


Article

Asymptotics of Subsampling for Generalized Linear Regression Models under Unbounded Design

Guangqiang Teng¹, Boping Tian^{1,*}, Yuanyuan Zhang^{2,*}  and Sheng Fu³

¹ School of Mathematics, Harbin Institute of Technology, Harbin 150001, China

² School of Mathematical Sciences, Soochow University, Suzhou 215006, China

³ Department of Industrial Systems Engineering & Management, National University of Singapore, 21 Lower Kent Ridge Road, Singapore 119077, Singapore

* Correspondence: bopingt361147@hit.edu.cn (B.T.); zhangyy@suda.edu.cn (Y.Z.)

† These authors contributed equally to this work.

Abstract: The optimal subsampling is an statistical methodology for generalized linear models (GLMs) to make inference quickly about parameter estimation in massive data regression. Existing literature only considers bounded covariates. In this paper, the asymptotic normality of the subsampling M-estimator based on the Fisher information matrix is obtained. Then, we study the asymptotic properties of subsampling estimators of unbounded GLMs with nonnatural links, including conditional asymptotic properties and unconditional asymptotic properties.

Keywords: generalized linear models; massive data; nonnatural links; unbounded covariates; unconditional subsampling estimator



Citation: Teng, G.; Tian, B.; Zhang, Y.; Fu, S. Asymptotics of Subsampling for Generalized Linear Regression Models under Unbounded Design. *Entropy* **2023**, *25*, 84. <https://doi.org/10.3390/e25010084>

Academic Editors: Augustine Wong and Xiaoping Shi

Received: 8 November 2022

Revised: 27 December 2022

Accepted: 28 December 2022

Published: 31 December 2022



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In recent years, the amount of information that people need to process is increasing dramatically. It is of great challenge to directly process massive data for statistical analysis. The divide-and-conquer strategy can mitigate the challenge of directly processing such big data [1], but it still consumes considerable computing resources. As a cheaper alternative in computing, subsampling gains its value in the case of limited computing resources.

To reduce the burden on the machine, the subsampling strategy based on big data has been given more attention in recent years. Ref. [2] proposes simple necessary and sufficient conditions for a convolved subsampling estimator to produce a normal limit that matches the target of bootstrap estimation; Ref. [3] provides an optimally distributed subsampling for maximum quasi-likelihood estimators with massive data; Ref. [4] studies some adaptive optimal subsampling algorithms; and Ref. [5] describes a subdata selection method based on leverage scores which conduct the linear model selection on a small subdata set.

GLM is a kind of statistical model with a wide range of applications such as [6–8]. Many subsampling studies are based on GLMs such as [3,9,10]. However, the covariates of the subsampled GLMs in the literature are bounded. When dealing with some big data problems, the size of covariate is not strictly bounded, such as the number of clicks on a web page, which can grow infinitely. This requires the extension of existing theories to the unbounded design. To fill this gap, this paper aims to study asymptotic properties of the subsampled GLMs with unbounded covariates based on empirical process and martingale technology.

Our three contributions are shown as follows: (1) we describe the asymptotic property of subsampled M-estimator using Fisher information matrix; (2) we give the conditional consistency and asymptotic normality of unbounded GLMs subsampling estimator; (3) we provide the unconditional consistency and asymptotic normality of unbounded GLMs subsampling estimator.

The rest of the paper is organized as follows. Section 2 introduces the basic concepts in GLMs and subsampling M-estimation problem. Section 3 presents the asymptotical properties for unbounded GLMs subsampling estimators. Section 4 gives the conclusion and discussion, as well as future research directions. All the technical proofs are collected in the Appendix A.

2. Preliminaries

This section introduces the subsampling M-estimation problem and GLMs.

2.1. Subsampling M-Estimation

Let $\{l(\beta; \mathbf{Z}) \in \mathbb{R} | \mathbf{Z} \in \mathcal{Z}\}$ be a set of loss functions with a finite dimensional convex set $\beta \in \Theta \subset \mathbb{R}^p$, and $U = \{1, 2, \dots, N\}$ be the index of the full large dataset with σ -algebra $\mathcal{F}_N = \sigma(\mathbf{Z}_1, \dots, \mathbf{Z}_N)$, where for each $i \in U$, the random data point $\mathbf{Z}_i \in \mathcal{Z}$ (some probability space) is observed. The empirical risk $L_N : \Theta \rightarrow \mathbb{R}$ is given by $L_N(\beta) = \frac{1}{N} \sum_{i \in U} l(\beta; \mathbf{Z}_i)$.

The goal is to get the solution $\hat{\beta}_N$ to minimize the risk, namely

$$\hat{\beta}_N = \arg \min_{\beta \in \Theta} L_N(\beta). \tag{1}$$

To solve Equation (1), we need $\hat{\beta}_N$ satisfy: $\nabla L_N(\beta) = \frac{1}{N} \sum_{i \in U} \nabla l(\beta; \mathbf{Z}_i) = \mathbf{0}$, and let $\Sigma_N := \nabla^2 L_N(\hat{\beta}_N)$. This is an M-estimation problem; see [11]. For fast solving large-scale estimation in Equation (1), we propose the subsampling M-estimation. Consider an index set $S = \{i_1, i_2, \dots, i_n\}$ with replacement from U according to the sampling probability $\{\pi_i\}_{i=1}^N$ such that $\sum_{i=1}^N \pi_i = 1$. The subsampling M-estimation problem is to obtain the solution $\hat{\beta}_n$ satisfying

$$\nabla L_n(\beta) = \mathbf{0} \text{ with } L_n(\beta) = \frac{1}{Nn} \sum_{i \in S} \frac{1}{\pi_i^*} l(\beta; \mathbf{Z}_i^*),$$

where \mathbf{Z}_i^* is the i -th time subsample with replacement and π_i^* is the subsampling probability of \mathbf{Z}_i^* . For example, if $\mathbf{Z}_1^* = \mathbf{Z}_1$, then $\pi_1^* = \pi_1$; if $\mathbf{Z}_2^* = \mathbf{Z}_1$, then $\pi_2^* = \pi_1$. Denote a_i as the number of i -th subsampled data such that $\sum_{i \in U} a_i = n$. And $L_n(\beta)$ is constructed by inverse probability weighting skill such that $E[L_n(\beta) | \mathcal{F}_N] = L_N(\beta)$; see [12]. Details about properties of conditional expectation are shown in [13].

2.2. Generalized Linear Models

Let the random variable Y be the distribution of the natural exponential families P_α indexed by parameter α ,

$$P_\alpha(dy) = dF_Y(y) = c(y) \exp\{y\alpha - b(\alpha)\} \nu(dy), c(y) > 0, \tag{2}$$

where α is often referred to as the canonical parameter belonging to its natural space

$$\Lambda = \{\alpha : \int c(y) \exp\{y\alpha\} \nu(dy) < \infty\}.$$

$\nu(\cdot)$ is the Lebesgue measure for continuous distributions (Normal, Gamma) or counting measure for discrete distributions (binomial, Poisson, negative binomial). The $c(y)$ is free of α .

Let $\{(Y_i, \mathbf{X}_i)\}_{i=1}^N$ be N independent sample data pairs. Here the $\mathbf{X}_i \in \mathbb{R}^p$ is covariates and we assume that the response Y_i follows the distribution of the natural exponential families with the parameter $\alpha_i \in \Lambda$. The covariates $\mathbf{X}_i := (x_{i1}, \dots, x_{ip})^T, (i = 1, 2, \dots, N)$ are supposed to be deterministic.

The conditional expectation of Y_i for a given X_i is defined as a function of $\beta^T X_i$ after a transformation by a link function $\alpha_i = \psi(\beta^T X_i)$. The mean value denoted as $\mu_i := E(Y_i)$ is mostly considered for regression.

If $\alpha_i = \beta^T X_i$ then we call that $\psi(\beta^T X_i) = \beta^T X_i$ is canonical (or natural) link function, and corresponding model is canonical (or natural) GLMs; see Page 32 in [14]. Sometimes the assumption $\alpha_i = \beta^T X_i$ is somewhat strong and not very suitable in practice, while nonnatural link GLMs allow more flexible choices for the link function. We can further assume that α_i and $\beta^T X_i$ can be related by a nonnatural link function $\alpha_i = \psi(\beta^T X_i)$.

Let $f_\beta(Y_i|X_i)$ be the joint density function of the i.i.d. data $\{(Y_i, X_i)\}_{i=1}^N$ from the exponential family with a link function $\psi(\cdot)$. Then the nonnatural GLMs [15] is defined by

$$Y_i|X_i \sim f_\beta(Y_i|X_i) = c(Y_i) \exp\left\{Y_i \psi(\beta^T X_i) - b(\psi(\beta^T X_i))\right\}, \quad i = 1, 2, \dots, N. \quad (3)$$

Here is a classic result for the exponential family (3),

$$E(Y_i|X_i) := \mu_i = \dot{b}(\alpha_i) = \dot{b}(\psi(\beta^T X_i)) \quad \text{and} \quad \text{Var}(Y_i|X_i) := \text{Var}(Y_i) = \ddot{b}(\alpha_i), \quad (4)$$

where $i = 1, 2, \dots, N$; see P280 in [16].

3. Main Results

3.1. Subsampling M-Estimation Problem

In this part we first look at the term $\Sigma_N^{-1} \nabla L_n(\hat{\beta}_N)$. Define an independent random vector sequence $\{\zeta_j\}_{j=1}^N$ and the subsampled $\{\zeta_j^*\}_{j=1}^n$, such that each vector ζ takes the value among $\{\frac{1}{N\pi_i} \Sigma_N^{-1} \nabla l(\hat{\beta}_N; Z_i)\}_{i=1}^N$, and let

$$V_M(\hat{\beta}_N; n) = \frac{1}{N^2 n} \Sigma_N^{-1} \left[\sum_{i \in U} \frac{1}{\pi_i} \nabla l(\hat{\beta}_N; Z_i) \nabla^T l(\hat{\beta}_N; Z_i) \right] \Sigma_N^{-1}.$$

From the definition of $\nabla L_N(\beta)$, we have $E(\zeta|\mathcal{F}_N) = \Sigma^{-1} \nabla L_N(\hat{\beta}_N) = \mathbf{0}$ and $\text{Var}(\zeta|\mathcal{F}_N) = nV_M(\hat{\beta}_N; n)$. Then we have the asymptotic property of subsampled M-estimator.

Theorem 1. Suppose that the risk function $L_N(\beta)$ is twice differentiable and λ -strongly convex over Θ , that is, for $\beta \in \Theta$, $\nabla^2 L_N(\beta) \geq \lambda \mathbf{I}$, where \geq denotes the semidefinite positive ordering; and the sampling-based moment condition,

$$\frac{1}{N^4} \sum_{i=1}^N \frac{1}{\pi_i^3} \left\| \nabla l(\hat{\beta}_N; Z_i) \right\|^4 = O_P(1).$$

Then we can obtain: As $n \rightarrow \infty$, conditioning on \mathcal{F}_N ,

$$V_M(\hat{\beta}_N; n)^{-\frac{1}{2}} (\hat{\beta}_n - \hat{\beta}_N) \xrightarrow{d} N(0, \mathbf{I}_p), \quad (5)$$

where \xrightarrow{d} means convergence in distribution.

Theorem 1 reveals that the subsampling M-estimation scheme is theoretically feasible under mild conditions. In addition, the existence of the estimator is given by the Fisher information matrix.

3.2. Conditional Asymptotic Properties of Subsampled GLMs with Unbounded Covariates

The exponential family is very versatile for containing many common light-tail distributions such as binomial, Poisson, negative binomial, normal and Gamma. Along with their attendant convexity properties which leads to finite variance property for log-density, they can serve for a large amount of popular and effective statistical models. It is pre-

cisely because of the commonality of these distributions so that we study the subsampling problem for GLMs.

