Electronic Journal of Statistics
Vol. 13 (2019) 2898–2920
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1580

Sparse Poisson regression with penalized weighted score function*

Jinzhu Jia†

School of Public Health and Center for Statistical Science
Peking University
Beijing, China
e-mail: jzjia@pku.edu.cn

Fang Xie‡

School of Mathematics and Statistics
Wuhan University
Wuhan, Hubei 430072, China
e-mail: fangxie219@foxmail.com

and

Lihu Xu§

1. Department of Mathematics
Faculty of Science and Technology
University of Macau
Av. Padre Tomás Pereira, Taipa Macau, China
2. UMacau Zhuhai Research Institute
Zhuhai, China
e-mail: lihuxu@um.edu.mo

Abstract: By introducing a weighted score function, we propose a new penalized method, similar to square root lasso, to study sparse Poisson regression problems. The corresponding new estimator not only has ℓ_1 consistency but also enjoys the tuning free property. We further verify our theoretical results by numerical simulations and apply them to an image reconstruction problem.

Keywords and phrases: Poisson regression, ℓ_1 penalization, ℓ_1 consistency, moderate deviation, tuning-free, image reconstruction.

Received February 2018.

*Jinzhu Jia and Fang Xie contributed equally, they are co-first authors and are listed in alphabetical orders, Lihu Xu is the corresponding author.

†Jinzhu Jia is supported by the grant NSFC 11571021.

‡Fang Xie is supported by the grant NSFC 11571390.

§Lihu Xu is supported by the grants Macao S.A.R FDCT 030/2016/A1, 049/2014/A1 and NSFC 11571390 and University of Macau MYRG2015-00021-FST, 2016-00025-FST.

1. Introduction

Variable selection problems have been extensively studied in the past twenty years, for instance, lasso [5, 24], adaptive lasso [27], SCAD [7] and MCP [26], square root lasso [5]. Lasso is one of the most powerful methods in linear regressions, not only because of its prediction performance [24] but also due to its convexity and efficient computability. However, as pointed out by [5], its optimal tuning parameter λ depends on a usually unknown noise standard deviation σ, one has to first estimate it in many applications. To overcome this problem, [4] proposed the square root lasso by replacing the lasso's loss function with its square root. As a result, the corresponding tuning parameter λ does not depend on σ at all, which is often called tuning free property.

The lasso method has been widely applied in generalized linear models, such as ℓ_1 regularized logistic regression [20] and ℓ_1 penalized Poisson regression [14]. Poisson regression is a popular generalized linear model [17] for modeling count data, its high dimensional versions have received more and more attention, see [14] for the consistency of ℓ_1 penalized Poisson regression models, and [19] for the performance bounds of compressed sensing under Poisson models. Variable selection for sparse Poisson model can be performed via ℓ_1 penalized log-likelihood method, in which the penalty parameter λ depends on the variance of Poisson noise, like lasso. For more results on variable selection in generalized linear models, we refer the reader to [8, 11, 13, 15] and the literature therein.

In this paper, we consider a high-dimensional heteroscedastic Poisson model. Let $(x_1, y_1), \ldots, (x_n, y_n)$ be n observed data from independent random vectors $(X_1, Y_1), \ldots, (X_n, Y_n)$, where X_i is an $\mathbb{R}^p$-valued random vector and Y_i is an $\mathbb{R}$-valued random variable for each $i \in [n]$. We assume that for $i \in [n]$ and $x_i \in \mathbb{R}^p$

$$Y_i | X_i = x_i \sim \text{Poisson}(\mu(x_i)) \quad \text{with} \quad \log(\mu(x_i)) = x_i^T \beta^*,$$

where $\beta^* \in \mathbb{R}^p$ is an unknown parameter vector to be estimated. Inspired by the square root lasso, we study the sparse Poisson regression by optimizing a new ℓ_1 *penalized weighted score function* as the following:

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^{n} 2(y_i e^{-\frac{1}{2} x_i^T \beta} + e^{\frac{1}{2} x_i^T \beta}) + \lambda \|\beta\|_1 \right\},$$

where $\lambda > 0$ is the penalty level, see more details in Section 2. Note that [27] and [11] suggested taking into account the heteroscedasticity of Poisson observations by introducing adapted weights in the penalty rather than a score function.

Let us first have a brief discussion on the theoretical results in this paper. We show that the new estimator $\hat{\beta}$ is consistent and that the penalty level λ enjoys the tuning free property in the sense that it does not depend on the rate parameter $e^{x_i^T \beta^*}$. By a moderate deviation technique due to [22] (see also [10, 16]), we get a Gaussian approximation for λ, which provides a way to choose penalty level directly rather than by cross-validation. Simulations show that the estimator with this Gaussian approximated penalty performs very well.

As an application, we apply our method to image reconstruction based on a simulated limited-angle emission tomography data set and compare our algorithm with the well-known emission tomography reconstruction algorithms. The results show that our proposed method performs much better than others under Poisson noise.

The rest of the paper is organized as follows. In Section 2, we introduce the ℓ_1 *penalized weighted score function* for sparse Poisson regression. In Section 3, under three regular conditions and two choices of penalty level, we give the main theorems about the consistency of our estimator and the non-asymptotic error bounds. By numerical simulations in Section 4.1, we compare the estimation errors of our methods with three different selections of λ, and with the results of lasso and adaptive lasso [11] for Poisson regression. Section 4.2 gives the details of the application of our method to image reconstructions. We make a conclusion remark in Section 5 and give the detailed proofs of our theoretical results in Section A.

Notations and definitions

Now we introduce the notations and definitions which will be used throughout the paper. For any positive integer n, we denote $[n] = \{1, 2, \ldots, n\}$. For any d-dimensional vector $v = (v_1, \ldots, v_d)^T$, denote $\|v\|_q^q = \sum_{i=1}^{d} |v_i|^q$ for any $q \in [1, +\infty)$ and $\|v\|_\infty = \max_{i\in[d]} |v_i|$.

Write $X = (x_1, \ldots, x_n)^T \in \mathbb{R}^{n\times p}$ and $Y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$. Let $\beta^* \in \mathbb{R}^p$ be the parameter vector to be estimated, denote by $T = supp(\beta^*) = \{j \in [p] : \beta_j^* \neq 0\}$ the non-zero coordinates of β^*, and let $s = |T|$ be the number of non-zero elements of β^*.

Denote by $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$ two sequences, the notation $b_n = O(a_n)$ means that there exists a constant $C > 0$ such that $b_n \leq Ca_n$ for all $n \geq 1$ and the notation $b_n = o(a_n)$ means that $\lim_{n\to\infty} b_n/a_n = 0$. If $f : \mathbb{R}^p \to \mathbb{R}$ is a differentiable function, we denote by ∇f the gradient of f.

Definition 1.1 (Score function)**.** Let Z be a random variable with a likelihood function $L(\theta, Z)$, where θ is a vector parameter. *The score function* with respect to $L(\theta, Z)$ is defined by

$$u(\theta, Z) := \nabla_\theta \log L(\theta, Z).$$

2. ℓ_1 penalized weighted score function method

2.1. Square-root lasso revisited

Before introducing our ℓ_1 *penalized weighted score function*, we first briefly review the square root lasso and explain it from the point of view of weighted

score function. Recall the classical lasso for linear model [24]:

$$Y = X\beta^* + \epsilon, \qquad \epsilon \sim N(0, \sigma^2 I_n),$$

with

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{n} \|Y - X\beta\|_2^2 + 2\lambda \|\beta\|_1 \right\}, \tag{2.1}$$

where $Y \in \mathbb{R}^n$ is the response vector, $X \in \mathbb{R}^{n \times p}$ is the design matrix, $\beta^* \in \mathbb{R}^p$ is the parameter to be estimated, I_n is an $n \times n$ identity matrix, and $\lambda > 0$ is the penalty level. The lasso estimator satisfies Karush-Kuhn-Tucker(KKT) condition as follows:

$$-\frac{1}{n} X^T (Y - X\beta) + \lambda\kappa = 0, \tag{2.2}$$

where $\kappa = (\kappa_1, \ldots, \kappa_p)^T$ and κ_j is the subgradient of $|\beta_j|$ for each j. Note that κ_j is the sign of β_j for $\beta_j \neq 0$ and $\beta_j \in [-1, 1]$ for $\beta_j = 0$.

