Sample Size Determination For Multidimensional Parameters And Optimal Sampling In A Big Data Linear Regression Model

A Preprint

Sheng Zhang
Department of Mathematical Sciences, IUPUI
402 N Blackford St., LD 270, Indianapolis, IN 46202, USA
shezhang@iupui.edu

Fei Tan
Department of Mathematical Sciences, IUPUI
402 N Blackford St., LD 270, Indianapolis, IN 46202, USA
feitan@iupui.edu

Hanxiang Peng
Department of Mathematical Sciences, IUPUI
402 N Blackford St., LD 270, Indianapolis, IN 46202, USA
hanxpeng@iupui.edu

March 28, 2023

Abstract

To fast approximate the least squares estimator efficiently in a Big Data linear regression by a subsampling estimator, numerous optimal sampling distributions are derived based on the criterion of minimizing the trace norm of the variance-covariance matrix of the subsampling estimator. Relative error bounds and conditions for subsample sizes to be bounded are provided. A scoring algorithm is constructed with far less running time than the full-sample LSE. An almost sure asymptotic normality result is proved for the subsampling estimator for an arbitrary sampling distribution. Motivated by subsampling and data-splitting in machine learning, sample size determination for multidimensional parameters is presented. The numerical performance of the results is studied through large simulated and real data.

Keywords Asymptotic normality; Least squares estimator; Big data; Optimal sampling; Sample size determination

1 Introduction

In a linear regression model, the response y_i and covariate vector $\mathbf{x}_i$ satisfy

$$y_i = \boldsymbol{\beta}^\top \mathbf{x}_i + \varepsilon_i, \quad i = 1, \dots, n, \tag{1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown parameter and $\varepsilon_1, \dots, \varepsilon_n$ are independent and identically (i.i.d.) random errors with zero mean and finite positive variance $\sigma^2 = \mathrm{Var}(\varepsilon_i)$. Assume that $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top$ is a nonrandom $n \times p$ matrix of full rank p.

The parameter vector $\boldsymbol{\beta}$ can be estimated by the ordinary least squares estimator (LSE) $\hat{\boldsymbol{\beta}}_{\mathrm{ols}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$, where $\mathbf{y} = (y_1, \dots, y_n)^\top$. Consider the case of data of massive size in which $\hat{\boldsymbol{\beta}}_{\mathrm{ols}}$ is not available. One may draw a subsample $(\mathbf{X}^*, \mathbf{y}^*)$ of small size $r << n$ using a sampling distribution $\boldsymbol{\pi}_n = (\pi_1, \ldots, \pi_n)$ as a surrogate for the full sample, and calculate the subsampling weighted LSE $\hat{\boldsymbol{\beta}}_r^*$ to approximate $\hat{\boldsymbol{\beta}}_{\text{ols}}$,

$$\hat{\boldsymbol{\beta}}_r^* = (\mathbf{X}^{*\top}\mathbf{W}^*\mathbf{X}^*)^{-1}\mathbf{X}^{*\top}\mathbf{W}^*\mathbf{y}^*. \tag{2}$$

where $\mathbf{W}^* = \text{diag}(1/r\boldsymbol{\pi}^*)$ is the diagonal matrix with $\boldsymbol{\pi}^*$ equal to the vector of the corresponding sampling probabilities. Here we adopt the componentwise division $\mathbf{a}/\mathbf{b} = (a_1/b_1, \ldots, a_n/b_n)^\top$ for vectors $\mathbf{a}, \mathbf{b}$. This is a Hansen-Hurwitz estimator and could also be viewed as a weighted bootstrap estimator based on a subsample. Full sample weighted bootstrap estimators were well studied in the literature, see the monograph by Barbe and Bertail (1995)[2].

Over the past two decades, there have been considerable progresses on subsampling, see Liang, *et al.* (2013)[11], Kleiner, *et al.* (2014)[9], Wang, *et al.* (2015)[20], Wang, *et al.* (2019)[19] among others. Algorithms for fast computing the LSE were constructed, see the monograph by Mahoney (2011)[14] and the references therein. A key feature of these results is the nonuniform sampling. While these results were mainly focused on the algorithmic properties, we shall be concerned with statistical inference. Zhu, *et al.* (2015)[22] pioneered in this aspect and their work is influential in our work. They obtained several A-optimal distributions and proved asymptotic normality in probability. We give the A-optimal distributions for approximating a smooth function $\mathbf{g}(\hat{\boldsymbol{\beta}}_{\text{ols}})$ of $\hat{\boldsymbol{\beta}}_{\text{ols}}$ (the choice of $\mathbf{g}(\hat{\boldsymbol{\beta}}_{\text{ols}}) = \mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{ols}} = \mathbf{X}^\top\mathbf{y}$ yields their results), and prove an almost sure asymptotic normality result. We present the relative error bounds for $\hat{\boldsymbol{\beta}}_r^*$ in Section 4. Such bounds, as pointed out in Mahoney (p.17, 2011)[14], are gold standard and provide much stronger notion of approximation than additive bounds.

In textbooks, sample sizes are generally bounded for given margin of error (MOE) and confidence level. We acknowledge that sample sizes may be unbounded. One might wonder that under what conditions subsample sizes are bounded uniformly in n for given MOE and confidence level. The result presented here for subsampling in a linear regession model is that the leverage scores $h_{i,i}$ of the hat matrix must stay away from its boundary 0 and 1, the covariate matrix $\mathbf{X}$ must be well-conditioned, and truncation from below of the sampling distribution is required. As a consquence, for the uniform sampling (bootstrapping) the boundedness requires that the covariate matrix $\mathbf{X}$ must be well-conditioned.

It is obvious that a suitable subsample size is key for obtaining a desired result within a desired peroid of time. Sample size determination (SSD) for scalar parameters is a melody. In this article, we extend SSD to multidimensional parameters and study the numerical behavior through simulations. The result may also be useful for data spliting in machine learning.

The statistical leverage scores based distribution $\boldsymbol{\ell}$ has played a central role in the development of randomized matrix algorithms, see e e.g. Candés and Tao (2009)[3]; Drineas *et al.* (2012)[7]; Ma and Sun (2014)[12]; Ma, *et al.* (2015)[13]; Xu, *et al.* (2016)[21]. Interestingly, $\boldsymbol{\ell}$ and the A-optimal distribution $\hat{\boldsymbol{\pi}}_2$ draw data points in a totally opposite way. Specifically, the former draws points close to the regression hyperplane, whereas the latter does away from the hyperplane.

While classic methods compute the LSE $\hat{\boldsymbol{\beta}}_{\text{ols}}$ in $O(np^2)$ time, randomized methods usually take $o(np^2)$ time. Typically, the bottleneck is to compute the appropriate sampling distributions, and the A-optimal distributions fall in with this category. As the LSE $\hat{\boldsymbol{\beta}}_{\text{ols}}$ and $\boldsymbol{\ell}$ are fundamental and ubiquitous, there have been developed randomized algorithms on rapidly approximating them, see e.g. Drineas, *et al.* (2006)[6]. These algorithms can be utilized to fast compute the optimal distributions. In the spirit of the scoring method for improving estimation efficiency, we construct the *Scoring Algorithm* in Fig. 2 with running time $O(rp^2)$ where $r << n$. Our extensive simulations indicated that the algorithm worked particularly well.

The article is organized as follows. In Section 2, we define SSD for multidimensional parameters and proivde the formula. In Section 3, we prove an asymptotic normality result, give the A-optimal distributiogns, construct the Scoring Algorithm, and discuss trunction and the raltationship between the leverage scores based distribution and the A-optimal distributions. The relative error bounds and the boundedness conditions for subsample sizes are offered in Section 4. Some simulations are reported in Section 5. The proofs are collected in Sections 6–7.

2 SSD for multidimensional parameters

Let P be a probability measure on some measurable space. Let m the Borel measure on $\mathbb{R}^p$. Typically, m is the volume measure on $\mathbb{R}^p$. Consider a parameter vector $\boldsymbol{\theta} \in \mathbb{R}^p$, and a random region $\mathbb{R}_n$ on $\mathbb{R}^p$. Given $\epsilon > 0$ and $\alpha \in (0, 1)$, we seek a minimum sample size n such that at the level $1 - \alpha$ of confidence, $\mathbb{R}_n$ catches $\boldsymbol{\theta}$ within the "range of error" (ROE) ϵ, that is, $m(\mathbb{R}_n) \leq \epsilon$. We now introduce the definition. Let $\boldsymbol{\theta}_0$ denote the true value of parameter.

Definition 1. *Given $\epsilon > 0$ and $\alpha \in (0,\, 1)$, the sample size with the range of error (ROE) $\epsilon > 0$ at the level $1-\alpha$ of confidence is defined as*

$$n(\epsilon, \alpha) = \min\left\{n :\ P(\boldsymbol{\theta}_0 \in \mathbb{R}_n,\, m(\mathbb{R}_n) \leq \epsilon^p) \geq 1-\alpha\right\}.$$

We give two examples below, using the following two-step method.

Step 1 Construct a $(1-\alpha)$- level confidence region $\mathbb{R}_n$ for $\boldsymbol{\theta}$.

Step 2 Find the minimum sample size n such that the ROE is ϵ, that is, $m(\mathbb{R}_n) \leq \epsilon^p$.

Example 1. (Ellipsoid) Let $\hat{\boldsymbol{\theta}}$ be an estimator of a parameter $\boldsymbol{\theta}_0 \in \Theta \subset \mathbb{R}^p$ such that the variance-covariance matrix Σ_n of $\hat{\boldsymbol{\theta}}$ satisfies that $n\Sigma_n$ converges in probability to some positive definite matrix Σ. Let m be the volume measure on $\mathbb{R}^p$, and $\mathbb{R}_n$ be the $(1-\alpha)$- level confidence ellipsoid centered at $\hat{\boldsymbol{\theta}}_n$,

$$\mathbb{R}_n = \left\{\boldsymbol{\theta} \in \Theta :\ T_n(\boldsymbol{\theta}) \leq q_\alpha(p)\right\},$$

where $T_n(\boldsymbol{\theta}) = n(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \Sigma^{-1}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$, and $q_\alpha(p)$ denotes the upper α-percentile of the distribution of $T(\boldsymbol{\theta}_0)$, that is, $P(T(\boldsymbol{\theta}_0) > q_\alpha(p)) = \alpha$. By definition, the sample size is determined by

$$n_p(\epsilon, \alpha) = \min\left\{n :\ P(\boldsymbol{\theta}_0 \in \mathbb{R}_n,\, m(\mathbb{R}_n) \leq \epsilon^p) \geq 1-\alpha\right\}.$$

The volume of the ellipsoid is

$$m(\mathbb{R}_n) = \frac{n^{-p/2}\pi^{p/2}}{\Gamma(p/2+1)} q_\alpha^{p/2}(p) \prod_{d=1}^{p} \sqrt{\lambda_d},$$

where $\lambda_d, d = 1, \ldots, p$ are the eigenvalues of Σ. Solving $m(CR_n) \leq \epsilon$ about n yields the sample size $n_p(\epsilon, \alpha)$ with ROE ϵ at the level of $1-\alpha$, given by

$$n \geq n_p(\epsilon, \alpha) = \frac{\pi q_\alpha(p)}{\Gamma^{2/p}(p/2+1)} \frac{\sqrt[p]{\det(\Sigma)}}{\epsilon^2}, \tag{3}$$

where $\det(\Sigma) = \prod_{d=1}^{p} \lambda_d$ is the determinant of Σ. Often Σ is unknown, one uses an estimator $\hat{\Sigma}$ of Σ (or estimators $\hat{\lambda}_d$ of the eigenvalues λ_d). For $p = 1$, as $\Gamma(3/2) = \sqrt{\pi}/2$, the sample size with ROE 2ϵ (margin of error (MOE) ϵ) at the level $1-\alpha$ boils down to the formula $n_1(\epsilon, \alpha) = q_\alpha(1)\sigma^2/\epsilon^2$ found in textbooks.

Example 2. (Bonferroni) Consider the same problem as in Example 1, but now based on Bonferroni's method. We take $\mathbb{R}_n$ to be the p-dimensional $(1-\alpha)$-confidence hyperrectangle,

$$\mathbb{R}_n = \prod_{d=1}^{p} (\hat{\boldsymbol{\theta}}_{d,n} - q_{\alpha/p}\sigma_d/\sqrt{n}, \quad \hat{\boldsymbol{\theta}}_{d,n} + q_{\alpha/p}\sigma_d/\sqrt{n}),$$

where q_α denotes the upper α-percentile of the distribution of $\Sigma^{-1/2}\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\beta}_0)$, and $\hat{\boldsymbol{\theta}}_{d,n}$ and σ_d^2 denote the d-th component of $\hat{\boldsymbol{\theta}}_n$ and the d-th diagonal entry of Σ, respectively. As the volume of the hyperrectangle $\mathbb{R}_n$ is

$$m(\mathbb{R}_n) = 2^p n^{-p/2} q_{\alpha/p}^p \sigma_1 \cdots \sigma_p,$$

solving $m(\mathbb{R}_n) \leq \epsilon$ about n yields the sample size,

$$n \geq n_p^{bon}(\epsilon, \alpha) = 4q_{\alpha/p}^2 \sigma_1^{2/p} \cdots \sigma_p^{2/p}/\epsilon^2. \tag{4}$$

For unknown parameters σ_d's, one uses estimators $\hat{\sigma}_d$'s of them.