From the loss function introduced in Section 2.1, we set $l(\beta; \mathbf{Z}_i) := -\log f_\beta(Y_i|\mathbf{X}_i)$ where $f_\beta(Y_i|\mathbf{X}_i)$ is defined by Equation (2), then the problem solving the minimum of the loss function is equivalent to solve the maximum of the likelihood function. For simplicity, we assume that $c(y) = 1$, then

$$\nabla l(\beta; \mathbf{Z}_i) := -\frac{\partial \log f_\beta(Y_i|\mathbf{X}_i)}{\partial \beta} = -\left[Y_i - b\left(\psi\left(\beta^T \mathbf{X}_i\right)\right)\right] \psi\left(\beta^T \mathbf{X}_i\right) \mathbf{X}_i$$

with the nonnatural link function $\alpha_i = \psi(\beta^T \mathbf{X}_i)$. We also use this idea in Section 3.3.

More generally, we consider a wider class saying quasi-GLMs, rather than GLMs, which assumes that Equation (4) holds for a certain function $\mu(\cdot)$. Strong consistency and asymptotic normality of quasi maximum likelihood estimate in GLMs with bounded covariates are proved in [17]. For unbounded covariates, adopting the subsampled estimation of GLMs in [9], we calculate the inverse probability weighted estimator of β by solving the estimating equation based on the subsampled index set S ,

$$-\frac{1}{Nn} \sum_{i \in S} \frac{1}{\pi_i^*} \left[Y_i^* - \mu\left(\psi\left(\beta^T \mathbf{X}_i^*\right)\right) \right] \psi\left(\beta^T \mathbf{X}_i^*\right) \mathbf{X}_i^* = \mathbf{0}.$$

where $\{(Y_i^*, \mathbf{X}_i^*)\}_{i \in S}$ is subsampled data. Equivalently, we have

$$s_n(\beta) = \sum_{i \in S} \frac{1}{\pi_i^*} \left[Y_i^* - \mu\left(\psi\left(\beta^T \mathbf{X}_i^*\right)\right) \right] \psi\left(\beta^T \mathbf{X}_i^*\right) \mathbf{X}_i^* = \mathbf{0}. \tag{6}$$

Equation (6) is called quasi-GLMs since Equation (4) is given instead of the distribution function.

Let $\hat{\beta}_n$ be the estimator of the real parameter β_0 in subsampled quasi-GLMs and $\hat{\beta}_N$ be the estimator of β_0 in quasi-GLMs with full data. For the unbounded quasi-GLMs with full data, $\hat{\beta}_N$ is asymptotic unbiased with respect to β_0 ; see [18]. Next, we focus on the asymptotical properties of $\hat{\beta}_n$, as shown in the following theorems.

Theorem 2. Let $\{(Y_i^*, \mathbf{X}_i^*)\}_{i \in S}$ be subsampled from i.i.d. full data $\{(Y_i, \mathbf{X}_i)\}_{i \in U}$. Consider the Equation (4) and (6) where $\psi(\cdot)$ is three times continuously differentiable whose every derivative is bounded, and $b(\cdot)$ is twice continuously differentiable whose every derivative is also bounded. Assume that:

(A.1) The range of the unknown parameter β is an open subset of \mathbb{R}^p .

(A.2) For any $i \in S$, $E \sup_{\beta \in \Theta} \left[\frac{1}{\pi_i^*} |Y_i^* - \mu(\psi(\beta^T \mathbf{X}_i^*))| \mid \mathcal{F}_N \right] = O(1)$.

(A.3) For any $\beta \in \Theta$ and $i \in S$, $0 < \inf_i \varphi(\beta^T \mathbf{X}_i^*) \leq \sup_i \varphi(\beta^T \mathbf{X}_i^*) < \infty$, where $\varphi(t) = [\psi(t)]^2 \ddot{b}(\psi(t))$.

(A.4) For any $\beta_1 \in \Theta$ and $\beta_2 \in \Theta$, there exists a function $|m(\mathbf{X}_i^*)| < \infty$ such that

$$|\varphi(\beta_1^T \mathbf{X}_i^*) - \varphi(\beta_2^T \mathbf{X}_i^*)| \leq |m(\mathbf{X}_i^*)| |\beta_1^T \mathbf{X}_i^* - \beta_2^T \mathbf{X}_i^*|.$$

(A.5) When $n \rightarrow \infty$, $\max_{i \in S} \mathbf{X}_i^{*T} (\mathbf{X}^* \mathbf{X}^{*T})^{-1} \mathbf{X}_i^* = O(n^{-1})$ and $\lambda_{\min}[\mathbf{X}^* \mathbf{X}^{*T}] \rightarrow \infty$, where $\mathbf{X}^* = (\mathbf{X}_1^*, \dots, \mathbf{X}_n^*)$ and $\lambda_{\min}[\mathbf{A}]$ is the smallest eigenvalue of the matrix \mathbf{A} .

(A.6) $\min_{i=1, \dots, N} (N\pi_i) = O(1)$, $\max_{i=1, \dots, N} (N\pi_i) = O(1)$.

Then $\hat{\beta}_n$ is consistent with $\hat{\beta}_N$, i.e.,

$$\|\hat{\beta}_n - \hat{\beta}_N\| = o_{P|\mathcal{F}_N}(1)$$

where $o_{P|\mathcal{F}_N}(1)$ means $o(1)$ conditioning on \mathcal{F}_N in probability.

Theorem 3. Under the conditions in Theorem 2, as $N \rightarrow \infty$ and $n \rightarrow \infty$, conditional on \mathcal{F}_N in probability,

$$\sqrt{n}(\hat{\beta}_n - \hat{\beta}_N) \rightarrow N(\mathbf{0}, \mathbf{V}_s),$$

in distribution, where

$$\begin{aligned} \mathbf{V}_s &= \Sigma_N^{-1} \mathbf{V}_N \Sigma_N^{-1}, \\ \Sigma_N &= \sum_{i \in U} a_i \left[Y_i - b(\psi(\hat{\beta}_N^T \mathbf{X}_i)) \right] \ddot{\psi}(\hat{\beta}_N^T \mathbf{X}_i) \mathbf{X}_i \mathbf{X}_i^T - \sum_{i \in U} a_i \ddot{b}(\psi(\hat{\beta}_N^T \mathbf{X}_i)) \psi(\hat{\beta}_N^T \mathbf{X}_i)^2 \mathbf{X}_i \mathbf{X}_i^T, \\ \mathbf{V}_N &= \sum_{i \in U} \frac{a_i}{\pi_i} \left[Y_i - b(\psi(\hat{\beta}_N^T \mathbf{X}_i)) \right]^2 \ddot{\psi}(\hat{\beta}_N^T \mathbf{X}_i)^2 \mathbf{X}_i \mathbf{X}_i^T. \end{aligned}$$

In this part, we complete the asymptotic properties without the moment condition of the covariates $\{\mathbf{X}_i\}_{i=1}^N$ which is used in [9], and that means \mathbf{X}_i 's are unbounded. Here we only provide the theoretical asymptotic results. Furthermore, the subsampling probability can be derived by A-optimal criterion like [10].

3.3. Unconditional Asymptotic Properties of Subsampled GLMs with Unbounded Covariates

In real engineering, the measurement of some response variable data is very expensive, such as superconductor data, deep space exploration data, etc. The accuracy of estimating the target parameters under measurement constraints of responses is a very important issue. Ref. [19] completed the unconditional asymptotic properties of parameter estimation in bounded GLMs with canonical link. But the unbounded GLMs with nonnatural link situation has not been discussed yet.

In this section, we continue to use the notations of Section 3.2. Through the theory of empirical process [11], we obtain the unconditional consistency of $\hat{\beta}_n$ in the following theorem.

Theorem 4. (Unconditional subsampled consistency) Assume the conditions:

(B.1) $\lambda_{\min}(\mathbf{E}\mathbf{X}\mathbf{X}^T) > 0$ where \mathbf{X} is the unbounded covariate of GLMs.

(B.2) For $\forall u_1, u_2 \in [0, 1]$,

$$\inf_{\beta \in \Theta \setminus \{\beta_0\}} \frac{\mathbf{E}\{\ddot{b}(\tilde{\psi}_{u_1}) \psi[(1 - u_2)(\beta_0^T \mathbf{X}) + u_2(\beta^T \mathbf{X})]^2 (\beta^T \mathbf{X} - \beta_0^T \mathbf{X})^2\}}{\mathbf{E}(\beta^T \mathbf{X} - \beta_0^T \mathbf{X})^2} \geq C_1 > 0,$$

where $\tilde{\psi}_{u_1} = (1 - u_1)\psi(\beta_0^T \mathbf{X}) + u_1\psi(\beta^T \mathbf{X})$ and $\ddot{b}(\cdot)$ is the second derivative with respect to β .

(B.3)

$$\mathbf{E}_{\beta_0} \sup_{\beta \in \Theta} [|Y - b(\psi(\beta^T \mathbf{X}))| \cdot \|\mathbf{X}\|^2] < \infty,$$

where $\dot{b}(\cdot)$ is the first derivative with respect to β .

(B.4) $\psi(\cdot)$ in (3) is twice continuously differentiable and its every derivative has a positive minimum.

(B.5) $b(\cdot)$ in (3) is twice continuously differentiable and its every derivative has a positive minimum.

Then $\|\hat{\beta}_n - \beta_0\| = o_P(1)$.

Theorem 4 directly obtains the unconditional consistency of the subsampling estimator with respect to the true parameters under the unbounded assumption.

To prove the asymptotic normality of $\hat{\beta}_n$ with respect to β_0 , we briefly review the subsampled score function in Section 3.2

$$s_n(\beta) = \sum_{i \in S} \frac{1}{\pi_i^*} \left[Y_i^* - \mu(\psi(\beta^T \mathbf{X}_i^*)) \right] \dot{\psi}(\beta^T \mathbf{X}_i^*) \mathbf{X}_i^* := \sum_{i \in S} \frac{1}{\pi_i^*} \phi_{\beta}(\mathbf{X}_i^*, Y_i^*).$$

Next we will apply a multivariate martingale central limit theorem (Lemma 4 in [19]), which is the extension of Theorem A.1 in [20], to show the asymptotic normality of $\hat{\beta}_n$. Let

$\{\mathcal{F}_{N,i}\}_{i=1}^n$ be a filtration adaptive to the sampling: $\mathcal{F}_{N,0} = \sigma(\mathbf{X}_1^N, Y_1^N)$; $\mathcal{F}_{N,1} = \sigma(\mathbf{X}_1^N, Y_1^N) \vee \sigma(*_1); \dots; \mathcal{F}_{N,i} = \sigma(\mathbf{X}_1^N, Y_1^N) \vee \sigma(*_1) \vee \dots \vee \sigma(*_i); \dots$, where $\sigma(*_i)$ is the σ -algebra generated by i th sampling step. The subsample of size n is assumed to increase with N . By the filtration, we define the martingale

$$\bar{M} := \sum_{i=1}^n \bar{M}_i := \sum_{i=1}^n \left[\frac{1}{\pi_i^*} \phi_{\beta}(\mathbf{X}_i^*, Y_i^*) - \sum_{j=1}^N \phi_{\beta}(\mathbf{X}_j, Y_j) \right],$$

where $\{\bar{M}_i\}_{i=1}^n$ is a martingale difference sequence adapted to $\{\mathcal{F}_{N,i}\}_{i=1}^n$. In addition, define $Q := n \sum_{j=1}^N \phi_{\beta}(\mathbf{X}_j, Y_j)$; $T := s_n(\beta) = \bar{M} + Q$; $\zeta_{Ni} := \text{Var}^{-1/2}(T)\bar{M}_i$ and $B_N := \text{Var}^{-1/2}(T)\text{Var}(\bar{M})\text{Var}^{-1/2}(T)$, where matrix $A^{1/2}$ is the symmetric square root of A , i.e., $A = (A^{1/2})^2$, and $A^{-1/2} = (A^{1/2})^{-1} = (A^{-1})^{1/2}$. B_N is the variance of $\text{Var}^{-1/2}(T)\bar{M}$.

The following theorem shows the asymptotic normality of the estimator $\hat{\beta}_n$.

