mathematics

MDPI

Article

Heterogeneous Overdispersed Count Data Regressions via Double-Penalized Estimations

Shaomin Li [1], Haoyu Wei [2,*] and Xiaoyu Lei [3]

[1] Center for Statistics and Data Science, Beijing Normal University, Zhuhai 516087, China; lsmjim@bnu.edu.cn
[2] Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA
[3] Department of Statistics, University of Chicago, Chicago, IL 60637, USA; leixy@uchicago.edu
* Correspondence: hwei4@ncsu.edu

Abstract: Recently, the high-dimensional negative binomial regression (NBR) for count data has been widely used in many scientific fields. However, most studies assumed the dispersion parameter as a constant, which may not be satisfied in practice. This paper studies the variable selection and dispersion estimation for the heterogeneous NBR models, which model the dispersion parameter as a function. Specifically, we proposed a double regression and applied a double ℓ_1-penalty to both regressions. Under the restricted eigenvalue conditions, we prove the oracle inequalities for the lasso estimators of two partial regression coefficients for the first time, using concentration inequalities of empirical processes. Furthermore, derived from the oracle inequalities, the consistency and convergence rate for the estimators are the theoretical guarantees for further statistical inference. Finally, both simulations and a real data analysis demonstrate that the new methods are effective.

Keywords: negative binomial regressions; heterogeneous count data regression; estimation of dispersion parameter; oracle inequalities

MSC: 62E17; 62E20; 62F07

check for updates

Citation: Li, S.; Wei, H.; Lei, X. Heterogeneous Overdispersed Count Data Regressions via Double-Penalized Estimations. *Mathematics* **2022**, *10*, 1700. https://doi.org/10.3390/math10101700

Academic Editor: Francisco-José Vázquez-Polo

Received: 21 April 2022
Accepted: 11 May 2022
Published: 16 May 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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 many scientific fields, such as biomedical science, ecology, and economics, experimental and observational studies often yield count data, a type of data in which the observations can take only the non-negative integer values. The Poisson regression models are commonly used for count data. However, it needs a restrictive assumption that the variance equals the mean. For many count data, the variance is often larger than the mean [1], which is called overdispersion. Because the Poisson regression model is invalid under the overdispersion case, a more general and flexible regression model, the negative binomial regression, has attracted lots of research attention and become popular in analyzing count data [2–4].

With the advance of modern data collection techniques, high-dimensional data are becoming increasingly common in scientific studies. The widely used estimations for the high-dimensional parameter include the lasso [5], the scad [6], the elastic net [7], the adaptive lasso [8], and so on. Recently, there has been much research on the high-dimensional NBR model, such as [9–14]. All of these works assumed the dispersion parameter as a constant. In practice, however, not all models satisfy the assumption. If the dispersion parameter is wrongly assumed to be a constant, the estimation of the mean regression will perform poorly as shown in the simulation in Section 4.1, thus the need to model the dispersion parameter as a function of some covariates. The heterogeneous negative binomial regression (HNBR) extends the NBR by observation-specific parameterization of the dispersion parameter [3]. The HNBR is a valuable tool for assessing the source of overdispersion. It belongs to the double-generalized linear models (DGLMs) or vector-generalized linear models (VGLMs), which are very useful in fitting more complex and

potentially realistic models [15–18]. However, it appears that there is no study on selecting the dispersion explanation variables in the HNBR model.

In this paper, we study the variable selection and dispersion estimation for the heterogeneous NBR models. To the best of our knowledge and based on the literature, this study is the first. Specifically, we propose a double regression to estimate the coefficients of NB dispersion and NBR simultaneously. Because of the high dimension of the covariates, we apply a double ℓ_1 penalty to both regressions. The two adjustment parameters we set are different because the first-order conditions for estimating the regression coefficients are entirely different from those for estimating the dispersion parameters. We construct an algorithm to perform variable selection and dispersion estimation simultaneously. Similar studies on high-dimensional NBR models include [19], which assumed the dispersion parameter as a constant. Their method requires an iterative algorithm to estimate the mean regression and dispersion alternatively and implement a lasso in each iteration. If there are many iterations, such an algorithm is a waste of computing resources.

The rest of the paper is organized as follows. Section 2 introduces the heterogeneous overdispersed count data model and defines the double ℓ_1-penalized estimators for the mean and dispersion regressions. Then we use a technique called the stochastic Lipschitz condition to derive the asymptotic results in Section 3. Simulation studies and a real data application are given in Section 4. Finally, Section 5 concludes the article with a discussion. All proofs and technical details are provided in Appendix A.

2. Double ℓ_1-Penalized NBR

2.1. Heterogeneous Overdispersed Count Data Regressions

Suppose we have n count responses Y_i and p-dimensional covariates $X_i = (x_{i1}, \cdots, x_{ip})$, $i \in [n] := \{1, 2, \ldots, n\}$. For the Poisson regression models, the response obeys the Poisson distribution

$$\mathrm{P}(Y_i = y_i|\lambda_i) = \frac{\lambda_i^{y_i}}{y_i!} e^{-\lambda_i}, \qquad i \in [n].$$

with $\lambda_i = \mathrm{E}(Y_i)$, we require that the positive parameter λ_i is related to a linear combination of p covariates. A plausible assumption for the link function is $\eta(\lambda_i) = \log(\lambda_i) = X_i^\top \beta$. It is worth noting that $\mathrm{E}(Y_i|X_i) = \mathrm{var}(Y_i|X_i) = \exp(X_i^\top \beta) > 0$.

For the traditional negative binomial regression, it assumes that the count data response obeys the NB distribution with overdispersion:

$$\mathrm{P}(Y_\mathrm{i} = y_i|X_i) =: f(y_i; k, \mu_i) = \frac{\Gamma(k + y_i)}{\Gamma(k) y_i!} (\frac{\mu_i}{k + \mu_i})^{y_i} (\frac{k}{k + \mu_i})^k, \qquad i \in [n], \tag{1}$$

with $\mathrm{E}(Y_i|X_i) = \mu_i = \exp(\beta^\top X_i)$ and k is an unknown qualification of the overdispersion level. When $k \to \infty$, we have $\mathrm{var}(Y_i|X_i) = \mu_i + \frac{\mu_i^2}{k} \to \mu_i = \mathrm{E}(Y_i|X_i)$, the Poisson regression for the mean parameter μ_i. Thus, the Poisson regression is a limiting case of negative binomial regression when the dispersion parameter k tends to infinite.

In the heterogeneous negative binomial regression, k is proposed as a specific parameterization, i.e., $k = k(X_i)$. More specifically, we assume in this paper that

$$\mu(x) = \exp\{\theta^{(1)\top} x\}, \qquad k(x) = \exp\{\theta^{(2)\top} x\}.$$

For notation simplicity, we denote

$$Pf := \mathrm{E} f(X_i, Y_i), \qquad \mathbb{P}_n f := \frac{1}{n} \sum_{i=1}^n f(X_i, Y_i), \qquad \mathbb{G}_n f := \sqrt{n}(\mathbb{P}_n - P) f,$$

for any measurable and integrable function f.

Let $\theta = (\theta^{(1)\top}, \theta^{(2)\top})^\top \in \mathbb{R}^{2p}$, the log-likelihood is

$$
\begin{aligned}
n\ell(\theta) = \log\prod_{i=1}^{n} f(y_i, k_i, \mu_i) &= \sum_{i=1}^{n} \log\left\{ \frac{\Gamma(k_i + Y_i)}{\Gamma(k_i) Y_i!} \left(\frac{\mu_i}{k_i + \mu_i}\right)^{Y_i} \left(\frac{k_i}{k_i + \mu_i}\right)^{k_i} \right\} \\
&= \sum_{i=1}^{n} \Bigg[- \log \frac{\Gamma\left(\exp\{X_i^\top \theta^{(2)}\}\right)}{\Gamma\left(Y_i + \exp\{X_i^\top \theta^{(2)}\}\right)} + Y_i X_i^\top \left(\theta^{(1)} - \theta^{(2)}\right) \\
&\qquad - \left[Y_i + \exp\{X_i^\top \theta^{(2)}\}\right] \log\left(1 + \exp\{X_i^\top \left(\theta^{(1)} - \theta^{(2)}\right)\}\right) - \log Y_i! \Bigg]
\end{aligned}
$$

We use the negative log-likelihood as the loss function γ, and define

$$\gamma(\theta) := -\log f(y \mid x, \theta) + \log y!.$$

Denote $\partial_j := \frac{\partial}{\partial \theta^{(j)}}$, $j = 1, 2$, the score function for $\theta^{(1)}$ is

$$\partial_1 \ell(\theta) = -\mathbb{P}_n \partial_1 \gamma(\theta) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - e^{X_i^\top \theta^{(1)}}) \frac{e^{X_i^\top \theta^{(2)}} X_i}{e^{X_i^\top \theta^{(1)}} + e^{X_i^\top \theta^{(2)}}}.$$

Furthermore, fix $\theta^{(1)}$, the score function for $\theta^{(2)}$ is

$$\partial_2 \ell(\theta) = \mathbb{P}_n \partial_2 \gamma(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \left[\log\left(1 + e^{X_i^\top (\theta^{(1)} - \theta^{(2)})}\right) - \sum_{j=0}^{Y_i - 1} \frac{1}{j + e^{X_i^\top \theta^{(2)}}} \right] + \frac{Y_i - e^{X_i^\top \theta^{(1)}}}{e^{X_i^\top \theta^{(1)}} + e^{X_i^\top \theta^{(2)}}} \right\} e^{X_i^\top \theta^{(2)}} X_i.$$

It is easy to verify that

$$P\partial_1 \ell(\theta) = P\partial_2 \ell(\theta) = 0.$$

Thus, from now, we will suppose the true value of parameter θ is θ^*.

2.2. Heterogeneous Overdispersed NBR via Double ℓ_1 Penalty

The weighted lasso estimator under our circumstance is defined as

$$\widehat{\theta}_n = \underset{\theta \in \Theta}{\operatorname{argmin}} \left(\mathbb{P}_n \gamma(\theta) + \lambda \|\theta\|_{\omega,1}\right), \tag{2}$$

where $\lambda > 0$ is the tuning parameter and the weighted norm is defined by

$$\lambda \|\theta\|_{\omega,1} = \lambda_1 \|\theta^{(1)}\|_1 + \lambda_2 \|\theta^{(2)}\|_1 = \lambda \left(\omega_1 \|\theta^{(1)}\|_1 + \omega_2 \|\theta^{(2)}\|_1\right),$$

and $\omega = (\omega_1, \omega_2)^\top = (\lambda_1/\lambda, \lambda_2/\lambda)^\top \in [0,1] \times [0,1]$ is the weight, $\|\cdot\|_1$ means the ℓ_1-norm. This technique is also used in [20]. Equation (2) is a weighted double ℓ_1-penalized problem, which is a kind of convex penalty optimization, and when $\lambda_1 = \lambda_2$, it becomes a single-penalized problem. In this paper, we use different λ_1 and λ_2, as the first-order conditions for estimating the regression coefficients are entirely different from those for estimating the dispersion parameters, and take $\lambda = \lambda_1 \vee \lambda_2$.

Because the weighted group lasso estimator $\widehat{\theta}_n$ has no closed-form solution, we need to use iterative methods such as quasi-Newton or coordinate descent methods. We use BIC to choose the parameter λ_1 and λ_2.

$$\mathrm{BIC}(\lambda_1, \lambda_2) = -2\ell(\widehat{\theta}_n) + \frac{\log n}{n} k,$$

where k is the number of nonzero estimated coefficients. To illustrate the algorithm explicitly, we rewrite $\gamma(\theta)$ as $\gamma(\theta^{(1)\top} x,\ \theta^{(2)\top} x)$ and define $\theta^{(3)} = \lambda_2/\lambda_1 \theta^{(2)}$, $\theta^\dagger = (\theta^{(1)\top}, \theta^{(3)\top})^\top$. Converting $\theta^{(2)}$ into $\theta^{(3)}$ turns the double ℓ_1-penalized problem into a single penalized one, which can be solved through some R packages, such as "lbfgs". The algorithm is formally given in Algorithm 1.

Algorithm 1 Double ℓ_1-Penalized Optimization

Input: the set of tuning parameters $\Lambda = \{(\lambda_{1,i}, \lambda_{2,i})\}_{i=1}^{m}$
Output: the estimate $\widehat{\theta}_n$
for $i = 1, \ldots, m,$ **do**
let $x^* = \frac{\lambda_{1,i}}{\lambda_{2,i}} x$;
solve $\widehat{\theta^\dagger} = (\widehat{\theta}^{(1)\top}, \widehat{\theta}^{(3)\top})^\top = \operatorname{argmin}_{\theta^\dagger \in \Theta} \left(\mathbb{P}_n \gamma(\theta^{(1)\top} x,\ \theta^{(3)\top} x^*) \right) + \lambda_{1,i} \|\theta^\dagger\|_1 \big)$;
obtain the estimate $\widehat{\theta}_{n,i} = (\widehat{\theta}^{(1)\top}, \frac{\lambda_{2,i}}{\lambda_{1,i}} \widehat{\theta}^{(3)\top})^\top$;
compute $\mathrm{BIC}(\lambda_{1,i}, \lambda_{2,i}) = -2\ell(\widehat{\theta}_{n,i}) + \dfrac{\log n}{n} k_i$;
end for
find $i_{opt} = \operatorname{argmin}_{i=1,\ldots,m} \mathrm{BIC}(\lambda_{1,i}, \lambda_{2,i})$;
return $\widehat{\theta}_{n,i_{opt}}$

The proposed algorithm can perform variable selection and dispersion estimation simultaneously. Similar studies on high-dimensional NBR models include [19], which assumed the dispersion parameter as a constant. However, their method requires an iterative algorithm to estimate the mean regression and dispersion alternatively and implement lasso in each iteration. If there are many iterations, such an algorithm is a waste of computing resources.

