# Asymptotic Analysis of Sampling Estimators for Randomized Numerical Linear Algebra Algorithms

**Ping Ma**
University of Georgia

**Xinlian Zhang**
University of California at San Diego

**Xin Xing**
Harvard University

**Jingyi Ma**
Central University of Finance and Economics

**Michael W. Mahoney**
ICSI and University of California at Berkeley

## Abstract

The statistical analysis of Randomized Numerical Linear Algebra (RandNLA) algorithms within the past few years has mostly focused on their performance as point estimators. However, this is insufficient for conducting statistical inference, e.g., constructing confidence intervals and hypothesis testing, since the distribution of the estimator is lacking. In this article, we develop asymptotic analysis to derive the distribution of RandNLA sampling estimators for the least-squares problem. In particular, we derive the asymptotic distribution of a general sampling estimator with arbitrary sampling probabilities. The analysis is conducted in two complementary settings, i.e., when the objective of interest is to approximate the full sample estimator or is to infer the underlying ground truth model parameters. For each setting, we show that the sampling estimator is asymptotically normally distributed under mild regularity conditions. Moreover, the sampling estimator is asymptotically unbiased in both settings. Based on our asymptotic analysis, we use two criteria, the Asymptotic Mean Squared Error (AMSE) and the Expected Asymptotic Mean Squared Error (EAMSE), to identify optimal sampling probabilities. Several of these optimal sampling probability distributions are new to the literature, e.g., the root leverage sampling estimator and the predictor length sampling estimator. Our theoretical results clarify the role of leverage in the sampling process, and our empirical results demonstrate improvements over existing methods.

Proceedings of the 23[rd]International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. 

## 1 Introduction

Recent work in Randomized Numerical Linear Algebra (RandNLA) focuses on using random sketches of the input data in order to construct approximate solutions more quickly than with traditional deterministic algorithms. In this article, we consider *statistical* aspects of recently-developed fast RandNLA algorithms for the least-squares (LS) linear regression problem. Given $\mathbf{Y} = (Y_1, \ldots, Y_n)^T \in \mathbb{R}^n$ and $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T \in \mathbb{R}^{n \times p}$, we consider the model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is the coefficient vector, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T \in \mathbb{R}^n$, where $\varepsilon_i$s are i.i.d random errors with mean 0 and variance $\sigma^2 < \infty$. We assume the sample size $n$ is large and that $\mathbf{X}$ has full column rank. The ordinary least squares (OLS) estimator of $\boldsymbol{\beta}_0$ is

$$\hat{\boldsymbol{\beta}}_{OLS} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}, \tag{2}$$

where $\|\cdot\|$ is the Euclidean norm. While the OLS estimate is optimal in several senses, the algorithmic complexity for computing it with direct methods is $O(np^2)$, which can be daunting when $n$ and/or $p$ are large.

Motivated by these algorithmic considerations, randomized sketching methods have been developed within RandNLA to achieve improved computational efficiency (Mahoney, 2011; Drineas and Mahoney, 2016; Halko et al., 2011; Woodruff et al., 2014; Mahoney and Drineas, 2016; Drineas and Mahoney, 2018). With these methods, one takes a (usually nonuniform) random sample of the full data (perhaps after preprocessing or preconditioning with a random projection

matrix (Drineas and Mahoney, 2016)), and then the sample is retained as a surrogate for the full data for subsequent computation. Here is an example of this approach for the LS problem.

**Step 1: Sampling.** Draw a random sample of size $r \ll n$ with replacement from the full data using probabilities $\{\pi_i\}_{i=1}^n$. Denote the resulting sample and probabilities as $(\mathbf{X}^*, \mathbf{Y}^*)$ and $\{\pi_i^*\}_{i=1}^r$.
**Step 2: Estimation.** Calculate the weighted LS solution, using the random sample, by solving

$$\begin{aligned}\tilde{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{\Phi}^*\mathbf{Y}^* - \boldsymbol{\Phi}^*\mathbf{X}^*\boldsymbol{\beta}\|^2 \\ &= (\mathbf{X}^{*T}\boldsymbol{\Phi}^{*2}\mathbf{X}^*)^{-1}\mathbf{X}^{*T}\boldsymbol{\Phi}^{*2}\mathbf{Y}^*\end{aligned}$$

where $\boldsymbol{\Phi}^* = \mathrm{diag}(1/\sqrt{r\pi_i^*})$.

Popular RandNLA sampling approaches include the uniform sampling estimator (UNIF), the basic leverage-based sampling estimator (BLEV), where $\pi_i^{BLEV} = h_{ii}/\sum_{i=1}^n h_{ii}$, where $h_{ii} = \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i$ are the leverage scores of $\mathbf{X}$, and the shrinkage leverage estimator (SLEV), which involves sampling probabilities $\pi_i^{SLEV} = \lambda h_{ii}/\sum_{i=1}^n h_{ii} + (1-\lambda)/n$, where $\lambda \in (0,1)$ (Drineas et al., 2006, 2008, 2012; Ma et al., 2014).

In this article, we study the *statistical* properties of these and other estimators. Substantial evidence has shown the practical effectiveness of core RandNLA methods (Ma et al., 2014, 2015; Drineas and Mahoney, 2016) (as well as other randomized approximating methods, including the Hessian sketch (Wang et al., 2017; Pilanci and Wainwright, 2016) and iterative/divide-and-conquer methods (Avron et al., 2010; Meng et al., 2014)) in providing point estimators. However, this is not sufficient for statistical analysis since the uncertainty of the estimator is lacking. In statistics, uncertainty assessment can be conducted through confidence interval construction and significance testing. It is well-known that the construction of confidence intervals and significance testing are interrelated with each other (Lehmann and Romano, 2006). Performing these two analyses is more difficult than point estimation, since it requires the distributional results of the estimator, rather than just moment conditions or concentration bounds. In the RandNLA literature, distribution results of estimators are still lacking.