Remark 1. *If $T_n(\boldsymbol{\theta}_0)$ is chisquare distributed with p degrees of freedom (often approximately), one then takes $q_\alpha(p) = \chi^2_\alpha(p)$, the upper α-percentile of the chisquare distribution $\chi^2_\alpha(p)$ with d degrees of freedom. Similarly for Bonferroni, $q_\alpha = Z_\alpha$, the upper α-percentile of the standard normal $\mathscr{N}(0,1)$. Alternatively, one can get an estimate of $q_\alpha(p)$ by bootstrapping or pre-subsampling in the Scoring Algorithm 2 in the case of Big Data.*

Remark 2. *In nonuniform subsampling for data of massive size, a sampling distribution $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$ must be computed before actually subsampling. An optimal sampling distribution π typically has the same computational complexity as the original problem. To tackle this problem, one may take a uniform pre-subsample of small size and compute an approximation $\tilde{\boldsymbol{\pi}}_0$ to $\boldsymbol{\pi}$ as described in the Scoring Algorithm 2, choosing suitable values of ϵ, α and $q_\alpha(p) = \chi^2_\alpha(p)$. To determine the pre-subsample size, one may take Σ to be the identity matrix in (3) to get a pre-subsample size,*

$$n \geq n_{p,0}(\epsilon, \alpha) = \frac{\pi q_\alpha(p)}{\Gamma^{2/p}(p/2+1)} \frac{1}{\epsilon^2}. \tag{5}$$

The formula can be used for SSD in the uniform sampling (bootstrapping) and data splitting in machine learning.

Figure 1: Algorithm 1 (Computing the subsampling estimator $\hat{\boldsymbol{\beta}}_r^*$)

1. Construct a distribution $\boldsymbol{\pi}$ on the data points $(\mathbf{x}_i, y_i)$'s, use it to draw a subsample $(\mathbf{X}^*, \mathbf{y}^*)$ of size $r << n$ and formulate the diagonal matrix $\mathbf{W}^* = \text{diag}(1/r\boldsymbol{\pi}^*)$ with $\boldsymbol{\pi}^*$ the corresponding probability vector.
2. Calculate the weighted least squares estimator $\hat{\boldsymbol{\beta}}_r^* = (\mathbf{X}^{*\top}\mathbf{W}^*\mathbf{X}^*)^{-1}\mathbf{X}^{*\top}\mathbf{W}^*\mathbf{y}^*$.

3 Almost sure ASN and the A-optimal distributions

In this section, we prove ASN, derive the optimal distributions, construct the Scoring Algorithm and discuss truncation and the relationship of the leverage scores based and the A-optimal distributions.

3.1 Asymptotic normality

We give a set of conditions on $\boldsymbol{\pi}$ below for the almost sure asymptotic normality of $\hat{\boldsymbol{\beta}}_r^*$ for an arbitrary sampling distribution. Occasionally, we write $\boldsymbol{\pi} = \boldsymbol{\pi}_n$ and $\pi_i = \pi_{n,i}$ to stress their dependene on the sample size n.

(M1)

$$\frac{1}{n^2}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^\top(\varepsilon_i^2 - \sigma^2)}{\pi_{n,i}} = O(1), \quad a.s.$$

(M2) There is a $p \times p$ symmetric matrix Γ whose smallest eigenvalue is bounded away from zero, i.e., $\lambda_{\min}(\Gamma) \geq b_0 > 0$ for some constant b_0, such that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top = \Gamma + o(1).$$

(M3)

$$\frac{1}{n^2}\sum_{i=1}^{n}\frac{\|\mathbf{x}_i\|^4}{\pi_{n,i}} = O(1) \quad a.s.$$

(M4) $\mathbb{L}_n(\boldsymbol{\pi}) =: \sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top/n^2\pi_{n,i}$ satisfies $0 < b \leq \lambda_{\min}\mathbb{L}_n(\boldsymbol{\pi}) \leq \lambda_{\max}\mathbb{L}_n(\boldsymbol{\pi}) \leq B < \infty$ a.s. for constants b, B, where $\lambda_{\min}$ and $\lambda_{\max}$ denote the maximum and minimum eigenvalues, respectively.

(M5) Lindeberg condition: the double array $\boldsymbol{\eta}_{n,i} := \mathbf{x}_i\varepsilon_i/n\pi_{n,i}$, $i = 1, 2, \ldots, n,\ n \geq 1$ satisfies that for any $t > 0$,

$$\sum_{i=1}^{n}\pi_{n,i}\|\boldsymbol{\eta}_{n,i}\|^2\mathbf{1}[\|\boldsymbol{\eta}_{n,i}\| \geq \sqrt{r}t] = o(1), \quad a.s. \quad r \to \infty.$$

(D1) Condition (M1) can be verified using the result on the SLLN for weighted i.i.d. rv's of Baxter, *et al.* (2004)[1]. Specifically, for a sequence $\{a_i\}$, $\frac{1}{n}\sum_{j=1}^{n}|a_i|^q = O(1)$ for some $q > 1$ implies $\frac{1}{n}\sum_{j=1}^{n}a_i\xi_i \to 0$ a.s. for an i.i.d. $\{\xi_n\}$ with $\mathrm{E}(\xi_1) = 0$ and $E(|\xi_1|) < \infty$.

(D2) Condition (M2) was used in Lemma 3.1 of Portnoy (1984)[15].

Theorem 1. *Assume (M1)–(M5). Suppose that for every $\varrho > 0$,*

$$\max_{1\leq i\leq n}\|\mathbf{x}_i\| = o(n^{1/2}\log^{-\varrho}(n)), \quad a.s. \tag{6}$$

Suppose that there exists some $\rho > 2$ such that

$$E(|\varepsilon_1|^\rho) < \infty. \tag{7}$$

Then $\hat{\boldsymbol{\beta}}_r^*$ is asymptotically normal along almost all the sample paths of the sequence $\{(\mathbf{x}_i, y_i)\}$ as $r \to \infty$, i.e.,

$$\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\pi})\sqrt{r}(\hat{\boldsymbol{\beta}}_r^* - \hat{\boldsymbol{\beta}}_{\text{ols}}) \implies \mathcal{N}(0, \mathbf{I}_p), \quad a.s. \quad r \to \infty, \tag{8}$$

where $\boldsymbol{\Sigma}(\boldsymbol{\pi}) = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\text{Diag}(\hat{\boldsymbol{\varepsilon}}^2/\boldsymbol{\pi})\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}$.

3.2 The A-optimal distributions

For a $q \times p$ matrix $\mathbf{A}$, we minimize the trace norm $\mathrm{Tr}(\mathbf{\Sigma_A})$ over distributions supported on the data points, where

$$\text{(9)} \qquad \mathbf{\Sigma_A}(\boldsymbol{\pi}) = \mathbf{A}\Sigma(\boldsymbol{\pi})\mathbf{A}^\top = \mathbf{A}(\mathbf{X}^\top\mathbf{X})^{-1}\Sigma_c(\boldsymbol{\pi})(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{A}^\top,$$

with $\Sigma_c(\boldsymbol{\pi}) = \mathbf{X}^\top\mathrm{Diag}(\hat{\varepsilon}^2/r\boldsymbol{\pi})\mathbf{X}$. Let $\hat{\boldsymbol{\theta}} = \mathbf{A}\hat{\boldsymbol{\beta}}_{\text{ols}}$. The plug-in estimate $\hat{\boldsymbol{\theta}}^* = \mathbf{A}\hat{\boldsymbol{\beta}}_r^*$ of $\hat{\boldsymbol{\theta}}$ has $\mathrm{Var}^*(\hat{\boldsymbol{\theta}}^*) = \mathbf{\Sigma_A}(\boldsymbol{\pi})$. Consider $\boldsymbol{\theta} = \mathbf{g}(\boldsymbol{\beta})$, where $\mathbf{g}$ has the continuous partial derivative $\dot{\mathbf{g}}$. Then $\hat{\boldsymbol{\theta}}^* = \mathbf{g}(\hat{\boldsymbol{\beta}}_r^*)$ is a subsampling estimator to approximate $\hat{\boldsymbol{\theta}} = \mathbf{g}(\hat{\boldsymbol{\beta}}_{\text{ols}})$, and an A-optimal distribution for $\hat{\boldsymbol{\theta}}^*$ to approximate $\hat{\boldsymbol{\theta}}$ is given by taking $\mathbf{A} = \dot{\mathbf{g}}(\bar{\boldsymbol{\beta}})$ for some pilot estimator $\bar{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$.

An A-optimal distribution depends on data, parameters, and the estimation method. With these in mind and for convenience, we introduce the following definition.

Definition 2. *Given a σ-field $\mathcal{F}$, a distribution $\boldsymbol{\pi}$ supported on the data points is said to be A-optimal for the subsampling estimate $\hat{\boldsymbol{\theta}}^*$ to approximate an estimate $\hat{\boldsymbol{\theta}}$ of parameter $\boldsymbol{\theta}$ if $\boldsymbol{\pi}$ asymptotically minimizes the trace norm of the conditional variance-covariance matrix $\mathrm{Var}(\hat{\boldsymbol{\theta}}^*|\mathcal{F})$ of $\hat{\boldsymbol{\theta}}^*$ given $\mathcal{F}$.*

If $\mathcal{F}$ is the σ-field generated by $\{(\mathbf{x}_i, y_i)\}$ $(\{\mathbf{x}_i\})$, then $\boldsymbol{\pi}$ is referrred to as $\hat{A}$ $(\bar{A})$-optimal. Note that the plug-in estimtor $\mathbf{g}(\hat{\boldsymbol{\theta}}^*)$ is not A-optimal for it to approximate $\mathbf{g}(\hat{\boldsymbol{\theta}})$.

The $\hat{A}$-optimal distributions $\hat{\boldsymbol{\pi}}_2$. Minimizing the trace norm of the variance-covariance matrix $\mathbf{\Sigma_A}$ in (9), we obtain the $\hat{A}$-optimalizer $\hat{\boldsymbol{\pi}}_\mathbf{A}$. Let $\hat{\mathbf{H}}_{2,\mathbf{A}} = \mathrm{Diag}(\hat{\varepsilon})\mathbf{H}_{2,\mathbf{A}}\mathrm{Diag}(\hat{\varepsilon})$, where

$$\text{(10)} \qquad \mathbf{H}_{2,\mathbf{A}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{A}^\top\mathbf{A}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top.$$

We now invoke the Lagrange multipliers to get

Proposition 1. *Let $\mathbf{A}$ be a $q \times p$ matrix which is independent of $\boldsymbol{\pi}$. Assume that $\mathbf{A}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i \neq 0$ and $h_{i,i} \neq 1$ for all i. Then the square roots of the diagonal entries of $\hat{\mathbf{H}}_{2,\mathbf{A}}$ induce the unique $\hat{A}$-optimal distribution $\hat{\boldsymbol{\pi}}_\mathbf{A}$ for $\mathbf{A}\hat{\boldsymbol{\beta}}_r^*$ to approximate $\mathbf{A}\hat{\boldsymbol{\beta}}_{\text{ols}}$.*

We shall refer to $\hat{\mathbf{H}}_{2,\mathbf{A}}$ as the $\hat{A}$*-optimal score matrix.* Write $p_i \propto b_i$ if $p_i = b_i / \sum_j b_j$ for $\forall i$. Then $\hat{\boldsymbol{\pi}}_\mathbf{A}$ is given by

$$\text{(11)} \qquad \hat{\pi}_{\mathbf{A},i} \propto \|\mathbf{A}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i\|\,|\hat{\varepsilon}_i|.$$

For $\mathbf{A} = (\mathbf{X}^\top\mathbf{X})^{1-\alpha/2}$, set $\mathbf{H}_\alpha = \mathbf{H}_{2,\mathbf{A}}$ and $\hat{\mathbf{H}}_\alpha = \hat{\mathbf{H}}_{2,\mathbf{A}}$, so that

$$\mathbf{H}_\alpha = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\alpha}\mathbf{X}^\top, \quad \hat{\mathbf{H}}_\alpha = \mathrm{Diag}(\hat{\varepsilon})\mathbf{H}_\alpha\mathrm{Diag}(\hat{\varepsilon}), \quad \alpha \in \mathbb{R}.$$

It then follows $\hat{\mathbf{H}}_\alpha$ is the $\hat{A}$-optimal score matrix for $\hat{\boldsymbol{\theta}}_\alpha^* = (\mathbf{X}^\top\mathbf{X})^{1-\alpha/2}\hat{\boldsymbol{\beta}}_r^*$ to approximate $\hat{\boldsymbol{\theta}}_\alpha = (\mathbf{X}^\top\mathbf{X})^{1-\alpha/2}\hat{\boldsymbol{\beta}}_{\text{ols}} = (\mathbf{X}^\top\mathbf{X})^{-\alpha/2}\mathbf{X}^\top\mathbf{y}$, with the unique $\hat{A}$-optimal distribution $\hat{\boldsymbol{\pi}}_\alpha$ given by

$$\hat{\pi}_{\alpha,i} \propto \sqrt{h_{\alpha,i,i}}|\hat{\varepsilon}_i|, \quad \text{where} \quad h_{\alpha,i,i} = \mathbf{x}_i^\top(\mathbf{X}^\top\mathbf{X})^{-\alpha}\mathbf{x}_i.$$

Consequently, $\hat{\boldsymbol{\pi}}_2$ is the unique $\hat{A}$-optimal distribution for $\hat{\boldsymbol{\beta}}_r^*$ to approximate $\hat{\boldsymbol{\beta}}_{\text{ols}}$.

