full-sample estimator  $\hat{\beta}$  in terms of MSE.

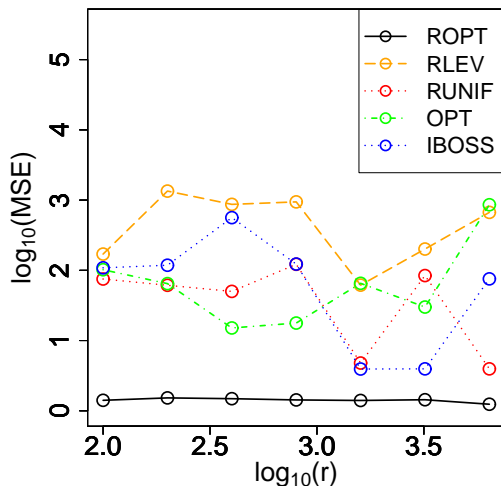


Figure 3: Comparison of different subsampling estimators:  $x$ -axis is the logarithm of the subsample size,  $y$ -axis is the logarithm of the mean squared error of the estimator  $\tilde{\beta}$  compared with full-sample estimator  $\hat{\beta}$ .

In Figure 3, we plot the MSE of the estimators calculated by different methods. Our method has the best performance at various subsample sizes compared with the competing methods. Finally, to better compare the performance of different methods, we report the test error under various subsample sizes in Table 1. Our method keeps the advantage compared with other methods as the subsample size grows.

r	100	200	400	800	1600	3200	6400
ROPT	<b>0.007</b>	<b>0.043</b>	<b>0.015</b>	<b>0.012</b>	<b>0.013</b>	<b>0.015</b>	<b>0.029</b>
RLEV	1.560	1.754	0.594	1.032	0.555	0.516	0.285
RUNIF	1.317	1.389	1.062	1.160	0.350	0.653	0.289
OPT	1.819	1.571	0.978	0.633	1.684	1.412	2.761
IBOSS	1.425	1.030	1.892	1.261	0.402	0.349	1.137

Table 1: The logarithm of the test error comparison under different subsample sizes.

## References

- AI, M., YU, J., ZHANG, H. and WANG, H. (2018). Optimal subsampling algorithms for big data generalized linear models. *arXiv preprint arXiv:1806.06761*.
- ALAOUI, A. and MAHONEY, M. W. (2015). Fast randomized kernel ridge regression with statistical guarantees. *Advances in Neural Information Processing Systems*, **28** 775–783.
- BATSON, J., SPIELMAN, D. A. and SRIVASTAVA, N. (2012). Twice-ramanujan sparsifiers. *SIAM Journal on Computing*, **41** 1704–1721.
- COHEN, M. B., MUSCO, C. and MUSCO, C. (2017). Input sparsity time low-rank approximation via ridge leverage score sampling. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 1758–1777.
- DEREZINSKI, M., KHANNA, R. and MAHONEY, M. W. (2020). Improved guarantees and a multiple-descent curve for column subset selection and the nystrom method. In *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan and H. Lin, eds.), vol. 33. Curran Associates, Inc., 4953–4964.
- DRINEAS, P., MAHONEY, M. W., MUTHUKRISHNAN, S. and SARLÓS, T. (2011). Faster least squares approximation. *Numerische Mathematik*, **117** 219–249.
- FERNANDES, K., VINAGRE, P. and CORTEZ, P. (2015). A proactive intelligent decision support

- system for predicting the popularity of online news. In *Portuguese Conference on Artificial Intelligence*. Springer, 535–546.
- GOLUB, G. H., HEATH, M. and WAHBA, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, **21** 215–223.
- HASTIE, T. (2020). Ridge regularization: An essential concept in data science. *Technometrics*, **62** 426–433.
- HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.
- HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12** 55–67.
- HOMRIGHAUSEN, D. and McDONALD, D. J. (2020). Compressed and penalized linear regression. *Journal of Computational and Graphical Statistics*, **29** 309–322.
- KACHAM, P. and WOODRUFF, D. (2020). Optimal deterministic coresets for ridge regression. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 4141–4150.
- KUMAR, S., MOHRI, M. and TALWALKAR, A. (2012). Sampling methods for the Nyström method. *The Journal of Machine Learning Research*, **13** 981–1006.
- MA, P., MAHONEY, M. W. and YU, B. (2015). A statistical perspective on algorithmic leveraging. *The Journal of Machine Learning Research*, **16** 861–911.
- MA, P., ZHANG, X., XING, X., MA, J. and MAHONEY, M. (2020). Asymptotic analysis of sampling estimators for randomized numerical linear algebra algorithms. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1026–1035.
- MAHONEY, M. W. (2011). Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, **3** 123–224.
- WANG, H. and MA, Y. (2020). Optimal subsampling for quantile regression in big data. *Biometrika*, **108** 99–112.



- WANG, H., YANG, M. and STUFKEN, J. (2019). Information-based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association*, **114** 393–405.
- WANG, H., ZHU, R. and MA, P. (2018). Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association*, **113** 829–844.
- WANG, S., GITTENS, A. and MAHONEY, M. W. (2017). Sketched ridge regression: Optimization perspective, statistical perspective, and model averaging. *The Journal of Machine Learning Research*, **18** 8039–8088.
- WOODRUFF, D. P. (2014). Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, **10** 1–157.