Model Constraints Independent Optimal Subsampling Probabilities for Softmax Regression

Yaqiong Yao[a,*], Jiahui Zou[b], HaiYing Wang[a]

[a]*Department of Statistics, University of Connecticut, Storrs, CT 06269, USA*
[b]*School of Statistics, Capital University of Economics and Business, Beijing 100070, China*

Abstract

A prevailing method to alleviate the computational cost is to perform analysis on a subsample of the full data. Optimal subsampling algorithm utilizes non-uniform subsampling probabilities, derived through minimizing the asymptotic mean squared error of the subsample estimator, to acquire a higher estimation efficiency for a given subsample size. The optimal subsampling probabilities for softmax regression have been studied under the baseline constraint which treats one dimension of the multivariate response differently from other dimensions. In this paper, we show that different model constraints lead to different optimal subsampling probabilities, and the summation constraint corresponds to a better subsampling strategy than the baseline constraint in terms of balancing the responses among all categories. Furthermore, we derive the asymptotic distribution of the mean squared prediction error, and minimize its asymptotic expectation to define the optimal subsampling probabilities that are invariant to model constraints. Simulations and a real data example are provided to show the effectiveness of the proposed optimal subsampling probabilities.

Keywords: Mean squared prediction error, Model constraints, Optimal subsampling probabilities, Softmax regression.

1. Introduction

Analyzing massive datasets challenges statisticians in two ways. Firstly, analyzing a massive data set requires a large computer memory to read in the data, and secondly, a long CPU time is needed for calculating the results. Subsampling is a practical solution to reduce the computational time by using a sample of the full data in the analysis process. For softmax regression, the optimal subsampling algorithm has been investigated in [1] under the baseline constraint, where one dimension of the multivariate response variable is set as the baseline and the corresponding parameter is set to be a vector of zeros. With this constraint, the resulting optimal subsampling probabilities treat the

*Corresponding author. Email address: yaqiong.yao@uconn.edu

baseline category differently from other categories, and this may cause imbalanced responses in a resulting subsample. To solve this, we construct the optimal subsampling probabilities under a summation constraint to ensure that all dimensions are treated equally. It can be seen that optimal subsampling probabilities, constructed by minimizing the asymptotic mean squared error of the subsample estimator, vary with the choice of model constraints. To deal with this problem, we formulate novel optimal subsampling probabilities that are independent to all model constraints by minimizing the asymptotic expectation of the mean squared prediction error in this paper.

Subsampling methods draw observations based on the pre-specified subsampling probabilities. Uniform subsampling probabilities are the easiest ones to compute; however, each observation is of equal significance in the sampling procedure regardless of how important one observation is. To avoid this problem, random projection method utilizes randomized Hadamard transform to integrate the information of all observations and then conduct uniform subsampling [2, 3]. Another way doing random sampling is to assign non-uniform subsampling probabilities to every observation. Algorithmic leveraging uses the statistical leveraging scores as the non-uniform subsampling probabilities for linear regression [4, 5], and its statistical properties and asymptotic distributions of the resulting estimators were studied in [6] and [7], respectively. Local case-control sampling utilizes both covariates and responses to formulate the subsampling probabilities for logistic regression with imbalanced responses [8]. Local uncertainty sampling generalizes this idea to softmax regression and derives the conditional maximum likelihood estimator for the sampled data [9]. Optimal subsampling was first proposed by [10] for logistic regression, which defines the optimal subsampling probabilities by minimizing the asymptotic variance-covariance matrix of the subsample estimator under A- and L- optimality criteria. Besides logistic regression, the optimal subsampling approach shows superior performance on other models, e.g. generalized linear models [11], quasi likelihood estimator [12] and quantile regression [13].

Softmax regression is used to model the relationship between multiple negative correlated binary outcomes and covariates. Suppose that an experiment has $K+1$ possible outcomes. Consider a dataset $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$, where $\{\mathbf{x}_i\}_{i=1}^N$ are d dimensional covariates, and $\{\mathbf{y}_i\}_{i=1}^N$ are $K+1$ dimensional multivariate responses. Each element of $\mathbf{y}_i$ is an indicator for the corresponding outcome with $y_{i,k} = 1$ if the k-th category occurs for $k \in \{0, 1, 2, ..., K\}$, and $\{y_{i,k}\}_{k=0}^K$ are dependent such that $\sum_{k=0}^K y_{i,k} = 1$. A softmax regression model assumes that for each observation,

$$\Pr(y_{i,k} = 1|\mathbf{x}_i) = p_k(\mathbf{x}_i, \boldsymbol{\beta}) = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta}_k)}{\sum_{l=0}^K \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_l)}, \tag{1}$$

where $\boldsymbol{\beta}_k, k = 0, 1, ..., K$, are d dimensional regression coefficients, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_0^\top, \boldsymbol{\beta}_1^\top, ..., \boldsymbol{\beta}_K^\top)^\top$ is the $(K+1)d$ dimensional vector. The mean response vector is denoted as $\boldsymbol{p}_i(\boldsymbol{\beta}) = \{p_0(\mathbf{x}_i, \boldsymbol{\beta}), p_1(\mathbf{x}_i, \boldsymbol{\beta}), ..., p_K(\mathbf{x}_i, \boldsymbol{\beta})\}^\top$, that is $\mathbb{E}(\mathbf{y}_i|\mathbf{x}_i) = \boldsymbol{p}_i(\boldsymbol{\beta})$. The maximum likelihood estimator (MLE), $\hat{\boldsymbol{\beta}}_{\text{full}}$, is often considered to

estimate $\boldsymbol{\beta}$ and obtained by maximizing the log-likelihood function

$$\ell(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N}\Big[\sum_{k=0}^{K} y_{i,k}\mathbf{x}_i^\top\boldsymbol{\beta}_k - \ln\Big\{\sum_{l=0}^{K}\exp(\mathbf{x}_i^\top\boldsymbol{\beta}_l)\Big\}\Big]$$

subject to a model constraint because the model (1) is not identifiable. As there is no general closed form solution to the MLE, the Newton-Raphson method is implemented to acquire $\hat{\boldsymbol{\beta}}_{\text{full}}$ iteratively, which takes $O(\eta NK^2d^2)$ time for $N > Kd$ where η is the number of iterations. Subsampling method can effectively reduce the computational burden for datasets with extremely large N by using a subsample estimator to approximate the full data MLE.

Let $\{\pi_i\}_{i=1}^N$ be subsampling probabilities for each observation in the full dataset with $\sum_{i=1}^N \pi_i = 1$. Draw a subsample of size n with replacement based on $\{\pi_i\}_{i=1}^N$. The samples and the corresponding subsampling probabilities are denoted as $\{\mathbf{y}_i^*, \mathbf{x}_i^*, \pi_i^*\}_{i=1}^n$. The subsample estimator $\hat{\boldsymbol{\beta}}_{\text{sub}}$ is obtained by maximizing

$$\ell_s^*(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{n}\frac{1}{n\pi_i^*}\Big[\sum_{k=0}^{K} y_{i,k}^*{\boldsymbol{\beta}_k}^\top\mathbf{x}_i^* - \ln\Big\{\sum_{l=0}^{K}\exp({\boldsymbol{\beta}_l}^\top\mathbf{x}_i^*)\Big\}\Big],$$

subject to a model constraint. The optimal subsampling probabilities under the baseline constraint were formulated in [1] for softmax regression.

However, the optimal subsampling probabilities in [1] handle the baseline category differently from other categories. As a result the subsample could be very imbalanced which causes potential problems [14, 15, 16]. To solve this, we construct optimal subsampling probabilities based on the summation constraint where all categories are treated equally. Furthermore, different model constraints give the same mean responses and only lead to different interpretations of the parameters when it comes to parameter estimation. Utilizing this, we propose to formulate the optimal subsampling probabilities by minimizing the asymptotic expectation of the mean squared prediction error, which focus on enhancing prediction ability, and the novel optimal subsampling probabilities are independent to the model constraints.

Even though the subsampling method used in this paper is subsampling with replacement, the problems mentioned above, including the differentiation to the baseline category under the baseline constraint and the variation of the optimal subsampling probabilities with different model constraints, also exist for any other subsampling methods, e.g. Poisson subsampling [17]. Nonuniform subsampling without replacement for a fixed subsample size is seldom considered under the optimal subsampling framework due to its low computational efficiency.

This paper is organized as follows. Section 2 reviews the optimal subsampling probabilities under the baseline constraint and derives the optimal subsampling probabilities under the summation constraint. The comparison between these two optimal subsampling probabilities is also given to show that

the baseline constraint does not handle all categories equally while the summation constraint does. In Section 3, we minimize the asymptotic expectation of the mean squared prediction error to obtain the optimal subsampling probabilities that are independent to model constraints. Section 4 gives a practical implementation of the optimal subsampling algorithm to the softmax regression. Simulations and real data examples are presented in Section 5. The theoretical proofs along with required assumptions and extra numerical results are given in the appendix.

Here are some notations used throughout the paper. We use $\hat{\boldsymbol{\beta}}$ to denote the coefficient estimator obtained under any model constraints. Superscripts ^{b} and ^{s} associated with $\hat{\boldsymbol{\beta}}$ represent the coefficient estimators under the baseline constraint and the summation constraint, respectively. Since the mean response vector is independent of different model constraints, the superscripts ^{b} and ^{s} of $\hat{\boldsymbol{\beta}}$ are omitted for quantities calculated based on $\boldsymbol{p}_i(\hat{\boldsymbol{\beta}})$.