Remark 3. *While $\hat{\pi}_{0,i} \propto \|\mathbf{x}_i\||\hat{\varepsilon}_i|$ has less computational cost than $\hat{\boldsymbol{\pi}}_\alpha$ $(\alpha \neq 0)$ (as only $\|\mathbf{x}_i\|$ and $|\hat{\varepsilon}_i|$ must be computed), $\hat{\pi}_{1,i} \propto \sqrt{h_{i,i}}|\hat{\varepsilon}_i|$ can be computed using the fast algorithm given in Drineas,* et al. *(2006)[6].*

The $\bar{A}$-optimal $\bar{\boldsymbol{\pi}}_2$ and its approximation $\tilde{\boldsymbol{\pi}}_2$. Consider minimizing the trace norm of the conditional variance-covariance matrix given $\mathbf{X}$. Since $\hat{\tau}_\mathbf{A}(\boldsymbol{\pi}) = \mathrm{Tr}(\mathbf{\Sigma_A}(\boldsymbol{\pi})) = r^{-1}\sum_{i=1}^n \|\mathbf{a}_i\|^2\hat{\varepsilon}_i^2/\pi_i$ and $\mathrm{Var}(\hat{\boldsymbol{\varepsilon}}|\mathbf{X}) = (\mathbf{I}_n - \mathbf{H})\sigma^2$, we integrate out the squared residuals in the trace $\hat{\tau}_\mathbf{A}(\boldsymbol{\pi})$ to get

$$\text{(12)} \qquad \bar{\tau}_\mathbf{A}(\boldsymbol{\pi}) = \mathrm{E}(\tau_\mathbf{A}(\boldsymbol{\pi})|\mathbf{X}) = \frac{\sigma^2}{r}\sum_{i=1}^n \frac{\|\mathbf{a}_i\|^2(1-h_{i,i})}{\pi_i}, \quad \mathbf{a}_i = \mathbf{A}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i.$$

Suppose that $h_{i,i}$'s satisfy $\max_{i=1,\ldots,n} h_{i,i} = o(1)$. One then obtains an approximation to the trace as follows:

$$\tilde{\tau}_\mathbf{A}(\boldsymbol{\pi}) = \frac{\sigma^2}{r}\sum_{i=1}^n \frac{\|\mathbf{a}_i\|^2}{\pi_i}.$$

Figure 2: The Scoring Algorithm

1. Take a uniform pre-subsample $(\mathbf{X}_0^*, \mathbf{y}_0^*)$ of size r_0 from $(\mathbf{X}, \mathbf{y})$, and use it to compute $\mathbf{H}_{0,\alpha}^*$ ($\bar{\mathbf{H}}_{0,\alpha}^*$ or $\hat{\mathbf{H}}_{0,\alpha}^*$) given in (14).
2. Call Algorithm 1 in Fig. 1 with the subsample size r and the A-optimal distribution $\boldsymbol{\pi}$.

Minimizing $\bar{\tau}_{\mathbf{A}}(\boldsymbol{\pi})$ and $\tilde{\tau}_{\mathbf{A}}(\boldsymbol{\pi})$ yields the sampling distributions $\bar{\boldsymbol{\pi}}_{\mathbf{A}}$ and $\tilde{\boldsymbol{\pi}}_{\mathbf{A}}$, respectively. Note the conditional version of $\hat{\mathbf{H}}_{2,\mathbf{A}}$ in (10) takes the form,

$$\bar{\mathbf{H}}_{2,\mathbf{A}} = \mathrm{Diag}((1 - h_{i,i})^{1/2})\mathbf{H}_{2,\mathbf{A}}\mathrm{Diag}((1 - h_{i,i})^{1/2}).$$

Thus $\bar{\boldsymbol{\pi}}_{\mathbf{A}}$ can be expressed as $\bar{\pi}_{\mathbf{A},i} \propto \|\mathbf{a}_i\|\sqrt{1 - h_{i,i}}$. For $\mathbf{A} = (\mathbf{X}^\top \mathbf{X})^{1-\alpha/2}$, let $\bar{\mathbf{H}}_\alpha = \bar{\mathbf{H}}_{2,\mathbf{A}}$. The $\bar{A}$-optimal $\bar{\boldsymbol{\pi}}_\alpha$ can be written as

$$\bar{\pi}_{\alpha,i} \propto \sqrt{h_{\alpha,i,i}}\sqrt{1 - h_{i,i}}.$$

Hence $\bar{\boldsymbol{\pi}}_2$ is the unique $\bar{A}$-optimal distribution for $\hat{\boldsymbol{\beta}}_r^*$ to approximate $\hat{\boldsymbol{\beta}}_{\mathrm{ols}}$. Likewise, $\tilde{\boldsymbol{\pi}}_\alpha$ is given by

$$\tilde{\pi}_{\alpha,i} \propto \sqrt{h_{\alpha,i,i}}. \tag{13}$$

Remark 4. *As in Remark 3, while $\bar{\boldsymbol{\pi}}_1, \tilde{\boldsymbol{\pi}}_1$ can be fast computed, $\bar{\boldsymbol{\pi}}_0, \tilde{\boldsymbol{\pi}}_0$ enjoy computational ease. The latter are, respectively, the optimal sampling (OPT) and predictor-length (PL) sampling given in Zhu,* et al. *(2015).*

Comparison and truncation. Since $\hat{\boldsymbol{\pi}}_{\mathbf{A}}$ minimizes $\hat{\tau}_{\mathbf{A}}(\boldsymbol{\pi})$, it follows from Proposition 1 that $\hat{\tau}_{\mathbf{A}}(\hat{\boldsymbol{\pi}}_{\mathbf{A}}) \leq \hat{\tau}_{\mathbf{A}}(\bar{\boldsymbol{\pi}}_{\mathbf{A}})$. Hence, by (12), we obtain

$$\mathrm{E}(\hat{\tau}_{\mathbf{A}}(\hat{\boldsymbol{\pi}}_{\mathbf{A}})) \leq \mathrm{E}(\hat{\tau}_{\mathbf{A}}(\bar{\boldsymbol{\pi}}_{\mathbf{A}})) = \bar{\tau}_{\mathbf{A}}(\bar{\boldsymbol{\pi}}_{\mathbf{A}}).$$

This shows that $\hat{\boldsymbol{\pi}}_{\mathbf{A}}$ is, on average, better A-optimizing than $\bar{\boldsymbol{\pi}}_{\mathbf{A}}$. Our extensive simulations and real data applications exhibited that $\hat{\boldsymbol{\pi}}_{\mathbf{A}}$ was much better optimizing than both $\bar{\boldsymbol{\pi}}_{\mathbf{A}}$ and $\tilde{\boldsymbol{\pi}}_{\mathbf{A}}$.

TRUNCATION. Observe that (11) implies that $(\mathbf{x}_i, y_i)$ must be drawn with probability $\hat{\pi}_{\mathbf{A},i}$ proportional to $|\hat{\varepsilon}_i|$. Since each probability is inversely used in constructing $\hat{\boldsymbol{\beta}}_r^*$, $\hat{\pi}_{\mathbf{A}}$ must be truncated from below in order to guarantee appropriate statistical properties for $\hat{\boldsymbol{\beta}}_r^*$. Truncation was used in constructing the generalized bootstrap estimator by Chatterjee and Bose (2002)[4]. Specifically, we truncate $\hat{\boldsymbol{\pi}}_{\mathbf{A}}$ from below by L/n and define $\hat{\boldsymbol{\pi}}_{\mathbf{A}}(l)$ by

$$\hat{\pi}_{\mathbf{A},i}(l) \propto \hat{\pi}_{\mathbf{A},i}\mathbf{1}[\hat{\pi}_{\mathbf{A},i} \geq L/n] + (l/n)\mathbf{1}[\hat{\pi}_{\mathbf{A},i} < L/n], \quad i = 1, 2, \ldots, n,$$

where L is a threshold value. Typically $0 < L \leq 1$. This is, in fact, a mixture distribution of the optimal and the uniform distributions. For fast computing, we may drop "unimportant" observations by taking $l = 0$, otherwise $l = L$. See p. 18 (Tropp, 2019)[17] for further discussion. As $\bar{\pi}_{\mathbf{A},i} = 0$ at $h_{i,i} = 1$, we truncate $\bar{\pi}_{\mathbf{A},i}$ similarly from below by $\bar{\pi}_{\mathbf{A},i}(l)$. Although $\tilde{\boldsymbol{\pi}}_{\mathbf{A}}$ is positive, we also truncate it and define the likewise $\tilde{\boldsymbol{\pi}}_{\mathbf{A}}(l)$.

To determine the value of L, we must take it into consideration the desired running time and the accuracy. Our extensive numerical results exhibited that even high percentages of truncation led to only slight loss of efficiency.

The Scoring Algorithm. Like a typical optimal sampling, the A-optimal sampling $\hat{\boldsymbol{\pi}}_2$, $\bar{\boldsymbol{\pi}}_2$ and $\tilde{\boldsymbol{\pi}}_2$ have the same running time $O(np^2)$ as the full data LSE $\hat{\boldsymbol{\beta}}_{\mathrm{ols}}$. We provide a fast algorithm in Fig. 2.

Since the computational bottleneck is to invert $\mathbf{X}^\top\mathbf{X}$, we shall approximate it by the subsampling $(\mathbf{X}_0^{*\top}\mathbf{X}_0^*)^{-1}$ based on a computationally easy pre-subsample $(\mathbf{X}_0^*, \mathbf{y}_0^*)$ from the data $(\mathbf{X}, \mathbf{y})$. Let the resulting estimator and residuals be

$$\hat{\boldsymbol{\beta}}_0^* = (\mathbf{X}_0^{*\top}\mathbf{X}_0^*)^{-1}\mathbf{X}_0^{*\top}\mathbf{y}_0^*, \quad \hat{\boldsymbol{\varepsilon}}_0^* = \mathbf{y}_1 - \mathbf{X}_1\hat{\boldsymbol{\beta}}_0^*,$$

where $(\mathbf{X}_1, \mathbf{y}_1)$ is the remaining observations in $(\mathbf{X}, \mathbf{y})$. Compute one of

$$\mathbf{H}_{0,\alpha}^* = \mathbf{X}_1(\mathbf{X}_0^{*\top}\mathbf{X}_0^*)^{-\alpha}\mathbf{X}_1^\top, \quad \hat{\mathbf{H}}_{0,\alpha}^*, \quad \text{and} \quad \bar{\mathbf{H}}_{0,\alpha}^*, \quad \alpha = 1, 2. \tag{14}$$

Our simulations in Section 5 exhibited that the Scoring Algorithm performed paticularly well.

Remark 5. *The Algorithm in Fig. 2 can be implemented in $O(\max(r_0, r)\ p^2)$ much faster than the original running time $O(np^2)$ as $\max(r_0, r) << n$.*

The leverage scores based distribution.

The formula $\ell_i = \mathbf{u}_i^\top\mathbf{u}_i/p$ indicates that ℓ_i depends *only* on the singular vector $\mathbf{u}_i$ of $\mathbf{X}$. Meanwhile, since the $\hat{A}$-optimal $\hat{\pi}_{2,i}$ depends on $h_{2,i,i}$, which can be written as

$$h_{2,i,i} = \mathbf{u}_i^\top\mathrm{Diag}(1/\sigma_1^2, \ldots, 1/\sigma_p^2)\mathbf{u}_i,$$

it follows that $\bar{\pi}_{2,i}$ depends on not only $\mathbf{u}_i$ but all the singular values σ_i's of $\mathbf{X}$. These suggest that $\boldsymbol{\ell}$ is not efficient in extracting information as it ignores the singular value information.

Suppose that $\mathbf{X}$ is column-orthonormal. Then $h_{i,i} = \|\mathbf{x}_i\|^2$ and

$$\bar{\pi}_{2,i} \propto \begin{cases} \sqrt{h_{i,i}} + o(1), & h_{i,i} = o(1), \\ \sqrt{1 - h_{i,i}} + o(1), & h_{i,i} = 1 - o(1). \end{cases}$$

When sampling according to $\boldsymbol{\ell}$, the ith observation is drawn with probability proportional to $h_{i,i}$, especially in the vicinity of $h_{i,i} = 1$. The $\bar{A}$-optimality, however, dictates that in this vicinity the ith observation must be drawn with the probability proportional to $\sqrt{1 - h_{i,i}}$ — decreasing with $h_{i,i}$. In fact, the increasing relationship occurs in the vicinity of $h_{i,i} = 0$ with the probability proportional to $\sqrt{h_{i,i}}$. Similarly, $\bar{\pi}_{2,i} \propto h_{2,i,i}^{1/2}|\hat{\varepsilon}_i|$, suggesting data points closer to the regression hyperplane is less informative than those farther away.

4 Relative error bounds and boundedness of subsample sizes

In this section, we give non-asymptotic error bounds, and conditions for the subsample sizes to be bounded.

Relative error bounds. Drineas, *et al.* (2006)[6] established relative error bounds for a subsampling estimator in a linear model for an arbitrary distribution $\{p_i\}$, and utilized the results to study stochastic algorithms. We now apply their results to specific distributions and, as a result, we obtain explicit formulas for determining subsample sizes. The authors assumed the existence of constants b_1, b_2, b_3 in their conditions (3.8)–(3.10). Specifically, in statistical terms, these conditions can be, respectively, reformulated as

$$(15) \qquad p_i \geq b_1 \ell_{1,i}, \quad p_i \geq b_2 \ell_{2,i}, \quad p_i \geq b_3 \ell_{3,i}, \quad i = 1, \ldots, n,$$

where $b_k > 0$ and $\boldsymbol{\ell}_k = (\ell_{k,i}), k = 1, 2, 3$ are the distributions defined by

$$\ell_{1,i} = \ell_i \propto h_{i,i}, \quad \ell_{2,i} \propto \sqrt{h_{i,i}}|\hat{\varepsilon}_i|, \quad \ell_{3,i} \propto \hat{\varepsilon}_i^2.$$

Here we used the identity $\mathbf{U}^{\perp}\mathbf{U}^{\perp\top}\mathbf{y} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{ols}} = \hat{\boldsymbol{\varepsilon}}$, where $\mathbf{U}^{\perp}$ is the orthogonal matrix whose column space is the ortho-complement of the column space of $\mathbf{U}$.

