Unweighted estimation based on optimal sample under measurement constraints

Jing Wang [1,2,3], HaiYing Wang [3]*, Shifeng Xiong [1]

July 3, 2022

NCMIS, KLSC, Academy of Mathematics and Systems Science, CAS, Beijing 100190, China [1]
School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China [2]
Department of Statistics, University of Connecticut, Storrs, CT 06269, U.S.A. [3]

Abstract

To tackle massive data, subsampling is a practical approach to sift more informative data points. However, when responses are expensive to measure, developing efficient subsampling schemes is challenging, and the optimal sampling approach under measurement constraints was developed to meet this challenge. This method uses the inverses of optimal sampling probabilities to reweight the objective function, which assigns smaller weights on more important data points. Thus the estimation efficiency of the resulting estimator can be improved. In this paper, we propose an unweighted estimating procedure based on optimal subsamples to obtain a more efficient estimator. We obtain the unconditional asymptotic distribution of the estimator via martingale techniques without conditioning on the pilot estimate, which has been less investigated in existing subsampling literature. Both asymptotic results and numerical results show that the unweighted estimator is more efficient in parameter estimation.

keywords: Generalized Linear Models; Massive Data; Martingale Central Limit Theorem

MSC2020: Primary 62D05; secondary 62J12

*Author to whom correspondence may be addressed. Email:haiying.wang@uconn.edu

1 INTRODUCTION

Data acquisition is becoming easier nowadays, and big data bring new challenges to data storage and processing. Conventional statistical models may not be applicable due to limited computational resources. Facing such problems, subsampling has become a popular approach to reduce computational burdens. The key idea of subsampling is to collect more informative data points from the full data and perform calculations on a smaller data set, see Drineas *et al.* (2006); Drineas *et al.* (2011); Mahoney (2011). In some circumstances, only covariates X_i's are available for all the data points, whereas responses Y_i's can merely be obtained on a small portion because they are expensive to measure. For example, extremely large size of modern galaxy datasets has made visual classification of galaxy impractical. Most subsampling probabilities developed recently for generalized linear models (GLMs) rely on all responses of the full data set, see Wang *et al.* (2018); Wang (2019); Ai *et al.* (2021). In order to handle the difficulty when responses are hard to measure, Zhang *et al.* (2021) proposed a response-free optimal sampling scheme under measurement constraints (OSUMC) for GLMs. However, their method uses the reweighted estimator which prevents the estimator from being the most efficient one, since it assigns smaller weights to more informative data points in the objective function. The robust sampling probabilities proposed in Nie *et al.* (2018) do not depend on the responses either, but their investigation focused on linear regression models.

In this paper, we focus on the subsampling method under measurement constraints and propose a more efficient estimator based on the same subsamples taken according to OSUMC for GLMs. We use martingale techniques to derive the unconditional asymptotic distribution of the unweighted estimator and show that its asymptotic covariance matrix is smaller, in the Loewner ordering, than that of the weighted estimator. Before showing the structure of the paper, we first give a short overview of the emerging field of subsampling methods.

Various subsampling methods have been studied in recent years. For linear regression, Drineas *et al.* (2006) developed a subsampling method based on statistical leveraging scores. Drineas *et al.* (2011) developed an algorithm using randomized Hardamard transform. Ma *et al.* (2015) investigated the statistical perspective of leverage sampling. Wang *et al.* (2019) developed an information-based procedure to select optimal subdata for linear regression deteministically. Zhang and Wang (2021) proposed a distributed sampling-based approach for linear models. Ma *et al.* (2020) studied the statistical properties of sampling estimators and proposed several estimators based on asymptotic results which are related to leveraging scores. Beyond linear models, Fithian and Hastie (2014) proposed a local case-control subsampling method to handle imbalanced data sets for logistic regression. Wang *et al.*

(2018) developed an optimal sampling method under A-optimality criterion (OSMAC) for logistic regression. Their estimator can be improved because inverse probability reweighting is applied on the objective function, and Wang (2019) developed a more efficient estimator for logistic regression based on optimal subsample. They proposed an unweighted estimator with bias correction using a similar idea in Fithian and Hastie (2014). They also introduced a Poisson sampling algorithm to reduce RAM usage when calculating optimal sampling probabilities. Ai *et al.* (2021) generalized OSMAC to GLMs and obtained optimal subsampling probabilities under A- and L-optimality criterion for GLMs. These optimal sampling methods require all the responses to construct optimal probabilities, which is not possible under measurement constraints. Zhang *et al.* (2021) developed an optimal sampling method under measurement constraints. Their estimator is also based on the weighted objective function, and thus the performance can be improved. Recently, Cheng *et al.* (2020) extended information-based data selection approach for linear model to logistic regression. Yu *et al.* (2022) derived optimal Poisson subsampling probabilies under the A- and L-optimality criterion for quasi-likelihood estimation, and developed a distributed subsampling framwork to deal with data stored in different machines. Wang and Ma (2020) developed an optimal sampling method for quantile regression. Pronzato and Wang (2021) proposed a sequential online subsampling procedure based on optimal bounded design measures.

We focus on GLMs in this paper, which include commonly used models such as linear, logistic and Poisson regressions. The rest of the paper is organized as follows. Section 2 presents the model setup and briefly reviews the OSUMC method. The more efficient estimator and asymptotic properties are presented in Section 3. Section 4 provides numerical simulations. We summarize our paper in Section 5. Proofs and technical details are presented in the Supplementary Material.

2 BACKGROUND AND MODEL SETUP

We start with reviewing GLMs. Consider independent and identically distributed (i.i.d) data (X_1, Y_1), (X_2, Y_2),..., $(X_n, Y_n) \sim (X, Y)$, where $X \in \mathbb{R}^p$ is the covariate vector and Y is the response variable. Assume that the conditional density of Y given X satisfies that

$$f(y|x, \beta_0, \sigma) \propto \exp\left\{\frac{yx^T\beta_0 - b(x^T\beta_0)}{c(\sigma)}\right\},$$

where β_0 is the unknown parameter we need to estimate from data, $b(\cdot)$ and $c(\cdot)$ are known functions, and σ is the dispersion parameter. In this paper, we are only interested in estimating β_0. Thus, we take $c(\sigma) = 1$ without loss of generality. We also include an intercept in

the model, which is almost always the case in practice. We obtain the maximum likelihood estimator (MLE) of β_0 through maximizing the loglikelihood function, namely,

$$\hat{\beta}_{\text{MLE}} := \arg\max_{\beta} \frac{1}{n} \sum_{i=1}^{n} \left\{ Y_i X_i^T \beta - b(X_i^T \beta) \right\}, \tag{1}$$

which is the same as solving the following score equation:

$$\Psi_n(\beta) := \frac{1}{n} \sum_{i=1}^{n} \{ b'(X_i^T \beta) - Y_i \} X_i = 0,$$

where $b'(\cdot)$ is the derivative of $b(\cdot)$. There is no general closed-form solution to $\hat{\beta}_{\text{MLE}}$, and iterative algorithms such as Newton's method are often used. Therefore, when the data are massive, the computation burden of estimating β_0 is very heavy. To handle this problem, Ai *et al.* (2021) proposed a subsampling-based approach, which constructs the sampling probability $\{\pi_i\}_{i=1}^n$ that depends on both the covariates X_i's and the responses Y_i's. However, it is infeasible to obtain all the responses under measurement constraints. For example, it costs both considerable money and time to synthesize superconducts. When we use data-driven methods to predict the critical temperature with the chemical composition of superconductors, it may be more pratical to measure on a small number of materials to build a data-driven model. To tackle this type of "many X, few Y" scenario, Zhang *et al.* (2021) developed the OSUMC subsampling probabilities.

