Noname manuscript No.
(will be inserted by the editor)

Optimal Subsampling for Softmax Regression

Yaqiong Yao · HaiYing Wang

Received: date / Accepted: date

Abstract To meet the challenge of massive data, Wang et al (2018b) developed an optimal subsampling method for logistic regression. The purpose of this paper is to extend their method to softmax regression, which is also called multinomial logistic regression and is commonly used to model data with multiple categorical responses. We first derive the asymptotic distribution of the general subsampling estimator, and then derive optimal subsampling probabilities under the A-optimality criterion and the L-optimality criterion with a specific L matrix. Since the optimal subsampling probabilities depend on the unknowns, we adopt a two-stage adaptive procedure to address this issue and use numerical simulations to demonstrate its performance.

Keywords Massive data · Subsampling · Optimality criterion · Softmax regression

1 Introduction

With the rapid development of science and technology, extremely large datasets are ubiquitous. How to extract useful information with limited computing resources from these massive datasets is a common challenge. To meet this challenge, the emerging subsampling-based methods have demonstrated promising

This work was supported by NSF grant 1812013, a UCONN REP grant, and a GPU grant from NVIDIA Corporation.

Yaqiong Yao
E-mail: yaqiong.yao@uconn.edu

HaiYing Wang
E-mail: haiying.wang@uconn.edu

Department of Statistics
University of Connecticut

performance. These methods randomly select subsamples from the full data and perform calculations on the subsamples to approximate the quantities of interest based on the full data. This idea has attracted a lot of attention in the fields of theoretical computer science and machine learning. However, studies from the statistical point of view are limited. The long history of investigation in statistics, especially in the field of experimental design and survey sampling, has accumulated various techniques to obtain useful information as much as possible with a fixed budget. These techniques can provide us with valuable guidance to design more efficient subsampling methods for massive datasets.

Investigations on subsampling-based methods are fruitful such as in matrix operation approximations (Frieze et al, 2004; Drineas et al, 2006a,b,c) and matrix decompositions (Drineas et al, 2008; Mahoney and Drineas, 2009). To solve ordinary least-squares, Drineas et al (2006d) developed a subsampling method focusing on influential data points. Drineas et al (2011) developed an algorithm that processes the data using a randomized Hadamard transform and then takes a subsample by uniform subsampling. Look up Mahoney (2011) for a systematic overview of the emerging field. Existing studies mostly focus on fast calculation and available results are about algorithmic properties of the proposed methods. Ma et al (2015) considered some statistical properties of subsampling-based algorithms for linear regression, and proposed to combine the uniform subsampling probability and statistical leverage scores for better performance. Raskutti and Mahoney (2016) assessed statistical properties of randomized sketching for least-squares estimators in linear regression. By using some basic indication of optimal design of experiments, Wang et al (2018a) proposed a novel method called Information-Based Optimal Subdata Selection for linear regression which outperforms other existing methods significantly.

The aforementioned studies focus on linear regression. Wang et al (2018b) first considered logistic regression and introduced the idea of optimal design of experiments into subsampling scheme to develop optimal subsampling methods. They derived the asymptotic distribution of the general subsampling estimator and then obtained subsampling probabilities that minimize the asymptotic variance-covariance matrix. This paper is closely related to the work of Wang et al (2018b). We consider the softmax regression model, which is also called multinomial logistic regression and is often used for multi-label classification. We will derive optimal subsampling probabilities for this model under the A-optimality criterion and the L-optimality criterion with a specific L matrix (Atkinson et al, 2007). We will present the model setup and our main results in Section 2, and present some simulation results in Section 3. Section 4 summarizes the paper and the proofs of our main results are provided in the appendix.

2 Model setup and optimal subsampling

Let $y \in \{c_0, ..., c_K\}$ be a multiclass categorical response variable and $\mathbf{x}$ be a d dimensional covariate. A softmax regression model assumes that given $\mathbf{x}$,

$$\mathbb{P}(y = c_k|\mathbf{x}) = p_k(\mathbf{x}, \boldsymbol{\beta}) = \frac{e^{\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_k}}{\sum_{l=0}^{K} e^{\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_l}}, \quad k = 0, ..., K, \tag{1}$$

where $\boldsymbol{\beta}_k$, $k = 0, ..., K$, are d dimensional regression coefficients belonging to a compact subset of $\mathbb{R}^d$. For identifiability, we assume that $\boldsymbol{\beta}_0 = \mathbf{0}$, so the whole vector of unknown parameters is $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, ..., \boldsymbol{\beta}_K^{\mathrm{T}})^{\mathrm{T}}$, a Kd d dimensional vector.

Let independent full data of size N be $\mathcal{D}_N = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)\}$. The unknown parameter $\boldsymbol{\beta}$ can be estimated by the maximum likelihood estimator (MLE). For ease of presentation, define $\delta_{i,k} = I(y_i = c_k)$, $k = 0, ...K$, where $I()$ is the indicator function. The MLE based on the full data, denoted as $\hat{\boldsymbol{\beta}}_{\text{full}}$, is the maximizer of the following log-likelihood

$$\begin{aligned}\ell_f(\boldsymbol{\beta}) &= \frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K}\delta_{i,k}\log\{p_k(\mathbf{x}_i, \boldsymbol{\beta})\} \\ &= \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{k=1}^{K}\delta_{i,k}\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_k - \log\left\{1 + \sum_{l=1}^{K} e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_l}\right\}\right].\end{aligned}$$

There is no general closed-form solution to this optimization problem, and an iterative algorithm must be used. A commonly used iterative algorithm is the Newton-Raphson method. To introduce this method, we derive the gradient (first order partial derivatives) and the Hessian matrix (second order partial derivatives) of the log-likelihood with respect to the parameter $\boldsymbol{\beta}$:

$$\frac{\partial \ell_f(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{s}_i(\boldsymbol{\beta}) \otimes \mathbf{x}_i, \quad \text{and} \quad \frac{\partial^2 \ell_f(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} = -\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\phi}_i(\boldsymbol{\beta}) \otimes (\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}),$$

respectively, where $\mathbf{s}_i(\boldsymbol{\beta})$ is a K dimensional vector with the kth element being $\mathbf{s}_{i,k}(\boldsymbol{\beta}) = \delta_{i,k} - p_k(\mathbf{x}_i, \boldsymbol{\beta})$, $\boldsymbol{\phi}_i(\boldsymbol{\beta})$ is a $K \times K$ matrix with the kth diagonal element being $\phi_{i,k}(\boldsymbol{\beta}) = p_k(\mathbf{x}_i, \boldsymbol{\beta}) - p_k^2(\mathbf{x}_i, \boldsymbol{\beta})$ and the k_1k_2th off diagonal element being $\phi_{i,k_1k_2}(\boldsymbol{\beta}) = -p_{k_1}(\mathbf{x}_i, \boldsymbol{\beta})p_{k_2}(\mathbf{x}_i, \boldsymbol{\beta})$, and $\otimes$ is the Kronecker product. The Newton-Raphson method finds the full data MLE, $\hat{\boldsymbol{\beta}}_{\text{full}}$, by iteratively applying

$$\hat{\boldsymbol{\beta}}_{\text{full}}^{(t+1)} = \hat{\boldsymbol{\beta}}_{\text{full}}^{(t)} + \left\{\sum_{i=1}^{N}\boldsymbol{\phi}_i(\hat{\boldsymbol{\beta}}_{\text{full}}^{(t)}) \otimes (\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}})\right\}^{-1}\sum_{i=1}^{N}\mathbf{s}_i(\hat{\boldsymbol{\beta}}_{\text{full}}^{(t)}) \otimes \mathbf{x}_i$$

until convergence. However, for massive data, iterative calculations on the full data may take too long time, and sometimes this is not feasible. A popular and practical solution in this case is to take a subsample, and use the estimator calculated from the subsample to approximate the full data MLE.