The key is to determine the values of b_k's. As mentioned by the authors (see also below), almost all distributions will satisfy (15) if one chooses sufficiently small values of b_k's. The small values, nevertheless, will have a direct adverse effect on the sampling complexity. Our goal is, therefore, to find the largest possible values of b_k's. Consider a distribution $\{p_i\}$ of the form

$$(16) \qquad p_i \propto v_{n,i}, \quad i = 1, \ldots, n,$$

where $\mathbf{v} = (v_{n,i}, i = 1, \ldots, n)$ are nonnegative rv's. Let $S_{\mathbf{v}} = \sum_{i=1}^{n} v_{n,i}$. Evidently, the largest values of b_k's are given by

$$(17) \qquad b_k = (S_{\boldsymbol{\ell}_k}/S_{\mathbf{v}}) \min_{1 \leq i \leq n} (v_{n,i}/\ell_{k,i}), \quad k = 1, 2, 3.$$

Observe that the LSE $\hat{\boldsymbol{\beta}}_{\text{ols}}$ and the residuals $\hat{\boldsymbol{\varepsilon}}$ satisfy

$$\inf\{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\| : \boldsymbol{\beta} \in \mathbb{R}^p\} = \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\| = \|\hat{\boldsymbol{\varepsilon}}\|.$$

Let $\tilde{\varepsilon}^* = \sqrt{\mathbf{W}^*}\mathbf{y}^* - \sqrt{\mathbf{W}^*}\mathbf{X}^*\hat{\boldsymbol{\beta}}_r^*$. Obviously, it differs from $\hat{\boldsymbol{\varepsilon}}^*$ and satisfies

$$\inf\{\|\sqrt{\mathbf{W}^*}\mathbf{y}^* - \sqrt{\mathbf{W}^*}\mathbf{X}^*\boldsymbol{\beta}\| : \boldsymbol{\beta} \in \mathbb{R}^p\} = \|\sqrt{\mathbf{W}^*}\mathbf{y}^* - \sqrt{\mathbf{W}^*}\mathbf{X}^*\hat{\boldsymbol{\beta}}_r^*\| = \|\tilde{\boldsymbol{\varepsilon}}^*\|.$$

Define $c(\epsilon, \delta, p) = p^2 \log(3/\delta)/\epsilon^4$. Recognizing the formulae for b_k's in (17), Theorem 3.1 of Drineas, *et al.* can be re-stated in statistical terms as follows:

Theorem 2. *Consider $\{p_i\}$ of the form (16). Let $\epsilon > 0$ and $0 < \delta < 1$. If $r \geq r_1 =: 64c(\epsilon, \delta, p)/\min(b_1^2, b_3^2)$, then with P^*-probability at least $1 - \delta$,*

$$(18) \qquad \|\tilde{\boldsymbol{\varepsilon}}^*\| \leq (1 + \epsilon)\|\hat{\boldsymbol{\varepsilon}}\|.$$

If $r \geq r_2 =: 388c(\sqrt{\epsilon}, \delta, p)/\min(b_1^2, b_2^2, b_3^2)$, then with P^*-probability at least $1 - \delta$,

$$(19) \qquad \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_r^*\| \leq (1 + \epsilon)\|\hat{\boldsymbol{\varepsilon}}\|,$$

$$(20) \qquad \|\hat{\boldsymbol{\beta}}_r^* - \hat{\boldsymbol{\beta}}_{\text{ols}}\| \leq \sigma_{\min}^{-1}(\mathbf{X})\epsilon\|\hat{\boldsymbol{\varepsilon}}\|.$$

If, in addition, $\|\hat{\boldsymbol{\varepsilon}}\| \geq \rho\|\mathbf{y}\|$ for $\rho \in (0, 1)$, then with P^*-probability at least $1 - \delta$,

$$(21) \qquad \|\hat{\boldsymbol{\beta}}_r^* - \hat{\boldsymbol{\beta}}_{\text{ols}}\|/\|\hat{\boldsymbol{\beta}}_{\text{ols}}\| \leq \rho(1 - \rho^2)^{-1/2}\kappa(\mathbf{X})\epsilon.$$

Boundedness of subsample sizes. By Theorem 2, given $\epsilon > 0$ and confidence level $1 - \delta$, subsample sizes r_1 and r_2 given in Theorem 2 can be calculated for a distribution $p_i \propto \mathbf{v}_i$ in (16), although these sizes are far from sharp and improvement can certainly be made. Our interest here is that under what conditions r_1 and r_2 are bounded. In the case of q-stable random errors ε_i's, using a result about Orlicz norms by Gordon, *et al.* (2002)[8], we obtained the following boundedness conditions.

Noting that r_1, r_2 are inversely proportional to b_k^2 for some k, the boundedness is equivalent to $0 < b \le \max_i |c_i \hat{\varepsilon}_i| \le B < \infty$ in probability for constants c_i's provided that the results in Theorem 3 hold. Condition (24) restricts the truncation, which is inevitable from the following Proposition 2, whose proof is clear in view of

$$\min_i |c_{n,i}\hat{\varepsilon}_i| \le \min_i |c_{n,i}\varepsilon_i| + \max_i |c_{n,i}\mathbf{x}_i^\top(\hat{\mathbf{b}} - \boldsymbol{\beta}_0)|.$$

Proposition 2. *Assume that $\varepsilon_1, \dots, \varepsilon_n$ are arbitrary random errors satisfying $\min_i |c_{n,i}\varepsilon_i| = o_P(1)$ for some constants $c_{n,i}$'s. Let $\hat{\mathbf{b}}$ be an estimator of $\boldsymbol{\beta}_0$ such that $\max_i |c_{n,i}\mathbf{x}_i^\top(\hat{\mathbf{b}} - \boldsymbol{\beta}_0)| = o_P(1)$. Then $\min_i |c_{n,i}\hat{\varepsilon}_i| = o_P(1)$.*

Recall a rv Z is q-stable with $q \in (0, 2]$ if its characteristic function satisfies $\mathrm{E}(\exp(\sqrt{-1}tZ)) = \exp(-c|t|^q), t \in \mathbb{R}$ for some constant $c > 0$. The normal variable with zero mean corresponds to $q = 2$.

Theorem 3. *Let $\varepsilon_1, \dots, \varepsilon_n$ be i.i.d. with a q-stable distribution for $q \in (1, 2]$. Suppose that $g_1, \dots, g_n$ satisfy*

$$\frac{1}{n}\sum_{i=1}^n \frac{1}{\sqrt{g_i}} + \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n \frac{|h_{i,j}|}{\sqrt{g_i}} + \sum_{j=1}^n \max_i \frac{h_{i,j}^2}{\sqrt{g_i}} = O(1/q_n), \tag{22}$$

$$\sum\sum_{j_1 \ne j_2} \Big(\max_i \frac{h_{i,j_1}h_{i,j_2}}{\sqrt{g_i}} - \min_i \frac{h_{i,j_1}h_{i,j_2}}{\sqrt{g_i}} \Big) = O(1/q_n), \tag{23}$$

where $q_n = n^{-1}\sum_i \sqrt{g_i}$. We have

1. *For $g_i = 1$, if $\max_i(h_{i,i}) = O(1/n)$, then the uniform $\mathscr{U}$ satisfies $1/b_k(\mathscr{U}) = O_P(1), k = 1, 2, 3$.*
2. *For $g_i = h_{i,i}^2$, the leverage scores based $\boldsymbol{\ell}$ satisfies $1/b_k(\boldsymbol{\ell}) = O_P(1), k = 2, 3$.*

Assume, further, $\kappa(\mathbf{X}) = O(1)$. For $\alpha = 0, 1, 2$, we have

3 *For $g_i = h_{i,i}^2$, if there exists $\mathbf{l}_n = (l_{n,i})$ such that*

$$\sup_n \max_{1 \le i \le n} (n^{1+\alpha} h_{i,i} h_{\alpha,i,i,}/l_{n,i}) < \infty, \tag{24}$$

then the truncated $\hat{\boldsymbol{\pi}}_\alpha(\mathbf{l}_n)$ satisfies $1/b_k(\hat{\boldsymbol{\pi}}_\alpha(\mathbf{l}_n)) = O_P(1)$, $k = 1, 3$.

4 *For $g_i = h_{i,i}(1 - h_{i,i})$, if $\max_i \sqrt{h_{i,i}/(1 - h_{i,i})} = O(1/nq_n)$, then $\bar{\boldsymbol{\pi}}_\alpha$ satisfies $1/b_k(\bar{\boldsymbol{\pi}}_\alpha) = O_P(1)$, $k = 1, 2, 3$.*

5 *For $g_i = h_{i,i}$, if $\max_i(\sqrt{h_{i,i}}) = O(1/nq_n)$, then $\tilde{\boldsymbol{\pi}}_\alpha$ satisfies $1/b_k(\tilde{\boldsymbol{\pi}}_\alpha) = O_P(1)$, $k = 1, 2, 3$.*

It is worth mentioning that the above conditions are necessary for the validity of the results, in view of the ranges $0 \le h_{i,i} \le 1$ and $-1/2 + 1/n \le h_{i,j} \le 1/2$ for $i \ne j$.

5 Simulations

In this Section, we report some simulation results about the numerical behaviors of the A-optimal distributions and their comparison with the uniform and the leverage scores (lev) based distributions.

Simulated "efficiency" of the uniform sampling. Reported on Table 2 are the simulated relative frequencies of the diagonal entries $h_{i,i}^*$ of $\mathbf{H}^* = \mathbf{X}(\mathbf{X}^{*\top}\mathbf{X}^*)^{-1}\mathbf{X}^\top$ falling in [0, 1] based on $n = 10^4$ and 500 repetitions. Here the r rows $\mathbf{x}_j^*$ of $\mathbf{X}^*$ is uniform (Unif) and A-optimal (Aopt) random samples from the n rows $\mathbf{x}_i$ of $\mathbf{X}$, where the rows were generated from the mixture $0.5N(0, \Sigma) + 0.5LN(0, \Sigma)$. Note that $\mathbf{H}^*$ approximates $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ whose diagonal entries satisfy $h_{i,i} \in [0, 1]$. The results indicated that the uniform sampling was inefficient, with its "efficiency" equal to only $1/15$ of the A-optimal sampling considered.

Simulated MSE. As in Zhu, *et al.* (2015)[22], we chose the coefficient $\boldsymbol{\beta} = (\mathbf{1}_{30}^\top, 0.1 \cdot \mathbf{1}_{20}^\top)^\top$, generated $p = 50$-dimensional covariate vector $\mathbf{x}$ (treated as non-random) from Gaussian $\mathrm{N}(\mathbf{0}, \Sigma)$ (GA), Log-normal $\exp(\mathrm{N}(\mathbf{0}, \Sigma))$(LN), and Mixing Gaussian $0.5\mathrm{N}(\mathbf{0}, \Sigma) + 0.5\mathrm{N}(\mathbf{0}, 625\Sigma)$(MG) with $\Sigma_{ij} = 2 * 0.8^{|i-j|}$. The random error ε was generated from the normal ($\mathscr{N}$) and the logistic ($\mathscr{L}$), both with zero mean and unit standard deviation. For sample size $n = 10^5$ and a few subsample sizes r, we calculated the empirical mean squared errors of $\hat{\boldsymbol{\beta}}_r^*$ as follows:

$$\mathrm{EMSE}(\hat{\boldsymbol{\beta}}_r^*) = \frac{1}{M} \sum_{m=1}^{M} \|\hat{\boldsymbol{\beta}}_m^* - \hat{\boldsymbol{\beta}}_{\mathrm{ols}}\|^2, \quad M = 500.$$

Reported on Tables 3–6 are the ratios of the EMSE of $\hat{\boldsymbol{\beta}}_r^*$ to that of the uniform subsampling estimator, where the sampling distributions are untruncated in Table 3 and truncated in Tables 4-6; the residual $\hat{\boldsymbol{\varepsilon}}$ was computed based on the full sample $(\mathbf{X}, \mathbf{y})$ in Tables 3-4 and on a uniform pre-subsample $(\mathbf{X}_0^*, \mathbf{y}_0^*)$ of size $0.1n$ in Tables 5-6. In addition, the Scoring Algorithm in Fig. 2 was used in Table 6.

Observe first that the ratios in all the tables are almost all less than one, indicating that the uniform sampling is ineffective in extracting information. This is most noticeable for $\hat{A}$-optimal sampling, and for the LN covariate in which some of the ratios were as low as 25%. Note that the LN is skewed, whereas both GA and MG are symmetric in which the uniform sampling had better performance. Second, the small differences of the ratios in all the tables indicated that the uniform pre-subsampling of a small size resulted in small loss of efficiency, and that the Scoring Algorithm worked well. Third, the $\hat{A}$-optimal sampling performed the best, and gave substantially smaller EMSE ratios than $\bar{A}$-, $\tilde{A}$- and the leverage scores based sampling. In particular, $\hat{\boldsymbol{\pi}}_2$ gave the smallest EMSE ratios in Table 3, when the subsample size reached half the full sample size, which was mostly kept for the truncated sampling distributions in Tables 4-6.