Assume we obtain a subsample of size r by sampling with replacement according to the probabilities $\pi = \{\pi_i\}_{i=1}^n$. A reweighted estimator is often used in subsample literature, which is defined as the minimizer of the reweighted target function, namely:

$$\hat{\beta}_{\text{w}} := \arg\max_{\beta} \frac{1}{r} \sum_{i=1}^{r} \frac{Y_i^* X_i^{*T} \beta - b(X_i^{*T} \beta)}{n \pi_i^*}, \tag{2}$$

where (X_i^*, Y_i^*) is the data sampled in the ith step, and π_i^* denotes the corresponding sampling probability. Equivalently, we can solve the reweighted score function:

$$\Psi_{\text{w}}^*(\beta) := \frac{1}{r} \sum_{i=1}^{r} \frac{b'(X_i^{*T} \beta) - Y_i^*}{n \pi_i^*} X_i^* = 0,$$

to obtain the reweighted estimator. Zhang *et al.* (2021) proposed a scheme to derive the optimal subsampling probabilities for GLMs under measurement constraints. They first proved that $\hat{\beta}_{\text{w}}$ is asymptotically normal:

$$\mathbb{V}\{\Psi_{\text{w}}^*(\beta_0)\}^{-\frac{1}{2}} \Phi(\hat{\beta}_{\text{w}} - \beta_0) \xrightarrow{d} N(0, I),$$

where the notation "$\xrightarrow{d}$" denotes convergence in distribution,

$$\mathbb{V}\{\Psi_{\mathrm{w}}^*(\beta_0)\} := \mathbb{E}\left[\mathbb{V}\{\Psi_{\mathrm{w}}^*(\beta_0)|X_1^n\}\right] = \mathbb{E}\left\{\frac{1}{n^2}\sum_{i=1}^{n} b''(X_i^T\beta_0)X_iX_i^T\left(\frac{1}{r\pi_i} - \frac{1}{r} + 1\right)\right\},$$

$X_1^n := (X_1, X_2, ..., X_n)$, $b''(\cdot)$ is the second derivative of $b(\cdot)$, and

$$\Phi := \mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^{n} b''(X_i^T\beta_0)X_iX_i^T\right\}. \tag{3}$$

Since the matrix $\Phi^{-1}\mathbb{V}\{\Psi_{\mathrm{w}}^*(\beta_0)|X_1^n\}\Phi^{-1}$ converges to the asymptotic variance of $\hat{\beta}_{\mathrm{w}}$, Zhang *et al.* (2021) minimized its trace, $\mathrm{tr}(\Phi^{-1}\mathbb{V}\{\Psi_{\mathrm{w}}^*(\beta_0)|X_1^n\}\Phi^{-1})$, to obtain the optimal sampling probabilities which depend only on covariate vectors X_1 ,..., X_n:

$$\pi_i^{\mathrm{A-OS}}(\beta_0, \Phi) = \frac{\sqrt{b''(X_i^T\beta_0)}\|\Phi^{-1}X_i\|}{\sum_{j=1}^{n}\sqrt{b''(X_j^T\beta_0)}\|\Phi^{-1}X_j\|}. \tag{4}$$

To avoid the matrix multiplication in $\|\Phi^{-1}X_i\|$ in (4), we can consider a variant of (4) which omits the inverse matrix Φ^{-1}:

$$\pi_i^{\mathrm{L-OS}}(\beta_0) = \frac{\sqrt{b''(X_i^T\beta_0)}\|X_i\|}{\sum_{j=1}^{n}\sqrt{b''(X_j^T\beta_0)}\|X_j\|}. \tag{5}$$

Here, $\{\pi_i^{\mathrm{L-OS}}\}_{i=1}^n$ are another widely used optimal probabilities, and they are the minimizer of $\mathrm{tr}(L\Phi^{-1}\mathbb{V}\{\Psi_{\mathrm{w}}^*(\beta_0)|X_1^n\}\Phi^{-1}L^T)$ with $L = \Phi$. This is a special case of using L-optimality criterion to obtain optimal subsampling probabilities (see Wang *et al.*, 2018; Ai *et al.*, 2021). The probabilities in (4) and (5) are useful when the responses are not available, as we discussed before. However as pointed out in Wang (2019) under the logistic model framework, the weighting scheme adopted in (2) does not bring us the most efficient estimator. Intuitively, if a data point (X_i, Y_i) has a larger sampling probability, it contains more information about β_0. However, data points with higher sampling probabilities have smaller weights in (2). This will reduce the efficiency of the estimator. We propose a more efficient estimator based on the unweighted target function.

3 UNWEIGHTED ESTIMATION AND ASYMPTOTIC THEORY

In this section, we present an algorithm with unweighted estimator and derive its asymptotic property. As we discussed before, the reweighted estimation reduces the importance of

more informative data points. To overcome this problem, Wang (2019) developed a method to correct the bias of the unweighted estimator in logistic regression. In this section, we show that, using the optimal probabilities under measurement constraints, the unweighted estimator is asymptotically unbiased and therefore it is a better estimator since it has smaller asymptotic variance matrix in the Loewner ordering. To make our investigation more general and put the probabilities in (4) and (5) in an unified class, we consider the following general class of the subsampling probabilities in the rest of the paper:

$$\pi_i^{\mathrm{OS}}(\beta_0, \Phi) = \frac{\sqrt{b''(X_i^T\beta_0)}\|L\Phi^{-1}X_i\|}{\sum_{j=1}^n \sqrt{b''(X_j^T\beta_0)}\|L\Phi^{-1}X_j\|}, \tag{6}$$

where L is a fixed matrix. Here the probabilities $\{\pi_i^{\mathrm{OS}}(\beta_0, \Phi)\}_{i=1}^n$ are optimal in terms of that they minimize the asymptotic variance of $L\hat{\beta}_{\mathrm{w}}$. Specifically, when $L = I$, the probabilities in (6) reduce to those in (4) and when $L = \Phi$, they reduce to those in (5).

We define our unweighted estimator as:

$$\hat{\beta}_{\mathrm{uw}} := \arg\max_{\beta} \frac{1}{r}\sum_{i=1}^{r}\left\{Y_i^* X_i^{*T}\beta - b(X_i^{*T}\beta)\right\}, \tag{7}$$

where (X_i^*, Y_i^*)'s are sampled according to the probabilities in (6).

3.1 Notation and main algorithm

We first introduce some notations and the main algorithm. Recall that $X_1^n := (X_1, X_2, ..., X_n)$ and denote $Y_1^n := (Y_1, Y_2, ..., Y_n)$. For a vector $X \in \mathbb{R}^p$, we use $\|X\|$ to denote its Euclidean norm. For a matrix $A \in \mathbb{R}^{p\times p}$, we use $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ to denote its minimum and maxmum eigenvalues, respectively, use $\|A\|_F$ to denote its Frobenius norm, and use $\mathrm{tr}(A)$ to denote its trace. For two positive semi-definite matrices A and B, $A \geq B$ if and only if $A - B$ is positive semi-definite, which is known as the Loewner ordering. For parameter β, we assume that β takes values in a compact set: $\beta \in \mathbb{B}$. Now, we present the main algorithm in Algorithm 1. Since the probabilities in (6) involve unknown quantities, β_0 and Φ, we use pilot estimates to replace them in Algorithm 1.

Algorithm 1 Unweighted estimation for GLM under measurement constraints

1: Take a pilot subsample of size r_{p}: $\{(X_i^{*\mathrm{p}}, Y_i^{*\mathrm{p}})\}_{i=1}^{r_{\mathrm{p}}}$ with simple random sampling from the full data set $\{(X_i, Y_i)\}_{i=1}^n$. Calculate the pilot estimate of β_0:

$$\hat{\beta}_{\mathrm{p}} := \arg\max_{\beta} \frac{1}{r_{\mathrm{p}}} \sum_{i=1}^{r_{\mathrm{p}}} \left\{ Y_i^{*\mathrm{p}} X_i^{*\mathrm{p}T} \beta - b(X_i^{*\mathrm{p}T} \beta) \right\},$$

and the pilot estimate of Φ:

$$\hat{\Phi}_{\mathrm{p}} := \frac{1}{r_{\mathrm{p}}} \sum_{i=1}^{r_{\mathrm{p}}} b''(X_i^{*\mathrm{p}T} \hat{\beta}_{\mathrm{p}}) X_i^{*\mathrm{p}} X_i^{*\mathrm{p}T}.$$

2: Use $\hat{\beta}_{\mathrm{p}}$ and $\hat{\Phi}_{\mathrm{p}}$ to replace β_0 and Φ in (6), respectively, and calculate the sampling probabilities $\{\pi_i^{\mathrm{OS}}(\hat{\beta}_{\mathrm{p}}, \hat{\Phi}_{\mathrm{p}})\}_{i=1}^n$.

3: Obtain a subsample $\{(X_i^*, Y_i^*)\}_{i=1}^r$ of size r according to the sampling probabilities $\{\pi_i^{\mathrm{OS}}(\hat{\beta}_{\mathrm{p}}, \hat{\Phi}_{\mathrm{p}})\}_{i=1}^n$ using sampling with replacement, and solve the estimation equation:

$$\Psi_{\mathrm{uw}}^*(\beta) := \frac{1}{r} \sum_{i=1}^{r} \{b'(X_i^{*T}\beta) - Y_i^*\} X_i^* = 0,$$

to obtain the unweighted estimator defined in (7).