3. Main Results

3.1. Stochastic Lipschitz Conditions

We write the maximum of Y_i from the sample of size n as $M_{Y,n}$, then the sample space for $\{Y_i\}_{i=1}^n$ is $\mathcal{Y} := \{y \in \mathbb{N}, y \leq M_{y,n}\}$., i.e., $M_{y,n} = \max_{i \in [n]} Y_i$. Note that $\lim_{n\to\infty} \mathrm{P}(M_{y,n} = \infty) = 1$; what we need to tackle is actually an unbounded empirical process. However, for $z := \begin{pmatrix} x^\top & \\ & x^\top \end{pmatrix} \in \mathbb{R}^{2\times 2p}$, we can assume the value space $\mathcal{S}$ for $s := z\theta$ is bounded and satisfies

$$\mathcal{S} := \left\{ s = (s_1, s_2)^\top \in \mathbb{R}^2,\ -\infty < m_{s,n} \leq s_j \leq |s_j| \leq M_{s,n} < \infty,\ j = 1, 2 \right\}.$$

As we can see, the most significant difference between this article and other conventional literature about lasso estimators is that we use $s = z\theta$ rather than θ as the explanatory variable to analyze the properties of the loss function γ. This is not a traditional way. At first glimpse, the combination may complicate the analysis in the next step because the KKT condition requires the story about $\frac{\partial}{\partial \theta}\gamma$, which is critical for the traditional convex penalty problem. However, this article will try a different approach, the stochastic Lipschitz conditions introduced in the event $\mathcal{A}$ of Proposition 1 in [14], to solve the ℓ_1-penalization problem. Define the *local stochastic Lipschitz constant* by

$$\mathrm{Lip}(f; \theta^*) := \sup_{\theta \in \Theta / \{\theta^*\}} \left| \frac{\sqrt{n}\mathbb{G}_n \big(f(\theta) - f(\theta^*) \big)}{\|\theta - \theta^*\|_1} \right|.$$

The most apparent advantage of the stochastic Lipschitz conditions over the KKT condition is that it can easily deal with the several parameters involved in different locations of the model that need to impose the same penalty on them, which is why we do not need to derive the KKT condition in this paper.

To establish the stochastic Lipschitz conditions for this unbounded counting process, another assumption, called the *strongly midpoint log-convex*, for some positive γ should be satisfied, which states for the joint density from the sample $\mathbb{Y} := (Y_1, \ldots, Y_n)^\top \in \mathbb{Z}^n$'s negative log-density of n independent NB responses $\psi(y) := -\log p_{\mathbb{Y}}(y)$ satisfies

$$\psi(x)+\psi(y)-\psi\left(\left\lceil\frac{1}{2}x+\frac{1}{2}y\right\rceil\right)-\psi\left(\left\lfloor\frac{1}{2}x+\frac{1}{2}y\right\rfloor\right)\geq\frac{\gamma}{4}\|x-y\|_2^2,\qquad\forall x,y\in\mathbb{Z}^n.$$

This assumption is a condition that ensures that the suprema of the multiplier empirical processes of n independent responses have sub-exponential concentration phenomena, which can be alternatively checked by the tail inequality for the suprema of the empirical processes corresponding to classes of unbounded functions ([21]).

Theorem 1. *Suppose* $\max_{i\in[n],1\leq k\leq p}|X_{ik}|\leq M_x<\infty$, *the parameter space* Θ *is convex and its diameter* $D_\Theta<\infty$. *If* $\{Y_i\}_{i=1}^n$ *and* $\{Z_i\theta\}_{i\in[n],\theta\in\Theta}$ *are both in the value space* $\mathcal{Y}$ *and* $\mathcal{S}$ *defined as previous, then for any* $\theta\in\Theta$,

$$\begin{aligned}\mathrm{Lip}(\gamma;\theta^*)&=\sup_{\theta\in\Theta/\{\theta^{T}*\}}\left|\frac{\sqrt{n}\mathbb{G}_n\big(\gamma(\theta)-\gamma(\theta^*)\big)}{\|\theta-\theta^*\|_1}\right|\\&\leq\sqrt{n}M_q:=\Big(A_1\sqrt{\log(2p/q_2)}+A_2\sqrt{\log p}+A_3\sqrt{\log(p/q_3)}\Big)\sqrt{\max_{1\leq k\leq p}\sum_{i=1}^n X_{ik}^2}\\&\quad+B\sqrt{\log(2p/q_1)}\sqrt{\Big(\max_{1\leq k\leq p}\sum_{i=1}^n X_{ik}^4\Big)^{1/2}}\vee C\log(2p/q_1)+D\log(p/q_3),\end{aligned}$$

with probability at least $1-q_0$, *where* $q_1,q_2,q_3\in(0,1)$ *satisfy* $q_1+q_2+q_3=q_0$, *and the constants are as follows:*

$$\begin{gathered}A_1=\sqrt{2}F_1,\qquad A_2=32\sqrt{2}M_xF_2D_\Theta,\qquad A_3=\sqrt{2}\big(2(F_1+M_{y,n})\vee F_2M_xD_\Theta\big),\\B=6\sqrt{2\big(w^{(1)}\vee w^{(2)}\big)\Big(\sum_{i=1}^n a(\mu_i,k_i)^4\Big)^{1/2}},\qquad C=12M_x\big(w^{(1)}\vee w^{(2)}\big)\max_{1\leq i\leq n}a(\mu_i,k_i),\\D=8\big(2(F_1+M_{y,n})\vee F_2M_xD_\Theta\big)M_x,\qquad w^{(1)}=\frac{e^{M_{s,n}}}{e^{m_{s,n}}+e^{M_{s,n}}},\\w^{(2)}=\frac{e+e^{M_{s,n}-m_{s,n}}}{1+e^{m_{s,n}-M_{s,n}}}+\frac{1}{1+e^{m_{s,n}-M_{s,n}}},\end{gathered}$$

where $M_{y,n}=\max_{i\in[n]}Y_i$ *is the suprema empirical process.*

It is worthy to note that the $M_{y,n}$ in Theorem 1 is a random process; hence, the bound above is not deterministic. Fortunately, $M_{y,n}$ can use the strongly midpoint log-convex condition to be bounded, which we state in Lemma A3. Theorem 1 combined with Lemma A3 will give the following result as a step more.

Theorem 2. *Assume the conditions are the same as that in Theorem 1, then the stochastic Lipschitz constant has a nonrandom upper bound:*

$$\begin{aligned}\mathrm{Lip}(\gamma;\theta^*)\leq\sqrt{n}M_q':=&\Big(A_1\sqrt{\log(2p/q_2)}+A_2\sqrt{\log p}+2A_3'\big(\log(2n/q_4)+\sqrt{\log(np/q_3)}\big)\Big)\sqrt{\max_{1\leq k\leq p}\sum_{i=1}^n X_{ik}^2}\\&+B\sqrt{\log(2p/q_1)}\sqrt{\Big(\max_{1\leq k\leq p}\sum_{i=1}^n X_{ik}^4\Big)^{1/2}}\vee C\log(2p/q_1)+D\log(p/q_3),\end{aligned}$$

with probability at least $1-q_0$, *where* $q_1,q_2,q_3,q_4\in(0,1)$ *satisfy* $q_1+q_2+q_3+q_4=q_0$, *and*

$$A_3'=2\sqrt{2}\left(F_1+\left(2\gamma\max_{i\in[n]}\left[a(\mu_i,k_i)-\frac{\mu_i}{\log 2}\right]\right)\vee F_2M_xD_\Theta\right).$$

Theorem 1 gives us a different sight of the loss function far more than KKT conditions. However, the stochastic Lipschitz condition above does not compare the estimated and true values directly. We can resolve this issue by using an eigenvalue condition on the design matrix consisting of X_i. Because the design matrix X is fixed, the eigenvalue condition in the next section is reasonable. It is worthy to note that this inequality is an oracle because it involves an unknown empirical process on the right side.

3.2. ℓ_2-Estimation Error Oracle Inequalities RE Conditions

As we said previously, although we use stochastic Lipschitz conditions instead of KKT conditions, the restricted eigenvalue conditions (RE conditions) are still required. We denote by δ_J the vector in $\mathbb{R}^p$ with the same coordinates as v on J and zero coordinates on the complement J^c of J, and $\mathrm{spt}(v) = \{j : v_j \neq 0\}$. We will assume that the minima in (2) can always be obtained in the following setting, but it may not be unique. In general, to bound $\widehat{\theta} - \theta^*$, some conditions on the design matrix $\mathrm{X} \in \mathbb{R}^{n\times p}$ are needed for obtaining abound in terms of the ℓ_2 norm of $\theta - \theta^*$. Here, we will utilize the restricted eigenvalue condition introduced in [22], which says that for some $1 \leq s \leq p$ and $K > 0$,

$$\kappa(s, K) = \min\left\{ \frac{\|\mathrm{X}v\|_2}{\sqrt{n}\|v_J\|_2} : 1 \leq |J| \leq s,\, v \in \mathbb{R}^p/\{0\},\, \|v_{J^c}\|_1 \leq K\|v_J\|_1 \right\} > 0. \tag{3}$$

It should be noted that omitting the weight ω and the sparse restricted set $\|v_{J^c}\|_1 \leq K\|v_J\|_1$ leads to $v^\top\left[\frac{1}{n}\mathrm{X}^\top\mathrm{X}\right]v/v^\top v \geq \kappa^2(s, K)$. Thus, it means that the smallest eigenvalue of the sample covariance matrix $\frac{1}{n}\mathrm{X}^\top\mathrm{X}$ is positive, which is impossible when $p > n$ because $\frac{1}{n}\mathrm{X}^\top\mathrm{X}$ is not full rank. To avoid this problem, ref. [22] consider the restricted eigenvalue condition under the sparse restricted set $\|v_{J^c}\|_1 \leq K\|v_J\|_1$ as a considerable relation in sparse high-dimensional estimation. The restricted eigenvalue is from the restricted strong convexity, which enforces a strong convexity condition for the negative log-likelihood function of linear models under a certain sparse restrict set.

Due to the double penalty, besides the RE condition, we also require another condition similar to the RE condition, the so-called l-restricted isometry constant defined in [23], as follows

$$\sigma^2_{\mathrm{X},l} = \max\left\{ \|\mathrm{X}v\|_2^2/\|v\|_2^2 : v \in \mathbb{R}^p,\, 1 \leq \mathrm{spt}(v) \leq l \right\} \in (0, \infty),$$

which essentially requires the eigenvalue of the sample covariance matrix under every vector with cardinality less than l (l should be no more than n) approximately behaves normally like the low-dimensional case.

With the RE condition and l-restricted isometry constant, and the two theorems we established before, the lasso estimator in (2) can guarantee a good consistent property.

Lemma 1 (see Lemma 3.1 in [23]). *Suppose T_0 is a set of cardinality S. For a vector $h \in \mathbb{R}^p$, we let T_1 be the S largest positions of h outside of T_0. Put $T_{01} = T_0 \cup T_1$, then*

$$\|h\|_2^2 \leq \|h_{T_{01}}\|_2^2 + S^{-1}\|h_{T_0^c}\|_1^2.$$

Theorem 3. *Suppose the condition is the same as that in Theorem A1. Furthermore, assume $p_1 = \mathrm{spt}\left(\theta^{(*1)}\right) \vee \mathrm{spt}\left(\theta^{(*2)}\right) \leq p/2$, and there exists some $K > 1$, $\kappa := \kappa(2p_1, K) > 0$. Let $\lambda = \frac{(K+1)M_q}{n(K-1)}$, then using this λ in (2), with probability at least $1 - q$,*

$$\|\widehat{\theta} - \theta^*\|_2^2 \leq \frac{8p_1M_q'^2K^2}{\kappa^4 n^2 C_\gamma^2 (K-1)^2}\left[2 + K^2 + \frac{2(1 + 2p_1K^2)(n\kappa^2 + 2\sigma^2_{\mathrm{X},p_1})}{n\kappa^2}\right],$$

where M_q, C_γ are defined in Theorems 1 and A1, respectively.

Remark 1. *Compared to the single lasso problem, in which we only have one unknown vectorized parameter, the oracle inequality in Theorem 3 has an extra term* $\frac{2(1+2p_1K^2)(n\kappa^2+2\sigma^2_{\bar{X},p_1})}{n\kappa^2}$.

Remark 2. *From Theorem 3, we know that the* ℓ_2 *convergence rate is minimax optimal, as studied in [14].*

Remark 3. *In this study, we use the lasso estimators of two partial regression coefficients because it is one of the most popular techniques for high-dimensional data. It is worth mentioning that the algorithms and theoretical results could be similarly generalized to other shrinkage estimators, such as the elastic net [7], the adaptive lasso [8], and so on.*

4. Numerical Studies

4.1. Simulations

In this section, we evaluate the finite sample performance of the proposed method. The response is generated from the negative binomial regression model (1) with

$$\mu(x) = \exp\{\theta^{(1)\top}x\}, \text{and } k(x) = \exp\{\theta^{(2)\top}x\},$$

where $\theta^{(1)}$ and $\theta^{(2)}$ are two p-dimensional parameters. The explanatory variables are generated from the multivariate normal distributions with mean vector 0 and $Cov(x_i, x_j) = \rho^{|i-j|}$, where $\rho = 0, 0.5$. The following two examples show the performance of the proposed estimator for the low-dimensional heterogeneous negative binomial regression and the variable selection in the high-dimensional case, respectively. The R package "lbfgs" is required to solve the optimization problem.

Example 1 (Low dimension)**.** *We set* $p = 3$ *and* $n = 100, 200, 400$. *The true parameters are* $\theta^{(1)} = (1, 2, -1)$ *and* $\theta^{(2)} = (-1, 0.5, 1)$, *and their maximum likelihood estimators are denoted as* $\hat{\theta}^{(1)}$ *and* $\hat{\theta}^{(2)}$, *respectively. We compare the estimator* $\hat{\theta}^{(1)}$ *with* $\hat{\theta}^{(1)*}$, *which ignores the heterogeneity of the overdispersion and treats* $k(x)$ *as a constant. Table 1 displays the average squared estimation errors* $\|\hat{\theta} - \theta\|_2^2$ *based on 200 repetitions.*

We can make the following observations from the table. Firstly, the performances of the three estimators become better and better as n increases. Secondly, the estimator $\hat{\theta}^{(1)}$, which estimates the parameter in the mean function $\mu(x)$, performs better than $\hat{\theta}^{(2)}$, which estimates the parameter in the overdispersion function $k(x)$. Last, but the most important, $\hat{\theta}^{(1)*}$ performs much worse than $\hat{\theta}^{(1)}$. For example, the average squared estimation error of $\hat{\theta}^{(1)*}$ is about 5 times of $\hat{\theta}^{(1)}$'s when $n = 100$, and 10 times of $\hat{\theta}^{(1)}$'s when $n = 400$. The comparison between $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(1)*}$ indicates the necessity of considering the heterogeneity of the overdispersion.

Table 1. The average squared estimation errors of the estimators.

n	$\rho = 0$			$\rho = 0.5$		
	$\hat{\theta}^{(1)*}$	$\hat{\theta}^{(1)}$	$\hat{\theta}^{(2)}$	$\hat{\theta}^{(1)*}$	$\hat{\theta}^{(1)}$	$\hat{\theta}^{(2)}$
100	0.1597	0.0335	0.72414	0.1809	0.0397	0.68904
200	0.0862	0.01	0.22149	0.0837	0.0169	0.33048
400	0.05	0.0047	0.08847	0.0619	0.0067	0.15066

Example 2 (High dimension)**.** *The sample sizes are chosen to be* $n = 100, 200, 400$, *with dimension* $p \in (25, 50, 150)$, $(50, 100, 250)$ *and* $(100, 200, 500)$, *respectively. We set* $\theta^{(1)} = (1, 2, -1, 0, \ldots, 0)$ *and* $\theta^{(2)} = (-1, 0.5, 1, 0, \ldots, 0)$. *The unknown tuning parameters* (λ_1, λ_2) *for the penalty functions are chosen by BIC criterion in the simulation. Results over 200 repetitions are reported. We compared the variable selection performance of the proposed method to the previous*

method, which ignores the heterogeneity of the overdispersion and treats $k(x)$ as a constant. For each case, Table 2 reports the number of repetitions that each important explanatory variable is selected in the final model and also the average number of unimportant explanatory variables being selected.