Now we introduce the general subsampling procedure. Let π_1, ..., π_N be subsampling probabilities such that $\sum_{i=1}^N \pi_i = 1$. Using subsampling with replacement, draw a random subsample of size n ($\ll N$) from the full data, according to the probabilities $\{\pi_i\}_{i=1}^N$. We use * to indicate quantities for the subsample. For example, the covariates, responses, and subsampling probabilities in the subsample are denoted as $\mathbf{x}_i^*$, y_i^*, and π_i^*, respectively, for $i = 1, ..., n$. Here, since the subsampling probabilities π_i's are allowed to depend on the full data, including the responses, we need to use inverses of π_i's as weights in the log-likelihood for the subsample. Otherwise, the resulting estimator is in general biased. Therefore, the subsample estimator, say $\hat{\boldsymbol{\beta}}_{\text{sub}}$, is the maximizer of

$$\ell_s^*(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \frac{1}{N\pi_i^*}\left[\sum_{k=1}^K \delta_{i,k}^* \boldsymbol{\beta}_k^{\mathrm{T}}\mathbf{x}_i^* - \log\left\{1 + \sum_{l=1}^K e^{\boldsymbol{\beta}_l^{\mathrm{T}}\mathbf{x}_i^*}\right\}\right], \tag{2}$$

where $\delta_{i,k}^* = I(y_i^* = c_k)$. To maximize (2), the Newton-Raphson iterator for the subsample is

$$\hat{\boldsymbol{\beta}}_{\text{sub}}^{(t+1)} = \hat{\boldsymbol{\beta}}_{\text{sub}}^{(t)} + \left\{\sum_{i=1}^n \frac{\boldsymbol{\phi}_i(\hat{\boldsymbol{\beta}}_{\text{sub}}^{(t)}) \otimes (\mathbf{x}_i^*\mathbf{x}_i^{*\mathrm{T}})}{N\pi_i^*}\right\}^{-1} \sum_{i=1}^n \frac{\mathbf{s}_i(\hat{\boldsymbol{\beta}}_{\text{sub}}^{(t)}) \otimes \mathbf{x}_i^*}{N\pi_i^*}.$$

Here, iterative calculations are performed on the subsample, and thus the required computational resources are in the scale of the subsample size. If $\sum_{i=1}^n \pi_i^{*-1}\boldsymbol{\phi}_i(\boldsymbol{\beta}^{(t)}) \otimes (\mathbf{x}_i^*\mathbf{x}_i^{*\mathrm{T}})$ is positive definite, then the objective function in (2) is concave, and under some conditions, $\hat{\boldsymbol{\beta}}_{\text{sub}}^{(t+1)}$ converges to $\hat{\boldsymbol{\beta}}_{\text{sub}}$ in a quadratic rate (Ortega and Rheinboldt, 1970).

Now we investigate asymptotic properties of $\hat{\boldsymbol{\beta}}_{\text{sub}}$ under some regularity assumptions listed below.

Assumption 1 *As $N \to \infty$, $\mathbf{M}_N = N^{-1}\sum_{i=1}^N \boldsymbol{\phi}_i(\hat{\boldsymbol{\beta}}_{\text{full}}) \otimes (\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}})$ goes to a positive-definite matrix in probability and $N^{-1}\sum_{i=1}^N \|\mathbf{x}_i\|^3 = O_P(1)$, where $O_P(1)$ means bounded in probability.*

Assumption 2 *For $k = 2, 4$, $N^{-2}\sum_{i=1}^N \pi_i^{-1}\|\mathbf{x}_i\|^k = O_P(1)$; and there exists some $\delta > 0$ such that $N^{-(2+\delta)}\sum_{i=1}^N \pi_i^{-1-\delta}\|\mathbf{x}_i\|^{2+\delta} = O_P(1)$.*

Assumption 1 essentially requires that the observed information matrix is asymptotically non-singular and the third moment of the full data covariates is bounded in probability. Assumption 2 imposes some conditions on the subsampling probabilities.

Theorem 1 *Under Assumptions 1 and 2, given the full data $\mathcal{D}_N$ in probability, as $n \to \infty$ and $N \to \infty$, the approximation error $\hat{\boldsymbol{\beta}}_w^{\pi} - \hat{\boldsymbol{\beta}}_{\text{full}}$ converges to zero in probability and its conditional distribution is asymptotically normal, namely*

$$\sqrt{n}\mathbf{V}_N^{-1/2}(\hat{\boldsymbol{\beta}}_{\text{sub}} - \hat{\boldsymbol{\beta}}_{\text{full}}) \xrightarrow{D|\mathcal{D}_N} \mathbb{N}(\mathbf{0}, \mathbf{I}), \tag{3}$$

where "$\xrightarrow{D|\mathcal{D}_N}$" means convergence in distribution conditional on the full data; $\mathbf{V}_N = \mathbf{M}_N^{-1}\mathbf{V}_{Nc}\mathbf{M}_N^{-1}$*;*

$$\mathbf{M}_N = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\phi}_i(\hat{\boldsymbol{\beta}}_{\text{full}}) \otimes (\mathbf{x}_i\mathbf{x}_i^{\text{T}}); \quad \mathbf{V}_{Nc} = \frac{1}{N}\sum_{i=1}^{N} \frac{\boldsymbol{\psi}_i(\hat{\boldsymbol{\beta}}_{\text{full}}) \otimes (\mathbf{x}_i\mathbf{x}_i^{\text{T}})}{N\pi_i}; \quad (4)$$

and $\boldsymbol{\psi}_i(\boldsymbol{\beta})$ *is a* $K \times K$ *matrix with the* k_1k_2*th element* $\psi_{i,k_1k_2}(\boldsymbol{\beta}) = \{\delta_{i,k_1} - p_{k_1}(\mathbf{x}_i, \boldsymbol{\beta})\}\{\delta_{i,k_2} - p_{k_2}(\mathbf{x}_i, \boldsymbol{\beta})\}$.

Remark 1 In this theorem, both n and N go to infinity, but there are no restrictions on their relative orders. Even if n is larger than N, the theorem is still valid. However, there is no computational benefit for oversampling, so it is typical that $n \ll N$ in practice. In addition, if one focuses on estimating the true parameter and replace $\hat{\boldsymbol{\beta}}_{\text{full}}$ with the true parameter in (3), then the asymptotic distribution is valid only if $n = o(N)$.