Remark 1. *Our Algorithm 1 is different from the subsampling algorithm in Zhang* et al. *(2021) at step 3 of obtaining the subsampling estimators. There are two types of weights in the subsampling algorithms: the one is the sampling weights which we call subsampling probabilities in this paper, and the other is the estimation weights used to construct the target function. Algorithm 1 and Zhang* et al. *(2021)'s algorithm share the same sampling probabilities (sampling weights) but they use different estimation weights. Zhang* et al. *(2021) use the estimation weights of* $1/\pi_i^{\mathrm{OS}}(\hat{\beta}_{\mathrm{p}}, \hat{\Phi}_{\mathrm{p}})$ *while we set the estimation weights to be uniformly one, i.e., the target function is unweighted. We will show in Section 3.3 that our estimator improve the estimation efficiency. This does not contradict the fact that* $\{\pi_i^{\mathrm{OS}}\}_{i=1}^n$ *are optimal for the algorithm in Zhang* et al. *(2021), because they force the estimation weights to be the inverses of the sampling weights while we do not enforce this requirement.*

Remark 2. *The computational complexity of our two-step Algorithm 1 is the same as the OSUMC estimator in Zhang* et al. *(2021), because we use the same sampling probabilities and the two methods only differ in the weights of the target function. With Newton's method, it requires* $O(\zeta_{\mathrm{p}} r_{\mathrm{p}} p^2)$ *time to compute the pilot estimates, where* ζ_{p} *is the number of iterations for the algorithm to convergence based on the pilot sample. The time complexities of calcu-*

lating sampling probabilities $\{\pi^{\mathrm{A-OS}}\}_{i=1}^n$ *and* $\{\pi^{\mathrm{L-OS}}\}_{i=1}^n$ *are* $O(np^2)$ *and* $O(np)$*, respectively. After obtaining the second stage subsample with the optimal sampling probabilities, it takes* $O(\zeta rp^2)$ *time to solve the unweighted target function where* ζ *is the number of iterations of Newton's algorithm. Thus, the total computational time is* $O(np^2 + \zeta_{\mathrm{p}} r_{\mathrm{p}} p^2 + \zeta rp^2)$ *for A-optimality and* $O(np + \zeta_{\mathrm{p}} r_{\mathrm{p}} p^2 + \zeta rp^2)$ *for L-optimality. The computational complexity of our algorithm based on the A-optimality criterion is the same as the OSUMC algorithm in Zhang* et al. *(2021). Therefore, our method increase the estimation efficiency without increasing the computational burden.*

3.2 Asymptotic normality of $\hat{\beta}_{\mathrm{uw}}$

We focus on unconditional asymptotic results of the unweighted algorithm, and use martingale techniques to prove theorems. To present the asymptotic results, we summarize some regularity conditions first.

Assumption 1. *The second derivative* $b''(\cdot)$ *is bounded and continuous.*

Assumption 2. *The forth moment of the covariate is finite, i.e.,* $\mathbb{E}\left(\|X\|^4\right) < \infty$.

Assumption 3. *Let* $g(x) := \inf_{\beta \in \mathbb{B}} b''(x^T\beta)$*. Assume that* $\lambda_{\min}[\mathbb{E}\{g(X)XX^T\}] > 0$*. Assume that there exists a function* $h(x)$ *such that* $|b'''(x^T\beta)| \leq h(x)$ *and* $\mathbb{E}\{h(X)\|X\|^4\} < \infty$*, where* $b'''(\cdot)$ *denotes the third derivative of* $b(\cdot)$.

Assumption 1 is a commonly used assumption in GLM literature, e.g., Zhang *et al.* (2021). Assumption 2 is a moment condition on X. The second part of Assumption 3 is similar to the third-derivative condition used in the classical theory of MLE. However, here we need a stronger moment condition, $\mathbb{E}[h(X)\|X\|^4] < \infty$, since we use an unequal probability sampling method. Before we prove the asymptotic normality of $\hat{\beta}_{\mathrm{uw}}$, we need to prove some lemmas. First, we present the convergence of $\dot{\Psi}^*_{\mathrm{uw}}(\beta)$.

Lemma 1. *Under assumptions A1-A3, for every sequence* $\beta_n \xrightarrow{p} \beta_0$*,*

$$\hat{m}\dot{\Psi}^*_{\mathrm{uw}}(\beta_n) \xrightarrow{p} \Gamma := \mathbb{E}\left[\left\{b''(X^T\beta_0)\right\}^{\frac{3}{2}} \|L\Phi^{-1}X\| XX^T\right],$$

where $\hat{m} = (1/n)\sum_{i=1}^n \sqrt{b''(X_i^T\hat{\beta}_{\mathrm{p}})}\|L\hat{\Phi}_{\mathrm{p}}^{-1}X_i\|$ *and the notation "*$\xrightarrow{p}$*" denotes convergence in probability.*

Furthermore, to establish the asymptotic normality of $\hat{\beta}_{\mathrm{uw}}$, we present the asymptotic normality of $\Psi^*_{\mathrm{uw}}(\beta_0)$.

Lemma 2. *Under assumptions A1-A3, if* $r/n \to \rho \in [0,1)$, $r_{\mathrm{p}}/\sqrt{n} \to 0$ *and* $\exists \delta > 0$ *such that*

$$\mathbb{E}\left\{\left|b'(X^T\beta_0) - Y\right|^{4+2\delta} \|X\|^{8+4\delta}\right\} < \infty, \tag{8}$$

then

$$\sqrt{r}\hat{m}\Psi^*_{\mathrm{uw}}(\beta_0) \xrightarrow{d} N(0, m\Gamma + \rho\Omega),$$

where $m := \mathbb{E}\left\{\sqrt{b''(X^T\beta_0)}\|L\Phi^{-1}X\|\right\}$ *and* $\Omega = \mathbb{E}\left[\left\{b''(X^T\beta_0)\right\}^2 \|L\Phi^{-1}X\|^2 XX^T\right]$. *If specifically* $\rho = 0$, *then the required condition in* (8) *can be weakened to*

$$\mathbb{E}\left\{\left|b'(X^T\beta_0) - Y\right|^{2+\delta} \|X\|^{4+2\delta}\right\} < \infty.$$

We will use the central limit theorem for martingales described in Jakubowski (1980) and Zhang *et al.* (2021) to prove this Lemma in the supplementary material. In Algorithm 1, the pilot subsample and the optimal subsample are from the same full data so the unconditional distributions of the two subsamples are not independent and it is possible to have overlaps. The assumption $r_{\mathrm{p}}/\sqrt{n} \to 0$ is to ensure that the data points used in the pilot subsample are asymptotically negligible when deriving the unconditional asymptotic distribution of $\Psi^*_{\mathrm{uw}}(\beta_0)$ which depends on both subsamples. This assumption can be replaced by other alternatives such as that the pilot estimator is independent of the full data (e.g., Fithian and Hastie, 2014) and this is appropriate if we modify step 3 of Algorithm 1 to sample from the rest of the data with the pilot subsample data points removed.

Now, we are ready to show the asymptotic normality of the unweighted estimator.

Theorem 1. *Under assumptions A1-A3, assuming that* Γ *is positive-definite, we have*

$$\hat{\beta}_{\mathrm{uw}} - \beta_0 = -\Gamma^{-1}\hat{m}\Psi^*_{\mathrm{uw}}(\beta_0) + o_p\left(1/\sqrt{r}\right).$$

In addition, under the conditions of Lemma 2

$$\sqrt{r}(\hat{\beta}_{\mathrm{uw}} - \beta_0) \xrightarrow{d} N(0, \Sigma^{\rho}_{\mathrm{uw}}),$$

where

$$\Sigma^{\rho}_{\mathrm{uw}} := m\Gamma^{-1} + \rho\Gamma^{-1}\Omega\Gamma^{-1}. \tag{9}$$

Theorem 1 shows that $\hat{\beta}_{\mathrm{uw}}$ is asymptotically unbiased, and from (9) we see that the asymptotic variance of $\hat{\beta}_{\mathrm{uw}}$ can be splitted into two parts: $m\Gamma^{-1}$ and $\rho\Gamma^{-1}\Omega\Gamma^{-1}$. Here, $m\Gamma^{-1}$ is the contribution from the randomness of subsampling and $\rho\Gamma^{-1}\Omega\Gamma^{-1}$ is due to the randomness of the full data. If the subsample size r is of a smaller order of the full data sample size n, i.e., $\rho = 0$, then the randomness of the full data is negligible. If r is of the same

order as n, we need a stronger moment condition (as stated in Lemma 2) to establish the asymptotic normality. In the subsampling setting, we usually expect $r \ll n$, and therefore $m\Gamma^{-1}$ is the dominating term of the asymptotic variance of $\hat{\beta}_{\text{uw}}$.