There are two main challenges in studying the statistical and distributional properties of RandNLA algorithms. The first challenge is that there are two sources of randomness contributing to the statistical performance of RandNLA sampling estimators: one source is the random errors in the model, i.e., the $\varepsilon_i$s, which are typically attributed to measurement error or random noise inherited by $\mathbf{Y}$; and the other source is the randomness in the random sampling procedure within the approximation algorithm. The second challenge is that these two sources of randomness couple together within the estimator in a nontrivial way. More formally, the sampling estimator can be expressed as $\tilde{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$, where $\mathbf{W}$ is a random diagonal matrix, with the $i^{th}$ diagonal element being related to the probability of choosing the $i^{th}$ sample. The random variable used to denote the random sampling procedure, i.e., $\mathbf{W}$, is involved in the sampling estimator in a nonlinear fashion, and it pre-multiplies $\mathbf{Y}$, which contains randomness from the $\varepsilon_i$s.

We address these challenges to studying the asymptotic distribution of a general RandNLA sampling estimator for LS problems. Our results are fundamentally different from previous results on the statistical properties of RandNLA algorithms (e.g., Ma et al. (2014, 2015); Raskutti and Mahoney (2015); Chen et al. (2016); Wang et al. (2017); Dereziński et al. (2019)), in that we provide asymptotic distribution analysis, rather than finite-sample concentration inequalities. The resulting asymptotic distributions open the possibility of performing statistical inference tasks such as hypothesis testing and constructing confidence intervals, whereas finite sample concentration inequality results may not. It is worth mentioning that the results of asymptotic analysis are usually practically valid as long as the sample size is only moderately large.

**Main Results.** Our main theoretical contribution is to derive the asymptotic distribution of RandNLA estimators in two complementary settings.

**Data are a random sample.** We first consider the data as a random sample from a population, in which case the goal is to estimate the parameters of the population model. In this case, for this *unconditional inference*, we establish the asymptotic normality, i.e., deriving the asymptotic distribution, of sampling estimators for the linear model under general regularity conditions. We show that sampling estimators are asymptotically unbiased estimators with respect to the true model coefficients, and we obtain an explicit form for the asymptotic variance, for both fixed number of predictors (Theorem 1) and diverging number of predictors (Theorem 2). **Sampling Estimators.** Using these distributional results, we propose several efficient and asymptotically optimal estimators. Depending on the quantity of interest (e.g., $\boldsymbol{\beta}_0$ versus some linear function of $\boldsymbol{\beta}_0$ such as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}_0$ or $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0$), we obtain different optimal sampling probabilities (Propositions 1, 2, and 3) that lead to sampling estimators that minimize the Asymptotic Mean Squared Error (AMSE) in the respective context. None of these distributions is proportional to the leverage scores, but one (RL of Proposition 2) is constructed using the square roots of

the leverage scores, and another (PL of Proposition 3) is constructed using the row norms of predictor matrix.

**Data are given and fixed.** We then consider the data as given/fixed, in which case the goal is to approximate the full sample OLS estimate. In this case, for this *conditional inference*, we establish the asymptotic normality, i.e., deriving the asymptotic distribution, of sampling estimators for the linear model under general regularity conditions. We show that sampling estimators are asymptotically unbiased with respect to the OLS estimate, and we obtain an explicit form of the asymptotic variance and the Expected Asymptotic Mean Squared Error (EAMSE) of sampling estimators (Theorem 3). **Sampling Estimators**. Using these results, we construct sampling probability distributions that lead to sampling estimators that minimize the EAMSE. Depending on the quantity of interest (here, $\hat{\boldsymbol{\beta}}_{OLS}$ versus some linear function of $\hat{\boldsymbol{\beta}}_{OLS}$ such as $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$ or $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$, we obtain different optimal sampling probabilities (Propositions 4, 5, and 6).

**Related Work.** There is a large body of related work in RandNLA (Mahoney, 2011; Drineas and Mahoney, 2016; Halko et al., 2011; Woodruff et al., 2014; Mahoney and Drineas, 2016; Drineas and Mahoney, 2018). However, very little of this work addresses statistical aspects of the methods. Recently, significant progress has been made in the study of the statistical properties of RandNLA sampling estimators (Ma et al., 2014, 2015; Raskutti and Mahoney, 2015; Chen et al., 2016; Wang et al., 2017; Dereziński et al., 2019). The work most related to ours is that of Ma et al. (2014, 2015), who employed a Taylor series expansion up to a linear term to study the MSE of RandNLA sampling estimators. Ma et al. (2014, 2015) failed to characterize the detailed convergence performance of the remainder term. They concluded that neither leverage-based sampling (BLEV) nor uniform sampling (UNIF) dominates the other in terms of variance; and they proposed and demonstrated the superiority of the SLEV sampling method. To find the sampling distribution of estimators, leading to statistically-better RandNLA sampling estimators, it is important to examine the convergence properties of the remainder term. To accomplish this, we consider the asymptotic distribution of the sampling estimator. Such asymptotic analysis is common in statistics, and it can substantially simplify the derivation of complicated random variables, leading to simpler analytic expressions (Le Cam, 1986).

Chen et al. (2016) proposed optimal estimators minimizing the variance that accounts for the randomness of sampling and model error. Our results and those of Chen et al. (2016) have similar goals, but they are different. First, Chen et al. (2016) used bias and variance, while we use AMSE and EAMSE. Second, we consider the asymptotic distribution of the sampling estimators, going beyond just the bias and variance of Chen et al. (2016). Thus, our results could be used for downstream statistical inferences, e.g., constructing confidence intervals and hypothesis testing, while those of Chen et al. (2016) could not. Third, the exact expression of optimal sampling probabilities in Chen et al. (2016) depends on the unknown true parameter of the model, $\beta_0$ and $\sigma^2$ (Eqn (4) in Chen et al. (2016)), while our optimal sampling probabilities (see Section 2) are readily computed from the data. Fourth, Chen et al. (2016) only studied properties of sampling estimators for estimating true model parameters, while we consider both estimating the true parameter and approximating the full sample estimate.

Wang et al. (2017) proposed an approximated A-optimality criterion, which is based on the conditional variance of the sampling estimator given a subsample. Since the randomness of sampling is not considered in the criterion, they obtained a simple analytic expressions of the optimal results. Dereziński et al. (2019) also consider experimental design from the RandNLA perspective, and they propose a framework for experimental design where the responses are produced by an arbitrary unknown distribution. Their main result yields nearly tight bounds for the classical A-optimality criterion, as well as improved bounds for worst-case responses. In addition, they propose a minimax-optimality criterion (which can be viewed as an extension of both A-optimal design and RandNLA sampling for worst-case regression). Related works on the asymptotic properties of subsampling estimators in logistic regression can be found in Wang et al. (2018) and Wang (2019).