We see from the table that our method performs much better than the previous method that treats $k(x)$ as a constant. Specifically, our method correctly selects important variables more times than the previous method, and it is less likely to select unimportant variables. Furthermore, the variable selection procedure performs better and better as the sample size n increases. When $n = 400$, the important explanatory variables in $\mu(x)$ and $k(x)$ are correctly selected in almost every repetitions. When the dimension p increases, the procedure may select more unimportant explanatory variables, but the average numbers are less than 1.3. The important variables in $k(x)$ are less likely to be selected than the important variables in $\mu(x)$ especially when the sample size is small, as well as the unimportant variables.

Table 2. The results of variable selection.

		Previous Method				**Proposed Method**							
		$\mu(x)$				$\mu(x)$				$k(x)$			
n	**p**	$\theta_1^{(1)}$	$\theta_2^{(1)}$	$\theta_3^{(1)}$	**Other $\theta^{(1)}$s**	$\theta_1^{(1)}$	$\theta_2^{(1)}$	$\theta_3^{(1)}$	**Other $\theta^{(1)}$s**	$\theta_1^{(2)}$	$\theta_2^{(2)}$	$\theta_3^{(2)}$	**Other $\theta^{(2)}$s**
$\rho = 0$													
100	25	173	198	171	2.33	192	200	190	0.37	180	184	180	0.32
	50	164	197	147	2.885	196	200	193	0.52	182	180	188	0.41
	150	136	182	111	2.725	194	194	192	1.02	188	182	186	0.41
200	50	196	200	192	1.435	200	200	200	0.59	200	190	198	0.53
	100	193	200	193	2.05	200	200	200	0.91	196	186	196	0.69
	250	162	198	155	1.5	199	199	198	1.18	198	198	198	0.69
400	100	200	200	200	0.605	200	200	200	0.4	200	198	200	0.55
	200	200	200	199	0.88	200	200	200	0.6	200	200	200	0.51
	500	197	200	198	1.29	200	200	200	1.21	200	200	200	0.61
$\rho = 0.5$													
100	25	183	199	179	2.3	194	198	194	0.41	179	184	180	0.35
	50	172	197	150	2.66	196	196	190	0.63	178	182	180	0.42
	150	134	191	99	2.32	194	196	192	1.01	180	184	182	0.43
200	50	195	200	197	1.48	200	200	198	0.38	196	183	190	0.32
	100	189	200	179	1.52	199	200	198	0.53	194	186	194	0.44
	250	178	200	154	1.39	196	198	196	1.1	196	196	194	0.55
400	100	200	200	200	0.435	200	200	200	0.28	200	199	194	0.34
	200	200	200	199	0.675	200	200	198	0.47	200	198	196	0.36
	500	199	200	194	1.12	200	200	198	1.07	200	198	196	0.56

4.2. A Real Data Example

In this section, we apply the proposed method to the dataset of German health care demand. The data were employed in [24] and could be downloaded on http://qed.econ.queensu.ca/jae/2003-v18.4/riphahn-wambach-million/, accessed on 1 January 2022.

The data contain 27,326 observations on 25 variables, including 2 dependent variables, Docvis (number of doctor visits in the last three months) and Hospvis (number of hospital visits in the last calendar year). For conciseness, we focus on Docvis in this study. We build the HNBR model based on the proposed variable selection procedure and make the standard NBR model a comparison. Define the fitting errors (FE) as $n^{-1}\sum_{i=1}^{n}(y_i - \widehat{y}_i)$, where y_i denotes the raw data of Hospvis, $\hat{y}_i$ is the predicted value, and n is the sample

size. As the data are observed during 1984–1988, 1991, and 1994, we make the analysis for each observed year. Table 3 displays the variable selection results and fitting errors.

We have the following findings from the table. First, the important variables in the NBR are the same as HNBR models in each year, and the estimates are close. Second, the selected variables in $\mu(x)$ are almost the same every year, namely Age, Hsat (health satisfaction), Handper (degree of handicap), and Educ (years of schooling). Moreover, some of these variables still play an essential role in $k(x)$, and $k(x)$ contains no variables other than these. Moreover, we can see that the fitting errors of the HNBR is less than that of the NBR. All of these illustrate the advantage of our method.

Table 3. The variable selection results and the fitting errors (FE) of NBR and HNBR models. The variable Others = {Married, Haupts, Reals, Fachhs, Abitur, Univ, Working, Bluec, Whitec, Self, Beamt, Public, Addon}. Because these variables are not selected in any year, we put them in "Others" for brevity.

Variables	**1984**			**1985**			**1986**			**1987**		
	NBR	**HNBR**		**NBR**	**HNBR**		**NBR**	**HNBR**		**NBR**	**HNBR**	
		$\mu(x)$	$k(x)$		$\mu(x)$	$k(x)$		$\mu(x)$	$k(x)$		$\mu(x)$	$k(x)$
Female	0	0	0	0	0	0	0	0	0	0	0	0
Age	−0.013	−0.013	−0.012	−0.009	−0.01	−0.007	−0.006	−0.006	−0.013	−0.002	−0.001	−0.018
Hsat	−0.205	−0.2	−0.025	−0.244	−0.237	0	−0.188	−0.195	−0.045	−0.158	−0.153	−0.043
Handdum	0	0	0	0	0	0	0	0	0	0	0	0
Handper	0.005	0.005	0.004	0.007	0.006	0.007	0.007	0.007	0	0.007	0.007	0.01
Hhninc	0	0	0	0	0	0	0	0	0	0	0	0
Hhkids	0	0	0	0	0	0	0	0	0	0	0	0
Educ	0	0	−0.027	0	0	−0.064	−0.035	−0.038	0	−0.095	−0.106	−0.003
Others	0	0	0	0	0	0	0	0	0	0	0	0
FE	0.798	0.602		2.203	1.874		0.735	0.581		1.314	1.027	
Variables	**1988**			**1991**			**1994**					
	NBR	**HNBR**		**NBR**	**HNBR**		**NBR**	**HNBR**				
		$\mu(x)$	$k(x)$		$\mu(x)$	$k(x)$		$\mu(x)$	$k(x)$			
Female	0	0	0	0	0	0	0	0	0			
Age	−0.015	−0.014	−0.012	−0.022	−0.019	−0.003	−0.005	−0.004	−0.011			
Hsat	−0.191	−0.187	−0.015	−0.112	−0.132	−0.049	−0.226	−0.224	−0.06			
Handdum	0	0	0	0	0	0	0	0	0			
Handper	0.011	0.009	0.006	0.014	0.013	0	0.007	0.008	0.004			
Hhninc	0	0	0	0	0	0	0	0	0			
Hhkids	0	0	0	0	0	0	0	0	0			
Educ	−0.016	−0.023	−0.002	−0.074	−0.068	0	−0.064	−0.069	0			
Others	0	0	0	0	0	0	0	0	0			
FE	1.144	0.912		1.007	0.787		0.713	0.58				

5. Conclusions and Future Study

We study the high-dimensional heterogeneous overdispersed count data via negative binomial regression models and propose a double ℓ_1-regularized method for simultaneous variable selection and dispersion estimation. Under the restricted eigenvalue conditions, we prove the oracle inequalities with lasso estimators of two partial regression coefficients for the first time, using concentration inequalities of empirical processes. Furthermore, we derive the consistency and convergence rate for the estimators, which are the theoretical guarantees for further statistical inference. Simulation studies and a real example from the German health care demand data indicate that the proposed method works satisfactorily.

There are some limitations of this study. First, we assume that the responses are independent in this work. However, the NB responses are temporal dependent in the time-series data [25]. Thus, weak dependence conditions, including ρ-mixing, m-dependent

types, could be considered in the future. Second, this study focuses little on the statistical inference, such as testing heterogeneous

$$H_0 : \theta^{(2)} = 0 \text{ vs. } H_1 : \theta^{(2)} \neq 0.$$

The issues concerning the hypothesis testing are via the debiased lasso estimator; see [26] and references therein. This will comprise our future research work. Another possible study is the false discovery rate (FDR) control, which aims to identify some small number of statistically significantly nonzero results after obtaining the sparse penalized estimation of the HNBR; see [27,28].

Author Contributions: Conceptualization, S.L. and H.W.; methodology, H.W.; software, S.L.; validation, S.L., H.W. and X.L.; data curation, S.L.; writing—original draft preparation, S.L., H.W. and X.L.; writing—review and editing, S.L. and H.W.; supervision, S.L. and H.W.; funding acquisition, S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 12101056.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in Section 4.2 could be downloaded on http://qed.econ.queensu.ca/jae/2003-v18.4/riphahn-wambach-million/, accessed on 1 January 2022.

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

Appendix A. Proofs

The first step is giving the property of the loss function. From mathematical analysis, we prefer bounded things to unlimited things. Denote ∂_j is the first partial differentiation with respect to s_j. The bounded aspect for y and s gives a nice property for the loss function $\gamma(s, y) = \gamma(z\theta, y)$.

Lemma A1. *We have*

$$\partial_1\gamma(s,y) = -\frac{ye^{s_2}}{e^{s_1}+e^{s_2}} + \frac{e^{s_1+s_2}}{e^{s_1}+e^{s_2}}, \qquad \partial_2\gamma(s,y) = \nu(s,y) + \frac{ye^{s_2}}{e^{s_1}+e^{s_2}}$$

where $\nu(s,y) = e^{s_2}\left(\psi(e^{s_2}) - \psi(y+e^{s_2})\right) + e^{s_2}\log\left(1+e^{s_1-s_2}\right) - \frac{e^{s_1+s_2}}{e^{s_1}+e^{s_2}}$ *satisfying*

$$\sup_{s\in\mathcal{S}, y\in\mathcal{Y}} |\nu(s,y)| \leq F_1, \qquad \sup_{s\neq t\in\mathcal{S}, y\in\mathcal{Y}} \frac{|\nu(s,y)-\nu(t,y)|}{\|s-t\|_\infty} \leq F_2,$$

with $F_1 = M_{y,n}\left(1+e^{-m_{s,n}}\right) + e^{M_{s,n}} + \frac{e^{2M_{s,n}}}{2e^{m_{s,n}}}$, *and*

$$F_2 = 2\left[\left|e^{M_{s,n}}\left(1+\log(M_{y,n}+e^{M_{s,n}}) - \frac{1}{2(M_{y,n}+e^{M_{s,n}})}\right)\right| \vee \left|1-m_{s,n}e^{m_{s,n}}\right|\right] + \left[\frac{e^{2M_{s,n}}}{e^{m_{s,n}}} + \frac{2e^{2M_{s,n}}}{e^{m_{s,n}}+e^{M_{s,n}}}\right] + \frac{3}{2}e^{M_{s,n}}.$$

Proof. We will use the properties of the psi function, the logarithmic derivative of the gamma function, to prove this lemma. Write $\psi(x) = \Gamma'(x)/\Gamma(x)$. For any $s \in \mathcal{S}$, $y \in \mathcal{Y}$, using the Binet 's formula (see p. 18 of [29])

$$\psi(x) = \log x - \int_0^\infty \varphi(t)e^{-tx}\,dt,$$

where $\varphi(t) = 1/(1-e^{-t}) - 1/t$ is strictly increasing on $(0,\infty)$, it gives

$$0 < \psi'(x) = \frac{1}{x} + \int_0^\infty t\varphi(t)e^{-tx}\,dt \le \frac{1}{x} + \int_0^\infty te^{-tx}\,dt = \frac{1}{x} + \frac{1}{x^2}.$$

and $y \ge 0$, we have

$$\begin{aligned}|\nu(s,y)| &= \left|e^{s_2}\big(\psi(e^{s_2}) - \psi(y+e^{s_2})\big) + e^{s_2}\log\big(1+e^{s_1-s_2}\big) - \frac{e^{s_1+s_2}}{e^{s_1}+e^{s_2}}\right| \\ &\le e^{s_2}y\left(\frac{1}{e^{s_2}} + \frac{1}{e^{s_2}}\right) + e^{s_2}e^{s_1-s_2} + \frac{e^{s_1+s_2}}{e^{s_1}+e^{s_2}} \le M_{y,n}\big(1+e^{-m_{s,n}}\big) + e^{M_{s,n}} + \frac{e^{2M_{s,n}}}{2e^{m_{s,n}}}.\end{aligned}$$

Then, the first inequality in the lemma has been verified. On the other hand, by using the fact that (see (2.2) in [30])

$$\frac{1}{2x} < \log(x) - \psi(x) < \frac{1}{x}, \qquad x > 0,$$

for the function $f_1(x) = e^x\psi(e^x)$ and $f_2(x) = e^x\psi(y+e^x)$,

$$\begin{aligned}f_1'(x) = e^x\psi(e^x) + e^{2x}\psi'(e^x) &\le e^x\left(x - \frac{1}{2e^x}\right) + e^{2x}\left(\frac{1}{e^x} + \frac{1}{e^{2x}}\right) = (x+1)e^x + \frac{1}{2}, \\ f_1'(x) = e^x\psi(e^x) - e^{2x}\psi'(e^x) &\ge e^x\psi(e^x) \ge xe^x - 1, \\ f_2'(x) = e^x\psi(y+e^x) + e^{2x}\psi'(y+e^x) &\le e^x\left(1 + \log(y+e^x) - \frac{1}{2(y+e^x)}\right) + 1, \\ f_2'(x) &\ge e^x\left(\log(y+e^x) - \frac{1}{y+e^x}\right) \ge xe^x - 1,\end{aligned}$$

and for any $s \ne t \in \mathcal{S}$, $y \in \mathcal{Y}$, we conclude that

$$|\psi(e^{s_2})e^{s_2} - \psi(e^{t_2})e^{t_2}| = |f_1(s_2) - f_1(t_2)| \le \Big(\Big|(M_{s,n}+1)e^{M_{s,n}} + 1/2\Big| \vee |1 - m_{s,n}e^{m_{s,n}}|\Big)\|s-t\|_\infty,$$

and

$$\begin{aligned}&|\psi(y+e^{s_2})e^{s_2} - \psi(y+e^{t_2})e^{t_2}| = |f_2(s_2) - f_2(t_2)| \\ &\le \left[\left|e^{M_{s,n}}\left(1 + \log(M_{y,n}+e^{M_{s,n}}) - \frac{1}{2(M_{y,n}+e^{M_{s,n}})}\right)\right| \vee |1 - m_{s,n}e^{m_{s,n}}|\right]\|s-t\|_\infty.\end{aligned}$$