The Running Time. Reported on Table 7 are the running times of the Scoring Algorithm and the LSE. They were measured on a computing cluster with 16 processors running at 2.60GHz with 250GB of memory. The *R* package (ver 3.3.1) was used to carry out the numerical computations. Since $\mathbf{X}^\top \mathbf{X}$ was approximated by the subsampling $\mathbf{X}_0^{*\top} \mathbf{X}_0^*$, the time-consuming part is the matrix multiplications in $\hat{\mathbf{H}}_2^*$. Instead of using *solve* to find the inverse, we called *svd* to obtain a singular value decomposition of $\mathbf{X}_0^*$ to compute the sampling distribution $\hat{\boldsymbol{\pi}}_2$, and called *lm* to compute both the subsampling estimator $\hat{\boldsymbol{\beta}}_r^*$ and the full data $\hat{\boldsymbol{\beta}}_{\mathrm{ols}}$. The Scoring Algorithm saved time in comparison with the LSE. The times spent on the matrix multiplications were found to be about 30% of the total running times, which can be improved by fast matrix multiplication. Here $\mathbf{x}$ was generated from GA and ε from $\mathscr{N}(0, 1)$. The results for the other distributions of $\mathbf{x}$ and ε considered in Table 3 are similar (not reported here).

6 Proof for boundedness of subsample sizes

We need the following result about Orlicz norms of a sequence of random variables, which combines Example 14 and 17 of Gordon, *et al.* (2002)[8]. A convex function $\mathscr{O} : \mathbb{R}^+ \to \mathbb{R}^+$ is Orlicz if it satisfies $\mathscr{O}(0) = 0$ and $\mathscr{O}(t) > 0$ for $t > 0$. The Orlicz norm of $\mathbf{x} \in \mathbb{R}^n$ is $|\mathbf{x}|_{\mathscr{O}} = \inf \{\rho > 0 : \sum_{i=1}^n \mathscr{O}(|x_i|/\rho) \le 1\}$.

Lemma 1. *Let $Z_1, \ldots, Z_n$ be i.i.d with a $q \in (1, 2]$-stable distribution. Let $\mathscr{O}$ be the Orlicz function given by $\mathscr{O}(0) = 0$ and*

$$\mathscr{O}(t) = \begin{cases} c_1(q)t^p + c_2(q)\exp(-3/2t^2), & t \in (0, 1), \\ d_1(q)t + d_2(q), & t \ge 1, \end{cases}$$

where $c_k(q) \ge 0, d_k(q)$ are absolute constants with $c_1(2) = 0$ and $c_2(q) = 0$ for $q \in (1, 2)$ and $d_1(q) > 0$. Then for every $\mathbf{x} = (x_1, \ldots, x_n)$ there are positive constants c, C such that

$$c|\mathbf{x}|_{\mathscr{O}} \le \mathrm{E}(\max_{1 \le i \le n} |x_i Z_i|) \le C|\mathbf{x}|_{\mathscr{O}}.$$

PROOF (of Theorem 3). Consider $v_{\alpha,n,i} = \sqrt{h_{\alpha,i,i}}e_{n,i}$ for $e_{n,i} \ge 0$. Let $b_k(\mathbf{v}_\alpha)$ be the corresponding b_k in (17). Let $S(\mathbf{e}) = \sum_{i=1}^n \sqrt{h_{i,i}}e_{n,i}$, and let

$$\check{b}_1(\mathbf{e}) = \frac{p}{S(\mathbf{e})} \min_i \frac{e_{n,i}}{\sqrt{h_{i,i}}}, \ \check{b}_2(\mathbf{e}) = \frac{S(\hat{\boldsymbol{\varepsilon}})}{S(\mathbf{e})} \min_i \frac{e_{n,i}}{|\hat{\varepsilon}_i|}, \ \check{b}_3(\mathbf{e}) = \frac{\|\hat{\boldsymbol{\varepsilon}}\|^2}{S(\mathbf{e})} \min_i \frac{\sqrt{h_{i,i}}e_{n,i}}{\hat{\varepsilon}_i^2}.$$

These expressions are independent of α. Recalling $h_{\alpha,i,i}$, one gets

$$\sigma_{\max}^{-|1-\alpha|}(\mathbf{X})\sqrt{h_{i,i}} \le \sqrt{h_{\alpha,i,i}} \le \sigma_{\min}^{-|1-\alpha|}(\mathbf{X})\sqrt{h_{i,i}}, \quad \alpha = 1, 2,$$

whereas for $\alpha = 0$, we have

$$\sigma_{\min}(\mathbf{X})\sqrt{h_{i,i}} \leq \sqrt{h_{0,i,i}} = \|\mathbf{x}_i\| \leq \sigma_{\max}(\mathbf{X})\sqrt{h_{i,i}}, \quad i = 1, \dots, n.$$

As a consequence,

$$\kappa^{-|1-\alpha|}(\mathbf{X}) \leq b_k(\mathbf{v}_\alpha)/\check{b}_k(\mathbf{e}) \leq \kappa^{|1-\alpha|}(\mathbf{X}), \quad \alpha = 0, 1, 2, \; k = 1, 2, 3. \tag{25}$$

Setting now $e_{n,i} = h_{i,i}^{-1/2}$, $\sqrt{h_{i,i}}$, $|\hat{\varepsilon}_i|$, $\sqrt{1 - h_{i,i}}$ and 1, we obtain the upper and lower bounds, respectively, for b_k's in terms of $\check{b}_k(\mathbf{e})$ and $\kappa^{-|1-\alpha|}(\mathbf{X})$ for the distributions $\mathscr{U}$ ($\alpha = 1$), $\boldsymbol{\ell}$ ($\alpha = 1$), $\hat{\boldsymbol{\pi}}_\alpha$, $\bar{\boldsymbol{\pi}}_\alpha$ and $\tilde{\boldsymbol{\pi}}_\alpha$.

We shall prove case 1 ($\boldsymbol{\ell}$) and the rest are similar. It suffices to show $\check{b}_3^{-1}(\boldsymbol{\ell}) = O_P(1)$ because

$$\check{b}_2^{-1}(\boldsymbol{\ell}) = \frac{p}{\sum \sqrt{h_{i,i}}|\hat{\varepsilon}_i|} \max_i\left(\frac{|\hat{\varepsilon}_i|}{\sqrt{h_{i,i}}}\right), \quad \check{b}_3^{-1}(\boldsymbol{\ell}) = \frac{p}{\|\hat{\boldsymbol{\varepsilon}}\|^2} \max_i\left(\frac{\hat{\varepsilon}_i^2}{h_{i,i}}\right).$$

Let $\mathbf{h}_i = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i$. Then $\mathbf{h}_i = (h_{i,j}, j = 1, \dots, n)^\top$ and $\hat{\varepsilon}_i = \varepsilon_i - \mathbf{h}_i^\top\boldsymbol{\varepsilon}$. As a result,

$$\mathrm{E}\Big(\max_i \frac{\hat{\varepsilon}_i^2}{h_{i,i}}\Big) \leq \mathrm{E}\Big(\max_i \frac{\varepsilon_i^2}{h_{i,i}}\Big) + \mathrm{E}\Big(\max_i \frac{(\mathbf{h}_i^\top\boldsymbol{\varepsilon})^2}{h_{i,i}}\Big) - 2\mathrm{E}\Big(\min_i(T_i)\Big), \tag{26}$$

where $T_i = \sum_{j:j\neq i}(h_{i,j}/h_{i,i})\varepsilon_i\varepsilon_j$. By Lemma 1,

$$\mathrm{E}\Big(\max_i \frac{\varepsilon_i^2}{h_{i,i}}\Big) \leq \frac{c_1\sigma^2}{n}\sum_i \frac{1}{h_{i,i}}, \tag{27}$$

where $c_1 > 0$ is a constant. Let $\varepsilon^+, \varepsilon^-$ be the positive, negative parts of ε and $v = \mathrm{E}(\varepsilon^+)$. Then $\mathrm{E}(\varepsilon) = 0$ implies $v = \mathrm{E}(\varepsilon^-)$. As a consequence,

$$\begin{aligned}
\mathrm{E}(\min_i(T_i)) &\geq \sum_{j:j\neq i} \mathrm{E}((\min_i(\varepsilon_i\varepsilon_j h_{i,j}/h_{i,i})) \\
&\geq v\sum_{j:j\neq i} \mathrm{E}(\min_i(h_{i,j}\varepsilon_i/h_{i,i}) - \max_i(h_{i,j}\varepsilon_i/h_{i,i})) \\
&\geq -2v\sum_{j:j\neq i} \mathrm{E}(\max_i(|h_{i,j}\varepsilon_i|/h_{i,i})).
\end{aligned}$$

By Lemma 1 again, there is a constant $c_2 > 0$ such that uniformly in j,

$$\mathrm{E}(\max_i(|h_{i,j}\varepsilon_i|/h_{i,i})) \leq \frac{c_2\sigma}{n}\sum_{i=1}^n \frac{|h_{i,j}|}{h_{i,i}}.$$

This implies

$$-\mathrm{E}(\min_i(T_i)) \leq \frac{2c_2v\sigma}{n}\sum_{i=1}^n\sum_{j=1}^n \frac{|h_{i,j}|}{h_{i,i}}. \tag{28}$$

Noting $\mathrm{E}(\max_i(h_{i,j_1}h_{i,j_2}\varepsilon_{j_1}\varepsilon_{j_2})) = 2v^2(\max_i(h_{i,j_1}h_{i,j_2}) - \min_i(h_{i,j_1}h_{i,j_2}))$ for $j_1 \neq j_2$, and using $(h_i^\top\boldsymbol{\varepsilon})^2 = \sum_j h_{i,j}^2\varepsilon_j^2 + \sum_{j_1\neq j_2} h_{i,j_1}h_{i,j_2}\varepsilon_{j_1}\varepsilon_{j_2}$, we get

$$\begin{aligned}
\mathrm{E}\Big(\max_i \frac{(h_i^\top\boldsymbol{\varepsilon})^2}{h_{i,i}}\Big) &\leq \sigma^2\sum_j \max_i \frac{h_{i,j}^2}{h_{i,i}} + \sum_{j_1\neq j_2} \mathrm{E}\Big(\max_i\Big(\frac{h_{i,j_1}h_{i,j_2}\varepsilon_{j_1}\varepsilon_{j_2}}{h_{i,i}}\Big)\Big) \\
&\leq \sigma^2\sum_j \max_i \frac{h_{i,j}^2}{h_{i,i}} + 2v^2\sum_{j_1\neq j_2}\Big(\max_i \frac{h_{i,j_1}h_{i,j_2}}{h_{i,i}} - \min_i \frac{h_{i,j_1}h_{i,j_2}}{h_{i,i}}\Big).
\end{aligned} \tag{29}$$

The desired result now follows from (25)–(29). □

7 Proof for asymptotic normality

A rv $\mathbf{w} = (w_1, \cdots, w_n)^\top \sim \mathrm{sMult}(\boldsymbol{\pi}, r)$ (the scaled multinomial distribution) for $\boldsymbol{\pi} \in [0,1]^n$ with $\sum_{i=1}^n \pi_i = 1$ if

$$\mathrm{P}\Big(w_1 = \frac{k_1}{r\pi_1}, \ldots, w_n = \frac{k_n}{r\pi_n}\Big) = \frac{r!}{\prod_{i=1}^n k_i!} \prod_{i=1}^n \pi_i^{k_i}, \quad k_i \geq 0, \sum_{i=1}^n k_i = r. \tag{30}$$

It is customary to express $\hat{\boldsymbol{\beta}}_r^*$ in the full data using $\mathbf{w}$, decoupling the resampling scheme from the data. Stochastically equivalently,

$$\hat{\boldsymbol{\beta}}_r^* \overset{d}{=} (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{y}, \quad \mathbf{W} = \mathrm{Diag}(\mathbf{w}), \tag{31}$$

where $\mathbf{x} \overset{d}{=} \mathbf{y}$ denotes $\mathbf{x}$ and $\mathbf{y}$ have the same distribution. Note that the laws $\mathrm{P}_{\mathbf{w}}$ and P^* governed by $\mathrm{sMult}(\boldsymbol{\pi}, r)$ and $\boldsymbol{\pi}$, respectively, are stochastically equivalent, see, e.g., page 2055, Præstgaard and Wellner (1993)[16] and Zhu, *et al.* (2015)[22]. Such equivalence is commonly used in the bootstrap theory, see Sections 3.5–3.6, Van de Vaart and Wellner (1996)[18]. We shall use P^* also for $\mathrm{P}_{\mathbf{w}}$, and write E^*, Var^*, etc. for the expected value, variance, etc. It is easy to check

$$\mathrm{E}^*(\mathbf{w}) = \mathbf{1}, \quad \mathrm{Cov}^*(\mathbf{w}) = (1/r)(\mathrm{Diag}(1/\boldsymbol{\pi}) - \mathbf{1}\mathbf{1}^\top). \tag{32}$$

Lemma 2. *Assume (M2). Suppose (6) holds for all $\varrho > 0$ and (7) holds for some $\rho > 2$. Then*

$$\|\hat{\boldsymbol{\beta}}_{\mathrm{ols}} - \boldsymbol{\beta}_0\| = O(n^{-1/2} \log_2^{1/2}(n)), \quad a.s. \tag{33}$$