2. Optimal Subsampling under Different Model Constraints

The baseline constraint for identifiability of model (1) assumes the coefficient for the baseline category to be $\mathbf{0}$. Here we let $\boldsymbol{\beta}_0 = \mathbf{0}$. The unknown parameter under the baseline constraint is a Kd dimensional vector, denoted as $\boldsymbol{\beta}^{\mathrm{b}} = ({\boldsymbol{\beta}_1^{\mathrm{b}}}^\top, ..., {\boldsymbol{\beta}_K^{\mathrm{b}}}^\top)^\top$. The optimal subsampling probabilities are formulated by minimizing the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\beta}}_{\mathrm{sub}}^{\mathrm{b}}$ via A-optimality and L-optimality criteria in [1]. They proved that, as $N \to \infty$ and $n \to \infty$, given the full data, the conditional distribution of the approximation error $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{sub}}^{\mathrm{b}} - \hat{\boldsymbol{\beta}}_{\mathrm{full}}^{\mathrm{b}})$ is asymptotically normal with asymptotic variance-covariance matrix being

$$\mathbf{V}_N^\circ = {\mathbf{M}_N^\circ}^{-1} \mathbf{D}_N^\circ {\mathbf{M}_N^\circ}^{-1},$$

where

$$\mathbf{M}_N^\circ = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\phi}_i^\circ(\hat{\boldsymbol{\beta}}_{\mathrm{full}}) \otimes (\mathbf{x}_i \mathbf{x}_i^\top),$$

$$\mathbf{D}_N^\circ = \frac{1}{N^2} \sum_{i=1}^{N} \frac{\boldsymbol{\psi}_i^\circ(\hat{\boldsymbol{\beta}}_{\mathrm{full}}) \otimes (\mathbf{x}_i \mathbf{x}_i^\top)}{\pi_i},$$

$\boldsymbol{\phi}_i^\circ(\boldsymbol{\beta}) = \mathrm{diag}\{\boldsymbol{p}_i^\circ(\boldsymbol{\beta})\} - \{\boldsymbol{p}_i^\circ(\boldsymbol{\beta})\}\{\boldsymbol{p}_i^\circ(\boldsymbol{\beta})\}^\top$, $\boldsymbol{\psi}_i^\circ(\boldsymbol{\beta}) = \mathbf{s}_i^\circ(\boldsymbol{\beta})\mathbf{s}_i^\circ(\boldsymbol{\beta})^\top$, $\boldsymbol{s}_i^\circ(\boldsymbol{\beta}) = \mathbf{y}_i^\circ - \boldsymbol{p}_i^\circ(\boldsymbol{\beta})$, $\boldsymbol{p}_i^\circ(\boldsymbol{\beta}) = \{p_1(\mathbf{x}_i, \boldsymbol{\beta}), p_2(\mathbf{x}_i, \boldsymbol{\beta}), ..., p_K(\mathbf{x}_i, \boldsymbol{\beta})\}^\top$ and $\mathbf{y}_i^\circ = \{y_{i,1}, y_{i,2}, ..., y_{i,K}\}$ is a K dimensional vector.

Under the A-optimality criterion, the trace of the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\beta}}_{\mathrm{sub}}^{\mathrm{b}}$, $\mathrm{tr}(\mathbf{V}_N^\circ)$, is minimized. As for the L-optimality criterion, the trace of the asymptotic variance-covariance matrix of a linear transformed $\hat{\boldsymbol{\beta}}_{\mathrm{sub}}^{\mathrm{b}}$ is minimized. Different linear transformations of $\hat{\boldsymbol{\beta}}_{\mathrm{sub}}^{\mathrm{b}}$ contribute to different L-optimal subsampling probabilities. To achieve the computational benefits, the trace of asymptotic variance-covariance matrix of $\mathbf{M}_N^\circ \hat{\boldsymbol{\beta}}_{\mathrm{sub}}^{\mathrm{b}}$,

$\text{tr}(\mathbf{D}_N^\circ)$, is chosen to be minimized in [1]. Based on these two optimality criteria, the derived optimal subsampling probabilities are

$$\pi_i^{\text{opt}} = \frac{\|\boldsymbol{\mathcal{L}}\{\mathbf{s}_i^\circ(\hat{\boldsymbol{\beta}}_{\text{full}}) \otimes \mathbf{x}_i\}\|}{\sum_{j=1}^N \|\boldsymbol{\mathcal{L}}\{\mathbf{s}_j^\circ(\hat{\boldsymbol{\beta}}_{\text{full}}) \otimes \mathbf{x}_j\}\|}, \quad i \in \{1, \ldots, N\}, \tag{2}$$

where $\boldsymbol{\mathcal{L}} = {\mathbf{M}_N^\circ}^{-1}$ for A-optimality criterion and $\boldsymbol{\mathcal{L}} = \mathbf{I}$ for L-optimality criterion when $\text{tr}(\mathbf{D}_N^\circ)$ is minimized. For easy presentation, we use $\pi_i^{\text{b,A}}$ to represent π_i^{opt} with $\boldsymbol{\mathcal{L}} = {\mathbf{M}_N^\circ}^{-1}$ and $\pi_i^{\text{b,L}}$ to represent π_i^{opt} with $\boldsymbol{\mathcal{L}} = \mathbf{I}$.

2.1. Summation Constraint

Summation constraint assumes the sum of unknown coefficients for all categories to be $\mathbf{0}$, say $\sum_{k=0}^K \boldsymbol{\beta}_k = \mathbf{0}$. Under this constraint, the unknown parameter is denoted as $\boldsymbol{\beta}^{\text{s}} = ({\boldsymbol{\beta}_0^{\text{s}}}^\top, {\boldsymbol{\beta}_1^{\text{s}}}^\top, ..., {\boldsymbol{\beta}_K^{\text{s}}}^\top)^\top$. Models under two constraints are equivalent in that $\boldsymbol{p}_i(\boldsymbol{\beta}^{\text{b}}) = \boldsymbol{p}_i(\boldsymbol{\beta}^{\text{s}})$ if

$$\boldsymbol{\beta}^{\text{s}} = \left\{ \begin{pmatrix} -(K+1)^{-1}\mathbf{1}_K^\top \\ \mathbf{I}_K - (K+1)^{-1}\mathbf{J}_K \end{pmatrix} \otimes \mathbf{I}_d \right\} \boldsymbol{\beta}^{\text{b}} \equiv \mathbf{G}\boldsymbol{\beta}^{\text{b}}, \tag{3}$$

where $\mathbf{1}_K$ is a K dimensional vector of ones, $\mathbf{I}_K$ is a $K \times K$ dimensional identity matrix and $\mathbf{J}_K = \mathbf{1}_K\mathbf{1}_K^\top$. Models under two constraints have no difference in terms of modelling $\mathbb{E}(\mathbf{y}_i|\mathbf{x}_i)$ using available data. Due to (3), we could compute the full data MLE $\hat{\boldsymbol{\beta}}_{\text{full}}^{\text{s}}$ by premultiplying $\mathbf{G}$ to $\hat{\boldsymbol{\beta}}_{\text{full}}^{\text{b}}$. Similarly, the subsample estimator $\hat{\boldsymbol{\beta}}_{\text{sub}}^{\text{s}}$ for a subsample drawn by arbitrary subsampling probabilities is obtained by $\mathbf{G}\hat{\boldsymbol{\beta}}_{\text{sub}}^{\text{b}}$. The asymptotic distribution of $\hat{\boldsymbol{\beta}}_{\text{sub}}^{\text{s}}$ is investigated in Theorem 1.

Theorem 1. *Under Assumptions 1 and 2 in Appendix A, given the full data $\mathcal{D}_N$, when $N \to \infty$ and $n \to \infty$, $\hat{\boldsymbol{\beta}}_{\text{sub}}^{\text{s}} - \hat{\boldsymbol{\beta}}_{\text{full}}^{\text{s}}$ satisfies*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{sub}}^{\text{s}} - \hat{\boldsymbol{\beta}}_{\text{full}}^{\text{s}}) \overset{a}{\sim} \mathbb{N}(\mathbf{0}, \mathbf{V}_G),$$

where $\overset{a}{\sim}$ means that two quantities have the same asymptotic distribution,

$$\begin{aligned} \mathbf{V}_G &= (\mathbf{M}_N)^+\mathbf{D}_N(\mathbf{M}_N)^+, \\ \mathbf{M}_N &= \frac{1}{N}\sum_{i=1}^N \boldsymbol{\phi}_i(\hat{\boldsymbol{\beta}}_{\text{full}}) \otimes (\mathbf{x}_i\mathbf{x}_i^\top), \\ \mathbf{D}_N &= \frac{1}{N^2}\sum_{i=1}^N \frac{\boldsymbol{\psi}_i(\hat{\boldsymbol{\beta}}_{\text{full}}) \otimes (\mathbf{x}_i\mathbf{x}_i^\top)}{\pi_i}, \end{aligned}$$

$(\cdot)^+$ represents the Moore-Penrose inverse, $\boldsymbol{\phi}_i(\boldsymbol{\beta}) = \text{diag}\{\boldsymbol{p}_i(\boldsymbol{\beta})\} - \{\boldsymbol{p}_i(\boldsymbol{\beta})\}\{\boldsymbol{p}_i(\boldsymbol{\beta})\}^\top$, $\boldsymbol{\psi}_i(\boldsymbol{\beta}) = \mathbf{s}_i(\boldsymbol{\beta})\mathbf{s}_i(\boldsymbol{\beta})^\top$, and $\mathbf{s}_i(\boldsymbol{\beta}) = \mathbf{y}_i - \boldsymbol{p}_i(\boldsymbol{\beta})$. Note that $\mathbf{V}_G$ is a singular matrix.

Since the regression coefficients under the summation constraint are linearly dependent, the matrices $\mathbf{M}_N$, $\mathbf{D}_N$, and $\mathbf{V}_G$ are all singular. Theorem 1 shows that the unique Moore-Penrose inverse of $\mathbf{M}_N$ exhibits in the sandwich form of $\mathbf{V}_G$. By minimizing $\mathbf{V}_G$ under the A- and L- optimality criteria, the optimal subsampling probabilities are obtained and presented in Theorem 2.

Theorem 2. *The optimal subsampling probabilities under A-optimality criterion are*

$$\pi_i^{\mathrm{s,A}} = \frac{\|(\mathbf{M}_N)^+\{\mathbf{s}_i(\hat{\boldsymbol{\beta}}_{\mathrm{full}}) \otimes \mathbf{x}_i\}\|}{\sum_{j=1}^N \|(\mathbf{M}_N)^+\{\mathbf{s}_j(\hat{\boldsymbol{\beta}}_{\mathrm{full}}) \otimes \mathbf{x}_j\}\|}. \tag{4}$$

The optimal subsampling probabilities under L-optimality criterion are

$$\pi_i^{\mathrm{s,L}} = \frac{\|\mathbf{s}_i(\hat{\boldsymbol{\beta}}_{\mathrm{full}}) \otimes \mathbf{x}_i\|}{\sum_{j=1}^N \|\mathbf{s}_j(\hat{\boldsymbol{\beta}}_{\mathrm{full}}) \otimes \mathbf{x}_j\|} = \frac{\|\mathbf{s}_i(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\|\|\mathbf{x}_i\|}{\sum_{j=1}^N \|\mathbf{s}_j(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\|\|\mathbf{x}_j\|}. \tag{5}$$

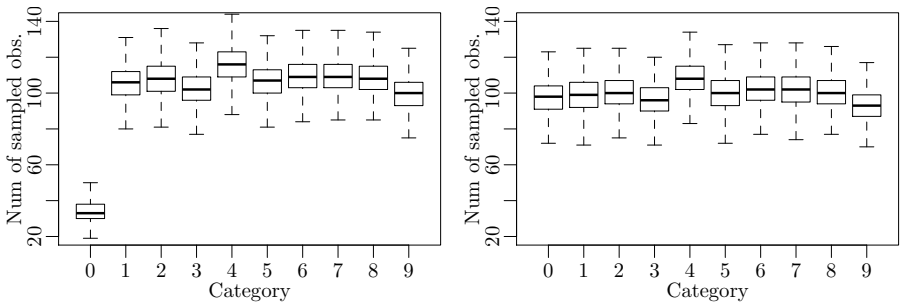
The A-optimal subsampling probabilities shown in (4) are obtained by minimizing $\mathrm{tr}(\mathbf{V}_G)$ and (4) is equal to (2) for $\boldsymbol{\mathcal{L}} = \mathbf{G}\mathbf{M}_N^{\circ\,-1}$. Similarly, (5) is obtained by minimizing the trace of the asymptotic variance-covariance matrix of $\mathbf{M}_N\hat{\boldsymbol{\beta}}^{\mathrm{s}}_{\mathrm{sub}}$, and is equivalent to (2) when $\boldsymbol{\mathcal{L}} = \mathbf{G}(\mathbf{G}^\top\mathbf{G})^{-1}$. Both of $\pi_i^{\mathrm{s,A}}$ and $\pi_i^{\mathrm{s,L}}$ can be viewed as L-optimal subsampling probabilities under the baseline constraint. For logistic regression, $\pi_i^{\mathrm{s,A}}$ and $\pi_i^{\mathrm{s,L}}$ are equivalent to $\pi_i^{\mathrm{b,A}}$ and $\pi_i^{\mathrm{b,L}}$, respectively.