**Technical Report.** A longer and much more detailed version of this short conference publication, with additional results, proofs of our main results, and additional discussion, is available (Ma et al., 2020)

## 2 Asymptotic analysis of RandNLA

### 2.1 Unconditional Inference: Estimating Model Parameters

For Model (1), from the traditional statistical perspective of using the data to perform inference, one major goal is to estimate the underlying true model parameters, i.e., $\boldsymbol{\beta}_0$. We refer to this as *unconditional inference*. For unconditional inference, both randomness in the data and randomness in the algorithm contribute to randomness in the RandNLA sampling estimators.

**Theorem 1** (**Unconditional inference, fixed** $p$). *Assume the number of predictors $p$ is fixed and the following regularity conditions hold.*

*(A1)[Data condition]. There exist positive constants $b$ and $B$ such that $b \leq \lambda_{min} \leq \lambda_{max} \leq B$, where $\lambda_{max}$ and $\lambda_{min}$ are the maximum and minimum eigenvalues of matrix $\mathbf{X}^T\mathbf{X}/n$, respectively.*
*(A2)[Sampling condition]. The sample size $r = O(n^{1-\alpha})$, where $0 \leq \alpha < 1$ and where the minimum sampling probability $\pi_{min} = O(n^{-\gamma_0})$, where $\gamma_0 \geq 1$ satisfy $\gamma_0 + \alpha < 2$.*
*Under these assumptions, as $n \to \infty$, we have*

$$(\sigma^2\mathbf{\Sigma}_0)^{-\frac{1}{2}}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} \boldsymbol{N}(\mathbf{0}, \mathbf{I}_p)$$

*where $\mathbf{\Sigma}_0 = (\mathbf{X}^T\mathbf{X})^{-1}\left(\mathbf{X}^T(\mathbf{I}_n + \mathbf{\Omega})\mathbf{X}\right)(\mathbf{X}^T\mathbf{X})^{-1}$, $\mathbf{\Omega} = diag\{1/r\pi_i\}_{i=1}^n$, and $\mathbf{I}_p$ is the $p \times p$ identity matrix. Thus, in this unconditional inference, the asymptotic mean of $\tilde{\boldsymbol{\beta}}$ is $AE(\tilde{\boldsymbol{\beta}}) = \boldsymbol{\beta}_0$, i.e., $\tilde{\boldsymbol{\beta}}$ is an asymptotically unbiased estimator of $\boldsymbol{\beta}_0$, and the asymptotic variance of $\tilde{\boldsymbol{\beta}}$ is $AVar(\tilde{\boldsymbol{\beta}}) = \sigma^2\mathbf{\Sigma}_0$.*

Condition (A1) in Theorem 1 indicates that $\mathbf{X}^T\mathbf{X}/n$ is positive definite. This condition requires the predictor matrix $\mathbf{X}$ to be of full column rank and that the values of elements in $\mathbf{X}$ are not over-dispersed. This condition ensures the consistency of full sample OLS estimator (Lai et al., 1978). Condition (A2) in Theorem 1, which can be rewritten as $n^{-\gamma_0} > n^{-(2-\alpha)}$, provides a lower bound on the *smallest* sampling probability. Bounding sampling probabilities from below mitigates the inflation of the variance $\mathbf{\Sigma}_0$, which is proportional to the reciprocal sampling probability. The importance of this condition for establishing *statistical* properties of RandNLA algorithms was highlighted by Ma et al. (2014, 2015). Condition (A2) can also be rewritten as $n^{1-\alpha}n^{-\gamma_0} > n^{-1}$, which states that when the smallest sampling probability is small, one compensates by making the sample size large.

In Theorem 1, the asymptotic variance $AVar(\tilde{\boldsymbol{\beta}})$ can be written as

$$AVar(\tilde{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} + \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Omega}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1},$$

where the first term is the variance of the full sample OLS, and the second term is the variation related to the sampling process. The second term of $AVar(\tilde{\boldsymbol{\beta}})$ has a "sandwich-type" expression. The center term, $\mathbf{\Omega}$, depends on the reciprocal sampling probabilities, suggesting that extremely small probabilities will result in large asymptotic variance and large AMSE of the corresponding estimator. This was observed previously by Ma et al. (2015).

**Remark.** In light of efficient estimation methods such as iterative Hessian sketch and dual random projection, we emphasize that besides estimation, our distribution results can be used for performing additional inference analysis, e.g., constructing a confidence interval and conducting hypothesis testing. These inference analyses cannot be achieved by other iterative methods as far as we know.

Given Theorem 1, it is natural to ask whether there is an optimal estimator, i.e., one with the smallest AMSE for estimating $\boldsymbol{\beta}_0$. Using the asymptotic results in Theorem 1, we propose the following three estimators.

**Estimating $\boldsymbol{\beta}_0$.** By Theorem 1, we could express the $AMSE(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}_0)$ as a function of $\{\pi_i\}_{i=1}^n$. Then, it is straightforward to employ the method of Lagrange multipliers to find the minimizer subject to the constraint $\sum_{i=1}^n \pi_i = 1$. The minimizer is then the optimal sampling probabilities for estimating $\boldsymbol{\beta}_0$.

**Proposition 1.** *The inverse-covariance (IC) sampling estimator, with the sampling probabilities*

$$\pi_i = \frac{\|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}{\sum_{i=1}^n \|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}, \quad i = 1, \ldots, n, \tag{3}$$

*has the smallest $AMSE(\tilde{\boldsymbol{\beta}}; \boldsymbol{\beta}_0) = \sigma^2 tr\{(\mathbf{X}^T\mathbf{X})^{-1}\} + \frac{1}{r}\sum_{i=1}^n \frac{\sigma^2}{\pi_i}||(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i||^2$.*

**Remark.** The implication of this optimal estimator is two-fold. On the one hand, as defined, the proposed IC estimator has the smallest AMSE. On the other hand, if given the same tolerance of uncertainty, i.e., to achieve a certain small standard error, the IC estimator requires the smallest sample size.

**Estimating linear functions of $\boldsymbol{\beta}_0$.** In addition to making inference on $\boldsymbol{\beta}_0$, one may also be interested in linear functions of $\boldsymbol{\beta}_0$, i.e., $\mathbf{L}\boldsymbol{\beta}_0$, where $\mathbf{L}$ is any constant matrix of suitable dimension. Here, we present results for $\mathbf{X}\boldsymbol{\beta}_0$ and $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0$.