To estimate the asymptotic variance of $\hat{\beta}_{\text{uw}}$, we propose the following formulas involving only the selected subsample:

$$\hat{\mathbb{V}}(\hat{\beta}_{\text{uw}}) = \frac{1}{r}\hat{m}\hat{\Gamma}^{-1} + \frac{1}{n}\hat{\Gamma}^{-1}\hat{\Omega}\hat{\Gamma}^{-1}, \tag{10}$$

where

$$\hat{\Gamma} = \frac{\hat{m}}{r}\sum_{i=1}^{r} b''(X_i^{*T}\hat{\beta}_{\text{uw}})X_i^* X_i^{*T},$$

and

$$\hat{\Omega} = \frac{n\hat{m}^2}{r}\sum_{i=1}^{r} \pi_i^* b''(X_i^{*T}\hat{\beta}_{\text{uw}})X_i^* X_i^{*T}.$$

Our estimator of the asymptotic variance follows the similar idea that is described in Wang *et al.* (2018) and Wang (2019).

3.3 Efficiency of the unweighted estimator

In this section, we compare the efficiency of the unweighted estimator $\hat{\beta}_{\text{uw}}$ with the weighted estimator $\hat{\beta}_{\text{w}}$ defined in (2). We first restate the asymptotic result in Zhang *et al.* (2021). In their paper, they proved that under some regularity conditions, $\hat{\beta}_{\text{w}}$ is asymptotically normal:

$$\mathbb{V}\{\Psi_{\text{w}}^*(\beta_0)\}^{-\frac{1}{2}}\Phi(\hat{\beta}_{\text{w}} - \beta_0) \xrightarrow{d} N(0, I),$$

where

$$\mathbb{V}\{\Psi_{\text{w}}^*(\beta_0)\} = \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n} b''(X_i^T\beta_0)X_i X_i^T\left\{\frac{1}{r\pi_i} - \frac{1}{r} + 1\right\}\right].$$

Denote

$$\Lambda := \mathbb{E}\left\{\frac{b''(X^T\beta_0)XX^T}{\sqrt{b''(X^T\beta_0)}\|L\Phi^{-1}X\|}\right\},$$

and replace $\pi = \{\pi_i\}_{i=1}^n$ in $\mathbb{V}\{\Psi_{\text{w}}^*(\beta_0)\}$ with the optimal sampling probabilities defined in (6). We then have that

$$\mathbb{V}\{\Psi_{\text{w}}^*(\beta_0)\} = \frac{1}{r}\frac{n-1}{n}m\Lambda + \frac{1}{n}\Phi, \tag{11}$$

where m is define in Lemma 2 and Φ is defined in (3). The details of the calculation are presented in the supplementary material. From (11), if $r/n \to \rho$, the asymptotic variance of $\sqrt{r}(\hat{\beta}_{\text{w}} - \beta_0)$ is

$$\Sigma_{\text{w}}^{\rho} := m\Phi^{-1}\Lambda\Phi^{-1} + \rho\Phi^{-1}. \tag{12}$$

The asymptotic variance $\Sigma_{\mathrm{w}}^{\rho}$ consists of two parts: the term $m\Phi^{-1}\Lambda\Phi^{-1}$ is due to the randomness of subsampling while the term $\rho\Phi^{-1}$ is due to the randomness of the full data. Similarly, in the asymptotic variance $\Sigma_{\mathrm{uw}}^{\rho}$ defined in (9) for the unweighted estimator, $m\Gamma^{-1}$ is due to the randomness of subsampling and $\rho\Gamma^{-1}\Omega\Gamma^{-1}$ is due to the randomness of the full data. We have the following results comparing the aforementioned terms for the weighted and unweighted estimators.

Theorem 2. *If Φ, Γ and Λ are finite and positive-definite, then*

$$\Gamma^{-1} \leq \Phi^{-1}\Lambda\Phi^{-1}, \quad \textit{and} \quad \Gamma^{-1}\Omega\Gamma^{-1} \geq \Phi^{-1},$$

where the inequalities are in the Loewner ordering.

From Theorem 2, $m\Gamma^{-1} \leq m\Phi^{-1}\Lambda\Phi^{-1}$. Thus compared with the weighted estimator, the unweighted estimator has a smaller asymptotic variance component cased by the randomness of subsampling, On the other hand, since $\rho\Gamma^{-1}\Omega\Gamma^{-1} \geq \rho\Phi^{-1}$, the asymptotic variance component due to the full data randomness is larger for the unweighted estimator. A major motivation of subsampling is to significantly reduce the computational or data measurement cost, so it is typical that $r \ll n$ and therefore ρ is typically very small. In this scenario, the asymptotic variance component due to subsampling is the dominating term, and the unweighted estimator has a higher estimation efficiency than the weighted estimator. In the case that $r/n \to 0$, the asymptotic variance component due to the full data randomness is negligible.

We can also get some insights on the difference between the weighted and unweighted estimators by considering them conditionally on the full data. Given the full data, the subsample weighted estimator $\hat{\beta}_{\mathrm{w}}$ is asymptotically unbiased towards the full data unweighted MLE $\hat{\beta}_{\mathrm{MLE}}$ in (1), while the subsample unweighted estimator $\hat{\beta}_{\mathrm{uw}}$ is asymptotically towards to the full data weighted MLE defined as

$$\hat{\beta}_{\mathrm{wMLE}} := \arg\max_{\beta} \frac{1}{n}\sum_{i=1}^{n} w_i \left\{ Y_i X_i^T \beta - b(X_i^T\beta) \right\},$$

where $w_i = \sqrt{b''(X_i^T\beta_0)}\|L\Phi^{-1}X_i\|$ does not depend on Y_i's. Here, $\hat{\beta}_{\mathrm{wMLE}}$ is asymptotically unbiased towards the true parameter because the weights w_i's are only related to X_i's. We see that $\hat{\beta}_{\mathrm{w}}$ and $\hat{\beta}_{\mathrm{uw}}$ essentially approximate different full data estimators $\hat{\beta}_{\mathrm{MLE}}$ and $\hat{\beta}_{\mathrm{wMLE}}$, respectively. It is well known that $\hat{\beta}_{\mathrm{MLE}}$ is more efficient than $\hat{\beta}_{\mathrm{wMLE}}$ based on the full data, but their variation is much smaller than that of $\hat{\beta}_{\mathrm{w}}$ or $\hat{\beta}_{\mathrm{uw}}$, and it is negligible if $r/n \to 0$. Thus the variation of $\hat{\beta}_{\mathrm{w}}$ around $\hat{\beta}_{\mathrm{MLE}}$ and the variation of $\hat{\beta}_{\mathrm{uw}}$ around $\hat{\beta}_{\mathrm{wMLE}}$ are

the major components of the asymptotic variances of $\hat{\beta}_{\text{w}}$ and $\hat{\beta}_{\text{uw}}$ in terms of estimating the true parameter.

When the model is correctly specified, then $\hat{\beta}_{\text{w}}$ and $\hat{\beta}_{\text{uw}}$ are consistent to the same true parameter. However, if the model is mis-specified, then $\hat{\beta}_{\text{w}}$ and $\hat{\beta}_{\text{uw}}$ will typically converge to different limits. Heuristically, $\hat{\beta}_{\text{w}}$ will converge to the solution of $\mathbb{E}[X\{Y - b'(X^T\beta)\}]$ while $\hat{\beta}_{\text{uw}}$ will converge to the solution of $\mathbb{E}[wX\{Y - b'(X^T\beta)\}]$ with $w = \sqrt{b''(X^T\beta_0)}\|L\Phi^{-1}X\|$. In this scenario, it is difficult to compare the efficiency of $\hat{\beta}_{\text{w}}$ with that of $\hat{\beta}_{\text{uw}}$, because it is unknown which solution is closer to the true data-generating parameter.

4 NUMERICAL RESULTS

We investigate the efficiency of the unweighted estimator in parameter estimation through numerical experiments in this section. We present simulation results in Section 4.1 and experiments for real data in Section 4.2.

4.1 Simulation Results

In this section, we evaluate the performance of the more efficient estimator we proposed with simulations. To compare with the original OSUMC estimator, we use the same setups described in Section 5 and appendix of Zhang *et al.* (2021), and show the numerical results for logistic, Poisson and linear regressions.