Observe that $-\frac{1}{n}\|Y - X\beta\|_2^2$ is (proportional to) the log-likelihood of a Gaussian distribution, whose score function is (proportional to) $\frac{1}{n} X^T (Y - X\beta)$, to get a bound for $\hat{\beta} - \beta^*$, the following relation needs to hold with high probability [4]:

$$\lambda > c \| \frac{1}{n} X^T (Y - X\beta^*) \|_\infty = \frac{c}{n} \|X^T \epsilon\|_\infty,$$

where $c > 1$ is some constant. Since ϵ has a variance σ^2, whose exact value is often not known, to tune λ, one has to firstly estimate σ^2. To overcome this disadvantage, [4] proposed the square root lasso by replacing $\frac{1}{n} X^T (Y - X\beta)$ with the following weighted score function:

$$\frac{\frac{1}{n} X^T (Y - X\beta)}{\frac{1}{\sqrt{n}} \|Y - X\beta\|_2},$$

whose distribution at β^* does not depend on σ^2 anymore. Replacing $-\frac{1}{n} X^T (Y - X\beta)$ in (2.2) with this new score function, we get

$$-\frac{X^T (Y - X\beta)}{\sqrt{n} \|Y - X\beta\|_2} + \lambda\kappa = 0. \tag{2.3}$$

Since the gradient of $\frac{1}{\sqrt{n}} \|Y - X\beta\|_2$ is $-\frac{X^T (Y - X\beta)}{\sqrt{n} \|Y - X\beta\|_2}$, (2.3) is the KKT condition of the following optimization problem:

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{\sqrt{n}} \|Y - X\beta\|_2 + \lambda \|\beta\|_1 \right\}. \tag{2.4}$$

The above idea of treating the square root lasso as a penalized weighted score function can be generalized to other regression problems such as heteroscedastic models, we will give details about the application of this idea to sparse Poisson models in the next subsection.

2.2. ℓ_1 penalized weighted score function method for sparse Poisson regression

Let us briefly recall the Poisson model. Let (x_1, y_1), …, (x_n, y_n) be n observed data from independent random vectors (X_1, Y_1), …, (X_n, Y_n), where X_i is an $\mathbb{R}^p$-valued random vector and Y_i is an $\mathbb{R}$-valued random variable for each $i \in [n]$. We assume that for $i \in [n]$ and $x_i \in \mathbb{R}^p$

$$Y_i | X_i = x_i \sim \text{Poisson}(\mu(x_i)) \quad \text{with} \quad \log(\mu(x_i)) = x_i^T \beta^*,$$

where $\beta^* \in \mathbb{R}^p$ is an unknown parameter vector to be estimated. Without loss of generality, we assume that x_1, …, x_n satisfy

$$\frac{1}{n}\sum_{i=1}^n x_{ij} = 0 \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^n x_{ij}^2 = 1, \quad \text{for all } j \in [p].$$

Under the above settings, the log-likelihood (up to a scale and a constant shift) of Poisson distribution is defined as follows:

$$\ell(\beta) = \frac{1}{n}\sum_{i=1}^n (y_i x_i^T \beta - e^{x_i^T \beta}). \tag{2.5}$$

Sparse Poisson regression can be solved via ℓ_1 penalized log-likelihood method [14] and the corresponding estimator is defined by

$$\bar{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \Big\{ -\frac{1}{n}\sum_{i=1}^n (y_i x_i^T \beta - e^{x_i^T \beta}) + \bar{\lambda}\|\beta\|_1 \Big\},$$

where $\bar{\lambda} > 0$ is the penalty level. The corresponding KKT condition reads as

$$-\frac{1}{n}\sum_{i=1}^n x_i (y_i - e^{x_i^T \bar{\beta}}) + \bar{\lambda}\kappa = 0, \tag{2.6}$$

where κ is the sub-gradient of $\|\beta\|_1$. Like the lasso, to get a good estimator of β^* requires that

$$\bar{\lambda} \geq c\|\nabla \ell(\beta^*)\|_\infty \tag{2.7}$$

holds with high probability for some constant $c > 1$ [14]. The score function valued at $\beta = \beta^*$ is

$$\nabla \ell(\beta^*) = \frac{1}{n}\sum_{i=1}^n x_i (y_i - e^{x_i^T \beta^*}),$$

in which $y_i - e^{x_i^T \beta^*}$ is a random variable with variance $e^{x_i^T \beta^*}$. This means that $\bar{\lambda}$ depends on β^* and is sensitive to the value $e^{x_i^T \beta^*}$.

Like (2.3), our idea is to divide $y_i - e^{x_i^T\beta^*}$ by $\sqrt{e^{x_i^T\beta^*}}$ to make the penalty level λ not depend on $e^{x_i^T\beta^*}$. More precisely, we replace $y_i - e^{x_i^T\beta^*}$ in (2.6) with $\frac{(y_i - e^{x_i^T\beta})}{\sqrt{e^{x_i^T\beta}}}$ and obtain

$$-\frac{1}{n}\sum_{i=1}^{n}\frac{x_i(y_i - e^{x_i^T\beta})}{\sqrt{e^{x_i^T\beta}}} + \lambda\kappa = 0. \tag{2.8}$$

Note that $\frac{(y_i - e^{x_i^T\beta^*})}{\sqrt{e^{x_i^T\beta^*}}}$ has mean 0 and variance 1 for each $i \in [n]$. It is easy to check that the following relation holds:

$$\nabla f(\beta) = -\frac{1}{n}\sum_{i=1}^{n}\frac{x_i(y_i - e^{x_i^T\beta})}{\sqrt{e^{x_i^T\beta}}} \quad \text{with} \quad f(\beta) = \frac{1}{n}\sum_{i=1}^{n} 2(y_i e^{-\frac{1}{2}x_i^T\beta} + e^{\frac{1}{2}x_i^T\beta}),$$

so (2.8) is the KKT condition of the convex optimization problem:

$$\hat{\beta} = \arg\min_{\beta\in\mathbb{R}^p}\left\{f(\beta) + \lambda\|\beta\|_1\right\}, \tag{2.9}$$

where $\lambda > 0$ is a new penalty level and we will see it does not depend on $e^{x_i^T\beta^*}$. To get a bound for $\hat{\beta} - \beta^*$, we require that the following relation holds with high probability:

$$\lambda \geq c\|\nabla f(\beta^*)\|_\infty, \tag{2.10}$$

with some constant $c > 1$.

3. Bounds on $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta^*}\|_1$

Recall $T = supp(\beta^*) = \{j \in [p] : \beta_j^* \neq 0\}$, define the set

$$\Delta = \{\delta \in \mathbb{R}^p \backslash \{0\} : \|\delta_{T^c}\|_1 \leq L\|\delta_T\|_1 \text{ with some } L > 1\}. \tag{3.1}$$

For theoretical analysis, we need a few regularity conditions as the following.

Condition (I). *There is some* $R \in (0, \infty)$ *such that* $\sup_{i\in[n], j\in[p]} |x_{ij}| \leq R$.

Condition (II). *There exists some constant* $\kappa > 0$ *such that for any* $\delta \in \Delta$

$$\left\langle \delta, \nabla^2 f(\beta^*)\delta \right\rangle \geq \kappa^2\|\delta_T\|_2^2.$$

Remark 3.1. (i) Condition (I) is a reasonable condition since the data we observed are usually finite and [25] also assumed this condition when studying the penalized generalized linear regressions.

(ii) Observe that

$$\nabla^2 f(\beta^*) = \frac{1}{2n}\sum_{i=1}^{n} x_i x_i^T (y_i e^{-\frac{1}{2}x_i^T\beta^*} + e^{\frac{1}{2}x_i^T\beta^*}) = \frac{1}{2n}X^T D X,$$

where $X = (x_1, \ldots, x_n)^T \in \mathbb{R}^{n\times p}$ and $D = \mathrm{diag}\{y_1 e^{-\frac{1}{2}x_1^T\beta^*} + e^{\frac{1}{2}x_1^T\beta^*}, \ldots, y_n e^{-\frac{1}{2}x_n^T\beta^*} + e^{\frac{1}{2}x_n^T\beta^*}\} \in \mathbb{R}^{n\times n}$. Condition (II) is a type of restricted eigenvalue condition [3, 5], κ can take

$$\kappa = \min_{\delta\in\Delta} \sqrt{\frac{\delta^T X^T D X \delta}{2n\|\delta_T\|_2^2}} > 0. \tag{3.2}$$

Since we have assumed $\frac{1}{n}XX^T = I_n$ in our model, if $\min_{i\in[n]}(y_i e^{-\frac{1}{2}x_i^T\beta^*} + e^{\frac{1}{2}x_i^T\beta^*}) > c_0$ for some $c_0 > 0$, then (3.2) clearly holds true.

Now we give a deterministic bound on the estimation error $\|\hat\beta - \beta^*\|_1$ under the two conditions above.

Theorem 3.2. *Let $\hat\beta$ be the estimator defined by (2.9), let Condition (I) hold and let Condition (II) hold with $L = \frac{c+1}{c-1}$ for some $c > 1$. If λ is chosen to satisfy $\lambda > cH$ with $H = \|\nabla f(\beta^*)\|_\infty$ and $\lambda s \le \frac{2\kappa^2}{3L(1+L)}$ with $s = |T|$, then*

$$\|\hat\beta - \beta^*\|_1 \le \frac{3L(1+L)}{\kappa^2}\lambda s, \tag{3.3}$$

$$|f(\hat\beta) - f(\beta^*)| \le \frac{3L(1+L)}{\kappa^2}\lambda^2 s. \tag{3.4}$$

Remark 3.3. From (3.3), we can see that $\|\hat\beta - \beta^*\|_1 = O(\lambda s)$, having the same order as the error bound of ℓ_1 penalized logistic regression [3]. $|f(\hat\beta) - f(\beta^*)|$ is not a prediction error, but we may take it as an error measured by the function f. For the square root lasso, using a triangle inequality, the corresponding $|f(\hat\beta) - f(\beta^*)|$ is bounded by $\|X^T(\hat\beta - \beta^*)\|_2$.