**Proposition 2.** *The root leverage (RL) sampling estimator, with the sampling probabilities*

$$\pi_i = \frac{\|\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}{\sum_{i=1}^n \|\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|} = \frac{\sqrt{h_{ii}}}{\sum_{i=1}^n \sqrt{h_{ii}}}, \tag{4}$$

*for $i = 1, .., n$, has the smallest $AMSE(\mathbf{X}\tilde{\boldsymbol{\beta}}; \mathbf{X}\boldsymbol{\beta}_0) = p\sigma^2 + \frac{1}{r}\sum_{i=1}^n \frac{\sigma^2}{\pi_i}||\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i||^2$.*

The probabilities in RL are a nonlinear transformation of the probabilities in BLEV. Comparing to the BLEV estimator, the RL estimator shrinks the large probabilities and pulls up the small probabilities. Thus we expect RL to provide an estimator with smaller variances in a way similar to SLEV.

**Remark.** Chen et al. (2016) proposed an optimal sampling estimators for estimating $\boldsymbol{\beta}_0$ and predicting $\mathbf{Y}$. Their sampling probabilities depend on the unknown parameters, and they proposed the probabilities in (4) as a rough approximation of their proposed probabilities without demonstration.

**Proposition 3.** *The predictor-length (PL) sampling estimator, with the sampling probabilities*

$$\pi_i = \frac{\|\mathbf{x}_i\|}{\sum_{i=1}^n \|\mathbf{x}_i\|}, \quad i = 1, \ldots, n, \tag{5}$$

*has the smallest* $AMSE(\mathbf{X}^T\mathbf{X}\tilde{\boldsymbol{\beta}}; \mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0) = \sigma^2 tr(\mathbf{X}^T\mathbf{X}) + \frac{\sigma^2}{r}\sum_{i=1}^n \frac{1}{\pi_i}||\mathbf{x}_i||^2$.

**Remark.** All these proposed metrics can be computed in the time it takes to approximate leverage scores (or faster, for PL), i.e., the time to implement a random projection, since they are essentially strongly related to leverage scores. In particular, all metrics involving leverages scores can be approximated using the algorithm of Drineas et al. (2012). Detailed implementations/comments may be found in Ma et al. (2014, 2015), as well as in the Blendenpik/LSRN papers.

**Diverging number of predictors, $p \to \infty$.** Theorem 1 considers the number of predictors/features, $p$, as fixed. It is also of interest to study the asymptotic properties of RandNLA estimators in the scenario that $p$ diverges with $n \to \infty$ (at a suitable rate relative to $n$). The following theorem states our results concerning this case. Observe that, in the case of a divergent $p$, the $(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ is of divergent dimension. Thus, we characterize its asymptotic distribution via the scalar $\boldsymbol{a}^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$, where $\boldsymbol{a}$ is an arbitrary bounded-norm vector.

**Theorem 2** (**Unconditional inference, diverging** $p$)**.** *In addition to condition (A1) in Theorem 1, assume the following regularity conditions hold.*
*(B1)[Data condition]. The number of predictors $p$ diverges at a rate $p = n^{1-\kappa}$, $0 < \kappa < 1$; and $\frac{\max_i \|\mathbf{x}_i\|^2}{n} = O(\frac{p}{n})$, where $\mathbf{x}_i$ is the $i^{th}$ row of $\mathbf{X}$.*
*(B2)[Sampling condition]: The parameters $\alpha$, $\gamma_0$, and $\kappa$ satisfy $\alpha + \gamma_0 - \kappa < 1$.*
*Under these assumptions, as $n \to \infty$, we have*

$$(\sigma^2 \boldsymbol{a}^T \boldsymbol{\Sigma}_0 \boldsymbol{a})^{-\frac{1}{2}} \boldsymbol{a}^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} \boldsymbol{N}(0, 1),$$

*where $\boldsymbol{a} \in \mathbb{R}^p$ satisfies $\|\boldsymbol{a}\|^2 < \infty$.*

Condition (B2) is more stringent than Condition (A2), and this is required for accommodating a divergent $p$. It is easy to verify that the sampling estimators in Propositions 1, 2, and 3 are still the optimal sampling estimators for their respective purposes, thus we omit restating the results.

### 2.2 Conditional Inference: Approximating the Full Sample OLS Estimate

For Model (1), a second goal is to approximate the full sample calculations, say the $\hat{\boldsymbol{\beta}}_{OLS}$, regardless of the underlying true model parameter $\boldsymbol{\beta}_0$. We refer to this as *conditional inference*. For conditional inference, we consider the full sample as given, and thus the only source of randomness contributing to the RandNLA sampling estimators is the randomness in the sampling algorithm. The following theorem states that, in conditional inference, the asymptotic distribution of the sampling estimator $\tilde{\boldsymbol{\beta}}$ is a normal distribution (with mean $\boldsymbol{\beta}_{OLS}$ and variance $\sigma^2\boldsymbol{\Sigma}_c$).

**Theorem 3** (**Conditional inference**)**.** *Assume the following regularity conditions hold.*
*(C1)[Data condition]. The full data $\{\mathbf{X}, \mathbf{Y}\}$, i.e., $n$ and $p$ are considered fixed; $\mathbf{X}$ is of full column rank, and $\|\mathbf{x}_i\| < \infty$, for $i = 1, \ldots, n$, where $\mathbf{x}_i$ is the $i^{th}$ row of $\mathbf{X}$.*
*(C2)[Sampling condition]. The sampling probabilities $\{\pi_i\}_{i=1}^n$ are nonzero.*
*Under these assumptions, as $r \to \infty$, we have*

$$(\sigma^2\boldsymbol{\Sigma}_c)^{-\frac{1}{2}}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{OLS}) \xrightarrow{d} \boldsymbol{N}(\mathbf{0}, \mathbf{I}_p)\,,$$