4.1.1 Logistic Regression and Poisson regression

We first present simulations for logistic regression for which the conditional density of the response has a form of:

$$f(y|x, \beta_0) = \exp\left\{yx^T\beta_0 - \log(1 + e^{x^T\beta_0})\right\}, \text{ for } y = 0, 1.$$

This model implies that the probability of $Y = 1$ given X is:

$$P(Y = 1|X, \beta_0) = \frac{e^{X^T\beta_0}}{1 + e^{X^T\beta_0}}.$$

To generate the full data, we set the true parameter β_0 as a 20 dimensional vector with all entries being equal to 1. The full data sample size is $n = 100,000$ and four distributions of X are considered, which are exactly the same distributions used in Zhang *et al.* (2021). We present these four covariate distributions below for completeness:

1. **mzNormal**: The covariate X follows a multivariate normal distribution $N(0, \Sigma)$, where $\Sigma_{ij} = 0.5^{I(i \neq j)}$ and $I(\cdot)$ represents the indicator function. We have almost equal numbers of 1's and 0's in this scenario.

2. **nzNormal**: The covariate X follows a multivariate normal distribution $N(0.5, \Sigma)$, where Σ is defined in mzNormal. In this scenario, roughly 75% of the responses are 1's.

3. **unNormal**: The covariate X follows a multivariate normal distribution $N(0, \Sigma_1)$, where $\Sigma_1 = U_1 \Sigma U_1$, $U_1 = diag(1, 1/2, ..., 1/20)$ and Σ is the same covariance matrix as we used in mzNormal. For this case, the components of X have different variances.

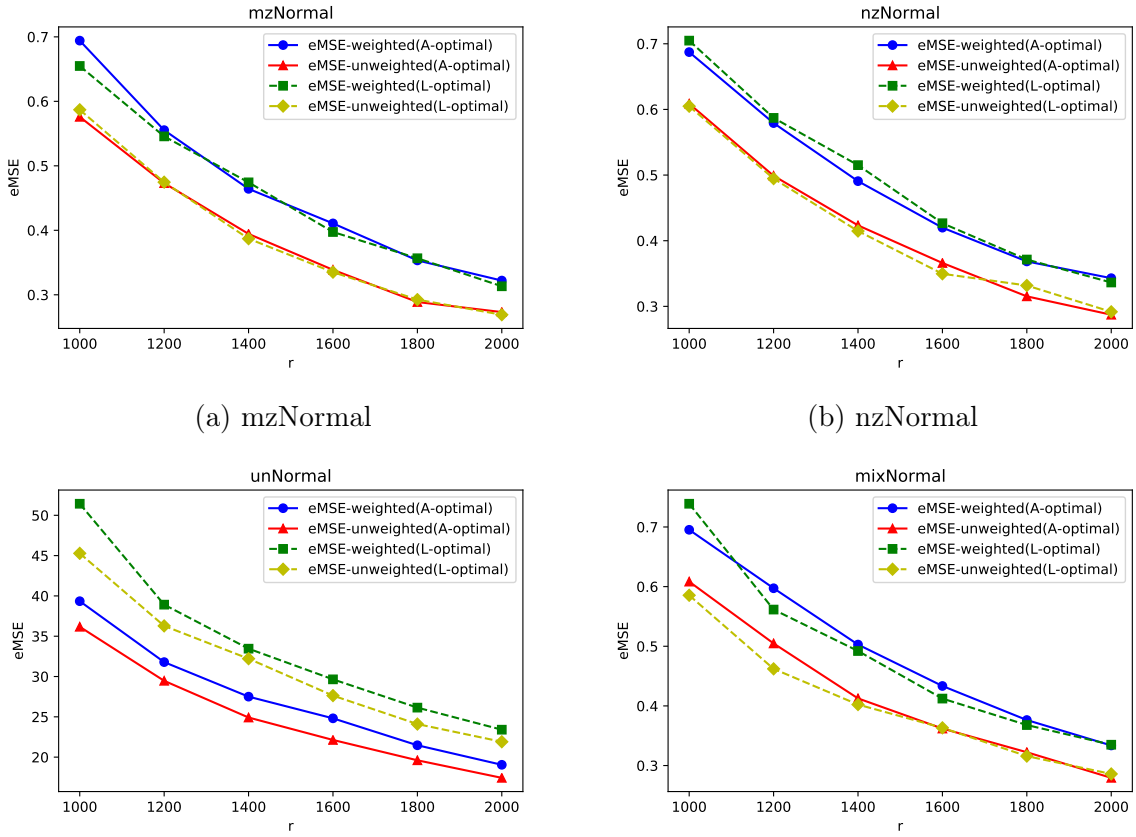
4. **mixNormal**: The covariate X follows a mixed multivariate normal distribution, namely, $X \sim 0.5N(0.5, \Sigma) + 0.5N(-0.5, \Sigma)$, where Σ is the same as what we used in mzNormal.

To compare the performance of the new estimator with the weighted one, we use the empirical MSE of $\hat{\beta}$:

$$\text{eMSE}(\hat{\beta}) = \frac{1}{S} \sum_{s=1}^{S} \|\hat{\beta}^{(s)} - \beta_0\|. \tag{13}$$

Here, $\hat{\beta}^{(s)}$ is the estimated parameter we obtained in the s-th repetition of the simulation. We repeated the simulation for $S = 500$ times to calculate $\text{eMSE}(\hat{\beta})$. For the pilot estimate, we used $r_{\text{p}} = 500$ for both weighted and unweighted methods. In every repetition, we generated the full data, which means we focus on the unconditional empirical MSE. Figure 1 shows that our unweighted estimator performs better than the original OSUMC weighted estimator under each setting when applied to logistic regression. This is true for both A-optimality and L-optimality criteria. For instance, when X has a mixNormal distribution, the emprical MSE of the weighted estimator is over 1.15 times as large as that of the unweighted one. In most cases, $\pi_i^{\text{A-OS}}$ and $\pi_i^{\text{L-OS}}$ perform similarly. When X has a unNormal design, $\pi_i^{\text{A-OS}}$ performs significantly better than $\pi_i^{\text{L-OS}}$ because the A-optimality aims to directly minimize asymptotic MSE.

To evaluate the performance of (10) in estimating the asymptotic variance, we compare $\text{tr}\{\hat{\mathbb{V}}(\hat{\beta}_{\text{uw}})\}$ with the empirical variance. Figure 2 shows that the estimated variances are very close to the empirical variances under the logistic regression model.



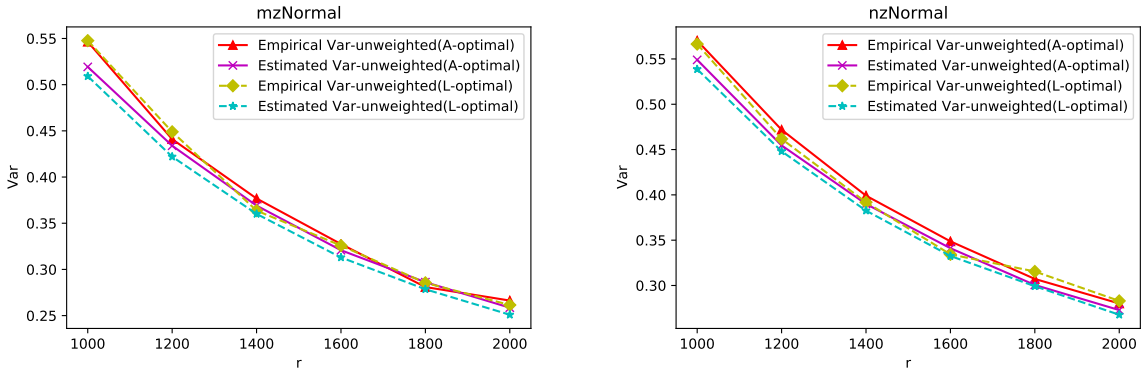
(a) mzNormal

(b) nzNormal

(c) unNormal

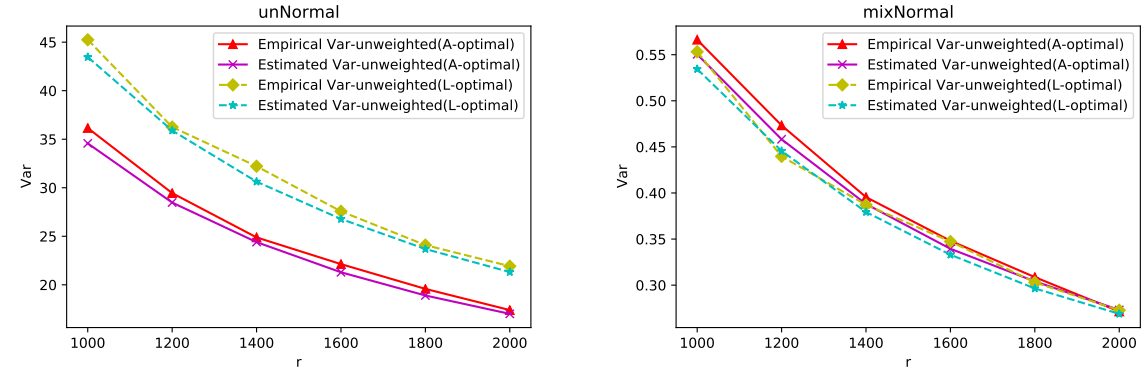
(d) mixNormal

Figure 1: eMSE for different subsampe sizes r with a pilot sample size $r_{\mathrm{p}} = 500$ for logistic regression under different settings.