In addition, using the median value theorem again, we also have

$$\begin{aligned}&\big|e^{s_2}\log(1+e^{s_1-s_2}) - e^{t_2}\log(1+e^{t_1-t_2})\big| \\ \le{}& \log(1+e^{s_1-s_2})|e^{s_2} - e^{t_2}| + e^{t_2}\big|\log(1+e^{s_1-s_2}) - \log(1+e^{t_1-t_2})\big| \\ \le{}& e^{2M_{s,n}-m_{s,n}}|s_2 - t_2| + e^{M_{s,n}}\frac{1}{1+e^{-(M_{s,n}-m_{s,n})}}|(s_1-s_2) - (t_1-t_2)| \\ \le{}& \left[\frac{e^{2M_{s,n}}}{e^{m_{s,n}}} + \frac{2e^{2M_{s,n}}}{e^{m_{s,n}}+e^{M_{s,n}}}\right]\|s-t\|_\infty,\end{aligned}$$

and

$$\begin{aligned}\left|\frac{e^{s_1+s_2}}{e^{s_1}+e^{s_2}} - \frac{e^{t_1+t_2}}{e^{t_1}+e^{t_2}}\right| &\le e^{s_1}\left|\frac{1}{1+e^{s_1-s_2}} - \frac{1}{1+e^{t_1-t_2}}\right| + \frac{1}{1+e^{t_1-t_2}}|e^{s_1} - e^{t_1}| \\ &\le e^{M_{s,n}}\frac{1}{4}|(s_1-s_2) - (t_1-t_2)| + 1 \times e^{M_{s,n}}|s_1 - t_1| \le \frac{3}{2}e^{M_{s,n}}\|s-t\|_\infty,\end{aligned}$$

where the fact used is that $f_3(x) = 1/(1+e^x)$ satisfies $|f_3'(x)| = 1/(e^x + e^{-x} + 2) \le 1/4$. Because

$$|\partial_2\gamma(s,y) - \partial_2\gamma(t,y)| \le \left|\psi(e^{s_2})e^{s_2} - \psi(e^{t_2})e^{t_2}\right| + \left|\psi(y+e^{s_2})e^{s_2} - \psi(y+e^{t_2})e^{t_2}\right| + \left|e^{s_2}\log(1+e^{s_1-s_2}) - e^{t_2}\log(1+e^{t_1-t_2})\right| + \left|\frac{e^{s_1+s_2}}{e^{s_1}+e^{s_2}} - \frac{e^{t_1+t_2}}{e^{t_1}+e^{t_2}}\right|,$$

we can conclude the second inequality in the lemma. □

The Lemma separates the partial derivative of γ into two parts: the first part is the linear about the response variable y (say $-ye^{s_2}/(e^{s_1}+e^{s_2})$, $e^{s_1+s_2}/(e^{s_1}+e^{s_2})$, and $ye^{s_2}/(e^{s_1}+e^{s_2})$), the second part is other complicated functions (not linear function) about y. The first part is relatively easy to analyze because the following concentration inequality gives a measure of dispersion about the weighted summation of negative binomial variables. This concentration inequality is a special case for the weighted summation of a series of random variables, which can be proved by sub-exponential concentration results in Proposition 4.2 in [31].

Lemma A2. *Suppose* $\{Y_i\}_{i=1}^n$ *are independently distributed as* $\mathrm{NB}(\mu_i, k_i)$. *Then, for any nonrandom weights* $w = (w_1, \cdots, w_n)^\top \in \mathbb{R}^n$ *independent with* $\{Y_i\}_{i=1}^n$ *and* $t \ge 0$,

$$\mathrm{P}\left(\left|\sum_{i=1}^n w_i(Y_i - \mathrm{E}Y_i)\right| \ge t\right) \le 2\exp\left\{-\frac{1}{4}\left(\frac{t^2}{2\sum_{i=1}^n w_i^2 a^2(\mu_i,k_i)} \wedge \frac{t}{\max\limits_{1\le i\le n}|w_i|a(\mu_i,k_i)}\right)\right\},$$

where $q_i := \frac{\mu_i}{k_i+\mu_i} \in (0,1)$ *and* $a(\mu,k) := \left[\log\frac{1-(1-q)/\sqrt[k]{2}}{q}\right]^{-1} + \frac{\mu}{\log 2}$.

Proof. We will use the sub-exponential norm. The moment-generating function (MGF) for Y_i is

$$\mathrm{E}e^{sY_i} = \left(\frac{1-q_i}{1-q_ie^s}\right)^{k_i}.$$

Then, by letting $\mathrm{E}\exp(|Y_i|/t) \le 2$, we have

$$2 \ge \mathrm{E}\exp(|Y_i|/t) = \mathrm{E}\exp(Y_i/t) = \left(\frac{1-q_i}{1-q_ie^{1/t}}\right)^{k_i},$$

which implies the sub-exponential norm for Y_i is

$$\|Y_i\|_{\psi_1} = \inf\{t>0 : \mathrm{E}\exp(|Y_i|/t) \le 2\} = \left[\log\frac{1-(1-q_i)/\sqrt[k_i]{2}}{q_i}\right]^{-1}.$$

Using the definition of a_i, from Proposition 4.2 in [31], we can immediately obtain the result in the Lemma. □

It should be noted that $a_i = a(\mu_i, k_i)$ naturally has a lower and upper bound for any $i \in [n]$ because μ_i and k_i are both bounded between $e^{m_{s,n}}$ and $e^{M_{s,n}}$.

Note that Y_i is an unbounded random variable; the next step is to find a probabilistic bound for $M_{y,n} = \max_{i\in[n]} Y_i$. We will cite an important lemma for this type of problem. We say a distribution P_γ is strongly discrete log-concave with $\gamma > 0$ if its density is strongly midpoint log-convex with the same $\gamma > 0$.

Lemma A3 (Concentration for strongly log-concave discrete distributions)**.** *Let P_γ be any strongly log-concave discrete distribution index by $\gamma > 0$ on $\mathbb{Z}^n$. Then, for any function $f : \mathbb{R}^n \to \mathbb{R}$ that is L-Lipschitz with respect to Euclidean norm, we have for $X \sim P_\gamma$,*

$$\mathrm{P}_{P_\gamma}\left(|f(X) - \mathrm{E}f(X)| \geq t\right) \leq 2\exp\left\{-\frac{\gamma t^2}{4L^2}\right\}$$

for any $t > 0$.

Lemma A4. *The maximal of the response $M_{y,n} = \max_{i\in[n]} Y_i$ has the concentration*

$$\mathrm{P}\left\{M_{y,n} - \left(2\max_{i\in[n]}\left[a(\mu_i, k_i) - \frac{\mu_i}{\log 2}\right]\left[\log(2n) + \sqrt{2\log(2n)}\right] + \max_{i\in[n]}\mu_i\right) > t\right\} \leq e^{-\gamma t^2/4}$$

for any $t > 0$.

Proof. For the upper bound of expectation, we first note that $Y_i - \mathrm{E}Y_i \sim \mathrm{subE}(2\|Y_i\|_{\psi_1})$ with $\|Y_i\|_{\psi_1}$ has calculated in Lemma A2, then we have $Y_i - \mathrm{E}Y_i \sim \mathrm{sub}\,\Gamma(4\|Y_i\|^2_{\psi_1}, 2\|Y_i\|_{\psi_1})$ by Example 5.3 in [31], which further gives

$$\begin{aligned}
\mathrm{E}M_{y,n} &\leq \mathrm{E}\max_{i\in[n]}(Y_i - \mathrm{E}Y_i) + \max_{i\in[n]}\mathrm{E}Y_i \\
&\leq \left(2\cdot\max_{i\in[n]} 4\|Y_i\|^2_{\psi_1}\cdot\log(2n)\right)^{\frac{1}{2}} + \max_{i\in[n]} 2\|Y_i\|_{\psi_1}\cdot\log(2n) + \max_{i\in[n]}\mu_i \\
&= 2\max_{i\in[n]}\|Y_i\|_{\psi_1}\left[\log(2n) + \sqrt{2\log(2n)}\right] + \max_{i\in[n]}\mu_i,
\end{aligned}$$

where the second $\leq$ is by Corollary 7.3 in [31] and the bound in the lemma comes from the explicit expression in Lemma A2.

By implementing Lemma A3, it remains that we need to verify that $\mathbb{Y} := (Y_1, \ldots, Y_n)^\top \in \mathbb{Z}^n$ belongs to some strongly log-concave discrete distribution P_γ with the specifying $\gamma > 0$ after we take $f : (x_1, \ldots, x_n) \mapsto \max_{i\in[n]} x_i$ which is 1-Lipschitz. By the definition, the derivative of log-density for $y := (y_1, \ldots, y_n)^\top$ is

$$\psi'(y_i) := \left.\frac{\partial \log p(y)}{\partial y}\right|_{y_i} = \log\frac{\Gamma(k_i + y_i)}{\Gamma(1 + y_i)} - y_i\log(k_i + \mu_i),$$

then the Taylor expansion gives

$$\begin{aligned}
\psi(y) &= \psi\left(\left[\frac{1}{2}x + \frac{1}{2}y\right]\right) + \frac{1}{2}\psi'\left(\left[\frac{1}{2}x + \frac{1}{2}y\right]\right)(y - x) + \frac{1}{8}(y - x)^2\psi''(a_1), \\
\psi(x) &= \psi\left(\left\lfloor\frac{1}{2}x + \frac{1}{2}y\right\rfloor\right) + \frac{1}{2}\psi'\left(\left\lfloor\frac{1}{2}x + \frac{1}{2}y\right\rfloor\right)(x - y) + \frac{1}{8}(y - x)^2\psi''(a_2)
\end{aligned}$$

where $a_1 = t_1 y + (1 - t_1)(x + y)/2$, $a_2 = t_2 y + (1 - t_1)(x + y)/2$ with $t_1, t_2 \in [0, 1]$. Define the difference function

$$\Delta(x, y) := \frac{x - y}{4}\left[\psi'\left(\left[\frac{1}{2}x + \frac{1}{2}y\right]\right) - \psi'\left(\left\lfloor\frac{1}{2}x + \frac{1}{2}y\right\rfloor\right)\right] + \frac{\psi''(a_1) + \psi''(a_2)}{16}(y - x)^2,$$

the Taylor expression above immediately implies

$$\Delta(x, y) \geq |x - y|^2\left\{\frac{\psi''(a_1) + \psi''(a_2)}{16} - \sup_{x\neq y; x, y\in\mathbb{Z}^n}\frac{|[\psi'(\lfloor(x + y)/2\rfloor) - \psi'(\lceil(x + y)/2\rceil)]|}{4|x - y|}\right\}.$$

Let

$$C_\psi := \sup_{x\neq y; x,y\in\mathbb{Z}^n} \frac{|[\psi'(\lfloor (x+y)/2\rfloor) - \psi'(\lceil (x+y)/2\rceil)]|}{4|x-y|}$$
$$= \sup_{x\neq y; x,y\in\mathbb{Z}^n} \left| \left[\log \frac{\Gamma(k_i+\lfloor (x+y)/2\rfloor)\Gamma(\lceil (x+y)/2\rceil+1)}{\Gamma(k_i+\lceil (x+y)/2\rceil)\Gamma(\lfloor (x+y)/2\rfloor+1)} - \frac{(\lfloor (x+y)/2\rfloor - \lceil (x+y)/2\rceil)}{\log^{-1}(k_i+\mu_i)} \right] \right| /4|x-y|,$$

and it is not hard to see $C_\psi \approx \frac{|\log(k_i+\mu_i)|}{4}$ or 0. Besides,

$$\psi''(y) := \frac{\partial^2 \log p(y)}{\partial y^2}\bigg|_{y=y_i} = \frac{d}{dy_i} \log \frac{\Gamma(\theta+y_i)}{\Gamma(y_i+1)} = \sum_{m=1}^{\infty}\left(\frac{1}{m+1} - \frac{1}{m+k_i+y_i}\right) - \sum_{m=1}^{\infty}\left(\frac{1}{m+1} - \frac{1}{m+y_i+1}\right)$$
$$= \sum_{m=1}^{\infty}\left(\frac{1}{m+y_i+1} - \frac{1}{m+k_i+y_i}\right) \geq \inf_{y_i\in\mathbb{Z}} \sum_{m=1}^{\infty}\left(\frac{1}{m+y_i+1} - \frac{1}{m+k_i+y_i}\right) = C_{\psi''}.$$

Now, we have obtained

$$\Delta(x,y) \geq |x-y|^2\left\{\frac{\psi''(a_1)+\psi''(a_2)}{16} - C_\psi\right\} \geq |x-y|^2\left(\frac{C_{\psi''}}{8} - C_\psi\right)$$

which gives $\gamma =: \frac{C_{\psi''}}{8} - C_\psi > 0$ from the strong log-concave assumption for $\mathbb{Y}$, if $C_\psi \approx \frac{|\log(k_i+\mu_i)|}{4}$ is small. Hence, we can conclude from Lemma A3 and the upper bound of $\mathrm{E}M_{y,n}$

$$\mathrm{P}\left\{M_{y,n} - \left(2\max_{i\in[n]}\left[a(\mu_i,k_i) - \frac{\mu_i}{\log 2}\right]\left[\log(2n) + \sqrt{2\log(2n)}\right] + \max_{i\in[n]}\mu_i\right) > t\right\}$$
$$\leq \mathrm{P}(M_{y,n} - \mathrm{E}M_{y,n} > t) \leq e^{-\gamma t^2/4}$$

which is exactly the result in the lemma. □

Remark A1. *For $M_{y,n}$, it is distributed as sub-Gumbel, which is rarely studied by research. Another way to deal with using the extreme value theory (EVT) technique, we note that for any $t \in \mathbb{R}$*

$$\mathrm{P}(M_{y,n} - \mathrm{E}M_{y,n} > t) = 1 - \prod_{i=1}^{n}\mathrm{P}\big(Y_i \leq t + \mathrm{E}M_{y,n}\big)$$
$$= 1 - \prod_{i=1}^{n}\left[1 - \exp\left\{-\frac{1}{4}\Big(\frac{(t+\mathrm{E}M_{y,n}-\mu_i)^2}{2a_i^2} \wedge \frac{t+\mathrm{E}M_{y,n}-\mu_i}{a_i}\Big)\right\}\right].$$

If Y_i is i.i.d., then in asymptotic sense,

$$\mathrm{P}(M_{y,n} - \mathrm{E}M_{y,n} > t) = 1 - \big[1 - \mathrm{P}\big(Y_1 > t + \mathrm{E}M_{y,n}\big)\big]^n$$
$$\sim 1 - \exp\big\{-n\mathrm{P}\big(Y_1 > t + \mathrm{E}M_{y,n}\big)\big\} + o(1)$$
$$\sim 1 - \exp\left\{-n\exp\Big(-\frac{1}{4}\frac{(t+\mathrm{E}M_{y,n}-\mu_1)^2}{2a_1^2} \wedge \frac{t+\mathrm{E}M_{y,n}-\mu_1}{a_1}\Big)\right\}.$$

Unfortunately, this technique cannot be used in the above lemma because: (i) we need non-asymptotic version inequality instead of a vague expression with $n \to \infty$ and (ii) $\{Y_i\}$ is not an i.i.d. series, and then EVT theory will not be easily used in this particular setting. Hence, we adopt a discrete technique which has been used in [32] and fully illustrated in [14].