*where $\boldsymbol{\Sigma}_c = \frac{1}{r}(\mathbf{X}^T\mathbf{X})^{-1}\left(\sum_{i=1}^n \frac{e_i^2}{\pi_i}\mathbf{x}_i\mathbf{x}_i^T\right)(\mathbf{X}^T\mathbf{X})^{-1}$, $e_i = Y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_{OLS}$, and $\mathbf{I}_p$ is the $p \times p$ identity. Thus, for conditional inference, the asymptotic mean of $\tilde{\boldsymbol{\beta}}$ is $AE(\tilde{\boldsymbol{\beta}}) = \hat{\boldsymbol{\beta}}_{OLS}$, i.e., $\tilde{\boldsymbol{\beta}}$ is an asymptotically unbiased estimator of $\boldsymbol{\beta}_{OLS}$, and the asymptotic variance of $\tilde{\boldsymbol{\beta}}$ is $AVar(\tilde{\boldsymbol{\beta}}) = \sigma^2\boldsymbol{\Sigma}_c$.*

Theorem 3 shows that as the sample size $r$ gets larger, the distribution of $\tilde{\boldsymbol{\beta}}$ is well-approximated by a normal distribution, with mean $\hat{\boldsymbol{\beta}}_{OLS}$ and variance $\sigma^2\boldsymbol{\Sigma}_c$. Similar to unconditional inference, the asymptotic variance $AVar(\tilde{\boldsymbol{\beta}})$ here also has "sandwich-type" expression, where the center term (here, $\left(\sum_{i=1}^n \frac{e_i^2}{\pi_i}\mathbf{x}_i\mathbf{x}_i^T\right)$) depends on the reciprocal sampling probabilities. Thus, we also expect that extremely small probabilities will result in large variances of the corresponding estimators.

In Theorem 3, $AVar(\tilde{\boldsymbol{\beta}})$ depends on the full sample least square residuals, i.e., the $e_i$s. These are not readily available from the sample. To solve this problem and to obtain meaningful results, we take the expectation of the $e_i^2$s. The metric we use is thus the EAMSE,

$$EAMSE(\tilde{\boldsymbol{\beta}}; \hat{\boldsymbol{\beta}}_{OLS}) = E_{\mathbf{Y}}(AMSE(\tilde{\boldsymbol{\beta}}; \hat{\boldsymbol{\beta}}_{OLS})).$$

For this result, we denote that $\mathrm{E}_{\mathbf{Y}}(e_i^2) = (1 - h_{ii})\sigma^2$.

**Approximating $\hat{\boldsymbol{\beta}}_{OLS}$.** By Theorem 3, we could express $EAMSE(\tilde{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}_{OLS})$ as a function of $\{\pi_i\}_{i=1}^n$. It is straightforward to find the minimizer of the $EAMSE(\tilde{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}_{OLS})$ subject to $\sum_{i=1}^n \pi_i = 1$. The minimizer is then the optimal sampling probabilities for approximating $\hat{\boldsymbol{\beta}}_{OLS}$.

**Proposition 4.** *The inverse-covariance negative-leverage (ICNLEV) sample estimator, with the sam-*

*pling probabilities*

$$\pi_i = \frac{\sqrt{1-h_{ii}}\|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}{\sum_{i=1}^n \sqrt{1-h_{ii}}\|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}, i = 1, \ldots, n, \quad (6)$$

*has the smallest* $EAMSE(\tilde{\boldsymbol{\beta}}; \hat{\boldsymbol{\beta}}_{OLS}) = E_{\boldsymbol{Y}}(tr(AVar(\tilde{\boldsymbol{\beta}}))) = \frac{1}{r}\sum_{i=1}^n \frac{(1-h_{ii})\sigma^2}{\pi_i}||(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i||^2$.

**Approximating linear functions of $\hat{\boldsymbol{\beta}}_{OLS}$.** In addition to approximating $\hat{\boldsymbol{\beta}}_{OLS}$, one may also be interested in linear functions of $\hat{\boldsymbol{\beta}}_{OLS}$, i.e., $\mathbf{L}\boldsymbol{\beta}_0$, where $\mathbf{L}$ is any constant matrix of suitable dimension. Here, we present results for $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$ and $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$.

**Proposition 5.** *The root leveraging negative-leverage (RLNLEV) sample estimator, with the sampling probabilities*

$$\begin{aligned}\pi_i &= \frac{\sqrt{1-h_{ii}}\|\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}{\sum_{i=1}^n \sqrt{1-h_{ii}}\|\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|} \\ &= \frac{\sqrt{(1-h_{ii})h_{ii}}}{\sum_{i=1}^n \sqrt{(1-h_{ii})h_{ii}}}, \quad i = 1, \ldots, n, \quad (7)\end{aligned}$$

*has the smallest* $EAMSE(\mathbf{X}\tilde{\boldsymbol{\beta}}; \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}) = \frac{1}{r}\sum_{i=1}^n \frac{(1-h_{ii})\sigma^2}{\pi_i}||\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i||^2$.

**Proposition 6.** *The predictor-length negative-leverage (PLNLEV) sampling estimator, with the sampling probabilities*

$$\pi_i = \frac{\sqrt{1-h_{ii}}\|\mathbf{x}_i\|}{\sum_{i=1}^n \sqrt{1-h_{ii}}\|\mathbf{x}_i\|}, i = 1, \ldots, n, \quad (8)$$

*has the smallest* $EAMSE(\mathbf{X}^T\mathbf{X}\tilde{\boldsymbol{\beta}}; \mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}) = \frac{1}{r}\sum_{i=1}^n \frac{(1-h_{ii})\sigma^2}{\pi_i}||\mathbf{x}_i||^2$.

**Remark.** All these proposed metrics can be computed in the time it takes to approximate leverage scores, i.e., the time to implement a random projection, using the algorithm of Drineas et al. (2012), since they are essentially strongly related to leverage scores.

### 2.3 Relationship of the Sampling Estimators

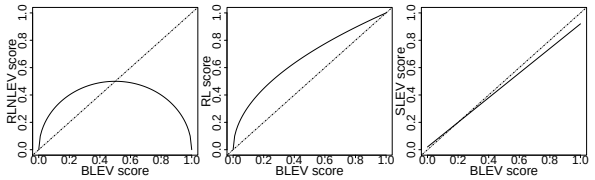
We use simple examples to show the relationship among various sampling probabilities, each of which is optimal with respect to a different statistical criterion.