Remark 3.4. If x_i is assumed to be drawn from a p-dimensional normal distribution, the Condition (I) may be replaced by $\sup_{i\in[n],j\in[p]} |x_{ij}| \le c_0\sqrt{\log p}$ with some positive constant c_0 [18]. Let us roughly explain it as below. For $A > 0$, by Chebyshev inequality,

$$\begin{aligned}\mathbb{P}\left(\sup_{i\in[n],j\in[p]} |x_{ij}| > A\right) &\le e^{-\theta^2A^2}\mathbb{E}\exp\left(\theta^2 \sup_{i\in[n],j\in[p]} |x_{ij}|^2\right)\\ &\le npe^{-\theta^2A^2}\mathbb{E}\exp\left(\theta^2|x_{ij}|^2\right) = O(p^{-k}),\end{aligned}$$

for any $k > 0$, as long as we choose $A = c_0\sqrt{\log p}$ with $c_0 > \frac{\sqrt{k+2}}{\theta}$. The corresponding results on estimation errors like (3.3) and (3.4) can be obtained by the same argument as proving Theorem 3.2 (replacing R with $c_0\sqrt{\log p}$ in (A.3)).

Theorem 3.2 tells us that if we can choose a λ satisfying $\lambda > cH = c\|\nabla f(\beta^*)\|_\infty$ and $\lambda s \le \frac{2\kappa^2}{3L(1+L)}$ with high probability, then conclusions (3.3) and (3.4) hold with high probability. This motivates us to study how to select a good tuning parameter λ.

Recall $H = \|\nabla f(\beta^*)\|_\infty$ where $\nabla f(\beta^*) = -\frac{1}{n}\sum_{i=1}^{n} \frac{x_i(y_i - e^{x_i^T\beta^*})}{\sqrt{e^{x_i^T\beta^*}}}$ is a p-dimensional random vector. Define by $H(1-\alpha|X)$ the $1-\alpha$ quantile of $H|X$ for any $\alpha \in (0,1)$. If we choose λ as follows,

$$\text{exact choice}: \lambda = cH(1-\alpha|X), \tag{3.5}$$

which implies $\mathbb{P}(\lambda \geq cH) \geq 1-\alpha$. To estimate this λ, we further assume

Condition (III). *n and p satisfy that $\sqrt{n} \ll p \leq e^{o(n^{1/5})}$ and $p/\alpha > 8$ for some $\alpha \in (0,1)$.*

The upper bound $p \leq e^{o(n^{1/5})}$ is required for proving by a moderate deviation technique that (2.10) holds with high probability, see more details in Lemma 3.6 below.

The next lemma gives an upper bound for the λ chosen by (3.5), and will be proven in Appendix.

Lemma 3.5. *Let λ be chosen by (3.5), then the following statements hold*

(i) $\mathbb{P}(\lambda \geq cH) \geq 1-\alpha$.

(ii) Under Conditions (I) and (III), we have $\lambda \leq c(\sqrt{n})^{-1}\Phi^{-1}(1-\frac{\alpha}{4p}) < c\sqrt{\frac{2\log(4p/\alpha)}{n}}$ *with* $c > 1$.

In practice, it is often hard to find the exact value of the λ, because the distribution of H can be extremely complicated for large n. To overcome this difficulty, we propose a Gaussian approximation to λ by a moderate deviation technique. Recall $H = \|\nabla f(\beta^*)\|_\infty$ with $\nabla f(\beta^*) = -\frac{1}{n}\sum_{i=1}^{n} \frac{x_i(y_i - e^{x_i^T\beta^*})}{\sqrt{e^{x_i^T\beta^*}}}$, note that $(\frac{y_i - e^{x_i^T\beta^*}}{\sqrt{e^{x_i^T\beta^*}}})_{1\leq i\leq n}$ are i.i.d. random variables with mean 0 and variance 1. Inspired by Lindeberg principle, replacing these random variables by i.i.d. $N(0,1)$-distributed $(z_i)_{1\leq i\leq n}$, we get an $\tilde{H} = \left\|\frac{1}{n}\sum_{i=1}^{n} x_i z_i\right\|_\infty$ and can expect that H weakly converges to $\tilde{H}$ as $n \to \infty$.

Motivated by the above observation, we propose an asymptotic choice of λ as the following:

$$\text{asymptotic choice}: \lambda = c(\sqrt{n})^{-1}\Phi^{-1}(1-\frac{\alpha}{2p}), \tag{3.6}$$

where $\Phi(\cdot)$ is the cumulative distribution function of N(0,1) and $\Phi^{-1}(\cdot)$ is its inverse function, and $c > 1$ is a constant. The following lemma gives the properties of the asymptotic choice of λ, which will be proven in Appendix.

Lemma 3.6. *(i) Suppose that Conditions (I) and (III) are satisfied and λ is*

chosen by (3.6). We have

$$\mathbb{P}(\lambda \geq cH) \geq 1 - \alpha(1 + O(1)(\sqrt{2\log(2p/\alpha)} - \sqrt{n}b)^3 n^{-1/2}(3K_1 \log p + b)) \times (1 + \frac{1}{\log(p/\alpha)}) \frac{\exp\{-2(n\log(p/\alpha))^{1/2}b + nb^2\}}{1 - \sqrt{n}b/(\log(p/\alpha))^{1/2}} + C_1 n/p^2,$$

where $b = 2C_1K_1/p^3$ with some positive constants C_1 and K_1. In particular, as $n, p \to \infty$, we have

$$\mathbb{P}(\lambda \geq cH) \geq 1 - \alpha(1 + o(1)).$$

(ii) For $\alpha \in (0,1)$, we have $\lambda < c\sqrt{\frac{2\log(2p/\alpha)}{n}}$ with $c > 1$.

Combining Theorem 3.2 and Lemmas 3.5 and 3.6, we have the following non-asymptotic results.

Theorem 3.7 (Finite-sample)**.** *Let $\hat{\beta}$ be the estimator defined by (2.9), let Conditions (I) and (III) both hold and let Condition (II) hold with $L = \frac{c+1}{c-1}$ for some $c > 1$.*

(i) If λ is chosen as (3.5) with the above c and the condition $\lambda s \leq \frac{2\kappa^2}{3L(1+L)}$ holds, then with probability at least $1 - \alpha$, we have

$$\|\hat{\beta} - \beta^*\|_1 \leq \frac{3L(1+L)}{\kappa^2}\lambda s, \tag{3.7}$$

$$|f(\hat{\beta}) - f(\beta^*)| \leq \frac{3L(1+L)}{\kappa^2}\lambda^2 s. \tag{3.8}$$

(ii) If λ is chosen as (3.6) with the above c and the condition $\lambda s \leq \frac{2\kappa^2}{3L(1+L)}$ holds, then for large enough n, with probability at least

$$1 - \alpha\left(1 + O(1)\frac{(\log p)^{5/2}}{n^{1/2}}\right), \tag{3.9}$$

the above inequalities (3.7) and (3.8) hold.

Remark 3.8. From Lemmas 3.5 and 3.6, we know that $\lambda = O(\sqrt{\frac{\log p}{n}})$ for both exact and asymptotic choices. Hence, we have $\|\hat{\beta} - \beta^*\|_1 = O(s\sqrt{\frac{\log p}{n}})$ and $|f(\hat{\beta}) - f(\beta^*)| = O(\frac{s\log p}{n})$. If we assume $s\sqrt{\frac{\log p}{n}} \to 0$ as $n, p \to \infty$ instead of $\lambda s \leq \frac{2\kappa^2}{3L(1+L)}$, then the results of (i) and (ii) in Theorem 3.7 can also hold and the probability term (3.9) of (ii) becomes $1 - \alpha(1 + o(1))$. In addition, the condition $s\sqrt{\frac{\log p}{n}} \to 0$ implies that $s = o(\sqrt{\frac{n}{\log p}})$. In general high-dimensional linear regression [5] and ℓ_1 penalized logistic regression [3], they also required the same sparsity condition.

Remark 3.9. If we replace Condition (I) by $\sup_{i\in[n],j\in[p]} |x_{ij}| \le c_0\sqrt{\log p}$ with some positive constant c_0, the results in Theorem 3.7 also hold with high probability for both exact and asymptotic choices, but it requires a stronger condition on s, that is $s = o(\frac{\log p}{\sqrt{n}})$. This proof is the same as ours in Section A only by replacing all R with $c_0 \log p$.