The stochastic Lipschitz conditions are established by using the properties of $\partial_1\gamma(s,y)$ and $\partial_2\gamma(s,y)$. As we said before, they are divided into two parts. The linear parts in them can be solved by the concentration inequality for NB variables given in Lemma A2, but the

non-linear part $\nu(s,y)$ needs some more advanced tools regarding the empirical process. They are given as the following lemmas.

Lemma A5 (The (3.12) in [33]). *Suppose $X_1(\omega),\cdots,X_n(\omega)\in\mathbb{R}$ are zero-mean independent stochastic processes indexed by $\omega\in\Omega$. If there exist M_0 and S_0 satisfying $|X_i(\omega)|\le M_0$ and $\sum_{i=1}^n \mathrm{var}\big(X_i(\omega)\big)\le S_0^2$ for all $\omega\in\Omega$. Denote $S_n=\sup_{\omega\in\Omega}\big|\sum_{i=1}^n X_i(\omega)\big|$, then for any $t>0$,*

$$\mathrm{P}\big(S_n\ge 2\mathrm{E}S_n+S_0\sqrt{2t}+4M_0t\big)\le e^{-t}.$$

A map $\phi:\mathbb{R}\to\mathbb{R}$ is called a contraction if $|\phi(s)-\phi(t)|\le|s-t|$ for all $s,t\in\mathbb{R}$. In addition, in the following lemmas, $\varepsilon_1,\cdots,\varepsilon_n$ are always i.i.d. Rademacher variables.

Lemma A6 (Theorem 2.2 in [34]). *Let $\mathcal{T}\subseteq V^n$ be a bounded set and $f_1,\cdots,f_n$ be functions $V\to\mathbb{R}$ such that f_i is (M_i,ℓ_∞)-Lipschitz with $f_i(0)=0$. For $j=1,\cdots,k\in\mathbb{N}$, let $T_j=\{(t_{1j},\cdots,t_{nj}):(t_1,\cdots,t_n)\in\mathcal{T}\}\subseteq\mathbb{R}^n$. Then,*

$$\mathrm{E}\sup_{t\in\mathcal{T}}\left|\sum_{i=1}^n\varepsilon_i f_i(t_i)\right|\le\beta_k\sum_{j=1}^k\mathrm{E}\sup_{s\in T_j}\left|\sum_{i=1}^n\varepsilon_i M_i s_i\right|,$$

where β_k is a universal constant that can be set no greater than $3^k+3^{k-1}-2^k$.

Lemma A7 (Theorem 4.12 in [35]). *Let $F:\mathbb{R}_+\to\mathbb{R}_+$ be convex and increasing. Let further $\phi_i:\mathbb{R}\to\mathbb{R}, i\le n$ be contractions such that $\phi_i(0)=0$. Then, for any bounded subset $\mathcal{T}$ in $\mathbb{R}^n$,*

$$\mathrm{E}F\left(\frac{1}{2}\sup_{\mathcal{T}}\left|\sum_{i=1}^n\varepsilon_i\phi_i(t_i)\right|\right)\le\mathrm{E}F\left(\sup_{\mathcal{T}}\left|\sum_{i=1}^n\varepsilon_i t_i\right|\right).$$

Lemma A8 (Lemma 5.2 in [36]). *Let $\mathcal{A}$ be some finite subset of $\mathbb{R}^n$, let $R=\sup_{a\in\mathcal{A}}\big[\sum_{i=1}^n a_i^2\big]^{1/2}$, then*

$$\mathrm{E}\left[\sup_{a\in\mathcal{A}}\sum_{i=1}^n\varepsilon_i a_i\right]\le R\sqrt{2\log\big(\mathrm{card}(\mathcal{A})\big)}.$$

With the assistance of these powerful tools, we can establish the stochastic Lipschitz condition as follows, which is one of the most important points in this article for establishing the oracle inequality of the ℓ_2 distance between the estimated value $\widehat{\theta}$ and the real value θ^*.

The proof of Theorem 1. Denote $c_i=Z_i\theta^*$. For $\theta\in\Theta$, denote $t_i=Z_i(\theta-\theta^*)=Z_i\theta-c_i$. We also define the map $\bar{\pi}_j:(x_1,\cdots,x_p)^\top\mapsto(x_1,\cdots,x_j,0,\cdots,0)^\top$ and the function

$$\varphi_{ij}(s)=\begin{cases}\dfrac{\gamma(c_i+\bar{\pi}_j s,Y_i)-\gamma(c_i+\bar{\pi}_{j-1}s,Y_i)}{s_j}-\partial_j\gamma(c_i,Y_i), & \text{if } s_j\neq 0;\\ \partial_j\gamma(c_i+\bar{\pi}_{j-1}s,Y_i)-\partial_j\gamma(c_i,Y_i), & \text{if } s_j=0.\end{cases}$$

Thus, $\varphi_{ij}:\mathbb{R}^2\to\mathbb{R}$ is a real-value function for $i=1,\cdots,n$, $j=1,2$. Then, it is easy to check that

$$\gamma(Z_i\theta,Y_i)-\gamma(Z_i\theta^*,Y_i)=\sum_{j=1}^2\big(\partial_j\gamma(c_i,Y_i)+\varphi_{ij}(t_i)\big)t_{ij},$$

and $n\mathbb{P}_n\big(\gamma(\theta)-\gamma(\theta^*)\big)=\sum_{i=1}^n\sum_{j=1}^2\big(\partial_j\gamma(c_i,Y_i)+\varphi_{ij}(t_i)\big)X_i^\top\big(\theta^{(j)}-\theta^{*(j)}\big)$ in turn. It gives

$$\begin{aligned}\sqrt{n}\mathbb{G}_n\big(\gamma(\theta)-\gamma(\theta^*)\big)=&\sum_{i=1}^n\sum_{j=1}^2\big(\partial_j\gamma(c_i,Y_i)-\mathrm{E}\partial_j\gamma(c_i,Y_i)\big)X_i^\top\big(\theta^{(j)}-\theta^{*(j)}\big)\\&+\sum_{i=1}^n\sum_{j=1}^2\big(\varphi_{ij}(t_i)-\mathrm{E}\varphi_{ij}(t_i)\big)X_i^\top\big(\theta^{(j)}-\theta^{*(j)}\big).\end{aligned}$$

First, we would like to give the explicit formula for φ_{i1} and obtain an upper bound as well as a Lipschitz parameter for φ_{i2}. Denote $h_i(\cdot)=\gamma(\cdot,Y_i)$, then

$$\varphi_{ij}(s)=\int_0^1\big(\partial_jh_i(c_i+\bar{\pi}_{j-1}s+s_jue_j)-\partial_jh_i(c_i)\big)\,du,$$

where e_j is the j-th basis vector of $\mathbb{R}^2$. Hence, for $j=1$,

$$\begin{aligned}\varphi_{i1}(s)&=-Y_i\int_0^1\left[\frac{e^{s_2}}{e^{s_1+u}+e^{s_2}}-\frac{e^{s_2}}{e^{s_1}+e^{s_2}}\right]du+\int_0^1\left[\frac{e^{s_1+s_2+u}}{e^{s_1+u}+e^{s_2}}-\frac{e^{s_1+s_2}}{e^{s_1}+e^{s_2}}\right]du\\&=\left[\log\frac{e^{s_1+1}+e^{s_2}}{e^{s_1}+e^{s_2}}-\frac{e^{s_1}}{e^{s_1}+e^{s_2}}\right]Y_i+C_1(s),\end{aligned}$$

in which $C_1(s)$ is a function only related to s and free of Y and the index i. Using Lemma A1, for $j=2$, write $F_3=F_1+M_{y,n}$,

$$|\varphi_{i2}(s)|\le\int_0^1\big|\partial_2h_i(c_i+\bar{\pi}_1s+s_2ue_2)-\partial_2h_i(u)\big|\,du\le 2F_3,$$

and

$$\begin{aligned}\big|\varphi_{i2}(s)-\varphi_{i2}(t)\big|&\le\int_0^1\big|\partial_2h_i(c_i+\bar{\pi}_1s+s_2ue_2)-\partial_2h_i(c_i+\bar{\pi}_1t+t_2ue_2)\big|\,du\\&\le\int_0^1F_2\|\bar{\pi}_1(s-t)+(s_2-t_2)ue_2\|_\infty\,du\le F_2\|s-t\|_\infty.\end{aligned}$$

This implies φ_{i2} is (F_2,ℓ_∞) Lipschitz. In particular, letting $s=Z_i(\theta-\theta^*)$ and $t=0$,

$$\big|\varphi_{i2}\big(Z_i(\theta-\theta^*)\big)\big|\le\|Z_i(\theta-\theta^*)\|_\infty\le F_2M_xD_\Theta.$$

Hence, we obtain an upper bound for φ_{i2} that

$$\big|\varphi_{i2}\big(Z_i(\theta-\theta^*)\big)\big|\le 2F_3\vee F_2M_xD_\Theta:=M_1 \tag{A1}$$

Now, for $k=1,\cdots,p$, define

$$\xi_{ik}(\theta):=\big(\varphi_{i2}(t_i)-\mathrm{E}\varphi_{i2}(t_i)\big)X_{ik},\qquad S_k=\sup_{\theta\in\Theta}\left|\sum_{i=1}^n\xi_{ik}(\theta)\right|.$$

Then, we can approach the final conclusion in the theorem by

$$\begin{aligned}\sup_{\theta\in\Theta/\{\theta^*\}}\left|\frac{\sqrt{n}\mathbb{G}_n\big(\gamma(\theta)-\gamma(\theta^*)\big)}{\|\theta-\theta^*\|_1}\right| &\le \max_{1\le k\le p}\left|\sum_{i=1}^n\big(\partial_1\gamma(c_i,Y_i)-\mathrm{E}\partial_1\gamma(c_i,Y_i)\big)X_{ik}\right| \\ &+ \sup_{\theta\in\Theta/\{\theta^*\}}\max_{1\le k\le p}\left|\sum_{i=1}^n\big(\varphi_{i1}(t_i)-\mathrm{E}\varphi_{i1}(t_i)\big)X_{ik}\right| \\ &+ \max_{1\le k\le p}\left|\sum_{i=1}^n\frac{e^{c_{i2}}}{e^{c_{i1}}+e^{c_{i2}}}\big(Y_i-\mathrm{E}Y_i\big)X_{ik}\right| \\ &+ \max_{1\le k\le p}\left|\sum_{i=1}^n\big(\nu(c_i,Y_i)-\mathrm{E}\nu(c_i,Y_i)\big)X_{ik}\right| + \sup_{\theta\in\Theta/\{\theta^*\}}\max_{1\le k\le p}S_k.\end{aligned} \tag{A2}$$

We will tickle with (A2) term by term.

(i). The first three terms in (A2):

We will use concentration inequality to deal with these terms. For any $1\le k\le p$ and $t\ge 0$, by Lemma A2 and Cauchy–Schwartz inequality,

$$\begin{aligned}&\mathrm{P}\left(\left|\sum_{i=1}^n\big(\partial_1\gamma(c_i,Y_i)-\mathrm{E}\partial_1\gamma(c_i,Y_i)\big)X_{ik}\right|\ge t\right)=\mathrm{P}\left(\left|\frac{e^{c_{i2}}}{e^{c_{i1}}+e^{c_{i2}}}X_{ik}\big(Y_i-\mathrm{E}Y_i\big)\right|\ge t\right)\\ &\le 2\exp\left\{-\frac{1}{4}\left(\frac{t^2}{2\sum_{i=1}^n(w_i^{(1)})^2X_{ik}^2a_i^2}\wedge\frac{t}{\max\limits_{1\le i\le n}|w_i^{(1)}X_{ik}|a_i}\right)\right\}\\ &\le 2\exp\left\{-\frac{1}{4}\left(\frac{t^2}{2\sqrt{\sum_{i=1}^n(w_i^{(1)})^4a_i^4}\max\limits_{1\le k\le p}\sqrt{\sum_{i=1}^nX_{ik}^4}}\wedge\frac{t}{M_x\max\limits_{1\le i\le n}|w_i^{(1)}|a_i}\right)\right\},\end{aligned}$$

where $w_i^{(1)}=e^{c_{i2}}/(e^{c_{i1}}+e^{c_{i2}})$ and $a_i=a(\mu_i,k_i)$ is defined in Lemma A2; they are both determined and free of θ and the index k. Hence,

$$\begin{aligned}&\mathrm{P}\left(\max_{1\le k\le p}\left|\sum_{i=1}^n\big(\partial_1\gamma(c_i,Y_i)-\mathrm{E}\partial_1\gamma(c_i,Y_i)\big)X_{ik}\right|\ge t\right)\\ &\qquad\le 2p\exp\left\{-\frac{1}{4}\left(\frac{t^2}{2\sqrt{\sum_{i=1}^n(w_i^{(1)})^4a_i^4}\max\limits_{1\le k\le p}\sqrt{\sum_{i=1}^nX_{ik}^4}}\wedge\frac{t}{M_x\max\limits_{1\le i\le n}|w_i^{(1)}|a_i}\right)\right\}.\end{aligned}$$

By letting the right side of the above display be $q_1\in(0,1)$, we can obtain

$$\begin{aligned}&\mathrm{P}\left(\max_{1\le k\le p}\left|\sum_{i=1}^n\big(\partial_1\gamma(c_i,Y_i)-\mathrm{E}\partial_1\gamma(c_i,Y_i)\big)X_{ik}\right|\right.\\ &\left.\ge 2\sqrt{2\Big(\sum_{i=1}^n(w_i^{(1)})^4a_i^4\Big)^{1/2}\Big(\max_{1\le k\le p}\sum_{i=1}^nX_{ik}^4\Big)^{1/2}\log(2p/q_1)}\vee 4M_x\max_{1\le i\le n}|w_i^{(1)}|a_i\log(2p/q_1)\right)\le q_1.\end{aligned}$$