(a) mzNormal

(b) nzNormal



(c) unNormal

(d) mixNormal

Figure 2: Empirical variance and estimated variance, $\mathrm{tr}\{\hat{\mathbb{V}}(\hat{\beta}_{\mathrm{uw}})\}$, for different subsampe sizes r with a pilot sample size $r_{\mathrm{p}} = 500$ for the unweighted estimator under different settings.

Performances of the unweighted estimator under the Poisson regression are also investigated. The Poisson regression model has a form of

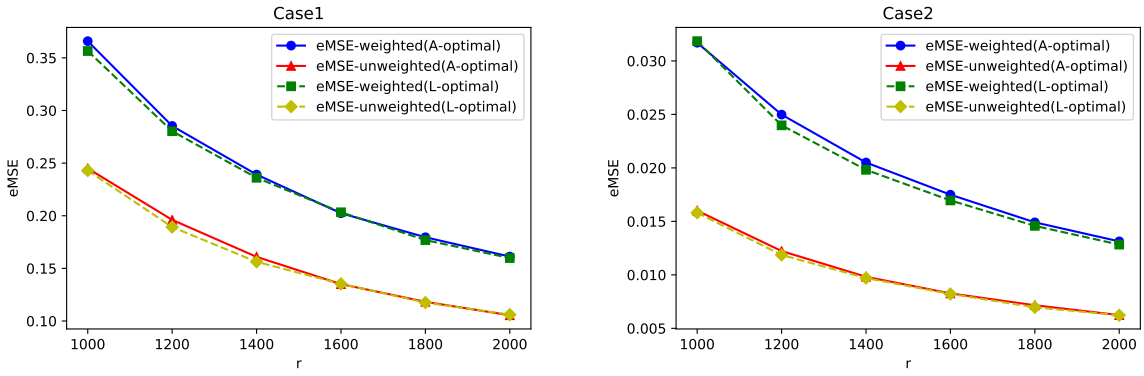
$$f(y|x, \beta_0) = \exp\left\{yx^T\beta_0 - e^{x^T\beta_0} - \log(y!)\right\}, \text{ for } y = 0, 1, 2, ...$$

We generated $n = 100,000$ data points. A 100×1 vector of 0.5 is used as the true value of the parameter, β_0, in this scenario. We use the same settings discussed in the appendix of Zhang *et al.* (2021). Specifically, covariates are generated using the following two settings:

1. **Case 1**: Each component of X are generated from the uniform distribution over $[-0.5, 0.5]$ independently.

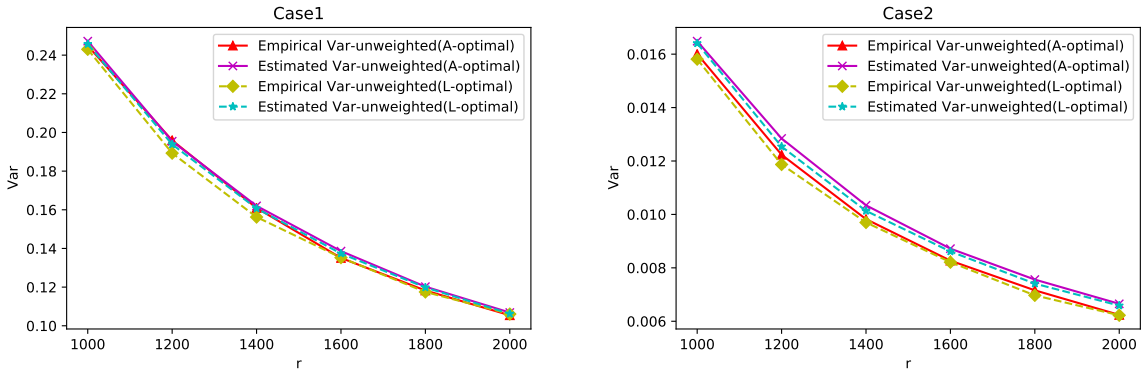
2. **Case 2**: First half of the components of X are generated from the uniform distribution over $[-0.5, 0.5]$ independently, and the other half of the components of X are generated from the uniform distribution over $[-1, 1]$ independently.

Again we repeated the experiment for $S = 500$ times and in each repetition we sampled $r_{\mathrm{p}} = 500$ data points to obtain polit estimates. We also compared the empirical MSE defined in (13) and calculated $\mathrm{tr}\{\hat{\mathbb{V}}(\hat{\beta}_{\mathrm{uw}})\}$ to investigate the performance of the estimated variance defined in (10). Empirical MSEs of the unweighted and weighted estimators are presented in Figure 3. For Poisson regression, our unweighted estimator also outperforms the weighted OSUMC estimator under both criteria we considered, and $\pi_i^{\mathrm{A-OS}}$ and $\pi_i^{\mathrm{L-OS}}$ perform similarly. For Case 1, The empirical MSE of the weighted estimator is around 1.5 times as large as that of the unweighted estimator we proposed. For Case 2, the empirical MSE of our estimator is about half of that of the weighted estimator. The results for the estimated variances are presented in Figure 4. Estimated variance we proposed in (10) also works well under Poisson model.



(a) Case1 (b) Case2

Figure 3: eMSE for different subsampe sizes r with a pilot sample size $r_{\mathrm{p}} = 500$ for Poisson regression under different settings.



(a) Case1 (b) Case2

Figure 4: Empirical variance and estimated variance, $\mathrm{tr}\{\hat{\mathbb{V}}(\hat{\beta}_{\mathrm{uw}})\}$, for different subsampe sizes r with a pilot sample size $r_{\mathrm{p}} = 500$ for the unweighted estimator under different settings.

4.1.2 Linear Model

We now present simulation results for linear regression. We also used the settings in Zhang *et al.* (2021) which generated full data of size $n = 100,000$ from the following model:

$$Y = X\beta_0 + \epsilon,$$

where $\beta_0 = (\underbrace{0.1, ..., 0.1}_{5}, \underbrace{10, ..., 10}_{20}, \underbrace{0.1, ..., 0.1}_{5})^T$ is a 30 dimensional vector, and $\epsilon \sim N(0, 9I_n)$. We used the following distributions of X:

1. **GA**: The covariate X follows a multivariate normal distribution $N(1_p, \Sigma_2)$, where $p = 30$, $\Sigma_2 = U_2 \Sigma U_2$ and $U_2 = diag(5, 5/2, ..., 5/30)$. The entries of Σ are $\Sigma_{ij} = 0.5^{I(i \neq j)}$, which is the same as we defined before.

2. **T3**: The covariate X follows a multivariate t-distribution which has degrees of freedom 3, $T_3(0, \Sigma_2)$, and Σ_2 is defined in GA above.

3. **T1**: The covariate X follows a multivariate t-distribution which has degrees of freedom 1, $T_1(0, \Sigma_2)$, and Σ_2 is the same as GA.

4. **EXP**: Components of X are i.i.d. from an exponential distribution with a rate parameter of 2.

The first three settings are exactly the same settings used in Zhang *et al.* (2021). The last setting is used in Wang *et al.* (2019) and Wang (2019). Since the sampling probabilities are not related to the responses for linear models, Algorithm 1 can be simplified. For completeness, we present the simplified algorithm as Algorithm 2, which is similar to the algorithm used in Ma *et al.* (2015).

We also repeated the simulation for $S = 500$ times and compared the empirical MSEs. In this section, we present the numerical results under A-optimality only. The results under L-optimality are similar and we present them in the supplementary material. Simulation results of unconditionally empirical MSE are presented in Figure 5. We see that the unweighted estimator is more efficient in every case. Especially, when X has a T_3 or T_1 distribution, the unweighted estimator performs significantly better than the weighted estimator. As described in Zhang *et al.* (2021), the OSUMC estimator outperforms other sampling methods more obviously when X is heavy-tailed. We notice that using the unweighted estimator, the advantage of OSUMC can be significantly reinforced when the design is heavy-tailed, despite that X does not meet the regularity conditions we present in Section 3.