2.2. Comparison between Baseline Constraint and Summation Constraint

One advantage of the summation constraint over the baseline constraint is that the former treats all categories equally whereas the latter deals with the baseline category differently, because $\pi_i^{\mathrm{b,L}}$ relates to $y_{i,k} - p_k(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_{\mathrm{full}})$ only for $k \in \{1, ..., K\}$ without $k = 0$. We generate a balanced synthetic dataset by setting all coefficients to be $\mathbf{0}$ and record the number of sampled observations for different categories using $\pi_i^{\mathrm{b,L}}$ and $\pi_i^{\mathrm{s,L}}$. In part (a) of Figure 1, the average number of sampled observations for the baseline category is only one third of the average number of sampled observations for other categories. This difference will be more evident when K goes larger, because $\pi_i^{\mathrm{b,L}} \propto \|x_i\|\sqrt{K}/(K+1)$ when $y_{i,0} = 1$ is smaller than $\pi_i^{\mathrm{b,L}} \propto \|x_i\|\sqrt{K^2+K-1}/(K+1)$ when $y_{i,k} = 1$ for any $k \in \{1, 2, ..., K\}$. Subsampling with $\pi_i^{\mathrm{b,A}}$ tends to have more observations drawn from the baseline category as shown in Figure B.5 of Appendix B. Part (b) of Figure 1 shows that roughly equal numbers of observations are sampled from all categories, suggesting that the summation constraint treats all categories equally in formulating the optimal subsampling probabilities.

3. Optimal Subsampling Probabilities by minimizing the Mean Squared Prediction Error

Different model constraints lead to different forms of optimal subsampling probabilities when they are formulated by minimizing the asymptotic variance-covariance matrix of the subsample estimators. For any constraint imposing



(a) Baseline Constraint **(b)** Summation Constraint

Figure 1: Boxplot for average number of sampled observations for all categories under L-optimal subsampling probabilities for the synthetic data when $N = 10000$ and $n = 1000$ of 1000 iterations. The covariates are generated from $\mathbb{N}_3(\mathbf{0}, \mathbf{I}_3)$ and the response has 10 different categories. The true coefficient is $\boldsymbol{\beta}^{\mathrm{s}} = 0 \times \mathbf{1}_{30}$. The number of observations in each category for the full data are roughly equal.

to model (1) to make it identifiable, new optimal subsampling probabilities are obtained, but they are only optimal under the given model constraint. Thus, constructing optimal subsampling probabilities that are immune to the choice of model constraints is desirable.

Besides the estimation efficiency of the subsample estimator, its prediction ability is another fundamental criterion to assess the effectiveness of the subsample, as classification is the primary goal of the softmax regression. Remember that the mean response $\boldsymbol{p}_i(\boldsymbol{\beta})$ stays invariant no matter which model constraint applies. The mean squared prediction error

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\boldsymbol{p}_i(\hat{\boldsymbol{\beta}}_{\mathrm{sub}}) - \boldsymbol{p}_i(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\right\|^2$$

is independent to the choice of model constraints and quantifies the prediction ability of the subsample estimator. In this section, we consider constructing optimal subsampling probabilities by minimizing the mean squared prediction error. To do that, we investigate the asymptotic distribution of the mean squared prediction error in Theorem 3.

Theorem 3. *Under Assumptions 1 and 2 in Appendix A, given the full data $\mathcal{D}_N$, when $N \to \infty$ and $n \to \infty$,*

$$\frac{n}{N}\sum_{i=1}^{N}\left\|\boldsymbol{p}_i(\hat{\boldsymbol{\beta}}_{\mathrm{sub}}) - \boldsymbol{p}_i(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\right\|^2 \overset{a}{\sim} \mathbf{z}^\top {\mathbf{V}_N^\circ}^{1/2} \boldsymbol{\Omega}_N^\circ {\mathbf{V}_N^\circ}^{1/2} \mathbf{z},$$

where $\mathbf{z} \sim \mathbb{N}(\mathbf{0}, \mathbf{I})$,

$$\boldsymbol{\Omega}_N^\circ = \frac{1}{N}\sum_{i=1}^{N}\left\{\mathbf{B}_i^\top(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\mathbf{B}_i(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\right\},$$

$$\mathbf{B}_i(\boldsymbol{\beta}) = \begin{pmatrix} -p_0(\mathbf{x}_i,\boldsymbol{\beta})p_1(\mathbf{x}_i,\boldsymbol{\beta}) & ... & -p_0(\mathbf{x}_i,\boldsymbol{\beta})p_K(\mathbf{x}_i,\boldsymbol{\beta}) \\ p_1(\mathbf{x}_i,\boldsymbol{\beta}) - p_1^2(\mathbf{x}_i,\boldsymbol{\beta}) & ... & -p_1(\mathbf{x}_i,\boldsymbol{\beta})p_K(\mathbf{x}_i,\boldsymbol{\beta}) \\ ... & ... & ... \\ -p_1(\mathbf{x}_i,\boldsymbol{\beta})p_K(\mathbf{x}_i,\boldsymbol{\beta}) & ... & p_K(\mathbf{x}_i,\boldsymbol{\beta}) - p_K^2(\mathbf{x}_i,\boldsymbol{\beta}) \end{pmatrix} \otimes \mathbf{x}_i^\top.$$

From Theorem 3, the asymptotic expectation of the mean squared prediction error is

$$\mathbb{E}\{\mathbf{z}^\top {\mathbf{V}_N^\circ}^{1/2} \boldsymbol{\Omega}_N^\circ {\mathbf{V}_N^\circ}^{1/2} \mathbf{z} | \mathcal{D}_N\} = \operatorname{tr}(\mathbf{V}_N^\circ \boldsymbol{\Omega}_N^\circ)$$

where $\mathbf{V}_N^\circ$ depends on the subsampling probabilities π_i. It is natural to think about finding optimal subsampling probabilities by minimizing the asymptotic expectation of the mean squared prediction error.

Theorem 4. *The optimal subsampling probabilities minimizing the asymptotic expectation of the mean squared prediction error are*

$$\pi_i^{\mathrm{P}} = \frac{\|{\boldsymbol{\Omega}_N^\circ}^{1/2} {\mathbf{M}_N^\circ}^{-1} \{\mathbf{s}_i^\circ(\hat{\boldsymbol{\beta}}_{\mathrm{full}}) \otimes \mathbf{x}_i\}\|}{\sum_{j=1}^N \|{\boldsymbol{\Omega}_N^\circ}^{1/2} {\mathbf{M}_N^\circ}^{-1} \{\mathbf{s}_j^\circ(\hat{\boldsymbol{\beta}}_{\mathrm{full}}) \otimes \mathbf{x}_j\}\|}. \tag{6}$$

Remark 1. Since $\boldsymbol{p}_i(\hat{\boldsymbol{\beta}})$ keeps constant no matter which model constraint is used, the value of the asymptotic expectation of the mean squared prediction error should also be the same under all model constraints for given $\{\pi_i\}_{i=1}^N$. Adopting the expression for the summation constraint,

$$\pi_i^{\mathrm{P}} = \frac{\|\boldsymbol{\Omega}_G^{1/2} {\mathbf{M}_N}^{+} \{\mathbf{s}_i(\hat{\boldsymbol{\beta}}_{\mathrm{full}}) \otimes \mathbf{x}_i\}\|}{\sum_{j=1}^N \|\boldsymbol{\Omega}_G^{1/2} {\mathbf{M}_N}^{+} \{\mathbf{s}_j(\hat{\boldsymbol{\beta}}_{\mathrm{full}}) \otimes \mathbf{x}_j\}\|} \tag{7}$$

where

$$\boldsymbol{\Omega}_G = \frac{1}{N} \sum_{i=1}^N \left\{ \boldsymbol{\Gamma}_i^\top(\hat{\boldsymbol{\beta}}_{\mathrm{full}}) \boldsymbol{\Gamma}_i(\hat{\boldsymbol{\beta}}_{\mathrm{full}}) \right\},$$

$$\boldsymbol{\Gamma}_i(\boldsymbol{\beta}) = \begin{pmatrix} p_0(\mathbf{x}_i,\boldsymbol{\beta}) - p_0^2(\mathbf{x}_i,\boldsymbol{\beta}) & ... & -p_0(\mathbf{x}_i,\boldsymbol{\beta})p_K(\mathbf{x}_i,\boldsymbol{\beta}) \\ -p_0(\mathbf{x}_i,\boldsymbol{\beta})p_1(\mathbf{x}_i,\boldsymbol{\beta}) & ... & -p_1(\mathbf{x}_i,\boldsymbol{\beta})p_K(\mathbf{x}_i,\boldsymbol{\beta}) \\ ... & ... & ... \\ -p_0(\mathbf{x}_i,\boldsymbol{\beta})p_K(\mathbf{x}_i,\boldsymbol{\beta}) & ... & p_K(\mathbf{x}_i,\boldsymbol{\beta}) - p_K^2(\mathbf{x}_i,\boldsymbol{\beta}) \end{pmatrix} \otimes \mathbf{x}_i^\top.$$

Here (6) and (7) give the same value, and they are only in different expressions.

We can see that (6) is a variant of the L-optimal probabilities under baseline constraint, which is acquired by minimizing the trace of the asymptotic variance-covariance matrix of ${\boldsymbol{\Omega}_N^\circ}^{1/2} \hat{\boldsymbol{\beta}}_{\mathrm{sub}}^{\mathrm{b}}$, and is equivalent to (2) for $\boldsymbol{\mathcal{L}} = {\boldsymbol{\Omega}_N^\circ}^{1/2} {\mathbf{M}_N^\circ}^{-1}$.