4. Experiments

We use the R package "lbfgs" to solve ℓ_1 penalized convex optimization problems [6]. The lbfgs package implements both the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) and the Orthant-Wise Quasi-Newton Limited-Memory (OWL-QN) optimization algorithms. The L-BFGS algorithm solves the problem of minimizing an objective, given its gradient, by iteratively computing approximations of the inverse Hessian matrix. The OWL-QN algorithm finds the optimum of an objective plus the ℓ_1-norm of the problem's parameters. The package offers a fast and memory-efficient implementation of these optimization routines, which is particularly suited for high-dimensional problems.

4.1. Toy Example

We conduct a simulation to compare the estimation errors by THREE different ways of selecting a tuning parameter for our ℓ_1 *penalized weighted score function* (LPWSF) method and also compare these results with the lasso for Poisson regression [14] and the adaptive lasso for Poisson regression [11]. We use R package "glmnet" to solve the sparse Poisson regression which returns ℓ_1 penalized log-likelihood estimator and R package "grplasso" to obtain the adaptive ℓ_1 penalized log-likelihood estimator. For data setting, we first generate a design matrix $X \in \mathbb{R}^{n\times p}$ with $n = 100$, $p = 1000$ and each element x_{ij} i.i.d. from the standard normal distribution. By a linear transform, we can make X satisfy $\sum_{i=1}^n x_{ij} = 0$, and $\frac{1}{n}\sum_{i=1}^n x_{ij}^2 = 1$, $j = 1, 2, \ldots, p$. We set the number of nonzero elements of β^* as 10 and each element randomly from $N(0,1) \times 0.5$ and $y_i \sim Poisson(\exp\{\sum_{j=1}^{10} x_{ij}\beta_j^*\})$.

Recall $H = \left\| \frac{1}{n} \sum_{i=1}^n \frac{x_i(y_i - e^{x_i^T\beta^*})}{\sqrt{e^{x_i^T\beta^*}}} \right\|_\infty$. Let $\alpha = 0.05$ and consider the following THREE different choices of λ. (1) As defined in (3.5), $\lambda = H(1-\alpha)$. This tuning parameter depends on the true parameter β^*, which is unknown in real applications, but we still list it here as a benchmark. (2) As defined in (3.6), $\lambda = 1.1 \times (\sqrt{n})^{-1}\Phi^{-1}(1 - \frac{\alpha}{2p})$, this is the asymptotic choice of the tuning parameter. (3) We use normal approximation of H defined as $\tilde{H} = \|\frac{1}{n}\sum_{i=1}^n x_i z_i\|_\infty$ with $z_i, i = 1, 2, \ldots, n$ i.i.d. from $N(0,1)$, and define $\lambda = \tilde{H}(1-\alpha)$. This is an approximation of the exact choice of tuning parameter defined in (3.5). For comparison, we also present the result of ℓ_1 penalized log-likelihood with λ selected by cross-

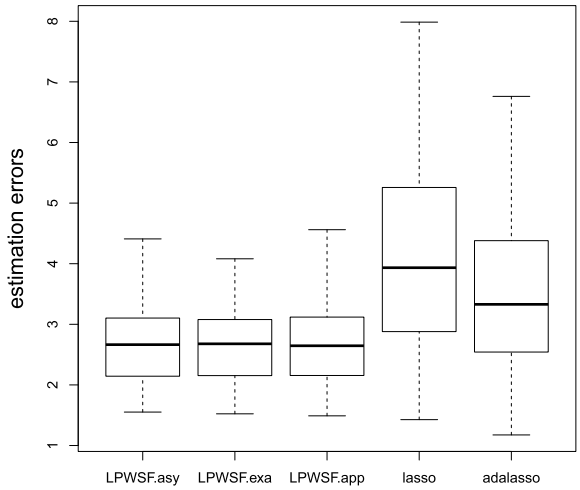

FIG 1. *The errors for different tuning parameters and methods. "LPWSF.asy" is the estimation errors for our proposed method with* $\lambda = 1.1 \times (\sqrt{n})^{-1}\Phi^{-1}(1 - \frac{\alpha}{2p})$. *"LPWSF.exa" is the estimation errors for our proposed method with exact choice of* λ *and "LPWSF.app" is the estimation errors for the new proposed method with an approximate of the exact choice. "lasso" denotes the estimation errors for* ℓ_1 *penalized log-likelihood method with* λ *selected by cross-validation. "adalasso" denotes the estimation errors for adaptive* ℓ_1 *penalized log-likelihood method with a theoretical penalty level* λ *(see equation (2.3) in [11] with* $\gamma = 1.01$*).*

validation and adaptive ℓ_1 penalized log-likelihood with a theoretical penalty level λ (see equation (2.3) in [11] with $\gamma = 1.01$).

We repeat each case 100 times, and compute the corresponding estimation error $\|\hat{\beta} - \beta^*\|_1$ in every time. All the results are reported in Figure 1, from which we see that our proposed method not only has a better performance than the traditional ℓ_1 penalized log-likelihood method and the adaptive ℓ_1 penalized log-likelihood method but also does not need a heavy procedure like cross-validation. Hence, our pre-specified tuning parameter works.

4.2. Image Reconstruction

Now we apply our method to an image reconstruction. The experiments are based on a simulated limited-angle emission tomography dataset. We compare our algorithm with the well-known emission tomography reconstruction algorithms and apply the well known Shepp and Logan "head phantom" shown in Figure 2, which consists of several ellipses with different sizes and densities.

Our goal is to reconstruct this "head phantom" by using limited-angle tomographic projection data, which correspond to parallel beam geometry with 384 detector pixels in a single projection and 180 angular samples spaced uniformly over 180 degrees. The projection data is called sinogram which can be seen in Figure 3. Mathematically, the relationship between sinogram and the

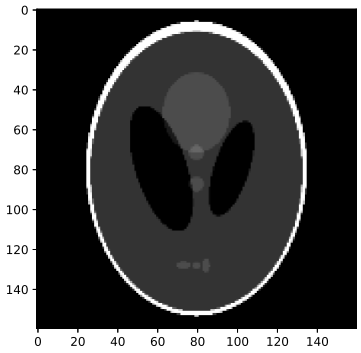

FIG 2. *Head phantom*

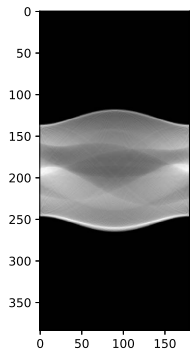

FIG 3. *Sinogram*

original source (the "head phantom") could be described using a group of linear equations

$$y = Ax, \tag{4.1}$$

where x is the source and y is the observed sinogram. A lot of researches have been done to reconstruct the source x through the sinogram y. The difficulty is that this is a singular equation with a very large number of unknown elements in x and only a few observations in y. When noise does not exist, algorithms such as Simultaneous Algebraic Reconstruction Technique (SART) [1], Simultaneous Iterative Reconstruction Technique (SIRT) [9], Conjugate Gradient method for Least Squares (CGLS) [2] and Filtered Backprojection (FBP) [23] were used. The reconstruction can be seen from Figure 4, from which we see that CGLS and FBP work quite well for this data.

It is also well known that limited-angle tomographic projection data consist of Poisson noise [21, 12]. A few models have been studied for noisy data. One of the models for data with noise is via the Poisson regression model. That is, the projection data y follows a Poisson distribution with parameter λ that satisfies

$$\log(\lambda) = Bx.$$

In our simulation, we take $B = A/5.0$, where A is the same as in Equation (4.1) which corresponds to the previous project data without noise. For this noisy project data, the sinogram is shown in Figure 5.

For this projection data, if we still use the linear model as in algorithms such as SART, SIRT, CGLS and FBP, the reconstruction will be very bad, as shown in Figure 6. By comparisons, we use our proposed method (LPWSF) to reconstruct the original source. The reconstruction is shown in Figure 7. It is clear that our proposed method works much better. In the experiments we take the tuning parameter as a theoretical one $\lambda = 1.1(\sqrt{n})^{-1}\Phi^{-1}(1 - \frac{\alpha}{2p})$, with $n = 69120$, $p = 160 \times 160 = 25600$ and $\alpha = 0.05$.