Algorithm 2 Unweighted estimation for linear model under measurement constraints

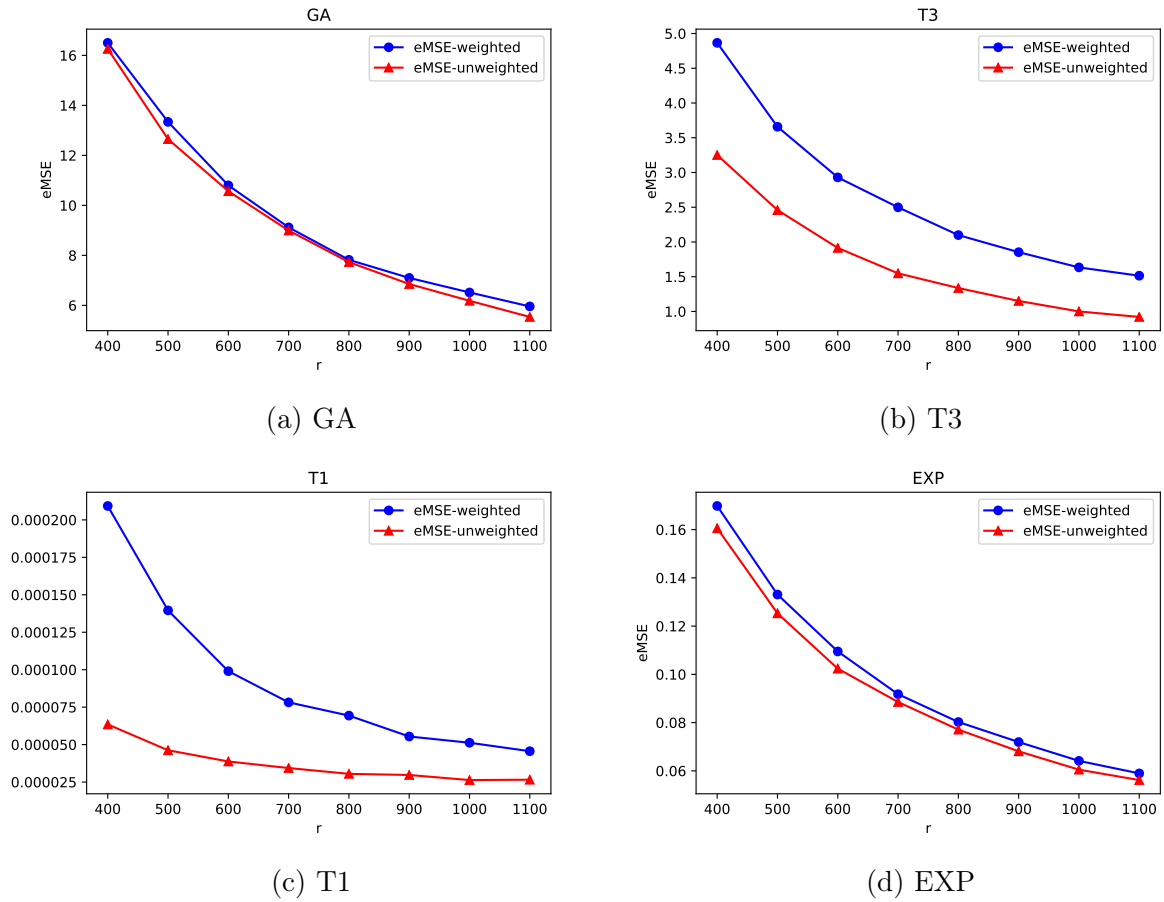
1: Caculate the sampling probabilities $\{\pi_i\}_{i=1}^n$ using the following formula:

$$\pi_i^{\mathrm{A-OS}} = \frac{\left\| \left(\sum_{j=1}^n X_j X_j^T \right)^{-1} X_i \right\|}{\sum_{k=1}^n \left\| \left(\sum_{j=1}^n X_j X_j^T \right)^{-1} X_k \right\|}$$

2: Obtain a subsample $\{(X_i^*, Y_i^*)\}_{i=1}^r$ of size r according to the sampling probabilities $\{\pi_i^{\mathrm{A-OS}}\}_{i=1}^n$ using sampling with replacement, and solve the estimation equation:

$$\Psi_{\mathrm{uw}}^*(\beta) := \frac{1}{r} \sum_{i=1}^r (X_i^{*T}\beta - Y_i^*) X_i^* = 0,$$

to obtain the unweighted estimator.



(a) GA (b) T3

(c) T1 (d) EXP

Figure 5: eMSE for different subsampe sizes r for linear regression under different settings.

4.1.3 Computational Complexity

We present the results of the computational time for the simulation study based on logistic regression in Table 1. We uses the same four settings for the logistic regression in Section 4.1.1, and repeated the experiments for $S = 500$ times. We recorded the computing time for both the weighted and unweighted procedures and implemented both $\pi^{\mathrm{A-OS}}$ and $\pi^{\mathrm{L-OS}}$ using Python. Our computation were carried out on a laptop running Windows 10 with an Intel i5 processor and 8GB memory, and the optimization was done using the `sklearn.linear_model.LogisticRegression` package. We present the results with subsample size $r = 1000$. The results for other subsample size are similar and thus are omitted.

Table 1: Computational time (seconds)

	A-optimiality		L-optimality		Full data
	weighted	unweighted	weighted	unweighted	
mzNormal	42.38	36.70	32.15	26.42	177.10
nzNormal	39.98	36.38	30.31	26.89	165.15
unNormal	41.20	37.47	32.61	28.69	256.45
mixNormal	40.97	36.14	31.92	27.45	162.44

In Table 1, both the weighted and unweighted subsample estimators significantly reduce the computational time compared with the MLE. The computational time of the unweighted estimator is not significant different from that of the weighted estimator. The probabilities based on L-optimality reduce more computational time compared with the probabilities based on the A-optimality, which agrees with the analysis in Section 3. Interestingly, we see that the unweighted estimators is faster than the weighted estimator. This is because the target function of the unweighted estimator is usually smoother than that of the weighted estimator, and thus it takes fewer number of iterations for the algorithm to converge with the unweighted estimator. To confirm this, we present the average numbers of iterations in optimizing the weighted and unweighted target functions in Table 2.

Table 2: Average number of iterations of the optimization algorithm

	A-optimiality		L-optimality	
	weighted	unweighted	weighted	unweighted
mzNormal	18.53	10.68	18.51	10.77
nzNormal	19.08	10.85	18.74	10.88
unNormal	22.71	11.81	22.77	12.42
mixNormal	19.04	10.85	18.88	10.82

4.2 Experiments for real data

We apply our more efficient unweighted estimator to real data and evaluate its performance in this section.

4.2.1 Superconductivty Data Set

In this section, we apply our more efficient estimator to the superconductivty data set, which is also used in Zhang *et al.* (2021). The data set is available from the Machine Learning Repository at `https://archive.ics.uci.edu/ml/datasets/Superconductivty+Data#`. It contains 21,263 different data points, and every data point has 81 features with one continuous response. Each data point represents a superconductor. The response is the superconductor's critical temperature and the features are extracted from its chemical formula. For example, the 81st column is the number of elements of the superconductor. We use standardized features as covariate variables and adopted a multiple linear regression model to fit the critical temperature from the chemical formula of the superconductor. Specially, the linear regression model is

$$Y = \beta^{(0)} + \beta^{(1)} Z_1 + \beta^{(2)} Z_2 + ... + \beta^{(81)} Z_{81} + \epsilon,$$

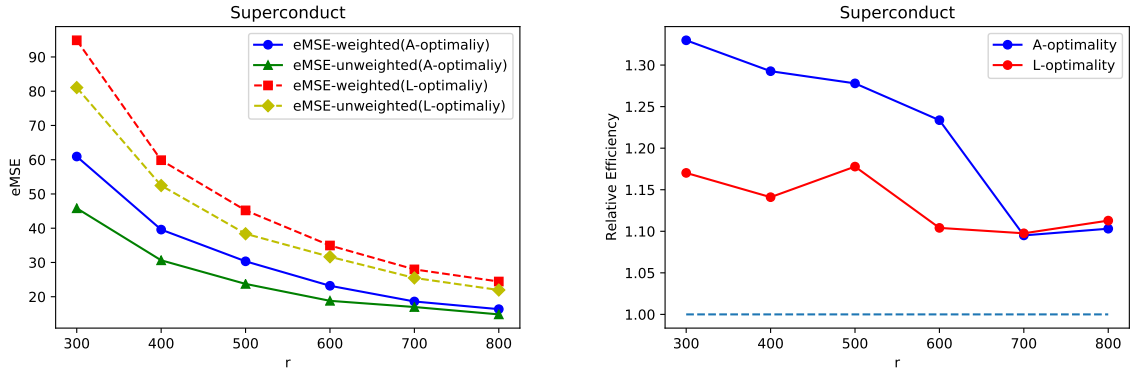
where Z_i's represent the standardize features, Y is the critical temperature, and ϵ is the normally distributed error. To measure the performances of sampling methods in parameter estimation, we use the empirical MSE of the estimator:

$$\text{eMSE}(\hat{\beta}) = \frac{1}{S} \sum_{s=1}^{S} \|\hat{\beta}^{(s)} - \hat{\beta}_{\text{MLE}}\|, \tag{14}$$

and the relative efficiency:

$$\text{Relative Efficiency} = \frac{\text{eMSE}(\hat{\beta}_{\text{w}})}{\text{eMSE}(\hat{\beta}_{\text{uw}})}, \tag{15}$$

where $\hat{\beta}^{(s)}$ represents the estimate in the s-th repetition. Here we use the full data estimator $\hat{\beta}_{\text{MLE}}$ instead of the "true" parameter β_0 to calculate eMSE because the true parameter is unknown for any real data sets. We repeated the experiment for $S = 1000$ times, and present the numerical results in Figure 6. Our unweighted estimator also outperforms the weighted estimator when applied to the Superconduct data set and $\pi_i^{\text{A-OS}}$ result in smaller eMSE than $\pi_i^{\text{L-OS}}$ for both the weighted and unweighted estimators.