Exactly the same, we can obtain for any $q_3\in(0,1)$, regarding to the third term,

$$\begin{aligned}&\mathrm{P}\left(\max_{1\le k\le p}\left|\sum_{i=1}^n\frac{e^{c_{i2}}}{e^{c_{i1}}+e^{c_{i2}}}\big(Y_i-\mathrm{E}Y_i\big)X_{ik}\right|\right.\\ &\left.\ge 2\sqrt{2\Big(\sum_{i=1}^n(w_i^{(1)})^4a_i^4\Big)^{1/2}\Big(\max_{1\le k\le p}\sum_{i=1}^nX_{ik}^4\Big)^{1/2}\log(2p/q_3)}\vee 4M_x\max_{1\le i\le n}|w_i^{(1)}|a_i\log(2p/q_3)\right)\le q_3.\end{aligned}$$

The situation is slightly different for the second term. Indeed,

$$\mathrm{P}\left(\Big|\sum_{i=1}^{n}\left(\varphi_{i1}(t_i)-\mathrm{E}\varphi_{i1}(t_i)\right)X_{ik}\Big|\geq t\right)=\mathrm{P}\left(\left|\sum_{i=1}^{n}\left[\log\frac{e^{t_{i1}+1}+e^{t_{i2}}}{e^{t_{i1}}+e^{t_{i2}}}-\frac{e^{t_{i1}}}{e^{t_{i1}}+e^{t_{i2}}}\right]X_{ik}(Y_i-\mathrm{E}Y_i)\right|\geq t\right)$$
$$:=\mathrm{P}\left(\Big|\sum_{i=1}^{n}w_i^{(2)}(\theta)X_{ik}\big(Y_i-\mathrm{E}Y_i\big)\Big|\geq t\right)$$

Because t_i is a function of θ, so as the weights $w_i^{(2)}(\theta)$, we cannot use the exact same method as previously. However, because Θ is convex, we have $\{t_i\}_{i=1}^n\subseteq\mathcal{S}$. Then, it only needs to note that,

$$\left|w_i^{(2)}(\theta)\right|=\left|\log\frac{e^{t_{i1}+1}+e^{t_{i2}}}{e^{t_{i1}}+e^{t_{i2}}}-\frac{e^{t_{i1}}}{e^{t_{i1}}+e^{t_{i2}}}\right|\leq\log\frac{e+e^{M_{s,n}-m_{s,n}}}{1+e^{m_{s,n}-M_{s,n}}}+\frac{1}{1+e^{m_{s,n}-M_{s,n}}}:=w^{(2)},$$

which gives

$$\mathrm{P}\Bigg(\max_{1\leq k\leq p}\Big|\sum_{i=1}^{n}\left(\varphi_{i1}(t_i)-\mathrm{E}\varphi_{i1}(t_i)\right)X_{ik}\Big|$$
$$\geq 2\sqrt{2\sqrt{n}w^{(2)2}\big(\sum_{i=1}^{n}a_i^4\big)^{1/2}\big(\max_{1\leq k\leq p}\sum_{i=1}^{n}X_{ik}^4\big)^{1/2}\log(2p/q_2)}\vee 4M_xw^{(2)}\max_{1\leq i\leq n}|a_i|\log(2p/q_2)\Bigg)\leq q_2.$$

for any $\theta\in\Theta$ and $q_2\in(0,1)$.

(ii). The fourth term in (A2):

From Lemma A1, we know that $|\nu(c_i,Y_i)|\leq F_1$. Thus, simply by Hoeffding inequality (see Corollary 2.1 (b) in [31]), for any $t\geq 0$ and $1\leq k\leq p$,

$$\mathrm{P}\left(\Big|\sum_{i=1}^{n}\left(\nu(c_i,Y_i)-\mathrm{E}\nu(c_i,Y_i)\right)X_{ik}\Big|\geq t\right)\leq 2\exp\left\{-\frac{t^2}{2F_1^2\sum_{i=1}^{n}X_{ik}^2}\right\}\leq 2\exp\left\{-\frac{t^2}{2F_1^2\max\limits_{1\leq k\leq p}\sum_{i=1}^{n}X_{ik}^2}\right\}.$$

For arbitrary $q_4\in(0,1)$, let $t=F_1\sqrt{2\log(2p/q_4)\max_{1\leq k\leq p}\sum_{i=1}^{n}X_{ik}^2}$, we obtain

$$\mathrm{P}\left(\max_{1\leq k\leq p}\Big|\sum_{i=1}^{n}\left(\nu(c_i,Y_i)-\mathrm{E}\nu(c_i,Y_i)\right)X_{ik}\Big|\geq F_1\sqrt{2\log(2p/q_4)\max_{1\leq k\leq p}\sum_{i=1}^{n}X_{ik}^2}\right)\leq q_4.$$

(iii). The last term in (A2):

For any $i=1,\cdots,n$ and $k=1,\cdots,p$, by (A1), $|\xi_{ik}(\theta)|\leq 2M_1M_x:=M_0$. In addition, for any $\theta\in\Theta$, (A1) also implies

$$\sum_{i=1}^{n}\mathrm{var}\left(\xi_{ik}(\theta)\right)=\sum_{i=1}^{n}\mathrm{E}\big(\varphi_{i2}(t_i)X_{ik}\big)^2\leq A_1^2\sum_{i=1}^{n}X_{ik}^2\leq A_1^2\max_{1\leq k\leq p}\sum_{i=1}^{n}X_{ik}^2:=S_0^2$$

Therefore, from Lemma A5, it follows that

$$\mathrm{P}\Big(S_k\geq 2\mathrm{E}S_k+S_0\sqrt{2t}+4M_0t\Big)\leq e^{-t}. \tag{A3}$$

Thus, the last task is giving an upper bound for $\mathrm{E}S_k$. Note that $\mathrm{E}\xi_{ik}(\theta)=0$, by symmetrization,

$$\mathrm{E}S_k=\mathrm{E}\sup_{\theta\in\Theta}\Big|\sum_{i=1}^{n}\left(\varphi_{i2}(t_i)-\mathrm{E}\varphi_{i2}(t_i)\right)X_{ik}\Big|\leq 2\mathrm{E}\sup_{\theta\in\Theta}\Big|\sum_{i=1}^{n}\varepsilon_i\varphi_{i2}(t_i)X_{ik}\Big|=2\mathrm{E}\sup_{t\in\mathcal{T}}\Big|\sum_{i=1}^{n}\varepsilon_i\varphi_{i2}(t_i)X_{ik}\Big|,$$

where $\mathcal{T} = \{t_i = Z_i(\theta - \theta^*) : \theta \in \Theta, i = 1, \cdots, n\}$, and $\varepsilon_1, \cdots, \varepsilon_n$ are i.i.d. Rademacher variables independent of $Y_1, \cdots, Y_n$. Here, using the fact $\varphi_{i2}(\cdot)X_{ik}$ is (M_xF_2, ℓ_∞)-Lipschitz and Lemmas A6–A8,

$$\begin{aligned}\mathrm{E}\sup_{t\in\mathcal{T}}\left|\sum_{i=1}^{n}\varepsilon_i\varphi_{i2}(t_i)X_{ik}\right| &\le 8M_xF_2\sum_{j=1}^{2}\mathrm{E}\sup_{t\in\mathcal{T}}\left|\varepsilon_i t_{ij}\right| = 8M_xF_2\sum_{j=1}^{2}\mathrm{E}\sup_{\theta\in\Theta}\left|\sum_{i=1}^{n}\varepsilon_i X_i^\top\left(\theta^{(j)} - \theta^{*(j)}\right)\right| \\ &\le 16M_xF_2D_\Theta\mathrm{E}\max_{1\le k\le p}\left|\sum_{i=1}^{n}\varepsilon_i X_{ik}\right| \le 16\sqrt{2\log p}M_xF_2D_\Theta\sqrt{\max_{1\le k\le p}\sum_{i=1}^{n}X_{ik}^2}.\end{aligned}$$

Then, by (A3),

$$\mathrm{P}\left(S_k \ge 32\sqrt{2\log p}M_xF_2D_\Theta\sqrt{\max_{1\le k\le p}\sum_{i=1}^{n}X_{ik}^2} + M_1\sqrt{2t\max_{1\le k\le p}\sum_{i=1}^{n}X_{ik}^2} + 8M_1M_xt\right) \le e^{-t}.$$

Note that the right side of the inequality is free of θ, let $t = \log(p/q_5)$ in the above inequality, and use the same technique as previous, we obtain the uniform bound for it. The Theorem is proved by letting $q_2 = q_3 = q_1$, $q_4 = q_2$, $q_5 = q_3$, and $|w_i^{(1)}| \le w^{(1)}$. □

The lower bound of the likelihood-based divergence

Recall the standard steps for establishing the oracle inequality for a lasso estimator are (see [37] for example):

I. To avoid the ill behavior of Hessian, propose the restricted eigenvalue condition or other analogous conditions about the design matrix.
II. Find the tuning parameter based on the high-probability event, i.e., the KKT conditions.
III. According to some restricted eigenvalue assumptions and tuning parameter selection, derive the oracle inequalities via the definition of the lasso optimality and the minimizer under unknown expected risk function and some basic inequalities. There are three sub-steps:
   (i) Under the KKT conditions, show that the error vector $\widehat{\theta} - \theta^*$ is in some restricted set with structure sparsity, and check that $\widehat{\theta} - \theta^*$ is in a big compact set;
   (ii) Show that the likelihood-based divergence of $\widehat{\theta}$ and θ^* can be lower bounded by some quadratic distance between $\widehat{\theta}$ and θ^*;
   (iii) By some elementary inequalities and (ii), show that $\|\widehat{\theta} - \theta^*\|_1$ is in a smaller compact set with a radius of optimal rate (proportional to λ).

Under our approach, the KKT condition with a high probability is replaced by the stochastic Lipschitz condition, while other steps should remain the same. For most models belonging to the canonical exponential family, the step III.(ii) is quite trivial, see Lemma 1 in [38] for example. Nonetheless, it is worthy to note that our loss function is not in the canonical exponential family, so there is no extended discussion about the lower bound of the likelihood-based divergence of $\widehat{\theta}$ and θ^* in our setting. We will use the following theorem to clarify this thing.

Theorem A1. *Suppose the condition is the same as that in Theorem 1. Denote the true parameter for Y_i is μ^* and k^*. If $\{Z_i\theta\}_{i=1,\cdots,n,\theta\in\Theta} \subseteq \mathcal{S} \cap \{s \in \mathbb{R}^2 : 2s_1 + (1 + s_2(1 - k^*)k^{*\mu^*})\mu^* \le \frac{s_1+\mu^*}{2s_2^2}\}$ and $\mu^* \ge 1$, then*

$$\mathrm{E}\gamma(Z_i\theta, Y_i) - \mathrm{E}\gamma(Z_i\theta^*, Y_i) \ge C_\gamma\|Z_i(\theta - \theta^*)\|_2^2,$$

where C_γ is a positive constant and its exact definition is in the proof.

Proof. For simplicity, we drop the index i. By the definition and the notation in Theorem 1,

$$\mathrm{E}\gamma(Z\theta, Y) - \mathrm{E}\gamma(Z\theta^*, Y) = D_{\mathrm{KL}}(s, c),$$

where D_{KL} is the Kullback–Leibler divergence from the Y_i's density $f(y \mid Z\theta)$ to $f(y \mid Z\theta^*)$, i.e.,

$$D_{\mathrm{KL}}(s,c) := \int f(y \mid c) \log \frac{f(y \mid c)}{f(y \mid s)}\, dy.$$

Due to the identification of the negative binomial distribution, we have $D_{\mathrm{KL}}(s,c) \geq 0$ with equality if and only if $s = c$. Using the Taylor theorem,

$$\begin{aligned} D_{\mathrm{KL}}(s,c) &= D_{\mathrm{KL}}(c,c) + \frac{\partial}{\partial s} D_{\mathrm{KL}}(s,c)\Big|_{s=c} + \frac{1}{2}(s-c)^\top \left[\frac{\partial^2}{\partial s \partial s^\top} D_{\mathrm{KL}}(s,c)\right]_{s=c+\rho(s-c)} (s-c) \\ &= \frac{1}{2}(s-c)^\top \left[\frac{\partial^2}{\partial s \partial s^\top} D_{\mathrm{KL}}(s,c)\right]_{s=c+\rho(s-c)} (s-c) \\ &\geq \frac{1}{2} \inf_{\rho \in [0,1]} \lambda_{min} \left[\frac{\partial^2}{\partial s \partial s^\top} D_{\mathrm{KL}}(s,c)\right]_{s=c+\rho(s-c)} \|s-c\|_2^2 \end{aligned}$$

where $\rho \in [0,1]$ and $\lambda_{min}(\mathrm{M})$ is the smallest eigenvalue of the matrix M. Thus, it is enough to show that $\left[\frac{\partial^2}{\partial s \partial s^\top} D_{\mathrm{KL}}(s,c)\right]_{s=c+\rho(s-c)}$ is strictly positive define for any $\rho \in [0,1]$. First, calculate directly,

$$\begin{aligned} \frac{\partial^2}{\partial s \partial s^\top} D_{\mathrm{KL}}(s,c) &= \int f(y \mid c) \left[\frac{\partial^2}{\partial s \partial s^\top} \gamma(s,y)\right] dy \\ &= \int f(y \mid c) \begin{bmatrix} \frac{e^{s_1+s_2}}{(e^{s_1}+e^{s_2})^2}(e^{s_2}+y) & \frac{e^{s_1+s_2}}{(e^{s_1}+e^{s_2})^2}(e^{s_1}-y) \\ \frac{e^{s_1+s_2}}{(e^{s_1}+e^{s_2})^2}(e^{s_1}-y) & \partial_2 \nu(s,y) + \frac{e^{s_1+s_2}}{(e^{s_1}+e^{s_2})^2} y \end{bmatrix} dy =: \begin{bmatrix} a_{11}+b & a_{12}-b \\ a_{21}-b & a_{22}+b \end{bmatrix}, \end{aligned}$$

where $a_{11} = \frac{e^{s_1+2s_2}}{(e^{s_1}+e^{s_2})^2}$, $a_{12} = \frac{e^{2s_1+s_2}}{(e^{s_1}+e^{s_2})^2}$, $b = \frac{e^{s_1+2s_2}}{(e^{s_1}+e^{s_2})^2}\mathrm{E}Y$, and

$$\begin{aligned} a_{22} = \mathrm{E}\partial_2 v(s,Y) = e^{s_2}\Big[\psi(e^{s_2}) + e^{s_2}\psi'(e^{s_2}) + \log(1+e^{s_1-s_2}) - \frac{e^{s_1}}{e^{s_1}+e^{s_2}} - \Big(\frac{e^{s_1}}{e^{s_1}+e^{s_2}}\Big)^2\Big] \\ - e^{s_2}\big[\mathrm{E}\psi(Y+e^{s_2}) + e^{s_2}\mathrm{E}\psi'(Y+e^{s_2})\big]. \end{aligned}$$