**Example 1: "Shrinkage" Properties of Proposed Estimators.** We illustrate the "shrinkage" property of proposed optimal sampling probabilities compared to the BLEV sampling probabilities. For convenience, we refer to the numerators of the sampling probabilities in a sampling estimator as the scores, e.g., the RL score is $\sqrt{h_{ii}}$ and the RLNLEV score is $\sqrt{(1-h_{ii})h_{ii}}$. In Figure 1a, we plot RL score, RLNLEV score, and SLEV score ($0.9h_{ii} + 0.1p/n$ with $p/n = 0.2$) as functions of leverage score $h_{ii}$ (i.e., BLEV score in Figure 1(a)). Observe that the RLNLEV score amplifies small $h_{ii}$s but shrinks large $h_{ii}$s. Both RLNLEV and RL scores provide nonlinear shrinkage of the BLEV. The SLEV scores also shrink large $h_{ii}$s and amplify small $h_{ii}$s, but in a linear fashion. The advantage of such "shrinkage" is two-fold. On the one hand, the data with high leverage scores could be "outliers." Shrinking the sampling probabilities of high leverage data points reduces the risk of selecting outliers into the sample. On the other hand, amplifying the sampling probabilities of low leverage data points reduces the variance of the resulting sampling estimators.

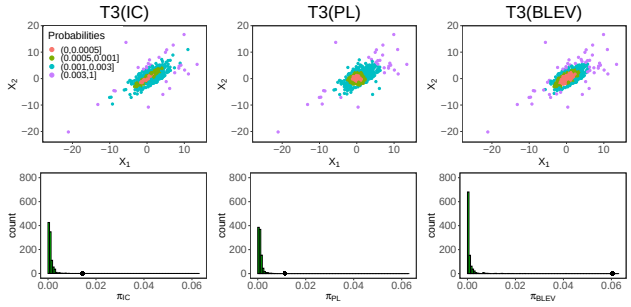
**Example 2: The Role of $h_{ii}$s.** On the one hand, if the $h_{ii}$s are homogeneous, then the sampling probabilities of the ICNLEV estimator ($\frac{\sqrt{1-h_{ii}}\|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}{\sum_{i=1}^n \sqrt{1-h_{ii}}\|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}$) and those of the IC estimator ($\frac{\|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}{\sum_{i=1}^n \|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}$) will be similar to each other. On the other hand, since $\sum_{i=1}^n h_{ii} = p$, given a fixed value of $p$, we expect that $h_{ii}$s are small when sample size $n$ is large. When $h_{ii} = o(1)$ for all $i = 1, \ldots, n$, i.e., $h_{ii}$s are extremely small compared to 1, the sampling probabilities of the ICNLEV estimator and those of the IC estimator will also be similar. Analogous arguments also apply to PLNLEV and PL.

**Example 3: Orthogonal predictor matrix, i.e., $\mathbf{X}^T\mathbf{X} = \mathbf{I}$.** In this case, $h_{ii} = \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i = \|\mathbf{x}_i\|^2$, and the ICNLEV score, RLNLEV score, and PLNLEV score are the same and equal $\sqrt{(1-h_{ii})h_{ii}}$. Analogously, the IC score coincides with the RL score and the PL score, and all equal $\|\mathbf{x}_i\|$.

**Example 4: A two dimensional example**. We generated 1000 data points for two predictors from a multivariate normal distribution, a multivariate noncentral $t$ distribution with three degrees of freedom, and a multivariate noncentral $t$ distribution with one degree of freedom. In Figure 1(b), we present scatterplots of these data points. In each scatterplot, the color of points indicates the magnitude of sampling probabilities in IC, PL and BLEV methods. Below each scatterplot, we also present histograms of the corresponding sampling probabilities. Examination of Figure 1(b) reveals one pattern shared by all sampling distributions, i.e., the sampling probabilities of data points in the center are smaller than those of data points at the boundary. In addition, note that, compared to $\pi_i^{PL} \propto \|\mathbf{x}_i\|$, both $\pi_i^{IC} \propto \|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|$ and $\pi_i^{BLEV} \propto \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i$ depends on $(\mathbf{X}^T\mathbf{X})^{-1}$, which normalizes the scale of predictors. Thus, we notice that data points with high probabilities in PL scatter around the upper right and lower left corner. However, the data points with high probabilities in IC and BLEV form a contour toward the exterior of the data




(a) Example 2 in Section 2.3.



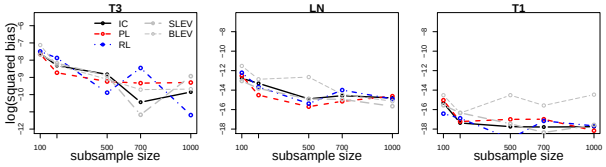

(b) Example 4 in Section 2.3.

Figure 1: (a) Relationship between RLNLEV, RL, SLEV and BLEV scores. (b) The scatterplots of data points generated from a bivariate $t$ distribution with 3 degrees of freedom, with colors coding the probabilities in IC, PL, and BLEV, and the corresponding histogram of sampling probabilities (bottom row). The dot on $x$-axis of histograms indicate the position of maximum.

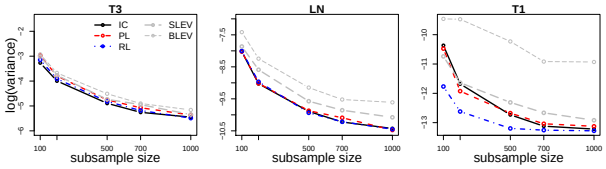
cloud. This difference is caused by the effect of the normalization using $(\mathbf{X}^T\mathbf{X})^{-1}$. The histograms in each row also show the key difference between the sampling probabilities of BLEV and those of IC and PL, i.e., the sampling probability distribution of BLEV is more dispersed than others. In other words, there are a significant number of data points with either extremely large or extremely small probabilities in BLEV.

## 3 Empirical Results

We generated synthetic data from Model (1) with $p = 10$, $n = 5000$, and random error $\epsilon_i \overset{iid}{\sim} N(0,1)$. We set the first and last two entries of $\boldsymbol{\beta}_0$ to be 1 and the rest to be 0.1. We generated the predictors from the following distributions. (1) Multivariate noncentral $t$-distribution $t_3(\mathbf{1}, \mathbf{D})$, where $\mathbf{1}$ is a vector of 1, and the $(i,j)^{th}$ element of $\mathbf{D}$ is set to $2 \times 0.7^{|i-j|}$ for $i, j = 1, \ldots, p$. We refer to this as T3 data. (2) Multivariate noncentral $t$-distribution $t_1(\mathbf{1}, \mathbf{D})$. We refer to this as T1 data. (3) Log-normal distribution $\mathbf{LN}(\mathbf{1}, \mathbf{D})$. We refer to this as LN data. For $t_1(\mathbf{1}, \mathbf{D})$, the expectation and variance do not exist. This violates Condition (A1) in Theorem 1. Thus, asymptotic squared bias and asymptotic variance of proposed estimators might not converge fast to 0 as $r$ increases.