From Theorem 1, we see that the asymptotic variance-covariance matrix $n^{-1}\mathbf{V}_N$ depends on the subsampling probabilities. Thus, we can derive optimal subsampling probabilities which minimize the asymptotic variance-covariance matrix. There is no complete ordering for matrices, so we adopt the idea of optimal design of experiments and use some criterion function to induce a complete ordering. Specifically, we consider A-optimality and L-optimality (see Atkinson et al, 2007). A-optimality minimizes the trace of the variance-covariance matrix and "A" means this criterion minimizes the average of the variances for all parameter components. For our case, this is to minimize the trace of $\mathbf{V}_N$, $\text{tr}(\mathbf{V}_N)$. L-optimality minimizes the trace of the variance-covariance matrix for some linear transformation, say L, of the parameter estimator; here "L" stands for linear transformation. For our case, we take a specific case of $\text{L} = \mathbf{M}_N$, because with this choice the resultant criterion is to minimize the trace of $\mathbf{V}_{Nc}$, $\text{tr}(\mathbf{V}_{Nc})$, and the resultant optimal subsampling probabilities require less time to compute. Here $n^{-1}\mathbf{V}_{Nc}$ is the asymptotic variance-covariance matrix when we use $\mathbf{M}_N\hat{\boldsymbol{\beta}}_{\text{sub}}$ to approximate $\mathbf{M}_N\hat{\boldsymbol{\beta}}_{\text{full}}$.

The following theorem shows optimal subsampling probabilities under A- and L- optimality criteria.

Theorem 2 *The A-optimal subsampling probabilities that minimize* $\text{tr}(\mathbf{V}_N)$ *are*

$$\pi_i^{\text{optA}} = \frac{\|\mathbf{M}_N^{-1}\{\mathbf{s}_i(\hat{\boldsymbol{\beta}}_{\text{full}}) \otimes \mathbf{x}_i\}\|}{\sum_{j=1}^{N}\|\mathbf{M}_N^{-1}\{\mathbf{s}_j(\hat{\boldsymbol{\beta}}_{\text{full}}) \otimes \mathbf{x}_j\}\|} \quad i = 1, ..., N. \quad (5)$$

The L-optimal subsampling probabilities with $L = \mathbf{M}_N$ *that minimize* $\text{tr}(\mathbf{V}_{Nc})$ *are*

$$\pi_i^{\text{optL}} = \frac{\sqrt{\sum_{k=1}^{K}\{\delta_{i,k} - p_k(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_{\text{full}})\}^2}\|\mathbf{x}_i\|}{\sum_{j=1}^{N}\sqrt{\sum_{k=1}^{K}\{\delta_{j,k} - p_k(\mathbf{x}_j, \hat{\boldsymbol{\beta}}_{\text{full}})\}^2}\|\mathbf{x}_j\|} \quad i = 1, ..., N. \quad (6)$$

The L-optimal subsampling probabilities have a computational advantage because it takes $O(NKd)$ time to compute π_i^{optL} for $i = 1, ..., N$, while computing π_i^{optA} for $i = 1, ..., N$ takes $O(NK^2d^2)$ time.

Both π_i^{optA} and π_i^{optL} depend on the full data estimator $\hat{\boldsymbol{\beta}}_{\text{full}}$, which is the statistic we are approximating. This is similar to the dilemma in optimal design of experiments: one has to know the value of the unknown parameter to find an optimal design which is to be used to collect data to estimate the unknown parameter. We adopt the idea of two-stage adaptive optimal design (Lane et al, 2014), and take a first stage subsample to approximate optimal subsampling probabilities. An easy way to take the first stage subsample is to use uniform subsampling in which $\pi_i^{\text{uni}} = N^{-1}$. However, if the numbers of observations for different categories are very imbalanced, uniform subsampling may not work well because the probability of including no observation in some category can be high. In this scenario, we use subsampling probabilities proportional to the inverses of the numbers of observations in the corresponding categories. To be specific, let m_k be the number of observations for which the responses are in the kth category; that is, $m_k = \sum_{i=1}^{N} \delta_{i,k}$ for $k = 0, 1, ..., K$. Proportional subsampling probabilities are proportional to $\sum_{k=0}^{K} \delta_{i,k} m_k^{-1}$, i.e., $\pi_i^{\text{prop}} = \sum_{k=0}^{K} \delta_{i,k}\{(K+1)m_k\}^{-1}$.

For clear presentation, we describe the two-stage adaptive procedure in Algorithm 1.

Algorithm 1 Two-stage adaptive algorithm

(1) Randomly draw a subsample of size n_0 with replacement according to proportional subsampling probabilities π_i^{prop}. Use the Newton-Raphson method to obtian $\hat{\boldsymbol{\beta}}_{\text{sub}}^{0} = \arg\max_{\boldsymbol{\beta}} \ell_s^{*0}(\boldsymbol{\beta})$ where $\ell_s^{*0}(\boldsymbol{\beta})$ has the same expression as (2) with n and π_i replaced by n_0 and π_i^{prop}, respectively. Replace $\hat{\boldsymbol{\beta}}_{\text{full}}$ with $\hat{\boldsymbol{\beta}}_{\text{sub}}^{0}$ in (5) or (6) to obtain approximated optimal subsampling probabilities $\tilde{\pi}_i^{\text{optA}}$ or $\tilde{\pi}_i^{\text{optL}}$.

(2) According to the approximated optimal subsampling probabilities $\tilde{\pi}_i^{\text{optA}}$ or $\tilde{\pi}_i^{\text{optL}}$, randomly select a subsample of size n with replacement. Combine the subsamples and the corresponding probabilities in these two steps and implement the Newton-Raphson method to obtain

$$\hat{\boldsymbol{\beta}}_{\text{sub}}^{\text{ada}} = \arg\max_{\boldsymbol{\beta}} \Big\{ \frac{n_0}{n_0+n} \ell_s^{*0}(\boldsymbol{\beta}) + \frac{n}{n_0+n} \ell_s^{\text{opt}*}(\boldsymbol{\beta}) \Big\}, \tag{7}$$

where $\ell_s^{\text{opt}*}(\boldsymbol{\beta})$ has the same expression as (2) with π_i replaced by $\tilde{\pi}_i^{\text{optA}}$ or $\tilde{\pi}_i^{\text{optL}}$.

3 Simulation

This section uses numerical simulations to evaluate the performance of the two-stage adaptive procedure in Algorithm 1. Suppose that in model (1), the response y has three possible outcomes, c_0, c_1 and c_2, and the dimension d of the covariate $\mathbf{x}$ is three. Thus the dimension of the unknown parameter

vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}}$ is six. We set the true value of the parameter vector to $\boldsymbol{\beta} = (1, 1, 1, 2, 2, 2)^{\mathrm{T}}$.

We simulate full data sets of size $N = 10000$ for different distributions of $\mathbf{x}$. Specifically, we consider the following cases.

Case 1. The covariate $\mathbf{x}$ follows a multivariate normal distribution $\mathbb{N}_3(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a matrix with all diagonal elements equal to one and off-diagonal elements equal to 0.5. For this case, the proportions of observations for the three categories in the responses are around 0.415, 0.166, and 0.420.

Case 2. The covariate $\mathbf{x}$ follows a multivariate normal distribution $\mathbb{N}_3(\mathbf{1.5}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is the same as in Case 1. For this case, the responses in the full dataset are heavily imbalanced. More than 91% of responses are c_2, around 3% responses are c_0, and around 5% responses are c_1.