Theorem 5. Assume the conditions,

(C.1)

$$\Phi = E(\nabla s_n(\beta)) = E \left[- \sum_{i \in S} \frac{1}{\pi_i^*} \dot{\mu}(\psi(\beta^T \mathbf{X}_i^*)) [\psi(\beta^T \mathbf{X}_i^*)]^2 \mathbf{X}_i^* \mathbf{X}_i^{*T} \right]$$

is finite and nonsingular.

$$(C.2) \ E \left\{ \left[\sum_{i \in U} \frac{a_i}{\pi_i} \dot{\mu}(\psi(\beta^T \mathbf{X}_i)) [\psi(\beta^T \mathbf{X}_i)]^2 X_{ik} X_{ij} \right]^2 \right\} = o_P(1), \text{ for } 1 \leq k, j \leq p,$$

where X_{ik} means k -th element of vector \mathbf{X}_i and X_{ij} means j -th element of vector \mathbf{X}_i .

(C.3) $\psi(x)$ is three-times continuously differentiable for every x with its domain.

(C.4) For any $i \in S$, $\|\ddot{\phi}_{\beta}(\mathbf{X}_i^*, Y_i^*)\| < \infty$.

(C.5) $\min_{i=1, \dots, N} (N\pi_i) = \max_{i=1, \dots, N} (N\pi_i) = O(1)$ and $n/N = o(1)$.

$$(C.6) \ \lim_{N \rightarrow \infty} \sum_{i=1}^n E[|\zeta_{Ni}|^4] = 0,$$

$$(C.7) \ \lim_{N \rightarrow \infty} E \left[\left\| \sum_{i=1}^n E[\zeta_{Ni} \zeta_{Ni}^T | \mathcal{F}_{N,i-1}] - B_N \right\|^2 \right] = 0.$$

Then

$$\text{Var}(T)^{-1/2} \Phi (\hat{\beta}_n - \beta_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_p).$$

Here, we establish the unconditional asymptotic properties of subsampling estimator for unbounded GLMs. The condition $n/N = o(1)$ ensures that small-scale subsamples also have expected performance, which greatly release the computational cost. We also present the theoretical asymptotic results, which leads to the subsampling probability using the A-optimal criterion in [10].

4. Conclusions and Future Work

In this paper, we derive the asymptotic normality of the subsampling M-estimator by Fisher information. In the unbounded GLMs with nonnatural link function, we separately obtain the conditional and unconditional asymptotic properties of subsampling estimator.

For future study, it is meaningful to apply the sub-Weibull concentration inequalities in [21] to make nonasymptotic inference. The importance sampling is not ideal, since it tends to assign high sampling probability to the observed samples. Hence, effective subsampling methods are considered for GLMs, such as Markov subsampling in [22]. Moreover, high-dimensional methods in [23,24] for subsampling need further studies.

Author Contributions: Conceptualization, B.T.; Methodology, Y.Z.; Validation, G.T.; Writing—original draft, G.T.; Writing—review & editing, B.T., Y.Z. and S.F.; Supervision, B.T.; Funding acquisition, Y.Z. and B.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Key University Science Research Project of Jiangsu Province 21KJB110023 and National Natural Science Foundation of China 91646106.

Acknowledgments: We would like to thank Huiming Zhang for helpful discussions on large sample theory.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Technical Details

Lemma A1 (Theorem 4.17 in [16]). *Let X_1, \dots, X_N be i.i.d. from a p.d.f. f_β w.r.t. a σ -finite measure ν on $(\mathbb{R}, \mathcal{B}_\mathbb{R})$, where $\beta \in \Theta$ and Θ is an open set in \mathbb{R}^p . Suppose that for every x in the range of X_1 , $f_\beta(x)$ is twice continuously differentiable on β and satisfies*

$$(D.1) \quad \frac{\partial}{\partial \beta} \int \psi_\beta(x) d\nu = \int \frac{\partial}{\partial \beta} \psi_\beta(x) d\nu$$

for $\psi_\beta(x) = f_\beta(x)$ and $\psi_\beta(x) = \frac{\partial f_\beta(x)}{\partial \beta}$.

(D.2) *The Fisher information matrix*

$$I_1(\beta) = E \left\{ \frac{\partial}{\partial \beta} \log f_\beta(X_1) \left[\frac{\partial}{\partial \beta} \log f_\beta(X_1) \right]^T \right\}$$

is positive definite.

(D.3) *For any given $\beta \in \Theta$, there exists a positive number C_β and a positive function h_β such that $E[h_\beta(X_1)] < \infty$ and*

$$\sup_{\gamma: \|\gamma - \beta\| < C_\beta} \left\| \frac{\partial^2 \log f_\gamma(x)}{\partial \gamma \partial \gamma^T} \right\| \leq h_\beta(x)$$

for all x in the range of X_1 , where $\|\cdot\|$ is Euclidean norm and $\|\mathbf{A}\| = \sqrt{\text{tr}(\mathbf{A}^T \mathbf{A})}$ for any matrix \mathbf{A} . Then there exist a sequence of estimators $\hat{\beta}_N$ (based on $\{X_i, i \in U\}$) such that

$$P(s_a(\hat{\beta}_N) = \mathbf{0}) \rightarrow 1 \quad \text{and} \quad \hat{\beta}_N \xrightarrow{P} \beta_0, \tag{A1}$$

where $s_a(\gamma) = \frac{\partial \log \tilde{L}_N(\gamma)}{\partial \gamma}$ and $\tilde{L}_N(\gamma)$ is the likelihood function of full data and β_0 is the real parameter. Meanwhile, there exist a sequence of estimators $\hat{\beta}_n$ (based on $\{X_i, i \in S\}$) such that

$$P(s_s(\hat{\beta}_n) = \mathbf{0}) \rightarrow 1 \quad \text{and} \quad \hat{\beta}_n \xrightarrow{P} \beta_0, \tag{A2}$$

where $s_s(\gamma) = \frac{\partial \log \tilde{L}_n(\gamma)}{\partial \gamma}$ and $\tilde{L}_n(\gamma)$ is the likelihood function of subsampled data and β_0 is the real parameter.

Let a_i be the number of i -th subsampled data such that $\sum_{i \in U} a_i = n$.

Lemma A2. $E[L_n(\beta) | \mathcal{F}_N] = L_N(\beta)$.

Proof. From the definition of a_i , one has

$$\begin{aligned}
 E[L_n(\boldsymbol{\beta})|\mathcal{F}_N] &= E\left[\frac{1}{Nn} \sum_{i \in S} \frac{1}{\pi_i^*} l(\boldsymbol{\beta}; \mathbf{Z}_i^*) \middle| \mathcal{F}_N\right] \\
 &= E\left[\frac{1}{Nn} \sum_{i \in U} \frac{1}{\pi_i} l(\boldsymbol{\beta}; \mathbf{Z}_i) a_i \middle| \mathcal{F}_N\right] \\
 &= \frac{1}{Nn} \sum_{i \in U} a_i E\left[\frac{1}{\pi_i} l(\boldsymbol{\beta}; \mathbf{Z}_i) \middle| \mathcal{F}_N\right] \\
 &= \frac{1}{Nn} \sum_{i \in U} a_i \frac{\sum_{i \in U} \frac{1}{\pi_i} l(\boldsymbol{\beta}; \mathbf{Z}_i) \pi_i}{\sum_{i \in U} \pi_i} \\
 &= \frac{1}{Nn} \sum_{i \in U} a_i \sum_{i \in U} l(\boldsymbol{\beta}; \mathbf{Z}_i) \\
 &= \frac{1}{Nn} n \sum_{i \in U} l(\boldsymbol{\beta}; \mathbf{Z}_i) \\
 &= \frac{1}{N} \sum_{i \in U} l(\boldsymbol{\beta}; \mathbf{Z}_i) \\
 &= L_N(\boldsymbol{\beta}).
 \end{aligned}$$

□

Proposition A1. Under the conditions of Lemma A1 and

$$\min_i(N\pi_i) = \max_i(N\pi_i) = O(1)(i = 1, \dots, N).$$

Assume that $\hat{\boldsymbol{\beta}}_N$ based on $\{\mathbf{X}_i, i \in U\}$ is an estimator of $\boldsymbol{\beta}$, and $\hat{\boldsymbol{\beta}}_n$ based on $\{\mathbf{X}_i, i \in S\}$ is also an estimator of $\boldsymbol{\beta}$, then

$$\hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_N = -\boldsymbol{\Sigma}_N^{-1} \nabla L_n(\hat{\boldsymbol{\beta}}_N) + o_{P|\mathcal{F}_N}(1). \tag{A3}$$

Proof. Taking Taylor series expansion of $\nabla L_n(\hat{\boldsymbol{\beta}}_n)$ around $\hat{\boldsymbol{\beta}}_N$, we have

$$\begin{aligned}
 \mathbf{0} &= \nabla L_n(\hat{\boldsymbol{\beta}}_n) \\
 &= \nabla L_n(\hat{\boldsymbol{\beta}}_N) + \nabla^2 L_n(\hat{\boldsymbol{\beta}}_N)(\hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_N) + o(\|\hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_N\|) \\
 &= \nabla L_n(\hat{\boldsymbol{\beta}}_N) + \nabla^2 L_n(\hat{\boldsymbol{\beta}}_N)(\hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_N) + \nabla^2 L_N(\hat{\boldsymbol{\beta}}_N)(\hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_N) \\
 &\quad - \nabla^2 L_N(\hat{\boldsymbol{\beta}}_N)(\hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_N) + o(\|\hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_N\|) \\
 &= \nabla L_n(\hat{\boldsymbol{\beta}}_N) + \nabla^2 L_N(\hat{\boldsymbol{\beta}}_N)(\hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_N) + (\nabla^2 L_n(\hat{\boldsymbol{\beta}}_N) - \nabla^2 L_N(\hat{\boldsymbol{\beta}}_N))(\hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_N) \\
 &\quad + o(\|\hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_N\|).
 \end{aligned} \tag{A4}$$

From the definition of a_i , one has

$$\begin{aligned}
 (\nabla^2 L_n(\hat{\beta}_N) - \nabla^2 L_N(\hat{\beta}_N))(\hat{\beta}_n - \hat{\beta}_N) &= \left(\frac{1}{Nn} \sum_{i \in S} \frac{1}{\pi_i^*} \nabla^2 l(\hat{\beta}_N; \mathbf{Z}_i^*) - \frac{1}{N} \sum_{i \in U} \nabla^2 l(\hat{\beta}_N; \mathbf{Z}_i) \right) \\
 &\quad \cdot (\hat{\beta}_n - \hat{\beta}_N) \\
 &= \left(\sum_{i \in U} \frac{1}{Nn\pi_i} \nabla^2 l(\hat{\beta}_N; \mathbf{Z}_i) a_i - \frac{1}{N} \sum_{i \in U} \nabla^2 l(\hat{\beta}_N; \mathbf{Z}_i) \right) \\
 &\quad \cdot (\hat{\beta}_n - \hat{\beta}_N) \\
 &= \left(\sum_{i \in U} \frac{a_i - n\pi_i}{Nn\pi_i} \nabla^2 l(\hat{\beta}_N; \mathbf{Z}_i) \right) (\hat{\beta}_n - \hat{\beta}_N) \\
 &\leq \left(\sum_{i \in U} \frac{a_i}{Nn\pi_i} \nabla^2 l(\hat{\beta}_N; \mathbf{Z}_i) \right) (\hat{\beta}_n - \hat{\beta}_N) \\
 &= o_{P|\mathcal{F}_N}(1).
 \end{aligned}
 \tag{A5}$$

Combine Equations (A1), (A2) and (A5) into Equation (A4), one has

$$\mathbf{0} = \nabla L_n(\hat{\beta}_N) + \nabla^2 L_N(\hat{\beta}_N)(\hat{\beta}_n - \hat{\beta}_N) + o_{P|\mathcal{F}_N}(1).$$

This can be transformed to

$$\hat{\beta}_n - \hat{\beta}_N = -\Sigma_N^{-1} \nabla L_n(\hat{\beta}_N) + o_{P|\mathcal{F}_N}(1).
 \tag{A6}$$

The proposition is proved. \square

Remark A1. The last equation in the proof ensures that $\hat{\beta}_n - \hat{\beta}_N + \Sigma_N^{-1} \nabla L_n(\hat{\beta}_N)$ is a higher-order infinitesimal with respect to 1, which is true according to conditional probability with \mathcal{F}_N . $o_{P|\mathcal{F}_N}(1)$ in Equation (A6) is denoted as the higher order infinitesimal of 1 according to conditional probability with \mathcal{F}_N .