(a) Estimating $\boldsymbol{\beta}_0$: squared biases.



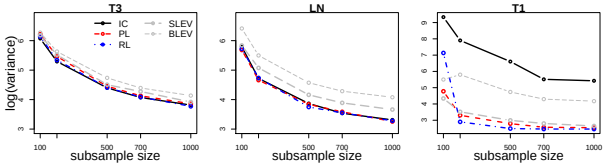

(b) Estimating $\boldsymbol{\beta}_0$: variances.

Figure 2: IC, RL, PL, SLEV, and BLEV for estimating $\boldsymbol{\beta}_0$.

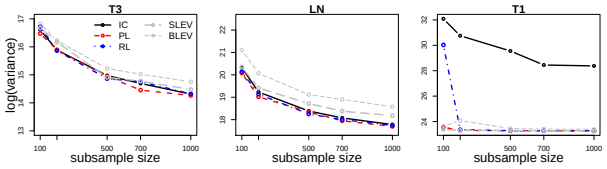
### 3.1 Estimating Model Parameters

Here, we evaluate the performance of proposed sampling estimators in estimating $\boldsymbol{\beta}_0$. We generated 100 replicates of T3, LN, and T1 data, applied IC, RL, PL, SLEV(with $\lambda = 0.9$ here and after), and BLEV to each replicated dataset to obtain $\tilde{\boldsymbol{\beta}}_b$ at sample sizes $r = 100, 200, 500, 700, 1000$, and calculated squared bias and variance with respect to $\boldsymbol{\beta}_0$ for each method.

In Figure 2, we plot the squared biases (in (a)) and the variances (in (b)) of all the methods in estimating $\boldsymbol{\beta}_0$. First, both the squared biases and the variances show decreasing patterns as $r$ increases, and the squared biases are much smaller than the corresponding variances, consistent with Theorem 1 stating that the RandNLA estimators are asymptotically unbiased and consistent estimators of $\boldsymbol{\beta}_0$. Second, the variances of estimates using IC, whose sampling probabilities minimize $AMSE(\tilde{\boldsymbol{\beta}}; \boldsymbol{\beta}_0)$, among other proposed estimators in this paper, are smaller than the variances of BLEV and SLEV when subsample sizes are greater than 200 in T3 and LN. The larger variances of BLEV estimates are due to the existence of extremely small sampling probabilities in BLEV. Taking a weighted average of the sampling probability distribution of BLEV and that of UNIF show a beneficial effect on the variances for SLEV estimators. However, the variances of SLEV estimators are still larger than those of all other proposed estimators, as the sample size $r$ increases in LN and T3. Third, the squared biases and variances of all estimates are smaller in LN and T1 compared to T3. Fourth, despite the violation of the regularity condition in T1, proposed estimators still show good performance and have consistently smaller variances than BLEV at larger $r$s. For estimating $\mathbf{Y}$ and $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0$, the biases of all sampling estimators are very similar to each other

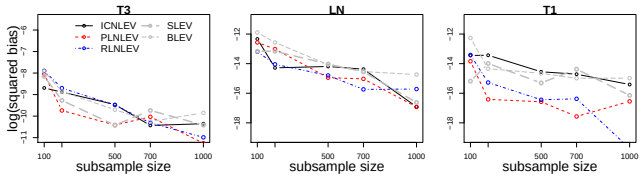



(a) Estimating $\mathbf{Y}$: variances.

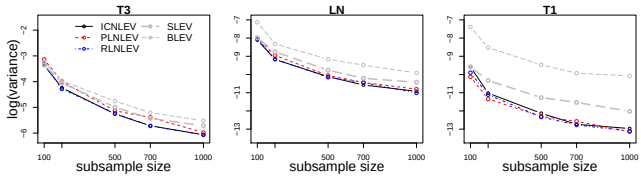



(b) Estimating $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0$ : variances.

Figure 3: IC, RL, PL, SLEV, and BLEV for estimating $\mathbf{Y}$ and $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0$.




(a) Approximating $\hat{\boldsymbol{\beta}}_{OLS}$: squared biases.




(b) Approximating $\hat{\boldsymbol{\beta}}_{OLS}$: variances.

Figure 4: ICNLEV, RLNLEV, PLNLEV, SLEV, and BLEV for approximating $\hat{\boldsymbol{\beta}}_{OLS}$.

and are much smaller than the corresponding variances. This observation is consistent with what we observed in estimating $\boldsymbol{\beta}_0$. We thus only present the variances of different estimators in Figure 3. As shown, the variances of estimates using PL, IC, and RL are smaller than the variances of estimates using BLEV and SLEV in T3 and LN at most sample sizes for estimating both $\mathbf{Y}$ and $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0$.

### 3.2 Approximating the Full Sample OLS Estimate

Here, we evaluate the performance of the RandNLA sampling estimators for conditional inference. We generated one T3 dataset, one LN dataset, and one T1 dataset. For each dataset, the full sample OLS estimates were calculated. Then, we set samples sizes at $r = 100, 200, 500, 700, 1000$ and repeatedly applied ICNLEV, RLNLEV, PLNLEV, SLEV and BLEV methods 100 times at each $r$ to get $\tilde{\boldsymbol{\beta}}_b$, $b = 1, \ldots, 100$. Using these estimates, we calculated the squared bias and variance of each method. In Figure 4, we plot the squared biases (in (a)) and the variances (in (b)) of ICNLEV, RLNLEV, PLNLEV, SLEV, and BLEV estimates for approximating $\hat{\boldsymbol{\beta}}_{OLS}$ at different sample sizes in all datasets. Several observations are worth noting in Figure 4. First, the squared biases are negligible compared to the corresponding variances. For all methods, both squared biases and variances decrease as sample size $r$ increases in all datasets. These observations corroborates Theorem 3, which states that sampling estimators are asymptotically unbiased and consistent approximations to the $\hat{\boldsymbol{\beta}}_{OLS}$. Second, the variances of estimates using ICNLEV, whose probabilities minimize $EAMSE(\tilde{\boldsymbol{\beta}}; \hat{\boldsymbol{\beta}}_{OLS})$, and other proposed estimators, get smaller than the variances of estimates using SLEV and BLEV, as sample size $r$ increases. Third, for all estimators the squared biases and variances of all estimates in LN and T1 are smaller than in T3. Fourth, despite the violation of the regularity condition in T1, proposed estimators still have consistently smaller variances than BLEV and SLEV.