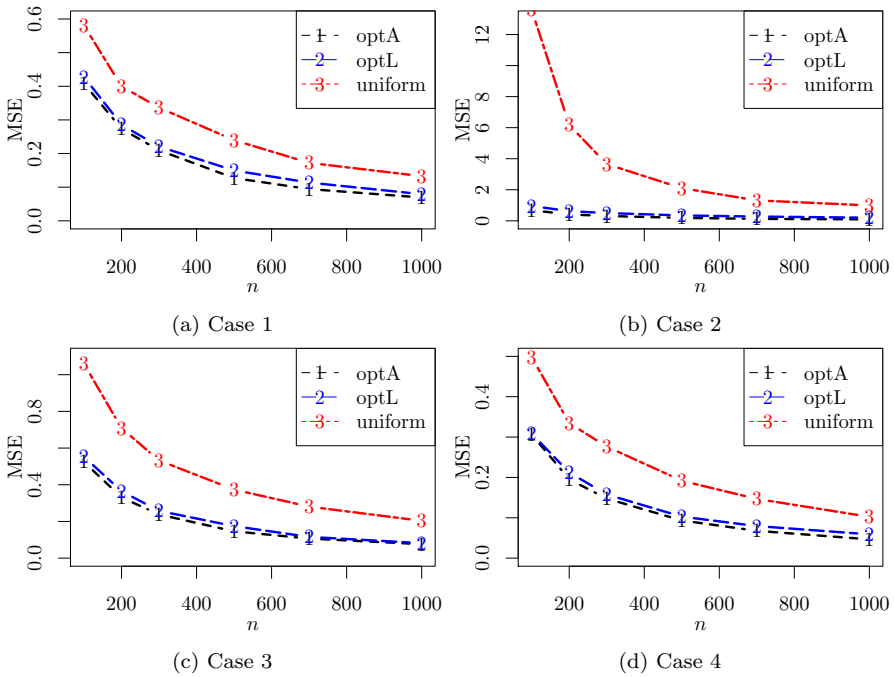
Case 3. The covariate $\mathbf{x}$ follows a mixture of two multivariate normal distributions $0.5\mathbb{N}_3(\mathbf{1}, \boldsymbol{\Sigma}) + 0.5\mathbb{N}_3(\mathbf{-1}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is the same as in Case 1. For this case, the proportions of observations for the three categories in the responses are around 0.448, 0.0968, and 0.455.

Case 4. The covariate $\mathbf{x}$ follows a multivariate t distribution with degrees of freedom of 3, $t_3(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is the same as in Case 1. This is a case that the covariate distribution has a heavier tail than normal distribution. The proportions of responses in the three categories are around 0.429, 0.152, and 0.419.

To evaluate the accuracy of Algorithm 1 in approximating the full data MLE, we implement it on the four full datasets corresponding to the aforementioned four cases. We repeat the implementation for $S = 1000$ times and calculate the empirical mean squared error (MSE) from $S^{-1}\sum_{s=1}^{S} ||\hat{\boldsymbol{\beta}}_{\mathrm{sub}}^{\mathrm{ada},(s)} - \hat{\boldsymbol{\beta}}_{\mathrm{full}}||^2$. For comparison, we also calculate the empirical MSE from uniform subsampling with a subsample size of $n_0 + n$.

Simulation results are presented in Figure 1. It shows that the two-stage adaptive procedure based on optimal subsampling probabilities is uniformly more efficient in approximating the full data MLE than the uniform subsampling method for all the four cases of covariate distributions. For the two-stage adaptive procedure, the performance based on π_i^{optA} is better than that based on π_i^{optL}, which is consistent with the theoretical results that π_i^{optA} is derived to minimize the asymptotic MSE of the subsample estimator.

We also perform simulations with a fixed total subsample size $n_0 + n$ and varying proportions of the two-stage subsample sizes. This gives us information about how different subsample size allocations between the two stages affect the performance of the two-stage procedure. Figure 2 gives results with $n_0 + n = 2000$ for Case 1. It shows that the two-stage algorithm does not have the best performance if n_0 is too small or too large. If n_0 is too small, the first stage estimator $\boldsymbol{\beta}_{\mathrm{sub}}^0$ may not be very accurate and thus the optimal probabilities may not be approximated well; if n_0 is relatively large and n is relatively small, then the more informative second stage sample is dominated by the first stage sample. The best approximation result is obtained when $\frac{n_0}{n_0+n}$ is around 0.1. Results in other cases are similar and are omitted to save space.



(a) Case 1 (b) Case 2

(c) Case 3 (d) Case 4

Fig. 1: Empirical MSEs of different methods for different second stage subsample sizes n when the first stage subsample size is fixed at $n_0 = 200$.

We also record the CPU times for the simulation studies. Results for Case 1 are given in Table 1, and results for other cases are omitted due to similarity. All computations were done using the R programming language (R Core Team, 2017) on a MacBook Pro with 2.5 GHz Intel Core i7 processor and 16 GB memory. Uniform subsampling uses less time than the two-stage algorithm based on π_i^{optA} or π_i^{optL} because it is a one step procedure and does not need to calculate the approximated optimal subsampling probabilities. As expected, the algorithm based on π_i^{optL} is faster than that based on π_i^{optA} due to its lower time complexity. The last row gives the CPU seconds for performing full data Newton-Raphson method for 1000 times, which is the longest one and confirms that the two-stage algorithm reduces computation burden.

Table 1: CPU seconds for different methods for Case 1 with $n_0 = 200$ and different n for 1000 repetitions.

Method	n					
	100	200	300	500	700	1000
optA	20.934	21.137	21.429	21.971	22.962	23.494
optL	15.437	15.991	15.909	16.687	17.653	18.283
Uniform	3.029	3.323	3.720	4.116	5.186	5.752
Full data CPU seconds: 33.286						

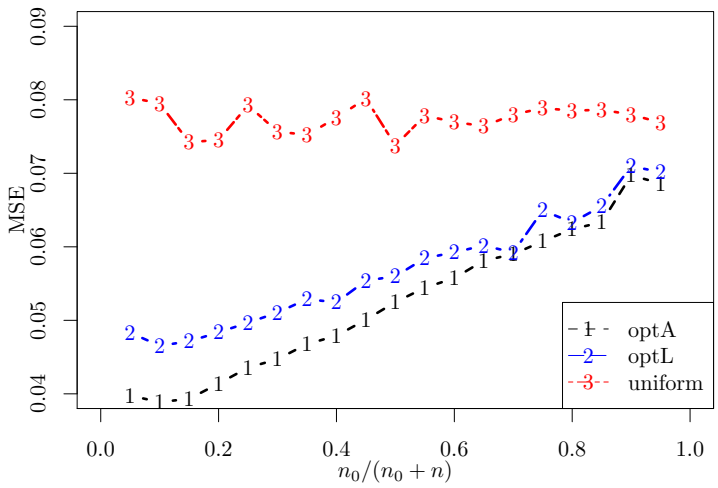


Fig. 2: MSEs of different ratios of first step subsample size to total subsample size for Case 1 when $n_0 + n = 2000$ is fixed.