Hence,

$$\max_{1 \leq i \leq n} |\mathbf{x}_i^\top (\hat{\boldsymbol{\beta}}_{\mathrm{ols}} - \boldsymbol{\beta}_0)| = o(1), \quad a.s. \tag{34}$$

PROOF. We show without loss of generality that (33) holds for the first component $\hat{\beta}_1$ of $\hat{\boldsymbol{\beta}}_{\mathrm{ols}}$. To do so, we shall apply Theorem 2 of Lai and Wei (1982)[10], for which we need to verify

$$\lim_{n\to\infty} A_n = \infty, \quad \limsup_{n\to\infty} A_{n+1}/A_n < \infty, \quad \text{and} \tag{35}$$

$$\max_{1\leq i\leq n} |x_{i,1} - \mathbf{k}_n^\top \mathbf{H}_n^{-1} \mathbf{t}_i| = o(n^{1/2} \log^{-\varrho}(n)) \tag{36}$$

for all $\varrho > 0$, where $\mathbf{x}_i = (x_{i,1}, \mathbf{t}_i^\top)^\top$, $\mathbf{k}_n = \sum_{i=1}^n x_{i,1} \mathbf{t}_i$, $\mathbf{H}_n = \sum_{i=1}^n \mathbf{t}_i \mathbf{t}_i^\top$, and $A_n = \sum_{i=1}^n (x_{i,1} - \mathbf{k}_n^\top \mathbf{H}_n^{-1} \mathbf{t}_i)^2$. Partition $\mathbf{M}_0$ as follows:

$$\mathbf{M}_0 = \begin{pmatrix} m_{1,1} & \mathbf{m}_1^\top \\ \mathbf{m}_1 & \mathbf{M}_{1,1} \end{pmatrix}.$$

It follows from (M2) that

$$\frac{1}{n} \sum_{i=1}^n x_{i,1}^2 = m_{1,1} + o(1), \quad \frac{\mathbf{k}_n}{n} = \mathbf{m}_1 + o(1), \quad \frac{\mathbf{H}_n}{n} = \mathbf{M}_{1,1} + o(1). \tag{37}$$

The last two equalities imply $\mathbf{k}_n^\top \mathbf{H}_n^{-1} = \mathbf{m}_1^\top \mathbf{M}_{1,1} + o(1)$. Hence,

$$n^{-1} A_n = m_{1,1} - \mathbf{m}_1^\top \mathbf{M}_{1,1}^{-1} \mathbf{m}_1 + o(1).$$

Since the above difference is positive as it is the inverse of the positive definite matrix $\mathbf{M}_0$, it follows that (35) holds, while (36) follows from the triangle inequality, $\|\mathbf{t}_i\| \leq \max_{1\leq i\leq n} \|\mathbf{x}_i\|$ and (6). Apply now Theorem 2 of Lai and Wei (1982)[10] to finish the proof. □

PROOF (of Theorem 1). Let

$$\bar{\mathbf{w}} = \mathbf{w} - \mathbf{1}, \quad \bar{\mathbf{W}} = \mathbf{W} - \mathbf{I}, \quad \boldsymbol{\Delta}^* = (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} - (\mathbf{X}^\top \mathbf{X})^{-1}. \tag{38}$$

Then $\mathrm{E}_{\mathbf{w}}(\bar{\mathbf{w}}) = 0$, $\mathrm{E}_{\mathbf{w}}(\bar{\mathbf{W}}) = 0$, and stochastically equivalently,

$$\boldsymbol{\Delta}^* \overset{d}{=} (\mathbf{X}^{*\top} \mathbf{W}^* \mathbf{X}^*)^{-1} - (\mathbf{X}^\top \mathbf{X})^{-1}, \ \mathbf{X}^\top \bar{\mathbf{W}} \mathbf{y} \overset{d}{=} \mathbf{X}^{*\top} \mathbf{W}^* \mathbf{y}^* - \mathbf{X}^\top \mathbf{y}. \tag{39}$$

Let $\mathbf{\Delta}_1^* = -(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\bar{\mathbf{W}}\mathbf{X})$. Stochastically equivalently,

$$\bar{\mathbf{\Delta}}_1^* =: \mathbf{I} - \mathbf{\Delta}_1^* = (\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{W}\mathbf{X}). \tag{40}$$

Recall $\bar{\mathbf{W}}$ and $\mathbf{\Delta}^*$ in (38) and write

$$(\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1} = (\mathbf{X}^\top\mathbf{X})^{-1} + \mathbf{\Delta}^*, \quad \mathbf{W}\mathbf{y} = \mathbf{y} + \bar{\mathbf{W}}\mathbf{y}.$$

Substitution of them in the full-data formula (31) of $\hat{\boldsymbol{\beta}}_r^*$ yields

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_r^* &= (\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{W}\mathbf{y} \\
&= \big((\mathbf{X}^\top\mathbf{X})^{-1} + \mathbf{\Delta}^*\big)\mathbf{X}^\top\big(\mathbf{y} + \bar{\mathbf{W}}\mathbf{y}\big) \\
&= \hat{\boldsymbol{\beta}}_{\text{ols}} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\bar{\mathbf{W}}\mathbf{y} + \mathbf{\Delta}^*\mathbf{X}^\top\mathbf{y} + \mathbf{\Delta}^*\mathbf{X}^\top\bar{\mathbf{W}}\mathbf{y} \\
&= \hat{\boldsymbol{\beta}}_{\text{ols}} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\bar{\mathbf{W}}\hat{\boldsymbol{\varepsilon}} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\bar{\mathbf{W}}\hat{\mathbf{y}} + \mathbf{\Delta}^*\mathbf{X}^\top\mathbf{y} + \mathbf{\Delta}^*\mathbf{X}^\top\bar{\mathbf{W}}\mathbf{y} \\
&= \hat{\boldsymbol{\beta}}_{\text{ols}} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\bar{\mathbf{W}}\hat{\boldsymbol{\varepsilon}} + \mathbf{\Delta}^*\mathbf{X}^\top\bar{\mathbf{W}}\hat{\boldsymbol{\varepsilon}} + [(\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top\bar{\mathbf{W}}\hat{\mathbf{y}} + \mathbf{\Delta}^*\mathbf{X}^\top\mathbf{y}].
\end{aligned}$$

Substituting $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ in the square bracket, the sum inside it is identically zero. Since all the preceding statements hold on the subspace in which $\mathbf{X}^\top\mathbf{W}\mathbf{X}$ is invertible, we show (41)-(42),

$$\hat{\boldsymbol{\beta}}_r^* = \hat{\boldsymbol{\beta}}_{\text{ols}} + \frac{1}{r}\sum_{j=1}^{r}(\mathbf{X}^\top\mathbf{X})^{-1}\frac{\mathbf{x}_j^*\hat{\varepsilon}_j^*}{\pi_j^*} + \mathbf{r}^*, \tag{41}$$

valid on the subspace in which $\mathbf{X}^{*\top}\mathbf{W}^*\mathbf{X}^*$ is invertible, where $\mathbf{r}^*$ is given by

$$\mathbf{r}^* = \big((\mathbf{X}^{*\top}\mathbf{W}^*\mathbf{X}^*)^{-1} - (\mathbf{X}^\top\mathbf{X})^{-1}\big)(\mathbf{X}^{*\top}\mathbf{W}^*\hat{\boldsymbol{\varepsilon}}^*). \tag{42}$$

Let A_n^* be the event on which $\bar{\mathbf{\Delta}}_1^*$ is nonsingular. Using $\mathbf{\Delta}_1^*(\bar{\mathbf{\Delta}}_1^*)^{-1} = (\bar{\mathbf{\Delta}}_1^*)^{-1}\mathbf{\Delta}_1^*$, we express

$$\mathbf{\Delta}^* = \mathbf{\Delta}_1^*(\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1} = \mathbf{\Delta}_1^*(\bar{\mathbf{\Delta}}_1^*)^{-1}(\mathbf{X}^\top\mathbf{X})^{-1} = (\bar{\mathbf{\Delta}}_1^*)^{-1}\mathbf{\Delta}_1^*(\mathbf{X}^\top\mathbf{X})^{-1},$$

valid on A_n^*. Recalling $\boldsymbol{\delta}^* = (\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\bar{\mathbf{W}}\hat{\boldsymbol{\varepsilon}})$, we thus obtain

$$\mathbf{r}^* = \mathbf{\Delta}^*\mathbf{X}^\top\bar{\mathbf{W}}\hat{\boldsymbol{\varepsilon}} = (\bar{\mathbf{\Delta}}_1^*)^{-1}\mathbf{\Delta}_1^*\boldsymbol{\delta}^* \quad \text{valid on } A_n^*. \tag{43}$$

By the second equality in (32), one gets

$$\mathrm{E}^*(\|\mathbf{\Delta}_1^*\|^2) \le \frac{1}{r}\sum_{i=1}^{n}\frac{h_{2,i,i}}{\pi_i}\|\mathbf{x}_i\|^2.$$

Using $\hat{\varepsilon}_i^2 \le 2\varepsilon_i^2 + 2\|\boldsymbol{\beta}_0\|^2\|\mathbf{x}_i\|^2$, one has

$$\mathrm{E}^*(\|\boldsymbol{\delta}^*\|^2) \le \frac{1}{r}\sum_{i=1}^{n}\frac{h_{2,i,i}}{\pi_i}\hat{\varepsilon}_i^2 \le \frac{2}{r}\sum_{i=1}^{n}\frac{h_{2,i,i}}{\pi_i}\varepsilon_i^2 + \frac{2\|\boldsymbol{\beta}_0\|^2}{r}\sum_{i=1}^{n}\frac{h_{2,i,i}}{\pi_i}\|\mathbf{x}_i\|^2.$$

It thus follows from (M1) and (M3) that

$$r[\mathrm{E}^*(\|\mathbf{\Delta}_1^*\boldsymbol{\delta}^*\|)]^2 \le r\mathrm{E}^*(\|\mathbf{\Delta}_1^*\|^2)\mathrm{E}^*(\|\boldsymbol{\delta}^*\|^2) = o(1), \quad a.s.$$

This, $\bar{\mathbf{\Delta}}^* = \mathbf{I} + o_{P^*}(1)$ a.s. and the expression (43) for the remainder $\mathbf{r}^*$ prove $\sqrt{r}\mathbf{r}^* = o_{P^*}(1)$ a.s. Consequently, by (41), it suffices to show for any $\mathbf{t} \in \mathbb{R}^p$ with $\|\mathbf{t}\| = 1$,

$$\frac{\sigma_n^{-1}(\mathbf{t})}{\sqrt{r}}\sum_{j=1}^{r}\mathbf{t}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\frac{\mathbf{x}_j^*\hat{\varepsilon}_j^*}{\pi_{nj}^*} \Longrightarrow \mathcal{N}(0,1), \quad a.s. \quad r \to \infty, \tag{44}$$

where $\sigma_n^2(\mathbf{t}) = \mathbf{t}^\top\mathbf{\Sigma}(\boldsymbol{\pi})\mathbf{t}$. As $\mathbf{X}^\top\hat{\boldsymbol{\varepsilon}} = 0$, we have

$$\mathrm{E}^*\big(\mathbf{x}_j^*\hat{\varepsilon}_j^*/\pi_{nj}^*\big) = \mathbf{X}^\top\hat{\boldsymbol{\varepsilon}} = 0, \quad \mathrm{Var}^*(\mathbf{x}_j^*\hat{\varepsilon}_j^*/\pi_{nj}^*) = \mathbf{X}^\top\mathrm{Diag}(\hat{\boldsymbol{\varepsilon}}^2/\boldsymbol{\pi})\mathbf{X}. \tag{45}$$

Let $\xi_j^* = \mathbf{t}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_j^*\hat{\varepsilon}_j^*/\pi_{nj}^*$. It is shown below for every $\eta > 0$,

$$\sigma_n^{-2}(\mathbf{t})\mathrm{E}^*(|\xi_1^*|^2\mathbf{1}[|\xi_1^*| > \sqrt{r}\sigma_n(\mathbf{t})\eta]) \to 0, \quad a.s. \quad r \to \infty. \tag{46}$$

We now apply the Lindeberg-Feller theorem (e.g. Theorem 7.2.1. of Chung (2001)[5]) to claim (44). To show (46), we prove below

$$\frac{1}{n^2}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^\top}{\pi_i}(\hat{\varepsilon}_i^2-\varepsilon_i^2)=o(1), \quad a.s. \tag{47}$$

Let $\mathbf{\Sigma}_c = n^{-2}\mathbf{X}^\top\mathrm{Diag}(\hat{\boldsymbol{\varepsilon}}^2/\boldsymbol{\pi})\mathbf{X}$. Then $\mathbf{\Sigma}(\boldsymbol{\pi}) = (n^{-1}\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{\Sigma}_c(n^{-1}\mathbf{X}^\top\mathbf{X})^{-1}$. It follows from (47) and (M1) that

$$\begin{aligned}\mathbf{\Sigma}_c &= \frac{1}{n^2}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^\top}{\pi_i}\sigma^2+\frac{1}{n^2}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^\top}{\pi_i}(\varepsilon_i^2-\sigma^2)+\frac{1}{n^2}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^\top}{\pi_i}(\hat{\varepsilon}_i^2-\varepsilon_i^2)\\ &= \frac{1}{n^2}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^\top}{\pi_i}\sigma^2+o(1), \quad a.s.\end{aligned}$$