4. Practical Implemention

From the previous discussions, all optimal subsampling probabilities depend on the unknown full data MLE. To solve this problem, we use a two-step algorithm, which first obtains a pilot sample and uses the pilot sample estimator to

substitute the full data estimator when calculating optimal subsampling probabilities [10]. The two-step approximately optimal subsampling algorithm is introduced in Algorithm 1. Here $\hat{\boldsymbol{\beta}}$ represents the general coefficient estimator which is generated under any pre-specified model constraint.

Algorithm 1 Two-Step Approximately Optimal Subsampling Algorithm

(i) Draw n_1 samples with replacement by uniform subsampling or case-control sampling. Denote the drawn samples and their subsampling probabilities as $\{\mathbf{y}_i^{*1}, \mathbf{x}_i^{*1}, \pi_i^{*1}\}_{i=1}^{n_1}$. Calculate the subsample estimator $\hat{\boldsymbol{\beta}}_{\text{sub}}^1$ based on this subsample.

(ii) Use $\hat{\boldsymbol{\beta}}_{\text{sub}}^1$ to substitute $\hat{\boldsymbol{\beta}}_{\text{full}}$ to calculate the approximated optimal subsampling probabilities as

$$\tilde{\pi}_i = \frac{\|\boldsymbol{\mathcal{L}}(\hat{\boldsymbol{\beta}}_{\text{sub}}^1)\{\mathbf{s}_i^{\circ}(\hat{\boldsymbol{\beta}}_{\text{sub}}^1) \otimes \mathbf{x}_i\}\|}{\sum_{j=1}^{N} \|\boldsymbol{\mathcal{L}}(\hat{\boldsymbol{\beta}}_{\text{sub}}^1)\{\mathbf{s}_j^{\circ}(\hat{\boldsymbol{\beta}}_{\text{sub}}^1) \otimes \mathbf{x}_j\}\|} \tag{8}$$

for $i = 1, 2, ..., N$ and $\boldsymbol{\mathcal{L}}(\hat{\boldsymbol{\beta}}_{\text{sub}}^1)$ means substituting $\hat{\boldsymbol{\beta}}_{\text{full}}$ with $\hat{\boldsymbol{\beta}}_{\text{sub}}^1$ when calculating $\boldsymbol{\mathcal{L}}$. Use $\{\tilde{\pi}_i\}_{i=1}^{N}$ to draw a second subsample with size n_2 and denote them as $\{\mathbf{y}_i^{*2}, \mathbf{x}_i^{*2}, \pi_i^{*2}\}_{i=1}^{n_2}$.

(iii) Combine the pilot subsample and the second stage subsample. Calculate the subsample estimator $\hat{\boldsymbol{\beta}}_{\text{sub}}^{\text{cmb}}$ using the combined sample.

Remark 2. When approximating the optimal subsampling probabilities under the A-optimality or those obtained by minimizing the mean squared prediction error in (8), computing $\boldsymbol{\mathcal{L}}(\hat{\boldsymbol{\beta}}_{\text{sub}}^1)$ takes $O(NK^2d^2)$ time. To reduce this computational burden, we calculate $\boldsymbol{\mathcal{L}}(\hat{\boldsymbol{\beta}}_{\text{sub}}^1)$ using the first stage sample $\{\mathbf{y}_i^{*1}, \mathbf{x}_i^{*1}, \pi_i^{*1}\}_{i=1}^{n_1}$. Specifically, when approximating $\pi_i^{\text{b,A}}$, we use $\boldsymbol{\mathcal{L}}(\hat{\boldsymbol{\beta}}_{\text{sub}}^1) = \mathbf{M}_N^{\circ *\,-1}$ where

$$\mathbf{M}_N^{\circ *} = \frac{1}{n_1 N} \sum_{i=1}^{n_1} \frac{\boldsymbol{\phi}_i^{\circ}(\hat{\boldsymbol{\beta}}_{\text{sub}}^1) \otimes (\mathbf{x}_i^{*1} \mathbf{x}_i^{*1\top})}{\pi_i^{*1}}; \tag{9}$$

when approximating $\pi_i^{\text{s,A}}$, we use $\boldsymbol{\mathcal{L}}(\hat{\boldsymbol{\beta}}_{\text{sub}}^1) = \mathbf{G}\mathbf{M}_N^{\circ *\,-1}$; and when approximating π_i^{P}, we use $\boldsymbol{\mathcal{L}}(\hat{\boldsymbol{\beta}}_{\text{sub}}^1) = \boldsymbol{\Omega}_N^{\circ *\,1/2} \mathbf{M}_N^{\circ *\,-1}$ where

$$\boldsymbol{\Omega}_N^{\circ *} = \frac{1}{n_1 N} \sum_{i=1}^{n_1} \frac{\mathbf{B}_i^{\top}(\hat{\boldsymbol{\beta}}_{\text{sub}}^1) \mathbf{B}_i(\hat{\boldsymbol{\beta}}_{\text{sub}}^1)}{\pi_i^{*1}}.$$

5. Numerical Results

We use simulations and real data examples in this section to demonstrate the effectiveness of the proposed subsampling probabilities with finite sample sizes.

5.1. Simulations

To compare the performances of different optimal subsampling probabilities, we simulate four synthetic datasets with full data size $N = 10000$. The dimension of covariates is 3 and the response has 6 different outcomes, 0, 1, 2, 3, 4 and 5. The true parameters are set as $\boldsymbol{\beta}^{\mathrm{s}} = 0.2\times(5\mathbf{1}_3^\top, -\mathbf{1}_3^\top, -\mathbf{1}_3^\top, -\mathbf{1}_3^\top, -\mathbf{1}_3^\top, -\mathbf{1}_3^\top)^\top$. The covariates for each data are generated by the following four distributions:

Multivariate Normal: $\mathbf{x}_i \sim \mathbb{N}_3(\mathbf{0}, \Sigma)$ where Σ is a 3×3 matrix with diagonal elements being 1 and off-diagonal elements being 0.5. Around 42% observations fall in category 0, and the other 58% observations fall in category 1, 2, 3, 4 and 5 nearly evenly.

Shifted Multivariate Normal: $\mathbf{x}_i \sim \mathbb{N}_3(1.5\mathbf{1}_3, \Sigma)$. This is a very unbalanced dataset with nearly 95% observations falling in category 0. Each of the other five categories has 1% observations to fall in.

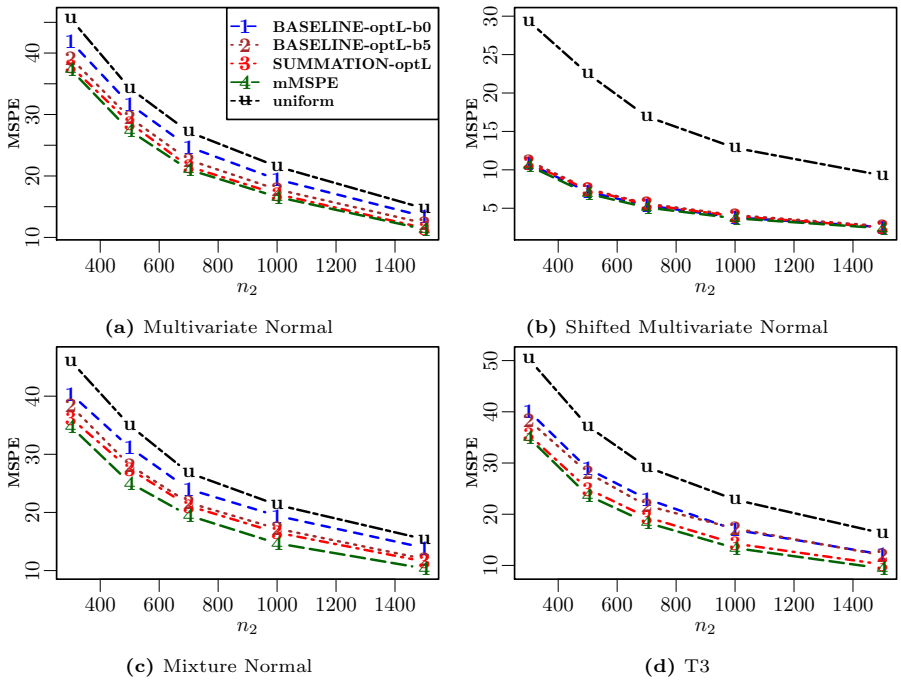
Mixture Normal: $\mathbf{x}_i \sim 0.5\mathbb{N}_3(\mathbf{1}_3, \Sigma) + 0.5\mathbb{N}_3(-\mathbf{1}_3, \Sigma)$. Around 46% observations fall in category 0, and the other five categories share the remaining 54% observations nearly equally.

T3: $\mathbf{x}_i \sim T_3(\mathbf{0}, \Sigma)$, where T_3 means t distribution with degree of freedom three. Around 46% observations fall in category 0, and the other 54% observations fall in category 1, 2, 3, 4 and 5 roughly equally.

To compare the performances of different optimal subsampling probabilities, we use the empirical mean squared prediction error

$$\frac{1}{S}\sum_{s=1}^{S}\sum_{i=1}^{N}\left\|\boldsymbol{p}_i(\hat{\boldsymbol{\beta}}_{\mathrm{sub},s}^{\mathrm{cmb}}) - \boldsymbol{p}_i(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\right\|^2$$

as the measurement criterion, where $\hat{\boldsymbol{\beta}}_{\mathrm{sub},s}^{\mathrm{cmb}}$ is the subsample estimator obtained by Algorithm 1 for s-th replication. Figure 2 compares the empirical mean squared prediction error for different subsampling probabilities and shows that except the one with covariates generated by shifted location multivariate normal distribution, all other three cases indicate that using the optimal subsampling probabilities obtained by minimizing mean squared prediction error gives the best prediction performance, matching the theoretical result. Under the baseline constraint, treating different category as the baseline category results in different prediction accuracy. All algorithms based on the optimal subsampling probabilities beat the uniform subsampling in terms of prediction ability. The optimal subsampling probabilities under A-optimality criterion have similar performance as that under L-optimality criterion, and the results are shown in Figure B.6 of Appendix B.



(a) Multivariate Normal

(b) Shifted Multivariate Normal

(c) Mixture Normal

(d) T3

Figure 2: Average mean squared prediction error for different types of optimal subsampling probabilities for all four cases with 1000 replicates when $n_1 = 300$. BASELINE-optL-b0 uses Algorithm 1 with $\boldsymbol{\mathcal{L}}(\hat{\boldsymbol{\beta}}^1_{\text{sub}}) = \mathbf{I}$ for (8) and treats the outcome 0 as the baseline category. BASELINE-optL-b5 uses Algorithm 1 with $\boldsymbol{\mathcal{L}}(\hat{\boldsymbol{\beta}}^1_{\text{sub}}) = \mathbf{I}$ for (8) and treats the outcome 5 as the baseline category. SUMMATION-optL means Algorithm 1 with $\boldsymbol{\mathcal{L}}(\hat{\boldsymbol{\beta}}^1_{\text{sub}}) = \mathbf{G}(\mathbf{G}^\top\mathbf{G})^{-1}$ for (8) and mMSPE stands for Algorithm 1 with $\boldsymbol{\mathcal{L}}(\hat{\boldsymbol{\beta}}^1_{\text{sub}})$ replaced by $\boldsymbol{\Omega}_N^{\circ *1/2}\mathbf{M}_N^{\circ * -1}$ in (8). The sample size for the uniform subsampling is $n_1 + n_2$.