## 4 Conclusion

We have studied the asymptotic properties of RandNLA sampling estimators in linear regression models. We showed that under certain regularity conditions on the sampling probability distributions, the sampling estimators are asymptotically normally distributed. Moreover, the sampling estimators are asymptotically unbiased for approximating the full sample OLS estimate and for estimating true coefficients. Based on asymptotic results, we proposed optimality criteria to assess the performance of the sampling estimators, based on AMSE and EAMSE. In particular, we developed six sampling estimators, i.e., IC, RLEV, PL, ICNLEV, RLNLEV, and PLNLEV, for minimizing AMSE and EAMSE under a variety of settings. Empirical studies demonstrated that these new sampling estimators outperform the conventional ones in the literature. For generalization, depending on the application, one may consider other criteria than AMSE and EAMSE. For example, when hypothesis testing problems are of primary interest, the power of the test is a more reasonable choice to serve as a criterion. Developing scalable sampling methods to optimize criteria such as this are of interest.

**Acknowledgement.** PM, XZ, and XX acknowledge NSF and NIH for providing partial support of this work. MWM acknowledges ARO, DARPA, NSF, and ONR for providing partial support of this work.

## References


Avron, H., P. Maymounkov, and S. Toledo (2010). Blendenpik: Supercharging LAPACK's least-squares solver. *SIAM Journal on Scientific Computing 32*, 1217–1236.

Chen, S., R. Varma, A. Singh, and J. Kovačcević (2016). A statistical perspective of sampling scores for linear regression. In *Information Theory (ISIT), 2016 IEEE International Symposium*, pp. 1556–1560. IEEE.

Dereziński, M., K. L. Clarkson, M. W. Mahoney, and M. K. Warmuth (2019). Minimax experimental design: Bridging the gap between statistical and worst-case approaches to least squares regression. Technical report. Preprint: arXiv:1902.00995.

Drineas, P., M. Magdon-Ismail, M. W. Mahoney, and D. P. Woodruff (2012). Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research 13*, 3475–3506.

Drineas, P. and M. W. Mahoney (2016). RandNLA: Randomized Numerical Linear Algebra. *Communications of the ACM 59*(6), 80–90.

Drineas, P. and M. W. Mahoney (2018). Lectures on randomized numerical linear algebra. In M. W. Mahoney, J. C. Duchi, and A. C. Gilbert (Eds.), *The Mathematics of Data*, IAS/Park City Mathematics Series, pp. 1–48. AMS/IAS/SIAM.

Drineas, P., M. W. Mahoney, and S. Muthukrishnan (2006). Sampling algorithms for $l_2$ regression and applications. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1127–1136.

Drineas, P., M. W. Mahoney, and S. Muthukrishnan (2008). Relative-error CUR matrix decomposition. *SIAM Journal on Matrix Analysis and Applications 30*, 844–881.

Halko, N., P.-G. Martinsson, and J. A. Tropp (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review 53*(2), 217–288.

Lai, T. L., H. Robbins, and C. Z. Wei (1978). Strong consistency of least squares estimates in multiple regression. *Proceedings of the National Academy of Sciences 75*(7), 3034–3036.

Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag.

Lehmann, E. L. and J. P. Romano (2006). *Testing Statistical Hypotheses*. Springer Science & Business Media.

Ma, P., M. Mahoney, and B. Yu (2014). A statistical perspective on algorithmic leveraging. In *Proceedings of the 31th ICML Conference*, pp. 91–99.

Ma, P., M. Mahoney, and B. Yu (2015). A statistical perspective on algorithmic leveraging. *Journal of Machine Learning Research 16*, 861–911.

Ma, P., X. Zhang, X. Xing, J. Ma, and M. W. Mahoney (2020). Asymptotic analysis of sampling estimators for randomized numerical linear algebra algorithms. Technical report. Preprint: arXiv:2002.10526.

Mahoney, M. (2011). *Randomized Algorithms for Matrices and Data*. Foundations and Trends in Machine Learning. Boston: NOW Publishers. Also available at: arXiv:1104.5557.

Mahoney, M. W. and P. Drineas (2016). Structural properties underlying high-quality randomized numerical linear algebra algorithms. In P. Bühlmann, P. Drineas, M. Kane, and M. van de Laan (Eds.), *Handbook of Big Data*, pp. 137–154. CRC Press.

Meng, X., M. A. Saunders, and M. W. Mahoney (2014). LSRN: A parallel iterative solver for strongly over- or under-determined systems. *SIAM Journal on Scientific Computing 36*(2), C95–C118.

Pilanci, M. and M. J. Wainwright (2016). Iterative Hessian sketch: Fast and accurate solution approximation for constrained least-squares. *Journal of Machine Learning Research 17*(53), 1–38.

Raskutti, G. and M. Mahoney (2015). A statistical perspective on randomized sketching for ordinary least-squares. In *Proceedings of the 32th ICML Conference*, pp. 617–625.

Wang, H. (2019). More efficient estimation for logistic regression with optimal subsamples. *Journal of Machine Learning Research 20*(132), 1–59.

Wang, H., R. Zhu, and P. Ma (2018). Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association 113*(522), 829–844.

Wang, J., J. D. Lee, M. Mahdavi, M. Kolar, and N. Srebro (2017). Sketching meets random projection in the dual: A provable recovery algorithm for big and high-dimensional data. *Electronic Journal of Statistics 11*(2), 4896–4944.

Wang, Y., A. W. Yu, and A. Singh (2017). On computationally tractable selection of experiments in measurement-constrained regression models. *Journal of Machine Learning Research 18*(143), 1–41.

Woodruff, D. P. et al. (2014). Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science 10*(1–2), 1–157.