To further investigate the performance of the two-stage adaptive algorithm in larger datasets, we enlarge the dimension of covariate to $d = 10$ (20 unknown parameters) and set the full data size to $N = 10^4$, 10^5 and 10^6. The covariate $\mathbf{x}$ is generated from the multivariate normal distribution as in Case 1. We set the mean of $\mathbf{x}$ to be $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Sigma}$ to be a matrix with diagonal elements equal to one and off-diagonal elements equal to 0.5. Table 2 presents the CPU seconds for repeating different methods for 200 times. The results indicate that the two-stage adaptive algorithm reduces computation burden dramatically compared with the full data MLE, and its advantage is more significant as the full data size increases.

Table 2: CPU seconds for 200 replications for different methods with a fixed $n_0 = 200$, $n = 1000$ and different N when $d = 10$.

Method	N		
	10^4	10^5	10^6
optA	12.29	89.99	665.55
optL	6.56	24.95	217.87
Uniform	3.44	4.07	8.77
Full	25.64	258.46	2636.99

4 Summary and future work

For the softmax regression model with massive data, we have established the asymptotic normality of the general subsampling estimator, and then derived

optimal subsampling probabilities under the A-optimality criterion and the L-optimality with a specific L. We have used the two-stage adaptive procedure to address the issue that the optimal subsampling probabilities depend on the full data estimator, and used numerical simulations to evaluate its performance.

There are some important questions we will investigate further. 1) We use a first stage subsample to obtain a pilot estimator to approximate optimal subsampling probabilities. We believe that, under reasonable assumptions, the resulting estimator is also asymptotically normal with the optimal asymptotic variance-covariance matrix. 2) Wang (2018) proposed to use the unweighted MLE of an optimal subsample for logistic regression and then correct the bias. This method depends on the special structure of the binary logistic regression model. Whether it can be extended to multinomial logistic regression requires further investigations.

Acknowledgments

We gratefully acknowledge the comments from two referees that helped improve the paper.

Appendix: Proofs

We prove the two theorems in this section. We use $O_{P|\mathcal{D}_N}(1)$ and $o_{P|\mathcal{D}_N}(1)$ to denote boundedness and convergence to zero, respectively, in conditional probability given the full data. Specifically for a sequence of random vector $\mathbf{v}_{n,N}$, as $n \to \infty$ and $N \to \infty$, $\mathbf{v}_{n,N} = O_{P|\mathcal{D}_N}(1)$ means that for any $\epsilon > 0$, there exists a finite $C_\epsilon > 0$ such that

$$\mathbb{P}\Big\{ \sup_n \mathbb{P}(\|\mathbf{v}_{n,N}\| > C_\epsilon | \mathcal{D}_N) \le \epsilon \Big\} \to 1;$$

$v_{n,N} = o_{P|\mathcal{D}_N}(1)$ means that for any $\epsilon > 0$ and δ,

$$\mathbb{P}\Big\{ \mathbb{P}(\|\mathbf{v}_{n,N}\| > \delta | \mathcal{D}_N) \le \epsilon \Big\} \to 1.$$

Proof (**Theorem 1**) By direct calculation under the conditional distribution of the subsample given $\mathcal{D}_N$, we have

$$\begin{aligned} \mathbb{E}\left\{ \ell_s^*(\boldsymbol{\beta}) \middle| \mathcal{D}_N \right\} &= \ell_f(\boldsymbol{\beta}), \\ \mathbb{E}\left\{ \ell_s^*(\boldsymbol{\beta}) - \ell_f(\boldsymbol{\beta}) \middle| \mathcal{D}_N \right\}^2 &= \frac{1}{n} \left[\frac{1}{N^2} \sum_{i=1}^N \frac{t_i^2(\boldsymbol{\beta})}{\pi_i} - \ell_f^2(\boldsymbol{\beta}) \right], \end{aligned} \tag{8}$$

where $t_i(\boldsymbol{\beta}) = \sum_{k=1}^K \delta_{i,k} \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_k - \log\left\{ 1 + \sum_{l=1}^K e^{\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_l} \right\}$. Note that

$$
\begin{aligned}
|t_i(\boldsymbol{\beta})| &\leq \sum_{k=1}^{K} \|\mathbf{x}_i\| \|\boldsymbol{\beta}_k\| + \log\left(1 + \sum_{k=1}^{K} e^{\|\mathbf{x}_i\| \|\boldsymbol{\beta}_k\|}\right) \\
&\leq K\|\mathbf{x}_i\| \|\boldsymbol{\beta}\| + \log\left(1 + Ke^{\|\mathbf{x}_i\| \|\boldsymbol{\beta}\|}\right) \\
&\leq K\|\mathbf{x}_i\| \|\boldsymbol{\beta}\| + 1 + \log K + \|\mathbf{x}_i\| \|\boldsymbol{\beta}\| \\
&= (K+1)\|\mathbf{x}_i\| \|\boldsymbol{\beta}\| + 1 + \log K,
\end{aligned}
$$

where the second inequality is from the fact that $\|\boldsymbol{\beta}_k\| \leq \|\boldsymbol{\beta}\|$, and the third inequality is from the fact that $\log(1+x) < 1 + \log x$ for $x \geq 1$. Therefore, from Assumption 2,

$$
\frac{1}{n^2}\sum_{i=1}^{N} \frac{t_i^2(\boldsymbol{\beta})}{\pi_i} - \ell_f^2(\boldsymbol{\beta}) \leq \frac{1}{n^2}\sum_{i=1}^{N} \frac{t_i^2(\boldsymbol{\beta})}{\pi_i} + \left(\frac{1}{n}\sum_{i=1}^{N} |t_i(\boldsymbol{\beta})|\right)^2 = O_P(1). \tag{9}
$$

Combining (8) and (9), $\ell_s^*(\boldsymbol{\beta}) - \ell_f(\boldsymbol{\beta}) \to 0$ in conditional probability given $\mathcal{D}_N$. Note that the parameter space is compact, and $\hat{\boldsymbol{\beta}}_{\text{sub}}$ and $\hat{\boldsymbol{\beta}}_{\text{full}}$ are the unique global maximums of the continuous concave functions $\ell_s^*(\boldsymbol{\beta})$ and $\ell_f(\boldsymbol{\beta})$, respectively. Thus, from Theorem 5.9 and its remark of van der Vaart (1998), we obtain that conditionally on $\mathcal{D}_N$ in probability,

$$
\|\hat{\boldsymbol{\beta}}_{\text{sub}} - \hat{\boldsymbol{\beta}}_{\text{full}}\| = o_{P|\mathcal{D}_N}(1). \tag{10}
$$