5.2. Real Data Analysis

We apply the two-stage approximately optimal subsampling algorithm to the cover type data set to compare the performances of different optimal subsampling probabilities in prediction accuracy. The cover type dataset [18, 19] is available at `https://archive.ics.uci.edu/ml/datasets/covertype`. It records seven forest types for given locations, including Spruce/Fir, Lodgepole Pine, Ponderosa Pine, Cottonwood/Willow, Aspen, Douglas-fir and Krummholz, and their corresponding proportions are 36.46%, 48.76%, 6.15%, 0.427%, 1.63%, 2.99% and 3.53%, respectively. The total number of observations is 581012 and 10 continuous covariates are used in the data analysis, including elevation, aspect, slope, horizontal distance and vertical distances to the nearest surface water, horizontal distance to the nearest roadway, hill shades measured at 9 AM, noon and 3PM, and horizontal distance to the nearest wildfire ignition point. All covariates are normalized to have mean 0 and variance 1. Figure 3 compares different optimal subsampling probabilities in empirical mean squared prediction error, and it indicates that using optimal subsampling probabilities obtained by minimizing the mean squared prediction error produces the least prediction error.

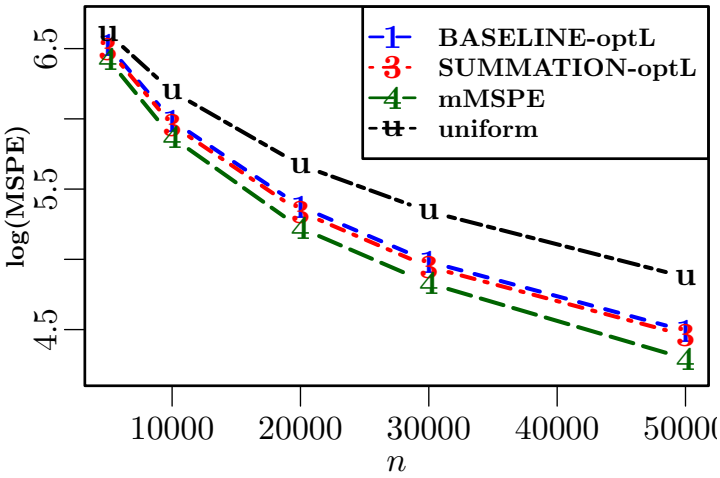


Figure 3: Log of average mean squared prediction error for different types of optimal subsampling probabilities for cover type dataset for 1000 replicates when $n_1 = 5000$. For BASELINE-optL, Cotttonwood/Willow is treated as the baseline category.

Besides the cover type dataset, we test the performance of the proposed optimal subsampling probabilities with the character font images dataset [19] that is available at `https://archive.ics.uci.edu/ml/datasets/Character+Font+Images`. This dataset contains image information for 153 fonts and all images are scaled into 20×20 pixel squares. It has 410 covariates, indicating the value of the character, size and style of the character, and pixel values ranging from 0 to 255 for 400 pixels. Here we use the data for five fonts and the

total number of observations is 124,817. The five fonts and their corresponding percentages are Agency FB (0.80%), Arial (21.02%), Mongolian Baiti (1.32%), Bank Gothic (1.79%) and OCR-B (75.06%). In this analysis, twenty principle components with the largest eigenvalues are used as the covariates. Figure 4 also demonstrates the effectiveness of using optimal subsampling probabilities obtained by minimizing the mean squared prediction error in prediction ability.

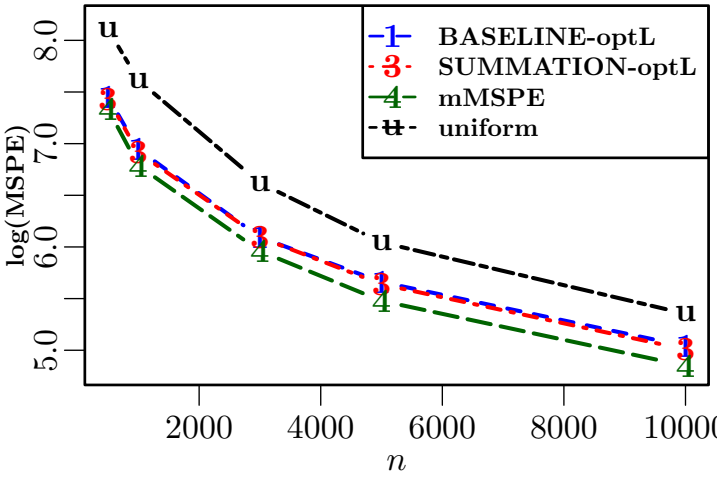


Figure 4: Log of average mean squared prediction error for different types of optimal subsampling probabilities for character font images dataset for 1000 replicates when $n_1 = 500$. For BASELINE-optL, AGENCY FB is treated as the baseline category.

Acknowledgements

We gratefully acknowledge the insightful comments and suggestions from the reviewers that helped improve the paper. Jiahui Zou's research was supported by the National Natural Science Foundation of China (Grant No. 12201431). HaiYing Wang's research was supported by NSF grant CCF-2105571.

Appendix A. Assumptions and Theoretical Proofs

Assumptions

The following assumptions are required by Theorem 1 and Theorem 3.

Assumption 1. As N goes to ∞, $\mathbf{M}_N^{\circ}$ 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 $k = 0, 4$, $N^{-2}\sum_{i=1}^{N}\pi_i^{-1}\|\mathbf{x}_i\|^k = O_P(1)$; 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 requires the observed information matrix to be invertible as N goes to ∞ and the third moment of the covariates is bounded in probability. Assumption 2 restricts the distribution of subsampling probabilities.

Proof of Theorem 1

Proof. Note that

$$\boldsymbol{\beta}^s = \left\{ \begin{pmatrix} -(K+1)^{-1}\mathbf{1}_K^\top \\ \mathbf{I}_K - (K+1)^{-1}\mathbf{J}_K \end{pmatrix} \otimes \mathbf{I}_d \right\} \boldsymbol{\beta}^{\mathrm{b}} \equiv \begin{pmatrix} \mathbf{G}_1 \\ \mathbf{G}_2 \end{pmatrix} \boldsymbol{\beta}^{\mathrm{b}} \equiv \mathbf{G}\boldsymbol{\beta}^{\mathrm{b}}. \tag{A.1}$$

Denote

$$\begin{aligned} \mathbf{G}_1^\top \mathbf{G}_1 &= \frac{1}{(K+1)^2}\mathbf{J}_K \otimes \mathbf{I}_d, \\ \mathbf{G}_2^\top \mathbf{G}_2 &= \frac{1}{(K+1)^2}\{(K+1)^2\mathbf{I}_K - (K+2)\mathbf{J}_K\} \otimes \mathbf{I}_d, \end{aligned}$$

and it can be shown that $\mathbf{G}^\top\mathbf{G} = \mathbf{G}_1^\top\mathbf{G}_1 + \mathbf{G}_2^\top\mathbf{G}_2 = \mathbf{G}_2$. Thus

$$\mathbf{G}(\mathbf{G}^\top\mathbf{G})^{-1} = \begin{pmatrix} \mathbf{G}_1 \\ \mathbf{G}_2 \end{pmatrix} \mathbf{G}_2^{-1} = \begin{pmatrix} -\mathbf{1}_K^\top \\ \mathbf{I}_K \end{pmatrix} \otimes \mathbf{I}_d. \tag{A.2}$$

By (A.2), we have

$$\begin{aligned} &\mathbf{G}(\mathbf{G}^\top\mathbf{G})^{-1}[\{\mathbf{y}_i^\circ - \boldsymbol{p}_i^\circ(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\} \otimes \mathbf{x}_i] \\ &= \left[\begin{pmatrix} -\mathbf{1}_K^\top \\ \mathbf{I}_K \end{pmatrix} \{\mathbf{y}_i^\circ - \boldsymbol{p}_i^\circ(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\} \right] \otimes \mathbf{x}_i \\ &= \{\mathbf{y}_i - \boldsymbol{p}_i(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\} \otimes \mathbf{x}_i \end{aligned}$$

and

$$\mathbf{D}_N = \mathbf{G}(\mathbf{G}^\top\mathbf{G})^{-1}\mathbf{D}_N^\circ(\mathbf{G}^\top\mathbf{G})^{-1}\mathbf{G}^\top = \frac{1}{N^2}\sum_{i=1}^{N}\frac{\boldsymbol{\psi}_i(\hat{\boldsymbol{\beta}}_{\mathrm{full}}) \otimes (\mathbf{x}_i\mathbf{x}_i^\top)}{\pi_i}, \tag{A.3}$$

where $\boldsymbol{\psi}_i(\boldsymbol{\beta}) = \{\mathbf{y}_i - \boldsymbol{p}(\boldsymbol{\beta})\}\{\mathbf{y}_i - \boldsymbol{p}(\boldsymbol{\beta})\}^\top$.

Further, denote $\mathbf{H} \equiv \mathbf{G}(\mathbf{M}_N^\circ)^{-1}\mathbf{G}^\top$ and we will prove $\mathbf{H} = \mathbf{M}_N^+$ in the following. Note that

$$\begin{aligned} \mathbf{G} &= \begin{pmatrix} \mathbf{G}_1 \\ \mathbf{G}_2 \end{pmatrix} = \begin{pmatrix} -(K+1)^{-1}\mathbf{1}^\top \\ \mathbf{I}_K - (K+1)^{-1}\mathbf{J}_K \end{pmatrix} \otimes \mathbf{I}_d \\ &= \begin{pmatrix} \mathbf{0}_K^\top \\ \mathbf{I}_K \end{pmatrix} \otimes \mathbf{I}_d - \frac{1}{K+1}(\mathbf{1}_{K+1}\mathbf{1}_K^\top) \otimes \mathbf{I}_d \end{aligned}$$

and

$$\mathbf{H} = \mathbf{G}(\mathbf{M}_N^{\circ})^{-1}\mathbf{G}^{\top} = \mathbf{G}(\mathbf{G}^{\top}\mathbf{M}_N\mathbf{G})^{-1}\mathbf{G}^{\top}, \tag{A.4}$$

where

$$\mathbf{M}_N = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\phi}_i(\hat{\boldsymbol{\beta}}_{\text{full}}) \otimes (\mathbf{x}_i\mathbf{x}_i^{\top}),$$

and $\boldsymbol{\phi}(\boldsymbol{\beta}) = \text{diag}\{\boldsymbol{p}_i(\boldsymbol{\beta})\} - \{\boldsymbol{p}_i(\boldsymbol{\beta})\}^{\otimes 2}$.