Proof of Theorem 1. For every constant $\hat{\gamma} > 0$, one has

$$\begin{aligned}
 \sum_{j \in S} \mathbb{E} \left\{ \left\| n^{-\frac{1}{2}} \zeta_j^* \right\|^2 I(\|\zeta_j^*\| > n^{\frac{1}{2}} \hat{\gamma}) \middle| \mathcal{F}_N \right\} &\leq \sum_{j \in S} \mathbb{E} \left\{ \frac{\left\| n^{-\frac{1}{2}} \zeta_j^* \right\|^4}{n \hat{\gamma}^2} I(\|\zeta_j^*\| > n^{\frac{1}{2}} \hat{\gamma}) \middle| \mathcal{F}_N \right\} \\
 &= \frac{1}{n^2 \hat{\gamma}^2} \sum_{j \in S} \mathbb{E} \left\{ \left\| \zeta_j^* \right\|^4 I(\|\zeta_j^*\| > n^{\frac{1}{2}} \hat{\gamma}) \middle| \mathcal{F}_N \right\} \\
 &\leq \frac{1}{n^2 \hat{\gamma}^2} \sum_{j \in S} \mathbb{E} \left\{ \left\| \zeta_j^* \right\|^4 \middle| \mathcal{F}_N \right\} \\
 &= \frac{1}{n^2 \hat{\gamma}^2} \sum_{i \in U} a_i \mathbb{E} \left\{ \left\| \zeta_i \right\|^4 \middle| \mathcal{F}_N \right\} \\
 &= \frac{1}{n^2 \hat{\gamma}^2} n \frac{\sum_{i \in U} \left\| \zeta_i \right\|^4 \pi_i}{\sum_{i \in U} \pi_i} \\
 &= \frac{1}{n \hat{\gamma}^2} \sum_{i \in U} \left\| \frac{1}{N \pi_i} \Sigma_N^{-1} \nabla l(\hat{\beta}_N; \mathbf{Z}_i) \right\|^4 \pi_i \\
 &= \frac{1}{n \hat{\gamma}^2} \frac{1}{N^4} \sum_{i \in U} \frac{1}{\pi_i^3} \left\| \Sigma_N^{-1} \nabla l(\hat{\beta}_N; \mathbf{Z}_i) \right\|^4 \\
 &\leq \frac{1}{n \hat{\gamma}^2} \frac{1}{N^4} \sum_{i \in U} \frac{1}{\pi_i^3} \frac{1}{\lambda^4} \left\| \nabla l(\hat{\beta}_N; \mathbf{Z}_i) \right\|^4 \\
 &= \frac{1}{n \hat{\gamma}^2} \frac{1}{\lambda^4} O(1) \\
 &= o(1).
 \end{aligned}$$

Furthermore,

$$\begin{aligned}
 \sum_{j \in S} \text{Cov}(n^{-\frac{1}{2}} \zeta_j^* | \mathcal{F}_N) &= \sum_{j \in S} \mathbb{E} \left\{ \left[n^{-\frac{1}{2}} \zeta_j^* - \mathbb{E}(n^{-\frac{1}{2}} \zeta_j^* | \mathcal{F}_N) \right] \left[n^{-\frac{1}{2}} \zeta_j^* - \mathbb{E}(n^{-\frac{1}{2}} \zeta_j^* | \mathcal{F}_N) \right]^T \middle| \mathcal{F}_N \right\} \\
 &= \sum_{j \in S} \mathbb{E} \left[(n^{-\frac{1}{2}} \zeta_j^*) (n^{-\frac{1}{2}} \zeta_j^*)^T \middle| \mathcal{F}_N \right] \\
 &= \frac{1}{n} \sum_{j \in S} \mathbb{E}(\zeta_j^* \zeta_j^{*T} | \mathcal{F}_N) \\
 &= \frac{1}{n} n \mathbb{E}(\zeta \zeta^T | \mathcal{F}_N) \\
 &= \text{Var}(\zeta | \mathcal{F}_N).
 \end{aligned}$$

Then, by the Lindeberg-Feller central limit theorem (Proposition 2.27 of [11]), conditional on \mathcal{F}_N ,

$$\sum_{j \in S} n^{-\frac{1}{2}} \zeta_j^* \xrightarrow{d} N(0, \text{Var}(\zeta | \mathcal{F}_N)).$$

Therefore, combining the above and Proposition A1, Equation (5) holds. Thus, the proof is completed. \square

Proof of Theorem 2. Next, one needs to show convexity (i.e., uniqueness and maximum value) due to the existence of the estimators from [25]. Let

$$\begin{aligned} I_1(\boldsymbol{\beta}) &= E(\nabla s_n(\boldsymbol{\beta})) \\ &= E\left(\frac{\partial\left(\sum_{i \in S} \frac{1}{\pi_i^*} [Y_i^* - \mu(\psi(\boldsymbol{\beta}^T \mathbf{X}_i^*))]\psi(\boldsymbol{\beta}^T \mathbf{X}_i^*)\mathbf{X}_i^{*T}\right)}{\partial \boldsymbol{\beta}}\right) \\ &= E\left(-\sum_{i \in S} \frac{1}{\pi_i^*} \dot{\mu}(\psi(\boldsymbol{\beta}^T \mathbf{X}_i^*))[\psi(\boldsymbol{\beta}^T \mathbf{X}_i^*)]^2 \mathbf{X}_i^* \mathbf{X}_i^{*T} \right. \\ &\quad \left. + \sum_{i \in S} \frac{1}{\pi_i^*} [Y_i^* - \mu(\psi(\boldsymbol{\beta}^T \mathbf{X}_i^*))]\ddot{\psi}(\boldsymbol{\beta}^T \mathbf{X}_i^*)\mathbf{X}_i^* \mathbf{X}_i^{*T}\right) \\ &= -\sum_{i \in S} \frac{1}{\pi_i^*} \dot{\mu}(\psi(\boldsymbol{\beta}^T \mathbf{X}_i^*))[\psi(\boldsymbol{\beta}^T \mathbf{X}_i^*)]^2 \mathbf{X}_i^* \mathbf{X}_i^{*T}, \end{aligned}$$

where

$$s_n(\boldsymbol{\beta}) = \sum_{i \in S} \frac{1}{\pi_i^*} [Y_i^* - \mu(\psi(\boldsymbol{\beta}^T \mathbf{X}_i^*))]\psi(\boldsymbol{\beta}^T \mathbf{X}_i^*)\mathbf{X}_i^{*T}.$$

From [16] in Theorem 4.17, one needs to show

$$\max_{\gamma \in G(C_0)} \left\| \left[I_1(\hat{\boldsymbol{\beta}}_N) \right]^{-1/2} \nabla s_n(\gamma) \left[I_1(\hat{\boldsymbol{\beta}}_N) \right]^{-1/2} + \mathbf{I}_p \right\| \xrightarrow{P|\mathcal{F}_N} 0,$$

where $G(C_0) = \left\{ \gamma : \left\| \left[I_1(\hat{\boldsymbol{\beta}}_N) \right]^{1/2} (\gamma - \hat{\boldsymbol{\beta}}_N) \right\| \leq C_0 \right\}$ and $\mathbf{I}_p = \text{diag}(1, 1, \dots, 1)$ is a p -dimensional identity matrix.

Let

$$M_n(\gamma) = \sum_{i \in S} \frac{1}{\pi_i^*} [\psi(\gamma^T \mathbf{X}_i^*)]^2 \dot{b}(\psi(\gamma^T \mathbf{X}_i^*)) \mathbf{X}_i^* \mathbf{X}_i^{*T} \tag{A7}$$

and

$$R_n(\gamma) = \sum_{i \in S} \frac{1}{\pi_i^*} [Y_i^* - \mu(\psi(\gamma^T \mathbf{X}_i^*))]\ddot{\psi}(\gamma^T \mathbf{X}_i^*)\mathbf{X}_i^* \mathbf{X}_i^{*T}. \tag{A8}$$

Then

$$\nabla s_n(\gamma) = R_n(\gamma) - M_n(\gamma) \tag{A9}$$

and

$$I_1(\gamma) = -E(\nabla s_n(\gamma)) = M_n(\gamma) \tag{A10}$$

Thus, one only needs to prove

$$\max_{\gamma \in G(C_0)} \left\| \left[M_n(\hat{\boldsymbol{\beta}}_N) \right]^{-1/2} \left[M_n(\gamma) - M_n(\hat{\boldsymbol{\beta}}_N) \right] \left[M_n(\hat{\boldsymbol{\beta}}_N) \right]^{-1/2} \right\| \xrightarrow{P|\mathcal{F}_N} 0, \tag{A11}$$

and

$$\max_{\gamma \in G(C_0)} \left\| \left[M_n(\hat{\boldsymbol{\beta}}_N) \right]^{-1/2} R_n(\gamma) \left[M_n(\hat{\boldsymbol{\beta}}_N) \right]^{-1/2} \right\| \xrightarrow{P|\mathcal{F}_N} 0 \tag{A12}$$

for any $C_0 > 0$. From the definition of $M_n(\gamma)$, and the property of trace in P288 of [16], the left-hand side of Equation (A11) can be bounded by

$$\sqrt{p} \max_{\gamma \in G(C_0), i \in S} \left| 1 - \varphi(\gamma^T \mathbf{X}_i^*) / \varphi(\hat{\boldsymbol{\beta}}_N^T \mathbf{X}_i^*) \right|.$$

From condition (A.4), one needs to prove $|\gamma^T \mathbf{X}_i^* - \hat{\beta}_N^T \mathbf{X}_i^*|$ converges to 0 so that Equation (A11) holds, and one has

$$\begin{aligned} |\gamma^T \mathbf{X}_i^* - \hat{\beta}_N^T \mathbf{X}_i^*|^2 &= \left| (\gamma^T - \hat{\beta}_N^T) [\mathbf{I}_1(\hat{\beta}_N)]^{1/2} [\mathbf{I}_1(\hat{\beta}_N)]^{-1/2} \mathbf{X}_i^* \right|^2 \\ &\leq \left\| [\mathbf{I}_1(\hat{\beta}_N)]^{1/2} (\gamma - \hat{\beta}_N) \right\|^2 \left\| [\mathbf{I}_1(\hat{\beta}_N)]^{-1/2} \mathbf{X}_i^* \right\|^2 \\ &\leq C_0^2 \max_{i \in S} \mathbf{X}_i^{*T} [\mathbf{I}_1(\hat{\beta}_N)]^{-1} \mathbf{X}_i^* \\ &= C_0^2 \max_{i \in S} \mathbf{X}_i^{*T} [M_n(\hat{\beta}_N)]^{-1} \mathbf{X}_i^* \\ &= C_0^2 \max_{i \in S} \mathbf{X}_i^{*T} \left[\sum_{i \in S} \frac{1}{\pi_i^*} [\psi(\hat{\beta}_N^T \mathbf{X}_i^*)]^2 \ddot{b}(\psi(\hat{\beta}_N^T \mathbf{X}_i^*)) \mathbf{X}_i^* \mathbf{X}_i^{*T} \right]^{-1} \mathbf{X}_i^* \\ &= C_0^2 \max_{i \in S} \mathbf{X}_i^{*T} \left[\sum_{i \in S} \frac{1}{\pi_i^*} \varphi(\hat{\beta}_N^T \mathbf{X}_i^*) \mathbf{X}_i^* \mathbf{X}_i^{*T} \right]^{-1} \mathbf{X}_i^* \\ &= C_0^2 \max_{i \in S} \mathbf{X}_i^{*T} \left[\sum_{i \in S} N \frac{1}{N \pi_i^*} \varphi(\hat{\beta}_N^T \mathbf{X}_i^*) \mathbf{X}_i^* \mathbf{X}_i^{*T} \right]^{-1} \mathbf{X}_i^* \\ &\leq C_0^2 \left[N \min_{i \in S} \frac{1}{N \pi_i^*} \inf_{i \in S} \varphi(\hat{\beta}_N^T \mathbf{X}_i^*) \right]^{-1} \max_{i \in S} \mathbf{X}_i^{*T} \left(\sum_{i \in S} \mathbf{X}_i^* \mathbf{X}_i^{*T} \right)^{-1} \mathbf{X}_i^* \\ &= C_0^2 \left[N \min_{i \in S} \frac{1}{N \pi_i^*} \inf_{i \in S} \varphi(\hat{\beta}_N^T \mathbf{X}_i^*) \right]^{-1} \max_{i \in S} \mathbf{X}_i^{*T} (\mathbf{X}^* \mathbf{X}^{*T})^{-1} \mathbf{X}_i^* \\ &\xrightarrow{P|\mathcal{F}_N} 0. \end{aligned}$$