From Taylor's theorem (c.f. Chapter 4 of Ferguson, 1996),

$$
0 = \dot{\ell}_{s,j}^*(\hat{\boldsymbol{\beta}}_{\text{sub}}) = \dot{\ell}_{s,j}^*(\hat{\boldsymbol{\beta}}_{\text{full}}) + \frac{\partial \dot{\ell}_{s,j}^*(\hat{\boldsymbol{\beta}}_{\text{full}})}{\partial \boldsymbol{\beta}^T}(\hat{\boldsymbol{\beta}}_{\text{sub}} - \hat{\boldsymbol{\beta}}_{\text{full}}) + R_j \tag{11}
$$

where $\dot{\ell}_{s,j}^*(\boldsymbol{\beta})$ is the partial derivative of $\ell_s^*(\boldsymbol{\beta})$ with respect to β_j, and

$$
R_j = (\hat{\boldsymbol{\beta}}_{\text{sub}} - \hat{\boldsymbol{\beta}}_{\text{full}})^T \int_0^1 \int_0^1 \frac{\partial^2 \dot{\ell}_{s,j}^*\{\hat{\boldsymbol{\beta}}_{\text{full}} + uv(\hat{\boldsymbol{\beta}}_{\text{sub}} - \hat{\boldsymbol{\beta}}_{\text{full}})\}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} v \mathrm{d}u \mathrm{d}v \, (\hat{\boldsymbol{\beta}}_{\text{sub}} - \hat{\boldsymbol{\beta}}_{\text{full}}).
$$

Note that the third partial derivative of the log-likelihood for the subsample takes the form of

$$
\frac{\partial^3 \ell_s^*(\boldsymbol{\beta})}{\partial \beta_{j_1} \partial \beta_{j_2} \partial \beta_{j_3}} = \sum_{i=1}^{n} \frac{\alpha_{i,j_1 j_2 j_3} x_{i,j_1'} x_{i,j_2'} x_{i,j_3'}}{nN\pi_i^*},
$$

where $j_l' = \text{Rem}(j_l/d) + dI\{\text{Rem}(j_l/d) = 0\}$, $l = 1, 2, 3$, $\text{Rem}(j_l/d)$ is the reminder of the integer division j_l/d, and $\alpha_{i,j_1 j_2 j_3}$ satisfies that $|\alpha_{j_1 j_2 j_3}| \leq 2$. Here $|\alpha_{j_1 j_2 j_3}| \leq 2$ because it has a form of $p_{k'}(\mathbf{x}_i, \boldsymbol{\beta})\{1 - p_{k'}(\mathbf{x}_i, \boldsymbol{\beta})\}\{1 - 2p_{k'}(\mathbf{x}_i, \boldsymbol{\beta})\}$, $p_{k_1'}(\mathbf{x}_i, \boldsymbol{\beta})p_{k_2'}(\mathbf{x}_i, \boldsymbol{\beta})\{2p_{k_2'}(\mathbf{x}_i, \boldsymbol{\beta}) - 1\}$, or $2p_{k_1'}(\mathbf{x}_i, \boldsymbol{\beta})p_{k_2'}(\mathbf{x}_i, \boldsymbol{\beta})p_{k_3'}(\mathbf{x}_i, \boldsymbol{\beta})$ for some k' and $k_1' \neq k_2' \neq k_3'$. Thus,

$$
\left\| \frac{\partial^2 \dot{\ell}_{s,j}^*(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right\| \leq \frac{2}{n} \sum_{i=1}^{n} \frac{K\|\mathbf{x}_i^*\|^3}{N\pi_i^*}
$$

for all $\boldsymbol{\beta}$. This gives us that

$$\sup_{u,v}\left\|\frac{\partial^2\dot{\ell}^*_{s,j}\{\hat{\boldsymbol{\beta}}_{\text{full}}+uv(\hat{\boldsymbol{\beta}}_{\text{sub}}-\hat{\boldsymbol{\beta}}_{\text{full}})\}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right\|=O_{P|\mathcal{D}_N}(1), \tag{12}$$

because

$$P\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\|\mathbf{x}_i^*\|^3}{N\pi_i^*}\geq\tau\middle|\mathcal{D}_N\right)\leq\frac{1}{nN\tau}\sum_{i=1}^{n}\mathbb{E}\left(\frac{\|\mathbf{x}_i^*\|^3}{\pi_i^*}\middle|\mathcal{D}_N\right)=\frac{1}{N\tau}\sum_{i=1}^{N}\|\mathbf{x}_i\|^3\to 0,$$

in probability as $\tau\to\infty$ by Assumption 1. From (12), we have that

$$R_j=O_{P|\mathcal{D}_N}(\|\hat{\boldsymbol{\beta}}_{\text{sub}}-\hat{\boldsymbol{\beta}}_{\text{full}}\|^2). \tag{13}$$

Denote $\mathbf{M}_n^*=\partial^2\dot{\ell}_s^*(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T=n^{-1}\sum_{i=1}^{n}(N\pi_i^*)^{-1}\boldsymbol{\phi}_i^*(\hat{\boldsymbol{\beta}}_{\text{full}})\otimes(\mathbf{x}_i^*\mathbf{x}_i^{*\text{T}})$. From (11) and (12), we have

$$\hat{\boldsymbol{\beta}}_{\text{sub}}-\hat{\boldsymbol{\beta}}_{\text{full}}=-\mathbf{M}_n^{*-1}\left\{\dot{\ell}_s^*(\hat{\boldsymbol{\beta}}_{\text{full}})+O_{P|\mathcal{D}_N}(\|\hat{\boldsymbol{\beta}}_{\text{sub}}-\hat{\boldsymbol{\beta}}_{\text{full}}\|^2)\right\}. \tag{14}$$

By direct calculation, we know that

$$\mathbb{E}(\mathbf{M}_n^*|\mathcal{D}_N)=\mathbf{M}_N. \tag{15}$$

For any component $\mathbf{M}_n^{*j_1j_2}$ of $\mathbf{M}_n^*$ where $1\leq j_1,j_2\leq d$,

$$\begin{aligned}\mathbb{V}\left(\mathbf{M}_n^{*j_1j_2}|\mathcal{D}_N\right)=&\frac{1}{n}\sum_{i=1}^{N}\pi_i\left\{\frac{\{\boldsymbol{\phi}_i(\hat{\boldsymbol{\beta}}_{\text{full}})\otimes(\mathbf{x}_i\mathbf{x}_i^{\text{T}})\}^{j_1j_2}}{N\pi_i}-\mathbf{M}_N^{j_1j_2}\right\}^2\\=&\frac{1}{nN^2}\sum_{i=1}^{N}\frac{[\{\boldsymbol{\phi}_i(\hat{\boldsymbol{\beta}}_{\text{full}})\otimes(\mathbf{x}_i\mathbf{x}_i^{\text{T}})\}^{j_1j_2}]^2}{\pi_i}-\frac{1}{n}(\mathbf{M}_N^{j_1j_2})^2\\\leq&\frac{1}{nN^2}\sum_{i=1}^{N}\frac{\|\mathbf{x}_i\|^4}{\pi_i}=O_P(n^{-1}),\end{aligned}$$

where the second last inequality holds by the fact that all elements of $\boldsymbol{\phi}_i$ are between 0 and 1, and the last equality is from Assumption 2. This result combined with Markov's inequality and (15), implies that

$$\mathbf{M}_n^*-\mathbf{M}_N=O_{P|\mathcal{D}_N}(n^{-1/2}). \tag{16}$$

By direct calculation, we have

$$\mathbb{E}\left\{\frac{\partial\ell_s^*(\hat{\boldsymbol{\beta}}_{\text{full}})}{\partial\boldsymbol{\beta}}\middle|\mathcal{D}_N\right\}=\frac{\partial\ell_f(\hat{\boldsymbol{\beta}}_{\text{full}})}{\partial\boldsymbol{\beta}}=0. \tag{17}$$