Note that

$$\mathbf{1}_{K+1}^{\top}\boldsymbol{\phi}_i(\hat{\boldsymbol{\beta}}_{\text{full}}) = \mathbf{0}_{K+1}^{\top},$$

and we have

$$(\mathbf{1}_{K+1} \otimes \mathbf{I}_d)^{\top}\mathbf{M}_N = (\mathbf{1}_{K+1} \otimes \mathbf{I}_d)^{\top} \begin{pmatrix} m_N & (\boldsymbol{m}_N)^{\top} \\ \boldsymbol{m}_N & \mathbf{M}_N^{\circ} \end{pmatrix} = \mathbf{0}_{d(K+1)}^{\top},$$

which indicates

$$\begin{aligned}
m_N &= -(\mathbf{1}_K \otimes \mathbf{I}_d)^{\top}\boldsymbol{m}_N, && \text{(A.5)}\\
\boldsymbol{m}_N &= -\mathbf{M}_N^{\circ}(\mathbf{1}_K \otimes \mathbf{I}_d), && \text{(A.6)}\\
(\boldsymbol{m}_N)^{\top}(\mathbf{M}_N^{\circ})^{-1}\boldsymbol{m}_N &= -(\boldsymbol{m}_N)^{\top}(\mathbf{M}_N^{\circ})^{-1}(\mathbf{M}_N^{\circ})(\mathbf{1}_K \otimes \mathbf{I}_d) \\
&= -(\boldsymbol{m}_N)^{\top}(\mathbf{1}_K \otimes \mathbf{I}_d) = m_N, && \text{(A.7)}\\
\mathbf{G}^{\top}\mathbf{M}_N &= \begin{pmatrix} \boldsymbol{m}_N & \mathbf{M}_N^{\circ} \end{pmatrix}. && \text{(A.8)}
\end{aligned}$$

Based on (A.7), we have

$$\begin{aligned}
\mathbf{M}_N\mathbf{H}\mathbf{M}_N &= \mathbf{M}_N\mathbf{G}(\mathbf{G}^{\top}\mathbf{M}_N\mathbf{G})^{-1}\mathbf{G}^{\top}\mathbf{M}_N \\
&= \mathbf{M}_N\mathbf{G}(\mathbf{M}_N^{\circ})^{-1}\mathbf{G}^{\top}\mathbf{M}_N \\
&= \begin{pmatrix} (\boldsymbol{m}_N)^{\top} \\ \mathbf{M}_N^{\circ} \end{pmatrix} (\mathbf{M}_N^{\circ})^{-1} \begin{pmatrix} \boldsymbol{m}_N & \mathbf{M}_N^{\circ} \end{pmatrix} \\
&= \begin{pmatrix} (\boldsymbol{m}_N)^{\top}(\mathbf{M}_N^{\circ})^{-1}\boldsymbol{m}_N & (\boldsymbol{m}_N)^{\top} \\ \boldsymbol{m}_N & \mathbf{M}_N^{\circ} \end{pmatrix} \\
&= \mathbf{M}_N, \\
\mathbf{H}\mathbf{M}_N\mathbf{H} &= \mathbf{G}(\mathbf{G}^{\top}\mathbf{M}_N\mathbf{G})^{-1}\mathbf{G}^{\top}\mathbf{M}_N\mathbf{G}(\mathbf{G}^{\top}\mathbf{M}_N\mathbf{G})^{-1}\mathbf{G}^{\top} \\
&= \mathbf{G}(\mathbf{G}^{\top}\mathbf{M}_N\mathbf{G})^{-1}\mathbf{G}^{\top} \\
&= \mathbf{H}.
\end{aligned}$$

Based on (A.2), (A.6) and (A.8), we have

$$\begin{aligned}
\mathbf{H}\mathbf{M}_N &= \mathbf{G}(\mathbf{G}^{\top}\mathbf{M}_N\mathbf{G})^{-1}\mathbf{G}^{\top}\mathbf{M}_N \\
&= \mathbf{G}(\mathbf{M}_N^{\circ})^{-1}\mathbf{G}^{\top}\mathbf{M}_N
\end{aligned}$$

$$
\begin{aligned}
&= \mathbf{G}(\mathbf{M}_N^{\circ})^{-1}\begin{pmatrix}\boldsymbol{m}_N & \mathbf{M}_N^{\circ}\end{pmatrix} \\
&= \mathbf{G}\begin{pmatrix}(\mathbf{M}_N^{\circ})^{-1}\boldsymbol{m}_N & \mathbf{I}_{dK}\end{pmatrix} \\
&= \mathbf{G}\left\{\begin{pmatrix}-\mathbf{1}_K & \mathbf{I}_K\end{pmatrix} \otimes \mathbf{I}_d\right\} \\
&= \mathbf{G}(\mathbf{G}^\top\mathbf{G})^{-1}\mathbf{G}^\top,
\end{aligned}
$$

which implies that $\mathbf{M}_N\mathbf{H} = (\mathbf{M}_N\mathbf{H})^\top$ and $\mathbf{H}\mathbf{M}_N = (\mathbf{H}\mathbf{M}_N)^\top$. According to the definition of Moore-Penrose inverse [20], we know $\mathbf{H} = (\mathbf{M}_N)^+$. Finally, combining with (A.1), (A.3) and (A.4), the asymptotic variance-covariance matrix of $\sqrt{n}(\hat{\boldsymbol{\beta}}^{\mathrm{s}}_{\mathrm{sub}} - \hat{\boldsymbol{\beta}}^{\mathrm{s}}_{\mathrm{full}})$ is

$$
\begin{aligned}
\mathbf{V}_G &= \mathbf{G}(\mathbf{M}_N^{\circ})^{-1}\mathbf{D}_N^{\circ}(\mathbf{M}_N^{\circ})^{-1}\mathbf{G}^\top \\
&= \mathbf{G}(\mathbf{M}_N^{\circ})^{-1}\mathbf{G}^\top\mathbf{G}(\mathbf{G}^\top\mathbf{G})^{-1}\mathbf{D}_N^{\circ}(\mathbf{G}^\top\mathbf{G})^{-1}\mathbf{G}^\top\mathbf{G}(\mathbf{M}_N^{\circ})^{-1}\mathbf{G}^\top \\
&= \mathbf{H}\mathbf{G}(\mathbf{G}^\top\mathbf{G})^{-1}\mathbf{D}_N^{\circ}(\mathbf{G}^\top\mathbf{G})^{-1}\mathbf{G}^\top\mathbf{H} \\
&= \mathbf{H}\mathbf{D}_N\mathbf{H} \\
&= (\mathbf{M}_N)^+\mathbf{D}_N(\mathbf{M}_N)^+.
\end{aligned}
$$

□

Proof of Theorem 2

Proof. Under A-optimality criteria,

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

The last step is based on the Cauchy-Schwarz inequality. The equality holds when π_i is proportional to $\|\mathbf{M}_N^+\{\mathbf{s}_i(\hat{\boldsymbol{\beta}}^{\mathrm{s}}_{\mathrm{full}}) \otimes \mathbf{x}_i\}\|$. Thus (4) is proved, and (5) is obtained in a similar way.

□

Proof of Theorem 3

Proof. By Taylor's series, we have

$$
\frac{1}{N}\sum_{i=1}^{N}\left\|\boldsymbol{p}(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_{\mathrm{sub}}) - \boldsymbol{p}(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_{\mathrm{full}})\right\|^2
$$

$$= \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \left\| \boldsymbol{p}(\mathbf{x}_i, \boldsymbol{\beta}^{\mathrm{b}}) - \boldsymbol{p}(\mathbf{x}_i, \hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{full}}) \right\|^2}{\partial \boldsymbol{\beta}^{\mathrm{b}}}\bigg|_{\boldsymbol{\beta}^{\mathrm{b}}=\hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{full}}} \left(\hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{sub}} - \hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{full}}\right)$$

$$+ \left(\hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{sub}} - \hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{full}}\right)^{\top} \frac{1}{2N}\sum_{i=1}^{N} \frac{\partial^2 \left\| \boldsymbol{p}(\mathbf{x}_i, \boldsymbol{\beta}^{\mathrm{b}}) - \boldsymbol{p}(\mathbf{x}_i, \hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{full}}) \right\|^2}{\partial \boldsymbol{\beta}^{\mathrm{b}} \partial \boldsymbol{\beta}^{\mathrm{b}\top}}\bigg|_{\boldsymbol{\beta}^{\mathrm{b}}=\hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{full}}} \left(\hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{sub}} - \hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{full}}\right)$$

$$+ R_N. \tag{A.9}$$

Direct calculation yields

$$\frac{\partial \left\| \boldsymbol{p}(\mathbf{x}_i, \boldsymbol{\beta}^{\mathrm{b}}) - \boldsymbol{p}(\mathbf{x}_i, \hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{full}}) \right\|^2}{\partial \boldsymbol{\beta}^{\mathrm{b}}} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_K \end{pmatrix} \otimes \mathbf{x}_i,$$

where

$$e_k = 2\left\{\boldsymbol{p}(\mathbf{x}_i, \boldsymbol{\beta}^{\mathrm{b}}) - \boldsymbol{p}(\mathbf{x}_i, \hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{full}})\right\}^{\top} \begin{pmatrix} -p_0(\mathbf{x}_i, \boldsymbol{\beta}^{\mathrm{b}}) p_k(\mathbf{x}_i, \boldsymbol{\beta}^{\mathrm{b}}) \\ \vdots \\ p_k(\mathbf{x}_i, \boldsymbol{\beta}^{\mathrm{b}}) - p_k^2(\mathbf{x}_i, \boldsymbol{\beta}^{\mathrm{b}}) \\ \vdots \\ -p_K(\mathbf{x}_i, \boldsymbol{\beta}^{\mathrm{b}}) p_k(\mathbf{x}_i, \boldsymbol{\beta}^{\mathrm{b}}) \end{pmatrix}$$

for $k = \{1, ..., K\}$. Thus we know that

$$\frac{1}{N}\sum_{i=1}^{N} \frac{\partial \left\| \boldsymbol{p}(\mathbf{x}_i, \boldsymbol{\beta}^{\mathrm{b}}) - \boldsymbol{p}(\mathbf{x}_i, \hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{full}}) \right\|^2}{\partial \boldsymbol{\beta}^{\mathrm{b}}}\bigg|_{\boldsymbol{\beta}^{\mathrm{b}}=\hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{full}}} = 0. \tag{A.10}$$