We now use (M2) to get

$$\mathbf{\Sigma}(\boldsymbol{\pi}) = \sigma^2\Gamma_n^{-1}\frac{1}{n^2}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^\top}{\pi_i}\Gamma_n^{-1}+o(1), \quad a.s. \tag{48}$$

This immediately yields for any unit vector $\mathbf{t}$,

$$\sigma_n^2(\mathbf{t}) = \sigma^2\mathbf{t}^\top\Gamma_n^{-1}\frac{1}{n^2}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^\top}{\pi_i}\Gamma_n^{-1}\mathbf{t}+o(1), \quad a.s. \tag{49}$$

By (M2)–(M4), there are constants b_0, B_0 such that

$$0 < b_0 \le \sup \sup_{\|\mathbf{t}\|=1} \sigma_n^2(\mathbf{t}) \le B_0 < \infty, \quad a.s.$$

This shows that (46) is implied by the following (shown below)

$$L(r,n) := \mathrm{E}^*(|\xi_1^*|^2\mathbf{1}[|\xi_1^*| > \sqrt{r}b_0\eta]) \to 0, \quad a.s. \quad r\to\infty. \tag{50}$$

To prove (47), we use (M1) and (M3) to get

$$\frac{1}{n^2}\sum_{i=1}^{n}\frac{\|\mathbf{x}_i\|^2}{\pi_i}\varepsilon_i^2 = \frac{1}{n^2}\sum_{i=1}^{n}\frac{\|\mathbf{x}_i\|^2}{\pi_i}(\varepsilon_i^2-\sigma^2)+O(1)=O(1),\ a.s. \tag{51}$$

By (34), we have uniformly in $i=1,\dots,n$,

$$\hat{\varepsilon}_i-\varepsilon_i = \mathbf{x}_i^\top(\hat{\boldsymbol{\beta}}_{\mathrm{ols}}-\boldsymbol{\beta}_0)=o(1), \quad \hat{\varepsilon}_i+\varepsilon_i = 2\varepsilon_i+o(1), \quad a.s. \tag{52}$$

Thus $\hat{\varepsilon}_i^2-\varepsilon_i^2 = o(1)\varepsilon_i$ a.s. uniformly in i. This yields (47) in view of

$$\Big\|\frac{1}{n^2}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^\top}{\pi_i}\varepsilon_i\Big\|^2 \le \frac{1}{n^2}\sum_{i=1}^{n}\frac{\|\mathbf{x}_i\|^2}{\pi_i}\frac{1}{n^2}\sum_{i=1}^{n}\frac{\|\mathbf{x}_i\|^2}{\pi_i}\varepsilon_i^2 = O(1), \quad a.s.$$

where (51) and (M3) were used. To finish, it remains to prove (50). This follows from (M2), (M5), the first equality in (52), and

$$\begin{aligned}L(r,n) &= \sum_{i=1}^{n}\frac{|\mathbf{t}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i|^2}{\pi_i}\hat{\varepsilon}_i^2\mathbf{1}\big[\frac{|\mathbf{t}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i|}{\pi_i}|\hat{\varepsilon}_i| \ge \sqrt{r}b_0\eta\big]\\ &\le 2\|\Gamma_n^{-1}\|_o^2\frac{1}{n^2}\sum_{i=1}^{n}\frac{\|\mathbf{x}_i\|^2\hat{\varepsilon}_i^2}{\pi_i}\mathbf{1}\big[\frac{\|\mathbf{x}_i\||\hat{\varepsilon}_i|}{n\pi_i} \ge \frac{\sqrt{r}b_0\eta}{\|\Gamma_n^{-1}\|_o}\big]\\ &\le 4\|\Gamma_n^{-1}\|_o^2\frac{1}{n^2}\sum_{i=1}^{n}\frac{\|\mathbf{x}_i\|^2\varepsilon_i^2}{\pi_i}\mathbf{1}\big[\frac{\|\mathbf{x}_i\||\varepsilon_i|}{n\pi_i} \ge \frac{\sqrt{r}b_0\eta}{2\|\Gamma_n^{-1}\|_o}\big]\\ &\longrightarrow 0, \quad a.s. \quad r\to\infty. \qquad\qquad \square\end{aligned}$$

See Table 2 in which the "efficiency" of the uniform sampling is only 1/15 of the optimal sampling for $n = 10^4$ and 500 repetitions.

Table 1: Simulated relative frequencies of the diagonal entries $h^*_{i,i}$ falling in $[0,\ 1]$.

r	0.1n	0.3n	0.5n	0.9n	1.2n	1.5n	2.0n	3.0n	3.5n	4.5n
Aopt	0.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Unif	0.00	0.00	0.00	0.07	0.18	0.33	0.60	0.91	0.98	1.00

Table 2: Simulated relative frequencies of the diagonal entries $h^*_{i,i}$ of $\mathbf{H}^* = \mathbf{X}(\mathbf{X}^{*\top}\mathbf{X}^*)^{-1}\mathbf{X}^\top$ falling in $[0,\ 1]$ for a uniform (Unif) and A-optimal (Aopt) random sample $\mathbf{X}^*$.

r	0.1n	0.3n	0.5n	0.9n	1.2n	1.5n	2.0n	3.0n	3.5n	4.5n
Aopt	0.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Unif	0.00	0.00	0.00	0.07	0.18	0.33	0.60	0.91	0.98	1.00

Here $\mathbf{X}^* = (\mathbf{x}^*_1, \ldots, \mathbf{x}^*_r)^\top$ was of size r drawn from $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ with $\mathbf{x}_i$'s generated from the mixture distribution $0.5N(0,\Sigma) + 0.5LN(0,\Sigma)$. Note $\mathbf{H}^*$ approximates $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ whose diagonal entries satisfy $h_{i,i} \in [0,1]$.

Table 3: Simulated ratios of the MSE of the subsampling estimator $\hat{\boldsymbol{\beta}}^*_r$ according to the optimal sampling distributions and the leverage scores to the MSE of Efron's (uniform) subsampling estimator with sample size $n = 10^5$ and subsample sizes r. The residual $\hat{\boldsymbol{\varepsilon}}$ was computed based on the full sample.

$\mathbf{x}$	ε	$r:n$	$\hat{\boldsymbol{\pi}}_2$	$\hat{\boldsymbol{\pi}}_1$	$\hat{\boldsymbol{\pi}}_0$	$\bar{\boldsymbol{\pi}}_2$	$\bar{\boldsymbol{\pi}}_1$	$\bar{\boldsymbol{\pi}}_0$	$\tilde{\boldsymbol{\pi}}_2$	$\tilde{\boldsymbol{\pi}}_1$	$\tilde{\boldsymbol{\pi}}_0$	Lev
GA	$\mathscr{N}$	.5%	.823	.832	.884	.992	.975	.999	.968	.960	1.03	.979
		1%	.784	.783	.813	1.01	.995	1.07	.994	1.01	1.06	1.02
		10%	.649	.658	.685	.981	.983	1.03	.985	.997	1.02	1.00
		45%	.653	.638	.651	.983	.991	1.03	.994	.982	1.06	1.01
		50%	.620	.629	.660	.961	.971	1.04	.965	.964	1.02	.995
	$\mathscr{L}$	.5%	.795	.813	.873	.990	1.02	1.03	.988	.998	1.04	1.02
		1%	.728	.716	.752	1.00	.985	1.02	.990	.987	1.02	.999
		10%	.618	.615	.661	1.03	1.03	1.06	1.01	1.02	1.07	1.04
		45%	.565	.588	.610	.980	.975	1.01	.987	.989	1.04	.998
		50%	.586	.599	.613	1.00	1.00	1.03	.990	.988	1.04	.983
LN	$\mathscr{N}$	.5%	.302	.303	.322	.333	.327	.352	.328	.332	.360	.493
		1%	.281	.278	.306	.338	.334	.366	.338	.331	.360	.599
		10%	.262	.267	.282	.381	.387	.401	.379	.389	.404	.851
		45%	.276	.278	.286	.419	.425	.447	.415	.425	.453	.952
		50%	.280	.280	.293	.430	.428	.450	.431	.435	.441	.977
	$\mathscr{L}$	.5%	.283	.284	.315	.324	.333	.361	.330	.335	.361	.486
		1%	.256	.253	.279	.331	.330	.361	.332	.331	.361	.576
		10%	.238	.238	.254	.382	.388	.404	.382	.385	.402	.848
		45%	.253	.253	.266	.412	.422	.450	.428	.426	.444	.942
		50%	.253	.253	.268	.420	.425	.450	.418	.427	.446	.959
MG	$\mathscr{N}$	.5%	.558	.551	.593	.644	.651	.675	.633	.636	.687	.900
		1%	.515	.506	.542	.655	.662	.709	.649	.651	.690	.948
		10%	.451	.454	.476	.682	.695	.723	.685	.683	.714	1.02
		45%	.438	.446	.458	.684	.692	.719	.694	.682	.698	1.01
		50%	.433	.438	.459	.671	.680	.721	.667	.697	.710	1.00
	$\mathscr{L}$	.5%	.554	.555	.562	.664	.658	.696	.648	.670	.690	.933
		1%	.500	.500	.509	.662	.685	.706	.672	.659	.698	.953
		10%	.399	.408	.428	.658	.654	.713	.673	.660	.690	.971
		45%	.395	.397	.417	.666	.684	.699	.673	.692	.712	.974
		50%	.407	.410	.428	.710	.685	.712	.690	.683	.722	.995

References

[1] BAXTER, J., JONES, R., LIN, M. and OLSEN, J. (2004). SLLN for Weighted Independent Identically Distributed Random Variables. *J. Theoret. Probab.*, **17**: 165–181. doi:10.1023/B:JOTP.0000020480.84425.8d.

[2] BARBE, P. AND BERTAIL, P. (1995). *Weighted bootstrap*. Lecture Notes in Statist. Vol. 98, Springer, New York.

[3] CANDÉS, E.J. and TAO, T. (2009). Exact Matrix Completion via Convex Optimization. *Found Comput Math* **9**:

Table 4: Same as Table 3 except that the sampling distributions are truncated.

$\mathbf{x}$	ε	$r:n$	$\hat{\boldsymbol{\pi}}_2$	$\hat{\boldsymbol{\pi}}_1$	$\hat{\boldsymbol{\pi}}_0$	$\bar{\boldsymbol{\pi}}_2$	$\bar{\boldsymbol{\pi}}_1$	$\bar{\boldsymbol{\pi}}_0$	$\tilde{\boldsymbol{\pi}}_2$	$\tilde{\boldsymbol{\pi}}_1$	$\tilde{\boldsymbol{\pi}}_0$	Lev
Truncation 10%												
GA	$\mathscr{N}$	.5%	.800	.812	.852	.985	.964	1.03	1.00	.994	1.02	.976
		1%	.718	.740	.755	.961	1.00	1.03	.985	1.01	1.04	.994
		10%	.646	.646	.680	.982	1.01	1.04	.989	1.00	1.04	1.00
	$\mathscr{L}$	.5%	.744	.775	.788	.977	.981	1.02	.952	.985	1.02	.973
		1%	.668	.686	.714	.964	.999	1.03	.960	.983	1.02	.996
		10%	.595	.588	.625	.995	.999	1.04	1.01	.995	1.02	.998
LN	$\mathscr{N}$	.5%	.305	.302	.322	.337	.330	.353	.323	.320	.361	.458
		1%	.269	.275	.286	.339	.329	.362	.331	.336	.359	.569
		10%	.260	.263	.278	.384	.392	.402	.386	.390	.409	.813
	$\mathscr{L}$	.5%	.287	.279	.303	.328	.327	.370	.324	.331	.358	.462
		1%	.253	.258	.277	.335	.340	.369	.331	.334	.364	.580
		10%	.247	.247	.258	.396	.406	.425	.398	.391	.424	.840
MG	$\mathscr{N}$	.5%	.545	.559	.572	.637	.632	.656	.633	.656	.670	.888
		1%	.522	.514	.556	.656	.674	.708	.661	.688	.705	.964
		10%	.455	.449	.477	.692	.685	.715	.695	.681	.724	.985
	$\mathscr{L}$	.5%	.527	.534	.558	.653	.638	.676	.636	.650	.684	.905
		1%	.478	.476	.504	.664	.663	.697	.649	.665	.687	.955
		10%	.412	.416	.430	.680	.671	.718	.676	.664	.695	.959
Truncation 30%												
GA	$\mathscr{N}$	.5%	.753	.749	.802	.980	.983	1.02	.971	1.01	1.02	1.01
		1%	.705	.689	.726	.970	.974	1.01	.967	.980	1.01	.971
		10%	.664	.684	.708	.995	1.01	1.06	.991	1.03	1.03	1.00
	$\mathscr{L}$	.5%	.701	.712	.730	.978	.983	1.01	.980	.989	.999	.990
		1%	.658	.673	.694	1.00	1.01	1.02	.991	1.00	1.02	1.01
		10%	.612	.619	.638	.989	.983	1.00	.987	.994	1.03	.998
LN	$\mathscr{N}$	.5%	.295	.301	.330	.340	.334	.373	.344	.342	.384	.422
		1%	.269	.266	.295	.326	.332	.356	.323	.341	.356	.500
		10%	.263	.264	.280	.391	.394	.416	.393	.391	.414	.741
	$\mathscr{L}$	.5%	.290	.291	.309	.348	.344	.390	.350	.344	.376	.434
		1%	.258	.257	.276	.331	.342	.364	.337	.340	.362	.515
		10%	.247	.251	.263	.403	.399	.426	.395	.398	.423	.747
MG	$\mathscr{N}$	.5%	.560	.546	.580	.646	.645	.659	.652	.651	.666	.866
		1%	.504	.510	.532	.659	.657	.698	.667	.656	.692	.886
		10%	.456	.466	.481	.685	.681	.733	.685	.677	.709	.945
	$\mathscr{L}$	.5%	.524	.535	.546	.655	.650	.642	.652	.658	.670	.861
		1%	.468	.481	.500	.660	.663	.688	.663	.670	.689	.926
		10%	.421	.420	.439	.693	.692	.710	.675	.693	.724	.939

717. doi:10.1007/s10208-009-9045-5.