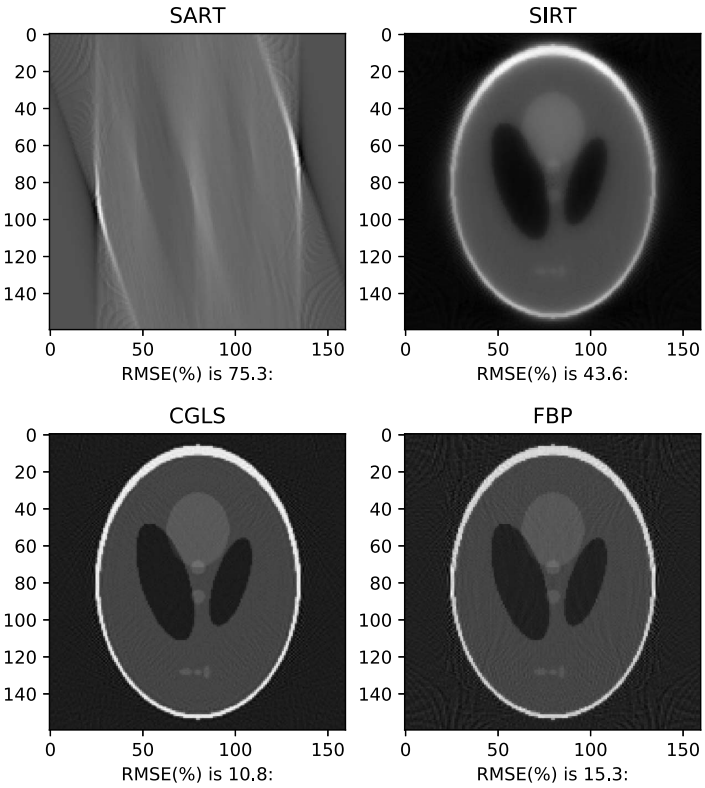

FIG 4. *Reconstruction from sinogram using different algorithms*

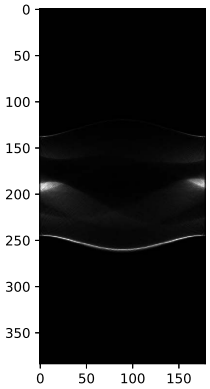

FIG 5. *Sinogram with Poisson noise*

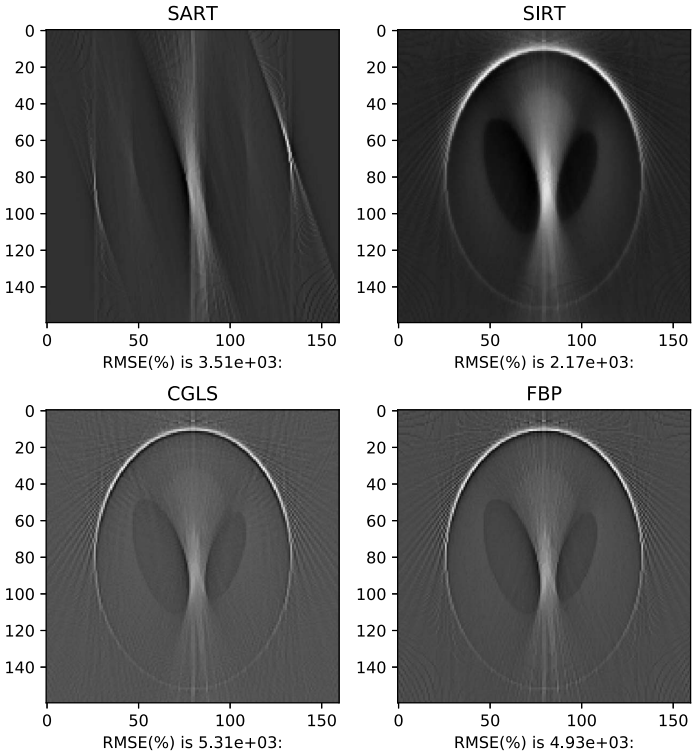

FIG 6. *Reconstruction from noisy sinogram using different algorithms*

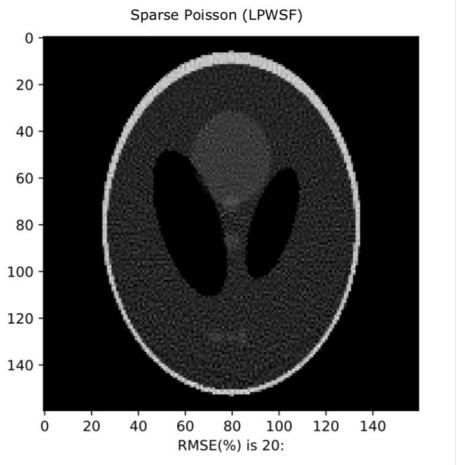

FIG 7. *Reconstruction from noisy sinogram using different algorithms*

5. Conclusion

We proposed an ℓ_1 *penalized weighted score function* method for sparse Poisson regression. After adding a weight on the score function, the penalty coefficient λ does not depend on the variance of Poisson noise anymore. A direct extension is to generalize this idea to other types of generalized linear models, such as negative binomial regression, exponential regression. We prove the estimator of our new method is consistent and give its explicit convergence rate to the true parameter. Simulations and a real application in image reconstruction indicate that our method can be computed efficiently and perform very well.

Appendix A: Appendix

Proof of Theorem 3.2. Let $\delta = \hat{\beta} - \beta^*$. Recall that $T = \{j : \beta_j^* \neq 0\}$. By definition of $\hat{\beta}$, we have

$$\begin{aligned} f(\hat{\beta}) - f(\beta^*) &\leq \lambda(\|\beta^*\|_1 - \|\hat{\beta}\|_1) \\ &= \lambda[(\|\beta_T^*\|_1 - \|\hat{\beta}_T\|_1) + (\|\beta_{T^c}^*\| - \|\hat{\beta}_{T^c}\|_1)] \\ &\leq \lambda(\|\delta_T\|_1 - \|\delta_{T^c}\|_1). \end{aligned} \tag{A.1}$$

Since $f(\beta)$ is a convex function, we have

$$f(\hat{\beta}) - f(\beta^*) \geq \delta^T \nabla f(\beta^*) \geq -\|\nabla f(\beta^*)\|_\infty \|\delta\|_1 \geq -\frac{\lambda}{c}\|\delta\|_1, \tag{A.2}$$

where the last inequality used the choice of λ such that $\lambda > cH = c\|\nabla f(\beta^*)\|_\infty$. Combining (A.1) and (A.2), we obtain that

$$\lambda(\|\delta_T\|_1 - \|\delta_{T^c}\|_1) \geq -\frac{\lambda}{c}(\|\delta_T\|_1 + \|\delta_{T^c}\|_1),$$

i.e.

$$\|\delta_{T^c}\|_1 \leq \frac{c+1}{c-1}\|\delta_T\|_1 = L\|\delta_T\|_1.$$

Defining a new function $\tilde{f}(t) = f(\beta^* + tv)$ from $\mathbb{R}$ to $\mathbb{R}$ for any vector $v \in \mathbb{R}^p$, we compute its second and third order derivatives and denote them as following

$$\begin{aligned} \tilde{f}''(t) &= \frac{\mathrm{d}^2 \tilde{f}(t)}{\mathrm{d}t^2} = \frac{1}{2n}\sum_{i=1}^n (x_i^T v)^2 (y_i e^{-x_i^T(\beta^*+tv)/2} + e^{x_i^T(\beta+tv)/2}), \\ \tilde{f}'''(t) &= \frac{\mathrm{d}^3 \tilde{f}(t)}{\mathrm{d}t^3} = -\frac{1}{4n}\sum_{i=1}^n (x_i^T v)^3 (y_i e^{-x_i^T(\beta^*+tv)/2} - e^{x_i^T(\beta^*+tv)/2}). \end{aligned}$$

By Condition (I), we obtain that

$$|\tilde{f}'''(t)| \leq \frac{1}{2}\sup_{i\in[n]}|x_i^T v|\tilde{f}''(t) \leq \frac{1}{2}\sup_{i\in[n],j\in[p]}|x_{ij}|\|v\|_1 \tilde{f}''(t) \leq \frac{1}{2}R\|v\|_1 \tilde{f}''(t).$$

Setting $v = \delta = \hat{\beta} - \beta^*$, we have

$$|\tilde{f}'''(t)| \leq \frac{1}{2} R\|\delta\|_1 \tilde{f}''(t). \tag{A.3}$$

Denote $\tilde{R} = \frac{1}{2}R$, by Proposition 1 of [3], (A.3) and Condition (II), we have

$$\begin{aligned} f(\hat{\beta}) - f(\beta^*) &\geq \delta^T \nabla f(\beta^*) + \frac{\delta^T \nabla^2 f(\beta^*)\delta}{\tilde{R}^2\|\delta\|_1^2}(e^{-\tilde{R}\|\delta\|_1} + \tilde{R}\|\delta\|_1 - 1) \\ &\geq -\|\nabla f(\beta^*)\|_\infty \|\delta\|_1 + \frac{\delta^T \nabla^2 f(\beta^*)\delta}{\tilde{R}^2\|\delta\|_1^2}(e^{-\tilde{R}\|\delta\|_1} + \tilde{R}\|\delta\|_1 - 1) \\ &\geq -\frac{\lambda}{c}\|\delta\|_1 + \frac{\kappa^2\|\delta_T\|_2^2}{\tilde{R}^2\|\delta\|_1^2}(e^{-\tilde{R}\|\delta\|_1} + \tilde{R}\|\delta\|_1 - 1). \end{aligned} \tag{A.4}$$