Moreover, we have

$$\frac{1}{N}\sum_{i=1}^{N} \frac{\partial^2 \left\| \boldsymbol{p}(\mathbf{x}_i, \boldsymbol{\beta}^{\mathrm{b}}) - \boldsymbol{p}(\mathbf{x}_i, \hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{full}}) \right\|^2}{\partial \boldsymbol{\beta}^{\mathrm{b}} \partial \boldsymbol{\beta}^{\mathrm{b}\top}}\bigg|_{\boldsymbol{\beta}^{\mathrm{b}}=\hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{full}}} = \frac{2}{N}\sum_{i=1}^{N} \mathbf{B}_i^{\top}(\hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{full}}) \mathbf{B}_i(\hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{full}}), \tag{A.11}$$

where

$$\mathbf{B}_i(\boldsymbol{\beta}) = \begin{pmatrix} -p_0(\mathbf{x}_i, \boldsymbol{\beta}) p_1(\mathbf{x}_i, \boldsymbol{\beta}) & ... & -p_0(\mathbf{x}_i, \boldsymbol{\beta}) p_K(\mathbf{x}_i, \boldsymbol{\beta}) \\ p_1(\mathbf{x}_i, \boldsymbol{\beta}) - p_1^2(\mathbf{x}_i, \boldsymbol{\beta}) & ... & -p_1(\mathbf{x}_i, \boldsymbol{\beta}) p_K(\mathbf{x}_i, \boldsymbol{\beta}) \\ ... & ... & ... \\ -p_1(\mathbf{x}_i, \boldsymbol{\beta}) p_K(\mathbf{x}_i, \boldsymbol{\beta}) & ... & p_K(\mathbf{x}_i, \boldsymbol{\beta}) - p_K^2(\mathbf{x}_i, \boldsymbol{\beta}) \end{pmatrix} \otimes \mathbf{x}_i^{\top}.$$

The reminder term of (A.9) is

$$R_N = \frac{3}{N} \sum_{a_1+a_2+...+a_{Kd}=3} \frac{(\hat{\beta}^b_{\mathrm{sub},1} - \hat{\beta}^b_{\mathrm{full},1})^{a_1} (\hat{\beta}^b_{\mathrm{sub},2} - \hat{\beta}^b_{\mathrm{full},2})^{a_2} \dots (\hat{\beta}^b_{\mathrm{sub},Kd} - \hat{\beta}^b_{\mathrm{full},Kd})^{a_{Kd}}}{a_1! a_2! \dots a_{Kd}!}$$

$$\int_0^1 (1-t)^2 \frac{\partial^3 \sum_{i=1}^N \left\| \boldsymbol{p}(\mathbf{x}_i, \boldsymbol{\beta}^b) - \boldsymbol{p}(\mathbf{x}_i, \hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{full}}) \right\|^2}{\partial \beta_1^{b\,a_1} \partial \beta_2^{b\,a_2} \ldots \partial \beta_{Kd}^{b\;a_{Kd}}} \Bigg|_{\boldsymbol{\beta}^b = \hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{full}} + t(\hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{sub}} - \hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{full}})} dt, \tag{A.12}$$

where $\boldsymbol{a} = (a_1, a_2, \ldots, a_{Kd})$, $a_1, a_2, \ldots, a_{Kd} \geq 0$. By the fact that

$$\left\| \frac{1}{N} \frac{\partial^3 \sum_{i=1}^N \left\| \boldsymbol{p}(\mathbf{x}_i, \boldsymbol{\beta}^{\mathrm{b}}) - \boldsymbol{p}(\mathbf{x}_i, \hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{full}}) \right\|^2}{\partial \beta_1^{b\,a_1} \partial \beta_2^{b\,a_2} \ldots \partial \beta_{Kd}^{b\;a_{Kd}}} \right\| \leq \frac{L}{N} \sum_{i=1}^N \|\mathbf{x}_i\|^3,$$

where L is a positive integer, and combined with Assumption 1, we know that

$$\sup_t \left\| \frac{1}{N} \frac{\partial^3 \sum_{i=1}^N \left\| \boldsymbol{p}(\mathbf{x}_i, \boldsymbol{\beta}^{\mathrm{b}}) - \boldsymbol{p}(\mathbf{x}_i, \hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{full}}) \right\|^2}{\partial \beta_1^{a_1} \partial \beta_2^{a_2} \ldots \partial \beta_{Kd}^{a_{Kd}}} \Bigg|_{\boldsymbol{\beta}^b = \hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{full}} + t(\hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{sub}} - \hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{full}})} \right\| = O_{P|\mathcal{D}_N}(1). \tag{A.13}$$

From (A.12) and (A.13), we obtain

$$R_N = O_{P|\mathcal{D}_N}(\|\hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{sub}} - \hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{full}}\|^3). \tag{A.14}$$

According to (A.9), (A.10), (A.11) and (A.14), we have

$$\begin{aligned}
&\frac{1}{N} \sum_{i=1}^N \left\| \boldsymbol{p}(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_{\mathrm{sub}}) - \boldsymbol{p}(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_{\mathrm{full}}) \right\|^2 \\
&= \left(\hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{sub}} - \hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{full}} \right)^\top \frac{1}{N} \sum_{i=1}^N \left\{ \mathbf{B}_i^\top(\hat{\boldsymbol{\beta}}_{\mathrm{full}}) \mathbf{B}_i(\hat{\boldsymbol{\beta}}_{\mathrm{full}}) \right\} \left(\hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{sub}} - \hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{full}} \right) \\
&\quad + O_{P|\mathcal{D}_N} \left(\left\| \hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{sub}} - \hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{full}} \right\|^3 \right).
\end{aligned} \tag{A.15}$$

For clear presentation, denote

$$\boldsymbol{\Omega}^{\circ}_N = \frac{1}{N} \sum_{i=1}^N \left\{ \mathbf{B}_i^\top(\hat{\boldsymbol{\beta}}_{\mathrm{full}}) \mathbf{B}_i(\hat{\boldsymbol{\beta}}_{\mathrm{full}}) \right\}.$$

From Theorem 1 of [1], we have

$$\hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{sub}} - \hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{full}} = O_{P|\mathcal{D}_N}(n^{-1/2}). \tag{A.16}$$

From (A.15) and (A.16),

$$\frac{n}{N} \sum_{i=1}^N \left\| \boldsymbol{p}(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_{\mathrm{sub}}) - \boldsymbol{p}(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_{\mathrm{full}}) \right\|^2$$

$$
\begin{aligned}
&= n\left(\hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{sub}} - \hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{full}}\right)^{\top} \boldsymbol{\Omega}^{\circ}_{N} \left(\hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{sub}} - \hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{full}}\right) + O_{P|\mathcal{D}_N}\left(n^{-1/2}\right) \\
&= \sqrt{n}\left(\hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{sub}} - \hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{full}}\right)^{\top} \mathbf{V}^{\circ}_{N}{}^{-1/2}\mathbf{V}^{\circ}_{N}{}^{1/2}\boldsymbol{\Omega}^{\circ}_{N}\mathbf{V}^{\circ}_{N}{}^{1/2}\sqrt{n}\mathbf{V}^{\circ}_{N}{}^{-1/2}\left(\hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{sub}} - \hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{full}}\right) \\
&\quad + O_{P|\mathcal{D}_N}\left(n^{-1/2}\right).
\end{aligned}
$$

By continuous-mapping theorem, we have that

$$
\begin{aligned}
&\sqrt{n}\left(\hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{sub}} - \hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{full}}\right)^{\top} \mathbf{V}^{\circ}_{N}{}^{-1/2}\mathbf{V}^{\circ}_{N}{}^{1/2}\boldsymbol{\Omega}^{\circ}_{N}\mathbf{V}^{\circ}_{N}{}^{1/2}\sqrt{n}\mathbf{V}^{\circ}_{N}{}^{-1/2}\left(\hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{sub}} - \hat{\boldsymbol{\beta}}^{\mathrm{b}}_{\mathrm{full}}\right) \\
&\quad \overset{a}{\sim} \mathbf{z}^{\top}\mathbf{V}^{\circ}_{N}{}^{1/2}\boldsymbol{\Omega}^{\circ}_{N}\mathbf{V}^{\circ}_{N}{}^{1/2}\mathbf{z}
\end{aligned}
$$

conditionally on the full data, where $\mathbf{z} \sim \mathbb{N}(0, \mathbf{I})$ and $\overset{a}{\sim}$ means the two quantities have the same asymptotic distribution.

□

Proof of Theorem 4

Proof. The asymptotic mean of the mean squared prediction error is

$$
\mathbb{E}\{\mathbf{z}^{\top}\mathbf{V}^{\circ}_{N}{}^{1/2}\boldsymbol{\Omega}^{\circ}_{N}\mathbf{V}^{\circ}_{N}{}^{1/2}\mathbf{z}|\mathcal{D}_N\} = \mathrm{tr}\left(\mathbf{V}^{\circ}_{N}\boldsymbol{\Omega}^{\circ}_{N}\right),
$$

which follows

$$
\begin{aligned}
\mathrm{tr}\left(\mathbf{V}^{\circ}_{N}\boldsymbol{\Omega}^{\circ}_{N}\right) &= \mathrm{tr}\left\{\mathbf{V}^{\circ}_{N}\frac{1}{N}\sum_{i=1}^{N}\mathbf{B}_i^{\top}(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\mathbf{B}_i(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\right\} \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathrm{tr}\left[\{\mathbf{M}^{\circ}_{N}\}^{-1}\mathbf{D}^{\circ}_{N}\{\mathbf{M}^{\circ}_{N}\}^{-1}\mathbf{B}_i^{\top}(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\mathbf{B}_i(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathrm{tr}\left[\left\{\frac{1}{N}\sum_{j=1}^{N}\frac{\boldsymbol{\psi}^{\circ}_{j}(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\otimes(\mathbf{x}_j\mathbf{x}_j^{\top})}{N\pi_j}\right\}\{\mathbf{M}^{\circ}_{N}\}^{-1}\mathbf{B}_i^{\top}(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\mathbf{B}_i(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\{\mathbf{M}^{\circ}_{N}\}^{-1}\right] \\
&= \frac{1}{N^3}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathrm{tr}\left\{\frac{\boldsymbol{\psi}^{\circ}_{j}(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\otimes(\mathbf{x}_j\mathbf{x}_j^{\top})}{\pi_j}\{\mathbf{M}^{\circ}_{N}\}^{-1}\mathbf{B}_i^{\top}(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\mathbf{B}_i(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\{\mathbf{M}^{\circ}_{N}\}^{-1}\right\} \\
&= \frac{1}{N^3}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\|\mathbf{B}_i(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\{\mathbf{M}^{\circ}_{N}\}^{-1}\{\mathbf{s}^{\circ}_{j}(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\otimes\mathbf{x}_j\}\|^2}{\pi_j}\times\sum_{j=1}^{N}\pi_j \\
&= \frac{1}{N^3}\sum_{j=1}^{N}\frac{\sum_{i=1}^{N}\|\mathbf{B}_i(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\{\mathbf{M}^{\circ}_{N}\}^{-1}\{\mathbf{s}^{\circ}_{j}(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\otimes\mathbf{x}_j\}\|^2}{\pi_j}\times\sum_{j=1}^{N}\pi_j \\
&\geq \left\{\frac{1}{N^{3/2}}\sum_{j=1}^{N}\sqrt{\sum_{i=1}^{N}\|\mathbf{B}_i(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\{\mathbf{M}^{\circ}_{N}\}^{-1}\{\mathbf{s}^{\circ}_{j}(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\otimes\mathbf{x}_j\}\|^2}\right\}^2,
\end{aligned}
$$

where the last inequality is from the Cauchy-Schwarz inequality.