For a 2×2 matrix M, it is strictly positive define if and only if $\mathrm{tr}(\mathrm{M}) > 0$ and $\det(\mathrm{M}) > 0$. Denote $\mu = e^{s_1}, k = e^{s_2}$, and $\mu^* = e^{c_1}, k^* = e^{c_2}$ are true parameters for Y. Then,

$$\begin{aligned} \mathrm{tr}\left[\frac{\partial^2}{\partial s \partial s^\top} D_{\mathrm{KL}}(s,c)\right] &= \frac{\mu k^2}{(\mu+k)^2} + 2\frac{\mu k^2}{(\mu+k)^2}\mu^* - k\left[\frac{\mu}{\mu+k} + \Big(\frac{\mu}{\mu+k}\Big)^2\right] \\ &\quad + k\big[\log(1+\mu/k) + (\psi(k) - \mathrm{E}\psi(Y+k)) + k(\psi'(k) - \mathrm{E}\psi'(Y+k))\big] \\ &= \frac{2(\mu^*-1)\mu k^2}{(\mu+k)^2} + k\big[\log(1+\mu/k) + g_1(k) + k g_2(k)\big] \\ &\geq k\big[\log(1+\mu/k) + g_1(k) + k g_2(k)\big]. \end{aligned} \tag{A4}$$

Now, we are going to deal with $g_1(k) = \psi(k) - \mathrm{E}\psi(Y+k)$ and $g_2(k) = \psi'(k) - \mathrm{E}\psi'(Y+k)$. For $\psi(x)$,

$$0 > \psi''(x) = -\frac{1}{x^2} - \int_0^\infty t^2 \varphi(t) e^{-tx}\, dt \geq -\frac{1}{x^2} - \frac{2}{x^3}.$$

Therefore, $\psi(\cdot)$ is concave. Using Jensen inequality and median value theorem

$$g_1(k) = \psi(k) - \mathrm{E}\psi(Y+k) \geq \psi(k) - \psi\big(\mathrm{E}Y + k\big) \geq -\left(\frac{1}{k} + \frac{1}{k^2}\right)\mathrm{E}Y = -\mu^*\left(\frac{1}{k} + \frac{1}{k^2}\right).$$

Similarly, for $g_2(k)$, by using the fact that $\mathrm{E}(1/Y) = (1-k^*)k^{*\mu^*}\mu^*$ and the assumption,

$$\begin{aligned} g_2(k) &= \mathrm{E}\big[\psi'(k) - \psi'(Y+k)\big] \geq \mathrm{E}\bigg[Y\bigg(\frac{1}{(\xi(Y)+k)^2} + \frac{2}{(\xi(Y)+k)^3}\bigg)\bigg] \\ &\geq \mathrm{E}\bigg[\frac{Y}{(Y+k)^2}\bigg] + 2\mathrm{E}\bigg[\frac{Y}{(Y+k)^3}\bigg] \geq \bigg[\mathrm{E}\frac{(Y+k)^2}{Y}\bigg]^{-1} + 2\bigg[\mathrm{E}\frac{(Y+k)^3)}{Y}\bigg]^{-1} \\ &= \frac{1}{2k + (1+k^2(1-k^*)k^{*\mu^*})\mu^*} + \frac{2}{k^{*2}(k^*+\mu^*)/\mu^{*2} + \mu^{*2} + 3k\mu^* + 3k^2 + k^3(1-k^*)k^{*\mu^*}\mu^*} \\ &\geq (\mu+\mu^*)\bigg(\frac{1}{2k^2} + \frac{1}{k^3}\bigg). \end{aligned}$$

where $\xi(Y)$ lies between 0 and Y. The lower bounds for g_1 and g_2, together with the fact that $\log(1+x) \geq x - x^2/2$ for $x \geq 0$, we conclude that $\mathrm{tr}\Big[\frac{\partial^2}{\partial s \partial s^\top} D_{\mathrm{KL}}(s,c)\Big] > 0$. Similarly, we can also prove $\det\Big[\frac{\partial^2}{\partial s \partial s^\top} D_{\mathrm{KL}}(s,c)\Big] > 0$, so the theorem holds. □

The proof of Theorem 3. The proof follows the idea in [22]. First, by the definition of $\widehat{\theta}$,

$$\begin{aligned} P\big(\gamma(\widehat{\theta}) - \gamma(\theta^*)\big) &\leq P\big(\gamma(\widehat{\theta}) - \gamma(\theta^*)\big) + \big(\mathbb{P}_n\gamma(\theta^*) + \lambda\|\theta^*\|_{\omega,1}\big) - \big(\mathbb{P}_n\gamma(\widehat{\theta}) + \lambda\|\widehat{\theta}\|_{\omega,1}\big) \\ &\leq \frac{1}{\sqrt{n}}\mathbb{G}_n\big(\gamma(\theta^*) - \gamma(\widehat{\theta})\big) + \lambda\big(\|\theta^*\|_{\omega,1} - \|\widehat{\theta}\|_{\omega,1}\big). \end{aligned}$$

From Theorem A1, we also have

$$P\big(\gamma(\widehat{\theta}) - \gamma(\theta^*)\big) \geq \frac{C_\gamma}{n}\sum_{i=1}^{n}\|Z_i(\widehat{\theta} - \theta^*)\|_2^2 = \frac{C_\gamma}{n}\sum_{j=1}^{2}\|\mathrm{X}(\widehat{\theta}^{(j)} - \theta^{*(j)})\|_2^2.$$

Then, by Theorem 1 and the definition of λ,

$$\begin{aligned} C_\gamma\sum_{j=1}^{2}\|\mathrm{X}(\widehat{\theta}^{(j)} - \theta^{*(j)})\|_2^2 &\leq \sqrt{n}\mathbb{G}_n\big(\gamma(\theta^*) - \gamma(\widehat{\theta})\big) + n\lambda\big(\|\theta^*\|_{\omega,1} - \|\widehat{\theta}\|_{\omega,1}\big) \\ &\leq M_q\|\theta^* - \widehat{\theta}\|_1 + (1+1/a)M_q\big(\|\theta^*\|_{\omega,1} - \|\widehat{\theta}\|_{\omega,1}\big) \\ &= M_q\sum_{j=1}^{2}\bigg[\|\widehat{\theta}^{(j)} - \theta^{*(j)}\|_1 + (1+1/a)\omega_j\big(\|\theta^{*(j)}\|_1 - \|\widehat{\theta}^{(j)}\|_1\big)\bigg] \end{aligned}$$

holds with probability at least $1-q$, where $a = (K-1)/2$. Now, let $J_1, J_2 \subseteq \{1, \cdots, p\}$ be any sets with $J_j \supseteq \mathrm{spt}\,(\theta^{*(j)})$. It is easy to check

$$\begin{aligned} &\|\widehat{\theta}^{(j)} - \theta^{*(j)}\|_1 + (1+1/a)\omega_j\big(\|\theta^{*(j)}\|_1 - \|\widehat{\theta}^{(j)}\|_1\big) \\ &\qquad = \|\widehat{\theta}^{(j)}_{J_j} - \theta^{*(j)}\|_1 + \|\widehat{\theta}^{(j)}_{J_j^c}\|_1 + (1+1/a)\omega_j\big(\|\theta^{*(j)}\|_1 - \|\widehat{\theta}^{(j)}_{J_j}\|_1 - \|\widehat{\theta}^{(j)}_{J_j^c}\|_1\big) \\ &\qquad \leq (K/a)\|\widehat{\theta}^{(j)}_{J_j} - \theta^{*(j)}\|_1 - (1/a)\|\widehat{\theta}^{(j)}_{J_j^c}\|_1. \end{aligned}$$

by the fact $\omega_j \in [0,1]$. It gives that with probability at least $1-q$,

$$\sum_{j=1}^{2}\|\mathrm{X}(\widehat{\theta}^{(j)} - \theta^{*(j)})\|_2^2 \leq \frac{M_q}{aC_\gamma}\sum_{j=1}^{2}\big(K\|\widehat{\theta}^{(j)}_{J_j} - \theta^{*(j)}\|_1 - \|\widehat{\theta}^{(j)}_{J_j^c}\|_1\big). \tag{A5}$$

Let $A_1, A_2 \subseteq \{1, \cdots, p\}$ satisfying spt $\left(\theta^{*(j)}\right) \subseteq A_j$ and card$(A_j) = p_1$, and we also let B_j be the union of A_j and the indices of p_1 largest $\widehat{\theta}^{(j)}$. Then, A_j and B_j also guarantee (A5). In addition, from Lemma 1, they also give

$$\|\widehat{\theta}^{(j)}_{B_j^c}\|_2^2 \leq p_1^{-1}\|\widehat{\theta}^{(j)}_{A_j^c}\|_1^2.$$

In addition, from the definition of A_j and B_j, we know that $\|\widehat{\theta}^{(j)}_{A_j^c}\|_1 \geq \|\widehat{\theta}^{(j)}_{B_j^c}\|_1$ and $\|\widehat{\theta}^{(j)}_{A_j} - \theta^{*(j)}\|_1 \leq \|\widehat{\theta}^{(j)}_{B_j} - \theta^{*(j)}\|_1$.

Unlike the single lasso question, here we need to define $I := \{j = 1, 2 : K\|\widehat{\theta}^{(j)}_{A_j} - \theta^{*(j)}\|_1 \geq \|\widehat{\theta}^{(j)}_{A_j^c}\|_1\}$, and consider $j \in I$ and $j \notin I$ separately. Obviously, $I \neq \emptyset$, or (A5) cannot be beholden. For $j \in I$, we have

$$K\|\widehat{\theta}^{(j)}_{B_j} - \theta^{*(j)}\|_1 - \|\widehat{\theta}^{(j)}_{B_j^c}\|_1 \geq K\|\widehat{\theta}^{(j)}_{A_j} - \theta^{*(j)}\|_1 - \|\widehat{\theta}^{(j)}_{A_j^c}\|_1 \geq 0.$$

Then, by the restricted eigenvalue condition,

$$n\kappa^2\|\widehat{\theta}^{(j)}_{J_j} - \theta^{*(j)}\|_2^2 \leq \|\mathrm{X}(\widehat{\theta}^{(j)} - \theta^{*(j)})\|_2^2$$

holds for $J_j = A_j$ or $J_j = B_j$. Note that from (A5),

$$\sum_{j\in I}\|\mathrm{X}(\widehat{\theta}^{(j)} - \theta^{*(j)})\|_2^2 \leq \frac{M_q}{aC_\gamma}\sum_{j\in I}(\|\widehat{\theta}^{(j)}_{A_j} - \theta^{*(j)}\|_1 - \|\widehat{\theta}^{(j)}_{A_j^c}\|_1) \leq \frac{M_q}{aC_\gamma}\sum_{j\in I}(\|\widehat{\theta}^{(j)}_{B_j} - \theta^{*(j)}\|_1 - \|\widehat{\theta}^{(j)}_{B_j^c}\|_1),$$

then by Cauchy–Schwartz inequality,

$$\begin{aligned} n\kappa^2\sum_{j\in I}\|\widehat{\theta}^{(j)}_{A_j} - \theta^{*(j)}\|_2^2 \leq \|\mathrm{X}(\widehat{\theta}^{(j)}_{A_j} - \theta^{*(j)})\|_2^2 &\leq \frac{M_q K}{aC_\gamma}\sum_{j\in I}\|\widehat{\theta}^{(j)}_{A_j} - \theta^{*(j)}\|_1 \\ &\leq \frac{M_q K\sqrt{p_1}}{aC_\gamma}\sum_{j\in I}\|\widehat{\theta}^{(j)}_{A_j} - \theta^{*(j)}\|_2 \leq \frac{M_q K\sqrt{2p_1}}{aC_\gamma}\left[\sum_{j\in I}\|\widehat{\theta}^{(j)}_{A_j} - \theta^{*(j)}\|_2^2\right]^{1/2}. \end{aligned}$$

It gives

$$\sum_{j\in I}\|\widehat{\theta}^{(j)}_{A_j} - \theta^{*(j)}\|_2^2 \leq \frac{2p_1M_q^2K^2}{a^2\kappa^4n^2C_\gamma^2}, \qquad \sum_{j\in I}\|\widehat{\theta}^{(j)}_{B_j} - \theta^{*(j)}\|_2^2 \leq \frac{4p_1M_q^2K^2}{a^2\kappa^4n^2C_\gamma^2},$$

where we use that fact card$(B_j) = 2p_1$. Furthermore, because

$$\|\widehat{\theta}^{(j)}_{B_j^c}\|_2^2 \leq \sum_{j\in I}p_1^{-1}\|\widehat{\theta}^{(j)}_{A_j^c}\|_1^2 \leq \frac{K^2}{p_1}\sum_{j\in I}\|\widehat{\theta}^{(j)}_{A_j} - \theta^{*(j)}\|_1^2 \leq K^2\sum_{j\in I}\|\widehat{\theta}^{(j)}_{A_j} - \theta^{*(j)}\|_2^2,$$

we can conclude that

$$\begin{aligned} \sum_{j\in I}\|\widehat{\theta}^{(j)} - \theta^{*(j)}\|_2^2 &= \sum_{j\in I}(\|\widehat{\theta}^{(j)}_{B_j} - \theta^{*(j)}\|_2^2 + \|\widehat{\theta}^{(j)}_{B_j^c}\|_2^2) \\ &\leq \sum_{j\in I}(\|\widehat{\theta}^{(j)}_{B_j} - \theta^{*(j)}\|_2^2 + K^2\|\widehat{\theta}^{(j)}_{A_j} - \theta^{*(j)}\|_2^2) = \frac{2p_1M_q^2(2+K^2)K^2}{a^2\kappa^4n^2C_\gamma^2}. \end{aligned} \tag{A6}$$

Now, we will tickle the situation that $j \notin I$. For $j \notin I$, $K\|\widehat{\theta}^{(j)}_{A_j} - \theta^{*(j)}\|_1 < \|\widehat{\theta}^{(j)}_{A_j^c}\|_1$. Again from (A5), we have

$$\sum_{j\notin I} \|\mathrm{X}(\widehat{\theta}^{(j)} - \theta^{*(j)})\|_2^2 \leq \frac{M_q K}{aC_\gamma} \sum_{j\in I} \|\widehat{\theta}^{(j)}_{A_j} - \theta^{*(j)}\|_1$$

and

$$0 \leq \sum_{j\notin I} \left(\|\widehat{\theta}^{(j)}_{A_j^c}\|_1 - K\|\widehat{\theta}^{(j)}_{A_j} - \theta^{*(j)}\|_1\right) \leq K \sum_{j\in I} \|\widehat{\theta}^{(j)}_{A_j} - \theta^{*(j)}\|_1.$$