Combining (A.1) and (A.4), we have

$$\frac{\kappa^2\|\delta_T\|_2^2}{\tilde{R}^2\|\delta\|_1^2}(e^{-\tilde{R}\|\delta\|_1} + \tilde{R}\|\delta\|_1 - 1) \leq \lambda\|\delta_T\|_1 + \frac{\lambda}{c}\|\delta\|_1 \leq L\lambda\sqrt{s}\|\delta_T\|_2, \tag{A.5}$$

where the last inequality utilizes the relation $\|\delta\|_1 \leq (1+L)\|\delta_T\|_1 \leq (1+L)\sqrt{s}\|\delta_T\|_2$. Using this relation again, we have

$$e^{-\tilde{R}\|\delta\|_1} + \tilde{R}\|\delta\|_1 - 1 \leq \frac{L(1+L)\lambda s\tilde{R}}{\kappa^2}\tilde{R}\|\delta\|_1. \tag{A.6}$$

Setting

$$h = \frac{L(1+L)\lambda s\tilde{R}}{\kappa^2},$$

then (A.6) becomes

$$e^{-\tilde{R}\|\delta\|_1} + \tilde{R}\|\delta\|_1 - 1 \leq h\tilde{R}\|\delta\|_1. \tag{A.7}$$

According to the condition on λ such that $\lambda s \leq \frac{2\kappa^2}{3L(1+L)R}$, we have $h \leq \frac{1}{3}$. Denote $w = \tilde{R}\|\delta\|_1 \geq 0$, then to solve (A.7) is equivalent to solve the inequality $e^{-w} + w - 1 \leq hw$ with $w \geq 0$. By Taylor formula, we have $\frac{w^2}{2} - \frac{w^3}{6} \leq e^{-w} + w - 1 \leq hw$ which implies $\{w \geq 0 : e^{-w} + w - 1 \leq hw, h \leq \frac{1}{3}\} \subseteq \{w \geq 0 : \frac{w^2}{2} - \frac{w^3}{6} \leq hw, h \leq \frac{1}{3}\}$. Since under the condition $h \leq \frac{1}{3}$ and $w \geq 0$, the solution of inequality $\frac{w^2}{2} - \frac{w^3}{6} \leq hw$ is $0 \leq w \leq \frac{3-\sqrt{9-24h}}{2} \leq 1$ or $w \geq \frac{3-\sqrt{9-24h}}{2} \geq 2$, then

$$\begin{aligned} &\{w \geq 0 : e^{-w} + w - 1 \leq hw, h \leq \frac{1}{3}\} \\ &\subseteq \{w : 0 \leq w \leq \frac{3-\sqrt{9-24h}}{2}, h \leq \frac{1}{3}\} \cup \{w : w \geq \frac{3-\sqrt{9-24h}}{2}, h \leq \frac{1}{3}\}. \end{aligned} \tag{A.8}$$

Define $g(w) = e^{-w} + w - 1 - hw$. By the knowledge of derivative and monotonicity, we have

$$\{w \geq 0 : g(w) \leq 0, h \leq \frac{1}{3}\} \subseteq \{w : 0 \leq w < 2\},$$

which and (A.8) imply

$$\begin{aligned}\{w \geq 0 : e^{-w} + w - 1 \leq hw, h \leq \frac{1}{3}\} &\subseteq \{w : 0 \leq w \leq \frac{3 - \sqrt{9 - 24h}}{2}, h \leq \frac{1}{3}\} \\ &\subseteq \{w \geq 0 : w \leq 3h, h \leq \frac{1}{3}\}.\end{aligned}$$

So, from (A.7), we obtain

$$\tilde{R}\|\delta\|_1 \leq \frac{3L(1+L)\lambda s\tilde{R}}{\kappa^2},$$

that is,

$$\|\delta\|_1 \leq \frac{3L(1+L)\lambda s}{\kappa^2}.$$

Furthermore, by (A.1) and (A.2), we obtain

$$|f(\hat{\beta}) - f(\beta^*)| \leq \lambda\|\delta\|_1 \leq \frac{3L(1+L)\lambda^2 s}{\kappa^2}.$$

We finish the proof. □

Proof of Lemma 3.5. (i) By the definition of quantile, it is easy to obtain that

$$\mathbb{P}(cH > \lambda) = \mathbb{P}(cH > cH(1-\alpha|X)) < \alpha.$$

Then $\mathbb{P}(\lambda \geq cH) \geq 1 - \alpha$.

(ii) Let $t_n = (\sqrt{n})^{-1}\Phi^{-1}(1 - \frac{\alpha}{4p})$. If we can prove that $\mathbb{P}(H > t_n) < \alpha$, then by definition of quantile we have $H(1-\alpha|X) \leq t_n$. By the proof of (ii) of Lemma 3.6, we can obtain that $\Phi^{-1}(1 - \frac{\alpha}{4p}) < \sqrt{2\log(4p/\alpha)}$. Hence, we can get the desired result. The rest is to show

$$\mathbb{P}(H > t_n) < \alpha,$$

as $n, p \to \infty$ with $n \leq p \leq e^{o(n^{1/5})}$.

Recall $H = \max_{j\in[p]} |\frac{1}{n}\sum_{i=1}^{n} x_{ij}\epsilon_i|$ and $\epsilon_i = (y_i - e^{x_i^T\beta^*})/\sqrt{e^{x_i^T\beta^*}}$. Denote $t = \Phi^{-1}(1 - \frac{\alpha}{4p})$ and then $t_n = (\sqrt{n})^{-1}t$. Observe that

$$\begin{aligned}\mathbb{P}(H > t_n) &= \mathbb{P}(\max_{j\in[p]} |\frac{1}{n}\sum_{i=1}^{n} x_{ij}\epsilon_i| > (\sqrt{n})^{-1}t) \\ &\leq p \max_{j\in[p]} \mathbb{P}(|\sum_{i=1}^{n} x_{ij}\epsilon_i| > \sqrt{n}t).\end{aligned} \tag{A.9}$$

Repeating the argument below (A.10), we get

$$\mathbb{P}(H > t_n) \le \frac{\alpha}{2}(1 + O(1)(\sqrt{2\log(4p/\alpha)} - \sqrt{n}b)^3 n^{-1/2}(3K_1 \log p + b)) \times (1 + \frac{1}{\log(2p/\alpha)}) \frac{\exp\{-2(n\log(2p/\alpha))^{1/2}b + nb^2\}}{1 - \sqrt{n}b/(\log(2p/\alpha))^{1/2}} + C_1 n/p^2,$$

where as $n, p \to \infty$ with $n \le p \le e^{o(n^{1/5})}$, notice that b, $\sqrt{n}b$ and nb^2 are $o(n^{-2})$, we have

$$\mathbb{P}(H > t_n) \le \frac{\alpha}{2}(1 + o(1)) < \alpha.$$

We finish the proof. □

Proof of Lemma 3.6. (i) By easy calculation, we obtain

$$\nabla f(\beta^*) = -\frac{1}{n}\sum_{i=1}^n x_i(y_i - e^{x_i^T\beta^*})/\sqrt{e^{x_i^T\beta^*}}.$$

Denote $\epsilon_i = (y_i - e^{x_i^T\beta^*})/\sqrt{e^{x_i^T\beta^*}}$, then

$$H = \|\nabla f(\beta^*)\|_\infty = \max_{j\in[p]} |\frac{1}{n}\sum_{i=1}^n x_{ij}\epsilon_i|.$$

Denote $a = \Phi^{-1}(1 - \frac{\alpha}{2p})$, then $\lambda = c(\sqrt{n})^{-1}a$. Hence

$$\begin{aligned} \mathbb{P}(cH > \lambda) &= \mathbb{P}(\max_{j\in[p]} |\frac{1}{n}\sum_{i=1}^n x_{ij}\epsilon_i| > (\sqrt{n})^{-1}a) \\ &\le p\max_{j\in[p]} \mathbb{P}(|\sum_{i=1}^n x_{ij}\epsilon_i| > \sqrt{n}a). \end{aligned} \tag{A.10}$$

Since $y_i|x_i \sim Possion(\mu(x_i))$ with $\mu_i = \mu(x_i) = e^{x_i^T\beta^*}$, then $\mathbb{E}(e^{\theta\epsilon_i}) = \exp\{\mu_i e^{\theta/\sqrt{\mu_i}} - \mu_i - \theta\sqrt{\mu_i}\}$ is a positive constant for any $\theta < \infty$. By the exponential Chebyshev's inequality, we have

$$\mathbb{P}(|\epsilon_i| > M) < e^{-M/K_1}\left[\mathbb{E}(e^{\epsilon_i/K_1}) + \mathbb{E}(e^{-\epsilon_i/K_1})\right] = C_1 e^{-M/K_1} \tag{A.11}$$

with some constant $C_1 = \mathbb{E}(e^{\epsilon_i/K_1}) + \mathbb{E}(e^{-\epsilon_i/K_1}) > 0$ and $K_1 > 0$. Denote $\hat{\epsilon}_i = \epsilon_i 1_{\{|\epsilon_i|\le M\}}$ and $\check{\epsilon}_i = \epsilon_i 1_{\{|\epsilon_i|>M\}}$. Taking $M = 3K_1\log p$, we have

$$\begin{aligned} \mathbb{P}(|\sum_{i=1}^n x_{ij}\epsilon_i| > \sqrt{n}a) &= \mathbb{P}(|\sum_{i=1}^n x_{ij}(\hat{\epsilon}_i + \check{\epsilon}_i)| > \sqrt{n}a, \sup_{i\in[n]}|\epsilon_i| \le M) \\ &\quad + \mathbb{P}(|\sum_{i=1}^n x_{ij}(\hat{\epsilon}_i + \check{\epsilon}_i)| > \sqrt{n}a, \sup_{i\in[n]}|\epsilon_i| > M) \\ &\le \mathbb{P}(|\sum_{i=1}^n x_{ij}\hat{\epsilon}_i| > \sqrt{n}a) + \mathbb{P}(\sup_{i\in[n]}|\epsilon_i| > M). \end{aligned}$$