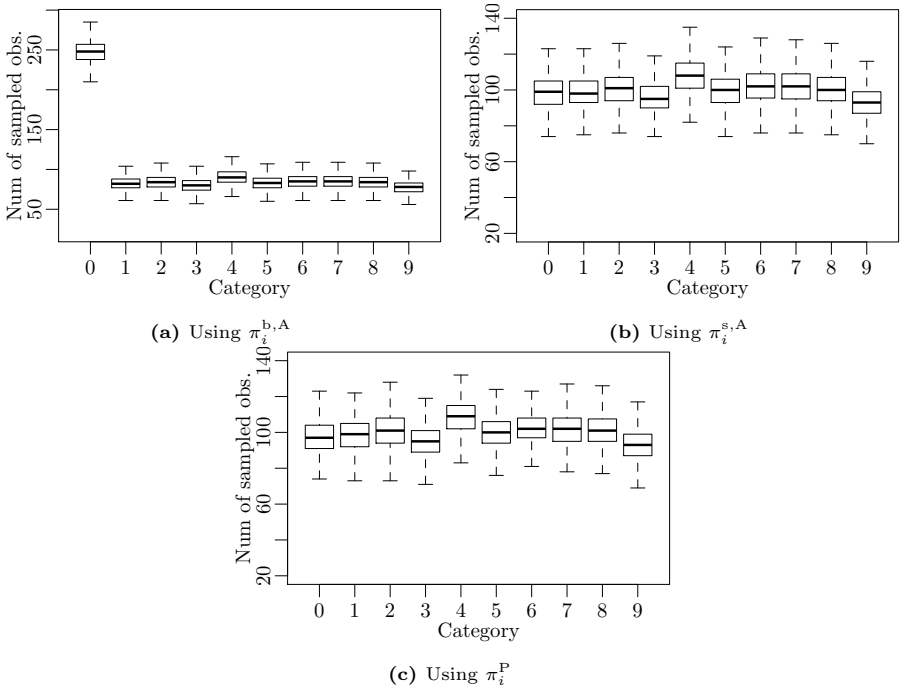
Thus the optimal subsampling probabilities are

$$\pi_i^{\mathrm{P}} = \frac{\|{\boldsymbol{\Omega}_N^{\circ}}^{1/2}{\mathbf{M}_N^{\circ}}^{-1}\{\mathbf{s}_i^{\circ}(\hat{\boldsymbol{\beta}}_{\mathrm{full}}) \otimes \mathbf{x}_i\}\|}{\sum_{i=1}^{N} \|{\boldsymbol{\Omega}_N^{\circ}}^{1/2}{\mathbf{M}_N^{\circ}}^{-1}\{\mathbf{s}_i^{\circ}(\hat{\boldsymbol{\beta}}_{\mathrm{full}}) \otimes \mathbf{x}_i\}\|}, \quad i \in \{1, \ldots, n\}.$$

□

Appendix B. Numerical Results

Figure B.5 compares the number of sampled observations for different categories drawn by different optimal subsampling probabilities with the synthetic data generated in Section 2. More observations are drawn from the baseline category with $\pi_i^{\mathrm{b,A}}$. For summation constraint, we have roughly equal samples chosen from all 10 categories under the A-optimality criterion. Figure B.5 also shows that samples drawn by π_i^{P} are evenly distributed among all categories with this balanced synthetic dataset.



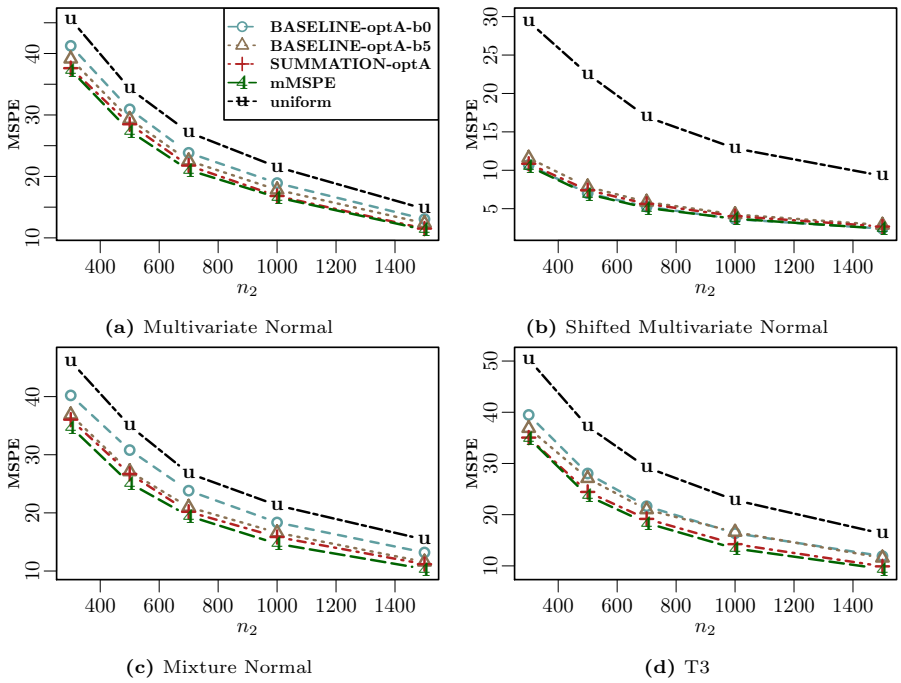
(a) Using $\pi_i^{\mathrm{b,A}}$ **(b)** Using $\pi_i^{\mathrm{s,A}}$

(c) Using π_i^{P}

Figure B.5: Boxplot for average number of sampled observations for all categories using different subsampling probabilities for the synthetic data when $N = 10000$ and $n = 1000$ of 1000 iterations. The covariates are generated from $\mathbb{N}_3(\mathbf{0}, \mathbf{I}_3)$ and the response has 10 different categories. The true coefficient is $\boldsymbol{\beta}^{\mathrm{b}} = \mathbf{0}_{27}$.

Figure B.6 compares the prediction efficiency among different algorithms with different subsampling probabilities. Obviously the optimal subsampling

algorithms have smaller empirical mean squared prediction error compared to the uniform subsampling. Using optimal subsampling probabilities obtained by minimizing the mean squared prediction error offers the highest prediction accuracy.



(a) Multivariate Normal

(b) Shifted Multivariate Normal

(c) Mixture Normal

(d) T3

Figure B.6: Average mean squared prediction error for different types of optimal subsampling probabilities for all four cases with 1000 replicates when $n_1 = 300$. BASELINE-optA-b0 uses Algorithm 1 with $\boldsymbol{\mathcal{L}}(\hat{\boldsymbol{\beta}}^1_{\text{sub}}) = \mathbf{M}_N^{\circ *\,-1}$ for (8) and treats the outcome 0 as the baseline category. BASELINE-optA-b5 uses Algorithm 1 with $\boldsymbol{\mathcal{L}}(\hat{\boldsymbol{\beta}}^1_{\text{sub}}) = \mathbf{M}_N^{\circ *\,-1}$ for (8) and treats the outcome 5 as the baseline category. SUMMATION-optL means Algorithm 1 with $\boldsymbol{\mathcal{L}}(\hat{\boldsymbol{\beta}}^1_{\text{sub}}) = \mathbf{M}_N^{*\,-1}$ for (8) and mMSPE stands for Algorithm 1 with $\boldsymbol{\mathcal{L}}(\hat{\boldsymbol{\beta}}^1_{\text{sub}})$ replaced by $\boldsymbol{\Omega}_N^{\circ *1/2}\mathbf{M}_N^{\circ *\,-1}$ for (8).

References

[1] Y. Yao, H. Wang, Optimal subsampling for softmax regression, Statistical Papers 60 (2) (2019) 585–599.

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

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

[4] P. Drineas, M. W. Mahoney, S. Muthukrishnan, 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, 2006, pp. 1127–1136.

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

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

[7] P. Ma, X. Zhang, X. Xing, J. Ma, M. Mahoney, Asymptotic analysis of sampling estimators for randomized numerical linear algebra algorithms, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 1026–1035.

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

[9] L. Han, K. M. Tan, T. Yang, T. Zhang, et al., Local uncertainty sampling for large-scale multiclass logistic regression, Annals of Statistics 48 (3) (2020) 1770–1788.

[10] H. Wang, R. Zhu, P. Ma, Optimal subsampling for large sample logistic regression, Journal of the American Statistical Association 113 (522) (2018) 829–844. doi:10.1080/01621459.2017.1292914.

[11] M. Ai, J. Yu, H. Zhang, H. Wang, Optimal subsampling algorithms for big data regressions, Statistica Sinica 31 (2021) 749–772. doi:10.5705/ss.202018.0439.

[12] J. Yu, H. Wang, M. Ai, H. Zhang, Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data, Journal of the American Statistical Association (2020) 1–12.

[13] H. Wang, Y. Ma, Optimal subsampling for quantile regression in big data, Biometrika 108 (1) (2021) 99–112.

[14] N. V. Chawla, Data mining for imbalanced datasets: An overview, in: Data mining and knowledge discovery handbook, Springer, 2009, pp. 875–886.

[15] H. Wang, Logistic regression for massive data with rare events, in: H. D. III, A. Singh (Eds.), Proceedings of the 37th International Conference on Machine Learning, Vol. 119 of Proceedings of Machine Learning Research, PMLR, 2020, pp. 9829–9836.
URL http://proceedings.mlr.press/v119/wang20a.html

[16] H. Wang, A. Zhang, C. Wang, Nonuniform negative sampling and log odds correction with rare events data, in: Proceedings of The 35 Conference on Neural Information Processing Systems (NeurIPS 2021)., Proceedings of Machine Learning Research, PMLR, 2021.

[17] J. Wang, J. Zou, H. Wang, Sampling with replacement vs poisson sampling: a comparative study in optimal subsampling, IEEE Transactions on Information Theory (2022) 10.1109/TIT.2022.3176955.
URL https://doi.org/10.1109/TIT.2022.3176955

[18] J. A. Blackard, Comparison of neural networks and discriminant analysis in predicting forest cover types, Colorado State University, 1998.

[19] D. Dheeru, E. Karra Taniskidou, UCI machine learning repository (2019).
URL http://archive.ics.uci.edu/ml

[20] R. A. Horn, C. R. Johnson, Matrix analysis, Cambridge university press, 2012.