Indeed, if the two inequalities above have the opposite direction, then for the first one, one can find that

$$\sum_{j\in I} \|\mathrm{X}(\widehat{\theta}^{(j)} - \theta^{*(j)})\|_2^2 \leq \frac{M_q}{aC_\gamma} \left[\sum_{j\notin I} \left(K\|\widehat{\theta}^{(j)}_{A_j} - \theta^{*(j)}\|_1 - \|\widehat{\theta}^{(j)}_{A_j^c}\|_1\right) - \sum_{j\in I} \|\widehat{\theta}^{(j)}_{A_j^c}\|_1\right] < 0,$$

and

$$\sum_{j=1}^{2} \|\mathrm{X}(\widehat{\theta}^{(j)} - \theta^{*(j)})\|_2^2 \leq -\frac{M_q}{aC_\gamma} \sum_{j\in I} \|\widehat{\theta}^{(j)}_{A_j^c}\|_1 < 0.$$

Once again, by Cauchy–Schwartz inequality,

$$\sum_{j\in I} \|\widehat{\theta}^{(j)}_{A_j} - \theta^{*(j)}\|_1 \leq \sqrt{p_1} \sum_{j\in I} \|\widehat{\theta}^{(j)}_{A_j} - \theta^{*(j)}\|_2 \leq \sqrt{2p_1} \left[\sum_{j\in I} \|\widehat{\theta}^{(j)}_{A_j} - \theta^{*(j)}\|_2^2\right]^{1/2} \leq \frac{2p_1 M_q K}{a\kappa^2 n C_\gamma}.$$

Denote $\Delta_j := \|\widehat{\theta}^{(j)}_{A_j^c}\|_1 - K\|\widehat{\theta}^{(j)}_{A_j} - \theta^{*(j)}\|_1$. Then, for $j \notin J$, $\Delta_j > 0$, and

$$\sum_{j\notin I} \Delta_j \leq K \sum_{j\in I} \|\widehat{\theta}^{(j)}_{A_j} - \theta^{*(j)}\|_1 \leq \frac{2p_1 M_q K^2}{a\kappa^2 n C_\gamma}.$$

For any $j \notin I$, define

$$\widetilde{\theta}^{(j)} = \widehat{\theta}^{(j)} + \frac{\Delta_j}{p_1 K} \sum_{k\in A_j} \mathrm{sgn}\left(\widehat{\theta}^{(j)}_k - \theta^{*(j)}_k\right) e_k.$$

Then, for $k \in A_j$,

$$|\widetilde{\theta}^{(j)}_k - \theta^{*(j)}_k| = |\widehat{\theta}^{(j)}_k - \theta^{*(j)}_k| + \frac{\Delta_j}{p_1 K},$$

while for $k \notin I$, $\widetilde{\theta}^{(j)}_k = \widehat{\theta}^{(j)}_k$. Therefore,

$$K\|\widetilde{\theta}^{(j)}_{A_j} - \theta^{*(j)}\|_1 = K\left[\|\widehat{\theta}^{(j)}_{A_j} - \theta^{*(j)}\|_1 + \sum_{k\in A_j} \frac{\Delta_j}{p_1 K}\right] = \|\widehat{\theta}^{(j)}_{A_j^c}\|_1 = \|\widetilde{\theta}^{(j)}_{A_j^c}\|_1,$$

and consequently $\|\widetilde{\theta}^{(j)}_{B_j^c}\|_1 \leq K\|\widetilde{\theta}^{(j)}_{B_j} - \theta^{*(j)}\|_1$. Once again, by the restricted eigenvalue condition,

$$\|\mathrm{X}(\widetilde{\theta}^{(j)} - \theta^{*(j)})\|_2^2 \geq n\kappa^2 \|\widetilde{\theta}^{(j)}_{B_j} - \theta^{*(j)}\|_2^2 \geq n\kappa^2 \|\widetilde{\theta}^{(j)}_{A_j} - \theta^{*(j)}\|_2^2. \tag{A7}$$

On the other hand, note that for any $s, t \in \mathbb{R}^m$ inequality $\|s+t\|_2^2 \le 2(\|s\|_2^2 + \|t\|_2^2)$ and $\|s\|_2 \le \|s\|_1 \le \sqrt{m}\|s\|_2$ hold, we conclude

$$\begin{aligned}\sum_{j\notin I}\|\mathrm{X}(\widetilde{\theta}^{(j)}-\theta^{*(j)})\|_2^2 &\le 2\sum_{j\notin I}\left(\|\mathrm{X}(\widehat{\theta}^{(j)}-\theta^{*(j)})\|_2^2+\|\mathrm{X}(\widehat{\theta}^{(j)}-\widetilde{\theta}^{(j)})\|_2^2\right)\\ &\le \frac{2M_qK}{aC_\gamma}\sum_{j\in I}\|\widehat{\theta}^{(j)}_{A_j}-\theta^{*(j)}\|_1+2\sum_{j\notin I}\|\mathrm{X}(\widehat{\theta}^{(j)}-\widetilde{\theta}^{(j)})\|_2^2\\ &\le \frac{4p_1M_q^2K^2}{na^2\kappa^2C_\gamma^2}+2\sum_{j\notin I}\|\mathrm{X}(\widehat{\theta}^{(j)}-\widetilde{\theta}^{(j)})\|_2^2.\end{aligned} \tag{A8}$$

Next, we will use the definition of the p_1-restricted isometry constant $\sigma^2_{\mathrm{X},I}$. Because $\mathrm{spt}\left(\widetilde{\theta}^{(j)}-\widehat{\theta}^{(j)}\right) \le \mathrm{card}(A_j) = p_1$, then

$$\begin{aligned}\sum_{j\notin I}\|\mathrm{X}(\widehat{\theta}^{(j)}-\widetilde{\theta}^{(j)})\|_2^2 &\le \sigma^2_{\mathrm{X},p_1}\sum_{j\notin I}\|\widehat{\theta}^{(j)}-\widetilde{\theta}^{(j)}\|_2^2\\ &= \sigma^2_{\mathrm{X},p_1}\sum_{j\notin I}\sum_{k\in A_j}\left(\frac{\Delta_j}{p_1K}\right)^2=\frac{\sigma^2_{\mathrm{X},p_1}}{p_1K^2}\sum_{j\notin I}\Delta_j^2\\ &\le \frac{\sigma^2_{\mathrm{X},p_1}}{p_1K^2}\left(\sum_{j\notin I}\Delta_j\right)^2\le\frac{4p_1\sigma^2_{\mathrm{X},p_1}K^2}{a^2\kappa^4n^2C_\gamma^2}.\end{aligned}$$

The above inequality together with (A7) and (A8) gives

$$\sum_{j\notin I}\|\widetilde{\theta}^{(j)}_{A_j}-\theta^{*(j)}\|_2^2\le\sum_{j\notin I}\|\widetilde{\theta}^{(j)}_{B_j}-\theta^{*(j)}\|_2^2\le\frac{4p_1(n\kappa^2+2\sigma^2_{\mathrm{X},p_1})M_q^2K^2}{a^2C_\gamma^2n^3\kappa^6}.$$

Finally, because

$$\|\widetilde{\theta}^{(j)}_{B_j^c}\|_2^2\le\|\widetilde{\theta}^{(j)}_{B_j^c}\|_1^2\le K^2\|\widetilde{\theta}^{(j)}_{B_j}-\theta^{*(j)}\|_1^2\le 2p_1K^2\|\widetilde{\theta}^{(j)}_{B_j}-\theta^{*(j)}\|_2^2,$$

we obtain that

$$\begin{aligned}\sum_{j\notin I}\|\widehat{\theta}^{(j)}-\theta^{*(j)}\|_2^2\le\sum_{j\notin I}\|\widetilde{\theta}^{(j)}-\theta^{*(j)}\|_2^2&=\sum_{j\notin I}\left(\|\widetilde{\theta}^{(j)}_{B_j}-\theta^{*(j)}\|_2^2+\|\widetilde{\theta}^{(j)}_{B_j^c}\|_2^2\right)\\ &\le(1+2p_1K)\sum_{j\notin I}\|\widetilde{\theta}^{(j)}_{B_j}-\theta^{*(j)}\|_2^2\le\frac{4p_1(1+2p_1K)(n\kappa^2+2\sigma^2_{\mathrm{X},p_1})M_q^2K^2}{a^2C_\gamma^2n^3\kappa^6}.\end{aligned} \tag{A9}$$

Combining (A6) and (A9), it is easy to see what remains. □

References

1. Dai, H.; Bao, Y.; Bao, M. Maximum likelihood estimate for the dispersion parameter of the negative binomial distribution. *Stat. Probab. Lett.* **2013**, *83*, 21–27. [CrossRef]
2. Allison, P.D.; Waterman, R.P. Fixed–effects negative binomial regression models. *Sociol. Methodol.* **2002**, *32*, 247–265. [CrossRef]
3. Hilbe, J.M. *Negative Binomial Regression*; Cambridge University Press: Cambridge, UK, 2011.
4. Weißbach, R.; Radloff, L. Consistency for the negative binomial regression with fixed covariate. *Metrika* **2020**, *83*, 627–641. [CrossRef]
5. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Statal Soc. Ser. B* **1996**, *58*, 267–288. [CrossRef]
6. Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [CrossRef]
7. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2005**, *67*, 301–320. [CrossRef]
8. Zou, H. The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* **2006**, *101*, 1418–1429. [CrossRef]

9. Qiu, Y.; Chen, S.X.; Nettleton, D. Detecting rare and faint signals via thresholding maximum likelihood estimators. *Ann. Stat.* **2018**, *46*, 895–923. [CrossRef]
10. Xie, F.; Xiao, Z. Consistency of l1 penalized negative binomial regressions. *Stat. Probab. Lett.* **2020**, *165*, 108816. [CrossRef]
11. Li, Y.; Rahman, T.; Ma, T.; Tang, L.; Tseng, G.C. A sparse negative binomial mixture model for clustering RNA-seq count data. *Biostatistics* **2021**, kxab025. [CrossRef]
12. Jankowiak, M. Fast Bayesian Variable Selection in Binomial and Negative Binomial Regression. *arXiv* **2021**, arXiv:2106.14981.
13. Lisawadi, S.; Ahmed, S.; Reangsephet, O. Post estimation and prediction strategies in negative binomial regression model. *Int. J. Model. Simul.* **2021**, *41*, 463–477. [CrossRef]
14. Zhang, H.; Jia, J. Elastic-net Regularized High-dimensional Negative Binomial Regression: Consistency and Weak Signals Detection. *Stat. Sin.* **2022**, *32*, 181–207. [CrossRef]
15. Xu, D.; Zhang, Z.; Wu, L. Variable selection in high-dimensional double generalized linear models. *Stat. Pap.* **2014**, *55*, 327–347. [CrossRef]
16. Yee, T.W. *Vector Generalized Linear and Additive Models: With an Implementation in R*; Springer: Berlin/Heidelberg, Germany, 2015.
17. Nguelifack, B.M.; Kemajou-Brown, I. Robust rank-based variable selection in double generalized linear models with diverging number of parameters under adaptive Lasso. *J. Stat. Comput. Simul.* **2019**, *89*, 2051–2072. [CrossRef]
18. Cavalaro, L.L.; Pereira, G.H. A procedure for variable selection in double generalized linear models. *J. Stat. Comput. Simul.* **2022**, 1–18. [CrossRef]
19. Wang, Z.; Ma, S.; Zappitelli, M.; Parikh, C.; Wang, C.Y.; Devarajan, P. Penalized count data regression with application to hospital stay after pediatric cardiac surgery. *Stat. Methods Med. Res.* **2016**, *25*, 2685–2703. [CrossRef]
20. Huang, H.; Zhang, H.; Li, B. Weighted Lasso estimates for sparse logistic regression: Non-asymptotic properties with measurement errors. *Acta Math. Sci.* **2021**, *41*, 207–230. [CrossRef]
21. Adamczak, R. A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.* **2008**, *13*, 1000–1034. [CrossRef]
22. Bickel, P.J.; Ritov, Y.; Tsybakov, A.B. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Stat.* **2009**, *37*, 1705–1732. [CrossRef]
23. Candes, E.; Tao, T. The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Stat.* **2007**, *35*, 2313–2351.
24. Riphahn, R.T.; Wambach, A.; Million, A. Incentive effects in the demand for health care: A bivariate panel count data estimation. *J. Appl. Econom.* **2003**, *18*, 387–405. [CrossRef]
25. Yang, X.; Song, S.; Zhang, H. Law of iterated logarithm and model selection consistency for generalized linear models with independent and dependent responses. *Front. Math. China* **2021**, *16*, 825–856. [CrossRef]
26. Shi, C.; Song, R.; Chen, Z.; Li, R. Linear hypothesis testing for high dimensional generalized linear models. *Ann. Stat.* **2019**, *47*, 2671. [CrossRef] [PubMed]
27. Xie, F.; Lederer, J. Aggregating Knockoffs for False Discovery Rate Control with an Application to Gut Microbiome Data. *Entropy* **2021**, *23*, 230. [CrossRef]
28. Cui, C.; Jia, J.; Xiao, Y.; Zhang, H. Directional FDR Control for Sub-Gaussian Sparse GLMs. *arXiv* **2021**, arXiv:2105.00393.
29. Bateman, H. *Higher Transcendental Functions [Volumes i–iii]*; McGraw-Hill Book Company: New York, NY, USA, 1953; Volume 1.
30. Alzer, H. On some inequalities for the gamma and psi functions. *Math. Comput.* **1997**, *66*, 373–389. [CrossRef]
31. Zhang, H.; Chen, S.X. Concentration inequalities for statistical inference. *Commun. Math. Res.* **2021**, *37*, 1–85.
32. Moriguchi, S.; Murota, K.; Tamura, A.; Tardella, F. Discrete midpoint convexity. *Math. Oper. Res.* **2020**, *45*, 99–128. [CrossRef]
33. Sen, B. *A Gentle Introduction to Empirical Process Theory and Applications*; Columbia University: New York, NY, USA, 2018.
34. Chi, Z. Stochastic Lipschitz continuity for high dimensional Lasso with multiple linear covariate structures or hidden linear covariates. *arXiv* **2010**, arXiv:1011.1384.
35. Ledoux, M.; Talagrand, M. *Probability in Banach Spaces: Isoperimetry and Processes*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013.
36. Massart, P. Some applications of concentration inequalities to statistics. *Ann. Fac. Sci. Toulouse Math.* **2000**, *9*, 245–303. [CrossRef]
37. 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*, 1–33. [CrossRef]
38. Abramovich, F.; Grinshtein, V. Model selection and minimax estimation in generalized linear models. *IEEE Trans. Inf. Theory* **2016**, *62*, 3721–3730. [CrossRef]