Hence Equation (A11) holds. Let $e_i^* = Y_i^* - \mu(\psi(\hat{\beta}_N^T \mathbf{X}_i^*))$, and

$$\begin{aligned} U_n(\gamma) &= \sum_{i \in S} \frac{1}{\pi_i^*} \left[\mu(\psi(\hat{\beta}_N^T \mathbf{X}_i^*)) - \mu(\psi(\gamma^T \mathbf{X}_i^*)) \right] \ddot{\psi}(\gamma^T \mathbf{X}_i^*) \mathbf{X}_i^* \mathbf{X}_i^{*T}, \\ V_n(\gamma) &= \sum_{i \in S} \frac{e_i^*}{\pi_i^*} \left[\ddot{\psi}(\gamma^T \mathbf{X}_i^*) - \ddot{\psi}(\hat{\beta}_N^T \mathbf{X}_i^*) \right] \mathbf{X}_i^* \mathbf{X}_i^{*T}, \\ W_n(\hat{\beta}_N) &= \sum_{i \in S} \frac{e_i^*}{\pi_i^*} \ddot{\psi}(\hat{\beta}_N^T \mathbf{X}_i^*) \mathbf{X}_i^* \mathbf{X}_i^{*T}. \end{aligned}$$

Then $R_n(\gamma) = U_n(\gamma) + V_n(\gamma) + W_n(\hat{\beta}_N)$. In the same way as proving Equation (A11), we have

$$\max_{\gamma \in G(C_0)} \left\| [M_n(\hat{\beta}_N)]^{-1/2} U_n(\gamma) [M_n(\hat{\beta}_N)]^{-1/2} \right\| \xrightarrow{P|\mathcal{F}_N} 0.$$

Note that $\left\| [M_n(\hat{\beta}_N)]^{-1/2} V_n(\gamma) [M_n(\hat{\beta}_N)]^{-1/2} \right\|$ is bounded by the product of

$$\left\| [M_n(\hat{\beta}_N)]^{-\frac{1}{2}} \sum_{i \in S} \frac{e_i^*}{\pi_i^*} \mathbf{X}_i^* \mathbf{X}_i^{*T} [M_n(\hat{\beta}_N)]^{-\frac{1}{2}} \right\| \tag{A13}$$

and

$$\max_{\gamma \in G(C_0), i \in S} \left| \ddot{\psi}(\gamma^T \mathbf{X}_i^*) - \ddot{\psi}(\hat{\beta}_N^T \mathbf{X}_i^*) \right|. \tag{A14}$$

Equation (A13) can be bounded as

$$\begin{aligned}
 & \left\| \left[M_n(\hat{\beta}_N) \right]^{-\frac{1}{2}} \sum_{i \in S} \frac{e_i^*}{\pi_i^*} \mathbf{X}_i^* \mathbf{X}_i^{*T} \left[M_n(\hat{\beta}_N) \right]^{-\frac{1}{2}} \right\| \\
 &= \left\| \left[\mathbf{I}_1(\hat{\beta}_N) \right]^{-\frac{1}{2}} \sum_{i \in S} \frac{e_i^*}{\pi_i^*} \mathbf{X}_i^* \mathbf{X}_i^{*T} \left[\mathbf{I}_1(\hat{\beta}_N) \right]^{-\frac{1}{2}} \right\| \\
 &\leq \left\| \sum_{i \in S} \frac{e_i^*}{\pi_i^*} \left\| \left[\mathbf{I}_1(\hat{\beta}_N) \right]^{-\frac{1}{2}} \mathbf{X}_i^* \right\|^2 \right\| \\
 &\leq \left\| \sum_{i \in S} \frac{e_i^*}{\pi_i^*} \left[N \min_{i \in S} \frac{1}{N\pi_i^*} \inf_{i \in S} \varphi(\hat{\beta}_N^T \mathbf{X}_i^*) \right]^{-1} \max_{i \in S} \mathbf{X}_i^{*T} (\mathbf{X}^* \mathbf{X}^{*T})^{-1} \mathbf{X}_i^* \right\| \\
 &\leq \left\| \sum_{i \in S} e_i^* \left| \max_{i \in S} \frac{1}{N\pi_i^*} \left[\min_{i \in S} \frac{1}{N\pi_i^*} \inf_{i \in S} \varphi(\hat{\beta}_N^T \mathbf{X}_i^*) \right]^{-1} \max_{i \in S} \mathbf{X}_i^{*T} (\mathbf{X}^* \mathbf{X}^{*T})^{-1} \mathbf{X}_i^* \right| \right\| \\
 &\leq \sum_{i \in U} |Y_i - \mu(\psi(\hat{\beta}_N^T \mathbf{X}_i))| \left| \max_{i \in S} \frac{1}{N\pi_i^*} \left[\min_{i \in S} \frac{1}{N\pi_i^*} \inf_{i \in S} \varphi(\hat{\beta}_N^T \mathbf{X}_i^*) \right]^{-1} \max_{i \in S} \mathbf{X}_i^{*T} (\mathbf{X}^* \mathbf{X}^{*T})^{-1} \mathbf{X}_i^* \right| \\
 &\leq \left[\sum_{i \in U} \sup_{\beta \in \Theta} |Y_i - \mu(\psi(\beta^T \mathbf{X}_i))| \right] \left| \max_{i \in S} \frac{1}{N\pi_i^*} \left[\min_{i \in S} \frac{1}{N\pi_i^*} \inf_{i \in S} \varphi(\hat{\beta}_N^T \mathbf{X}_i^*) \right]^{-1} \right. \\
 &\quad \cdot \max_{i \in S} \mathbf{X}_i^{*T} (\mathbf{X}^* \mathbf{X}^{*T})^{-1} \mathbf{X}_i^* \left. \right| \\
 &= \frac{1}{n} \sum_{i \in S} \mathbb{E} \sup_{\beta \in \Theta} \left[\frac{1}{\pi_i^*} |Y_i^* - \mu(\psi(\beta^T \mathbf{X}_i^*))| \mid \mathcal{F}_N \right] \left| \max_{i \in S} \frac{1}{N\pi_i^*} \left[\min_{i \in S} \frac{1}{N\pi_i^*} \inf_{i \in S} \varphi(\hat{\beta}_N^T \mathbf{X}_i^*) \right]^{-1} \right. \\
 &\quad \cdot \max_{i \in S} \mathbf{X}_i^{*T} (\mathbf{X}^* \mathbf{X}^{*T})^{-1} \mathbf{X}_i^* \left. \right| \\
 &= O_{P|\mathcal{F}_N}(1/n),
 \end{aligned}$$

where the penultimate equal sign applies the Lemma A2 with

$$l(\beta) = \sup_{\beta \in \Theta} |Y_i - \mu(\psi(\beta^T \mathbf{X}_i))|.$$

Equation (A14) can be bounded as

$$\max_{\gamma \in G(C_0), i \in S} \left| \ddot{\psi}(\gamma^T \mathbf{X}_i^*) - \ddot{\psi}(\hat{\beta}_N^T \mathbf{X}_i^*) \right| \xrightarrow{P|\mathcal{F}_N} 0,$$

which can be proved as the same argument of Equation (A11) by Lagrange mean value theorem. Combine the bounds of Equations (A13) and (A14), one obtains

$$\max_{\gamma \in G(C_0)} \left\| \left[M_n(\hat{\beta}_N) \right]^{-1/2} V_n(\gamma) \left[M_n(\hat{\beta}_N) \right]^{-1/2} \right\| \xrightarrow{P|\mathcal{F}_N} 0.$$

Let $\delta \in (0, 1)$ be a constant. Since $\sup_{i \in S} \mathbb{E}(|e_i^*|^{1+\delta} | \mathcal{F}_N) < \infty$, one has

$$\begin{aligned}
 & \sum_{i \in S} \mathbb{E} \left(\left| \frac{e_i^*}{\pi_i^*} \dot{\psi}(\hat{\beta}_N^T \mathbf{X}_i^*) \mathbf{X}_i^{*T} [M_n(\hat{\beta}_N)]^{-1} \mathbf{X}_i^* \right|^{1+\delta} \middle| \mathcal{F}_N \right) \\
 & \leq \sum_{i \in S} \mathbb{E} \left(\left| \frac{e_i^*}{\pi_i^*} \right|^{1+\delta} \middle| \mathcal{F}_N \right) \cdot \max_{i \in S} \left| \dot{\psi}(\hat{\beta}_N^T \mathbf{X}_i^*) \right|^{1+\delta} \cdot \left| \mathbf{X}_i^{*T} [M_n(\hat{\beta}_N)]^{-1} \mathbf{X}_i^* \right|^{1+\delta} \\
 & \leq \sum_{i \in S} \left(\frac{1}{\pi_i^*} \right)^{1+\delta} \mathbb{E} \left(|e_i^*|^{1+\delta} \middle| \mathcal{F}_N \right) \max_{i \in S} \left| \dot{\psi}(\hat{\beta}_N^T \mathbf{X}_i^*) \right|^{1+\delta} \\
 & \quad \cdot \left| \mathbf{X}_i^{*T} \left[\sum_{i \in S} N \frac{1}{N \pi_i^*} \varphi(\hat{\beta}_N^T \mathbf{X}_i^*) \mathbf{X}_i^* \mathbf{X}_i^{*T} \right]^{-1} \mathbf{X}_i^* \right|^{1+\delta} \\
 & = \sum_{i \in S} \left(\frac{1}{N \pi_i^*} \right)^{1+\delta} \mathbb{E} \left(|e_i^*|^{1+\delta} \middle| \mathcal{F}_N \right) \max_{i \in S} \left| \dot{\psi}(\hat{\beta}_N^T \mathbf{X}_i^*) \right|^{1+\delta} \\
 & \quad \cdot \left| \mathbf{X}_i^{*T} \left[\sum_{i \in S} \frac{1}{N \pi_i^*} \varphi(\hat{\beta}_N^T \mathbf{X}_i^*) \mathbf{X}_i^* \mathbf{X}_i^{*T} \right]^{-1} \mathbf{X}_i^* \right|^{1+\delta} \\
 & \leq C_\delta \sum_{i \in S} \left| \mathbf{X}_i^{*T} (\mathbf{X}^* \mathbf{X}^{*T})^{-1} \mathbf{X}_i^* \right|^{1+\delta} \\
 & \leq C_\delta \sum_{i \in S} \left| \mathbf{X}_i^{*T} (\mathbf{X}^* \mathbf{X}^{*T})^{-1} \mathbf{X}_i^* \right| \max_{i \in S} \left| \mathbf{X}_i^{*T} (\mathbf{X}^* \mathbf{X}^{*T})^{-1} \mathbf{X}_i^* \right|^\delta \\
 & = C_\delta \max_{i \in S} \left| \mathbf{X}_i^{*T} (\mathbf{X}^* \mathbf{X}^{*T})^{-1} \mathbf{X}_i^* \right|^\delta \sum_{i \in S} \text{tr} \left[\mathbf{X}_i^{*T} (\mathbf{X}^* \mathbf{X}^{*T})^{-1} \mathbf{X}_i^* \right] \\
 & = C_\delta \max_{i \in S} \left| \mathbf{X}_i^{*T} (\mathbf{X}^* \mathbf{X}^{*T})^{-1} \mathbf{X}_i^* \right|^\delta \sum_{i \in S} \text{tr} \left[(\mathbf{X}^* \mathbf{X}^{*T})^{-1} \mathbf{X}_i^* \mathbf{X}_i^{*T} \right] \\
 & = C_\delta \max_{i \in S} \left| \mathbf{X}_i^{*T} (\mathbf{X}^* \mathbf{X}^{*T})^{-1} \mathbf{X}_i^* \right|^\delta \text{tr} \left[(\mathbf{X}^* \mathbf{X}^{*T})^{-1} \sum_{i \in S} \mathbf{X}_i^* \mathbf{X}_i^{*T} \right] \\
 & = C_\delta \max_{i \in S} \left| \mathbf{X}_i^{*T} (\mathbf{X}^* \mathbf{X}^{*T})^{-1} \mathbf{X}_i^* \right|^\delta \text{tr} \left[(\mathbf{X}^* \mathbf{X}^{*T})^{-1} (\mathbf{X}^* \mathbf{X}^{*T}) \right] \\
 & = C_\delta \max_{i \in S} \left| \mathbf{X}_i^{*T} (\mathbf{X}^* \mathbf{X}^{*T})^{-1} \mathbf{X}_i^* \right|^\delta \text{tr} \mathbf{I}_p \\
 & = p C_\delta \max_{i \in S} \left| \mathbf{X}_i^{*T} (\mathbf{X}^* \mathbf{X}^{*T})^{-1} \mathbf{X}_i^* \right|^\delta \\
 & \xrightarrow{P|\mathcal{F}_N} 0,
 \end{aligned}$$