Note that

$$\mathbb{V}\left\{\frac{\partial\ell_s^*(\hat{\boldsymbol{\beta}}_{\text{full}})}{\partial\boldsymbol{\beta}}\middle|\mathcal{D}_N\right\}=\frac{1}{nN^2}\sum_{i=1}^{N}\frac{\boldsymbol{\psi}_i(\hat{\boldsymbol{\beta}}_{\text{full}})\otimes(\mathbf{x}_i\mathbf{x}_i^{\text{T}})}{\pi_i}, \tag{18}$$

whose elements are bounded by $(nN^2)^{-1}\sum_{i=1}^{N}\pi_i^{-1}\|\mathbf{x}_i\|^2$ which is of order $O_P(n^{-1})$ by Assumption 2. From (17), (18) and Markov's inequality, we know that

$$\frac{\partial \ell_s^*(\hat{\boldsymbol{\beta}}_{\text{full}})}{\partial \boldsymbol{\beta}} = O_{P|\mathcal{D}_N}(n^{-1/2}). \tag{19}$$

Note that (16) indicates that $\mathbf{M}_n^{*-1} = O_{P|\mathcal{D}_N}(1)$. Combining this with (10), (14) and (19), we have

$$\hat{\boldsymbol{\beta}}_{\text{sub}} - \hat{\boldsymbol{\beta}}_{\text{full}} = O_{P|\mathcal{D}_N}(n^{-1/2}) + o_{P|\mathcal{D}_N}(\|\hat{\boldsymbol{\beta}}_{\text{sub}} - \hat{\boldsymbol{\beta}}_{\text{full}}\|),$$

which implies that

$$\hat{\boldsymbol{\beta}}_{\text{sub}} - \hat{\boldsymbol{\beta}}_{\text{full}} = O_{P|\mathcal{D}_N}(n^{-1/2}). \tag{20}$$

Note that

$$\dot{\ell}_s^*(\hat{\boldsymbol{\beta}}_{\text{full}}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbf{s}_i^*(\hat{\boldsymbol{\beta}}_{\text{full}})\otimes\mathbf{x}_i^*}{N\pi_i^*} \equiv \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\eta}_i \tag{21}$$

Given $\mathcal{D}_N$, $\boldsymbol{\eta}_1, ..., \boldsymbol{\eta}_n$ are i.i.d, with mean $\mathbf{0}$ and variance,

$$\mathbb{V}(\boldsymbol{\eta}_i|\mathcal{D}_N) = \mathbf{V}_{Nc} = O_P(1). \tag{22}$$

Meanwhile, for every $\varepsilon > 0$ and some $\rho > 0$,

$$\begin{aligned}
&\sum_{i=1}^{n}\mathbb{E}\{\|n^{-1/2}\boldsymbol{\eta}_i\|^2 I(\|\boldsymbol{\eta}_i\| > n^{1/2}\varepsilon)|\mathcal{D}_N\} \\
&\le \frac{1}{n^{1+\rho/2}\varepsilon^\rho}\sum_{i=1}^{n}\mathbb{E}\{\|\boldsymbol{\eta}_i\|^{2+\rho} I(\|\boldsymbol{\eta}_i\| > n^{1/2}\varepsilon)|\mathcal{D}_N\} \\
&\le \frac{1}{n^{1+\rho/2}\varepsilon^\rho}\sum_{i=1}^{n}\mathbb{E}(\|\boldsymbol{\eta}_i\|^{2+\rho}|\mathcal{D}_N) \\
&= \frac{1}{n^{\rho/2}}\frac{1}{N^{2+\rho}}\frac{1}{\varepsilon^\rho}\sum_{i=1}^{N}\frac{\|\mathbf{s}_i(\hat{\boldsymbol{\beta}}_{\text{full}})\|^{2+\rho}\|\mathbf{x}_i\|^{2+\rho}}{\pi_i^{1+\rho}} \\
&\le \frac{1}{n^{\rho/2}}\frac{1}{N^{2+\rho}}\frac{1}{\varepsilon^\rho}\sum_{i=1}^{N}\frac{\|\mathbf{x}_i\|^{2+\rho}}{\pi_i^{1+\rho}} = o_P(1)
\end{aligned}$$

where the last equality is from Assumption 2. This and (22) show that the Lindeberg-Feller conditions are satisfied in the conditional distribution in probability. From (21) and (22), by the Lindeberg-Feller central limit theorem (Proposition 2.27 of van der Vaart, 1998), conditionally on $\mathcal{D}_N$ in probability,

$$n^{1/2}\mathbf{V}_{Nc}^{-1/2}\dot{\ell}_s^*(\hat{\boldsymbol{\beta}}_{\text{full}}) = \frac{1}{n^{1/2}}\{\mathbb{V}(\boldsymbol{\eta}_i|\mathcal{D}_N)\}^{-1/2}\sum_{i=1}^{n}\boldsymbol{\eta}_i \to \mathbb{N}(0, \mathbf{I}),$$

in distribution. From (14) and (20),

$$\hat{\boldsymbol{\beta}}_{\text{sub}} - \hat{\boldsymbol{\beta}}_{\text{full}} = -\mathbf{M}_n^{*-1}\dot{\ell}_s^*(\hat{\boldsymbol{\beta}}_{\text{full}}) + O_{P|\mathcal{D}_N}(n^{-1}). \tag{23}$$

From (16),

$$\mathbf{M}_n^{*-1} - \mathbf{M}_N^{-1} = -\mathbf{M}_N^{-1}(\mathbf{M}_n^* - \mathbf{M}_N)\mathbf{M}_n^{*-1} = O_{P|\mathcal{D}_N}(n^{-1/2}). \tag{24}$$

Based on Assumption 1 and (22), it is verified that,

$$\mathbf{V} = \mathbf{M}_N^{-1}\mathbf{V}_{Nc}\mathbf{M}_N^{-1} = O_P(1). \tag{25}$$

Thus, (23), (24) and (25) yield,

$$\begin{aligned}
&n^{1/2}\mathbf{V}^{-1/2}(\hat{\boldsymbol{\beta}}_{\text{sub}} - \hat{\boldsymbol{\beta}}_{\text{full}}) \\
&= -n^{1/2}\mathbf{V}^{-1/2}\mathbf{M}_n^{*-1}\dot{\ell}_s^*(\hat{\boldsymbol{\beta}}_{\text{full}}) + O_{P|\mathcal{D}_N}(n^{-1/2}) \\
&= -\mathbf{V}^{-1/2}\mathbf{M}_N^{-1}n^{1/2}\dot{\ell}_s^*(\hat{\boldsymbol{\beta}}_{\text{full}}) - \mathbf{V}^{-1/2}(\mathbf{M}_n^{*-1} - \mathbf{M}_N^{-1})n^{1/2}\dot{\ell}_s^*(\hat{\boldsymbol{\beta}}_{\text{full}}) + O_{P|\mathcal{D}_N}(n^{-1/2}) \\
&= -\mathbf{V}^{-1/2}\mathbf{M}_N^{-1}\mathbf{V}_{Nc}^{1/2}\mathbf{V}_{Nc}^{-1/2}n^{-1/2}\dot{\ell}_s^*(\hat{\boldsymbol{\beta}}_{\text{full}}) + O_{P|\mathcal{D}_N}(n^{-1/2}).
\end{aligned}$$