(a) eMSE

(b) Relative Efficiency

Figure 6: eMSE and Relative Efficiency for Superconductivty data set with different subsample data sizes r.

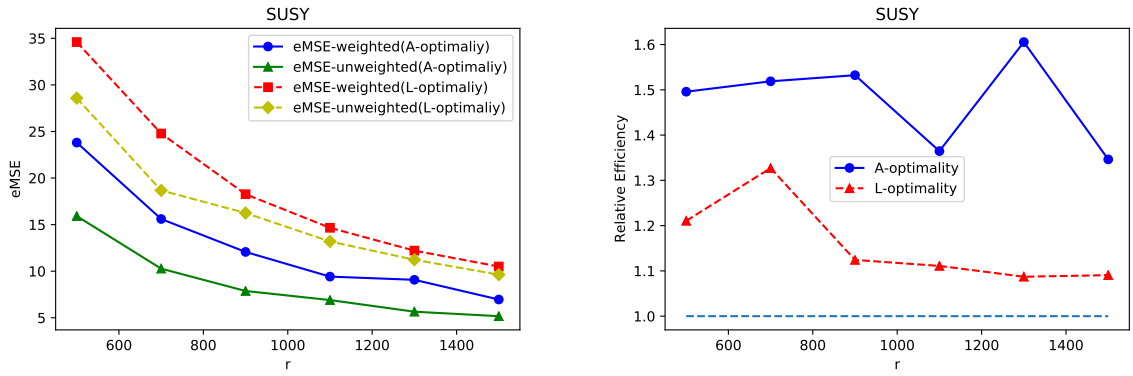
4.2.2 Supersymmetric Data Set

In this section, the supersymmetric(SUSY) benchmark data set is used to evaluate the performance of the unweighted estimator when applied to real data under logistic model. The SUSY data set is available from the Machine Learning Repository at `https://archive.ics.uci.edu/ml/datasets/SUSY`, which is also used in Wang *et al.* (2018) and Wang (2019). The data is composed of $n = 5,000,000$ data points. Each data point represents a process and has one binary response with 18 covariates. The response variable represents whether the process produces new supersymmetric particles or the process is a background process. The kinematic features of the process are used as covariates. We used a logistic regression model to fit the data. Specifically, model the probability that a process produces new supersymmetric particles as

$$P(Y = 1|Z, \beta) = \frac{e^{\beta^{(0)} + \sum_{i=1}^{18} \beta^{(i)} Z_i}}{1 + e^{\beta^{(0)} + \sum_{i=1}^{18} \beta^{(i)} Z_i}},$$

where Z_i's are the kinematic features of a process. In order to compare the efficiency of parameter estimation, we still use the regression coefficients $\hat{\beta}_{\text{MLE}}$ derived from the full data

as the "true parameter". The empirical MSE of the estimator defined in (14) and the relative efficiency defined in (15) are also considered. We still repeated the experiment for $S = 1000$ times here and drew a pilot subsample of size $r_{\mathrm{p}} = 500$ in each repetition. Figure 7 shows that the unweighted estimator is over 130% more efficient than the weighted one when applied to the SUSY data set when using $\pi_i^{\mathrm{A-OS}}$, and over 110% more efficient when using $\pi^{\mathrm{L-OS}}$. Also, $\pi_i^{\mathrm{A-OS}}$ performs better than $\pi_i^{\mathrm{L-OS}}$ for the SUSY data set.



(a) eMSE

(b) Relative Efficiency

Figure 7: eMSE and Relative Efficiency for SUSY data set with different subsample data sizes r and a pilot sample size $r_{\mathrm{p}} = 500$.

5 CONCLUSION

In this paper, we proposed a novel unweighted estimator based on OSUMC subsample for GLMs. It can be used to reduce computational burdens when responses are hard to acquire. A two-step scheme was proposed and we showed the asymptotic normality of the estimator unconditionally. We proved asymptotic results under martingale framwork without conditional on pilot estimates. Furthermore, we showed that our new estimator is more efficient than the original OSUMC estimator in parameter estimation under subsampling settings. Several numerical experiments are implemented to demonstrate the better performance of our unweighted estimator over the original weighted esitmator.

Some extensions may be interesting for future research. Sampling with replacement is used for both the weighted estimator and our new unweighted estimator based on OSUMC. Poisson sampling is an alternative that reduces the RAM usage for subsampling methods. Therefore, Poisson sampling is worthy to develop under measurement constraint circumstances. Extending our subsampling procedure to models beyond GLMs is also an interesting topic for future studies.

ACKNOWLEDGEMENTS

The authors are very grateful to two anonymous referees and the editor for constructive comments that helped to improve the paper. Wang and Wang's research is partially supported by US NSF grant CCF-2105571. Xiong's work is supported by the National Science Foundation of China (Grant No. 12171462).

References

Ai, M., Yu, J., Zhang, H., and Wang, H. (2021). Optimal subsampling algorithms for big data regressions. *Statistica Sinica* **31**, 2, 749–772.

Cheng, Q., Wang, H., and Yang, M. (2020). Information-based optimal subdata selection for big data logistic regression. *Journal of Statistical Planning and Inference* **209**, 112 – 122.

Drineas, P., Mahoney, M., Muthukrishnan, S., and Sarlos, T. (2011). Faster least squares approximation. *Numerische Mathematik* **117**, 219–249.

Drineas, P., Mahoney, M. W., and Muthukrishnan, S. (2006). Sampling algorithms for l_2 regression and applications. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, 1127–1136. Society for Industrial and Applied Mathematics.

Fithian, W. and Hastie, T. (2014). Local case-control sampling: Efficient subsampling in imbalanced data sets. *Annals of statistics* **42**, 5, 1693.

Jakubowski, A. (1980). On limit theorems for sums of dependent hilbert space valued random variables. In *Mathematical Statistics and Probability Theory*, 178–187. Springer.

Ma, P., Mahoney, M. W., and Yu, B. (2015). A statistical perspective on algorithmic leveraging. *The Journal of Machine Learning Research* **16**, 1, 861–911.

Ma, P., Zhang, X., Xing, X., Ma, J., and Mahoney, M. (2020). Asymptotic analysis of sampling estimators for randomized numerical linear algebra algorithms. vol. 108 of *Proceedings of Machine Learning Research*, 1026–1035, Online. PMLR.

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

Nie, R., Wiens, D. P., and Zhai, Z. (2018). Minimax robust active learning for approximately specified regression models. *Canadian Journal of Statistics* **46**, 1, 104–122.

Pronzato, L. and Wang, H. (2021). Sequential online subsampling for thinning experimental designs. *Journal of Statistical Planning and Inference* **212**, 169–193.

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

Wang, H. and Ma, Y. (2020). Optimal subsampling for quantile regression in big data. *Biometrika* **108**, 1, 99–112.

Wang, H., Yang, M., and Stufken, J. (2019). Information-based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association* **114**, 525, 393–405.

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

Yu, J., Wang, H., Ai, M., and Zhang, H. (2022). Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data. *Journal of the American Statistical Association* **117**, 537, 265–276.

Zhang, H. and Wang, H. (2021). Distributed subdata selection for big data via sampling-based approach. *Computational Statistics & Data Analysis* **153**, 107072.

Zhang, T., Ning, Y., and Ruppert, D. (2021). Optimal sampling for generalized linear models under measurement constraints. *Journal of Computational and Graphical Statistics* **30**, 1, 106–114.