where $C_\delta > 0$ is a constant. Under the definition of $W_n(\hat{\beta}_N)$ and $\mathbb{E}(e_i^* | \mathcal{F}_N) = 0$, together with Theorem 1.14(ii) in [16], one obtains

$$\left\| [M_n(\hat{\beta}_N)]^{-\frac{1}{2}} W_n(\hat{\beta}_N) [M_n(\hat{\beta}_N)]^{-\frac{1}{2}} \right\| \xrightarrow{P|\mathcal{F}_N} 0.$$

Hence, Equation (A12) holds and the proof is completed. \square

Proof of Theorem 3. According to the mean value theorem, one has

$$\mathbf{0} = s_n(\hat{\beta}_n) = s_n(\hat{\beta}_N) + \nabla s_n(\bar{\beta})(\hat{\beta}_n - \hat{\beta}_N),$$

where $\bar{\beta}$ is between $\hat{\beta}_n$ and $\hat{\beta}_N$, then

$$\sqrt{n}(\hat{\beta}_n - \hat{\beta}_N) = -\sqrt{n}\nabla s_n(\bar{\beta})^{-1}s_n(\hat{\beta}_N). \tag{A15}$$

Let $q_i^*(\hat{\beta}_N) = \frac{1}{\pi_i^*} [Y_i^* - \mu(\psi(\hat{\beta}_N^T X_i^*))] \psi(\hat{\beta}_N^T X_i^*) X_i^*$, then

$$\sum_{i \in S} q_i^*(\hat{\beta}_N) = \sum_{i \in S} \frac{1}{\pi_i^*} [Y_i^* - \mu(\psi(\hat{\beta}_N^T X_i^*))] \psi(\hat{\beta}_N^T X_i^*) X_i^* = s_n(\hat{\beta}_N).$$

According to $E(Y_i^* | \mathcal{F}_N) = \mu(\psi(\hat{\beta}_N^T X_i^*))$ in Equation (4), one obtains

$$E(q_i^*(\hat{\beta}_N) | \mathcal{F}_N) = \frac{1}{\pi_i^*} [E(Y_i^* | \mathcal{F}_N) - \mu(\psi(\hat{\beta}_N^T X_i^*))] \psi(\hat{\beta}_N^T X_i^*) X_i^* = \mathbf{0}.$$

Applying Lindeberg-Lévy CLT, one has

$$\frac{s_n(\hat{\beta}_N)}{\sqrt{n}} \xrightarrow{d} N(0, \text{Var}(q_i^*(\hat{\beta}_N) | \mathcal{F}_N)), \tag{A16}$$

where

$$\begin{aligned} \text{Var}(q_i^*(\hat{\beta}_N) | \mathcal{F}_N) &= E(q_i^*(\hat{\beta}_N) q_i^*(\hat{\beta}_N)^T | \mathcal{F}_N) \\ &= \sum_{i \in U} \frac{a_i}{\pi_i} [Y_i - b(\psi(\hat{\beta}_N^T X_i))]^2 \psi(\hat{\beta}_N^T X_i)^2 X_i X_i^T. \end{aligned}$$

Applying [26] in Theorem 2, one has

$$\frac{\nabla s_n(\bar{\beta})}{n} = \frac{1}{n} \sum_{i \in S} \frac{\partial q_i^*(\bar{\beta})}{\partial \bar{\beta}} \xrightarrow{a.s.} E\left(\frac{\partial q_i^*(\bar{\beta})}{\partial \bar{\beta}} \middle| \mathcal{F}_N\right),$$

where

$$\begin{aligned} E\left(\frac{\partial q_i^*(\bar{\beta})}{\partial \bar{\beta}} \middle| \mathcal{F}_N\right) &= \sum_{i \in U} a_i [Y_i - b(\psi(\bar{\beta}^T X_i))] \dot{\psi}(\bar{\beta}^T X_i) X_i X_i^T \\ &\quad - \sum_{i \in U} a_i \dot{b}(\psi(\bar{\beta}^T X_i)) \psi(\bar{\beta}^T X_i)^2 X_i X_i^T. \end{aligned}$$

Since $\bar{\beta}$ is between $\tilde{\beta}_n$ and $\hat{\beta}_N$, and $\tilde{\beta}_n$ is consistent with $\hat{\beta}_N$ with respect to \mathcal{F}_N in probability, then

$$\frac{\nabla s_n(\bar{\beta})}{n} \xrightarrow{P|\mathcal{F}_N} E\left(\frac{\partial q_i^*(\hat{\beta}_N)}{\partial \hat{\beta}_N} \middle| \mathcal{F}_N\right), \tag{A17}$$

where

$$\begin{aligned} E\left(\frac{\partial q_i^*(\hat{\beta}_N)}{\partial \hat{\beta}_N} \middle| \mathcal{F}_N\right) &= \sum_{i \in U} a_i [Y_i - b(\psi(\hat{\beta}_N^T X_i))] \dot{\psi}(\hat{\beta}_N^T X_i) X_i X_i^T \\ &\quad - \sum_{i \in U} a_i \dot{b}(\psi(\hat{\beta}_N^T X_i)) \psi(\hat{\beta}_N^T X_i)^2 X_i X_i^T. \end{aligned}$$

At last, combining Equations (A15)–(A17) by Slutsky’s theorem, one obtains

$$\sqrt{n}(\hat{\beta}_n - \hat{\beta}_N) \xrightarrow{d} N(0, V_s),$$

where $V_s = \left[E\left(\frac{\partial q_i^*(\hat{\beta}_N)}{\partial \hat{\beta}_N} \middle| \mathcal{F}_N\right) \right]^{-1} \text{Var}(q_i^*(\hat{\beta}_N) | \mathcal{F}_N) \left[E\left(\frac{\partial q_i^*(\hat{\beta}_N)}{\partial \hat{\beta}_N} \middle| \mathcal{F}_N\right) \right]^{-1} = \Sigma_N^{-1} V_N \Sigma_N^{-1}$. The proof is completed. \square

Proof of Theorem 4. Here, one needs to prove the consistency of $\hat{\beta}_n$ with respect to β_0 due to the existence of $\hat{\beta}_n$; see [27].

Denote $p_\beta(\mathbf{X}, y) := \exp\{y\psi(\beta^T \mathbf{X}) - b(\psi(\beta^T \mathbf{X}))\}$, $m_\beta(\mathbf{X}, y) = \log p_\beta(\mathbf{X}, y) := y\psi(\beta^T \mathbf{X}) - b(\psi(\beta^T \mathbf{X}))$ and $\tilde{\varphi}(\beta^T \mathbf{X}) = \dot{b}[\psi(\beta^T \mathbf{X})]\psi(\beta^T \mathbf{X})$. Then the negative K-L divergence in [28] is bounded,

$$\begin{aligned} -D_{KL}(P_{\beta_0} || P_\beta) &:= E_{\beta_0}(m_\beta - m_{\beta_0}) \\ &= E\{(E_{\beta_0} y | \mathbf{X})[\psi(\beta^T \mathbf{X}) - \psi(\beta_0^T \mathbf{X})] - b(\psi(\beta^T \mathbf{X})) + b(\psi(\beta_0^T \mathbf{X}))\} \\ &= E\{\dot{b}[\psi(\beta_0^T \mathbf{X})][\psi(\beta^T \mathbf{X}) - \psi(\beta_0^T \mathbf{X})] - b(\psi(\beta^T \mathbf{X})) + b(\psi(\beta_0^T \mathbf{X}))\} \\ (\exists t_1 \in [0, 1]) &= E\{\dot{b}[\psi(\beta_0^T \mathbf{X})][\psi(\beta^T \mathbf{X}) - \psi(\beta_0^T \mathbf{X})] \\ &\quad - \dot{b}[(1 - t_1)\psi(\beta^T \mathbf{X}) + t_1\psi(\beta_0^T \mathbf{X})][\psi(\beta^T \mathbf{X}) - \psi(\beta_0^T \mathbf{X})]\} \\ (\exists t_2 \in [0, 1]) &= E\{\dot{b}[(1 - t_2)\psi(\beta_0^T \mathbf{X}) + (1 - t_1)t_2\psi(\beta^T \mathbf{X}) + t_1t_2\psi(\beta_0^T \mathbf{X})] \\ &\quad \cdot [\psi(\beta_0^T \mathbf{X}) - (1 - t_1)\psi(\beta^T \mathbf{X}) - t_1\psi(\beta_0^T \mathbf{X})][\psi(\beta^T \mathbf{X}) - \psi(\beta_0^T \mathbf{X})]\} \\ &= - (1 - t_1) E\{\dot{b}[(1 - t_3)\psi(\beta_0^T \mathbf{X}) + t_3\psi(\beta^T \mathbf{X})][\psi(\beta^T \mathbf{X}) - \psi(\beta_0^T \mathbf{X})]^2\} \\ (\exists t_4 \in [0, 1]) &= - (1 - t_1) E\{\dot{b}[(1 - t_3)\psi(\beta_0^T \mathbf{X}) + t_3\psi(\beta^T \mathbf{X})] \\ &\quad \cdot \psi[(1 - t_4)(\beta_0^T \mathbf{X}) + t_4(\beta^T \mathbf{X})]^2 (\beta^T \mathbf{X} - \beta_0^T \mathbf{X})^2\} \\ [\text{By (B.4) and (B.5)}] &\leq - (1 - t_1) C_1 E(\beta^T \mathbf{X} - \beta_0^T \mathbf{X})^2 \\ &= - (1 - t_1) C_1 (\beta - \beta_0)^T (E\mathbf{X}\mathbf{X}^T) (\beta - \beta_0) \\ &\leq - (1 - t_1) C_1 \lambda_{\min}(E\mathbf{X}\mathbf{X}^T) \|\beta - \beta_0\|^2 \\ [\text{By (B.1)}] &\leq - (1 - t_1) C_1 C_2 \|\beta - \beta_0\|^2, \end{aligned}$$

where $t_3 = t_2 - t_1 t_2 \in [0, 1]$ and $C_2 > 0$. Then for any $\varepsilon > 0$, one has the well-separation condition

$$\sup_{\|\beta - \beta_0\|^2 \geq \varepsilon} E_{\beta_0} m_\beta(\mathbf{X}, y) < E_{\beta_0} m_{\beta_0}(\mathbf{X}, y).$$

Let $\tilde{M}_n(\beta) := \frac{1}{n} \sum_{i=1}^n m_\beta(\mathbf{X}_i^*, Y_i^*)$, which is essentially a logarithmic likelihood function of subsampled GLMs, and $\hat{\beta}_n$ is the function's maximum point. Thus, one has the nearly maximization $\tilde{M}_n(\hat{\beta}_n) \geq \tilde{M}_n(\beta_0) \geq \tilde{M}_n(\beta_0) - o_P(1)$.