The result in Theorem 1 follows from Slutsky's Theorem(Theorem 6 of Ferguson, 1996) and the fact that

$$\mathbf{V}^{-1/2}\mathbf{M}_N^{-1}\mathbf{V}_{Nc}^{1/2}(\mathbf{V}^{-1/2}\mathbf{M}_N^{-1}\mathbf{V}_{Nc}^{1/2})^T = \mathbf{V}^{-1/2}\mathbf{M}_N^{-1}\mathbf{V}_{Nc}^{1/2}\mathbf{V}_{Nc}^{1/2}\mathbf{M}_N^{-1}\mathbf{V}^{-1/2} = \mathbf{I}.$$

Proof (**Theorem 2**) Note that $\boldsymbol{\psi}_i(\hat{\boldsymbol{\beta}}_{\text{full}}) = \mathbf{s}_i(\hat{\boldsymbol{\beta}}_{\text{full}})\mathbf{s}_i^{\mathrm{T}}(\hat{\boldsymbol{\beta}}_{\text{full}})$, so

$$\boldsymbol{\psi}_i(\hat{\boldsymbol{\beta}}_{\text{full}}) \otimes (\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}) = \{\mathbf{s}_i(\hat{\boldsymbol{\beta}}_{\text{full}}) \otimes \mathbf{x}_i\}\{\mathbf{s}_i^{\mathrm{T}}(\hat{\boldsymbol{\beta}}_{\text{full}}) \otimes \mathbf{x}_i^{\mathrm{T}}\}. \tag{26}$$

Therefore, for the A-optimality,

$$\begin{aligned}
\text{tr}(\mathbf{V}_N) &= \text{tr}(\mathbf{M}_N^{-1}\mathbf{V}_{Nc}\mathbf{M}_N^{-1}) \\
&= \frac{1}{N^2}\text{tr}\left\{\mathbf{M}_N^{-1}\sum_{i=1}^N \frac{\boldsymbol{\psi}_i(\hat{\boldsymbol{\beta}}_{\text{full}}) \otimes (\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}})}{\pi_i}\mathbf{M}_N^{-1}\right\} \\
&= \frac{1}{N^2}\sum_{i=1}^N \frac{\text{tr}\left[\mathbf{M}_N^{-1}\{\mathbf{s}_i(\hat{\boldsymbol{\beta}}_{\text{full}}) \otimes \mathbf{x}_i\}\{\mathbf{s}_i^{\mathrm{T}}(\hat{\boldsymbol{\beta}}_{\text{full}}) \otimes \mathbf{x}_i^{\mathrm{T}}\}\mathbf{M}_N^{-1}\right]}{\pi_i} \\
&= \frac{1}{N^2}\sum_{i=1}^N \frac{\|\mathbf{M}_N^{-1}\{\mathbf{s}_i(\hat{\boldsymbol{\beta}}_{\text{full}}) \otimes \mathbf{x}_i\}\|^2}{\pi_i} \times \sum_{i=1}^N \pi_i \\
&\geq \left\{\frac{1}{N}\sum_{i=1}^N \|\mathbf{M}_N^{-1}\{\mathbf{s}_i(\hat{\boldsymbol{\beta}}_{\text{full}}) \otimes \mathbf{x}_i\}\|\right\}^2.
\end{aligned}$$

Here, the last step due to the Cauchy-Schwarz inequality, and the equality holds if and only if π_i is proportional to $\|\mathbf{M}_N^{-1}\{\mathbf{s}_i(\hat{\boldsymbol{\beta}}_{\text{full}}) \otimes \mathbf{x}_i\}\|$. Thus, the A-optimal subsampling probabilities take the form of (5).

For the L-optimality, the proof is similar by noticing that

$$\begin{aligned}
\mathrm{tr}(\mathbf{V}_{Nc}) &= \frac{1}{N^2}\sum_{i=1}^{N}\frac{\mathrm{tr}\{\boldsymbol{\psi}_i(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\otimes(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}})\}}{\pi_i} \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\frac{\mathrm{tr}\{\boldsymbol{\psi}_i(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\}\mathrm{tr}(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}})}{\pi_i} = \frac{1}{N^2}\sum_{i=1}^{N}\frac{\|\mathbf{s}_i(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\|^2\|\mathbf{x}_i\|^2}{\pi_i} \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\frac{\big[\sum_{k=1}^{K}\{\delta_{i,k}-p_k(\mathbf{x}_i,\hat{\boldsymbol{\beta}}_{\mathrm{full}})\}^2\big]\|\mathbf{x}_i\|^2}{\pi_i}\times\sum_{i=1}^{N}\pi_i \\
&\geq \left(\frac{1}{N}\sum_{i=1}^{N}\left[\sum_{k=1}^{K}\{\delta_{i,k}-p_k(\mathbf{x}_i,\hat{\boldsymbol{\beta}}_{\mathrm{full}})\}\right]^{1/2}\|\mathbf{x}_i\|\right)^2.
\end{aligned}$$

References

Atkinson A, Donev A, Tobias R (2007) Optimum experimental designs, with SAS, vol 34. Oxford University Press

Drineas P, Kannan R, Mahoney MW (2006a) Fast monte carlo algorithms for matrices i: Approximating matrix multiplication. SIAM Journal on Computing 36(1):132–157

Drineas P, Kannan R, Mahoney MW (2006b) Fast monte carlo algorithms for matrices ii: Computing a low-rank approximation to a matrix. SIAM Journal on Computing 36(1):158–183

Drineas P, Kannan R, Mahoney MW (2006c) Fast monte carlo algorithms for matrices iii: Computing a compressed approximate matrix decomposition. SIAM Journal on Computing 36(1):184–206

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

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

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

Ferguson TS (1996) A Course in Large Sample Theory. Chapman and Hall

Frieze A, Kannan R, Vempala S (2004) Fast monte-carlo algorithms for finding low-rank approximations. Journal of the ACM 51:1025–1041

Lane A, Yao P, Flournoy N (2014) Information in a two-stage adaptive optimal design. Journal of Statistical Planning and Inference 144:173–187

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

Mahoney MW (2011) Randomized algorithms for matrices and data. Foundations and Trends® in Machine Learning 3(2):123–224

Mahoney MW, Drineas P (2009) CUR matrix decompositions for improved data analysis. Proceedings of the National Academy of Sciences 106(3):697–702

Ortega JM, Rheinboldt WC (1970) Iterative solution of nonlinear equations in several variables, vol 30. Siam

R Core Team (2017) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/

Raskutti G, Mahoney M (2016) A statistical perspective on randomized sketching for ordinary least-squares. Journal of Machine Learning Research 17:1–31

van der Vaart A (1998) Asymptotic Statistics. Cambridge University Press, London

Wang H (2018) More efficient estimation for logistic regression with optimal subsample. arXiv preprint arXiv:180202698

Wang H, Yang M, Stufken J (2018a) Information-based optimal subdata selection for big data linear regression. Journal of the American Statistical Association 0(0):1–13, URL https://doi.org/10.1080/01621459.2017.1408468

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