[4] CHATTERJEE, S. AND BOSE, A. (2002). Dimension asymptotics for generalized bootstrap in linear regression. *Ann. Inst. Statist. Math.* **54** (2): 367–381.

[5] CHUNG, K.L. (2001). *A Course in Probability Theory*. Academic Press, San Diego, CA.

[6] DRINEAS P., KANNAN R. and MAHONEY M.W. (2006). Fast Monte Carlo algorithms for matrices I: Approximating matrix multiplication. *SIAM Journal on Computing*, **36**: 132–157.

[7] DRINEAS P., MAGDON-ISMAIL, M., MAHONEY M.W. and WOODRUFF, D.P. (2012). Fast approximation of matrix coherence and statistical leverage. *The Journal of Machine Learning Research*, **13**: 3475–3506.

[8] GORDON, R., LITVAK, A., SCHÜTT, C. AND WERNER, E. (2002). Orlicz norms of sequences of random variables. *Ann. Probab.* **30**(4): 1833-1853.

[9] KLEINER, A., TALWALKAR, A., SARKAR, P. AND JORDAN, M. I. (2014). A scalable bootstrap for massive data. *J. R. Stat. Soc. Series B Stat. Methodol.* **76**(4), 795–816.

[10] LAI, T. L. AND C. Z. WEI (1982). A Law of the Iterated Logarithm for Double Arrays of Independent Random Variables with Applications to Regression and Time Series Models. *Ann. Probab.* **10**(2): 320–335.

[11] LIANG, F., CHENG, Y., SONG, Q., PARK, J., AND YANG, P. (2013). stochastic approximation method for analysis of large geostatistical data. *J. Am. Stat. Assoc.* **108**(501): 325–339.

[12] MA, P. AND SUN, X. (2014). Leveraging for big data regression. *Computational Statistics*. **7** (1): 70-76.

Table 5: Same as Table 4 except that the residual $\hat{\varepsilon}$ was approximated based on a uniform pre-subsample $\mathbf{X}_0^*$ of size $r_0 : n = 10\%$.

$\mathbf{x}$	ε	$r : n$	$\hat{\boldsymbol{\pi}}_2$	$\hat{\boldsymbol{\pi}}_1$	$\hat{\boldsymbol{\pi}}_0$	$\bar{\boldsymbol{\pi}}_2$	$\bar{\boldsymbol{\pi}}_1$	$\bar{\boldsymbol{\pi}}_0$	$\tilde{\boldsymbol{\pi}}_2$	$\tilde{\boldsymbol{\pi}}_1$	$\tilde{\boldsymbol{\pi}}_0$	Lev
Truncation 10%												
GA	$\mathscr{N}$	.5%	.843	.814	.868	.985	1.01	1.06	1.00	1.01	1.03	1.02
		1%	.730	.724	.781	1.00	.984	1.05	.975	.996	1.04	.994
		10%	.649	.651	.699	1.02	.986	1.04	.996	1.01	1.05	1.00
	$\mathscr{L}$	.5%	.808	.798	.846	.988	1.01	1.04	1.02	1.02	1.05	1.02
		1%	.690	.694	.715	.982	.992	1.02	.977	.990	1.03	1.00
		10%	.593	.603	.631	.991	1.01	1.05	.984	1.02	1.03	.986
LN	$\mathscr{N}$	.5%	.288	.287	.311	.320	.324	.365	.322	.325	.353	.474
		1%	.268	.269	.298	.336	.335	.365	.338	.342	.364	.585
		10%	.279	.281	.295	.391	.387	.415	.401	.399	.417	.834
	$\mathscr{L}$	.5%	.256	.266	.286	.318	.316	.349	.312	.320	.341	.455
		1%	.249	.255	.277	.339	.340	.355	.335	.326	.349	.572
		10%	.253	.258	.270	.382	.391	.407	.383	.390	.409	.828
MG	$\mathscr{N}$	.5%	.555	.554	.587	.636	.643	.665	.628	.638	.673	.879
		1%	.527	.533	.547	.660	.669	.708	.676	.681	.690	.944
		10%	.460	.466	.491	.714	.701	.743	.697	.707	.745	1.02
	$\mathscr{L}$	.5%	.531	.516	.549	.630	.633	.652	.632	.640	.661	.888
		1%	.471	.483	.501	.650	.652	.674	.643	.639	.679	.922
		10%	.422	.420	.442	.670	.673	.691	.666	.676	.709	.972
Truncation 30%												
GA	$\mathscr{N}$	.5%	.772	.781	.828	1.01	.989	1.01	1.00	1.01	1.04	1.00
		1%	.694	.718	.743	.978	1.01	1.02	.972	.978	1.04	1.00
		10%	.683	.686	.706	1.02	1.01	1.03	1.03	1.00	1.04	1.01
	$\mathscr{L}$	.5%	.722	.714	.739	.983	.996	1.03	.985	1.01	1.03	.996
		1%	.652	.666	.691	.996	.988	1.04	.993	.990	1.02	1.01
		10%	.614	.613	.623	.987	.998	1.02	.982	.995	1.03	.998
LN	$\mathscr{N}$	.5%	.288	.294	.309	.328	.333	.368	.329	.337	.366	.413
		1%	.263	.264	.283	.321	.346	.364	.336	.338	.368	.510
		10%	.279	.275	.297	.396	.398	.419	.383	.395	.416	.732
	$\mathscr{L}$	.5%	.258	.270	.294	.332	.336	.365	.332	.331	.362	.408
		1%	.240	.236	.260	.325	.328	.376	.328	.326	.355	.493
		10%	.258	.258	.275	.394	.406	.429	.395	.400	.420	.753
MG	$\mathscr{N}$	.5%	.555	.557	.576	.644	.656	.677	.656	.645	.675	.862
		1%	.517	.510	.548	.677	.675	.689	.688	.675	.691	.925
		10%	.469	.477	.491	.703	.706	.716	.694	.696	.711	.977
	$\mathscr{L}$	.5%	.515	.522	.546	.627	.640	.665	.655	.633	.667	.846
		1%	.486	.483	.508	.664	.663	.703	.663	.671	.684	.943
		10%	.438	.438	.458	.717	.701	.746	.701	.713	.729	.972

[13] MA, P. , MAHONEY, M.W, AND YU, B. (2015). A statistical perspective on algorithmic leveraging. *Journal of Machine Learning Research*. **16** (April): 861–911.

[14] MAHONEY, M. W. (2011). Randomized algorithms for matrices and data. *arXiv:1104.5557v3* [cs.DS]

[15] PORTNOY, S. (1984). Asymptotic Behavior of M-Estimators of p Regression Parameters when p^2/n is Large. I. Consistency. *Ann. Statist.* **12** (4): 1298–1309.

[16] PRÆSTGAARD, J. AND WELLNER, J. A. (1993). Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.*, **21** (4): 2053–2086.

[17] TROPP, J.A. (2019). Matrix Concentration & Computational Linear Algebra. https://resolver.caltech.edu/CaltechAUTHORS:20190715-125341188.

[18] VAN DE VAART AND WELLNER (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag New York,Inc. 1996.

[19] WANG, H., YANG, M. and STUFKEN, J. (2019). Information-Based Optimal Subdata Selection for Big Data Linear Regression. *J. Am. Stat. Assoc.* **114** (52): 393–405.

[20] WANG, H., ZHU, R., AND MA, P. (2015). Optimal subsampling for large sample logistic regression. *J. Am. Stat. Assoc.* **113** (522): 829–844.

Table 6: The Scoring Algorithm in Fig. 2 was used, i.e. same as Table 5 except that the sampling distributions were also approximated by replacing $(\mathbf{X}^\top\mathbf{X})^{-1}$ by $(\mathbf{X}_0^{*\top}\mathbf{X}_0^*)^{-1}$.

$\mathbf{x}$	ε	$r:n$	$\hat{\boldsymbol{\pi}}_2$	$\hat{\boldsymbol{\pi}}_1$	$\hat{\boldsymbol{\pi}}_0$	$\bar{\boldsymbol{\pi}}_2$	$\bar{\boldsymbol{\pi}}_1$	$\bar{\boldsymbol{\pi}}_0$	$\tilde{\boldsymbol{\pi}}_2$	$\tilde{\boldsymbol{\pi}}_1$	$\tilde{\boldsymbol{\pi}}_0$	Lev
Truncation 10%												
GA	$\mathscr{N}$	.5%	.834	.844	.869	.999	1.01	1.06	1.01	1.01	1.05	.986
		1%	.740	.749	.777	.965	1.01	1.03	.977	1.03	1.05	1.01
		10%	.663	.670	.709	1.03	1.02	1.04	1.02	1.01	1.04	1.02
	$\mathscr{L}$	.5%	.764	.779	.813	.976	1.01	1.02	.970	.995	1.02	.981
		1%	.680	.688	.734	.987	.978	1.02	.994	.969	1.01	.986
		10%	.604	.605	.641	.987	.982	1.05	1.01	1.02	1.03	1.02
LN	$\mathscr{N}$	.5%	.288	.280	.313	.340	.335	.364	.338	.331	.357	.485
		1%	.274	.262	.280	.358	.344	.360	.338	.337	.351	.612
		10%	.292	.285	.301	.409	.404	.424	.394	.403	.413	.858
	$\mathscr{L}$	.5%	.284	.275	.308	.357	.348	.369	.341	.335	.370	.486
		1%	.242	.242	.258	.338	.334	.359	.332	.319	.343	.588
		10%	.259	.254	.262	.394	.385	.402	.381	.387	.394	.852
MG	$\mathscr{N}$	.5%	.555	.550	.591	.619	.639	.670	.632	.630	.667	.889
		1%	.534	.539	.556	.687	.686	.712	.682	.676	.713	.979
		10%	.458	.454	.475	.668	.686	.705	.664	.688	.713	.973
	$\mathscr{L}$	.5%	.528	.537	.560	.635	.641	.702	.639	.643	.669	.919
		1%	.481	.497	.509	.657	.662	.681	.653	.659	.690	.930
		10%	.419	.420	.442	.684	.672	.713	.676	.687	.712	.980
Truncation 30%												
GA	$\mathscr{N}$	.5%	.782	.776	.800	.999	1.02	1.06	1.00	1.01	1.03	.990
		1%	.726	.735	.747	.971	1.01	1.03	.992	1.02	1.03	1.00
		10%	.676	.691	.715	1.03	1.03	1.03	1.01	1.03	1.03	1.01
	$\mathscr{L}$	.5%	.719	.716	.746	.978	1.00	1.01	.976	.996	1.01	.975
		1%	.655	.670	.711	.992	.982	1.01	.992	.978	1.01	.987
		10%	.615	.617	.642	.991	.985	1.04	1.00	1.01	1.01	1.01
LN	$\mathscr{N}$	.5%	.302	.285	.303	.357	.349	.379	.342	.337	.358	.438
		1%	.274	.276	.291	.355	.357	.382	.342	.345	.358	.544
		10%	.288	.289	.303	.414	.412	.431	.410	.404	.420	.780
	$\mathscr{L}$	.5%	.287	.279	.302	.368	.352	.383	.359	.343	.377	.438
		1%	.244	.244	.253	.346	.342	.364	.335	.331	.343	.528
		10%	.260	.250	.268	.395	.399	.407	.392	.394	.401	.773
MG	$\mathscr{N}$	.5%	.547	.547	.574	.629	.643	.676	.652	.632	.665	.866
		1%	.528	.529	.536	.677	.683	.713	.681	.677	.704	.940
		10%	.462	.455	.483	.679	.695	.689	.674	.676	.731	.936
	$\mathscr{L}$	.5%	.528	.523	.556	.634	.642	.674	.647	.639	.662	.879
		1%	.474	.471	.497	.656	.676	.672	.657	.664	.682	.887
		10%	.422	.429	.446	.675	.665	.712	.679	.683	.722	.939

Table 7: The running times (in seconds) of the Scoring Algorithm in Fig. 2 and the LSE for sample sizes n and subsample sizes r with $\mathbf{x} \sim$ GA and $\varepsilon \sim \mathscr{N}(0,1)$.

	The Scoring Algorithm						LSE
$n\backslash r$	$0.05n$	$0.10n$	$0.20n$	$0.30n$	$0.40n$	$0.50n$	n
$6*10^6$	11.807	12.576	18.671	23.276	29.296	30.050	36.344
$6*10^5$	0.882	0.981	1.502	1.896	2.266	2.784	3.809
$6*10^4$	0.116	0.134	0.161	0.175	0.173	0.201	0.234
$6*10^3$	0.012	0.013	0.017	0.018	0.030	0.029	0.027

[21] XU, P., YANG, J., ROOSTA-KHORASANI, F., RÉ, C. AND MAHONEY, M.W. (2016). Subsampled Newton Methods with Non-uniform Sampling. *arXiv:1607.00559.v2* [math.OC].

[22] ZHU, R., MA, P., MAHONEY, M.W. AND YU, B. (2015). Optimal subsampling Approaches for Large Sample Linear Regression. *arXiv:1509.0511.v1* [stat.ME].