Let $\mathcal{F} := \{m_\beta(\mathbf{X}, y) = -y\psi(\beta^T \mathbf{X}) + b(\psi(\beta^T \mathbf{X}))\}$, $\beta \in \Theta$. Now one obtains

$$\begin{aligned} |m_{\beta_1}(\mathbf{X}, y) - m_{\beta_2}(\mathbf{X}, y)| &= |-y\psi(\beta_1^T \mathbf{X}) + b(\psi(\beta_1^T \mathbf{X})) + y\psi(\beta_2^T \mathbf{X}) - b(\psi(\beta_2^T \mathbf{X}))| \\ &= |y\psi(\beta_1^T \mathbf{X}) - b(\psi(\beta_1^T \mathbf{X})) - y\psi(\beta_2^T \mathbf{X}) + b(\psi(\beta_2^T \mathbf{X}))| \\ &= |y\dot{\psi}(\xi_{(5)}^T \mathbf{X})(\beta_1^T \mathbf{X} - \beta_2^T \mathbf{X})\mathbf{X} \\ &\quad - \dot{b}(\psi(\xi_{(6)}^T \mathbf{X}))\dot{\psi}(\xi_{(6)}^T \mathbf{X})(\beta_1^T \mathbf{X} - \beta_2^T \mathbf{X})\mathbf{X}| \\ &\leq C_4 |y - \dot{b}(\psi(\xi_{(6)}^T \mathbf{X}))| \cdot |\beta_1^T \mathbf{X} - \beta_2^T \mathbf{X}| \cdot \|\mathbf{X}\| \\ &\leq C_4 |y - \dot{b}(\psi(\xi_{(6)}^T \mathbf{X}))| \cdot \|\mathbf{X}\|^2 \cdot \|\beta_1 - \beta_2\|, \forall \beta_1, \beta_2 \in \Theta, \end{aligned}$$

where $\xi_{(5)}$ and $\xi_{(6)}$ are both between β_1 and β_2 and $C_4 > 0$.

Let $\tilde{m}(\mathbf{X}, y) = |y - \dot{b}(\psi(\xi_{(6)}^T \mathbf{X}))| \cdot \|\mathbf{X}\|^2$ and by (B.3), one has

$$\|\tilde{m}(\mathbf{X}, y)\|_{P,1} := E_{\beta_0} |\tilde{m}(\mathbf{X}, Y)| \leq E_{\beta_0} \sup_{\beta \in \Theta} [|y - \dot{b}(\psi(\beta^T \mathbf{X}))| \cdot \|\mathbf{X}\|^2] < \infty,$$

where $\|\cdot\|_{\tilde{P},1} = \tilde{P}|\cdot|$ is the $L_1(\tilde{P})$ -norm in P269-P270 of [11] and $\tilde{P} := E_{\beta_0}$. And then from the Example 19.7 in [11], one obtains

$$N_{[\cdot]}(\varepsilon, \mathcal{F}, L_1(E_{\beta_0})) \leq K \left(\frac{\text{diam}\Theta}{\varepsilon / \|\tilde{m}\|_{E_{\beta_0},1}} \right)^p < \infty, \text{ every } 0 < \varepsilon < \text{diam}\Theta < \infty$$

where $N_{[\cdot]}(\varepsilon, \mathcal{F}, L_1(E_{\beta_0}))$ is called *bracketing number* which is the minimum number of ε -brackets needed to cover \mathcal{F} ; see P270 in [11]. And K is a constant, and $\text{diam}\Theta = \sup_{\beta_1, \beta_2 \in \Theta} \|\beta_1 - \beta_2\|$.

Therefore, the class \mathcal{F} is P-Glivenko-Cantelli by Theorem 19.4 in [11]. And from the definition of P-Glivenko-Cantelli in P269 of [11], we have

$$\sup_{\beta \in \Theta} |\tilde{M}_n(\beta) - E_{\beta_0} m_{\beta}(\mathbf{X}, \mathbf{y})| \xrightarrow{a.s.} 0.$$

Finally, according to Theorem 5.7 in [11], we get $\|\hat{\beta}_n - \beta_0\| = o_P(1)$. The proof is then completed. \square

Recall (A7) and (A8) respectively, then $\nabla s_n(\gamma) = R_n(\gamma) - M_n(\gamma)$. Let $\Phi = E(\nabla s_n(\beta))$, then we have the following lemma.

Lemma A3. For $\beta \in \mathbb{R}^p$, assume that

(E.1) $R_n(\beta)$ is finite and nonsingular.

(E.2) For $1 \leq k, j \leq p$,

$$E \left[\sum_{i \in S} \frac{a_i}{\pi_i} [Y_i - \mu(\psi(\beta^T \mathbf{X}_i))] \ddot{\psi}(\beta^T \mathbf{X}_i) x_{ik} x_{ij} \right]^2 = o(1).$$

(E.3) For $1 \leq k, j \leq p$,

$$\text{Var} \left[\sum_{i \in U} \frac{a_i}{\pi_i} [\dot{\psi}(\beta^T \mathbf{X}_i)]^2 \dot{\mu}(\psi(\beta^T \mathbf{X}_i)) x_{ik} x_{ij} \right] = o(1).$$

Then,

$$\nabla s_n(\beta) \rightarrow \Phi.$$

Proof. One derives each entry in the matrix by

$$\begin{aligned} (\nabla s_n(\beta))_{kj} &= (R_n(\beta))_{kj} - (M_n(\beta))_{kj} \\ &= \sum_{i \in S} \frac{1}{\pi_i^*} [Y_i^* - \mu(\psi(\beta^T \mathbf{X}_i^*))] \dot{\psi}(\beta^T \mathbf{X}_i^*) x_{ik}^* x_{ij}^* \\ &\quad - \sum_{i \in S} \frac{1}{\pi_i^*} [\dot{\psi}(\beta^T \mathbf{X}_i^*)]^2 \dot{\mu}(\psi(\beta^T \mathbf{X}_i^*)) x_{ik}^* x_{ij}^*. \end{aligned}$$

By the definition of Φ , one has

$$\Phi_{kj} = E(\nabla s_n(\beta))_{kj} = E \left[- \sum_{i \in S} \frac{1}{\pi_i^*} \dot{\mu}(\psi(\beta^T \mathbf{X}_i^*)) [\dot{\psi}(\beta^T \mathbf{X}_i^*)]^2 x_{ik}^* x_{ij}^* \right].$$

Next, one obtains

$$\begin{aligned}
 & E\left[(\nabla s_n(\boldsymbol{\beta}))_{kj} - \Phi_{kj}\right]^2 \\
 &= E\left\{\left[(\nabla s_n(\boldsymbol{\beta}))_{kj} - \Phi_{kj}\right]^2 \middle| (\mathbf{X}_i, Y_i)_{i=1}^N\right\} \\
 &= E\left\{\left[(\nabla s_n(\boldsymbol{\beta}))_{kj} - E(\nabla s_n(\boldsymbol{\beta}))_{kj}\right]^2 \middle| (\mathbf{X}_i, Y_i)_{i=1}^N\right\} \\
 &= E\left\{\left[(R_n(\boldsymbol{\beta}))_{kj} - (M_n(\boldsymbol{\beta}))_{kj} + E(M_n(\boldsymbol{\beta}))_{kj}\right]^2 \middle| (\mathbf{X}_i, Y_i)_{i=1}^N\right\} \\
 &= E\left\{(R_n(\boldsymbol{\beta}))_{kj}^2 + \left[E(M_n(\boldsymbol{\beta}))_{kj} - (M_n(\boldsymbol{\beta}))_{kj}\right]^2 \middle| (\mathbf{X}_i, Y_i)_{i=1}^N\right\} \\
 &\quad + E\left\{2(R_n(\boldsymbol{\beta}))_{kj}\left[E(M_n(\boldsymbol{\beta}))_{kj} - (M_n(\boldsymbol{\beta}))_{kj}\right] \middle| (\mathbf{X}_i, Y_i)_{i=1}^N\right\} \\
 &= E\left\{(R_n(\boldsymbol{\beta}))_{kj}^2 \middle| (\mathbf{X}_i, Y_i)_{i=1}^N\right\} + \text{Var}\left\{(M_n(\boldsymbol{\beta}))_{kj} \middle| (\mathbf{X}_i, Y_i)_{i=1}^N\right\} \\
 &= o(1),
 \end{aligned}$$

where the first equality is based on the fact that after conditioning on the N data points, the n repeating sampling steps should be independent and identically distributed in each step. The last equality holds by the conditions (E.2) and (E.3). \square

Lemma A4. Under the conditions (C.1)–(C.5) in Theorem 5, if $s_n(\hat{\boldsymbol{\beta}}_n) = 0$ for all large n and $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| = O_P(1/N)$, then

$$s_n(\boldsymbol{\beta}_0) = -\Phi(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + o_P(1).$$

Proof. By Taylor’s expansion:

$$0 = s_n(\hat{\boldsymbol{\beta}}_n) = s_n(\boldsymbol{\beta}_0) + \nabla s_n(\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + \frac{1}{2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^T \Sigma(\tilde{\boldsymbol{\beta}}_n)(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0),$$

where $\Sigma(\tilde{\boldsymbol{\beta}}_n) = \nabla^2 s_n(\tilde{\boldsymbol{\beta}}_n)$ and $\tilde{\boldsymbol{\beta}}_n$ is between $\boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\beta}}_n$. From assumption (C.3), (C.4) and (C.5) in Theorem 5, we have

$$\begin{aligned}
 \|\Sigma(\tilde{\boldsymbol{\beta}}_n)\| &= \left\| \sum_{i \in S} \frac{1}{\pi_i^*} \ddot{\phi}_{\boldsymbol{\beta}}(\mathbf{X}_i^*, Y_i^*) \right\| \\
 &\leq \sum_{i \in S} \frac{1}{\pi_i^*} \cdot \|\ddot{\phi}_{\boldsymbol{\beta}}(\mathbf{X}_i^*, Y_i^*)\| = O(nN).
 \end{aligned}$$

Then $\frac{1}{2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^T \Sigma(\tilde{\boldsymbol{\beta}}_n)(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) = o_P(1)$. Therefore, by Lemma A3, one has

$$0 = s_n(\boldsymbol{\beta}_0) + (\Phi + o(1))(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + o_P(1),$$

which implies

$$s_n(\boldsymbol{\beta}_0) = -\Phi(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + o_P(1).$$

Hence, the proof is completed. \square

Lemma A5. $\{\bar{M}_i\}_{i=1}^n$ is a martingale difference sequence adapt to the filtration $\{\mathcal{F}_{N,i}\}_{i=1}^n$.

Proof. The \bar{M}_i 's are $\mathcal{F}_{N,i}$ -measurable by the definition of \bar{M}_i and the definition of the filtration $\{\mathcal{F}_{N,i}\}_{i=1}^n$. Then we obtain

$$\begin{aligned} E[\bar{M}_i | \mathcal{F}_{N,i-1}] &= E \left[\frac{1}{\pi_i^*} \phi_\beta(\mathbf{X}_i^*, Y_i^*) - \sum_{j=1}^N \phi_\beta(\mathbf{X}_j, Y_j) \middle| \mathcal{F}_{N,i-1} \right] \\ &= E \left[\frac{1}{\pi_i^*} \phi_\beta(\mathbf{X}_i^*, Y_i^*) \middle| \mathcal{F}_{N,i-1} \right] - E \left[\sum_{j=1}^N \phi_\beta(\mathbf{X}_j, Y_j) \middle| \mathcal{F}_{N,i-1} \right] \\ &= \frac{\sum_{i=1}^N \pi_i \frac{1}{\pi_i} \phi_\beta(\mathbf{X}_i, Y_i)}{\sum_{i=1}^N \pi_i} - \frac{\sum_{i=1}^N \pi_i \sum_{j=1}^N \phi_\beta(\mathbf{X}_j, Y_j)}{\sum_{i=1}^N \pi_i} \\ &= \sum_{i=1}^N \phi_\beta(\mathbf{X}_i, Y_i) - \sum_{j=1}^N \phi_\beta(\mathbf{X}_j, Y_j) \\ &= 0. \end{aligned}$$

By the definition of martingale difference sequence in P230 of [29], the proof is completed. \square

Under the definition of T, \bar{M}, Q , it is obvious that $\text{Var}(T) = \text{Var}(\bar{M}) + \text{Var}(Q)$.