Denote $P_1 = \mathbb{P}(|\sum_{i=1}^n x_{ij}\hat{\epsilon}_i| > \sqrt{n}a)$ and $P_2 = \mathbb{P}(\sup_{i\in[n]} |\epsilon_i| > M)$, then the above inequality can be written as

$$\mathbb{P}(|\sum_{i=1}^n x_{ij}\epsilon_i| > \sqrt{n}a) \le P_1 + P_2. \tag{A.12}$$

By inequality (A.11) with $M = 3K_1 \log p$, we obtain that

$$P_2 \le \sum_{i=1}^n \mathbb{P}(|\epsilon_i| > M) \le C_1 n e^{-3\log p} = C_1 n/p^3. \tag{A.13}$$

To estimate the P_1, we need the following moderate deviation theorem for standardized sum due to [22](see also [10, 16]), i.e.

Lemma A.1. *Let $\eta_1, \ldots, \eta_n$ be independent random variables with $\mathbb{E}\eta_i = 0$ and $|\eta_i| \le 1$ for all $i \in [n]$. Denote $\sigma_n^2 = \sum_{i=1}^n \mathbb{E}\eta_i^2$ and $L_n = \sum_{i=1}^n \mathbb{E}|\eta_i|^3/\sigma_n^3$. Then there exists a positive constant A such that for all $x \in [1, \frac{1}{A}\min\{\sigma_n, L_n^{-1/3}\}]$*

$$\mathbb{P}(\sum_{i=1}^n \eta_i > x\sigma_n) = (1 + O(1)x^3 L_n)\bar{\Phi}(x),$$

where $\bar{\Phi}(x) = 1 - \Phi(x)$ and $\Phi(x)$ is the cumulative distribution function of standard normal distribution.

Since $\mathbb{E}(\epsilon_i) = \mathbb{E}(\hat{\epsilon}_i) + \mathbb{E}(\check{\epsilon}_i) = 0$, we obtain that

$$\begin{aligned}
|\mathbb{E}(\hat{\epsilon}_i)| &= |\mathbb{E}(\check{\epsilon}_i)| \le \mathbb{E}|\check{\epsilon}_i| = \mathbb{E}(|\hat{\epsilon}_i| 1_{\{|\epsilon_i|>M\}}) \\
&= \int_M^{+\infty} x \, \mathrm{d}\Phi(x) + \int_{-\infty}^{-M} -x \, \mathrm{d}\Phi(x) \\
&= \int_M^{+\infty} \int_0^x 1 \, \mathrm{d}z \, \mathrm{d}\Phi(x) + \int_{-\infty}^{-M} \int_x^0 1 \, \mathrm{d}z \, \mathrm{d}\Phi(x) \\
&= \int_M^{+\infty} \int_z^{+\infty} 1 \, \mathrm{d}\Phi(x) \, \mathrm{d}z + \int_{-\infty}^{-M} \int_{-\infty}^z 1 \, \mathrm{d}\Phi(x) \, \mathrm{d}z \\
&\le \int_M^{+\infty} C_1 e^{-z/K_1} \, \mathrm{d}z + \int_{-\infty}^{-M} C_1 e^{z/K_1} \, \mathrm{d}z \\
&= 2C_1 K_1 e^{-M/K_1} \\
&= 2C_1 K_1/p^3,
\end{aligned} \tag{A.14}$$

where the fifth equality follows Fubini theorem and the last inequality follows (A.11).

Denote $b = 2C_1K_1/p^3$, then $|\mathbb{E}(\hat{\epsilon}_i)| \le b$ and

$$|\sum_{i=1}^n x_{ij}\mathbb{E}\hat{\epsilon}_i| \le \sqrt{(\sum_{i=1}^n x_{ij}^2)(\sum_{i=1}^n |\mathbb{E}\hat{\epsilon}_i|^2)} \le nb.$$

Furthermore, with Condition (I) we have

$$|x_{ij}(\hat{\epsilon}_i - \mathbb{E}\hat{\epsilon}_i)| \le (\sup_{i\in[n],j\in[p]} |x_{ij}|)(|\epsilon_i| + |\mathbb{E}\epsilon_i|) \le R(M+b).$$

Notice that for all $i \in [n]$,

$$\begin{aligned}
\mathbb{E}(\hat{\epsilon}_i - \mathbb{E}\hat{\epsilon}_i)^2 &= \mathbb{E}\hat{\epsilon}_i^2 - (\mathbb{E}\hat{\epsilon}_i)^2 \\
&= \mathbb{E}\epsilon_i^2 - \mathbb{E}\check{\epsilon}_i^2 - (\mathbb{E}\hat{\epsilon}_i)^2 \\
&\le \mathbb{E}\epsilon_i^2 = 1, \\
\mathbb{E}(\hat{\epsilon}_i - \mathbb{E}\hat{\epsilon}_i)^2 &= \mathbb{E}\epsilon_i^2 - \mathbb{E}\check{\epsilon}_i^2 - (\mathbb{E}\hat{\epsilon}_i)^2 \\
&\ge 1 - 4C_1K_1(M+K_1)e^{-M/K_1} - 4C_1^2K_1^2e^{-2M/K_1} \\
&= 1 - 4C_1K_1^2(3\log p + 1)/p^3 - 4C_1^2K_1^2/p^3
\end{aligned} \tag{A.15}$$

where the inequality $(\mathbb{E}\hat{\epsilon}_i)^2 \le 4C_1^2K_1^2e^{-2M/K_1}$ follows (A.14) and the inequality $\mathbb{E}\check{\epsilon}_i^2 \le 4C_1K_1(M+K_1)e^{-M/K_1}$ can be obtained by the same arguments in (A.14). For simplicity, denote $r = 1 - 4C_1K_1^2(3\log p + 1)/p^3 - 4C_1^2K_1^2/p^3$. For large enough p, we have $r \in (0,1)$ and is far away from 0.

Let $\eta_{ij} = x_{ij}(\hat{\epsilon}_i - \mathbb{E}\hat{\epsilon}_i)/R(M+b)$, then we have $\mathbb{E}\eta_{ij} = 0$, $|\eta_{ij}| < 1$. We define σ_{nj}^2 and L_{nj} and calculate them as following

$$\begin{aligned}
\sigma_{nj}^2 &= \sum_{i=1}^n \mathbb{E}\eta_{ij}^2 = \frac{1}{R^2(M+b)^2}\sum_{i=1}^n \mathbb{E}(x_{ij}^2(\hat{\epsilon}_i - \mathbb{E}\hat{\epsilon}_i)^2) \\
&= \frac{1}{R^2(M+b)^2}\sum_{i=1}^n x_{ij}^2\mathbb{E}(\hat{\epsilon}_i - \mathbb{E}\hat{\epsilon}_i)^2, \\
L_{nj} &= \sum_{i=1}^n \mathbb{E}|\eta_{ij}|^3/\sigma_{nj}^3 \le \sum_{i=1}^n \mathbb{E}|\eta_{ij}|^2/\sigma_{nj}^3 = \frac{1}{\sigma_{nj}}.
\end{aligned}$$

By using the bounds of $\mathbb{E}(\hat{\epsilon}_i - \mathbb{E}\hat{\epsilon}_i)^2$ in (A.15) and $\sum_{i=1}^n x_{ij}^2 = n$, we can estimate the upper and lower bounds of σ_{nj}^2 as follows

$$\frac{rn}{R^2(M+b)^2} \le \sigma_{nj}^2 \le \frac{n}{R^2(M+b)^2},$$

where $r \in (0,1)$ and is far away from 0.