Lemma A6. $\sup_N \lambda_{\max}(B_N) \leq 1$.

Proof. By symmetry of B_N , we only to show for any $N, \mathbf{I} - B_N$ is positive definite.

$$\begin{aligned} \mathbf{I} - B_N &= \text{Var}(T)^{-\frac{1}{2}} (\text{Var}(T) - \text{Var}(\bar{M})) \text{Var}(T)^{-\frac{1}{2}} \\ &= \text{Var}(T)^{-\frac{1}{2}} \text{Var}(Q) \text{Var}(T)^{-\frac{1}{2}}. \end{aligned}$$

Therefore, $\mathbf{I} - B_N$ is equivalent to the positive definite matrix $\text{Var}(Q)$. The proof is completed. \square

Lemma A7 (Multivariate version of martingale CLT, Lemma 4 in [19]). For $k = 1, 2, 3, \dots$, let $\{\xi_{ki}; i = 1, 2, \dots, N_k\}$ be a martingale difference sequence in \mathbb{R}^p relative to the filtration $\{\mathcal{F}_{ki}; i = 0, 1, \dots, N_k\}$ and let $Y_k \in \mathbb{R}^p$ be an \mathcal{F}_{k0} -measurable random vector. Set $S_k = \sum_{i=1}^{N_k} \xi_{ki}$.

Assume that

(F.1) $\lim_{k \rightarrow \infty} \sum_{i=1}^{N_k} E[\|\xi_{ki}\|^4] = 0;$

(F.2) $\lim_{k \rightarrow \infty} E \left[\left\| \sum_{i=1}^{N_k} [\xi_{ki} \xi_{ki}^T | \mathcal{F}_{k,i-1}] - B_k \right\|^2 \right] = 0$ for some sequence of positive definite matrices $\{B_k\}_{k=1}^\infty$ with $\sup_k \lambda_{\max}(B_k) < \infty$ i.e., the largest eigenvalue is uniformly bounded;

(F.3) For some probability distribution L_0 , $*$ denotes convolution and $L(\cdot)$ denotes the law of random variates:

$$L(Y_k) * N(0, B_k) \xrightarrow{d} L_0.$$

Then

$$L(Y_k + S_k) \xrightarrow{d} L_0.$$

Lemma A8 (Asymptotic normality of $s_n(\beta_0)$). Assume that

(G.1) $\lim_{N \rightarrow \infty} \sum_{i=1}^n E[\|\zeta_{Ni}\|^4] = 0;$

$$(G.2) \lim_{N \rightarrow \infty} E \left[\left\| \sum_{i=1}^n E[\zeta_{Ni} \tilde{\zeta}_{Ni}^T | \mathcal{F}_{N,i-1}] - B_N \right\|^2 \right] = 0.$$

Then

$$\text{Var}(T)^{-\frac{1}{2}} \cdot T \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_p).$$

Proof. The conditions in lemma A7 can be substituted with

$$\begin{aligned} \tilde{\zeta}_{ki} &= \zeta_{Ni}, & Y_k &= \text{Var}(T)^{-\frac{1}{2}} \cdot Q, \\ B_k &= B_N, & L_0 &\sim N(\mathbf{0}, \mathbf{I}_p). \end{aligned}$$

By Lemma A5, conditions (F.1) and (F.2) of Lemma A7 are satisfied. Next we only need to show the third condition in Lemma A7 holds. According to central limit theorem we have

$$\text{Var}(Q)^{-\frac{1}{2}} \cdot Q \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_p).$$

For any $t \in \mathbb{R}^p$, let $\tilde{t} = \text{Var}(T)^{-\frac{1}{2}}t$, $\tilde{X} = iQ$, due to the properties of the complex multivariate normal distributions are equivalent to the properties of real multivariate normal distributions in P222 of [30], and $EQ = 0$, one has

$$\begin{aligned} \text{Var}(\tilde{X}) &= \text{Var}(iQ) = E[(iQ)(iQ)] - [E(iQ)]^2 \\ &= -EQ^2 = -EQ^2 + (EQ)^2 = -\text{Var}(Q). \end{aligned}$$

Thus, according to Equations (45.4)–(45.6) in P108 of [30], one has

$$Ee^{i^T \tilde{X}} = e^{i^T E(\tilde{X}) + \frac{1}{2} i^T \text{Var}(\tilde{X}) i} = e^{-\frac{1}{2} t^T \text{Var}(T)^{-\frac{1}{2}} \text{Var}(Q) \text{Var}(T)^{-\frac{1}{2}} t}.$$

Further, we obtain

$$E[e^{it^T \text{Var}(T)^{-\frac{1}{2}} Q}] \cdot e^{-\frac{1}{2} t^T \text{Var}(T)^{-\frac{1}{2}} \text{Var}(\tilde{M}) \text{Var}(T)^{-\frac{1}{2}} t} = e^{-\frac{1}{2} t^T t}.$$

Therefore, condition (F.3) in Lemma A7 is verified. Then one obtains

$$\text{Var}(T)^{-\frac{1}{2}} T = \text{Var}(T)^{-\frac{1}{2}} \cdot Q + \text{Var}(T)^{-\frac{1}{2}} \cdot \tilde{M} \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_p).$$

The proof is completed. □

Proof of Theorem 5. According to Lemma A4,

$$\Phi(\hat{\beta}_n - \beta_0) + o_p(1) = -s_n(\beta_0) = -T. \tag{A18}$$

Multiplying with $\text{Var}(T)^{-\frac{1}{2}}$ in (A18), one obtains

$$\text{Var}(T)^{-\frac{1}{2}} \Phi(\hat{\beta}_n - \beta_0) + o_p(\|\text{Var}(T)^{-\frac{1}{2}}\|) = -\text{Var}(T)^{-\frac{1}{2}} T.$$

Applying Lemma A8, one obtains

$$\text{Var}(T)^{-\frac{1}{2}} \Phi(\hat{\beta}_n - \beta_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_p).$$

The proof is completed. □

References

1. Xi, R.; Lin, N. Direct regression modelling of high-order moments in big data. *Stat. Its Interface* **2016**, *9*, 445–452. [\[CrossRef\]](#)
2. Tewes, J.; Politis, D.N.; Nordman, D.J. Convolved subsampling estimation with applications to block bootstrap. *Ann. Stat.* **2019**, *47*, 468–496. [\[CrossRef\]](#)

3. Yu, J.; Wang, H.; Ai, M.; Zhang, H. Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data. *J. Am. Stat. Assoc.* **2022**, *117*, 265–276. [[CrossRef](#)]
4. Yao, Y.; Wang, H. A review on optimal subsampling methods for massive datasets. *J. Data Sci.* **2021**, *19*, 151–172. [[CrossRef](#)]
5. Yu, J.; Wang, H. Subdata selection algorithm for linear model discrimination. *Stat. Pap.* **2021**, *63*, 1883–1906. [[CrossRef](#)]
6. Fu, S.; Chen, P.; Liu, Y.; Ye, Z. Simplex-based Multinomial Logistic Regression with Diverging Numbers of Categories and Covariates. *Stat. Sin.* **2022**, *in press*. [[CrossRef](#)]
7. Ma, J.; Xu, J.; Maleki, A. Analysis of sensing spectral for signal recovery under a generalized linear model. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 22601–22613.
8. Mahmood, T. Generalized linear model based monitoring methods for high-yield processes. *Qual. Reliab. Eng. Int.* **2020**, *36*, 1570–1591. [[CrossRef](#)]
9. Ai, M.; Yu, J.; Zhang, H.; Wang, H. Optimal Subsampling Algorithms for Big Data Regressions. *Stat. Sin.* **2021**, *31*, 749–772. [[CrossRef](#)]
10. Wang, H.; Zhu, R.; Ma, P. Optimal subsampling for large sample logistic regression. *J. Am. Stat. Assoc.* **2018**, *113*, 829–844. [[CrossRef](#)]
11. van der Vaart, A.W. *Asymptotic Statistics*; Cambridge University Press: London, UK, 1998.
12. Wooldridge, J.M. Inverse probability weighted M-estimators for sample selection, attrition, and stratification. *Port. Econ. J.* **2002**, *1*, 117–139. [[CrossRef](#)]
13. Durrett, R. *Probability: Theory and Examples*, 5th ed.; Cambridge University Press: Cambridge, UK, 2019.
14. McCullagh, P.; Nelder, J. *Generalized Linear Models*, 2nd ed.; Chapman and Hall/CRC: Boca Raton, FL, USA, 1989.
15. Fahrmeir, L.; Kaufmann, H. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Ann. Stat.* **1985**, *13*, 342–368. [[CrossRef](#)]
16. Shao, J. *Mathematical Statistics*, 2nd ed.; Springer: New York, NY, USA, 2003.
17. Yin, C.; Zhao, L.; Wei, C. Asymptotic normality and strong consistency of maximum quasi-likelihood estimates in generalized linear models. *Sci. China Ser. A* **2006**, *49*, 145–157. [[CrossRef](#)]
18. Rigollet, P. Kullback-Leibler aggregation and misspecified generalized linear models. *Ann. Stat.* **2012**, *40*, 639–665. [[CrossRef](#)]
19. Zhang, T.; Ning, Y.; Ruppert, D. Optimal sampling for generalized linear models under measurement constraints. *J. Comput. Graph. Stat.* **2021**, *30*, 106–114. [[CrossRef](#)]
20. Ohlsson, E. Asymptotic normality for two-stage sampling from a finite population. *Probab. Theory Relat. Fields* **1989**, *81*, 341–352. [[CrossRef](#)]
21. Zhang, H.; Wei, H. Sharper Sub-Weibull Concentrations. *Mathematics* **2022**, *10*, 2252. [[CrossRef](#)]
22. Gong, T.; Dong, Y.; Chen, H.; Dong, B.; Li, C. Markov Subsampling Based on Huber Criterion. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *in press*. [[CrossRef](#)]
23. Xiao, Y.; Yan, T.; Zhang, H.; Zhang, Y. Oracle inequalities for weighted group lasso in high-dimensional misspecified Cox models. *J. Inequalities Appl.* **2020**, *2020*, 252. [[CrossRef](#)]
24. Zhang, H.; Jia, J. Elastic-net regularized high-dimensional negative binomial regression: Consistency and weak signals detection. *Stat. Sin.* **2022**, *32*, 181–207. [[CrossRef](#)]
25. Ding, J.L.; Chen, X.R. Large-sample theory for generalized linear models with non-natural link and random variates. *Acta Math. Appl. Sin.* **2006**, *22*, 115–126. [[CrossRef](#)]
26. Jennrich, R.I. Asymptotic properties of non-linear least squares estimators. *Ann. Math. Stat.* **1969**, *40*, 633–643. [[CrossRef](#)]
27. White, H. Maximum likelihood estimation of misspecified models. *Econom. J. Econom. Soc.* **1982**, *50*, 1–25. [[CrossRef](#)]
28. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [[CrossRef](#)]
29. Davidson, J. *Stochastic Limit Theory: An Introduction for Econometricians*; OUP Oxford: Oxford, UK, 1994.
30. Kotz, S.; Balakrishnan, N.; Johnson, N.L. *Continuous Multivariate Distributions, Volume 1: Models and Applications*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2000.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.