Then, $\sigma_{nj}^2 = O(\frac{n}{(M+b)^2})$ and $L_{nj} = O(\frac{(M+b)}{\sqrt{n}})$. Using Lemma A.1, for large enough n, p such that $\sqrt{n} \ll p \le e^{o(n^{1/5})}$ (Condition (III)), we have

$$\begin{aligned}
P_1 &= \mathbb{P}(|\sum_{i=1}^n x_{ij}(\hat{\epsilon}_i - \mathbb{E}\hat{\epsilon}_i + \mathbb{E}\hat{\epsilon}_i)| > \sqrt{n}a) \\
&\le \mathbb{P}(|\sum_{i=1}^n x_{ij}(\hat{\epsilon}_i - \mathbb{E}\hat{\epsilon}_i))| > \sqrt{n}a - |\sum_{i=1}^n x_{ij}\mathbb{E}\hat{\epsilon}_i|)
\end{aligned} \tag{A.16}$$

$$\begin{aligned}
&\le \mathbb{P}(|\sum_{i=1}^{n} \frac{x_{ij}(\hat{\epsilon}_i - \mathbb{E}\hat{\epsilon}_i))}{R(M+b)}| > \frac{\sqrt{n}}{R(M+b)}(a-\sqrt{n}b)) \\
&\le \mathbb{P}(|\sum_{i=1}^{n} \eta_{ij}| > \sigma_{nj}(a-\sqrt{n}b)) \\
&= 2(1+O(1)(a-\sqrt{n}b)^3 L_{nj})\bar{\Phi}(a-\sqrt{n}b)
\end{aligned}$$

with $a-\sqrt{n}b$ uniformly in $[1, O(n^{1/6}(\log p)^{-1/3})]$. Notice that $\log(p/\alpha) < a^2 < 2\log(2p/\alpha)$ when $p/\alpha > 8$ (Condition (III)) and for all $u > 0$ the inequality $\frac{u}{1+u^2}\phi(u) \le \bar{\Phi}(u) \le \frac{\phi(u)}{u}$ holds where $\phi(\cdot)$ is the density function of standard normal distribution. Then,

$$\begin{aligned}
\bar{\Phi}(a-\sqrt{n}b) &\le \frac{\phi(a-\sqrt{n}b)}{a-\sqrt{n}b} = \phi(a)\frac{\exp\{-2a\sqrt{n}b+nb^2\}}{a-\sqrt{n}b} \\
&= \frac{a}{1+a^2}\phi(a)\frac{1+a^2}{a(a-\sqrt{n}b)}\exp\{-2a\sqrt{n}b+nb^2\} \\
&\le \bar{\Phi}(a)\frac{1+a^2}{a(a-\sqrt{n}b)}\exp\{-2a\sqrt{n}b+nb^2\} \qquad \text{(A.17)} \\
&= \frac{\alpha}{2p}(1+\frac{1}{a^2})\frac{1}{1-\sqrt{n}b/a}\exp\{-2a\sqrt{n}b+nb^2\} \\
&\le \frac{\alpha}{2p}(1+\frac{1}{\log(p/\alpha)})\frac{\exp\{-2(n\log(p/\alpha))^{1/2}b+nb^2\}}{1-\sqrt{n}b/(\log(p/\alpha))^{1/2}}
\end{aligned}$$

and

$$(a-\sqrt{n}b)^3 L_{nj} = O(1)(\sqrt{2\log(2p/\alpha)}-\sqrt{n}b)^3 n^{-1/2}(3K_1\log p + b). \qquad \text{(A.18)}$$

Combining (A.16), (A.17) and (A.18), we have

$$\begin{aligned}
P_1 \le \frac{\alpha}{p}&(1+O(1)(\sqrt{2\log(2p/\alpha)}-\sqrt{n}b)^3 n^{-1/2}(3K_1\log p+b)) \\
&\times(1+\frac{1}{\log(p/\alpha)})\frac{1}{1-\sqrt{n}b/(\log(p/\alpha))^{1/2}}\exp\{-2(n\log(p/\alpha))^{1/2}b+nb^2\}.
\end{aligned} \qquad \text{(A.19)}$$

Thus, combining (A.10), (A.13) and (A.19), we obtain that

$$\begin{aligned}
\mathbb{P}(c\|\nabla f(\beta^*)\|_\infty > \lambda) &\le p(P_1+P_2) \\
&\le \alpha(1+O(1)(\sqrt{2\log(2p/\alpha)}-\sqrt{n}b)^3 n^{-1/2}(3K_1\log p+b)) \\
&\quad\times(1+\frac{1}{\log(p/\alpha)})\frac{\exp\{-2(n\log(p/\alpha))^{1/2}b+nb^2\}}{1-\sqrt{n}b/(\log(p/\alpha))^{1/2}} \\
&\quad + C_1 n/p^2.
\end{aligned}$$

As $n, p \to \infty$ with $\sqrt{n} \ll p \le e^{o(n^{1/5})}$, notice that b, $\sqrt{n}b$ and nb^2 are all $o(n^{-2})$, hence, we have

$$\mathbb{P}(c\|\nabla f(\beta^*)\|_\infty > \lambda) \le \alpha(1+o(1)).$$

(ii) Notice the fact that for any $u > 0$, the inequality

$$1 - \Phi(u) \leq \frac{\phi(u)}{u}$$

holds where the $\phi(\cdot)$ is the density function of standard normal distribution. Let $u = \Phi^{-1}(1 - \frac{\alpha}{2p})$. If $p/\alpha > 8$, it is easy to see $u > 3/2$. Then the above inequality becomes

$$\frac{\alpha}{2p} = 1 - \Phi(u) \leq \frac{\phi(u)}{u} = \frac{\exp\{-u^2/2\}}{\sqrt{2\pi}u} < \exp\{-u^2/2\},$$

i.e. $u < \sqrt{2\log(2p/\alpha)}$. Thus $\Phi^{-1}(1 - \frac{\alpha}{2p}) < \sqrt{2\log(2p/\alpha)}$ and

$$\lambda = c(\sqrt{n})^{-1}\Phi^{-1}(1 - \frac{\alpha}{2p}) < c\sqrt{\frac{2\log(2p/\alpha)}{n}}.$$

□

References

[1] A. H. Andersen and A. C. Kak. Simultaneous algebraic reconstruction technique (sart): A superior implementation of the art algorithm. *Ultrasonic Imaging*, 6(1):81–94, 1984.

[2] K. E. Atkinson. *An Introduction to Numerical Analysis*. John Wiley & Sons, 2nd edition, 1989. MR1007135

[3] F. Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4(3):384–414, 2010. MR2645490

[4] A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011. MR2860324

[5] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2008. MR2533469

[6] A. Coppola and B. M. Stewart. lbfgs: Efficient l-bfgs and owl-qn optimization in r. 2014. Software.

[7] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001. MR1946581

[8] J. Fan and J. Lv. Nonconcave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory*, 57(8):5467–5484, 2011. MR2849368

[9] P. Gilbert. Iterative methods for the three-dimensional reconstruction of an object from projections. *Journal of Theoretical Biology*, 36(1):105–117, 1972.

[10] L. Heinrich. Non-uniform estimates, moderate and large deviations in the central limit theorem for m-dependent random variables. *Mathematische Nachrichten*, 121(1):107–121, 1985. MR0809317

[11] S. Ivanoff, F. Picard, and V. Rivoirard. Adaptive lasso and group-lasso for functional Poisson regression. *Journal of Machine Learning Research*, 17(55):1–46, 2016. MR3504615

[12] K. Lange and R. Carson. Em reconstruction algorithms for emission and transmission tomography. *Journal of Computer Assisted Tomography*, 8(2):306–316, 1984.

[13] J. D. Lee, Y. Sun, and J. E. Taylor. On model selection consistency of regularized m-estimators. *Electronic Journal of Statistics*, 9(1):608–642, 2014. MR3331852

[14] Y. H. Li and V. Cevher. Consistency of ℓ_1-regularized maximum-likelihood for compressive Poisson regression. *International Conference on Acoustics, Speech, and Signal Processing*, 2015.

[15] Y.-H. Li, J. Scarlett, P. Ravikumar, and V. Cevher. Sparsistency of ℓ_1-regularized m-estimators. In *AISTATS*, 2015.

[16] W. Liu, Q. M. Shao, and Q. Wang. Self-normalized cramér type moderate deviations for the maximum of sums. *Bernoulli*, 19(3):1006–1027, 2013. MR3079304

[17] J. A. Nelder and R. W. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972. MR0375592

[18] D. Pollard. Empirical processes: Theory and applications. *NSF-CBMS Regional Conference Series in Probability and Statistics*, 2:i–86, 1990. MR1089429

[19] M. Raginsky, R. Willett, Z. T. Harmany, and R. F. Marcia. Compressed sensing performance bounds under Poisson noise. *IEEE Transactions on Signal Processing*, 58(8):3990–4002, 2010. MR2780163

[20] P. Ravikumar, M. J. Wainwright, and J. Lafferty. High-dimensional ising model selection using ℓ_1-regularized logistic regression. *Annals of Statistics*, 38(3):1287–1319, 2010. MR2662343

[21] A. J. Rockmore and A. Macovski. A maximum likelihood approach to transmission image reconstruction from projections. *IEEE Transactions on Nuclear Science*, 24(3):1929–1935, 2007.

[22] A. I. Sakhanenko. Berry-esseen type estimates for large deviation probabilities. *Siberian Mathematical Journal*, 32(4):647–656, 1991. MR1142075

[23] L. A. Shepp and B. F. Logan. The Fourier reconstruction of a head section. *IEEE Transactions on Nuclear Science*, 21(3):21–43, 1974.

[24] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. MR1379242

[25] S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42(3):1166–1202, 2014. MR3224285

[26] C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010. MR2604701

[27] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006. MR2279469